Variational and Dynamical Perspectives On Learning and Optimization

Andre Wibisono

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2016-78
http://www.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-78.html

May 13, 2016

Variational and Dynamical Perspectives On Learning and Optimization

by

Andre Yohannes Wibisono

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Jordan, Chair
Professor Peter Bartlett
Professor Martin Wainwright
Assistant Professor Adityanand Guntuboyina

Spring 2016

Variational and Dynamical Perspectives On Learning and Optimization

Abstract

Variational and Dynamical Perspectives On Learning and Optimization

by

Andre Yohannes Wibisono

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Michael Jordan, Chair

The problem of learning from data is prevalent in the modern scientific age, and optimization provides a natural mathematical language for describing learning problems. We study some problems in learning and optimization from variational and dynamical perspectives, by identifying the optimal structure in the problems and leveraging the parallel results between continuous and discrete-time problems.

We begin by studying the class of accelerated methods in optimization from a continuous-time perspective. We show that there is a Lagrangian functional that we call the *Bregman Lagrangian*, which generates a family of dynamics via the variational principle of least action, and these dynamics are related via speeding up time. Furthermore, we provide a systematic methodology for discretizing the dynamics into the family of accelerated higher-order algorithms with matching convergence rates in discrete time. Our work illuminates two classes of natural dynamics for optimization, the gradient and Lagrangian dynamics.

Next, we study the problem of approximate inference in graphical models. We analyze reweighted Kikuchi approximation for estimating the log partition function, which approximates the entropy in the variational representation with a region graph decomposition. We establish sufficient conditions for the concavity of the objective function in terms of weight assignments in the Kikuchi expansion, and characterize the polytope of concavity in terms of the cycle structure of the region graph. We also provide an algorithm to find the global optimum and simulations to demonstrate the advantages of the reweighted Kikuchi approach.

Finally, we study the problem of minimax option pricing as an online learning game between Nature and an Investor. Whereas the classical Black-Scholes model assumes the price fluctuates continuously following the geometric Brownian motion, we consider a worst-case model in which Nature chooses a sequence of price fluctuations under a cumulative quadratic volatility constraint, possibly with jumps, while the Investor makes a sequence of hedging decisions. We show that even in this adversarial, non-stochastic framework, the value of the game converges to the Black-Scholes option price, and the Black-Scholes hedging strategy is near-optimal for the Investor.

To bee and bun

Contents

List of Figures

Acknowledgments

First, I would like to thank my advisor Michael Jordan for being a great advisor, for encouraging me to freely explore my interests and always believing in me, even when I did not. Grad school has been a (long and hard) journey, but it also has been deeply transformative. I thank Mike for giving me the opportunity to explore the landscape of ideas in my mind and discover my own perspective.

I'm also very thankful to my family, Mama and Papa and Anton and Anita, for their constant support and love, even when we are far away. I miss you all.

I would like to especially thank Ashia Wilson for being my best friend and collaborator, and for sharing the amazing awe and wonder in this journey. I also thank my good friends who made Berkeley a fun experience even through the difficult times, especially Po-Ling and Purna. I also would like to thank my colleagues and collaborators, from whom I have learned a lot: Po-Ling Loh, Tamara Broderick, John Duchi, Nick Boyd, Jake Abernethy, Raf Frongillo, Chris Hillar, Shaowei Lin, Martin Wainwright, Peter Bartlett.

Thank you also to Meryl, Pierre, Fran, for being there and keeping me strong.

Finally, I thank Teddy and Ashia for the inexplicably wonderful journey that is Berkeley.

Chapter 1

Introduction

The problem of *learning from data* is prevalent and essential in the current era of Big Data. Recent advances in measurement techniques, as well as the increasingly significant parts of our lives that are conducted online, have resulted in the proliferation of massive, complex data sets. This leads to an exciting array of possibilities, because with this deluge of data comes—in principle—a wealth of information. Thus, for example, we now have attempts to decode the functionalities of brain regions from neuron connectivity patterns; to discover evidence of new particles from collision data in experimental physics; to decipher the social network structure of billions of people online; to extract contextual understanding from real-time sensory inputs to enable autonomous vehicles; to understand weather patterns at scale to predict climate change more accurately; etc. In particular, the field of machine learning (and data science more generally) has experienced a rapid growth in the development of both theoretical and computational tools in order to tackle the challenges introduced by the increasing amount of data.

In recent years, *optimization* has emerged as a natural language for learning problems. Indeed, to evaluate the quality of a learning strategy we typically compare its outputs or predictions against ground truth, and we measure these comparisons via a cost function. Thus, the problem of learning can be abstractly formulated as optimizing a certain objective function (which encodes the goals of learning as well as constraints from data), and learning strategies correspond to optimization algorithms. This formulation of learning as optimization has introduced a shift in research perspective. Whereas previously learning problems were studied via the lens of complexity theory (e.g., via the notions of Kolmogorov complexity, VC dimension [66], or the framework of PAC-learning [65]), the emphasis now is on the development of fast practical algorithms for optimization. The synthesis of optimization and machine learning has been a very active research area with rapid progress, resulting in the development of a wide array of algorithms for various problem settings that arise in practice along with some understanding of their theoretical guarantees (see, e.g., [59] for a review).

In this thesis, we study some problems in learning and optimization from *variational* and *dynamical* perspectives. Here, a *variational* perspective means we cast our problem instance as part of a larger optimization structure, and we choose our action (algorithm) as the optimal

answer to the variational problem. This provides a certificate of optimality for our action, and it also serves a guiding principle for designing algorithms. A *dynamical* perspective means we view our problem or algorithm in discrete time as a version (discretization) of a corresponding problem or dynamics in continuous time. The two time domains often have matching parallel structures, so we can use results from continuous time to guide the development in discrete time, or vice versa.

We elaborate further on these points below, which also serves as a brief summary and context for the remainder of this thesis. In particular, we demonstrate that the synthesis of both variational and dynamical perspectives provides a rich prism to understand the structure of learning and optimization problems.

1.1 On accelerated methods in optimization

In convex optimization, there is a phenomenon of *acceleration* in which we can boost the convergence rates of certain algorithms. This is first observed in the accelerated gradient descent algorithm proposed by Nesterov in 1983, and has since been extended to various settings. Accelerated methods achieve faster convergence rates than gradient methods and indeed, under certain conditions, they achieve optimal rates. However, accelerated methods are not descent methods and remain somewhat of a conceptual mystery. Furthermore, while many interpretations of Nesterov's acceleration technique have been proposed, it is not yet clear what is the natural scope of the acceleration concept.

In Chapter 2, we propose a variational, continuous-time framework for understanding accelerated methods. We show that there is a Lagrangian functional that we call the *Bregman Lagrangian* which generates a large family of second-order dynamics in continuous time via the principle of least action. We provide a systematic methodology for converting the dynamics in continuous time to accelerated higher-order methods in discrete time. Furthermore, we show that the continuous-time limit of all of these methods correspond to traveling the same curve in spacetime at different speeds. From this perspective, Nesterov's technique and many of its generalizations can be viewed as a systematic way to go from the continuous-time dynamics generated by the Bregman Lagrangian to a family of discrete-time algorithms. This chapter is adapted from the work published as [75].

Our work illuminates a new class of Lagrangian dynamics which may be useful for designing better algorithms for optimization. Furthermore, our work also provides an interesting perspective on the problem of optimization in continuous time. Whereas there is a clear complexity theory of optimization in discrete time governing the trade-off between convergence rates and problem assumptions (e.g., [44, 48]), it is a priori unclear if there is a similar theory of optimization in continuous time. Indeed, the naive definition of the optimization problem—to find dynamics that optimizes a function as fast as possible—suffers from the problem of *speeding-up time*. Namely, once we have a curve (dynamics) that works to minimize the function in continuous time, then we can speed it up to get any arbitrarily fast rate, so there is no notion of a "fastest" algorithm.

Nevertheless, our results in Chapter 2 give evidence that there is a structure of optimization in continuous time, which is interestingly tied to the matching structure of optimization in discrete time. Indeed, since our ultimate interest is in discrete-time algorithms (that we can implement in computers), our notion of continuous-time dynamics needs to be constrained by being *implementable* as a discrete-time algorithm with a matching convergence guarantee. From this perspective, our results demonstrate that we have two natural classes of dynamics for optimization. First, we have the class of gradient dynamics, which are first-order differential equations, and have simple variational interpretations as greedy steepest descent flows that are locally optimal in space. Second, we have the class of Bregman Lagrangian dynamics, which are second-order Euler-Lagrange equations, and have more complicated variational interpretations as the locally optimal curve in spacetime via the principle of least action. Both families have the nice property of being implementable in discrete time: Gradient dynamics give rise to the family of higher-order gradient methods, while Lagrangian dynamics give rise to the family of accelerated higher-order methods, both with matching convergence rates under increasingly strict assumptions. Furthermore, while the gradient dynamics are related by changing how we measure space, the Lagrangian dynamics are related by changing how we measure time. Thus, these two families have many nice properties for optimization, both in continuous and discrete time, and we conjecture that they are the optimal dynamics for optimization.

1.2 Kikuchi approximation for graphical models

When we have complex data with many covariates, we may wish to build a model to incorporate the structure of the data into our learning problem. Probabilistic graphical models are a familiar framework with diverse application domains including computer vision, statistical physics, coding theory, social science, and epidemiology. Graphical models capture the dependency structure of the covariates as a graph, where nodes represent variables and absent edges represent conditional independence. In this setting, the problem of learning from data becomes the problem of inferring some states or parameters of the joint distribution.

A crucial step in probabilistic inference is to compute the log partition function of the distribution based on the potential functions and the structure of the graph. However, computing the log partition function either exactly or approximately is NP-hard in general. An active area of research involves finding accurate approximations of the log partition function and characterizing the graph structures for which such approximations work well. A key technique is *variational inference*, which uses the variational representation of the log partition function as the dual function of the negative entropy, and replaces some terms in the variational problem with more tractable approximations [70]. In particular, the Kikuchi approximation method replaces the entropy term in the variational representation with an expression that decomposes with respect to a region graph. Kikuchi approximations were previously introduced in the physics literature and have been reformalized in the language of graphical models. The special case when the region graph has only two layers is known as

the Bethe approximation, which has been studied extensively, and there is a tree-reweighted version [69] that provides an upper bound on the true log partition function.

In Chapter 3, we analyze a reweighted version of the Kikuchi approximation, which generalizes the standard Kikuchi approximation by assigning arbitrary weights to individual terms in the Kikuchi entropy expansion. We establish necessary and sufficient conditions under which this class of objective functions is concave, so a global optimum may be found efficiently. Our theoretical results synthesize known results on Kikuchi and Bethe approximations, and our main theorem concerning concavity conditions for the reweighted Kikuchi entropy recovers existing results when specialized to the unweighted Kikuchi or reweighted Bethe case. Furthermore, we provide a valuable converse result in the reweighted Bethe case, showing that when our concavity conditions are violated, the entropy function cannot be concave over the whole feasible region. As demonstrated by our experiments, a message-passing algorithm designed to optimize the Kikuchi objective may terminate in local optima for weights outside the concave region. Furthermore, generating weight vectors in the Kikuchi region of concavity may yield closer approximations to the log partition function. In the reweighted Bethe setting, we also present a useful characterization of the concave region of the Bethe entropy function in terms of the geometry of the graph. This chapter is adapted from the work published as [39].

1.3 Minimax option pricing meets Black-Scholes

We now consider a popular problem in finance, option pricing, via the lens of online learning. An *option* is a contract that gives the right—but not the obligation—to buy an asset (e.g., a stock) for a given price on a given date. This allows firms to hedge against risk exposure, and the problem of option pricing is to determine the fair price for an option, namely, one that provides no opportunity for arbitrage (risk-free profit). This question is inherently difficult because while we know the asset's previous prices, we are uncertain as to its future price.

The classical model of Black-Scholes option pricing from 1973 assumes that the underlying asset's price fluctuates continuously following the geometric Brownian motion (GBM) stochastic process in continuous time, which means the firm should be able to buy and sell continuously until the option's expiration date. This is not true in practice, as the stock market is open only eight hours per day, and stock prices can make significant jumps even during regular trading. Nevertheless, the Black-Scholes model provides a very useful baseline model that allows fast and explicit calculations based on the properties of GBM.

In [1], we have studied the option pricing problem from the lens of regret minimization in online learning, by modeling option pricing as a sequential ("discrete-time") zero-sum game being played between an Investor, who is attempting to replicate the option payoff, and Nature, who is sequentially setting the price changes of the underlying asset. The value of this game is the *minimax option price*, since it is what the Investor should pay for the option against an adversarially chosen price path. Our main result in [1] was to show that the game value approaches the Black-Scholes option price as the Investor's trading frequency

increases (i.e., in the continuous-time limit). Thus, the minimax price tends to the option price under the GBM assumption; this result lends further credibility to the Black-Scholes model, as it suggests that the GBM assumption may already be a "worst-case model" in a certain sense.

In Chapter 4, we study a generalization of the minimax option pricing problem with weaker constraints. We consider a worst-case model, in which Nature chooses a sequence of price fluctuations under a *cumulative* quadratic volatility constraint, and allowing price jumps, while the Investor can make a sequence of hedging decisions to try to replicate the option payoff. The cumulative volatility constraint means the price path of Nature no longer converges to GBM. However, we show that the value of the proposed game, which is the *regret* of the hedging strategy, still converges to the Black-Scholes option price. Furthermore, we show that the Black-Scholes hedging strategy is near-optimal for the Investor, even in this non-stochastic framework. This chapter is adapted from the work published as [2].

Chapter 2

A Variational Perspective on Accelerated Methods in Optimization

Optimization lies at the core of many fields concerned with data analysis. It provides a mathematical language in which both computational and statistical concepts can be expressed and it delivers practical data analysis algorithms that can scale to the enormous data sets that are increasingly the norm in science and technology. The recent literature on data analysis and optimization has focused on gradient-based optimization methods, given their low per-iteration cost and the relative ease with which they can be deployed on parallel and distributed processing architectures. Establishing that such methods do indeed address the scalability problems inherent in large-scale data analysis raises fundamental questions concerning the convergence rate of gradient-based methods, the extent to which those rates can be increased systematically and whether there are upper bounds on achievable rates.

In the body of theory and practice built up to answer such questions, the phenomenon of *acceleration* plays a key role. In 1983, Nesterov introduced acceleration in the context of gradient descent for convex functions [45], showing that it achieves an improved convergence rate with respect to gradient descent, and moreover that it achieves an optimal convergence rate under an oracle model of optimization complexity [44]. The acceleration idea has since been extended to a wide range of other settings, including composite optimization [47, 64, 11], stochastic optimization [25, 34], nonconvex optimization [20, 37], and conic programming [35]. There have been generalizations to non Euclidean optimization [49, 33] and higher-order algorithms [46, 9], and there have been numerous applications that further extend the reach of the idea [27, 28, 29, 43].

Despite this compelling evidence of the value of the idea of acceleration, it remains something of a conceptual mystery. Derivations of accelerated methods do not flow from a single underlying principle, but tend to rely on case-specific algebra [30]. The basic Nesterov technique is often explained intuitively in terms of momentum, but this intuition does not easily carry over to non-Euclidean settings [4]. In recent years, the number of explanations and interpretations of acceleration has increased [4, 6, 18, 36, 14], but these explanations have been focused on restrictive instances of acceleration, such as first-order algorithms, the

Euclidean setting, or cases in which the objective function is strongly convex or quadratic. It is not yet clear what the natural scope of the acceleration concept is and indeed whether it is a single phenomenon.

In this chapter we study acceleration from a continuous-time, variational point of view. We build on recent work by [61], who show that the continuous-time limit of Nesterov's accelerated gradient descent is a second-order differential equation, and we take inspiration from continuous time analysis of mirror descent [44]. In our approach, rather than starting from existing discrete-time accelerated gradient methods and deriving differential equations by taking limits, we take as our point of departure a variational formulation in which we define a functional on continuous-time curves that we refer to as a *Bregman Lagrangian*. Next, we calculate and discretize the Euler-Lagrange equation corresponding to the Bregman Lagrangian. It turns out that naive discretization (the Euler method) does not yield a stable discrete-time algorithm that retains the rate of the underlying differential equation; rather, a more elaborate discretization involving an auxiliary sequence is necessary. This auxiliary sequence is essentially that used by Nesterov in his constructions of accelerated mirror descent [49] and accelerated cubic-regularized Newton's method [46], and later generalized by Baes [9]. Thus, from our perspective, Nesterov's approach can be viewed as a methodology for the discretization of a certain class of differential equations. Given the complexities associated with the discretization of differential equations, it is perhaps not surprising that it has been difficult to perceive the generality and scope of the acceleration concept in a discrete-time framework.

The Bregman-Lagrangian framework permits a systematic understanding of the matching rates associated with higher-order gradient methods in discrete and continuous time. In the case of gradient descent, Su et al. show that the discrete and continuous-time dynamics have convergence rates of $O(1/(\epsilon k))$ and $O(1/t)$, respectively, and that these match using the identification $t = \epsilon k$; for accelerated gradient descent, the convergence rates are $O(1/(\epsilon k^2))$ and $O(1/t^2)$ respectively, which match using the identification $t = \sqrt{\epsilon} k$ [61]. This result has been extended to the non-Euclidean case by Krichene et al. [33]. Higher-order gradient descent is a descent method which minimizes a regularized $(p-1)$-st order Taylor approximation of the objective function f, generalizing gradient descent ($p = 2$) and Nesterov and Polyak's cubic-regularized Newton's method ($p = 3$) [50]. The p-th order gradient algorithm with a constant step size ϵ has convergence rate $O(1/(\epsilon k^{p-1}))$ when $\nabla^{p-1} f$ is $(1/\epsilon)$-Lipchitz and, in continuous time, as $\epsilon \to 0$, this algorithm corresponds to the p-th *rescaled gradient flow*, which is a first-order differential equation with a matching convergence rate $O(1/t^{p-1})$. Thus, the p-th order gradient algorithm can be seen as a discretization $t = \delta k$ of the rescaled gradient flow with time step $\delta = \epsilon^{1/(p-1)}$. Similarly, we show that the accelerated higher-order gradient algorithm achieves an improved convergence rate $O(1/(\epsilon k^p))$ under the same assumption (i.e., $\nabla^{p-1} f$ is $(1/\epsilon)$-Lipschitz). In continuous time, as $\epsilon \to 0$, this corresponds to the second-order Euler-Lagrange curve of the Bregman Lagrangian with a matching convergence rate $O(1/t^p)$. Thus, the p-th order accelerated algorithm can be seen as a discretization $t = \delta k$ of the Euler-Lagrange equation of the Bregman Lagrangian with time step $\delta = \epsilon^{1/p}$.

In addition to its value in relating continuous-time and discrete-time acceleration, the

study of the Bregman Lagrangian can provide further insights into the nature of acceleration. For instance, it is noteworthy that the Bregman Lagrangian is closed under time dilation. This means that if we take an Euler-Lagrange curve of a Bregman Lagrangian and reparameterize time so we travel the curve at a different speed, then the resulting curve is also the Euler-Lagrange curve of another Bregman Lagrangian, with appropriately modified parameters. Thus, the entire family of accelerated methods correspond to a single curve in spacetime and can be obtained by speeding up (or slowing down) any single curve. Another insight is obtained by noting that from the discrete-time point of view, an interpretation of acceleration starts with a base algorithm, which we can accelerate by coupling with a suitably weighted mirror descent step. From the continuous-time point of view, however, it is the weighted mirror descent step that is important since the base gradient algorithm operates on a smaller time scale. Thus, Nesterov's accelerated gradient methods are but one possible implementation of second-order Bregman-Lagrangian curves as a discrete-time algorithm.

The remainder of the chapter is organized as follows. In Section 2.1, we introduce the general family of Bregman Lagrangians and study its properties. In Section 2.2, we demonstrate how to discretize the Euler-Lagrange equations corresponding to the polynomial subfamily of Bregman Lagrangians to obtain discrete-time accelerated algorithms. In particular, we introduce the family of higher-order gradient methods which can be used to complete the discretization. In Section 2.3, we discuss additional properties of the Bregman Lagrangian, including gauge-invariance properties, connection to classical gradient flows, and the correspondence with a functional that we refer to as a Bregman Hamiltonian. Finally, we end in Section 2.4 with a brief discussion.

Problem setting

We consider the optimization problem

$$\min_{x \in \mathcal{X}} \; f(x),$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set and $f \colon \mathcal{X} \to \mathbb{R}$ is a continuously differentiable convex function. To simplify the presentation in this chapter we focus on the case $\mathcal{X} = \mathbb{R}^d$. We also assume f has a unique minimizer, $x^* \in \mathcal{X}$, satisfying the optimality condition $\nabla f(x^*) = 0$. We use the inner product norm $\|x\| = \langle x, x \rangle^{1/2}$.

We consider the general non-Euclidean setting in which the space $\mathcal{X}$ is endowed with a distance-generating function $h \colon \mathcal{X} \to \mathbb{R}$ that is convex and essentially smooth (i.e., h is continuously differentiable in $\mathcal{X}$, and $\|\nabla h(x)\|_* \to \infty$ as $\|x\| \to \infty$). The function h can also be used to define an alternative measure of distance in $\mathcal{X}$ via its Bregman divergence:

$$D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle,$$

which is nonnegative since h is convex. When x and y are nearby the Bregman divergence is an approximation to the Hessian metric,

$$D_h(y,x) \approx \frac{1}{2}\langle y-x, \nabla^2 h(x)(y-x)\rangle := \frac{1}{2}\|y-x\|^2_{\nabla^2 h(x)}.$$

The *Euclidean setting* is obtained when $h(x) = \frac{1}{2}\|x\|^2$, in which case the Bregman divergence and Hessian metric coincide since $\nabla^2 h(x)$ is the identity matrix.

In continuous time, the Hessian metric is generally studied rather than the more general Bregman divergence; this is the case, for instance, in the case of natural gradient flow, which is the continuous-time limit of mirror descent [5, 54]. By way of contrast, we shall see that our continuous-time, Lagrangian framework crucially employs the Bregman divergence.

In this chapter we denote a discrete-time sequence in lower case, e.g., x_k with $k \geq 0$ an integer. We denote a continuous-time curve in upper case, e.g., X_t with $t \in \mathbb{R}$. An over-dot means derivative with respect to time, i.e., $\dot{X}_t = \frac{d}{dt}X_t$.

2.1 The Bregman Lagrangian

We define the *Bregman Lagrangian*

$$\mathcal{L}(x,\dot{x},t) = e^{\alpha_t+\gamma_t}\left(D_h(x+e^{-\alpha_t}\dot{x},x) - e^{\beta_t}f(x)\right) \tag{2.1}$$

which is a function of position $x \in \mathcal{X}$, velocity $\dot{x} \in \mathbb{R}^d$, and time $t \in \mathbb{T}$, where $\mathbb{T} \subseteq \mathbb{R}$ is an interval of time. The functions $\alpha, \beta, \gamma \colon \mathbb{T} \to \mathbb{R}$ are arbitrary smooth (continuously differentiable) functions of time that determine the weighting of the velocity, the potential function, and the overall damping of the Lagrangian. We also define the following *ideal scaling conditions*:

$$\dot{\beta}_t \leq e^{\alpha_t} \tag{2.2a}$$

$$\dot{\gamma}_t = e^{\alpha_t}; \tag{2.2b}$$

these conditions will be justified in the following section.

Convergence rates of the Euler-Lagrange equation

In this section we show that—under the ideal scaling assumption (2.2)—the Bregman Lagrangian (2.1) defines a variational problem the solutions to which minimize the objective function f at an exponential rate.

Given a general Lagrangian $\mathcal{L}(x,\dot{x},t)$, we define a functional on curves $\{X_t : t \in \mathbb{T}\}$ via integration of the Lagrangian: $J(X) = \int_{\mathbb{T}} \mathcal{L}(X_t, \dot{X}_t, t)dt$. From the calculus of variations, a necessary condition for a curve to minimize this functional is that it solve the Euler-Lagrange equation:

$$\frac{d}{dt}\left\{\frac{\partial \mathcal{L}}{\partial \dot{x}}(X_t,\dot{X}_t,t)\right\} = \frac{\partial \mathcal{L}}{\partial x}(X_t,\dot{X}_t,t). \tag{2.3}$$

Specifically, for the Bregman Lagrangian (2.1), the partial derivatives are

$$\frac{\partial \mathcal{L}}{\partial x}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(\nabla h(x + e^{-\alpha_t} \dot{x}) - \nabla h(x) - e^{-\alpha_t} \nabla^2 h(x)\, \dot{x} - e^{\beta_t} \nabla f(x) \right) \tag{2.4a}$$

$$\frac{\partial \mathcal{L}}{\partial \dot{x}}(x, \dot{x}, t) = e^{\gamma_t} \left(\nabla h(x + e^{-\alpha_t} \dot{x}) - \nabla h(x) \right). \tag{2.4b}$$

Thus, for general functions $\alpha_t, \beta_t, \gamma_t$, the Euler-Lagrange equation (2.3) for the Bregman Lagrangian (2.1) is a second-order differential equation given by

$$\begin{aligned} &\ddot{X}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t} \left[\nabla^2 h(X_t + e^{-\alpha_t} \dot{X}_t) \right]^{-1} \nabla f(X_t) \\ &\quad + e^{\alpha_t}(\dot{\gamma}_t - e^{\alpha_t}) \left[\nabla^2 h(X_t + e^{-\alpha_t} \dot{X}_t) \right]^{-1} (\nabla h(X_t + e^{-\alpha_t} \dot{X}_t) - \nabla h(X_t)) = 0. \end{aligned} \tag{2.5}$$

We now impose the ideal scaling condition (2.2b). In this case the last term in (2.5) vanishes, so the Euler-Lagrange equation simplifies to

$$\ddot{X}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t} \left[\nabla^2 h(X_t + e^{-\alpha_t} \dot{X}_t) \right]^{-1} \nabla f(X_t) = 0. \tag{2.6}$$

In (2.6), we have assumed the Hessian matrix $\nabla^2 h(X_t + e^{-\alpha_t} \dot{X}_t)$ is invertible. But we can also write the equation (2.6) in the following way, which only requires that ∇h be differentiable,

$$\frac{d}{dt} \nabla h(X_t + e^{-\alpha_t} \dot{X}_t) = -e^{\alpha_t + \beta_t} \nabla f(X_t). \tag{2.7}$$

To establish a convergence rate associated with solutions to the Euler-Lagrange equation—under the ideal scaling conditions—we take a Lyapunov function approach. Defining the following energy functional:

$$\mathcal{E}_t = D_h \left(x^*,\, X_t + e^{-\alpha_t} \dot{X}_t \right) + e^{\beta_t} (f(X_t) - f(x^*)), \tag{2.8}$$

we immediately obtain a convergence rate, as shown in the following theorem.

Theorem 2.1. *If the ideal scaling* (2.2) *holds, then solutions to the Euler-Lagrange equation* (2.7) *satisfy*

$$f(X_t) - f(x^*) \leq O(e^{-\beta_t}).$$

Proof. The time derivative of the energy functional is

$$\dot{\mathcal{E}}_t = -\left\langle \frac{d}{dt} \nabla h(X_t + e^{-\alpha_t} \dot{X}_t),\, x^* - X_t - e^{-\alpha_t} \dot{X}_t \right\rangle + \dot{\beta}_t e^{\beta_t} (f(X_t) - f(x^*)) + e^{\beta_t} \langle \nabla f(X_t), \dot{X}_t \rangle.$$

If X_t satisfies the Euler-Lagrange equation (2.7), then the time derivative simplifies to

$$\dot{\mathcal{E}}_t = -e^{\alpha_t + \beta_t} D_f(x^*, X_t) + (\dot{\beta}_t - e^{\alpha_t}) e^{\beta_t} (f(X_t) - f(x^*))$$

where $D_f(x^*, X_t) = f(x^*) - f(X_t) - \langle \nabla f(X_t), x^* - X_t \rangle$ is the Bregman divergence of f. Note that $D_f(x^*, X_t) \geq 0$ since f is convex, so the first term in $\dot{\mathcal{E}}_t$ is nonpositive. Furthermore, if the ideal scaling condition (2.2a) holds, then the second term is also nonpositive, so $\dot{\mathcal{E}}_t \leq 0$. Since $D_h(x^*, X_t + e^{-\alpha_t}\dot{X}_t) \geq 0$, this implies that for any $t \geq t_0 \in \mathbb{T}$, $e^{\beta_t}(f(X_t) - f(x^*)) \leq \mathcal{E}_t \leq \mathcal{E}_{t_0}$. Thus, $f(X_t) - f(x^*) \leq \mathcal{E}_{t_0} e^{-\beta_t} = O(e^{-\beta_t})$, as desired. □

For a given α_t, which determines γ_t by (2.2a), the optimal convergence rate is achieved by setting $\dot{\beta}_t = e^{\alpha_t}$, resulting in convergence rate $O(e^{-\beta_t}) = O(\exp(-\int_{t_0}^{t} e^{\alpha_s}\, ds))$. In Section 2.2 we study a subfamily of Bregman Lagrangians that have a polynomial convergence rate, and we show how we can discretize the resulting Euler-Lagrange equations to obtain discrete-time methods that have a matching, accelerated convergence rate. In Section 2.3 we study another subfamily of Bregman Lagrangians that have an exponential convergence rate, and discuss its connection to a generalization of Nesterov's restart scheme. In the Euclidean setting, our derivations simplify. We present these derivations at the end of Section 2.12, and comment on the insight that they provide into the question posed by Su et al. [61] on the significance of the value 3 in the damping coefficient for Nesterov's accelerated gradient descent.

Time dilation

A notable property of the Bregman Lagrangian family is that it is closed under time dilation. This means if we take the Euler-Lagrange equation (2.5) of the Bregman Lagrangian (2.1) and reparameterize time to travel the curve at a different speed, the resulting curve is also the Euler-Lagrange equation of a Bregman Lagrangian with a suitably modified set of parameters.

Concretely, let $\tau \colon \mathbb{T} \to \mathbb{T}'$ be a smooth (twice-continuously differentiable) increasing function, where $\mathbb{T}' = \tau(\mathbb{T}) \subseteq \mathbb{R}$ is the image of $\mathbb{T}$. Given a curve $X \colon \mathbb{T}' \to \mathcal{X}$, we consider the reparameterized curve $Y \colon \mathbb{T} \to \mathcal{X}$ defined by

$$Y_t = X_{\tau(t)}. \tag{2.9}$$

That is, the new curve Y is obtained by traversing the original curve X at a new speed of time determined by τ. If $\tau(t) > t$, then we say that Y is the *sped-up version* of X, because the curve Y at time t has the same value as the original curve X at the future time $\tau(t)$.

For clarity, we let $\mathcal{L}_{\alpha,\beta,\gamma}$ denote the Bregman Lagrangian (2.1) parameterized by α, β, γ. Then we have the following result whose proof is provided in Section 2.5.

Theorem 2.2. *If X_t satisfies the Euler-Lagrange equation* (2.5) *for the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$, then the reparameterized curve $Y_t = X_{\tau(t)}$ satisfies the Euler-Lagrange equation for the Bregman Lagrangian $\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}$, with modified parameters*

$$\tilde{\alpha}_t = \alpha_{\tau(t)} + \log \dot{\tau}(t) \tag{2.10a}$$

$$\tilde{\beta}_t = \beta_{\tau(t)} \tag{2.10b}$$

$$\tilde{\gamma}_t = \gamma_{\tau(t)}. \tag{2.10c}$$

Furthermore, α, β, γ satisfy the ideal scaling (2.2) if and only if $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ do.

We note that in general, when we reparameterize time by a time-dilation function $\tau(t)$, the Lagrangian functional transforms to $\tilde{\mathcal{L}}(x, \dot{x}, t) = \dot{\tau}(t)\, \mathcal{L}\left(x, \frac{1}{\dot{\tau}(t)}\dot{x}, \tau(t)\right)$. Thus, another way of stating the result in Theorem 2.2 is to claim that

$$\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}(x, \dot{x}, t) = \dot{\tau}(t)\, \mathcal{L}_{\alpha,\beta,\gamma}\left(x, \frac{1}{\dot{\tau}(t)}\dot{x}, \tau(t)\right), \tag{2.11}$$

which we can easily verify by directly substituting the definition of the Lagrangian (2.1) and the modified parameters $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ (2.10).

In Section 2.2, we show that the Bregman Lagrangian generates the family of higher-order accelerated methods in discrete time. Thus, the time-dilation property means that the entire family of curves for accelerated methods in continuous time corresponds to a single curve in spacetime, which is traveled at different speeds. This suggests that the underlying solution curve has a more fundamental structure that is worth exploring further.

2.2 Polynomial convergence rates and accelerated methods

In this section, we study a subfamily of Bregman Lagrangians (2.1) with the following choice of parameters, indexed by a parameter $p > 0$,

$$\alpha_t = \log p - \log t \tag{2.12a}$$
$$\beta_t = p \log t + \log C \tag{2.12b}$$
$$\gamma_t = p \log t, \tag{2.12c}$$

where $C > 0$ is a constant. The parameters α, β, γ satisfy the ideal scaling condition (2.2) (with an equality on the first condition (2.2a)). The Bregman Lagrangian (2.1) becomes

$$\mathcal{L}(x, \dot{x}, t) = p t^{p-1}\left(D_h\left(x + \frac{t}{p}\dot{x}, x\right) - C t^{p-2} f(x)\right). \tag{2.13}$$

Its Euler-Lagrange equation (2.6) is given by

$$\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + C p^2 t^{p-2}\left[\nabla^2 h\left(X_t + \frac{t}{p}\dot{X}_t\right)\right]^{-1} \nabla f(X_t) = 0 \tag{2.14}$$

and, by Theorem 2.1, it has an $O(1/t^p)$ rate of convergence. As direct result of the time-dilation property (Theorem 2.2), the entire family of curves (2.14) can be obtained by speeding up the curve in the case $p = 2$ by the time-dilation function $\tau(t) = t^{p/2}$. In Section 2.5 we discuss the issue of the existence and uniqueness of the solution to the differential equation (2.14).

The case $p = 2$ of the equation (2.14) is the continuous-time limit of Nesterov's accelerated mirror descent [49], and the case $p = 3$ is the continuous-time limit of Nesterov's accelerated cubic-regularized Newton's method [46]. The case $p = 2$ has also been derived independently in a recent work of Krichene et al. [33]; in the Euclidean case, when the Hessian $\nabla^2 h$ is the identity matrix, we recover the differential equation of Su et al. [61].

Naive discretization

We now turn to the challenge of discretizing the differential equation in (2.14), with the goal of obtaining a discrete-time algorithm whose convergence rate matches that of the underlying differential equation. As we show in this section, a naive Euler method is not able to match the underlying rate. To match the rate a more sophisticated approach is needed, and it is at this juncture that Nesterov's three-sequence idea makes its appearance.

We first write the second-order equation (2.14) as the following system of first-order equations:

$$Z_t = X_t + \frac{t}{p}\dot{X}_t \tag{2.15a}$$

$$\frac{d}{dt}\nabla h(Z_t) = -Cpt^{p-1}\nabla f(X_t). \tag{2.15b}$$

Now we discretize X_t and Z_t into sequences x_k and z_k with time step $\delta > 0$. That is, we make the identification $t = \delta k$ and set $x_k = X_t$, $x_{k+1} = X_{t+\delta} \approx X_t + \delta\dot{X}_t$ and $z_k = Z_t$, $z_{k+1} = Z_{t+\delta} \approx Z_t + \delta\dot{Z}_t$. Applying the forward-Euler method to (2.15a) gives the equation $z_k = x_k + \frac{\delta k}{p}\frac{1}{\delta}(x_{k+1} - x_k)$, or equivalently,

$$x_{k+1} = \frac{p}{k}z_k + \frac{k-p}{k}x_k. \tag{2.16}$$

Similarly, applying the backward-Euler method to equation (2.15b) gives $\frac{1}{\delta}(\nabla h(z_k) - \nabla h(z_{k-1})) = -Cp(\delta k)^{p-1}\nabla f(x_k)$, which we can write as the optimality condition of the following weighted mirror descent step:

$$z_k = \arg\min_z \left\{ Cpk^{p-1}\langle \nabla f(x_k), z\rangle + \frac{1}{\epsilon}D_h(z, z_{k-1}) \right\}, \tag{2.17}$$

with step size $\epsilon = \delta^p$. In principle, the two updates (2.16), (2.17) define an algorithm that implements the dynamics (2.15) in discrete time. However, we cannot establish a convergence rate for the algorithm (2.16), (2.17); indeed, empirically, we find that the algorithm is unstable. Even for the simple case in which f is a quadratic function in two dimensions, the iterates of the algorithm initially approach and oscillate near the minimizer, but eventually the oscillation increases and the iterates shoot off to infinity.

A rate-matching discretization

We now discuss how to modify the naive discretization scheme (2.16), (2.17) into an algorithm whose rate matches that of the underlying differential equation. Our approach is inspired by Nesterov's constructions of accelerated mirror descent [49] and accelerated cubic-regularized Newton's method [46], which maintain three sequences in the algorithms and use the estimate sequence technique to prove convergence. Indeed, from our point of view, Nesterov's methodology can be viewed as a rate-matching discretization methodology.

Specifically, we consider the following scheme, in which we introduce a third sequence y_k to replace x_k in the updates,

$$x_{k+1} = \frac{p}{k+p} z_k + \frac{k}{k+p} y_k \tag{2.18a}$$

$$z_k = \arg\min_z \left\{ Cpk^{(p-1)} \langle \nabla f(y_k), z \rangle + \frac{1}{\epsilon} D_h(z, z_{k-1}) \right\}, \tag{2.18b}$$

where $k^{(p-1)} := k(k+1)\cdots(k+p-2)$ is the rising factorial. A sufficient condition for the algorithm (2.18) to have an $O(1/(\epsilon k^p))$ convergence rate is that the new sequence y_k satisfy the inequality

$$\langle \nabla f(y_k), x_k - y_k \rangle \geq M \epsilon^{\frac{1}{p-1}} \|\nabla f(y_k)\|_*^{\frac{p}{p-1}}, \tag{2.19}$$

for some constant $M > 0$. Note that in going from (2.16) to (2.18a) we have replaced the weight $\frac{p}{k}$ by $\frac{p}{k+p}$; this is only for convenience in the proof given below, and does not change the asymptotics since $\frac{p}{k} = \Theta(\frac{p}{k+p})$ as $k \to \infty$. Similarly, we replace k^{p-1} in (2.17) by the rising factorial $k^{(p-1)}$ in (2.18b) to make the algebra easier, but we still have $k^{(p-1)} = \Theta(k^{p-1})$.

The following result also requires a uniform convexity assumption on the distance-generating function h. Recall that h is *σ-uniformly convex of order $p \geq 2$* if its Bregman divergence is lower bounded by the p-th power of the norm,

$$D_h(y, x) \geq \frac{\sigma}{p} \|y - x\|^p. \tag{2.20}$$

The case $p = 2$ is the usual definition of strong convexity. An example of a uniformly convex function is the p-th power of the norm, $h(x) = \frac{1}{p}\|x - w\|^p$ for any $w \in \mathcal{X}$, which is σ-uniformly convex of order p with $\sigma = 2^{-p+2}$ [46, Lemma 4].

Theorem 2.3. *Assume h is 1-uniformly convex of order $p \geq 2$, and the sequence y_k satisfies the inequality (2.19) for all $k \geq 0$. Then the algorithm (2.18) with the constant $C \leq M^{p-1}/p^p$ and initial condition $z_0 = x_0 \in \mathcal{X}$ has the convergence rate*

$$f(y_k) - f(x^*) \leq \frac{D_h(x^*, x_0)}{C\epsilon k^{(p)}} = O\left(\frac{1}{\epsilon k^p}\right). \tag{2.21}$$

The proof of Theorem 2.3 uses a generalization of Nesterov's estimate sequence technique, and can be found in Section 2.5. We note that with the scaling $\epsilon = \delta^p$ as in the previous section, the convergence rate $O(1/(\epsilon k^p))$ matches the $O(1/t^p)$ rate in continuous time for the differential equation (2.14). We also note that the result in Theorem 2.3 does not require any assumptions on f beyond the ability to construct a sequence y_k satisfying (2.19). In the next section, we will see that we can satisfy (2.19) using the higher-order gradient method, which requires a higher-order smoothness assumption on f; the resulting algorithm is then the accelerated higher-order gradient method.

Higher-order gradient method

We study the higher-order gradient update, which minimizes a regularized higher-order Taylor approximation of the objective function f.

Recall that for an integer $p \geq 2$, the $(p-1)$-st order Taylor approximation of f centered at $x \in \mathcal{X}$ is the $(p-1)$-st degree polynomial

$$\begin{aligned} f_{p-1}(y; x) &= \sum_{i=0}^{p-1} \frac{1}{i!} \nabla^i f(x)(y-x)^i \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \cdots + \frac{1}{(p-1)!} \nabla^{p-1} f(x)(y-x)^{p-1}. \end{aligned}$$

We say that f is *L-smooth of order $p-1$* if f is p-times continuously differentiable and $\nabla^{p-1} f$ is L-Lipschitz, which means for all $x, y \in \mathcal{X}$,

$$\|\nabla^{p-1} f(y) - \nabla^{p-1} f(x)\|_* \leq L \|y - x\|. \tag{2.22}$$

For a constant $N > 0$ and step size $\epsilon > 0$, we define the update operator $G_{p,\epsilon,N} \colon \mathcal{X} \to \mathcal{X}$ by

$$G_{p,\epsilon,N}(x) = \arg\min_y \left\{ f_{p-1}(y; x) + \frac{N}{\epsilon p} \|y - x\|^p \right\}. \tag{2.23}$$

When f is smooth of order $p-1$, the operator $G_{p,\epsilon,N}$ has the following property, which generalizes [46, Lemma 6]. We provide the proof in Section 2.5.

Lemma 2.4. *Let $x \in \mathcal{X}$, $y = G_{p,\epsilon,N}(x)$, and $N > 1$. If f is $L = \frac{(p-1)!}{\epsilon}$-smooth of order $p-1$, then*

$$\langle \nabla f(y), x - y \rangle \geq \frac{(N^2-1)^{\frac{p-2}{2p-2}}}{2N} \epsilon^{\frac{1}{p-1}} \|\nabla f(y)\|_*^{\frac{p}{p-1}}. \tag{2.24}$$

Furthermore,

$$\frac{(N^2-1)^{\frac{p-2}{2p-2}}}{2N} \epsilon^{\frac{1}{p-1}} \|\nabla f(y)\|_*^{\frac{1}{p-1}} \leq \|x - y\| \leq \frac{1}{(N-1)^{\frac{1}{p-1}}} \epsilon^{\frac{1}{p-1}} \|\nabla f(y)\|_*^{\frac{1}{p-1}}. \tag{2.25}$$

The inequality (2.24) means that we can use the update operator $G_{p,\epsilon,N}$ to produce a sequence y_k satisfying the requirement (2.19) under a higher-order order smoothness condition on f. We state the resulting algorithm in the next section.

Higher-order gradient method. In this section, we study the following higher-order gradient algorithm defined by the update operator $G_{p,\epsilon,N}$:

$$x_{k+1} = G_{p,\epsilon,N}(x_k). \tag{2.26}$$

The case $p = 2$ is the usual gradient descent algorithm, and the case $p = 3$ is Nesterov and Polyak's cubic-regularized Newton's method [50].

If f is smooth of order $p-1$, then the algorithm (2.26) is a descent method. Furthermore, we can prove the following rate of convergence, which generalizes the results for gradient descent and the cubic-regularized Newton's method. We provide the proof in Section 2.5.

Theorem 2.5. *If f is $\frac{(p-1)!}{\epsilon}$-smooth of order $p-1$, then the algorithm* (2.26) *with constant $N > 0$ and initial condition $x_0 \in \mathcal{X}$ has the convergence rate*

$$f(x_k) - f(x^*) \leq \frac{p^{p-1}(N+1)R^p}{\epsilon k^{p-1}} = O\left(\frac{1}{\epsilon k^{p-1}}\right), \tag{2.27}$$

where $R = \sup_{x:\, f(x)\leq f(x_0)} \|x - x^*\|$ is the radius of the level set of f from the initial point x_0.

Rescaled gradient flow. We can take the continuous-time limit of the higher-order gradient algorithm as the step size $\epsilon \to 0$. The resulting curve is a first-order differential equation that is a rescaled version of gradient flow. We show that it minimizes f with a matching convergence rate. In the following, we take $N = 1$ in (2.26) for simplicity (the general N simply scales the vector field by a constant). We provide the proof of Theorem 2.6 in Section 2.5.

Theorem 2.6. *The continuous-time limit of the algorithm* (2.26) *is the rescaled gradient flow*

$$\dot{X}_t = -\frac{\nabla f(X_t)}{\|\nabla f(X_t)\|_*^{\frac{p-2}{p-1}}}, \tag{2.28}$$

where we define the right-hand side to be the zero if $\nabla f(X_t) = 0$. Furthermore, the rescaled gradient flow has convergence rate

$$f(X_t) - f(x^*) \leq \frac{(p-1)^{p-1}R^p}{t^{p-1}} = O\left(\frac{1}{t^{p-1}}\right), \tag{2.29}$$

where $R = \sup_{x:\, f(x)\leq f(X_0)} \|x - x^*\|$ is the radius of the level set of f from the initial point X_0.

Equivalently, we can interpret the higher-order gradient algorithm (2.26) as a discretization of the rescaled gradient flow (2.28) with time step $\delta = \epsilon^{\frac{1}{p-1}}$, so $t = \delta k = \epsilon^{\frac{1}{p-1}} k$. With this identification, the convergence rates in discrete time, $O(1/(\epsilon k^{p-1}))$, and in continuous time, $O(1/t^{p-1})$, match. The convergence rate for the continuous-time dynamics does not require any assumption beyond the convexity and differentiability of f (as in the case of the Lagrangian flow (2.6)), whereas the convergence rate for the discrete-time algorithm requires the higher-order smoothness assumption on f. We note that the limiting case $p \to \infty$ of (2.28) is the *normalized gradient flow*, which has been shown to converge to the minimizer of f in finite time [15]. We also note that unlike the Lagrangian flow, the family of rescaled gradient flows is *not* closed under time dilation.

Accelerated higher-order gradient method

By the result of Lemma 2.4, we see that we can use the higher-order gradient update $G_{p,\epsilon,N}$ to produce a sequence y_k satisfying the inequality (2.19), to complete the algorithm (2.26) that implements the polynomial family of the Bregman-Lagrangian flow (2.14). Explicitly, the resulting algorithm is as follows,

$$x_{k+1} = \frac{p}{k+p} z_k + \frac{k}{k+p} y_k \tag{2.30a}$$

$$y_k = \arg\min_y \left\{ f_{p-1}(y; x_k) + \frac{N}{\epsilon p} \|y - x_k\|^p \right\} \tag{2.30b}$$

$$z_k = \arg\min_z \left\{ Cpk^{(p-1)} \langle \nabla f(y_k), z \rangle + \frac{1}{\epsilon} D_h(z, z_{k-1}) \right\}. \tag{2.30c}$$

By Theorem 2.3 and Lemma 2.4, we have the following guarantee for this algorithm.

Corollary 2.7. *Assume f is $\frac{(p-1)!}{\epsilon}$-smooth of order $p-1$, and h is 1-uniformly convex of order p. Then the algorithm (2.30) with constants $N > 1$ and $C \leq (N^2-1)^{\frac{p-2}{2}}/((2N)^{p-1}p^p)$ and initial conditions $z_0 = x_0 \in \mathcal{X}$ has an $O(1/(\epsilon k^p))$ convergence rate.*

The resulting algorithm (2.30) and its convergence rate recovers the results of Baes [9], who studied a generalization of Nesterov's estimate sequence technique to higher-order algorithms. We note that the convergence rate $O(1/(\epsilon k^p))$ of algorithm (2.30) is better than the $O(1/(\epsilon k^{p-1}))$ rate of the higher-order gradient algorithm (2.26), under the same assumption of the $(p-1)$-st order smoothness of f. This gives the interpretation of the algorithm (2.30) as "accelerating" the higher-order gradient method. Indeed, in this view the "base algorithm" that we start with is the higher-order gradient algorithm in the y-sequence (2.30b), and the acceleration is obtained by coupling it with a suitably weighted mirror descent step in (2.30a) and (2.30c).

However, from the continuous-time point of view, where our starting point is the polynomial Lagrangian flow (2.14), we see that the algorithm (2.30) is only one possible implementation of the flow as a discrete-time algorithm. As we saw previously, it is only the

x- and z-sequences (2.30a) and (2.30c) that play a role in the correspondence between the continuous-time dynamics and its discrete-time implementation, and the requirement (2.19) in the y-update is only needed to complete the convergence proof. Indeed, the higher-order gradient update (2.30b) does not change the continuous-time limit, since from (2.25) in Lemma 2.4 we have that $\|x_k - y_k\| = \Theta(\epsilon^{\frac{1}{p-1}})$, which is smaller than the $\delta = \epsilon^{\frac{1}{p}}$ time step in the discretization of (2.14). Therefore, the x and y sequences in (2.30) coincide in continuous time as $\epsilon \to 0$. Thus, from this point of view, Nesterov's accelerated methods (for the cases $p = 2$ and $p = 3$) are one of possibly many discretizations of the polynomial Lagrangian flow (2.14). For instance, in the case $p = 2$, Krichene et al. [33, Section 4.1] show that we can use a general regularizer in the gradient step (2.30b) under some additional smoothness assumptions. If there are other implementations, it would be interesting to see if the higher-gradient methods have some distinguishing property, such as computational efficiency.

2.3 Further explorations of the Bregman Lagrangian

In addition to providing a unifying framework for the generation of accelerated gradient-based algorithms, the Bregman Lagrangian has mathematical structure that can be investigated directly. In this section we briefly discuss some of the additional perspective that can be obtained from the Bregman Lagrangian. We elaborate on these results in Sections 2.6–2.12.

Hessian vs. Bregman Lagrangian. It is important to note the presence of the Bregman divergence in the Bregman Lagrangian (2.1). In the non-Euclidean setting, intuition might suggest using the Hessian metric $\nabla^2 h$ to measure a "kinetic energy," and thereby obtain a *Hessian Lagrangian*. This approach turns out to be unsatisfying, however, because the resulting differential equation does not yield a convergence rate and the Euler-Lagrange equation involves the third-order derivative $\nabla^3 h$, posing serious difficulties for discretization. As we have seen, the Bregman Lagrangian, on the other hand, readily provides a rate of convergence via a Lyapunov function; moreover, the resulting discrete-time algorithm in (2.30) involves only the gradient ∇h via the weighted mirror descent update.

Gradient vs. Lagrangian flows. In the Euclidean case, it is known classically that we can view gradient flow as the strong-friction limit of a damped Lagrangian flow [67, p. 646]. We show that the same interpretation holds for natural gradient flow and rescaled gradient flow. In particular, we show in Section 2.8 that we can recover natural gradient flow as the strong-friction limit of a Bregman Lagrangian flow with an appropriate choice of parameters. Similarly, we can recover the rescaled gradient flow (2.28) as the strong-friction limit of a Lagrangian flow that uses the p-th power of the norm as the kinetic energy. Therefore, the general family of second-order Lagrangian flows is more general, and includes first-order gradient flows in its closure. From this point of view, a particle with gradient-flow dynamics is operating in the regime of high friction. The particle simply rolls downhill and stops at the

equilibrium point as soon as the force $-\nabla f$ vanishes; there is no oscillation since it is damped by the infinitely strong friction. Thus, the effect of moving from a first-order gradient flow to a second-order Lagrangian flow is to reduce the friction from infinity to a finite amount; this permits oscillation [51, 61, 33], but also allows faster convergence.

Bregman Hamiltonian. One way to understand a Lagrangian is to study its Hamiltonian, which is the Legendre conjugate (dual function) of the Lagrangian. Typically, when the Lagrangian takes the form of the difference between kinetic and potential energy, the Hamiltonian is the sum of the kinetic and potential energy. The Hamiltonian is often easier to study than the Lagrangian, since its second-order Euler-Lagrangian equation is transformed into a pair of first-order equations. In our case, the Hamiltonian corresponding to the Bregman Lagrangian (2.1) is the following *Bregman Hamiltonian*,

$$\mathcal{H}(x,p,t) = e^{\alpha_t+\gamma_t}\left(D_{h^*}\left(\nabla h(x) + e^{-\gamma_t}p,\, \nabla h(x)\right) + e^{\beta_t}f(x)\right)$$

which indeed has the form of the sum of the kinetic and potential energy. Here the kinetic energy is measured using the Bregman divergence of h^*, which is the convex dual function of h.

Gauge invariance. The Euler-Lagrange equation of a Lagrangian is gauge-invariant, which means it does not change when we add a total time derivative to the Lagrangian. For the Bregman Lagrangian with the ideal scaling condition (2.2b), this property implies that we can replace the Bregman divergence $D_h(x+e^{-\alpha_t}\dot{x}, x)$ in (2.1) by its first term $h(x+e^{-\alpha_t}\dot{x})$. This might suggest a different interpretation of the role of h in the Lagrangian, one that is more symmetric to f.

Natural motion. The *natural motion* of the Bregman Lagrangian (i.e., the motion when there is no force, $-\nabla f \equiv 0$) is given by $X_t = ae^{-\gamma_t} + b$, for some constants $a, b \in \mathcal{X}$. Notice that even though the Bregman Lagrangian still involves the distance-generating function h, its natural motion is actually independent of h. Thus, the effect of h is felt only via its interaction with f—this can also be seen in (2.6) where h and f only appear together in the final term. Furthermore, assuming $e^{\gamma_t} \to \infty$, the natural motion always converges to a limit point, which a priori can be anything. However, as we see from Theorem 2.1, as soon as we introduce a convex potential function f, all motions converge to the minimizer x^* of f.

Exponential convergence rate via uniform convexity In addition to the polynomial family in Section 2.2, we can also study the subfamily of Bregman Lagrangians that have exponential convergence rates $O(e^{-ct})$, $c > 0$. As we discuss in Section 2.6, in this case the link to discrete-time algorithms is not as clear. Using the same discretization technique as in Section 2.2 suggests that to get a matching convergence rate, constant progress is needed at each iteration.

From the discrete-time perspective, we show that the higher-order gradient algorithm (2.26) achieves an exponential convergence rate when the objective function f is uniformly convex. Furthermore, we show that a restart scheme applied to the accelerated method (2.30) achieves a better dependence on the condition number; this generalizes Nesterov's restart scheme for the case $p = 3$ [46, Section 5].

It is an open question to understand if there is a better connection between the discrete-time restart algorithms and the continuous-time exponential Lagrangian flows. In particular, it is of interest to consider whether a restart scheme is necessary to achieve exponential convergence in discrete time; we know it is not needed for the special case $p = 2$, since a variant of Nesterov's accelerated gradient descent [48] that incorporates the condition number also achieves the optimal convergence rate.

2.4 Discussion

In this chapter, we have presented a variational framework for understanding accelerated methods from a continuous-time perspective. We presented the general family of Bregman Lagrangian, which generates a family of second-order Lagrangian dynamics that minimize the objective function at an accelerated rate compared to gradient flows. These dynamics are related to each other by the operation of speeding up time, because the Bregman Lagrangian family is closed under time dilation. In the polynomial case, we showed how to discretize the second-order Lagrangian dynamics to obtain an accelerated algorithm with a matching convergence rate. The resulting algorithm accelerates a base algorithm by coupling it with a weighted mirror descent step. An example of a base algorithm is a higher-order gradient method, which in continuous time corresponds to a first-order rescaled gradient flow with a matching convergence rate. Our continuous-time perspective makes clear that it is the mirror descent coupling that is more important for the acceleration phenomenon rather than the base algorithm. Indeed, the higher-order gradient algorithm operates on a smaller timescale than the enveloping mirror descent coupling step, so it makes no contribution in the continuous-time limit, and in principle we can use other base algorithms.

Our work raises many questions for further research. First, the case $p = 2$ is worthy of further investigation. In particular, the assumptions needed to show convergence of the discrete-time algorithm ($\nabla^{p-1} f$ is Lipschitz) are different than those required to show existence and uniqueness of solutions of the continuous-time dynamics (∇f is Lipschitz). In the case $p = 2$ however, these assumptions match. This suggests a strong link between the discrete- and continuous-time dynamics that might help us understand why several results seem to be unique to the special case $p = 2$. Second, in discrete time, Nesterov's accelerated methods have been extended to various settings, for example to the stochastic setting. An immediate question is whether we can extend our Lagrangian framework to these settings. Third, we would like to understand better the transition from continuous-time dynamics to discrete-time algorithms, and whether we can establish general assumptions that preserve desirable properties (e.g., convergence rate). In Section 2.2 we saw that the polynomial

convergence rate requires a higher-order smoothness assumption in discrete time, and in Section 2.3 we discussed whether the exponential case requires a uniform convexity assumption. Finally, our work to date focuses on the convergence rates of the function values rather than the iterates. Recently there has been some work extending [61] to study the convergence of the iterates [8] and some perturbative aspects [7]; it would be interesting to extend these results to the general Bregman Lagrangian.

At an abstract level, the general family of Bregman Lagrangian has a rich mathematical structure that deserves further study; we discussed some of these properties in Section 2.3. We hope that doing so will give us new insights into the nature of the optimization problem in continuous time, and help us design better dynamics with matching discrete-time algorithms. For example, we can study how to use some of the appealing properties of the Hamiltonian formalism (e.g., volume preservation in phase space) to help us discretize the dynamics. We also wish to understand where the Bregman Lagrangian itself comes from, why it works so well, and whether there are other Lagrangian families with similarly favorable properties.

2.5 Proofs of results

Proof of Theorem 2.2

The velocity and acceleration of the reparameterized curve $Y_t = X_{\tau(t)}$ are given by

$$\begin{aligned} \dot{Y}_t &= \dot{\tau}(t)\, \dot{X}_{\tau(t)} \\ \ddot{Y}_t &= \ddot{\tau}(t)\, \dot{X}_{\tau(t)} + \dot{\tau}(t)^2\, \ddot{X}_{\tau(t)}. \end{aligned}$$

Inverting these relations, we get

$$\dot{X}_{\tau(t)} = \frac{1}{\dot{\tau}(t)}\, \dot{Y}_t \tag{2.31a}$$

$$\ddot{X}_{\tau(t)} = \frac{1}{\dot{\tau}(t)^2} \ddot{Y}_t - \frac{\ddot{\tau}(t)}{\dot{\tau}(t)^3} \dot{Y}_t. \tag{2.31b}$$

By assumption, the original curve X_t satisfies the Euler-Lagrange equation (2.5) for the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$. At time $\tau(t)$, this equation reads

$$\ddot{X}_{\tau(t)} + (e^{\alpha_{\tau(t)}} - \dot{\alpha}_{\tau(t)})\dot{X}_{\tau(t)} + e^{2\alpha_{\tau(t)}+\beta_{\tau(t)}} \left[\nabla^2 h(X_{\tau(t)} + e^{-\alpha_{\tau(t)}} \dot{X}_{\tau(t)})\right]^{-1} \nabla f(X_{\tau(t)})$$
$$+ e^{\alpha_{\tau(t)}}(\dot{\gamma}_{\tau(t)} - e^{\alpha_{\tau(t)}}) \left[\nabla^2 h(X_{\tau(t)} + e^{-\alpha_{\tau(t)}} \dot{X}_{\tau(t)})\right]^{-1} (\nabla h(X_{\tau(t)} + e^{-\alpha_{\tau(t)}} \dot{X}_{\tau(t)}) - \nabla h(X_{\tau(t)})) = 0.$$

We now use the relations (2.31). After multiplying by $\dot{\tau}(t)^2$ and collecting terms, we get

$$\ddot{Y}_t + \left(\dot{\tau}(t) e^{\alpha_{\tau(t)}} - \dot{\tau}(t)\dot{\alpha}_{\tau(t)} - \frac{\ddot{\tau}(t)}{\dot{\tau}(t)}\right) \dot{Y}_t + \dot{\tau}(t)^2 e^{2\alpha_{\tau(t)}+\beta_{\tau(t)}} \left[\nabla^2 h\left(Y_t + \frac{e^{-\alpha_{\tau(t)}}}{\dot{\tau}(t)} \dot{Y}_t\right)\right]^{-1} \nabla f(Y_t)$$
$$+ \dot{\tau}(t)^2 e^{\alpha_{\tau(t)}} (\dot{\gamma}_{\tau(t)} - e^{\alpha_{\tau(t)}}) \left[\nabla^2 h\left(Y_t + \frac{e^{-\alpha_{\tau(t)}}}{\dot{\tau}(t)} \dot{Y}_t\right)\right]^{-1} \left(\nabla h\left(Y_t + \frac{e^{-\alpha_{\tau(t)}}}{\dot{\tau}(t)} \dot{Y}_t\right) - \nabla h(Y_t)\right) = 0.$$

Finally, with the definition of the modified parameters $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ (2.10), we can write this equation as

$$\ddot{Y}_t + (e^{\tilde{\alpha}_t} - \dot{\tilde{\alpha}}_t)\dot{Y}_t + e^{2\tilde{\alpha}_t + \tilde{\beta}_t}\left[\nabla^2 h(Y_t + e^{-\tilde{\alpha}_t}\dot{Y}_t)\right]^{-1}\nabla f(Y_t) \\ + e^{\tilde{\alpha}_t}(\dot{\tilde{\gamma}}_t - e^{\tilde{\alpha}_t})\left[\nabla^2 h(Y_t + e^{-\tilde{\alpha}_t}\dot{Y}_t)\right]^{-1}(\nabla h(Y_t + e^{-\tilde{\alpha}_t}\dot{Y}_t) - \nabla h(Y_t)) = 0,$$

which we recognize as the Euler-Lagrange equation (2.5) for the Bregman Lagrangian $\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}$.

Furthermore, suppose α, β, γ satisfy the ideal scaling (2.2). Then

$$\dot{\tilde{\beta}}_t = \frac{d}{dt}\beta_{\tau(t)} = \dot{\tau}(t)\dot{\beta}_{\tau(t)} \overset{(2.2a)}{\leq} \dot{\tau}(t)e^{\alpha_{\tau(t)}} = e^{\alpha_{\tau(t)} + \log\dot{\tau}(t)} = e^{\tilde{\alpha}_t}$$
$$\dot{\tilde{\gamma}}_t = \frac{d}{dt}\gamma_{\tau(t)} = \dot{\tau}(t)\dot{\gamma}_{\tau(t)} \overset{(2.2b)}{=} \dot{\tau}(t)e^{\alpha_{\tau(t)}} = e^{\alpha_{\tau(t)} + \log\dot{\tau}(t)} = e^{\tilde{\alpha}_t},$$

which means that the modified parameters $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ also satisfy the ideal scaling (2.2). The converse follows by considering the inverse function $\tau^{-1}(t)$ in place of $\tau(t)$.

□

Existence and uniqueness of solution to the polynomial family

In this section we discuss the existence and uniqueness of solution to the differential equation (2.14) arising from the polynomial family of Bregman Lagrangian. We begin by writing the second-order equation (2.14) as the pair of first-order equations (2.15). We also write $W_t = \nabla h(Z_t)$, so we can write (2.15) as

$$\dot{X}_t = \frac{p}{t}\left(\nabla h^*(W_t) - X_t\right) \tag{2.32a}$$
$$\dot{W}_t = -Cpt^{p-1}\nabla f(X_t). \tag{2.32b}$$

Here $h^*\colon \mathcal{X}^* \to \mathbb{R}$ is the Legendre conjugate function of h, defined by

$$h^*(w) = \sup_{z\in\mathcal{X}}\left\{\langle w, z\rangle - h(z)\right\}, \tag{2.33}$$

where $\mathcal{X}^*$ is the dual space of $\mathcal{X}$, i.e., the space of all linear functionals over $\mathcal{X}$. Under the assumption that h be essentially smooth, the supremum in (2.33) is achieved by $z = \nabla h^*(w)$, and we have the relation that ∇h and ∇h^* are inverses of each other, i.e., $z = \nabla h^*(w) \Leftrightarrow w = \nabla h(z)$. Thus, with the definition $W_t = \nabla h(Z_t)$, we can write $Z_t = \nabla h^*(W_t)$, which gives us (2.32).

Now assume ∇f and ∇h^* are Lipschitz continuous functions. Then over any bounded time intervals $[t_0, t_1]$ with $0 < t_0 < t_1$, the right-hand side of (2.32) is a Lipschitz continuous vector field. Thus, by the Cauchy-Lipschitz theorem, for any given initial conditions $(X_{t_0}, W_{t_0}) = (x_0, w_0)$ at time $t = t_0$, the system of differential equations (2.32) has a unique

solution over the time interval $[t_0, t_1]$. Furthermore, the solution does not blow up in any finite time, since from Theorem 2.1 we know that the energy functional $\mathcal{E}_t$ (2.8) is non-increasing, so in particular, the Bregman divergence $D_h(x^*, X_t + \frac{t}{p}\dot{X}_t)$ is bounded above by a constant. Since t_1 is arbitrary, this shows that (2.32) has a unique maximal solution, i.e., t_1 can be extended to $t_1 \to +\infty$.

In the above argument we have started at time $t_0 > 0$, because the vector field in (2.32) has a singularity at $t = 0$. For $p = 2$, Su et al. [61] and Krichene et al. [33] treat the case when we start at $t = 0$ with initial condition $(X_0, W_0) = (x_0, \nabla h(x_0))$, so that $\dot{X}_0 = 0$. In that case, they show that the system (2.32) still has a unique solution for all time $[0, \infty)$, by replacing the p/t coefficient by the approximation $p/\max\{t, \delta\}$ for $\delta > 0$ and letting $\delta \to 0$. We can adapt this technique to the more general case (2.32); alternatively, we can appeal to the time dilation property and state that since the general system (2.32) is the result of speeding up the $p = 2$ case by time dilation function $\tau(t) = t^{p/2}$, once we know a unique solution exists for $p = 2$, we can also conclude that it exists for all $p > 0$.

Proof of Theorem 2.3

We define the following function, which is a generalization of Nesterov's *estimate function* from [46],

$$\psi_k(x) = Cp \sum_{i=0}^{k} i^{(p-1)} \big[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle \big] + \frac{1}{\epsilon} D_h(x, x_0). \tag{2.34}$$

The estimate function ψ_k arises as the objective function that the sequence z_k is optimizing in (2.18b). Indeed, the optimality condition for the z_k update (2.18b) is

$$\nabla h(z_k) = \nabla h(z_{k-1}) - \epsilon C p k^{(p-1)} \nabla f(y_k).$$

By unrolling the recursion, we can write

$$\nabla h(z_k) = \nabla h(z_0) - \epsilon C p \sum_{i=0}^{k} i^{(p-1)} \nabla f(y_i),$$

and since $x_0 = z_0$, we can write this equation as $\nabla \psi_k(z_k) = 0$. Since ψ_k is a convex function, this means z_k is the minimizer of ψ_k. Thus, we can equivalently write the update for z_k as

$$z_k = \arg\min_z \ \psi_k(z). \tag{2.35}$$

For proving the convergence rate for the algorithm (2.18), we have the following property.

Lemma 2.8. *For all $k \geq 0$, we have*

$$\psi_k(z_k) \ \geq \ C k^{(p)} f(y_k). \tag{2.36}$$

Proof. We proceed via induction on $k \geq 0$. The base case $k = 0$ is true since both sides equal zero. Now assume (2.36) holds for some $k \geq 0$; we will show it also holds for $k+1$.

Since h is 1-uniformly convex of order p, the rescaled Bregman divergence $\frac{1}{\epsilon}D_h(x, x_0)$ is $(\frac{1}{\epsilon})$-uniformly convex. Thus, the estimate function ψ_k (2.34) is also $(\frac{1}{\epsilon})$-uniformly convex of order p. Since z_k is the minimizer of ψ_k, $\nabla\psi_k(z_k) = 0$, so for all $x \in \mathcal{X}$ we have

$$\psi_k(x) \;=\; \psi_k(z_k) + D_{\psi_k}(x, z_k) \;\geq\; \psi_k(z_k) + \frac{1}{\epsilon p}\|x - z_k\|^p.$$

Applying the inductive hypothesis (2.36) and using the convexity of f gives us

$$\psi_k(x) \;\geq\; Ck^{(p)}\big[f(y_{k+1}) + \langle \nabla f(y_{k+1}), y_k - y_{k+1}\rangle\big] + \frac{1}{\epsilon p}\|x - z_k\|^p.$$

We now add $Cp(k+1)^{(p-1)}[f(y_{k+1}) + \langle \nabla f(y_{k+1}), x - y_{k+1}\rangle]$ to both sides of the equation to obtain

$$\psi_{k+1}(x) \;\geq\; C(k+1)^{(p)}\big[f(y_{k+1}) + \big\langle \nabla f(y_{k+1}),\, x_{k+1} - y_{k+1} + \tau_k(x - z_k)\big\rangle\big] + \frac{1}{\epsilon p}\|x - z_k\|^p, \tag{2.37}$$

where $\tau_k = \frac{p(k+1)^{(p-1)}}{(k+1)^{(p)}} = \frac{p}{k+p}$, and where we have also used the definition of x_{k+1} as a convex combination of y_k and z_k with weight τ_k (2.18a).

Note that the first term in (2.37) gives our desired inequality (2.36) for $k+1$. So to finish the proof, we have to prove the remaining terms in (2.37) are nonnegative. We do so by applying two inequalities. We first apply the inequality (2.19) to the term $\langle \nabla f(y_{k+1}), x_{k+1} - y_{k+1}\rangle$, so from (2.37) we have

$$\begin{aligned}\psi_{k+1}(x) \;\geq\; C(k+1)^{(p)} f(y_{k+1}) \;+\; C(k+1)^{(p)} M\epsilon^{\frac{1}{p-1}}\,\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} \\ +\; Cp(k+1)^{(p-1)}\langle \nabla f(y_{k+1}), x - z_k\rangle \;+\; \frac{1}{\epsilon p}\|x - z_k\|^p.\end{aligned} \tag{2.38}$$

Next, we apply the Fenchel-Young inequality [46, Lemma 2]

$$\langle s, u\rangle + \frac{1}{p}\|u\|^p \geq -\frac{p-1}{p}\|s\|_*^{\frac{p}{p-1}} \tag{2.39}$$

with the choices $u = \epsilon^{-\frac{1}{p}}(x - z_k)$ and $s = \epsilon^{\frac{1}{p}} Cp(k+1)^{(p-1)}\nabla f(y_{k+1})$. Then from (2.38), we obtain

$$\psi_{k+1}(x) \;\geq\; C(k+1)^{(p)}\left[f(y_{k+1}) + \left(M - \frac{p-1}{p}\,p^{\frac{p}{p-1}}\,C^{\frac{1}{p-1}}\,\frac{\big\{(k+1)^{(p-1)}\big\}^{\frac{p}{p-1}}}{(k+1)^{(p)}}\right)\epsilon^{\frac{1}{p-1}}\,\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}}\right].$$

Notice that $\big\{(k+1)^{(p-1)}\big\}^{\frac{p}{p-1}} \leq (k+1)^{(p)}$. Then from the assumption $C \leq M^{p-1}/p^p$, we see that the second term inside the parentheses is nonnegative. Hence we conclude the desired inequality $\psi_{k+1}(x) \geq C(k+1)^{(p)} f(y_{k+1})$. Since $x \in \mathcal{X}$ is arbitrary, it also holds for the minimizer $x = z_{k+1}$ of ψ_{k+1}, finishing the induction. $\square$

With Lemma 2.8 in hand, we can complete the proof of Theorem 2.3.

Proof of Theorem 2.3. Since f is convex, we can bound the estimate sequence ψ_k by

$$\psi_k(x) \;\leq\; Cp\sum_{i=0}^{k} i^{(p-1)} f(x) + \frac{1}{\epsilon} D_h(x, x_0) \;=\; Ck^{(p)} f(x) + \frac{1}{\epsilon} D_h(x, x_0).$$

This holds for all $x \in \mathcal{X}$, and in particular for the minimizer x^* of f. Combining the bound with the result of Lemma 2.8, and recalling that z_k is the minimizer of ψ_k, we get

$$Ck^{(p)} f(y_k) \;\leq\; \psi_k(z_k) \;\leq\; \psi_k(x^*) \;\leq\; Ck^{(p)} f(x^*) + \frac{1}{\epsilon} D_h(x^*, x_0).$$

Rearranging and dividing by $Ck^{(p)}$ gives us the desired convergence rate (2.21). □

Proof of Lemma 2.4

We follow the approach of [46, Lemma 6]. Since y solves the optimization problem (2.23), it satisfies the optimality condition

$$\sum_{i=1}^{p-1} \frac{1}{(i-1)!} \nabla^i f(x)\,(y-x)^{i-1} + \frac{N}{\epsilon}\|y-x\|^{p-2}\,(y-x) = 0. \tag{2.40}$$

Furthermore, since $\nabla^{p-1} f$ is $\frac{(p-1)!}{\epsilon}$-Lipschitz, we have the following error bound on the $(p-2)$-nd order Taylor expansion of ∇f,

$$\left\| \nabla f(y) - \sum_{i=1}^{p-1} \frac{1}{(i-1)!} \nabla^i f(x)\,(y-x)^{i-1} \right\|_* \leq \frac{1}{\epsilon}\|y-x\|^{p-1}. \tag{2.41}$$

Substituting (2.40) to (2.41) and writing $r = \|y-x\|$, we obtain

$$\left\| \nabla f(y) + \frac{N r^{p-2}}{\epsilon}\,(y-x) \right\|_* \;\leq\; \frac{r^{p-1}}{\epsilon}. \tag{2.42}$$

Squaring both sides, expanding, and rearranging the terms, we get the inequality

$$\langle \nabla f(y), x-y \rangle \;\geq\; \frac{\epsilon}{2Nr^{p-2}}\|\nabla f(y)\|_*^2 + \frac{(N^2-1)r^p}{2N\epsilon}. \tag{2.43}$$

Note that if $p = 2$, then the first term in (2.43) already implies the desired bound (2.24). Now assume $p \geq 3$. The right-hand side of (2.43) is of the form $A/r^{p-2} + Br^p$, which is a convex function of $r > 0$ and minimized by $r^* = \left\{ \frac{(p-2)}{p}\frac{A}{B} \right\}^{\frac{1}{2p-2}}$, yielding a minimum value of

$$\frac{A}{(r^*)^{p-2}} + B(r^*)^p \;=\; A^{\frac{p}{2p-2}} B^{\frac{p-2}{2p-2}} \left[\left(\frac{p}{p-2} \right)^{\frac{p-2}{2p-2}} + \left(\frac{p-2}{p} \right)^{\frac{p}{p-2}} \right] \;\geq\; A^{\frac{p}{2p-2}} B^{\frac{p-2}{2p-2}}.$$

Substituting the values $A = \frac{\epsilon}{2N}\|\nabla f(y)\|_*^2$ and $B = \frac{1}{2N\epsilon}(N^2-1)$ from (2.43), we obtain

$$\langle \nabla f(y), x-y\rangle \;\geq\; \left(\frac{\epsilon}{2N}\|\nabla f(y)\|_*^2\right)^{\frac{p}{2p-2}} \left(\frac{1}{2N\epsilon}(N^2-1)\right)^{\frac{p-2}{2p-2}} \;=\; \frac{(N^2-1)^{\frac{p-2}{2p-2}}}{2N}\epsilon^{\frac{1}{p-1}}\|\nabla f(y)\|_*^{\frac{p}{p-1}}$$

which proves (2.24).

To obtain the first inequality of (2.25), we use Cauchy-Schwarz inequality on (2.24),

$$\frac{(N^2-1)^{\frac{p-2}{2p-2}}}{2N}\epsilon^{\frac{1}{p-1}}\|\nabla f(y)\|_*^{\frac{p}{p-1}} \;\leq\; \langle \nabla f(y), x-y\rangle \;\leq\; \|\nabla f(y)\|_*\,\|x-y\|$$

and cancel out $\|\nabla f(y)\|_*$ from both sides. For the second inequality of (2.25), we use triangle inequality on the left hand side of (2.42),

$$\frac{Nr^{p-1}}{\epsilon} - \|\nabla f(y)\|_* \;\leq\; \left\|\nabla f(y) + \frac{Nr^{p-2}}{\epsilon}(y-x)\right\|_* \;\leq\; \frac{r^{p-1}}{\epsilon}.$$

Rearranging the terms and taking the $(p-1)$-st root of both sides gives us the result (2.25). □

Proof of Theorem 2.5

This proof follows the approach in the proof of [46, Theorem 1]. We first prove the following lemma. Here $\delta_k = f(x_k) - f(x^*) \geq 0$ denotes the residual value at iteration k.

Lemma 2.9. *Under the setting of Theorem 2.5, we have*

$$\delta_{k+1} \leq \delta_k - \frac{(p-1)}{p}\cdot\left(\frac{\epsilon\delta_k^p}{(N+1)R^p}\right)^{\frac{1}{p-1}}. \tag{2.44}$$

Proof. Since f is $\frac{(p-1)!}{\epsilon}$-smooth of order $p-1$, by the Taylor remainder theorem we have the bound

$$|f_{p-1}(x;x_k) - f(x)| \leq \frac{1}{\epsilon p}\|x-x_k\|^p.$$

Then from the definition of x_{k+1} (2.26), we have

$$f(x_{k+1}) \;=\; \min_{x\in\mathcal{X}}\left\{f_{p-1}(x;x_k) + \frac{N}{\epsilon p}\|x-x_k\|^p\right\} \leq \min_{x\in\mathcal{X}}\left\{f(x) + \frac{N+1}{\epsilon p}\|x-x_k\|^p\right\}. \tag{2.45}$$

Plugging in $x = x_k$ on the right-hand side of (2.45) shows that $f(x_{k+1}) \leq f(x_k)$; that is, the algorithm (2.26) is a descent method. In particular, for all $k \geq 0$ we have $\|x_k - x^*\| \leq R$,

where $R = \sup_{x:\, f(x)\le f(x_0)} \|x - x^*\|$ is the radius of the level set as defined in Theorem 2.5. Moreover, plugging in $x = x^*$ on the right-hand side of (2.45) gives us

$$f(x_{k+1}) - f(x^*) \le \frac{N+1}{\epsilon p}\|x_k - x^*\|^p \le \frac{N+1}{\epsilon p}R^p. \tag{2.46}$$

Now for any $\lambda \in [0,1]$, consider the midpoint

$$x_\lambda = x^* + (1-\lambda)(x_k - x^*) = \lambda x^* + (1-\lambda)x_k.$$

By Jensen's inequality, $f(x_\lambda) \le \lambda f(x^*) + (1-\lambda) f(x_k)$. We also have $\|x_\lambda - x_k\| = \lambda\|x_k - x^*\| \le \lambda R$. Plugging in the point x_λ to the right-hand side of (2.45) gives

$$f(x_{k+1}) \;\le\; f(x_\lambda) + \frac{N+1}{\epsilon p}\|x_\lambda - x_k\|^p \;\le\; \lambda f(x^*) + (1-\lambda) f(x_k) + \frac{N+1}{\epsilon p}R^p\lambda^p.$$

With the notation $\delta_k = f(x_k) - f(x^*)$, we can write the last inequality as

$$\delta_{k+1} \;\le\; (1-\lambda)\delta_k + \frac{N+1}{\epsilon p}R^p\lambda^p. \tag{2.47}$$

The right-hand side is a convex function of λ, which is minimized at $\lambda^* = \left\{\frac{\epsilon}{N+1}\frac{\delta_k}{R^p}\right\}^{\frac{1}{p-1}}$. Note that $\lambda^* \in [0,1]$ by (2.46). Plugging in λ^* to (2.47) yields the desired bound (2.44). □

With Lemma 2.9, we can complete the proof of Theorem 2.5.

Proof of Theorem 2.5. Define the energy functional $e_k = \delta_k^{-\frac{1}{p-1}}$. We can write

$$e_{k+1} - e_k \;=\; \frac{1}{\delta_{k+1}^{\frac{1}{p-1}}} - \frac{1}{\delta_k^{\frac{1}{p-1}}} \;=\; \frac{\delta_k^{\frac{1}{p-1}} - \delta_{k+1}^{\frac{1}{p-1}}}{\delta_{k+1}^{\frac{1}{p-1}} \cdot \delta_k^{\frac{1}{p-1}}} \;=\; \frac{\delta_k - \delta_{k+1}}{\delta_{k+1}^{\frac{1}{p-1}} \cdot \delta_k^{\frac{1}{p-1}}} \cdot \frac{1}{\left(\sum_{i=0}^{p-2} \delta_k^{\frac{i}{p-1}} \cdot \delta_{k+1}^{\frac{p-2-i}{p-1}}\right)}. \tag{2.48}$$

Since $\delta_{k+1} \le \delta_k$, we can upper bound the summation in the denominator of (2.48) by $(p-1)\delta_k^{\frac{p-2}{p-1}}$. We use Lemma 2.9 to lower bound $\delta_k - \delta_{k+1}$, obtaining

$$e_{k+1} - e_k \;\ge\; \frac{(p-1)}{p} \cdot \left(\frac{\epsilon\delta_k^p}{(N+1)R^p}\right)^{\frac{1}{p-1}} \cdot \frac{1}{\delta_k^{\frac{2}{p-1}}} \cdot \frac{1}{(p-1)\delta_k^{\frac{p-2}{p-1}}} \;=\; \frac{1}{p} \cdot \left(\frac{\epsilon}{(N+1)R^p}\right)^{\frac{1}{p-1}}. \tag{2.49}$$

Summing (2.49) and telescoping the terms, we get

$$\frac{1}{(f(x_k) - f(x^*))^{\frac{1}{p-1}}} \;=\; e_k \;\ge\; e_k - e_0 \;\ge\; \frac{k}{p} \cdot \left(\frac{\epsilon}{(N+1)R^p}\right)^{\frac{1}{p-1}}$$

which gives us the desired conclusion (2.27). □

Proof of Theorem 2.6

We write the higher-order gradient algorithm (2.26) (with $N = 1$) as

$$x_{k+1} - x_k = \arg\min_u \left\{ f(x_k) + \langle \nabla f(x_k), u \rangle + \cdots + \frac{1}{(p-1)!} \nabla^{p-1} f(x_k) u^{p-1} + \frac{1}{\epsilon p} \|u\|^p \right\}. \tag{2.50}$$

Our goal is to express the sequence x_k as a discretization $x_k = X_t$, $x_{k+1} = X_{t+\delta} \approx X_t + \delta \dot{X}_t$ of some continuous-time curve X_t with time step $\delta > 0$, which will be a function of ϵ. To that end, we write $u = \delta v$, so (2.50) becomes

$$\frac{x_{k+1} - x_k}{\delta} = \arg\min_v \left\{ f(x_k) + \delta \langle \nabla f(x_k), v \rangle + \cdots + \frac{\delta^{p-1}}{(p-1)!} \nabla^{p-1} f(x_k) v^{p-1} + \frac{\delta^p}{\epsilon p} \|v\|^p \right\}.$$

Eliminating the constant term $f(x_k)$ from the right-hand side, which does not change the minimizer, and canceling a factor of δ, we get

$$\frac{x_{k+1} - x_k}{\delta} = \arg\min_v \left\{ \langle \nabla f(x_k), v \rangle + \frac{\delta}{2} \nabla^2 f(x_k) v^2 + \cdots + \frac{\delta^{p-2}}{(p-1)!} \nabla^{p-1} f(x_k) v^{p-1} + \frac{\delta^{p-1}}{\epsilon p} \|v\|^p \right\}.$$

We see that the first term in the objective function does not depend on δ. As $\epsilon \to 0$, for the equation to have a meaningful limit, we have to set $\delta^{p-1} = \epsilon$, so the last term in the objective function becomes a constant. On the other hand, the middle terms all have dependence on $\delta = \epsilon^{\frac{1}{p-1}}$, so as $\epsilon \to 0$, those terms vanish. Thus, the limit as $\epsilon \to 0$ is

$$\dot{X}_t = \arg\min_v \left\{ \langle \nabla f(X_t), v \rangle + \frac{1}{p} \|v\|^p \right\}. \tag{2.51}$$

Equivalently, $\dot{X}_t$ satisfies the optimality condition

$$\nabla f(X_t) + \|\dot{X}_t\|^{p-2} \dot{X}_t = 0. \tag{2.52}$$

This gives us the relation $\|\nabla f(X_t)\|_* = \|\dot{X}_t\|^{p-1}$, so we can also write (2.52) as

$$\dot{X}_t = -\frac{\nabla f(X_t)}{\|\dot{X}_t\|^{p-2}} = -\frac{\nabla f(X_t)}{\|\nabla f(X_t)\|_*^{\frac{p-2}{p-1}}},$$

which is the rescaled gradient flow as claimed in (2.28).

We note that the rescaled gradient flow (2.28) is a descent method, since

$$\frac{d}{dt} f(X_t) = \langle \nabla f(X_t), \dot{X}_t \rangle = -\|\nabla f(X_t)\|_*^{\frac{p}{p-1}} \leq 0.$$

Now to establish the convergence rate of the rescaled gradient flow (2.28), we consider the energy functional

$$\mathcal{E}_t = (f(X_t) - f(x^*))^{-\frac{1}{p-1}} \tag{2.53}$$

which is the same energy functional as in the discrete-time convergence proof in Section 2.5. The energy functional $\mathcal{E}_t$ has time derivative

$$\dot{\mathcal{E}}_t = -\frac{1}{(p-1)} \frac{\langle \nabla f(X_t), \dot{X}_t \rangle}{(f(X_t) - f(x^*))^{\frac{p}{p-1}}}.$$

If X_t satisfies the rescaled gradient flow equation (2.28), then $\dot{\mathcal{E}}_t$ simplifies to

$$\dot{\mathcal{E}}_t = \frac{1}{(p-1)} \left(\frac{\|\nabla f(X_t)\|_*}{f(X_t) - f(x^*)} \right)^{\frac{p}{p-1}}. \tag{2.54}$$

By the convexity of f and the Cauchy-Schwarz inequality, we have

$$0 \leq f(X_t) - f(x^*) \leq \langle \nabla f(X_t), X_t - x^* \rangle \leq \|\nabla f(X_t)\|_* \, \|X_t - x^*\|.$$

Since the rescaled gradient flow is a descent method, we have $\|X_t - x^*\| \leq R$. Therefore, from (2.54) we get the bound

$$\dot{\mathcal{E}}_t \geq \frac{1}{(p-1)} \frac{1}{\|X_t - x^*\|^{\frac{p}{p-1}}} \geq \frac{1}{(p-1) R^{\frac{p}{p-1}}}.$$

This means that $\mathcal{E}_t$ increases at least linearly, so

$$\frac{1}{(f(X_t) - f(x^*))^{\frac{1}{p-1}}} \;=\; \mathcal{E}_t \;\geq\; \mathcal{E}_0 + \frac{t}{(p-1)\, R^{\frac{p}{p-1}}} \;\geq\; \frac{t}{(p-1)\, R^{\frac{p}{p-1}}},$$

which gives us the desired result (2.29). □

Remark: From the proof above, we see that rescaled gradient flow (2.28) is a generalization of the usual gradient flow (the case $p = 2$) which is obtained by replacing the squared norm by the p-th power of the norm in the variational formulation (2.51). It turns out that when the objective function is the p-th power of the norm, $f(x) = \frac{1}{p}\|x\|^p$, the rescaled gradient flow (2.28) reduces to an explicit equation. Specifically, in this case we have $\nabla f(x) = \|x\|^{p-2}x$, so $\|\nabla f(x)\|_* = \|x\|^{p-1}$. Therefore, the rescaled gradient flow equation (2.28) becomes

$$\dot{X}_t = -\frac{\nabla f(X_t)}{\|\nabla f(X_t)\|_*^{\frac{p-2}{p-1}}} = -\frac{\|X_t\|^{p-2} X_t}{\|X_t\|^{p-2}} = -X_t,$$

which is now independent of p, and has an explicit solution $X_t = e^{-t} X_0$.

Alternative proof of convergence rate. In the proof above, we can also use the following alternative energy functional,

$$\tilde{\mathcal{E}}_t = t^p(f(X_t) - f(x^*)). \tag{2.55}$$

Its time derivative is

$$\begin{aligned}
\dot{\tilde{\mathcal{E}}}_t &= pt^{p-1}(f(X_t) - f(x^*)) + t^p\langle \nabla f(X_t), \dot{X}_t\rangle \\
&\leq pt^{p-1}\langle \nabla f(X_t), X_t - x^*\rangle + t^p\langle \nabla f(X_t), \dot{X}_t\rangle && \text{(2.56a)} \\
&= pt^{p-1}\langle \nabla f(X_t), X_t - x^*\rangle - t^p\|\nabla f(X_t)\|_*^{\frac{p}{p-1}}, && \text{(2.56b)}
\end{aligned}$$

where (2.56a) follows from the convexity of f, and in (2.56b) we have substituted the rescaled gradient flow dynamic (2.28). We now apply the Fenchel-Young inequality (2.39) with $s = t^{p-1}\nabla f(X_t)$ and $u = -(p-1)(X_t - x^*)$, to obtain

$$\dot{\tilde{\mathcal{E}}}_t \leq \frac{1}{p-1}\|(p-1)(X_t - x^*)\|^p \leq (p-1)^{p-1}R^p, \tag{2.57}$$

where in the last step we have used the fact that $\|X_t - x^*\| \leq R$ since rescaled gradient flow is a descent method. Integrating (2.57) and plugging in the definition of $\tilde{\mathcal{E}}_t$ (2.55), we obtain

$$f(X_t) - f(x^*) \leq \frac{(p-1)^{p-1}\,R^p}{t^{p-1}},$$

which is exactly the same bound as claimed in (2.29).

2.6 Exponential convergence rate via uniform convexity

Similar to the polynomial case in Section 2.2, in this section we study the subfamily of Bregman Lagrangian (2.1) with the following choice of parameters, parameterized by $c > 0$,

$$\begin{aligned}
\alpha_t &= \log c && \text{(2.58a)} \\
\beta_t &= ct && \text{(2.58b)} \\
\gamma_t &= ct. && \text{(2.58c)}
\end{aligned}$$

The parameters (2.58) satisfy the ideal scaling condition (2.2), with an equality on the first condition (2.2a). The Bregman Lagrangian (2.1) becomes

$$\mathcal{L}(x, \dot{x}, t) = ce^{ct}\left(D_h\left(x + \frac{1}{c}\dot{x}, x\right) - e^{ct}f(x)\right). \tag{2.59}$$

The Euler-Lagrange equation (2.6) in this case is given by

$$\ddot{X}_t + c\dot{X}_t + c^2 e^{ct}\left[\nabla^2 h\left(X_t + \frac{1}{c}\dot{X}_t\right)\right]^{-1}\nabla f(X_t) = 0, \tag{2.60}$$

and by Theorem 2.1, it has an $O(e^{-ct})$ rate of convergence. Thus, whereas the polynomial Lagrangian flow (2.14) has a polynomial rate of convergence, the exponential Lagrangian flow (2.60) has an exponential rate of convergence. Furthermore, from the time-dilation property in Theorem 2.2, we see that we can obtain the exponential curve (2.60) by speeding up the polynomial curve (2.14) using a time-dilation function $\tau(t) = e^{ct/p}$.

However, unlike the polynomial Lagrangian flow (2.14), the process of discretizing the exponential Lagrangian flow (2.60) is not as straightforward. Following the same approach as the polynomial family, we write the second-order equation (2.60) as the following pair of first-order equations:

$$Z_t = X_t + \frac{1}{c}\dot{X}_t \tag{2.61a}$$

$$\frac{d}{dt}\nabla h(Z_t) = -ce^{ct}\nabla f(X_t). \tag{2.61b}$$

Now we discretize X_t and Z_t into sequences x_k and z_k with time step $\delta > 0$, so that $t = \delta k$ as before. In doing so, we can write (2.61) as the following discrete-time equations similar to (2.16) and (2.17):

$$x_{k+1} = c\delta z_k + (1 - c\delta)x_k \tag{2.62a}$$

$$z_{k+1} = \arg\min_z \left\{ ce^{c\delta k}\langle \nabla f(x_k), z\rangle + \frac{1}{\delta}D_h(z, z_k)\right\}. \tag{2.62b}$$

Note that the weight in (2.62a) is independent of time, but depends on δ, and (2.62b) suggests the step size $\epsilon = \delta$ in the algorithm. If our analogy between continuous and discrete-time convergence holds, then given the $O(e^{-ct})$ convergence rate in continuous time, we expect a matching $O(\frac{1}{\epsilon}e^{-ck})$ convergence rate in discrete time. However, it is not clear how to obtain that rate via (2.62). If we try to adapt the proof of Theorem 2.3, we find that in order to conclude a convergence rate $O(\delta e^{-c\delta k})$, we need to introduce a sequence y_k satisfying the following analog of inequality (2.19) (with the ideal choice $p = \infty$):

$$\langle \nabla f(y_k), x_k - y_k\rangle \;\geq\; M\|\nabla f(y_k)\|_*. \tag{2.63}$$

Notice that the rates are consistent if we set $\epsilon = \delta = 1$. However, the condition (2.63) means we need to make a constant improvement in each iteration from x_k to y_k, although we are also free on how we choose to construct y_k and impose any assumptions on f.

In the remainder of this section, we approach this problem from a discrete-time perspective, and study the performance of the higher-order gradient algorithm (2.26) and its accelerated variant (2.30) when f is uniformly convex.

Exponential convergence rate of higher-order gradient algorithm

In this section we show that the higher-order gradient algorithm (2.26) has an exponential convergence rate when the objective function f is uniformly convex of order $p \geq 2$; this generalizes the results in [46, Section 5] for the case $p = 3$, and the classical result of gradient descent for the case $p = 2$ [48].

Specifically, we have the following result. Recall the definition of smoothness in (2.22), and the definition of uniform convexity in (2.20).

Theorem 2.10. *Suppose f is $\frac{(p-1)!}{\epsilon}$-smooth of order $p-1$, and σ-uniformly convex of order p. Then the p-th order gradient algorithm (2.26) with $N > 1$ has convergence rate*

$$f(x_{k+1}) - f(x^*) \leq \frac{(N+1)\|x_0 - x^*\|^p}{\epsilon p\big(1 + L\kappa^{\frac{1}{p-1}}\big)^k} = O\left(\frac{1}{\epsilon}\exp(-L\kappa^{\frac{1}{p-1}}k)\right), \tag{2.64}$$

where $L = (N^2-1)^{\frac{p-2}{2p-2}}/(2N)$, and $\kappa = \epsilon\sigma$ is the inverse condition number (which we assume is small).

Proof. By inequality (2.24) from Lemma 2.4, we know that since f is $\frac{(p-1)!}{\epsilon}$-smooth of order $p-1$,

$$\langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \geq L\epsilon^{\frac{1}{p-1}}\|\nabla f(x_{k+1})\|_*^{\frac{p}{p-1}},$$

where $L = (N^2-1)^{\frac{p-2}{2p-2}}/(2N)$. Since f is convex, we have $f(x_k) - f(x_{k+1}) \geq \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle$. Furthermore, since f is σ-uniformly convex of order p, from [46, Lemma 3] we also have

$$\|\nabla f(x_{k+1})\|_*^{\frac{p}{p-1}} \geq \frac{p}{p-1}\sigma^{\frac{1}{p-1}}(f(x_{k+1}) - f(x^*)) \geq \sigma^{\frac{1}{p-1}}(f(x_{k+1}) - f(x^*)). \tag{2.65}$$

Combining these inequalities and recalling the definition $\kappa = \epsilon\sigma$ gives us

$$f(x_k) - f(x_{k+1}) \geq L\kappa^{\frac{1}{p-1}}(f(x_{k+1}) - f(x^*)),$$

or equivalently,

$$f(x_{k+1}) - f(x^*) \leq \frac{f(x_k) - f(x^*)}{1 + L\kappa^{\frac{1}{p-1}}} \leq \frac{f(x_1) - f(x^*)}{\big(1 + L\kappa^{\frac{1}{p-1}}\big)^k}. \tag{2.66}$$

Note that by the smoothness of f, as in (2.45), we can write $f(x_1) \leq \min_x\{f(x) + \frac{N+1}{\epsilon p}\|x - x_0\|^p\} \leq f(x^*) + \frac{N+1}{\epsilon p}\|x_0 - x^*\|^p$. Furthermore, since we assume the inverse condition number $\kappa = \epsilon\sigma$ is small, we can write $1 + L\kappa^{\frac{1}{p-1}} \approx \exp(L\kappa^{\frac{1}{p-1}})$. Therefore, (2.66) yields the desired convergence rate (2.64). □

Notice that the result of Theorem 2.10 matches the desired convergence rate $O(\frac{1}{\epsilon}e^{-ck})$ discussed above, with $c = L\kappa^{\frac{1}{p-1}}$.

Exponential convergence rate of rescaled gradient flow

As a side remark, we note that the rescaled gradient flow also has an exponential convergence rate when the objective function f is uniformly convex. However, notice that the following continuous-time convergence rate only depends on the uniform convexity constant of f, whereas the discrete-time convergence rate above also depends on the Lipschitz constant for the higher-order smoothness of f.

Theorem 2.11. *If f is σ-uniformly convex of order p, then the rescaled gradient flow* (2.28) *has convergence rate*

$$f(X_t) - f(x^*) \leq (f(X_0) - f(x^*)) \exp\left(-\sigma^{\frac{1}{p-1}} t\right). \tag{2.67}$$

Proof. As we saw in (2.65), the uniform convexity of f implies the inequality

$$\|\nabla f(X_t)\|_*^{\frac{p}{p-1}} \geq \sigma^{\frac{1}{p-1}} (f(X_t) - f(x^*)).$$

Using this inequality and plugging in the rescaled gradient flow equation (2.28), we have

$$\frac{d}{dt}(f(X_t) - f(x^*)) \;=\; \langle \nabla f(X_t), \dot{X}_t \rangle \;=\; -\|\nabla f(X_t)\|_*^{\frac{p}{p-1}} \;\leq\; -\sigma^{\frac{1}{p-1}} (f(X_t) - f(x^*)).$$

Dividing both sides by $f(X_t) - f(x^*)$ and integrating, we get the desired convergence rate (2.67). □

Exponential convergence rate of accelerated method with restart scheme

We now show that a variant of the accelerated gradient method (2.30) with a restart scheme also attains an exponential convergence rate, with a better dependence on the condition number κ than the higher-order gradient method as in the previous section.

Specifically, we consider the following variant of the accelerated gradient method (2.30),

$$x_{k+1} = \frac{p}{k+p} z_k + \frac{k}{k+p} y_k \tag{2.68a}$$

$$y_k = \arg\min_y \left\{ f_{p-1}(y; x_k) + \frac{2}{\epsilon p} \|y - x_k\|^p \right\} \tag{2.68b}$$

$$z_k = \arg\min_z \left\{ \frac{p}{(4p)^p} \sum_{i=0}^{k} i^{(p-1)} \langle \nabla f(y_i), z \rangle + \frac{2^{p-2}}{\epsilon p} \|z - x_0\|^p \right\}. \tag{2.68c}$$

In (2.68), for simplicity we have explicitly set the constant N in (2.30b) to be $N = 2$, and set C in (2.30c) to be $C = 1/(4p)^p$, which satisfies the condition $C \leq (N^2-1)^{\frac{p-2}{2}}/((2N)^{p-1} p^p)$.

Furthermore, for the z-update (2.68c) we have used the equivalent version (2.35) where we unroll the recursion, and we have also replaced the Bregman divergence in the z-update (2.30c) by the rescaled p-th power $d_p(z) = \frac{2^{p-2}}{p}\|z - x_0\|^p$, which is 1-uniformly convex of order p. The proof of Theorem 2.3 still holds in this case, so we have the guarantee

$$f(y_k) - f(x^*) \;\le\; \frac{(4p)^p \cdot 2^{p-2}\|x_0 - x^*\|^p}{\epsilon p k^{(p)}} \;\le\; \frac{2^{3p-2}p^{p-1}\|x_0 - x^*\|^p}{\epsilon k^p}. \tag{2.69}$$

Then we define the following restart scheme, which proceeds by running the accelerated method (2.68) for some number of iterations at each step,

$$\hat{x}_k \;=\; \big(\text{the output } y_m \text{ of running (2.68) for } m \text{ iterations with input } x_0 = \hat{x}_{k-m}\big). \tag{2.70}$$

Our main result is the following.

Theorem 2.12. *Suppose f is $\frac{(p-1)!}{\epsilon}$-smooth of order $p-1$, and σ-uniformly convex of order p. Let $\hat{x}_k$ be the output of running the restart scheme (2.70) for k/m times with $m = 8p/\kappa^{\frac{1}{p}}$, where $\kappa = \epsilon\sigma$ is the inverse condition number, and let $\hat{y}_k = G_{p,\epsilon,2}(\hat{x}_k)$ be the output of running one step of the gradient update (2.23) with input $\hat{x}_k$. Then we have the convergence rate*

$$f(\hat{y}_k) - f(x^*) \;\le\; \frac{3\|\hat{x}_0 - x^*\|^p}{\epsilon p\, e^{k/m}} \;=\; O\left(\frac{1}{\epsilon}\exp\left(-\frac{\kappa^{\frac{1}{p}}k}{8p}\right)\right). \tag{2.71}$$

Proof. Since f is σ-uniformly convex of order p, and by the bound (2.69), we have

$$\frac{\sigma}{p}\|\hat{x}_k - x^*\|^p \;\le\; f(\hat{x}_k) - f(x^*) \;\le\; \frac{2^{3p-2}p^{p-1}\|\hat{x}_{k-m} - x^*\|^p}{\epsilon m^p} \;\le\; \frac{\sigma}{pe}\|\hat{x}_{k-m} - x^*\|^p, \tag{2.72}$$

where the last inequality follows from our choice of m. Thus, an execution of (2.70) with m iterations of the accelerated method reduces the distance to optimum by a factor of at least $1/e$. Iterating (2.72), we obtain $\|\hat{x}_k - x^*\|^p \le e^{-k/m}\|\hat{x}_0 - x^*\|^p$. To convert this into a bound on the function value, we use the smoothness of f. As noted in (2.45), since $\hat{y}_k$ is the output of one step of the gradient update (2.23) with input $\hat{x}_k$, we have $f(\hat{y}_k) - f(x^*) \le \frac{3}{\epsilon p}\|\hat{x}_k - x^*\|^p$. This gives the desired bound (2.71). □

The result of Theorem 2.12 matches the desired convergence rate $O(\frac{1}{\epsilon}e^{-ck})$ as discussed in Section 2.6 with $c = \frac{1}{8p}\kappa^{\frac{1}{p}}$. Note that this convergence rate has a better dependence on the inverse condition number $\kappa = \epsilon\sigma$ than the higher-order gradient algorithm as in Theorem 2.10, because $\kappa^{\frac{1}{p}} > \kappa^{\frac{1}{p-1}}$ for small κ. This generalizes the conclusion of [46, Section 5] for the case $p = 3$. However, as noted previously, the link to continuous time is not as clear as that of the polynomial family.

2.7 Hessian vs. Bregman Lagrangian

In a Hessian manifold, the metric is generated by the Hessian $\nabla^2 h$ of the distance-generating function h. So for example, the gradient flow equation in the Euclidean case, $\dot{X}_t = -\nabla f(X_t)$, which can be written as

$$\dot{X}_t = \arg\min_{\dot{x}} \left\{ \langle \nabla f(X_t), \dot{x} \rangle + \frac{1}{2}\|\dot{x}\|^2 \right\},$$

in general becomes the natural gradient flow $\dot{X}_t = -[\nabla^2 h(X_t)]^{-1} \nabla f(X_t)$, or equivalently,

$$\dot{X}_t = \arg\min_{\dot{x}} \left\{ \langle \nabla f(X_t), \dot{x} \rangle + \frac{1}{2}\|\dot{x}\|^2_{\nabla^2 h(X_t)} \right\},$$

which is obtained by replacing the Euclidean squared norm $\|v\|^2 = \langle v, v \rangle$ by the Hessian metric

$$\|v\|^2_{\nabla^2 h(x)} := \langle \nabla^2 h(x) v, v \rangle.$$

At the Lagrangian level, recall that a starting point of our work is the differential equation $\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0$ for accelerated gradient descent [61], which we observe is the Euler-Lagrange equation for the damped Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = t^3 \left(\frac{1}{2}\|\dot{x}\|^2 - f(x) \right). \tag{2.73}$$

How should we generalize this Lagrangian to the non-Euclidean case? From our discussion on natural gradient flow, a natural guess is to replace the Euclidean metric in (2.73) by the Hessian metric. Thus, we are led to consider the following family of *Hessian Lagrangians*:

$$\mathcal{L}_{\text{Hess}}(x, \dot{x}, t) \;=\; e^{\gamma_t} \left(\frac{1}{2}\|\dot{x}\|^2_{\nabla^2 h(x)} - e^{\beta_t} f(x) \right) \tag{2.74}$$

where we have also introduced arbitrary weighting functions $\beta_t, \gamma_t \in \mathbb{R}$ ((2.73) is the Euclidean case with $\beta_t = 0, \gamma_t = 3\log t$). However, the Hessian Lagrangian (2.74) turns out to be unsuitable for our optimization purposes. This is because the Euler-Lagrange equation for the Hessian Lagrangian (2.74),

$$\frac{1}{2}\nabla^3 h(X_t)\, \dot{X}_t\, \dot{X}_t \;+\; \nabla^2 h(X_t) \left(\ddot{X}_t + \dot{\gamma}_t\, \dot{X}_t \right) + e^{\beta_t} \nabla f(X_t) = 0, \tag{2.75}$$

involves the third-order derivative $\nabla^3 h$ (which comes from being the derivative of the metric tensor $\nabla^2 h$). This makes the analysis difficult, preventing us from obtaining a convergence rate for (2.75). Furthermore, the presence of $\nabla^3 h$ in the equation makes it difficult to implement as an efficient discrete-time algorithm.

On the other hand, our work shows that the "correct" way to generalize (2.73) to the non-Euclidean case is to use the Bregman divergence, rather than Hessian metric. This results in the general Bregman Lagrangian family (2.1), which requires an additional parameter α_t controlling the amount of interaction between the position x and velocity $\dot{x}$. When the parameters are coupled in an ideal scaling, the Bregman Lagrangian produces dynamics that converge at a provable rate. This is achieved via the design of a corresponding Lyapunov function (the energy functional $\mathcal{E}_t$ (2.8)), whose form is intimately tied to the use of the Bregman divergence in the Lagrangian. Furthermore, for the polynomial family, we can discretize the resulting dynamics as a discrete-time algorithm (2.30) that does not require the Hessian $\nabla^2 h$, but only the gradient ∇h.

It is interesting to consider whether the Hessian Lagrangian (2.74) has useful properties, and how it relates to the Bregman Lagrangian. For a small displacement $\varepsilon > 0$ we know that Bregman divergence approximates the Hessian metric, i.e., $D(x + \varepsilon v, x) \approx \frac{\varepsilon^2}{2}\|v\|^2_{\nabla^2 h(x)}$. Setting $\varepsilon = e^{-\alpha_t}$, this suggests that the Bregman Lagrangian (2.1) is approximating the Hessian Lagrangian $\mathcal{L}_{\text{Hess}}(x, \dot{x}, t) = e^{\gamma_t - \alpha_t}\left(\frac{1}{2}\|\dot{x}\|^2_{\nabla^2 h(x)} - e^{2\alpha_t + \beta_t} f(x)\right)$. However, this argument assumes ϵ is small, whereas in our particular case of interest (the polynomial subfamily in Section 2.2) the value of $\epsilon = e^{-\alpha_t} = \frac{t}{p}$ is growing over time.

2.8 Gradient vs. Lagrangian flows

In the Euclidean case, we can think of gradient flow as describing the behavior of a damped Lagrangian system "in an asymptotic regime in which dissipative effects play such an important role, that the effects of forcing and dissipation compensate each other" [67, p. 646]. That is, the gradient flow equation $\dot{X}_t = -\nabla f(X_t)$ can be seen as the strong-friction limit $\lambda \to \infty$ of the equation $\ddot{X}_t + \lambda \dot{X}_t + \lambda \nabla f(X_t) = 0$.

This is perhaps more apparent if we define $m = 1/\lambda$ to be the "mass" of the fictitious particle, so the equation of motion becomes

$$m\ddot{X}_t + \dot{X}_t + \nabla f(X_t) = 0, \tag{2.76}$$

which is the Euler-Lagrange equation of the damped Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = e^{t/m}\left(\frac{m}{2}\|\dot{x}\|^2 - f(x)\right), \tag{2.77}$$

where the damping factor $e^{t/m}$ also scales with m. In the massless limit $m \to 0$, we indeed recover gradient flow from (2.76). In the following, we show that this result also holds more generally, both for natural gradient flow (as the massless limit of a Bregman Lagrangian flow) and for the rescaled gradient flow (as the massless limit of a Lagrangian flow which uses the p-th power of the norm).

However, notice that in all these cases, the momentum variable $p = \frac{\partial \mathcal{L}}{\partial \dot{x}}$ becomes infinite as $m \to 0$. For instance, $p = me^{t/m}\dot{x}$ for (2.77), and $me^{t/m} \to \infty$. This means as $m \to 0$, the

particle also becomes more massive and has more inertia. Thus, gradient flow is the limiting case where the infinitely massive particle simply rolls downhill and stops at the minimum x^* as soon as the force $-\nabla f$ vanishes, without oscillation (which is damped by the infinitely strong friction). In this view, moving from a first-order gradient algorithm to a second-order Lagrangian (accelerated) algorithm does *not* amount to preventing oscillation; rather, it is the opposite, by unwinding the curve to finite momentum where it can travel faster, albeit now with some oscillation.

Natural gradient flow as massless limit

Consider the following Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = \frac{e^{t/m}}{m} \left(D_h(x + m\dot{x}, x) - m f(x) \right), \tag{2.78}$$

which is the Bregman Lagrangian (2.1) with parameters $\alpha_t = -\log m$, $\beta_t = \log m$, and $\gamma_t = t/m$ (which satisfy the ideal scaling (2.2)). Note that (2.78) recovers (2.77) in the Euclidean case. The Euler-Lagrange equation (2.6) for the Lagrangian (2.78) is given by

$$\ddot{X}_t + \frac{1}{m}\dot{X}_t + \frac{1}{m}\left[\nabla^2 h(X_t + m\dot{X}_t)\right]^{-1} \nabla f(X_t) = 0.$$

Multiplying the equation by m and letting $m \to 0$, we recover

$$\dot{X}_t + \left[\nabla^2 h(X_t)\right]^{-1} \nabla f(X_t) = 0,$$

which is the natural gradient flow equation. In this case the momentum variable is $p = \frac{\partial \mathcal{L}}{\partial \dot{x}} = e^{t/m}(\nabla h(x + m\dot{x}) - \nabla h(x)) \approx m e^{t/m} \nabla^2 h(x)\dot{x}$, so we still have $p \to \infty$ as $m \to 0$.

Rescaled gradient flow as massless limit

Consider the following Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = e^{t/m} \left(\frac{m}{p} \|\dot{x}\|^p - f(x) \right), \tag{2.79}$$

where we use the p-th power of the norm to measure the kinetic energy. Note that (2.79) recovers (2.77) in the case $p = 2$. The Euler-Lagrange equation is

$$\|\dot{X}_t\|^{p-2}\big(m\ddot{X}_t + \dot{X}_t\big) + m(p-2)\|\dot{X}_t\|^{p-4}\langle \ddot{X}_t, \dot{X}_t\rangle \, \dot{X}_t + \nabla f(X_t) = 0.$$

So as $m \to 0$, this equation recovers

$$\|\dot{X}_t\|^{p-2}\dot{X}_t + \nabla f(X_t) = 0 \tag{2.80}$$

which is equivalent to the rescaled gradient flow (2.28). In this case the momentum variable is $p = \frac{\partial \mathcal{L}}{\partial \dot{x}} = m e^{t/m} \|\dot{x}\|^{p-2}\dot{x}$, which still goes to infinity as $m \to 0$.

2.9 Bregman Hamiltonian

In this section we define and compute the Bregman Hamiltonian corresponding to the Bregman Lagrangian. In general, given a Lagrangian $\mathcal{L}(x, \dot{x}, t)$, its Hamiltonian is defined by

$$\mathcal{H}(x, p, t) = \langle p, \dot{x} \rangle - \mathcal{L}(x, \dot{x}, t) \tag{2.81}$$

where $p = \frac{\partial \mathcal{L}}{\partial \dot{x}}$ is the momentum variable conjugate to position.

For the Bregman Lagrangian (2.1), the momentum variable is given by

$$p = \frac{\partial \mathcal{L}}{\partial \dot{x}} = e^{\gamma_t} \left(\nabla h(x + e^{-\alpha_t} \dot{x}) - \nabla h(x) \right). \tag{2.82}$$

We can invert this equation to solve for the velocity $\dot{x}$,

$$\dot{x} \;=\; e^{\alpha_t} \left(\nabla h^*(\nabla h(x) + e^{-\gamma_t} p) - x \right), \tag{2.83}$$

where h^* is the conjugate function to h (recall the definition in (2.33)), and we have used the property that $\nabla h^* = [\nabla h]^{-1}$. So for the first term in the definition (2.81) we have

$$\langle p, \dot{x} \rangle = e^{\alpha_t} \big\langle p,\; \nabla h^*(\nabla h(x) + e^{-\gamma_t} p) - x \big\rangle.$$

Next, we write the Bregman Lagrangian $\mathcal{L}(x, \dot{x}, t)$ in terms of (x, p, t). We can directly substitute (2.83) to the definition (2.1) and calculate the result. Alternatively, we can use the property that the Bregman divergences of h and h^* satisfy $D_h(y, x) = D_{h^*}(\nabla h(x), \nabla h(y))$. Therefore, we can write the Bregman Lagrangian (2.1) as

$$\begin{aligned}
\mathcal{L}(x, \dot{x}, t) &= e^{\alpha_t + \gamma_t} \left(D_{h^*} \left(\nabla h(x),\, \nabla h(x + e^{-\alpha_t} \dot{x}) \right) - e^{\beta_t} f(x) \right) \\
&= e^{\alpha_t + \gamma_t} \left(D_{h^*} \left(\nabla h(x),\, \nabla h(x) + e^{-\gamma_t} p \right) - e^{\beta_t} f(x) \right) \\
&= \; e^{\alpha_t + \gamma_t} \left(h^*(\nabla h(x)) - h^*(\nabla h(x) + e^{-\gamma_t} p) + e^{-\gamma_t} \langle \nabla h^*(\nabla h(x) + e^{-\gamma_t} p), p \rangle - e^{\beta_t} f(x) \right),
\end{aligned}$$

where in the second step we have used the relation $\nabla h(x + e^{-\alpha_t} \dot{x}) = \nabla h(x) + e^{-\gamma_t} p$ from (2.82), and in the last step we have expanded the Bregman divergence.

Substituting these calculations into (2.81) and simplifying, we get the Hamiltonian

$$\mathcal{H}(x, p, t) \;=\; e^{\alpha_t + \gamma_t} \left(h^*(\nabla h(x) + e^{-\gamma_t} p) - h^*(\nabla h(x)) - \langle x,\, e^{-\gamma_t} p \rangle + e^{\beta_t} f(x) \right).$$

Since $x = \nabla h^*(\nabla h(x))$, we can also write this result in terms of the Bregman divergence of h^*,

$$\mathcal{H}(x, p, t) \;=\; e^{\alpha_t + \gamma_t} \left(D_{h^*}(\nabla h(x) + e^{-\gamma_t} p,\, \nabla h(x)) + e^{\beta_t} f(x) \right). \tag{2.84}$$

We call the Hamiltonian (2.84) the *Bregman Hamiltonian*. Notice that whereas the Bregman Lagrangian takes the form of the difference between the kinetic and potential energy, the Bregman Hamiltonian takes the form of the sum of the kinetic and potential energy. (However, note that the kinetic energy is slightly different: it is $D_{h^*}(\nabla h(x) + e^{-\gamma_t} p,\, \nabla h(x)) = D_h(x, x + e^{-\alpha_t} \dot{x})$ in the Hamiltonian (2.84), while it is $D_h(x + e^{-\alpha_t} \dot{x}, x)$ in the Lagrangian (2.1).)

Hamiltonian equations of motion

The second-order Euler-Lagrange equation of a Lagrangian can be equivalently written as a pair of first-order equations

$$\dot{X}_t \;=\; \frac{\partial \mathcal{H}}{\partial p}(X_t, P_t, t), \qquad \dot{P}_t \;=\; -\frac{\partial \mathcal{H}}{\partial x}(X_t, P_t, t). \tag{2.85}$$

For the Bregman Hamiltonian (2.84), the equations of motion are given by

$$\dot{X}_t \;=\; e^{\alpha_t}\left(\nabla h^*(\nabla h(X_t) + e^{-\gamma_t} P_t) - X_t\right) \tag{2.86a}$$

$$\dot{P}_t \;=\; -e^{\alpha_t+\gamma_t}\nabla^2 h(X_t)\left(\nabla h^*(\nabla h(X_t) + e^{-\gamma_t} P_t) - X_t\right) + e^{\alpha_t} P_t - e^{\alpha_t+\beta_t+\gamma_t}\nabla f(X_t). \tag{2.86b}$$

Notice that the first equation (2.86a) recovers the definition of momentum (2.82). Furthermore, when $\dot{\gamma}_t = e^{\alpha_t}$, by substituting (2.86a) to (2.86b) we can write (2.86) as

$$\frac{d}{dt}\left\{\nabla h(X_t) + e^{-\gamma_t} P_t\right\} \;=\; \nabla^2 h(X_t)\,\dot{X}_t - \dot{\gamma}_t e^{-\gamma_t} P_t + e^{-\gamma_t}\dot{P}_t \;=\; -e^{\alpha_t+\beta_t}\nabla f(X_t).$$

Since $\nabla h(X_t) + e^{-\gamma_t} P_t = \nabla h(X_t + e^{-\alpha_t}\dot{X}_t)$ by (2.86a), this indeed recovers the Euler-Lagrange equation (2.7).

A Lyapunov function for the Hamiltonian equations of motion (2.86) is the following, which is simply the energy functional (2.8) written in terms of (X_t, P_t, t),

$$\mathcal{E}_t \;=\; D_{h^*}\left(\nabla h(X_t) + e^{-\gamma_t} P_t,\, \nabla h(x^*)\right) + e^{\beta_t}(f(X_t) - f(x^*)).$$

The Hamiltonian formulation of the dynamics has appealing properties that seem worthy of further exploration. For example, Hamiltonian flow preserves volume in phase space (Liouville's theorem); this property has been used in the context of sampling to develop the technique of Hamiltonian Markov chain Monte-Carlo, and may also be useful to help us design better algorithms for optimization. Furthermore, the Hamilton-Jacobi-Bellman equation (which is a reformulation of the Hamiltonian dynamics) is a central object of study in the field of optimal control theory, and it would be interesting to study the Bregman Hamiltonian framework from that perspective.

2.10 Gauge invariance

The Euler-Lagrange equation of a Lagrangian is gauge-invariant, which means it does not change when we transform the Lagrangian by adding a total time derivative,

$$\mathcal{L}'(X_t, \dot{X}_t, t) = \mathcal{L}(X_t, \dot{X}_t, t) + \frac{d}{dt} G(X_t, t) \tag{2.87}$$

for any smooth function G. We can show this by directly checking that the Euler-Lagrange equation of $\mathcal{L}'$ is the same as that of $\mathcal{L}$. Alternatively, this follows from the formulation

of the principle of least action, where we fix two points (x_0, t_0) and (x_1, t_1), and ask for a curve X joining the two endpoints ($X_{t_0} = x_0$ and $X_{t_1} = x_1$) that minimizes the action $J(X) = \int_{t_0}^{t_1} \mathcal{L}(X_t, \dot{X}_t, t)dt$. Thus, when the Lagrangian transforms as (2.87), the action only changes to $J'(X) = J(X) + \int_{t_0}^{t_1} \frac{d}{dt} G(X_t, t)dt = J(X) + G(x_1, t_1) - G(x_0, t_0)$. Since (x_0, t_0) and (x_1, t_1) are fixed, this means the new action only differs from the old action by a constant; this implies that the optimal least action curve—namely, the Euler-Lagrange equation—does not change.

In our case, under the ideal scaling condition $\dot{\gamma}_t = e^{\alpha_t}$ (2.2b), this property implies that the Bregman Lagrangian (2.1) is equivalent to the following Lagrangian

$$\mathcal{L}'(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(h(x + e^{-\alpha_t} \dot{x}) - e^{\beta_t} f(x) \right), \tag{2.88}$$

where we have replaced the Bregman divergence $D_h(x + e^{-\alpha_t}\dot{x}, x)$ by its first term $h(x + e^{-\alpha_t}\dot{x})$. Indeed, we can check that the difference between the Bregman Lagrangian (2.1) and the reduced form (2.88) is a total time derivative,

$$\mathcal{L}'(X_t, \dot{X}_t, t) - \mathcal{L}(X_t, \dot{X}_t, t) = e^{\gamma_t + \alpha_t} \left(h(X_t) + \langle \nabla h(X_t), e^{-\alpha_t} \dot{X}_t \rangle \right) = \frac{d}{dt} \left\{ e^{\gamma_t} h(X_t) \right\}$$

where the last step follows from the ideal scaling $e^{\alpha_t} = \dot{\gamma}_t$.

The reduced Lagrangian (2.88) is slightly simpler than the Bregman Lagrangian (2.1), and in a sense it makes the roles of h and f more symmetric. It also suggests that the role of h is not so much as measuring the distance via the Hessian metric or Bregman divergence, but rather, as evaluating the extrapolated future point $X_t + e^{-\alpha_t} \dot{X}_t$.

2.11 Natural motion

A natural motion is the motion of a particle when it experiences no force. In the physical world, the natural motion of a particle is a straight-line motion with constant velocity. But for the Bregman Lagrangian, which describes a dissipative system, the natural motion always converges.

Specifically, the Bregman Lagrangian (2.1) in the case of zero (or constant) potential function $f \equiv 0$ is $\mathcal{L}(x, \dot{x}, t) = e^{\alpha_t + \gamma_t} D_h(x + e^{-\alpha_t}\dot{x}, x)$. Assuming the ideal scaling $\dot{\gamma}_t = e^{\alpha_t}$ (2.2b), its Euler-Lagrange equation is given by (2.7), which in this case is

$$\frac{d}{dt} \nabla h(X_t + e^{-\alpha_t} \dot{X}_t) = 0. \tag{2.89}$$

This means $\nabla h(X_t + e^{-\alpha_t}\dot{X}_t)$ is a constant, say $\nabla h(X_t + e^{-\alpha_t}\dot{X}_t) = \nabla h(b)$ for some $b \in \mathcal{X}$. Applying $\nabla h^* = [\nabla h]^{-1}$ to both sides gives us $X_t + e^{-\alpha_t}\dot{X}_t = b$. Since $e^{\alpha_t} = \dot{\gamma}_t$, we can write this as

$$\frac{d}{dt} \left\{ e^{\gamma_t} (X_t - b) \right\} = e^{\alpha_t + \gamma_t} (X_t - b) + e^{\gamma_t} \dot{X}_t = 0.$$

This means $e^{\gamma_t}(X_t - b)$ is a constant, say $e^{\gamma_t}(X_t - b) = a$ for some $a \in \mathcal{X}$. Thus, we conclude that the natural motion of the Bregman Lagrangian is

$$X_t = ae^{-\gamma_t} + b. \tag{2.90}$$

Notice that the natural motion is independent of h, although the Lagrangian still depends on h. Furthermore, in contrast with the straight-line motion, the natural motion (2.90) always converges; in particular, if we assume $e^{\gamma_t} \to \infty$ as $t \to \infty$, then $X_t \to b$.

The natural motion (2.90) has simple explicit invariance and symmetry properties. Indeed, (2.89) states that $\nabla h(X_t + e^{-\alpha_t}\dot{X}_t)$ is a conserved quantity, which is always equal to $\nabla h(b)$. By Noether's theorem, any conservation law corresponds to a symmetry of the Lagrangian. In our case, the corresponding symmetry is the transformation

$$X_t' = X_t + e^{-\gamma_t}u, \tag{2.91}$$

for any $u \in \mathcal{X}$. Under this transformation, $\dot{X}_t$ changes to $\dot{X}_t' = \dot{X}_t - \dot{\gamma}_t e^{-\gamma_t}u$. Since $\dot{\gamma}_t = e^{\alpha_t}$, this implies $X_t' + e^{-\alpha_t}\dot{X}_t' = X_t + e^{-\alpha_t}\dot{X}_t$. This means the reduced Lagrangian $\mathcal{L}(X_t, \dot{X}_t, t) = e^{\gamma_t + \alpha_t}h(X_t + e^{-\alpha_t}\dot{X}_t)$ is invariant under the transformation (2.91). Therefore, the Bregman Lagrangian (which is gauge-equivalent to the reduced Lagrangian) is also invariant. So indeed (2.91) is a symmetry of the Bregman Lagrangian when $f = 0$.

2.12 The Euclidean case

In the Euclidean case many of our results and equations simplify, as we summarize in this section. When h is the squared Euclidean norm, $h(x) = \frac{1}{2}\|x\|^2$, the Bregman divergence is also the squared norm and it coincides with the Hessian metric, $D_h(y, x) = \frac{1}{2}\|y - x\|^2 = \frac{1}{2}\|y - x\|^2_{\nabla^2 h(x)}$. Furthermore, $h^* = h$ and both $\nabla h, \nabla h^*$ are the identity function.

In the Euclidean case, the Bregman Lagrangian (2.1) becomes

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t - \alpha_t}\left(\frac{1}{2}\|\dot{x}\|^2 - e^{2\alpha_t + \beta_t}f(x)\right).$$

For general $\alpha_t, \beta_t, \gamma_t$, the Euler-Lagrange equation (2.5) is given by

$$\ddot{X}_t + (\dot{\gamma}_t - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t}\nabla f(X_t) = 0.$$

When the ideal scaling $\dot{\gamma}_t = e^{\alpha_t}$ (2.2b) holds, this equation becomes

$$\ddot{X}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t}\nabla f(X_t) = 0, \tag{2.92}$$

which we can equivalently write as $\frac{d}{dt}(X_t + e^{-\alpha_t}\dot{X}_t) = -e^{\alpha_t + \beta_t}\nabla f(X_t)$. The energy functional (2.8) for proving the rate of convergence becomes

$$\mathcal{E}_t = \frac{1}{2}\|X_t + e^{-\alpha_t}\dot{X}_t - x^*\|^2 + e^{\beta_t}(f(X_t) - f(x^*)).$$

The Bregman Hamiltonian (2.84) becomes

$$\mathcal{H}(x,p,t) = e^{\alpha_t - \gamma_t}\left(\frac{1}{2}\|p\|^2 + e^{2\gamma_t+\beta_t}f(x)\right),$$

where the momentum variable (2.82) is given by $p = e^{\gamma_t-\alpha_t}\dot{x}$. The Hamiltonian equations of motion (2.86) simplify to

$$\dot{X}_t = e^{\alpha_t-\gamma_t}P_t \tag{2.93}$$

$$\dot{P}_t = -e^{\alpha_t+\beta_t+\gamma_t}\nabla f(X_t). \tag{2.94}$$

In particular, for the polynomial case with the parameters (2.12), the Euler-Lagrange equation (2.92) is given by

$$\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + Cp^2t^{p-2}\nabla f(X_t) = 0, \tag{2.95}$$

with an $O(1/t^p)$ rate of convergence. For $p = 2$, this recovers the differential equation $\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0$ corresponding to Nesterov's accelerated gradient descent, as derived in [61].

Su et al. [61] observed that the generalized equation $\ddot{X}_t + \frac{r}{t}\dot{X}_t + \nabla f(X_t) = 0$ still has convergence rate $O(1/t^2)$ whenever $r \geq 3$, and they posed the question on the significance of the threshold $r = 3$. Our results give the following perspective: The equation $\ddot{X}_t + \frac{r}{t}\dot{X}_t + \nabla f(X_t) = 0$ is the case of (2.92) with parameters $\alpha_t = \log(r-1) - \log t$, $\gamma_t = (r-1)\log t$, and $\beta_t = 2\log t - 2\log(r-1)$. These parameters satisfy the ideal scaling condition (2.2) when $r \geq 3$, so Theorem 2.1 guarantees a convergence rate of $O(e^{-\beta_t}) = O(1/t^2)$. However, for a fixed $r > 3$, the choice of $\beta_t = 2\log t - 2\log(r-1)$ is suboptimal, since from the ideal scaling condition $\dot{\beta}_t \leq e^{\alpha_t}$ we know we can increase β_t up to $(r-1)\log t$. This will introduce a factor of t^{r-3} on the force term, as in (2.95), but it will also yield a faster convergence rate of $O(1/t^{r-1})$.

Chapter 3

Concavity of Reweighted Kikuchi Approximation

Undirected graphical models are a familiar framework in diverse application domains such as computer vision, statistical physics, coding theory, social science, and epidemiology. In certain settings of interest, one is provided with potential functions defined over nodes and (hyper)edges of the graph. A crucial step in probabilistic inference is to compute the log partition function of the distribution based on these potential functions for a given graph structure. However, computing the log partition function either exactly or approximately is NP-hard in general [10, 56]. An active area of research involves finding accurate approximations of the log partition function and characterizing the graph structures for which such approximations may be computed efficiently [76, 69, 22, 62, 72, 57].

When the underlying graph is a tree, the log partition function may be computed exactly via the sum product algorithm in time linear in the number of nodes [53]. However, when the graph contains cycles, a generalized version of the sum product algorithm known as loopy belief propagation may either fail to converge or terminate in local optima of a nonconvex objective function [73, 63, 26, 42].

In this chapter, we analyze the Kikuchi approximation method, which is constructed from a variational representation of the log partition function by replacing the entropy with an expression that decomposes with respect to a region graph. Kikuchi approximations were previously introduced in the physics literature [31] and reformalized by Yedidia et al. [77, 76] and others [3, 52] in the language of graphical models. The Bethe approximation, which is a special case of the Kikuchi approximation when the region graph has only two layers, has been studied by various authors [12, 77, 24, 72]. In addition, a reweighted version of the Bethe approximation was proposed by Wainwright et al. [69, 55]. As described in Vontobel [68], computing the global optimum of the Bethe variational problem may in turn be used to approximate the permanent of a nonnegative square matrix.

The particular objective function that we study generalizes the Kikuchi objective appearing in previous literature by assigning arbitrary weights to individual terms in the Kikuchi entropy expansion. We establish necessary and sufficient conditions under which this class of

objective functions is concave, so a global optimum may be found efficiently. Our theoretical results synthesize known results on Kikuchi and Bethe approximations, and our main theorem concerning concavity conditions for the reweighted Kikuchi entropy recovers existing results when specialized to the unweighted Kikuchi [52] or reweighted Bethe [69] case. Furthermore, we provide a valuable converse result in the reweighted Bethe case, showing that when our concavity conditions are violated, the entropy function cannot be concave over the whole feasible region. As demonstrated by our experiments, a message-passing algorithm designed to optimize the Kikuchi objective may terminate in local optima for weights outside the concave region. Watanabe and Fukumizu [71, 72] provide a similar converse in the unweighted Bethe case, but our proof is much simpler and our result is more general.

In the reweighted Bethe setting, we also present a useful characterization of the concave region of the Bethe entropy function in terms of the geometry of the graph. Specifically, we show that if the region graph consists of only singleton vertices and pairwise edges, then the region of concavity coincides with the convex hull of incidence vectors of single-cycle forest subgraphs of the original graph. When the region graph contains regions with cardinality greater than two, the latter region may be strictly contained in the former; however, our result provides a useful way to generate weight vectors within the region of concavity. Whereas Wainwright et al. [69] establish the concavity of the reweighted Bethe objective on the spanning forest polytope, that region is contained within the single-cycle forest polytope, and our simulations show that generating weight vectors in the latter polytope may yield closer approximations to the log partition function.

The remainder of the chapter is organized as follows: In Section 3.1, we review background information about the Kikuchi and Bethe approximations. In Section 3.2, we provide our main results on concavity conditions for the reweighted Kikuchi approximation, including a geometric characterization of the region of concavity in the Bethe case. Section 3.3 outlines the reweighted sum product algorithm and proves that fixed points correspond to global optima of the Kikuchi approximation. Section 3.4 presents experiments showing the improved accuracy of the reweighted Kikuchi approximation over the region of concavity. Technical proofs and additional simulations are contained in Sections 3.6–3.10.

3.1 Background and problem setup

In this section, we review basic concepts of the Kikuchi approximation and establish some terminology to be used in the chapter.

Let $G = (V, R)$ denote a *region graph* defined over the vertex set V, where each region $r \in R$ is a subset of V. Directed edges correspond to inclusion, so $r \to s$ is an edge of G if $s \subseteq r$. We use the following notation, for $r \in R$:

$$\begin{aligned} \mathcal{A}(r) &:= \{s \in R \colon r \subsetneq s\} && (\textit{ancestors} \text{ of } r) \\ \mathcal{F}(r) &:= \{s \in R \colon r \subseteq s\} && (\textit{forebears} \text{ of } r) \\ N(r) &:= \{s \in R \colon r \subseteq s \text{ or } s \subseteq r\} && (\textit{neighbors} \text{ of } r). \end{aligned}$$

For $R' \subseteq R$, we define $\mathcal{A}(R') = \bigcup_{r \in R'} \mathcal{A}(r)$, and we define $\mathcal{F}(R')$ and $N(R')$ similarly.

We consider joint distributions $x = (x_s)_{s \in V}$ that factorize over the region graph; i.e.,

$$p(x) = \frac{1}{Z(\alpha)} \prod_{r \in R} \alpha_r(x_r), \tag{3.1}$$

for potential functions $\alpha_r > 0$. Here, $Z(\alpha)$ is the normalization factor, or partition function, which is a function of the potential functions α_r, and each variable x_s takes values in a finite discrete set $\mathcal{X}$. One special case of the factorization (3.1) is the pairwise Ising model, defined over a graph $G = (V, E)$, where the distribution is given by

$$p_\gamma(x) = \exp\Big(\sum_{s \in V} \gamma_s(x_s) + \sum_{(s,t) \in E} \gamma_{st}(x_s, x_t) - A(\gamma) \Big), \tag{3.2}$$

and $\mathcal{X} = \{-1, +1\}$. Our goal is to analyze the log partition function

$$\log Z(\alpha) = \log \Big\{ \sum_{x \in \mathcal{X}^{|V|}} \prod_{r \in R} \alpha_r(x_r) \Big\}. \tag{3.3}$$

Variational representation

It is known from the theory of graphical models [52] that the log partition function (3.3) may be written in the variational form

$$\log Z(\alpha) = \sup_{\{\tau_r(x_r)\} \in \Delta_R} \Big\{ \sum_{r \in R} \sum_{x_r} \tau_r(x_r) \log(\alpha_r(x_r)) + H(p_\tau) \Big\}, \tag{3.4}$$

where p_τ is the maximum entropy distribution with marginals $\{\tau_r(x_r)\}$ and

$$H(p) := -\sum_x p(x) \log p(x)$$

is the usual entropy. Here, Δ_R denotes the R-marginal polytope; i.e., $\{\tau_r(x_r) \colon r \in R\} \in \Delta_R$ if and only if there exists a distribution $\tau(x)$ such that $\tau_r(x_r) = \sum_{x_{\setminus r}} \tau(x_r, x_{\setminus r})$ for all r. For ease of notation, we also write $\tau \equiv \{\tau_r(x_r) \colon r \in R\}$. Let $\theta \equiv \theta(x)$ denote the collection of log potential functions $\{\log(\alpha_r(x_r)) \colon r \in R\}$. Then equation (3.4) may be rewritten as

$$\log Z(\theta) = \sup_{\tau \in \Delta_R} \left\{ \langle \theta, \tau \rangle + H(p_\tau) \right\}. \tag{3.5}$$

Specializing to the Ising model (3.2), equation (3.5) gives the variational representation

$$A(\gamma) = \sup_{\mu \in \mathbb{M}} \left\{ \langle \gamma, \mu \rangle + H(p_\mu) \right\}, \tag{3.6}$$

which appears in Wainwright and Jordan [70]. Here, $\mathbb{M} \equiv \mathbb{M}(G)$ denotes the marginal polytope, corresponding to the collection of mean parameter vectors of the sufficient statistics in the exponential family representation (3.2), ranging over different values of γ, and p_μ is the maximum entropy distribution with mean parameters μ.

Reweighted Kikuchi approximation

Although the set Δ_R appearing in the variational representation (3.5) is a convex polytope, it may have exponentially many facets [70]. Hence, we replace Δ_R with the set

$$\Delta_R^K = \Big\{\tau \colon \forall t, u \in R \text{ s.t. } t \subseteq u, \sum_{x_{u\setminus t}} \tau_u(x_t, x_{u\setminus t}) = \tau_t(x_t) \quad \text{and} \quad \forall u \in R, \sum_{x_u} \tau_u(x_u) = 1\Big\}$$

of *locally consistent* R-pseudomarginals. Note that $\Delta_R \subseteq \Delta_R^K$ and the latter set has only polynomially many facets, making optimization more tractable.

In the case of the pairwise Ising model (3.2), we let $\mathbb{L} \equiv \mathbb{L}(G)$ denote the polytope Δ_R^K. Then $\mathbb{L}$ is the collection of nonnegative functions $\tau = (\tau_s, \tau_{st})$ satisfying the marginalization constraints

$$\begin{array}{ll} \sum_{x_s} \tau_s(x_s) = 1, & \forall s \in V, \\ \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \;\text{ and }\; \sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t), & \forall (s,t) \in E. \end{array}$$

Recall that $\mathbb{M}(G) \subseteq \mathbb{L}(G)$, with equality achieved if and only if the underlying graph G is a tree. In the general case, we have $\Delta_R = \Delta_R^K$ when the Hasse diagram of the region graph admits a minimal representation that is loop-free (cf. Theorem 2 of Pakzad and Anantharam [52]).

Given a collection of R-pseudomarginals τ, we also replace the entropy term $H(p_\tau)$, which is difficult to compute in general, by the approximation

$$H(p_\tau) \approx \sum_{r \in R} \rho_r H_r(\tau_r) := H(\tau; \rho), \tag{3.7}$$

where $H_r(\tau_r) := -\sum_{x_r} \tau_r(x_r) \log \tau_r(x_r)$ is the entropy computed over region r, and $\{\rho_r \colon r \in R\}$ are weights assigned to the regions. Note that in the pairwise Ising case (3.2), with $p := p_\gamma$, we have the equality

$$H(p) = \sum_{s \in V} H_s(p_s) - \sum_{(s,t) \in E} I_{st}(p_{st})$$

when G is a tree, where $I_{st}(p_{st}) = H_s(p_s) + H_t(p_t) - H_{st}(p_{st})$ denotes the mutual information and p_s and p_{st} denote the node and edge marginals. Hence, the approximation (3.7) is exact with

$$\rho_{st} = 1, \quad \forall (s,t) \in E, \qquad \text{and} \qquad \rho_s = 1 - \deg(s), \quad \forall s \in V.$$

Using the approximation (3.7), we arrive at the following *reweighted Kikuchi approximation*:

$$B(\theta; \rho) := \sup_{\tau \in \Delta_R^K} \underbrace{\big\{\langle \theta, \tau \rangle + H(\tau; \rho)\big\}}_{B_{\theta,\rho}(\tau)}. \tag{3.8}$$

Note that when $\{\rho_r\}$ are the *overcounting numbers* $\{c_r\}$, defined recursively by

$$c_r = 1 - \sum_{s \in \mathcal{A}(r)} c_s, \tag{3.9}$$

the expression (3.8) reduces to the usual (unweighted) Kikuchi approximation considered in Pakzad and Anantharam [52].

3.2 Main results and consequences

In this section, we analyze the concavity of the Kikuchi variational problem (3.8). We derive a sufficient condition under which the function $B_{\theta,\rho}(\tau)$ is concave over the set Δ_R^K, so global optima of the reweighted Kikuchi approximation may be found efficiently. In the Bethe case, we also show that the condition is necessary for $B_{\theta,\rho}(\tau)$ to be concave over the entire region Δ_R^K, and we provide a geometric characterization of Δ_R^K in terms of the edge and cycle structure of the graph.

Sufficient conditions for concavity

We begin by establishing sufficient conditions for the concavity of $B_{\theta,\rho}(\tau)$. Clearly, this is equivalent to establishing conditions under which $H(\tau;\rho)$ is concave. Our main result is the following:

Theorem 3.1. *If $\rho \in \mathbb{R}^{|R|}$ satisfies*

$$\sum_{s \in \mathcal{F}(S)} \rho_s \geq 0, \qquad \forall S \subseteq R, \tag{3.10}$$

then the Kikuchi entropy $H(\tau;\rho)$ is strictly concave on Δ_R^K.

The proof of Theorem 3.1 is presented in Section 3.6, and makes use of a generalization of Hall's marriage lemma for weighted graphs (cf. Lemma 3.8 in Section 3.6).

The condition (3.10) depends heavily on the structure of the region graph. For the sake of interpretability, we now specialize to the case where the region graph has only two layers, with the first layer corresponding to vertices and the second layer corresponding to hyperedges. In other words, for $r, s \in R$, we have $r \subseteq s$ only if $|r| = 1$, and $R = V \cup F$, where F is the set of hyperedges and V denotes the set of singleton vertices. This is the *Bethe case*, and the entropy

$$H(\tau;\rho) = \sum_{s \in V} \rho_s H_s(\tau_s) + \sum_{\alpha \in F} \rho_\alpha H_\alpha(\tau_\alpha) \tag{3.11}$$

is consequently known as the Bethe entropy.

The following result is proved in Section 3.6:

Corollary 3.2. *Suppose $\rho_\alpha \geq 0$ for all $\alpha \in F$, and the following condition also holds:*

$$\sum_{s \in U} \rho_s + \sum_{\alpha \in F \colon \alpha \cap U \neq \emptyset} \rho_\alpha \geq 0, \qquad \forall U \subseteq V. \tag{3.12}$$

Then the Bethe entropy $H(\tau; \rho)$ is strictly concave over Δ_R^K.

Necessary conditions for concavity

We now establish a converse to Corollary 3.2 in the Bethe case, showing that condition (3.12) is also necessary for the concavity of the Bethe entropy. When $\rho_\alpha = 1$ for $\alpha \in F$ and $\rho_s = 1 - |N(s)|$ for $s \in V$, we recover the result of Watanabe and Fukumizu [72] for the unweighted Bethe case. However, our proof technique is significantly simpler and avoids the complex machinery of graph zeta functions. Our approach proceeds by considering the Bethe entropy $H(\tau; \rho)$ on appropriate slices of the domain Δ_R^K so as to extract condition (3.12) for each $U \subseteq V$. The full proof is provided in Section 3.7.

Theorem 3.3. *If the Bethe entropy $H(\tau; \rho)$ is concave over Δ_R^K, then $\rho_\alpha \geq 0$ for all $\alpha \in F$, and condition* (3.12) *holds.*

Indeed, as demonstrated in the simulations of Section 3.4, the Bethe objective function $B_{\theta,\rho}(\tau)$ may have multiple local optima if ρ does *not* satisfy condition (3.12).

Polytope of concavity

We now characterize the polytope defined by the inequalities (3.12). We show that in the pairwise Bethe case, the polytope may be expressed geometrically as the convex hull of single-cycle forests formed by the edges of the graph. In the more general (non-pairwise) Bethe case, however, the polytope of concavity may strictly contain the latter set.

Note that the Bethe entropy (3.11) may be written in the alternative form

$$H(\tau; \rho) = \sum_{s \in V} \rho'_s H_s(\tau_s) - \sum_{\alpha \in F} \rho_\alpha \widetilde{I}_\alpha(\tau_\alpha), \tag{3.13}$$

where $\widetilde{I}_\alpha(\tau_\alpha) := \{\sum_{s \in \alpha} H_s(\tau_s)\} - H_\alpha(\tau_\alpha)$ is the KL divergence between the joint distribution τ_α and the product distribution $\prod_{s \in \alpha} \tau_s$, and the weights ρ'_s are defined appropriately.

We show that the polytope of concavity has a nice geometric characterization when $\rho'_s = 1$ for all $s \in V$, and $\rho_\alpha \in [0, 1]$ for all $\alpha \in F$. Note that this assignment produces the expression for the reweighted Bethe entropy analyzed in Wainwright et al. [69] (when all elements of F have cardinality two). Equation (3.13) then becomes

$$H(\tau; \rho) = \sum_{s \in V} \Big(1 - \sum_{\alpha \in N(s)} \rho_\alpha\Big) H_s(\tau_s) + \sum_{\alpha \in F} \rho_\alpha H_\alpha(\tau_\alpha), \tag{3.14}$$

and the inequalities (3.12) defining the polytope of concavity are

$$\sum_{\alpha \in F\colon \alpha \cap U \neq \emptyset} (|\alpha \cap U| - 1)\rho_\alpha \leq |U|, \quad \forall U \subseteq V. \tag{3.15}$$

Consequently, we define

$$\mathbb{C} := \Big\{ \rho \in [0,1]^{|F|} \colon \sum_{\alpha \in F\colon \alpha \cap U \neq \emptyset} (|\alpha \cap U| - 1)\rho_\alpha \leq |U|, \quad \forall U \subseteq V \Big\}.$$

By Theorem 3.3, the set $\mathbb{C}$ is the region of concavity for the Bethe entropy (3.14) within $[0,1]^{|F|}$.

We also define the set

$$\mathbb{F} := \{1_{F'} \colon F' \subseteq F \text{ and } F' \cup N(F') \text{ is a single-cycle forest in } G\} \subseteq \{0,1\}^{|F|},$$

where a *single-cycle forest* is defined to be a subset of edges of a graph such that each connected component contains at most one cycle. (We disregard the directions of edges in G.) The following theorem gives our main result. The proof is presented in Section 3.8.

Theorem 3.4. *In the Bethe case (i.e., the region graph G has two layers), we have the containment* $\operatorname{conv}(\mathbb{F}) \subseteq \mathbb{C}$. *If in addition $|\alpha| = 2$ for all $\alpha \in F$, then* $\operatorname{conv}(\mathbb{F}) = \mathbb{C}$.

The significance of Theorem 3.4 is that it provides us with a convenient graph-based method for constructing vectors $\rho \in \mathbb{C}$. From the inequalities (3.15), it is not even clear how to efficiently verify whether a given $\rho \in [0,1]^{|F|}$ lies in $\mathbb{C}$, since it involves testing $2^{|V|}$ inequalities.

Comparing Theorem 3.4 with known results, note that in the pairwise case ($|\alpha| = 2$ for all $\alpha \in F$), Theorem 1 of Wainwright et al. [69] states that the Bethe entropy is concave over $\operatorname{conv}(\mathbb{T})$, where $\mathbb{T} \subseteq \{0,1\}^{|E|}$ is the set of edge indicator vectors for spanning forests of the graph. It is trivial to check that $\mathbb{T} \subseteq \mathbb{F}$, since every spanning forest is also a single-cycle forest. Hence, Theorems 3.3 and 3.4 together imply a stronger result than in Wainwright et al. [69], characterizing the precise region of concavity for the Bethe entropy as a superset of the polytope $\operatorname{conv}(\mathbb{T})$ analyzed there. In the unweighted Kikuchi case, it is also known [3, 52] that the Kikuchi entropy is concave for the assignment $\rho = 1_F$ when the region graph G is connected and has at most one cycle. Clearly, $1_F \in \mathbb{C}$ in that case, so this result is a consequence of Theorems 3.3 and 3.4, as well. However, our theorems show that a much more general statement is true.

It is tempting to posit that $\operatorname{conv}(\mathbb{F}) = \mathbb{C}$ holds more generally in the Bethe case. However, as the following example shows, settings arise where $\operatorname{conv}(\mathbb{F}) \subsetneq \mathbb{C}$. Details are contained in Section 3.8.

Example 3.5. Consider a two-layer region graph with vertices $V = \{1,2,3,4,5\}$ and factors $\alpha_1 = \{1,2,3\}$, $\alpha_2 = \{2,3,4\}$, and $\alpha_3 = \{3,4,5\}$. Then $(1, \frac{1}{2}, 1) \in \mathbb{C} \setminus \operatorname{conv}(\mathbb{F})$.

In fact, Example 3.5 is a special case of a more general statement, which we state in the following proposition. Here, $\mathfrak{F} := \{F' \subseteq F \colon 1_{F'} \in \mathbb{F}\}$, and an element $F^* \in \mathfrak{F}$ is *maximal* if it is not contained in another element of $\mathfrak{F}$.

Proposition 3.6. *Suppose (i) G is not a single-cycle forest, and (ii) there exists a maximal element $F^* \in \mathfrak{F}$ such that the induced subgraph $F^* \cup N(F^*)$ is a forest. Then $conv(\mathbb{F}) \subsetneq \mathbb{C}$.*

The proof of Proposition 3.6 is contained in Section 3.8. Note that if $|\alpha| = 2$ for all $\alpha \in F$, then condition (ii) is violated whenever condition (i) holds, so Proposition 3.6 provides a partial converse to Theorem 3.4.

3.3 Reweighted sum product algorithm

In this section, we provide an iterative message passing algorithm to optimize the Kikuchi variational problem (3.8). As in the case of the generalized belief propagation algorithm for the unweighted Kikuchi approximation [77, 76, 40, 52, 41, 74] and the reweighted sum product algorithm for the Bethe approximation [69], our message passing algorithm searches for stationary points of the Lagrangian version of the problem (3.8). When ρ satisfies condition (3.10), Theorem 3.1 implies that the problem (3.8) is strictly concave, so the unique fixed point of the message passing algorithm globally maximizes the Kikuchi approximation.

Let $G = (V, R)$ be a region graph defining our Kikuchi approximation. Following Pakzad and Anantharam [52], for $r, s \in R$, we write $r \prec s$ if $r \subsetneq s$ and there does not exist $t \in R$ such that $r \subsetneq t \subsetneq s$. For $r \in R$, we define the parent set of r to be $\mathcal{P}(r) = \{s \in R \colon r \prec s\}$ and the child set of r to be $\mathcal{C}(r) = \{s \in R \colon s \prec r\}$. With this notation, $\tau = \{\tau_r(x_r) \colon r \in R\}$ belongs to the set Δ_R^K if and only if $\sum_{x_{s\setminus r}} \tau_s(x_r, x_{s\setminus r}) = \tau_r(x_r)$ for all $r \in R$, $s \in \mathcal{P}(r)$.

The message passing algorithm we propose is as follows: For each $r \in R$ and $s \in \mathcal{P}(r)$, let $M_{sr}(x_r)$ denote the message passed from s to r at assignment x_r. Starting with an arbitrary positive initialization of the messages, we repeatedly perform the following updates for all $r \in R$, $s \in \mathcal{P}(r)$:

$$M_{sr}(x_r) \leftarrow C \left[\frac{\sum_{x_{s\setminus r}} \exp\big(\theta_s(x_s)/\rho_s\big) \prod_{v \in \mathcal{P}(s)} M_{vs}(x_s)^{\rho_v/\rho_s} \prod_{w \in \mathcal{C}(s)\setminus r} M_{sw}(x_w)^{-1}}{\exp\big(\theta_r(x_r)/\rho_r\big) \prod_{u \in \mathcal{P}(r)\setminus s} M_{ur}(x_r)^{\rho_u/\rho_r} \prod_{t \in \mathcal{C}(r)} M_{rt}(x_t)^{-1}} \right]^{\frac{\rho_r}{\rho_r+\rho_s}}. \quad (3.16)$$

Here, $C > 0$ may be chosen to ensure a convenient normalization condition; e.g., $\sum_{x_r} M_{sr}(x_r) = 1$. Upon convergence of the updates (3.16), we compute the pseudomarginals according to

$$\tau_r(x_r) \propto \exp\left(\frac{\theta_r(x_r)}{\rho_r}\right) \prod_{s \in \mathcal{P}(r)} M_{sr}(x_r)^{\rho_s/\rho_r} \prod_{t \in \mathcal{C}(r)} M_{rt}(x_t)^{-1}, \quad (3.17)$$

and we obtain the corresponding Kikuchi approximation by computing the objective function (3.8) with these pseudomarginals. We have the following result, which is proved in Section 3.9:

Theorem 3.7. *The pseudomarginals τ specified by the fixed points of the messages $\{M_{sr}(x_r)\}$ via the updates* (3.16) *and* (3.17) *correspond to the stationary points of the Lagrangian associated with the Kikuchi approximation problem* (3.8).

As with the standard belief propagation and reweighted sum product algorithms, we have several options for implementing the above message passing algorithm in practice. For example, we may perform the updates (3.16) using serial or parallel schedules. To improve the convergence of the algorithm, we may damp the updates by taking a convex combination of new and previous messages using an appropriately chosen step size. As noted by Pakzad and Anantharam [52], we may also use a minimal graphical representation of the Hasse diagram to lower the complexity of the algorithm.

Finally, we remark that although our message passing algorithm proceeds in the same spirit as classical belief propagation algorithms by operating on the Lagrangian of the objective function, our algorithm as presented above does not immediately reduce to the generalized belief propagation algorithm for unweighted Kikuchi approximations or the reweighted sum product algorithm for tree-reweighted pairwise Bethe approximations. Previous authors use algebraic relations between the overcounting numbers (3.9) in the Kikuchi case [77, 76, 40, 52] and the two-layer structure of the Hasse diagram in the Bethe case [69] to obtain a simplified form of the updates. Since the coefficients ρ in our problem lack the same algebraic relations, following the message-passing protocol used in previous work [40, 77] leads to more complicated updates, so we present a slightly different algorithm that still optimizes the general reweighted Kikuchi objective.

3.4 Experiments

In this section, we present empirical results to demonstrate the advantages of the reweighted Kikuchi approximation that support our theoretical results. For simplicity, we focus on the binary pairwise Ising model given in equation (3.2). Without loss of generality, we may take the potentials to be $\gamma_s(x_s) = \gamma_s x_s$ and $\gamma_{st}(x_s, x_t) = \gamma_{st} x_s x_t$ for some $\gamma = (\gamma_s, \gamma_{st}) \in \mathbb{R}^{|V|+|E|}$. We run our experiments on two types of graphs: (1) K_n, the complete graph on n vertices, and (2) T_n, the $\sqrt{n} \times \sqrt{n}$ toroidal grid graph where every vertex has degree four.

Bethe approximation. We consider the pairwise Bethe approximation of the log partition function $A(\gamma)$ with weights $\rho_{st} \geq 0$ and $\rho_s = 1 - \sum_{t \in N(s)} \rho_{st}$. Because of the regularity structure of K_n and T_n, we take $\rho_{st} = \rho \geq 0$ for all $(s,t) \in E$ and study the behavior of the Bethe approximation as ρ varies. For this particular choice of weight vector $\vec{\rho} = \rho 1_E$, we define

$$\rho_{\text{tree}} = \max\{\rho \geq 0 \colon \vec{\rho} \in \text{conv}(\mathbb{T})\},$$

and

$$\rho_{\text{cycle}} = \max\{\rho \geq 0 \colon \vec{\rho} \in \text{conv}(\mathbb{F})\}.$$

It is easily verified that for K_n, we have $\rho_{\text{tree}} = \frac{2}{n}$ and $\rho_{\text{cycle}} = \frac{2}{n-1}$; while for T_n, we have $\rho_{\text{tree}} = \frac{n-1}{2n}$ and $\rho_{\text{cycle}} = \frac{1}{2}$.

Our results in Section 3.2 imply that the Bethe objective function $B_{\gamma,\rho}(\tau)$ in equation (3.8) is concave if and only if $\rho \leq \rho_{\text{cycle}}$, and Wainwright et al. [69] show that we have the bound $A(\gamma) \leq B(\gamma; \rho)$ for $\rho \leq \rho_{\text{tree}}$. Moreover, since the Bethe entropy may be written in terms of the edge mutual information (3.13), the function $B(\gamma; \rho)$ is decreasing in ρ. In our results below, we observe that we may obtain a tighter approximation to $A(\gamma)$ by moving from the upper bound region $\rho \leq \rho_{\text{tree}}$ to the concavity region $\rho \leq \rho_{\text{cycle}}$. In addition, for $\rho > \rho_{\text{cycle}}$, we observe multiple local optima of $B_{\gamma,\rho}(\tau)$.

Procedure. We generate a random potential $\gamma = (\gamma_s, \gamma_{st}) \in \mathbb{R}^{|V|+|E|}$ for the Ising model (3.2) by sampling each potential $\{\gamma_s\}_{s\in V}$ and $\{\gamma_{st}\}_{(s,t)\in E}$ independently. We consider two types of models:

$$\textit{Attractive}: \quad \gamma_{st} \sim \text{Uniform}[0, \omega_{st}],$$

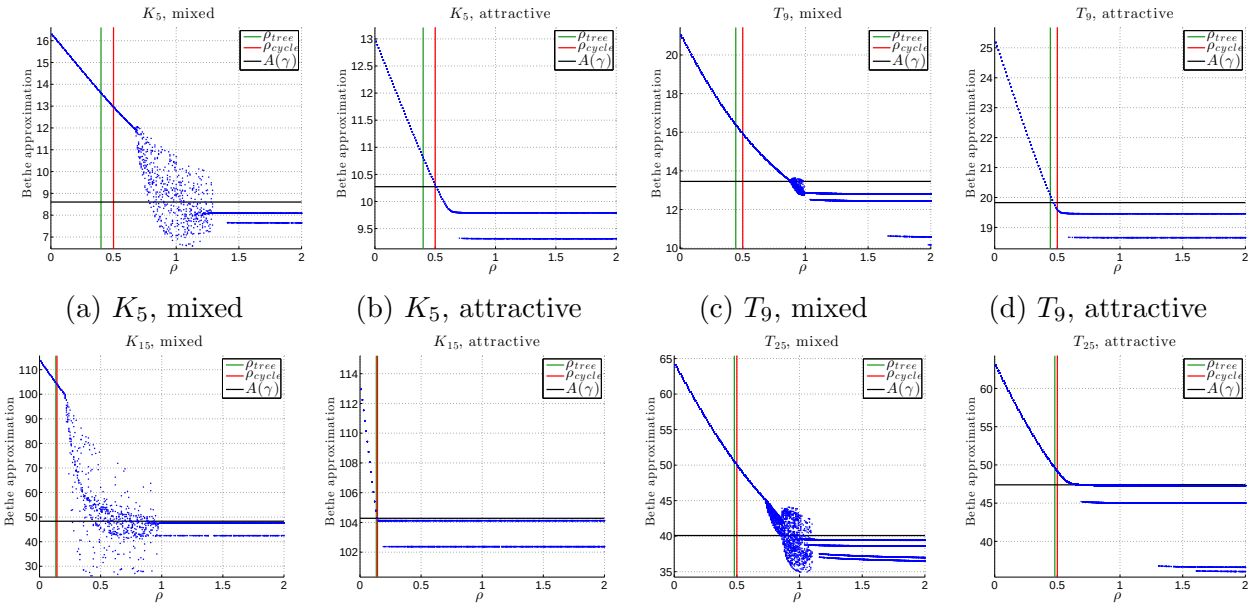
and

$$\textit{Mixed}: \quad \gamma_{st} \sim \text{Uniform}[-\omega_{st}, \omega_{st}].$$

In each case, $\gamma_s \sim \text{Uniform}[0, \omega_s]$. We set $\omega_s = 0.1$ and $\omega_{st} = 2$. Intuitively, the attractive model encourages variables in adjacent nodes to assume the same value, and it has been shown [57, 62] that the ordinary Bethe approximation ($\rho_{st} = 1$) in an attractive model lower-bounds the log partition function. For $\rho \in [0, 2]$, we compute stationary points of $B_{\gamma,\rho}(\tau)$ by running the reweighted sum product algorithm of Wainwright et al. [69]. We use a damping factor of $\lambda = 0.5$, convergence threshold of 10^{-10} for the average change of messages, and at most 2500 iterations. We repeat this process with at least 8 random initializations for each value of ρ. Figure 3.1 shows the scatter plots of ρ and the Bethe approximation $B_{\gamma,\rho}(\tau)$. In each plot, the two vertical lines are the boundaries $\rho = \rho_{\text{tree}}$ and $\rho = \rho_{\text{cycle}}$, and the horizontal line is the value of the true log partition function $A(\gamma)$.

Results. Figures 3.1a–3.1d show the results of our experiments on small graphs (K_5 and T_9) for both attractive and mixed models. We see that the Bethe approximation with $\rho \leq \rho_{\text{cycle}}$ generally provides a better approximation to $A(\gamma)$ than the Bethe approximation computed over $\rho \leq \rho_{\text{tree}}$. However, in general we cannot guarantee whether $B(\gamma; \rho)$ will give an upper or lower bound for $A(\gamma)$ when $\rho \leq \rho_{\text{cycle}}$. As noted above, we have $B(\gamma; 1) \leq A(\gamma)$ for attractive models.

We also observe from Figures 3.1a–3.1d that shortly after ρ leaves the concavity region $\{\rho \leq \rho_{\text{cycle}}\}$, multiple local optima emerge for the Bethe objective function. The presence of the point clouds near $\rho = 1$ in Figures 3.1a and 3.1c arises because the sum product algorithm has not converged after 2500 iterations. Indeed, the same phenomenon is true for all our results: in the region where multiple local optima begin to appear, it is more difficult for the algorithm to converge. See Figure 3.2 and the accompanying text in Section 3.10 for a plot of the points $(\rho, \log_{10}(\Delta))$, where Δ is the final average change in the messages at



(a) K_5, mixed (b) K_5, attractive (c) T_9, mixed (d) T_9, attractive

(e) K_{15}, mixed (f) K_{15}, attractive (g) T_{25}, mixed (h) T_{25}, attractive

Figure 3.1: Values of the reweighted Bethe approximation as a function of ρ.

termination of the algorithm. From Figure 3.2, we see that the values of Δ are significantly higher for the values of ρ near where multiple local optima emerge. We suspect that for these values of ρ, the sum product algorithm fails to converge since distinct local optima are close together, so messages oscillate between the optima. For larger values of ρ, the local optima become sufficiently separated and the algorithm converges to one of them. However, it is interesting to note that this point cloud phenomenon does not appear for attractive models, despite the presence of distinct local optima.

Simulations for larger graphs are shown in Figures 3.1e–3.1h. If we zoom into the region near $\rho \leq \rho_{\text{cycle}}$, we still observe the same behavior that $\rho \leq \rho_{\text{cycle}}$ generally provides a better Bethe approximation than $\rho \leq \rho_{\text{tree}}$. Moreover, the presence of the point clouds and multiple local optima are more pronounced, and we see from Figures 3.1c, 3.1g, and 3.1h that new local optima with even worse Bethe values arise for larger values of ρ. Finally, we note that the same qualitative behavior also occurs in all the other graphs that we have tried (K_n for $n \in \{5, 10, 15, 20, 25\}$ and T_n for $n \in \{9, 16, 25, 36, 49, 64\}$), with multiple random instances of the Ising model p_γ.

3.5 Discussion

In this chapter, we have analyzed the reweighted Kikuchi approximation method for estimating the log partition function of a distribution that factorizes over a region graph. We have characterized necessary and sufficient conditions for the concavity of the variational

objective function, generalizing existing results in literature. Our simulations demonstrate the advantages of using the reweighted Kikuchi approximation and show that multiple local optima may appear outside the region of concavity.

An interesting future research direction is to obtain a better understanding of the approximation guarantees of the reweighted Bethe and Kikuchi methods. In the Bethe case with attractive potentials θ, several recent results [69, 62, 57] establish that the Bethe approximation $B(\theta;\rho)$ is an upper bound to the log partition function $A(\theta)$ when ρ lies in the spanning tree polytope, whereas $B(\theta;\rho) \leq A(\theta)$ when $\rho = 1_F$. By continuity, we must have $B(\theta;\rho^*) = A(\theta)$ for some values of ρ^*, and it would be interesting to characterize such values where the reweighted Bethe approximation is exact.

Another interesting direction is to extend our theoretical results on properties of the reweighted Kikuchi approximation, which currently depend solely on the structure of the region graph and the weights ρ, to incorporate the effect of the model potentials θ. For example, several authors [63, 23] present conditions under which loopy belief propagation applied to the unweighted Bethe approximation has a unique fixed point. The conditions for uniqueness of fixed points slightly generalize the conditions for convexity, and they involve both the graph structure and the strength of the potentials. We suspect that similar results would hold for the reweighted Kikuchi approximation.

3.6 Proofs for the sufficient condition

Proof of Theorem 3.1

We use the proof technique of Theorem 1 in Pakzad and Anantharam [52] for the unweighted Bethe entropy, together with Lemma 3.8 in Section 3.6, which provides a generalization of Hall's marriage lemma for weighted bipartite graphs.

We construct a bipartite graph according to

$$V_1 := \{r \in R\colon \rho_r < 0\}, \quad \text{and} \quad V_2 := \{r \in R\colon \rho_r > 0\},$$

where $(s,t) \in E$ for $s \in V_1$ and $t \in V_2$ when $s \subset t$. Let weights w be defined such that $w(s) = -\rho_s$ for $s \in V_1$ and $w(s) = \rho_s$ for $s \in V_2$. We claim that condition (3.19) of Lemma 3.8 is satisfied. Indeed, for $U \subseteq V_1$, we have

$$w(U) = -\sum_{s\in U} \rho_s \leq \sum_{s\in A(U)} \rho_s = \sum_{s\in A(U):\rho_s>0} \rho_s + \sum_{s\in A(U):\rho_s<0} \rho_s \leq \sum_{s\in A(U):\rho_s>0} \rho_s = w(N(U)),$$

where the first inequality is a direct application of the assumption (3.10). Hence, by Lemma 3.8, we have a saturating edge labeling γ.

For each $t \in V_2$, define

$$\rho'_t := \rho_t - \sum_{s\in N(t)} \gamma_{st} \geq 0.$$

We may then write

$$\begin{aligned} H(\tau;\rho) &= \sum_{s\in V_1} \rho_s H_s(\tau_s) + \sum_{t\in V_2} \rho_t H_t(\tau_t) \\ &= \sum_{(s,t)\in E} \gamma_{st} \left\{ -H_s(\tau_s) + H_t(\tau_t) \right\} + \sum_{t\in V_2} \rho'_t H_t(\tau_t) \\ &= \sum_{(s,t)\in E} \gamma_{st} \left\{ \sum_{x_s} \tau_s(x_s) \log \tau_s(x_s) - \sum_{x_t} \tau_t(x_t) \log \tau_t(x_t) \right\} + \sum_{t\in V_2} \rho'_t H_t(\tau_t) \\ &= \sum_{(s,t)\in E} \gamma_{st} \sum_{x_t} \tau_t(x_t) \log \left(\frac{\tau_s(x_s)}{\tau_t(x_t)} \right) + \sum_{t\in V_2} \rho'_t H_t(\tau_t), \end{aligned} \tag{3.18}$$

where we have used the fact that $\sum_{x_{t\setminus s}} \tau_t(x_s, x_{t\setminus s}) = \tau_s(x_s)$, since $\tau \in \Delta_R^K$, to obtain the last equality.

Note that for each pair (s,t), we have

$$\sum_{x_t} \tau_t(x_t) \log \left(\frac{\tau_s(x_s)}{\tau_t(x_t)} \right) = -D_{\mathrm{KL}}(\tau_t \| \tau_s),$$

which is strictly concave in the pair (τ_t, τ_s). Furthermore, each term $H_t(\tau_t)$ is concave in τ_t. It follows by the expansion (3.18) that $H(\tau;\rho)$ is strictly concave, as wanted.

Generalization of Hall's marriage lemma

In this section, we prove a generalization of Hall's marriage lemma, which is useful in proving concavity of the Bethe entropy function $H(\tau;\rho)$.

Let $G = (V_1 \cup V_2, E)$ be a bipartite graph, where each vertex $v \in V := V_1 \cup V_2$ is assigned a weight $w(v) > 0$. For a set $U \subseteq V$, define

$$w(U) := \sum_{s\in U} w(s).$$

Also define the neighborhood set

$$N(U) := \bigcup_{s\in U} N(s),$$

where $N(s) := \{t \colon (s,t) \in E\}$ is the usual neighborhood set of a single node.

We say that an edge labeling $\gamma = (\gamma_{st} \colon (s,t) \in E) \in \mathbb{R}_{\geq 0}^{|E|}$ *saturates* V_1 if the following conditions hold:

1. For all $s \in V_1$, we have $\sum_{t\in N(s)} \gamma_{st} = w(s)$.

2. For all $t \in V_2$, we have $\sum_{s \in N(t)} \gamma_{st} \leq w(t)$.

Lemma 3.8. *Suppose*

$$w(U) \leq w(N(U)), \quad \forall U \subseteq V_1. \tag{3.19}$$

Then there exists an edge labeling γ that saturates V_1.

Proof. We prove the lemma in stages. First, assume $w(v) \in \mathbb{Q}$ for all $v \in V$ and condition (3.19) holds. With an appropriate rescaling, we may assume that all weights are integers. Call the new weights w'. We then construct a graph G' such that each node $v \in V$ is expanded into a set U_v of $w'(v)$ nodes, and edges of G' are constructed by connecting all nodes in U_s to all nodes in U_t, for each $(s,t) \in E$. By the usual version of Hall's marriage lemma [21], there exists a matching of G' that saturates $V_1' := \bigcup_{v \in V_1} U_v$. Indeed, it follows immediately from condition (3.19) that

$$w'(U) \leq w'(N(U)), \qquad \forall U \subseteq V_1.$$

Suppose $T' \subseteq V_1'$, and let $T := \{s \in V_1 \colon U_s \cap T' \neq \emptyset\}$. Then

$$|T'| \leq \left| \bigcup_{s \in T} U_s \right| = w'(T) \leq w'(N(T)) = |N(T')|,$$

so the sufficient condition of Hall's marriage lemma is met, implying the existence of a matching. The edge labeling γ is obtained by setting

$$\gamma_{st} = \{\#\text{ of edges between } U_s \text{ and } U_t \text{ in matching}\}$$

and rescaling.

Next, suppose $w(v) \in \mathbb{R}$ for all $v \in V$ and condition (3.19) holds with *strict* inequality; i.e.,

$$w(U) < w(N(U)), \qquad \forall U \subseteq V_1. \tag{3.20}$$

We claim that there exists an edge labeling γ that saturates V_1. Indeed, let

$$\epsilon := \min_{U \subseteq V_1} \{w(N(U)) - w(U)\} > 0.$$

Define a new weighting w' with only rational values, such that

$$w'(s) \in \left[w(s), \quad w(s) + \frac{\epsilon}{2 \cdot \deg(G)} \right), \qquad \forall s \in V_1,$$
$$w'(t) \in \left(w(t) - \frac{\epsilon}{2 \cdot \deg(G)}, \quad w(t) \right], \qquad \forall t \in V_2,$$

where $\deg(G) = |E|$ is the number of edges in G. It is clear that Hall's condition (3.19) still holds for w'. Hence, by the result of the last paragraph, there exists an edge labeling γ' that

saturates V_1 with respect to w'. Observe that by decreasing the weights of γ' slightly, we easily obtain an edge labeling γ that saturates V_1 with respect to the original weighting w.

Finally, consider the most general case: condition (3.19) holds and $w(v) \in \mathbb{R}$ for all $v \in V$. Note that the problem of finding an edge labeling that saturates V_1 may be rephrased as follows. Let $b_1 \in \mathbb{R}^{|V_1|}$ be the vector of weights $(w(s)\colon s \in V_1)$. Then for an appropriate choice of the matrix $A_1 \in \{0,1\}^{|V_1|\times|E|}$, the conditions

$$\sum_{t\in N(s)} \gamma_{st} = w(s), \qquad \forall s \in V_1,$$

may be expressed as a system of linear equations,

$$A_1\gamma = b_1. \tag{3.21}$$

Similarly, letting $b_2 = (w(t)\colon t \in V_2) \in \mathbb{R}^{|V_2|}$, the conditions

$$\sum_{s\in N(t)} \gamma_{st} \le w(t), \qquad \forall t \in V_2,$$

may be expressed in the form

$$A_2\gamma \le b_2, \tag{3.22}$$

where $A_2 \in \{0,1\}^{|V_2|\times|E|}$. A saturating edge labeling exists if and only if there exists $\gamma \in \mathbb{R}^{|E|}_{\ge 0}$ that simultaneously satisfies conditions (3.21) and (3.22). Now consider a sequence of weight vectors $\{b_1^n\}_{n\ge 1}$, such that $b_1^n \to b_1$ and the convergence is from below and strictly monotone for each component. Let $w^n = (b_1^n, b_2)$ denote the full sequence of weights. Then

$$w^n(U) < w(U) \le w(N(U)) = w^n(N(U)), \qquad \forall U \subseteq V.$$

It follows by the result of the previous paragraph that there exists an edge labeling $\gamma^n \in \mathbb{R}^{|E|}_{\ge 0}$ such that

$$A_1\gamma^n = b_1^n, \qquad \text{and} \qquad \gamma^n \in D := \Big\{\gamma \in \mathbb{R}^{|E|}_{\ge 0}\colon A_2\gamma \le b_2\Big\}.$$

Clearly, D is a closed set; furthermore, it is easy to see that the constraint $A_2\gamma \le b_2$ implies that each component of γ is bounded from above, since A_2 contains only nonnegative entries. It follows that the sequence $\{\gamma_n\}_{n\ge 1}$ has a limit point $\gamma^* \in D$. By continuity of the linear map A_1, we must have

$$A_1\gamma^* = \lim_{n\to\infty} A_1\gamma_n = \lim_{n\to\infty} b_1^n = b_1.$$

Hence, γ^* is a valid edge labeling that saturates V_1.

□

Proof of Corollary 3.2

By Theorem 3.1, $H(\tau;\rho)$ is strictly concave provided condition (3.10) holds. Note that

$$\mathcal{F}(\alpha) = \{\alpha\}, \qquad \forall \alpha \in F,$$

whereas

$$\mathcal{F}(s) = \{s\} \cup N(s), \qquad \forall s \in V.$$

Condition (3.10) applied to the set $S = \{\alpha\}$ gives the inequality

$$\rho_\alpha \geq 0, \qquad \forall \alpha \in F. \tag{3.23}$$

For a subset $U \subseteq V$, we can write

$$\mathcal{F}(U) = \bigcup_{s \in U} \mathcal{F}(s) = U \cup N(U) = U \cup \{\alpha \in F \colon \alpha \cap U \neq \emptyset\},$$

so (3.10) translates into

$$\sum_{s \in U} \rho_s + \sum_{\alpha \in F \colon \alpha \cap U \neq \emptyset} \rho_\alpha \geq 0, \qquad \forall U \subseteq V, \tag{3.24}$$

which is condition (3.12). It is easy to see that conditions (3.23) and (3.24) together also imply the validity of condition (3.10) for any other set of regions $S \subseteq R$.

3.7 Proofs for the necessary condition

Proof of Theorem 3.3

Our result relies on the property that if the Bethe entropy $H(\tau;\rho)$ is concave over Δ_R^K, then $H(\tau;\rho)$ is also concave over any subset $\Delta' \subseteq \Delta_R^K$. In particular, it is sufficient to assume that $\mathcal{X}$ is binary, say $\mathcal{X} = \{-1, +1\}$; the general multinomial case $|\mathcal{X}| > 2$ follows by restricting the distribution of X_s to be supported on only two points.

The first lemma shows that $\rho_\alpha \geq 0$ for all $\alpha \in F$. The proof is given in the next section.

Lemma 3.9. *If the Bethe entropy $H(\tau;\rho)$ is concave over Δ_R^K, then $\rho_\alpha \geq 0$ for all $\alpha \in F$.*

To establish the necessity of condition (3.12), consider a nonempty subset $U \subseteq V$ and the corresponding sub-region graph $R_U = U \cup F_U$, where $F_U = \{\alpha \cap U \colon \alpha \in F, \alpha \cap U \neq \emptyset\}$. From the original weights $\rho \in \mathbb{R}^{|V|+|F|}$, construct the sub-region weights $\rho^U \in \mathbb{R}^{|U|+|F_U|}$ given by

$$\rho_s^U = \rho_s, \quad \forall s \in U, \qquad \text{and} \qquad \rho_{\alpha \cap U}^U = \rho_\alpha, \quad \forall \alpha \cap U \in F_U.$$

For simplicity, we consider R_U to be a multiset by remembering which factor $\alpha \in F$ each $\beta = \alpha \cap U \in F_U$ comes from; we can equivalently work with R_U as a set by defining the

weights ρ^U to be the sum over the pre-images of the factors in R_U. Consider the set of locally consistent R_U-pseudomarginals $\Delta^K_{R_U}$. Define a map that sends $\tilde{\tau} \in \Delta^K_{R_U}$ to $\tau \in \Delta^K_R$ defined by

$$\tau_s(x_s) = \begin{cases} \tilde{\tau}_s(x_s) & \text{if } s \in U, \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

$$\tau_\alpha(x_\alpha) = \begin{cases} \tilde{\tau}_{\alpha\cap U}(x_{\alpha\cap U}) \cdot \prod_{s\in\alpha\setminus U} \tau_s(x_s) & \text{if } \alpha \cap U \neq \emptyset \ \ (\text{so } \alpha \cap U \in F_U), \\ \prod_{s\in\alpha} \tau_s(x_s) & \text{otherwise.} \end{cases}$$

Let Δ_U denote the image of $\Delta^K_{R_U}$ under the mapping above, and note that $\Delta_U \subseteq \Delta^K_R$. Therefore, $H(\tau;\rho)$ is concave over Δ_U. Now let $\tau \in \Delta_U$ and let $\tilde{\tau} \in \Delta^K_{R_U}$ be a pre-image of τ. With this construction, we have the following lemma:

Lemma 3.10. *The entropy $H(\tau;\rho)$ differs from $H_U(\tilde{\tau};\rho^U)$ by a constant, where $H_U(\tilde{\tau};\rho^U)$ is the Bethe entropy defined over the sub-region graph R_U.*

Finally, we have a lemma showing that we can extract condition (3.12) for $U = V$. The proof is provided below.

Lemma 3.11. *If the Bethe entropy $H(\tau;\rho)$ is concave over Δ^K_R, then*

$$\sum_{s\in V} \rho_s + \sum_{\alpha\in F} \rho_\alpha \geq 0.$$

By Lemma 3.10, the concavity of $H(\tau;\rho)$ over Δ_U implies the concavity of $H_U(\tilde{\tau};\rho^U)$ over $\Delta^K_{R_U}$. Then by Lemma 3.11 applied to R_U, we have

$$\sum_{s\in U} \rho_s + \sum_{\alpha\in F:\ \alpha\cap U\neq\emptyset} \rho_\alpha = \sum_{s\in U} \rho^U_s + \sum_{\beta\in F_U} \rho^U_\beta \geq 0,$$

finishing the proof.

Proof of Lemma 3.9

Fix $\alpha \in F$, and let Δ_α be the set of pseudomarginals $\tau \in \Delta^K_R$ with the property that for all $s \in V$ and $\beta \in F \setminus \{\alpha\}$, τ_s and τ_β are uniform distributions over X_s and X_β, respectively, while τ_α is an arbitrary distribution on X_α with uniform single-node marginals. Then $H(\tau;\rho)$ is concave over Δ_α. On the other hand, note that for $\tau \in \Delta_\alpha$, $H_s(\tau_s) = \log 2$ and $H_\beta(\tau_\beta) = |\beta| \log 2$ are constants for $s \in V$ and $\beta \in F \setminus \{\alpha\}$, so we can write

$$H(\tau;\rho) = \rho_\alpha H_\alpha(\tau_\alpha) + \text{constant}.$$

Since $H_\alpha(\tau_\alpha)$ is concave in τ_α, this implies $\rho_\alpha \geq 0$, as claimed.

Proof of Lemma 3.10

By construction, for $s \in V \setminus U$, we have $H_s(\tau_s) = \log 2$; and for $\alpha \in F$ with $\alpha \cap U = \emptyset$, we have $H_\alpha(\tau_\alpha) = |\alpha| \log 2$. Moreover, for $\alpha \in F$ with $\alpha \cap U \neq \emptyset$, we have

$$H_\alpha(\tau_\alpha) = H_{\alpha\cap U}(\tilde{\tau}_{\alpha\cap U}) + \sum_{s\in\alpha\setminus U} H_s(\tau_s) = H_{\alpha\cap U}(\tilde{\tau}_{\alpha\cap U}) + |\alpha \setminus U| \log 2.$$

Therefore, for $\tau \in \Delta_U$, we can write

$$\begin{aligned}
H(\tau;\rho) &= \sum_{s\in V} \rho_s H_s(\tau_s) + \sum_{\alpha\in F} \rho_\alpha H_\alpha(\tau_\alpha) \\
&= \sum_{s\in U} \rho_s H_s(\tilde{\tau}_s) + \Big(\sum_{s\in V\setminus U} \rho_s\Big) \log 2 \\
&\qquad + \sum_{\alpha\in F\colon\, \alpha\cap U\neq\emptyset} \rho_\alpha \Big(H_{\alpha\cap U}(\tilde{\tau}_{\alpha\cap U}) + |\alpha\setminus U| \log 2\Big) + \sum_{\alpha\in F\colon\, \alpha\cap U=\emptyset} \rho_\alpha |\alpha| \log 2 \\
&= \sum_{s\in U} \rho_s^U H_s(\tilde{\tau}_s) + \sum_{\beta\in F_U} \rho_\beta^U H_\beta(\tilde{\tau}_\beta) + \text{constant} \\
&= H_U(\tilde{\tau};\rho^U) + \text{constant},
\end{aligned}$$

as wanted.

Proof of Lemma 3.11

Given $m_o, m_e \in \mathbb{R}$, we define a pseudomarginal $\tau = (\tau_s, \tau_\alpha)$ by

$$\tau_s(x_s) = \frac{1 + x_s m_o}{2}, \qquad \forall s \in V,\ x_s \in X = \{-1, +1\},$$

and for $\alpha \in F$ with $|\alpha| = k$,

$$\tau_\alpha(x_\alpha) = \begin{cases} 2^{-k}\left(1 + 2^{k-1} m_o + (2^{k-1}-1) m_e\right) & \text{if } x_\alpha = (1,\dots,1), \\ 2^{-k}\left(1 - 2^{k-1} m_o + (2^{k-1}-1) m_e\right) & \text{if } x_\alpha = (-1,\dots,-1), \\ 2^{-k}(1 - m_e) & \text{otherwise.} \end{cases}$$

Notice that the definition of τ above is equivalent to imposing the conditions

$$\mathbb{E}_{\tau_\alpha}\Big[\prod\nolimits_{s\in\beta} X_s\Big] = m_o \ \text{ if } |\beta| \text{ is odd}$$

and

$$\mathbb{E}_{\tau_\alpha}\Big[\prod\nolimits_{s\in\beta} X_s\Big] = m_e \ \text{ if } |\beta| \text{ is even},$$

for all $\alpha \in V \cup F$ and $\emptyset \neq \beta \subseteq \alpha$.

It is easy to see that $\sum_{x_s} \tau_s(x_s) = \sum_{x_\alpha} \tau_\alpha(x_\alpha) = 1$, and that τ_s is the single-node marginal of τ_α. Thus, for τ to lie in Δ_R^K, we only need to ensure that $\tau_s(x_s) \geq 0$ and $\tau_\alpha(x_\alpha) \geq 0$, or equivalently,

$$-1 \leq m_o \leq 1, \qquad \frac{1 + 2^{k-1}|m_o|}{2^{k-1} - 1} \leq m_e \leq 1, \qquad \forall\, 2 \leq k \leq K,$$

where $K = \max\{|\alpha| \colon \alpha \in F\}$. Let M denote the set of (m_o, m_e) satisfying the constraints above, and let Δ_M denote the set of pseudomarginals $\tau[m_o, m_e] \in \Delta_R^K$ given by the construction above for each $(m_o, m_e) \in M$.

Observe that the function $(m_o, m_e) \mapsto \tau[m_o, m_e]$ is additive for convex combinations; i.e., for $(m_o^{(1)}, m_e^{(1)}), \ldots, (m_o^{(j)}, m_e^{(j)}) \in M$ and $\lambda_1, \ldots, \lambda_j \geq 0$ with $\lambda_1 + \cdots + \lambda_j = 1$, we have

$$\sum_{i=1}^{j} \lambda_i \tau[m_o^{(i)}, m_e^{(i)}] = \tau\Big[\sum_{i=1}^{j} \lambda_i m_o^{(i)},\ \sum_{i=1}^{j} \lambda_i m_e^{(i)}\Big].$$

Since M is convex, this shows that Δ_M is a convex subset of Δ_R^K. Therefore, $H(\tau; \rho)$ is concave over Δ_M, and the additivity property above implies that the function

$$\zeta(m_o, m_e) := H(\tau[m_o, m_e]; \rho)$$

is concave over M. We now compute the Hessian of ζ and show how it relates to the required quantity that we want to prove is nonnegative.

Fix $(m_o, m_e) \in M$, and note that $\tau \equiv \tau[m_o, m_e]$ has the property that $\tau_\alpha = \tau_\beta$ whenever $|\alpha| = |\beta|$. Therefore, we can collect the terms in $H(\tau; \rho)$ based on the cardinality of $\alpha \in V \cup F$. The single-node entropy is, as a function of m_o,

$$\zeta_1(m_o) := H_s(\tau_s) = -\eta\left(\frac{1 + m_o}{2}\right) - \eta\left(\frac{1 - m_o}{2}\right),$$

where $\eta(t) := t \log t$. For $\alpha \in F$ with $|\alpha| = k \geq 2$, the entropy corresponding to τ_α is

$$\begin{aligned}\zeta_k(m_o, m_e) := H_\alpha(\tau_\alpha) = &-\eta\left(\frac{1 + 2^{k-1} m_o + (2^{k-1} - 1) m_e}{2^k}\right) - \eta\left(\frac{1 - 2^{k-1} m_o + (2^{k-1} - 1) m_e}{2^k}\right) \\ &- (2^k - 2)\, \eta\left(\frac{1 - m_e}{2^k}\right).\end{aligned}$$

The Bethe entropy can then be written as

$$\zeta(m_o, m_e) = H(\tau; \rho) = c_1 \zeta_1(m_o) + \sum_{k=2}^{K} c_k \zeta_k(m_o, m_e),$$

where $c_1 = \sum_{s \in V} \rho_s$ and $c_k = \sum_{\alpha \in F \colon |\alpha| = k} \rho_\alpha$ for $k \geq 2$.

Let us compute the Hessian matrix of $\zeta(m_o, m_e)$ along the axis $m_o = 0$. The function ζ_1 has second derivative $\zeta_1''(m_o) = -1/(1 - m_o^2)$, so at $m_o = 0$, the contribution of ζ_1 to the Hessian of ζ is

$$\nabla^2 \zeta_1(0, m_e) = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}.$$

For $k \geq 2$, the first partial derivatives of ζ_k are

$$\begin{aligned} \frac{\partial \zeta_k}{\partial m_o} &= -\frac{1}{2}\left\{\log\left(1 + 2^{k-1}m_o + (2^{k-1} - 1)m_e\right) - \log\left(1 - 2^{k-1}m_o + (2^{k-1} - 1)m_e\right)\right\}, \\ \frac{\partial \zeta_k}{\partial m_e} &= -\frac{(2^{k-1} - 1)}{2^k}\left\{\log\left(1 + 2^{k-1}m_o + (2^{k-1} - 1)m_e\right) + \log\left(1 - 2^{k-1}m_o + (2^{k-1} - 1)m_e\right)\right. \\ &\qquad \left. -2\log\left(1 - m_e\right)\right\}. \end{aligned}$$

The Hessian $\nabla^2 \zeta_k$ at $m_o = 0$ is then given by

$$\nabla^2 \zeta_k(0, m_e) = \begin{pmatrix} -\dfrac{2^{k-1}}{1 + (2^{k-1} - 1)m_e} & 0 \\ 0 & -\dfrac{2^{k-1} - 1}{(1 + (2^{k-1} - 1)m_e)(1 - m_e)} \end{pmatrix}.$$

Therefore, the Hessian of ζ at $m_o = 0$ is the diagonal matrix

$$\nabla^2 \zeta(0, m_e) = \begin{pmatrix} -c_1 - \displaystyle\sum_{k=2}^{K} \frac{2^{k-1} c_k}{1 + (2^{k-1} - 1)m_e} & 0 \\ 0 & -\displaystyle\sum_{k=2}^{K} \frac{(2^{k-1} - 1)c_k}{(1 + (2^{k-1} - 1)m_e)(1 - m_e)} \end{pmatrix}.$$

In particular, the eigenvalues of $\nabla^2 \zeta(0, m_e)$ are its diagonal entries. Taking $m_e \to 1$, we see that the eigenvalue corresponding to the first diagonal entry satisfies

$$\lim_{m_e \to 1} \lambda_1(m_e) = \lim_{m_e \to 1} \left\{ -c_1 - \sum_{k=2}^{K} \frac{2^{k-1} c_k}{1 + (2^{k-1} - 1)m_e} \right\} = -\sum_{k=1}^{K} c_k.$$

Since $(0, m_e) \in M$ as $m_e \to 1$ and $\zeta(m_o, m_e)$ is concave over M, we see that the eigenvalue above is nonpositive, which implies

$$\sum_{s \in V} \rho_s + \sum_{\alpha \in F} \rho_\alpha = \sum_{k=1}^{K} c_k \geq 0,$$

as desired.

3.8 Proofs for the polytope of concavity

Proof of Theorem 3.4

We first show that $\mathrm{conv}(\mathbb{F}) \subseteq \mathbb{C}$ in the general Bethe case. Since $\mathbb{C}$ is convex, it suffices to show that $\mathbb{F} \subseteq \mathbb{C}$, so consider $1_{F'} \in \mathbb{F}$. We need to show that inequality (3.15) holds for $\rho = 1_{F'}$.

Let $W_1, \ldots, W_m$ denote the connected components of $F' \cup N(F')$ in G. Consider an arbitrary $U \subseteq V$, and define $U_i := W_i \cap U$ for $1 \leq i \leq m$, and $U_0 := U \backslash \{U_1, \ldots, U_m\}$. Then each W_i has at most one cycle. Furthermore, we may write

$$\sum_{\substack{\alpha \in F: \\ \alpha \cap U \neq \emptyset}} (|\alpha \cap U| - 1)\rho_\alpha = \sum_{\substack{\alpha \in F': \\ \alpha \cap U \neq \emptyset}} (|\alpha \cap U| - 1) = \sum_{i=1}^{m} \Big\{ \sum_{\substack{\alpha \in W_i: \\ \alpha \cap U_i \neq \emptyset}} (|\alpha \cap U_i| - 1) \Big\}. \tag{3.25}$$

We claim that

$$\sum_{\alpha \in W_i \colon \alpha \cap U_i \neq \emptyset} (|\alpha \cap U_i| - 1) \leq |U_i|, \qquad \forall 1 \leq i \leq m. \tag{3.26}$$

Indeed, consider the induced subgraph W_i' of W_i with vertex set

$$V_i := U_i \cup \{\alpha \in W_i \colon \alpha \cap U_i \neq \emptyset\}.$$

Since W_i has at most one cycle, W_i' has at most one cycle, as well. Furthermore, the number of edges of W_i' is given by

$$|E(W_i')| = \sum_{\alpha \in W_i \colon \alpha \cap U_i \neq \emptyset} |\alpha \cap U_i|,$$

and the number of vertices is $|V_i| = |U_i| + |\{\alpha \in W_i \colon \alpha \cap U_i \neq \emptyset\}|$.

We have the following simple lemma:

Lemma 3.12. *A connected graph G has at most one cycle if and only if*

$$|E(U)| \leq |U|, \qquad \forall U \subseteq V.$$

Proof. First suppose G has at most one cycle. For any subset $U \subseteq V$, the induced subgraph H clearly also contains at most one cycle. Hence, we may remove at most one edge to obtain a graph H' which is a forest. Then

$$|E(H')| \leq |V(H')| - 1 = |U| - 1. \tag{3.27}$$

Furthermore, $|E(U)| \leq |E(H')| + 1$. It follows that $|E(U)| \leq |U|$.

Conversely, if G is a connected graph with more than one cycle, we may pick U to be the union of vertices in the two cycles, along with a path connecting the two cycles (in case the cycles are disconnected). It is easy to check that condition (3.27) is violated in this case. □

Applying Lemma 3.12 to the graph W_i' and rearranging then yields inequality (3.26). Combining with equation (3.25) then yields

$$\sum_{\alpha \in F\colon \alpha \cap U \neq \emptyset} (|\alpha \cap U| - 1)\rho_\alpha \leq \sum_{i=1}^{m} |U_i| = |U| - |U_0| \leq |U|,$$

proving the condition (3.15).

We now specialize to the case where $|\alpha| = 2$ for all $\alpha \in F$. Note that in this case, we may identify the region graph with an ordinary graph $\bar{G} = (V, E)$, where the edge set E is given by F. It is easy to check that $1_{F'} \in \mathbb{F}$ if and only if the subgraph of $\bar{G}$ with edge set F' is a single-cycle forest. In the following argument, we abuse notation and refer to $\bar{G}$ as G.

Recall that a *rational polyhedron* is a set of the form $\{x \in \mathbb{R}^p\colon Ax \leq b\}$, such that A and b have rational entries. Clearly, $\mathbb{C}$ is a rational polyhedron. Furthermore, a polyhedron is *integral* if all vertices are elements of the integer lattice $\mathbb{Z}^p$. The following result is standard in integer programming:

Lemma 3.13. *[Theorem 5.12, [32]] Let P be a rational polyhedron. Then P is integral if and only if $\max\{c^T x\colon x \in P\}$ is attained by an integral vector for each c for which the maximum is finite.*

We have already established that $1_{F'} \in \mathbb{C}$ for all $1_{F'} \in \mathbb{F}$. Furthermore, any lattice point in $\mathbb{C}$ is of the form 1_H, where $H \subseteq E$. By Lemma 3.12, each connected component of H must contain at most one cycle, implying that H is a single-cycle forest. Hence, $1_H \in \mathbb{F}$, as well. We then combine Lemma 3.13 with the following proposition to obtain the desired result.

Proposition 3.14. *Let $G = (V, E)$ be a graph. For any set of weights $c = (c_{st}) \in \mathbb{R}^{|E|}$, the LP*

$$\max \sum_{(s,t)\in E} c_{st} x_{st} \tag{3.28}$$

$$s.t. \sum_{(s,t)\in E(U)} x_{st} \leq |U|, \qquad \forall U \subseteq V, \tag{3.29}$$

$$0 \leq x_{st} \leq 1, \qquad \forall (s,t) \in E,$$

attains its maximum value at an integral vector x^*.

Proof. We first argue that it suffices to consider rational weights $c \in \mathbb{Q}^{|E|}$. Let X denote the feasible set of the LP, and let $F(c) = \max_{x\in X} c^\top x$ denote the maximum value of the LP. Note that $F(c)$ is continuous in c.

Suppose the claim in the proposition holds for $c \in \mathbb{Q}^{|E|}$. Given $c \in \mathbb{R}^{|E|}$, let $x^* \in \arg\max_{x\in X} c^T x$. Let $(c^{(n)})_{n\geq 1}$ be a sequence of weights in $\mathbb{Q}^{|E|}$ converging to c elementwise

as $n \to \infty$. Given $\epsilon > 0$, choose n sufficiently large such that $\|c^{(n)} - c\|_1 < \epsilon$ and $|F(c) - F(c^{(n)})| < \epsilon$. Applying our hypothesis, we know there exists an integral vector $z^* \in X$ such that $F(c^{(n)}) = (c^{(n)})^\top z^*$. Then

$$|F(c) - c^\top z^*| \le |F(c) - F(c^{(n)})| + |(c^{(n)} - c)^\top z^*| \le \epsilon + \|c^{(n)} - c\|_1 \, \|z^*\|_\infty \le 2\epsilon.$$

Thus, we can find an integral vector $z^* \in X$ that achieves the objective function that is within 2ϵ from the optimal value. Since $\epsilon > 0$ is arbitrary, we conclude by continuity that we may find an integral vector in X arbitrarily close to x^*. This implies that x^* is an integral vector.

It now remains to prove the claim in the proposition for $c \in \mathbb{Q}^{|E|}$. If $c_{st} < 0$ for some $(s,t) \in E$, then any optimal solution x^* will have $x^*_{st} = 0$. If $c_{st} = 0$, then we can set $x^*_{st} = 0$ without changing the objective value. Thus, we can assume $c_{st} > 0$ for all $(s,t) \in E$. By scaling the weights, we can further assume that $c_{st} \in \{1, \dots, K\}$ for all $(s,t) \in E$, for some $K \in \mathbb{N}$.

We first upper-bound the objective function. For $1 \le i \le K$, let $E_i = \{(s,t) \in E \colon c_{st} \ge i\}$ denote the set of edges with weights at least i, and let V_i denote the set of vertices in E_i. By construction, we have

$$V = V_1 \supset \cdots \supset V_K, \qquad \text{and} \qquad E = E_1 \supset \cdots \supset E_K.$$

Suppose the subgraph $G_i = (V_i, E_i)$ is decomposed into connected components

$$G_i = T_{i1} \cup \cdots \cup T_{i\alpha_i} \cup H_{i1} \cup \cdots \cup H_{i\beta_i}, \tag{3.30}$$

where each $T_{ij} = (V(T_{ij}), E(T_{ij}))$ is a tree and each $H_{i\ell} = (V(H_{i\ell}), E(H_{i\ell}))$ is a connected graph with at least one loop. Thus, we have the disjoint partitions

$$V_i = \bigcup_{j=1}^{\alpha_i} V(T_{ij}) \cup \bigcup_{\ell=1}^{\beta_i} V(H_{i\ell}),$$

and

$$E_i = \bigcup_{j=1}^{\alpha_i} E(T_{ij}) \cup \bigcup_{\ell=1}^{\beta_i} E(H_{i\ell}).$$

Then we can write the objective function of the LP as

$$\sum_{(s,t)\in E} c_{st} x_{st} = \sum_{i=1}^{K} \sum_{(s,t)\in E_i} x_{st} = \sum_{i=1}^{K} \left(\sum_{j=1}^{\alpha_i} \sum_{(s,t)\in E(T_{ij})} x_{st} + \sum_{\ell=1}^{\beta_i} \sum_{(s,t)\in E(H_{i\ell})} x_{st} \right). \tag{3.31}$$

For $i = 1, \dots, K$ and $j = 1, \dots, \alpha_i$, since T_{ij} is a tree, we have

$$\sum_{(s,t)\in E(T_{ij})} x_{st} \le |E(T_{ij})| = |V(T_{ij})| - 1, \qquad \forall x \in X. \tag{3.32}$$

For $\ell = 1, \ldots, \beta_i$, note that the set $E(H_{i\ell})$ of edges in $H_{i\ell}$ is contained within the set $E(V(H_{i\ell}))$ of edges in the subgraph of G induced by $V(H_{i\ell})$. Thus, by inequality (3.29), we have

$$\sum_{(s,t)\in E(H_{i\ell})} x_{st} \leq \sum_{(s,t)\in E(V(H_{i\ell}))} x_{st} \leq |V(H_{i\ell})|. \tag{3.33}$$

Plugging in the bounds (3.32) and (3.33) to inequality (3.31), we arrive at the upper bound

$$\sum_{(s,t)\in E} c_{st}x_{st} \leq \sum_{i=1}^{K} \left(\sum_{j=1}^{\alpha_i} \big\{ |V(T_{ij})| - 1 \big\} + \sum_{\ell=1}^{\beta_i} |V(H_{i\ell})| \right) = \sum_{i=1}^{K} \left(|V_i| - \alpha_i \right). \tag{3.34}$$

We now prove the claim in the proposition by explicitly constructing an integral vector x^* that achieves the upper bound (3.34). Since $x^* \in \{0,1\}^{|E|}$, it is the indicator vector of a subset $E^* \subseteq E$.

Our approach is to construct, for each $1 \leq i \leq K$, a spanning single-cycle forest $F_i = (V_i, C_i)$ of $G_i = (V_i, E_i)$ with the following properties:

1. The restriction of F_i to $V_{i+1} \subseteq V_i$ is equal to $F_{i+1} = (V_{i+1}, C_{i+1})$, or equivalently, $C_i \cap E_{i+1} = C_{i+1}$. By induction, this implies $C_1 \cap E_i = C_i$, for $1 \leq i \leq K$.

2. For $1 \leq i \leq K$, we have $|C_i| = |V_i| - \alpha_i$.

Suppose we can construct such F_i's. Setting $E^* = C_1$, we see that this construction yields a vector $x^* = 1_{E^*}$ satisfying

$$\begin{aligned} \sum_{(s,t)\in E} c_{st}x^*_{st} = \sum_{i=1}^{K} \sum_{(s,t)\in E_i} x^*_{st} &= \sum_{i=1}^{K} \sum_{(s,t)\in E_i} \mathbf{1}\{(s,t) \in C_1\} \\ &= \sum_{i=1}^{K} |C_1 \cap E_i| = \sum_{i=1}^{K} |C_i| = \sum_{i=1}^{K} \big(|V_i| - \alpha_i \big), \end{aligned}$$

so x^* achieves the bound (3.34), as desired.

It now remains to construct the F_i's. We start by taking F_K to be a spanning single-cycle forest of G_K. Specifically, for each connected component H of G_K, we do the following: If H is a tree, we take H to be in F_K. If H contains at least one loop, then we take an arbitrary spanning single-cycle subgraph (i.e., a spanning tree with an additional edge to form one cycle) of H to be in F_K. Then $F_K = (V_K, C_K)$ satisfies $|C_K| = |V_K| - \alpha_K$, since there are α_K trees among the connected components of G_K.

Suppose that for some $1 \leq i \leq K-1$, we have constructed a spanning single-cycle forest F_{i+1} satisfying the desired properties. Now consider $G_i = (V_i, E_i)$, and construct $F_i = (V_i, C_i)$ as follows: Consider each connected component of G_i in the decomposition (3.30).

(a) For each tree $T_{ij} = (V(T_{ij}), E(T_{ij}))$, for all $1 \leq j \leq \alpha_i$, take T_{ij} to be in F_i. This component of F_i is clearly consistent with F_{i+1}, and the contribution to the total number of edges $|C_i|$ is

$$\sum_{j=1}^{\alpha_i} |E(T_{ij})| = \sum_{j=1}^{\alpha_i} \big(|V(T_{ij})| - 1\big) = \sum_{j=1}^{\alpha_i} |V(T_{ij})| - \alpha_i.$$

(b) Consider $H_{i\ell} = (V(H_{i\ell}), E(H_{i\ell}))$, for some $1 \leq \ell \leq \beta_i$, so $H_{i\ell}$ has at least one loop. There may be several connected components of F_{i+1} in $H_{i\ell}$; suppose there are $\gamma_{i\ell}$ trees and $\delta_{i\ell}$ single-cycle graphs from F_{i+1} in $H_{i\ell}$. From each of the $\delta_{i\ell}$ single-cycle graphs, remove one edge to reduce it to a tree, and complete the $\gamma_{i\ell} + \delta_{i\ell}$ trees into a spanning tree of $H_{i\ell}$. Add the $\delta_{i\ell}$ edges back, so the spanning tree now has $\delta_{i\ell}$ cycles. Remove $\delta_{i\ell} - 1$ edges to break this graph into $\delta_{i\ell}$ connected components, such that each of the original $\delta_{i\ell}$ single-cycle graphs is in a separate connected components, and the last connected component is a tree. Set this new graph to be in F_i. It is clear by construction that this component of F_i is consistent with F_{i+1} since we keep all the edges from F_{i+1}. Moreover, its contribution to the total number of edges C_i is precisely

$$\sum_{\ell=1}^{\beta_i} \big(\{|V(H_{i\ell})| - 1\} + \delta_{i\ell} - \{\delta_{i\ell} - 1\}\big) = \sum_{\ell=1}^{\beta_i} |V(H_{i\ell})|.$$

Combining the two cases above, for each $1 \leq i \leq K$ we have constructed a spanning single-cycle forest F_i that is consistent with F_{i+1} and satisfies $|C_i| = \sum_{j=1}^{\alpha_i} |V(T_{ij})| - \alpha_i + \sum_{\ell=1}^{\beta_i} |V(H_{i\ell})| = |V_i| - \alpha_i$, as desired. This completes the proof of the proposition.

□

Details for Example 3.5

It is easy to check that $\mathbb{F} = \{0,1\}^3 \backslash (1,1,1)$. Hence, $(1, \frac{1}{2}, 1) \notin \text{conv}(\mathbb{F})$. By enumerating the inequalities defining the boundary of $\mathbb{C}$ for different values of $U \subseteq V$, one may check that the only inequalities that are not trivially satisfied by $\rho \in [0,1]^3$ are

$$\begin{aligned} \rho_1 + 2\rho_2 + \rho_3 &\leq 3, \\ 2\rho_1 + 2\rho_2 + \rho_3 &\leq 4, \\ \rho_1 + 2\rho_2 + 2\rho_3 &\leq 4, \\ 2\rho_1 + 2\rho_2 + 2\rho_3 &\leq 5. \end{aligned}$$

The first inequality together with the condition $\rho \in [0,1]^3$ implies the remaining three inequalities, so

$$\mathbb{C} = \left\{\rho \in [0,1]^3 \colon \rho_1 + 2\rho_2 + \rho_3 \leq 3\right\}.$$

Clearly, $(1, \frac{1}{2}, 1) \in \mathbb{C}$.

Proof of Proposition 3.6

The first condition implies $F \notin \mathfrak{F}$. In particular, $F^* \neq F$ and we can find $\alpha^* \in F \setminus F^*$. Since F^* is maximal, $\tilde{F} = F^* \cup \{\alpha^*\} \notin \mathfrak{F}$. This means $1_{F^*} \in \mathbb{F}$ but $1_{\tilde{F}} = 1_{F^*} + 1_{\{\alpha^*\}} \notin \mathbb{F}$. Define

$$\rho = 1_{F^*} + \epsilon 1_{\{\alpha^*\}}, \qquad \text{with} \qquad \epsilon = \frac{1}{|\alpha^*| - 1} \in (0, 1).$$

We claim that $\rho \in \mathbb{C}$, which will give us the desired conclusion since $\rho \notin \text{conv}(\mathbb{F})$.

To show $\rho \in \mathbb{C}$, since we already know that $1_{F^*} \in \mathbb{F} \subseteq \mathbb{C}$, we only need to verify inequality (3.15) for $U \subseteq V$ with $U \cap \alpha^* \neq \emptyset$. Given such a subset U, note that since $F^* \cup N(F^*)$ is a forest, the subgraph induced by the nodes $U \cup \{\alpha \in F^* \colon \alpha \cap U \neq \emptyset\}$ is also a forest, so

$$\sum_{\alpha \in F^* \colon \alpha \cap U \neq \emptyset} (|\alpha \cap U| - 1) \leq |U| - 1.$$

Therefore,

$$\sum_{\alpha \in F \colon \alpha \cap U \neq \emptyset} (|\alpha \cap U| - 1)\rho_\alpha = \sum_{\alpha \in F^* \colon \alpha \cap U \neq \emptyset} (|\alpha \cap U| - 1) + \frac{|\alpha^* \cap U| - 1}{|\alpha^*| - 1} \leq |U| - 1 + 1 = |U|,$$

verifying condition (3.15), as desired.

3.9 Proof of Theorem 3.7

For $r \in R$ and $s \in \mathcal{P}(r)$, let $\lambda_{sr}(x_r)$ be a Lagrange multiplier associated with the consistency constraint $\sum_{x_{s \setminus r}} \tau_s(x_r, x_{s \setminus r}) = \tau_r(x_r)$. We enforce the nonnegativity constraint $\tau_r(x_r) \geq 0$ and normalization constraint $\sum_{x_r} \tau_r(x_r) = 1$ explicitly. Then the Lagrangian associated with the optimization problem (3.8) is

$$\begin{aligned} \mathcal{L}_{\theta,\rho}(\tau; \lambda) = &\sum_{r \in R} \sum_{x_r} \tau_r(x_r)\theta_r(x_r) - \sum_{r \in R} \rho_r \sum_{x_r} \tau_r(x_r) \log \tau_r(x_r) \\ &+ \sum_{r \in R} \sum_{t \in \mathcal{C}(r)} \sum_{x_t} \lambda_{rt}(x_t) \left(\tau_t(x_t) - \sum_{x_{r \setminus t}} \tau_r(x_t, x_{r \setminus t}) \right). \end{aligned} \tag{3.35}$$

Setting the partial derivatives of $\mathcal{L}_{\theta,\rho}$ with respect to the Lagrange multipliers equal to zero recovers the consistency constraints. Taking the derivative of $\mathcal{L}_{\theta,\rho}$ with respect to $\tau_r(x_r)$ and setting it equal to zero yields

$$\log \tau_r(x_r) = C + \frac{\theta_r(x_r)}{\rho_r} + \sum_{s \in \mathcal{P}(r)} \frac{\lambda_{sr}(x_r)}{\rho_r} - \sum_{t \in \mathcal{C}(r)} \frac{\lambda_{rt}(x_t)}{\rho_r},$$

where C is a constant that enforces the normalization condition $\sum_{x_r} \tau_r(x_r) = 1$. Defining the messages by

$$\log M_{sr}(x_r) = \frac{\lambda_{sr}(x_r)}{\rho_s},$$

we can write the equation above as

$$\tau_r(x_r) \propto \exp\left(\frac{\theta_r(x_r)}{\rho_r}\right) \frac{\prod_{s\in\mathcal{P}(r)} M_{sr}(x_r)^{\rho_s/\rho_r}}{\prod_{t\in\mathcal{C}(r)} M_{rt}(x_t)},$$

recovering equation (3.17).

For $s \in R$ and $r \in \mathcal{C}(s)$, enforcing the consistency condition $\sum_{x_{s\setminus r}} \tau_s(x_r, x_{s\setminus r}) = \tau_r(x_r)$ gives us

$$\exp\left(\frac{\theta_r(x_r)}{\rho_r}\right) \frac{M_{sr}(x_r)^{\rho_s/\rho_r} \prod_{u\in\mathcal{P}(r)\setminus s} M_{ur}(x_r)^{\rho_u/\rho_r}}{\prod_{t\in\mathcal{C}(r)} M_{rt}(x_t)} \propto \sum_{x_{s\setminus r}} \exp\left(\frac{\theta_s(x_s)}{\rho_s}\right) \frac{\prod_{v\in\mathcal{P}(s)} M_{vs}(x_s)^{\rho_v/\rho_s}}{M_{sr}(x_r) \prod_{w\in\mathcal{C}(s)\setminus r} M_{sw}(x_w)}.$$

Rearranging the equation to collect $M_{sr}(x_r)$ on the left hand side and taking the $(1+\rho_s/\rho_r)$-th root on both sides gives us the update equation (3.16).

From the derivation above, it is clear that if $\{M_{sr}(x_r)\}$ is a fixed point of the update equation (3.16), then the collection τ of pseudomarginals defined by (3.17) is a stationary point of the Lagrangian (3.35), since it sets the derivatives of $\mathcal{L}_{\theta,\rho}$ equal to zero.

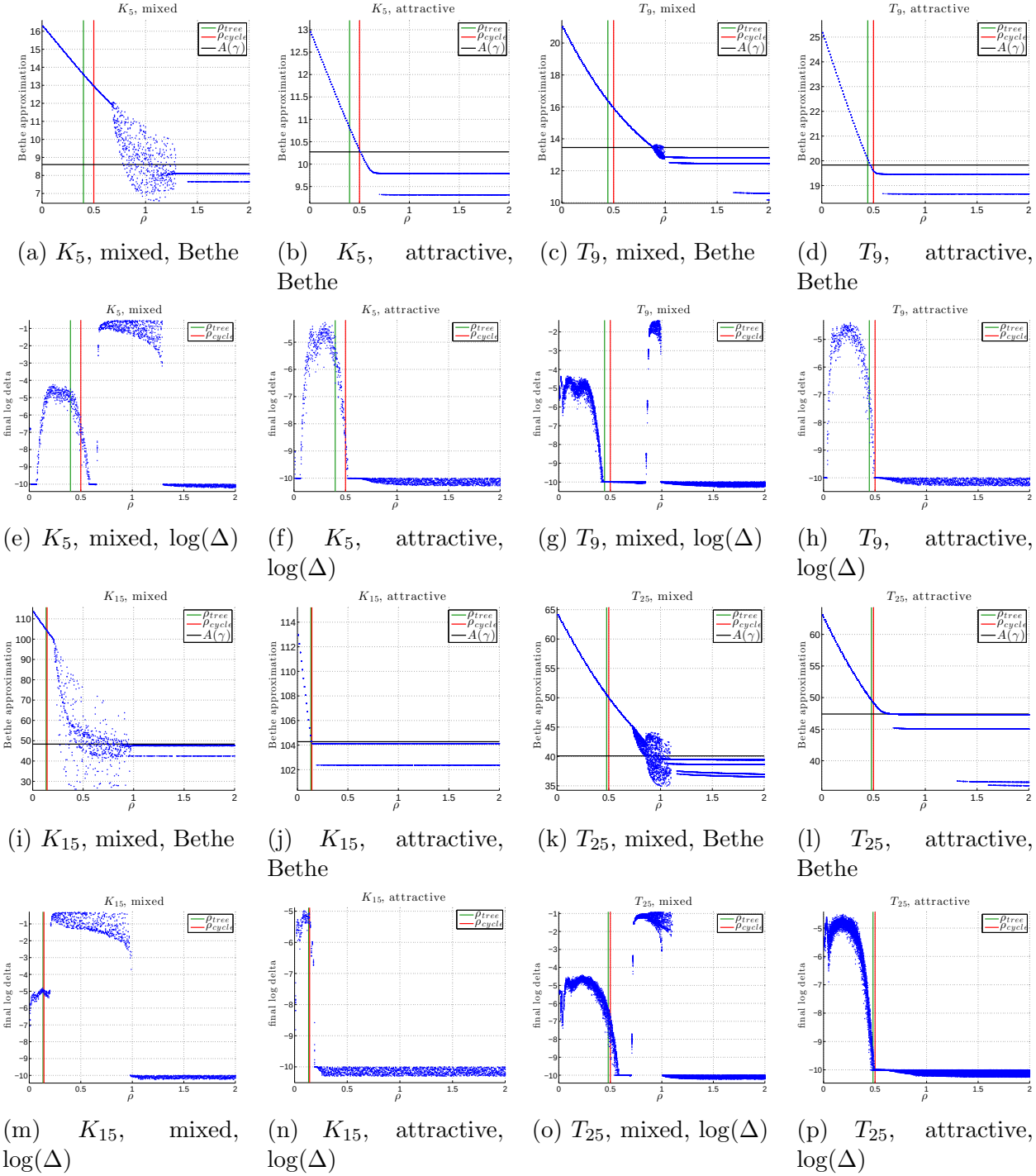
3.10 Additional simulation results

In this section, we provide additional plots to better illustrate the observations that we make in Section 3.4. For convenience, Figures 3.2a–3.2d and Figures 3.2i–3.2l show the same plots as in Figure 3.1. Figures 3.2e–3.2h show the plots of $(\rho, \log_{10}(\Delta))$ for the Ising models in Figures 3.2a–3.2d, and similarly for Figures 3.2m–3.2p. Here, Δ is the final average change of the messages in the sum product algorithm at termination; i.e., either when $\Delta \leq 10^{-10}$ or after 2500 iterations of the algorithm with parallel updates.

For $\rho \leq \rho_{\text{cycle}}$, in which the Bethe variational problem (3.8) is concave, there is a unique optimal value for the Bethe approximation. The values of Δ in this region are slightly higher than the convergence threshold, which means sum product has not converged after 2500 iterations, but the final value of Δ is sufficiently small that the messages have stabilized.

Shortly after ρ becomes larger than ρ_{cycle}, the curve of the Bethe values splits into multiple lines, which indicates that the Bethe objective function has multiple local optima. These lines are evidently distinct local optima since the values of Δ are at the convergence threshold, which means sum product converges and yields stationary points of the Lagrangian.

In the models with mixed potentials, we observe that for the values of ρ where the multiple local optima begin to emerge, the values of Δ are significantly higher and sum product does not converge. This behavior is reflected in the presence of the point cloud in the plots of the Bethe values. As noted in Section 3.4, we suspect that this behavior arises because distinct local optima are initially close together, so messages oscillate between them. For larger values of ρ, however, the local optima are sufficiently separated, so sum product converges and there are multiple lines in the graphs of the Bethe values.



(a) K_5, mixed, Bethe (b) K_5, attractive, Bethe (c) T_9, mixed, Bethe (d) T_9, attractive, Bethe

(e) K_5, mixed, $\log(\Delta)$ (f) K_5, attractive, $\log(\Delta)$ (g) T_9, mixed, $\log(\Delta)$ (h) T_9, attractive, $\log(\Delta)$

(i) K_{15}, mixed, Bethe (j) K_{15}, attractive, Bethe (k) T_{25}, mixed, Bethe (l) T_{25}, attractive, Bethe

(m) K_{15}, mixed, $\log(\Delta)$ (n) K_{15}, attractive, $\log(\Delta)$ (o) T_{25}, mixed, $\log(\Delta)$ (p) T_{25}, attractive, $\log(\Delta)$

Figure 3.2: Values of the reweighted Bethe approximation and the final $\log_{10}(\Delta)$ as a function of ρ.

Chapter 4

How to Hedge an Option Against an Adversary: Black-Scholes Pricing is Minimax Optimal

An *option* is a financial contract that allows the purchase or sale of a given asset, such as a stock, bond, or commodity, for a predetermined price on a predetermined date. The contract is named as such because the transaction in question is optional for the purchaser of the contract. Options are bought and sold for any number of reasons, but in particular they allow firms and individuals with risk exposure to hedge against potential price fluctuations. Airlines, for example, have heavy fuel costs and hence are frequent buyers of oil options.

What ought we pay for the privilege of purchasing an asset at a fixed price on a future expiration date? The difficulty with this question, of course, is that while we know the asset's previous prices, we are uncertain as to its future price. In a seminal paper from 1973, Fischer Black and Myron Scholes introduced what is now known as the Black-Scholes Option Pricing Model, which led to a boom in options trading as well as a huge literature on the problem of derivative pricing [13]. Black and Scholes had a key insight that a firm which had sold/purchased an option could "hedge" against the future cost/return of the option by buying and selling the underlying asset as its price fluctuates. Their model is based on stochastic calculus and requires a critical assumption that the asset's price behaves according to a *Geometric Brownian Motion* (GBM) with known drift and volatility.

The GBM assumption in particular implies that (almost surely) an asset's price fluctuates continuously. The Black-Scholes model additionally requires that the firm be able to buy and sell continuously until the option's expiration date. Neither of these properties are true in practice: the stock market is only open eight hours per day, and stock prices are known to make significant jumps even during regular trading. These and other empirical observations have led to much criticism of the Black-Scholes model.

An alternative model for option pricing was considered[1] by DeMarzo et al. [16], who posed

[1] A similar approach with a slightly distinct flavor was explored in the book of Vovk and Shafer [58].

the question: "Can we construct hedging strategies that are robust to *adversarially chosen* price fluctuations?" Essentially, the authors asked if we may consider hedging through the lens of *regret minimization in online learning*, an area that has proved fruitful, especially for obtaining guarantees robust to worst-case conditions. Within this minimax option pricing framework, DeMarzo et al. provided a particular algorithm resembling the Weighted Majority and Hedge algorithms [19, 38] with a nice bound.

In an earlier work [1], we have taken the minimax option pricing framework a step further by analyzing the zero-sum game being played between an Investor, who is attempting to replicate the option payoff, and Nature, who is sequentially setting the price changes of the underlying asset. The Investor's goal is to "hedge" the payoff of the option as the price fluctuates, whereas Nature attempts to foil the Investor by choosing a challenging sequence of price fluctuations. The *value* of this game can be interpreted as the "minimax option price," since it is what the Investor should pay for the option against an adversarially chosen price path. Our main result in [1] was to show that the game value approaches the Black-Scholes option price as the Investor's trading frequency increases. Put another way, the minimax price tends to the option price under the GBM assumption. This lends significant further credibility to the Black-Scholes model, as it suggests that the GBM assumption may already be a "worst-case model" in a certain sense.

The previous result, while useful and informative, left two significant drawbacks. First, our techniques used minimax duality to compute the value of the game, but no particular hedging algorithm for the Investor is given. This is in contrast to the Black-Scholes framework (as well as to the DeMarzo et al.'s result [16]) in which a hedging strategy is given explicitly. Second, the result depended on a strong constraint on Nature's choice of price path: the multiplicative price variance is uniformly constrained, which forbids price jumps and other large fluctuations.

In this work, we resolve these two drawbacks. We consider the problem of minimax option pricing with much weaker constraints: we restrict the *sum* over the length of the game of the squared price fluctuations to be no more than a constant c, and we allow arbitrary price jumps, up to a bound ζ. We show that the minimax option price is exactly the Black-Scholes price of the option, up to an additive term of $O(c\zeta^{1/4})$. Furthermore, we give an explicit hedging strategy: this upper bound is achieved when the Investor's strategy is essentially a version of the Black-Scholes hedging algorithm.

4.1 The Black-Scholes formula

Let us now briefly review the Black-Scholes pricing formula and hedging strategy. The derivation requires some knowledge of continuous random walks and stochastic calculus—Brownian motion, Itô's Lemma, a second-order partial differential equation—which can be found in standard references on stochastic calculus, e.g., [60].

Let us imagine we have an underlying asset A whose price is fluctuating. We let $W(t)$ be a *Brownian motion*, also known as a *Wiener process*, with zero drift and unit variance;

in particular, $W(0) = 0$ and $W(t) \sim N(0, t)$ for $t > 0$. We shall imagine that A's price path $G(t)$ is described by a *geometric Brownian motion* with drift μ and volatility σ, which we can describe via the definition of a Brownian motion:

$$G(t) \stackrel{d}{=} \exp\{(\mu - \frac{1}{2}\sigma^2)t + \sigma W(t)\}.$$

If an Investor purchases a *European call* option on some asset A (say, MSFT stock) with a *strike price* of $K > 0$ that matures at time T, then the Investor has the right to buy a share of A at price K at time T. Of course, if the market price of A at T is $G(T)$, then the Investor will only "exercise" the option if $G(T) > K$, since the Investor has no benefit of purchasing the asset at a price higher than the market price. Hence, the payoff of a European call option has a profit function of the form $\max\{0, G(T) - K\}$. Throughout the chapter we shall use

$$g_{\text{EC}}(x) := \max\{0, x - K\}$$

to refer to the payout of the European call when the price of asset A at time T is x (the parameter K is implicit).

We assume the current time is t. The Black-Scholes derivation begins with a guess: assume that the "value" of the European call option can be described by a smooth function $\mathcal{V}(G(t), t)$, depending only on the current price of the asset $G(t)$ and the time to expiration $T - t$. We can immediately define a boundary condition on $\mathcal{V}$, since at the expiration time T the value of the option is

$$\mathcal{V}(G(T), 0) = g_{\text{EC}}(G(T)).$$

So how do we arrive at a value for the option at another time point t? We assume the Investor has a *hedging strategy*, $\Delta(x, t)$ that determines the amount to invest when the current price is x and the time is t. Notice that if the asset's current price is $G(t)$ and the Investor purchases $\Delta(G(t), t)$ dollars of asset A at t, then the incremental amount of money made in an infinitesimal amount of time is $\Delta(G(t), t)\ dG/G(t)$, since $dG/G(t)$ is the instantaneous multiplicative price change at time t. Of course, if the earnings of the Investor are guaranteed to exactly cancel out the infinitesimal change in the value of the option $d\mathcal{V}(G(t), t)$, then the Investor is totally hedged with respect to the option payout for any sample of G for the remaining time to expiration. In other words, we hope to achieve

$$d\mathcal{V}(G, t) = \frac{\Delta(G, t)}{G}\ dG.$$

However, by Itô's Lemma [60] we have the following useful identity:

$$d\mathcal{V}(G, t) = \frac{\partial \mathcal{V}}{\partial x} dG + \frac{\partial \mathcal{V}}{\partial t} dt + \frac{1}{2}\sigma^2 G^2 \frac{\partial^2 \mathcal{V}}{\partial x^2} dt. \tag{4.1}$$

Black and Scholes proposed a generic hedging strategy, that the investor should invest

$$\Delta(x, t) = x \frac{\partial \mathcal{V}}{\partial x} \tag{4.2}$$

dollars in the asset A when the price of A is x at time t. As mentioned, the goal of the Investor is to hedge out risk so that it is always the case that $d\mathcal{V}(G,t) = \Delta(G,t)\, dG/G$. Combining this goal with Equations (4.1) and (4.2), we have

$$\frac{\partial \mathcal{V}}{\partial t} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 \mathcal{V}}{\partial x^2} = 0. \tag{4.3}$$

Notice the latter is an entirely non-stochastic PDE, and indeed it can be solved explicitly:

$$\mathcal{V}(x,t) = \mathbb{E}_Y[g_{\mathrm{EC}}(x \cdot \exp(Y))] \quad \text{where} \quad Y \sim \mathcal{N}(-\tfrac{1}{2}\sigma^2(T-t), \sigma^2(T-t)) \tag{4.4}$$

Remark: While we have described the derivation for the European call option, with payoff function g_{EC}, the analysis above does not rely on this specific choice of g. We refer the reader to a standard text on asset pricing for more on this [60].

4.2 The minimax hedging game

We now describe a sequential decision protocol in which an Investor makes a sequence of trading decisions on some underlying asset, with the goal of hedging away the risk of some option (or other financial derivative) whose payout depends on the final price of the asset at the expiration time T. We assume the Investor is allowed to make a trading decision at each of n time periods, and before making this trade the investor observes how the price of the asset has changed since the previous period. Without loss of generality, we can assume that the current time is 0 and the trading periods occur at $\{T/n,\ 2T/n, \ldots, 1\}$, although this will not be necessary for our analysis.

The protocol is as follows.

Initial price of asset is $S = S_0$.
for $i = 1, 2, \ldots, n$ **do**
    Investor hedges, invests $\Delta_i \in \mathbb{R}$ dollars in asset.
    Nature selects a price fluctuation r_i and updates price $S \leftarrow S(1 + r_i)$.
    Investor receives (potentially negative) profit of $\Delta_i r_i$.
end for
Investor is charged the cost of the option, $g(S) = g\left(S_0 \cdot \prod_{i=1}^{n}(1 + r_i)\right)$.

Stepping back for a moment, we see that the Investor is essentially trying to minimize the following objective:

$$g\left(S_0 \cdot \prod_{i=1}^{n}(1 + r_i)\right) - \sum_{i=1}^{n} \Delta_i r_i.$$

We can interpret the above expression as a form of *regret*: the Investor chose to execute a trading strategy, earning him $\sum_{i=1}^{n} \Delta_i r_i$, but in hindsight might have rather purchased the option instead, with a payout of $g\left(S_0 \cdot \prod_{i=1}^{n}(1 + r_i)\right)$. What is the best hedging strategy the Investor can execute to minimize the difference between the option payoff and the gains/losses from hedging? Indeed, how much regret may be suffered against a worst-case sequence of price fluctuations?

Constraining Nature. The cost of playing the above sequential game is clearly going to depend on how much we expect the price to fluctuate. In the original Black-Scholes formulation, the price volatility σ is a major parameter in the pricing function. In our earlier work [1], a key assumption was that Nature may choose any $r_1, \ldots, r_n$ with the constraint that $\mathbb{E}[r_i^2 \mid r_1, \ldots, r_{i-1}] = O(1/n)$.[2] Roughly, this constraint means that in any ϵ-sized time interval, the price fluctuation variance shall be no more than ϵ. This constraint, however, does not allow for large price jumps during trading. In the present work, we impose a much weaker set of constraints, described as follows:[3]

- **TotVarConstraint**: The total price fluctuation is bounded by a constant c, that is, $\sum_{i=1}^{n} r_i^2 \leq c$.
- **JumpConstraint**: Every price jump $|r_i|$ is no more than ζ, for some $\zeta > 0$ (which may depend on n).

The first constraint above says that Nature is bounded by how much, in total, the asset's price path can fluctuate. The latter says that at no given time can the asset's price jump more than a given value. It is worth noting that if $c \geq n\zeta^2$ then **TotVarConstraint** is superfluous, whereas **JumpConstraint** becomes superfluous if $c < \zeta^2$.

The Minimax Option Price We are now in a position to define the *value* of the sequential option pricing game using a minimax formulation. That is, we shall ask how much the Investor loses when making optimal trading decisions against worst-case price fluctuations chosen by Nature.

Let $V_\zeta^{(n)}(S; c, m)$ be the value of the game, measured by the investor's *loss*, when the asset's current price is $S \geq 0$, the **TotVarConstraint** is $c \geq 0$, the **JumpConstraint** is $\zeta > 0$, the total number of trading rounds are $n \in \mathbb{N}$, and there are $0 \leq m \leq n$ rounds remaining. We define recursively:

$$V_\zeta^{(n)}(S; c, m) = \inf_{\Delta \in \mathbb{R}} \sup_{r:\, |r| \leq \min\{\zeta, \sqrt{c}\}} -\Delta r + V_\zeta^{(n)}((1+r)S;\; c - r^2, m-1), \tag{4.5}$$

with the base case $V_\zeta^{(n)}(S; c, 0) = g(S)$. Notice that the constraint under the supremum enforces both **TotVarConstraint** and **JumpConstraint**. For simplicity, we will write $V_\zeta^{(n)}(S; c) := V_\zeta^{(n)}(S; c, n)$. This is the value of the game that we are interested in analyzing.

Towards establishing an upper bound on the value (4.5), we shall discuss the question of how to choose the hedge parameter Δ on each round. We can refer to a "hedging strategy" in this game as a function of the tuple $(S, c, m, n, \zeta, g(\cdot))$ that returns hedge position. In our upper bound, in fact we need only consider hedging strategies $\Delta(S, c)$ that depend on S and c; there certainly will be a dependence on $g(\cdot)$ as well but we leave this implicit.

[2] The constraint in [1] was $\mathbb{E}[r_i^2 \mid r_1, \ldots, r_{i-1}] \leq \exp(c/n) - 1$, but this is roughly equivalent.

[3] We note that in [1] we also assumed that the multiplicative price jumps $|r_i|$ are bounded by $\hat{\zeta}_n = \Omega(\sqrt{(\log n)/n})$; this is a stronger assumption than what we impose on (ζ_n) in Theorem 4.1.

4.3 Asymptotic results: Convergence to the Black-Scholes price

The central focus of the present work is the following questions:

For fixed c and S, what is the asymptotic behavior of the value $V_\zeta^{(n)}(S;c)$?

and

Is there a natural hedging strategy $\Delta(S,c)$ that (roughly) achieves this value?

In other words, what is the minimax value of the option, as well as the optimal hedge, when we fix the variance budget c and the asset's current price S, but let the number of rounds tend to ∞? We now give answers to these questions, and devote the remainder of the chapter to developing the results in detail.

We consider payoff functions $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ satisfying three constraints:

1. g is convex.
2. g is L-Lipschitz, i.e. $|g(x) - g(y)| \le L|x - y|$.
3. g is *eventually linear*, i.e. there exists $K > 0$ such that $g(x)$ is a linear function for all $x \ge K$; in this case we also say g is *K-linear*.

We believe the first two conditions are strictly necessary to achieve the desired results. The K-linearity may not be necessary but makes our analysis possible. We note that the constraints above encompass the standard European call and put options.

Henceforth we shall let G be a *zero-drift* GBM with unit volatility. In particular, we have that

$$\log G(t) \sim \mathcal{N}\left(-\frac{1}{2}t, t\right).$$

For $S, c \ge 0$, define the function

$$U(S,c) = \mathbb{E}_G[g(S \cdot G(c))],$$

and observe that $U(S,0) = g(S)$. Our goal will be to show that U is asymptotically the minimax price of the option. Most importantly, this function $U(S,c)$ is *identical* to $\mathcal{V}(S, \frac{1}{\sigma^2}(T-c))$, the Black-Scholes value of the option in (4.4) when the GBM volatility parameter is σ in the Black-Scholes analysis. In particular, analogous to to (4.3), $U(S,c)$ satisfies a differential equation:

$$\frac{1}{2}S^2\frac{\partial^2 U}{\partial S^2} - \frac{\partial U}{\partial c} = 0. \tag{4.6}$$

The following is our main result of this chapter.

Theorem 4.1. *Let $S > 0$ be the initial asset price and let $c > 0$ be the variance budget. Assume we have a sequence $\{\zeta_n\}$ with $\lim_{n\to\infty} \zeta_n = 0$ and $\liminf_{n\to\infty} n\zeta_n^2 > c$. Then*

$$\lim_{n\to\infty} V_{\zeta_n}^{(n)}(S;c) = U(S,c).$$

This statement tells us that the minimax price of an option, when Nature has a total fluctuation budget of c, approaches the Black-Scholes price of the option when the time to expiration is c. This is particularly surprising since our minimax pricing framework made no assumptions as to the stochastic process generating the price path. This is the same conclusion as in [1], but we obtained our result with a significantly weaker assumption. Unlike [1], however, we do not show that the adversary's minimax optimal stochastic price path necessarily converges to a GBM. The convergence of Nature's price path to GBM in [1] was made possible by the uniform per-round variance constraint.

The previous theorem is the result of two main technical contributions. First, we prove a lower bound on the limiting value of $V_{\zeta_n}^{(n)}(S;c)$ by exhibiting a simple randomized strategy for Nature in the form of a stochastic price path, and appealing to the Lindeberg-Feller central limit theorem. Second, we prove an $O(c\zeta^{1/4})$ upper bound on the deviation between $V_{\zeta}^{(n)}(S;c)$ and $U(S,c)$. The upper bound is obtained by providing an explicit strategy for the Investor:

$$\Delta(S,c) = S\frac{\partial U(S,c)}{\partial S}$$

and carefully bounding the difference between the output using this strategy and the game value. In the course of doing so, because we are invoking Taylor's remainder theorem, we need to bound the first few derivatives of $U(S,c)$. Bounding these derivatives turns out to be the crux of the analysis; in particular, it uses the full force of the assumptions on g, including that g is eventually linear, to avoid the pathological cases when the derivative of g converges to its limiting value very slowly.

4.4 Lower bound

In this section we prove that $U(S,c)$ is a lower bound to the game value $V_{\zeta_n}^{(n)}(S;c)$. We note that the result in this section does not use the assumptions in Theorem 4.1 that $\zeta_n \to 0$, nor that g is convex and eventually linear. In particular, the following result also applies when the sequence (ζ_n) is a constant $\zeta > 0$.

Theorem 4.2. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be an L-Lipschitz function, and let $\{\zeta_n\}$ be a sequence of positive numbers with $\liminf_{n\to\infty} n\zeta_n^2 > c$. Then for every $S, c > 0$,*

$$\liminf_{n\to\infty} V_{\zeta_n}^{(n)}(S;c) \geq U(S,c).$$

The proof of Theorem 4.2 (given below) is based on correctly "guessing" a randomized strategy for Nature. For each $n \in \mathbb{N}$, define the i.i.d. random variables

$$R_{1,n}, \ldots, R_{n,n} \sim \text{Uniform}\left\{\pm\sqrt{\frac{c}{n}}\right\}.$$

Note that $(R_{i,n})_{i=1}^n$ satisfies **TotVarConstraint** by construction. Moreover, the assumption $\liminf_{n\to\infty} n\zeta_n^2 > c$ implies $\zeta_n > \sqrt{c/n}$ for all sufficiently large n, so eventually $(R_{i,n})$ also satisfies **JumpConstraint**. We have the following convergence result for $(R_{i,n})$, which we prove in Section 4.6.

Lemma 4.3. *Under the same setting as in Theorem 4.2, we have the convergence in distribution*

$$\prod_{i=1}^n (1 + R_{i,n}) \xrightarrow{d} G(c) \quad \text{as} \quad n \to \infty.$$

Moreover, we also have the convergence in expectation

$$\lim_{n\to\infty} \mathbb{E}\left[g\left(S \cdot \prod_{i=1}^n (1 + R_{i,n})\right)\right] = U(S, c). \tag{4.7}$$

With the help of Lemma 4.3, we are now ready to prove Theorem 4.2.

Proof of Theorem 4.2: Let n be sufficiently large such that $n\zeta_n^2 > c$. Let $R_{i,n} \sim \text{Uniform}\{\pm\sqrt{c/n}\}$ i.i.d., for $1 \le i \le n$. As noted above, $(R_{i,n})$ satisfies both **TotVarConstraint** and **JumpConstraint**. Then we have

$$\begin{aligned}
V_{\zeta_n}^{(n)}(S; c) &= \inf_{\Delta_1} \sup_{r_1} \cdots \inf_{\Delta_n} \sup_{r_n} \; g\Big(S \cdot \prod_{i=1}^n (1 + r_i)\Big) - \sum_{i=1}^n \Delta_i r_i \\
&\ge \inf_{\Delta_1} \cdots \inf_{\Delta_n} \; \mathbb{E}\Big[g\Big(S \cdot \prod_{i=1}^n (1 + R_{i,n})\Big) - \sum_{i=1}^n \Delta_i R_{i,n}\Big] \\
&= \mathbb{E}\Big[g\Big(S \cdot \prod_{i=1}^n (1 + R_{i,n})\Big)\Big].
\end{aligned}$$

The first line follows from unrolling the recursion in the definition (4.5); the second line from replacing the supremum over (r_i) with expectation over $(R_{i,n})$; and the third line from $\mathbb{E}[R_{i,n}] = 0$. Taking limit on both sides and using (4.7) from Lemma 4.3 give us the desired conclusion. □

4.5 Upper bound

In this section we prove that $U(S, c)$ is an upper bound to the limit of $V_\zeta^{(n)}(S; c)$.

Theorem 4.4. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be a convex, L-Lipschitz, K-linear function. Let $0 < \zeta \leq 1/16$. Then for any $S, c > 0$ and $n \in \mathbb{N}$, we have*

$$V_\zeta^{(n)}(S;c) \leq U(S,c) + \left(18c + \frac{8}{\sqrt{2\pi}}\right) LK\,\zeta^{1/4}.$$

We remark that the right-hand side of the above bound does not depend on the number of trading periods n. The key parameter is ζ, which determines the size of the largest price jump of the stock. However, we expect that as the trading frequency increases, the size of the largest price jump will shrink. Plugging a sequence $\{\zeta_n\}$ in place of ζ in Theorem 4.4 gives us the following corollary.

Corollary 4.5. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be a convex, L-Lipschitz, K-linear function. Let $\{\zeta_n\}$ be a sequence of positive numbers with $\zeta_n \to 0$. Then for $S, c > 0$,*

$$\limsup_{n\to\infty} V_{\zeta_n}^{(n)}(S;c) \leq U(S,c).$$

Note that the above upper bound relies on the convexity of g, for if g were concave, then we would have the reverse conclusion:

$$V_\zeta^{(n)}(S;c) \geq g(S) = g(S \cdot \mathbb{E}[G(c)]) \geq \mathbb{E}[g(S \cdot G(c))] = U(S,c).$$

Here the first inequality follows from setting all $r = 0$ in (4.5), and the second is by Jensen's inequality.

Intuition

For brevity, we write the partial derivatives as

$$U_c(S,c) = \frac{\partial U(S,c)}{\partial c}, \qquad U_S(S,c) = \frac{\partial U(S,c)}{\partial S}, \qquad \text{and} \qquad U_{S^2}(S,c) = \frac{\partial^2 U(S,c)}{\partial S^2}.$$

The proof of Theorem 4.4 proceeds by providing a "guess" for the Investor's action, which is a modification of the original Black-Scholes hedging strategy. Specifically, when the current price is S and the remaining budget is c, then the Investor invests

$$\Delta(S,c) := SU_S(S,c).$$

We now illustrate how this strategy gives rise to a bound on $V_\zeta^{(n)}(S;c)$ as stated in Theorem 4.4. First suppose for some $m \geq 1$ we know that $V_\zeta^{(n)}(S;c,m-1)$ is a rough approximation to $U(S,c)$. Note that a Taylor approximation of the function $r_m \mapsto U(S + Sr_m, c - r_m^2)$ around $U(S,c)$ gives us

$$\begin{aligned} U(S + Sr_m, c - r_m^2) &= U(S,c) + r_m SU_S(S,c) - r_m^2 U_c(S,c) + \frac{1}{2} r_m^2 S^2 U_{S^2}(S,c) + O(r_m^3) \\ &= U(S,c) + r_m SU_S(S,c) + O(r_m^3), \end{aligned}$$

where the last line follows from the Black-Scholes equation (4.6). Now by setting $\Delta = SU_S(S,c)$ in the definition (4.5), and using the assumption and the Taylor approximation above, we obtain

$$\begin{aligned} V_\zeta^{(n)}(S;c,m) &= \inf_{\Delta\in\mathbb{R}} \sup_{|r_m|\le\min\{\zeta,\sqrt{c}\}} -\Delta r_m + V_\zeta^{(n)}(S+Sr_m; c-r_m^2, m-1) \\ &\le \sup_{r_m} -r_m SU_S(S,c) + V_\zeta^{(n)}(S+Sr_m; c-r_m^2, m-1) \\ &= \sup_{r_m} -r_m SU_S(S,c) + U(S+Sr_m, c-r_m^2) + (\text{approx terms}) \\ &= U(S,c) + O(r_m^3) + (\text{approx terms}). \end{aligned}$$

In other words, on each round of the game we add an $O(r_m^3)$ term to the approximation error. By the time we reach $V_\zeta^{(n)}(S;c,n)$ we will have an error term that is roughly on the order of $\sum_{m=1}^n |r_m|^3$. Since $\sum_{m=1}^n r_m^2 \le c$ and $|r_m| \le \zeta$ by assumption, we get $\sum_{m=1}^n |r_m|^3 \le \zeta c$.

The details are more intricate because the error $O(r_m^3)$ from the Taylor approximation also depends on S and c. Trading off the dependencies of c and ζ leads us to the bound stated in Theorem 4.4.

Proof of Theorem 4.4

In this section we describe an outline of the proof of Theorem 4.4. Throughout, we assume g is a convex, L-Lipschitz, K-linear function, and $0 < \zeta \le 1/16$. The proofs of Lemma 4.6 and Lemma 4.8 are provided in Section 4.7, and Lemma 4.7 is proved in Section 4.8.

For $S, c > 0$ and $|r| \le \sqrt{c}$, we define the *(single-round) error term* of the Taylor approximation,

$$\epsilon_r(S,c) := U(S+Sr, c-r^2) - U(S,c) - rSU_S(S,c). \tag{4.8}$$

We also define a sequence $\{\alpha^{(n)}(S,c,m)\}_{m=0}^n$ to keep track of the cumulative errors. We define this sequence by setting $\alpha^{(n)}(S,c,0) = 0$, and for $1 \le m \le n$,

$$\alpha^{(n)}(S,c,m) := \sup_{|r|\le\min\{\zeta,\sqrt{c}\}} \epsilon_r(S,c) + \alpha^{(n)}(S+Sr, c-r^2, m-1). \tag{4.9}$$

For simplicity, we write $\alpha^{(n)}(S,c) \equiv \alpha^{(n)}(S,c,n)$. Then we have the following result, which formalizes the notion from the preceding section that $V_\zeta^{(n)}(S;c,m)$ is an approximation to $U(S,c)$.

Lemma 4.6. *For $S, c > 0$, $n \in \mathbb{N}$, and $0 \le m \le n$, we have*

$$V_\zeta^{(n)}(S;c,m) \le U(S,c) + \alpha^{(n)}(S,c,m). \tag{4.10}$$

It now remains to bound $\alpha^{(n)}(S,c)$ from above. A key step in doing so is to show the following bounds on ϵ_r. This is where the assumptions that g be L-Lipschitz and K-linear are important.

Lemma 4.7. *For $S, c > 0$, and $|r| \le \min\{1/16,\ \sqrt{c}/8\}$, we have*

$$\epsilon_r(S,c) \le 16LK\ \left(\max\{c^{-3/2},\ c^{-1/2}\}\ |r|^3 + \max\{c^{-2},\ c^{-1/2}\}\ r^4\right). \tag{4.11}$$

Moreover, for $S > 0$, $0 < c \le 1/4$, and $|r| \le \sqrt{c}$, we also have

$$\epsilon_r(S,c) \le \frac{4LK}{\sqrt{2\pi}} \cdot \frac{r^2}{\sqrt{c}}. \tag{4.12}$$

Using Lemma 4.7, we have the following bound on $\alpha^{(n)}(S,c)$.

Lemma 4.8. *For $S, c > 0$, $n \in \mathbb{N}$, and $0 < \zeta \le 1/16$, we have*

$$\alpha^{(n)}(S,c) \le \left(18c + \frac{8}{\sqrt{2\pi}}\right)\ LK\ \zeta^{1/4}.$$

Proof (sketch). By unrolling the inductive definition (4.9), we can write $\alpha^{(n)}(S,c)$ as the supremum of $f(r_1,\ldots,r_n)$, where

$$f(r_1,\ldots,r_n) = \sum_{m=1}^{n} \epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c - \sum_{i=1}^{m-1} r_i^2\Big).$$

Let $(r_1,\ldots,r_n)$ be such that $|r_m| \le \zeta$ and $\sum_{m=1}^{n} r_m^2 \le c$. We will show that $f(r_1,\ldots,r_n) \le (18c + 8/\sqrt{2\pi})\ LK\ \zeta^{1/4}$. Let $0 \le n_* \le n$ be the largest index such that $\sum_{m=1}^{n_*} r_m^2 \le c - \sqrt{\zeta}$. We split the analysis into two parts.

1. **For $1 \le m \le \min\{n,\ n_*+1\}$:** By (4.11) from Lemma 4.7 and a little calculation, we have

   $$\epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c - \sum_{i=1}^{m-1} r_i^2\Big) \le 18LK\ \zeta^{1/4}\ r_m^2.$$

   Summing over $1 \le m \le \min\{n,\ n_*+1\}$ then gives us

   $$\sum_{m=1}^{\min\{n,\ n_*+1\}} \epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c - \sum_{i=1}^{m-1} r_i^2\Big) \le 18LK\ \zeta^{1/4} \sum_{m=1}^{\min\{n,\ n_*+1\}} r_m^2 \le 18LK\ \zeta^{1/4}\ c.$$

2. **For $n_*+2 \le m \le n$ (if $n_* \le n-2$):** By (4.12) from Lemma 4.7, we have

   $$\epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c - \sum_{i=1}^{m-1} r_i^2\Big) \le \frac{4LK}{\sqrt{2\pi}} \cdot \frac{r_m^2}{\sqrt{\sum_{i=m}^{n} r_i^2}}.$$

   Therefore,

   $$\sum_{m=n_*+2}^{n} \epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c - \sum_{i=1}^{m-1} r_i^2\Big) \le \frac{4LK}{\sqrt{2\pi}} \sum_{m=n_*+2}^{n} \frac{r_m^2}{\sqrt{\sum_{i=m}^{n} r_i^2}} \le \frac{8LK}{\sqrt{2\pi}}\ \zeta^{1/4},$$

   where the last inequality follows from Lemma 4.9 in Section 4.7.

Combining the two cases above gives us the desired conclusion. □

Proof of Theorem 4.4: Theorem 4.4 follows immediately from Lemma 4.6 and Lemma 4.8. □

4.6 Proof of Lemma 4.3

For each $1 \le i \le n$, the random variable $\log(1 + R_{i,n})$ has mean and variance, respectively,

$$\mu_n = \frac{1}{2}\log\left(1 - \frac{c}{n}\right) \quad \text{and} \quad \sigma_n^2 = \frac{1}{4}\log^2\left(\frac{\sqrt{n}+\sqrt{c}}{\sqrt{n}-\sqrt{c}}\right).$$

We now define

$$X_{i,n} := \frac{\log(1 + R_{i,n}) - \mu_n}{\sigma_n\sqrt{n}}, \tag{4.13}$$

so $X_{1,n}, \ldots, X_{n,n}$ are i.i.d. random variables with $\mathbb{E}[X_{i,n}] = 0$ and $\sum_{i=1}^n \mathbb{E}[X_{i,n}^2] = 1$. Recalling that $R_{i,n} \in \{\pm\sqrt{c/n}\}$, we see that the two possible values for $X_{i,n}$ both approach 0 as $n \to \infty$. This means for any $\epsilon > 0$ we can find a sufficiently large n such that $|X_{i,n}| < \epsilon$ for all $1 \le i \le n$. In particular, this implies the Lindeberg condition for the triangular array $(X_{i,n}, 1 \le i \le n)$: for all $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}[X_{i,n}^2\, \mathbf{1}\{|X_{i,n}| > \epsilon\}] = 0.$$

Thus, by the Lindeberg central limit theorem [17, Theorem 3.4.5], we have the convergence in distribution $\sum_{i=1}^n X_{i,n} \xrightarrow{d} Z$, where $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable.

Clearly $\mu_n \to 0$ as $n \to \infty$. Furthermore, one can easily verify that by the L'Hôpital's rule,

$$\lim_{n\to\infty} \sigma_n\sqrt{n} = \sqrt{c} \quad \text{and} \quad \lim_{n\to\infty} n\mu_n = -\frac{c}{2}.$$

Therefore, from the convergence $\sum_{i=1}^n X_{i,n} \xrightarrow{d} Z$ and recalling the definition (4.13) of $X_{i,n}$, we also obtain

$$\sum_{i=1}^n \log(1 + R_{i,n}) = \sum_{i=1}^n \big(\mu_n + \sigma_n\sqrt{n}X_{i,n}\big) = n\mu_n + \sigma_n\sqrt{n}\sum_{i=1}^n X_{i,n} \xrightarrow{d} \frac{-c}{2} + \sqrt{c}Z.$$

In particular, by the continuous mapping theorem,

$$\prod_{i=1}^n (1 + R_{i,n}) \xrightarrow{d} \exp\Big(-\frac{c}{2} + \sqrt{c}Z\Big) \stackrel{d}{=} G(c).$$

We now want to show that we also have convergence in expectation when g is an L-Lipschitz function, namely, that $\mathbb{E}[g(S \cdot \prod_{i=1}^n (1+R_{i,n}))] \to \mathbb{E}[g(S \cdot G(c))]$. Without loss of generality (by replacing $g(x)$ by $\hat{g}(x) = g(S \cdot x) - g(0)$) we may assume $S = 1$ and $g(0) = 0$. For simplicity, let $S_n = \prod_{i=1}^n (1+R_{i,n})$. For each $M > 0$ define the continuous bounded function $g_M(x) = \min\{g(x), M\}$. The convergence in distribution $S_n \xrightarrow{d} G(c)$ gives us

$$\lim_{n\to\infty} \mathbb{E}[g_M(S_n))] = \mathbb{E}[g_M(G(c))] \quad \text{for all } M > 0. \tag{4.14}$$

Since $g_M \uparrow g$ pointwise, by the monotone convergence theorem we also have

$$\lim_{M\to\infty} \mathbb{E}[g_M(G(c))] = \mathbb{E}[g(G(c))]. \tag{4.15}$$

Now observe that $\mathbb{E}[S_n] = 1$ and $\mathbb{E}[S_n^2] = (1+c/n)^n \le \exp(c)$. Since $g(0) = 0$ and g is L-Lipschitz, we have $g(x) \le Lx$ for all $x \ge 0$. In particular, $\mathbb{E}[g(S_n)^2] \le L^2 \mathbb{E}[S_n^2] \le L^2 \exp(c)$. Moreover, by Markov's inequality,

$$\mathbb{P}(g(S_n) > M) \le \mathbb{P}\left(S_n > \frac{M}{L}\right) \le \frac{\mathbb{E}[S_n]}{M/L} = \frac{L}{M}.$$

Therefore, for each n and for all $M > 0$, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\big|\mathbb{E}[g(S_n)] - \mathbb{E}[g_M(S_n)]\big| &= \mathbb{E}[(g(S_n) - M) \cdot \mathbf{1}\{g(S_n) > M\}] \\
&\le \mathbb{E}[g(S_n) \cdot \mathbf{1}\{g(S_n) > M\}] \\
&\le \mathbb{E}[g(S_n)^2]^{1/2}\, \mathbb{P}(g(S_n) > M)^{1/2} \\
&\le \big(L^3 \exp(c)/M\big)^{1/2}.
\end{aligned}$$

Since the final bound does not involve n, this shows that $\lim_{M\to\infty} \mathbb{E}[g_M(S_n)] \to \mathbb{E}[g(S_n)]$ uniformly in n. This allows us to interchange the order of the limit operations below, which, together with (4.14) and (4.15), give us our desired result:

$$\lim_{n\to\infty} \mathbb{E}[g(S_n)] = \lim_{n\to\infty} \lim_{M\to\infty} \mathbb{E}[g_M(S_n)] = \lim_{M\to\infty} \lim_{n\to\infty} \mathbb{E}[g_M(S_n)] = \lim_{M\to\infty} \mathbb{E}[g_M(G(c))] = \mathbb{E}[g(G(c))].$$

This completes the proof of Lemma 4.3.

4.7 Proofs of Lemma 4.6 and Lemma 4.8

Proof of Lemma 4.6

Lemma 4.6 essentially follows from the definition of $\alpha^{(n)}$.

Proof of Lemma 4.6: We proceed by induction on m. For the base case $m = 0$, we use Jensen's inequality and the fact that $\mathbb{E}[G(c)] = 1$:

$$V_\zeta^{(n)}(S; c, 0) = g(S) = g(S \cdot \mathbb{E}[G(c)]) \le \mathbb{E}[g(S \cdot G(c))] = U(S, c).$$

Now assume the statement (4.10) holds for $m-1$. Then for m,

$$\begin{aligned}
V_\zeta^{(n)}(S;c,m) &= \inf_{\Delta\in\mathbb{R}} \sup_{|r|\le\min\{\zeta,\sqrt{c}\}} -\Delta r + V_\zeta^{(n)}(S+Sr; c-r^2, m-1) \\
&\le \inf_{\Delta\in\mathbb{R}} \sup_{|r|\le\min\{\zeta,\sqrt{c}\}} -\Delta r + U(S+Sr, c-r^2) + \alpha^{(n)}(S+Sr, c-r^2, m-1) \\
&\le \sup_{|r|\le\min\{\zeta,\sqrt{c}\}} -rSU_S(S,c) + U(S+Sr, c-r^2) + \alpha^{(n)}(S+Sr, c-r^2, m-1) \\
&= \sup_{|r|\le\min\{\zeta,\sqrt{c}\}} U(S,c) + \epsilon_r(S,c) + \alpha^{(n)}(S+Sr, c-r^2, m-1) \\
&= U(S,c) + \alpha^{(n)}(S,c,m).
\end{aligned}$$

The first line is from the definition (4.5); the second line is using the inductive hypothesis that (4.10) holds for $m-1$; the third line is from substituting the choice $\Delta = SU_S(S,c)$; the fourth line is from the definition of ϵ_r; and the last line is from the definition of $\alpha^{(n)}(S,c,m)$. □

Proof of Lemma 4.8

For completeness, we provide a more detailed proof of Lemma 4.8.

Proof of Lemma 4.8: Unrolling the inductive definition (4.9), we can write

$$\alpha^{(n)}(S,c) = \sup_{\substack{r_1,\ldots,r_n \\ |r_m|\le\zeta,\ \sum_{m=1}^n r_m^2\le c}} f(r_1,\ldots,r_n),$$

where f is the function

$$f(r_1,\ldots,r_n) = \sum_{m=1}^{n} \epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c-\sum_{i=1}^{m-1} r_i^2\Big).$$

Let $(r_1,\ldots,r_n)$ be such that $|r_m|\le\zeta$ and $\sum_{m=1}^n r_m^2 \le c$. We will show that $f(r_1,\ldots,r_n) \le (18c + 8/\sqrt{2\pi})\, LK\, \zeta^{1/4}$.

Assume for now that $\zeta \le c^2$. Let $0 \le n_* \le n$ be the largest index such that

$$\sum_{m=1}^{n_*} r_m^2 \le c - \sqrt{\zeta}.$$

We split the analysis into two parts.

For $1 \le m \le \min\{n,\ n_*+1\}$: We want to apply the bound in Lemma 4.7, so let us verify that the conditions in Lemma 4.7 are satisfied. Clearly $|r_m| \le \zeta \le 1/16$. Moreover, since $c - \sum_{i=1}^{m-1} r_i^2 \ge c - \sum_{i=1}^{n_*} r_i^2 \ge \sqrt{\zeta}$ and $\zeta \le 1/16$, we also have

$$|r_m| \le \zeta \le \frac{\zeta^{1/4}}{8} \le \frac{\sqrt{c - \sum_{i=1}^{m-1} r_i^2}}{8}.$$

Therefore, by (4.11) from Lemma 4.7,

$$\begin{aligned}
\epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\, c-\sum_{i=1}^{m-1}r_i^2\Big) &\le 16LK\left(\max\left\{\Big(c-\sum_{i=1}^{m-1}r_i^2\Big)^{-3/2},\,\Big(c-\sum_{i=1}^{m-1}r_i^2\Big)^{-1/2}\right\}|r_m|^3\right.\\
&\qquad\left.+\max\left\{\Big(c-\sum_{i=1}^{m-1}r_i^2\Big)^{-2},\,\Big(c-\sum_{i=1}^{m-1}r_i^2\Big)^{-1/2}\right\}r_m^4\right)\\
&\le 16LK\,\left(\max\{\zeta^{-3/4},\,\zeta^{-1/4}\}\,|r_m|^3+\max\{\zeta^{-1},\,\zeta^{-1/4}\}\,r_m^4\right)\\
&= 16LK\,\left(\zeta^{-3/4}\,|r_m|^3+\zeta^{-1}\,r_m^4\right) && \text{(since } \zeta<1\text{)}\\
&\le 16LK\,\left(\zeta^{1/4}\,r_m^2+\zeta\,r_m^2\right) && \text{(since } |r_m|\le\zeta\text{)}\\
&\le 16LK\,\left(\zeta^{1/4}\,r_m^2+\zeta^{1/4}\,\frac{1}{16^{3/4}}\,r_m^2\right) && \text{(since } \zeta\le 1/16\text{)}\\
&= 18LK\,\zeta^{1/4}\,r_m^2.
\end{aligned}$$

Summing over $1\le m\le \min\{n,\,n_*+1\}$ gives us

$$\sum_{m=1}^{\min\{n,\,n_*+1\}}\epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\,c-\sum_{i=1}^{m-1}r_i^2\Big)\le 18LK\,\zeta^{1/4}\sum_{m=1}^{\min\{n,\,n_*+1\}}r_m^2\le 18LK\,\zeta^{1/4}\,c. \quad (4.16)$$

For $n_*+2\le m\le n$, if $n_*\le n-2$: Without loss of generality we may assume $r_n\neq 0$, for if $r_n=0$, then the term depending on r_n does not affect $f(r_1,\ldots,r_n)$ since

$$\epsilon_{r_n}\Big(S\prod_{i=1}^{n-1}(1+r_i),\,c-\sum_{i=1}^{n-1}r_i^2\Big)=0,$$

so we can remove r_n and only consider $n_*+2\le m\le n-1$. From the definition of n_* we see that $\sum_{m=1}^{n_*+1}r_m^2>c-\sqrt{\zeta}$, and since $\sum_{m=1}^{n}r_m^2\le c$, this implies

$$\sum_{m=n_*+2}^{n}r_m^2\le c-\sum_{m=1}^{n_*+1}r_m^2<c-(c-\sqrt{\zeta})=\sqrt{\zeta}. \quad (4.17)$$

Note also that for each $n_*+2\le m\le n$,

$$0<r_n^2\le\sum_{i=m}^{n}r_i^2\le c-\sum_{i=1}^{m-1}r_i^2\le c-\sum_{i=1}^{n_*+1}r_i^2\le\sqrt{\zeta}\le\frac{1}{4},$$

so by (4.12) from Lemma 4.7,

$$\epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\,c-\sum_{i=1}^{m-1}r_i^2\Big)\le\frac{4LK}{\sqrt{2\pi}}\cdot\frac{r_m^2}{\sqrt{c-\sum_{i=1}^{m-1}r_i^2}}\le\frac{4LK}{\sqrt{2\pi}}\cdot\frac{r_m^2}{\sqrt{\sum_{i=m}^{n}r_i^2}}.$$

Therefore, by applying Lemma 4.9 below to $x_i = r_{n_*+1+i}^2$, we see that

$$\begin{aligned}\sum_{m=n_*+2}^{n} \epsilon_{r_m}\Big(S\prod_{i=1}^{m-1}(1+r_i),\ c-\sum_{i=1}^{m-1} r_i^2\Big) &\le \frac{4LK}{\sqrt{2\pi}}\sum_{m=n_*+2}^{n}\frac{r_m^2}{\sqrt{\sum_{i=m}^{n} r_i^2}} \\ &\le \frac{8LK}{\sqrt{2\pi}}\left(\sum_{m=n_*+2}^{n} r_m^2\right)^{1/2} \le \frac{8LK}{\sqrt{2\pi}}\,\zeta^{1/4},\end{aligned} \tag{4.18}$$

where the last inequality follows from (4.17). Combining (4.16) and (4.18) gives us the desired conclusion.

Now if $\zeta > c^2$, then the argument in the second case above (for $n_* + 2 \le m \le n$) still holds with n_* set to be -1, so we still get the same conclusion. □

It now remains to prove the following result, which we use at the end of the proof of Lemma 4.8.

Lemma 4.9. *For $x_1, \dots, x_k \ge 0$ with $x_k > 0$, we have*

$$\sum_{i=1}^{k}\frac{x_i}{\sqrt{x_i + x_{i+1} + \cdots + x_k}} \le 2\left(\sum_{i=1}^{k} x_i\right)^{1/2}.$$

Proof. Let $\mathcal{L}_k$ denote the objective function that we wish to bound,

$$\mathcal{L}_k(x_1, \dots, x_k) = \sum_{i=1}^{k}\frac{x_i}{\sqrt{x_i + x_{i+1} + \cdots + x_k}},$$

and note that for any $t > 0$,

$$\mathcal{L}_k(tx_1, \dots, tx_k) = \sqrt{t}\,\mathcal{L}_k(x_1, \dots, x_k), \tag{4.19}$$

For each $k \in \mathbb{N}$, let Δ_k denote the unit simplex in $\mathbb{R}^k$ with $x_k > 0$,

$$\Delta_k = \left\{(x_1, \dots, x_k) :\ x_1, \dots, x_{k-1} \ge 0,\ x_k > 0,\ \sum_{i=1}^{k} x_i = 1\right\},$$

and let η_k denote the supremum of the function $\mathcal{L}_k$ over $x \in \Delta_k$. Given $x = (x_1, \dots, x_k) \in \Delta_k$, define $y = (y_1, \dots, y_{k-1})$ by $y_i = x_{i+1}/(1 - x_1)$, so $y \in \Delta_{k-1}$. Then we can write

$$\begin{aligned}\mathcal{L}_k(x_1, \dots, x_k) &= \frac{x_1}{\sqrt{x_1 + \cdots + x_k}} + \mathcal{L}_{k-1}(x_2, \dots, x_k) \\ &= x_1 + \sqrt{1 - x_1}\,\mathcal{L}_{k-1}(y_1, \dots, y_{k-1}) \\ &\le x_1 + \sqrt{1 - x_1}\,\eta_{k-1},\end{aligned}$$

where the second equality is from (4.19) and the last inequality is from the definition of η_{k-1}. The function $x_1 \mapsto x_1 + \sqrt{1-x_1}\,\eta_{k-1}$ is concave and maximized at $x_1^* = 1 - \eta_{k-1}^2/4$, giving us

$$\mathcal{L}_k(x_1,\dots,x_k) \le x_1^* + \sqrt{1-x_1^*}\,\eta_{k-1} = 1 - \frac{\eta_{k-1}^2}{4} + \sqrt{\frac{\eta_{k-1}^2}{4}}\,\eta_{k-1} = 1 + \frac{\eta_{k-1}^2}{4}.$$

Taking the supremum over $x \in \Delta_k$ gives us the recursion

$$\eta_k \le 1 + \frac{\eta_{k-1}^2}{4},$$

which, along with the base case $\eta_1 = 1$, easily implies $\eta_k \le 2$ for all $k \in \mathbb{N}$. Now given $x_1,\dots,x_k \ge 0$ with $x_k > 0$, let $x' = (tx_1,\dots,tx_k)$ with $t = 1/(x_1+\cdots+x_k)$, so $x' \in \Delta_k$. Then using (4.19) and the bound $\eta_k \le 2$, we get

$$\mathcal{L}_k(x_1,\dots,x_k) = \frac{1}{\sqrt{t}}\,\mathcal{L}_k(tx_1,\dots,tx_k) \le \eta_k \left(\sum_{i=1}^k x_i\right)^{1/2} \le 2\left(\sum_{i=1}^k x_i\right)^{1/2},$$

as desired. □

4.8 Proof of Lemma 4.7

In this section we provide a proof of Lemma 4.7. Throughout the rest of this chapter, we use the following notation for the higher-order partial derivatives of U,

$$U_{S^a c^b}(S,c) = \frac{\partial^{a+b} U(S,c)}{\partial S^a \partial c^b}, \quad a,b \in \mathbb{N}_0.$$

We will use the following bounds on U_{S^2}, U_{S^3}, and U_{S^4}, which we prove in Section 4.9. These bounds are where we use the crucial assumptions that the payoff function g is convex, L-Lipschitz, and K-linear.

Lemma 4.10. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be a convex, L-Lipschitz, K-linear function. Then for all $S, c > 0$,*

$$|U_{S^2}(S,c)| \le \frac{2LK}{\sqrt{2\pi}} \cdot \frac{1}{S^2\sqrt{c}} \tag{4.20}$$

$$|U_{S^3}(S,c)| \le 7LK \cdot \frac{\max\{c^{-3/2},\, c^{-1/2}\}}{S^3}, \tag{4.21}$$

$$|U_{S^4}(S,c)| \le 28LK \cdot \frac{\max\{c^{-2},\, c^{-1/2}\}}{S^4}. \tag{4.22}$$

We will also use the following property of the function U.

Lemma 4.11. *The function $U(S,c)$ is convex in S and non-decreasing in c.*

Proof. For each fixed $c \geq 0$ and for each realization of the random variable $G(c) > 0$, the function $S \mapsto g(S \cdot G(c))$ is convex. Therefore, $U(S, c)$ is convex in S, being a nonnegative linear combination of convex functions. In particular, this implies $U_{S^2}(S, c) \geq 0$. So by the Black-Scholes equation (4.6), we also have $U_c(S, c) = \frac{1}{2}S^2 U_{S^2}(S, c) \geq 0$. □

We are now ready to prove Lemma 4.7. For clarity, we divide the proof into two parts: we first prove the bound (4.12), then prove the bound (4.11).

Proof of (4.12) **in Lemma 4.7:** Recall that $U(S, c)$ is non-decreasing in c by Lemma 4.11. Then by the Taylor remainder theorem, we can write

$$\begin{aligned}\epsilon_r(S, c) &= U(S + Sr, c - r^2) - U(S, c) - rSU_S(S, c) \\ &\leq U(S + Sr, c) - U(S, c) - rSU_S(S, c) \\ &= \frac{1}{2}r^2 S^2 U_{S^2}(S + S\xi, c)\end{aligned}$$

where ξ is some value between 0 and r. Since $|\xi| \leq |r| \leq \sqrt{c} \leq 1/2$, we have $(1 + \xi)^2 \geq 1/4$. Moreover, from (4.20) in Lemma 4.10, we have

$$\left|(1 + \xi)^2 S^2 U_{S^2}(S + S\xi, c)\right| \leq \frac{2LK}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{c}}.$$

Combining the bounds above gives us

$$\epsilon_r(S, c) \leq \frac{1}{2}\frac{r^2}{(1 + \xi)^2}\left|(1 + \xi)^2 S^2 U_{S^2}(S + S\xi, c)\right| \leq \frac{4LK}{\sqrt{2\pi}} \cdot \frac{r^2}{\sqrt{c}},$$

as desired. □

Proof of (4.11) **in Lemma 4.7:** Fix $S, c > 0$, and consider the function

$$f(r) = U(S + Sr, c - r^2), \quad |r| \leq \sqrt{c}.$$

By repeatedly applying the Black-Scholes differential equation (4.6), we can easily verify that $f(0) = U(S, c)$, $f'(0) = SU_S(S, c)$, and

$$\begin{aligned}f''(r) = p_2(r)\, r\, S^2 U_{S^2}(S + Sr, c - r^2) \;+\; p_3(r)\,(1 + r)^2 r\, S^3 U_{S^3}(S + Sr, c - r^2) \\ +\; (1 + r)^4 r^2\, S^4 U_{S^4}(S + Sr, c - r^2),\end{aligned} \tag{4.23}$$

where p_2, p_3 are the polynomials $p_2(r) = 2r^3 + 4r^2 - 3r - 6$ and $p_3(r) = 4r^2 + 4r - 2$.

Noting that we can write

$$\epsilon_r(S, c) = f(r) - f(0) - f'(0)r,$$

another application of Taylor's remainder theorem allows us to write

$$\epsilon_r(S, c) = \frac{1}{2}f''(\xi)r^2$$

for some ξ lying between 0 and r. It is easy to verify that we have

$$\left|\frac{p_2(\xi)}{(1+\xi)^2}\right| \le 7, \qquad \left|\frac{p_3(\xi)}{(1+\xi)}\right| \le 3 \qquad \text{for all } |\xi| \le |r| \le \frac{1}{16}.$$

Moreover, since $\xi^2 \le r^2 \le c/64$, we have $c - \xi^2 \ge \frac{63}{64}c$. Then from the bound (4.20) in Lemma 4.10, we have

$$\left|(1+\xi)^2 S^2 U_{S^2}(S+S\xi, c-\xi^2)\right| \le \frac{2LK}{\sqrt{2\pi}} \cdot \frac{1}{(c-\xi^2)^{1/2}} \le \frac{2LK}{\sqrt{2\pi}} \cdot \frac{1}{(\frac{63}{64}c)^{1/2}} \le LK\, c^{-1/2}.$$

We also get from the bound (4.21) in Lemma 4.10,

$$\begin{aligned}
\left|(1+\xi)^3 S^3 U_{S^3}(S+S\xi, c-\xi^2)\right| &\le 7LK\ \max\{(c-\xi^2)^{-3/2},\ (c-\xi^2)^{-1/2}\} \\
&\le 7LK\ \max\left\{\left(\frac{63}{64}c\right)^{-3/2},\ \left(\frac{63}{64}c\right)^{-1/2}\right\} \\
&\le 7LK\ \left(\frac{64}{63}\right)^{3/2} \max\{c^{-3/2},\ c^{-1/2}\} \\
&\le 8LK\ \max\{c^{-3/2},\ c^{-1/2}\}.
\end{aligned}$$

Similarly, the bound (4.22) in Lemma 4.10 gives us

$$\left|(1+\xi)^4 S^4 U_{S^4}(S+S\xi, c-\xi^2)\right| \le 29LK\ \max\{c^{-2},\ c^{-1/2}\}.$$

Applying the bounds above to (4.23) gives us

$$\begin{aligned}
|f''(\xi)| &\le \left|\frac{p_2(\xi)}{(1+\xi)^2}\right| \cdot |\xi| \cdot \left|(1+\xi)^2 S^2 U_{S^2}(S+S\xi, c-\xi^2)\right| \\
&\quad + \left|\frac{p_3(\xi)}{(1+\xi)}\right| \cdot |\xi| \cdot \left|(1+\xi)^3 S^3 U_{S^3}(S+S\xi, c-\xi^2)\right| \\
&\quad + \xi^2 \cdot \left|(1+\xi)^4 S^4 U_{S^4}(S+S\xi, c-\xi^2)\right| \\
&\le 7LK\,|r|\, c^{-1/2} + 24LK\,|r|\ \max\{c^{-3/2},\ c^{-1/2}\} + 29LK\, r^2\ \max\{c^{-2},\ c^{-1/2}\} \\
&\le 31LK\,|r|\ \max\{c^{-3/2},\ c^{-1/2}\} + 29LK\, r^2\ \max\{c^{-2},\ c^{-1/2}\}.
\end{aligned}$$

Therefore, we obtain

$$|\epsilon_r(S,c)| = \frac{1}{2}\,|f''(\xi)| \cdot r^2 \le 16LK\ \left(|r|^3\ \max\{c^{-3/2},\ c^{-1/2}\} + r^4\ \max\{c^{-2},\ c^{-1/2}\}\right),$$

as desired. □

4.9 Proof of Lemma 4.10

In this section we prove the bounds on the higher-order derivatives $U_{S^a}(S,c)$, $a \geq 0$. Proving the bounds in Lemma 4.10 is more difficult than the analysis that we have done so far, and uses the full force of the assumptions that the payoff function g is convex, L-Lipschitz, and K-linear.

The outline of the proof is as follows. By writing $U(S,c)$ as a convolution, we can write its derivatives $U_{S^a}(S,c)$ as an expectation of $g(S \cdot G(c))$ modulated by certain polynomials. The K-linearity of g allows us to approximate g by the European-option payoff function g_{EC} that we encountered in Section 4.1, so we first prove Lemma 4.10 for the specific case when the payoff function is g_{EC}. We extend the bound on $U_{S^2}(S,c)$ to the general case by dominating the function inside the expectation by another carefully constructed function. Finally, we use the approximation of g by g_{EC} to prove the bounds on the higher-order derivatives U_{S^3} and U_{S^4}. In particular, Lemma 4.17 proves the bound (4.20), and Lemma 4.20 proves the bounds (4.21) and (4.22).

Throughout the rest of this section, $Z \sim \mathcal{N}(0,1)$ denotes a standard Gaussian random variable, and Φ and ϕ denote the cumulative distribution function and the probability density function, respectively, of the standard Gaussian distribution. The symbol $*$ denotes the convolution operator on $\mathbb{R}$. We also use the fact that $G(c) \stackrel{d}{=} \exp(-\frac{1}{2}c + \sqrt{c}Z)$. Recall that the convexity of g implies differentiability almost everywhere, so we can work with its derivative g', which is necessarily increasing (since g is convex) and satisfies $|g'(x)| \leq L$ (since g is L-Lipschitz).

Finally, in the proofs below we use the following easy property, which we state without proof.

Lemma 4.12. *For $f\colon \mathbb{R} \to \mathbb{R}$, $Z \sim \mathcal{N}(0,1)$, and $c \geq 0$, we have*

$$\mathbb{E}[f(Z)\, \exp(\sqrt{c}Z)] = \exp\left(\frac{c}{2}\right)\, \mathbb{E}[f(Z+\sqrt{c})],$$

provided all the expectations above exist.

Formulae for the Derivatives

In this section we show that the partial derivative $U_{S^a}(S,c)$ can be expressed as an expectation of a polynomial modulated by the payoff function g. We define the family of polynomials $p^{[a]}(x,y)$, $a \geq 0$, as follows:

$$\begin{aligned} p^{[0]}(x,y) &= 1 \\ p^{[a+1]}(x,y) &= (x - ay)\, p^{[a]}(x,y) - p_x^{[a]}(x,y) \quad \text{for } a \geq 1, \end{aligned} \tag{4.24}$$

where $p_x^{[a]}(x,y) = \partial p^{[a]}(x,y)/\partial x$.

The following is the main result in this section; note that we only assume that g is Lipschitz.

Lemma 4.13. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be an L-Lipschitz function. For $a \geq 0$ and $S, c > 0$,*

$$U_{S^a}(S,c) = \frac{1}{S^a c^{a/2}}\,\mathbb{E}\left[p^{[a]}(Z,\sqrt{c})\cdot g\left(S\cdot\exp\left(-\frac{c}{2}+\sqrt{c}Z\right)\right)\right],$$

where $Z \sim \mathcal{N}(0,1)$.

In proving Lemma 4.13 we will need the following result, which allows us to differentiate the convolution.

Lemma 4.14. *Fix $c > 0$. Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be an L-Lipschitz function, and let $\tilde{g}(x) = g(\exp(x))$. Let $\omega\colon \mathbb{R} \to \mathbb{R}$ be given by*

$$\omega(x) = p(x)\,\phi\left(\frac{x}{\sqrt{c}}\right)$$

where $p(x)$ is a polynomial in x with coefficients involving c. Finally, let $f\colon \mathbb{R} \to \mathbb{R}$ be given by $f(r) = (\tilde{g} * \omega)(r)$. Then the derivative $f'(r) = df(r)/dr$ can be written as the derivative of the convolution, $f'(r) = (\tilde{g} * \omega')(r)$.

Proof. Fix $r \in \mathbb{R}$. For $h \neq 0$, consider the quantity $\rho_h = \frac{1}{h}(f(r+h) - f(r))$, and note that $f'(r) = \lim_{h\to 0}\rho_h$. Recalling the definition of f as a convolution and using the mean-value theorem, we can write ρ_h as

$$\rho_h = \int_{-\infty}^{\infty}\tilde{g}(x)\left(\frac{\omega(r-x+h)-\omega(r-x)}{h}\right)dx = \int_{-\infty}^{\infty} g(x)\,\omega'(r-x+\xi_h)\,dx,$$

for some ξ_h between 0 and h. Let

$$\rho_0 := \int_{-\infty}^{\infty}\tilde{g}(x)\,\omega'(r-x)\,dx = (\tilde{g}*\omega')(r).$$

Then by another application of the mean-value theorem, we can write

$$\begin{aligned}\Delta_h := \rho_h - \rho_0 &= \xi_h\int_{-\infty}^{\infty}\tilde{g}(x)\left(\frac{\omega'(r-x+\xi_h)-\omega'(r-x)}{\xi_h}\right)dx \\ &= \xi_h\int_{-\infty}^{\infty}\tilde{g}(x)\,\omega''(r-x+\xi_h^{(2)})\,dx,\end{aligned} \tag{4.25}$$

for some $\xi_h^{(2)}$ lying between 0 and ξ_h. One can easily verify that the second derivative of ω is given by

$$\omega''(x) = \frac{q(x)}{c^2}\,\phi\left(\frac{x}{\sqrt{c}}\right),$$

where $q(x)$ is the polynomial $q(x) = (x^2-c)p(x) - 2cxp'(x) + c^2p''(x)$. Since g is L-Lipschitz, for each $x \in \mathbb{R}$ we have

$$0 \leq \tilde{g}(x) = g(\exp(x)) \leq g(0) + |g(\exp(x)) - g(0)| \leq g(0) + L\exp(x)$$

This gives us the estimate

$$\begin{aligned}
&\left|\int_{-\infty}^{\infty} \tilde{g}(x)\,\omega''(r-x+\xi_h^{(2)})\,dx\right| \\
&\qquad \le \frac{1}{c^2}\int_{-\infty}^{\infty} \big(g(0)+L\exp(x)\big)\cdot|q(r-x+\xi_h^{(2)})|\cdot\phi\left(\frac{r-x+\xi_h^{(2)}}{\sqrt{c}}\right)\,dx \\
&\qquad = \frac{1}{c^{3/2}}\int_{-\infty}^{\infty} \big(g(0)+L\exp(r+\xi_h^{(2)}-\sqrt{c}y)\big)\cdot|q(\sqrt{c}y)|\cdot\phi(y)\,dy<\infty,
\end{aligned}$$

where in the computation above we have used the substitution $y=(r-x+\xi_h^{(2)})/\sqrt{c}$. The last expression above shows that the integral is finite, since we are integrating exponential and polynomial functions against the Gaussian density. Plugging this bound to (4.25) and recalling that $|\xi_h|\le|h|$, we obtain

$$|\Delta_h| \le |h|\cdot\left|\int_{-\infty}^{\infty}\tilde{g}(x)\,\omega''(r-x+\xi_h^{(2)})\,dx\right| \to 0 \quad \text{as} \quad h\to 0.$$

Since $\Delta_h=\rho_h-\rho_0$, this implies our desired conclusion,

$$f'(r)=\lim_{h\to 0}\rho_h=\rho_0=(\tilde{g}*\omega')(r).$$

□

We are now ready to prove Lemma 4.13.

Proof of Lemma 4.13: We proceed by induction on a. The base case $a=0$ follows from the definition of U. Assume the statement holds for some $a\ge 0$; we prove it also holds for $a+1$. Our strategy is to express U_{S^a} as a convolution, use Lemma 4.14 to differentiate the convolution, and write the result back as an expectation.

Fix $S,c>0$ for the rest of this proof. Let $\tilde{g}(x)=g(\exp(x))$ and $\phi_c(x)=\phi(x/\sqrt{c})$. From the inductive hypothesis and the fact that $-Z\overset{d}{=}Z$, we have

$$\begin{aligned}
U_{S^a}(S,c) &= \frac{1}{S^a c^{a/2}}\,\mathbb{E}\left[p^{[a]}(-Z,\sqrt{c})\cdot\tilde{g}\left(\log S-\frac{c}{2}-\sqrt{c}Z\right)\right] \\
&= \frac{1}{S^a c^{a/2}}\int_{-\infty}^{\infty} p^{[a]}(-x,\sqrt{c})\cdot\tilde{g}\left(\log S-\frac{c}{2}-\sqrt{c}x\right)\cdot\phi(x)\,dx \\
&= \frac{1}{S^a c^{(a+1)/2}}\int_{-\infty}^{\infty} p^{[a]}\left(-\frac{y}{\sqrt{c}},\sqrt{c}\right)\cdot\tilde{g}\left(\log S-\frac{c}{2}-y\right)\cdot\phi_c(y)\,dy \\
&= \frac{1}{S^a c^{(a+1)/2}}\int_{-\infty}^{\infty} \tilde{g}\left(\log S-\frac{c}{2}-y\right)\cdot\omega(y)\,dy \\
&= \frac{1}{S^a c^{(a+1)/2}}\,(\tilde{g}*\omega)\left(\log S-\frac{c}{2}\right),
\end{aligned}$$

where in the computation above we have used the substitution $y = \sqrt{c}x$, and we have defined the function

$$\omega(y) = p^{[a]}\left(-\frac{y}{\sqrt{c}}, \sqrt{c}\right) \cdot \phi\left(\frac{y}{\sqrt{c}}\right).$$

In particular, ω has derivative

$$\omega'(y) = -\frac{1}{\sqrt{c}}\left(p_x^{[a]}\left(-\frac{y}{\sqrt{c}}, \sqrt{c}\right) + \frac{y}{\sqrt{c}} p^{[a]}\left(-\frac{y}{\sqrt{c}}, \sqrt{c}\right)\right) \phi\left(\frac{y}{\sqrt{c}}\right).$$

Differentiating U_{S^a} with respect to S and using the result of Lemma 4.14 give us

$$\begin{aligned}
U_{S^{a+1}}(S,c) &= -\frac{a}{S^{a+1}c^{(a+1)/2}} \left(\tilde{g} * \omega\right)\left(\log S - \frac{c}{2}\right) + \frac{1}{S^{a+1}c^{(a+1)/2}} \left(\tilde{g} * \omega'\right)\left(\log S - \frac{c}{2}\right) \\
&= \frac{1}{S^{a+1}c^{(a+1)/2}} \int_{-\infty}^{\infty} \tilde{g}\left(\log S - \frac{c}{2} - y\right) \left(\omega'(y) - a\omega(y)\right) dy \\
&= \frac{1}{S^{a+1}c^{a/2}} \int_{-\infty}^{\infty} \tilde{g}\left(\log S - \frac{c}{2} - \sqrt{c}x\right) \left(\omega'(\sqrt{c}x) - a\omega(\sqrt{c}x)\right) dx \\
&= \frac{1}{S^{a+1}c^{a/2}} \int_{-\infty}^{\infty} \tilde{g}\left(\log S - \frac{c}{2} - \sqrt{c}x\right) \frac{\left((-x - a\sqrt{c})p^{[a]}(-x,\sqrt{c}) - p_x^{[a]}(-x,\sqrt{c})\right)}{\sqrt{c}} \phi(x)\, dx \\
&= \frac{1}{S^{a+1}c^{(a+1)/2}} \int_{-\infty}^{\infty} \tilde{g}\left(\log S - \frac{c}{2} - \sqrt{c}x\right) p^{[a+1]}(-x, \sqrt{c})\, \phi(x)\, dx \\
&= \frac{1}{S^{a+1}c^{(a+1)/2}} \mathbb{E}\left[p^{[a+1]}(-Z, \sqrt{c}) \cdot \tilde{g}\left(\log S - \frac{c}{2} - \sqrt{c}Z\right)\right] \\
&= \frac{1}{S^{a+1}c^{(a+1)/2}} \mathbb{E}\left[p^{[a+1]}(Z, \sqrt{c}) \cdot g\left(S \cdot \left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right],
\end{aligned}$$

as desired. In the computation above we have again used the substitution $x = y/\sqrt{c}$ and the fact that $-Z \stackrel{d}{=} Z$. This completes the induction step and the proof of the lemma. □

As an example, the first few polynomials $p^{[a]}(x, y)$ are

$$\begin{aligned}
p^{[0]}(x,y) &= 1 \\
p^{[1]}(x,y) &= x \\
p^{[2]}(x,y) &= x^2 - yx - 1 \\
p^{[3]}(x,y) &= x^3 - 3yx^2 + (2y^2 - 3)x + 3y,
\end{aligned}$$

giving us the formulae

$$\begin{aligned}
U(S,c) &= \mathbb{E}\left[g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \\
U_S(S,c) &= \frac{1}{S\sqrt{c}} \mathbb{E}\left[Z \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \\
U_{S^2}(S,c) &= \frac{1}{S^2 c} \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \\
U_{S^3}(S,c) &= \frac{1}{S^3 c^{3/2}} \mathbb{E}\left[\left(Z^3 - 3\sqrt{c}Z^2 + (2c - 3)Z + 3\sqrt{c}\right) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right].
\end{aligned}$$

We also have the following corollary.

Corollary 4.15. *For $a \geq 1$, we have*

$$\mathbb{E}[p^{[a]}(Z, \sqrt{c})] = 0.$$

For $a \geq 2$, we also have

$$\mathbb{E}[p^{[a]}(Z + \sqrt{c}, \sqrt{c})] = 0.$$

Proof. First assume $a \geq 1$, and take g to be the constant function $g(x) = 1$. In this case $U(S, c) = 1$ and $U_{S^a}(S, c) = 0$, so by the result of Lemma 4.13,

$$\mathbb{E}[p^{[a]}(Z, \sqrt{c})] = S^a c^{a/2}\, U_{S^a}(S, c) = 0.$$

Next, assume $a \geq 2$, and take g to be the linear function $g(x) = x$. In this case $U(S, c) = \mathbb{E}[S \cdot G(c)] = S$, so $U_{S^a}(S, c) = 0$. Then using the results of Lemma 4.12 and Lemma 4.13,

$$\mathbb{E}\left[p^{[a]}(Z + \sqrt{c}, \sqrt{c})\right] = \exp\left(-\frac{c}{2}\right) \mathbb{E}\left[p^{[a]}(Z, \sqrt{c})\, \exp(\sqrt{c}Z)\right] = S^{a-1} c^{a/2}\, U_{S^a}(S, c) = 0.$$

□

Calculations for the European-Option Payoff Function

In this section, we bound the derivatives $U_{S^a}(S, c)$ for the special case when g is the payoff function of the European call function, $g(x) = \max\{0, x - K\}$, where $K > 0$ is a constant. Note that the bounds on U_{S^3} and U_{S^4} are slightly stronger than the stated bounds (4.21) and (4.22), because in this case we are able to compute the derivatives exactly.

Lemma 4.16. *Let $g(x) = \max\{0, x - K\}$. Then for all $S, c > 0$,*

$$|U_{S^2}(S, c)| \leq \frac{K}{\sqrt{2\pi}} \cdot \frac{1}{S^2\sqrt{c}}$$
$$|U_{S^3}(S, c)| \leq \frac{K}{\sqrt{2\pi}} \cdot \frac{(2\sqrt{c} + 1)}{S^3 c}$$
$$|U_{S^4}(S, c)| \leq \frac{K}{\sqrt{2\pi}} \cdot \frac{(6c + 5\sqrt{c} + 2)}{S^4 c^{3/2}}$$

Proof. We first compute the Black-Scholes value $U(S, c)$. Define

$$\alpha \equiv \alpha(S, c) = -\frac{1}{\sqrt{c}} \log \frac{S}{K} + \frac{\sqrt{c}}{2},$$

and observe that $S \cdot \exp(-c/2 + \sqrt{c}Z) \geq K$ if and only if $Z \geq \alpha$. Then using the result of Lemma 4.12, we have

$$\begin{aligned} U(S,c) &= \mathbb{E}\left[g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \\ &= \mathbb{E}\left[\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right) - K\right) \cdot \mathbf{1}\{Z \geq \alpha\}\right] \\ &= S \cdot \exp\left(-\frac{c}{2}\right)\, \mathbb{E}\left[\exp(\sqrt{c}Z) \cdot \mathbf{1}\{Z \geq \alpha\}\right] - K\, \mathbb{P}(Z \geq \alpha) \\ &= S\, \mathbb{P}(Z \geq \alpha - \sqrt{c}) - K\, \mathbb{P}(Z \geq \alpha) \\ &= S\, \Phi(-\alpha + \sqrt{c}) - K\, \Phi(-\alpha). \end{aligned}$$

Differentiating the formula above with respect to c and applying the Black-Scholes differential equation (4.6), we get

$$U_{S^2}(S,c) = \frac{2}{S^2}\, U_c(S,c) = \frac{1}{S^2 c}\left[S\alpha\phi\left(-\alpha + \sqrt{c}\right) + K(-\alpha + \sqrt{c})\phi\left(\alpha\right)\right] = \frac{K}{S^2\sqrt{c}}\, \phi(\alpha),$$

where the last equality follows from the relation $S\phi(-\alpha + \sqrt{c}) = K\phi(\alpha)$. In particular, we have the bound $0 \leq U_{S^2}(S,c) \leq K/(S^2\sqrt{2\pi c})$. A direct calculation reveals that the higher order derivatives of U are given by

$$U_{S^3}(S,c) = \frac{K}{S^3 c}\left(\alpha - 2\sqrt{c}\right)\, \phi(\alpha)$$

and

$$U_{S^4}(S,c) = \frac{K}{S^4 c^{3/2}}\left(\alpha^2 - 5\sqrt{c}\alpha + 6c - 1\right)\, \phi(\alpha).$$

It is not difficult to see that we have $|\alpha \exp(-\alpha^2/2)| \leq 1$ and $|\alpha^2 \exp(-\alpha^2/2)| \leq 1$. Applying these bounds to the formulae above gives us the desired conclusion. □

Bounding the Second Derivative $U_{S^2}(S,c)$

We now bound the second-order derivative $U_{S^2}(S,c)$ in the general case.

Lemma 4.17. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be a convex, L-Lipschitz, K-linear function. Then for all $S, c > 0$,*

$$0 \leq U_{S^2}(S,c) \leq \frac{2LK}{\sqrt{2\pi}} \cdot \frac{1}{S^2\sqrt{c}}.$$

Proof. Recall that $U(S,c)$ is convex in S (Lemma 4.11), so $U_{S^2}(S,c) \geq 0$. If g is a linear function, say $g(x) = \gamma x$ for some $0 \leq \gamma \leq L$, then $U(S,c) = \mathbb{E}[\gamma S \cdot G(c)] = \gamma S$. In this case $U_{S^2}(S,c) = 0$, and we are done.

Now assume g is not a linear function. Since g is non-negative, L-Lipschitz, and K-linear, we can find $0 \leq \gamma \leq L$ such that $g'(x) = \gamma$ for $x \geq K$. Moreover, since g is convex and not a linear function, we also have that $\gamma > g'(0)$. Define the function $\tilde{g}\colon \mathbb{R}_0 \to \mathbb{R}_0$ by

$$\tilde{g}(x) = \frac{g(x) - g(0) - xg'(0)}{\gamma - g'(0)}, \tag{4.26}$$

and note that $\tilde{g}$ is an increasing, 1-Lipschitz convex function with $\tilde{g}(0) = \tilde{g}'(0) = 0$, $0 \leq \tilde{g}'(x) \leq 1$, and $\tilde{g}'(x) = 1$ for $x \geq K$.

Consider the quantity $V(S, c) = \mathbb{E}[\tilde{g}(S \cdot G(c))]$, and note that we can write

$$V(S,c) = \frac{\mathbb{E}\left[g(S \cdot G(c)) - g(0) - g'(0) \cdot S \cdot G(c)\right]}{\gamma - g'(0)} = \frac{U(S,c) - g(0) - g'(0) \cdot S}{\gamma - g'(0)}.$$

Taking second derivative with respect to S on both sides and using the fact that $0 \leq \gamma - g'(0) \leq 2L$, we obtain

$$0 \leq U_{S^2}(S,c) = (\gamma - g'(0)) \cdot V_{S^2}(S,c) \leq 2L \cdot V_{S^2}(S,c).$$

We already know that $V_{S^2}(S, c) \geq 0$ since $\tilde{g}$ is convex, so we only need to show that

$$V_{S^2}(S,c) \leq \frac{K}{\sqrt{2\pi}} \cdot \frac{1}{S^2\sqrt{c}}.$$

For $0 < S \leq K$, using the formula from Lemma 4.13 and the result of Lemma 4.18 below, we obtain

$$V_{S^2}(S,c) = \frac{1}{S^2 c}\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot \tilde{g}\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \leq \frac{1}{S^2 c} \cdot \frac{S\sqrt{c}}{\sqrt{2\pi}} \leq \frac{K}{\sqrt{2\pi}} \cdot \frac{1}{S^2\sqrt{c}},$$

and for $S \geq K$, we use the result of Lemma 4.19 to obtain

$$V_{S^2}(S,c) = \frac{1}{S^2 c}\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot \tilde{g}\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \leq \frac{1}{S^2 c} \cdot \frac{K\sqrt{c}}{\sqrt{2\pi}} = \frac{K}{\sqrt{2\pi}} \cdot \frac{1}{S^2\sqrt{c}}.$$

This completes the proof of the lemma. □

It remains to prove the following two results, which we use in the proof of Lemma 4.17 above with $\tilde{g}$ in place of g. Note that the first result below does not use the assumption that g is eventually linear.

Lemma 4.18. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be an increasing, nonnegative, convex, 1-Lipschitz function. Then for all $S, c > 0$,*

$$\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \leq \frac{S\sqrt{c}}{\sqrt{2\pi}}.$$

Proof. Fix $S, c > 0$, and define the following quantities:

$$\begin{aligned}
t_1 &= \frac{\sqrt{c} - \sqrt{c+4}}{2} \\
t_2 &= \frac{\sqrt{c} + \sqrt{c+4}}{2} \\
\lambda_1 &= S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_1\right) \\
\lambda_2 &= S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_2\right) \\
g_1 &= g(\lambda_1) \\
g_2 &= g(\lambda_2) \\
t_* &= \frac{1}{\sqrt{c}} \log\left(\exp(\sqrt{c} t_2) - \frac{1}{S} \cdot \exp\left(\frac{c}{2}\right) \cdot (g_2 - g_1)\right).
\end{aligned}$$

Furthermore, define the function $h\colon \mathbb{R} \to \mathbb{R}_0$ by

$$h(x) = g_1 + \left(g_2 - g_1 - \lambda_2 + S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, x\right)\right) \cdot \mathbf{1}\{x \geq t_*\}.$$

We will show that

$$\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \leq \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot h(Z)\right], \tag{4.27}$$

and furthermore, we can evaluate the latter expectation explicitly:

$$\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot h(Z)\right] = S\sqrt{c}\, \phi(t_* - \sqrt{c}) \leq \frac{S\sqrt{c}}{\sqrt{2\pi}}.$$

We begin by noting that t_1 and t_2 are the two roots of the polynomial $x^2 - \sqrt{c}x - 1$. Since g is increasing and 1-Lipschitz,

$$g_2 - g_1 = g(\lambda_2) - g(\lambda_1) \leq \lambda_2 - \lambda_1 = S \cdot \exp\left(-\frac{c}{2}\right) \left(\exp(\sqrt{c}\, t_2) - \exp(\sqrt{c}\, t_1)\right).$$

Therefore, from the definition of t_*, we see that

$$\exp(\sqrt{c}\, t_2) - \exp(\sqrt{c}\, t_*) = \frac{1}{S} \cdot \exp\left(\frac{c}{2}\right) \cdot (g_2 - g_1) \leq \exp(\sqrt{c}\, t_2) - \exp(\sqrt{c}\, t_1),$$

so $t_1 \leq t_* \leq t_2$. Furthermore, by construction,

$$S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_*\right) = S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_2\right) - (g_2 - g_1) = \lambda_2 - g_2 + g_1,$$

so $h(t_*) = g_1$. This means h is a continuous convex function of x (although we will not actually use this property). We will now show that pointwise,

$$\phi(x) \cdot (x^2 - \sqrt{c}x - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \leq \phi(x) \cdot (x^2 - \sqrt{c}x - 1) \cdot h(x). \tag{4.28}$$

We consider four cases:

- Suppose $x \le t_1$, so $x^2 - \sqrt{c}x - 1 \ge 0$. Since g is increasing and nonnegative,

$$0 \le g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \le g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_1\right)\right) = g_1 = h(x).$$

- Suppose $t_1 \le x \le t_*$, so $x^2 - \sqrt{c}x - 1 \le 0$. Since g is increasing,

$$g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \ge g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_1\right)\right) = g_1 = h(x) \ge 0.$$

- Suppose $t_* \le x \le t_2$, so $x^2 - \sqrt{c}x - 1 \le 0$. Since g is increasing and 1-Lipschitz,

$$\begin{aligned} &g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \\ &\quad \ge g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_2\right)\right) + S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right) - S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_2\right) \\ &\quad = g_2 - \lambda_2 + S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right) = h(x) \ge 0. \end{aligned}$$

- Suppose $x \ge t_2$, so $x^2 - \sqrt{c}x - 1 \ge 0$. Since g is increasing and 1-Lipschitz,

$$\begin{aligned} &g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \\ &\quad \le g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_2\right)\right) + S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right) - S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_2\right) \\ &\quad = g_2 - \lambda_2 + S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right) = h(x). \end{aligned}$$

Integrating (4.28) over $x \in \mathbb{R}$ gives us the desired inequality (4.27). Let us now evaluate the expectation on the right hand side of (4.27). A simple computation using the properties of $Z \sim \mathcal{N}(0,1)$ gives us

$$\begin{aligned} \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot h(Z)\right] &= g_1\, \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1)\right] \\ &\quad + (g_2 - g_1 - \lambda_2) \cdot \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot \mathbf{1}\{Z \ge t_*\}\right] \\ &\quad + S \cdot \exp\left(-\frac{c}{2}\right) \cdot \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1)\, \exp(\sqrt{c}Z) \cdot \mathbf{1}\{Z \ge t_*\}\right] \\ &= (g_2 - g_1 - \lambda_2) \cdot (t_* - \sqrt{c})\, \phi(t_*) + S t_*\, \phi(t_* - \sqrt{c}) \\ &= -S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\, t_*\right) \cdot (t_* - \sqrt{c})\, \phi(t_*) + S t_*\, \phi(t_* - \sqrt{c}) \\ &= -S(t_* - \sqrt{c})\, \phi(t_* - \sqrt{c}) + S t_*\, \phi(t_* - \sqrt{c}) \\ &= S\sqrt{c}\, \phi(t_* - \sqrt{c}), \end{aligned}$$

as desired. □

The following result is similar to Lemma 4.18, except that this result assumes the K-linearity of g and achieves a stronger result.

Lemma 4.19. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be an increasing, nonnegative, convex, 1-Lipschitz function with the property that $g'(x) = 1$ for $x \geq K$. Then for all $S \geq K$ and $c > 0$,*

$$\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \leq \frac{K\sqrt{c}}{\sqrt{2\pi}}.$$

Proof. This proof is similar in nature to the proof of Lemma 4.18, and we omit some of the details.

Case 1: Suppose $S \geq K\exp(\sqrt{c(c+4)}/2)$. Recall the European-option payoff function $g_{\mathrm{EC}}(x) = \max\{0, x - K\}$ from Section 4.1, and note that the K-linearity of g implies $g(x) = g(K) + g_{\mathrm{EC}}(x)$ for $x \geq K$. Using the fact that g is increasing and K-linear, we can show that for all $x \in \mathbb{R}$ we have

$$(x^2 - \sqrt{c}x - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \leq (x^2 - \sqrt{c}x - 1) \cdot \left\{g(K) + g_{\mathrm{EC}}\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right)\right\}.$$

Integrating both sides above with $Z \sim \mathcal{N}(0,1)$ in place of x and using the result of Lemma 4.16, we obtain

$$\begin{aligned}
&\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \\
&\qquad \leq \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot \left\{g(K) + g_{\mathrm{EC}}\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right\}\right] \\
&\qquad = \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g_{\mathrm{EC}}\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \\
&\qquad \leq \frac{K\sqrt{c}}{\sqrt{2\pi}}.
\end{aligned}$$

Case 2: Suppose $K \leq S \leq K\exp(\sqrt{c(c+4)}/2)$. Define the following quantities:

$$\begin{aligned}
t_0 &= \frac{\sqrt{c}}{2} - \frac{1}{\sqrt{c}} \log \frac{S}{K - g(K) + g_1} \\
\lambda_1 &= S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\left(\frac{\sqrt{c} - \sqrt{c+4}}{2}\right)\right)
\end{aligned}$$

and $g_1 = g(\lambda_1)$. Consider the function $h_2\colon \mathbb{R} \to \mathbb{R}_0$ given by

$$h_2(x) = g_1 + \left(g(K) - g_1 - K + S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}\,x\right)\right) \cdot \mathbf{1}\{x \geq t_0\}.$$

Using the fact that g is increasing, 1-Lipschitz, and K-linear, we can show that for all $x \in \mathbb{R}$,

$$(x^2 - \sqrt{c}x - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}x\right)\right) \leq (x^2 - \sqrt{c}x - 1) \cdot h_2(x).$$

Integrating both sides above with $Z \sim \mathcal{N}(0,1)$ in place of x, we get

$$\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot g\left(S \cdot \exp\left(-\frac{c}{2} + \sqrt{c}Z\right)\right)\right] \leq \mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1) \cdot h_2(Z)\right].$$

Following the same calculation as in the proof of Lemma 4.18, we can evaluate the latter expectation to be

$$\mathbb{E}\left[(Z^2 - \sqrt{c}Z - 1)\cdot h_2(Z)\right] = (K - g(K) + g_1)\,\sqrt{c}\,\phi(t_0) \le \frac{K\sqrt{c}}{\sqrt{2\pi}},$$

where the last inequality follows from the relation $0 \le g(K) - g_1 \le K - \lambda_1$, since g is increasing and 1-Lipschitz. □

Bounding the Higher-Order Derivatives

We now turn to bounding the higher-order derivatives $U_{S^3}(S,c)$ and $U_{S^4}(S,c)$. Our strategy is to approximate the eventually linear payoff function g by the European-option payoff g_{EC} and applying the bounds for g_{EC} developed in Lemma 4.16.

Lemma 4.20. *Let $g\colon \mathbb{R}_0 \to \mathbb{R}_0$ be a convex, L-Lipschitz, K-linear function. Then for all $S, c > 0$,*

$$|U_{S^3}(S,c)| \le 7LK \cdot \frac{\max\{c^{-3/2},\ c^{-1/2}\}}{S^3},$$
$$|U_{S^4}(S,c)| \le 28LK \cdot \frac{\max\{c^{-2},\ c^{-1/2}\}}{S^4}.$$

Proof. Since g is L-Lipschitz and K-linear, we can find $0 \le \gamma \le L$ such that $g(x) = g(K) + \gamma(x - K)$ for $x \ge K$. We decompose g into two parts,

$$g(x) = \gamma g_{\mathrm{EC}}(x) + g^*(x),$$

where $g_{\mathrm{EC}}(x) = \max\{0,\ x - K\}$ is the European-option payoff function, and $g^*\colon \mathbb{R}_0 \to \mathbb{R}_0$ is given by $g^*(x) = g(x)$ for $0 \le x \le K$, and $g^*(x) = g(K)$ otherwise.

Then the Black-Scholes value $U(S,c)$ also decomposes,

$$U(S,c) = \mathbb{E}[g(S\cdot G(c))] = \gamma\,\mathbb{E}[g_{\mathrm{EC}}(S\cdot G(c))] + \mathbb{E}[g^*(S\cdot G(c))] \equiv \gamma\,U^{\mathrm{EC}}(S,c) + U^*(S,c),$$

and similarly for the derivatives,

$$U_{S^a}(S,c) = \gamma\,U^{\mathrm{EC}}_{S^a}(S,c) + U^*_{S^a}(S,c), \qquad a \ge 0. \tag{4.29}$$

For the function g_{EC}, Lemma 4.16 tells us that for all $S, c > 0$,

$$\begin{aligned} |U^{\mathrm{EC}}_{S^3}(S,c)| &\le \frac{3K}{\sqrt{2\pi}} \cdot \frac{\max\{c^{1/2},\ 1\}}{S^3 c}, \\ |U^{\mathrm{EC}}_{S^4}(S,c)| &\le \frac{13K}{\sqrt{2\pi}} \cdot \frac{\max\{c,\ 1\}}{S^4 c^{3/2}}. \end{aligned} \tag{4.30}$$

Now for the second function g^*, we use Lemma 4.13 to write

$$U^*_{S^a}(S,c) = \frac{1}{S^a c^{a/2}}\,\mathbb{E}\left[p^{[a]}(Z,\sqrt{c})\cdot g^*\left(S\cdot\exp\left(-\frac{c}{2}+\sqrt{c}Z\right)\right)\right]. \tag{4.31}$$

Since $\mathbb{E}[p^{[a]}(Z,\sqrt{c})] = 0$ for $a \geq 1$ (Corollary 4.15), we may assume that $g(0) = 0$, so $g^*(0) = 0$ as well. Since g is L-Lipschitz, this implies

$$\sup_{x\in\mathbb{R}}\left|g^*(x)\right| = \max_{0\leq x\leq K}|g(x)| \leq \max_{0\leq x\leq K} Lx = LK.$$

Therefore, by applying triangle inequality and Cauchy-Schwarz inequality to (4.31), we get for $a \geq 1$,

$$|U^*_{S^a}(S,c)| \leq \frac{1}{S^a c^{a/2}}\,\mathbb{E}\left[\left|p^{[a]}(Z,\sqrt{c})\right|\cdot LK\right] \leq \frac{LK}{S^a c^{a/2}}\,\mathbb{E}\left[\left(p^{[a]}(Z,\sqrt{c})\right)^2\right]^{1/2}. \tag{4.32}$$

For $a = 3, 4$, we use the recursion (4.24) to compute the polynomials $p^{[a]}(Z,\sqrt{c})$, and we evaluate the expectation $\mathbb{E}[(p^{[a]}(Z,\sqrt{c}))^2]$. Plugging in this expectation to (4.32) with $a = 3$ gives us

$$|U^*_{S^3}(S,c)| \leq \frac{LK}{S^3c^{3/2}}\cdot\left(4c^2+18c+6\right)^{1/2} \leq \sqrt{28}\cdot LK\cdot\frac{\max\{c,\,1\}}{S^3c^{3/2}}. \tag{4.33}$$

Therefore, by combining the bound above with the first inequality in (4.30) and using the decomposition (4.29), we get the first part of our lemma,

$$|U_{S^3}(S,c)| \leq \frac{3}{\sqrt{2\pi}}\cdot LK\cdot\frac{\max\{c^{1/2},\,1\}}{S^3c} + \sqrt{28}\cdot LK\cdot\frac{\max\{c,\,1\}}{S^3c^{3/2}} \leq 7LK\cdot\frac{\max\{c,\,1\}}{S^3c^{3/2}}.$$

A similar computation with $a = 4$ yields the second part of the lemma,

$$|U_{S^4}(S,c)| \leq \frac{13}{\sqrt{2\pi}}\cdot LK\cdot\frac{\max\{c,\,1\}}{S^4c^{3/2}} + \sqrt{518}\cdot LK\cdot\frac{\max\{c^{3/2},\,1\}}{S^4c^2} \leq 28LK\cdot\frac{\max\{c^{3/2},\,1\}}{S^3c^2}.$$

□

Bibliography

[1] Jacob Abernethy, Rafael M. Frongillo, and Andre Wibisono. "Minimax option pricing meets Black-Scholes in the limit". In: *STOC*. Ed. by Howard J. Karloff and Toniann Pitassi. ACM, 2012, pp. 1029–1040. ISBN: 978-1-4503-1245-5.

[2] Jacob Abernethy, Peter L. Bartlett, Rafael M. Frongillo, and Andre Wibisono. "How to Hedge an Option Against an Adversary: Black-Scholes Pricing is Minimax Optimal". In: *Advances in Neural Information Processing Systems (NIPS) 26*. 2013.

[3] S. M. Aji and R. J. McEliece. "The generalized distributive law and free energy minimization". In: *Proceedings of the 39th Allerton Conference*. 2001.

[4] Zeyuan Allen-Zhu and Lorenzo Orecchia. "Linear coupling: An ultimate unification of gradient and mirror descent". In: *ArXiv preprint arXiv:1407.1537* (2014).

[5] Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. "Hessian Riemannian Gradient Flows in Convex Programming". In: *SIAM Journal on Control and Optimization* 43.2 (2004), pp. 477–501.

[6] Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. "On Lower and Upper Bounds for Smooth and Strongly Convex Optimization Problems". In: *ArXiv preprint arXiv:1503.06833* (2015).

[7] Hedy Attouch and Zaki Chbani. "Fast Inertial Dynamics and FISTA Algorithms in Convex Optimization: Perturbation Aspects". In: *ArXiv preprint arXiv:1507.01367* (2015).

[8] Hedy Attouch, Juan Peypouquet, and Patrick Redont. "On the Fast Convergence of an Inertial Gradient-like System with Vanishing Viscosity". In: *ArXiv preprint arXiv:1507.04782* (2015).

[9] Michel Baes. *Estimate sequence methods: Extensions and approximations*. Manuscript, available at `http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf`. Aug. 2009.

[10] F. Barahona. "On the computational complexity of Ising spin glass models". In: *Journal of Physics A: Mathematical and General* 15.10 (1982), p. 3241.

[11] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (Mar. 2009), pp. 183–202. ISSN: 1936-4954.

[12] H. A. Bethe. "Statistical Theory of Superlattices". In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 150.871 (1935), pp. 552–575.

[13] F. Black and M. Scholes. "The pricing of options and corporate liabilities". In: *The Journal of Political Economy* (1973), pp. 637–654.

[14] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. "A geometric alternative to Nesterov's accelerated gradient descent". In: *ArXiv preprint arXiv:1506.08187* (2015).

[15] Jorge Cortés. "Finite-time convergent gradient flows with applications to network consensus". In: *Automatica* 42.11 (2006), pp. 1993–2000.

[16] P. DeMarzo, I. Kremer, and Y. Mansour. "Online trading algorithms and robust option pricing". In: *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*. ACM. 2006, pp. 477–486.

[17] R. Durrett. *Probability: Theory and Examples (Fourth Edition)*. Cambridge University Press, 2010.

[18] Nicolas Flammarion and Francis R. Bach. "From Averaging to Acceleration, There is Only a Step-size". In: *Proceedings of the 28th Conference on Learning Theory (COLT)*. 2015.

[19] Y. Freund and R. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Computational learning theory*. Springer. 1995, pp. 23–37.

[20] Saeed Ghadimi and Guanghui Lan. "Accelerated gradient methods for nonconvex nonlinear and stochastic programming". English. In: *Mathematical Programming* 156.1 (2015), pp. 59–99. ISSN: 0025-5610.

[21] P. Hall. "On representatives of subsets". In: *Journal of the London Mathematical Society* 10 (1935), pp. 26–30.

[22] T. Heskes. "Convexity Arguments for Efficient Minimization of the Bethe and Kikuchi Free Energies." In: *Journal of Artificial Intelligence Research* 26 (2006), pp. 153–190.

[23] T. Heskes. "On the Uniqueness of Loopy Belief Propagation Fixed Points". In: *Neural Computation* 16.11 (2004), pp. 2379–2413.

[24] T. Heskes. "Stable fixed points of loopy belief propagation are minima of the Bethe free energy". In: *Advances in Neural Information Processing Systems 15*. 2002.

[25] Chonghai Hu, James T. Kwok, and Weike Pan. "Accelerated Gradient Methods for Stochastic Optimization and Online Learning". In: *Advances in Neural Information Processing Systems (NIPS) 22*. 2009.

[26] A. T. Ihler, J. W. Fischer III, and A. S. Willsky. "Loopy Belief Propagation: Convergence and Effects of Message Errors". In: *Journal of Machine Learning Research* 6 (Dec. 2005), pp. 905–936. ISSN: 1532-4435.

[27] Shuiwang Ji and Jieping Ye. "An Accelerated Gradient Method for Trace Norm Minimization". In: *Proceedings of the 26th International Conference on Machine Learning (ICML)*. Montreal, Quebec, Canada, 2009.

[28] Shuiwang Ji, Liang Sun, Rong Jin, and Jieping Ye. "Multi-label Multiple Kernel Learning". In: *Advances in Neural Information Processing Systems (NIPS) 21*. 2009.

[29] Vladimir Jojic, Stephen Gould, and Daphne Koller. "Accelerated dual decomposition for MAP inference". In: *Proceedings of the 27th International Conference on Machine Learning (ICML)*. 2010.

[30] Anatoli Juditsky. *Convex Optimization II: Algorithms*. Lecture notes, 2013.

[31] R. Kikuchi. "A Theory of Cooperative Phenomena". In: *Phys. Rev.* 81 (6 Mar. 1951), pp. 988–1003.

[32] B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms*. 4th. Springer, 2007.

[33] Walid Krichene, Alexandre Bayen, and Peter Bartlett. "Accelerated Mirror Descent in Continuous and Discrete Time". In: *Advances in Neural Information Processing Systems (NIPS) 29*. 2015.

[34] Guanghui Lan. "An optimal method for stochastic composite optimization". English. In: *Mathematical Programming* 133.1-2 (2012), pp. 365–397.

[35] Guanghui Lan, Zhaosong Lu, and Renato Monteiro. "Primal-dual First-order Methods with $O(1/\epsilon)$ Iteration-complexity for Cone Programming". In: *Mathematical Programming* 126.1 (2011), pp. 1–29. ISSN: 0025-5610.

[36] Laurent Lessard, Benjamin Recht, and Andrew Packard. "Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints". In: *SIAM Journal on Optimization* 26.1 (2016), pp. 57–95.

[37] Huan Li and Zhouchen Lin. "Accelerated Proximal Gradient Methods for Nonconvex Programming". In: *Advances in Neural Information Processing Systems (NIPS) 28*. 2015.

[38] N. Littlestone and M. K. Warmuth. "The weighted majority algorithm". In: *Information and Computation* 108.2 (1994), pp. 212–261. ISSN: 0890-5401.

[39] Po-Ling Loh and Andre Wibisono. "Convexity of Reweighted Kikuchi Approximation". In: *Advances in Neural Information Processing Systems (NIPS) 27*. 2014.

[40] R. J. McEliece and M. Yildirim. "Belief Propagation On Partially Ordered Sets". In: *Mathematical Systems Theory in Biology, Communications, Computation, and Finance*. 2002, pp. 275–300.

[41] T. Meltzer, A. Globerson, and Y. Weiss. "Convergent message passing algorithms: a unifying view". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09. 2009.

[42] J. M. Mooij and H. J. Kappen. "Sufficient Conditions for Convergence of the Sum-Product Algorithm". In: *IEEE Transactions on Information Theory* 53.12 (Dec. 2007), pp. 4422–4437.

[43] Indraneel Mukherjee, Kevin Canini, Rafael Frongillo, and Yoram Singer. "Parallel boosting with momentum". In: *Machine Learning and Knowledge Discovery in Databases*. Springer, 2013.

[44] Arkadi Nemirovskii and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

[45] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.

[46] Yurii Nesterov. "Accelerating the cubic regularization of Newton's method on convex problems". English. In: *Mathematical Programming* 112.1 (2008), pp. 159–181. ISSN: 0025-5610.

[47] Yurii Nesterov. *Gradient methods for minimizing composite objective function*. CORE Discussion Papers 2007076. Université Catholique de Louvain, 2007.

[48] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Boston: Kluwer, 2004.

[49] Yurii Nesterov. "Smooth Minimization of Non-smooth Functions". In: *Mathematical Programming* 103.1 (2005), pp. 127–152.

[50] Yurii Nesterov and Boris T. Polyak. "Cubic regularization of Newton's method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

[51] Brendan O'Donoghue and Emmanuel Candès. "Adaptive Restart for Accelerated Gradient Schemes". English. In: *Foundations of Computational Mathematics* 15.3 (2015), pp. 715–732.

[52] P. Pakzad and V. Anantharam. "Estimation and marginalization using Kikuchi approximation methods". In: *Neural Computation* 17 (2003), pp. 1836–1873.

[53] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[54] Garvesh Raskutti and Sayan Mukherjee. "The Information Geometry of Mirror Descent". In: *IEEE Transactions on Information Theory* 61.3 (2015), pp. 1451–1457.

[55] T. Roosta, M. J. Wainwright, and S. S. Sastry. "Convergence Analysis of Reweighted Sum-Product Algorithms." In: *IEEE Transactions on Signal Processing* 56.9 (2008), pp. 4293–4305.

[56] D. Roth. "On the hardness of approximate reasoning". In: *Artificial Intelligence* 82.1–2 (1996), pp. 273–302.

[57] N. Ruozzi. "The Bethe Partition Function of Log-supermodular Graphical Models". In: *Advances in Neural Information Processing Systems 25*. 2012.

[58] G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!* Vol. 373. Wiley-Interscience, 2001.

[59] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. MIT Press, 2012.

[60] J. M. Steele. *Stochastic Calculus and Financial Applications*. Vol. 45. Springer Verlag, 2001.

[61] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *Advances in Neural Information Processing Systems (NIPS) 27*. 2014.

[62] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky. "Loop Series and Bethe Variational Bounds in Attractive Graphical Models." In: *Advances in Neural Information Processing Systems 20*. 2007.

[63] S. C. Tatikonda and M. I. Jordan. "Loopy Belief Propagation and Gibbs Measures". In: *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*. UAI '02. 2002.

[64] Paul Tseng. "On accelerated proximal gradient methods for convex-concave optimization". In: *SIAM Journal on Optimization* (2008).

[65] Leslie G. Valiant. "A Theory of the Learnable". In: *Communications of the ACM* 27.11 (Nov. 1984), pp. 1134–1142.

[66] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[67] Cedric Villani. *Optimal Transport, Old and New*. Springer, 2008.

[68] P. O. Vontobel. "The Bethe Permanent of a Nonnegative Matrix". In: *IEEE Transactions on Information Theory* 59.3 (2013), pp. 1866–1901.

[69] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. "A New Class of Upper Bounds on the Log Partition Function". In: *IEEE Transactions on Information Theory* 51.7 (2005), pp. 2313–2335.

[70] M. J. Wainwright and M. I. Jordan. "Graphical Models, Exponential Families, and Variational Inference". In: *Foundations and Trends in Machine Learning* 1.1–2 (Jan. 2008), pp. 1–305. ISSN: 1935-8237.

[71] Y. Watanabe and K. Fukumizu. "Graph Zeta Function in the Bethe Free Energy and Loopy Belief Propagation". In: *Advances in Neural Information Processing Systems 22*. 2009.

[72] Y. Watanabe and K. Fukumizu. "Loopy belief propagation, Bethe free energy and graph zeta function". In: *arXiv preprint arXiv:1103.0605* (2011).

[73] Y. Weiss. "Correctness of Local Probability Propagation in Graphical Models with Loops". In: *Neural Computation* 12.1 (2000), pp. 1–41.

[74] T. Werner. "Primal View on Belief Propagation". In: *UAI 2010: Proceedings of the Conference of Uncertainty in Artificial Intelligence*. Corvallis, Oregon: AUAI Press, July 8–11, 2010, pp. 651–657.

[75] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. "A Variational Perspective on Accelerated Methods in Optimization". In: *ArXiv e-prints arXiv:1603.04245* (2016). arXiv: `1603.04245 [math.OC]`.

[76] J. S. Yedidia, W. T. Freeman, and Y. Weiss. "Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms". In: *IEEE Transactions on Information Theory* 51 (2005), pp. 2282–2312.

[77] J. S. Yedidia, W. T. Freeman, and Y. Weiss. "Generalized Belief Propagation". In: *Advances in Neural Information Processing Systems 13*. 2000.