Variational and Bayesian Methods for Solving Hamilton–Jacobi Equations in Machine Learning and Imaging Science

by

Gabriel Provencher Langlois

B.Sc., McGill University; Montreal, QC, Canada, 2013

M.Sc., ETH Zürich; Zürich, Switzerland, 2015

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2022

This dissertation by Gabriel Provencher Langlois is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Jérôme Darbon, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Stuart Geman, Ph.D., Reader

Date_____ _____

Paul Dupuis, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Andrew G. Campbell, Dean of the Graduate School

Curriculum Vitae

Gabriel Provencher Langlois graduated from McGill University where he received a B. Sc. in Honours Applied Mathematics and Honours Physics, both with First Class Honours. He also graduated from ETH Zürich where he received a M. Sc. in Applied Mathematics.

Publications

Referred Journal Articles

Jérôme Darbon and **Gabriel P. Langlois**. On Bayesian posterior mean estimators in imaging sciences and Hamilton–Jacobi partial differential equations. *Journal of Mathematical Imaging and Vision* 63: 821-854 (2021). DOI: 10.1007/s10851-021-01036-0.

Jérôme Darbon, **Gabriel P. Langlois**, and Tingwei Meng. Overcoming the curse of dimensionality for some Hamilton–Jacobi partial differential equations via neural network architectures. *Research in the Mathematical Sciences* 7, no. 3: 1-50 (2020). DOI: 10.1007/s40687-020-00215-6.

Tiemo Pedergnana, David Oettinger, **Gabriel P. Langlois**, and George Haller. Explicit unsteady Navier—Stokes solutions and their analysis via local vortex criteria. *Physics of Fluids* 32, no 4: 046603 (2020). DOI: 10.1007/s40687-020-00215-6.

Gabriel P. Langlois, Donald M. Arnold, Jayson Potts, Brian Leber, David C. Dale, and Michael C. Mackey. Cyclic thrombocytopenia with statistically significant neutrophil oscillations. *Clinical Case Reports* 6: 1347-1352 (2018). DOI: 10.1002/ccr3.1611.

Gabriel P. Langlois, Morgan Craig, Antony Humphries, Michael C. Mackey, Joseph M.

Mahaffy, Jacques Bélair, Thibault Moulin, Sean R. Sinclair, and Liangliang Wang. Normal and pathological dynamics of platelets in humans. *Journal of Mathematical Biology* 75: 1-52 (2017). DOI: 10.1007/s00285-017-1125-6.

Gabriel P. Langlois, Mohammad Farazmand, and George Haller. Asymptotic dynamics of inertial particles with memory. *Journal of Nonlinear Science* 25, no. 6: 1225-1255 (2015). DOI: 10.1007/s00332-015-9250-0.

Grace Brooks, **Gabriel P. Langlois**, Jinzhi Lee, and Michael C. Mackey. Neutrophil dynamics after chemotherapy and G-CSF: The role of pharmacokinetics in shaping the response. *Journal of Theoretical Biology* 315: 97-109 (2012). DOI: 10.1016/j.jtbi.2012.08.028.

Book Chapters

Jérôme Darbon, **Gabriel P. Langlois**, and Tingwei Meng. Connecting Hamilton-Jacobi Partial Differential Equations with Maximum a Posteriori and Posterior Mean Estimators for Some Non-convex Priors. *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision* 1-25 (2021). DOI: 10.1007/978-3-030-03009-4_56-1.

Submitted manuscripts or manuscripts in preparation

Jérôme Darbon and **Gabriel P. Langlois**. Accelerated nonlinear primal-dual hybrid gradient algorithms with applications to machine learning. Submitted to the *Journal of Machine Learning Research*.

Jérôme Darbon and **Gabriel P. Langlois**. Efficient and robust high-dimensional sparse logistic regression via nonlinear primal-dual hybrid gradient algorithms. A preprint is available on arXiv (https://arxiv.org/abs/2111.15426).

Jérôme Darbon and **Gabriel P. Langlois**. Efficient and robust high-dimensional maximum entropy estimation via nonlinear primal-dual hybrid gradient algorithms. In preparation.

Preface and Acknowledgments

I would like to thank my advisor Jérôme Darbon for his guidance and support throughout my doctoral studies. His insight and frequent, frank feedback have significantly helped me develop as a mathematician and scientist.

I would like to thank my readers Paul Dupuis and Stuart Geman for agreeing to serve as readers on my thesis committee. I appreciate their support, advice and insights in our interactions.

I would like to thank my research colleagues Tingwei Meng, Paula Chen, and Taewoo Kim with whom I had the pleasure to work and collaborate it.

I would like to thank the staff at the Division of Applied Mathematics for their infallible organizational help and support.

I would like to give an heartfelt thank you to all my friends and mentors that supported me during this journey. I have fond memories of you and I will treasure these for the rest of my life.

I would like to thank my parents, Richard and Elisabeth, my siblings, Marion and Dominic, and my extended family for their love and support.

Finally, I give my most deepest thanks to my beloved and inspirational partner Ritu for her love, support, and encouragement. I doubt I would have made it this far without you.

Abstract of "Variational and Bayesian Methods for Solving Hamilton–Jacobi Equations in Machine Learning and Imaging Science", by Gabriel Provencher Langlois, Ph.D., Brown University, May 2022.

The growing computational demands of data science applications pose a significant challenge to machine learning and the applied sciences. These applications have relied mainly on increases in computing power to improve performance, but the computational power required to manage growing data sets and continue progress is soon expected to become economically and environmentally unsustainable. The design of efficient algorithms from problem domains (e.g., machine learning, imaging science, and optimal control) that take advantage of emerging hardware (e.g., field-programmable gate arrays architectures) has accordingly been identified as crucial to meet this challenge. Many traditional algorithms, however, were not developed to handle big data sets efficiently in this way. In this dissertation, I contribute innovative variational and Bayesian methods for large-scale machine learning and imaging science to try and meet this challenge. The focus is mathematical and supplemented with numerical examples. Chapter 2 of this dissertation introduces novel accelerated nonlinear primal-dual hybrid gradient methods tailored for efficiently solving a broad class of convex-concave saddle-point problems. I prove rigorous convergence results, including results for strongly convex or smooth problems posed on infinite-dimensional reflexive Banach spaces. Moreover, I establish novel connections between supervised learning tasks in machine learning and a broad class of first-order Hamilton–Jacobi partial differential equations with initial data. Chapter 3 of this dissertation applies the optimization methods developed in Chapter 2 to sparse logistic regression, regularized maximum entropy estimation, and entropy-regularized matrix games. I discuss each problem in detail, and I propose an explicit accelerated nonlinear primal-dual hybrid gradient method to solve each problem efficiently. I also present some numerical experiments to illustrate that my novel accelerated nonlinear primal-dual hybrid gradient methods are considerably faster than competing optimization methods. Finally, Chapter 4 of this dissertation presents new theoretical connections between a broad class of Bayesian posterior mean estimators for imaging science and viscous Hamilton–Jacobi partial differential equations with initial data. I use these connections to establish novel representation formulas and various properties of Bayesian posterior estimators.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

Introduction

1.1 Overview

The growing computational demands of data science applications pose a significant challenge to machine learning and the applied sciences. These applications have relied mainly on increases in computing power to improve performance, but the computational power required to manage growing data sets and continue progress is soon expected to become economically and environmentally unsustainable. The design of efficient algorithms from problem domains (e.g., machine learning, imaging science, and optimal control) that take advantage of emerging hardware (e.g., field-programmable gate arrays architectures) has accordingly been identified as crucial to meet this challenge. Many traditional algorithms, however, were not developed to handle big data sets efficiently in this way.

This dissertation aims to try and meet this challenge and focuses on variational and Bayesian methods for solving large-scale problems in machine learning and imaging science as well as on connections between these methods to a broad class of Hamilton–Jacobi partial differential equations with initial data. Specifically, its goals are to develop innovative optimization methods for creating efficient machine learning algorithms to leverage emerging hardware for big data applications, to create new insights and novel methods for high-dimensional Bayesian estimation in imaging science, and to establish novel connections between Hamilton–Jacobi partial differential equations to be leveraged in future applications and research, e.g., in statistics, optimal control and scientific computing.

The next three chapters collects the results that I published, submitted for publication, or are in preprint form. Each chapter includes an introduction that motivates the problem(s) at hand, the formulation of said problem(s), the results and a discussion. The material in chapter 3 applies the optimization methods developed in chapter 2 to several supervised machine learning problems, but it can be read separately from chapter 2 if one accepts on faith that the optimization methods work as intended. The key mathematical concepts, symbols and notation are introduced in Section 1.2 of this chapter. The notation may vary a little across chapters, but I tried to define these terms explicitly whenever they appeared and be as consistent as possible to avoid any confusion.

Chapter 2 presents my mathematical work on nonlinear primal-dual hybrid gradient optimization methods. Motivated by machine learning and statistics problems, I introduce accelerated nonlinear PDHG optimization methods that use Bregman proximal operators to be highly flexible and efficient in a way that is not currently possible in the optimization literature. The introduction presents an example with sparse logistic regression to illustrate this point. Sparse logistic regression, it turns out, admits a formulation in terms of an Hamilton–Jacobi partial differential equation. The introduction briefly and formally explains this. In addition, I discuss in Section 2.5 of the chapter how several supervised machine learning algorithms correspond to solutions to Hamilton–Jacobi partial differential equations. To the best of my knowledge, these results are novel and this dissertation presents connections between logistic regression (and other supervised machine learning problems described in chapter 2 and 3) and Hamilton–Jacobi partial differential equations for the first time in the mathematical and scientific literature. In the rest of chapter 2, I prove rigorous convergence results for accelerated nonlinear primal-dual hybrid gradient methods, including results for strongly convex or smooth problems posed on infinite-dimensional reflexive Banach spaces.

Chapter 3 applies the accelerated nonlinear primal-dual hybrid gradient optimization methods described in chapter 2 to several supervised machine learning algorithms. It presents detailed treatments of sparse logistic regression and regularized maximum entropy estimation problems, including novel connections between these problems and Hamilton–Jacobi partial differential equations. I also discuss other applications to regression and classification problems defined on the unit simplex. Finally, this chapter presents numerical experiments to illustrate that the accelerated nonlinear primal-dual hybrid gradient methods I introduce are considerably faster than competing methods.

Chapter 4 presents my work on Bayesian methods in imaging science and Hamilton–Jacobi partial differential equations. This work establishes that solutions to some viscous Hamilton–Jacobi partial differential equations with initial data describe extensively a broad class of posterior mean estimators with quadratic fidelity term and log-concave prior. It also uses these connections to establish representation formulas and various properties of posterior mean estimators, and it describes the practical consequences for posterior mean estimators used in imaging science. Notably,

I use these connections to prove that some posterior mean estimators can be expressed as proximal mappings of smooth functions and derive representation formulas for these functions. This result, in particular, shows that posterior mean estimators correspond to maximum a posterior estimators (or modes) of appropriately smooth posterior distributions. It also explains why the posterior mean estimator in imaging science avoids image denoising staircasing effects. Finally, I also present some extensions of these results to a class of posterior mean estimators whose priors are sums of log-concave priors, that is, to posterior mean estimators of mixture distributions.

The last chapter of this dissertation summarizes the intellectual merit of this work, outlines future work and directions of the work presented in this dissertation, and describes the broader impact of this work in the context of the broader optimization, machine learning, optimal control, scientific computing, and imaging science communities.

The publications (refereed journal articles, book chapters, or submitted to a journal or in preparation) on which each chapter is based are detailed below. I am the primary contributor to all the works below except for the work based on the book chapter [72] presented in Section 4.4, which is joint work with Tingwei Meng. Although not presented in this dissertation, I also co-authored a paper [71] with Jérôme Darbon and Tingwei Meng on overcoming the curse of dimensionality for some Hamilton–Jacobi partial differential equations via neural network architectures.

Chapters 2 and 3:

- *Accelerated nonlinear primal-dual hybrid gradient methods with applications to supervised machine learning*, Darbon, Jérôme and Langlois, P. Gabriel. Submitted to the *Journal of Machine Learning Research* [67]
- *Efficient and robust high-dimensional sparse logistic regression via nonlinear primal-dual hybrid gradient algorithms*, Darbon, Jérôme and Langlois P. Gabriel. Preprint on arXiv [65]
- *Efficient and robust nonlinear high-dimensional maximum entropy estimation via nonlinear primal-dual hybrid gradient algorithms*, Darbon, Jérôme and Langlois P. Gabriel. In preparation.

Chapter 4:

- *On Bayesian posterior mean estimators in imaging sciences and Hamilton–Jacobi partial differential equations*, Darbon, Jérôme and Langlois, P. Gabriel, Published in *Journal of Mathematical Imaging and Vision* [66]
- *Connecting Hamilton–Jacobi partial differential equations with maximum a posteriori and posterior mean estimators for some non-convex priors*, Darbon, Jérôme, Langlois, P. Gabriel, and Tingwei Meng. Book chapter [72]

1.2 Background

This section introduces the mathematical concepts from real, convex and functional analysis that are used throughout the dissertation. For comprehensive references, see [33, 92, 108, 136, 137, 214, 215]. The key symbols and notation are summarized in Table 1.1.

In all definitions and facts below, the spaces $\mathcal{X}$ and $\mathcal{Y}$ denote two real reflexive Banach spaces endowed with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$. The interior of a non-empty subset C of $\mathcal{X}$ or $\mathcal{Y}$ is denoted by $\operatorname{int} C$. The set of proper, convex and lower semicontinuous functions defined on $\mathcal{X}$ and $\mathcal{Y}$ are denoted by $\Gamma_0(\mathcal{X})$ and $\Gamma_0(\mathcal{Y})$. The dual spaces of all continuous linear functionals defined on $\mathcal{X}$ and $\mathcal{Y}$ are denoted by $\mathcal{X}^*$ and $\mathcal{Y}^*$. For a linear functional $\boldsymbol{x}^* \in \mathcal{X}^*$ and an element $\boldsymbol{x} \in \mathcal{X}$, the bilinear form $\langle \boldsymbol{x}^*, \boldsymbol{x} \rangle$ gives the value of $\boldsymbol{x}^*$ at $\boldsymbol{x}$. Likewise, for a linear functional $\boldsymbol{y}^* \in \mathcal{Y}^*$ and an element $\boldsymbol{y} \in \mathcal{Y}$, the bilinear form $\langle \boldsymbol{y}^*, \boldsymbol{y} \rangle$ gives the value of $\boldsymbol{y}^*$ at $\boldsymbol{y}$. The norms associated to $\mathcal{X}^*$ and $\mathcal{Y}^*$ are defined as

$$\|\boldsymbol{x}^*\|_{\mathcal{X}^*} = \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}=1} \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle \quad \text{and} \quad \|\boldsymbol{y}^*\|_{\mathcal{Y}^*} = \sup_{\|\boldsymbol{y}\|_{\mathcal{Y}}=1} \langle \boldsymbol{y}^*, \boldsymbol{y} \rangle .$$

Let $\boldsymbol{A} \colon \mathcal{X} \to \mathcal{Y}$ denote a bounded linear operator. Its corresponding adjoint operator $\boldsymbol{A}^* \colon \mathcal{Y}^* \to \mathcal{X}^*$ is defined so as to satisfy

$$\langle \boldsymbol{A}^* \boldsymbol{y}^*, \boldsymbol{x} \rangle = \langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x} \rangle$$

for every $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y^*} \in \mathcal{Y}^*$. The operator norm associated to $\boldsymbol{A}$ is defined as

$$\|\boldsymbol{A}\|_{\mathrm{op}} = \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}=1} \|\boldsymbol{A}\boldsymbol{x}\|_{\mathcal{Y}} = \|\boldsymbol{A}^*\|_{\mathrm{op}} = \sup_{\|\boldsymbol{y}^*\|_{\mathcal{Y}^*}=1} \|\boldsymbol{A}^*\boldsymbol{y}^*\|_{\mathcal{X}^*} .$$

These definitions imply the Cauchy–Schwartz inequality

$$|\langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x}\rangle| \leqslant \|\boldsymbol{A}\|_{\mathrm{op}} \, \|\boldsymbol{x}\|_{\mathcal{X}} \, \|\boldsymbol{y}^*\|_{\mathcal{Y}^*} .$$

Table 1.1: List of key symbols and notation used throughout the dissertation

Notation	Meaning
$\mathcal{X}$	Real reflexive Banach space endowed with norm $\|\cdot\|_{\mathcal{X}}$
$\mathcal{X}^*$	Dual space of all continuous linear functionals defined on $\mathcal{X}$
$\boldsymbol{x} \mapsto \langle \boldsymbol{x}^*, \boldsymbol{x}\rangle$	Value of the functional $\boldsymbol{x}^*$ at $\boldsymbol{x}$
$\|\cdot\|_{\mathcal{X}^*}$	Norm over the dual space $\mathcal{X}^*$: $\|\boldsymbol{x}^*\|_{\mathcal{X}^*} = \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}=1} \langle \boldsymbol{x}^*, \boldsymbol{x}\rangle$
$\boldsymbol{A}\colon \mathcal{X} \to \mathcal{Y}$	Bounded linear operator between two reflexive Banach spaces $\mathcal{X}$ and $\mathcal{Y}$
$\boldsymbol{A}^*\colon \mathcal{Y}^* \to \mathcal{X}^*$	Adjoint operator of $\boldsymbol{A}$
$\|\boldsymbol{A}\|_{\mathrm{op}}$	Operator norm of $\boldsymbol{A}$: $\|\boldsymbol{A}\|_{\mathrm{op}} = \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}=1} \|\boldsymbol{A}\boldsymbol{x}\|_{\mathcal{Y}} = \sup_{\|\boldsymbol{y}^*\|_{\mathcal{Y}^*}=1} \|\boldsymbol{A}^*\boldsymbol{y}^*\|_{\mathcal{X}^*}$
$\|\boldsymbol{A}\|_{2,2}$	Largest singular value of an $m \times n$ real matrix $\boldsymbol{A}$
$\|\boldsymbol{A}\|_{1,2}$	Maximum ℓ_2 norm of a column of an $m \times n$ real matrix $\boldsymbol{A}$
$\|\boldsymbol{A}\|_{1,\infty}$	Maximum ℓ_∞ norm of a column of an $m \times n$ real matrix $\boldsymbol{A}$
$(\boldsymbol{A} \mid \boldsymbol{B})$	Horizontal concatenation of two $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$
int C	Interior of a non-empty subset C
ri C	Interior of a non-empty subset C relative to the affine hull of C
cl C	Closure of a non-empty subset C
bd C	Boundary of a non-empty subset C: bd C = cl $C \setminus$ int C
dom g	Domain of a function g
$\Gamma_0(\mathcal{X})$	Set of proper, convex and lower semicontinuous functions defined on $\mathcal{X}$
$\partial g(\boldsymbol{x})$	Subdifferential of a function $g \in \Gamma_0(\mathcal{X})$ at $\boldsymbol{x} \in \mathcal{X}$
dom ∂f	The set of points $\boldsymbol{x} \in$ dom f for which the subdifferential $\partial f(\boldsymbol{x})$ is non-empty
g^*	Convex conjugate of a function g
$\boldsymbol{I}_{n\times n}$	$n \times n$ identity matrix
Δ_n	Unit simplex over $\mathbb{R}^n$: $\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n : \sum_{j=1}^n [\boldsymbol{x}]_j = 1\}$
$\mathcal{H}_n(\boldsymbol{x})$	Negative entropy of $\boldsymbol{x} \in \Delta_n$: $\mathcal{H}_n(\boldsymbol{x}) = \sum_{j=1}^n [\boldsymbol{x}]_j \log([\boldsymbol{x}]_j)$
$\pi_C(\boldsymbol{x})$	Projection of $\boldsymbol{x} \in \mathcal{X}$ onto a closed convex set C: $\pi_C(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}\in C} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2$
$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, t)$	Gradient vector with respect to $\boldsymbol{x}$ of the function $(\boldsymbol{x}, t) \mapsto f(\boldsymbol{x}, t)$
$\nabla_{\boldsymbol{x}} \cdot f(\boldsymbol{x}, t)$	Divergence with respect to $\boldsymbol{x}$ of the function $(\boldsymbol{x}, t) \mapsto f(\boldsymbol{x}, t)$
$\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}, t)$	Laplacian with respect to $\boldsymbol{x}$ of the function $(\boldsymbol{x}, t) \mapsto f(\boldsymbol{x}, t)$

1.2.1 Definitions

Definition 1 (Convex sets)**.** *A subset $C \subset \mathcal{X}$ is convex if for every pair $(\boldsymbol{x}, \boldsymbol{x}') \in C \times C$ and every scalar $\lambda \in (0,1)$, the point $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{x}'$ is contained in C.*

Definition 2 (Proper functions)**.** *A function f defined on $\mathcal{X}$ is proper if its domain*

$$\text{dom } f = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) < +\infty\}$$

is non-empty and $f(\boldsymbol{x}) > -\infty$ for every $\boldsymbol{x} \in \text{dom } f$.

Definition 3 (Lower semicontinuous functions)**.** *A proper function $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous at a point $\boldsymbol{x} \in \mathcal{X}$ if for every sequence $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ in $\mathcal{X}$ that converges to $\boldsymbol{x}$,*

$$\liminf_{k \to +\infty} f(\boldsymbol{x}_k) \geqslant f(\boldsymbol{x}).$$

We say that f is lower semicontinuous if it is lower semicontinuous at every $\boldsymbol{x} \in \text{dom } f$.

Definition 4 (Convex functions)**.** *A proper function $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is convex if its domain* dom f *is convex and if for every pair $(\boldsymbol{x}, \boldsymbol{x}') \in \text{dom } f \times \text{dom } f$ and every scalar $\lambda \in [0,1]$,*

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{x}') \leqslant \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{x}').$$

It is strictly convex if the inequality above is strict whenever $\boldsymbol{x} \neq \boldsymbol{x}'$ and $\lambda \in (0,1)$, and it is α-strongly convex (with $\alpha > 0$) if for every pair $(\boldsymbol{x}, \boldsymbol{x}') \in \text{dom } f \times \text{dom } f$ and every scalar $\lambda \in [0,1]$.

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{x}') \leqslant \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{x}') - \frac{\alpha}{2}\lambda(1-\lambda)\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}}^2.$$

Definition 5 (Coercive functions)**.** *A proper function $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is coercive if for every sequence $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ in $\mathcal{X}$ such that $\lim_{k \to +\infty} \|\boldsymbol{x}_k\|_{\mathcal{X}} = +\infty$,*

$$\lim_{k \to +\infty} f(\boldsymbol{x}_k) = +\infty.$$

A proper function $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is supercoercive if for every sequence $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ in $\mathcal{X}$ such that

$\lim_{k\to+\infty} \|\boldsymbol{x}_k\|_{\mathcal{X}} = +\infty$,

$$\lim_{k\to+\infty} \frac{f(\boldsymbol{x}_k)}{\|\boldsymbol{x}_k\|_{\mathcal{X}}} = +\infty.$$

Definition 6 (Weak convergence)**.** *A sequence $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ of points in $\mathcal{X}$ converges weakly to $\boldsymbol{x} \in \mathcal{X}$ if for every linear functional $\boldsymbol{x}^* \in \Gamma_0(\mathcal{X})$,*

$$\lim_{k\to+\infty} \langle \boldsymbol{x}^*, \boldsymbol{x}_k \rangle = \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle .$$

Definition 7 (Differentiability)**.** *A proper function $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ with $\operatorname{int}(\operatorname{dom} f) \neq \varnothing$ is differentiable at a point $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom} f)$ if there exists a linear functional $\boldsymbol{x}^* \in \mathcal{X}^*$ such that for every $\boldsymbol{x}' \in \mathcal{X}$,*

$$\lim_{\substack{\lambda\to 0 \\ \lambda>0}} \frac{f(\boldsymbol{x}+\lambda\boldsymbol{x}') - f(\boldsymbol{x})}{\lambda} = \left\langle \boldsymbol{x}^*, \boldsymbol{x}' \right\rangle .$$

This linear functional, when it exists, is unique. It is called the gradient of f at $\boldsymbol{x}$ and is denoted by $\nabla f(\boldsymbol{x})$.

Definition 8 (Subdifferentiability and subgradients)**.** *A function $f \in \Gamma_0(\mathcal{X})$ is subdifferentiable at a point $\boldsymbol{x} \in \mathcal{X}$ if there exists a linear functional $\boldsymbol{x}^* \in \mathcal{X}^*$ such that for every $\boldsymbol{x}' \in \operatorname{dom} f$,*

$$f(\boldsymbol{x}') - f(\boldsymbol{x}) - \left\langle \boldsymbol{x}^*, \boldsymbol{x}' - \boldsymbol{x} \right\rangle \geqslant 0. \tag{1.1}$$

In this case, $\boldsymbol{x}^*$ is called a subgradient of the function f at $\boldsymbol{x}$. The set of subgradients at $\boldsymbol{x} \in \mathcal{X}$ is called the subdifferential of f at $\boldsymbol{x}$, and it is denoted by $\partial f(\boldsymbol{x})$. The subdifferential $\partial f(\boldsymbol{x})$ is a closed convex subset of $\mathcal{X}$ whenever it is non-empty, and f has a unique subgradient at $\boldsymbol{x}$ if and only if f is differentiable at $\boldsymbol{x}$ [214, Theorem 25.1].

The set of points $\boldsymbol{x} \in \operatorname{dom} f$ at which the subdifferential $\partial f(\boldsymbol{x})$ is non-empty is denoted by $\operatorname{dom} \partial f$.

If f is strictly convex, then for $\boldsymbol{x} \neq \boldsymbol{x}'$ the inequality in (1.1) is strict. If f is m-strongly convex and $\boldsymbol{x} \in \operatorname{dom} \partial f$, then for every $\boldsymbol{x}' \in \operatorname{dom} f$ the subgradients $\boldsymbol{x}^* \in \partial f(\boldsymbol{x})$ satisfy the inequality

$$f(\boldsymbol{x}') - f(\boldsymbol{x}) - \left\langle \boldsymbol{x}^*, \boldsymbol{x}' - \boldsymbol{x} \right\rangle \geqslant \frac{m}{2} \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}}^2 . \tag{1.2}$$

Definition 9 (Convex conjugates)**.** *Let $f \in \Gamma_0(\mathcal{X})$. The convex conjugate $f^*\colon \mathcal{X}^* \to \mathbb{R} \cup \{+\infty\}$ of f is defined by*

$$f^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x} \in \operatorname{dom} f} \{\langle \boldsymbol{x}^*, \boldsymbol{x} \rangle - f(\boldsymbol{x})\}. \tag{1.3}$$

By definition, the function f^* is in $\Gamma_0(\mathcal{X}^*)$ [92, Page 17, Definition 4.1]. The supremum in (1.3) is attained if and only if there exists $\boldsymbol{x} \in \operatorname{dom} \partial f$ such that $\boldsymbol{x}^* \in \partial f(\boldsymbol{x})$.

Definition 10 (Essential smoothness)**.** *A function $f \in \Gamma_0(\mathcal{X})$ is essentially smooth if* $\operatorname{dom} \partial f \neq \varnothing$, $\operatorname{dom} \partial f = \operatorname{int}(\operatorname{dom} f)$, *$f$ is differentiable on* $\operatorname{int}(\operatorname{dom} f)$, *and $\|\nabla f(\boldsymbol{x}_k)\|_{\mathcal{X}} \to +\infty$ for every sequence $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ in* $\operatorname{int}(\operatorname{dom} f)$ *converging to some boundary point of* $\operatorname{dom} f$.

Definition 11 (Essential strict convexity)**.** *A function $f \in \Gamma_0(\mathcal{X})$ is essentially strictly convex if f is strictly convex on every convex subset of* $\operatorname{dom} \partial f$ *and the subdifferential mapping ∂f^* is locally bounded on its domain.*

Definition 12 (Strong smoothness)**.** *Let $f\colon \mathcal{X} \to \mathbb{R}$. The function f is β-strongly smooth (with $\beta > 0$) if it is continuously differentiable everywhere on its domain and if for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,*

$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{x}') + \left\langle \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \right\rangle + \frac{\beta}{2} \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_{\mathcal{X}}^2.$$

If f is also convex, then this characterization is equivalent to the uniform Lipschitz continuity of the gradient $\boldsymbol{x} \mapsto \nabla f(\boldsymbol{x})$ with constant β [261, Lemma 4]:

$$\left\| \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}') \right\|_{\mathcal{X}} \leqslant \beta \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_{\mathcal{X}}.$$

Definition 13 (Saddle points)**.** *Let $\mathcal{L}\colon \mathcal{X} \times \mathcal{Y}^* \to \mathbb{R} \cup \{+\infty\}$ be a proper function. A pair of points $(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \in \mathcal{X} \times \mathcal{Y}^*$ is a saddle point of $\mathcal{L}$ if for every $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y}^* \in \mathcal{Y}^*$,*

$$\mathcal{L}(\boldsymbol{x}_s, \boldsymbol{y}^*) \leqslant \mathcal{L}(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \leqslant \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_s^*).$$

Definition 14 (Bregman divergences)**.** *Let $\phi \in \Gamma_0(\mathcal{X})$ with* $\operatorname{int}(\operatorname{dom} \phi) \neq \varnothing$. *The Bregman divergence of the function ϕ is the function $D_\phi\colon \mathcal{X} \times \operatorname{int}(\operatorname{dom} \phi) \to [0, +\infty]$ defined as*

$$D_\phi(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \max_{\boldsymbol{x}^* \in \partial \phi(\boldsymbol{x}')} \left\{ \left\langle \boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}' \right\rangle \right\}. \tag{1.4}$$

Note that Bregman divergences are sometimes defined differently in the convex analysis literature. The definition above is that from Bauschke et al. [20, Definition 7.1 and Lemma 7.3(i)].

A more general definition, which will be used in Chapter 3, expresses the Bregman divergence in terms of the spaces $\mathcal{X}$ and $\mathcal{X}^*$: Let $\phi \in \Gamma_0(\mathcal{X})$. The Bregman divergence of ϕ is the function $D_\phi \colon \mathcal{X} \times \mathcal{X}^* \to [0, +\infty]$ defined by

$$D_\phi(\boldsymbol{x}, \boldsymbol{x}^*) = \phi(\boldsymbol{x}) - \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle + \phi^*(\boldsymbol{x}^*). \tag{1.5}$$

Note that for every $\boldsymbol{x}' \in \operatorname{int}(\operatorname{dom} \phi)$, there exists some $\boldsymbol{x}^* \in \mathcal{X}^*$ for which $\boldsymbol{x}^* \in \partial\phi(\boldsymbol{x}')$. Hence in this case one may write

$$\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \max_{\boldsymbol{x}^* \in \partial\phi(\boldsymbol{x}')} \left\{ \left\langle \boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}' \right\rangle \right\} = \phi(\boldsymbol{x}) - \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle + \phi^*(\boldsymbol{x}^*),$$

which is precisely the right hand side of (1.5). Thus one can always express the Bregman divergence in (1.4) as the one in (1.5) for some appropriate $\boldsymbol{x}^* \in \mathcal{X}^*$. The converse does not hold in general because $\operatorname{int} \operatorname{dom} \phi \subset \operatorname{dom} \partial\phi$ *and the inclusion may be strict [214, Theorem 23.4].*

Definition 15 (Bregman proximity operators)**.** *Let $f, \phi \in \Gamma_0(\mathcal{X})$ with $\operatorname{int}(\operatorname{dom} \phi) \neq \varnothing$ and let $t > 0$. The Bregman D_ϕ-proximal operator $\operatorname{prox}_{(tf, D_\phi)}(\cdot)$ is a set-valued mapping defined for every $\boldsymbol{x}' \in \operatorname{int}(\operatorname{dom} \phi)$ as*

$$\operatorname{prox}_{(tf, D_\phi)}(\boldsymbol{x}') = \left\{ \hat{\boldsymbol{x}} \in \operatorname{dom} f \cap \operatorname{dom} \phi : tf(\hat{\boldsymbol{x}}) + D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{x}') = \inf_{\boldsymbol{x} \in \mathcal{X}} \left\{ tf(\boldsymbol{x}) + D_\phi(\boldsymbol{x}, \boldsymbol{x}') \right\} < +\infty \right\}. \tag{1.6}$$

Definition 16 (Projections)**.** *Let C denote a closed convex subset of $\mathbb{R}^n$. To every $\boldsymbol{x} \in \mathbb{R}^n$, there exists a unique element $\pi_C(\boldsymbol{x}) \in C$ called the projection of $\boldsymbol{x}$ onto C that is closest to $\boldsymbol{x}$ in Euclidean norm, i.e.,*

$$\pi_C(\boldsymbol{x}) \coloneqq \arg\min_{\boldsymbol{u} \in C} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 . \tag{1.7}$$

This correspondence defines a map $\boldsymbol{x} \mapsto \pi_C(\boldsymbol{x})$ from $\mathbb{R}^n$ to C called the projector onto C [10,

Chapter 0.6, Corollary 1]. It satisfies the characterization

$$\left\langle \boldsymbol{x} - \pi_C(\boldsymbol{x}), \boldsymbol{x}' - \pi_C(\boldsymbol{x}) \right\rangle \leqslant 0, \quad \forall \boldsymbol{x}' \in C. \tag{1.8}$$

Definition 17 (Infimal convolutions)**.** *Let $f_1 \in \Gamma_0(\mathbb{R}^n)$ and $f_2 \in \Gamma_0(\mathbb{R}^n)$. The infimal convolution of f_1 and f_2 is the function*

$$\mathbb{R}^n \ni \boldsymbol{x} \mapsto (f_1 \square f_2)(\boldsymbol{x}) = \inf_{\boldsymbol{x}_1 + \boldsymbol{x}_2 = \boldsymbol{x}} \{f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2)\}. \tag{1.9}$$

The infimal convolution is exact if the infimum is attained at $\boldsymbol{x}_1 \in \operatorname{dom} f_1$ and $\boldsymbol{x}_2 \in \operatorname{dom} f_2$, and in that case the infimum in (1.9) can be replaced by a minimum.

Definition 18 (Moreau–Yosida envelopes and proximal mappings)**.** *Let $t > 0$ and $J \in \Gamma_0(\mathbb{R}^n)$. The functions*

$$\boldsymbol{x} \mapsto \left(\frac{1}{2t} \|\cdot\|_2^2 \square J \right)(\boldsymbol{x}) = \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} \tag{1.10}$$

and

$$\boldsymbol{x} \mapsto \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} \tag{1.11}$$

are called the Moreau–Yosida envelope and proximal mapping of J, respectively [137, 182, 215].

1.2.2 Facts

Fact 1.2.1. *Let $\alpha > 0$ and let $\boldsymbol{A} \colon \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator. For every $(\boldsymbol{x}, \boldsymbol{y}^*), (\boldsymbol{x}', \boldsymbol{y}^{*\prime}) \in \mathcal{X} \times \mathcal{Y}^*$, the following auxiliary inequality holds:*

$$\left| \left\langle \boldsymbol{y}^* - \boldsymbol{y}^{*\prime}, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}') \right\rangle \right| \leqslant \|\boldsymbol{A}\|_{\mathrm{op}} \left(\frac{\alpha}{2} \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_{\mathcal{X}}^2 + \frac{1}{2\alpha} \left\| \boldsymbol{y}^* - \boldsymbol{y}^{*\prime} \right\|_{\mathcal{Y}^*}^2 \right). \tag{1.12}$$

Proof. From the Cauchy–Schwartz inequality,

$$\begin{aligned}
\left|\left\langle \boldsymbol{y}^* - \boldsymbol{y}^{*\prime}, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}')\right\rangle\right| &\leqslant \|\boldsymbol{A}\|_{\mathrm{op}} \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}} \left\|\boldsymbol{y}^* - \boldsymbol{y}^{*\prime}\right\|_{\mathcal{Y}^*} \\
&= \|\boldsymbol{A}\|_{\mathrm{op}} \left(\frac{\alpha}{2}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}}^2 + \frac{1}{2\alpha}\left\|\boldsymbol{y}^* - \boldsymbol{y}^{*\prime}\right\|_{\mathcal{Y}^*}^2\right) \\
&\quad - \|\boldsymbol{A}\|_{\mathrm{op}} \left(\sqrt{\frac{\alpha}{2}}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}} - \sqrt{\frac{1}{2\alpha}}\left\|\boldsymbol{y}^* - \boldsymbol{y}^{*\prime}\right\|_{\mathcal{Y}^*}\right)^2 \\
&\leqslant \|\boldsymbol{A}\|_{\mathrm{op}} \left(\frac{\alpha}{2}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}}^2 + \frac{1}{2\alpha}\left\|\boldsymbol{y}^* - \boldsymbol{y}^{*\prime}\right\|_{\mathcal{Y}^*}^2\right).
\end{aligned}$$

□

Fact 1.2.2 (Weighted averages of a convergent sequence)**.** *Let $\{\boldsymbol{x}_k\}_{k=1}^{+\infty} \subset \mathcal{X}$ be a sequence converging strongly to some $\boldsymbol{x} \in \mathcal{X}$, let $\{\lambda_k\}_{k=1}^{+\infty} \subset (0, +\infty)$ be a divergent sequence, i.e., $\sum_{k=1}^{+\infty} \lambda_k = +\infty$, and set $T_k = \sum_{j=1}^{k} \lambda_j$. Then*

$$\lim_{k \to +\infty} \left\| \frac{1}{T_k}\left(\sum_{j=1}^{k} \lambda_j \boldsymbol{x}_j\right) - \boldsymbol{x}\right\|_{\mathcal{X}} = 0.$$

Proof. Fix $\epsilon > 0$. Then there exists some $K_1 \in \mathbb{N}$ such that for every $k \geqslant K_1$, we have $\|\boldsymbol{x}_k - \boldsymbol{x}\|_{\mathcal{X}} < \epsilon/2$. Now, let $k \geqslant K_1$, take the difference between the weighted average $\frac{1}{T_k}\sum_{j=1}^{k} \lambda_j \boldsymbol{x}_j$ and $\boldsymbol{x}$, take the norm, use the triangle inequality and rearrange to get

$$\begin{aligned}
\left\| \frac{1}{T_k}\sum_{j=1}^{k} \lambda_j \boldsymbol{x}_j - \boldsymbol{x}\right\|_{\mathcal{X}} &= \left\| \frac{1}{T_k}\sum_{j=1}^{k} \lambda_j \left(\boldsymbol{x}_j - \boldsymbol{x}\right)\right\|_{\mathcal{X}} \\
&\leqslant \frac{1}{T_k}\sum_{j=1}^{k} \lambda_j \left\|\boldsymbol{x}_j - \boldsymbol{x}\right\|_{\mathcal{X}} \\
&= \frac{1}{T_k}\sum_{j=1}^{K_1-1} \lambda_j \left\|\boldsymbol{x}_j - \boldsymbol{x}\right\|_{\mathcal{X}} + \frac{1}{T_k}\sum_{j=K_1}^{k} \lambda_j \left\|\boldsymbol{x}_j - \boldsymbol{x}\right\|_{\mathcal{X}} \\
&\leqslant \frac{1}{T_k}\sum_{j=1}^{K_1-1} \lambda_j \left\|\boldsymbol{x}_j - \boldsymbol{x}\right\|_{\mathcal{X}} + \frac{1}{T_k}\sum_{j=K_1}^{k} \lambda_j \frac{\epsilon}{2} \\
&\leqslant \frac{1}{T_k}\sum_{j=1}^{K_1-1} \lambda_j \left\|\boldsymbol{x}_j - \boldsymbol{x}\right\|_{\mathcal{X}} + \frac{\epsilon}{2}.
\end{aligned}$$

The first term on the right hand side of the last line depends on k only through the term T_k. By

assumption, $T_k \to +\infty$ as $k \to +\infty$, and therefore there exists some $K_2 \in \mathbb{N}$ such that for $k \geqslant K_2$,

$$\frac{1}{T_k} \sum_{j=1}^{K_1-1} \lambda_j \left\| \boldsymbol{x}_j - \boldsymbol{x} \right\|_{\mathcal{X}} < \frac{\epsilon}{2}.$$

Taking $k \geqslant \max(K_1, K_2)$, we find

$$\left\| \frac{1}{T_k} \sum_{j=1}^{k} \lambda_j \boldsymbol{x}_j - \boldsymbol{x} \right\|_{\mathcal{X}} < \epsilon.$$

As ϵ was arbitrary positive number, we can take $\epsilon \to 0$ and obtain the desired result. □

Fact 1.2.3 (Supercoercivity)**.** *Let $f \in \Gamma_0(\mathcal{X})$ and suppose that f is supercoercive. Then for every $\alpha > 0$, there exists $\beta \in \mathbb{R}$ such that $f(\boldsymbol{x}) \geqslant \alpha \left\| \boldsymbol{x} \right\|_{\mathcal{X}} + \beta$ for every $\boldsymbol{x} \in \mathcal{X}$. In particular, a supercoercive function is always bounded from below.*

Proof. See [20, Lemma 3.2] for a proof. □

Fact 1.2.4 (Bounded sequences and weak convergence)**.** *Let $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ be a bounded sequence in $\mathcal{X}$. Then this sequence has a subsequence $\{\boldsymbol{x}_{k_l}\}_{l=1}^{+\infty}$ that converges weakly to some element in $\mathcal{X}$.*

Proof. See [33, Theorem 3.18]. □

Fact 1.2.5 (Strong convexity and strong smoothness)**.** *Let $f \in \Gamma_0(\mathcal{X})$. Then f is α-strongly convex (with $\alpha > 0$) if and only if its convex conjugate $f^* \in \Gamma_0(\mathcal{X}^*)$ is $\frac{1}{\alpha}$-strongly smooth.*

Proof. See [147, Theorem 6 and Appendix A.1]. □

Fact 1.2.6 (The primal problem and its dual problem)**.** *Let $g \in \Gamma_0(\mathcal{X})$, let $h \in \Gamma_0(\mathcal{Y})$, and let $\boldsymbol{A} \colon \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator. Assume the primal (minimization) problem*

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \{g(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})\} \tag{1.13}$$

has at least one solution and assume there exists $\boldsymbol{x} \in \mathcal{X}$ *such that* h *is continuous at* $\boldsymbol{A}\boldsymbol{x}$*. Then the dual (maximization) problem*

$$\sup_{\boldsymbol{y}^* \in \mathcal{Y}^*} \{-g^*(-\boldsymbol{A}^*\boldsymbol{y}^*) - h^*(\boldsymbol{y}^*)\} \tag{1.14}$$

is finite and has at least one solution. Moreover, if $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ *denotes a pair of solutions to the primal and dual problem then* $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ *satisfies the following optimality conditions*

$$-\boldsymbol{A}^*\boldsymbol{y}_s^* \in \partial g(\boldsymbol{x}_s) \quad \text{and} \quad \boldsymbol{y}_s^* \in \partial h(\boldsymbol{A}\boldsymbol{x}_s).$$

Proof. See [92, Theorem 4.1, Theorem 4.2, Equations (4.24)-(4.25)] for a proof. (Beware, in [92] the notation used for the solution $\boldsymbol{y}_s^*$ is flipped by a minus sign.) □

Fact 1.2.7 (Convex-concave saddle point problems)**.** *Let* $g \in \Gamma_0(\mathcal{X})$*, let* $h \in \Gamma_0(\mathcal{Y})$*, let* $\boldsymbol{A}\colon \mathcal{X} \to \mathcal{Y}$ *be a bounded linear operator, define the function* $\mathcal{L}\colon \mathcal{X} \times \mathcal{Y}^* \to \mathbb{R} \cup \{+\infty\}$ *as*

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) = g(\boldsymbol{x}) + \langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x} \rangle - h^*(\boldsymbol{y}^*).$$

Then the pair of points $(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \in \mathcal{X} \times \mathcal{Y}^*$ *is a saddle point of* $\mathcal{L}$ *if and only if* $\boldsymbol{x}_s$ *is a solution of the primal problem* (1.13) *and* $\boldsymbol{y}_s^*$ *is a solution of the dual problem* (1.14)*.*

Proof. See [92, Proposition 3.1, page 57]. □

Fact 1.2.8 (Properties of Bregman divergences)**.** *Let* $\phi \in \Gamma_0(\mathcal{X})$ *with* $\operatorname{int}(\operatorname{dom} \phi) \neq \varnothing$*, let* $\boldsymbol{x} \in \operatorname{dom} \phi$*, and let* $\boldsymbol{x}', \hat{\boldsymbol{x}} \in \operatorname{int}(\operatorname{dom} \phi)$*. Assume that* ϕ *is differentiable on* $\operatorname{int}(\operatorname{dom} \phi)$*. Then the Bregman divergence* D_ϕ *of* ϕ *satisfies the following properties:*

(i) The Bregman divergence can be written as $D_\phi(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \langle \nabla\phi(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle$*.*

(ii) The Bregman divergence D_ϕ *satisfies the three-point identity*

$$D_\phi(\boldsymbol{x}, \boldsymbol{x}') = D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{x}') + D_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \left\langle \nabla\phi(\boldsymbol{x}') - \nabla\phi(\hat{\boldsymbol{x}}), \hat{\boldsymbol{x}} - \boldsymbol{x} \right\rangle. \tag{1.15}$$

(iii) If ϕ *is essentially strictly convex, then* $D_\phi(\boldsymbol{x}, \boldsymbol{x}') = 0$ *if and only if* $\boldsymbol{x} = \boldsymbol{x}'$*.*

(iv) *If ϕ is essentially strictly convex, then the function $\boldsymbol{x} \mapsto D_\phi(\boldsymbol{x}, \boldsymbol{x}')$ is coercive for every $\boldsymbol{x}' \in \operatorname{int}(\operatorname{dom} \phi)$.*

(v) *If ϕ is supercoercive, then the function $\boldsymbol{x}' \mapsto D_\phi(\boldsymbol{x}, \boldsymbol{x}')$ is coercive for every $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom} \phi)$.*

(vi) *If $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ is a sequence in $\operatorname{int}(\operatorname{dom} \phi)$ converging to a point $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom} \phi)$, then*

$$\lim_{k \to +\infty} D_\phi(\boldsymbol{x}, \boldsymbol{x}_k) = 0.$$

(vii) *Assume that ϕ is essentially smooth. If $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ is a sequence in $\operatorname{int}(\operatorname{dom} \phi)$ converging to a point $\boldsymbol{x}_c \in \operatorname{int}(\operatorname{dom} \phi)$, then*

$$\lim_{k \to +\infty} D_\phi(\boldsymbol{x}, \boldsymbol{x}_k) = D_\phi(\boldsymbol{x}, \boldsymbol{x}_c).$$

(viii) *If ϕ is m-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$, then*

$$D_\phi(\boldsymbol{x}, \boldsymbol{x}') \geqslant \frac{m}{2} \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_{\mathcal{X}}^2.$$

Proof. See [20, Lemma 7.3] for the proof of (i) and (iii)-(vi). Statement (ii) follows from (i) and a straightforward calculation. Statement (vii) follows from (i) and the continuity of both ϕ and $\nabla\phi$ over $\operatorname{int}(\operatorname{dom} \phi)$. Statement (viii) follows from (i) and inequality (1.2). □

Fact 1.2.9 (Properties of Bregman proximity operators)**.** *Let $f, \phi \in \Gamma_0(\mathcal{X})$ be two functions such that $\operatorname{dom} f \cap \operatorname{int}(\operatorname{dom} \phi) \neq \varnothing$, let $t > 0$, and assume ϕ is essentially smooth and essentially strictly convex. In addition, assume that either f is bounded from below or ϕ is supercoercive. Then the following properties hold:*

(i) *The proximal operator $\boldsymbol{x}' \mapsto \operatorname{prox}_{(tf, D_\phi)}(\boldsymbol{x}')$ defined in (1.6) is single-valued on its domain $\operatorname{int}(\operatorname{dom} \phi)$. That is, for every $\boldsymbol{x}' \in \operatorname{int}(\operatorname{dom} \phi)$,*

$$\operatorname{prox}_{(tf, D_\phi)}(\boldsymbol{x}') = \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \left\{ tf(\boldsymbol{x}) + D_\phi(\boldsymbol{x}, \boldsymbol{x}') \right\}.$$

Moreover, $\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}') \in \operatorname{dom} \partial f \cap \operatorname{int}(\operatorname{dom} \phi)$.

(ii) For every $\boldsymbol{x} \in \operatorname{dom} f$ *and* $\boldsymbol{x}' \in \operatorname{int}(\operatorname{dom} \phi)$*, the proximal point* $\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')$ *satisfies the characterization*

$$f(\boldsymbol{x}) - f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) - \frac{1}{t}\left\langle \nabla\phi(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) - \nabla\phi(\boldsymbol{x}'), \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}') - \boldsymbol{x} \right\rangle \geqslant 0. \tag{1.16}$$

If, in addition, there exists $\gamma_f > 0$ *such that the function* $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - \gamma_f\phi(\boldsymbol{x})$ *is convex, then this characterization can be strengthened to*

$$\begin{aligned} f(\boldsymbol{x}) - f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) - \frac{1}{t}\left\langle \nabla\phi(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) - \nabla\phi(\boldsymbol{x}'), \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}') - \boldsymbol{x} \right\rangle \\ \geqslant \gamma_f D_\phi(\boldsymbol{x}, \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')). \end{aligned} \tag{1.17}$$

(iii) For every $\boldsymbol{x} \in \operatorname{dom} f$ *and* $\boldsymbol{x}' \in \operatorname{int}(\operatorname{dom} \phi)$*,*

$$\begin{aligned} f(\boldsymbol{x}) + \frac{1}{t}D_\phi(\boldsymbol{x}, \boldsymbol{x}') \geqslant f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) + \frac{1}{t}D_f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}'), \boldsymbol{x}') \\ + \frac{1}{t}D_f(\boldsymbol{x}, \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')). \end{aligned} \tag{1.18}$$

If, in addition, there exists $\gamma_f > 0$ *such that the function* $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - \gamma_f\phi(\boldsymbol{x})$ *is convex, then* (1.18) *can be strengthened to*

$$\begin{aligned} f(\boldsymbol{x}) + \frac{1}{t}D_\phi(\boldsymbol{x}, \boldsymbol{x}') \geqslant f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) + \frac{1}{t}D_f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}'), \boldsymbol{x}') \\ + \left(\frac{1}{t} + \gamma_f\right) D_f(\boldsymbol{x}, \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')). \end{aligned} \tag{1.19}$$

Proof. See [21, Proposition 3.21-3.23, Theorem 3.24, Corollary 3.25] for the proof of statements (i). Statement (ii) follows directly from [92, Proposition 2.2, page 38]. To prove inequality (1.18) in (iii), use the characterization (1.16) to write

$$\begin{aligned} f(\boldsymbol{x}) + \frac{1}{t}D_\phi(\boldsymbol{x}, \boldsymbol{x}') \geqslant f(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')) + \frac{1}{t}D_\phi(\boldsymbol{x}, \boldsymbol{x}') \\ - \frac{1}{t}\left\langle \nabla\phi(\boldsymbol{x}') - \nabla\phi(\operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')), \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}') - \boldsymbol{x} \right\rangle. \end{aligned}$$

Then use the three-point identity (1.15) with $\hat{\boldsymbol{x}} = \operatorname{prox}_{(tf,D_\phi)}(\boldsymbol{x}')$ to obtain (1.18). The proof of

inequality (1.19) in (iii) is nearly identical, with the exception that the characterization (1.17) is used in place of (1.16). □

Fact 1.2.10 (Monotone property of the subdifferential)**.** *Let $f \in \Gamma_0(\mathcal{X})$ and suppose that f is m-strongly convex over its domain. Then for every pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \operatorname{dom} \partial f \times \operatorname{dom} \partial f$ and $(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*) \in \partial f(\boldsymbol{x}_1) \times \partial f(\boldsymbol{x}_2)$,*

$$m \left\| \boldsymbol{x}_1 - \boldsymbol{x}_2 \right\|_2^2 \leqslant \left\langle \boldsymbol{x}_1^* - \boldsymbol{x}_2^*, \boldsymbol{x}_1 - \boldsymbol{x}_2 \right\rangle . \tag{1.20}$$

Proof. This follows directly from [214, Page 240, Corollary 31.5.2]. □

Fact 1.2.11 (Set-valued subdifferential mapping)**.** *Let $f \in \Gamma_0(\mathbb{R}^n)$. The mapping $\operatorname{dom} \partial f \ni \boldsymbol{x} \mapsto \pi_{\partial f(\boldsymbol{u})}(\mathbf{0})$, which selects the subgradient of minimal norm in the subdifferential $\partial f(\boldsymbol{x})$, is well-defined and defines a function continuous almost everywhere on $\operatorname{dom} \partial f$.*

Proof. This follows directly from the fact that the mapping agrees with the gradient of f over the set of points in int (dom J) at which f is differentiable [214, Theorem 25.5]. □

Fact 1.2.12 (Properties of convex conjugates)**.** *Let $f \in \Gamma_0(\mathcal{X})$. The mapping $f \mapsto f^*$ is one-to-one, $(f^*)^* = f$, and for every $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{x}^* \in \mathcal{X}$, the functions f and f^* satisfy Fenchel's inequality*

$$f(\boldsymbol{x}) + f^*(\boldsymbol{x}^*) \geqslant \left\langle \boldsymbol{x}^*, \boldsymbol{x} \right\rangle , \tag{1.21}$$

where equality holds if and only if $\boldsymbol{x}^* \in \partial f(\boldsymbol{x})$, if and only if $\boldsymbol{x} \in \partial f^*(\boldsymbol{x}^*)$.

Proof. See [137, Corollary 1.4.4]. □

Fact 1.2.13 (Properties of the infimal convolution)**.** *(i) Let $f_1 \in \Gamma_0(\mathbb{R}^n)$ and $f_2 \in \Gamma_0(\mathbb{R}^n)$ and suppose that the relative interiors of f_1 and f_2 have a point in common, i.e.,* ri dom $f_1 \cap$ ri dom $f_2 \neq \varnothing$. *Then the convex conjugate of the infimal convolution $f_1 \square f_2$ at $\boldsymbol{x}^* \in \mathbb{R}^n$ equals the sum of their respective convex conjugate, that is,*

$$\left(f_1 \square f_2\right)^* (\boldsymbol{p}) = f_1^*(\boldsymbol{p}) + f_2^*(\boldsymbol{p}).$$

(ii) [Moreau's decomposition] Let $f \in \Gamma_0(\mathbb{R}^n)$. Then the following decomposition holds:

$$\frac{1}{2}\|\cdot\|_2^2 \square f + \frac{1}{2}\|\cdot\|_2^2 \square f^* = \frac{1}{2}\|\cdot\|_2^2.$$

(iii) [Deconvolutions] Suppose f_1 and f_2 are two convex functions on $\mathbb{R}^n$ such that $f_1 + f_2 = \frac{1}{2}\|\cdot\|_2^2$. Then there exists a unique function $f \in \Gamma_0(\mathbb{R}^n)$ such that

$$f_1 = \frac{1}{2}\|\cdot\|_2^2 \square f \text{ and } f_2 = \frac{1}{2}\|\cdot\|_2^2 \square f^*,$$

where $f(\boldsymbol{x}) = f_2^*(\boldsymbol{x}) - \frac{1}{2}\|\boldsymbol{x}\|_2^2$ for every $\boldsymbol{x} \in \mathbb{R}^n$. Moreover, f_1 and f_2 are continuously differentiable and

$$\nabla f_1(\boldsymbol{x}) \in \partial f(\nabla h(\boldsymbol{x})) \quad \text{and} \quad \nabla f_2(\boldsymbol{x}) \in \partial f^*(\nabla g(\boldsymbol{x})).$$

Proof. See [214, Theorem 16.4] for the proof of (i). Item (ii) is known as the Moreau's decomposition theorem and the proof can be found in [138, 182]. See [138] for the proof of (iii). □

Fact 1.2.14 (Moreau–Yosida envelopes, proximal mappings and their connections to Hamilton—Jacobi PDEs)**.** *Let $J \in \Gamma_0(\mathbb{R}^n)$. Then the following statements hold.*

(i) The unique continuously differentiable and convex function $S_0 \colon \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}$ that satisfies the first-order Hamilton–Jacobi equation with initial data

$$\begin{cases} \dfrac{\partial S_0}{\partial t}(\boldsymbol{x}, t) + \dfrac{1}{2}\|\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t)\|_2^2 = 0 & \text{in } \mathbb{R}^n \times (0, +\infty), \\ S_0(\boldsymbol{x}, 0) = J(\boldsymbol{x}) & \text{in } \mathbb{R}^n, \end{cases} \tag{1.22}$$

is defined by

$$S_0(\boldsymbol{x}, t) = \left((\frac{1}{2t}\|\cdot\|_2^2) \square J \right)(\boldsymbol{x}) \qquad (\textit{Lax–Oleinik formula}) \tag{1.23}$$

$$= \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\}. \tag{1.24}$$

Furthermore, for every $\boldsymbol{x} \in \operatorname{dom} J$*, sequence* $\{t_k\}_{k=1}^{+\infty}$ *of positive real numbers converging to* 0*, and sequence* $\{\boldsymbol{d}_k\}_{k=1}^{+\infty}$ *of vectors converging to* $\boldsymbol{d} \in \mathbb{R}^n$*, the pointwise limit* $S_0(\boldsymbol{x} + t_k\boldsymbol{d}_k, t_k)$ *as* $k \to +\infty$ *exists and satisfies*

$$\lim_{k\to+\infty} S_0(\boldsymbol{x} + t_k\boldsymbol{d}_k, t_k) = J(\boldsymbol{x}).$$

(ii) For every $\boldsymbol{x} \in \mathbb{R}^n$ *and* $t > 0$*, the infimum in (1.24) exists and is attained at a unique point* $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) \in \operatorname{dom} \partial J$ *(see Equation (4.2)) In addition, the minimizer* $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$ *satisfies the formula*

$$\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t), \tag{1.25}$$

and

$$\frac{\boldsymbol{x} - \boldsymbol{u}_{MAP}(\boldsymbol{x}, t)}{t} \in \partial J(\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)).$$

(iii) Let $\{t_k\}_{k=1}^{+\infty}$ *be a sequence of positive real numbers converging to zero and let* $\{\boldsymbol{d}_k\}_{k=1}^{+\infty}$ *be a sequence of elements in* $\mathbb{R}^n$ *converging to some* $\boldsymbol{d} \in \mathbb{R}^n$*. Then, for every* $\boldsymbol{x} \in \operatorname{dom} J$ *the pointwise limit of* $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$ *as* $t \to 0$ *exists and satisfies*

$$\lim_{k\to+\infty} \boldsymbol{u}_{MAP}(\boldsymbol{x} + t_k\boldsymbol{d}_k, t_k) = \boldsymbol{x}.$$

(iv) Let $\boldsymbol{x} \in \operatorname{dom} \partial J$ *and let* $\{t_k\}_{k=1}^{+\infty}$ *be a sequence of positive real numbers converging to zero. Then the limit of* $\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t_k)$ *as* $k \to +\infty$ *exists and satisfies*

$$\lim_{k\to+\infty} \nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t_k) = \pi_{\partial J(\boldsymbol{x})}(\boldsymbol{0}). \tag{1.26}$$

Proof. See [64] for the proof of these statements. □

Chapter Two

Variational methods for machine learning algorithms and connections to Hamilton–Jacobi PDEs I: Theory

2.1 Introduction

2.1.1 Overview

This chapter and the following present variational methods, also called optimization methods, for solving certain types of Hamilton–Jacobi partial differential equations that appear in machine learning. As we will see later, several supervised machine learning algorithms for regression and classification admit formulations in terms of Hamilton–Jacobi partial differential equations. In that sense, this chapter and the following focus on novel optimization methods for solving several supervised machine learning problems, although, as will be argued later, in a far more efficient way than competing methods in the literature. The starting point of this chapter concerns first-order convex optimization methods for solving convex optimization problems with saddle-point structure, specifically the linear primal-dual hybrid gradient method.

The linear primal-dual hybrid gradient (PDHG) method is a first-order splitting method for minimizing the sum of two convex functions [46, 47, 95, 204, 205, 262]. It works by splitting the sum into smaller subproblems, each of which is easier to solve. These subproblems, unlike those obtained from most splitting methods, can generally be solved efficiently because they involve simple operations such as matrix-vector multiplications or proximal mappings that are fast to evaluate numerically. This makes the linear PDHG method flexible and easy to implement for solving a wide range of constrained and nondifferentiable optimization problems. Due to this advantage, the linear PDHG method is widely used for solving problems in imaging science [24, 32, 96, 119, 155, 157, 212], optimal control [98, 154], compressive sensing [110, 143], distributed optimization [197, 219, 220], and optimal transport [40, 90, 103, 113, 165, 194]. It is also used, to a limited extent, for solving large-scale problems in machine learning [8, 17, 43, 135, 206, 221, 253].

Despite its flexibility and ease of implementation, the linear PDHG method requires precise stepsize parameters for the problem at hand to achieve an optimal convergence rate. Unfortunately, these stepsize parameters are often prohibitively expensive to compute for large-scale optimization problems. This issue makes the otherwise simple linear PDHG method unsuitable for solving

large-scale optimization problems, such as those in machine learning. This issue is shared by most first-order optimization methods as well.

To illustrate this point, consider the ℓ_1-constrained logistic regression problem

$$\inf_{\substack{\boldsymbol{v}\in\mathbb{R}^d \\ \|\boldsymbol{v}\|_1\leqslant\lambda}} \frac{1}{m}\sum_{i=1}^{m} \log\left(1+e^{-[\boldsymbol{b}]_i\langle \boldsymbol{u}_i,\boldsymbol{v}\rangle}\right), \tag{2.1}$$

where $\{\boldsymbol{u}_i, [\boldsymbol{b}]_i\}_{i=1}^m$ denote a collection of m feature vectors $\boldsymbol{u}_i \in \mathbb{R}^d$ with labels $[\boldsymbol{b}]_i \in \{-1,+1\}$ and $\lambda > 0$ is a parameter. This problem can be solved using the linear PDHG method as follows. Let $\boldsymbol{B}$ denote the $m \times d$ matrix whose rows are the elements $-[\boldsymbol{b}]_i\boldsymbol{u}_i$, let $\boldsymbol{B}^*$ denote its matrix transpose, and let $\|\boldsymbol{B}\|_{2,2}$ denote the largest singular value of $\boldsymbol{B}$. Formally, the linear PDHG method computes a global minimum of problem (2.1) via the iterations [47, Algorithm 5][182]

$$\begin{aligned}
\boldsymbol{z}_k &= \boldsymbol{w}_k + \sigma_k\boldsymbol{B}(\boldsymbol{v}_k + \theta_k[\boldsymbol{v}_k - \boldsymbol{v}_{k-1}]),\\
\boldsymbol{w}_{k+1} &= \boldsymbol{z}_k - \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^m}\left\{\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{z}_k\|_2^2 + \frac{\sigma_k}{m}\sum_{i=1}^{m}\log\left(1+e^{[\boldsymbol{w}]_i/\sigma_k}\right)\right\},\\
\boldsymbol{v}_{k+1} &= \operatorname*{arg\,min}_{\substack{\boldsymbol{v}\in\mathbb{R}^n \\ \|\boldsymbol{v}\|_1\leqslant\lambda}} \frac{1}{2}\|\boldsymbol{v}-(\boldsymbol{v}_k-\tau_k\boldsymbol{B}\boldsymbol{w}_{k+1})\|_2^2,\\
\theta_{k+1} &= 1/\sqrt{1+4m\sigma_k}, \quad \tau_{k+1} = \tau_k/\theta_{k+1} \quad \text{and} \quad \sigma_{k+1} = \theta_{k+1}\sigma_k,
\end{aligned} \tag{2.2}$$

where $\boldsymbol{v}_{-1} = \boldsymbol{v}_0$ are vectors in the interior of the d-dimensional ℓ_1-ball of radius λ, $\boldsymbol{w}_0$ is a vector in $\mathbb{R}^m$, and $\tau_0 > 0$, $\sigma_0 = 1/(\|\boldsymbol{B}\|_{2,2}^2\,\tau_0)$ and $\theta_0 = 0$ are the initial stepsize parameters. The updates for $\boldsymbol{w}_{k+1}$ and $\boldsymbol{v}_{k+1}$ in (2.2) can be evaluated efficiently using standard first or second-order optimization methods and efficient ℓ_1-ball projection algorithms [60], respectively. The other operations in the updates can all be computed exactly in at most $O(md)$ operations. The convergence rate for this method is $O(1/k^2)$ in the number of iterations k, which is the best possible achievable rate of convergence for this problem in the Nesterov class of optimal first-order methods [188].

Attaining this optimal rate of convergence requires a precise estimate of the largest singular value $\|\boldsymbol{B}\|_{2,2}$ of the matrix $\boldsymbol{B}$. However, this quantity takes on the order of $O(\min(m^2d, md^2))$ operations to compute [130], which makes it essentially impossible to estimate the largest singular

value for large matrices. Line search methods and other heuristics are often used to bypass this issue, but they typically slow down the convergence too much to alleviate the problem. Most first-order optimization methods used for solving large-scale optimization problems share this issue as well.

To address this issue, we present novel accelerated nonlinear PDHG methods that can achieve an optimal rate of convergence with stepsize parameters that are simple and efficient to compute. Returning to the previous example, let $\|\boldsymbol{B}\|_{1,2}$ denote the maximum ℓ_2 norm of a column of the matrix $\boldsymbol{B}$ and define new parameters $\hat{\tau}_0 > 0$, $\hat{\sigma}_0 = 1/(\|\boldsymbol{B}\|_{1,2}^2\,\hat{\tau}_0)$ and $\hat{\theta} = 0$. In addition, let $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$ denote vectors contained in the interior of the $2d$-dimensional unit simplex Δ_{2d}, let $\boldsymbol{y}_0^*$ denote a vector in the m-dimensional cube $(0, 1/m)^m$, let $[\hat{\boldsymbol{w}}_0^*]_i = \log\left(m[\boldsymbol{y}_0^*]_i/(1 - m[\boldsymbol{y}_0^*]_i)\right)$ for $i \in \{1, \ldots, m\}$, and let $\boldsymbol{A} = \lambda(\boldsymbol{B}| - \boldsymbol{B})$ denote the horizontal concatenation of the matrices $\lambda\boldsymbol{B}$ and $-\lambda\boldsymbol{B}$. Then, we show in Sections 2.4.4 and 3.2 that the accelerated nonlinear PDHG method

$$
\begin{aligned}
\hat{\boldsymbol{w}}_{k+1} &= \left(4m\hat{\sigma}_k\boldsymbol{x}_k + 4m\hat{\sigma}_k\hat{\theta}\,(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) + \hat{\boldsymbol{w}}_k\right)/(1 + 4m\hat{\sigma}_k),\\
[\boldsymbol{y}_{k+1}^*]_i &= \frac{1}{m + me^{-[\boldsymbol{w}_{k+1}]_i}} \quad \text{for } i \in \{1, \ldots, m\},\\
[\boldsymbol{x}_{k+1}]_j &= \frac{[\boldsymbol{x}_k]_j e^{-\hat{\tau}_k[\boldsymbol{A}^*\boldsymbol{y}_{k+1}^*]_j}}{\sum_{j=1}^m [\boldsymbol{x}_k]_j e^{-\hat{\tau}_k[\boldsymbol{A}^*\boldsymbol{y}_{k+1}^*]_j}} \quad \text{for } j \in \{1, \ldots, n\},\\
\boldsymbol{v}_{k+1} &= \lambda(\boldsymbol{B} \mid -\boldsymbol{B})\boldsymbol{x}_{k+1}\\
\hat{\theta}_{k+1} &= 1/\sqrt{1 + 4m\hat{\sigma}_k}, \quad \hat{\tau}_{k+1} = \hat{\tau}_k/\hat{\theta}_{k+1}, \quad \text{and} \quad \hat{\sigma}_{k+1} = \hat{\theta}_{k+1}\hat{\sigma}_k,
\end{aligned}
\tag{2.3}
$$

computes a global minimum of problem (2.1) through the iterates $\boldsymbol{v}_k$. Moreover, the convergence rate is $O(1/k^2)$ in the number of iterations k, which is the best possible achievable rate of convergence for this problem in the Nesterov class of optimal first-order methods [188].

Unlike in the linear PDHG method (2.2), the stepsize parameters in the nonlinear PDHG method (2.3) are computed in optimal $\Theta(md)$ operations from the matrix norm $\|\boldsymbol{A}\|_{1,2}$. In addition, the computational bottleneck in the iterates consists of matrix-vector multiplications that can be computed in $O(md)$ operations or better with appropriate parallel algorithms. Thus all stepsize parameters and updates in the nonlinear method (2.3) are computed in quadratic $O(md)$ time, in contrast to the stepsize parameters in the linear method (2.2) which are computed in cubic

$O(\min(m^2d, md^2))$ time. This gain turns out to be considerable in practice: In Section (3.2.4) of Chapter 3, we will present some numerical experiments in which the nonlinear PDHG method (2.3) converges 5 to 10 times faster than the linear PDHG method (2.2).

2.1.2 Connections to Hamilton–Jacobi partial differential equations?

What does nonlinear PDHG optimization methods and ℓ_1-constrained logistic regression have to do with Hamilton–Jacobi partial differential equations (HJ PDEs)? As a first step to answer this question, let's write the ℓ_1-constraint in problem (2.1) in terms of its convex conjugate, which turns out to be the ℓ_∞ norm. Specifically, writing

$$\boldsymbol{v} \mapsto \|\boldsymbol{v}\|_\infty^* = \begin{cases} 0, \text{ if } \|\boldsymbol{v}\|_1 \leqslant 1, \\ +\infty, \text{ otherwise,} \end{cases}$$

the ℓ_1-constrained logistic regression problem (2.4) can then be written in terms of this convex conjugate as follows:

$$\inf_{\boldsymbol{v}\in\mathbb{R}^d} \left\{ \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-[\boldsymbol{b}]_i\langle \boldsymbol{u}_i, \boldsymbol{v}\rangle}\right) + \lambda \left\| \frac{\boldsymbol{v}}{\lambda} \right\|_\infty^* \right\}. \tag{2.4}$$

In this form, with the ℓ_∞ norm, the connection between ℓ_∞-constrained logistic regression and HJ PDEs can be made explicit. Suppose the feature vectors $\boldsymbol{u}_i$ are all linearly independent. Then Equation (2.4) turns out to be the Lax–Oleinik representation formula to the (classical) solution of a first-order HJ PDE with Hamiltonian equal to the ℓ_∞ norm and initial data equal to the logistic regression model. More precisely, let $S\colon \mathbb{R}^d \times [0,_\infty) \to \mathbb{R}$ denote the function defined by

$$S(\boldsymbol{v}', t) = \inf_{\boldsymbol{v}\in\mathbb{R}^d} \left\{ \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-[\boldsymbol{b}]_i\langle \boldsymbol{u}_i, \boldsymbol{v}\rangle}\right) + \lambda \left\| \frac{\boldsymbol{v}' - \boldsymbol{v}}{\lambda} \right\|_\infty^* \right\}.$$

Then the function $(\boldsymbol{v}', t) \mapsto S(\boldsymbol{v}', t)$ is the classical solution to the first-order HJ PDE [16, Propo-

sition 4.1]

$$\begin{cases} \dfrac{\partial S}{\partial t}(\boldsymbol{v}', t) + \left\| \nabla_{\boldsymbol{v}'} S(\boldsymbol{v}', t) \right\|_{\infty} = 0, & \boldsymbol{v}' \in \mathbb{R}^n,\ t \in [0, +\infty), \\ S(\boldsymbol{v}', 0) = \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \log\left(1 + e^{-[\boldsymbol{b}]_i \langle \boldsymbol{u}_i, \boldsymbol{v}' \rangle}\right), & \boldsymbol{v}' \in \mathbb{R}^n. \end{cases} \tag{2.5}$$

Solving the ℓ_1-constrained logistic regression problem (2.1) therefore amounts to evaluating the solution of the HJ PDE (2.5) at $\boldsymbol{v}' = \mathbf{0}$ and $t = \lambda$.

The formal derivation above shows how ℓ_1-constrained logistic regression connects to HJ PDEs. Similar connections also hold for a broad class of supervised machine learning algorithms; these will be made explicit in Section 2.5. The work presented in this chapter and the following will not leverage these connections explicitly, but in future work described in Section 5.1 of the last chapter will describe how these connections could be leveraged in practice for sparse logistic regression. Finally, note that the representation formulas for many first-order HJ PDEs can be cast as convex optimization problems with appropriate saddle-point structure. Hence the accelerated nonlinear PDHG optimization methods presented in this chapter should prove particularly efficient and robust for solving high-dimensional first-order HJ PDEs, including those that arise in optimal control and in imaging science [69, 68, 154].

2.1.3 Related work

The linear PDHG method was introduced at around the same time by Pock et al. [205] and Esser et al. [95] to solve problems in imaging science (see also earlier work from [207, 262]). The convergence of the linear PDHG method for problems posed on Euclidean spaces was later proven by Chambolle and Pock [46]. In addition to a proof of convergence, their work provided accelerated schemes of the linear PDHG method for problems with some degree of smoothness or strong convexity or both.

Since then, many variants and extensions of the linear PDHG method have been proposed; see [48, 47, 49] for further details and references. A partial list of these variants include: overrelaxed [59, 133], inertial [166], operator, forward-backward, and proximal-gradient split-

ting [29, 58, 75, 84, 245], multistep [56], stochastic [193, 49, 101, 197, 241, 248, 253], and nonlinear [47, 140] variants, including the mirror descent method [186]. Here, we focus on nonlinear PDHG methods.

The extension of the linear PDHG method to the nonlinear setting was first done, to our knowledge, by Hohage and Homann [140] to solve non-smooth convex optimization problems posed on Banach spaces. A nonlinear PDHG method for solving such problems using nonlinear proximity operators based on Bregman divergences was later proposed by Chambolle and Pock [47]. Their work also provided an accelerated and partially nonlinear scheme for solving strongly convex problems. Their scheme is not fully nonlinear, however, as it requires one of the Bregman divergence to be a quadratic function. Moreover, their work did not provide accelerated nonlinear schemes for smooth convex problems or smooth and strongly convex problems.

2.1.4 Contributions

This chapter contributes accelerated nonlinear PDHG methods that achieve an optimal rate of convergence in the Nesterov class of optimal first-order methods with stepsize parameters that are simple and efficient to compute. To do so, we extend the theory of accelerated nonlinear PDHG methods initiated in [47] to solve optimization problems on Banach spaces with nonlinear proximity operators based on Bregman divergences. The main theoretical results and accelerated nonlinear PDHG methods are described in Section 2.4. We prove rigorous convergence results, including results strongly convex or smooth problems posed on infinite-dimensional reflexive Banach spaces. The results we present are generally applicable to convex-concave saddle-point optimization problems posed on real reflexive Banach spaces. In addition, we present in Section 2.5 some novel theoretical connections between a broad class of supervised machine learning problems and first-order HJ PDEs with initial data. The following chapter will describe accelerated nonlinear PDHG methods for several of these problems, including regularized logistic regression, regularized maximum entropy estimation and entropy-regularized zero-sum matrix games.

2.2 Setup

We are interested here with convex-concave saddle-point problems posed on real reflexive Banach spaces. Concretely, let $\mathcal{X}$ and $\mathcal{Y}$ denote two real reflexive Banach spaces endowed with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, and let $\boldsymbol{A}\colon \mathcal{X} \to \mathcal{Y}$ denote a bounded linear operator between those two spaces. We consider the following convex-concave saddle-point problem

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{y}\in\mathcal{Y}^*} \{g(\boldsymbol{x}) + \langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x}\rangle - h^*(\boldsymbol{y}^*)\} \tag{2.6}$$

where $g \in \Gamma_0(\mathcal{X})$ and $h \in \Gamma_0(\mathcal{Y})$. Formally, this is the primal-dual formulation associated to the primal problem

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \{g(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})\} \tag{2.7}$$

and the dual problem

$$\sup_{\boldsymbol{y}^*\in\mathcal{Y}^*} \{-g^*(-\boldsymbol{A}^*\boldsymbol{y}^*) - h^*(\boldsymbol{y}^*)\}. \tag{2.8}$$

The objective function $\mathcal{L}\colon \mathcal{X} \times \mathcal{Y}^* \to \mathbb{R} \cup \{+\infty\}$ in the saddle-point problem (2.6), namely

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) = g(\boldsymbol{x}) + \langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x}\rangle - h^*(\boldsymbol{y}^*), \tag{2.9}$$

is called the Lagrangian of the primal and dual problems (2.7) and (2.8). Solutions to the saddle-point problem (2.6), when they exist, are saddle points of the Lagrangian (2.9) (see Definition (13) and Fact (1.2.6)).

This work focuses on accelerated nonlinear PDHG methods designed to compute saddle points of (2.6), and therefore solutions to the primal and dual problems (2.7) and (2.8). We describe below the formalism behind the nonlinear PDHG method. Let $\phi_{\mathcal{X}} \in \Gamma_0(\mathcal{X})$ and $\phi_{\mathcal{Y}^*} \in \Gamma_0(\mathcal{Y}^*)$ denote two essentially smooth and essentially strictly convex functions, and consider their corresponding Bregman divergences:

$$D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \phi_{\mathcal{X}}(\boldsymbol{x}) - \phi(\bar{\boldsymbol{x}}) - \langle \nabla\phi_{\mathcal{X}}(\bar{\boldsymbol{x}}), \boldsymbol{x} - \bar{\boldsymbol{x}}\rangle$$

$$D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) = \phi_{\mathcal{Y}^*}(\boldsymbol{y}^*) - \phi_{\mathcal{Y}^*}(\bar{\boldsymbol{y}}^*) - \langle \boldsymbol{y}^* - \bar{\boldsymbol{y}}^*, \nabla\phi_{\mathcal{Y}^*}(\bar{\boldsymbol{y}}^*)\rangle\,.$$

Formally, we propose using these Bregman divergence to alternate in (2.6) a nonlinear proximal descent step in the variable $\boldsymbol{x}$ and a nonlinear proximal ascent step in the variable $\boldsymbol{y}^*$ as follows:

$$\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \left\{ g(\boldsymbol{x}) + \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}\boldsymbol{x}\rangle + \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \right\} \\ \hat{\boldsymbol{y}}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^*\in\mathcal{Y}^*} \left\{ -h^*(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}\tilde{\boldsymbol{x}}\rangle - \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) \right\}. \end{aligned} \tag{2.10}$$

The iteration scheme (2.10) takes the stepsize parameters $\tau, \sigma > 0$, initial points $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}^*) \in \mathcal{X} \times \mathcal{Y}^*$, and intermediate points $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}^*) \in \mathcal{X} \times \mathcal{Y}^*$ to output the new points $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*)$. The nonlinear PDHG method consists of this iteration scheme with appropriate parameter values and initial and intermediate points to attain an optimal convergence rate.

Assumptions

(A1) The two functions g and h are proper, lower semicontinuous, and convex over their respective domains $\mathcal{X}$ and $\mathcal{Y}$. Moreover, the primal problem (2.7) has at least one solution and there exists a point $\boldsymbol{x} \in \operatorname{dom} g$ such that $\boldsymbol{A}\boldsymbol{x} \in \operatorname{dom} h$ and h is continuous at $\boldsymbol{A}\boldsymbol{x}$.

(A2) The two functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ are proper, lower semicontinuous, and convex over their respective domains $\mathcal{X}$ and $\mathcal{Y}^*$. Moreover, $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ are both essentially smooth and essentially strictly convex.

(A3) The domains of the two functions g and $\phi_{\mathcal{X}}$ satisfy the inclusion $\operatorname{dom} \partial g \subseteq \operatorname{int}(\operatorname{dom} \phi_{\mathcal{X}})$, and at least one of g and $\phi_{\mathcal{X}}$ is supercoercive.

(A4) The domains of the two functions h^* and $\phi_{\mathcal{Y}^*}$ satisfy the inclusion $\operatorname{dom} \partial h^* \subseteq \operatorname{int}(\operatorname{dom} \phi_{\mathcal{Y}^*})$, and at least one of h^* and $\phi_{\mathcal{Y}^*}$ is supercoercive.

(A5) The two functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ are 1-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}^*}$ on their respective domains.

Assumptions (A1) ensures that the primal problem (2.7) and dual problem (2.8) each has at least one solution [92, Theorem 4.1], and that the saddle-point problem (2.6) has at least one saddle

point [92, Proposition 3.1]. Assumptions (A1)-(A4) ensure that the Bregman divergences of $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ and the minimization problems in the iteration (2.10) satisfy the properties described by Facts 1.2.8 and 1.2.9 in Appendix 1.2. Finally, assumption (A5) is used later in Section 2.3 and 2.4 to prove the convergence of the nonlinear PDHG methods. We note that the domain inclusions in (A3) and (A4) are more restrictive than those assumed in [47] and are necessary for the optimization methods to work (see Fact 1.2.9).

Under assumptions (A1)-(A4) and an appropriate choice of stepsize parameters, initial points, and intermediate points, the iteration scheme (2.10) is well-defined and satisfies a descent rule:

Lemma 2.2.1. *Assume (A1)-(A4) hold, and assume the iteration scheme (2.10) takes as input the stepsize parameters $\tau, \sigma > 0$, initial points $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$, and intermediate points $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}^*) \in \mathcal{X} \times \mathcal{Y}^*$. Then the iteration scheme (2.10) generates a unique output $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$, and for every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ the output $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*)$ satisfies the descent rule*

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}^*) \leqslant & \frac{1}{\tau} \left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \right) \\ & + \frac{1}{\sigma} \left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^*) \right) \\ & + \langle \tilde{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}}) \rangle - \langle \boldsymbol{y}^* - \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}) \rangle \,. \end{aligned} \tag{2.11}$$

Proof. See Appendix 2.A. □

2.3 The basic nonlinear primal-dual hybrid gradient method

The basic nonlinear PDHG method takes two stepsize parameters $\tau, \sigma > 0$ and an initial pair of points $(\boldsymbol{x}_0, \boldsymbol{y}_0^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$ to generate the iterates

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \left\{ g(\boldsymbol{x}) + \langle \boldsymbol{y}_k^*, \boldsymbol{A}\boldsymbol{x} \rangle + \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) \right\} \\ \boldsymbol{y}_{k+1}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^* \in \mathcal{Y}^*} \left\{ -h^*(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}(2\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \rangle - \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \right\}. \end{aligned} \tag{2.12}$$

Under assumptions (A1)-(A4), Lemma 2.2.1 applies to method (2.12), and starting from the input $(\boldsymbol{x}_0, \boldsymbol{y}_0^*)$ the method (2.12) generates a unique output $(\boldsymbol{x}_1, \boldsymbol{y}_1^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$. A simple induction argument using Lemma 2.2.1 then shows that $(\boldsymbol{x}_k, \boldsymbol{y}_k^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$ for every $k \in \mathbb{N}$. As such, method (2.12) is well-defined. In addition, under assumption (A5) and appropriate conditions on the values of the stepsize parameters τ and σ, the nonlinear PDHG method (2.12) satisfies the following properties:

Proposition 2.3.1. *Assume (A1)-(A5) hold and assume $\tau, \sigma > 0$ satisfy the strict inequality*

$$\tau\sigma \left\|\boldsymbol{A}\right\|_{\mathrm{op}}^2 < 1. \tag{2.13}$$

Let $(\boldsymbol{x}_0, \boldsymbol{y}_0^*)$ be a pair of points contained in $\operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$, let $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ be a saddle point of the Lagrangian (2.9), and let $K \in \mathbb{N}$. Consider the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^K$ generated by the nonlinear PDHG method (2.12) from the initial points $(\boldsymbol{x}_0, \boldsymbol{y}_0^*)$, define the averages

$$\boldsymbol{X}_K = \frac{1}{K}\sum_{k=1}^K \boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{Y}_K^* = \frac{1}{K}\sum_{k=1}^K \boldsymbol{y}_k^*,$$

and for $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$, define the quantity

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) = \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) - \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle. \tag{2.14}$$

Then:

(a) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ and nonnegative integer k, the output $(\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}^*)$ of the nonlinear PDHG method (2.12) satisfies the descent rule

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*). \tag{2.15}$$

(b) For every $(\boldsymbol{x}, \boldsymbol{y}^{\boldsymbol{*}}) \in \operatorname{dom} g \times \operatorname{dom} h^*$ *and* $K \in \mathbb{N}$*, we have the estimate*

$$\begin{aligned}\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^{\boldsymbol{*}}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*) \leqslant \frac{1+\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{K}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_0) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{y}_0^*)\right) \\ - \frac{1-\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{K}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_K) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{y}_K^*)\right),\end{aligned} \tag{2.16}$$

and for $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_s, \boldsymbol{y}_s^*)$*, the global bound*

$$\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_K) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_K^*) \leqslant \frac{1+\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{1-\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*)\right). \tag{2.17}$$

(c) [Convergence properties] The sequences $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ *and* $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K^*)\}_{K=1}^{+\infty}$ *are bounded, and the latter has a subsequence that converges weakly to a saddle point of the Lagrangian (2.9). If, in addition, the spaces* $\mathcal{X}$ *and* $\mathcal{Y}^*$ *are finite-dimensional, then the sequences* $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ *and* $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K^*)\}_{K=1}^{+\infty}$ *both converge strongly to the same saddle point.*

Proof. See Appendix 2.B. □

2.4 Accelerated nonlinear primal-dual hybrid gradient methods

In this section, we describe accelerated nonlinear PDHG method (2.12) that are suitable when the functions g and h^* in the saddle-point problem (2.6) have additional structure beyond that stated in assumptions (A1)-(A5). Specifically, we assume either one or both of these statements:

(A6) There is a positive number γ_g such that the function $\boldsymbol{x} \mapsto g(\boldsymbol{x}) - \gamma_g \phi_{\mathcal{X}}(\boldsymbol{x})$ is convex.

(A7) There is a positive number γ_{h^*} such that the function $\boldsymbol{y}^{\boldsymbol{*}} \mapsto h^*(\boldsymbol{y}^{\boldsymbol{*}}) - \gamma_{h^*}\phi_{\mathcal{Y}^*}(\boldsymbol{y}^{\boldsymbol{*}})$ is convex.

Note that by Fact 1.2.5, assumption (A7) is equivalent to assuming that h is also continuously differentiable everywhere on $\mathcal{Y}$, with its gradient being uniformly Lipschitz continuous of parameter $1/gamma_{h^*}$.

With assumptions (A1)-(A7), the descent rule (2.11) in Lemma 2.2.1 is improved: For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$, the output $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*)$ satisfies

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}^*) \leqslant \frac{1}{\tau}\left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}})\right) - \left(\frac{1+\gamma_g\tau}{\tau}\right) D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \\ + \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*)\right) - \left(\frac{1+\gamma_{h^*}\sigma}{\sigma}\right) D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^*) \\ + \langle \tilde{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle - \langle \boldsymbol{y}^* - \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}})\rangle\,.\end{aligned} \tag{2.18}$$

The proof of inequality (2.18) is nearly identical to the proof of inequality (2.11), with the only difference that we use the stronger inequality (1.19) (see Fact 1.2.9(iii)) on each line of the iteration scheme (2.10) to get

$$g(\hat{\boldsymbol{x}}) - g(\boldsymbol{x}) \leqslant \frac{1}{\tau}\left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}})\right) - \left(\frac{1+\gamma_g\tau}{\tau}\right) D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \langle \boldsymbol{A}^*\tilde{\boldsymbol{y}}^*, \boldsymbol{x} - \hat{\boldsymbol{x}}\rangle\,.$$

and

$$h^*(\hat{\boldsymbol{y}}^*) - h^*(\boldsymbol{y}^*) \leqslant \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*)\right)\right) - \left(\frac{1+\gamma_{h^*}\sigma}{\sigma}\right) D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^*) - \langle \boldsymbol{y}^* - \hat{\boldsymbol{y}}^*, \boldsymbol{A}\tilde{\boldsymbol{x}}\rangle\,.$$

Inequality (2.18) then follows from these two inequalities and the same steps used to prove the descent rule (2.11) in Lemma 2.2.1.

Remark 2.4.1. *Assumptions (A1)-(A7) imply that g is γ_g-strongly convex over* $\operatorname{dom} g \cap \operatorname{dom} \phi_{\mathcal{X}}$ *and h^* is γ_{h^*}-strongly convex over* $\operatorname{dom} h^* \cap \phi_{\mathcal{Y}^*}$. *For example,*

$$g(\boldsymbol{x}) - \frac{\gamma_g}{2}\|\boldsymbol{x}\|_{\mathcal{X}}^2 = \left(g(\boldsymbol{x}) - \gamma_g\phi_{\mathcal{X}}(\boldsymbol{x})\right) + \gamma_g\left(\phi_{\mathcal{X}}(\boldsymbol{x}) - \frac{1}{2}\|\boldsymbol{x}\|_{\mathcal{X}}^2\right)$$

for every $\boldsymbol{x} \in \operatorname{dom} g \cap \operatorname{dom} \phi_{\mathcal{X}}$, and as the set $\operatorname{dom} g \cap \operatorname{dom} \phi_{\mathcal{X}}$ *is convex and the right hand side is the sum of two convex functions, the left hand side is also convex.*

Remark 2.4.2. *In light of Remark 2.4.1 and Fact 1.2.6, if assumptions (A1) and (A6) hold, then the primal problem (2.7) has a unique solution. Likewise, if assumptions (A1) and (A7) hold, then the dual problem (2.8) has a unique solution. Finally, if assumptions (A1) and (A6)-(A7) hold, then the Lagrangian (2.9) has a unique saddle point.*

The additional terms in (2.18) allow us to create accelerated methods with better convergence rate than the $O(1/K)$ rate for estimate (2.16). The first accelerated method, which we describe in Section 2.4.1, has a sublinear $O(1/K^2)$ convergence rate and is applicable if assumption (A6) hold. A variant of the first accelerated method, which we describe in Section 2.4.2, has a sublinear $O(1/K^2)$ convergence rate and is applicable if assumption (A7) hold. The second accelerated method, which we describe in Section 2.4.3, has a linear convergence rate and is applicable if both assumptions (A6) and (A7) hold. We also present another variant of this method in Section 2.4.4.

2.4.1 Accelerated nonlinear PDHG methods for strongly convex problems

This accelerated nonlinear PDHG method requires statement (A6) to hold with $\gamma_g > 0$. It takes two parameters $\theta_0 \in [0,1]$ and $\sigma_0 > 0$, a set parameter $\tau_0 = 1/(\|\boldsymbol{A}\|_{\text{op}}^2 \sigma_0)$, an initial point $\boldsymbol{x}_0 \in \text{dom } \partial g$, and the initial points $\boldsymbol{y}_{-1}^* = \boldsymbol{y}_0^* \in \text{dom } \partial h^*$ to generate the iterates

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\min} \left\{ g(\boldsymbol{x}) + \left\langle \boldsymbol{y}_k^* + \theta_k(\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*), \boldsymbol{A}\boldsymbol{x} \right\rangle + \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) \right\}, \\ \boldsymbol{y}_{k+1}^* &= \underset{\boldsymbol{y}^*\in\mathcal{Y}^*}{\arg\max} \left\{ -h^*(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x}_{k+1} \rangle - \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \right\}, \end{aligned} \tag{2.19}$$

where the parameters $\tau_k, \sigma_k, \theta_k$ for $k \in \mathbb{N}$ satisfy the recurrence relations

$$\theta_{k+1} = \frac{1}{\sqrt{1+\gamma_g \tau_k}}, \quad \tau_{k+1} = \theta_{k+1}\tau_k, \quad \text{and} \quad \sigma_{k+1} = \sigma_k/\theta_{k+1}. \tag{2.20}$$

Under assumptions (A1)-(A4), Lemma 2.2.1 applies to method (2.19), and the method generates points $(\boldsymbol{x}_k, \boldsymbol{y}_k^*)$ that are contained in dom $\partial g \times$ dom ∂h^*. As such, method (2.19) is well-defined. If, in addition, assumptions (A5)-(A6) hold, then this method satisfies the following properties:

Proposition 2.4.1. *Assume (A1)-(A6) hold. Let $\theta_0 \in [0,1]$, $\sigma_0 > 0$ and $\tau_0 = 1/(\|\boldsymbol{A}\|_{\text{op}}^2 \sigma_0)$, let $(\boldsymbol{x}_0, \boldsymbol{y}_0^*) \in$ dom $\partial g \times$ dom ∂h^*, let $\boldsymbol{y}_{-1}^* = \boldsymbol{y}_0^*$, and let $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ denote a saddle point of the Lagrangian (2.9). Consider the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^K$ with $K \in \mathbb{N}$ generated by the accelerated nonlinear PDHG method (2.19) and the recurrence relations (2.20) from the initial*

points $(\boldsymbol{x}_0, \boldsymbol{y}_0^*)$ *and* $\boldsymbol{y}_{-1}^*$ *and initial parameters* τ_0, σ_0 *and* θ_0. *Define the averages*

$$T_K = \sum_{k=1}^{K} \frac{\sigma_{k-1}}{\sigma_0}, \quad \boldsymbol{X}_K = \frac{1}{T_K} \sum_{k=1}^{K} \frac{\sigma_{k-1}}{\sigma_0} \boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{Y}_K^* = \frac{1}{T_K} \sum_{k=1}^{K} \frac{\sigma_{k-1}}{\sigma_0} \boldsymbol{y}_k^*$$

and for $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$, *the quantity*

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) = \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) + \theta_k \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle. \tag{2.21}$$

Then:

(a) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ *and nonnegative integer* k, *the output* $(\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}^*)$ *of the accelerated nonlinear PDHG method* (2.19) *satisfies the descent rule*

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta_{k+1}. \tag{2.22}$$

(b) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$, *we have the estimate*

$$T_K \left(\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*) \right) \leqslant \delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{\sigma_K}{\sigma_0} \delta_K(\boldsymbol{x}, \boldsymbol{y}^*) \tag{2.23}$$

and, for the choice of the saddle point $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_s, \boldsymbol{y}_s^*)$, *the global bound*

$$\frac{\gamma_g}{1 + \gamma_g \tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_K) + \frac{1}{\sigma_K} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_K^*) \leqslant \delta_K(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \leqslant \frac{\sigma_0}{\sigma_K} \delta_0(\boldsymbol{x}_s, \boldsymbol{y}_s). \tag{2.24}$$

(c) The average quantity T_K *satisfies the formula*

$$T_K = \|\boldsymbol{A}\|_{\text{op}}^2 \left(\sigma_K^2 - \sigma_0^2 \right) / (\gamma_g \sigma_0) \tag{2.25}$$

and, with $a = \gamma_g / (2 \|\boldsymbol{A}\|_{\text{op}}^2)$, *the bounds*

$$\frac{\sigma_0}{a + \sigma_0} K + \frac{a \sigma_0}{2(a + \sigma_0)^2} K^2 \leqslant T_K \leqslant K + \frac{a}{2\sigma_0} K^2. \tag{2.26}$$

(d) [Convergence properties] The sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ *is bounded and the individual sequence* $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ *converges strongly to the unique solution of the primal problem* (2.7)*. Moreover, the sequence of averages* $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K)\}_{K=1}^{+\infty}$ *is bounded, it has subsequence that converges weakly to a saddle point of the Lagrangian* (2.9)*, and the individual sequence* $\{\boldsymbol{X}_K\}_{K=1}^{+\infty}$ *converges strongly to the unique solution of the primal problem* (2.7)*. If, in addition, the space* $\mathcal{Y}^*$ *is finite-dimensional, then the individual sequences* $\{\boldsymbol{y}_k^*\}_{k=1}^{+\infty}$ *and* $\{\boldsymbol{Y}_k^*\}_{k=1}^{+\infty}$ *each has a subsequence that converges strongly to a solution* $\boldsymbol{y}_s^*$ *of the dual problem* (2.8)*.*

Proof. We divide the proof into five parts, first deriving an auxiliary result, and then proving in turn the descent rule (2.22) (Proposition 2.4.1(a)), the estimate (2.23) and global bound (2.24) (Proposition 2.4.1(b)), formula (2.25) and the bounds (2.26) (Proposition 2.4.1(c)), and the convergence properties of the accelerated nonlinear PDHG method (2.19) (Proposition 2.4.1(d)).

Part 1. We first show that for every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ and $k \in \mathbb{N}$, the quantity $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ satisfies the lower bound

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) \geqslant \frac{\gamma_g}{1 + \gamma_g \tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*). \tag{2.27}$$

To do so, use Fact 1.2.1 with $\alpha = \theta_k / (\tau_k \|\boldsymbol{A}\|_{\mathrm{op}})$ and assumption (A5) to get

$$\left|\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle\right| \leqslant \frac{\theta_k}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{\tau_k \|\boldsymbol{A}\|_{\mathrm{op}}^2}{\theta_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*).$$

By the second and third recurrence relations in (2.20), we have the identity

$$\tau_k \sigma_k \|\boldsymbol{A}\|_{\mathrm{op}}^2 = 1. \tag{2.28}$$

Use this identity in the previous inequality to find

$$\left|\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle\right| \leqslant \frac{\theta_k}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k \theta_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*).$$

Substitute in $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ to get

$$\begin{aligned}\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) &\geqslant \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \\ &\quad - \theta_k \left(\frac{\theta_k}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k \theta_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \right) \\ &= \frac{1 - \theta_k^2}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*).\end{aligned}$$

The first and second recurrence relations in (2.20) imply

$$\frac{1 - \theta_k^2}{\tau_k} = \frac{1}{\tau_k} \left(1 - \frac{1}{1 + \gamma_g \tau_{k-1}} \right) = \frac{1}{\tau_k} \left(\frac{\gamma_g \tau_{k-1}}{1 + \gamma_g \tau_{k-1}} \right) = \frac{1}{\theta_k} \left(\frac{\gamma_g}{1 + \gamma_g \tau_{k-1}} \right) = \frac{\gamma_g}{\theta_k + \gamma_g \tau_k} \geqslant \frac{\gamma_g}{1 + \gamma_g \tau_0}.$$

Hence

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) \geqslant \frac{\gamma_g}{1 + \gamma_g \tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*),$$

which proves the auxiliary result (2.27).

Part 2. Let $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$. By assumption (A1)-(A6), Lemma 2.2.1 holds, and we can apply the improved descent rule (2.18) to the $(k+1)^{\text{th}}$ iterate given by the accelerated nonlinear PDHG method (2.19) with the initial points $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}^*) = (\boldsymbol{x}_k, \boldsymbol{y}_k^*)$, intermediate points $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}^*) = (\boldsymbol{x}_{k+1}, \boldsymbol{y}_k^* + \theta_k(\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*))$, output points $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*) = (\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}^*)$, strong convexity constants $\gamma_g > 0$ and $\gamma_{h^*} = 0$, and parameters $\tau = \tau_k$, $\sigma = \sigma_k$ and $\theta = \theta_k$ to get

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant{}& \frac{1}{\tau_k} \left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) - (1 + \gamma_g \tau_k) D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_{k+1}) \right) \\ &+ \frac{1}{\sigma_k} \left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_{k+1}^*, \boldsymbol{y}_k^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_{k+1}^*) \right) \\ &+ \left\langle \boldsymbol{y}_k^* + \theta_k(\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*) - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle.\end{aligned} \tag{2.29}$$

We wish to bound the last line on the right hand side of (2.29) to eliminate the Bregman divergence term $D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k)$. To do so, first distribute the last line on the right hand side of (2.29) as

$$\theta_k \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle - \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle. \tag{2.30}$$

Write $\boldsymbol{x} - \boldsymbol{x}_{k+1} = (\boldsymbol{x} - \boldsymbol{x}_k) + (\boldsymbol{x}_k - \boldsymbol{x}_{k+1})$ and substitute in (2.30) to get

$$\begin{aligned} \theta_k \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle - \theta_k \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \right\rangle \\ - \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle . \end{aligned} \tag{2.31}$$

Next, use Fact 1.2.1 with $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_{k+1}, \boldsymbol{y}_k^*)$, $(\boldsymbol{x}', \boldsymbol{y}^{*\prime}) = (\boldsymbol{x}_k, \boldsymbol{y}_{k-1}^*)$ and $\alpha = 1/(\theta_k \tau_k \left\|\boldsymbol{A}\right\|_{\text{op}})$, and assumption (A5), identity (2.28) derived in Part 1 to bound the second bilinear form in (2.31) as follows:

$$\left|\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \right\rangle\right| \leqslant \frac{1}{\theta_k \tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + \frac{\theta_k}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*). \tag{2.32}$$

Finally, use (2.30), (2.31), and (2.32) to eliminate the Bregman divergence term $D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k)$ on the right hand side of the descent rule (2.29) and rearrange to get

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) + \left(\frac{1 + \gamma_g \tau_k}{\tau_k}\right) D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_{k+1}) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_{k+1}^*) \\ &\quad + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_{k+1}^*, \boldsymbol{y}_k^*) + \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle \\ &\leqslant \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \\ &\quad + \frac{\theta_k^2}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) + \theta_k \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle \end{aligned} \tag{2.33}$$

We now want to express both sides of inequality (2.33) in terms of $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ and $\delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)$, starting from the left hand side. Note that the recurrence relations (2.20) imply

$$\frac{1 + \gamma_g \tau_k}{\tau_k} = \frac{1}{\theta_{k+1} \tau_{k+1}} \quad \text{and} \quad \frac{1}{\sigma_k} = \frac{1}{\theta_{k+1} \sigma_{k+1}}.$$

As such, the left hand side of (2.33) admits the lower bound

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) + \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta_{k+1}. \tag{2.34}$$

Since $0 \leqslant \theta_k \leqslant 1$, we have

$$\frac{\theta_k^2}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \leqslant \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*).$$

Hence the right hand side of (2.33) is bounded from above by $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$. In summary, we find

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*_{k+1}) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta_{k+1}.$$

This proves the descent rule (2.22).

Part 3. Use (2.22), the third recurrence relation in (2.20), and the averages T_K, $\boldsymbol{X}_K$ and $\boldsymbol{Y}^*_K$ to compute the weighted sum

$$\begin{aligned} T_K\left(\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}^*_K)\right) &\leqslant \sum_{k=1}^{K} \frac{\sigma_{k-1}}{\sigma_0} \left(\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*_k)\right) \\ &\leqslant \sum_{k=1}^{K} \frac{\sigma_{k-1}}{\sigma_0} \left(\delta_{k-1}(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_k(\boldsymbol{x}, \boldsymbol{y}^*)/\theta_k\right) \\ &= \sum_{k=1}^{K} \left(\frac{\sigma_{k-1}}{\sigma_0}\delta_{k-1}(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{1}{\theta_k}\frac{\sigma_{k-1}}{\sigma_0}\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)\right) \\ &= \sum_{k=1}^{K} \left(\frac{\sigma_{k-1}}{\sigma_0}\delta_{k-1}(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{\sigma_k}{\sigma_0}\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)\right) \\ &= \delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{\sigma_K}{\sigma_0}\delta_K(\boldsymbol{x}, \boldsymbol{y}^*). \end{aligned} \tag{2.35}$$

This proves the estimate (2.23). Finally, substitute the saddle point $(\boldsymbol{x}_s, \boldsymbol{y}^*_s)$ for $(\boldsymbol{x}, \boldsymbol{y}^*)$ in inequality (2.35) and use the saddle-point property $\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}^*_s) - \mathcal{L}(\boldsymbol{x}_s, \boldsymbol{y}^*_k) \geqslant 0$ to get

$$\delta_K(\boldsymbol{x}_s, \boldsymbol{y}^*_s) \leqslant \frac{\sigma_0}{\sigma_K}\delta_0(\boldsymbol{x}_s, \boldsymbol{y}_s).$$

The global bound (2.24) follows from this upper bound and the lower bound (2.27) in Part 1.

Part 4. Substitute the identity (2.28) in the third recurrence relation of (2.20) and take the square to get the nonlinear recurrence relation

$$\sigma^2_{k+1} = \sigma^2_k + \gamma_g \sigma_k / \left\|\boldsymbol{A}\right\|^2_{\mathrm{op}}. \tag{2.36}$$

Use this to express the average quantity T_K as a telescoping sum:

$$T_K = \sum_{k=1}^{K} \frac{\sigma_k}{\sigma_0} = \sum_{k=1}^{K} \left(\|\boldsymbol{A}\|_{\text{op}}^2 \left(\sigma_k^2 - \sigma_{k-1}^2\right) / \gamma_g \sigma_0 \right) = \|\boldsymbol{A}\|_{\text{op}}^2 \left(\sigma_K^2 - \sigma_0^2\right) / (\gamma_g \sigma_0).$$

This proves formula (2.25).

We now compute the bounds in (2.26), starting with the upper bound. Let $a = \gamma_g / (2 \|\boldsymbol{A}\|_{\text{op}}^2)$, and use this quantity in equation (2.36) to derive a simple upper bound on σ_K:

$$\begin{aligned} \sigma_K^2 &= \sigma_{K-1}^2 + 2a\sigma_{K-1} \\ &\leqslant \sigma_{K-1}^2 + 2a\sigma_{K-1} + a^2 \\ &= (\sigma_{K-1} + a)^2. \end{aligned}$$

Take the square root to find

$$\sigma_K \leqslant \sigma_{K-1} + a.$$

A simple calculation gives

$$\sigma_K \leqslant \sigma_0 + aK.$$

Hence

$$T_K = \frac{\sigma_K^2 - \sigma_0^2}{2a\sigma_0} \leqslant \frac{(\sigma_0 + aK)^2 - \sigma_0^2}{2a\sigma_0} = K + \frac{a}{2\sigma_0} K^2.$$

which proves the upper bound in inequality (2.26).

The lower bound in inequality (2.26) is the same as derived in Chambolle and Pock [47, Section

5.2, Equation (41)], but here we give a different proof. Use (2.36) to derive a lower bound on σ_K:

$$\begin{aligned}
\sigma_K^2 &= \sigma_{K-1}^2 + 2a\sigma_{K-1} \\
&= \sigma_{K-1}^2 + 2a\sigma_{K-1}\left(\frac{\sigma_0}{a+\sigma_0} + \frac{a}{a+\sigma_0}\right) \\
&= \sigma_{K-1}^2 + \frac{2a\sigma_0\sigma_{K-1}}{a+\sigma_0} + \frac{2a^2\sigma_{K-1}}{a+\sigma_0} \\
&\geqslant \sigma_{K-1}^2 + \frac{2a\sigma_0\sigma_{K-1}}{a+\sigma_0} + \frac{2a^2\sigma_{K-1}}{(a+\sigma_0)}\frac{\sigma_0}{(a+\sigma_0)} \\
&\geqslant \sigma_{K-1}^2 + \frac{2a\sigma_0\sigma_{K-1}}{a+\sigma_0} + \frac{2a^2\sigma_0^2}{(a+\sigma_0)^2} \\
&= \left(\sigma_{K-1} + \frac{a\sigma_0}{a+\sigma_0}\right)^2,
\end{aligned}$$

where on the fifth line we used that $\sigma_k \geqslant \sigma_0$ for every nonnegative integer k, as per the third recurrence relation (2.20). We have found

$$\sigma_K \geqslant \sigma_{K-1} + \frac{a\sigma_0}{a+\sigma_0}$$

which implies, after a simple calculation,

$$\sigma_K \geqslant \sigma_0 + \frac{a\sigma_0}{a+\sigma_0}K.$$

Hence

$$\begin{aligned}
T_K = \frac{\sigma_K^2 - \sigma_0^2}{2a\sigma_0} &\geqslant \frac{\sigma_0^2}{2a\sigma_0}\left[\left(1 + \frac{a}{a+\sigma_0}K\right)^2 - 1\right] \\
&= \frac{\sigma_0}{2a}\left[\frac{2a}{a+\sigma_0}K + \frac{a^2}{(a+\sigma_0)^2}K^2\right] \\
&= \frac{\sigma_0}{a+\sigma_0}K + \frac{a\sigma_0}{2(a+\sigma_0)^2}K^2,
\end{aligned}$$

which proves the lower bound in inequality (2.26).

Part 5. First, combine the auxiliary result (2.27) and global bound (2.24) to get the inequality.

$$\frac{\gamma_g}{1+\gamma_g\tau_0}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_K) + \frac{1}{\sigma_k}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_K^*) \leqslant \frac{1}{\tau_0}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma_0}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*).$$

As a consequence, we have that

$$0 \leqslant \frac{\gamma_g}{1+\gamma_g\tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_K) \leqslant \frac{\sigma_0}{\sigma_K}\left(\frac{1}{\tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma_0} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*)\right) \tag{2.37}$$

and

$$0 \leqslant \frac{1}{\sigma_0} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_K^*) \leqslant \frac{1}{\tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma_0} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*). \tag{2.38}$$

These inequality immediately imply that the sequence of iterates $\{\boldsymbol{x}_K, \boldsymbol{y}_K^*\}_{K=1}^{+\infty}$ is bounded. It follows from the definitions of the averages $\boldsymbol{X}_K$ and $\boldsymbol{Y}_K^*$ that the sequence of averages $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K^*)\}_{K=1}^{+\infty}$ is also bounded.

Now, thanks to Fact 1.2.4 there is a subsequence $\{(\boldsymbol{X}_{K_l}, \boldsymbol{Y}_{K_l}^*)\}_{l=1}^{+\infty}$ that converges weakly to some point $(\boldsymbol{X}, \boldsymbol{Y}^*) \in \mathcal{X} \times \mathcal{Y}^*$. We claim that $(\boldsymbol{X}, \boldsymbol{Y}^*)$ is a saddle point of the Lagrangian (2.9). To see this, use inequality (2.23) with $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ and take the infimum limit $l \to +\infty$ to get

$$\liminf_{l\to+\infty} \mathcal{L}(\boldsymbol{X}_{K_{l+1}}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_{K_{l+1}}^*) \leqslant \liminf_{l\to+\infty} \frac{1}{T_{K_{l+1}}}\left(\frac{1}{\tau_0} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_0) + \frac{1}{\sigma_0} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_0^*)\right) = 0.$$

The lower semicontinuity property of the functions g and h^* implies

$$0 \geqslant \liminf_{l\to+\infty} \left(\mathcal{L}(\boldsymbol{X}_{K_l}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_{K_l}^*)\right) \geqslant \mathcal{L}(\boldsymbol{X}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}^*),$$

from which we find

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{y}^*) \leqslant \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}^*).$$

As the pair of points $(\boldsymbol{x}, \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}^*$ was arbitrary, we conclude that $(\boldsymbol{X}, \boldsymbol{Y}^*)$ is a saddle point of the Lagrangian (2.9). Moreover, we deduce from Remark 2.4.2 that $\boldsymbol{X}$ coincides with the unique solution $\boldsymbol{x}_s$ of the primal problem (2.7), i.e., $\boldsymbol{X} = \boldsymbol{x}_s$.

Next, we show that the individual sequences $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ and $\{\boldsymbol{X}_K\}_{K=1}^{+\infty}$ converge strongly to the unique solution $\boldsymbol{x}_s$ of the primal problem (2.7). The strong convergence of $\boldsymbol{x}_k$ is evident from (2.37),

the limit $\lim_{K\to+\infty}\sigma_K = +\infty$ from (2.25) and (2.26), and assumption (A5):

$$\begin{aligned} 0 \leqslant \lim_{K\to+\infty} \frac{1}{2}\left\|\boldsymbol{x}_s - \boldsymbol{x}_K\right\|_2^2 &\leqslant \lim_{K\to+\infty} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_K) \\ &\leqslant \lim_{K\to+\infty} \left(\frac{1+\gamma_g\tau_0}{\gamma_g}\right)\frac{\sigma_0}{\sigma_K}\left(\frac{1}{\tau_0}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma_0}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*)\right) \\ &= 0. \end{aligned}$$

We further deduce from Fact 1.2.2 that the sequence $\{\boldsymbol{X}_K\}_{K=1}^{+\infty}$ converges strongly to the same limit $\boldsymbol{x}_s$.

Suppose now that $\mathcal{Y}^*$ is finite-dimensional. Since the sequence $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ is bounded and $\boldsymbol{x}_k$ converges strongly to $\boldsymbol{x}_s$, there is some subsequence $\{(\boldsymbol{x}_{k_l}, \boldsymbol{y}_{k_l}^*)\}_{k=1}^{+\infty}$ that converges strongly to a point $(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \in \mathcal{X}\times\mathcal{Y}^*$. By Fact 1.2.2, the subsequence of averages $\{(\boldsymbol{X}_{k_l}, \boldsymbol{Y}_{k_l}^*)\}_{k=1}^{+\infty}$ also strongly converges to $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$. A similar argument as the one described two paragraphs before shows then that $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ is a saddle point of the Lagrangian (2.9), and moreover from Fact 1.2.7 we deduce that $\boldsymbol{y}_s^*$ is a solution to the dual problem (2.8). This concludes the proof.

□

Remark 2.4.3 (Choice of the free parameter σ_0)**.** *The accelerated nonlinear PDHG method* (2.19) *converges at a rate determined by the average quantity T_K, which depends on the stepsize parameter σ_0. One possible choice for σ_0 is to choose it so as to maximize the coefficient multiplying K^2 in the lower bound* (2.26) *of T_K. This coefficient is maximized for the choice of $\sigma_0 = \gamma_g/(2\left\|\boldsymbol{A}\right\|_{\text{op}}^2)$.*

2.4.2 Accelerated nonlinear PDHG methods for smooth convex problems

We present a variant of the first accelerated nonlinear PDHG method. It requires statement (A7) to hold with $\gamma_{h^*} > 0$, and it is similar to method (2.19); it takes two free parameters $\theta_0 \in [0,1]$ and $\tau_0 > 0$, a set parameter $\sigma_0 = 1/(\left\|\boldsymbol{A}\right\|_{\text{op}}^2\tau_0)$, an initial point $\boldsymbol{y}_0^* \in \operatorname{dom}\partial h^*$, and the initial points

$\boldsymbol{x}_{-1} = \boldsymbol{x}_0 \in \operatorname{dom} \partial g$ to generate the iterates

$$\begin{aligned} \boldsymbol{y}_{k+1}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^* \in \mathcal{Y}^*} \left\{ -h^*(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}(\boldsymbol{x}_k + \theta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})) \rangle - \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \right\}, \\ \boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \left\{ g(\boldsymbol{x}) + \left\langle \boldsymbol{y}_{k+1}^*, \boldsymbol{A}\boldsymbol{x} \right\rangle + \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) \right\}, \end{aligned} \tag{2.39}$$

where the parameters $\tau_k, \sigma_k, \theta_k$ for $k \in \mathbb{N}$ satisfy the recurrence relations

$$\theta_{k+1} = \frac{1}{\sqrt{1 + \gamma_{h^*}\sigma_k}}, \quad \tau_{k+1} = \tau_k/\theta_{k+1}, \quad \text{and} \quad \sigma_{k+1} = \theta_{k+1}\sigma_k. \tag{2.40}$$

Under assumptions (A1)-(A4), Lemma 2.2.1 applies to method (2.39), and the method generates points $(\boldsymbol{x}_k, \boldsymbol{y}_k^*)$ that are contained in $\operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$. As such, method (2.39) is well-defined. If, in addition, assumptions (A5) and (A7) hold, then this method satisfies the following properties:

Proposition 2.4.2. *Assume (A1)-(A5) and (A7) hold. Let $\theta_0 \in [0, 1]$, $\tau_0 > 0$ and $\sigma_0 = 1/(\|\boldsymbol{A}\|_{\text{op}}^2 \tau_0)$, let $(\boldsymbol{x}_0, \boldsymbol{y}_0^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$, let $\boldsymbol{x}_{-1}^* = \boldsymbol{x}_0^*$, and let $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ denote a saddle point of the Lagrangian (2.9). Consider the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^K$ with $K \in \mathbb{N}$ generated by the accelerated nonlinear PDHG method (2.39) and the recurrence relations (2.40) from the initial points $(\boldsymbol{x}_0, \boldsymbol{y}_0^*)$ and $\boldsymbol{x}_{-1}$ and initial parameters τ_0, σ_0, and θ_0, and define the averages*

$$T_K = \sum_{k=1}^K \frac{\tau_{k-1}}{\tau_0}, \quad \boldsymbol{X}_K = \frac{1}{T_K} \sum_{k=1}^K \frac{\tau_{k-1}}{\tau_0} \boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{Y}_K^* = \frac{1}{T_K} \sum_{k=1}^K \frac{\tau_{k-1}}{\tau_0} \boldsymbol{y}_k^*$$

and for $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$, the quantity

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) = \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma_k} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) + \frac{1}{\tau_k} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}) + \theta_k \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \right\rangle.$$

Then:

(a) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ and nonnegative integer k, the output $(\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}^*)$ of the accelerated nonlinear PDHG method (2.39) satisfies the descent rule

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta_{k+1}.$$

(b) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$*, we have the estimate*

$$T_K \left(\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*) \right) \leqslant \delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{\tau_K}{\tau_0} \delta_K(\boldsymbol{x}, \boldsymbol{y}^*)$$

and, for the choice of the saddle point $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_s, \boldsymbol{y}_s^*)$*, the global bound*

$$\frac{1}{\tau_K} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_K) + \frac{\gamma_{h^*}}{1 + \gamma_{h^*}\sigma_0} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_K^*) \leqslant \delta_K(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \leqslant \frac{\tau_0}{\tau_K} \delta_0(\boldsymbol{x}_s, \boldsymbol{y}_s).$$

(c) The average quantity T_K *satisfies the formula*

$$T_K = \|\boldsymbol{A}\|_{\mathrm{op}}^2 \, (\tau_K^2 - \tau_0^2)/(\gamma_{h^*}\tau_0)$$

and, with $a = \gamma_{h^*}/(2\,\|\boldsymbol{A}\|_{\mathrm{op}}^2)$*, the bounds*

$$\frac{\tau_0}{a + \tau_0} K + \frac{a\tau_0}{2(a + \tau_0)^2} K^2 \leqslant T_K \leqslant K + \frac{a}{2\tau_0} K^2.$$

(d) [Convergence properties] The sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ *is bounded and the individual sequence* $\{\boldsymbol{y}_k^*\}_{k=1}^{+\infty}$ *converges strongly to the unique solution of the dual problem (2.8). Moreover, the sequence of averages* $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K)\}_{K=1}^{+\infty}$ *is bounded, it has subsequence that converges weakly to a saddle point of the Lagrangian (2.9), and the individual sequence* $\{\boldsymbol{Y}_K^*\}_{K=1}^{+\infty}$ *converges strongly to the unique solution of the dual problem (2.8). If, in addition, the space* $\mathcal{X}$ *is finite-dimensional, then the individual sequences* $\{\boldsymbol{x}_k\}_{k=1}^{+\infty}$ *and* $\{\boldsymbol{X}_k\}_{k=1}^{+\infty}$ *each has a subsequence that converges strongly to a solution* $\boldsymbol{x}_s$ *of the primal problem (2.8).*

Proof. The proof is essentially the same as for Proposition 2.4.1 and is omitted. □

2.4.3 Accelerated nonlinear PDHG method for smooth and strongly convex problems I

The second accelerated nonlinear PDHG method requires statements (A6) and (A7) to hold with $\gamma_g > 0$ and $\gamma_{h^*} > 0$. It takes the parameters

$$\theta = 1 - \frac{\gamma_g \gamma_{h^*}}{2\left\|\boldsymbol{A}\right\|_{\text{op}}^2}\left(\sqrt{1 + \frac{4\left\|\boldsymbol{A}\right\|_{\text{op}}^2}{\gamma_g \gamma_{h^*}}} - 1\right), \quad \tau = \frac{1-\theta}{\gamma_g \theta}, \quad \text{and} \quad \sigma = \frac{1-\theta}{\gamma_{h^*}\theta}, \tag{2.41}$$

an initial point $\boldsymbol{x}_0 \in \operatorname{dom} \partial g$, and the initial points $\boldsymbol{y}_{-1}^* = \boldsymbol{y}_0^* \in \operatorname{dom} \partial h^*$ to generate the iterates

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \left\{ g(\boldsymbol{x}) + \left\langle \boldsymbol{y}_k^* + \theta(\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*), \boldsymbol{A}\boldsymbol{x} \right\rangle + \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) \right\}, \\ \boldsymbol{y}_{k+1}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^*\in\mathcal{Y}^*} \left\{ -h^*(\boldsymbol{y}^*) + \left\langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x}_{k+1} \right\rangle - \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \right\}. \end{aligned} \tag{2.42}$$

Under assumptions (A1)-(A4), Lemma 2.2.1 applies to method (2.42), and the method generates points $(\boldsymbol{x}_k, \boldsymbol{y}_k^*)$ that are contained in $\operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$. As such, method (2.42) is well-defined. If, in addition, assumptions (A5)-(A7) hold, then this method satisfies the following properties:

Proposition 2.4.3. *Assume (A1)-(A7) hold. Let $(\boldsymbol{x}_0, \boldsymbol{y}_0^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$, let $\boldsymbol{y}_{-1}^* = \boldsymbol{y}_0^*$, and let $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ denote the unique saddle point of the Lagrangian (2.9). Consider the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^K$ with $K \in \mathbb{N}$ generated by the accelerated nonlinear PDHG method (2.42) from the initial points $\boldsymbol{x}_0$, $\boldsymbol{y}_0^*$ and $\boldsymbol{y}_{-1}^*$, and the parameters θ, τ, and σ defined in (2.41). Define the averages*

$$T_K = \sum_{k=1}^K \frac{1}{\theta^{k-1}} = \frac{1-\theta^K}{(1-\theta)\theta^{K-1}}, \quad \boldsymbol{X}_K = \frac{1}{T_K}\sum_{k=1}^K \frac{1}{\theta^{k-1}}\boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{Y}_K^* = \frac{1}{T_K}\sum_{k=1}^K \frac{1}{\theta^{k-1}}\boldsymbol{y}_k^*,$$

and for $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$, the quantity

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) = \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) + \frac{\theta}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) + \theta\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle. \tag{2.43}$$

Then:

(a) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ *and nonnegative integer* k*, the output* $(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*_{k+1})$ *of the accelerated nonlinear PDHG method* (2.42) *satisfies the descent rule*

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*_{k+1}) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta. \tag{2.44}$$

(b) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$*, we have the estimate*

$$T_K \left(\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}^*_K)\right) \leqslant \delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{1}{\theta^K}\delta_K(\boldsymbol{x}, \boldsymbol{y}^*) \tag{2.45}$$

and, for the choice of the saddle point $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_s, \boldsymbol{y}^*_s)$*, the global bound*

$$\frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*_s, \boldsymbol{y}^*_K) \leqslant \delta_K(\boldsymbol{x}_s, \boldsymbol{y}^*_s) \leqslant \theta^K \delta_0(\boldsymbol{x}_s, \boldsymbol{y}^*_s). \tag{2.46}$$

(c) [Convergence properties] The sequences $\{(\boldsymbol{x}_k, \boldsymbol{y}^*_k)\}_{k=1}^{+\infty}$ *and* $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K)\}_{K=1}^{+\infty}$ *both converge strongly to the unique saddle point* $(\boldsymbol{x}_s, \boldsymbol{y}^*_s)$ *of the Lagrangian* (2.9)*.*

Proof. We divide the proof into four parts, first deriving an auxiliary result, and then proving in turn the descent rule (2.44) (Proposition 2.4.3(a)), the estimate (2.45) and global bound (2.46) (Proposition 2.4.3(b)), and the convergence properties of the accelerated nonlinear PDHG method (2.42) (Proposition 2.4.3(c)).

Part 1. First, we show that for every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ and $k \in \mathbb{N}$, the quantity $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ satisfies the lower bound

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) \geqslant \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}^*_k). \tag{2.47}$$

To do so, use Fact 1.2.1 with $\alpha = 1/(\tau\theta \left\|\boldsymbol{A}\right\|_{\mathrm{op}})$ and use assumption (A5) to find

$$\left|\left\langle \boldsymbol{y}^*_k - \boldsymbol{y}^*_{k-1}, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k)\right\rangle\right| \leqslant \frac{1}{\tau\theta} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \tau\theta \left\|\boldsymbol{A}\right\|^2_{\mathrm{op}} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*_k, \boldsymbol{y}^*_{k-1}).$$

From the choice of parameters in (2.41), we have the identity

$$\tau\sigma\theta \left\|\boldsymbol{A}\right\|_{\mathrm{op}}^2 = 1. \tag{2.48}$$

Use this identity in the previous inequality to find

$$\left|\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k)\right\rangle\right| \leqslant \frac{1}{\tau\theta} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*). \tag{2.49}$$

Substitute in $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ to get

$$\begin{aligned}
\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) &\geqslant \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) + \frac{\theta}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \\
&\quad - \theta \left(\frac{1}{\tau\theta} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \right) \\
&= \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*).
\end{aligned}$$

This proves the auxiliary result (2.47).

Part 2. Let $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$. By assumptions (A1)-(A7), Lemma 2.2.1 holds, and we can apply the improved descent rule (2.18) to the $(k+1)^{\text{th}}$ iterate given by the accelerated nonlinear PDHG method (2.42) with the initial points $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}^*) = (\boldsymbol{x}_k, \boldsymbol{y}_k^*)$, intermediate points $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}^*) = (\boldsymbol{x}_{k+1}, \boldsymbol{y}_k^* + \theta(\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*))$, output points $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*) = (\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}^*)$, the strong convexity constants $\gamma_g > 0$, $\gamma_{h^*} > 0$, and the parameters τ, σ, and θ defined in (2.41):

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) &\leqslant \frac{1}{\tau} \left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) - (1 + \gamma_g\tau) D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_{k+1}) \right) \\
&\quad + \frac{1}{\sigma} \left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_{k+1}^*, \boldsymbol{y}_k^*) - (1 + \gamma_{h^*}\sigma) D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_{k+1}^*) \right) \\
&\quad + \left\langle \boldsymbol{y}_k^* + \theta(\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*) - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle.
\end{aligned} \tag{2.50}$$

We wish to bound the last line on the right hand side of (2.50) to eliminate the Bregman divergence term $D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k)$. To do so, first distribute the last line on the right hand side of (2.50) as

$$\theta \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle - \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle. \tag{2.51}$$

Write $\boldsymbol{x} - \boldsymbol{x}_{k+1} = (\boldsymbol{x} - \boldsymbol{x}_k) + (\boldsymbol{x}_k - \boldsymbol{x}_{k+1})$ and substitute in (2.51) to get

$$\begin{aligned} \theta \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle - \theta \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \right\rangle \\ - \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle . \end{aligned} \tag{2.52}$$

Next, use inequality (2.49) with $\boldsymbol{x} = \boldsymbol{x}_{k+1}$ derived in Part 1 to bound the second bilinear form in (2.52) as follows:

$$\left| \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \right\rangle \right| \leqslant \frac{1}{\tau\theta} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \tag{2.53}$$

Finally, use (2.51), (2.52), and (2.53) to eliminate the Bregman divergence term $D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k)$ on the right hand side of the descent rule (2.50):

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) + \left(\frac{1 + \gamma_g \tau}{\tau} \right) D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_{k+1}) + \left(\frac{1 + \gamma_{h^*} \sigma}{\sigma} \right) D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_{k+1}^*) \\ &\quad + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_{k+1}^*, \boldsymbol{y}_k^*) + \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) \right\rangle \\ &\leqslant \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \\ &\quad + \frac{\theta}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) + \theta \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) \right\rangle \end{aligned} \tag{2.54}$$

We now want to express both sides of inequality (2.54) in terms of $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ and $\delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)$, starting from the left hand side. Note that the choice of parameters in (2.41) implies

$$1 + \gamma_g \tau = 1/\theta \quad \text{and} \quad 1 + \gamma_{h^*} \sigma = 1/\theta.$$

As such, the left hand side of (2.54) is equal to

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) + \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta,$$

and the right hand side of (2.54) is equal to $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$. Put together, we find

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta. \tag{2.55}$$

This proves the descent rule (2.44).

Part 3. Use (2.55) and the averages T_K, $\boldsymbol{X}_K$ and $\boldsymbol{Y}_K^*$ to compute the sum

$$\begin{aligned} T_K\mathcal{L}(\boldsymbol{X}_K,\boldsymbol{y}^*)-\mathcal{L}(\boldsymbol{x},\boldsymbol{Y}_K^*) &= \sum_{k=1}^{K}\frac{1}{\theta^{k-1}}\left(\mathcal{L}(\boldsymbol{x}_k,\boldsymbol{y}^*)-\mathcal{L}(\boldsymbol{x},\boldsymbol{y}_k^*)\right)\\ &\leqslant \sum_{k=1}^{K}\frac{1}{\theta^{k-1}}\left(\delta_{k-1}(\boldsymbol{x},\boldsymbol{y}^*)-\delta_k(\boldsymbol{x},\boldsymbol{y}^*)/\theta\right)\\ &= \sum_{k=1}^{K}\left(\delta_{k-1}(\boldsymbol{x},\boldsymbol{y}^*)/\theta^{k-1}-\delta_k(\boldsymbol{x},\boldsymbol{y}^*)/\theta^k\right)\\ &= \delta_0(\boldsymbol{x},\boldsymbol{y}^*)-\delta_K(\boldsymbol{x},\boldsymbol{y}^*)/\theta^K. \end{aligned} \tag{2.56}$$

This proves the estimate (2.45). Finally, substitute the saddle point $(\boldsymbol{x}_s,\boldsymbol{y}_s^*)$ for $(\boldsymbol{x},\boldsymbol{y}^*)$ in inequality (2.56) and use the saddle-point property $\mathcal{L}(\boldsymbol{x}_k,\boldsymbol{y}_s^*)-\mathcal{L}(\boldsymbol{x}_s,\boldsymbol{y}_k^*)\geqslant 0$ to get

$$\delta_K(\boldsymbol{x}_s,\boldsymbol{y}_s^*)\leqslant\theta^K\delta_0(\boldsymbol{x}_s,\boldsymbol{y}_s^*).$$

The global bound (2.46) follows from this upper bound and the lower bound (2.47) derived in Part 1.

Part 4. The global bound (2.46), assumption (A5), and Fact 1.2.8(viii) immediately imply that the sequence of iterates $\{\boldsymbol{y}_k^*\}_{k=1}^{+\infty}$ converges strongly to $\boldsymbol{y}_s$. It follows from this and Fact 1.2.2 that the sequence of averages $\{\boldsymbol{Y}_K^*\}_{K=1}^{+\infty}$ also converges strongly to $\boldsymbol{y}_s^*$.

Now, consider inequality (2.56) with $(\boldsymbol{x},\boldsymbol{y}^*)=(\boldsymbol{x}_s,\boldsymbol{y}_s^*)$ written in full:

$$\begin{aligned} \frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s,\boldsymbol{x}_k)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*,\boldsymbol{y}_k^*)+\frac{\theta}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*,\boldsymbol{y}_{k-1}^*)+\theta\left\langle\boldsymbol{y}_k^*-\boldsymbol{y}_{k-1}^*,\boldsymbol{A}(\boldsymbol{x}_s-\boldsymbol{x}_k)\right\rangle\\ \leqslant\theta^K\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s,\boldsymbol{x}_0)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*,\boldsymbol{y}_0^*)\right). \end{aligned} \tag{2.57}$$

We wish to bound the bilinear form on the left hand side to obtain a bound on $\boldsymbol{x}_k$. To do so, use

Fact 1.2.1 with $\alpha = 1/(2\tau\theta \|\boldsymbol{A}\|_{\text{op}})$ and identity (2.48) to obtain the bound

$$\left|\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*, \boldsymbol{A}(\boldsymbol{x}_s - \boldsymbol{x}_k)\right\rangle\right| \leqslant \frac{1}{4\tau\theta}\|\boldsymbol{x}_s - \boldsymbol{x}_k\|_{\mathcal{X}}^2 + \frac{1}{\sigma}\left\|\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*\right\|_{\mathcal{Y}^*}^2.$$

Substitute in (2.57) to get

$$\begin{aligned}\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_k) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_k^*) + \frac{\theta}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) - \frac{\theta}{\sigma}\left\|\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*\right\|_{\mathcal{Y}^*}^2 - \frac{1}{4\tau}\|\boldsymbol{x}_s - \boldsymbol{x}_k\|_{\mathcal{X}}^2 \\ \leqslant \theta^K\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*)\right).\end{aligned}$$

Finally, use the inequalities $D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_k^*) \geqslant 0$, $D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_k^*, \boldsymbol{y}_{k-1}^*) \geqslant 0$, and $D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_k) \geqslant \frac{1}{2}\|\boldsymbol{x}_s - \boldsymbol{x}_k\|_{\mathcal{X}}^2$ (thanks to assumption (A5) and Fact 1.2.8(viii) with $m = 1$) to obtain

$$\frac{1}{4\tau}\|\boldsymbol{x}_s - \boldsymbol{x}_k\|_{\mathcal{X}}^2 \leqslant \frac{\theta}{\sigma}\left\|\boldsymbol{y}_k^* - \boldsymbol{y}_{k-1}^*\right\|_{\mathcal{Y}^*}^2 + \theta^K\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_0) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*, \boldsymbol{y}_0^*)\right).$$

Taking the limit $k \to +\infty$ yields $\lim_{k\to+\infty}\boldsymbol{x}_k = \boldsymbol{x}_s$. It follows from this and Fact 1.2.2 that the sequence of averages $\{\boldsymbol{X}_K\}_{K=1}^{+\infty}$ also converges strongly to $\boldsymbol{x}_s$. This concludes the proof. □

2.4.4 Accelerated nonlinear PDHG method for smooth and strongly convex problems II

We present a variant of the accelerated nonlinear PDHG method (2.42). It requires statements (A6) and (A7) to hold with $\gamma_g > 0$ and $\gamma_{h^*} > 0$ and it takes the parameters

$$\theta = 1 - \frac{\gamma_g\gamma_{h^*}}{2\|\boldsymbol{A}\|_{\text{op}}^2}\left(\sqrt{1 + \frac{4\|\boldsymbol{A}\|_{\text{op}}^2}{\gamma_g\gamma_{h^*}}} - 1\right), \quad \tau = \frac{1-\theta}{\gamma_g\theta}, \quad \text{and} \quad \sigma = \frac{1-\theta}{\gamma_{h^*}\theta}, \tag{2.58}$$

the initial points $\boldsymbol{x}_{-1} = \boldsymbol{x}_0 \in \operatorname{dom}\partial g$ and an initial point $\boldsymbol{y}_0^* \in \operatorname{dom}\partial h^*$ to generate the iterates

$$\begin{aligned}\boldsymbol{y}_{k+1}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^*\in\mathcal{Y}^*}\left\{-h^*(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}(\boldsymbol{x}_k + \theta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}))\rangle - \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*)\right\} \\ \boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}}\left\{g(\boldsymbol{x}) + \left\langle \boldsymbol{y}_{k+1}^*, \boldsymbol{A}\boldsymbol{x}\right\rangle + \frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k)\right\}.\end{aligned} \tag{2.59}$$

method (2.59) and (2.42) differ only in that we update $\boldsymbol{y}^*$ first. This change nonetheless yields different global bounds and convergence estimates. Under assumptions (A1)-(A4), Lemma 2.2.1 applies to method (2.59), and the method generates points $(\boldsymbol{x}_k, \boldsymbol{y}_k^*)$ that are contained in dom $\partial g \times$ dom ∂h^*. As such, method (2.59) is well-defined. If, in addition, assumptions (A5)-(A7) hold, then this method satisfies the following properties:

Proposition 2.4.4. *Assume (A1)-(A7) hold. Let* $(\boldsymbol{x}_0, \boldsymbol{y}_0^*) \in \text{dom } \partial g \times \text{dom } \partial h^*$, *let* $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$, *and let* $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ *denote the unique saddle point of the Lagrangian* (2.9). *Consider the sequence of iterates* $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^K$ *with* $K \in \mathbb{N}$ *generated by the accelerated nonlinear PDHG method* (2.59) *from the initial points* $\boldsymbol{x}_0$, $\boldsymbol{y}_0^*$ *and* $\boldsymbol{x}_{-1}^*$, *and the parameters* θ, τ, *and* σ *defined in* (2.41). *Define the averages*

$$T_K = \sum_{k=1}^{K} \frac{1}{\theta^{k-1}} = \frac{1-\theta^K}{(1-\theta)\theta^{K-1}}, \quad \boldsymbol{X}_K = \frac{1}{T_K} \sum_{k=1}^{K} \frac{1}{\theta^{k-1}} \boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{Y}_K^* = \frac{1}{T_K} \sum_{k=1}^{K} \frac{1}{\theta^{k-1}} \boldsymbol{y}_k^*,$$

and for $(\boldsymbol{x}, \boldsymbol{y}^*) \in \text{dom } g \times \text{dom } h^*$, *the quantity*

$$\delta_k(\boldsymbol{x}, \boldsymbol{y}^*) = \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) + \frac{\theta}{\sigma} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_k^*, \boldsymbol{x}_{k-1}^*) + \theta \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \right\rangle.$$

Then:

(a) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \text{dom } g \times \text{dom } h^*$ *and nonnegative integer* k, *the output* $(\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}^*)$ *of the accelerated nonlinear PDHG method* (2.42) *satisfies the descent rule*

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*)/\theta.$$

(b) For every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \text{dom } g \times \text{dom } h^*$, *we have the estimate*

$$T_K \left(\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*) \right) \leqslant \delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \frac{1}{\theta^K} \delta_K(\boldsymbol{x}, \boldsymbol{y}^*)$$

and, for the choice of the saddle point $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_s, \boldsymbol{y}_s^*)$, *the global bound*

$$\frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s, \boldsymbol{x}_K) \leqslant \delta_K(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \leqslant \theta^K \delta_0(\boldsymbol{x}_s, \boldsymbol{y}_s^*).$$

(c) [Convergence properties] The sequences $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ *and* $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K)\}_{K=1}^{+\infty}$ *both converge strongly to the unique saddle point* $(\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ *of the Lagrangian* (2.9).

Proof. The proof is essentially the same as for Proposition 2.4.3 and is omitted. □

2.5 Connections between supervised machine learning algorithms and Hamilton–Jacobi PDEs

This section presents some novel connections between a broad class of supervised machine learning algorithms and first-order Hamilton–Jacobi partial differential equations. It is meant to be illustrative and is not exhaustive. In addition, we will only consider classical solutions of first-order HJ PDEs with initial data.

Our starting point are the primal and dual problems (2.7) defined on the real vector spaces $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$. We present two approaches for connecting supervised machine learning algorithms with classical solutions to first-order HJ PDEs with initial data. In the first approach, we will express the primal problem (2.7) in such a way that it will correspond to the Lax–Oleinik representation formula of the solution to an appropriate first-order HJ PDE. In the second approach, we will express the primal problem (2.7) in such a way that it will correspond to the Hopf representation formula of the solution to an appropriate first-order HJ PDE. The first approach will require us to impose strong conditions on the function h in the primal problem, while the second approach will require us to impose strong conditions on the function g in the primal problem. These approaches will lead to connections to constrained-type problems in machine learning, regularized maximum entropy estimation problems, and Maximum-likelihood type estimators of generalized linear models in statistics.

First approach: strong assumptions on the initial data

We start by imposing certain conditions on the functions g and h as well as the matrix $\boldsymbol{A}$ in the primal problem (2.7). First, for any $t > 0$ and $\boldsymbol{x}' \in \mathbb{R}^n$ we set

$$g(\boldsymbol{x}) = tH^*\left(\frac{\boldsymbol{x}' - \boldsymbol{x}}{t}\right)$$

for some supercoercive function H^* defined on $\mathbb{R}^n$. Its convex conjugate is denoted by H and is called the Hamiltonian. The proper, lower semicontinuous, convex, and supercoercive properties of H^* are equivalent to requiring the Hamiltonian H to be a convex function defined on $\mathbb{R}^n$ [215, Theorem 11.8]. Next, we suppose that h is convex, uniformly Lipschitz continuous function on $\mathbb{R}^m$, and twice continuously differentiable on $\mathbb{R}^m$ with uniformly bounded second partial derivatives. Finally, we assume that the matrix $\boldsymbol{A}$ has full row rank, that is, $\boldsymbol{A}\colon \mathbb{R}^n \to \mathbb{R}^m$ is surjective. We also define the function $J\colon \mathbb{R}^n \to \mathbb{R}$ through h by $J(\boldsymbol{x}) = h(\boldsymbol{A}\boldsymbol{x})$.

Now, let $S\colon \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}$ denote the value of the primal problem (2.7) parametrized in terms of $\boldsymbol{x}'$ and t. Under the assumptions above, $S(\boldsymbol{x}', t)$ can be expressed as follows:

$$\begin{aligned} S(\boldsymbol{x}', t) &= \inf_{\boldsymbol{x}\in\mathbb{R}^n} \{g(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})\} \\ &= \inf_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ tH^*\left(\frac{\boldsymbol{x}' - \boldsymbol{x}}{t}\right) + J(\boldsymbol{x}) \right\}. \end{aligned} \tag{2.60}$$

Moreover, this function is the classical solution to the first-order HJ PDE [16, Proposition 4.1]

$$\begin{cases} \dfrac{\partial S}{\partial t}(\boldsymbol{x}', t) + H\left(\nabla_{\boldsymbol{x}'} S(\boldsymbol{x}', t)\right) = 0, & \boldsymbol{x}' \in \mathbb{R}^n,\ t \in [0, +\infty) \\ S(\boldsymbol{x}', 0) = J(\boldsymbol{x}'), & \boldsymbol{x}' \in \mathbb{R}^n. \end{cases} \tag{2.61}$$

The Hamiltonian and initial data of the HJ PDE are the functions H and J. The second line in (2.60) is the Lax–Oleinik representation formula of the solution to the HJ PDE (2.61). The

solution $S(\boldsymbol{x}', t)$ can also be represented using the dual problem (2.8). We have

$$\begin{aligned} S(\boldsymbol{x}', t) &= \sup_{\boldsymbol{y}^* \in \mathbb{R}^m} \{-g^*(-\boldsymbol{A}^* \boldsymbol{y}^*) - h^*(\boldsymbol{y}^*)\} \\ &= \sup_{\boldsymbol{y}^* \in \mathbb{R}^m} \left\{ \left\langle \boldsymbol{A}^* \boldsymbol{y}^*, \boldsymbol{x}' \right\rangle - tH(\boldsymbol{A}^* \boldsymbol{y}^*) - h^*(\boldsymbol{y}^*) \right\}. \end{aligned} \tag{2.62}$$

The second equality in (2.62) is the Hopf representation formula of the solution to the HJ PDE (2.61), although written in a slightly unconventional form. It can be written in conventional form using the convex conjugate of the function J [137, Page 56, Theorem 2.2.1]:

$$J^*(\boldsymbol{x}^*) = \inf_{\substack{\boldsymbol{y}^* \in \mathbb{R}^m \\ \boldsymbol{A}^* \boldsymbol{y}^* = \boldsymbol{x}^*}} h^*(\boldsymbol{y}^*). \tag{2.63}$$

Since the matrix $\boldsymbol{A}$ has full row rank, its transpose $\boldsymbol{A}^*$ has full column rank. The mapping $\boldsymbol{A}^* \colon \mathbb{R}^m \to \mathbb{R}^n$ is therefore injective, and hence one may replace the supremum taken over $\boldsymbol{y}^* \in \mathbb{R}^m$ in (2.62) with the supremum taken over $\boldsymbol{x}^* \in \mathbb{R}^n$ under the constraint that $\boldsymbol{A}^* \boldsymbol{y}^* = \boldsymbol{x}^*$. That is to say, we can write

$$S(\boldsymbol{x}', t) = \sup_{\boldsymbol{x}^* \in \mathbb{R}^n} \left\{ \left\langle \boldsymbol{x}^*, \boldsymbol{x}' \right\rangle - tH(\boldsymbol{x}^*) - J^*(\boldsymbol{x}^*) \right\},$$

which is the Hopf representation formula of the solution to the HJ PDE (2.62).

Examples of supervised machine learning algorithms. Different choices of Hamiltonians and initial data yield different supervised machine learning algorithms. Here, the Hamiltonian H plays the role of the convex function determining the constraints or plays the role of the regularization term, the initial data J plays the role of the regression model or loss function, the matrix $\boldsymbol{A}$ encodes the data for the problem at hand, and the parameter $t > 0$ determines the constraint or the relative importance between the Hamiltonian and loss function in the optimization problem. For example, the choice

$$H(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty \quad \text{and} \quad J(\boldsymbol{x}) = h(\boldsymbol{A}\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-[\boldsymbol{A}\boldsymbol{x}]_i}\right)$$

with $\boldsymbol{x}' = \mathbf{0}$ and appropriate matrix $\boldsymbol{A}$ yields the ℓ_1-constrained logistic regression problem (2.1) considered in the introduction. The same Hamiltonian with initial data

$$J(\boldsymbol{x}) = h(\boldsymbol{A}\boldsymbol{x}) = \frac{1}{2m} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$$

with $\boldsymbol{x}' = \mathbf{0}$, $\boldsymbol{b} \in \mathbb{R}^m$ and appropriate matrix $\boldsymbol{A}$ yields ℓ_1-constrained least squared regression, which is also known as the constrained form of the Lasso in the machine learning literature [234]. The same initial data with quadratic Hamiltonian $H(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ and with $\boldsymbol{x}' = \mathbf{0}$ yield ℓ_2^2-regularized logistic and linear regression, respectively.

Second approach: strong assumptions on the Hamiltonian

We start by imposing certain conditions on the functions g and h as well as the matrix $\boldsymbol{A}$ in the primal problem (2.7). First, for any $t > 0$ we set

$$g(\boldsymbol{x}) = tH(\boldsymbol{x})$$

for some strictly convex and differentiable function H on $\mathbb{R}^n$. This function is, as before, called the Hamiltonian. Next, we suppose that

$$h(\boldsymbol{y}) = f(\boldsymbol{y}) - \langle \boldsymbol{b}, \boldsymbol{y} \rangle$$

for some $\boldsymbol{b} \in \mathbb{R}^m$ and $f \in \Gamma_0(\mathbb{R}^m)$. We also assume that the matrix $\boldsymbol{A}$ has full row rank, that is, $\boldsymbol{A}\colon \mathbb{R}^n \to \mathbb{R}^m$ is surjective. These conditions ensure that the function $J^*\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by $J^*(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x})$ is itself a proper, lower semicontinuous and convex function [137, Page 56, Theorem 2.2.1]. Finally, we assume that either a) the Hamiltonian H is supercoercive or that b) H^* is differentiable at any point where a subgradient exists, H is bounded from below by a constant, and the function f in h is coercive.

Now, let $\boldsymbol{p} = \boldsymbol{A}^*\boldsymbol{b}$, let J denote the convex conjugate of J^*, and let $S\colon \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}$

denote the negative value of the primal problem (2.7) parametrized in terms of $\boldsymbol{p}$ and $t > 0$. Under the assumptions above, $S(\boldsymbol{p}, t)$ can be expressed as

$$\begin{aligned} S(\boldsymbol{p}, t) &= -\inf_{\boldsymbol{x}\in\mathbb{R}^n} \{g(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})\} \\ &= -\inf_{\boldsymbol{x}\in\mathbb{R}^n} \{tH(\boldsymbol{x}) + f(\boldsymbol{A}\boldsymbol{x}) - \langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x}\rangle\} \\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x}\rangle - tH(\boldsymbol{x}) - f(\boldsymbol{A}\boldsymbol{x})\} \\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\langle \boldsymbol{p}, \boldsymbol{x}\rangle - tH(\boldsymbol{x}) - J^*(\boldsymbol{x})\}. \end{aligned} \tag{2.64}$$

Moreover, this function is the classical solution to the first-order HJ PDE [64, Lemma 2.1 and Theorem 2.6]

$$\begin{cases} \dfrac{\partial S}{\partial t}(\boldsymbol{p}, t) + H\left(\nabla_{\boldsymbol{p}} S(\boldsymbol{p}, t)\right) = 0, & \boldsymbol{p} \in \mathbb{R}^n,\ t \in [0, +\infty), \\ S(\boldsymbol{p}, 0) = J(\boldsymbol{p}), & \boldsymbol{p} \in \mathbb{R}^n, \end{cases} \tag{2.65}$$

The Hamiltonian and initial data of the HJ PDE are the functions H and J. The fourth line in (2.64) is the Hopf representation formula of the solution to the HJ PDE (2.65). The solution $S(\boldsymbol{p}, t)$ can also be represented using the negative value of the dual problem (2.8). Using the identity

$$\begin{aligned} h^*(\boldsymbol{y}^*) &= \sup_{\boldsymbol{y}\in\mathbb{R}^m} \{\langle \boldsymbol{y}^*, \boldsymbol{y}\rangle - f(\boldsymbol{y}) + \langle \boldsymbol{b}, \boldsymbol{y}\rangle\} \\ h^*(\boldsymbol{y}^*) &= f^*(\boldsymbol{b} + \boldsymbol{y}^*), \end{aligned}$$

we have

$$\begin{aligned} S(\boldsymbol{p}, t) &= -\sup_{\boldsymbol{y}^*\in\mathbb{R}^m} \{-g^*(-\boldsymbol{A}^*\boldsymbol{y}^*) - h^*(\boldsymbol{y}^*)\} \\ &= \inf_{\boldsymbol{y}^*\in\mathbb{R}^m} \{g^*(-\boldsymbol{A}^*\boldsymbol{y}^*) + h^*(\boldsymbol{y}^*)\} \\ &= \inf_{\boldsymbol{y}^*\in\mathbb{R}^m} \left\{tH^*\left(\frac{\boldsymbol{p} - \boldsymbol{A}^*\boldsymbol{y}^*}{t}\right) + f^*(\boldsymbol{y}^*)\right\}. \end{aligned} \tag{2.66}$$

The third equality in (2.66) is the Lax–Oleinik representation formula of the solution to the HJ PDE (2.65), although written in a slightly unconventional form. It can be written in conventional form using the convex conjugate of the function J [137, Theorem 2.2.1]:

$$J(\boldsymbol{x}^*) = \inf_{\substack{\boldsymbol{y}^*\in\mathbb{R}^m \\ \boldsymbol{A}^*\boldsymbol{y}^* = \boldsymbol{x}^*}} f^*(\boldsymbol{y}^*). \tag{2.67}$$

Since the matrix $\boldsymbol{A}$ has full row rank, its transpose $\boldsymbol{A}^*$ has full column rank. The mapping $\boldsymbol{A}^*\colon \mathbb{R}^m \to \mathbb{R}^n$ is therefore injective, and hence one may replace the infimum taken over $\boldsymbol{y}^* \in \mathbb{R}^m$ in (2.66) with the infimum taken over $\boldsymbol{x}^* \in \mathbb{R}^n$ under the constraint that $\boldsymbol{A}^*\boldsymbol{y}^* = \boldsymbol{x}^*$. That is to say, we can write

$$S(\boldsymbol{p}, t) = \inf_{\boldsymbol{x}^* \in \mathbb{R}^n} \left\{ tH^* \left(\frac{\boldsymbol{p} - \boldsymbol{x}^*}{t} \right) + J(\boldsymbol{x}^*) \right\},$$

which is the Lax–Oleinik representation formula of the solution to the HJ PDE (2.66).

Examples of supervised machine learning algorithms. Different choices of Hamiltonians and initial data yield different supervised machine learning algorithms. Here, the Hamiltonian H in the function g plays the role of a regularization term, the function h plays the role of the loss function and defines the form of the initial data J in the HJ PDE, the matrix $\boldsymbol{A}$ encodes the data for the problem at hand, and the parameter $t > 0$ determines the constraint or the relative importance between the Hamiltonian and loss function in the optimization problem. As a first example, the choice

$$H(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{x}\|_2^2 \quad \text{and} \quad h(\boldsymbol{A}\boldsymbol{x}) = \log \left(\frac{1}{m} \sum_{i=1}^{m} e^{\langle \boldsymbol{x}, \boldsymbol{\Phi}(i) \rangle} \right)$$

with appropriate values $\boldsymbol{p}$ in (2.64) and $\{\boldsymbol{\Phi}(j)\}_{i=1}^m \subset \mathbb{R}^n$ leads to ℓ_2^2-regularized maximum entropy estimation with uniform prior distribution. Regularized maximum entropy estimation will be covered in detail in Section 3.3. As a second, and more detailed example, we consider "Maximum-likelihood type" estimators arising from generalized linear models in statistics [185, 246]. The problem is detailed below.

Suppose we receive m samples $\{\boldsymbol{u}_i, b_i\}_{i=1}^m \subset \mathbb{R}^n \times \mathcal{I}$, where b_i is some output or label from some space $\mathcal{I}$. The samples are drawn independently and identically from an unknown distribution $\mathcal{P}$. In addition, we assume that $\mathcal{P}$ belongs to an indexed family of probability distributions $\{\mathcal{P}_{\boldsymbol{x}}, \boldsymbol{x} \in \mathbb{R}^n\}$. Our goal is to use the m samples $\{\boldsymbol{u}_i, b_i\}_{i=1}^m$ to estimate or approximate the unknown parameter $\boldsymbol{x}_{\text{unknown}} \in \mathbb{R}^n$.

One possible approach for estimating the parameter $\boldsymbol{x}_u$ is to construct a "maximum likelihood-type" estimator, or M-estimator for short [185, 246]. These estimators consists of the sum of a

loss function and a regularizer term. The loss function $L\colon \mathbb{R}^n \times (\mathbb{R}^n \times \mathcal{I}) \to \mathbb{R}$ measures the fit of a parameter $\boldsymbol{x} \in \mathbb{R}^n$ to the m samples while the regularizer $H\colon \mathbb{R}^n \to \mathbb{R}$ enforces a certain type of structure expected in the parameter $\boldsymbol{x}$, e.g., sparsity. More precisely, M-estimators are the solutions to the minimization problem

$$\boldsymbol{x}_M = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ L(\boldsymbol{x}, \{\boldsymbol{u}_i, b_i\}_{i=1}^m) + tH(\boldsymbol{x}) \right\},$$

whenever they exist. An M-estimator to the problem above, if it exists, may not be unique without assuming additional assumptions on the loss function L and the regularizer H. The parameter $t > 0$ weighs the relative importance between the loss function and the regularizer.

Suppose now that the conditional distribution of the response $b \in \mathcal{I}$ given the predictor variable $\boldsymbol{u} \in \mathbb{R}^n$ is modeled as an exponential family. That is,

$$\mathcal{P}_{\boldsymbol{x}}(b \mid \boldsymbol{u}) = C(b) e^{b\langle \boldsymbol{u}, \boldsymbol{x}\rangle - \psi(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)}$$

for some constant $C(b)$ that depends only on the response b. The function $\psi\colon \mathbb{R} \to \mathbb{R}$ is the log-partition function underlying the exponential family. The associated loss function L to this conditional distribution is the scaled negative log-likelihood

$$\begin{aligned} L(\boldsymbol{x}, \{\boldsymbol{u}_i, b_i\}_{i=1}^m) &= -\frac{1}{m} \log\Big(\prod_{i=1}^m \mathcal{P}_{\boldsymbol{x}}(b_i \mid \boldsymbol{u}_i)\Big) \\ &= \frac{1}{m}\sum_{i=1}^m \psi(\langle \boldsymbol{u}_i, \boldsymbol{x}\rangle) - \left\langle \frac{1}{m}\sum_{i=1}^m b_i \boldsymbol{u}_i, \boldsymbol{x} \right\rangle. \end{aligned}$$

The corresponding M-estimator is then

$$\boldsymbol{x}_M = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \frac{1}{m}\sum_{i=1}^m \psi(\langle \boldsymbol{u}_i, \boldsymbol{x}\rangle) - \left\langle \frac{1}{m}\sum_{i=1}^m b_i \boldsymbol{u}_i, \boldsymbol{x} \right\rangle + tH(\boldsymbol{x}) \right\}$$

for some appropriate regularizer term H and $t > 0$. As an example, the M-estimator with

$$H(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2 \quad \text{and} \quad \psi(z) = \frac{z^2}{2}$$

corresponds to ℓ_2^2-regularized linear regression. The same function H with $\psi(z) = \log(1 + e^z)$ and $\psi(z) = e^z$ correspond to ℓ_2^2-regularized logistic and Poisson regression.

The connections to HJ PDEs can now be made explicitly. Let $\boldsymbol{b} \in \mathbb{R}^m$ denote the m-dimensional vector with entries $(b_1/m, \dots, b_m/m)$ and let $\boldsymbol{A}$ denote the $m \times n$ matrix whose rows are the predictor variables $\boldsymbol{u}_k$. In this case, one may express the loss function L as

$$L(\boldsymbol{x}, \{\boldsymbol{u}_i, b_i\}_{i=1}^m) = \frac{1}{m} \sum_{i=1}^m \psi([\boldsymbol{A}\boldsymbol{x}]_i) - \langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x} \rangle .$$

In addition, let $f \colon \mathbb{R}^m \to \mathbb{R}$ denote the function given by $f(\boldsymbol{y}) = \frac{1}{m} \sum_{i=1}^m \psi([\boldsymbol{y}]_i)$. Then the first term in the loss function is $f(\boldsymbol{A}\boldsymbol{x})$. With this notation, the M-estimator can be expressed as

$$\begin{aligned} \boldsymbol{x}_M &= \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathbb{R}^n} \{ f(\boldsymbol{A}\boldsymbol{x}) - \langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x} \rangle + tH(\boldsymbol{x}) \} \\ &= \operatorname*{arg\,max}_{\boldsymbol{x} \in \mathbb{R}^m} \{ \langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x} \rangle - tH(\boldsymbol{x}) - f(\boldsymbol{A}\boldsymbol{x}) \} . \end{aligned}$$

This is exactly in the same form as the Hopf representation formula (2.64) (third line), except that we take the arg max instead of the supremum.

Hence under appropriate conditions on the regularizer H, the samples $\{\boldsymbol{u}_i, b_i\}_{i=1}^m$ and the log-partition function ψ in the exponential family, the M-estimator corresponds to the unique minimizer associated to the Hopf representation formula of the solution to an appropriate first-order HJ PDE. Specifically, the value function that the M-estimator minimizes is, as a function of $\boldsymbol{p}$ and t, the solution

$$S(\boldsymbol{p}, t) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \{ \langle \boldsymbol{p}, \boldsymbol{x} \rangle - tH(\boldsymbol{x}) - J(\boldsymbol{x}) \}$$

to the HJ PDE (2.65) with initial condition given by the convex conjugate J of

$$J^*(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^m \psi([\boldsymbol{A}\boldsymbol{x}]_i)$$

evaluated at the point

$$\boldsymbol{p} = \boldsymbol{A}^* \boldsymbol{b} = \frac{1}{m} \sum_{i=1}^m b_i \boldsymbol{u}_i .$$

The M-estimator $\boldsymbol{x}_M$, as a function of $\boldsymbol{p}$ and t, also admits the following characterization [64, Proposition 3.1]:

$$\boldsymbol{x}_M(\boldsymbol{p}, t) = \boldsymbol{p} - t\nabla H(\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, t))).$$

The conditions needed for the M-estimator to be connected to the solution of an HJ PDE are satisfied by several Hamiltonians H and log-partition functions ψ. For example, the choice

$$H(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2 \quad \text{and} \quad \psi(z) = \frac{z^2}{2}$$

satisfy these conditions, and the M-estimator corresponds to ℓ_2^2-regularized linear regression. For the same Hamiltonian and choice $\psi(z) = \log(1 + e^z)$ and $\psi(z) = e^z$, the conditions are satisfied and the M-estimators correspond to ℓ_2^2-regularized logistic and Poisson regression.

2.6 Discussion

This chapter introduced new accelerated nonlinear primal-dual hybrid gradient (PDHG) optimization methods for solving large-scale convex optimization problems with saddle-point structure. We proved rigorous convergence results, including results for strongly convex or smooth problems posed on infinite-dimensional reflexive Banach spaces. It also presented some connections between classical solutions to first-order Hamilton–Jacobi partial differential equations and some supervised machine learning algorithms.

The introduction presented an example with ℓ_1-constrained logistic regression to illustrate how the new accelerated nonlinear PDHG methods are advantageous; because they use Bregman proximal operators to be flexible and because they can achieve an optimal convergence rate with stepsize parameters that are simple and efficient to compute. For this example, and several other examples discussed in the following chapter, the stepsize parameters can be computed on the order of $O(mn)$ operations, where m and n denote the dimensions to the dual and primal problems at hand. In contrast, most first-order optimization methods, including the linear PDHG method, require on

the order of $O(\min(m^2n, mn^2))$ operations to compute all the parameters required to achieve an optimal convergence rate. This gain turns out to be considerable in practice: In Section (3.2.4) of Chapter 3, we will present some numerical experiments in which the nonlinear PDHG method for ℓ_1-logistic regression (2.3) converges 5 to 10 times faster than the linear PDHG method (2.2). Chapter 3 will also explore applications to maximum entropy estimation problems in detail, as well as zero-sum matrix games with entropy regularization.

Appendix

2.A Proof of Lemma 2.2.1

We divide the proof into two parts, first proving that the output $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}^*)$ is contained in the set $\operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$ and then deriving the descent rule (2.11).

Part 1. Consider the functions

$$f = g + \langle \boldsymbol{A}^* \tilde{\boldsymbol{y}}^*, \cdot \rangle \quad \text{and} \quad \phi = \phi_{\mathcal{X}}.$$

By assumptions (A1)-(A3), the function ϕ is essentially smooth and essentially strictly convex, we have that $\operatorname{dom} f \cap \operatorname{int}(\operatorname{dom} \phi) \neq \varnothing$, and at least one of g and ϕ is supercoercive. If g is supercoercive, then an elementary calculation shows that the function f is also supercoercive, and therefore bounded from below by Fact 1.2.3. Hence we are guaranteed that f is bounded from below or ϕ is supercoercive. In either case, we can invoke Fact 1.2.9(i) to conclude that the minimization problem

$$\arg\min_{\boldsymbol{x} \in \mathcal{X}} \left\{ g(\boldsymbol{x}) + \langle \boldsymbol{A}^* \tilde{\boldsymbol{y}}^*, \boldsymbol{x} \rangle + \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \right\}$$

has a unique solution $\hat{\boldsymbol{x}}$ that is contained in the set $\operatorname{dom} \partial g \cap \operatorname{dom} \partial \phi_{\mathcal{X}}$. A similar argument using

assumptions (A1)-(A2) and (A4) shows that the minimization problem

$$\arg\min_{\boldsymbol{y}^* \in \mathcal{Y}^*} \left\{ h^*(\boldsymbol{y}^*) - \langle \boldsymbol{y}^*, \boldsymbol{A}\tilde{\boldsymbol{x}} \rangle + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) \right\}$$

has a unique solution $\hat{\boldsymbol{y}}^*$ that is contained in the set $\operatorname{dom} \partial h^* \cap \operatorname{dom} \partial \phi_{\mathcal{Y}^*}$.

Part 2. To derive the descent rule (2.11), we apply inequality (1.18) to each minimization problem in the iteration scheme (2.10). Note that inequality (1.18) can be used here because we showed in Part 1 that the conditions of Fact 1.2.9 are satisfied by each minimization problem in (2.10).

First, use inequality (1.18) with the functions

$$f = g + \langle \boldsymbol{A}^* \tilde{\boldsymbol{y}}^*, \cdot \rangle \quad \text{and} \quad \phi = \phi_{\mathcal{X}},$$

the parameter $t = \tau$, the element $\boldsymbol{x}' = \bar{\boldsymbol{x}}$, and the proximal point $\operatorname{prox}_{(tf, D_\phi)}(\boldsymbol{x}') = \hat{\boldsymbol{x}}$ to get

$$g(\boldsymbol{x}) + \langle \boldsymbol{A}^* \tilde{\boldsymbol{y}}^*, \boldsymbol{x} \rangle + \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \geqslant g(\hat{\boldsymbol{x}}) + \langle \boldsymbol{A}^* \tilde{\boldsymbol{y}}^*, \hat{\boldsymbol{x}} \rangle + \frac{1}{\tau} \left(D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) + D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \right).$$

Rearrange this inequality in terms of the difference $g(\hat{\boldsymbol{x}}) - g(\boldsymbol{x})$ to get

$$g(\hat{\boldsymbol{x}}) - g(\boldsymbol{x}) \leqslant \frac{1}{\tau} \left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \right) + \langle \boldsymbol{A}^* \tilde{\boldsymbol{y}}^*, \boldsymbol{x} - \hat{\boldsymbol{x}} \rangle. \tag{2.68}$$

A similar application of inequality (1.18) to the second line of the iteration scheme (2.10) gives

$$h^*(\hat{\boldsymbol{y}}^*) - h^*(\boldsymbol{y}^*) \leqslant \frac{1}{\sigma} \left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^*) \right) - \langle \boldsymbol{y}^* - \hat{\boldsymbol{y}}^*, \boldsymbol{A}\tilde{\boldsymbol{x}} \rangle. \tag{2.69}$$

Second, add the difference of bilinear forms

$$\langle \boldsymbol{y}^*, \boldsymbol{A}\hat{\boldsymbol{x}} \rangle - \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}\boldsymbol{x} \rangle$$

to both sides of inequality (2.69) and rearrange to get

$$\begin{aligned}(\langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}\hat{\boldsymbol{x}}\rangle - h^*(\boldsymbol{y}^{\boldsymbol{*}})) - (\langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}\boldsymbol{x}\rangle - h^*(\hat{\boldsymbol{y}}^*)) &= \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \hat{\boldsymbol{y}}^*)\right) \\ &\quad + \langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}\hat{\boldsymbol{x}}\rangle - \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}\boldsymbol{x}\rangle - \langle \boldsymbol{y}^{\boldsymbol{*}} - \hat{\boldsymbol{y}}^*, \boldsymbol{A}\tilde{\boldsymbol{x}}\rangle \\ &= \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \hat{\boldsymbol{y}}^*)\right) \\ &\quad + \langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle - \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle\,.\end{aligned} \tag{2.70}$$

Next, we combine inequalities (2.68) and (2.70). Add the left hand sides of inequalities (2.68) and (2.70) and use the definition (2.9) of the Lagrangian function $\mathcal{L}(\cdot,\cdot)$ to get

$$(g(\hat{\boldsymbol{x}}) + \langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}\hat{\boldsymbol{x}}\rangle - h^*(\boldsymbol{y}^{\boldsymbol{*}})) - (g(\boldsymbol{x}) + \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}\boldsymbol{x}\rangle - h^*(\hat{\boldsymbol{y}}^*)) = \mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{y}^{\boldsymbol{*}}) - \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}^*). \tag{2.71}$$

Thanks to (2.71), the sum of inequalities (2.68) and (2.70) give

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{y}^{\boldsymbol{*}}) - \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}^*) \leqslant\ & \frac{1}{\tau}\left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}})\right) \\ & + \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^{\boldsymbol{*}}, \hat{\boldsymbol{y}}^*)\right) \\ & + \langle \boldsymbol{A}^*\tilde{\boldsymbol{y}}^*, \boldsymbol{x} - \hat{\boldsymbol{x}}\rangle + \langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle - \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle\,.\end{aligned} \tag{2.72}$$

Now, write

$$\begin{aligned}\langle \boldsymbol{A}^*\tilde{\boldsymbol{y}}^*, \boldsymbol{x} - \hat{\boldsymbol{x}}\rangle &= \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \hat{\boldsymbol{x}})\rangle \\ &= \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle \\ &= \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle - \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle\end{aligned}$$

and use this to express the last line on the right hand side of inequality (2.72) as

$$\begin{aligned}\langle \boldsymbol{A}^*\tilde{\boldsymbol{y}}^*, \boldsymbol{x} - \hat{\boldsymbol{x}}\rangle + \langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle - \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle &= \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle - \langle \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle \\ &\quad + \langle \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{A}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})\rangle - \langle \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle \\ &= \langle \tilde{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle - \langle \boldsymbol{y}^{\boldsymbol{*}} - \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}})\rangle\,.\end{aligned} \tag{2.73}$$

Finally, combine inequalities (2.72) and (2.73) to find

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}^*) \leqslant \frac{1}{\tau}\left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\hat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \hat{\boldsymbol{x}})\right) \\ + \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\hat{\boldsymbol{y}}^*, \bar{\boldsymbol{y}}^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \hat{\boldsymbol{y}}^*)\right) \\ + \langle \tilde{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}^*, \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\rangle - \langle \boldsymbol{y}^* - \tilde{\boldsymbol{y}}^*, \boldsymbol{A}(\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}})\rangle\,.\end{aligned}$$

which is the desired result.

2.B Proof of Proposition 2.3.1

We divide the proofs into four parts, first deriving an auxiliary result, and then proving in turn the descent rule (2.15) (Proposition 2.3.1(a)), the estimate (2.16) and the global bound (2.17) (Proposition 2.3.1(b)), and the convergence properties of the nonlinear PDHG method (2.12) (Proposition 2.3.1(c)).

Part 1. We first show that for every $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$ and nonnegative integer k, the quantity $\delta_k(\boldsymbol{x}, \boldsymbol{y}^*)$ satisfies the bounds

$$\begin{aligned}0 \leqslant (1 - \sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}})\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*)\right) \\ \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) \leqslant (1 + \sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}})\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*)\right).\end{aligned} \tag{2.74}$$

To derive this, use fact 1.2.1 with the choice of $\alpha = \sqrt{\sigma/\tau}$ and $(\boldsymbol{x}', \boldsymbol{y}^{*\prime}) = (\boldsymbol{x}_k, \boldsymbol{y}_k^*)$ to get

$$\begin{aligned}|\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k)\rangle| &\leqslant \|\boldsymbol{A}\|_{\mathrm{op}}\left(\frac{\sqrt{\sigma}}{2\sqrt{\tau}}\|\boldsymbol{x} - \boldsymbol{x}_k\|_{\mathcal{X}}^2 + \frac{\sqrt{\tau}}{2\sqrt{\sigma}}\|\boldsymbol{y}^* - \boldsymbol{y}_k^*\|_{\mathcal{Y}^*}^2\right) \\ &= \sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}\left(\frac{1}{2\tau}\|\boldsymbol{x} - \boldsymbol{x}_k\|_{\mathcal{X}}^2 + \frac{1}{2\sigma}\|\boldsymbol{y}^* - \boldsymbol{y}_k^*\|_{\mathcal{Y}^*}^2\right) \\ &\leqslant \sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*)\right),\end{aligned} \tag{2.75}$$

where in the last line we used assumption (A5) and Fact 1.2.8(viii) with $m = 1$. Inequality (2.74) then follows from equation (2.14) and inequalities (2.75) and (2.13).

Part 2. Let $(\boldsymbol{x}, \boldsymbol{y}^*) \in \operatorname{dom} g \times \operatorname{dom} h^*$. By assumption (A1)-(A4), Lemma 2.2.1 holds, and we can apply the descent rule (2.11) to the $(k+1)^{\text{th}}$ iterate given by (2.12) with initial points $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}^*) = (\boldsymbol{x}_k, \boldsymbol{y}_k^*)$ and intermediate points $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}^*) = (2\boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \boldsymbol{y}_k^*)$ to get

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \frac{1}{\tau}\left(D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_k) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) - D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_{k+1})\right) \\ + \frac{1}{\sigma}\left(D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_{k+1}^*, \boldsymbol{y}_k^*) - D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_{k+1}^*)\right) \\ + \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - (2\boldsymbol{x}_{k+1} - \boldsymbol{x}_k))\right\rangle \\ - \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}((2\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) - \boldsymbol{x}_{k+1})\right\rangle.\end{aligned} \tag{2.76}$$

To proceed, we want to rewrite the last two lines of (2.76) to simplify the analysis. First, write the penultimate line on the right hand side of (2.76) as

$$\begin{aligned}\left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - (2\boldsymbol{x}_{k+1} - \boldsymbol{x}_k))\right\rangle &= \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1}) - \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\right\rangle \\ &= \left\langle \boldsymbol{y}_k^* - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1})\right\rangle \\ &\quad + \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\right\rangle \\ &= \left\langle \boldsymbol{y}^* - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1})\right\rangle \\ &\quad - \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1})\right\rangle \\ &\quad + \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\right\rangle \\ &= \left\langle \boldsymbol{y}^* - \boldsymbol{y}_{k+1}^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{k+1})\right\rangle \\ &\quad + \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\right\rangle \\ &\quad - \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k)\right\rangle \\ &\quad + \left\langle \boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\right\rangle,\end{aligned} \tag{2.77}$$

The bilinear form on the last line of (2.76) simplifies to

$$\left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}((2\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) - \boldsymbol{x}_{k+1})\right\rangle = \left\langle \boldsymbol{y}^* - \boldsymbol{y}_k^*, \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\right\rangle. \tag{2.78}$$

Combine equations (2.14), (2.77) and (2.78) together to write inequality (2.76) as

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}_k, \boldsymbol{y}_k^*).$$

Thanks to inequality (2.74),

$$-\delta_{k+1}(\boldsymbol{x}_k, \boldsymbol{y}_k^*) \leqslant 0,$$

and hence

$$\mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k+1}^*) \leqslant \delta_k(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k+1}(\boldsymbol{x}, \boldsymbol{y}^*).$$

This proves the descent rule (2.15).

Part 3. Sum inequality (2.15) from $k = 1$ to K on both sides to obtain

$$\sum_{k=1}^{K} \mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_k^*) \leqslant \delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_K(\boldsymbol{x}, \boldsymbol{y}^*). \tag{2.79}$$

Use the averages

$$\boldsymbol{X}_K = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_k \quad \text{and} \quad \boldsymbol{Y}_K^* = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{y}_k^*,$$

the convexity and concavity in the first and second arguments of the Lagrangian (2.9), respectively, and inequality (2.79) to bound the difference of Lagrangians $\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*)$ as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*) &\leqslant \frac{1}{K}\sum_{k=1}^{K} \left(\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_k^*)\right) \\ &\leqslant \frac{1}{K}\left(\delta_0(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_K(\boldsymbol{x}, \boldsymbol{y}^*)\right). \end{aligned} \tag{2.80}$$

Finally, use the lower and upper bounds (2.74) in (2.80) to get

$$\begin{aligned} \mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{Y}_K^*) \leqslant{}& \frac{1 + \sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{K}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_0) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_0^*)\right) \\ &- \frac{1 - \sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{K}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_K) + \frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_K^*)\right). \end{aligned}$$

This proves the estimate (2.16).

Now, let $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_s, \boldsymbol{y}_s^*)$ in estimate (2.16) and use the saddle point property

$$\mathcal{L}(\boldsymbol{X}_K, \boldsymbol{y}_s^*) - \mathcal{L}(\boldsymbol{x}_s, \boldsymbol{Y}_K^*) \geqslant 0$$

and rearrange to get

$$(1-\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}})\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s,\boldsymbol{x}_K)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*,\boldsymbol{y}_K^*)\right)$$
$$\leqslant(1+\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}})\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s,\boldsymbol{x}_0)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*,\boldsymbol{y}_0^*)\right).$$

Since $\tau\sigma\,\|\boldsymbol{A}\|_{\mathrm{op}}^2<1$, the number $(1-\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}})$ is strictly positive and we can divide both sides of the previous inequality by $(1-\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}})$ to get

$$\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s,\boldsymbol{x}_K)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*,\boldsymbol{y}_K^*)\leqslant\frac{1+\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{1-\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_s,\boldsymbol{x}_0)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_s^*,\boldsymbol{y}_0^*)\right).$$

This proves inequality (2.17).

Part 4. First, note that the global bound (2.17) implies that the sequence of iterates $\{(\boldsymbol{x}_k,\boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ is bounded. It follows immediately from the definitions of the averages $\boldsymbol{X}_K$ and $\boldsymbol{Y}_K^*$ that the sequence of averages $\{(\boldsymbol{X}_K,\boldsymbol{Y}_K^*)\}_{K=1}^{+\infty}$ is also bounded.

From Fact 1.2.4, there is a subsequence $\{(\boldsymbol{X}_{K_l},\boldsymbol{Y}_{K_l}^*)\}_{l=1}^{+\infty}$ that converges weakly to some point $(\boldsymbol{X},\boldsymbol{Y}^*)\in\mathcal{X}\times\mathcal{Y}^*$. We claim that $(\boldsymbol{X},\boldsymbol{Y}^*)$ is a saddle point of the Lagrangian (2.9). To see this, use inequality (2.16) with $(\boldsymbol{x},\boldsymbol{y}^*)\in\operatorname{dom}g\times\operatorname{dom}h^*$ and take the infimum limit $l\to+\infty$ to get

$$\liminf_{l\to+\infty}\left[\mathcal{L}(\boldsymbol{X}_{K_l},\boldsymbol{y}^*)-\mathcal{L}(\boldsymbol{x},\boldsymbol{Y}_{K_l}^*)\right]\leqslant\liminf_{l\to+\infty}\left[\frac{1+\sqrt{\tau\sigma}\,\|\boldsymbol{A}\|_{\mathrm{op}}}{K_l}\left(\frac{1}{\tau}D_{\phi_{\mathcal{X}}}(\boldsymbol{x},\boldsymbol{x}_0)+\frac{1}{\sigma}D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*,\boldsymbol{y}_0^*)\right)\right]$$
$$=0.$$

The lower semicontinuity property of the functions g and h^* implies

$$0\geqslant\liminf_{l\to+\infty}\left(\mathcal{L}(\boldsymbol{X}_{K_l},\boldsymbol{y}^*)-\mathcal{L}(\boldsymbol{x},\boldsymbol{Y}_{K_l}^*)\right)\geqslant\mathcal{L}(\boldsymbol{X},\boldsymbol{y}^*)-\mathcal{L}(\boldsymbol{x},\boldsymbol{Y}^*),$$

from which we find

$$\mathcal{L}(\boldsymbol{X},\boldsymbol{y}^*)\leqslant\mathcal{L}(\boldsymbol{x},\boldsymbol{Y}^*).$$

As the pair of points $(\boldsymbol{x},\boldsymbol{y}^*)\in\mathcal{X}\times\mathcal{Y}^*$ was arbitrary, we conclude that $(\boldsymbol{X},\boldsymbol{Y}^*)$ is a saddle point

of the Lagrangian (2.9).

Assume now that the spaces $\mathcal{X}$ and $\mathcal{Y}^*$ are finite-dimensional. Since the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ is bounded, by Fact 1.2.4 there is a subsequence that converges strongly to some point $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$. Note that $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$ is a fixed point of the nonlinear PDHG method (2.12), and therefore we can invoke Lemma 2.2.1 to conclude that $(\boldsymbol{x}_c, \boldsymbol{y}_c^*) \in \operatorname{dom} \partial g \times \operatorname{dom} \partial h^*$.

We claim that $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$ is a saddle point of the Lagrangian (2.9). To see this, consider the descent rule (2.15) with arbitrary $(\boldsymbol{x}, \boldsymbol{y}^*)$ and the subsequence $\{(\boldsymbol{x}_{k_l}, \boldsymbol{y}_{k_l}^*)\}_{l=1}^{+\infty}$:

$$\mathcal{L}(\boldsymbol{x}_{k_l}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k_l}^*) \leqslant \delta_{k_l}(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k_{l+1}}(\boldsymbol{x}, \boldsymbol{y}^*).$$

By the strong convergence of the subsequence $\{(\boldsymbol{x}_{k_l}, \boldsymbol{y}_{k_l}^*)\}_{l=1}^{+\infty}$ to $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$ and Fact 1.2.8(vii), we have the limits

$$\lim_{l \to +\infty} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_{k_l}) = D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_c) \quad \text{and} \quad \lim_{l \to +\infty} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_{k_l}^*) = D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_c^*).$$

Hence

$$\lim_{l \to +\infty} \delta_{k_l}(\boldsymbol{x}, \boldsymbol{y}^*) = \frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{x}_c) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}^*, \boldsymbol{y}_c^*) - \langle \boldsymbol{y}^* - \boldsymbol{y}_c^*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_c) \rangle,$$

and we conclude, from the completeness property of the real numbers, that

$$\liminf_{l \to +\infty} \delta_{k_l}(\boldsymbol{x}, \boldsymbol{y}^*) - \delta_{k_{l+1}}(\boldsymbol{x}, \boldsymbol{y}^*) = 0.$$

We therefore deduce the infimum limit

$$\liminf_{l \to +\infty} \mathcal{L}(\boldsymbol{x}_{k_l}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k_l}^*) \leqslant 0.$$

The lower semicontinuity property of the functions g and h^* implies

$$0 \geqslant \liminf_{l \to +\infty} \left(\mathcal{L}(\boldsymbol{x}_{k_l}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{k_l}^*) \right) \geqslant \mathcal{L}(\boldsymbol{x}_c, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_c^*),$$

from which we find

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_c^*) \leqslant \mathcal{L}(\boldsymbol{x}_c, \boldsymbol{y}^*).$$

As the pair of points $(\boldsymbol{x}, \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}^*$ was arbitrary, we conclude that $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$ is a saddle point of the Lagrangian (2.9).

It remains to prove that the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{+\infty}$ converges strongly to the saddle point $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$. To do so, consider the descent rule (2.15) with the choice of saddle point $(\boldsymbol{x}, \boldsymbol{y}^*) = (\boldsymbol{x}_c, \boldsymbol{y}_c^*)$. From the saddle-point property $\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}_c^*) - \mathcal{L}(\boldsymbol{x}_c, \boldsymbol{y}_k^*) \geqslant 0$ and inequality (2.74), we have

$$0 \leqslant \delta_k(\boldsymbol{x}_c, \boldsymbol{y}_c^*) \leqslant \delta_{k-1}(\boldsymbol{x}_c, \boldsymbol{y}_c^*).$$

The sequence of real numbers $\{\delta_k(\boldsymbol{x}_c, \boldsymbol{y}_c^*)\}_{k=1}^{+\infty}$ is non-increasing in l, and as such, it has a limit. By Lemma 2.2.1, Fact 1.2.8(vi), and the strong convergence of the subsequence $\{(\boldsymbol{x}_{k_l}, \boldsymbol{y}_{k_l}^*)\}_{l=1}^{+\infty}$ to $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$, we have

$$\lim_{l \to +\infty} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_c, \boldsymbol{x}_{k_l}) = 0, \quad \lim_{l \to +\infty} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_c^*, \boldsymbol{y}_{k_l}^*) = 0, \quad \text{and} \quad \lim_{l \to +\infty} \left\langle \boldsymbol{y}_c^* - \boldsymbol{y}_{k_l}^*, \boldsymbol{A}(\boldsymbol{x}_c - \boldsymbol{x}_{k_l}) \right\rangle = 0.$$

We deduce the limit

$$\lim_{k \to +\infty} \delta_k(\boldsymbol{x}_c, \boldsymbol{y}_c^*) = 0.$$

Now, from this limit and the lower bound (2.74) with $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}_c, \boldsymbol{y}_c^*)$, we have

$$0 \leqslant \lim_{k \to +\infty} (1 - \sqrt{\tau\sigma}\, \|\boldsymbol{A}\|_{\text{op}}) \left(\frac{1}{\tau} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_c, \boldsymbol{x}_k) + \frac{1}{\sigma} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_c^*, \boldsymbol{y}_k^*) \right) \leqslant \lim_{k \to +\infty} \delta_k(\boldsymbol{x}_c, \boldsymbol{y}_c^*) = 0.$$

Since $(1 - \sqrt{\tau\sigma}\, \|\boldsymbol{A}\|_{\text{op}}) > 0$ and assumption (A5) holds, we deduce the limits

$$0 \leqslant \lim_{k \to +\infty} \frac{1}{2} \|\boldsymbol{x}_c - \boldsymbol{x}_k\|_{\mathcal{X}}^2 \leqslant \lim_{k \to +\infty} D_{\phi_{\mathcal{X}}}(\boldsymbol{x}_c, \boldsymbol{x}_k) = 0$$

and

$$0 \leqslant \lim_{k \to +\infty} \frac{1}{2} \|\boldsymbol{y}_c^* - \boldsymbol{y}_k^*\|_{\mathcal{Y}^*}^2 \leqslant \lim_{k \to +\infty} D_{\phi_{\mathcal{Y}^*}}(\boldsymbol{y}_c^*, \boldsymbol{y}_k^*) = 0.$$

This proves the strong convergence of the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{y}_k^*)\}_{k=1}^{+\infty}$ to the saddle point

$(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$. Finally, we deduce from Fact 1.2.2 that the sequence of averages $\{(\boldsymbol{X}_K, \boldsymbol{Y}_K^*)\}_{K=1}^{+\infty}$ converges strongly to the same limit $(\boldsymbol{x}_c, \boldsymbol{y}_c^*)$. This concludes the proof.

Chapter Three

Variational methods for machine learning algorithms and connections to Hamilton–Jacobi PDEs II: Applications

3.1 Introduction

This chapter applies the accelerated nonlinear PDHG methods developed in the previous chapter to sparse logistic regression, regularized maximum entropy estimation, and entropy-regularized matrix games. We discuss each problem in detail and provide an accelerated nonlinear PDHG optimization method to solve it efficiently. Finally, we present numerical experiments to illustrate that the accelerated nonlinear primal-dual hybrid gradient methods are considerably faster than competing methods.

In all problems, the real reflexive Banach spaces $\mathcal{X}$ and $\mathcal{Y}$ are taken to be finite dimensional Banach spaces with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ chosen suitably for each example. In each forthcoming example, we will choose the norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ and Bregman functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ to obtain an explicit accelerated nonlinear PDHG method for which the stepsize parameters and updates are simple and efficient to compute. To see how this works, consider the case where $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$. Table (3.1.1) below illustrates that with these spaces, certain combinations of norms make it possible to compute the induced operator norm in $O(mn)$ operations, down from $O(\min(m^2n, mn^2))$ complexity with the classical choice $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_{\mathcal{Y}} = \|\cdot\|_2$. In addition, we can choose the Bregman functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ in conjunction with these norms to ensure that assumptions (A1)–(A5) from the previous chapter hold. This is our strategy. For certain choices of norms adapted to the problem at hand, we will obtain an accelerated nonlinear PDHG method that is both explicit and significantly more efficient compared to other competing methods.

		Codomain		
		$(\mathbb{R}^m, \|\cdot\|_1)$	$(\mathbb{R}^m, \|\cdot\|_2)$	$(\mathbb{R}^m, \|\cdot\|_\infty)$
Domain	$(\mathbb{R}^n, \|\cdot\|_1)$	Max. ℓ_1 norm of a column ($\sim O(mn)$)	Max. ℓ_2 norm of a column ($\sim O(mn)$)	Max. ℓ_∞ norm of a column ($\sim O(mn)$)
	$(\mathbb{R}^n, \|\cdot\|_2)$	NP-hard	Largest singular value ($\sim O(\min(m^2n, mn^2))$)	Max. ℓ_2 norm of a row ($\sim O(mn)$)
	$(\mathbb{R}^n, \|\cdot\|_\infty)$	NP-hard	NP-hard	Max. ℓ_1 norm of a row ($\sim O(mn)$)

Table 3.1.1: Table of some operator norms of $\boldsymbol{A}$ with their associated computational complexity. Table extracted from [239, Section 4.3.1].

3.2 Sparse logistic regression

Logistic regression is a widely used statistical model to describe the relationship between a binary response variable and features in data sets [142]. It is often used in machine learning to identify important features [93, 259]. This task, variable selection, typically amounts to fitting a logistic regression model regularized by an ℓ_1 penalty. Variable selection is frequently applied to problems in medicine [11, 37, 123, 195, 210, 252, 260], natural language processing [26, 174, 118, 208, 230], economics [170, 232, 257, 258], and social science [1, 153, 183], among others.

Since modern big data sets can contain hundred of thousands to billions of features, variable selection methods require efficient and robust optimization algorithms to scale well [171]. State-of-the-art algorithms for variable selection methods, however, were not traditionally designed to handle big data sets; they either lack scalable parallelism or scale poorly in size [57, 161] or are prone to produce unreliable numerical results [34, 167, 255, 256]. These shortcomings in terms of efficiency and robustness make variable selection methods on big data sets essentially impossible without access to adequate and costly computational resources [77, 224].

This section proposes an accelerated nonlinear PDHG method designed to overcome the shortcomings described above for variable selection via ℓ_1-constrained logistic regression. This method, as we will describe soon, provably computes a solution to ℓ_1-constrained logistic regression in $O(md/\sqrt{\epsilon})$ operations, where $\epsilon \in (0, 1)$ denotes the tolerance and m and d denote the number of samples and features present in the logistic regression problem. This result improves on the known complexity bound of $O(\min(m^2d, md^2)/\sqrt{\epsilon})$ for first-order optimization methods such as the linear PDHG method or forward-backward splitting methods. This gain turns out to be considerable in practice: for instance, in Section 3.2.4 we present some numerical experiments in which the accelerated nonlinear PDHG method converges 5 to 10 times faster than its linear counterpart, an order of magnitude speed-up.

3.2.1 Description of the problem

Suppose we receive m independent samples $\{\boldsymbol{u}_i, b_i\}_{i=1}^m$, each comprising a d-dimensional vector of features $\boldsymbol{u}_i$ and a response variable $b_i \in \{-1, +1\}$. The goal of variable selection is to identify which of the d features best describe the m response variables. A common approach to do so is to fit a logistic regression model constrained by an ℓ_1 norm:

$$\inf_{\boldsymbol{v}\in\mathbb{R}^d} f(\boldsymbol{v};\lambda) = \inf_{\substack{\boldsymbol{v}\in\mathbb{R}^d \\ \|\boldsymbol{v}\|_1 \leqslant \lambda}} \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-b_i\langle \boldsymbol{u}_i, \boldsymbol{v}\rangle}\right) \tag{3.1}$$

where $\lambda > 0$ is a tuning parameter. The ℓ_1-constrained logistic regression problem is often rewritten in non-constrained form as

$$\inf_{\boldsymbol{v}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-b_i\langle \boldsymbol{u}_i, \boldsymbol{v}\rangle}\right) + t\,\|\boldsymbol{v}\|_1 \tag{3.2}$$

where $t > 0$ and there exists a one-to-one correspondence between λ and t such that the solution is the same for each problem. We shall focus in this section on the constrained-form of sparse logistic regression, that is, on problem (3.1).

The constraint on the ℓ_1 norm regularizes the logistic model in two ways. First, it ensures that problem (3.1) has at least one solution [92, Page 35, Proposition 1.2]. Second, it promotes solutions to have a number of entries that are identically zero [109, 93, 259]. The non-zero entries are identified as the important features, and the zero entries are discarded. The number of non-zero entries itself depends on the value of the tuning parameter λ.

In most applications, the appropriate value of λ proves challenging to estimate. To determine it, variable selection methods first compute a sequence of minimums $\boldsymbol{v}^*(\lambda)$ of problem (3.1) from a chosen sequence of values of the parameter λ and then choose the parameter that gives the preferred minimum [34, 112]. Variable selection methods differ in how they choose the sequence of parameters λ and how they repeatedly compute global minimums of problem (3.1), but the procedure is generally the same. The sequence of parameters thus computed is called a regularization path [112].

Unfortunately, computing a regularization path to problem (3.1) can be prohibitively expensive for big data sets. To see why, fix $\lambda > 0$ and let $\boldsymbol{v}_\epsilon(\lambda) \in \mathbb{R}^n$ with $\epsilon > 0$ denote an ϵ-approximate solution to a global minimum $\boldsymbol{v}^*(\lambda)$ in (3.1), that is, the objective function in (3.1) satisfies

$$f(\boldsymbol{v}_\epsilon(\lambda); \lambda) - f(\boldsymbol{v}^*(\lambda); \lambda) < \epsilon.$$

Then the best achievable convergence rate for computing $\boldsymbol{v}_\epsilon(\lambda)$ in the Nesterov class of optimal first-order methods is sublinear, that is, $O(1/\sqrt{\epsilon})$ in the number of iterations [187]. While optimal, this convergence rate is difficult to achieve in practice. Indeed, letting $\boldsymbol{B}$ denote the $m \times d$ matrix whose rows are the elements $-b_i \boldsymbol{u}_i$, one requires a precise estimate of the largest singular value of the matrix $\boldsymbol{B}$ to achieve an optimal convergence rate. This quantity, however, is essentially impossible to compute for large matrices due to its prohibitive computational cost of $O(\min(m^2 d, m d^2))$ operations [130]. This issue generally makes solving problem (3.1) difficult and laborious. As computing a regularization path entails repeatedly solving problem (3.1) for different values of λ, this process can become particularly time consuming and resource intensive for big data sets.

3.2.2 State-of-the-art optimization methods

The issue of solving sparse logistic regression and constructing regularization paths has driven much research in the development of robust and efficient methods to minimize costs and maximize performance. Most optimization methods developed in this vein focus on the equivalent non-constrained form (3.2) of sparse logistic regression. The state-of-the-art is based on coordinate descent algorithms [111, 112, 131, 227, 228, 235, 251, 256]. These algorithms are implemented, for example, in the popular glmnet software package [131], which is available in the Python, MATLAB, and R programming languages. Other widely used variable selection methods include those based on the least angle regression algorithm and its variants [91, 134, 161, 236, 263], and those based on the forward-backward splitting algorithm and its variants [23, 48, 74, 225, 226]. Here, we focus on these algorithms, but before doing so we wish to stress that many more algorithms have been developed for sparse logistic regression; see [27, 93, 163, 243, 259] for recent surveys and comparisons

of different methods and models.

Coordinate descent algorithms are considered the state of the art for computing regularization paths because they are scalable, with steps in the algorithms generally having an asymptotic space complexity of at most $O(md)$ operations. Some coordinate descent algorithms, such as the one implemented in the popular glmnet software [131], also implement selection rules to exploit the sparsity of the matrix $\boldsymbol{B}$ and a priori knowledge of the sparsity of the solution. Despite these advantages, coordinate descent algorithms for constructing regularization paths generally are non-parallelizable and generally lack robustness and good convergence properties. For example, the glmnet implementation offers no option for parallel computing. The implementation also depends on the sparsity of the matrix $\boldsymbol{B}$ and a priori knowledge of the sparsity of the solution to converge fast [263]. It is also known to be slowed down when the features are highly correlated [111]. It would be desirable to have a fast algorithm for when the matrix $\boldsymbol{B}$ is dense and for when the features are highly correlated, as these occur often in practice. Another issue is that the glmnet implementation approximates the logarithm term in sparse logistic regression with a quadratic in order to solve the problem efficiently. Without costly step-size optimization, which glmnet avoids to improve performance, the glmnet implementation may not converge [112, 161]. Case in point, Yuan et al. [255] provides two numerical experiments in which glmnet does not converge for the non-constrained problem (3.2). Although some coordinate descent algorithms recently proposed in [41] and in [102] can provably solve the logistic regression problem (3.2), in the first case, the convergence rate is strictly less than the achievable rate, and in the second case, the method fails to construct meaningful regularization paths to problem (3.2), in addition to having large memory requirements.

The least angle regression algorithm is another popular tool for computing regularization paths. This algorithm, however, scales poorly with the size of data sets because the entire sequence of steps for computing regularization paths has an asymptotic space complexity of at most $O(\min(m^2n + m^3, mn^2 + n^3))$ operations [91]. It also lacks robustness because, under certain conditions, it fails to compute meaningful regularization paths [34, 167]. Case in point, Bringmann et al. [34] provides an example for which the least angle regression algorithm fails to converge.

The forward-backward splitting algorithm and its variants are widely used because they are robust and can provably compute ϵ-approximate solutions of (3.1) and (3.2) in at most $O(1/\sqrt{\epsilon})$ iterations. To achieve this convergence rate, the stepsize parameter in the algorithm needs to be fine-tuned using a precise estimate of the largest singular value of the matrix $\boldsymbol{B}$. As mentioned before, however, computing this estimate is essentially impossible for large matrices due to its prohibitive computational cost, which has an asymptotic computational complexity of at most $O(\min(m^2d, md^2))$ operations. Line search methods and other heuristics are often employed to bypass this problem, but they come at the cost of slowing down the convergence of the forward-backward splitting algorithm. Another approach is to compute a crude estimate of the largest singular value of the matrix $\boldsymbol{B}$, but doing so dramatically reduces the speed of convergence of the algorithm. This problem makes regularization path construction methods based on the forward-backward splitting algorithm and its variants generally inefficient and impractical for big data sets.

In summary, state-of-the-art variable selection methods for sparse logistic regression and computing regularization paths to problems (3.1) and (3.2) either scale poorly in size or are prone to produce unreliable numerical results. These shortcomings in terms of efficiency and robustness make it challenging to perform variable selection on big data sets without access to adequate and costly computational resources. We shall present here an efficient and robust accelerated nonlinear PDHG optimization that addresses these shortcomings.

3.2.3 Derivation of the accelerated nonlinear PDHG method

To derive an appropriate accelerated nonlinear PDHG method for ℓ_1-constrained logistic regression, we will express problem (3.1) as an minimization problem over the $2d$-dimensional unit simplex Δ_{2d}. We can do so because every polytope, including the ℓ_1-ball, can be represented as a convex hull of its vertices in barycentric coordinates [128, 145]. Here this means for every $\boldsymbol{v}$ inside the

ℓ_1-ball of radius λ, there exists a point $\boldsymbol{x}$ in Δ_{2d} for which

$$\boldsymbol{v} = \lambda(\boldsymbol{I}_{d\times d} \mid -\boldsymbol{I}_{d\times d})\boldsymbol{x}, \tag{3.3}$$

where $(\boldsymbol{I}_{d\times d} \mid -\boldsymbol{I}_{d\times d})$ denotes the horizontal concatenation of the identity matrices $\boldsymbol{I}_{d\times d}$ and $-\boldsymbol{I}_{d\times d}$.

We now apply the change of variables (3.3) to problem (3.1). Let $\boldsymbol{B}$ denote the $m \times d$ matrix $\boldsymbol{B}$ whose rows are the elements $-b_i\boldsymbol{u}_i$, let $\boldsymbol{A} = \lambda(\boldsymbol{B} \mid -\boldsymbol{B})$, and let $n = 2d$. Then problem (3.1) becomes equivalent to

$$\inf_{\boldsymbol{x}\in\Delta_n} \frac{1}{m}\sum_{i=1}^{m} \log\left(1 + e^{[\boldsymbol{A}\boldsymbol{x}]_i}\right), \tag{3.4}$$

where Δ_n denotes the n-dimensional unit simplex. This is the primal problem of interest. Its associated convex-concave saddle-point problem is

$$\inf_{\boldsymbol{x}\in\Delta_n} \sup_{\boldsymbol{y}^*\in\mathbb{R}^m} \{\langle \boldsymbol{y}^*, \boldsymbol{A}\boldsymbol{x}\rangle - \psi(\boldsymbol{y}^*)\} \tag{3.5}$$

where $\psi\colon [0, 1/m]^m \to \mathbb{R}$ denotes the average negative sum of m binary entropy terms,

$$\psi(\boldsymbol{y}^*) = \begin{cases} \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} m[\boldsymbol{y}^*]_i \log\left(m[\boldsymbol{y}^*]_i\right) + \left(1 - m[\boldsymbol{y}^*]_i\right)\log\left(1 - m[\boldsymbol{y}^*]_i\right) & \text{if } \boldsymbol{y}^* \in [0, 1/m]^m, \\ +\infty, \quad \text{otherwise.} \end{cases} \tag{3.6}$$

The dual problem is

$$\sup_{\boldsymbol{y}^*\in\mathbb{R}^m} \{\mathrm{vecmax}(-\boldsymbol{A}^*\boldsymbol{y}^*) - \psi(\boldsymbol{y}^*)\} \tag{3.7}$$

where $\mathrm{vecmax}(\boldsymbol{y}) = \max\left([\boldsymbol{y}]_1, \ldots, [\boldsymbol{y}]_m\right)$ for $\boldsymbol{y} \in \mathbb{R}^m$. Due to the strong concavity of the dual problem (3.7), the convex-concave saddle-point problem (3.5) has at least one saddle point $(\boldsymbol{x}_s, \boldsymbol{y}_s^*) \in \Delta_n \times \mathbb{R}^m$, where $\boldsymbol{x}_s$ is a global solution to the primal problem (3.4) and $\boldsymbol{y}_s^*$ is the unique solution to the dual problem (3.7). They satisfy the optimality conditions

$$\boldsymbol{x}_s \in \partial\mathrm{vecmax}(-\boldsymbol{A}^*\boldsymbol{y}_s^*) \quad \text{and} \quad [\boldsymbol{y}_s^*]_i = \frac{1}{m + me^{-[\boldsymbol{A}\boldsymbol{x}_s]_i}} \text{ for } i \in \{1, \ldots, m\}. \tag{3.8}$$

The solution $\boldsymbol{v}_s$ of the original problem (3.1) follows from $\boldsymbol{x}_s$ and the change of variables for-

mula (3.3). In addition, the first optimality condition in (3.8) can be used to identify the zero entries of $\boldsymbol{x}_s$ as follows [215]: Let $J(-\boldsymbol{A}^*\boldsymbol{y}_s^*)$ denote the set of indices $j \in \{1,\dots,n\}$ with $\operatorname{vecmax}(-\boldsymbol{A}^*\boldsymbol{y}_s^*) = [-\boldsymbol{A}^*\boldsymbol{y}_s^*]_j$. Then $[\boldsymbol{x}_s]_j = 0$ whenever $j \notin J(-\boldsymbol{A}^*\boldsymbol{y}_s^*)$.

Accelerated nonlinear PDHG method

We propose to solve the ℓ_1-constrained logistic regression problem (3.1) through (3.4) and (3.3) using the accelerated nonlinear PDHG method (2.39) with the following choice of norms and Bregman functions:

$$\|\cdot\|_{\mathcal{X}} = \|\cdot\|_1\,, \quad \|\cdot\|_{\mathcal{Y}^*} = \|\cdot\|_2\,, \quad \phi_{\mathcal{X}} = \mathcal{H}_n, \quad \text{and} \quad \phi_{\mathcal{Y}^*} = \frac{1}{4m}\psi,$$

where $\mathcal{H}_n\colon \Delta_n \to (-\infty, 0]$ denotes the negative entropy function,

$$\mathcal{H}_n(\boldsymbol{x}) = \sum_{j=1}^{n} [\boldsymbol{x}]_j \log([\boldsymbol{x}]_j).$$

The negative entropy function induces the Bregman divergence $D_{\mathcal{H}_n}\colon \Delta_n \times \operatorname{int}\Delta_n \to [0,+\infty)$ given by

$$D_{\mathcal{H}_n}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \sum_{j=1}^{n} [\boldsymbol{x}]_j \log\left([\boldsymbol{x}]_j / [\bar{\boldsymbol{x}}]_j\right).$$

This Bregman divergence is the so-called Kullback–Leibler divergence or relative entropy. The Bregman function $\phi_{\mathcal{Y}^*}$ is, up to a factor of $1/4m$, the average negative sum of m binary entropy terms (3.6). It induces the Bregman divergence $D_{\psi/4m}\colon [0,1/m]^m \times (0,1/m)^m \to [0,+\infty)$ given by

$$D_{\psi/4m}(\boldsymbol{y}^*, \bar{\boldsymbol{y}}^*) = \frac{1}{4m^2} \sum_{i=1}^{m} m[\boldsymbol{y}^*]_i \log\left(\frac{[\boldsymbol{y}^*]_i}{[\bar{\boldsymbol{y}}^*]_i}\right) + (1 - m[\boldsymbol{y}^*]_i) \log\left(\frac{1 - m[\boldsymbol{y}^*]_i}{1 - m[\bar{\boldsymbol{y}}^*]_i}\right).$$

With these choices, assumptions (A1)-(A5) and (A7) hold with $\gamma_{h^*} = 4m$. In particular, assumption (A5) holds because $\mathcal{H}_n$ is 1-strongly convex with respect to the ℓ_1 norm over the unit simplex Δ_n. This fact is a direct consequence of a fundamental result in information theory known as Pinsker's inequality [22, 63, 152, 158, 203]. Moreover, the induced operator norm is the maximum ℓ_2 norm

of the columns of $\boldsymbol{A}$, i.e.,

$$\|\boldsymbol{A}\|_{\text{op}} = \|\boldsymbol{A}\|_{1,2} = \sup_{\|\boldsymbol{x}\|_1=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \max_{j\in\{1,\dots,n\}} \sqrt{\sum_{i=1}^{m} A_{ij}^2}.$$

For this method, we set the initial stepsize parameters to be $\theta_0 = 0$, $\tau_0 > 0$ and $\sigma_0 = 1/(\|\boldsymbol{A}\|_{1,2}^2\,\tau_0)$. Given $\boldsymbol{x}_{-1} = \boldsymbol{x}_0 \in \operatorname{int}\Delta_n$ and $\boldsymbol{y}_0^* \in (0, 1/m)^m$, the corresponding accelerated nonlinear PDHG method for problem (3.4) consists of the iterations

$$\begin{aligned}
\boldsymbol{y}_{k+1}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^*\in\mathbb{R}^m} \left\{ -\psi(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}(\boldsymbol{x}_k + \theta_k[\boldsymbol{x}_k - \boldsymbol{x}_{k-1}])\rangle - \frac{1}{\sigma_k} D_{\psi/4m}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \right\}, \\
\boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\Delta_n} \left\{ \left\langle \boldsymbol{A}^*\boldsymbol{y}_{k+1}^*, \boldsymbol{x}\right\rangle + \frac{1}{\tau_k} D_{\mathcal{H}_n}(\boldsymbol{x}, \boldsymbol{x}_k) \right\} \\
\theta_{k+1} &= 1/\sqrt{1 + 4m\sigma_k}, \quad \tau_{k+1} = \tau_k/\theta_{k+1}, \quad \text{and} \quad \sigma_{k+1} = \theta_{k+1}\sigma_k.
\end{aligned}$$

The updates $\boldsymbol{y}_{k+1}^*$ and $\boldsymbol{x}_{k+1}$ can be both computed explicitly. For the first update, define the auxiliary variable

$$[\boldsymbol{w}_k]_i = \log\left(m[\boldsymbol{y}_k^*]_i/(1 - m[\boldsymbol{y}_k^*]_i)\right) \text{ for } i \in \{1, \dots m\}.$$

Then we can update $\boldsymbol{y}_{k+1}^*$ in two steps:

$$\boldsymbol{w}_{k+1} = \left(4m\sigma_k\boldsymbol{x}_k + 4m\sigma_k\theta_k\left(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right) + \boldsymbol{w}_k\right)/(1 + 4m\sigma_k)$$

and

$$[\boldsymbol{y}_{k+1}^*]_i = \frac{1}{m + me^{-[\boldsymbol{w}_{k+1}]_i}} \quad \text{for } i \in \{1, \dots, m\}.$$

For the second update, a straightforward calculation gives

$$[\boldsymbol{x}_{k+1}]_j = \frac{[\boldsymbol{x}_k]_j e^{-\tau_k[\boldsymbol{A}^*\boldsymbol{y}_{k+1}^*]_j}}{\sum_{j=1}^m [\boldsymbol{x}_k]_j e^{-\tau_k[\boldsymbol{A}^*\boldsymbol{y}_{k+1}^*]_j}}$$

for $j \in \{1, \ldots, n\}$. Hence the iterations are given by

$$\begin{aligned}
\boldsymbol{w}_{k+1} &= \left(4m\sigma_k \boldsymbol{x}_k + 4m\sigma_k\theta_k \left(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right) + \boldsymbol{w}_k\right)/(1 + 4m\sigma_k) \\
[\boldsymbol{y}^*_{k+1}]_i &= [\boldsymbol{y}^*_{k+1}]_i = \frac{1}{m + me^{-[\boldsymbol{w}_{k+1}]_i}} \quad \text{for } i \in \{1, \ldots, m\} \\
[\boldsymbol{x}_{k+1}]_j &= \frac{[\boldsymbol{x}_k]_j e^{-\tau_k [\boldsymbol{A}^* \boldsymbol{y}^*_{k+1}]_j}}{\sum_{j=1}^m [\boldsymbol{x}_k]_j e^{-\tau_k [\boldsymbol{A}^* \boldsymbol{y}^*_{k+1}]_j}} \quad \text{for } j \in \{1, \ldots, n\} \\
\theta_{k+1} &= 1/\sqrt{1 + 4m\sigma_k}, \quad \tau_{k+1} = \tau_k/\theta_{k+1}, \quad \text{and} \quad \sigma_{k+1} = \theta_{k+1}\sigma_k.
\end{aligned} \tag{3.9}$$

All parameter calculations and updates can be performed in $O(mn)$ operations. According to Proposition 2.4.2 and the optimality conditions (3.5), we have the strong limits

$$\lim_{k \to +\infty} \boldsymbol{y}^*_k = \boldsymbol{y}^*_s \quad \text{and} \quad \lim_{k \to +\infty} [\boldsymbol{y}^*_k]_i = \frac{1}{m + me^{-[\boldsymbol{A}\boldsymbol{x}_s]_i}} \text{ for } i \in \{1, \ldots, m\}.$$

The convergence is $O(1/k^2)$ in the number of iterations k, which is the best possible achievable rate of convergence for this problem in the Nesterov class of optimal first-order methods [188]. In particular, this means that for a given $\lambda > 0$ and $\epsilon > 0$, the nonlinear PDHG method provably computes an ϵ-approximate solution to a global minimum $\boldsymbol{v}^*(\lambda)$ of (3.1) in $O(mn/\sqrt{\epsilon})$ operations.

3.2.4 Numerical experiments

We present here some numerical experiments to compare the running times of the accelerated nonlinear PDHG methods proposed for sparse logistic regression to other commonly-used first-order optimization methods. These methods include the accelerated linear PDHG method [46, 47] and the forward-backward splitting method [23, 48]. These methods are described below and were implemented in MATLAB. All numerical experiments were performed on a single core Intel(R) Core(TM) i7-10750H CPU @ 2.60 GHz.

Data generation and optimization methods

We consider the setting where the m vectors of features $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ are independent and the true solution is sparse. Specifically, we draw m independent samples $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ from a d-dimensional Gaussian distribution with zero mean and unit variance. Letting $\boldsymbol{v} \in \mathbb{R}^n$ denote the true solution to be estimated, we set 1% of the coefficients of $\boldsymbol{v}$ to be equal to 10 and the other coefficients to be zero. Finally, letting $\boldsymbol{\xi}$ denote d-dimensional Gaussian distribution with zero mean and unit variance, we define the response model as

$$[\boldsymbol{b}]_i = \begin{cases} +1 & \text{if } \langle \boldsymbol{u}_i, \boldsymbol{v} \rangle + [\boldsymbol{\xi}]_i \geqslant 0, \\ -1 & \text{otherwise.} \end{cases}$$

This setting allows us to process dense, large-scale data sets with sparsity structure. We choose the number of samples to be smaller than then number of features, with $m = 10000$, and $d = 10000$, 25000, 50000, 75000, 100000, 125000 and 150000. We set the tuning parameter to be $\lambda = 1$.

We perform simulations using the accelerated nonlinear PDHG method (3.9), the accelerated linear PDHG method (2.2) described in the introduction, and the forward-backward splitting method as applied to problem (3.1). The initial values, parameters and numerical criteria for convergence of each method are described below.

Accelerated nonlinear PDHG method (3.9). We set $[\boldsymbol{y}_0^*]_i = 1/2m$ for each $i \in \{1, \ldots, m\}$, we set $[\boldsymbol{x}_{-1}]_j = [\boldsymbol{x}_0]_j = 1/n$ for each $j \in \{1, \ldots, n\}$, and we set $\tau_0 = 2m/\|\boldsymbol{A}\|_{1,2}^2$, $\sigma_0 = 1/2m$ and $\theta_0 = 0$. We compute the time required for convergence in the dual variable $\boldsymbol{y}_k^*$ and also the time required for convergence in the average dual variable $\boldsymbol{Y}^*$ as defined in Proposition 2.4.3. The iterations were stopped once $\|\boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*\|_2 \leqslant 10^{-4} \|\boldsymbol{y}_{k+1}^*\|_2$ and $\|\boldsymbol{Y}_{K+1}^* - \boldsymbol{Y}_K^*\|_2 \leqslant 10^{-4} \|\boldsymbol{Y}_{K+1}^*\|_2$.

Accelerated PDHG method (2.2). We set $[\boldsymbol{y}_0^*]_i = 1/2m$ for each $i \in \{1, \ldots, m\}$, we set $[\boldsymbol{v}_{-1}]_j = [\boldsymbol{v}_0]_j = 1/d$ for each $j \in \{1, \ldots, n\}$, and we set $\tau_0 = 4m/2\|\boldsymbol{A}\|_{2,2}^2$, $\sigma_0 = 1/2m$ and $\theta_0 = 0$. We evaluate the update in $\boldsymbol{w}_{k+1}$ using the forward-backward splitting method and we evaluate

the update $\boldsymbol{v}_{k+1}$ using the ℓ_1-ball projection algorithm described in Condat [60]. We compute the time required for convergence in the dual variable $\boldsymbol{y}_k^*$ and also the time required for convergence in the average dual variable $\boldsymbol{Y}^*$ as defined in Proposition 2.4.3. The iterations were stopped once $\left\|\boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*\right\|_2 \leqslant 10^{-4} \left\|\boldsymbol{y}_{k+1}^*\right\|_2$ and $\left\|\boldsymbol{Y}_{K+1}^* - \boldsymbol{Y}_K^*\right\|_2 \leqslant 10^{-4} \left\|\boldsymbol{Y}_{K+1}^*\right\|_2$.

Forward-backward splitting method. We compute the iterates

$$\begin{aligned}
\boldsymbol{w}_k &= \boldsymbol{v}_k + \beta_k(\boldsymbol{v}_k - \boldsymbol{v}_{k-1}),\\
\boldsymbol{v}_{k+1} &= \underset{\|\boldsymbol{v}\|_1 \leqslant \lambda}{\arg\min} \left\{\boldsymbol{v} - \left[\boldsymbol{w}_k - \tau \boldsymbol{B}^*/(m + me^{-\boldsymbol{B}\boldsymbol{w}_k})\right]\right\},\\
t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \quad \text{and} \quad \beta_{k+1} = \frac{(t_k - 1)}{t_{k+1}},
\end{aligned}$$

where $\tau = 4m/\left\|\boldsymbol{B}\right\|_{2,2}^2$, $q = \lambda\tau/(1 + \lambda\tau)$, $[\boldsymbol{v}_{-1}]_j = [\boldsymbol{v}_0]_j = 1/d$ for $j \in \{1, \ldots, d\}$ and $t_0 = \beta_0 = 0$. We evaluate the update $\boldsymbol{v}_{k+1}$ using the ℓ_1-ball projection algorithm described in Condat [60]. We compute the time required for convergence in the variable $\boldsymbol{v}_k$. The iterations were stopped once $\left\|\boldsymbol{v}_{k+1} - \boldsymbol{v}_k\right\|_1 \leqslant 10^{-4} \left\|\boldsymbol{v}_{k+1}\right\|_1$.

Numerical results. Table 3.2.1 shows the time results for the forward-backward splitting, linear PDHG and nonlinear PDHG methods. For the linear and nonlinear PDHG methods, we also show the time results for convergence with the regular and ergodic sequences as described before. We observe that the nonlinear PDHG method is considerably faster than both the forward-backward splitting and linear PDHG methods; the nonlinear PDHG method achieves a speedup of about 4 to 6.

3.3 Regularized maximum entropy estimation problems

Maximum entropy (Maxent) models are widely-used statistical models for estimating probability distributions from data. These models use the maximum entropy principle [146] to construct

	Number of features d						
	10000	25000	50000	75000	100000	125000	150000
Methods	Timings (s)						
Forward-backward splitting	40.55	114.82	269.25	437.52	725.81	839.24	1281.77
Linear PDHG (Reg.)	40.56	111.60	254.41	408.79	670.57	739.37	1122.36
Linear PDHG (Erg.)	46.96	126.51	284.25	447.06	717.33	810.46	1180.35
Nonlinear PDHG (Reg.)	9.72	26.23	58.60	87.67	112.95	177.50	203.10
Nonlinear PDHG (Erg.)	13.52	32.47	64.71	93.97	125.65	190.25	193.36

Table 3.2.1: Time results (in seconds) for solving the ℓ_1-restricted logistic regression problem (3.1) with the forward-backward and linear PDHG methods and time results for solving the equivalent problem (3.4) with the nonlinear PDHG method.

probability distributions that reproduce key statistics of data sets as closely as possible. Historically applied to problems in physics [146], engineering [149, 129, 25] and statistics [247], Maxent models are now frequently applied to problems in natural language processing [26, 55, 76, 172, 211, 240], social science [132, 159, 184], neuroscience [122, 238, 218] and ecological modeling [87, 88, 94, 148, 180, 200, 201, 202, 223, 222], among others.

Large-scale Maxent models require estimating probability distributions from massive data sets comprising hundred of thousands to billions of features [171]. Due to this enormous number, large-scale Maxent models need efficient and robust algorithms to perform well. However, state-of-the-art algorithms for Maxent models were not originally designed to handle massive data sets. These algorithms either rely on technical devices that may yield unreliable numerical results [105], or lack scalable parallelism or scale poorly in size [73, 76] or depend on assumptions (e.g., smoothness properties) that are not satisfied by many Maxent models in practice [172]. These limitations make it practically impossible to construct large-scale Maxent models for applications without adequate and costly computational resources [77, 233].

Further exacerbating this issue is that machine learning applications, in general, strongly rely on increases in computing power to manage growing data sets and improve performance [80]. Without more efficient and robust algorithms to minimize monetary and energy costs, this progress will quickly become economically and environmentally unsustainable as computational requirements become a severe constraint [233]. This constraint on computational requirements, in particular, has been recently identified as a crucial challenge to overcome for Maxent models used in climate

change ecological studies in order to create realistic environmental predictors within a reasonable amount of run time [223, 222].

This sections proposes accelerated nonlinear PDHG methods designed to overcome the shortcomings of present state-of-the-art methods used for constructing large-scale Maxent models. These methods, as we will describe soon, provably compute solutions to a broad class of Maxent models, including models with ℓ_1 and ℓ_2^2 penalties, with computations requiring on the order of $O(mn/\sqrt{\epsilon})$ or $O(mn\log(1/\epsilon))$ operations (the order depending on strong convexity assumptions), where $\epsilon \in (0,1)$ denotes the tolerance, m denote the number of features in the Maxent model, and n denote the dimensionality of the Maxent model. This result improves on the known complexity bound of $O(\min(m^2n, mn^2)/\sqrt{\epsilon})$ and $O(\min(m^2n, mn^2)/\log(1/\epsilon))$ for first-order optimization methods such as the linear PDHG or forward-backward splitting methods. These gains turn out to be considerable in practice: for instance, in Section 3.3.4 we present some numerical experiments in which an accelerated nonlinear PDHG method tailored to ℓ_2^2-regularized maximum entropy estimation converges 3-4 times faster than the classical accelerated linear PDHG method.

3.3.1 Description of the problem

Suppose we receive l independent and identically distributed samples $\{v_1, ..., v_l\} \subset \mathcal{I}$ from an unknown distribution $\mathcal{D}$. We assume throughout this section that the input space $\mathcal{I}$ is discrete with n elements, and without loss of generality $\mathcal{I} = \{1, \ldots, n\}$. In addition, suppose we are given some prior probability distribution p_{prior} on $\mathcal{I}$ that encapsulates some prior knowledge about the samples or unknown distribution. Finally, suppose we have access to a set of features from the samples via a bounded feature map $\mathbf{\Phi}\colon \mathcal{I} \to \mathbb{R}^m$, with $\sup_{j\in\{1,\ldots,n\}} \|\mathbf{\Phi}(j)\|_2 \leqslant r$ for some $r > 0$. Then, how do we estimate the unknown distribution $\mathcal{D}$ from the prior distribution p_{prior}, the samples $\{v_1, ..., v_l\}$ and the feature map $\mathbf{\Phi}$?

The maximum entropy principle offers a way to answer this question. It states that the distribution that best estimates the unknown distribution $\mathcal{D}$ is the one that remains as close as possible to the prior probability p_{prior} while matching the features $\{\mathbf{\Phi}(v_1), \ldots, \mathbf{\Phi}(v_l)\}$ exactly or as closely

as possible, in some suitable sense. We measure closeness of a probability distribution $p \in \Delta_n$, where Δ_n denote the n-dimensional probability simplex, to the prior probability $p_{\text{prior}} \in \Delta_n$ using the Kullback–Leibler divergence:

$$D_{\mathcal{H}_n}(p, p_{\text{prior}}) = \sum_{j=1}^{n} p(j) \log \left(\frac{p(j)}{p_{\text{prior}}(j)} \right). \tag{3.10}$$

The symbol $\mathcal{H}_n$ stands for the negative entropy function with respect to the probability simplex Δ_n. We measure how the average of the features induced by a probability distribution p match the empirical average of the features $\{\mathbf{\Phi}(v_1), \ldots, \mathbf{\Phi}(v_l)\}$ as follows. Let $\hat{\mathcal{D}}$ denote the empirical distribution induced by samples $\{v_1, ..., v_l\}$, that is,

$$\hat{\mathcal{D}}(j) = \frac{1}{l} \left| \{1 \leqslant i \leqslant l \mid v_i = j\} \right|. \tag{3.11}$$

Let $\mathbb{E}_p[\mathbf{\Phi}]$ denote the average induced by the probability distribution p and let $\mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}]$ denote the empirical average induced by the samples $\{v_1, ..., v_l\}$, that is,

$$\mathbb{E}_p[\mathbf{\Phi}] = \sum_{j=1}^{n} p(j)\mathbf{\Phi}(j).$$

and

$$\mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}] = \sum_{j=1}^{n} \hat{\mathcal{D}}(j)\mathbf{\Phi}(j).$$

Formally, we measure how the averages $\mathbb{E}_p[\mathbf{\Phi}]$ and $\mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}]$ are close to each other via an arbitrary proper, lower-semicontinuous and convex function $H^* \colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$. Maxent models seek to minimize the sum

$$\inf_{p \in \Delta_n} f(p; t) = \inf_{p \in \Delta_n} \left\{ D_{\mathcal{H}_n}(p, p_{\text{prior}}) + tH^* \left(\frac{\mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}] - \mathbb{E}_p[\mathbf{\Phi}]}{t} \right) \right\} \tag{3.12}$$

where $t > 0$ is a free parameter that is typically either chosen by the user or estimated using cross-validation with testing data. The function H^* is also typically assumed to be finite at $\mathbf{0}$, meaning that the probability distribution $p = \hat{\mathcal{D}}$ is a feasible point of the optimization problem. Following the ecological modeling literature, we will call H^* the potential function of the Maxent

model (3.12).

Dual formulation

The generalized Maxent problem (3.12) admits a dual problem that corresponds to regularized maximum a posteriori estimation [6]. To derive the dual problem, we first write the second term on the right hand side of (3.12) in terms of its convex conjugate:

$$tH^*\left(\frac{\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]-\mathbb{E}_p[\boldsymbol{\Phi}]}{t}\right)=\sup_{\boldsymbol{w}\in\mathbb{R}^m}\left\{\left\langle \boldsymbol{w},\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]-\mathbb{E}_p[\boldsymbol{\Phi}]\right\rangle - tH(\boldsymbol{w})\right\},$$

where we abused the notation to write the convex conjugate of H^* as H. We can now express problem (3.12) in saddle-point form as

$$\inf_{p\in\Delta_n}\sup_{\boldsymbol{w}\in\mathbb{R}^m}\left\{\left\langle \boldsymbol{w},\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]-\mathbb{E}_p[\boldsymbol{\Phi}]\right\rangle - tH(\boldsymbol{w})+D_{\mathcal{H}_n}(p,p_{\text{prior}})\right\}. \tag{3.13}$$

Assuming that the potential function H^* is proper, lower semicontinuous, convex and finite at $\mathbf{0}$, the infimum and supremum can be swapped [92, Page 61, Statement (4.1)]. In that case, we can use the convex conjugate formula

$$\inf_{p\in\Delta_n}\left\{D_{\mathcal{H}_n}(p,p_{\text{prior}})-\langle \boldsymbol{w},\mathbb{E}_p[\boldsymbol{\Phi}]\rangle\right\}=-\log\left[\sum_{j=1}^{n}p_{\text{prior}}(j)e^{\langle \boldsymbol{w},\boldsymbol{\Phi}(j)\rangle}\right]$$

to obtain the dual problem of (3.12):

$$\sup_{\boldsymbol{w}\in\mathbb{R}^m}\left\{\left\langle \boldsymbol{w},\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]\right\rangle - tH(\boldsymbol{w})-\log\left[\sum_{j=1}^{n}p_{\text{prior}}(j)e^{\langle \boldsymbol{w},\boldsymbol{\Phi}(j)\rangle}\right]\right\}. \tag{3.14}$$

The dual problem (3.14) is a regularized maximum likelihood estimation problem over the family of Gibbs distributions [6, 181].

Examples of Maxent models

Different Maxent models vary in the choice of the prior distribution p_{prior}, the potential function H^*, and the free parameter t. In most models used in practice, the prior distribution is the uniform distribution and the free parameter is either pre-selected or trained over testing data. The choice of the potential function depends on the application. We give here three examples of potential functions. The first example is the indicator function

$$\boldsymbol{u} \mapsto H^*(\boldsymbol{u}) = \begin{cases} 0, \text{ if } \boldsymbol{u} = \mathbf{0}, \\ \\ +\infty \text{ otherwise.} \end{cases}$$

This potential function yields the classical maximum entropy estimation problem

$$\inf_{p \in \Delta_n} D_{\mathcal{H}_n}(p, p_{\text{prior}}) \quad \text{such that} \quad \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] = \mathbb{E}_p[\boldsymbol{\Phi}]. \tag{3.15}$$

The second example is the conjugate of the ℓ_1-norm, which is the characteristic set of the unit ball defined with respect to the ℓ_1 norm:

$$\boldsymbol{u} \mapsto H^*(\boldsymbol{u}) = \|\boldsymbol{u}\|_1^* = \{v \in \mathbb{R}^m \mid |u_j - v_j| \leqslant 1 \text{ for } j \in \{1, \ldots, m\}\}.$$

This potential function yields the ℓ_1-regularized maximum entropy estimation problem

$$\min_{p \in \Delta_n} \left\{ D_{\mathcal{H}_n}(p, p_{\text{prior}}) + t \left\| \frac{\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}]}{t} \right\|_1^* \right\}. \tag{3.16}$$

This Maxent model allows the average $\mathbb{E}_p[\boldsymbol{\Phi}]$ to be close to the empirical average $\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]$ without having to be equal to it. This model has been extensively studied and is frequently applied to problems in natural language processing and ecological modeling [240, 88, 61, 181]. The corresponding dual problem is

$$\sup_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \left\langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] \right\rangle - t \|\boldsymbol{w}\|_1 - \log \left[\sum_{j=1}^{n} p_{\text{prior}}(j) e^{\langle \boldsymbol{w}, \boldsymbol{\Phi}(j) \rangle} \right] \right\}. \tag{3.17}$$

The third example is the quadratic $\boldsymbol{u} \mapsto H^*(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{u}\|_2^2$. This potential function yields the ℓ_2^2-regularized maximum entropy estimation problem

$$\min_{p \in \Delta_n} \left\{ D_{\mathcal{H}_n}(p, p_{\text{prior}}) + \frac{1}{2t} \left\| \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}] \right\|_2^2 \right\}. \tag{3.18}$$

This Maxent model, like the ℓ_1-regularized Maxent model (3.16), has been extensively studied and applied to problems in natural language processing and ecological modeling [55, 160, 88, 181]. The corresponding dual problem is

$$\sup_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \left\langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] \right\rangle - \frac{t}{2}\|\boldsymbol{w}\|_2^2 - \log \left[\sum_{j=1}^{n} p_{\text{prior}}(j) e^{\langle \boldsymbol{w}, \boldsymbol{\Phi}(j) \rangle} \right] \right\}. \tag{3.19}$$

Challenges in computing large-scale Maxent models

Estimating a probability distribution from the Maxent model (3.12) can be prohibitively expensive for big data sets. To see why, fix $t > 0$ and suppose that the Maxent model (3.12) has a global solution $p_s(t)$. Let $p_\epsilon(t) \in \Delta_n$ with $\epsilon > 0$ denote an ϵ-approximate solution to the global solution $p_s(t)$, that is, the objective function f in (3.12) satisfies

$$f(p_\epsilon(t); t) - f(p_s(t); t) < \epsilon.$$

Then for a strongly convex potential function H^*, the best achievable rate of convergence for computing $p_\epsilon(t)$ in the Nesterov class of optimal first-order methods is linear, $O(\log(1/\epsilon))$, in the number of iterations [187]. Without strong convexity, the optimal rate convergence is sublinear, $O(1/\sqrt{\epsilon})$, in the number of iterations. While optimal, these rates of convergence can only be achieved if a precise estimate of the largest singular value of the linear operator $\boldsymbol{A}\colon \Delta_n \to \mathbb{R}^m$ defined by

$$\boldsymbol{A}p = \sum_{j=1}^{n} p(j)\boldsymbol{\Phi}(j) = \mathbb{E}_p[\boldsymbol{\Phi}] \tag{3.20}$$

is available.

Unfortunately, this quantity is essentially impossible to compute for large matrices due to its prohibitive computational cost of $O(\min(m^2n, mn^2))$ operations [130]. This issue generally makes solving the Maxent model (3.12) difficult and laborious. Even worse, in some applications the appropriate value of the free parameter t in (3.12) is difficult to guess and must be selected by repeatedly solving (3.12) from a large pool of values of t, a process that can become particularly time consuming and resource intensive for big data sets.

3.3.2 State-of-the-art optimization methods

Estimating probability distributions from large-scale Maxent models has driven much research in the development of robust and efficient algorithms to minimize computational costs and maximum model performance. The current state of the art is based on a technical device called infinitely weighted logistic regression (IWLR) [105, 202], a technical device that makes it possible to fit Maxent models using coordinate descent algorithms [111, 112, 131]. The IWLR method is implemented, for instance, in the Maxent package available in the R programming language [202], and it is widely used by the ecological modeling community. Other popular methods for solving Maxent models include those based on limited-memory BFGS algorithms [172, 7] and first-order optimization algorithms such as forward-backward splitting [23, 48, 74]. We focus here on these methods, but we wish to mention that many more algorithms have been developed to estimate probability densities from the general Maxent model (3.12) (see, e.g., [86, 172, 173, 181] for surveys and comparisons of different algorithms).

The IWLR method is considered the state of the art because it makes it possible to fit the Maxent model (3.12) as if it were a logistic regression model. This technical device makes it possible to fit Maxent models using coordinate descent algorithms [111, 112, 131]. Coordinate descent algorithms have been popular for fitting logistic regression models because they are generally scalable, with steps in the algorithms having at worse an asymptotic space complexity of $O(mn)$ operations. However, despite these advantages, IWLR is an approximate technical device that may yield unreliable numerical results and that largely depends on coordinate descent algorithms

to perform well. Coordinate descent algorithms themselves are generally non-parallelizable and may lack robustness and good convergence properties. The coordinate descent implementation in the popular glmnet software package, for instance, depends on the sparsity of the matrix $\boldsymbol{A}$ to converge quickly [263]. It would be desirable to have a fast optimization method for when the matrix $\boldsymbol{A}$ is dense, as this often occurs in practice. Another issue is that the glmnet implementation approximates the logarithm term in logistic regression models with a quadratic term to fit them efficiently. Without costly step-size optimization, which glmnet avoids to improve performance, the glmnet implementation may not converge [112, 161]. Case in point, Yuan et al. [255] provides two numerical experiments in which glmnet does not converge.

The limited-memory BFGS is an iterative algorithm that uses an estimate of the inverse Hessian matrix to solve sufficiently smooth optimization problems. This algorithm, in particular, has been called the algorithm of choice for solving the ℓ_2^2-regularized Maxent model (3.18) [172, 7]. It has been proposed as a faster alternative to iterative scaling methods [73, 76] for constructing Maxent models as well. The limited-memory BFGS method requires some degree of smoothness and differentiability to work, which may not be present in a given Maxent model, but some variants of this method do not require differentiability of the objective function [7]. The main disadvantage of the limited-memory BFGS method is that it requires to be initialized somewhat close to the true solution to converge quickly. Without this, it may converge slowly and fail to be competitive compared to other methods, such as coordinate descent, or fail outright to converge [254].

The forward-backward splitting algorithm and its variants are widely used because they are robust and can provably compute ϵ-approximate solutions of (3.12) (under appropriate conditions on the potential function H^*) with an optimal rate of convergence. To achieve this convergence rate, however, the step size parameter in the algorithm needs to be fine-tuned using a precise estimate of the largest singular value of the matrix of features $\boldsymbol{A}$ (3.20). As mentioned before, however, computing this estimate is essentially impossible for large matrices due to its prohibitive computational cost, which has an asymptotic computational complexity of at most $O(\min(m^2n, mn^2))$ operations. Line search methods and other heuristics are often employed to bypass this problem, but they slow down the convergence of the forward-backward splitting algorithm. Another ap-

proach is to compute a crude estimate of the largest singular value of the matrix $\boldsymbol{A}$, but doing so significantly reduces the speed of convergence. This problem makes the forward-backward splitting algorithm generally inefficient and impractical to estimate probability densities from large-scale Maxent models.

In summary, state-of-the-art and other widely used algorithms for estimating probability densities from Maxent models either scale poorly in size or may fail to converge or may be prone to produce unreliable numerical results. These shortcomings in terms of efficiency and robustness make it challenging to use large-scale Maxent models without access to adequate and costly computational resources. We shall present here efficient and robust accelerated nonlinear PDHG optimization methods that address these shortcomings.

3.3.3 Derivation of the accelerated nonlinear PDHG method

We present here accelerated nonlinear PDHG methods for solving the generalized maximum entropy estimation problem (3.12). In terms of the abstract primal and dual problems (2.7) and (2.8) from Chapter 2, we set the real reflexive Banach spaces to be $\mathcal{X} = (\mathbb{R}^n, \|\cdot\|_1)$, $\mathcal{Y} = (\mathbb{R}^m, \|\cdot\|_2)$, we set the functions g and h as

$$p \mapsto g(\boldsymbol{p}) = \begin{cases} D_{\mathcal{H}_n}(p, p_{\text{prior}}) & \text{if } p \in \Delta_n, \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$\boldsymbol{u} \mapsto h(\boldsymbol{u}) = tH^* \left(\frac{\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \boldsymbol{u}}{t} \right)$$

for arbitrary $t > 0$. We assume that H^* is a proper, lower semicontinuous, convex function that is finite at $\mathbf{0}$. This ensures that assumption (A1) from Chapter 2 holds for any $t > 0$ [92, Page 61, Statement (4.1)][136, Propositions 2.2.1 and 2.2.2]. For the Bregman functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$, we set

$$p \mapsto \phi_{\mathcal{X}}(p) = \sum_{j=1}^{n} p(j) \log p(j)$$

and

$$\boldsymbol{w} \mapsto \phi_{\mathcal{Y}^*}(\boldsymbol{w}) = \frac{1}{2} \|\boldsymbol{w}\|_2^2 .$$

Implicit in the choice of $\phi_{\mathcal{Y}^*}$ is that the linear proximal operator

$$\inf_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \frac{1}{2} \|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2^2 + tH(\boldsymbol{w}) \right\}$$

can be computed explicitly or can be quickly evaluated numerically. Depending on the form of the potential function H^*, it may be more advantageous to use a different Bregman function, potentially with a different norm on $\mathbb{R}^m$. Nonetheless, for simplicity and the forthcoming numerical experiments we fix $\phi_{\mathcal{Y}^*}$ to be a quadratic.

With these choices, assumptions (A1)–(A6) from Chapter 2 hold with $\gamma_g = 1$. Indeed, assumption (A6) holds because the Bregman function $\phi_{\mathcal{X}}$ naturally induces the Kullback–Leibler divergence as its Bregman divergence and because the Kullback–Leibler divergence is 1-strongly convex with respect to the ℓ_1 norm over the unit simplex Δ_n. As mentioned in Section 3.2, this fact is a direct consequence of a fundamental result in information theory known as Pinsker's inequality [22, 63, 152, 158, 203]. Moreover, the induced operator norm is the maximum ℓ_2 norm of the columns of $\boldsymbol{A}$, i.e.,

$$\|\boldsymbol{A}\|_{\text{op}} = \|\boldsymbol{A}\|_{1,2} = \sup_{\|\boldsymbol{p}\|_1 = 1} \|\boldsymbol{A}p\|_2 = \max_{j \in \{1,\dots,n\}} \|\boldsymbol{\Phi}(j)\|_2 .$$

Note that the induced operator norm is always bounded because the map $\boldsymbol{\Phi}$ defining the features is bounded with respect to the input space $\mathcal{I}$. It can also be computed in *optimal* $\Theta(mn)$ time. This is unlike in most first-order optimization methods, such as the forward-backward splitting algorithm, where instead the equivalent operator norm is the largest singular value of $\boldsymbol{A}$, which takes $O(\min(m^2 n, mn^2))$ operations to compute. This point is *crucial*: the smaller computational cost makes it efficient to compute the operator norm and all subsequent parameters of the accelerated nonlinear PDHG method, which is needed to achieve an optimal rate of convergence.

From there, the choice of accelerated nonlinear PDHG method depends on whether the potential

function H^* is strongly smooth or not. We describe below the appropriate method for each case.

Accelerated nonlinear PDHG method for non-strongly smooth potential function H^*

For this case, we solve the regularized Maxent problem (3.12) using the accelerated nonlinear PDHG method (2.19) with $\gamma_g = 1$. We set the initial stepsize parameters to be $\theta_0 = 0$, $\sigma_0 > 0$ and $\tau_0 = 1/(\|\boldsymbol{A}\|_{1,2}^2\,\sigma_0)$, and we pick an initial probability distribution $p_0 \in \operatorname{int}\Delta_n$ and initial values $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 \in \mathbb{R}^m$. The corresponding accelerated nonlinear PDHG method for (3.12) consists of the iterations

$$
\begin{aligned}
p_{k+1} &= \operatorname*{arg\,min}_{p\in\Delta_n} \left\{ D_{\mathcal{H}_n}(p, p_{\text{prior}}) + \left\langle \boldsymbol{w}_k + \theta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}), \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}] \right\rangle + \frac{1}{\tau_k} D_{\mathcal{H}_n}(p, p_k) \right\} \\
\boldsymbol{w}_{k+1} &= \operatorname*{arg\,max}_{\boldsymbol{w}\in\mathbb{R}^m} \left\{ -tH(\boldsymbol{w}) + \left\langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}] \right\rangle - \frac{1}{2\sigma_k} \|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2 \right\} \\
\theta_{k+1} &= 1/\sqrt{1+\tau_k}, \quad \tau_{k+1} = \theta_{k+1}\tau_k, \quad \sigma_{k+1} = \sigma_k/\theta_{k+1},
\end{aligned}
$$

where we omit the dependence of the probability distribution p on $j \in \{1, \ldots, n\}$.

The update p_{k+1} can be computed explicitly. First, let

$$
\boldsymbol{z}_k = \boldsymbol{w}_k + \theta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1})
$$

and introduce a Lagrangian variable ξ to express the first update as

$$
p_{k+1} = \operatorname*{arg\,min}_{\substack{p\in(0,+\infty)^n \\ \xi\in\mathbb{R}}} \left\{ D_{\mathcal{H}_n}(p, p_{\text{prior}}) - \langle \boldsymbol{z}_k, \mathbb{E}_p[\boldsymbol{\Phi}] \rangle + \frac{1}{\tau_k} D_{\mathcal{H}_n}(p, p_k) - \xi \left(1 - \sum_{j=1}^n p(j) \right) \right\}.
$$

Taking the gradient with respect to p and setting it to zero gives the optimality condition

$$
\log\left(p_{k+1}/p_{\text{prior}}\right) + 1 - \langle \boldsymbol{z}_k, \boldsymbol{\Phi} \rangle + \frac{1}{\tau_k} \left[\log\left(p_{k+1}/p_k\right) + 1 \right] + \xi = 0
$$

We can rearrange this as

$$(1+\tau_k)\log(p_{k+1}) = \tau_k \log(p_{\text{prior}}) + \log(p_k) - \tau_k(1+\xi - \langle \boldsymbol{z}_k, \boldsymbol{\Phi}\rangle) - 1.$$

Solving for p_{k+1} yields

$$p_{k+1} = p_{\text{prior}}^{\tau_k/(1+\tau_k)} p_k^{1/(1+\tau_k)} e^{-\tau_k(1+\xi-\langle \boldsymbol{z}_k,\boldsymbol{\Phi}\rangle)/(1+\tau_k)-1/1+\tau_k}.$$

Since the probabilities sum to one, we must have, for every $j \in \{1,\dots,n\}$,

$$p_{k+1}(j) = \frac{p_{\text{prior}}(j)^{\tau_k/(1+\tau_k)} p_k(j)^{1/(1+\tau_k)} e^{\tau_k\langle \boldsymbol{z}_k,\boldsymbol{\Phi}(j)\rangle/(1+\tau_k)}}{\sum_{j=1}^n p_{\text{prior}}(j)^{\tau_k/(1+\tau_k)} p_k(j)^{1/(1+\tau_k)} e^{\tau_k\langle \boldsymbol{z}_k,\boldsymbol{\Phi}(j)\rangle/(1+\tau_k)}}.$$

For the second update $\boldsymbol{w}_{k+1}$, we can express it as the proximal mapping

$$\boldsymbol{w}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^m} \left\{ \frac{1}{2} \left\| \boldsymbol{w} - \left(\boldsymbol{w}_k + \sigma_k \left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}] \right] \right) \right\|_2^2 + t\sigma_k H(\boldsymbol{w}) \right\}$$

Hence the iterations are given by

$$\begin{aligned}
\boldsymbol{z}_k &= \boldsymbol{w}_k + \theta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}), \\
p_{k+1}(j) &= \frac{p_{\text{prior}}(j)^{\tau_k/(1+\tau_k)} p_k(j)^{1/(1+\tau_k)} e^{\tau_k\langle \boldsymbol{z}_k,\boldsymbol{\Phi}(j)\rangle/(1+\tau_k)}}{\sum_{j=1}^n p_{\text{prior}}(j)^{\tau_k/(1+\tau_k)} p_k(j)^{1/(1+\tau_k)} e^{\tau_k\langle \boldsymbol{z}_k,\boldsymbol{\Phi}(j)\rangle/(1+\tau_k)}} \quad \text{for } j \in \{1,\dots,n\}, \\
\boldsymbol{w}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^m} \left\{ \frac{1}{2} \left\| \boldsymbol{w} - \left(\boldsymbol{w}_k + \sigma_k \left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}] \right] \right) \right\|_2^2 + t\sigma_k H(\boldsymbol{w}) \right\}, \\
\theta_{k+1} &= 1/\sqrt{1+\tau_k}, \quad \tau_{k+1} = \theta_{k+1}\tau_k, \quad \sigma_{k+1} = \sigma_k/\theta_{k+1}.
\end{aligned} \tag{3.21}$$

The iterations in (3.21) simplify for the choice of initial probability $p_0 = p_{\text{prior}}$. Let

$$\hat{\boldsymbol{z}}_k = \tau_k \left(\hat{\boldsymbol{z}}_{k-1} + (\boldsymbol{w}_k + \theta_k[\boldsymbol{w}_k - \boldsymbol{w}_{k-1}]) \right) / (1+\tau_k)$$

with $\hat{\boldsymbol{z}}_{-1} = \boldsymbol{0}$. Then we have

$$p_{k+1}(j) = \frac{p_{\text{prior}}(j) e^{\langle \hat{\boldsymbol{z}}_k,\boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle \hat{\boldsymbol{z}}_k,\boldsymbol{\Phi}(j)\rangle}} \quad \text{for } j \in \{1,\dots,n\}. \tag{3.22}$$

To show this, we induct on k. For the update p_1, a short calculation gives

$$p_1(j) = \frac{p_{\text{prior}}(j)e^{\langle \tau_0 \boldsymbol{w}_0/(1+\tau_0), \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_0, \boldsymbol{\Phi}(j)\rangle}} = \frac{p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_0, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_0, \boldsymbol{\Phi}(j)\rangle}} \quad \text{for } j \in \{1, \dots, n\}.$$

Suppose that this holds for $k \in \mathbb{N}$, namely

$$p_k(j) = \frac{p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_{k-1}, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_{k-1}, \boldsymbol{\Phi}(j)\rangle}} \quad \text{for } j \in \{1, \dots, n\}.$$

The next update is then proportional to

$$p_{k+1}(j) \propto p_{\text{prior}}(j)e^{\tau_k \langle (\hat{\boldsymbol{z}}_{k-1} + (\boldsymbol{w}_k + \theta_k[\boldsymbol{w}_k - \boldsymbol{w}_{k-1}]))/(1+\tau_k), \boldsymbol{\Phi}(j)\rangle} \quad \text{for } j \in \{1, \dots, n\}.$$

The term on the left hand side of the inner product is the term $\hat{\boldsymbol{z}}_k$. Since the probabilities sum to one, the update for the probabilities is then precisely (3.22). Hence, for the choice of $p_0 = p_{\text{prior}}$ the iterations for this generalized Maxent problem are given by

$$\begin{aligned}
\hat{\boldsymbol{z}}_k &= \tau_k \left(\hat{\boldsymbol{z}}_{k-1} + (\boldsymbol{w}_k + \theta_k[\boldsymbol{w}_k - \boldsymbol{w}_{k-1}])\right)/(1+\tau_k), \\
p_{k+1}(j) &= \frac{p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_k, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j)e^{\langle \hat{\boldsymbol{z}}_k, \boldsymbol{\Phi}(j)\rangle}} \quad \text{for } j \in \{1, \dots, n\}, \\
\boldsymbol{w}_{k+1} &= \underset{\boldsymbol{w} \in \mathbb{R}^m}{\arg\min} \left\{ \frac{1}{2} \left\| \boldsymbol{w} - \left(\boldsymbol{w}_k + \sigma_k \left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}] \right] \right) \right\|_2^2 + t\sigma_k H(\boldsymbol{w}) \right\}, \\
\theta_{k+1} &= 1/\sqrt{1+\tau_k}, \quad \tau_{k+1} = \theta_{k+1}\tau_k, \quad \sigma_{k+1} = \sigma_k/\theta_{k+1}.
\end{aligned} \tag{3.23}$$

where $\hat{\boldsymbol{z}}_{-1} = \mathbf{0}$. All parameters calculations and updates can be performed in $O(mn)$ operations. The convergence is $O(1/k^2)$ in the number of iterations k, which is the best possible achievable rate of convergence for this problem in the Nesterov class of optimal first-order methods [187]. In particular, this means that for a given $t > 0$ and $\epsilon > 0$, this nonlinear PDHG method provably computes an ϵ-approximation solution to a global minimum $p_\epsilon(t)$ of the regularized Maxent problem (3.12) in $O(mn/\sqrt{\epsilon})$ operations.

Accelerated nonlinear PDHG method for strongly smooth potential function H^*

For this case, we solve the regularized Maxent problem (3.12) using the accelerated nonlinear PDHG method (2.42) with $\gamma_g = 1$ and $\gamma_{h^*} > 0$. We set the initial stepsize parameters to be

$$\theta = 1 - \frac{\gamma_g \gamma_{h^*}}{2\left\|\boldsymbol{A}\right\|_{1,2}^2}\left(\sqrt{1 + \frac{4\left\|\boldsymbol{A}\right\|_{1,2}^2}{\gamma_g \gamma_{h^*}}} - 1\right), \quad \tau = \frac{1-\theta}{\gamma_g \theta}, \quad \text{and} \quad \sigma = \frac{1-\theta}{\gamma_{h^*}\theta},$$

and we pick an initial probability distribution $p_0 \in \operatorname{int} \Delta_n$ and initial values $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 \in \mathbb{R}^m$. The corresponding accelerated nonlinear PDHG method for (3.12) can be derived as in the non-strongly smooth case. It consists of the iterations

$$\begin{aligned}
\boldsymbol{z}_k &= \boldsymbol{w}_k + \theta(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}), \\
p_{k+1} &= \frac{p_{\text{prior}}(j)^{\tau/(1+\tau)} p_k(j)^{1/(1+\tau)} e^{\tau\langle \boldsymbol{z}_k, \boldsymbol{\Phi}(j)\rangle/(1+\tau)}}{\sum_{j=1}^n p_{\text{prior}}(j)^{\tau/(1+\tau)} p_k(j)^{1/(1+\tau)} e^{\tau\langle \boldsymbol{z}_k, \boldsymbol{\Phi}(j)\rangle/(1+\tau)}}, \\
\boldsymbol{w}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^m} \left\{ \frac{1}{2}\left\| \boldsymbol{w} - \left(\boldsymbol{w}_k + \sigma\left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}]\right]\right)\right\|_2^2 + t\sigma H(\boldsymbol{w}) \right\}.
\end{aligned} \tag{3.24}$$

As before, the iterations in (3.24) simplify for the choice of initial probability $p_0 = p_{\text{prior}}$. In this case, we have

$$\begin{aligned}
\hat{\boldsymbol{z}}_k &= \tau\left(\hat{\boldsymbol{z}}_{k-1} + (\boldsymbol{w}_k + \theta[\boldsymbol{w}_k - \boldsymbol{w}_{k-1}])\right)/(1+\tau), \\
p_{k+1}(j) &= \frac{p_{\text{prior}}(j) e^{\langle \hat{\boldsymbol{z}}_k, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle \hat{\boldsymbol{z}}_k, \boldsymbol{\Phi}(j)\rangle}} \quad \text{for } j \in \{1, \ldots, n\}, \\
\boldsymbol{w}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^m} \left\{ \frac{1}{2}\left\| \boldsymbol{w} - \left(\boldsymbol{w}_k + \sigma\left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}]\right]\right)\right\|_2^2 + t\sigma H(\boldsymbol{w}) \right\}
\end{aligned} \tag{3.25}$$

where $\hat{\boldsymbol{z}}_{-1} = \boldsymbol{0}$. All parameters calculations and updates can be performed in $O(mn)$ operations. The convergence is $O(\theta^k)$ in the number of iterations k, which is the best possible achievable rate of convergence for this problem in the Nesterov class of optimal first-order methods [187]. In particular, this means that for a given $t > 0$ and $\epsilon > 0$, this nonlinear PDHG method provably computes an ϵ-approximation solution to a global minimum $p_\epsilon(t)$ of the regularized Maxent problem (3.12) in $O(mn \log(1/\epsilon))$ operations.

3.3.4 Numerical experiments

We present some numerical experiments to compare the running times of the accelerated nonlinear PDHG methods proposed for regularized maximum entropy estimation to the accelerated linear PDHG method [46, 47]. We consider here regularized Maxent for which the potential function is a quadratic. All methods are described below and were implemented in MATLAB. All numerical experiments were performed on a Laptop with single core Intel(R) Core(TM) i7-10750H CPU @ 2.60 GHz.

Data generation and optimization methods

We generate $l = 200,000$ independent and identically distributed outcomes from a binomial distribution with parameters $(n-1)$ and $p = 0.5$. These outcomes are used to construct an empirical distribution $\hat{\mathcal{D}}$ (see (3.11)) with support in $\{1, \ldots, n\}$. For the feature maps $\boldsymbol{\Phi}(j)$, we generate n independent and identically distributed m-dimensional Gaussian vectors with zero mean and normalized so that $\|\boldsymbol{\Phi}(j)\|_2 = 1$ for every $j \in \{1, \ldots, n\}$. We select the prior probability p_{prior} to be uniform, that is, $p_{\text{prior}}(j) = 1/n$ for each $j \in \{1, \ldots, n\}$. We set the regularization parameter $t = 0.0025$ and we choose the dimensionality n to be smaller than the number of features m, with $n = 1000$ and $m = 100000$, 250000, 500000, 750000, 1000000, 1250000 and 1500000.

We perform simulations using the accelerated nonlinear PDHG method (3.25) with $H = \frac{1}{2}\|\cdot\|_2^2$ and its accelerated linear PDHG method counterpart. The initial values, parameters, and numerical criteria for convergence of each method are described below.

Accelerated nonlinear PDHG method (3.25) with $H = \frac{1}{2}\|\cdot\|_2^2$. We set $p_0(j) = p_{\text{prior}}(j) = 1/n$ for each $j \in \{1, \ldots, n\}$, we set $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 = \mathbf{0} \in \mathbb{R}^m$. For the parameters, we set

$$\theta = 1 - \frac{t}{2\|\boldsymbol{A}\|_{1,2}^2}\left(\sqrt{1 + \frac{4\|\boldsymbol{A}\|_{1,2}^2}{t}} - 1\right), \quad \tau = \frac{1-\theta}{t\theta}, \quad \text{and} \quad \sigma = \frac{1-\theta}{\theta},$$

The third update in (3.25), with $H = \frac{1}{2}\|\cdot\|_2^2$, can be expressed analytically as follows:

$$\boldsymbol{w}_{k+1} = \left(\boldsymbol{w}_k + \sigma\left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}]\right]\right)/(1+t\sigma). \tag{3.26}$$

We compute the time required for convergence in the primal variable p_{k+1} and also the time required for convergence with the ergodic average

$$P_K = \frac{1}{T_K}\sum_{k=1}^{K}\frac{1}{\theta^{k-1}}p_k \quad \text{with} \quad T_K = \sum_{k=1}^{K}\frac{1}{\theta^{k-1}} = \frac{1-\theta^K}{(1-\theta)\theta^{K-1}}$$

as defined in Proposition 2.4.3 (with $\boldsymbol{X}_k$ substituted for P_k). The iterations were stopped once $\|p_{k+1}-p_k\|_1 \leqslant 10^{-4}$ and $\|P_{k+1}-P_k\|_1 \leqslant 10^{-4}$.

Accelerated linear PDHG method. We compute the iterates

$$\begin{aligned}
p_{k+1} &= \operatorname*{arg\,min}_{p\in\Delta_n}\left\{D_{\mathcal{H}_n}(p,p_{\text{prior}}) + \left\langle \boldsymbol{w}_k + \theta(\boldsymbol{w}_k-\boldsymbol{w}_{k-1}), \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}]\right\rangle + \frac{1}{2\tau}\|p-p_k\|_2^2\right\},\\
\boldsymbol{w}_{k+1} &= \operatorname*{arg\,max}_{\boldsymbol{w}\in\mathbb{R}^m}\left\{-\frac{t}{2}\|\boldsymbol{w}\|_2^2 + \left\langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}]\right\rangle - \frac{1}{2\sigma}\|\boldsymbol{w}-\boldsymbol{w}_k\|_2^2\right\}.
\end{aligned}$$

To compute these iterates, we apply Moreau's identity (3.26) in the first update and use formula (3.26) to express these updates as

$$\begin{aligned}
\boldsymbol{x}_k &= p_k + \tau\boldsymbol{A}^*(\boldsymbol{w}_k + \theta[\boldsymbol{w}_k - \boldsymbol{w}_{k-1}])\\
p_{k+1} &= \boldsymbol{x}_k - \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2 + \tau\log\left(\sum_{j=1}^{n} p_{\text{prior}}(j)e^{[\boldsymbol{x}]_j/\tau}\right)\right\}\\
\boldsymbol{w}_{k+1} &= \left(\boldsymbol{w}_k + \sigma\left[\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^{(k+1)}}[\boldsymbol{\Phi}]\right]\right)/(1+t\sigma).
\end{aligned}$$

We use the forward-backward splitting method [48, Algorithm 5] to compute the second line. Here we use the same initial values as for the accelerated nonlinear PDHG method, and for the parameters we set

$$\theta = 1 - \frac{1}{2\|\boldsymbol{A}\|_{2,2}^2}\left(\sqrt{1+4\|\boldsymbol{A}\|_{2,2}^2}-1\right), \quad \tau = \frac{1-\theta}{\theta}, \quad \text{and} \quad \sigma = \frac{1-\theta}{\theta}.$$

We compute the time required for convergence in the primal variable p_{k+1} and also the time required for convergence with the ergodic average

$$P_K = \frac{1}{T_K} \sum_{k=1}^{K} \frac{1}{\theta^{k-1}} p_k \quad \text{with} \quad T_K = \sum_{k=1}^{K} \frac{1}{\theta^{k-1}} = \frac{1-\theta^K}{(1-\theta)\theta^{K-1}}$$

as defined in Proposition 2.4.3 (with $\boldsymbol{X}_k$ substituted for P_k). The iterations were stopped once $\|p_{k+1} - p_k\|_1 \leqslant 10^{-4}$ and $\|P_{k+1} - P_k\|_1 \leqslant 10^{-4}$.

Numerical results Table 3.3.1 shows the time results for the accelerated linear and nonlinear PDHG methods. For both methods, we show the time results for convergence with the regular and ergodic sequences as described before. We observe that the nonlinear PDHG method is considerably faster than its linear counterpart; the nonlinear PDHG method achieves a speedup of about 3-4.

	Numbers of features m						
	100000	250000	500000	750000	1000000	1250000	1500000
Methods	Timings (s)						
Linear PDHG (Reg.)	28.23	72.95	156.8	262.65	332.73	486.16	620.72
Linear PDHG (Erg.)	32.91	84.81	179.18	295.90	374.72	539.58	700.94
Nonlinear PDHG (Reg.)	10.86	28.85	62.76	93.39	129.37	185.14	217.86
Nonlinear PDHG (Erg.)	11.64	30.67	61.50	96.89	124.48	170.06	210.15

Table 3.3.1: Time results (in seconds) for solving ℓ_2^2-regularized maximum entropy estimation with the linear and nonlinear PDHG methods.

3.4 Zero-sum matrix games with entropy regularization

3.4.1 Description of the problem

Two-player zero-sum matrix games are a class of saddle-point optimization problems that model one of the basic forms of constrained competitive games [42]. We focus here on zero-sum matrix games with entropy regularization, the latter which models the imperfect knowledge of the payoff matrix $\boldsymbol{A}$ by the two players [179]. Let Δ_m and Δ_n denote the unit simplices on $\mathbb{R}^m$ and $\mathbb{R}^n$, and let $\boldsymbol{A}$ denote an $m \times n$ matrix, called the payoff matrix. Zero-sum matrix games with entropy

regularization are formulated as follow:

$$\min_{\boldsymbol{x}\in\Delta_n}\max_{\boldsymbol{y}^*\in\Delta_m}\left\{\lambda\mathcal{H}_n(\boldsymbol{x})+\langle\boldsymbol{y}^*,\boldsymbol{A}\boldsymbol{x}\rangle-\lambda\mathcal{H}_m(\boldsymbol{y}^*)\right\}, \tag{3.27}$$

where $\lambda>0$ and $\mathcal{H}_n(\boldsymbol{x})=\sum_{j=1}^n[\boldsymbol{x}]_j\log([\boldsymbol{x}]_j)$ and $\mathcal{H}_m(\boldsymbol{y}^*)=\sum_{i=1}^m[\boldsymbol{y}^*]_i\log([\boldsymbol{y}^*]_i)$ denote the negative entropies of the probability distributions $\boldsymbol{x}$ and $\boldsymbol{y}^*$.

The primal and dual problems associated to the entropy regularized zero-sum matrix game (3.27) are given by

$$\min_{\boldsymbol{x}\in\Delta_n}\left\{\lambda\mathcal{H}_n(\boldsymbol{x})+\lambda\log\left(\sum_{i=1}^m e^{[\boldsymbol{A}\boldsymbol{x}]_i/\lambda}\right)\right\} \tag{3.28}$$

and

$$\max_{\boldsymbol{y}^*\in\Delta_n}\left\{-\lambda\log\left(\sum_{j=1}^n e^{-[\boldsymbol{A}^*\boldsymbol{y}^*]_j/\lambda}\right)-\lambda\mathcal{H}_m(\boldsymbol{y}^*)\right\} \tag{3.29}$$

Due to the strong convexity of the primal problem (3.28) and strong concavity of the dual problem (3.29), the saddle-point problem (3.27) has a unique saddle point $(\boldsymbol{x}_s,\boldsymbol{y}_s^*)\in\mathbb{R}^n\times\mathbb{R}^m$, which are also the unique solutions to the primal and dual problems above. They satisfy the optimality conditions

$$-[\boldsymbol{A}^*\boldsymbol{y}_s^*]_j=\lambda(1+\log([\boldsymbol{x}_s]_j))\quad\text{and}\quad[\boldsymbol{y}_s^*]_i=\frac{e^{[\boldsymbol{A}\boldsymbol{x}_s]_i/\lambda}}{\sum_{i=1}^m e^{[\boldsymbol{A}\boldsymbol{x}_s]_i/\lambda}}. \tag{3.30}$$

Accelerated nonlinear PDHG method

We propose to solve the zero-sum matrix game with entropy regularization (3.27) using the accelerated PDHG method (2.59) with the following choice of norms and Bregman functions:

$$\|\cdot\|_{\mathcal{X}}=\|\cdot\|_1,\quad\|\cdot\|_{\mathcal{Y}}=\|\cdot\|_\infty\implies\|\cdot\|_{\mathcal{Y}^*}=\|\cdot\|_1,\quad\phi_{\mathcal{X}}=\mathcal{H}_n,\quad\text{and}\quad\phi_{\mathcal{Y}^*}=\mathcal{H}_m.$$

The Bregman divergences induced by $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}^*}$ are the Kullback–Leibler divergences

$$D_{\mathcal{H}_n}(\boldsymbol{x},\bar{\boldsymbol{x}})=\sum_{j=1}^n[\boldsymbol{x}]_j\log\left([\boldsymbol{x}]_j/[\bar{\boldsymbol{x}}]_j\right)\quad\text{and}\quad D_{\mathcal{H}_m}(\boldsymbol{y}^*,\bar{\boldsymbol{y}}^*)=\sum_{i=1}^m[\boldsymbol{y}^*]_i\log\left([\boldsymbol{y}^*]_i/[\bar{\boldsymbol{y}}^*]_i\right)$$

where $\boldsymbol{x} \in \Delta_n$, $\bar{\boldsymbol{x}} \in \operatorname{int} \Delta_n$, $\boldsymbol{y} \in \Delta_m$ and $\bar{\boldsymbol{y}}^* \in \operatorname{int} \Delta_m$. With these choices, assumptions (A1)-(A7) hold with the strong convexity parameters $\gamma_g = \gamma_{h^*} = \lambda$. In particular, assumption (A5) holds because both $\mathcal{H}_n$ and $\mathcal{H}_m$ are 1-strongly convex with respect to the l_1 norm over their respective unit simplices, due to Pinsker's inequality [22, 63, 152, 158, 203]. Moreover, the induced operator norm is the entry of the payoff matrix $\boldsymbol{A}$ with largest magnitude:

$$\|\boldsymbol{A}\|_{\mathrm{op}} = \|\boldsymbol{A}\|_{1,\infty} = \sup_{\|\boldsymbol{x}\|_1 = 1} \|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{\substack{i \in \{1,\dots,m\} \\ j \in \{1,\dots,n\}}} |A_{ij}|.$$

The stepsize parameters θ, τ, and σ are accordingly

$$\theta = 1 - \frac{\lambda^2}{2\|\boldsymbol{A}\|_{1,\infty}^2} \left(\sqrt{1 + \frac{4\|\boldsymbol{A}\|_{1,\infty}^2}{\lambda^2}} - 1 \right) \quad \text{and} \quad \tau = \sigma = \frac{1-\theta}{\lambda\theta}.$$

Given $\boldsymbol{y}_0^* \in \mathbb{R}^m$ and $\boldsymbol{x}_{-1}^* = \boldsymbol{x}_0^* \in \mathbb{R}^n$, the corresponding accelerated nonlinear PDHG method for the matrix game (3.27) consists of the iterations

$$\begin{aligned}
\boldsymbol{y}_{k+1} &= \operatorname*{arg\,max}_{\boldsymbol{y}^* \in \Delta_m} \left\{ -\lambda \mathcal{H}_m(\boldsymbol{y}^*) + \langle \boldsymbol{y}^*, \boldsymbol{A}(\boldsymbol{x}_k - \theta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})) \rangle - \frac{1}{\sigma} D_{\mathcal{H}_m}(\boldsymbol{y}^*, \boldsymbol{y}_k^*) \right\}, \\
\boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x} \in \Delta_n} \left\{ \lambda \mathcal{H}_n(\boldsymbol{x}) + \langle \boldsymbol{y}_{k+1}^*, \boldsymbol{A}\boldsymbol{x} \rangle + \frac{1}{\tau} D_{\mathcal{H}_n}(\boldsymbol{x}, \boldsymbol{x}_k) \right\}.
\end{aligned}$$

The updates $\boldsymbol{x}_{k+1}$ and $\boldsymbol{y}_{k+1}^*$ can be both computed explicitly. A straightforward calculation gives the updates

$$\begin{aligned}
[\boldsymbol{y}_{k+1}^*]_i &= \frac{\left([\boldsymbol{y}_k^*]_i e^{-\tau [\boldsymbol{A}(\boldsymbol{x}_k - \theta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}))]_i} \right)^{1/(1+\lambda\sigma)}}{\sum_{i=1}^m \left([\boldsymbol{y}_k^*]_i e^{-\tau [\boldsymbol{A}(\boldsymbol{x}_k - \theta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}))]_i} \right)^{1/(1+\lambda\sigma)}} \\
[\boldsymbol{x}_{k+1}]_j &= \frac{\left([\boldsymbol{x}_k^*]_j e^{-\tau [\boldsymbol{A}^* \boldsymbol{y}_{k+1}^*]_j} \right)^{1/(1+\lambda\tau)}}{\sum_{j=1}^n \left([\boldsymbol{x}_k^*]_j e^{-\tau [\boldsymbol{A}^* \boldsymbol{y}_{k+1}^*]_j} \right)^{1/(1+\lambda\tau)}}
\end{aligned} \tag{3.31}$$

for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. All parameter calculations and updates can be performed in $O(mn)$ operations. According to Proposition 2.4.4 and the optimality conditions (3.30), we have the strong limits

$$\lim_{k \to +\infty} \boldsymbol{x}_k = \boldsymbol{x}_s, \quad \lim_{k \to +\infty} \boldsymbol{y}_k^* = \boldsymbol{y}_s^*,$$

$$\lim_{k\to+\infty} -[\boldsymbol{A}^*\boldsymbol{y}_s^*]_j = \lambda(1+\log([\boldsymbol{x}_s]_j)), \quad \text{and} \quad \lim_{k\to+\infty}[\boldsymbol{y}_k^*]_i = \frac{e^{([\boldsymbol{A}\boldsymbol{x}_s]_i/\lambda)}}{\sum_{i=1}^m e^{([\boldsymbol{A}\boldsymbol{x}_s]_i/\lambda)}}.$$

3.4.2 Numerical experiments

We present here some numerical experiments to compare the running times of the accelerated nonlinear PDHG methods proposed entropy-regularized zero-sum matrix games to other commonly-used first-order optimization methods. These methods include the accelerated linear PDHG method [46, 47] and state-of-the-art Predictive Update (PU) and Optimistic Multiplicative Weights Update (OMWU) methods from Cen et al. [42]. These methods are described below and were implemented in MATLAB. All numerical experiments were performed on a single core Intel(R) Core(TM) i7-10750H CPU @ 2.60 GHz.

Data generation and optimization methods

Following the methodology described in [42, Section 2.3], we generate each entry of the payoff matrix $\boldsymbol{A}$ from the uniform distribution on $[-1,1]$ and we set $\lambda = 0.1$. Here, we set $m = n$, with $n = 10000$, 15000, 20000, 25000, 30000, 35000 and 40000.

We perform simulations using the accelerated nonlinear PDHG method (3.31), the accelerated linear PDHG method, and the Predictive Update (PU) and Optimistic Multiplicative Weights Update (OMWU) methods from Cen et al. [42]. The initial values, parameters and numerical criteria for convergence of each method are described below.

Accelerated nonlinear PDHG method (3.31). We generate the entries of the initial vectors $\boldsymbol{y}_0^*$ and $[\boldsymbol{x}_{-1}]_j = [\boldsymbol{x}_0]_j$ for $j \in \{1,\dots,n\}$ uniformly at random in $(0, 1/m)$ and $(0, 1/n)$, respectively,

and normalized their entries so that $\sum_{i=1}^{m}[\boldsymbol{y}_0^*]_i = 1$ and $\sum_{j=1}^{n}[\boldsymbol{x}_0]_j = 1$. For the parameters, we set

$$\theta = 1 - \frac{\lambda^2}{2\left\|\boldsymbol{A}\right\|_{1,\infty}^2}\left(\sqrt{1 + \frac{4\left\|\boldsymbol{A}\right\|_{1,\infty}^2}{\lambda^2}} - 1\right) \quad \text{and} \quad \tau = \sigma = \frac{1-\theta}{\lambda\theta}.$$

We compute the time required for convergence in the dual variable $\boldsymbol{y}_k^*$ and also the time required for convergence in the average dual variable $\boldsymbol{Y}^*$ as defined in Proposition 2.4.4. The iterations were stopped once $\left\|\boldsymbol{y}_{k+1}^* - \boldsymbol{y}_k^*\right\|_2 \leqslant 10^{-4}\left\|\boldsymbol{y}_{k+1}^*\right\|_2$ and $\left\|\boldsymbol{Y}_{K+1}^* - \boldsymbol{Y}_K^*\right\|_2 \leqslant 10^{-4}\left\|\boldsymbol{Y}_{K+1}^*\right\|_2$.

Accelerated linear PDHG method. We compute the iterates

$$\begin{aligned}
\boldsymbol{y}_{k+1}^* &= \operatorname*{arg\,max}_{\boldsymbol{y}^*\in\Delta_m}\left\{-\lambda\mathcal{H}_m(\boldsymbol{y}^*) + \langle\boldsymbol{y}^*, \boldsymbol{A}(\boldsymbol{x}_k + \theta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}))\rangle - \frac{1}{2\sigma}\left\|\boldsymbol{y}^* - \boldsymbol{y}_k^*\right\|_2^2\right\},\\
\boldsymbol{x}_{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\Delta_n}\left\{\lambda\mathcal{H}_n(\boldsymbol{x}) + \left\langle\boldsymbol{y}_{k+1}^*, \boldsymbol{A}\boldsymbol{x}\right\rangle + \frac{1}{2\tau}\left\|\boldsymbol{x} - \boldsymbol{x}_k\right\|_2^2\right\}.
\end{aligned}$$

To compute these iterates, we use Moreau's identity [182] to express them as follows:

$$\begin{aligned}
\boldsymbol{v}_k &= \boldsymbol{y}_k^* + \sigma\boldsymbol{A}(\boldsymbol{x}_k + \theta[\boldsymbol{x}_k - \boldsymbol{x}_{k-1}]),\\
\boldsymbol{y}_{k+1}^* &= \boldsymbol{v}_k - \operatorname*{arg\,min}_{\boldsymbol{z}\in\mathbb{R}^m}\left\{\frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{v}_k\right\|_2^2 + \lambda\sigma\log\left(\sum_{i=1}^{m} e^{[\boldsymbol{z}]_i/\lambda\sigma}\right)\right\},\\
\boldsymbol{w}_k &= \boldsymbol{x}_k - \tau\boldsymbol{A}^*\boldsymbol{y}_{k+1}^*,\\
\boldsymbol{x}_{k+1} &= \boldsymbol{w}_k - \operatorname*{arg\,min}_{\boldsymbol{z}\in\mathbb{R}^n}\left\{\frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{w}_k\right\|_2^2 + \lambda\tau\log\left(\sum_{i=1}^{m} e^{[\boldsymbol{z}]_i/\lambda\tau}\right)\right\}.
\end{aligned}$$

We use the forward-backward splitting method [48, Algorithm 5] to compute the second and fourth line. Here we use the same initial values as for the accelerated nonlinear PDHG method (3.31), and for the parameters we set

$$\theta = 1 - \frac{\lambda^2}{2\left\|\boldsymbol{A}\right\|_{2,2}^2}\left(\sqrt{1 + \frac{4\left\|\boldsymbol{A}\right\|_{2,2}^2}{\lambda^2}} - 1\right) \quad \text{and} \quad \tau = \sigma = \frac{1-\theta}{\lambda\theta}.$$

We compute the time required for convergence in the dual variable $\boldsymbol{y}_k^*$ and also the time required for convergence in the average dual variable $\boldsymbol{Y}^*$ as defined in Proposition 2.4.4. The iterations were

stopped once $\left\|\boldsymbol{y}_{k+1}^{*}-\boldsymbol{y}_{k}^{*}\right\|_{2} \leqslant 10^{-4}\left\|\boldsymbol{y}_{k+1}^{*}\right\|_{2}$ and $\left\|\boldsymbol{Y}_{K+1}^{*}-\boldsymbol{Y}_{K}^{*}\right\|_{2} \leqslant 10^{-4}\left\|\boldsymbol{Y}_{K+1}^{*}\right\|_{2}$.

Predictive Update and Optimistic Multiplicative Weights Update methods. For the PU and OMWU, we use Algorithms 1 and 2 as described in [42] with the learning rates

$$\eta_{\mathrm{PU}}=\frac{1}{2+\|\boldsymbol{A}\|_{1,\infty}} \quad \text{and} \quad \eta_{\mathrm{OMWU}}=\min\left\{\frac{1}{2+2\|\boldsymbol{A}\|_{1,\infty}}, \frac{1}{4\|\boldsymbol{A}\|_{1,\infty}}\right\}.$$

Numerical results. Table 3.4.1 shows the time results for the PU, OMWU, and linear and nonlinear PDHG methods. For the linear and nonlinear PDHG methods, we also show the time results for convergence with the regular and ergodic sequences as described before. We observe that the nonlinear PDHG method is considerable faster than both the linear PDHG method and the state-of-the-art methods PU and OMWU for solving the entropy regularized zero-sum matrix game (3.27); the nonlinear PDHG method achieves a speedup of 5 to 11 compared to linear PDHG method and a speedup of 3 to 5 compared to the state-of-the-art methods PU and OMWU.

	Numbers $m=n$						
	10000	15000	20000	25000	30000	35000	40000
Methods	Timings (s)						
PU	34.16	55.01	88.48	137.14	197.66	289.49	353.61
OMWU	44.89	81.67	142.53	221.23	318.65	493.73	568.94
Linear PDHG (Reg.)	42.47	100.65	218.26	366.52	601.11	938.18	1094.25
Linear PDHG (Erg.)	44.43	104.61	225.52	379.71	608.78	949.36	1114.24
Nonlinear PDHG (Reg.)	8.56	15.88	24.68	38.40	55.13	82.52	103.50
Nonlinear PDHG (Erg.)	12.02	19.05	31.45	48.70	70.27	105.26	129.28

Table 3.4.1: Time results (in seconds) for solving the entropy regularized zero-sum matrix game (3.27) with the PU, OMWU, and linear and nonlinear PDHG methods.

3.5 Discussion

This chapter presented practical implementations of accelerated nonlinear PDHG methods for sparse logistic regression, regularized maximum entropy estimation problems and entropy-regularized zero-sum matrix games. The accelerated nonlinear PDHG methods are particularly

useful to solve such problems because they involve a logistic regression model or are defined on the unit simplex or both. For these problems, one may choose to use a Bregman divergence defined in terms of the average negative sum of binary entropy terms or the relative entropy to arrive at a straightforward and efficient optimization method. Numerical experiments showed that the nonlinear PDHG methods are considerably faster than competing methods.

The new nonlinear PDHG methods are advantageous because they can achieve an optimal convergence rate with stepsize parameters that are simple and efficient to compute. They can be typically computed on the order of $O(mn)$ operations where m and n denote the dimensions to the dual and primal problems at hand. In contrast, most first-order optimization methods, including the linear PDHG method, require on the order of $O(\min(m^2n, mn^2))$ operations to compute all the parameters required to achieve an optimal convergence rate. This gain in efficiency can be considerable: in our numerical experiments for ℓ_1-constrained logistic regression, ℓ_2^2-constrained regularized maximum entropy estimation, and zero-sum matrix games with entropy regularization, we were able to get a speedup of 3 to 10 compared to other competing optimization methods.

We expect the accelerated nonlinear PDHG methods described in this work to provide efficient methods for solving large-scale supervised machine learning. In particular, these applications to strongly convex and smooth problems defined on the unit simplex, such as ν-support vector machines with squared loss, boosting and structured prediction problems in machine learning, will be pursued in future work. It would be interesting to extend the accelerated nonlinear PDHG methods described here to the stochastic case for problems that are separable in the dual variable, and to the non-convex case to deal with large-scale non-convex problems, such as those arising in deep learning. These extensions will be pursued in future work as well.

Chapter Four

Bayesian methods for imaging science and connections to Hamilton–Jacobi PDEs

4.1 Introduction

Overview

Image denoising problems aim to remove noise from noisy images while accounting for underlying uncertainties. Among several proposed approaches for denoising images, two fundamental ones are variational and Bayesian methods. Variational methods formulate image denoising problems as optimization problems. These problems typically minimize the weighted sum of a data fidelity term and a regularization term, where the former embeds properties of the noise corrupting the noisy image and the latter embeds properties of the image to denoise. The solution to such a problem then gives an estimate that hopefully accounts well for the data fidelity term and the regularization term [48, 64]. Bayesian methods formulate image denoising problems in a probabilistic framework that combines observed data through a likelihood function which models the noise corrupting the unknown image and prior knowledge through a prior distribution which models known properties of the unknown image to generate a posterior distribution. An appropriate decision rule then selects a meaningful estimate of the actual image from the posterior distribution that hopefully accounts well for both the prior knowledge and observed data [78, 229, 231, 237, 244]. This decision rule is usually chosen to minimize the posterior expected value of some loss function and is called a Bayes estimator. A standard example is the posterior mean estimator, the mean of the posterior distribution [150, pages 344-345], which minimizes the mean squared error and, more generally, Bregman loss functions [12].

In a Bayesian setting, variational and Bayesian methods for denoising images use maximum a posterior (MAP) estimators and posterior mean (PM) estimators. Variational methods are well-understood theoretically; for instance, it is known that a broad class of MAP estimators correspond to solutions of first-order Hamilton–Jacobi partial differential equations (HJ PDEs) [64, 68]. The image denoising properties of these MAP estimators, in particular, follow readily from the properties of the solutions to these HJ PDEs. Bayesian methods, in contrast, are less well-understood theoretically. This chapter aims to partially fill this gap. Specifically, we present novel theoretical

connections between viscous HJ PDEs and a broad class of Bayesian PM estimators and use these connections to clarify certain image denoising properties of this class of Bayesian PM estimators.

The work presented in this chapter focuses on the class of finite-dimensional image denoising problems

$$\boldsymbol{x} = \boldsymbol{u}_{\text{true}} + \boldsymbol{\eta}, \tag{4.1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the observed image, $\boldsymbol{u}_{\text{true}} \in \mathbb{R}^n$ is the true image and η is independent identically distributed Gaussian noise. These problems are well-known to be ill-posed in general, and variational and Bayesian approaches are popular methods to find meaningful solutions to these ill-posed problems [9, 78, 249]. Concretely, these methods compute the MAP estimate

$$\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) \coloneqq \underset{\boldsymbol{u}\in\mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} \tag{4.2}$$

and the MAP estimate

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \coloneqq \frac{\int_{\mathbb{R}^n} \boldsymbol{u} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}}. \tag{4.3}$$

The functions

$$\boldsymbol{u} \mapsto \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \quad \text{and} \quad J \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$$

in (4.2) are, respectively, the (quadratic) data fidelity and regularization terms associated to the variational method. The functions

$$\boldsymbol{u} \mapsto e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \quad \text{and} \quad \boldsymbol{u} \mapsto e^{-J(\boldsymbol{u})/\epsilon}$$

in (4.3) are, respectively, the (Gaussian) likelihood function and generalized prior distribution associated to the Bayesian method. The parameter $t > 0$ controls the relative importance of the data fidelity term over the regularization term, and the parameter ϵ controls the shape of the posterior distribution in (4.3), where small values of ϵ favor configurations close to the mode, which is the MAP estimate, of the posterior distribution.

To motivate the work presented here, let us illustrate the MAP and PM estimates and their denoising capabilities with an example. We consider an anisotropic version of the Rudin–Osher–Fatemi (ROF) image denoising model, which consists of an anisotropic total variation (TV) regularization term with quadratic data fidelity term [30, 45, 216]. We define anisotropic TV as follows

$$\mathrm{TV}(\boldsymbol{u}) = \sum_{i,j\in\{1,\dots n\}} w_{ij}|[\boldsymbol{u}]_i - [\boldsymbol{u}]_j|,$$

where $w_{ij} \geqslant 0$ and the value of an image $\boldsymbol{u}$ at the pixel i is denoted by $y_i \in \mathbb{R}$. For this example, we assume that a digital image is defined on a two-dimensional regular grid and only consider the 4-nearest neighbors interactions for defining TV (i.e., $w_{ij} = w_{ji} = \frac{1}{2}$ if i and j are neighbors, and $w_{ij} = w_{ji} = 0$ otherwise, see [70] for instance). Let $\boldsymbol{x}$ denote an observed noisy image and t and ϵ be parameters as previously defined. Then the associated anisotropic ROF problem [216] is

$$\min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \mathrm{TV}(\boldsymbol{u}) \right\}. \tag{4.4}$$

The MAP and PM estimates to the ROF problem (4.4) are given, respectively, by Equations (4.2) and (4.3) with $J(\boldsymbol{u}) = \mathrm{TV}(\boldsymbol{u})$, i.e.,

$$\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) = \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \mathrm{TV}(\boldsymbol{u}) \right\} \tag{4.5}$$

and

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \frac{\int_{\mathbb{R}^n} \boldsymbol{u} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + \mathrm{TV}(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + \mathrm{TV}(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}}. \tag{4.6}$$

We note here that the PM estimate (4.6) with total variation prior and its denoising properties were investigated in [168, 169].

Figure 4.1.1(a) depicts the image Barbara, which we corrupt with Gaussian noise (zero mean with standard deviation $\sigma = 10$) in Figure 4.1.1(b). We let $\boldsymbol{x}$ denote this corrupted image, and we choose the parameters $t = 16$ and $\epsilon = 6.25$ in the MAP and PM estimates. The MAP estimate can be computed up to the machine precision using maximum-flow based algorithms [44, 70, 139], and the PM estimate can be approximated using Markov Chain Monte Carlo methods. Here, we

approximated the PM estimate (4.6) using the variable-at-a-time Metropolis–Hastings algorithm with random scan detailed in [168, Page 42, Algorithm 2]. Specifically, for the parameters of algorithm 2 in [168], we used, in the terminology of their algorithm, the parameters $\sigma = 10$ and $\lambda = 32$ (corresponding here to the choice of $t = 16$ and $\epsilon = 6.25$ in (4.6)), we chose the initial point of the algorithm to be the MAP estimate $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$, and finally, we set the internal parameters of algorithm 2 in [168] as follows: $\alpha = 17.32$ (this values yields an acceptance rate in the algorithm close to the optimal value 0.234 suggested in [213]), 20,000 for the maximum number of iterations, and n for the subsampling rate.

The MAP and PM estimates associated to the ROF model with these parameters produce the denoised images illustrated in Figures 4.1.1(c) and (d). Figures 4.1.2(a)-(d) zoom-in on the face of Barbara in Figures 4.1.1 The denoised image of Barbara with the MAP estimate exhibits staircasing effects [51, 81, 89] that can be observed in Figure 4.1.2(c), whereas the denoised image of Barbara with the posterior mean estimate does not. In either case, the denoised images result in a lost of texture, as can be seen by comparing Figure 4.1.2(a) with (c) and (d).

Variational methods are popular in practice because they are well-understood and often lead to optimization problems that can be solved efficiently using robust numerical optimization methods [48]. Such problems include, for example, total variation minimization or ℓ_1-norm based minimization [30, 38, 45, 64, 68, 82, 216]. In particular, MAP estimates from variational methods are also significantly faster to compute than PM estimates because the latter require complex stochastic methods to compute. However, reconstructed images from variational methods with non-smooth and convex regularization terms often have undesirable and visually unpleasant staircasing effects due to the singularities of the non-smooth regularization terms [51, 81, 89, 190, 168, 250]. This is illustrated for example in Figure 1(c), which contains regions where the pixel values are equal and lead to staircasing effects. In contrast, posterior mean estimates with quadratic fidelity term and total variation regularization terms have been shown to avoid staircasing effects [168, 169]. This is illustrated for example in Figure 4.1.1 and 4.1.2(d), where the denoised image with posterior mean estimate does not contain visibly substantial regions where the pixel values are equal.

(a) (b)

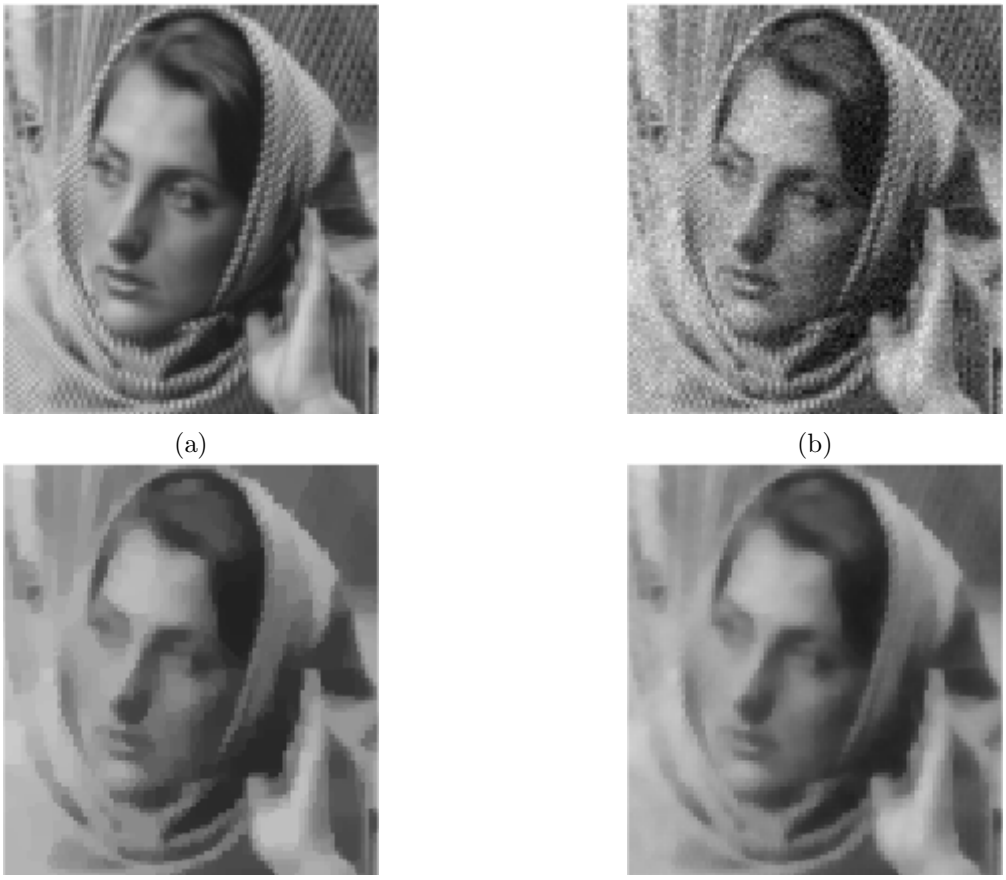
(c) (d)

Figure 4.1.1: The anisotropic ROF model endowed with 4-nearest neighbors is applied to the test image "Barbara". The original image is shown in (a). The image is corrupted by Gaussian noise (zero mean with standard deviation $\sigma = 10$) and is shown in (b). The corresponding minimizer $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$ given by (4.4) and posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ given by (4.6) with parameters $t = 16$ and $\epsilon = 6.25$ are illustrated in (c) and (d), respectively.

Related work

Several papers have proposed novel connections between MAP and Bayesian estimators, including PM estimators. First, [168, 169] showed that the class of Bayesian posterior mean estimates (4.3) with TV regularization term J can be expressed as minimizers to optimization problems involving a quadratic fidelity term and a smooth convex regularization term, i.e., there exists a smooth regularization term $f_{\text{reg}} \colon \mathbb{R}^n \to \mathbb{R}$ such that

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + f_{\text{reg}}(\boldsymbol{u}) \right\}. \tag{4.7}$$



(a) (b)

(c) (d)

Figure 4.1.2: The anisotropic ROF model endowed with 4-nearest neighbors is applied to the test image "Barbara". Images (a)-(d) are zoomed-in versions of the images illustrated in Figure 4.1.1.

This result was later extended to general priors [124], general Gaussian data fidelity terms [125], and to some non-quadratic data fidelity terms [126, 127]. To our knowledge, there is no representation formula for this smooth regularization term available in the literature.

Second, [35] showed that the MAP estimate (4.2) corresponds to a Bayes estimator when the regularization term J is convex and uniformly Lipschitz continuous on $\mathbb{R}^n$, that is, the MAP estimate (4.2) minimizes the posterior expected value of an appropriate loss function. This was later extended by [36] to some log-concave posterior distributions with non-quadratic fidelity term, and later studied from the point of view of differential geometry in [196] and also derived for posterior distributions that are strongly log-concave and at least three times differentiable.

In addition to these results, it is known that under certain assumptions on the regularization

term J, the value of the minimization problem

$$S_0(\boldsymbol{x}, t) \coloneqq \min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} \tag{4.8}$$

whose minimizer is the MAP estimate (4.2), satisfies the first-order HJ PDE

$$\begin{cases} \dfrac{\partial S_0}{\partial t}(\boldsymbol{x}, t) + \dfrac{1}{2} \|\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t)\|_2^2 = 0 & \text{in } \mathbb{R}^n \times (0, +\infty), \\ S_0(\boldsymbol{x}, 0) = J(\boldsymbol{x}) & \text{in } \mathbb{R}^n. \end{cases} \tag{4.9}$$

The properties of the minimizer $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$ follow from the properties of the solution to this HJ equation [64, 68]. In particular, the MAP estimate satisfies the representation formula $\boldsymbol{u}_{PM}(\boldsymbol{x}, t) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t)$.

We note that the results of [64, 68] considers only connections between a class of first-order HJ PDEs and MAP estimators. To our knowledge, connections between posterior estimators and HJ PDEs are not available in the literature.

Contributions

This chapter proposes novel theoretical connections between solutions to HJ PDEs and a broad class of Bayesian methods and posterior mean estimators. These connections are described in Proposition 4.2.1 and 4.2.2 for viscous HJ PDEs and first-order HJ PDEs, respectively. We show in Proposition 4.2.1 that the posterior mean estimate (4.3) is described by the solution to a viscous HJ with initial data corresponding to the convex regularization term J, which we characterize in detail in terms of the data $\boldsymbol{x}$ and parameters t and ϵ. In particular, the posterior mean estimate (4.3) satisfies the representation formula $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$. Next, we use the connections between viscous HJ PDEs and posterior mean estimates established in Proposition 4.2.1 to show in Proposition 4.2.2 that the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ can be expressed through the gradient of the solution to a first-order HJ PDE with twice continuously differentiable convex initial

data $\mathbb{R}^n \ni \boldsymbol{x} \mapsto K_\epsilon^*(\boldsymbol{x}, t) - \frac{1}{2}\|\boldsymbol{x}\|_2^2$, where

$$K_\epsilon(\boldsymbol{x}, t) = t\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\text{dom } J} e^{\left(\frac{1}{t}\langle \boldsymbol{x}, \boldsymbol{u}\rangle - \frac{1}{2t}\|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}\right)$$

and $\boldsymbol{x} \mapsto K_\epsilon^*(\boldsymbol{x}, t)$ is the convex conjugate of the function $\boldsymbol{x} \mapsto K_\epsilon(\boldsymbol{x}, t)$. In other words, we show

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \underset{\boldsymbol{u} \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \left(K_\epsilon^*(\boldsymbol{u}, t) - \frac{1}{2}\|\boldsymbol{u}\|_2^2\right)\right\}.$$

This formula gives the representation of the convex regularization term, enabling one to express the posterior mean estimate as the minimizer of a convex variational problem, and in fact in terms of the solution to a first-order HJ PDE. This thereby extends the results of [168, 124], who showed existence of this regularization term when the data fidelity term is quadratic, but not its representation. The second-order continuous differentiability of this regularization term, in particular, implies that the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ avoids image denoising staircasing effects as a consequence of the results derived in [189, Theorem 3].

We also present several topological properties of posterior mean estimators in Proposition 4.3.1 and we use these in conjunction with the connections between HJ PDEs and posterior mean estimators to derive representation and monotonicity properties of posterior mean estimators in Propositions 4.3.2 and 4.3.3, respectively. These properties are then used to derive an optimal upper bound on the mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right]$, an estimate of the squared difference between the MAP and posterior mean estimates, monotone and non-expansive properties of the posterior mean estimate, and the behavior of the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ in the limit $t \to 0$ (Proposition 4.3.4). In addition, we use the connections between both MAP and posterior mean estimates and HJ PDEs to characterize the MAP estimate (4.2) in the context of Bayesian estimation theory, and specifically in Proposition 4.3.5 to show that the MAP estimate (4.2) corresponds to the Bayes estimator of the Bayesian risk (4.37) whenever J is convex on $\mathbb{R}^n$ and bounded from below. When J is defined only on a strict subset of $\mathbb{R}^n$, we further show that the Bayesian risk (4.37) has a corresponding Bayes estimator that is described in terms of the solution to both the first-order HJ PDE (1.2.14) and the viscous HJ PDE (4.2.1). Finally, we present in 4.4 some extensions of these results to a class of posterior mean estimators whose priors

are sums of log-concave priors, that is, to posterior mean estimators of mixture distributions.

Organization

Section 4.2 establishes theoretical connections between a broad class of Bayesian posterior mean estimators and HJ PDEs. The mathematical set-up is described in Subsection 4.2.1, the connections of posterior mean estimators to viscous HJ PDEs are described in Subsection 4.2.2 and the connections of posterior mean estimators to first-order HJ PDEs are described in Subsection 4.2.3. Section 4.3 uses establishes various properties of posterior mean estimators using the aforementioned connections to HJ PDEs. Specifically, we present topological, representation, and monotonicity properties of posterior mean estimators in Subsection 4.3.1, an optimal upper bound on the mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right]$, an estimate of the squared difference between the MAP and posterior mean estimates, monotone and non-expansive properties of the posterior mean estimate, and the behavior of the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ in the limit $t \to 0$ in Subsection 4.3.2. Finally, we establish properties of MAP and posterior mean estimators in terms of Bayesian risks involving Bregman divergences in Subsection 4.3.3.

4.2 Connections between Bayesian posterior mean estimators and Hamilton–Jacobi partial differential equations

4.2.1 Set-up

To establish connections between Bayesian posterior mean estimators and Hamilton–Jacobi equations, we will assume that the regularization term J in the variational imaging model (4.8) satisfies the following assumptions:

(A1) $J \in \Gamma_0(\mathbb{R}^n)$,

(A2) int (dom J) $\neq \varnothing$,

(A3) $\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) \in \mathbb{R}$, and without loss of generality, $\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) = 0$.

Assumption (A1) ensures that the minimal value of the convex imaging problem (4.8) and its minimizer (4.2) are well-defined and enjoy several properties (see Section 1.2, Proposition 1.2.14). Assumption (A2) ensures that for every $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, the posterior distribution

$$\mathbb{R}^n \ni \boldsymbol{u} \mapsto \frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u}} \tag{4.10}$$

and its associated partition function

$$\mathbb{R}^n \times (0,+\infty) \times (0,+\infty) \ni (\boldsymbol{x}, t, \epsilon) \mapsto Z_J(\boldsymbol{x}, t, \epsilon) = \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \tag{4.11}$$

are well-defined, and finally, assumption (A3) guarantees that the partition function (4.11) is also bounded from above independently of $\boldsymbol{x} \in \mathbb{R}^n$. We will denote the posterior expectation (with respect to the posterior distribution (4.10)) of a measurable function $f\colon \Omega \mapsto \mathbb{R}$ with $\Omega \subset \text{dom } f$ integrable on the set dom $f \cap$ dom J by

$$\mathbb{E}_J\left[f(\boldsymbol{u})\right] = \frac{1}{Z_J(\boldsymbol{x}, t, \epsilon)} \int_{\Omega \cap \text{dom } J} f(\boldsymbol{u}) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u}. \tag{4.12}$$

Posterior expectations of vector quantities are defined similarly component-wise. Posterior expectations generally depends on $(\boldsymbol{x}, t, \epsilon)$ but we will omit writing this dependence explicitly.

4.2.2 Connections to viscous Hamilton–Jacobi partial differential equations

The next proposition establishes connections between viscous HJ PDEs with initial data J satisfying assumptions (A1)-(A3) and both the partition function (4.11) and the Bayesian posterior mean estimate (4.3). These connections mirror those between the first-order HJ PDE (1.22) with initial data J satisfying assumption (A1) and both the convex minimization problem (4.8) and the MAP estimate (4.2). The connections between viscous HJ PDEs and Bayesian posterior mean estimators

will be leveraged later to describe several properties of posterior mean estimators in terms of the observed image $\boldsymbol{x}$ and parameters t and ϵ, and in particular in Section 4.2.3 to show that the posterior mean estimate (4.3) can be expressed as the minimizer associated to the solution to a first-order HJ PDE (Proposition 4.2.2) with at least twice continuously differentiable and convex regularization term.

Proposition 4.2.1 (The viscous Hamilton–Jacobi equation with initial data in $\Gamma_0(\mathbb{R}^n)$)**.** *Suppose the function J satisfies assumptions (A1)-(A3). Then the following statements hold.*

(i) (Cole–Hopf transformation, [97, Section 4.4.1]) For every $\epsilon > 0$, the function $S_\epsilon \colon \mathbb{R}^n \times [0,+\infty) \to [0,+\infty)$ defined by

$$S_\epsilon(\boldsymbol{x},t) \coloneqq -\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}} Z_J(\boldsymbol{x},t,\epsilon)\right) = -\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u}\right) \tag{4.13}$$

is the unique smooth solution to the viscous HJ PDE with initial data

$$\begin{cases} \dfrac{\partial S_\epsilon}{\partial t}(\boldsymbol{x},t) + \dfrac{1}{2}\left\|\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)\right\|_2^2 = \dfrac{\epsilon}{2}\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x},t) & \text{in } \mathbb{R}^n \times (0,+\infty), \\ S_\epsilon(\boldsymbol{x},0) = J(\boldsymbol{x}) & \text{in } \mathbb{R}^n. \end{cases} \tag{4.14}$$

In addition, the domain of integration in (4.2.1) can be taken to be dom J *or, up to a set of Lebesgue measure zero,* int (dom J) *or* dom (∂J)*. Furthermore, for every $\boldsymbol{x} \in$ dom J and $\epsilon > 0$, except possibly at the boundary points $\boldsymbol{x} \in$ (dom J) $\setminus$ int (dom J) if such points exist, the pointwise limit $S_\epsilon(\boldsymbol{x},t)$ as $t \to 0$ exists and satisfies*

$$\lim_{\substack{t\to 0 \\ t>0}} S_\epsilon(\boldsymbol{x},t) = J(\boldsymbol{x}).$$

(ii) (Convexity and monotonicity properties).

(a) The function $\mathbb{R}^n \times (0,+\infty) \ni (\boldsymbol{x},t) \mapsto S_\epsilon(\boldsymbol{x},t) - \frac{n\epsilon}{2}\ln t$ is jointly convex.

(b) The function $(0,+\infty) \ni t \mapsto S_\epsilon(\boldsymbol{x},t) - \frac{n\epsilon}{2}\ln t$ is strictly monotone decreasing.

(c) The function $(0,+\infty) \ni \epsilon \mapsto S_\epsilon(\boldsymbol{x},t) - \frac{n\epsilon}{2}\ln \epsilon$ is strictly monotone decreasing.

(d) The function $\mathbb{R}^n \ni \boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_2^2 - tS_\epsilon(\boldsymbol{x},t)$ is strictly convex.

(iii) (Connections to the posterior mean and mean squared error) The posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ *and the mean squared error* $\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right]$ *satisfy the formulas*

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) \tag{4.15}$$

and

$$\begin{aligned} \mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right] &= t\epsilon\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \\ &= nt\epsilon - t^2\epsilon\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x}, t). \end{aligned} \tag{4.16}$$

Moreover, $\boldsymbol{x} \mapsto \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ *is a bijective function.*

(iv) (Vanishing $\epsilon \to 0$ *limit) Let* $S_0 : \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}$ *denote the continuously differentiable and convex solution to the first-order HJ PDE (1.22) with initial data* J*. For every* $\boldsymbol{x} \in \mathbb{R}^n$ *and* $t > 0$*, the following limit holds:*

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} -\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}\right) = \inf_{\boldsymbol{u} \in \mathbb{R}^n}\left\{\frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right\}, \tag{4.17}$$

that is,

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} S_\epsilon(\boldsymbol{x}, t) = S_0(\boldsymbol{x}, t),$$

and the limit converges uniformly over every compact set of $\mathbb{R}^n \times (0, +\infty)$ *in* $(\boldsymbol{x}, t)$*. In addition, the gradient* $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$*, the partial derivative* $\frac{\partial S_\epsilon(\boldsymbol{x}, t)}{\partial t}$*, and the Laplacian* $\frac{\epsilon}{2}\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x}, t)$ *satisfy the limits*

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t), \quad \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{\partial S_\epsilon}{\partial t}(\boldsymbol{x}, t) = \frac{\partial S_0}{\partial t}(\boldsymbol{x}, t),$$

and

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{\epsilon}{2}\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x}, t) = 0,$$

where each limit converges uniformly over every compact set of $\mathbb{R}^n \times (0, +\infty)$ *in* $(\boldsymbol{x}, t)$*. As a consequence, for every* $\boldsymbol{x} \in \mathbb{R}^n$ *and* $t > 0$*, the pointwise limit of* $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ *as* $\epsilon \to 0$ *exists and satisfy*

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{u}_{MAP}(\boldsymbol{x}, t),$$

and the limit converges uniformly over every compact set of $\mathbb{R}^n \times (0, +\infty)$ *in* $(\boldsymbol{x}, t)$.

Proof. See Appendix 4.A for the proof. □

To illustrate certain aspects of Proposition 4.2.1 and properties of posterior mean estimates, we give here two analytical examples.

Example 1 (Tikhonov–Phillips regularization)**.** *Let* $J(\boldsymbol{x}) = \frac{m}{2} \|\boldsymbol{x}\|_2^2$ *with* $m > 0$*, and consider the solution* $S_0(\boldsymbol{x}, t)$ *and* $S_\epsilon(\boldsymbol{x}, t)$ *to the first-order PDE* (1.22) *and viscous HJ PDE* (4.14) *with initial data* J*, respectively.*

The solution $S_0(\boldsymbol{x}, t)$ *is given by the Lax–Oleinik formula (Proposition 1.2.14, Equation* (1.24)*)*

$$\begin{aligned} S_0(\boldsymbol{x}, t) &= \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \frac{m}{2} \|\boldsymbol{u}\|_2^2 \right\} \\ &= \frac{m \|\boldsymbol{x}\|_2^2}{2(1 + mt)}. \end{aligned}$$

This minimization problem is a special case of Tikhonov–Phillips regularization (also known as ridge regression in statistics), a method for regularizing ill-posed problems in inverse problems and statistics using a quadratic regularization term [199, 237]. The corresponding minimizer can be computed using the gradient $\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t)$ *via equation* (1.25) *in Proposition 4.2.1:*

$$\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) = \boldsymbol{x} - t \nabla_{\boldsymbol{x}} S_0(\boldsymbol{x}, t) = \boldsymbol{x} - \frac{mt\boldsymbol{x}}{1 + mt} = \frac{\boldsymbol{x}}{1 + mt}.$$

The solution $S_\epsilon(\boldsymbol{x}, t)$ *is given by the integral*

$$\begin{aligned} S_\epsilon(\boldsymbol{x}, t) &= -\epsilon \ln \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \frac{m}{2} \|\boldsymbol{u}\|_2^2 \right)/\epsilon} \, d\boldsymbol{u} \right) \\ &= \frac{m \|\boldsymbol{x}\|^2}{2(1 + mt)} + \frac{n\epsilon}{2} \ln (1 + mt). \end{aligned}$$

The posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ *can be computed using the representation formula* (4.15)

in Proposition 4.2.1(iii) by calculating the gradient $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$*:*

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) = \boldsymbol{x} - \frac{mt\boldsymbol{x}}{1+mt} = \frac{\boldsymbol{x}}{1+mt}.$$

The mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right]$ *can be computed using the representation formula* (4.16) *in Proposition 4.2.1(iii) by calculating the divergence of* $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$*:*

$$\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right] = t\epsilon\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \frac{nt\epsilon}{1+mt}. \tag{4.18}$$

Comparing the solutions $S_0(\boldsymbol{x}, t)$ *and* $S_\epsilon(\boldsymbol{x}, t)$*, we see that* $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x}, t) = S_0(\boldsymbol{x}, t)$ *for every* $\boldsymbol{x} \in \mathbb{R}^n$ *and* $t > 0$*, in accordance to the result established in Proposition 4.2.1(iv). Note also that while* $(\boldsymbol{x}, t) \mapsto S_0(\boldsymbol{x}, t)$ *is jointly convex, its viscous counterpart* $(\boldsymbol{x}, t) \mapsto S_\epsilon(\boldsymbol{x}, t)$ *is not. Indeed,* $t \mapsto S_\epsilon(\boldsymbol{x}, t)$ *is not convex, and it is convex only after subtracting* $\frac{n\epsilon}{2}\ln t$ *from* $S_\epsilon(\boldsymbol{x}, t)$*.*

Example 2 (Soft thresholding)**.** *Let* $J(\boldsymbol{x}) = \sum_{i=1}^n \lambda_i |\boldsymbol{x}_i|$*, where* $\lambda_i > 0$ *for each* $i \in \{1, \ldots, n\}$*, and consider the solutions* $S_0(\boldsymbol{x}, t)$ *and* $S_\epsilon(\boldsymbol{x}, t)$ *to the first-order* (1.22) *and viscous HJ PDEs* (4.14) *with initial data* J*, respectively.*

The solution $S_0(\boldsymbol{x}, t)$ *is given by the Lax–Oleinik formula*

$$\begin{aligned} S_0(\boldsymbol{x}, t) &= \inf_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \sum_{i=1}^n \lambda_i |y_i|\right\} \\ &= \sum_{i=1}^n \left(\inf_{y_i\in\mathbb{R}}\left\{\frac{1}{2t}(x_i - y_i)^2 + \lambda_i |y_i|\right\}\right), \end{aligned}$$

where x_i *and* y_i *denote the* i^{th} *component of the vectors* $\boldsymbol{x}$ *and* $\boldsymbol{u}$*, respectively. In the context of imaging, this minimization problem corresponds to denoising an image with the weighted sum of a quadratic fidelity term and a weighted* l_1*-norm as the regularization term. This term is widely used in imaging to encourage sparsity of an image, and it has received considerable interest due to its connection with compressed sensing reconstruction [38, 82]. The solution to this minimization problem corresponds to a soft thresholding applied component-wise to the vector* $\boldsymbol{x}$ *[74, 104, 164].*

The soft thresholding operator is defined for any real number a and positive real number α as

$$\mathbb{R}\times(0,+\infty)\ni(a,\alpha)\mapsto T(a,\alpha)=\begin{cases}a-\alpha & \text{if } a>\alpha,\\ 0 & \text{if } a\in[-\alpha,\alpha],\\ a+\alpha & \text{if } a<-\alpha.\end{cases} \tag{4.19}$$

The minimizer in the Lax–Oleinik formula of $S_0(\boldsymbol{x},t)$ is then given component-wise for $i\in\{1,\dots,n\}$ by

$$(\boldsymbol{u}_{MAP}(\boldsymbol{x},t))_i=T(x_i,t\lambda_i),$$

so that

$$S_0(\boldsymbol{x},t)=\sum_{i=1}^{n}\left(\frac{1}{2t}(x_i-T(x_i,t\lambda_i))^2+\lambda_i\left|T(x_i,t\lambda_i)\right|\right).$$

The solution $S_\epsilon(\boldsymbol{x},t)$ is given by the integral

$$\begin{aligned}S_\epsilon(\boldsymbol{x},t)&=-\epsilon\ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}}\int_{\mathbb{R}^n}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+\sum_{k=1}^n\lambda_i|y_i|\right)/\epsilon}\,d\boldsymbol{u}\right)\\&=-\epsilon\sum_{i=1}^{n}\ln\left(\frac{1}{2}\sqrt{\frac{2}{\pi t\epsilon}}\int_{-\infty}^{+\infty}e^{-\left(\frac{1}{2t}(x_i-y_i)^2+\lambda_i|y_i|\right)/\epsilon}\,dy_i\right)\\&=-\epsilon\sum_{i=1}^{n}\ln\left(\frac{1}{2}\sqrt{\frac{2}{\pi t\epsilon}}\left(\int_0^{+\infty}e^{-\left(\frac{1}{2t}(x_i+y_i)^2+\lambda_iy_i\right)/\epsilon}\,dy_i+\int_0^{+\infty}e^{-\left(\frac{1}{2t}(x_i-y_i)^2+\lambda_iy_i\right)/\epsilon}\,dy_i\right)\right)\end{aligned}$$

To compute this integral, first define the function

$$\mathbb{R}\ni z\mapsto L(z)=\frac{1}{2}e^{z^2}\operatorname{erfc}(z),$$

where erfc *denotes the complementary error function. Then we have ([121], page 336, integral 3.332, 2., and page 887, integral 8.250, 1.)*

$$\frac{1}{2}\sqrt{\frac{2}{\pi t\epsilon}}\int_0^{+\infty}e^{-\left(\frac{1}{2t}(x_i+y_i)^2+\lambda_iy_i\right)/\epsilon}\,dy_i=e^{-\frac{x_i^2}{2t\epsilon}}L\left(\frac{x_i+t\lambda_i}{\sqrt{2t\epsilon}}\right)$$

and

$$\frac{1}{2}\sqrt{\frac{2}{\pi t\epsilon}}\int_0^{+\infty}e^{-\left(\frac{1}{2t}(x_i-y_i)^2+\lambda_iy_i\right)/\epsilon}\,dy_i=e^{-\frac{x_i^2}{2t\epsilon}}L\left(\frac{-x_i+t\lambda_i}{\sqrt{2t\epsilon}}\right),$$

from which we get

$$S_\epsilon(\boldsymbol{x},t) = \frac{\|\boldsymbol{x}\|_2^2}{2t} - \epsilon \sum_{i=1}^{n} \ln\left(L\left(\frac{x_i + t\lambda_i}{\sqrt{2t\epsilon}}\right) + L\left(\frac{-x_i + t\lambda_i}{\sqrt{2t\epsilon}}\right)\right).$$

Now, to find the posterior mean estimate it suffices to compute the gradient of $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)$ *and use formula* (4.15)*. To do so, we need the derivative of the function* L*. Since*

$$\frac{dL}{dz}(z) = 2zL(z) + \frac{1}{\sqrt{\pi}},$$

the chain rule gives

$$\frac{\partial}{\partial x_i}\left(L\left(\frac{x_i + t\lambda_i}{\sqrt{2t\epsilon}}\right) + L\left(\frac{-x_i + t\lambda_i}{\sqrt{2t\epsilon}}\right)\right) = \left(\frac{x_i + t\lambda_i}{t\epsilon}\right) L\left(\frac{x_i + t\lambda_i}{\sqrt{2t\epsilon}}\right) - \left(\frac{-x_i + t\lambda_i}{t\epsilon}\right) L\left(\frac{-x_i + t\lambda_i}{\sqrt{2t\epsilon}}\right).$$

The posterior mean estimate is therefore given component-wise by

$$\begin{aligned}(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon))_i &= x_i - t(\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t))_i \\ &= x_i + t\lambda_i \left(\frac{L\left(\frac{x_i+t\lambda_i}{\sqrt{2t\epsilon}}\right) + L\left(\frac{-x_i+t\lambda_i}{\sqrt{2t\epsilon}}\right)}{L\left(\frac{x_i+t\lambda_i}{\sqrt{2t\epsilon}}\right) - L\left(\frac{-x_i+t\lambda_i}{\sqrt{2t\epsilon}}\right)} \right)\end{aligned}$$

The posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ *yields a smooth analogue of the soft thresholding operator* T *(defined in* (4.19)*) evaluated at* $(x_i, t\lambda_i)$*, in the sense that* $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} (\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon))_i = T(x_i, t\lambda_i)$ *for every* $i \in \{1,\dots,n\}$ *by Proposition 4.2.1(iv). Figure 4.2.1 shows the MAP and posterior mean estimates in one dimension for the choice of* $t = 1.25$*,* $\epsilon = \{0.025, 0.1, 0.25, 0.5, 1\}$*, and* $\lambda_1 = 2$ *for* $x \in [-5,5]$*.*

4.2.3 Connections to first-order Hamilton–Jacobi equations

In this section, we use the connections between the posterior mean estimate (4.3) and viscous HJ PDEs established in Proposition 4.2.1 to show that the posterior mean estimate can be expressed

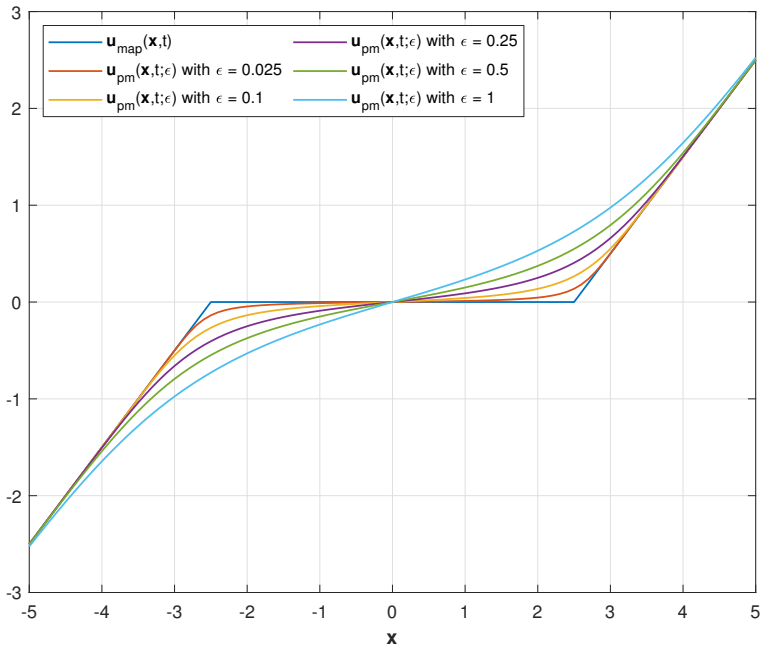


Figure 4.2.1: Numerical example of the MAP and posterior mean estimates in one dimension with $J(x) = \lambda_1 |x|$ for the choice of $t = 1.25$, $\epsilon = \{0.025, 0.1, 0.25, 0.5, 1\}$, and $\lambda_1 = 2$ for $x \in [-5, 5]$.

through the solution to a first-order HJ PDE with initial data of the form of (1.22). In particular, we show that the posterior mean estimate satisfies the proximal mapping formula

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \underset{\boldsymbol{u} \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \left(K_\epsilon^*(\boldsymbol{u}, t) - \frac{1}{2} \|\boldsymbol{u}\|_2^2 \right) \right\},$$

where the function $K_\epsilon \colon \mathbb{R}^n \times \times (0, +\infty) \to \mathbb{R}$ is defined through the solution $S_\epsilon(\boldsymbol{x}, t)$ to the viscous HJ PDE (4.14) via

$$K_\epsilon(\boldsymbol{x}, t) := \frac{1}{2} \|\boldsymbol{x}\|_2^2 - t S_\epsilon(\boldsymbol{x}, t) \equiv t\epsilon \ln \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{\text{dom } J} e^{\left(\frac{1}{t} \langle \boldsymbol{x}, \boldsymbol{u} \rangle - \frac{1}{2t} \|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u}) \right)/\epsilon} \, d\boldsymbol{u} \right),$$

which is convex by Proposition 4.2.1(ii)(d), and where $K_\epsilon^*(\boldsymbol{u}, t)$ denotes the convex conjugate of $\boldsymbol{u} \mapsto K_\epsilon(\boldsymbol{u}, t)$. This result gives the representation of the convex imaging regularization term whose existence was derived by [168, 124, 125, 169] (and later extended to non-quadratic data fidelity terms in [127, 126]). This representation result depends crucially on the connections established between the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ and the viscous HJ PDE (4.14) established in Proposition 4.2.1. Moreover, we also show that $\boldsymbol{u} \mapsto K_\epsilon^*(\boldsymbol{u}, t)$ is at least twice continuously differ-

entiable. This fact implies that the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ for image denoising does not suffer from staircasing effects thanks to a result established in [189, Theorem 3] as proven for Total Variation regularization terms in [168]. Here, our results are applicable to any regularization term J satisfying assumptions (A1)-(A3).

Proposition 4.2.2 (Connections between the posterior mean estimate and first-order HJ PDEs)**.** *Suppose the function J satisfies assumptions (A1)-(A3). For every $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, let $S_\epsilon(\boldsymbol{x}, t)$ denote the solution to the viscous HJ PDE (4.14) with initial data J and let $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ denote the posterior mean estimate (4.3). Consider the first-order HJ PDE*

$$\begin{cases} \dfrac{\partial \tilde{S}}{\partial s}(\boldsymbol{x}, s) + \dfrac{1}{2}\left\|\nabla_{\boldsymbol{x}} \tilde{S}(\boldsymbol{x}, s)\right\|_2^2 = 0 & \text{in } \mathbb{R}^n \times (0, +\infty), \\ \tilde{S}(\boldsymbol{x}, 0) = K_\epsilon^*(\boldsymbol{x}, t) - \dfrac{1}{2}\|\boldsymbol{x}\|_2^2 & \text{in } \mathbb{R}^n. \end{cases} \tag{4.20}$$

Then the initial data $\boldsymbol{x} \mapsto K_\epsilon^*(\boldsymbol{x}, t) - \frac{1}{2}\|\boldsymbol{x}\|_2^2$ is convex, the solution to the HJ PDE (4.20) satisfies the Lax–Oleinik formula

$$\tilde{S}(\boldsymbol{x}, s) = \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2s}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \left(K_\epsilon^*(\boldsymbol{u}, t) - \frac{1}{2}\|\boldsymbol{u}\|_2^2 \right) \right\},$$

and the corresponding minimizer at $s = 1$ is the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$:

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \left(K_\epsilon^*(\boldsymbol{u}, t) - \frac{1}{2}\|\boldsymbol{u}\|_2^2 \right) \right\}. \tag{4.21}$$

Moreover, for every $t > 0$ and $\epsilon > 0$ the function $\mathbb{R}^n \ni \boldsymbol{u} \mapsto K_\epsilon^*(\boldsymbol{u}, t)$ is at least twice continuously differentiable.

Proof. By definition of the function $(\boldsymbol{x}, t) \mapsto K_\epsilon(\boldsymbol{x}, t)$, we can write

$$tS_\epsilon(\boldsymbol{x}, t) + K_\epsilon(\boldsymbol{x}, t) = \frac{1}{2}\|\boldsymbol{x}\|_2^2.$$

As both $\boldsymbol{x} \mapsto tS_\epsilon(\boldsymbol{x}, t)$ and $\boldsymbol{x} \mapsto K_\epsilon(\boldsymbol{x}, t)$ are convex by Proposition 4.2.1(ii)(a) and (d), we can apply Fact 1.2.13(iii) in Section 1.2 to conclude that $\boldsymbol{x} \mapsto K_\epsilon^*(\boldsymbol{x}, t) - \frac{1}{2}\|\boldsymbol{x}\|_2^2$ is convex and to express

$tS_\epsilon(\boldsymbol{x}, t)$ as

$$tS_\epsilon(\boldsymbol{x}, t) = \inf_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \left(K_\epsilon^*(\boldsymbol{u}, t) - \frac{1}{2} \|\boldsymbol{u}\|_2^2 \right) \right\} \tag{4.22}$$

On the one hand, by Proposition 1.2.14 the right hand side of (4.22) is the solution $\tilde{S}_0(\boldsymbol{x}, s)$ to the first-order HJ PDE (4.20) at $s = 1$, and therefore its minimizer is given by $\boldsymbol{x} - \nabla_{\boldsymbol{x}} \tilde{S}_0(\boldsymbol{x}, 1)$. On the other hand, the gradient $\nabla_{\boldsymbol{x}} \tilde{S}_0(\boldsymbol{x}, 1)$ is equal to the left hand side of (4.22), that is, $\nabla_{\boldsymbol{x}} \tilde{S}_0(\boldsymbol{x}, 1) = t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$, which is equal to $\boldsymbol{x} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ by formula (4.3). As a result, the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ minimizes the right hand side of (4.22), that is,

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \left(K_\epsilon^*(\boldsymbol{u}, t) - \frac{1}{2}\|\boldsymbol{u}\|_2^2 \right) \right\}.$$

Now, using the strict convexity of $\boldsymbol{x} \mapsto K_\epsilon(\boldsymbol{x}, t)$ and that $\nabla K_\epsilon(\boldsymbol{x}, t) = \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ is a bijective function in $\boldsymbol{x}$ for every $t > 0$ and $\epsilon > 0$ by Proposition 4.2.1(iii) we can invoke [214, Theorem 26.5] to conclude that $\boldsymbol{u} \mapsto K_\epsilon^*(\boldsymbol{u}, t)$ is a continuously differentiable, strictly convex, and bijective function on $\mathbb{R}^n$, and moreover that $\boldsymbol{u} \mapsto \nabla_{\boldsymbol{u}} K_\epsilon^*(\boldsymbol{u}, t)$ corresponds to the inverse of $\boldsymbol{x} \mapsto \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$, i.e., $\nabla_{\boldsymbol{u}} K_\epsilon^*(\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon), t) = \boldsymbol{x}$. Finally, as $\boldsymbol{x} \mapsto K_\epsilon(\boldsymbol{x}, t)$ is twice differentiable and strictly convex on $\mathbb{R}^n$, the inverse function theorem (see, e.g.,[97, Appendix C, Theorem 7]) implies that $\boldsymbol{u} \mapsto \nabla_{\boldsymbol{u}} K_\epsilon^*(\boldsymbol{u}, t)$ is continuously differentiable on $\mathbb{R}^n$, whence $\boldsymbol{u} \mapsto K_\epsilon(\boldsymbol{u}, t)$. □

4.3 Properties of posterior mean and MAP estimators

In this section, we describe various properties of the Bayesian posterior mean estimate (4.3) in terms of the data $\boldsymbol{x} \in \mathbb{R}^n$, parameters $t > 0$ and $\epsilon > 0$, and the imaging regularization term J. Specifically, in Section 4.3.1, we derive topological, representation, and monotonicity properties of the posterior mean estimate, which we use in Section 4.3.2 to further derive an optimal upper bound on the mean squared error $\mathbb{E}_J \left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2 \right]$, an estimate of the squared difference between the MAP and posterior mean estimates, monotonicity and non-expansive properties of the posterior mean estimate, and the behavior of the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ in the limit $t \to 0$. Finally, we describe the MAP and posterior mean estimates in terms of Bayes risks and

their connections to HJ PDEs in Section 4.3.3.

4.3.1 Topological, representation, and monotonicity properties

This section describes the topological, representation, and monotonicity properties of the Bayesian posterior mean estimate (4.3), which are stated, respectively, in Propositions 4.3.1, 4.3.2, and 4.3.3.

The first result, Proposition 4.3.1, states that the posterior mean estimate belongs in the interior of the domain of J for all data $\boldsymbol{x} \in \mathbb{R}^n$ and parameters $t > 0$ and $\epsilon > 0$.

Proposition 4.3.1 (Topological properties)**.** *Suppose the function J satisfies assumptions (A1)-(A3). Then the following properties hold.*

(i) For every $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ is contained in int (dom J).

(ii) Let $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, and let $S_\epsilon \colon \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}$ denote the solution to the viscous HJ PDEs (4.14) with initial data J. Then the expected value of the initial data $\mathbb{E}_J[J(\boldsymbol{u})]$ satisfies the bounds

$$0 \leqslant J(\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)) \leqslant \mathbb{E}_J[J(\boldsymbol{u})] < \epsilon \left(e^{S_\epsilon(\boldsymbol{x}, t)/\epsilon} - 1 \right) < +\infty. \tag{4.23}$$

Proof. See Appendix 4.B. □

The second result, Proposition 4.3.2, gives representation formulas for the posterior mean estimate. In particular, when the regularization term J satisfies assumptions (A1)-(A3) and dom $J = \mathbb{R}^n$, the posterior mean estimate and mean squared error then satisfy representation formulas in terms of the mean minimal subgradient of J given by $\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right]$. These representation formulas are then used to show that when dom $J \neq \mathbb{R}^n$, the posterior mean estimate can nonetheless be approximated using the first-order HJ PDE (1.22) by smoothing the initial value J via a Moreau–Yosida approximation $S_0(\boldsymbol{x}, \mu)$ with $\mu > 0$.

Proposition 4.3.2 (Representation properties)**.** *Suppose the function J satisfies assumptions (A1)-(A3), let $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, and let $(\boldsymbol{x}, t) \mapsto S_0(\boldsymbol{x}, t)$ and $(\boldsymbol{x}, t) \mapsto S_\epsilon(\boldsymbol{x}, t)$ denote the solutions, respectively, to the first-order and viscous HJ PDEs (1.22) and (4.14) with initial data J.*

(i) (Representation formulas) If dom $J = \mathbb{R}^n$*, then* $\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right\|_2\right] < +\infty$ *and*

$$\mathbb{E}_J\left[\left\langle\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}), \boldsymbol{u}-\boldsymbol{u}_0\right\rangle\right] = n\epsilon, \tag{4.24}$$

for every $\boldsymbol{u}_0 \in \mathbb{R}^n$. In addition, the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ and mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right]$ of the Bayesian posterior distribution (4.10) satisfy the representation formulas

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{x} - t\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right] \tag{4.25}$$

and

$$\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right] = nt\epsilon - t\mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}), \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right]. \tag{4.26}$$

Moreover, the gradient $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$ and Laplacian $\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x}, t)$ satisfy the representation formulas

$$\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) = \mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right] \tag{4.27}$$

and

$$\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x}, t) = \frac{1}{t\epsilon}\mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}), \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right]. \tag{4.28}$$

(ii) (Limit formulas) Let $\{\mu_k\}_{k=1}^{+\infty}$ be a sequence of positive real numbers decreasing to zero. The solution $S_\epsilon(\boldsymbol{x}, t)$ to the viscous HJ PDE (4.14) and its gradient $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$ satisfy the limits

$$\begin{aligned} S_\epsilon(\boldsymbol{x}, t) &:= -\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}}\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u}\right) \\ &= \lim_{k\to+\infty} -\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}}\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\, d\boldsymbol{u}\right) \end{aligned} \tag{4.29}$$

and

$$\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) = \lim_{k \to +\infty} \left(\frac{\int_{\mathbb{R}^n} \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}} \right). \tag{4.30}$$

In particular, the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ *satisfies the limits*

$$\begin{aligned} \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) &= \lim_{k \to +\infty} \left(\frac{\int_{\mathbb{R}^n} \boldsymbol{u} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}} \right) \\ &= \boldsymbol{x} - t \lim_{k \to +\infty} \left(\frac{\int_{\mathbb{R}^n} \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}} \right), \end{aligned} \tag{4.31}$$

Proof. See Appendix 4.C for the proof. □

Remark 4.3.1. *Note that the representation formulas in Proposition 4.3.2(i) may not hold if* dom $J \neq \mathbb{R}^n$. *To see this, consider* $J : \mathbb{R}^n \to [0, +\infty]$ *defined by*

$$J(\boldsymbol{u}) = \begin{cases} 0, & \text{if } \|\boldsymbol{u}\|_2 \leqslant 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

The domain of J *is the unit sphere in* $\mathbb{R}^n$, *which is convex, and* J *satisfies assumptions (A1)-(A3). The function* J *is continuously differentiable on* int (dom J), *with* $\nabla J(\boldsymbol{u}) = \boldsymbol{0}$ *for every* $\boldsymbol{u} \in$ int (dom J). *Clearly,* $\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right] = 0$. *However, for every* $\boldsymbol{x} \neq \boldsymbol{0}$, *the posterior mean estimate* $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \neq \boldsymbol{x}$. *Hence, the representation formula* (4.25) *does not hold in that case.*

The next result, Proposition 4.3.3, uses the properties of solutions to first-order HJ PDEs presented in Proposition 1.2.14 together with the representation formulas (4.25) and (4.26) to describe monotonicity properties of the posterior mean estimate. Proposition 4.3.3 will be leveraged in the next subsection to derive an optimal upper bound for the mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2^2\right]$ and several estimates and limit results of $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ in terms of the observed image $\boldsymbol{x}$ and parameter $t > 0$.

For the statement and proof of Proposition 4.3.3, and later for Proposition 4.3.5, we define the function

$$\text{dom } \partial J \ni \boldsymbol{u} \mapsto \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t) = \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}),$$

which is a subgradient of the convex function $\boldsymbol{u} \ni \boldsymbol{u} \mapsto \frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})$ for every $\boldsymbol{u} \in \operatorname{dom} \partial J$.

Proposition 4.3.3 (Monotonicity property)**.** *Suppose the function J is m-strongly convex and satisfies assumptions (A1)-(A3). Let $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$. Then for every $\boldsymbol{u}_0 \in \operatorname{dom} \partial J$,*

$$\begin{aligned}\left(\frac{1+mt}{t}\right)\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2\right] &\leqslant \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle\right] \\ &\leqslant n\epsilon - \langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{u}_0\rangle\,.\end{aligned} \tag{4.32}$$

Moreover, $\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right\|_2\right] < +\infty$.

Proof. See Appendix 4.D for the proof. □

4.3.2 Error Bounds and limit properties

In this section, we derive an optimal bound for the mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right]$, a bound on the squared difference between the MAP and posterior mean estimates, monotone and non-expansive properties of the posterior mean estimate, and limiting results of the posterior mean estimate in terms of the parameters t.

Proposition 4.3.4 (Error Bounds and limit properties)**.** *Suppose the function J is m-strongly convex and satisfies assumptions (A1)-(A3).*

(i) For every $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, the mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right]$ of the Bayesian posterior distribution (4.10) satisfies the upper bound

$$\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right] \leqslant \frac{nt\epsilon}{1+mt}. \tag{4.33}$$

(ii) For every $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$, the squared difference between the MAP and posterior mean estimates satisfies the upper bound

$$\|\boldsymbol{u}_{MAP}(\boldsymbol{x},t) - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2 \leqslant \frac{nt\epsilon}{1+mt}. \tag{4.34}$$

(iii) The posterior mean estimate is monotone and non-expansive, that is, for every $\boldsymbol{x}$, $\boldsymbol{d} \in \mathbb{R}^n$, $t > 0$, *and* $\epsilon > 0$,

$$\langle \boldsymbol{u}_{PM}(\boldsymbol{x} + \boldsymbol{d}, t, \epsilon) - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon), \boldsymbol{d} \rangle \geqslant 0 \tag{4.35}$$

and

$$\|\boldsymbol{u}_{PM}(\boldsymbol{x} + \boldsymbol{d}, t, \epsilon) - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2 \leqslant \|\boldsymbol{d}\|_2 . \tag{4.36}$$

(iv) Let $\{t_k\}_{k=1}^{+\infty}$ *be a sequence of positive real numbers converging to* 0 *and let* $\{\boldsymbol{d}_k\}_{k=1}^{+\infty}$ *be a sequence of elements of* $\mathbb{R}^n$ *converging to* $\boldsymbol{d} \in \mathbb{R}^n$. *Then for every* $\boldsymbol{x} \in \operatorname{dom} J$ *and* $\epsilon > 0$, *the pointwise limit of* $\boldsymbol{u}_{PM}(\boldsymbol{x} + t_k \boldsymbol{d}_k, t_k, \epsilon)$ *as* $k \to +\infty$ *exists and satisfies*

$$\lim_{k \to +\infty} \boldsymbol{u}_{PM}(\boldsymbol{x} + t_k \boldsymbol{d}_k, t_k, \epsilon) = \boldsymbol{x}.$$

Proof. **Proof of (i):** Since $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \in \operatorname{int}(\operatorname{dom} J)$ by Proposition 4.3.1 and $\operatorname{int}(\operatorname{dom} J) \subset \operatorname{dom} \partial J$ (see Definition 8), we can set $\boldsymbol{u}_0 = \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ in the monotonicity inequality (4.32) in Proposition 4.3.3(i) and rearrange to get the upper bound (4.33).

Proof of (ii): Note that for every $\boldsymbol{u}_0 \in \operatorname{dom} \partial J$, the monotonicity inequality (4.32) in Proposition 4.3.3 yields

$$\mathbb{E}_J \left[\left\langle \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t} + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) \right), \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \right] \leqslant n\epsilon.$$

Choose $\boldsymbol{u}_0 = \boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$, which for every $\boldsymbol{x}$ and $t > 0$ is always an element of $\operatorname{dom} \partial J$ and also satisfies the inclusion $\left(\frac{\boldsymbol{x} - \boldsymbol{u}_{MAP}(\boldsymbol{x}, t)}{t} \right) \in \partial J(\boldsymbol{u}_{MAP}(\boldsymbol{x}, t))$ by part (ii) of Proposition 1.2.14. Hence the monotonicity of the subdifferential of $\boldsymbol{u} \mapsto \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u})$ and m-strong convexity of J implies

$$\left(\frac{1 + mt}{t} \right) \|\boldsymbol{u} - \boldsymbol{u}_{MAP}(\boldsymbol{x}, t)\|_2^2 \leqslant \left\langle \left(\frac{\boldsymbol{x} - \boldsymbol{u}}{t} + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) \right), \boldsymbol{u} - \boldsymbol{u}_{MAP}(\boldsymbol{x}, t) \right\rangle .$$

Combine these inequalities to get $\mathbb{E}_J \left[\|\boldsymbol{u} - \boldsymbol{u}_{MAP}(\boldsymbol{x}, t)\|_2^2 \right] \leqslant \frac{nt\epsilon}{1+mt}$, and use the convexity of the Euclidean norm to get inequality (4.34).

Proof of (iii): The convexity of $\boldsymbol{x} \mapsto K_\epsilon(\boldsymbol{x}, t)$ by Proposition 4.2.1(ii)(d) and $\nabla_{\boldsymbol{x}} K_\epsilon(\boldsymbol{x}, t) =$

$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ implies the monotonicity property (4.35) (see definition 8, equation (1.20), and [214, Page 240, Corollary 31.5.2]). Since both functions $\boldsymbol{x} \mapsto S_\epsilon(\boldsymbol{x}, t)$ and $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_2^2 - tS_\epsilon(\boldsymbol{x}, t)$ are convex by Proposition 4.2.1(ii)(a) and (d), the gradient of the function $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_2^2 - tS_\epsilon(\boldsymbol{x}, t)$, whose value is the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ by Proposition 4.2.1(iii), is Lipschitz continuous with unit constant (see [261] for a simple proof), that is,

$$\|(\boldsymbol{x} + \boldsymbol{d} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x} + \boldsymbol{d}, t)) - (\boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t))\|_2 \equiv \|\boldsymbol{u}_{PM}(\boldsymbol{x} + \boldsymbol{d}, t, \epsilon) - \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)\|_2 \leqslant \|\boldsymbol{d}\|_2 ,$$

which proves the non-expansive inequality (4.36).

Proof of (iv): Inequality (4.34) and the triangle inequality imply

$$\|(\boldsymbol{x} + t_k\boldsymbol{d}_k) - \boldsymbol{u}_{PM}(\boldsymbol{x} + t_k\boldsymbol{d}_k, t_k, \epsilon)\|_2 \leqslant \|(\boldsymbol{x} + t_k\boldsymbol{d}_k) - \boldsymbol{u}_{MAP}(\boldsymbol{x} + t_k\boldsymbol{d}_k, t_k)\|_2 + \sqrt{\frac{nt_k\epsilon}{1 + mt}}.$$

The limit $\lim_{k\to+\infty} \boldsymbol{u}_{PM}(\boldsymbol{x} + t_k\boldsymbol{d}_k, t_k, \epsilon) = \boldsymbol{x}$ then follows by Proposition 1.2.14(iii). □

Remark 4.3.2. *The upper bound for the mean squared error in (4.33) is optimal. As shown in Example 1, it is attained for the quadratic term* $J(\boldsymbol{x}) = \frac{m}{2}\|\boldsymbol{x}\|_2^2$.

4.3.3 Bayesian risks and Hamilton–Jacobi partial differential equations

In this section, we will consider the Bayesian risk associated to the following Bregman divergence (see Definition 14, equation (1.5))

$$\mathbb{R}^n \times \mathbb{R}^n \ni (\boldsymbol{v}, \boldsymbol{u}) \mapsto \begin{cases} D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t)) & \text{if } \boldsymbol{u} \in \text{dom } \partial J, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.37}$$

where

$$\text{dom } \partial J \ni \boldsymbol{u} \mapsto \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t) = \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}),$$

$$\mathbb{R}^n \ni \boldsymbol{u} \mapsto \Phi_J(\boldsymbol{u}|\boldsymbol{x}, t) = \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}).$$

The associated Bayesian risk to the posterior distribution (4.10) corresponds to the expected value $\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t))\right]$. We refer the reader to [12] and [151] for discussions on Bregman loss functions and Bayesian estimation theory.

Here, we will use the connections between maximum a posteriori and posterior mean estimates and Hamilton–Jacobi equations derived in Section 4.2 to show that when the regularization term J is convex on $\mathbb{R}^n$ and bounded from below, then the MAP estimate $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$ minimizes in expectation the Bregman loss function (4.37). We also show that when $\operatorname{dom} J \neq \mathbb{R}^n$ and satisfies assumptions (A1)-(A3). The results rely on the monotonicity property (4.32) established in Proposition 4.3.3.

Proposition 4.3.5 (Bregman divergences)**.** *Suppose the function J satisfies assumptions (A1)-(A3), and let $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $\epsilon > 0$.*

(i) The mean Bregman loss function $\operatorname{dom} J \ni \boldsymbol{v} \mapsto \mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t))\right] \in \mathbb{R}$ has a unique minimizer $\bar{\boldsymbol{v}} \in \operatorname{dom} \partial J$ that satisfies the inclusion

$$\left(\frac{\boldsymbol{x} - \bar{\boldsymbol{v}}}{t}\right) \in \partial J(\bar{\boldsymbol{v}}) + \left(\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) - \mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right]\right), \tag{4.38}$$

where addition in (4.38) is taken in the sense of sets.

(ii) If J is finite everywhere on $\mathbb{R}^n$, then the MAP estimate $\boldsymbol{u}_{MAP}(\boldsymbol{x}, t)$ is the unique global minimizer of the Bregman loss function $\mathbb{R}^n \ni \boldsymbol{v} \mapsto \mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t))\right] \in \mathbb{R}$, that is,

$$\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathbb{R}^n} \mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t))\right] \tag{4.39}$$

Proof. See Appendix 4.E for the proof. □

4.4 Extension to certain non log-concave priors

So far, we have assumed that the regularization term J in the posterior distribution (4.10) and Proposition (4.2.1) is convex. When J is non-convex, the solution to the first-order HJ PDEs (4.9) may not be classical, in the sense that it is not differentiable. For this class of HJ PDEs, the concept of viscosity solutions [15, 18, 19, 62, 97, 107] is generally the appropriate notion of solution. The corresponding Lax–Oleinik formula (4.8), however, becomes a non-convex optimization problem.

This section presents some extensions of the previous results to the case where the regularization term J takes the form

$$J(\boldsymbol{x}) = \min_{i\in\{1,\dots,m\}} J_i(\boldsymbol{x}) \tag{4.40}$$

where every J_i for $i \in \{1,\dots,m\}$ satisfies assumptions (A1)-(A3). Note that in general, this function J is non-convex. Nonetheless, the min-plus algebra technique [2, 3, 83, 106, 115, 156, 175, 176, 178, 177] can handle this case.

4.4.1 Min-plus algebra for first-order HJ PDEs

Let $S_{i,0}\colon \mathbb{R}^n \times [0,+\infty) \to \mathbb{R}$ denote the solution to the HJ PDE

$$\begin{cases} \dfrac{\partial S_{i,0}}{\partial t}(\boldsymbol{x},t) + \dfrac{1}{2}\left\|\nabla_{\boldsymbol{x}} S_{i,0}(\boldsymbol{x},t)\right\|_2^2 = 0 & \text{in } \mathbb{R}^n \times (0,+\infty), \\ S_{i,0}(\boldsymbol{x},0) = J_i(\boldsymbol{x}) & \text{in } \mathbb{R}^n, \end{cases}$$

which is given explicitly by the Lax–Oleinik formula

$$S_{i,0}(\boldsymbol{x},t) = \min_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J_i(\boldsymbol{u})\right\},$$

and let S_0 denote the solution to the HJ PDE (4.9) with initial condition J given by (4.40). By min-plus algebra theory, the semi-group of this HJ PDE is linear with respect to the min-plus

algebra. That is,

$$\begin{aligned} S_0(\boldsymbol{x},t) &= \min_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right\} \\ &= \min_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\min_{i\in\{1,\ldots,m\}}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J_i(\boldsymbol{u})\right\}\right\} \\ &= \min_{i\in\{1,\ldots,m\}}\left\{\min_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J_i(\boldsymbol{u})\right\}\right\} \\ &= \min_{i\in\{1,\ldots,m\}} S_{i,0}(\boldsymbol{x},t). \end{aligned} \tag{4.41}$$

Hence the solution $S_0(\boldsymbol{x},t)$ is given by the pointwise minimum of $S_{i,0}(\boldsymbol{x},t)$ for $i \in \{1,\ldots,m\}$. This approach seems particularly appropriate for solving this non-convex optimization problem and associated HJ PDE. Note that such an approach is embarrassingly parallel since we can solve the initial data J_i for each $i \in \{1,\ldots,m\}$ independently and compute in linear time the pointwise minimum. However, this approach is only feasible if m is not too big; we will show that robust edge preserving priors (e.g., truncated Total Variation or truncated quadratic) can be written in the form of (4.40) where m is exponential in n.

We can also compute the set of minimizers $\boldsymbol{u}(\boldsymbol{x},t)$ as follows. Here, we abuse notation and use $\boldsymbol{u}(\boldsymbol{x},t)$ to denote the set of minimizers, which may be not a singleton set when the minimizer is not unique. We can write

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{x},t) &= \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\min_{i\in\{1,\ldots,m\}}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J_i(\boldsymbol{u})\right\}\right\} \\ &= \bigcup_{i\in I(\boldsymbol{x},t)}\operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J_i(\boldsymbol{u})\right\}, \end{aligned} \tag{4.42}$$

where the index set $I(\boldsymbol{x},t)$ is defined by

$$I(\boldsymbol{x},t) = \operatorname*{arg\,min}_{i\in\{1,\ldots,m\}} S_i(\boldsymbol{x},t). \tag{4.43}$$

As an example, we can take the regularization term J to be the truncated regularization term

with pairwise interactions

$$J(\boldsymbol{x}) = \sum_{(i,j)\in E} w_{ij} f([\boldsymbol{x}]_i - [\boldsymbol{x}]_j), \text{ for each } \boldsymbol{x} = ([\boldsymbol{x}]_1, \ldots, [\boldsymbol{x}]_n) \in \mathbb{R}^n, \tag{4.44}$$

where $w_{ij} \geqslant 0$, $f(x) = \min\{g(x), 1\}$ for some convex function $g\colon \mathbb{R} \to \mathbb{R}$ and $E = \{1, \ldots, n\} \times \{1, \ldots, n\}$. This function can be written as the minimum of a collection of convex functions $J_\Omega\colon \mathbb{R}^n \to \mathbb{R}$ as

$$J(\boldsymbol{x}) = \min_{\Omega \subseteq E} J_\Omega,$$

with each J_Ω defined by

$$J_\Omega(\boldsymbol{x}) := \left\{ \sum_{(i,j)\in\Omega} w_{ij} + \sum_{(i,j)\notin\Omega} w_{ij} g([\boldsymbol{x}]_i - [\boldsymbol{x}]_j) \right\},$$

where Ω is any subset of E. The truncated regularization term (4.44) can therefore be written in the form of (4.40), and hence the minimizer to the corresponding optimization problem (4.41) with the non-convex regularization term J in (4.44) can be computed using (4.42). Concretely, we have

$$\begin{aligned}
\boldsymbol{u}(\boldsymbol{x}, t) &= \bigcup_{\Omega \in I(\boldsymbol{x},t)} \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J_\Omega(\boldsymbol{u}) \right\} \\
&= \bigcup_{\Omega \in I(\boldsymbol{x},t)} \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \sum_{(i,j)\notin\Omega} w_{ij} g([\boldsymbol{u}]_i - [\boldsymbol{u}]_j) \right\} \\
&= \bigcup_{\Omega \in I(\boldsymbol{x},t)} \{\boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\Omega(\boldsymbol{x}, t)\}
\end{aligned}$$

where

$$S_\Omega(\boldsymbol{x}, t) = \sum_{(i,j)\in\Omega} w_{ij} + \min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \sum_{(i,j)\notin\Omega} w_{ij} g([\boldsymbol{u}]_i - [\boldsymbol{u}]_j) \right\}$$

and

$$I(\boldsymbol{x}, t) = \operatorname*{arg\,min}_{\Omega\subseteq E} S_\Omega(\boldsymbol{x}, t).$$

We give here two examples of truncated regularization term with pairwise interactions in the form of (4.44). First, let g be the ℓ_1 norm. Then J is the truncated discrete Total Variation

regularization term defined by

$$J(\boldsymbol{x}) = \sum_{(i,j)\in E} w_{ij} \min\{|[\boldsymbol{x}]_i - [\boldsymbol{x}]_j|, 1\}, \text{ for each } \boldsymbol{x} = ([\boldsymbol{x}]_1, \ldots, [\boldsymbol{x}]_n) \in \mathbb{R}^n. \tag{4.45}$$

This function J can be written as the formula (4.44) with $f\colon \mathbb{R} \to \mathbb{R}$ given by

$$f(z) = \begin{cases} |z| & \text{if } |z| \leqslant 1, \\ 1 & \text{otherwise.} \end{cases}$$

. Second, let g be the quadratic function. Then J is the half-quadratic regularization term defined by

$$J(\boldsymbol{x}) = \sum_{(i,j)\in E} w_{ij} \min\{([\boldsymbol{x}]_i - [\boldsymbol{x}]_j)^2, 1\}, \text{ for each } \boldsymbol{x} = ([\boldsymbol{x}]_1, \ldots, [\boldsymbol{x}]_n) \in \mathbb{R}^n. \tag{4.46}$$

This function J can be written as the formula (4.44) with $f\colon \mathbb{R} \to \mathbb{R}$ given by

$$f(z) = \begin{cases} |z|^2 & \text{if } |z| \leqslant 1, \\ 1 & \text{otherwise.} \end{cases}$$

This specific form of edge-preserving prior was investigated in the seminal works of [52, 116, 117]. Several algorithms have been proposed to solve the resultant non-convex optimization problem (4.41), i.e., the solution to the corresponding HJ PDE, for some specific choice of data fidelity terms (e.g., [5, 144, 116, 117, 192, 50, 191]).

In general, however, there is a drawback to the min-plus algebra technique. To compute the minimizers using (4.42), we need to compute the index set $I(\boldsymbol{x}, t)$ defined in (4.43), which involves solving m HJ PDEs to obtain the solutions $S_{1,0}, \ldots, S_{m,0}$. When m is too large, this approach is impractical since it involves solving too many HJ PDEs. For instance, if J is the truncated Total Variation in (4.45), the number m equals the number of subsets of the set E, i.e., $m = 2^{|E|}$, which is computationally intractable. Hence, in general, it is impractical to use (4.42) to solve the problem (4.41) where the regularization term J is given by the truncated Total Variation. The same issue arises when the truncated Total Variation is replaced by half-quadratic regularization. Several authors attempted to address this intractability for half-quadratic regularizations by proposing

heuristic optimization methods that aim to compute a global minimizer [5, 144, 116, 117, 192, 50, 191].

4.4.2 Analogue of min-plus algebra for viscous HJ PDEs

We consider here an analogue of the min-plus algebra technique designed for certain first order HJ PDEs tailored to viscous HJ PDEs. This will enable us to derive representation formulas for posterior mean estimators whose priors are sums of log-concave priors, i.e., mixture distributions.

The min-plus algebra technique for first order HJ PDEs described in Section 4.4.1 involves initial data of the form $\min_{i\in\{1,\dots,m\}} J_i(\boldsymbol{x})$ where every $J_i\colon \mathbb{R}^n \to [0,+\infty]$ satisfies assumptions (A1)–(A3). Here we consider initial data of the form

$$J(\boldsymbol{x}) = -\epsilon \ln\left(\sum_{i=1}^{m} e^{-J_i(\boldsymbol{x})/\epsilon}\right). \tag{4.47}$$

Note that formula (4.47) approximates the non-convex term (4.40) in that

$$\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} -\epsilon \ln\left(\sum_{i=1}^{m} e^{-J_i(\boldsymbol{x})/\epsilon}\right) = \min_{i\in\{1,\dots,m\}} J_i(\boldsymbol{x}) \text{ for each } \boldsymbol{x}\in\mathbb{R}^n.$$

Now, let

$$S_{i,\epsilon}(\boldsymbol{x},t) = -\epsilon \ln\left(\frac{1}{(2\pi t\epsilon)^{n/2}}\int_{\mathbb{R}^n} e^{-\left(J_i(\boldsymbol{u})+\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon}\,d\boldsymbol{u}\right),$$

and

$$\boldsymbol{u}_{i,PM}(\boldsymbol{x},t,\epsilon) = \frac{\int_{\mathbb{R}^n} \boldsymbol{u}\, e^{-\left(J_i(\boldsymbol{u})+\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon}\,d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(J_i(\boldsymbol{u})+\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon}\,d\boldsymbol{u}}$$

denote, respectively, the solution to the viscous HJ PDE (4.13) with initial data J_i and its associated posterior mean. Then, a short calculation shows that for every $\epsilon > 0$, the function $S_\epsilon(\boldsymbol{x},t)\colon \mathbb{R}^n \times$

$(0, +\infty) \to \mathbb{R}$ defined by

$$\begin{aligned} S_\epsilon(\boldsymbol{x}, t) &= -\epsilon \ln \left(\sum_{i=1}^{m} \frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(J_i(\boldsymbol{u}) + \frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon} \, d\boldsymbol{u} \right) \\ &= -\epsilon \ln \left(\sum_{i=1}^{m} e^{-S_{i,\epsilon}(\boldsymbol{x},t)/\epsilon} \right) \end{aligned} \tag{4.48}$$

is the unique smooth solution to the viscous HJ PDE (4.13) with initial data (4.47). As stated in Section (4.2.2), the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ is given by the representation formula

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t), \tag{4.49}$$

which can be expressed in terms of the solutions $S_{i,\epsilon}(\boldsymbol{x}, t)$, their spatial gradients $\nabla_{\boldsymbol{x}} S_{i,\epsilon}(\boldsymbol{x}, t)$, and posterior mean estimates $\boldsymbol{u}_{i,PM}(\boldsymbol{x}, t, \epsilon)$ as the weighted sums

$$\begin{aligned} \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) &= \boldsymbol{x} - t \left(\frac{\sum_{i=1}^{m} \nabla_{\boldsymbol{x}} S_{i,\epsilon}(\boldsymbol{x}, t) e^{-S_{i,\epsilon}(\boldsymbol{x},t)/\epsilon}}{\sum_{i=1}^{m} e^{-S_{i,\epsilon}(\boldsymbol{x},t)/\epsilon}} \right) \\ &= \frac{\sum_{i=1}^{m} \boldsymbol{u}_{i,PM}(\boldsymbol{x}, t, \epsilon) e^{-S_{i,\epsilon}(\boldsymbol{x},t)/\epsilon}}{\sum_{i=1}^{m} e^{-S_{i,\epsilon}(\boldsymbol{x},t)/\epsilon}}. \end{aligned} \tag{4.50}$$

As an application of this result, consider the problem of denoising a noisy image $\boldsymbol{x} \in \mathbb{R}^n$ using a Gaussian mixture model [85]. Suppose $J_i(\boldsymbol{u}) = \frac{1}{2\sigma_i^2}\|\boldsymbol{u} - \boldsymbol{\mu}_i\|_2^2$, where $\boldsymbol{\mu}_i \in \mathbb{R}^n$ and $\sigma_i > 0$. The regularized minimization problem (4.41) is given by

$$\begin{aligned} S_0(\boldsymbol{x}, t) &= \min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \min_{i \in \{1,\dots,m\}} \left\{ \frac{1}{2\sigma_i^2}\|\boldsymbol{u} - \boldsymbol{\mu}_i\|_2^2 + \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \right\} \right\} \\ &= \min_{i \in \{1,\dots,m\}} \left\{ \min_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2\sigma_i^2}\|\boldsymbol{u} - \boldsymbol{\mu}_i\|_2^2 + \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \right\} \right\} \\ &= \min_{i \in \{1,\dots,m\}} \left\{ \frac{1}{2(\sigma_i^2 + t)}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|_2^2 \right\}. \end{aligned} \tag{4.51}$$

Letting $I(\boldsymbol{x}, t) = \arg\min_{i \in \{1,\dots,m\}} \left\{ \frac{1}{2(\sigma_i^2 + t)}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|_2^2 \right\}$, the MAP estimator is then the collection

$$\boldsymbol{u}_{MAP}(\boldsymbol{x}, t) = \bigcup_{i \in I(\boldsymbol{x},t)} \left\{ \frac{\sigma_i^2 \boldsymbol{x} + t\boldsymbol{\mu}_i}{\sigma_i^2 + t} \right\}.$$

Consider now the initial data (4.47):

$$J(\boldsymbol{u}) = -\epsilon \ln \left(\sum_{i=1}^{m} e^{-\frac{1}{2\sigma_i^2 \epsilon} \|\boldsymbol{u}-\boldsymbol{\mu}_i\|_2^2} \right).$$

The solution $S_\epsilon(\boldsymbol{x}, t)$ to the viscous HJ PDE (4.13) with initial data J is given by formula (4.48), which in this case can be computed analytically:

$$S_\epsilon(\boldsymbol{x}, t) = -\epsilon \ln \left(\sum_{i=1}^{m} \left(\frac{\sigma_i^2}{\sigma_i^2 + t} \right)^{n/2} e^{-\frac{1}{2(\sigma_i^2+t)\epsilon} \|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2} \right). \tag{4.52}$$

Since $e^{-S_{i,\epsilon}(\boldsymbol{x},t)/\epsilon} = \left(\frac{\sigma_i^2}{\sigma_i^2+t} \right)^{n/2} e^{-\frac{1}{2(\sigma_i^2+t)\epsilon} \|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2}$, we can write the corresponding posterior mean estimator (4.50) using the representation formulas (4.49) and (4.50):

$$\begin{aligned} \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) &= \boldsymbol{x} - t \nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t) \\ &= \frac{\sum_{i=1}^{m} \left(\frac{\sigma_i^2 \boldsymbol{x} + t\boldsymbol{\mu}_i}{\sigma_i^2+t} \right) \left(\frac{\sigma_i^2}{\sigma_i^2+t} \right)^{n/2} e^{-\frac{1}{2(\sigma_i^2+t)\epsilon} \|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2}}{\sum_{i=1}^{m} \left(\frac{\sigma_i^2}{\sigma_i^2+t} \right)^{n/2} e^{-\frac{1}{2(\sigma_i^2+t)\epsilon} \|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2}}. \end{aligned} \tag{4.53}$$

4.5 Discussion

This chapter presented novel theoretical connections between Hamilton–Jacobi partial differential equations and a broad class of Bayesian posterior mean estimators with quadratic data fidelity term and log-concave prior relevant to image denoising problems. We derived a representation formula for the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon)$ in terms of the spatial gradient of the solution to a viscous HJ PDE with initial data corresponding to the convex regularization term J. We used these connections to show that the posterior mean estimate can be expressed through the gradient of the solution to a first-order HJ PDE with twice continuously differentiable convex initial data. Furthermore, we derived a novel representation formula for this initial data that was not available in the literature.

In addition, we used the connections between HJ PDEs and Bayesian posterior mean estima-

tors to establish several topological, representation and monotonicity properties of posterior mean estimates. These properties were then used to derive an optimal upper bound on the mean squared error $\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right]$, an estimate of the squared difference between the MAP and posterior mean estimates, monotonicity and non-expansive properties of the posterior mean estimate, and the behavior of the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ in the limit $t\to 0$. We also used the connections between both MAP and posterior mean estimates and HJ PDEs to show that the MAP estimate (4.2) corresponds to the Bayes estimator of the Bayesian risk (4.37) whenever the regularization term J is convex on $\mathbb{R}^n$ and bounded from below and the data fidelity term is quadratic. We also show that when dom $J\neq\mathbb{R}^n$, the Bayesian risk (4.37) has still a Bayes estimator that is described in terms of the solution to both the first-order HJ PDE (1.2.14) and the viscous HJ PDE (4.2.1). Finally, we extended some of the results above to a class of posterior mean estimators whose priors are sums of log-concave priors, that is, to posterior mean estimators of mixture distributions.

We wish to note that in addition to its relevance to image denoising problems, the viscous HJ PDE (4.14) has recently received some attention in the deep learning literature, where its solution $\boldsymbol{x}\mapsto S_\epsilon(\boldsymbol{x},t)$ is known as the local entropy loss function and is a loss regularization effective at training deep networks [53, 54, 114, 242]. While this chapter focuses on HJ PDEs and Bayesian estimators in imaging sciences, the results are relevant to the deep learning literature and may give new theoretical understandings of the local entropy loss function in terms of the data $\boldsymbol{x}$ and parameters t and ϵ.

The results presented in this work crucially depend on the data fidelity term being quadratic and the generalized prior distribution $\boldsymbol{u}\mapsto e^{-J(\boldsymbol{u})}$ being log-concave. This chapter did not consider non-quadratic data fidelity terms (corresponding to non-Gaussian additive noise models) with log-concave priors, or non-additive noise models [28, 31]. These directions will be pursued elsewhere in the future.

Appendix

4.A Proof of Proposition 4.2.1

We will use the following lemma, which characterizes the partition function (4.11) in terms of the solution to a Cauchy problem involving the heat equation with initial data $J \in \Gamma_0(\mathbb{R}^n)$, to prove parts (i) and (ii)(a)-(d) of Proposition 4.2.1.

Lemma 4.A.1 (The heat equation with initial data in $\Gamma_0(\mathbb{R}^n)$)**.** *Suppose the function $J\colon \mathbb{R}^n \to [0,+\infty]$ satisfies assumptions (A1)-(A3).*

(i) For every $\epsilon > 0$, the function $w_\epsilon\colon \mathbb{R}^n \times [0,+\infty) \to (0,1]$ defined by

$$w_\epsilon(\boldsymbol{x},t) \coloneqq \frac{1}{(2\pi t\epsilon)^{n/2}} Z_J(\boldsymbol{x},t,\epsilon) = \frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \tag{4.54}$$

is the unique smooth solution to the Cauchy problem

$$\begin{cases} \dfrac{\partial w_\epsilon}{\partial t}(\boldsymbol{x},t) = \dfrac{\epsilon}{2}\nabla_{\boldsymbol{x}}^2 w_\epsilon(\boldsymbol{x},t) & \text{in } \mathbb{R}^n \times (0,+\infty), \\ w_\epsilon(\boldsymbol{x},0) = e^{-J(\boldsymbol{x})/\epsilon} & \text{in } \mathbb{R}^n. \end{cases} \tag{4.55}$$

In addition, the domain of integration of the integral (4.54) can be taken to be dom J *or, up to a set of Lebesgue measure zero,* int (dom J) *or* dom ∂J*. Furthermore, for every $\boldsymbol{x} \in \mathbb{R}^n$ and $\epsilon > 0$, except possibly at the points $\boldsymbol{x} \in (\text{dom } J) \setminus (\text{int dom } J)$ if such points exist, the*

pointwise limit of $w_\epsilon(\boldsymbol{x}, t)$ *as* $t \to 0$ *exists and satisfies*

$$\lim_{\substack{t\to 0 \\ t>0}} w_\epsilon(\boldsymbol{x}, t) = e^{-J(\boldsymbol{x})/\epsilon},$$

with the limit equal to 0 *whenever* $\boldsymbol{x} \notin \operatorname{dom} J$.

(ii) (Log-concavity and monotonicity properties).

(a) The function $\mathbb{R}^n \times (0, +\infty) \ni (\boldsymbol{x}, t) \mapsto t^{n/2} w_\epsilon(\boldsymbol{x}, t)$ *is jointly log-concave.*

(b) The function $(0, +\infty) \ni t \mapsto t^{n/2} w_\epsilon(\boldsymbol{x}, t)$ *is strictly monotone increasing.*

(c) The function $(0, +\infty) \ni \epsilon \mapsto \epsilon^{n/2} w_\epsilon(\boldsymbol{x}, t)$ *is strictly monotone increasing.*

(d) The function $\mathbb{R}^n \ni \boldsymbol{x} \mapsto e^{\frac{1}{2t\epsilon}\|\boldsymbol{x}\|_2^2} w_\epsilon(\boldsymbol{x}, t)$ *is strictly log-convex.*

The proof of (i) follows from classical PDEs arguments for the Cauchy problem (4.55) tailored to the initial data $(\boldsymbol{x}, \epsilon) \mapsto e^{-J(\boldsymbol{x})/\epsilon}$ with J satisfying assumptions (A1)-(A3), and the proof of log-concavity and monotonicity (ii)(a)-(d) follows from the Prékopa–Leindler and Hölder's inequalities [162, 209, 108]; we present the details below.

Proof. **Proof of Lemma 4.A.1 (i):** This result follows directly from the theory of convolution of Schwartz distributions ([141], Chapter 2, Section 2.1, Chapter 4, Sections 4.2 and 4.4., and in particular Theorem 4.4.1 on page 110). To see why this is the case, note that by assumptions (A1)-(A3) the initial condition $\boldsymbol{u} \mapsto e^{-J(\boldsymbol{u})}$ is a locally integrable function, and locally integrable functions are Schwartz distributions.

Proof of Lemma 4.A.1 (ii)(a): The log-concavity property will be shown using the Prékopa–Leindler inequality.

Theorem 4.A.1. *[Prékopa–Leindler inequality [162, 209]] Let* f, g, *and* h *be non-negative real-valued and measurable functions on* $\mathbb{R}^n$, *and suppose*

$$h(\lambda \boldsymbol{u}_1 + (1-\lambda)\boldsymbol{u}_2) \geqslant f(\boldsymbol{u}_1)^\lambda g(\boldsymbol{u}_2)^{(1-\lambda)}$$

for every $\boldsymbol{u}_1,\, \boldsymbol{u}_2 \in \mathbb{R}^n$ *and* $\lambda \in (0,1)$. *Then*

$$\int_{\mathbb{R}^n} h(\boldsymbol{u})\, d\boldsymbol{u} \geqslant \left(\int_{\mathbb{R}^n} f(\boldsymbol{u})\, d\boldsymbol{u}\right)^{\lambda} \left(\int_{\mathbb{R}^n} g(\boldsymbol{u})\, d\boldsymbol{u}\right)^{(1-\lambda)}.$$

Proof of Lemma 4.A.1 (ii)(a) (continued): Let $\epsilon > 0$, $\lambda \in (0,1)$, $\boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2$, $\boldsymbol{u} = \lambda \boldsymbol{u}_1 + (1-\lambda)\boldsymbol{u}_2$, and $t = \lambda t_1 + (1-\lambda) t_2$ for any $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{u}_1$, $\boldsymbol{u}_2 \in \mathbb{R}^n$ and t_1, $t_2 \in (0, +\infty)$. The joint convexity of the function $\mathbb{R}^n \times (0,+\infty) \ni (\boldsymbol{z}, t) \mapsto \frac{1}{2t} \|\boldsymbol{z}\|_2^2$ and convexity of J imply

$$\frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \leqslant \frac{\lambda}{2t_1} \|\boldsymbol{x}_1 - \boldsymbol{u}_1\|_2^2 + \frac{(1-\lambda)}{2t_2} \|\boldsymbol{x}_2 - \boldsymbol{u}_2\|_2^2 + \lambda J(\boldsymbol{u}_1) + (1-\lambda) J(\boldsymbol{u}_2),$$

This gives

$$\frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}}{(2\pi\epsilon)^{n/2}} \geqslant \left(\frac{e^{-\left(\frac{1}{2t_1}\|\boldsymbol{x}_1-\boldsymbol{u}_1\|_2^2 + J(\boldsymbol{u}_1)\right)/\epsilon}}{(2\pi\epsilon)^{n/2}}\right)^{\lambda} \left(\frac{e^{-\left(\frac{1}{2t_2}\|\boldsymbol{x}_2-\boldsymbol{u}_2\|_2^2 + J(\boldsymbol{u}_2)\right)/\epsilon}}{(2\pi\epsilon)^{n/2}}\right)^{1-\lambda}.$$

Applying the Prékopa–Leindler inequality with

$$h(\boldsymbol{u}) = \frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}}{(2\pi\epsilon)^{n/2}},$$

$$f(\boldsymbol{u}) = \frac{e^{-\left(\frac{1}{2t_1}\|\boldsymbol{x}_1-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}}{(2\pi\epsilon)^{n/2}},$$

and

$$g(\boldsymbol{u}) = \frac{e^{-\left(\frac{1}{2t_2}\|\boldsymbol{x}_2-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}}{(2\pi\epsilon)^{n/2}},$$

and using the definition (4.54) of $w_\epsilon(\boldsymbol{x}, t)$, we get

$$t^{n/2} w_\epsilon(\boldsymbol{x}, t) \geqslant \left(t_1^{n/2} w_\epsilon(\boldsymbol{x}_1, t_1)\right)^{\lambda} \left(t_2^{n/2} w_\epsilon(\boldsymbol{x}_2, t_2)\right)^{(1-\lambda)},$$

As a result, the function $(\boldsymbol{x}, t) \mapsto t^{n/2} w_\epsilon(\boldsymbol{x}, t)$ is jointly log-concave on $\mathbb{R}^n \times (0, +\infty)$.

Proof of Lemma 4.A.1 (ii)(b): Since $t \mapsto \frac{1}{t}$ is strictly monotone decreasing on $(0, +\infty)$,

then for every $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{u} \in \operatorname{dom} J$, $\epsilon > 0$, and $0 < t_1 < t_2$,

$$\frac{e^{-\left(\frac{1}{2t_1}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}}{(2\pi\epsilon)^{n/2}} < \frac{e^{-\left(\frac{1}{2t_2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}}{(2\pi\epsilon)^{n/2}}$$

whenever $\boldsymbol{x} \neq \boldsymbol{u}$. Integrating both sides of the inequality with respect to $\boldsymbol{u}$ over dom J yields

$$\frac{1}{(2\pi\epsilon)^{n/2}} \int_{\operatorname{dom}\ J} e^{-\left(\frac{1}{2t_1}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} < \frac{1}{(2\pi\epsilon)^{n/2}} \int_{\operatorname{dom}\ J} e^{-\left(\frac{1}{2t_2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u},$$

As a result, the function $t \mapsto t^{n/2} w_\epsilon(\boldsymbol{x}, t)$ is strictly monotone increasing on $(0, +\infty)$.

Proof of Lemma 4.A.1 (ii)(c): Since $\epsilon \mapsto \frac{1}{\epsilon}$ is strictly monotone decreasing on $(0, +\infty)$ and dom $J \ni \boldsymbol{u} \mapsto J(\boldsymbol{u})$ is non-negative by assumption (A3), then for every $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, and $0 < \epsilon_1 < \epsilon_2$ we have

$$e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon_1} < e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon_2}$$

whenever $\boldsymbol{x} \neq \boldsymbol{u}$. Integrating both sides of the inequality with respect to $\boldsymbol{u}$ over dom J yields

$$\int_{\operatorname{dom}\ J} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon_1} d\boldsymbol{u} < \int_{\operatorname{dom}\ J} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon_2}\, d\boldsymbol{u},$$

As a result, the function $\epsilon \mapsto \epsilon^{n/2} w_\epsilon(\boldsymbol{x}, t)$ is strictly monotone increasing on $(0, +\infty)$.

Proof of Lemma 4.A.1 (ii)(d): Let $\epsilon > 0$, $t > 0$, $\lambda \in (0, 1)$, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ with $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ and $\boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2$. Then

$$\begin{aligned} e^{\frac{1}{2t\epsilon}\|\boldsymbol{x}\|_2^2} w_\epsilon(\boldsymbol{x}, t) &= \frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\operatorname{dom}\ J} e^{\left(\langle \boldsymbol{x}, \boldsymbol{u}\rangle/t - \frac{1}{2t}\|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \\ &= \int_{\operatorname{dom}\ J} \left(\frac{e^{\left(\langle \boldsymbol{x}_1, \boldsymbol{u}\rangle/t - \frac{1}{2t}\|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u})\right)/\epsilon}}{(2\pi t\epsilon)^{n/2}}\right)^{\lambda} \left(\frac{e^{\langle \boldsymbol{x}_2, \boldsymbol{u}\rangle/t\epsilon - \frac{1}{2t}\|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u})/\epsilon}}{(2\pi t\epsilon)^{n/2}}\right)^{1-\lambda} d\boldsymbol{u}. \end{aligned}$$

Hölder's inequality ([108], Theorem 6.2) then implies

$$e^{\frac{1}{2t\epsilon}\|\boldsymbol{x}\|_2^2} w_\epsilon(\boldsymbol{x},t) \leqslant \left(\int_{\text{dom } J} \frac{e^{\left(\langle \boldsymbol{x}_1,\boldsymbol{u}\rangle/t - \frac{1}{2t}\|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u})\right)/\epsilon}}{(2\pi t\epsilon)^{n/2}}\, d\boldsymbol{u}\right)^{\lambda} \left(\int_{\text{dom } J} \frac{e^{\langle \boldsymbol{x}_2,\boldsymbol{u}\rangle/t\epsilon - \frac{1}{2t}\|\boldsymbol{u}\|_2^2 - J(\boldsymbol{u})/\epsilon}}{(2\pi t\epsilon)^{n/2}}\, d\boldsymbol{u}\right)^{1-\lambda}$$
$$= \left(e^{\frac{1}{2t\epsilon}\|\boldsymbol{x}_1\|_2^2} w_\epsilon(\boldsymbol{x}_1,t)\right)^{\lambda} \left(e^{\frac{1}{2t\epsilon}\|\boldsymbol{x}_2\|_2^2} w_\epsilon(\boldsymbol{x}_2,t)\right)^{1-\lambda},$$

where the inequality in the equation above is an equality if and only if there exists a constant $\alpha \in \mathbb{R}$ such that $\alpha e^{\langle \boldsymbol{x}_1,\boldsymbol{u}\rangle/t\epsilon} = e^{\langle \boldsymbol{x}_x,\boldsymbol{u}\rangle/t\epsilon}$ for almost every $\boldsymbol{u} \in \text{dom } J$. This does not hold here since $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$. As a result, the function $\mathbb{R}^n \ni \boldsymbol{x} \mapsto e^{\frac{1}{2t\epsilon}\|\boldsymbol{x}\|_2^2} w_\epsilon(\boldsymbol{x},t)$ is strictly log-convex. □

Proof of Proposition 4.2.1 (i) and (ii)(a)-(d): The proof of these statements follow from Lemma 4.A.1 and classic results about the Cole–Hopf transform (see, e.g., [97], Section 4.4.1), with $S_\epsilon(\boldsymbol{x},t) := -\epsilon \log(w_\epsilon(\boldsymbol{x},t))$.

Proof of Proposition 4.2.1 (iii): The formulas follow from a straightforward calculation of the gradient, divergence, and Laplacian of $S_\epsilon(\boldsymbol{x},t)$ that we omit here. Since the function $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_2^2 - tS_\epsilon(\boldsymbol{x},t)$ is strictly convex, we can invoke [214, Corollary 26.3.1] to conclude that its gradient $\boldsymbol{x} \mapsto \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)$, which gives the posterior mean $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$, is bijective.

Proof of Proposition 4.2.1 (iv): We will prove this result in three steps. First, we will show that

$$\limsup_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x},t) \leqslant \inf_{\boldsymbol{u}\in \text{int } (\text{dom } J)} \left\{\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right\}$$

and

$$\inf_{\boldsymbol{u}\in \text{int } (\text{dom } J)} \left\{\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right\} = \inf_{\boldsymbol{u}\in \text{dom } J} \left\{\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right\} \equiv S_0(\boldsymbol{x},t).$$

Next, we will show that $\liminf_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x},t) \geqslant S_0(\boldsymbol{x},t)$. Finally, we will use steps 1 and 2 to conclude that $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x},t) = S_0(\boldsymbol{x},t)$. Pointwise and local uniform convergence of the gradient $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) = \nabla_{\boldsymbol{x}} S_0(\boldsymbol{x},t)$, the partial derivative $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \frac{\partial S_\epsilon(\boldsymbol{x},t)}{\partial t} = \frac{\partial S_0(\boldsymbol{x},t)}{\partial t}$, and the Laplacian $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \frac{\epsilon}{2}\nabla_{\boldsymbol{x}}^2 S_\epsilon(\boldsymbol{x},t) = 0$ then follow from the convexity and differentiability of the solutions $(\boldsymbol{x},t) \mapsto S_0(\boldsymbol{x},t)$ and $(\boldsymbol{x},t) \mapsto S_\epsilon(\boldsymbol{x},t)$ to the HJ PDEs (1.22) and (4.14).

In what follows, we will use the following large deviation principle result [79]: For every Lebesgue measurable set $\mathcal{A} \in \mathbb{R}^n$,

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} -\epsilon \ln \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{\mathcal{A}} e^{-\frac{1}{2t\epsilon} \|\boldsymbol{x}-\boldsymbol{u}\|_2^2} \, d\boldsymbol{u} \right) = \operatorname*{ess\,inf}_{\boldsymbol{u} \in \mathcal{A}} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \right\},$$

where

$$\operatorname*{ess\,inf}_{\boldsymbol{u} \in \mathcal{A}} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \right\} = \sup \left\{ a \in \mathbb{R} \colon a \leqslant \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \text{, for a.e. } \boldsymbol{u} \in \mathbb{R}^n \right\}.$$

Step 1. (Adapted from Deuschel and Stroock [79, Lemma 2.1.7].) By convexity, the function J is continuous for every $\boldsymbol{u}_0 \in \text{int (dom } J)$, the latter set being open. Therefore, for every such $\boldsymbol{u}_0$ there exists a number $r_{\boldsymbol{u}_0} > 0$ such that for every $0 < r \leqslant r_{\boldsymbol{u}_0}$ the open ball $B_r(\boldsymbol{u}_0)$ is contained in int (dom J). Hence

$$\begin{aligned} S_\epsilon(\boldsymbol{x}, t) &\coloneqq -\epsilon \ln \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{\text{int (dom } J)} e^{-(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u}))/\epsilon} \, d\boldsymbol{u} \right) \\ &\leqslant -\epsilon \ln \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{B_r(\boldsymbol{u}_0)} e^{-(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u}))/\epsilon} \, d\boldsymbol{u} \right) \\ &\leqslant -\epsilon \ln \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{B_r(\boldsymbol{u}_0)} e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2} \, d\boldsymbol{u} \right) + \sup_{\boldsymbol{u} \in B_r(\boldsymbol{u}_0)} J(\boldsymbol{u}). \end{aligned}$$

Take $\limsup_{\substack{\epsilon \to 0 \\ \epsilon > 0}}$ and apply the large deviation principle to the term on the right to get

$$\limsup_{\substack{\epsilon \to 0 \\ \epsilon > 0}} S_\epsilon(\boldsymbol{x}, t) \leqslant \operatorname*{ess\,inf}_{\boldsymbol{u} \in B_r(\boldsymbol{u}_0)} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \right\} + \sup_{\boldsymbol{u} \in B_r(\boldsymbol{u}_0)} J(\boldsymbol{u}).$$

Take $\lim_{r \to 0}$ on both sides of the inequality to find

$$\limsup_{\substack{\epsilon \to 0 \\ \epsilon > 0}} S_\epsilon(\boldsymbol{x}, t) \leqslant \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}_0\|_2^2 + J(\boldsymbol{u}_0).$$

Since the inequality holds for every $\boldsymbol{u}_0 \in \text{int (dom } J)$, we can take the infimum over all $y \in \text{int (dom } J)$ on the right-hand-side of the inequality to get

$$\limsup_{\substack{\epsilon \to 0 \\ \epsilon > 0}} S_\epsilon(\boldsymbol{x}, t) \leqslant \inf_{\boldsymbol{u} \in \text{int (dom } J)} \left\{ \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\}. \tag{4.56}$$

By assumptions (A1) and (A2) that $J \in \Gamma_0(\mathbb{R}^n)$ and int (dom J) $\neq \varnothing$, the infimum on the right hand side is equal to that taken over dom J ([214], Corollary 7.3.2), i.e.,

$$\inf_{\boldsymbol{u}\in \text{int } (\text{dom } J)} \left\{ \frac{1}{2t} \|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} = \inf_{\boldsymbol{u}\in \text{dom } J} \left\{ \frac{1}{2t} \|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} \equiv S_0(\boldsymbol{x},t). \tag{4.57}$$

We combine (4.56) and (4.57) to obtain

$$\limsup_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x},t) \leqslant S_0(\boldsymbol{x},t),$$

which is the desired result.

Step 2. We can invoke Lemma 2.1.8 in [79] because its conditions are satisfied (in the notation of [79], $\Phi = -J$, which is upper semicontinuous, $\boldsymbol{u} \mapsto \frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2$ is the rate function, and note that the tail condition (2.1.9) is satisfied in that $\sup_{\boldsymbol{u}\in\mathbb{R}^n} -J(\boldsymbol{u}) = -\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) = 0$ by assumption (A3)) to get

$$\liminf_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x},t) \geqslant S_0(\boldsymbol{x},t).$$

Step 3. Combining the two limits derived in steps 1 and 2 yield

$$\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} S_\epsilon(\boldsymbol{x},t) = S_0(\boldsymbol{x},t)$$

for every $\boldsymbol{x} \in \mathbb{R}^n$ and $t > 0$, where the limit converges uniformly on every compact subset $(\boldsymbol{x}, t)$ of $\mathbb{R}^n \times (0, +\infty)$ ([214], Theorem 10.8).

By differentiability and joint convexity of both $\mathbb{R}^n \times (0,+\infty) \ni (\boldsymbol{x},t) \mapsto S_0(\boldsymbol{x},t)$ and $\mathbb{R}^n \times (0,+\infty) \ni (\boldsymbol{x},t) \mapsto S_\epsilon(\boldsymbol{x},t) - \frac{n\epsilon}{2}\ln t$ (Proposition 1.2.14 (i), and Proposition 4.2.1 (i) and (ii)(a)), we can invoke ([214], Theorem 25.7) to get

$$\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) = \nabla_{\boldsymbol{x}} S_0(\boldsymbol{x},t) \text{ and } \lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \left(\frac{\partial S_\epsilon(\boldsymbol{x},t)}{\partial t} - \frac{n\epsilon}{2t} \right) = \lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \frac{\partial S_\epsilon(\boldsymbol{x},t)}{\partial t} = \frac{\partial S_0(\boldsymbol{x},t)}{\partial t},$$

for every $\boldsymbol{x} \in \mathbb{R}^n$ and $t > 0$, where the limit converges uniformly on every compact subset of

$\mathbb{R}^n \times (0, +\infty)$. Furthermore, the viscous HJ PDE (4.14) for S_ϵ implies that

$$\begin{aligned}\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \frac{\epsilon}{2}\nabla^2_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) &= \lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \left(\frac{\partial S_\epsilon(\boldsymbol{x},t)}{\partial t} + \frac{1}{2}\left\|\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)\right\|^2\right),\\ &= \left(\frac{\partial S_0(\boldsymbol{x},t)}{\partial t} + \frac{1}{2}\left\|\nabla_{\boldsymbol{x}} S_0(\boldsymbol{x},t)\right\|^2\right)\\ &= 0,\end{aligned}$$

where the last equality holds thanks to the HJ PDE (1.22) (see Proposition 1.2.14). Here, again, the limit holds for every $\boldsymbol{x} \in \mathbb{R}^n$ and $t > 0$, and the limit converges uniformly over any compact subset of $\mathbb{R}^n \times (0, +\infty)$. Finally, the limit $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) = \boldsymbol{u}_{MAP}(\boldsymbol{x},t)$ holds directly as a consequence to the limit $\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) = \nabla_{\boldsymbol{x}} S_0(\boldsymbol{x},t)$ and the representation formulas (4.15) (see Proposition 4.2.1(iii)) and (1.25) (see Proposition 1.2.14(ii))for the posterior mean and MAP estimates, respectively.

4.B Proof of Proposition 4.3.1

Proof of (i): We will prove that $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) \in \text{int (dom } J)$ in two steps. First, we will use the projection operator (1.7) (see Definition 16) and the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ to prove by contradiction that $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) \in \text{cl (dom } J)$. Second, we will use the following variant of the Hahn–Banach theorem for convex bodies in $\mathbb{R}^n$ to show in fact that $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) \in \text{int (dom } J)$.

Theorem 4.B.1. *([214], Theorem 11.6 and Corollary 11.6.2) Let C be a convex set. A point $\boldsymbol{v} \in C$ is a relative boundary point of C if and only if there exist a vector $\boldsymbol{a} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ and a number $b \in \mathbb{R}$ such that*

$$\boldsymbol{v} = \arg\max_{\boldsymbol{u}\in C} \{\langle \boldsymbol{a}, \boldsymbol{u}\rangle + b\},$$

with $\langle \boldsymbol{a}, \boldsymbol{u}\rangle + b < \langle \boldsymbol{a}, \boldsymbol{v}\rangle + b$ for every $\boldsymbol{u} \in \text{int } (C)$.

Step 1. Suppose $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) \notin \text{cl (dom } J)$. Since the set cl (dom J) is closed and convex, the projection of $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ onto cl (dom J) given by $\pi_{\text{cl(dom } J)}(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) \equiv \bar{\boldsymbol{u}}$ is well-defined

and unique (see Definition 16), with $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \neq \bar{\boldsymbol{u}}$ by assumption. The projection $\bar{\boldsymbol{u}}$ also satisfies the characterization (1.8), namely

$$\langle \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) - \bar{\boldsymbol{u}}, \boldsymbol{u} - \bar{\boldsymbol{u}} \rangle \leqslant 0$$

for every $\boldsymbol{u} \in \text{cl (dom } J)$. Then, by linearity of the posterior mean estimate,

$$\begin{aligned} \|\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) - \bar{\boldsymbol{u}}\|_2^2 &= \langle \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) - \bar{\boldsymbol{u}}, \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) - \bar{\boldsymbol{u}} \rangle \\ &= \langle \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) - \bar{\boldsymbol{u}}, \mathbb{E}_J[\boldsymbol{u}] - \bar{\boldsymbol{u}} \rangle \\ &= \mathbb{E}_J \left[\langle \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) - \bar{\boldsymbol{u}}, \boldsymbol{u} - \bar{\boldsymbol{u}} \rangle \right] \\ &\leqslant 0, \end{aligned}$$

which implies that $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \bar{\boldsymbol{u}}$. This contradicts the assumption that $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \notin \text{cl (dom } J)$. Hence, it follows that $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \in \text{cl (dom } J)$.

Step 2. We now wish to prove that $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \in \text{int (dom } J)$. Note that this inclusion trivially holds if there are no boundary points, i.e., $(\text{cl (dom } J) \setminus \text{int (dom } J)) = \varnothing$. Now we consider the case $(\text{cl (dom } J) \setminus \text{int (dom } J)) \neq \varnothing$. Suppose that $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \in (\text{cl (dom } J) \setminus \text{int (dom } J))$. Then Theorem 4.B.1 applies and there exist a vector $\boldsymbol{a} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ and a number $b \in \mathbb{R}$ such that

$$\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \underset{\boldsymbol{u} \in \text{cl (dom } J)}{\arg\max} \{\langle \boldsymbol{a}, \boldsymbol{u} \rangle + b\},$$

with $\langle \boldsymbol{a}, \boldsymbol{u} \rangle + b < \langle \boldsymbol{a}, \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \rangle + b$ for every $\boldsymbol{u} \in \text{int (dom } J)$. By linearity of the posterior mean estimate,

$$\begin{aligned} \langle \boldsymbol{a}, \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \rangle + b &= \langle \boldsymbol{a}, \mathbb{E}_J[\boldsymbol{u}] \rangle + b \\ &= \mathbb{E}_J[\langle \boldsymbol{a}, \boldsymbol{u} \rangle + b] \\ &< \mathbb{E}_J[\langle \boldsymbol{a}, \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \rangle + b] \\ &= \langle \boldsymbol{a}, \boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \rangle + b, \end{aligned}$$

where the strict inequality in the third line follows from integrating over int (dom J). This contradicts the assumption that $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \in (\text{cl (dom } J) \setminus \text{int (dom } J))$. Hence, $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) \in \text{int (dom } J)$.

Proof of (ii): First, as a consequence that $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) \in \text{int (dom } J)$, the subdifferential of J at $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ is non-empty because the subdifferential ∂J is non-empty at every point $\boldsymbol{u} \in \text{int (dom } J)$ ([214], Theorem 23.4). Hence there exists a subgradient $\boldsymbol{p} \in \partial J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon))$ such that

$$J(\boldsymbol{u}) \geqslant J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) - \langle \boldsymbol{p}, \boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle \,. \tag{4.58}$$

Take the expectation $\mathbb{E}_J[\cdot]$ on both sides of inequality (4.58) to find

$$\begin{aligned}
\mathbb{E}_J[J(\boldsymbol{u})] &\geqslant \mathbb{E}_J[J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) - \langle \boldsymbol{p}, \boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle] \\
&= J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) - \mathbb{E}_J[\langle \boldsymbol{p}, \boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle] \\
&= J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) - \langle \boldsymbol{p}, \mathbb{E}_J[\boldsymbol{u}] - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle \\
&= J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) - \langle \boldsymbol{p}, \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle \\
&= J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)).
\end{aligned} \tag{4.59}$$

This gives the lower bound of inequality (4.23).

Second, use the convex inequality $1 + z \leqslant e^z$ that holds on $\mathbb{R}$ with $z \equiv J(\boldsymbol{u})/\epsilon$ for $\boldsymbol{u} \in \text{dom } J$. This gives the inequality $1 + \frac{1}{\epsilon}J(\boldsymbol{u}) \leqslant e^{J(\boldsymbol{u})/\epsilon}$. Multiply this inequality by $e^{-J(\boldsymbol{u})/\epsilon}$ and subtract by $e^{-J(\boldsymbol{u})/\epsilon}$ on both sides to find

$$\frac{1}{\epsilon}J(\boldsymbol{u})e^{-J(\boldsymbol{u})/\epsilon} \leqslant (1 - e^{-J(\boldsymbol{u})/\epsilon}). \tag{4.60}$$

Multiply both sides by $e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2}$, divide by the partition function $Z_J(\boldsymbol{x},t,\epsilon)$ (see Equation (4.11)), integrate with respect to $\boldsymbol{u} \in \text{dom } J$, and use

$$\frac{1}{Z_J(\boldsymbol{x},t,\epsilon)} \int_{\text{dom } J} \frac{1}{\epsilon}J(\boldsymbol{u})e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} d\boldsymbol{u} = \frac{1}{\epsilon}\mathbb{E}_J[J(\boldsymbol{u})]$$

to obtain

$$\frac{1}{\epsilon}\mathbb{E}_J[J(\boldsymbol{u})] \leqslant \frac{1}{Z_J(\boldsymbol{x},t,\epsilon)} \int_{\text{dom } J} \left(e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2/\epsilon} - e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\right) d\boldsymbol{u}. \tag{4.61}$$

Now, we can bound the right hand side of (4.61) as follows

$$\begin{aligned}
&\frac{1}{Z_J(\boldsymbol{x},t,\epsilon)}\int_{\mathrm{dom}\ J}\left(e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2/\epsilon}-e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\right)d\boldsymbol{u}\\
&\qquad=\frac{1}{Z_J(\boldsymbol{x},t,\epsilon)}\int_{\mathrm{dom}\ J}e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2/\epsilon}\,d\boldsymbol{u}\\
&\qquad\quad-\frac{1}{Z_J(\boldsymbol{x},t,\epsilon)}\int_{\mathrm{dom}\ J}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\
&\qquad\leqslant\frac{1}{Z_J(\boldsymbol{x},t,\epsilon)}\int_{\mathbb{R}^n}e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2/\epsilon}\,d\boldsymbol{u}-1\\
&\qquad=\frac{(2\pi t\epsilon)^{n/2}}{Z_J(\boldsymbol{x},t,\epsilon)}-1.
\end{aligned}\tag{4.62}$$

Combining (4.61) and (4.62), we get

$$\mathbb{E}_J\left[J(\boldsymbol{u})\right]\leqslant\epsilon\left(\frac{(2\pi t\epsilon)^{n/2}}{Z_J(\boldsymbol{x},t,\epsilon)}-1\right).\tag{4.63}$$

Using the representation formula (4.13) for the solution $(\boldsymbol{x},t)\mapsto S_\epsilon$ to the viscous HJ PDE (4.14), we have that $(2\pi t\epsilon)^{n/2}/Z_J(\boldsymbol{x},t,\epsilon)=e^{S_\epsilon(\boldsymbol{x},t)/\epsilon}$. We can therefore write (4.63) as follows

$$\mathbb{E}_J\left[J(\boldsymbol{u})\right]\leqslant\epsilon\left(e^{S_\epsilon(\boldsymbol{x},t)/\epsilon}-1\right)<+\infty.$$

Combining the latter inequalities with (4.59) we obtain the desired set of inequalities (4.23).

4.C Proof of Proposition 4.3.2

Proof of (i): We will show that $\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right\|_2\right]<+\infty$ and derive formulas (4.24)–(4.28) in four steps. To describe these steps, let us first introduce some notation. Recall that J satisfies assumptions (A1)-(A3) and dom $J=\mathbb{R}^n$. Define the set

$$D_J\coloneqq\left\{\boldsymbol{u}\in\mathbb{R}^n\mid\partial J(\boldsymbol{u})=\{\nabla J(\boldsymbol{u})\}\right\}.$$

We can invoke [214, Theorem 25.5] to conclude that D_J is a dense subset of $\mathbb{R}^n$, the n-dimensional Lebesgue measure of the set $(\mathbb{R}^n\setminus D_J)$ is zero, and the function $\boldsymbol{u}\mapsto\nabla J(\boldsymbol{u})$ is continuous on D_J.

Now, let $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, $\epsilon > 0$, and $\boldsymbol{u}_0 \in \mathbb{R}^n$. Define the function $\varphi_J \colon \mathbb{R}^n \to \mathbb{R}^n$ as

$$\varphi_J(\boldsymbol{u}|\boldsymbol{x}, t) = \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}).$$

Note that for every $\boldsymbol{u} \in \mathbb{R}^n$ we have $\varphi_J(\boldsymbol{u}|\boldsymbol{x}, t) \in \partial\left(\mathbb{R}^n \ni \boldsymbol{v} \mapsto \frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|_2^2 + J(\boldsymbol{v})\right)(\boldsymbol{u})$, i.e., $\varphi_J(\boldsymbol{u}|\boldsymbol{x}, t)$ is a subgradient of the function $\boldsymbol{v} \mapsto \frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|_2^2 + J(\boldsymbol{v})$ evaluated at $\boldsymbol{v} = \boldsymbol{u}$. Let

$$C_1(\boldsymbol{x}, \boldsymbol{u}_0, t, \epsilon) = \int_{\mathbb{R}^n} \left\|\boldsymbol{u} - \boldsymbol{u}_0\right\|_2 e^{-\frac{1}{2t\epsilon}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2} \, d\boldsymbol{u}, \tag{4.64}$$

and note that by assumption (A3), the expected value $\mathbb{E}_J\left[\left\|\boldsymbol{u} - \boldsymbol{u}_0\right\|_2\right]$ is bounded as follows

$$\begin{aligned}
\mathbb{E}_J\left[\left\|\boldsymbol{u} - \boldsymbol{u}_0\right\|_2\right] &= \frac{1}{Z_J(\boldsymbol{x}, t, \epsilon)} \int_{\mathbb{R}^n} \left\|\boldsymbol{u} - \boldsymbol{u}_0\right\|_2 e^{-\left(\frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u} \\
&\leqslant \frac{1}{Z_J(\boldsymbol{x}, t, \epsilon)} \int_{\mathbb{R}^n} \left\|\boldsymbol{u} - \boldsymbol{u}_0\right\|_2 e^{-\frac{1}{2t\epsilon}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2} \, d\boldsymbol{u} \\
&= \frac{C_1(\boldsymbol{x}, \boldsymbol{u}_0, t, \epsilon)}{Z_J(\boldsymbol{x}, t, \epsilon)}.
\end{aligned} \tag{4.65}$$

Define the vector field $\boldsymbol{V} \colon \mathbb{R}^n \to \mathbb{R}^n$ as

$$\boldsymbol{V}(\boldsymbol{u}) = (\boldsymbol{u} - \boldsymbol{u}_0) e^{-\left(\frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}, \tag{4.66}$$

which is continuous on $\mathbb{R}^n$. It is also bounded on $\mathbb{R}^n$; to see this, use the triangle inequality, assumption (A3), and the fact that the function $(0, +\infty) \ni r \mapsto r e^{-\frac{1}{2t\epsilon} r^2}$ attains its maximum at $r^* = \sqrt{t\epsilon}$ to get

$$\begin{aligned}
\left\|\boldsymbol{V}(\boldsymbol{u})\right\|_2 &= \left\|(\boldsymbol{u} - \boldsymbol{u}_0)\right\|_2 e^{-\left(\frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\
&= \left\|(\boldsymbol{u} - \boldsymbol{x} + \boldsymbol{x} - \boldsymbol{u}_0)\right\|_2 e^{-\left(\frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\
&\leqslant (\left\|\boldsymbol{x} - \boldsymbol{u}_0\right\|_2 + \left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2) e^{-\left(\frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\
&\leqslant \left\|\boldsymbol{x} - \boldsymbol{u}_0\right\|_2 + \left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2 e^{-\left(\frac{1}{2t}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\
&\leqslant \left\|\boldsymbol{x} - \boldsymbol{u}_0\right\|_2 + \left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2 e^{-\frac{1}{2t\epsilon}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2} \\
&\leqslant \left\|\boldsymbol{x} - \boldsymbol{u}_0\right\|_2 + \sup_{\boldsymbol{u} \in \mathbb{R}^n} \left(\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2 e^{-\frac{1}{2t\epsilon}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2}\right) \\
&\leqslant \left\|\boldsymbol{x} - \boldsymbol{u}_0\right\|_2 + (\sqrt{t\epsilon}) e^{-\frac{\sqrt{t\epsilon}}{2}}.
\end{aligned} \tag{4.67}$$

The divergence $\nabla_{\boldsymbol{u}} \cdot \boldsymbol{V}(\boldsymbol{u})$, which is well-defined and continuous on D_J, is given for every $\boldsymbol{u} \in D_J$ by

$$\begin{aligned}
\nabla_{\boldsymbol{u}} \cdot \boldsymbol{V}(\boldsymbol{u}) &= \nabla_{\boldsymbol{u}} \cdot \left((\boldsymbol{u} - \boldsymbol{u}_0) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \right) \\
&= (\nabla_{\boldsymbol{u}} \cdot (\boldsymbol{u} - \boldsymbol{u}_0)) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} + \left\langle \nabla_{\boldsymbol{u}} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}, \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \\
&= n e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} - \left\langle \frac{1}{\epsilon} \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t} + \nabla J(\boldsymbol{u}) \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}, \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \\
&= \left(n - \left\langle \frac{1}{\epsilon} \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t} + \nabla J(\boldsymbol{u}) \right), \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}.
\end{aligned} \tag{4.68}$$

We now outline the four steps that will be used to prove Proposition 4.3.2(i). In the first step, we will show that the divergence of the vector field $\boldsymbol{V}$ on D_J integrates to zero in the sense that

$$\lim_{r \to +\infty} \left| \int_{\left\{ \boldsymbol{u} \in \mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 \leqslant r \right\} \cap D_J} \nabla_{\boldsymbol{u}} \cdot \boldsymbol{V}(\boldsymbol{u}) \, d\boldsymbol{u} \right| = 0. \tag{4.69}$$

In the second step, we will show that $\mathbb{E}_J \left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t), \boldsymbol{u} - \boldsymbol{u}_0 \rangle \right] = n\epsilon$, hereby proving formula (4.24), using the convexity of the function $\boldsymbol{u} \mapsto \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u})$, Fatou's lemma ([108], Lemma 2.18), and Equation (4.69) derived in the first step. In the third step, we will combine the results from the first and second steps to show that $\mathbb{E}_J \left[\left\| \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) \right\|_2 \right] < +\infty$ and conclude that the representation formulas (4.25) and (4.26) hold. Finally, in the fourth step we will conclude that the representation formulas (4.27) and (4.28) hold using Equations (4.25) and (4.26) and Proposition (4.2.1)(iii).

Step 1. The proof of the limit result (4.69) that we present here is based on an application of Theorem 4.14 in [198] to the vector field $\boldsymbol{V}(\cdot)$. As this result is fairly technical, we first introduce some terminology and definitions that will be used exclusively in this part of the proof of (i).

Let C be a non-empty convex subset of $\mathbb{R}^n$. The *dimension* of the set C is defined as the smallest dimension of a non-empty affine set containing C, with the dimension of a non-empty affine set being the dimension of the subspace parallel to it [214, Pages 4 and 12]. If C consists of a single point then its dimension is taken to be zero.

Let $k \in \{0, \ldots, n\}$. Denote by $\mathcal{H}^{n-k}$ the $(n-k)$-dimensional outer Hausdorff measure in $\mathbb{R}^n$ as defined in [100, Page 171, Section 2.10.2]. The measure $\mathcal{H}^{n-k}$, in particular, is a constant multiple of the $(n-k)$-dimensional Lebesgue measure for every measurable subset $B \subset \mathbb{R}^n$ (see [99], Section 1.2, p.7, and Theorem 1.12, p.13).

A subset $S \subset \mathbb{R}^n$ is called *slight* if $\mathcal{H}^{n-1}(S) = 0$, and a subset $T \subset \mathbb{R}^n$ is called *thin* if T is σ-finite for $\mathcal{H}^{n-1}$, i.e., T can be expressed as a countable union of sets $T = \cup_{k=1}^{+\infty} T_k$ with $\mathcal{H}^{n-1}(T_k) < +\infty$ for each $k \in \mathbb{N}^+$ (see, e.g., [198]).

Let $k \in \{0, \ldots, n\}$. A non-empty, measurable subset $\Omega \subset \mathbb{R}^n$ is said to be *countably $\mathcal{H}^{n-k}$-rectifiable* if it is contained, up to a null set of $(n-k)$-dimensional outer Hausdorff measure $\mathcal{H}^{n-k}$ zero, in a countable union of continuously differentiable hypersurfaces of dimension $(n-k)$ (see, e.g., [4] and references therein). A non-empty, measurable and countably $\mathcal{H}^{n-k}$-rectifiable subset of $\mathbb{R}^n$, in particular, is σ-finite for $\mathcal{H}^{n-k}$.

A subset $A \subset \mathbb{R}^n$ is called *admissible* if its boundary bd A is thin and if the distributional gradient of the characteristic function of A is a vector measure on Borel subsets of $\mathbb{R}^n$ whose variation is finite (see [198] pp.151 and the reference therein). For the purpose of our proof, we will use the fact that the family of closed balls of radius $r > 0$, namely $\{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}$, are admissible sets (see [120], Example 1.10, and note that admissible sets are also called *Caccioppoli* sets [120, Pages 5-6]).

Let A be an admissible set and let $\boldsymbol{v} \colon A \to \mathbb{R}^n$ be a vector field. In the terminology of [198], we say that $\boldsymbol{v}$ is *integrable* over the admissible set A if $\boldsymbol{v}$ satisfies definition 4.1 of [198], and in that case, the number $I(\boldsymbol{v}, A)$ is called the integral of $\boldsymbol{v}$ over A. Note, here, that the notion of integrability considered in [198] is different from that of the usual Lebesgue integrability. Nevertheless, if $\boldsymbol{v}$ is integrable in the sense of [198], $\boldsymbol{v}$ is also Lebesgue measurable [198, Corollary 4.9], and if the Lebesgue integral $\int_A |\boldsymbol{v}(\boldsymbol{u})|\, d\boldsymbol{u}$ is finite, then $I(\boldsymbol{v}, A) = \int_A |\boldsymbol{v}(\boldsymbol{u})|\, d\boldsymbol{u}$ [198, Proposition 4.7].

Let E be an non-empty subset of $\mathbb{R}^n$, let $\boldsymbol{v}\colon E \to \mathbb{R}^n$ be a vector field, and let $D_{\boldsymbol{v}}$ denote the set of points at which $\boldsymbol{v}$ is differentiable in int E (for the definition of differentiability of vector fields, see [217, Page 150, Definition 7.22]). In the terminology of [198], we call a *divergence of* $\boldsymbol{v}$ any function $g\colon E \mapsto \mathbb{R}$ such that $g(\boldsymbol{u}) = \nabla \cdot \boldsymbol{v}(\boldsymbol{u})$ for each $\boldsymbol{u} \in (\text{int } E) \cap D_{\boldsymbol{v}}$.

In addition to these definitions, we will need the following two results due to, respectively, [4] and [198].

Theorem 4.C.1. *[[4], Theorem 4.1 (for convex functions)] Let Ω be a bounded, open, convex subset of $\mathbb{R}^n$, and let $f\colon \Omega \to \mathbb{R}$ be a convex and Lipschitz continuous function. Denote the subdifferential of f at $\boldsymbol{u} \in \Omega$ by $\partial f(\boldsymbol{u})$. Then, for each $k \in \{0, \dots, n\}$, the set*

$$\{\boldsymbol{u} \in \Omega \mid dim\,(\partial f(\boldsymbol{u})) \geqslant k\}$$

is countably $\mathcal{H}^{n-k}$-rectifiable.

Theorem 4.C.2. *[[198], Theorem 4.14] Let A be an admissible set, and let S and T be, respectively, a slight and thin subset of* cl A. *Let $\boldsymbol{v}$ be a bounded vector field in* cl A *that is continuous in* (cl A)$\setminus S$ *and differentiable in* (int A) $\setminus\, T$. *Then every divergence of $\boldsymbol{v}$ is integrable in A. Moreover, there exists a vector field* bd $A \ni \boldsymbol{u} \to n_A(\boldsymbol{u})$ *with* $\|n_A(\boldsymbol{u})\|_2 = 1$ *for every* $\boldsymbol{u} \in$ bd A *such that if div $\boldsymbol{v}$ denotes any divergence $\boldsymbol{v}$, then*

$$I(div\ \boldsymbol{v}, A) = \int_{\text{bd } A} \langle \boldsymbol{v}(\boldsymbol{u}), \boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u}) \rangle \, d\mathcal{H}^{n-1}\, d\boldsymbol{u}. \tag{4.70}$$

Step 1 (Continued). Fix $r > 0$ and let $A = \{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}$ denote the closed ball of radius r centered at the origin in $\mathbb{R}^n$. Note that A is bounded, convex, closed, and admissible. Consider now the restriction of the convex function J to int A. As int A is bounded, open and convex, the function J is Lipschitz continuous on int A [214, Theorem 10.4]. All conditions in Theorem (4.C.1) are satisfied (with $\Omega =$ int A and $f = J$), and we can invoke the theorem to conclude that the set

$$T = \{\boldsymbol{u} \in \text{int } A \mid \dim\,(\partial J(\boldsymbol{u})) \geqslant 1\}$$

is countably $\mathcal{H}^{n-1}$-rectifiable, and therefore σ-finite for $\mathcal{H}^{n-1}$. In particular, the set T is thin. Moreover, recalling the definition of the set $D_J \coloneqq \{\boldsymbol{u} \in \mathbb{R}^n \mid \partial J(\boldsymbol{u}) = \{\nabla J(\boldsymbol{u})\}\}$, we find that the set $(\operatorname{int} A) \setminus T$ comprises the points $\boldsymbol{u} \in \operatorname{int} A$ at which the subdifferential $\partial J(\boldsymbol{u})$ is a singleton, i.e., $T = (\operatorname{int} A) \cap (\mathbb{R}^n \setminus D_J)$.

Now consider the vector field $\boldsymbol{V}$ defined by (4.66). This vector field is continuous in $\mathbb{R}^n$ by convexity of J and $\operatorname{dom} J = \mathbb{R}^n$. It is also bounded by (4.67). Now define the function $g\colon A \to \mathbb{R}$ via

$$g(\boldsymbol{u}) = \begin{cases} \nabla \cdot \boldsymbol{V}(\boldsymbol{u}) \text{ if } \boldsymbol{u} \in A \cap D_J, \\ 0, \quad \text{if } \boldsymbol{u} \in A \cap (\mathbb{R}^n \setminus D_J). \end{cases} \tag{4.71}$$

The function g constitutes a divergence of the vector field $\boldsymbol{V}$ because it coincides with the divergence $\nabla \cdot \boldsymbol{V}(\boldsymbol{u})$ at every $\boldsymbol{u} \in (\operatorname{int} A) \cap D_J$. Moreover, its Lebesgue integral over A is finite; to see this, first note that for every $\boldsymbol{u} \in A \cap D_J$ the absolute value of $g(\boldsymbol{u})$ can be bounded using (4.68), the triangle inequality, the Cauchy–Schwarz inequality, and assumption (A3) as follows

$$\begin{aligned} |g(\boldsymbol{u})| = |\nabla \cdot \boldsymbol{V}(\boldsymbol{u})| &= \left| \left(n - \left\langle \frac{1}{\epsilon} \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t} + \nabla J(\boldsymbol{u}) \right), \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \right| \\ &\leqslant \left(n + \left| \left\langle \frac{1}{\epsilon} \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t} + \nabla J(\boldsymbol{u}) \right), \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \right| \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\ &\leqslant \left(n + \frac{1}{\epsilon} \left\| \frac{\boldsymbol{u}-\boldsymbol{x}}{t} + \nabla J(\boldsymbol{u}) \right\|_2 \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\ &\leqslant \left(n + \frac{1}{\epsilon} \left(\left\| \frac{\boldsymbol{u}-\boldsymbol{x}}{t} \right\|_2 + \|\nabla J(\boldsymbol{u})\|_2 \right) \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \\ &\leqslant \left(n + \frac{1}{\epsilon} \left(\left\| \frac{\boldsymbol{u}-\boldsymbol{x}}{t} \right\|_2 + \|\nabla J(\boldsymbol{u})\|_2 \right) \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \right) e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2} \end{aligned} \tag{4.72}$$

Second, as the set A is a closed bounded subset of $\operatorname{dom} J = \mathbb{R}^n$ the function J is Lipschitz continuous relative to A, and therefore there exists a number $L_A > 0$ such that $\|\nabla J(\boldsymbol{u})\|_2 \leqslant L_A$ for every $\boldsymbol{u} \in A \cap D_J$. As a consequence, we can further bound $g(\boldsymbol{u})$ for every $\boldsymbol{u} \in A \cap D_J$ in (4.72) as

$$|g(\boldsymbol{u})| \leqslant \left(n + \frac{1}{\epsilon} \left(\left\| \frac{\boldsymbol{u}-\boldsymbol{x}}{t} \right\|_2 + L_A \right) \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \right) e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2}. \tag{4.73}$$

In particular, using the definition of g given by (4.71), we have that (4.73) holds for every $\boldsymbol{u} \in A$.

We can now use (4.71) and (4.73) to get

$$\begin{aligned}\int_A |g(\boldsymbol{u})|\,d\boldsymbol{u} &= \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}} |g(\boldsymbol{u})|\,d\boldsymbol{u} \\ &= \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}\cap D_J} |\nabla\cdot\boldsymbol{V}(\boldsymbol{u})|\,d\boldsymbol{u} \\ &\leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}\cap D_J} \left(n+\frac{1}{\epsilon}\left(\left\|\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right\|_2 + L_A\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2\right) e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2}\,d\boldsymbol{u} \\ &\leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}} \left(n+\frac{1}{\epsilon}\left(\left\|\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right\|_2 + L_A\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2\right) e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2}\,d\boldsymbol{u}.\end{aligned} \tag{4.74}$$

Since the function

$$\boldsymbol{u}\mapsto \left(n+\frac{1}{\epsilon}\left(\left\|\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right\|_2 + L_A\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2\right) e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2}$$

is continuous, it is bounded on the compact set $A=\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2\leqslant r\}$. Its integral over A is therefore finite, and using (4.74) we find that $\int_A |g(\boldsymbol{u})|\,d\boldsymbol{u}$ is finite as well.

The previous considerations show that all conditions in Theorem (4.C.2) are satisfied (with $A=\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2\leqslant r\}$, $\boldsymbol{v}=\boldsymbol{V}$, $S=\varnothing$, $T=\{\boldsymbol{u}\in \operatorname{int} A \mid \dim(\partial J(\boldsymbol{u}))\geqslant 1\}=(\operatorname{int} A)\cap(\mathbb{R}^n\setminus D_J)$). We can therefore invoke [198, Theorem 4.14] to conclude that the divergence of g is integrable (in the sense described by [198]), with integral $I(g,A)$, and that there exists a vector field $\operatorname{bd} A\ni\boldsymbol{u}\to\boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})$ with $\|\boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\|_2=1$ for every $\boldsymbol{u}\in\operatorname{bd} A$ such that

$$I(g,A)=\int_{\operatorname{bd}\ A}\langle\boldsymbol{V}(\boldsymbol{u}),\boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle\ d\mathcal{H}^{n-1}\,d\boldsymbol{u}. \tag{4.75}$$

Since the Lebesgue integral of $|g|$ over A is finite, we also have [198, Proposition 4.7]

$$I(g,A)=\int_A g(\boldsymbol{u})\,d\boldsymbol{u}. \tag{4.76}$$

Using that $A=\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2\leqslant r\}$, Equations (4.71), (4.75), and (4.76), we obtain

$$\int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2\leqslant r\}} g(\boldsymbol{u})\,d\boldsymbol{u}=\int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2\leqslant r\}}\nabla\cdot\boldsymbol{V}(\boldsymbol{u})\,d\boldsymbol{u}=\int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2= r\}}\langle\boldsymbol{V}(\boldsymbol{u}),\boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle\ d\mathcal{H}^{n-1}. \tag{4.77}$$

As r was an arbitrary positive number, we can take the absolute value and then the limit $r \to +\infty$ on both sides of (4.77) to find

$$\lim_{r\to+\infty} \left| \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}\cap D_J} \nabla\cdot \boldsymbol{V}(\boldsymbol{u})\, d\boldsymbol{u} \right| = \lim_{r\to+\infty} \left| \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} \langle \boldsymbol{V}(\boldsymbol{u}), \boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle \; d\mathcal{H}^{n-1} \right|. \tag{4.78}$$

We will now show that the limit on the right side of (4.78) is equal to zero. To show this, first take the absolute value inside the integral on the right side of (4.78) to find

$$\left| \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} \langle \boldsymbol{V}(\boldsymbol{u}), \boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle \; d\mathcal{H}^{n-1} \right| \leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} |\langle \boldsymbol{V}(\boldsymbol{u}), \boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle| \; d\mathcal{H}^{n-1}. \tag{4.79}$$

Use the Cauchy–Schwarz inequality, Equation (4.66), assumption (A3) ($\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) = 0$) and $\|\boldsymbol{n}_{\boldsymbol{v}}\|_2 = 1$ to further bound the right side of (4.79) as follows

$$\begin{aligned}
\int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} |\langle \boldsymbol{V}(\boldsymbol{u}), \boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle| \; d\mathcal{H}^{n-1} &\leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} \|\boldsymbol{V}(\boldsymbol{u})\|_2 \, \|\boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\|_2 \; d\mathcal{H}^{n-1} \\
&\leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\mathcal{H}^{n-1} \\
&\leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} (\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}_0\|_2) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon} \, d\mathcal{H}^{n-1}.
\end{aligned} \tag{4.80}$$

Use the parallelogram law $2(\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{v}\|_2^2) = \|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \|\boldsymbol{x}+\boldsymbol{v}\|_2^2$ with $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{u}$ to bound the exponential $e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2/\epsilon}$ by

$$\begin{aligned}
e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2/\epsilon} &= e^{-\frac{1}{2t}\left(\frac{1}{2}(\|\boldsymbol{u}\|_2^2 + \|2\boldsymbol{x}-\boldsymbol{u}\|_2^2) - \|\boldsymbol{x}\|_2^2\right)/\epsilon} \\
&\leqslant e^{-\frac{1}{2t}\left(\frac{1}{2}\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{x}\|_2^2\right)/\epsilon}
\end{aligned} \tag{4.81}$$

and use it in (4.80) to get

$$\int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} |\langle \boldsymbol{V}(\boldsymbol{u}), \boldsymbol{n}_{\boldsymbol{v}}(\boldsymbol{u})\rangle| \; d\mathcal{H}^{n-1} \leqslant \int_{\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}\|_2 = r\}} (\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}\left(\frac{1}{2}\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{x}\|_2^2\right)/\epsilon} \, d\mathcal{H}^{n-1}. \tag{4.82}$$

Since the domain of integration in (4.82) is over the surface of an n-dimensional sphere of radius

$\|\boldsymbol{u}\|_2 = r$, the integral on the right side of (4.82) is given by

$$\begin{aligned}\int_{\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 = r\}} &(\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}(\frac{1}{2}\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{x}\|_2^2)/\epsilon}\, d\mathcal{H}^{n-1} \\ &= \int_{\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 = r\}} (r + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}(\frac{1}{2}r^2 - \|\boldsymbol{x}\|_2^2)/\epsilon}\, d\mathcal{H}^{n-1} \\ &= (r + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}(\frac{1}{2}r^2 - \|\boldsymbol{x}\|_2^2)/\epsilon} \int_{\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 = r\}} d\mathcal{H}^{n-1} \\ &= (r + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}\left(\frac{1}{2}r^2 - \|\boldsymbol{x}\|_2^2\right)/\epsilon} \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)},\end{aligned} \tag{4.83}$$

where $n\pi^{n/2}/\Gamma\left(\frac{n}{2}+1\right)$ is the area of an n-dimensional sphere of radius one, with $\Gamma\left(\frac{n}{2}+1\right)$ denoting the Gamma function evaluated at $\frac{n}{2}+1$. Since

$$\lim_{r\to+\infty} (r + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}\left(\frac{1}{2}r^2 - \|\boldsymbol{x}\|_2^2\right)/\epsilon} = 0,$$

the limit $r \to +\infty$ in (4.83) is equal to zero, i.e.,

$$\lim_{r\to+\infty} \int_{\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 = r\}} (\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}_0\|_2) e^{-\frac{1}{2t}(\frac{1}{2}\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{x}\|_2^2)/\epsilon}\, d\mathcal{H}^{n-1} = 0. \tag{4.84}$$

Combining (4.78), (4.79), (4.82) and (4.84) yield

$$\lim_{r\to+\infty} \left| \int_{\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 \leqslant r\} \cap D_J} \nabla_{\boldsymbol{u}} \cdot \boldsymbol{V}(\boldsymbol{u})\, d\boldsymbol{u} \right| = 0.$$

which proves the limit result (4.69).

Step 2. Recall that the divergence of the vector field $\boldsymbol{u} \mapsto \boldsymbol{V}(\boldsymbol{u})$ on D_J is given by (4.68). Combine (4.69) and (4.68) to conclude that

$$\lim_{r\to+\infty} \left| \int_{\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 \leqslant r\} \cap D_J} \left(n\epsilon - \left\langle \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t} + \nabla J(\boldsymbol{u}) \right), \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \right| = 0. \tag{4.85}$$

Note that the minimal subgradient $\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) = \nabla J(\boldsymbol{u})$ for every $\boldsymbol{u} \in D_J$. We can therefore substitute the minimal subgradient $\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})$ for the gradient $\nabla J(\boldsymbol{u})$ inside the integral in the limit (4.85) without changing its value. Moreover, since the set D_J is dense in $\mathbb{R}^n$ and the n-

dimensional Lebesgue measure of $(\mathbb{R}^n \setminus D_J)$ is zero, we can further substitute the domain of integration $\{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\} \cap D_J$ of the integral in the limit (4.85) with $\{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u}\|_2 \leqslant r\}$ without changing its value. With these two changes, the limit (4.85) can be written as

$$\lim_{r\to+\infty} \left| \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n \,\middle|\, \|\boldsymbol{u}\|_2 \leqslant r\right\}} \left(n\epsilon - \left\langle \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t} + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) \right), \boldsymbol{u}-\boldsymbol{u}_0 \right\rangle \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u} \right| = 0.$$

Using the notation $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t) = \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})$, we can write this limit more succinctly as

$$\lim_{r\to+\infty} \left| \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n \,\middle|\, \|\boldsymbol{u}\|_2 \leqslant r\right\}} \left(n\epsilon - \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0 \rangle \right) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u} \right| = 0. \tag{4.86}$$

Now, consider the function $\mathbb{R}^n \ni \boldsymbol{u} \mapsto \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0 \rangle$. Note here that as J is convex with dom $J = \mathbb{R}^n$, both $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)$ and $\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)$ are subgradients of the convex function $\boldsymbol{v} \mapsto \frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + J(\boldsymbol{v})$ at $\boldsymbol{v} = \boldsymbol{u}$ and $\boldsymbol{v} = \boldsymbol{u}_0$, respectively ([214], Theorem 23.4). We can therefore apply inequality (1.20) (with $p = \varphi_J(\boldsymbol{u}|\boldsymbol{x},t)$, $p_0 = \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)$, and $m = 0$) to find $\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0 \rangle \geqslant 0$. Define $F\colon \mathbb{R}^n \to \mathbb{R}$ and $G\colon \mathbb{R}^n \to \mathbb{R}$ as follows:

$$F(\boldsymbol{u}) = \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0 \rangle \, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}$$

and

$$G(\boldsymbol{u}) = \langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0 \rangle \, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}.$$

Note that $F(\boldsymbol{u}) - G(\boldsymbol{u}) = \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0 \rangle \, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \geqslant 0$ for every $\boldsymbol{u} \in \mathbb{R}^n$. Integrate $\boldsymbol{u} \mapsto F(\boldsymbol{u}) - G(\boldsymbol{u})$ over $\mathbb{R}^n$ and use Fatou's lemma to find

$$\begin{aligned} 0 \leqslant \int_{\mathbb{R}^n} F(\boldsymbol{u}) - G(\boldsymbol{u}) \, d\boldsymbol{u} &\leqslant \lim_{r\to+\infty} \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n \,\middle|\, \|\boldsymbol{u}\|_2 \leqslant r\right\}} F(\boldsymbol{u}) - G(\boldsymbol{u}) \, d\boldsymbol{u} \\ &= \lim_{r\to+\infty} \left(\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n \,\middle|\, \|\boldsymbol{u}\|_2 \leqslant r\right\}} F(\boldsymbol{u}) \, d\boldsymbol{u} + \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n \,\middle|\, \|\boldsymbol{u}\|_2 \leqslant r\right\}} (-G(\boldsymbol{u})) \, d\boldsymbol{u} \right) \end{aligned} \tag{4.87}$$

Use the Cauchy–Schwarz inequality assumption (A3) ($\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) = 0$) to bound the second

integral on the right hand side of (4.87) as follows

$$\begin{aligned}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(-G(\boldsymbol{u}))\,d\boldsymbol{u} &= \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}} -\langle\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &\leqslant \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}} \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\,\|\boldsymbol{u}-\boldsymbol{u}_0\|_2\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &\leqslant \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\int_{\mathbb{R}^n}\|\boldsymbol{u}-\boldsymbol{u}_0\|_2\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon}\,d\boldsymbol{u}\\ &= \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon),\end{aligned} \tag{4.88}$$

where $C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon)$ was defined in (4.64). Combine (4.87) and (4.88) to find

$$\begin{aligned}0\leqslant\int_{\mathbb{R}^n}F(\boldsymbol{u})-G(\boldsymbol{u})\,d\boldsymbol{u} &\leqslant \lim_{r\to+\infty}\left(\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}F(\boldsymbol{u})\,d\boldsymbol{u}+\|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\,C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon)\right)\\ &= \left(\lim_{r\to+\infty}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}F(\boldsymbol{u})\,d\boldsymbol{u}\right)+\|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\,C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon).\end{aligned} \tag{4.89}$$

The integral on the right hand side of (4.89) can be bounded using assumption (A3) as follows

$$\begin{aligned}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}F(\boldsymbol{u})\,d\boldsymbol{u} &= \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &= \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle+(n\epsilon-n\epsilon))e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &= \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle-n\epsilon)e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &\quad+n\epsilon\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &\leqslant \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle-n\epsilon)e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &\quad+n\epsilon\int_{\mathbb{R}^n}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon}\,d\boldsymbol{u}\\ &= \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle-n\epsilon)e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\ &\quad+n\epsilon(2\pi t\epsilon)^{n/2}.\end{aligned} \tag{4.90}$$

Combine (4.89) and (4.90) to get

$$\begin{aligned}0 \leqslant \int_{\mathbb{R}^n} F(\boldsymbol{u}) - G(\boldsymbol{u})\, d\boldsymbol{u} \leqslant \lim_{r\to+\infty} \int_{\left\{\boldsymbol{u}\in\mathbb{R}^n \,|\, \|\boldsymbol{u}\|_2 \leqslant r\right\}} (\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle - n\epsilon) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \\ + \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon) + n\epsilon(2\pi t\epsilon)^{n/2}.\end{aligned} \tag{4.91}$$

Combine (4.86) and (4.91) to get

$$\begin{aligned}0 \leqslant \int_{\mathbb{R}^n} F(\boldsymbol{u}) - G(\boldsymbol{u})\, d\boldsymbol{u} &= \int_{\mathbb{R}^n} \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \\ &\leqslant \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon) + n\epsilon(2\pi t\epsilon)^{n/2}.\end{aligned} \tag{4.92}$$

Divide (4.92) by the partition function $Z_J(\boldsymbol{x},t,\epsilon)$ (see Equation (4.11)) to get

$$0 \leqslant \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle\right] \leqslant \frac{\|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon) + n\epsilon(2\pi t\epsilon)^{n/2}}{Z_J(\boldsymbol{x},t,\epsilon)} < +\infty. \tag{4.93}$$

Now, using the Cauchy–Schwarz inequality and (4.65), we can bound $\mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right]$ as follows

$$\begin{aligned}\mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right] &= \frac{1}{Z_J(\boldsymbol{x},t,\epsilon)} \int_{\mathbb{R}^n} |\langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}\, d\boldsymbol{u} \\ &\leqslant \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, \frac{1}{Z_J(\boldsymbol{x},t,\epsilon)} \int_{\mathbb{R}^n} \|\boldsymbol{u}-\boldsymbol{u}_0\|_2\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2\right)/\epsilon}\, d\boldsymbol{u} \\ &= \frac{\|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon)}{Z_J(\boldsymbol{x},t,\epsilon)}.\end{aligned} \tag{4.94}$$

Use the triangle inequality and the upper bounds in (4.93) and (4.94) to obtain

$$\begin{aligned}\mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right] &= \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - (\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right] \\ &\leqslant \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle| + |\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right] \\ &= \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right] + \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{u}_0\rangle|\right] \\ &\leqslant \frac{2\,\|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2\, C_1(\boldsymbol{x},\boldsymbol{u}_0,t,\epsilon)}{Z_J(\boldsymbol{x},t,\epsilon)} + \frac{n\epsilon(2\pi t\epsilon)^{n/2}}{Z_J(\boldsymbol{x},t,\epsilon)} \\ &< +\infty.\end{aligned} \tag{4.95}$$

Since $\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle|\right]<+\infty$, we can use (4.86) to conclude that

$$\begin{aligned}
&\int_{\mathbb{R}^n}\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\
&\quad=\lim_{r\to+\infty}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n\,|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\, e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\
&\quad=\lim_{r\to+\infty}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n\,|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(n\epsilon-n\epsilon+\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle)e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\\
&\quad=n\epsilon Z_J(\boldsymbol{x},t,\epsilon)\\
&\qquad-\lim_{r\to+\infty}\left(\int_{\left\{\boldsymbol{u}\in\mathbb{R}^n\,|\,\|\boldsymbol{u}\|_2\leqslant r\right\}}(n\epsilon-\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle)e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}\right)\\
&\quad=n\epsilon.
\end{aligned}\tag{4.96}$$

Inequality (4.95) and equality (4.96) show the desired results $\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle|\right]<+\infty$ and $\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\right]=n\epsilon$, which, after recalling the definition $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)=\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)+\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})$, also proves formula (4.24).

Step 3. Thanks to Step 2, we have the inequalities

$$\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle|\right]<+\infty$$

and

$$\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\right]=n\epsilon$$

for every $\boldsymbol{u}_0\in\mathbb{R}^n$. In particular, the choice of $\boldsymbol{u}_0=\boldsymbol{0}$ yields $\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}\rangle|\right]<+\infty$ and $\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}\rangle\right]=n\epsilon$. As a consequence, we have that

$$\begin{aligned}
\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}_0\rangle|\right]&=\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}_0+(\boldsymbol{u}-\boldsymbol{u})\rangle|\right]\\
&\leqslant\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle|+|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}\rangle|\right]\\
&=\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle|\right]+\mathbb{E}_J\left[|\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}\rangle|\right]\\
&<\infty,
\end{aligned}\tag{4.97}$$

and

$$\begin{aligned}\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}_0\rangle\right] &= \mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),(\boldsymbol{u}-\boldsymbol{u})+\boldsymbol{u}_0\rangle\right]\\ &= \mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}\rangle\right]-\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\right]\\ &= n\epsilon - n\epsilon\\ &= 0,\end{aligned} \tag{4.98}$$

for every $\boldsymbol{u}_0 \in \mathbb{R}^n$. Now let $\{\boldsymbol{e}_i\}_{i=1}^n$ denote the standard basis in $\mathbb{R}^n$ and let $\{\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i\}_{i=1}^n$ denote the components of the vector $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)$, i.e., $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t) = (\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_1,\ldots,\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_n)$. Using (4.97) with the choice of $\boldsymbol{u}_0 = \boldsymbol{e}_i$ for $i \in \{1,\ldots,n\}$, we get $\mathbb{E}_J\left[|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|\right] < +\infty$ for every $i \in \{1,\ldots,n\}$. Using the norm inequality $\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2 \leqslant \sum_{i=1}^n |\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|$, we can bound $\mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right]$ as follows

$$\begin{aligned}\mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right] &\leqslant \mathbb{E}_J\left[\sum_{i=1}^n |\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|\right]\\ &= \sum_{i=1}^n \mathbb{E}_J\left[|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|\right]\\ &< +\infty.\end{aligned} \tag{4.99}$$

We can therefore combine (4.98) and (4.99) to get $\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}_0\rangle\right] = \langle\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\right],\boldsymbol{u}_0\rangle = 0$ for every $\boldsymbol{u}_0 \in \mathbb{R}^n$, which yields the following equality:

$$\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\right] = 0. \tag{4.100}$$

Moreover, recalling the definition $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t) = \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})$ and using (4.65) (with $\boldsymbol{u}_0 = \boldsymbol{x}$) and (4.99), we can bound $\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right\|_2\right]$ as follows

$$\begin{aligned}\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right\|_2\right] &= \mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) + \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right) - \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right\|_2\right]\\ &\leqslant \mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}) + \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right\|_2 + \left\|\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right\|_2\right]\\ &= \mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right] + \frac{1}{t}\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{x}\|_2\right]\\ &\leqslant \mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right] + \frac{C_1(\boldsymbol{x},\boldsymbol{x},t,\epsilon)}{tZ_J(\boldsymbol{x},t,\epsilon)}\\ &< +\infty.\end{aligned} \tag{4.101}$$

We can now combine (4.65), (4.100) and (4.101) to expand the expected value of $\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\right]$ as follows

$$\begin{aligned}\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\right] = \mathbb{E}_J\left[\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)+\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right] &= \mathbb{E}_J\left[\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right]+\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right] \\ &= \left(\frac{\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{x}}{t}\right)+\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right] \\ &= \mathbf{0}.\end{aligned} \tag{4.102}$$

Solving for $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ in (4.102) yields $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) = \boldsymbol{x} - t\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right]$, which gives the representation formula (4.25).

We now derive the second representation formula (4.26). Let $\boldsymbol{u}_0 = \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ in Equation (4.96) and use the representation formula (4.25) to find

$$\begin{aligned}\mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\mathbf{0}),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] &= \mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})+\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)-\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &= \mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})+\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &\quad -\mathbb{E}_J\left[\left\langle\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &= \mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})+\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &\quad -\mathbb{E}_J\left[\left\langle\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &= \mathbb{E}_J\left[\left\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &\quad -\mathbb{E}_J\left[\left\langle\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] \\ &= n\epsilon-\mathbb{E}_J\left[\left\langle\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right].\end{aligned} \tag{4.103}$$

We will use (4.103) to derive a representation formula for $\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right]$. Multiply (4.103) by t and rearrange to get

$$\mathbb{E}_J\left[\left\langle\boldsymbol{u}-\boldsymbol{x},\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right] = nt\epsilon - t\mathbb{E}_J\left[\left\langle\pi_{\partial J(\boldsymbol{u})}(\mathbf{0}),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right]. \tag{4.104}$$

The left hand side of (4.104) can be expressed as

$$\begin{aligned}\mathbb{E}_J\left[\langle \boldsymbol{u}-\boldsymbol{x}, \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle\right] &= \mathbb{E}_J\left[\langle \boldsymbol{u}-\boldsymbol{x}+(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)), \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle\right] \\ &= \mathbb{E}_J\left[\langle \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon), \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle\right] \\ &\quad + \mathbb{E}_J\left[\langle \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{x}, \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle\right] \\ &= \mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right].\end{aligned} \tag{4.105}$$

Combine Equations (4.104) and (4.105) to get

$$\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right] = nt\epsilon - t\mathbb{E}_J\left[\left\langle \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}), \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right],$$

which gives the representation formula (4.26).

Step 4. Thanks to Step 3, the representation formulas (4.25) and (4.26) hold. Recall that by Proposition 4.2.1(iii), the gradient $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)$ and Laplacian $\nabla^2_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)$ of the solution S_ϵ to the viscous HJ PDE (4.14) satisfy the representation formulas

$$\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) \tag{4.106}$$

and

$$\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2^2\right] = nt\epsilon - t^2\epsilon\nabla^2_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t). \tag{4.107}$$

Use (4.25) and (4.106) to get

$$\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) = \mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right],$$

which is the representation formula (4.27). Use (4.26) and (4.107) to get

$$t^2\epsilon\nabla^2_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) = t\mathbb{E}_J\left[\left\langle \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}), \boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle\right]$$

which is, after dividing by $t\epsilon$ on both sides, the representation formulas (4.28). This concludes Step 4.

Proof of (ii): Here we only assume that J satisfies assumptions (A1)-(A3); we do not assume that dom $J = \mathbb{R}^n$. Let $\{\mu_k\}_{k=1}^{+\infty}$ be a sequence of positive real numbers converging to zero. Define $f_k \colon \mathbb{R}^n \times (0, +\infty) \times (0, +\infty) \to \mathbb{R}$ by

$$f_\epsilon(\boldsymbol{x}, t, k) = -\epsilon \log \left(\frac{1}{(2\pi t \epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u}, \mu_k)\right)/\epsilon} \, d\boldsymbol{u} \right) \tag{4.108}$$

and let $S_0(\boldsymbol{x}, \mu_k)$ denote the solution to the first-order HJ PDE (1.22) with initial data J evaluated at $(\boldsymbol{x}, \mu_k)$, that is,

$$S_0(\boldsymbol{x}, \mu_k) = \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2\mu_k} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\}. \tag{4.109}$$

By Proposition 1.2.14(i), the function $\mathbb{R}^n \ni \boldsymbol{x} \mapsto S_0(\boldsymbol{x}, \mu_k)$ is continuously differentiable and convex for each $k \in \mathbb{N}$, and the sequence of real numbers $\{S_0(\boldsymbol{x}, \mu_k)\}_{k=1}^{+\infty}$ converges to $J(\boldsymbol{x})$ for every $\boldsymbol{x} \in$ dom J. Moreover, by assumption (A3) ($\inf_{\boldsymbol{u} \in \mathbb{R}^n} J(\boldsymbol{u}) = 0$) the sequence $\{S_0(\boldsymbol{x}, \mu_k)\}_{k=1}^{+\infty}$ is uniformly bounded from below by 0, that is,

$$\begin{aligned} S_0(\boldsymbol{x}, \mu_k) &= \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2\mu_k} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}) \right\} \\ &\geqslant \inf_{\boldsymbol{u} \in \mathbb{R}^n} \left\{ \frac{1}{2\mu_k} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \right\} + \inf_{\boldsymbol{u} \in \mathbb{R}^n} J(\boldsymbol{u}) \\ &= 0. \end{aligned}$$

As a consequence, we can invoke Proposition 4.2.1(i) to conclude that for each $k \in \mathbb{N}$, the function $(\boldsymbol{x}, t) \mapsto f_\epsilon(\boldsymbol{x}, t, k)$ corresponds to the solution to the viscous HJ PDE (4.14) with initial data $f_k(\boldsymbol{x}, 0, \epsilon) = S_0(\boldsymbol{x}, \mu_k)$. Moreover, $\mathbb{R}^n \ni \boldsymbol{x} \mapsto f_\epsilon(\boldsymbol{x}, t, k)$ is continuously differentiable and convex by Proposition 4.2.1(i) and (ii)(a). Finally, as the domain of the function $\boldsymbol{x} \mapsto S_0(\boldsymbol{x}, \mu_k)$ is $\mathbb{R}^n$, we can use the representation formula (4.27) in Proposition 4.3.2(i) (which was proven previously in this Appendix) to express the gradient $\nabla_{\boldsymbol{x}} f_k(\boldsymbol{x}, t, \epsilon)$ as follows

$$\nabla_{\boldsymbol{x}} f_\epsilon(\boldsymbol{x}, t, k) = \frac{\int_{\mathbb{R}^n} \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u}, \mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u}, \mu_k)\right)/\epsilon} \, d\boldsymbol{u}}. \tag{4.110}$$

Now, since $S_0(\boldsymbol{x}, \mu_k) \geqslant 0$ for every $k \in \mathbb{N}$, we can bound the integrand in (4.108) as follows

$$\frac{1}{(2\pi t\epsilon)^{n/2}} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \leqslant \frac{1}{(2\pi t\epsilon)^{n/2}} e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2}, \tag{4.111}$$

where $\int_{\mathbb{R}^n} \frac{1}{(2\pi t\epsilon)^{n/2}} e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2} \, d\boldsymbol{u} = 1$. We can therefore invoke the Lebesgue dominated convergence theorem ([108], Theorem 2.24) and use (4.108) and the limit $\lim_{k\to+\infty} e^{-S_0(\boldsymbol{x},\mu_k)/\epsilon} = e^{-J(\boldsymbol{x})/\epsilon}$ (with $\lim_{k\to+\infty} e^{-S_0(\boldsymbol{x},\mu_k)/\epsilon} = 0$ for every $\boldsymbol{x} \notin \operatorname{dom} J$) to find

$$\begin{aligned}
\lim_{k\to+\infty} f_\epsilon(\boldsymbol{x}, t, k) &= \lim_{k\to+\infty} -\epsilon \log\left(\frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}\right) \\
&= -\epsilon \log\left(\frac{1}{(2\pi t\epsilon)^{n/2}} \int_{\operatorname{dom} J} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} \, d\boldsymbol{u}\right) \\
&= S_\epsilon(\boldsymbol{x}, t),
\end{aligned} \tag{4.112}$$

which gives the limit (4.29). By continuous differentiability and convexity of $\mathbb{R}^n \ni \boldsymbol{x} \mapsto f_\epsilon(\boldsymbol{x}, t, k)$ and $\mathbb{R}^n \ni \boldsymbol{x} \mapsto S_\epsilon(\boldsymbol{x}, t)$ and the limit (4.112), we can invoke [214, Theorem 25.7] to conclude that the gradient $\nabla_{\boldsymbol{x}} f_k(\boldsymbol{x}, t, \mu_k)$ converges to the gradient $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$ as $k \to +\infty$. Hence we can take the limit $k \to +\infty$ in (4.110) to find

$$\begin{aligned}
\lim_{k\to+\infty} \nabla_{\boldsymbol{x}} f_\epsilon(\boldsymbol{x}, t, k) &= \lim_{k\to+\infty} \left(\frac{\int_{\mathbb{R}^n} \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}\right) \\
&= \nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t),
\end{aligned} \tag{4.113}$$

which gives the limit (4.30). Finally, using the definition of the posterior mean estimate (4.3), the limit (4.113), and the representation formula (4.15) derived in Proposition 4.2.1(iii), namely $\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) = \boldsymbol{x} - t\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x}, t)$, we find the two limits

$$\begin{aligned}
\boldsymbol{u}_{PM}(\boldsymbol{x}, t, \epsilon) &= \lim_{k\to+\infty} \left(\frac{\int_{\mathbb{R}^n} \boldsymbol{u} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}\right) \\
&= \boldsymbol{x} - t \lim_{k\to+\infty} \left(\frac{\int_{\mathbb{R}^n} \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k) e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} \, d\boldsymbol{u}}\right),
\end{aligned}$$

which establishes (4.31). This concludes the proof of (ii).

4.D Proof of Proposition 4.3.3

Let us first introduce some notation. Let $\boldsymbol{x} \in \mathbb{R}^n$, $t > 0$, $\epsilon > 0$, and $\boldsymbol{u}_0 \in \operatorname{dom} \partial J$. Define the functions

$$\operatorname{dom} \partial J \ni \boldsymbol{u} \mapsto \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t) = \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0}),$$

$$\operatorname{dom} \partial J \ni \boldsymbol{u} \mapsto \Phi_J(\boldsymbol{u}|\boldsymbol{x}, t) = \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + J(\boldsymbol{u}),$$

and

$$\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x}, t) = \left(\frac{\boldsymbol{u} - \boldsymbol{x}}{t}\right) + \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k).$$

Note that for every $\boldsymbol{u} \in \mathbb{R}^n$, $\varphi_J(\boldsymbol{u}|\boldsymbol{x}, t)$ is a subgradient of the function $\boldsymbol{v} \mapsto \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + J(\boldsymbol{v})$ evaluated at $\boldsymbol{v} = \boldsymbol{u}$ and $\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x}, t)$ is a subgradient of the function $\boldsymbol{v} \mapsto \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + S(\boldsymbol{v}, \mu_k)$ evaluated at $\boldsymbol{v} = \boldsymbol{u}$. Let $\{\mu_k\}_{k=1}^{+\infty}$ be a sequence of positive real numbers converging to zero and let $S_0 \colon \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}$ denote the solution to the first-order HJ PDE (1.22) with initial data J (see Proposition 1.2.14). Note that the sequence $\{S_0(\boldsymbol{u}, \mu_k)\}_{k=1}^{+\infty}$ is uniformly bounded from below since

$$\begin{aligned} S_0(\boldsymbol{u}, \mu_k) &= \inf_{\boldsymbol{v} \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 + J(\boldsymbol{v}) \right\} \\ &\geqslant J(\boldsymbol{u}) \\ &\geqslant 0. \end{aligned} \tag{4.114}$$

Now, define the function $F \colon \operatorname{dom} \partial J \times \operatorname{dom} \partial J \times \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}$ as

$$F(\boldsymbol{u}, \boldsymbol{u}_0, \boldsymbol{x}, t) = \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x}, t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x}, t), \boldsymbol{u} - \boldsymbol{u}_0 \rangle \frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + J(\boldsymbol{u})\right)/\epsilon} d\boldsymbol{u}} \tag{4.115}$$

and the sequence of functions $\{F_{\mu_k}\}_{k=1}^{+\infty}$ with $F_{\mu_k} \colon \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}$ as

$$F_{\mu_k}(\boldsymbol{u}, \boldsymbol{u}_0, \boldsymbol{x}, t) = \left\langle \varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x}, t) - \varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}_0|\boldsymbol{x}, t), \boldsymbol{u} - \boldsymbol{u}_0 \right\rangle \frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + S_0(\boldsymbol{u},\mu_k)\right)/\epsilon} d\boldsymbol{u}}. \tag{4.116}$$

Since $\lim_{k \to +\infty} S_0(\boldsymbol{u}, \mu_k) = J(\boldsymbol{u})$ and $\lim_{k \to +\infty} \nabla_{\boldsymbol{u}} S_0(\boldsymbol{u}, \mu_k) = \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})$ for every $\boldsymbol{u} \in \operatorname{dom} \partial J$ by

Proposition 1.2.14(i) and (iv), and

$$\lim_{k\to+\infty}\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u} = \int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u} \tag{4.117}$$

by (4.29) in Proposition 4.3.2(ii) and continuity of the logarithm, the limit

$$\lim_{k\to+\infty} F_\mu(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t) = F(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t)$$

holds for every $\boldsymbol{u}\in\operatorname{dom}\partial J$, $\boldsymbol{u}_0\in\operatorname{dom}\partial J$, $\boldsymbol{x}\in\mathbb{R}^n$ and $t>0$. Note that since J is m-strongly convex, the functions $\boldsymbol{u}\mapsto\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})$ and $\boldsymbol{u}\mapsto\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)$ are $\left(\frac{1+mt}{t}\right)$-strongly convex. As a consequence, for every pair $(\boldsymbol{u},\boldsymbol{u}_0)\in\operatorname{dom}\partial J\times\operatorname{dom}\partial J$, the following monotonicity inequalities hold (see Definition 8, Equation (1.20)):

$$0\leqslant\left(\frac{1+mt}{t}\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2\leqslant\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)-\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle \tag{4.118}$$

and

$$0\leqslant\left(\frac{1+mt}{t}\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2\leqslant\left\langle\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x},t)-\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\right\rangle. \tag{4.119}$$

Multiply the first set of inequalities by $e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}/\int_{\mathbb{R}^n}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}$ and the second set of inequalities by $e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}/\int_{\mathbb{R}^n}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}$ and use the definition of F and F_{μ_k} to get the inequalities

$$0\leqslant\left(\frac{1+mt}{t}\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2\,\frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}}{\int_{\mathbb{R}^n}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+J(\boldsymbol{u})\right)/\epsilon}\,d\boldsymbol{u}}\leqslant F(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t) \tag{4.120}$$

$$0\leqslant\left(\frac{1+mt}{t}\right)\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2\,\frac{e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}}{\int_{\mathbb{R}^n}e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}\leqslant F_{\mu_k}(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t).$$

These inequalities show, in particular, that F and F_{μ_k} are both non-negative functions for every $(\boldsymbol{u},\boldsymbol{u}_0)\in\operatorname{dom}\partial J\times\operatorname{dom}\partial J$, $\boldsymbol{x}\in\mathbb{R}^n$, and $t>0$. As a consequence, Fatou's lemma ([108], Lemma

2.18) applies to the sequence of functions $\{F_{\mu_k}\}_{k=1}^{+\infty}$, and hence

$$\begin{aligned}\int_{\text{dom }\partial J} F(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t)\,d\boldsymbol{u} &\leqslant \liminf_{k\to+\infty}\int_{\text{dom }\partial J} F_{\mu_k}(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t)\,d\boldsymbol{u}\\ &\leqslant \liminf_{k\to+\infty}\int_{\mathbb{R}^n} F_{\mu_k}(\boldsymbol{u},\boldsymbol{u}_0,\boldsymbol{x},t)\,d\boldsymbol{u}\\ &= \liminf_{k\to+\infty}\left(\frac{\int_{\mathbb{R}^n}\left\langle\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\right\rangle e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}\right.\\ &\quad\left.-\frac{\int_{\mathbb{R}^n}\left\langle\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\right\rangle e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}\right).\end{aligned} \tag{4.121}$$

We now wish to compute the limit in (4.121). On the one hand, we can apply formula (4.24) in Proposition 4.3.2(i) (with initial data $S_0(\cdot,\mu_k)$ and using $\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x},t)=\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)+\nabla_{\boldsymbol{u}}S_0(\boldsymbol{u},\mu_k)$) to the first integral on the right side on the last line of (4.121) to get

$$\frac{\int_{\mathbb{R}^n}\left\langle\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\right\rangle e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}=n\epsilon. \tag{4.122}$$

On the other hand, applying the limit result (4.31) in Proposition 4.3.2(ii) for the posterior mean estimate $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ and the limit $\lim_{k\to+\infty}\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}_0|\boldsymbol{x},t)=\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)=\left(\frac{\boldsymbol{u}_0-\boldsymbol{x}}{t}\right)+\pi_{\partial J(\boldsymbol{u}_0)}(\boldsymbol{0})$ to the second integral on the right side on the last line of (4.121), we get

$$\begin{aligned}\liminf_{k\to+\infty}\frac{\int_{\mathbb{R}^n}\left\langle\varphi_{S_0(\cdot,\mu_k)}(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\right\rangle e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}{\int_{\mathbb{R}^n} e^{-\left(\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2+S_0(\boldsymbol{u},\mu_k)\right)/\epsilon}\,d\boldsymbol{u}}&\\ =\left\langle\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{u}_0\right\rangle.&\end{aligned} \tag{4.123}$$

Combine (4.12), (4.115), (4.120), (4.121), (4.122), and (4.123) to get

$$\begin{aligned}\left(\frac{1+mt}{t}\right)\mathbb{E}_J\left[\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2\right]&\leqslant\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)-\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}-\boldsymbol{u}_0\rangle\right]\\ &\leqslant n\epsilon-\left\langle\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t),\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{u}_0\right\rangle.\end{aligned}$$

This establishes the set of inequalities (4.32).

Next, we show that $\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right\|_2\right]<+\infty$ indirectly using the set of inequalities (4.32). By

Proposition 4.3.1, $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) \in \text{int (dom } J)$. Hence there exists a number $\delta > 0$ such that the open ball $\{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2 < \delta\}$ is contained in int (dom J). Let $\boldsymbol{u}_0 \in \{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2 < \delta\}$ with $\boldsymbol{u}_0 \neq \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$. Recall int (dom J) $\subset$ dom ∂J, so that both $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$ and $\boldsymbol{u}_0$ are in the set dom ∂J. We claim that

$$\mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \boldsymbol{u}_0\rangle|\right] < +\infty.$$

Indeed, using the triangle inequality, the set of inequalities (4.32) proven previously, the Cauchy-Schwarz inequality, and that $\mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_0\|_2\right] \leqslant \left(\int_{\mathbb{R}^n} \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \, e^{-\frac{1}{2t\epsilon}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2} \, d\boldsymbol{u}\right) / Z_J(\boldsymbol{x},t,\epsilon) < +\infty$ by assumption (A3),

$$\begin{aligned}
\mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \boldsymbol{u}_0\rangle|\right] &= \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \boldsymbol{u}_0 + (\boldsymbol{u} - \boldsymbol{u})\rangle|\right] \\
&= \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle - \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle|\right] \\
&\leqslant \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle| + |\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle|\right] \\
&= \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle|\right] \\
&\quad + \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle|\right] \\
&\leqslant \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle|\right] + n\epsilon \\
&= \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) + (\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)), \boldsymbol{u} - \boldsymbol{u}_0\rangle|\right] \\
&\quad + n\epsilon \\
&\leqslant \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle|\right] \\
&\quad + \mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle|\right] + n\epsilon \\
&\leqslant \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u} - \boldsymbol{u}_0\rangle\right] \\
&\quad + \mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2 \, \|\boldsymbol{u} - \boldsymbol{u}_0\|_2\right] + n\epsilon \\
&\leqslant n\epsilon - \langle \varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t), \boldsymbol{u}_{PM}(x,t,\epsilon) - \boldsymbol{u}_0\rangle \\
&\quad + \|\varphi_J(\boldsymbol{u}_0|\boldsymbol{x},t)\|_2 \, \mathbb{E}_J\left[\|\boldsymbol{u} - \boldsymbol{u}_0\|_2\right] + n\epsilon \\
&< +\infty.
\end{aligned} \tag{4.124}$$

This shows that $\mathbb{E}_J\left[|\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \boldsymbol{u}_0\rangle|\right] < +\infty$ for every $\boldsymbol{u}_0 \in \{\boldsymbol{u} \in \mathbb{R}^n \mid \|\boldsymbol{u} - \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2 < \delta\}$ different from $\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)$. Now, let $\{\boldsymbol{e}_i\}_{i=1}^n$ denote the standard ba-

sis in $\mathbb{R}^n$ and let $\{\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i\}_{i=1}^n$ denote the components of the vector $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)$, i.e., $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t) = (\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_1,\ldots,\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_n)$. Using (4.124) with the choice of $\boldsymbol{u}_0 = \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \frac{\delta}{2}\boldsymbol{e}_i$, which is contained in the open ball $\{\boldsymbol{u}\in\mathbb{R}^n \mid \|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\|_2 < \delta\}$ for each $i\in\{1,\ldots,n\}$, we get

$$\begin{aligned}
\mathbb{E}_J\left[\left|\left\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - (\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \frac{\delta}{2}\boldsymbol{e}_i)\right\rangle\right|\right] &= \mathbb{E}_J\left[\left|\left\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \frac{\delta}{2}\boldsymbol{e}_i\right\rangle\right|\right] \\
&= \frac{\delta}{2}\mathbb{E}_J\left[|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|\right] \\
&\leqslant 2n\epsilon - \left\langle \varphi_J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \frac{\delta}{2}\boldsymbol{e}_i \mid \boldsymbol{x},t), \frac{\delta}{2}e_i\right\rangle \\
&\quad + \left\|\varphi_J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \frac{\delta}{2}\boldsymbol{e}_i|\boldsymbol{x},t)\right\|_2 \mathbb{E}_J\left[\left\|\boldsymbol{u} - (\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \frac{\delta}{2}\boldsymbol{e}_i)\right\|_2\right] \\
&< +\infty.
\end{aligned} \tag{4.125}$$

Using (4.125) and the norm inequality $\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2 \leqslant \sum_{i=1}^n |\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|$, we can bound $\mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right]$ as follows

$$\begin{aligned}
0 \leqslant \mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right] &\leqslant \mathbb{E}_J\left[\sum_{i=1}^n |\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|\right] \\
&= \sum_{i=1}^n \mathbb{E}_J\left[|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)_i|\right] \\
&\leqslant 2n^2\epsilon - \sum_{i=1}^n \left\langle \varphi_J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) - \frac{\delta}{2}\boldsymbol{e}_i|\boldsymbol{x},t), \frac{\delta}{2}e_i\right\rangle \\
&\quad + \sum_{i=1}^n \left\|\varphi_J(\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon) + \frac{\delta}{2}\boldsymbol{e}_i|\boldsymbol{x},t)\right\|_2 \mathbb{E}_J\left[\left\|\boldsymbol{u} - (\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\frac{\delta}{2}\boldsymbol{e}_i)\right\|_2\right] \\
&< +\infty.
\end{aligned} \tag{4.126}$$

This shows that $\mathbb{E}_J\left[\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\|_2\right] < +\infty$. Finally, use (4.126), $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t) = \frac{\boldsymbol{u}-\boldsymbol{x}}{t} + \pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})$, and

assumption (A3) to find

$$\begin{aligned}
\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right\|_2\right] &= \mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\mathbf{0}) + \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right) - \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right\|_2\right] \\
&\leqslant \mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\mathbf{0}) + \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right\|_2 + \left\|\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right\|_2\right] \\
&= \mathbb{E}_J\left[\left\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\right\|_2\right] + \mathbb{E}_J\left[\left\|\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right\|_2\right] \\
&\leqslant \mathbb{E}_J\left[\left\|\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)\right\|_2\right] + \frac{1}{tZ_J(\boldsymbol{x},t,\epsilon)}\int_{\mathbb{R}^n}\left\|\boldsymbol{u}-\boldsymbol{x}\right\|_2 e^{-\frac{1}{2t\epsilon}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2}\,d\boldsymbol{u} \\
&< +\infty.
\end{aligned}$$

This shows that $\mathbb{E}_J\left[\left\|\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right\|_2\right] < +\infty$.

4.E Proof of Proposition 4.3.5

Proof of (i): Let $\boldsymbol{x} \in \mathbb{R}^n$ and $t > 0$ and define the functions

$$\text{dom } \partial J \ni \boldsymbol{u} \mapsto \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) = \left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right) + \pi_{\partial J(\boldsymbol{u})}(\mathbf{0}),$$

$$\text{dom } \partial J \ni \boldsymbol{u} \mapsto \Phi_J(\boldsymbol{u}|\boldsymbol{x},t) = \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u}).$$

Note that for every $\boldsymbol{u} \in \mathbb{R}^n$, $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)$ is a subgradient of the function $\boldsymbol{v} \mapsto \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_2^2 + J(\boldsymbol{v})$ evaluated at $\boldsymbol{v} = \boldsymbol{u}$.

Let $\boldsymbol{v} \in \text{dom } \partial J$. The Bregman divergence of the function $\text{dom } \partial J \ni \boldsymbol{u} \mapsto \Phi_J(\boldsymbol{u}|\boldsymbol{x},t)$ at $(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x},t))$ is given by

$$\begin{aligned}
D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x},t)) &= \Phi_J(\boldsymbol{v}|\boldsymbol{x},t) - \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t)), \boldsymbol{v}\rangle + \Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)) \\
&\equiv \Phi_J(\boldsymbol{v}|\boldsymbol{x},t) - \Phi_J(\boldsymbol{u}|\boldsymbol{x},t) + \langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t)), \boldsymbol{u}-\boldsymbol{v}\rangle\,,
\end{aligned}$$

where the second equality follows the definition of the convex conjugate (1.3) and that $\varphi_J(\boldsymbol{u}|\boldsymbol{x},t) \in \partial\Phi_J(\boldsymbol{u},\boldsymbol{x},t)$.

Take the expected value with respect to the variable $\boldsymbol{u}$ over dom ∂J to find

$$\begin{aligned}\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v},\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right] &= \Phi_J(\boldsymbol{v}|\boldsymbol{x},t) - \mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)),\boldsymbol{v}\rangle + \Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right] \\ &\equiv \Phi_J(\boldsymbol{v}|\boldsymbol{x},t) - \mathbb{E}_J\left[\Phi_J(\boldsymbol{u}|\boldsymbol{x},t) + \langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)),\boldsymbol{u}-\boldsymbol{v}\rangle\right].\end{aligned} \tag{4.127}$$

We claim that the expected value $\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v},\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ is finite. We will show this by proving, in turn, that the expected values $\mathbb{E}_J\left[\Phi_J(\boldsymbol{u}|\boldsymbol{x},t)\right]$ and $\mathbb{E}_J\left[\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)),\boldsymbol{u}-\boldsymbol{v}\rangle\right]$ are finite. Establishing the finiteness of $\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v},\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ will enable us to conclude that the expected value $\mathbb{E}_J\left[\Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ on the right hand side of the first equality of (4.127) is also finite.

First, using the definition of $\Phi_J(\boldsymbol{u}|\boldsymbol{x},t)$ we have

$$\mathbb{E}_J\left[\Phi_J(\boldsymbol{u}|\boldsymbol{x},t)\right] \equiv \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2 + J(\boldsymbol{u})\right].$$

The expected value $\mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2\right]$ is finite because we can use the definitions of the posterior mean estimate and inequality (4.33) (with $m \equiv 0$ in (4.33)) to express it as

$$\begin{aligned}0 < \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2\right] &= \mathbb{E}_J\left[\frac{1}{2t}\left\|(\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)) - (\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon))\right\|_2^2\right] \\ &= \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2\right] + \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2\right] + \\ &\quad 2\mathbb{E}_J\left[\langle\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon),\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\rangle\right] \\ &= \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2 + \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2\right] \\ &\quad + 2\left\langle\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon),\mathbb{E}_J\left[\boldsymbol{u}\right]-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle \\ &= \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2 + \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2\right] \\ &\quad + 2\left\langle\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon),\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\rangle \\ &= \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2 + \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{u}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2\right] \\ &\leqslant \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)\right\|_2^2 + \frac{n\epsilon}{2}.\end{aligned}$$

The expected value $\mathbb{E}_J\left[J(\boldsymbol{u})\right]$ is also finite because it is bounded by the set of inequalities (4.23) in Proposition 4.3.1. Hence, the expected value $\mathbb{E}_J\left[\Phi_J(\boldsymbol{u}|\boldsymbol{x},t)\right] \equiv \mathbb{E}_J\left[\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{u}\right\|_2^2\right] + \mathbb{E}_J\left[J(\boldsymbol{u})\right]$ is

finite.

Second, note that the expected value $\mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right]$ can be written as

$$\begin{aligned}\mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] &= \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle + \langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] \\ &= \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] + \langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \mathbb{E}_J\left[\boldsymbol{u}\right]-\boldsymbol{v}\rangle \\ &= \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] + \langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{v}\rangle\,.\end{aligned} \tag{4.128}$$

Apply the monotonicity property (4.32) to the expected value $\mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right]$ (with $\boldsymbol{u}_0 \equiv \boldsymbol{v}$ in (4.32)) in the previous equation to find

$$0 \leqslant \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] \leqslant n\epsilon - \langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{v}\rangle\,.$$

Add the term $\langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{v}\rangle$ on both sides of these inequalities to get

$$\begin{aligned}\langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{v}\rangle &\leqslant \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t) - \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] \\ &\quad + \langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{v}\rangle \\ &\leqslant n\epsilon.\end{aligned} \tag{4.129}$$

Combine the inequalities (4.129) with the equality (4.128) to find

$$\langle \varphi_J(\boldsymbol{v}|\boldsymbol{x},t), \boldsymbol{u}_{PM}(\boldsymbol{x},t,\epsilon)-\boldsymbol{v}\rangle \leqslant \mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right] \leqslant n\epsilon.$$

These bounds prove that the expected value $\mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t), \boldsymbol{u}-\boldsymbol{v}\rangle\right]$ is finite.

The previous arguments show that the expected value $\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ is finite. Now, we claim that the expected value $\mathbb{E}_J\left[\langle \varphi_J(\boldsymbol{u}|\boldsymbol{x},t)), \boldsymbol{v}\rangle\right] \equiv \langle \mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right], \boldsymbol{v}\rangle$ is finite. Indeed, we can use the representation formula (4.15) for expressing the posterior mean estimate in terms of the gradient $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)$ of the solution to the viscous HJ PDE (4.14) and use that $\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right]$

is finite (Proposition 4.3.3) to write

$$\begin{aligned}\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right] &= \mathbb{E}_J\left[\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)+\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right]\\ &= \mathbb{E}_J\left[\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{t}\right)\right]+\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right]\\ &\equiv -\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)+\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right],\end{aligned}$$

where both terms on the right hand side are finite. This shows that $\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ is finite.

Using that $\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v},\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ and $\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ are finite in Equation (4.127), we conclude that the expected value $\mathbb{E}_J\left[\Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$ is also finite. We can now use the definitions of Φ_J and φ_J to express Equation (4.127) as

$$\begin{aligned}\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v},\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right] &= \mathbb{E}_J\left[\Phi_J(\boldsymbol{v}|\boldsymbol{x},t)-\langle\varphi_J(\boldsymbol{u}|\boldsymbol{x},t)),\boldsymbol{v}\rangle+\Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]\\ &= \frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_2^2+J(\boldsymbol{v})\\ &\quad+\left\langle\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)-\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right],\boldsymbol{v}\right\rangle+\mathbb{E}_J\left[\Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right],\end{aligned}\tag{4.130}$$

where, again, we used that $\mathbb{E}_J\left[\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]=-\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)+\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right]$. Now, let

$$\tilde{J}(\boldsymbol{v})=J(\boldsymbol{v})+\left\langle\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t)-\mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\boldsymbol{0})\right],\boldsymbol{v}\right\rangle.$$

Take the infimum over $\boldsymbol{v}\in\mathbb{R}^n$ on both sides of Equation (4.130) to find:

$$\inf_{\boldsymbol{v}\in\mathbb{R}^n}\mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v},\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]=\inf_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\frac{1}{2t}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_2^2+\tilde{J}(\boldsymbol{v})\right\}+\mathbb{E}_J\left[\Phi_J^*(\varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$$

Now, note that by assumption (A1), the function $\boldsymbol{v}\mapsto\tilde{J}(\boldsymbol{v})\in\Gamma_0(\mathbb{R}^n)$. Hence the function $\boldsymbol{v}\ni\mathbb{R}^n\to\mapsto\frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_2^2+\tilde{J}(\boldsymbol{v})$ is strictly convex and has a unique minimizer denoted by $\bar{\boldsymbol{v}}$. Therefore, the infimum in the equality above can be replaced by a minimum. In addition, recall that $\min_{\boldsymbol{v}\in\mathbb{R}^n}\frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_2^2+\tilde{J}(\boldsymbol{v})$ corresponds to the solution to the first-order HJ PDE (1.22) with initial condition $\tilde{J}$. Using Proposition 1.2.14(ii), the unique minimizer $\bar{\boldsymbol{v}}$ can be expressed using

the inclusion relation

$$\left(\frac{\boldsymbol{x}-\bar{\boldsymbol{v}}}{t}\right) \in \partial J(\bar{\boldsymbol{v}}) + \left(\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) - \mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right]\right). \tag{4.131}$$

Therefore, the minimizer $\bar{\boldsymbol{v}}$ is also the unique minimizer to $\boldsymbol{v} \mapsto \mathbb{E}_J\left[D_{\Phi_J}(\boldsymbol{v}, \varphi_J(\boldsymbol{u}|\boldsymbol{x},t))\right]$.

Proof of (ii): If dom $J = \mathbb{R}^n$, then the representation formula $\nabla_{\boldsymbol{x}} S_\epsilon(\boldsymbol{x},t) = \mathbb{E}_J\left[\pi_{\partial J(\boldsymbol{u})}(\mathbf{0})\right]$ derived in Proposition 4.3.2 holds and the characterization of the unique minimizer $\bar{\boldsymbol{v}}$ in equation (4.131) reduces to

$$\left(\frac{\boldsymbol{x}-\bar{\boldsymbol{v}}}{t}\right) \in \partial J(\bar{\boldsymbol{v}}).$$

By Proposition 1.2.14(ii), the unique minimizer that satisfies this characterization is the MAP estimate $\boldsymbol{u}_{MAP}(\boldsymbol{x},t)$, i.e., $\bar{\boldsymbol{v}} = \boldsymbol{u}_{MAP}(\boldsymbol{x},t)$.

Chapter Five

Discussion and future work

5.1 Future Work

Chapter 2 of this dissertation presented novel accelerated nonlinear primal-dual hybrid gradient (PDHG) methods to solve a broad class of convex optimization problems with saddle-point structure. These methods were used in Chapter 3 to solve certain sparse logistic regression, regularized maximum entropy estimation and entropy-regularized zero-sum matrix games problems in a way that is far more efficient than competing methods. Work that applies these methods to concrete problems will be pursued in the future. They should prove useful to a broad class of supervised machine learning not covered in this dissertation, including regression and classification problems defined on the unit simplex as well boosting and structured prediction algorithms for classification problems.

The latter, boosting and structured prediction algorithms, are central for solving classification problems in many applications, such as natural language processing and computational biology. Several of these algorithms, e.g., AdaBoost, LogitBoost, soft-margin LPBoost, and conditional random fields, correspond to entropy maximization problems via their dual problems. Most optimization methods for these algorithms, however, ignore these connections. As these connections can be leveraged by nonlinear PDHG methods (e.g., such as in sparse logistic regression and regularized maximum entropy methods) for speed and efficiency, one can anticipate that nonlinear PDHG methods would work particularly well for boosting and structured prediction algorithms.

It would be interesting and particularly useful to extend the accelerated nonlinear PDHG methods described here to the stochastic case for problems that are separable in the dual variable, and to the non-convex case to deal with large-scale non-convex problems, such as those arising in deep learning. These extensions will be pursued in in future work as well.

In addition, chapter 2 highlighted how a broad class of supervised machine learning algorithms correspond to solutions of first-order Hamilton–Jacobi partial differential equations (HJ PDEs) with initial data. As the representation formulas for many first-order HJ PDEs can be cast as convex optimization problems with appropriate saddle-point structure, the accelerated nonlinear

PDHG optimization methods presented in chapter 2 should prove particularly efficient and robust for solving high-dimensional first-order HJ PDEs, including those that arise in optimal control and in imaging science [69, 68, 154].

Chapter 2 of this dissertation presented connections between first-order HJ PDEs and a broad class of supervised machine learning algorithms, but it did not discuss how these connections could be applied or used in practice. It would be of interest to investigate how these connections could be used in practice. For sparse logistic regression, in particular, it would be of interest to use these connections to HJ PDEs for the problem of controlling the false discovery rate inherent to variable selection with logistic regression via knockoff statistics [13, 39, 14]. As these statistics can be built from sparse logistic regression, and hence from the solution to an HJ PDE, it may be possible to obtain new properties for these statistics that may suggest a novel way to create a statistic that is more desirable than other statistics to control the false discovery rate.

Chapter 4 presented connections between some viscous Hamilton–Jacobi partial differential equations and a broad class of posterior mean (PM) estimators with log-concave prior and quadratic data fidelity term. These connections were leveraged to establish representation formulas and various properties of these PM estimators. In particular, we used these connections to show that some Bayesian PM estimators can be expressed as proximal mappings of smooth functions and we derived representation formulas for these functions. Based on preliminary work, we expect that similar results can be established for certain posterior mean estimators with log-concave prior but with data fidelity term corresponding to Poisson noise and Speckle noise.

BIBLIOGRAPHY

[1] Thomas NO Achia, Anne Wangombe, and Nancy Khadioli. A logistic regression model to identify key determinants of poverty using demographic and health survey data. *European Journal of Social Sciences*, 13(1), 2010.

[2] Marianne Akian, Ravindra Bapat, and Stéphane Gaubert. Max-plus algebra. *Handbook of linear algebra*, 39, 2006.

[3] Marianne Akian, Stéphane Gaubert, and Asma Lakhoua. The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis. *SIAM Journal on Control and Optimization*, 47(2):817–848, 2008.

[4] Giovanni Alberti, Luigi Ambrosio, and Piermarco Cannarsa. On the singularities of convex functions. *Manuscripta Math*, 76(3-4):421–435, 1992.

[5] Marc Allain, Jérôme Idier, and Yves Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Transactions on Image Processing*, 15(5):1130–1142, 2006.

[6] Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pages 139–153. Springer, 2006.

[7] Galen Andrew and Jianfeng Gao. Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40, 2007.

[8] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[9] Gilles Aubert and Pierre Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations*, volume 147. Springer Science & Business Media, 2006.

[10] J-P Aubin and Arrigo Cellina. *Differential inclusions: set-valued maps and viability theory*, volume 264. Springer Science & Business Media, 2012.

[11] Steven C Bagley, Halbert White, and Beatrice A Golomb. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54(10):979–985, 2001. ISSN 0895-4356. doi: https://doi.org/10.1016/S0895-4356(01)00372-9.

[12] Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Trans. Inform. Theory*, 51(7):2664–2669, 2005.

[13] Rina Foygel Barber and Emmanuel Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[14] Rina Foygel Barber, Emmanuel Candès, and Richard J. Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431, 2020.

[15] Martino Bardi and Italo Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 1997. ISBN 0-8176-3640-4. doi: 10.1007/978-0-8176-4755-1. With appendices by Maurizio Falcone and Pierpaolo Soravia.

[16] Martino Bardi and Lawrence C. Evans. On Hopf's formulas for solutions of Hamilton-Jacobi equations. *Nonlinear Anal.*, 8(11):1373–1381, 1984.

[17] Michel Barlaud, Antonin Chambolle, and Jean-Baptiste Caillau. Classification and feature selection using a primal-dual method and projection on structured constraints. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6538–6545. IEEE, 2021.

[18] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*. Mathématiques et Applications. Springer-Verlag Berlin Heidelberg, 1994. ISBN 978-3-540-58422-3.

[19] Emmanuel Nicholas Barron, Lawrence C. Evans, and Robert Jensen. Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls. *Journal of Differential Equations*, 53(2):213 – 233, 1984. ISSN 0022-0396. doi: 10.1016/0022-0396(84)90040-8.

[20] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Essential smoothness, essential strict convexity, and legendre functions in banach spaces. *Communications in Contemporary Mathematics*, 3(04):615–647, 2001.

[21] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

[22] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[23] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[24] Martin Benning, Carola-Bibiane Schönlieb, Tuomo Valkonen, and Verner Vlačić. Explorations on anisotropic regularisation of dynamic inverse problems by bilevel optimisation. *arXiv preprint arXiv:1602.01278*, 2016.

[25] Przemysław Bereziński, Bartosz Jasiul, and Marcin Szpyrka. An entropy-based network anomaly detection method. *Entropy*, 17(4):2367–2408, 2015.

[26] Adam Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[27] Dimitris Bertsimas, Jean Pauphilet, and Bart Van Parys. Sparse regression: Scalable algorithms and empirical performance. *arXiv preprint arXiv:1902.06547*, 2019.

[28] Charles Boncelet. Image noise models. In *The Essential Guide to Image Processing*, pages 143–167. Elsevier, 2009.

[29] Radu Ioan Boţ, Ernö Robert Csetnek, André Heinrich, and Christopher Hendrich. On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. *Mathematical Programming*, 150(2):251–279, 2015.

[30] Charles Bouman and Ken Sauer. A generalized gaussian image model for edge-preserving map estimation. *IEEE Transactions on image processing*, 2(3):296–310, 1993.

[31] Ajay Kumar Boyat and Brijendra Kumar Joshi. A review paper: noise models in digital image processing. *Signal & Image Processing: An International Journal (SIPIJ)*, 6(2), 2015.

[32] Kristian Bredies and Martin Holler. A tgv-based framework for variational image decompression, zooming, and reconstruction. part i: Analytics. *SIAM Journal on Imaging Sciences*, 8 (4):2814–2850, 2015.

[33] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[34] Björn Bringmann, Daniel Cremers, Felix Krahmer, and Michael Möller. The homotopy method revisited: Computing solution paths of ℓ_1-regularized problems. *Mathematics of Computation*, 87(313):2343–2364, 2018. doi: 10.1090/mcom/3287. URL https://doi.org/10.1090/mcom/3287.

[35] Martin Burger and Felix Lucka. Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper bayes estimators. *Inverse Probl.*, 30(11):114004, 2014.

[36] Martin Burger, Yiqui. Dong, and Federica Sciacchitano. Bregman cost for non-Gaussian noise. *arXiv preprint arXiv:1608.07483*, 2016.

[37] Zoran Bursac, C Heath Gauss, David Keith Williams, and David W. Hosmer. Purposeful selection of variables in logistic regression. *Source code for biology and medicine*, 3(1):1–8, 2008.

[38] Emmanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[39] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[40] Jose A. Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for Wasserstein gradient flows. *Foundations of Computational Mathematics*, pages 1–55, 2021.

[41] Alejandro Catalina, Carlos M Alaíz, and José R Dorronsoro. scho. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[42] Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

[43] Volkan Cevher, Stephen Becker, and Mark Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, 2014.

[44] Antonin Chambolle and Jérôme Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International journal of computer vision*, 84(3):288, 2009.

[45] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

[46] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[47] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[48] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numer.*, 25:161–319, 2016.

[49] Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[50] Frédéric Champagnat and Jérôme Idier. A connection between half-quadratic criteria and em algorithms. *IEEE Signal Processing Letters*, 11(9):709–712, 2004.

[51] Tony Chan, Antonio Marquina, and Pep Mulet. High-order total variation-based image restoration. *SIAM Journal on Scientific Computing*, 22(2):503–516, 2000.

[52] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2):298–311, Feb 1997. ISSN 1941-0042. doi: 10.1109/83.551699.

[53] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Res. Math. Sci.*, 5(3):30, 2018.

[54] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[55] Stanley F. Chen and Ronald Rosenfeld. A survey of smoothing techniques for me models. *IEEE transactions on Speech and Audio Processing*, 8(1):37–50, 2000.

[56] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[57] Cheng Chu, Sang Kyun Kim, Yian Lin, YuanYuan Yu, Gary Bradski, Andrew Y Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. *Advances in neural information processing systems*, 19:281, 2007.

[58] Patrick L. Combettes, Laurent Condat, J-C Pesquet, and BC Vũ. A forward-backward view of some primal-dual optimization methods in image recovery. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4141–4145. IEEE, 2014.

[59] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2):460–479, 2013.

[60] Laurent Condat. Fast projection onto the simplex and the l_1 ball. *Math. Program.*, 158 (1-2, Ser. A):575–585, 2016. ISSN 0025-5610. doi: 10.1007/s10107-015-0946-6. URL https://doi.org/10.1007/s10107-015-0946-6.

[61] Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Structural maxent models. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 391–399. JMLR.org, 2015.

[62] Michael G. Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American mathematical society*, 27(1):1–67, 1992. doi: 10.1090/S0273-0979-1992-00266-5.

[63] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[64] Jérôme Darbon. On convex finite-dimensional variational methods in imaging sciences and hamilton–jacobi equations. *SIAM J. Imaging Sci.*, 8(4):2268–2293, 2015.

[65] Jérôme Darbon and Gabriel P. Langlois. Efficient and robust high-dimensional sparse logistic regression via nonlinear primal-dual hybrid gradient algorithms, 2021. URL https://arxiv.org/abs/2111.15426.

[66] Jérôme Darbon and Gabriel P. Langlois. On bayesian posterior mean estimators in imaging sciences and hamilton–jacobi partial differential equations. *Journal of Mathematical Imaging and Vision*, pages 1–34, 2021.

[67] Jérôme Darbon and Gabriel P. Langlois. Accelerated nonlinear primal-dual hybrid gradient methods with applications to supervised machine learning, 2022. URL https://arxiv.org/abs/2109.12222.

[68] Jérôme Darbon and Tingwei Meng. On decomposition models in imaging sciences and multi-time hamilton-jacobi partial differential equations. *arXiv preprint arXiv:1906.09502*, 2019.

[69] Jérôme Darbon and Stanley Osher. Algorithms for overcoming the curse of dimensionality for certain hamilton–jacobi equations arising in control theory and elsewhere. *Research in the Mathematical Sciences*, 3(1):1–26, 2016.

[70] Jérôme Darbon and Marc Sigelle. Image restoration with discrete constrained total variation part i: Fast and exact optimization. *Journal of Mathematical Imaging and Vision*, 26(3): 261–276, 2006.

[71] Jérôme Darbon, Gabriel P. Langlois, and Tingwei Meng. Overcoming the curse of dimensionality for some hamilton–jacobi partial differential equations via neural network architectures. *Research in the Mathematical Sciences*, 7(3):1–50, 2020.

[72] Jérôme Darbon, Gabriel P. Langlois, and Tingwei Meng. Connecting hamilton-jacobi partial differential equations with maximum a posteriori and posterior mean estimators for some non-convex priors. *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, pages 1–25, 2021.

[73] John N. Darroch and Douglas Ratcliff. Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480, 1972.

[74] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11): 1413–1457, 2004.

[75] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858, 2017.

[76] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence*, 19(4):380–393, 1997.

[77] Yuri Demchenko, Paola Grosso, Cees De Laat, and Peter Membrey. Addressing big data issues in scientific data infrastructure. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 48–55. IEEE, 2013.

[78] Guy Demoment. Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37 (12):2024–2036, 1989.

[79] Jean-Dominique Deuschel and Daniel W. Stroock. *Large deviations*, volume 342. American Mathematical Soc., 2001.

[80] Payal Dhar. The carbon impact of artificial intelligence. *Nat Mach Intell*, 2:423–5, 2020.

[81] David C. Dobson and Fadil Santosa. Recovery of blocky images from noisy and blurred data. *SIAM Journal on Applied Mathematics*, 56(4):1181–1198, 1996.

[82] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[83] Peter M. Dower, William M. McEneaney, and Huan Zhang. Max-plus fundamental solution semigroups for optimal control problems. In *2015 Proceedings of the Conference on Control and its Applications*, pages 368–375. SIAM, 2015.

[84] Yoel Drori, Shoham Sabach, and Marc Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

[85] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[86] Miroslav Dudík and Robert E Schapire. Maximum entropy distribution estimation with generalized regularization. In *International Conference on Computational Learning Theory*, pages 123–138. Springer, 2006.

[87] Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In John Shawe-Taylor and Yoram Singer, editors, *Learning Theory*, pages 472–486, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-27819-1.

[88] Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 2007.

[89] Sylvain Durand, François Malgouyres, and Bernard Rougé. Image deblurring, spectrum interpolation and application to satellite imaging. *ESAIM: Control, Optimisation and Calculus of Variations*, 5:445–475, 2000.

[90] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

[91] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[92] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

[93] M El Guide, K Jbilou, C Koukouvinos, and A Lappa. Comparative study of l 1 regularized logistic regression methods for variable selection. *Communications in Statistics-Simulation and Computation*, pages 1–16, 2020.

[94] Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. A statistical explanation of maxent for ecologists. *Diversity and distributions*, 17(1): 43–57, 2011.

[95] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[96] Virginia Estellers, Stefano Soatto, and Xavier Bresson. Adaptive regularization with the structure tensor. *IEEE Transactions on Image Processing*, 24(6):1777–1790, 2015.

[97] Lawrence C. Evans. *Partial differential equations*. American mathematical society (Providence, RI), 2010.

[98] Matthias Faessler, Flavio Fontana, Christian Forster, Elias Mueggler, Matia Pizzoli, and Davide Scaramuzza. Autonomous, vision-based flight and live dense 3d mapping with a quadrotor micro aerial vehicle. *Journal of Field Robotics*, 33(4):431–450, 2016.

[99] Kenneth J. Falconer. *The Geometry of Fractal Sets*. Cambridge Tracts in Mathematics. Cambridge University Press, 1985. doi: 10.1017/CBO9780511623738.

[100] Herbert Federer. *Geometric measure theory*. Springer, 1969.

[101] Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM Journal on Optimization*, 29(1):100–134, 2019.

[102] Olivier Fercoq and Peter Richtárik. Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent. *SIAM Review*, 58(4):739–771, 2016.

[103] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[104] Mário AT Figueiredo and Robert D. Nowak. Wavelet-based image estimation: an empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Transactions on Image Processing*, 10(9):1322–1331, 2001.

[105] William Fithian and Trevor Hastie. Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4):1917, 2013.

[106] Wendell H. Fleming and William M. McEneaney. A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering. *SIAM Journal on Control and Optimization*, 38(3):683–710, 2000. doi: 10.1137/S0363012998332433.

[107] Wendell H. Fleming and Halil Mete Soner. *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media, 2006.

[108] Gerald B. Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

[109] Simon Foucart and Holger Rauhut. *Sparse Solutions of Underdetermined Systems*, pages 41–59. Springer New York, New York, NY, 2013. doi: 10.1007/978-0-8176-4948-7_2.

[110] Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pages 1–39. Springer, 2013.

[111] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007.

[112] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[113] Wilfrid Gangbo, Wuchen Li, Stanley Osher, and Michael Puthawala. Unnormalized optimal transport. *Journal of Computational Physics*, 399:108940, 2019.

[114] Nicolas García Trillos, Zachary Kaplan, and Daniel Sanz-Alonso. Variational characterizations of local entropy and heat regularization in deep learning. *Entropy*, 21(5):511, 2019.

[115] Stephane Gaubert, William McEneaney, and Zheng Qu. Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 1054–1061. IEEE, 2011.

[116] Donald Geman and Yang Chengda. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, July 1995. ISSN 1941-0042. doi: 10.1109/83.392335.

[117] Donald Geman and George Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):367–383, March 1992. ISSN 1939-3539. doi: 10.1109/34.120331.

[118] Alexander Genkin, David D. Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[119] Guy Gilboa, Michael Moeller, and Martin Burger. Nonlinear spectral analysis via one-homogeneous functionals: Overview and future prospects. *Journal of Mathematical Imaging and Vision*, 56(2):300–319, 2016.

[120] Enrico Giusti and Graham Hale Williams. *Minimal surfaces and functions of bounded variation*, volume 80. Springer, 1984.

[121] Izrail Solomonovich Gradshteyn, Iosif Moiseevich Ryzhik, Alan Jeffrey, and Daniel Zwillinger. *Table of integrals, series and products*. Academic Press, 2007.

[122] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-dependent maximum entropy models of neural population codes. *PLoS computational biology*, 9(3):e1002922, 2013.

[123] Casey S. Greene, Jie Tan, Matthew Ung, Jason H Moore, and Chao Cheng. Big data bioinformatics. *Journal of cellular physiology*, 229(12):1896–1900, 2014.

[124] Rémi Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Trans. Signal Process.*, 59(5):2405–2410, 2011.

[125] Rémi Gribonval and Pierre Machart. Reconciling" priors" &" priors" without prejudice? In *Advances in Neural Information Processing Systems*, pages 2193–2201, 2013.

[126] Rémi Gribonval and Mila Nikolova. On Bayesian estimation and proximity operators. *Applied and Computational Harmonic Analysis*, 2019. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2019.07.002.

[127] Rémi Gribonval and Mila Nikolova. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, pages 1–25, 2020. doi: 10.1007/s10851-020-00951-y.

[128] Branko Grünbaum, Victor Klee, Micha A Perles, and Geoffrey Colin Shephard. *Convex polytopes*, volume 16. Springer, 1967.

[129] Yu Gu, Andrew McCallum, and Don Towsley. Detecting anomalies in network traffic using maximum entropy estimation. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, pages 32–32, 2005.

[130] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. doi: 10.1007/978-0-387-84858-7. Data mining, inference, and prediction.

[131] Trevor Hastie, Junyang Qian, and Kenneth Tay. An introduction to glmnet (2021). Available at "https://glmnet.stanford.edu/articles/glmnet.html". Accessed on 17 February 2021., 2021.

[132] Bruce Hayes and Colin Wilson. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440, 2008.

[133] Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5 (1):119–149, 2012.

[134] Tim Hesterberg, Nam Hee Choi, Lukas Meier, Chris Fraley, et al. Least angle and l^1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

[135] Le Thi Khanh Hien and Nicolas Gillis. Algorithms for nonnegative matrix factorization with the kullback–leibler divergence. *Journal of Scientific Computing*, 87(3):1–32, 2021.

[136] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305 of *Grundlehren Text Editions*. Springer Science & Business Media, 1993.

[137] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms II: Advanced Theory and Bundle Methods*, volume 306 of *Grundlehren Text Editions*. Springer Science & Business Media, 1993.

[138] Jean-Baptiste Hiriart-Urruty and Phillipe Plazanet. Moreau's decomposition theorem revisited. In *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, volume 6, pages 325–338. Elsevier, 1989.

[139] Dorit S. Hochbaum. An efficient algorithm for image segmentation, markov random fields and related problems. *Journal of the ACM (JACM)*, 48(4):686–701, 2001.

[140] Thorsten Hohage and Carolin Homann. A generalization of the chambolle-pock algorithm to banach spaces with applications to inverse problems. *arXiv preprint arXiv:1412.0126*, 2018.

[141] Lars Hörmander. *The analysis of linear partial differential operators. I.* Classics in Mathematics. Springer-Verlag, Berlin, 2003. URL https://doi.org/10.1007/978-3-642-61497-2. Distribution theory and Fourier analysis, Reprint of the second (1990) edition [Springer, Berlin; MR1065993 (91m:35001a)].

[142] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[143] Jingwei Hou, Christopher W Ashling, Sean M Collins, Andraž Krajnc, Chao Zhou, Louis Longley, Duncan N Johnstone, Philip A Chater, Shichun Li, Marie-Vanessa Coulet, et al. Metal-organic framework crystal-glass composites. *Nature communications*, 10(1):1–10, 2019.

[144] Jérôme Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE transactions on image processing*, 10(7):1001–1009, 2001.

[145] Martin Jaggi. An equivalence between the lasso and support vector machines. *Regularization, optimization, kernels, and support vector machines*, pages 1–26, 2013.

[146] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[147] Sham Kakade, Shai Shalev-Shwartz, Ambuj Tewari, et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1): 35, 2009.

[148] Cass E. Kalinski. *Building Better Species Distribution Models with Machine Learning: Assessing the Role of Covariate Scale and Tuning in Maxent Models.* PhD thesis, University of Southern California, 2019.

[149] Jagat Narain Kapur. *Maximum-entropy models in science and engineering.* John Wiley & Sons, 1989.

[150] Steven M. Kay. *Fundamentals of statistical signal processing.* Prentice Hall PTR, 1993.

[151] Robert W. Keener. *Theoretical statistics: Topics for a core course.* Springer, 2011.

[152] Johannes HB Kemperman. On the optimum rate of transmitting information. In *Probability and information theory*, pages 126–169. Springer, 1969.

[153] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9 (2):137–163, 2001.

[154] Matthew R Kirchner, Gary Hewer, Jérôme Darbon, and Stanley Osher. A primal-dual method for optimal control and trajectory generation in high-dimensional systems. In *2018 IEEE Conference on Control Technology and Applications (CCTA)*, pages 1583–1590. IEEE, 2018.

[155] Florian Knoll, Martin Holler, Thomas Koesters, Ricardo Otazo, Kristian Bredies, and Daniel K Sodickson. Joint mr-pet reconstruction using a multi-channel image regularizer. *IEEE transactions on medical imaging*, 36(1):1–16, 2016.

[156] Vassili N. Kolokoltsov and Victor P. Maslov. *Idempotent analysis and its applications*, volume 401 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1997. ISBN 0-7923-4509-6. doi: 10.1007/978-94-015-8901-7. URL https://doi.org/10.1007/978-94-015-8901-7. Translation of ıt Idempotent analysis and its application in optimal control (Russian), "Nauka" Moscow, 1994 [MR1375021 (97d:49031)], Translated by V. E. Nazaikinskii, With an appendix by Pierre Del Moral.

[157] Rasmus Dalgas Kongskov, Yiqiu Dong, and Kim Knudsen. Directional total generalized variation regularization. *BIT Numerical Mathematics*, 59(4):903–928, 2019.

[158] Solomon Kullback. A lower bound for discrimination information in terms of variation (corresp.). *IEEE transactions on Information Theory*, 13(1):126–127, 1967.

[159] Taek Mu Kwon. Tmc traffic data automation for mndot's traffic monitoring program. Technical report, University of Minnesota, Twin Cities, 2004.

[160] Guy Lebanon John Lafferty. Boosting and maximum likelihood for exponential models. *Advances in neural information processing systems*, 14:447, 2002.

[161] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l˜ 1 regularized logistic regression. In *Aaai*, volume 6, pages 401–408, 2006.

[162] L. Leindler. On a certain converse of holder inequality. In *Linear Operators and Approximation, Lineare Operatoren und Approximation*, pages 182–184. Springer, 1972.

[163] Xiaoping Li, Yadi Wang, and Rubén Ruiz. A survey on sparse learning models for feature selection. *IEEE Transactions on Cybernetics*, 2020.

[164] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[165] Jialin Liu, Wotao Yin, Wuchen Li, and Yat Tin Chow. Multilevel optimal transport: a fast approximation of wasserstein-1 distances. *SIAM Journal on Scientific Computing*, 43(1): A193–A220, 2021.

[166] Dirk A Lorenz and Thomas Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.

[167] Ignace Loris. L1packv2: A mathematica package for minimizing an l1-penalized functional. *Computer physics communications*, 179(12):895–902, 2008.

[168] Cécile Louchet. *Modèles variationnels et bayésiens pour le débruitage d'images: de la variation totale vers les moyennes non-locales*. PhD thesis, Université René Descartes-Paris V, 2008.

[169] Cécile Louchet and Lionel Moisan. Posterior expectation of the total variation model: properties and experiments. *SIAM J. Imaging Sci.*, 6(4):2640–2684, 2013.

[170] David J. Lowe and Jamshid Parvar. A logistic regression approach to modelling the contractor's decision to bid. *Construction Management and Economics*, 22(6):643–653, 2004.

[171] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5:7776–7797, 2017.

[172] Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[173] Gideon Mann, Ryan McDonald, Mehryar Mohri, Nathan Silberman, and Daniel Walker IV. Efficient large-scale distributed training of conditional maximum entropy models. Technical report, Google Research, 2009.

[174] Christopher Manning and Dan Klein. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials-Volume 5*, pages 8–8, 2003.

[175] William McEneaney. *Max-plus methods for nonlinear control and estimation*. Springer Science & Business Media, 2006.

[176] William McEneaney. A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. *SIAM Journal on Control and Optimization*, 46(4):1239–1276, 2007. doi: 10.1137/040610830.

[177] William M McEneaney and L Jonathan Kluberg. Convergence rate for a curse-of-dimensionality-free method for a class of HJB PDEs. *SIAM Journal on Control and Optimization*, 48(5):3052–3079, 2009.

[178] William M McEneaney, Ameet Deshpande, and Stephane Gaubert. Curse-of-complexity attenuation in the curse-of-dimensionality-free method for HJB PDEs. In *2008 American Control Conference*, pages 4684–4690. IEEE, 2008.

[179] Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[180] Cory Merow, Matthew J Smith, and John A Silander Jr. A practical guide to maxent for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069, 2013.

[181] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, second edition, 2018.

[182] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[183] David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, pages 87–103, 2016.

[184] Luthfi Alwi Muttaqin, Sigit Heru Murti, and Bowo Susilo. Maxent (maximum entropy) model for predicting prehistoric cave sites in karst area of gunung sewu, gunung kidul, yogyakarta. In *Sixth Geoinformation Science Symposium*, volume 11311, page 113110B. International Society for Optics and Photonics, 2019.

[185] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

[186] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[187] Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[188] Yurii Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, 2018.

[189] Mila Nikolova. Weakly constrained minimization: application to the estimation of images and signals involving constant regions. *Journal of Mathematical Imaging and Vision*, 21(2): 155–175, 2004.

[190] Mila Nikolova. Model distortions in bayesian map reconstruction. *Inverse Problems and Imaging*, 1(2):399, 2007.

[191] Mila Nikolova and Michael Ng. Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 1, pages 277–280. IEEE, 2001.

[192] Mila Nikolova and Michael K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3):937–966, 2005.

[193] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

[194] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[195] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209, 2009. ISSN 1053-8119. Mathematics in Brain Imaging.

[196] Marcelo Pereyra. Revisiting maximum-a-posteriori estimation in log-concave models. *SIAM J. Imaging Sci.*, 12(1):650–670, 2019.

[197] Jean-Christophe Pesquet and Audrey Repetti. A class of randomized primal-dual algorithms for distributed optimization. *arXiv preprint arXiv:1406.6404*, 2014.

[198] Washek F. Pfeffer. Divergence theorem for vector fields with singularities. In *New Integrals*, pages 150–166. Springer, 1990.

[199] David L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM (JACM)*, 9(1):84–97, 1962.

[200] Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83, 2004.

[201] Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

[202] Steven J. Phillips, Robert P. Anderson, Miroslav Dudík, Robert E Schapire, and Mary E Blair. Opening the black box: An open-source release of maxent. *Ecography*, 40(7):887–893, 2017.

[203] Mark S. Pinsker. *Information and information stability of random variables and processes*. Holden-Day, 1964.

[204] Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *2011 International Conference on Computer Vision*, pages 1762–1769. IEEE, 2011.

[205] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. An algorithm for minimizing the mumford-shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1133–1140. IEEE, 2009.

[206] Nicholas G. Polson, James G. Scott, Brandon T. Willard, et al. Proximal algorithms in statistics and machine learning. *Statistical Science*, 30(4):559–581, 2015.

[207] Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[208] Tomas Pranckevičius and Virginijus Marcinkevičius. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221, 2017.

[209] A. Prékopa. Logarithmic concave measures with application to stochastic programming. *Acta Sci. Math.*, 32:301–316, 1971.

[210] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael GB Blum. Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, 2018.

[211] Adwait Ratnaparkhi. *Maximum Entropy Models for Natural Language Processing*, pages 800–805. Springer US, Boston, MA, 2017. doi: 10.1007/978-1-4899-7687-1_525.

[212] David S. Rigie and Patrick J. La Rivière. Joint reconstruction of multi-channel, spectral ct data via constrained total nuclear variation minimization. *Physics in Medicine & Biology*, 60 (5):1741, 2015.

[213] Gareth O. Roberts, Jeffrey S. Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[214] Ralph T. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[215] Ralph T. Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[216] Leonid I. Rudin, Osher Stanley, and Fatemi Emad. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[217] Walter Rudin. *Real and complex analysis*. Tata McGraw-hill education, 2006.

[218] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.

[219] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for non-smooth distributed optimization in networks. *arXiv preprint arXiv:1806.00291*, 2018.

[220] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31, 2019.

[221] Hayden Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2197):20160446, 2017.

[222] John L Schnase and Mark L Carroll. Automatic variable selection in ecological niche modeling: A case study using cassin's sparrow (peucaea cassinii). *PloS one*, 17(1):e0257502, 2022.

[223] John L. Schnase, Mark L. Carroll, Roger L. Gill, Glenn S. Tamkin, Jian Li, Savannah L. Strong, Thomas P. Maxwell, Mary E. Aronne, and Caleb S. Spradlin. Toward a monte carlo approach to selecting climate variables in maxent. *PloS one*, 16(3):e0237208, 2021.

[224] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[225] Jianing Shi, Wotao Yin, Stanley Osher, and Paul Sajda. A fast hybrid algorithm for large-scale l1-regularized logistic regression. *The Journal of Machine Learning Research*, 11:713–741, 2010.

[226] Jianing Shi, Wotao Yin, and Stanley Osher. Linearized bregman for l1-regularized logistic regression. In *Proceedings of the 30th international conference on machine learning (ICML)*. Citeseer, 2013.

[227] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

[228] Noah Simon, Jerome Friedman, and Trevor Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*, 2013.

[229] Andrew M. Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[230] Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

[231] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. siam, 2005.

[232] Ioannis Theodossiou. The effects of low-pay and unemployment on psychological well-being: a logistic regression approach. *Journal of health economics*, 17(1):85–104, 1998.

[233] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. Deep learning's diminishing returns: The cost of improvement is becoming unsustainable. *IEEE Spectrum*, 58(10):50–55, 2021.

[234] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[235] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[236] Ryan J. Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

[237] Andrei Nikolaevich Tikhonov, AV Goncharsky, VV Stepanov, and Anatoly G Yagola. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer Science & Business Media, 1995.

[238] Gašper Tkačik, Olivier Marre, Thierry Mora, Dario Amodei, Michael J Berry II, and William Bialek. The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03011, 2013.

[239] Joel Aaron Tropp. *Topics in sparse approximation*. PhD thesis, The University of Texas at Austin, 2004.

[240] Jun'ichi Tsujii and Jun'ichi Kazama. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 137–144, 2003.

[241] Tuomo Valkonen. Block-proximal methods with spatially adapted acceleration. *arXiv preprint arXiv:1609.07373*, 2016.

[242] Rene Vidal, Joan Bruna, Raja Giryes, and Stefano Soatto. Mathematics of deep learning. *arXiv preprint arXiv:1712.04741*, 2017.

[243] Diego Vidaurre, Concha Bielza, and Pedro Larrañaga. A survey of l1 regression. *International Statistical Review*, 81(3):361–387, 2013.

[244] Curtis R. Vogel. *Computational methods for inverse problems*, volume 23. Siam, 2002.

[245] Bang Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[246] Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. doi: 10.1017/9781108627771. A non-asymptotic viewpoint.

[247] Martin J. Wainwright and Michael I. Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[248] Meng Wen, Shigang Yue, Yuchao Tan, and Jigen Peng. A randomized inertial primal-dual fixed point algorithm for monotone inclusions. *arXiv preprint arXiv:1611.05142*, 2016.

[249] G. Winkler. *Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction*, volume 27. Springer Science & Business Media, 2012.

[250] Oliver J. Woodford, Carsten Rother, and Vladimir Kolmogorov. A global perspective on map inference for low-level vision. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2319–2326. IEEE, 2009.

[251] Tong Tong Wu, Kenneth Lange, et al. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[252] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[253] Felipe Yanez and Francis Bach. Primal-dual algorithms for non-negative matrix factorization with the kullback-leibler divergence. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2257–2261. IEEE, 2017.

[254] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011.

[255] Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *The Journal of Machine Learning Research*, 11:3183–3234, 2010.

[256] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. An improved glmnet for l1-regularized logistic regression. *The Journal of Machine Learning Research*, 13:1999–2030, 2012.

[257] Taha Zaghdoudi. Bank failure prediction with logistic regression. *International Journal of Economics and Financial Issues*, 3(2):537, 2013.

[258] M Zaidi and A Amirat. Forecasting stock market trends by logistic regression and neural networks: Evidence from ksa stock market. *Int. J. Econ. Commer. Manag*, 4:4–7, 2016.

[259] Mattia Zanon, Giuliano Zambonin, Gian Antonio Susto, and Seán McLoone. Sparse logistic regression: Comparison of regularization and bayesian implementations. *Algorithms*, 13(6): 137, 2020.

[260] Zhongheng Zhang, Victor Trevino, Sayed Shahabuddin Hoseini, Smaranda Belciug, Arumugam Manivanna Boopathi, Ping Zhang, Florin Gorunescu, Velappan Subha, and Songshi Dai. Variable selection in logistic regression model with genetic algorithm. *Annals of translational medicine*, 6(3), 2018.

[261] Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

[262] Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 34, 2008.

[263] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.