Proximal Gradient Algorithms for Gaussian Variational Inference: Optimization in the Bures–Wasserstein Space

by

Michael Ziyang Diao

S.B., Electrical Engineering and Computer Science and Mathematics,
Massachusetts Institute of Technology (2023)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engingeering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Michael Ziyang Diao
Department of Electrical Engineering and Computer Science
May 12, 2023

Certified by: Ankur Moitra
Norbert Wiener Professor of Mathematics
Thesis Supervisor

Certified by: Sinho Chewi
Graduate Student, Mathematics
Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Proximal Gradient Algorithms for Gaussian Variational Inference:

Optimization in the Bures–Wasserstein Space

by

Michael Ziyang Diao

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engingeering and Computer Science

Abstract

Variational inference (VI) seeks to approximate a target distribution π by an element of a tractable family of distributions. Of key interest in statistics and machine learning is Gaussian VI, which approximates π by minimizing the Kullback–Leibler (KL) divergence to π over the space of Gaussians. In this work, we develop the (Stochastic) Forward-Backward Gaussian Variational Inference (FB–GVI) algorithm to solve Gaussian VI. Our approach exploits the composite structure of the KL divergence, which can be written as the sum of a smooth term (the potential) and a non-smooth term (the entropy) over the Bures–Wasserstein (BW) space of Gaussians endowed with the Wasserstein distance. For our proposed algorithm, we obtain state-of-the-art convergence guarantees when π is log-smooth and log-concave, as well as the first convergence guarantees to first-order stationary solutions when π is only log-smooth. Additionally, in the setting where the potential admits a representation as the average of many smooth component functionals, we develop and analyze a variance-reduced extension to (Stochastic) FB–GVI with improved complexity guarantees.

Thesis Supervisor: Ankur Moitra
Title: Norbert Wiener Professor of Mathematics

Thesis Supervisor: Sinho Chewi
Title: Graduate Student, Mathematics

Acknowledgments

My four short years at MIT have been a gift. The journey has been tumultuous yet thrilling, challenging yet freeing, and through it all I've learned and grown so much. So many people have left an indelible mark on me, and I'm the luckiest person in the world for having crossed paths with them. To all my mentors, teachers, friends, and family: you are what made this chapter of my life so profoundly special.

To my research mentors throughout the years: Sinho Chewi, Krishnakumar Balasubramanian, Adil Salim, Dylan Cable, Michael McAneny, Francisco Rivera, Dylan Grullon, and Pranav Chowdhary — thank you for taking a chance on me. Research is tough, and you believed in me even when I felt hopelessly stuck. The advice and wisdom you've imparted on me has been a compass, helping me navigate research — and life at large. Thank you especially to my advisor Sinho Chewi: you have taught me more than any other person or class at MIT. Often I would message you far past midnight with a nebulous, ill-formed random thought on my mind (sorry!). And yet somehow you would always respond instantly with clarity and deep insight. You are a gifted teacher and dedicated mentor. Without your guidance, I would be lost.

To Joe Blitzstein and Ankur Moitra: thank you for being such great professors. You both have this amazing skill for explaining things, crystallizing key insights in a way that makes them appear almost obvious in hindsight. Your classes were some of my absolute favorites at MIT (or Harvard!), and I'm so glad I had the chance to learn from the very best. Professor Moitra, thanks especially for agreeing to sign off on my thesis!

To Polina Golland, Lizhong Zheng, Greg Wornell, and Martin Wainwright: thank you for an amazing TA experience. Being a part of the 6.008 and 6.437 crew (still feels weird saying 6.3800/6.7800!) has been a highlight for me, and I couldn't ask for a better team and a cooler class to be a part of.

To Gao Laoshi, Zhou Laoshi, Liang Laoshi, Liao Laoshi, and the MIT Chinese department as a whole: thank you so much for the opportunity to connect with my roots. I never imagined that of all things, *Chinese class* would turn out to be one of the core experiences of my college career, spanning six semesters and a whole trip to Taiwan. Every Chinese class or office hours, I would come out feeling renewed because your joy and enthusiasm never failed to brighten my day. One

impromptu decision to take Chinese on a whim gave me so much happiness throughout these years, and the reason is because of you all.

To all my friends: you are a cherished part of me, and even though time and space might soon separate us, the memories I've made with you will stick with me for a lifetime. To Merrick Cai: thank you for being an amazing friend and roommate, and for making me laugh so hard I've somehow managed to nearly choke on water on MULTIPLE occasions. To Roger Jin and Jason Yang: your friendship — 5AM potato.land engineering sessions, late nights in 26-100, J-On! practices and all — is what got me through the pandemic and onwards. To Varkey Alumootil and Jason Lu: thanks for all the stories and advice, and for significantly enhancing my vocabulary. You guys are the greatest *philosopher kings*. To Cathy Yung and Joli Dou: thanks for making my last year in NH3 a hilariously fun time.

And of course, to my beloved family: Mom, Dad, Xavier, LaoLao, LaoYe, YeYe, NaiNai. You are my anchor and guiding light in this world. To bask in the warmth of your unwavering support and unconditional love — this is the best life I could ever live. I owe you everything. This thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sampling from a target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ is a problem of fundamental statistical interest, serving as an algorithmic primitive that powers generative machine learning [82, 38, 83], Bayesian inference [66], and differential privacy [33], among other fields of contemporary importance.

Lately, a surge of recent work has illuminated a deep connection between sampling and the better-understood field of Euclidean optimization [94, 25]. By endowing the space of probability measures with the right geometry — the Wasserstein geometry — a rich dictionary between sampling and optimization emerges: concepts like the "proximal operator" in Euclidean space admit synonyms like the "JKO operator" in the space of probability measures. In turn, this dictionary facilitates the design of algorithms for sampling by way of direct analogy, inspired by classical techniques in Euclidean optimization.

This thesis aims to expand this dictionary. We endeavor to further our understanding of what analogies we might be able to draw, and to see what gets lost in translation. We do this through the lens of a motivating case study with considerable practical relevance: Gaussian variational inference (VI). The fruit of our analysis is a suite of implementable, computationally efficient algorithms for Gaussian VI, with state-of-the-art convergence guarantees to boot.

1.1 Outline of the thesis

The rest of the thesis is organized as follows. In Section 1.2, we present background on the problem of Gaussian VI and place our work in the context of the larger literature on sampling and variational inference in Section 1.4. We outline our approach to solving Gaussian VI in Section 1.3 and clarify the notation used throughout the rest of the thesis in Section 1.5.

In Section 2.1, we recall and present core concepts in Euclidean optimization. We proceed to define the Wasserstein geometry and introduce the Bures–Wasserstein space of Gaussian probability measures in Section 2.2. The rich geometric structure of this space allows us to perform *calculus* and hence *optimization*. Then, in Section 2.3, we will articulate precise mappings between tools in Euclidean optimization versus Bures–Wasserstein optimization, laying the groundwork for our development of algorithms for Gaussian VI. We organize these analogies in Table 2.1.

In Section 3.1, we recall classical algorithms for Euclidean optimization, and provide crisp analyses of their convergence. Using the dictionary developed in Table 2.1, we translate these algorithms and analyses to the setting of the Bures–Wasserstein space in order to solve Gaussian VI. The result is the development and complexity analysis of (Stochastic) Forward–Backward Gaussian Variational Inference (FB–GVI) (based on the joint work [29]) as well as a variance-reduced version (VRFB–GVI). In our analysis, we encounter several complications owing to the "curved" structure of the BW space, and find that we must take care in handling key details not present in the Euclidean setting. We highlight these difficulties and elaborate on how to fix them, while deferring the details of technical proofs to Appendix A. Preliminary simulation results are provided in Appendix B, and a Jupyter notebook containing code for our experiments can be found at

`https://github.com/mzydiao/FBGVI/blob/main/FBGVI-Experiments.ipynb`.

Finally, we conclude in Chapter 4 and raise a number of questions and directions for future study.

1.2 Background

Suppose we wish to devise an algorithm to output a sample from a target probability distribution $\pi \propto \exp(-V)$, where $V\colon \mathbb{R}^d \to \mathbb{R}$ is a smooth function. To do this, we might imagine one

approach. Perhaps we don't know how to directly generate a sample from π, but we instead know:

1. how to generate a sample from a different distribution μ_0, and
2. how to apply an iterative procedure to update any sample so that its distribution eventually approaches π.

Then we could simply proceed by drawing a sample from μ_0 and repeatedly updating it according to the iterative procedure in (2). This is the idea of the Monte Carlo Markov Chain (MCMC) method [62, 63], which accomplishes (2) by forming a reversible Markov chain with stationary distribution π, and updating a sample by running one step of that Markov chain. As the number of iterations tends to infinity, the law of the iterate converges in distribution to π.

However, there is an issue: in the real world, we cannot simply send the number of iterations to infinity. We need to stop at some point. In order to know if our samples are any good, we need quantitative, non-asymptotic guarantees that bound the discrepancy between the target distribution and the distribution of our samples. And if convergence is slow, in order to draw many approximate samples from our target distribution, we would need to run this iterative procedure many times, leading to a blowup in computational complexity.

However, there is another approach. What if instead of taking a tractable distribution μ_0 and iteratively updating its samples to fit the shape of π, we simply found a tractable distribution μ_0 that was close to π in the first place? Then we could just directly draw samples from μ_0 and avoid running MCMC altogether. This is the fundamental idea of variational inference (VI) [14, 47]. Of particular interest is the problem of Gaussian VI, in which we approximate π by the solution to

$$\operatorname*{arg\,min}_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \mathsf{KL}\left(\mu \,\|\, \pi\right), \tag{1.1}$$

where KL denotes the Kullback–Leibler divergence and $\mathsf{BW}(\mathbb{R}^d)$ the set of Gaussian distributions over $\mathbb{R}^d$ (see Section 3.2.1 for formal definitions). Gaussian VI has demonstrated favorable performance to MCMC in important settings of practical interest, especially in the presence of large datasets: see, for example, Barber and Bishop [9], Seeger [80], Honkela and Valpola [39], Opper and Archambeau [68], and Quiroz, Nott, and Kohn [74].

Why focus on Gaussians? For one, they are indeed easy to sample from [16]. And in the literature on Gaussian VI, recent works have demonstrated strong *statistical* properties for the solutions to Problem (1.1): see, for example, Chérief-Abdellatif, Alquier, and Khan [24], Alquier and Ridgway [2], and Katsevich and Rigollet [45]. As further justification, consider the case where π represents the posterior distribution of a sufficiently regular Bayesian model. Then, the Bernstein–von Mises theorem (see Vaart [89, Chapter 10] and recent non-asymptotic results [44, 85]) state that π is well-approximated by a Gaussian distribution, with mean given by any asymptotically efficient estimator of the true parameter, and covariance matrix given by the inverse Fisher information matrix.

With abundant motivation in mind for efficiently computing the best Gaussian approximation of π, we seek to develop a principled approach for solving Problem (1.1). Several other existing methods have been proposed, which we summarize in the related works (Section 1.4). Of particular note is the approach of Lambert et al. [51], who recently proposed an algorithm for Gaussian VI that can be seen as an analog of stochastic gradient descent for Problem (1.1) over the space $\mathsf{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance, called the Bures–Wasserstein (BW) space. This viewpoint takes inspiration from the theory of gradient flows over the Wasserstein space of probability measures [43, 5], the machinery of which has been instrumental for many problems in probabilistic inference (see Section 1.4 for a discussion of related works).

However, from an optimization standpoint, the gradient descent-inspired approach of Lambert et al. [51] is not the most natural. Indeed, the objective functional $\mathsf{KL}\left(\cdot \,\|\, \pi\right)$ is composite: it can be canonically decomposed as the sum of a "smooth" term called the potential and a "non-smooth" term called the entropy. This key observation has given rise to more than two decades of research on forward-backward methods on the Wasserstein space [see, for example, 43, 11, 94, 77]. In Euclidean optimization, the correct approach for optimizing a composite objective consisting of a smooth term and a non-smooth term is not gradient descent, but rather the *proximal* gradient algorithm [71], which crucially relies on the use of the proximal operator in Euclidean space. In Wasserstein space, the natural analogue of the proximal operator is the so-called JKO operator [43], but unfortunately this operator is computationally intractable in general, hampering the implementation of analogous proximal gradient algorithms in Wasserstein space.

1.3 Our approach

In this thesis, we develop an approach to Gaussian VI that resolves the aforementioned issues faced by previous works, namely non-smoothness and non-implementability. The result is a novel, implementable and efficient algorithm called (Stochastic) Forward-Backward Gaussian Variational Inference (FB–GVI). In the same vein as Lambert et al. [51], the rich differential and geometric structure of the BW space comprises the linchpin of our approach, but we are able to fully exploit optimization tools tailored to handling the composite nature of the KL divergence, whereas previous approaches fail. A key insight in this work is that the JKO operator for the entropy, when restricted to the BW space, admits a closed form [94], and hence leads to a truly implementable (stochastic) forward-backward (or proximal gradient) algorithm for Gaussian VI. In turn, it yields new state-of-the-art computational guarantees for Gaussian VI under a variety of standard assumptions. In addition, we demonstrate that our approach is extensible to the setting where the potential admits a representation as the average of many component functions, a common scenario in machine learning [36]. Leveraging tools from the literature on variance reduction in stochastic optimization in Euclidean space [42, 95], we devise a variance-reduced version of (Stochastic) FB–GVI that comes with significantly improved complexity guarantees.

In summary, we highlight our contributions below.

- We propose a new (stochastic) forward-backward algorithm, (Stochastic) FB–GVI, to solve Problem (1.1). The algorithm relies on a closed-form formula for the JKO operator of the entropy over the BW space.
- We prove state-of-the-art convergence rates for Gaussian VI via our algorithm, leveraging recent techniques of optimization over the space of probability measures [5].
- We devise a variance-reduction method for (stochastic) FB–GVI that enjoys favorable complexity guarantees.

1.4 Related work

We now contextualize our work in the setting of the larger literature on sampling and variational inference, and discuss closely related streams of research.

Optimization algorithms for Gaussian VI. Algorithms for solving Gaussian VI have been considered in Paisley, Blei, and Jordan [70], Ranganath, Gerrish, and Blei [75], and Lambert, Bonnabel, and Bach [50]. The general approach is to parametrize the set of Gaussian distributions and to apply Euclidean optimization. In particular, Alquier and Ridgway [2] noticed that when π is the posterior distribution in a Bayesian logistic regression, Problem (1.1) becomes convex with a certain choice of parametrization. In this case, they also characterized the statistical properties of the iterates of gradient descent. Other settings in which the corresponding optimization problem is convex are provided in Challis and Barber [20] and Domke [30]. In particular, [30] showed under the parameterization of [2], the Euclidean smoothness/convexity properties of the negative log-density of the model give rise to the same smoothness/convexity properties on the parameter space. However, to obtain convergence rates in the stochastic setting, one needs to control the variance of the stochastic gradient. This non-trivial task is not carried out in [30]. In our work, the required variance control is established in Lemma 3.3.5, and forms a crucial step in obtaining our convergence rates.

Algorithms based on natural gradient methods [99, 54, 55] and normalizing flows [76, 46, 19] have also been proposed for variational inference. However, to the best of our knowledge, convergence results for such methods are lacking in the literature.

Finally, the closest related work to ours in this literature is that of Lambert et al. [51], who similarly proposed an optimization algorithm over the BW space, called Bures–Wasserstein Stochastic Gradient Descent (BW–SGD), to solve Problem (1.1). Their algorithm relies on taking the gradient of the non-smooth entropy, and in particular they were only able to provide a (suboptimal) rate of convergence when π is strongly log-concave. In this work, we not only improve upon their convergence rate in the strongly log-concave case, but also demonstrate a convergence rate for the log-concave case as well.

Minimization of KL over the Wasserstein space. As mentioned previously, our approach has roots in the recent literature on viewing sampling methods as optimization algorithms over the Wasserstein space.

For example, the Langevin Monte Carlo (LMC) algorithm [28] is an MCMC algorithm to sample from the target distribution π. The theory of Wasserstein gradient flows [5] provides

the mathematical tools to view LMC (and its many variants) as an optimization algorithm over Wasserstein space. In the case of LMC, the objective to minimize is $\mathsf{KL}\left(\cdot \,\|\, \pi\right)$. Therefore, one can use optimization analysis (over the Wasserstein space) to show convergence bounds for LMC [94, 32, 8, 23, 25].

Stein Variational Gradient Descent [57, 56] is another method that can be seen as an optimization algorithm for minimizing $\mathsf{KL}\left(\cdot \,\|\, \pi\right)$. SVGD is a deterministic algorithm that drives the empirical distribution of a set of particles to fit π. The iterations of SVGD are computed by iterating a well-chosen map T such that $T - I$ belongs to a Reproducing Kernel Hilbert Space (RKHS). Little is known about the convergence rates of SVGD [31, 58, 27, 48, 78, 37, 81]. However, there is an interesting connection between SVGD and BW–GD [51]: when the number of particles of SVGD tends to infinity (the "mean-field" limit), the iterations of BW–GD are equivalent to the iterations of SVGD if the RKHS is the set of affine functions with symmetric linear part. We also remark that other heuristic algorithms for particle-based optimization over Wasserstein spaces have been proposed, for example, in [18, 4, 7, 92, 96] without any non-asymptotic convergence guarantees.

Another closely related work to ours is that of Salim, Korba, and Luise [77]. In the same vein as our work, they view the objective $\mathsf{KL}\left(\cdot \,\|\, \pi\right)$ as a composite functional over the Wasserstein space. They propose a forward-backward algorithm, involving the JKO of the entropy, with strong convergence properties. However, they do not address the fact that the JKO of the entropy is not implementable in general: hence, to our knowledge, their algorithm as a whole is not implementable. On the contrary, our algorithm relies on the JKO of the entropy over the BW space, which is shown to admit a closed form.

(Non–smooth) manifold optimization. Our work is also related to recent works developing efficient algorithms for solving non-smooth optimization problems over certain manifolds. For example, in the deterministic setting over manifolds, Li et al. [53] analyzed the sub-gradient method, Chen et al. [22] and Huang and Wei [41] analyzed the proximal gradient method, Chen et al. [21] analyzed the proximal point method, and Wang et al. [93] analyzed the proximal linear method. Stochastic versions were considered in Li, Balasubramanian, and Ma [52] and Wang et al. [93]. We also refer to Zhang, Chen, and Ma [97], Hu et al. [40], Peng et al. [72], and Zhang and Davanloo Tajbakhsh [98] for other recent advances in non–smooth manifold optimization. Several

of the above works consider the general setting of a Riemannian manifold. While the BW space is a Riemannian manifold, it is also a subset of the Wasserstein space, a structure we leverage in our work to prove our convergence results.

The geometry of the BW space was investigated in Modin [64], Malagò, Montrucchio, and Pistone [60], and Bhatia, Jain, and Lim [12], and optimization over this space has proven to be fruitful for various applications, see, for example, Chewi et al. [26], Altschuler et al. [3], Han et al. [35], Lambert et al. [51], Luo and Trillos [59], and Maunu, Le Gouic, and Rigollet [61].

1.5 Notation

We will make use of standard complexity notation. We write $a_n \lesssim b_n$, or alternatively $a_n = O(b_n)$, to indicate that $a_n \leq Cb_n$ for a universal constant $C > 0$. Similarly, $a_n \gtrsim b_n$ and $a_n = \Omega(b_n)$ indicate that $a_n \geq cb_n$ for some constant $c > 0$, and $a_n \asymp b_n$ and $a_n = \Theta(b_n)$ indicate that both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. The notation $a \wedge b$ and $a \vee b$ are shorthand for $\min(a, b)$ and $\max(a, b)$, respectively.

We will denote the space of real symmetric $d \times d$ matrices by $\mathbf{S}^d$ and the space of real positive definite $d \times d$ matrices by $\mathbf{S}^d_{++}$. Additionally, we denote the $d \times d$ dimensional identity matrix by I. Throughout, $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures μ over $\mathbb{R}^d$ with finite second moment $\int \|x\|^2 \, \mathrm{d}\mu(x) < \infty$. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The space $L^2(\mu)$ is the Hilbert space of Borel functions $f : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\mathbb{E}_\mu \|f\|^2 = \int \|f(x)\|^2 \, \mathrm{d}\mu(x) < \infty,$$

endowed with the inner product

$$\langle f, g \rangle_\mu := \int \langle f(x), g(x) \rangle \, \mathrm{d}\mu(x)$$

and the associated norm $\|f\|_\mu = \sqrt{\langle f, f \rangle_\mu}$. In particular, the identity map $\mathrm{id} : \mathbb{R}^d \to \mathbb{R}^d$ belongs to $L^2(\mu)$. If $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $T \in L^2(\mu)$, the pushforward measure of μ by T is denoted by $T_\# \mu$. This pushforward measure satisfies $\int \varphi \, \mathrm{d}T_\# \mu = \int \varphi(T(x)) \, \mathrm{d}\mu(x)$ for any measurable function $\varphi : \mathbb{R}^d \to \mathbb{R}_+$. The subset of $\mathcal{P}_2(\mathbb{R}^d)$ of all Gaussian distributions with positive definite covariance matrix is denoted by $\mathsf{BW}(\mathbb{R}^d)$. For an element $\mu \in \mathsf{BW}(\mathbb{R}^d)$, we denote its mean by m_μ and its

covariance matrix by Σ_μ. The notation $\mathcal{N}(m, \Sigma)$ refers to the Gaussian distribution with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbf{S}^d_{++}$. Finally, the notation κ denotes the condition number $\frac{\beta}{\alpha}$.

Chapter 2

Sampling and Optimization

In this chapter, our aim is to provide exposition on stochastic and non-smooth composite optimization, and to explicitly draw parallels between Euclidean optimization and optimization over the Bures–Wasserstein (BW) space.

We first consider the setting of optimization over Euclidean space in Section 2.1, defining the key terms and tools used. With the Euclidean setting as a methodological anchor, in Section 2.2 we proceed to define the BW space — the primary object of our study — and describe its rich geometric and differential structure, which is key for performing optimization. Finally, we make explicit the relations between optimization-related concepts in Euclidean space and those in BW space, collecting them into a brief "dictionary" in Section 2.3. This dictionary lays the groundwork for the translation of Euclidean optimization algorithms to the setting of the BW space.

2.1 Euclidean optimization

Before delving into optimization on the space of probability measures, we review some details of stochastic and convex non-smooth optimization over the Euclidean space $\mathbb{R}^d$. First, a function $V\colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth if V is twice continuously differentiable and its Hessian $\nabla^2 V(x)$ is bounded by β in the operator norm, for every $x \in \mathbb{R}^d$. In particular, V is differentiable and its gradient ∇V is β-Lipschitz. In addition, for any $h \in \mathbb{R}^d$, V satisfies the Taylor inequality

$$|V(x+h) - V(x) - \langle \nabla V(x), h\rangle| \leq \frac{\beta}{2}\|h\|^2 . \tag{2.1}$$

Equivalently, for any $x_0, x_1 \in \mathbb{R}^d$, we have that

$$V(x_1) \leq V(x_0) + \langle \nabla V(x_0), x_1 - x_0 \rangle + \frac{\beta}{2} \|x_1 - x_0\|^2 . \tag{2.2}$$

Consider the optimization problem

$$\min_{x \in \mathbb{R}^d} \{V(x)\} , \tag{2.3}$$

where $V \colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth. Given an initial point $x_0 \in \mathbb{R}^d$ and a step size $\eta > 0$, the classical algorithm of *gradient descent* proceeds by iteratively applying the following update rule for $k = 0, \ldots, N-1$:

$$x_{k+1} = x_k - \eta \nabla V(x_k) . \tag{2.4}$$

In machine learning applications, the user often does not have direct access to $\nabla V(x_k)$ because computing the gradient of V is expensive. Instead, the user has access to a cheaper stochastic estimator $\hat{g}_k$ of $\nabla V(x_k)$. In this scenario, the *stochastic* gradient descent algorithm uses $\hat{g}_k$ in lieu of $\nabla V(x_k)$ in Equation (2.4), resulting in the update rule

$$x_{k+1} = x_k - \eta \, \hat{g}_k . \tag{2.5}$$

Gradient descent and its stochastic variant enjoy a number of favorable convergence guarantees in the setting where the step size η is sufficiently small and when the variance of the gradient estimate $\hat{g}_k$ is controlled. In fact, in the setting when V is α-*strongly convex*, meaning that:

1. $0 \prec \alpha \preceq \nabla^2 V$, or equivalently,
2. for any $x_0, x_1 \in \mathbb{R}^d$ we have

$$V(x_1) \geq V(x_0) + \langle \nabla V(x_0), x_1 - x_0 \rangle + \frac{\alpha}{2} \|x_1 - x_0\|^2 , \tag{2.6}$$

gradient descent in fact converges at a *linear* rate to the minimizer of the objective. We elaborate on these guarantees and provide crisp and concise proofs of convergence in Section 3.1.

Unfortunately, smoothness can be a restrictive condition. Not all relevant objectives are

smooth over the entire Euclidean space, meaning that we cannot naively apply the gradient descent algorithm to solve Problem (2.3). However, in a variety of practical problems, the objective has a special *composite* structure that may be exploited. In particular, consider the optimization problem

$$\min_{x \in \mathbb{R}^d} \{V(x) + H(x)\}, \tag{2.7}$$

where $V \colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth and $H : \mathbb{R}^d \to \mathbb{R}$ is convex but potentially non-smooth. This setting arises naturally in constrained optimization and classical problems such as LASSO [86, 71]. Because of the non-smoothness of H, the gradient descent algorithm applied to Problem (2.7) may not converge. But there is a fix. We introduce the *proximal operator* of H, defined by

$$\operatorname{prox}_H(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ H(y) + \frac{1}{2} \|x - y\|^2 \right\}. \tag{2.8}$$

In many situations of interest, there exist computationally tractable closed-form expressions for the proximal operator.[1] And given access to this proximal operator as well as the gradient of V, the proximal gradient (also called *forward-backward*) algorithm [10] provides a canonical approach to solve Problem (2.7). Starting with an initial point $x_0 \in \mathbb{R}^d$ and a step size $\eta > 0$, the proximal gradient algorithm computes the following iterative updates for $k = 0, \ldots, N-1$:

$$x_{k+1} = \operatorname{prox}_{\eta H}(x_k - \eta \nabla V(x_k)). \tag{2.9}$$

Analogously to gradient descent, the proximal gradient algorithm admits a stochastic variant based on access to a stochastic estimator $\hat{g}_k$ of $\nabla V(x_k)$. In this case, $\hat{g}_k$ is again substituted for $\nabla V(x_k)$ in Equation (2.9), resulting in the following update rule:

$$x_{k+1} = \operatorname{prox}_{\eta H}(x_k - \eta \hat{g}_k). \tag{2.10}$$

Similarly to gradient descent, the proximal gradient algorithm and its stochastic variant enjoy favorable convergence guarantees when η is sufficiently small and when the variance of the gradient estimate $\hat{g}_k$ is controlled [6, 13, 34]. When V is additionally α-convex, proximal gradient

[1] See `proximity-operator.net`.

also obtains a linear rate of convergence to the minimizer of Problem (2.7). Furthermore, proximal gradient descent is applicable to a strictly larger class of optimization problems, as it is able to handle non-smooth objectives with composite structure as discussed above.

We consider one more extension. In many settings of relevance in contemporary machine learning [36], V is more naturally expressed as the average of many other simpler smooth functions V_i. Specifically, for $i = 1, \dots, m$, we have that $V_i \colon \mathbb{R}^d \to \mathbb{R}$ is a β-smooth and α-strongly convex function, and are given that V can be expressed as

$$V(x) = \frac{1}{m} \sum_{i=1}^{m} V_i(x).$$

If m is large, the update equations of gradient descent (2.4) and proximal gradient (2.9) may be computationally expensive, as computing the gradient of V requires evaluating ∇V_i for each $i = 1, \dots, m$.

One might imagine applying the stochastic versions of gradient descent or proximal gradient using the unbiased estimator $\hat{g}_k = \nabla V_{i_k}$ for a randomly selected index $i_k \sim \text{Unif}[m]$. Although this method is cheap to compute as it now only requires 1 rather than m gradient evaluation per iteration, it suffers from a slow convergence rate, owing to the fact that the stochastic estimate $\hat{g}_k$ has a large variance.

Instead, the approach of Stochastic Variance Reduced Gradient (SVRG) [42] and its proximal analogue prox-SVRG [95] seeks to leverage the best of both worlds: low variance of the gradient estimate, as well as cheap iteration complexity. It achieves this by introducing a *centering step*, wherein the gradient estimate is instead taken to be

$$\hat{g}_k = \nabla V_{i_k}(x_k) - \nabla V_{i_k}(y) + \nabla V(y).$$

This is again an unbiased estimator of ∇V when $i_k \sim \text{Unif}[m]$. The key benefit of this alternative estimator is that its variance can be controlled: intuitively as x_k and y both approach the optimum of V, the values $\nabla V(y)$ and $\nabla V_{i_k}(x_k) - \nabla V_{i_k}(y)$ both shrink to zero. The *centering point* y is updated less frequently than at every step of the algorithm, so even though the gradient $\nabla V(y)$ is expensive to compute, its value can be reused for many iterations before y is updated, signifi-

cantly reducing the amortized iteration complexity compared to exact gradient descent. Combining these properties, SVRG and prox-SVRG achieve a significantly improved convergence rate per gradient evaluation, especially in the setting when m is large.

Having laid the foundational groundwork of optimization over Euclidean space, we now turn to the BW space, where the elements are no longer real-valued random vectors but rather *Gaussian probability measures*. Despite the fact that the BW space is not a Euclidean space, its inherent structure still enables us to perform optimization. We will define this structure — specifically, the *geometry* and *calculus* over the BW space. And in so doing, we will set the stage for the adaptation of gradient descent and proximal gradient to the BW space.

2.2 The Bures–Wasserstein space

A detailed presentation of the Wasserstein space and its geometry, which in turn enables optimization, can be found in Ambrosio, Gigli, and Savaré [5]. In this section, we provide an overview of the *Bures–Wasserstein* (BW) space and its geometry, hence providing the requisite tools to perform optimization over the BW space and solve Problem (1.1). We start with formal definitions of the Wasserstein and BW spaces.

2.2.1 Geometry of the BW space

The *Wasserstein space* is the metric space $\mathcal{P}_2(\mathbb{R}^d)$ endowed with the 2-Wasserstein distance W_2 (which we simply refer to as the Wasserstein distance). We recall that the Wasserstein distance is defined for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \|x - y\|^2 \, \mathrm{d}\gamma(x, y)\,, \tag{2.11}$$

where $\mathcal{C}(\mu, \nu)$ is the set of couplings between μ and ν. The BW space is the metric space $\mathsf{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance W_2. In other words, the BW space is the subset of the Wasserstein space consisting of all Gaussian distributions with positive definite covariance matrix.

Given $\mu, \nu \in \mathsf{BW}(\mathbb{R}^d)$, there exists a unique *optimal transport* map from μ to ν: that is, a map

$T\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $T_{\#}\mu = \nu$ and

$$W_2^2(\mu,\nu) = \int \|x - T(x)\|^2 \,\mathrm{d}\mu(x)\,. \tag{2.12}$$

In other words, the coupling $(\mathrm{id}, T)_{\#}\mu$ belongs to $\mathcal{C}(\mu,\nu)$ and attains the infimum in (2.11). Moreover, since μ and ν are both Gaussian measures, T is in fact an affine map with symmetric linear part, meaning that it can be written as $T(x) = Sx + b$, where $S \in \mathbf{S}^d$ and $b \in \mathbb{R}^d$ [67]. In particular, the BW space is a genuine *Riemannian manifold* where at each $\mu \in \mathsf{BW}(\mathbb{R}^d)$, the tangent space $\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ corresponds to the space of d-dimensional affine maps with symmetric linear part. Using the fact that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we can deduce that $T_{\#}\mu \in \mathcal{P}_2(\mathbb{R}^d)$ implies $T \in L^2(\mu)$. Therefore, $\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ is naturally endowed with the $L^2(\mu)$ inner product, making $\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ a finite-dimensional subspace of $L^2(\mu)$.

2.2.2 Optimization over the BW space

In this section, we review the differential structure of the BW space. Further background on differential calculus over the BW space is provided in Section 2.2.3.

Smoothness over BW space

Consider a functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$. We say that $\mathcal{F}$ is differentiable at μ if there exists $g_\mu \in \mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ such that for every affine map $h\colon \mathbb{R}^d \to \mathbb{R}^d$,

$$\mathcal{F}((\mathrm{id} + th)_{\#}\mu) = \mathcal{F}(\mu) + t\,\langle g_\mu, h\rangle_\mu + o(t)\,. \tag{2.13}$$

In this case, g_μ is unique, and called the *Bures–Wasserstein gradient* of $\mathcal{F}$ at μ. We denote this gradient by $\nabla_{\mathsf{BW}}\mathcal{F}(\mu) = g_\mu$.

Given a β-smooth function $V\colon \mathbb{R}^d \to \mathbb{R}$, the *potential energy* functional (or simply potential) $\mathcal{V}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ is defined by

$$\mathcal{V}(\mu) := \int V \,\mathrm{d}\mu \tag{2.14}$$

for every $\mu \in \mathsf{BW}(\mathbb{R}^d)$. The potential is a prototypical example of a differentiable functional over the BW space. Lemma 2.2.1 verifies that the potential is differentiable and gives a formula for its

BW gradient. The formula for the BW gradient can be obtained by a straightforward adaptation of Lambert et al. [51, Section C.1], but we give a self-contained derivation in Section 2.2.3. Notably, the differentiability of $\mathcal{V}$ is well-established in the literature on Wasserstein space [5, Theorem 10.4.13]; we adapt this result to the BW space.

Lemma 2.2.1 (BW gradient of the potential)**.** *Consider the potential $\mathcal{V}$ defined by (2.14) where V is β-smooth. Then, $\mathcal{V}$ is differentiable at $\mu \in \mathsf{BW}(\mathbb{R}^d)$ and the following Taylor inequality holds: for $h\colon \mathbb{R}^d \to \mathbb{R}^d$ affine,*

$$\left|\mathcal{V}((\mathrm{id}+h)_{\#}\mu) - \mathcal{V}(\mu) - \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_{\mu}\right| \leq \frac{\beta}{2}\|h\|_{\mu}^2 . \tag{2.15}$$

Moreover, the BW gradient of $\mathcal{V}$ is known in closed form:

$$\nabla_{\mathsf{BW}}\mathcal{V}(\mu) : x \mapsto \mathbb{E}_{\mu}\nabla V + (\mathbb{E}_{\mu}\nabla^2 V)(x - m_{\mu}) , \tag{2.16}$$

where $m_{\mu} = \int x \,\mathrm{d}\mu(x)$ is the mean of μ.

Proof. We defer these proofs to the next section, which give greater detail on the calculus of the BW space. The proof of Equation (2.16) can be found in Section 2.2.3, and the proof of Inequality (2.15) can be found in Section 2.2.4. □

Essentially, the potential $\mathcal{V}$ inherits the smoothness properties from V. Note the analogy with Inequality (2.1). Inequality (2.15) is stronger than differentiability and can be interpreted as the potential $\mathcal{V}$ being β-smooth over the BW space, giving a BW analogue of Euclidean smoothness.

Convexity over BW space

We say that $\mathcal{F}$ is *geodesically convex* if for all $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$,

$$\mathcal{F}(\mu_0) + \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu_0), T - \mathrm{id}\rangle_{\mu_0} \leq \mathcal{F}(\mu_1) , \tag{2.17}$$

where T is the optimal transport map from μ_0 to μ_1. In this case, we can introduce an analog of the proximal operator of $\mathcal{F}$ over the BW space, called the *BW JKO operator* of $\mathcal{F}$ [43].[2] The BW JKO

[2] In Jordan, Kinderlehrer, and Otto [43] the authors define the JKO operator as an analog of the proximal operator over the entire Wasserstein space. In our work, we define the JKO operator *over the BW space*, and we call it the BW JKO operator.

operator is defined by

$$\mathrm{BWJKO}_{\mathcal{F}}(\mu) \coloneqq \operatorname*{arg\,min}_{\nu \in \mathsf{BW}(\mathbb{R}^d)} \left\{ \mathcal{F}(\nu) + \frac{1}{2} W_2^2(\mu, \nu) \right\}. \tag{2.18}$$

This definition is exactly analogous to (2.8), but with $W_2^2(\mu, \nu)$ in the BW JKO operator taking the role of the squared Euclidean distance in the proximal operator.

The *entropy functional* (or simply "entropy") $\mathcal{H}$ is a crucial example of a geodesically convex functional over the BW space. More precisely, the entropy is defined by

$$\mathcal{H}(\mu) = \int \log \mu(x) \, \mathrm{d}\mu(x) \,, \tag{2.19}$$

for every $\mu \in \mathsf{BW}(\mathbb{R}^d)$, where we identify μ with its density w.r.t. Lebesgue measure.[3]

Lemma 2.2.2 verifies that the entropy is geodesically convex and gives a formula for its BW JKO. The formula for the BW JKO operator can be obtained by a straightforward adaptation of Wibisono [94, Example 7], although we replicate the argument here. Notably, the geodesic convexity of $\mathcal{H}$ is also well-established in the literature on the Wasserstein space [5, Remark 9.3.10]; we adapt this result to the BW space.

Lemma 2.2.2 (BW JKO of the entropy)**.** *Consider the entropy functional $\mathcal{H}$ defined in (2.19). Then, $\mathcal{H}$ is geodesically convex and the following stronger inequality holds: for all $\nu, \mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$,*

$$\mathcal{H}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu \leq \mathcal{H}(\mu_1) \,, \tag{2.20}$$

where T_0 (resp. T_1) is the optimal transport map from ν to μ_0 (resp. μ_1).

Moreover, for $\mu \in \mathsf{BW}(\mathbb{R}^d)$, the BW JKO operator of $\mathcal{H}$ is known in closed form: with $\eta > 0$, we have that $\mathrm{BWJKO}_{\eta\mathcal{H}}(\mu) = \mathcal{N}(m_\mu, \Sigma_\star)$, where m_μ is the mean of μ, and the covariance matrix $\Sigma_\star$ is given by

$$\Sigma_\star = \frac{1}{2} \left(\Sigma_\mu + 2\eta I + \left[\Sigma_\mu \left(\Sigma_\mu + 4\eta I \right) \right]^{1/2} \right), \tag{2.21}$$

[3] To avoid possible confusion, note that this definition differs from contexts such as information theory, where the entropy is typically defined as the expectation of the *negative* log-density. This is an intentional matter of convention, since in our scenario $\mathcal{H}$ remains a *convex* functional rather than a *concave* functional (as it would be in the information theoretic context).

where Σ_μ is the covariance matrix of μ.

Proof. We defer the proofs of Equation (2.21) and Inequality (2.20) to Section 2.2.4, after a detailed discussion of calculus over the BW space. □

Inequality (2.20) is stronger than geodesic convexity and is a consequence of the so-called *generalized geodesic convexity* of the entropy [5, Remark 9.3.10]. We elaborate on this property in our detailed discussion of calculus in the BW space (Section 2.2.3). It is remarkable that we can compute the BW JKO of the entropy in closed form, as the JKO operator of the entropy over all of Wasserstein space is intractable in general [43]. **The ability to compute the BW JKO in closed form is key to our approach to solving Gaussian VI**. As a comparison, Salim, Korba, and Luise [77] proposed a proximal gradient algorithm relying on the JKO of the entropy over the whole Wasserstein space, but due to the intractability of the JKO, their algorithm is *not implementable*.

2.2.3 Calculus over the BW space

In this section, we describe a *calculus* over the BW space, deriving a formula for the BW gradient of a generic functional. In doing so, we demonstrate computation rules for differentiating a functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ along a curve of measures $(\mu_t)_{t\geq 0} \subseteq \mathsf{BW}(\mathbb{R}^d)$, which will be essential for our proofs of smoothness and convexity inequalities later on. Our derivation relies on specializing *Otto calculus* [69], which deals with the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, to the BW space $\mathsf{BW}(\mathbb{R}^d)$.

Background on Otto calculus

We first give an informal overview of the computation rules of Otto calculus [69], which endows the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ with a formal Riemannian structure. We refer to Ambrosio, Gigli, and Savaré [5] for a more rigorous development of the mathematical theory.

Let μ be an arbitrary element of $\mathcal{P}_2(\mathbb{R}^d)$ admitting a density w.r.t. Lebesgue measure. The tangent space $\mathfrak{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ is identified as the space of gradients of scalar functions on $\mathbb{R}^d$, i.e.,

$$\mathfrak{T}_\mu\mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\psi \mid \psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}.$$

For a functional $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, we can formally define its W_2 gradient at μ as the mapping

$\nabla_{W_2}\mathcal{F}(\mu) \in \mathfrak{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ satisfying

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu \,,$$

for any sufficiently regular curve of measures $(\mu_t)_{t\in\mathbb{R}} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ with $\mu_0 = \mu$ and velocity vector fields $(v_t)_{t\in\mathbb{R}}$ with $v_t \in L^2(\mu_t)$ for a.e. t satisfying the continuity equation

$$\partial_t\mu_t + \operatorname{div}(\mu_t v_t) = 0\,. \tag{2.22}$$

In fact, we can compute this W_2 gradient via direct identification. Let $\delta\mathcal{F}(\mu)\colon \mathbb{R}^d \to \mathbb{R}$ denote a *first variation* of $\mathcal{F}$ at μ [see 79, Chapter 7], for which

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\,\partial_t|_{t=0}\mu_t\,.$$

Then, by Equation (2.22) and integration by parts,

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\,\partial_t|_{t=0}\mu_t = -\int \delta\mathcal{F}(\mu)\operatorname{div}(\mu v_0) = \int \langle\nabla\delta\mathcal{F}(\mu), v_0\rangle\,\mathrm{d}\mu = \langle\nabla\delta\mathcal{F}(\mu), v_0\rangle_\mu\,.$$

Hence, we conclude that

$$\nabla_{W_2}\mathcal{F}(\mu) \equiv \nabla\delta\mathcal{F}(\mu)\,. \tag{2.23}$$

Now we turn our attention to the BW space. The BW space $\mathsf{BW}(\mathbb{R}^d)$ is a submanifold of $\mathcal{P}_2(\mathbb{R}^d)$ [69, 51], and hence inherits the formal Riemannian structure described above.

Let μ be an arbitrary element of $\mathsf{BW}(\mathbb{R}^d)$. The tangent space $\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ is identified as the space of affine functions on $\mathbb{R}^d$ with symmetric linear term, i.e.,

$$\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d) = \{x \mapsto b + S\,(x - m_\mu) \mid b \in \mathbb{R}^d,\, S \in \mathbf{S}^d\}\,.$$

In analogy to the above, for a functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$, we can formally define its BW gradient

at μ as the element $\nabla_{\mathsf{BW}}\mathcal{F}(\mu) \in \mathfrak{T}_\mu \mathsf{BW}(\mathbb{R}^d)$ satisfying

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu), v_0 \rangle_\mu \,,$$

for any curve of measures $(\mu_t)_{t\in\mathbb{R}} \subseteq \mathsf{BW}(\mathbb{R}^d)$ with $\mu_0 = \mu$ and velocity vector fields $(v_t)_{t\in\mathbb{R}}$, with each v_t an affine map, satisfying Equation (2.22). Using Equation (2.23) and integration by parts, we compute an expression for the BW gradient of $\mathcal{F}$ in the following subsection.

BW gradient calculation

The BW gradient of a functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ can be derived analogously to Lambert et al. [51, Section C.1]. We present a self-contained derivation here for completeness, and in doing so we obtain a formula for the rate of change of $\mathcal{F}$ along a curve of Gaussians for which the corresponding velocity vector fields are affine maps with linear parts which are *not necessarily symmetric*; this will play an important role in later proofs. The key idea is to use integration by parts repeatedly, exploiting the fact that the gradient of a Gaussian density is simply that same density multiplied by an affine term.

Lemma 2.2.3. *Let $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional on the Wasserstein space with first variation $\delta\mathcal{F}$. Then, for $\mu \in \mathsf{BW}(\mathbb{R}^d)$, we have that $\nabla_{\mathsf{BW}}\mathcal{F}(\mu)$ is given by*

$$\nabla_{\mathsf{BW}}\mathcal{F}(\mu)\colon x \mapsto (\mathbb{E}_\mu \nabla^2 \delta\mathcal{F})(x - m_\mu) + \mathbb{E}_\mu \nabla \delta\mathcal{F} \,.$$

Proof. Let $(\mu_t)_{t\in\mathbb{R}} \subseteq \mathsf{BW}(\mathbb{R}^d)$ be a regular curve of Gaussians with $\mu_0 = \mu$ and $(v_t)_{t\in\mathbb{R}}$ be a family of affine maps satisfying Equation (2.22). Furthermore, suppose that v_0 is given by

$$v_0\colon x \mapsto a + M\,(x - m_\mu)\,, \qquad (a, M) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}\,,$$

and that $\nabla_{\mathsf{BW}}\mathcal{F}(\mu) \in \mathfrak{T}_\mu \mathsf{BW}(\mathbb{R}^d)$ is given by

$$\nabla_{\mathsf{BW}}\mathcal{F}(\mu)\colon x \mapsto b_{\mathcal{F}} + S_{\mathcal{F}}\,(x - m_\mu)\,, \qquad (b_{\mathcal{F}}, S_{\mathcal{F}}) \in \mathbb{R}^d \times \mathbf{S}^d\,.$$

Letting $X \sim \mu$, we find that

$$\begin{aligned}
\langle \nabla_{\mathsf{BW}} \mathcal{F}(\mu), v_0 \rangle_\mu &= \mathbb{E}\left\langle b_{\mathcal{F}} + S_{\mathcal{F}}\left(X - m_\mu\right), a + M\left(X - m_\mu\right)\right\rangle \\
&= \langle b_{\mathcal{F}}, a\rangle + \mathbb{E}\left\langle S_{\mathcal{F}}\left(X - m_\mu\right), M\left(X - m_\mu\right)\right\rangle \\
&= \langle b_{\mathcal{F}}, a\rangle + \mathbb{E}\langle S_{\mathcal{F}}, M\left(X - m_\mu\right)\left(X - m_\mu\right)^{\mathsf{T}}\rangle \\
&= \langle b_{\mathcal{F}}, a\rangle + \left\langle S_{\mathcal{F}}, M\Sigma_\mu\right\rangle \\
&= \langle b_{\mathcal{F}}, a\rangle + \langle S_{\mathcal{F}}, \Sigma_\mu M^{\mathsf{T}}\rangle\,. \qquad \text{(since } S_{\mathcal{F}} = S_{\mathcal{F}}^{\mathsf{T}} \text{ and } \langle A, B\rangle = \langle A^{\mathsf{T}}, B^{\mathsf{T}}\rangle\text{)}
\end{aligned}$$

On the other hand, from the definition of the W_2 gradient, we obtain that

$$\begin{aligned}
\partial_t|_{t=0}\mathcal{F}(\mu_t) &= \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu && \text{(definition of } \nabla_{W_2}\mathcal{F}\text{)} \\
&= \langle \nabla\delta\mathcal{F}(\mu), v_0\rangle_\mu && \text{(by Equation (2.23))} \\
&= \mathbb{E}\left\langle \nabla\delta\mathcal{F}(X), a + M\left(X - \mathbb{E}X\right)\right\rangle \\
&= \mathbb{E}\left\langle \nabla\delta\mathcal{F}(X), a\right\rangle + \mathbb{E}\langle \Sigma_\mu M^{\mathsf{T}}\nabla\delta\mathcal{F}(X), \Sigma_\mu^{-1}\left(X - \mathbb{E}X\right)\rangle \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle - \int \langle \Sigma_\mu M^{\mathsf{T}}\nabla\delta\mathcal{F}, \nabla\mu\rangle && \text{(since } \nabla\mu(x) = -\mu(x)\,\Sigma_\mu\left(x - \mathbb{E}X\right)\text{)} \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle + \mathbb{E}[\mathrm{div}(\Sigma_\mu M^{\mathsf{T}}\nabla\delta\mathcal{F})(X)] && \text{(integration by parts)} \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle + \langle \mathbb{E}_\mu\nabla^2\delta\mathcal{F}(X), \Sigma_\mu M^{\mathsf{T}}\rangle\,.
\end{aligned}$$

Hence, by direct identification, we conclude that

$$(b_{\mathcal{F}}, S_{\mathcal{F}}) = (\mathbb{E}_\mu\nabla\delta\mathcal{F},\, \mathbb{E}_\mu\nabla^2\delta\mathcal{F})\,,$$

proving our desired result. □

Examples of BW gradients and stationary condition for Problem (1.1)

Consider the functional $\mathcal{F} = \mathcal{V} + \mathcal{H}$ defined by the sum of the potential (associated to the function V) and the entropy, and recall that Problem (1.1) is equivalent to minimizing $\mathcal{F}$ over the BW space, *i.e.*, solving Problem (3.11). Using the result of Lambert et al. [51, Section C.1], we have the

following formulas for the BW gradients of $\mathcal{V}$ and $\mathcal{H}$:

$$\begin{aligned}\nabla_{\mathsf{BW}}\mathcal{V}(\mu) &: x \mapsto \mathbb{E}_\mu \nabla V + (\mathbb{E}_\mu \nabla^2 V)(x - m_\mu)\,,\\ \nabla_{\mathsf{BW}}\mathcal{H}(\mu) &: x \mapsto -\Sigma_\mu^{-1}(x - m_\mu)\,.\end{aligned} \tag{2.24}$$

We can also derive the above formulas from Lemma 2.2.3.

Moreover, by the proof of Lemma 2.2.3, we can compute $\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu_t), v_t\rangle_{\mu_t}$ along any curve of Gaussians $(\mu_t)_{t\in\mathbb{R}}$ and any family of affine maps $(v_t)_{t\in\mathbb{R}}$ which together satisfy the continuity equation.

In particular, if $\hat{\pi}$ is a minimizer of (1.1), the first-order stationary condition $\nabla_{\mathsf{BW}}\mathcal{F}(\hat{\pi}) = 0$ for Problem (1.1) reads:

$$\mathbb{E}_{\hat{\pi}} \nabla V = 0 \qquad \text{and} \qquad \mathbb{E}_{\hat{\pi}} \nabla^2 V = \hat{\Sigma}^{-1}\,, \tag{2.25}$$

where $\hat{\Sigma}$ is the covariance matrix corresponding to the distribution $\hat{\pi}$.

2.2.4 Convexity and smoothness inequalities in the BW space for the potential and the entropy

Having derived a formula for the BW gradient of a generic functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ in Section 2.2.3, we may now proceed to prove Lemma 2.2.1 (for the potential) and Lemma 2.2.2 (for the entropy).

For both lemmas, the key idea is to differentiate a functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ along a curve $(\mu_t)_{t\in[0,1]}$ with velocity vector fields $(v_t)_{t\in[0,1]}$ satisfying the continuity equation (2.22), utilizing our calculation rules laid out in Section 2.2.3. In particular, we will use that

$$\begin{aligned}\mathcal{F}(\mu_1) - \mathcal{F}(\mu_0) &= \int_0^1 \partial_t \mathcal{F}(\mu_t)\,\mathrm{d}t\\ &= \partial_t|_{t=0}\mathcal{F}(\mu_t) + \int_0^1 \int_0^t \partial_s^2 F(\mu_s)\,\mathrm{d}s\,\mathrm{d}t\\ &= \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu_0), v_0\rangle_{\mu_0} + \int_0^1 (1-t)\,\partial_t^2 \mathcal{F}(\mu_t)\,\mathrm{d}t\,,\end{aligned} \tag{2.26}$$

for both the entropy and the potential.

Proof of Lemma 2.2.1: smoothness of the potential

We prove the following result for the potential. This result is stronger than Lemma 2.2.1, and will be useful in our subsequent analysis.

Lemma 2.2.4. *Suppose that $\alpha I \preceq \nabla^2 V \preceq \beta I$. Let $\mu \in \mathsf{BW}(\mathbb{R}^d)$ and let $h\colon \mathbb{R}^d \to \mathbb{R}^d$ be an affine map. Then the following inequalities hold:*

$$\mathcal{V}((\mathrm{id}+h)_{\#}\mu) - \mathcal{V}(\mu) \geq \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_{\mu} + \frac{\alpha}{2}\,\|h\|_{\mu}^2\,,$$
$$\mathcal{V}((\mathrm{id}+h)_{\#}\mu) - \mathcal{V}(\mu) \leq \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_{\mu} + \frac{\beta}{2}\,\|h\|_{\mu}^2\,.$$

Proof. Let $X \sim \mu$. Note that regardless of μ, we have that $\delta\mathcal{V}(\mu) = V$. Hence, $\nabla_{W_2}\mathcal{V}(\mu) = \nabla V$. We thus compute that

$$\begin{aligned}
\mathcal{V}((\mathrm{id}+h)_{\#}\mu) - \mathcal{V}(\mu) &= \mathbb{E}[V(X+h(X)) - V(X)] \\
&\geq \mathbb{E}\left[\langle \nabla V(X), h(X)\rangle + \frac{\alpha}{2}\,\|h(X)\|^2\right] && (\text{since } \nabla^2 V \succeq \alpha I) \\
&= \langle \nabla_{\mathsf{W}_2}\mathcal{V}(\mu), h\rangle_{\mu} + \frac{\alpha}{2}\,\|h\|_{\mu}^2 \\
&= \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_{\mu} + \frac{\alpha}{2}\,\|h\|_{\mu}^2\,,
\end{aligned}$$

proving the first inequality. The second inequality follows similarly, using the fact that $\nabla^2 V \preceq \beta I$. □

Lemma 2.2.1 then follows as a corollary of the above lemma.

Proof of Lemma 2.2.1. Note that if V is β-smooth, then we have by definition that $-\beta I \preceq \nabla^2 V \preceq \beta I$. Hence, applying Lemma 2.2.4 with $\alpha = -\beta$, we obtain that

$$\left|\mathcal{V}((\mathrm{id}+h)_{\#}\mu) - \mathcal{V}(\mu) - \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_{\mu}\right| \leq \frac{\beta}{2}\,\|h\|_{\mu}^2\,.$$

Moreover, we have shown in Section 2.2.3 that $\nabla_{\mathsf{BW}}\mathcal{V}(\mu)$ is given by

$$\nabla_{\mathsf{BW}}\mathcal{V}(\mu) : x \mapsto \mathbb{E}_{\mu}\nabla V + (\mathbb{E}_{\mu}\nabla^2 V)(x - m_{\mu})\,,$$

completing the proof of our desired result. □

Proof of Lemma 2.2.2: convexity of the entropy

For the entropy, we follow the same strategy as in the previous proof, differentiating the entropy $\mathcal{H}$ along a particular curve. This time, we will differentiate along the *generalized geodesic* $(\mu_t^\nu)_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$, which we define as follows:

Definition 2.2.5. Let T_0, T_1 be the optimal transport maps for which $T_0 - \mathrm{id}, T_1 - \mathrm{id} \in T_\nu \mathsf{BW}(\mathbb{R}^d)$ and $(T_0)_\# \nu = \mu_0$ and $(T_1)_\# \nu = \mu_1$, respectively. Defining $T_t := (1-t)\, T_0 + t\, T_1$, the *generalized geodesic* with basepoint ν and endpoints μ_0, μ_1 is then the curve of measures $(\mu_t^\nu)_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$ with $\mu_t^\nu = (T_t)_\# \nu$.

We note that $\mu_0^\nu = \mu_0$ and $\mu_1^\nu = \mu_1$, and that $(\mu_t^\nu)_{t\in[0,1]}$ solves the continuity equation

$$\partial_t \mu_t^\nu + \operatorname{div}(\mu_t^\nu v_t) = 0, \qquad \text{where } v_t = (T_1 - T_0) \circ T_t^{-1}.$$

Generalized geodesics were used in the work of Ambrosio, Gigli, and Savaré [5] to study gradient flows in the Wasserstein space, and have since been useful for various applications of the theory of Wasserstein gradient flows (see, for instance, Chewi et al. [26], Ahn and Chewi [1], and Altschuler et al. [3]).

Proof of Lemma 2.2.2. First, we remark that the JKO operator of $\mathcal{H}$ over the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ is derived in Wibisono [94, Example 7] for a Gaussian measure $\mu = \mathcal{N}(\mu, \Sigma)$, and takes the form $\mu' = \mathcal{N}(\mu, \Sigma_\star)$ where $\Sigma_\star$ is defined in the same manner as Equation (2.21). Since μ' is also an element of $\mathsf{BW}(\mathbb{R}^d)$, we conclude that μ' is also the result of applying the BW JKO operator to μ, proving our desired closed form. Alternatively, we can consider the following derivation, taking advantage of the BW calculus we have just defined:

Let $\nu \in \mathsf{BW}(\mathbb{R}^d)$ be an arbitrary element of the BW space, and let $T_{\nu\to\mu} \colon \mathbb{R}^d \to \mathbb{R}^d$ be the optimal transport map taking ν to μ. The result of Villani [90, Theorem 23.9] states that

$$[\nabla_{W_2} W_2^2(\cdot, \mu)](\nu) = 2(\mathrm{id} - T_{\nu\to\mu}). \tag{2.27}$$

Since $\mu, \nu \in \mathsf{BW}(\mathbb{R}^d)$, we know that $T_{\nu\to\mu}$ is in fact an affine map with positive semidefinite linear

term, so we can express it in the form

$$T_{\nu\to\mu}(x) = S_{\nu\to\mu}(x - m_\nu) + b_{\nu\to\mu}$$

for some $S_{\nu\to\mu} \in \mathbf{S}^d_{++}$ and $b_{\nu\to\mu} \in \mathbb{R}^d$. Hence, we can identify the BW gradient of $W_2^2(\cdot,\mu)$ as an element of $\mathfrak{T}_\nu \mathsf{BW}(\mathbb{R}^d)$ as follows:

$$\begin{aligned}
[\nabla_{BW} W_2^2(\cdot,\mu)](\nu) &\equiv (\mathbb{E}_\nu \nabla[\delta W_2^2(\cdot,\mu)](\nu), \mathbb{E}_\nu \nabla^2[\delta W_2^2(\cdot,\mu)](\nu)) && \text{(by Lemma 2.2.3)}\\
&= \left(\mathbb{E}_\nu\left[[\nabla_{W_2} W_2^2(\cdot,\mu)](\nu)\right],\ \mathbb{E}_\nu\left[\nabla[\nabla_{W_2} W_2^2(\cdot,\mu)](\nu)\right]\right) && \text{(by Equation (2.23))}\\
&= \left(2\mathbb{E}_\nu\left[\mathrm{id} - T_{\nu\to\mu}\right],\ 2\mathbb{E}_\nu\left[\nabla(\mathrm{id} - T_{\nu\to\mu})\right]\right) && \text{(by Equation (2.27))}\\
&= \left(2(m_\nu - m_\mu),\ 2(I - S_{\nu\to\mu})\right)\ .
\end{aligned}$$

Define

$$\nu_\star := \mathrm{BWJKO}_{\eta\mathcal{H}}(\mu) = \operatorname*{arg\,min}_{\nu\in\mathsf{BW}(\mathbb{R}^d)} \left\{\mathcal{H}(\nu) + \frac{1}{2\eta} W_2^2(\mu,\nu)\right\}\ .$$

Since $\nu_\star$ attains optimality in the above objective, it satisfies the stationarity condition

$$\nabla_{\mathsf{BW}}\left(\mathcal{H}(\nu) + \frac{1}{2\eta} W_2^2(\mu,\nu)\right)\Big|_{\nu=\nu_\star} = 0.$$

Hence, using the formulas for BW gradients we have just derived, we have that

$$\begin{aligned}
(0,0) &\equiv \nabla_{\mathsf{BW}}\mathcal{H}(\nu_\star) + \frac{1}{2\eta}\left[\nabla_{\mathsf{BW}} W_2^2(\cdot,\mu)\right](\nu_\star)\\
&= (0, -\Sigma_{\nu_\star}^{-1}) + \frac{1}{\eta}(m_{\nu_\star} - m_\mu, I - S_{\nu_\star\to\mu}).
\end{aligned}$$

Rearranging, we find that

$$\begin{aligned}
m_{\nu_\star} &= m_\mu\\
S_{\nu_\star\to\mu} &= I - \eta\Sigma_{\nu_\star}^{-1}\ .
\end{aligned}$$

On the other hand, we have that $\Sigma_\mu = S_{\nu_\star\to\mu}\Sigma_{\nu_\star}S_{\nu_\star\to\mu}$, so combining this with the above we obtain

that

$$\Sigma_\mu = (I - \eta \Sigma_{\nu_\star}^{-1})^2 \Sigma_{\nu_\star}.$$

Solving for $\Sigma_{\nu_\star}$, we find that

$$\Sigma_{\nu_\star} = \frac{1}{2}\left(\Sigma_\mu + 2\eta I + \left[\Sigma_\mu\left(\Sigma_\mu + 4\eta I\right)\right]^{1/2}\right),$$

which precisely matches Equation (2.21). Having shown that $\nu_\star = \mathcal{N}(m_\mu, \Sigma_\star)$, we conclude our proof of the desired formula for the BW JKO.

Now we demonstrate the desired generalized geodesic convexity inequality for the entropy. In fact, this claim follows from general results on the Wasserstein space [see, e.g., 5, §9.4], but we give a proof here for completeness. As mentioned above, to do so we will differentiate $\mathcal{H}$ along the generalized geodesic $(\mu_t^\nu)_{t\in[0,1]}$ defined above. Abusing notation, we identify a distribution μ with its density with respect to Lebesgue measure. We then have that

$$\begin{aligned}
\partial_t^2 \mathcal{H}(\mu_t^\nu) &= \partial_t^2 \int \mu_t^\nu \ln \mu_t^\nu = \partial_t^2 \int \nu \ln(\mu_t^\nu \circ T_t) = \partial_t^2 \int \nu \ln \frac{\nu}{\det \nabla T_t} \\
&\qquad\qquad\qquad\qquad (\text{since } (T_t)_\# \nu = \mu_t^\nu, \text{ change of variable}) \\
&= -\int (\partial_t^2 \ln \det \nabla T_t)\, \mathrm{d}\nu = -\int \partial_t \left\langle [\nabla T_t]^{-1}, \partial_t \nabla T_t \right\rangle \mathrm{d}\nu \\
&= -\int \partial_t \left\langle [\nabla T_t]^{-1}, \nabla T_1 - \nabla T_0 \right\rangle \mathrm{d}\nu = \int \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \mathrm{d}\nu \\
&= \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle,
\end{aligned}$$

where the last line follows since T_t is an affine map, meaning that ∇T_t is constant on $\mathbb{R}^d$. In addition, by Brenier's theorem [91, Theorem 2.12], T_t is the gradient of a convex function for all $t \in [0,1]$. Hence, we know that $\nabla T_t \succeq 0$ for all $t \in [0,1]$, meaning that $\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \geq 0$. Hence, using Equation (2.26) applied to $\mathcal{H}$, we obtain that

$$\begin{aligned}
\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 (1-t) \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \mathrm{d}t \\
&\geq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu .
\end{aligned}$$

This proves the desired inequality for the entropy, and we conclude our proof. □

Remark 2.2.6. In fact, we can show a *strong convexity* inequality for the entropy along generalized geodesics connecting distributions $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ with the same mean. Let m_0, m_1 be the means of μ_0, μ_1 respectively, and suppose that $\Sigma_{\mu_0}, \Sigma_{\mu_1} \preceq \lambda I$. We compute that

$$\begin{aligned}
\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle &= \left\langle I, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\
&\geq \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle \Sigma_{\mu_t^\nu}, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\
&= \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle [\nabla T_t]^{-1} \Sigma_{\mu_t^\nu} [\nabla T_t]^{-1}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\
&= \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle.
\end{aligned}$$

Since T_0 is an affine map, we know that $T_0(x) - (\nabla T_0)\, x$ is a constant for all $x \in \mathbb{R}^d$, and similarly for T_1. Hence, we find that if $Y \sim \nu$, then

$$\begin{aligned}
\|T_1 - T_0\|_\nu^2 &= \mathrm{Tr}(\mathrm{Cov}_\nu[T_1 - T_0, T_1 - T_0]) + \|\mathbb{E}_\nu[T_1 - T_0]\|^2 \qquad \text{(by bias-variance decomposition)} \\
&= \mathrm{Tr}(\mathrm{Cov}[(\nabla T_1 - \nabla T_0)(Y), (\nabla T_1 - \nabla T_0)(Y)]) + \|m_1 - m_0\|^2 \quad \text{(since } T_0, T_1 \text{ are affine)} \\
&= \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle + \|m_1 - m_0\|^2 . \qquad (2.28)
\end{aligned}$$

In addition, from Chewi et al. [26, Lemma 10], we know that the operator norm of the covariance matrix is convex along generalized geodesics in $\mathsf{BW}(\mathbb{R}^d)$, implying that $\Sigma_{\mu_t^\nu} \preceq \lambda I$ for all $t \in [0, 1]$. Thus, we obtain

$$\begin{aligned}
\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle &\leq \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\
&= \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right) \qquad \text{(by Equation (2.28))} \\
&\geq \frac{1}{\lambda} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right). \qquad \text{(by Chewi et al. [26, Lemma 10])}
\end{aligned}$$

Hence, using Equation (2.26) applied to $\mathcal{H}$, we obtain that

$$\begin{aligned}\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 (1-t)\, \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \mathrm{d}t \\ &\geq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \frac{1}{2\lambda} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right).\end{aligned}$$

This implies that the entropy is strongly convex along generalized geodesics between two Gaussians $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ with the same mean. Similarly, the same computation can be used to show a *smoothness* inequality for the entropy along *geodesics*. As before, we compute that

$$\begin{aligned}\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle &= \left\langle I, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\ &\leq \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left\langle \Sigma_{\mu_t}^\nu, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\ &= \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left\langle [\nabla T_t]^{-1} \Sigma_{\mu_t^\nu} [\nabla T_t]^{-1}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\ &= \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\ &= \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right) \\ &\leq \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \|T_1 - T_0\|_\nu^2 .\end{aligned}$$

Once again using Equation (2.26) applied to $\mathcal{H}$, we obtain that

$$\begin{aligned}\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 (1-t)\, \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \mathrm{d}t \\ &\leq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 \frac{1-t}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \|T_1 - T_0\|_\nu^2 \, \mathrm{d}t . \qquad (2.29)\end{aligned}$$

As a corollary of Inequality 2.29, we obtain a smoothness inequality for the entropy along geodesics, which will be useful for our subsequent analysis.

Lemma 2.2.7 (Smoothness of entropy along geodesics)**.** *Suppose that $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ satisfy $\Sigma_{\mu_0}^{-1}, \Sigma_{\mu_1}^{-1} \preceq \gamma I$. Then if T is the optimal transport map from μ_0 to μ_1, we have that*

$$\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) \leq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0), T - \mathrm{id} \rangle_{\mu_0} + \frac{\gamma}{2} \|T - \mathrm{id}\|_{\mu_0}^2 .$$

Proof. We apply Inequality 2.29 with $\nu = \mu_0$, noting in this case that $T_1 = T$ and $T_0 = \mathrm{id}$, and that $(\mu_t^\nu)_{t\in[0,1]}$ is precisely the constant-speed geodesic $(\mu_t)_{t\in[0,1]}$ connecting μ_0, μ_1. Furthermore, by Altschuler et al. [3, Appendix B], we know that $\lambda_{\min}$ is concave along geodesics, so $\lambda_{\min}(\Sigma_{\mu_t}) \geq \gamma^{-1} I$ for all t. Hence, we obtain

$$\begin{aligned}
\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &\leq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0), T - \mathrm{id} \rangle_{\mu_0} + \int_0^1 \frac{1-t}{\lambda_{\min}(\Sigma_{\mu_t})} \, \|T - \mathrm{id}\|_{\mu_0}^2 \, \mathrm{d}t \\
&\leq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0), T - \mathrm{id} \rangle_{\mu_0} + \int_0^1 \frac{1-t}{\gamma^{-1}} \, \|T - \mathrm{id}\|_{\mu_0}^2 \, \mathrm{d}t \\
&= \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0), T - \mathrm{id} \rangle_{\mu_0} + \frac{\gamma}{2} \, \|T - \mathrm{id}\|_{\mu_0}^2 \, ,
\end{aligned}$$

proving the desired result. □

2.3 From Euclidean to Bures–Wasserstein

Having defined the core vocabulary of optimization, we are now well-positioned to translate the Euclidean optimization toolkit to the setting of the BW space. Here in Table 2.1, we collect the relations between key concepts covered in Sections 2.1 and 2.2, comprising a "dictionary" between the Euclidean and Bures–Wasserstein spaces.

Concept	Euclidean space	BW space
squared distance	vectors $x, y \in \mathbb{R}^n$ squared Euclidean dist $\|x-y\|^2$	measures $\mu, \nu \in \mathsf{BW}(\mathbb{R}^d)$ squared W_2 dist $W_2^2(\mu,\nu)$
geodesic Section 2.2.2	curve $(x_t)_{t\in[0,1]} \subseteq \mathbb{R}^d$ $x_t = tx_1 + (1-t)x_0$	curve $(\mu_t)_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$ endpoints $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ T is OT map from μ_0 to μ_1 $\mu_t = (T_t)_\# \mu_0$, where $T_t := (1-t)\,\mathrm{id} + t\,T$
generalized geodesic Section 2.2.4	same as above	curve $(\mu_t^\nu)_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$ basepoint $\nu \in \mathsf{BW}(\mathbb{R}^d)$, endpoints $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ T_0, T_1 are OT maps from ν to μ_0, μ_1 $\mu_t^\nu = (T_t)_\# \nu$, where $T_t := (1-t)\,T_0 + t\,T_1$
constant-speed curve Section 2.2.4	same as above	curve $(\mu_t)_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$ endpoints $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ h is affine map, $h_\# \mu_0 = \mu_1$ $\mu_t = (h_t)_\# \mu_0$, where $h_t := (1-t)\,\mathrm{id} + t\,h$
β-smoothness Equation (2.1) Equation (2.15) Lemma 2.2.7	$V\colon \mathbb{R}^d \to \mathbb{R}$ diff'able *Equivalent forms:* $-\beta I \preceq \nabla^2 V \preceq \beta I$ for all $x, h \in \mathbb{R}^d$, $\lvert V(x+h) - V(x) - \langle \nabla V(x), h\rangle\rvert \le \frac{\beta}{2}\|h\|^2$ for all $x_0, x_1 \in \mathbb{R}^d$, $V(x_1) \le V(x_0) + \langle \nabla V(x_0), x_1 - x_0\rangle + \frac{\beta}{2}\|x_1 - x_0\|^2$	$\mathcal{V}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ diff'able *Over constant-speed curves:* $\lvert \mathcal{V}((\mathrm{id}+h)_\#\mu) - \mathcal{V}(\mu) - \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu), h\rangle_\mu\rvert \le \frac{\beta}{2}\|h\|_\mu^2$ for all $\mu \in \mathsf{BW}(\mathbb{R}^d)$ and $h \in \mathbb{R}^d \to \mathbb{R}^d$ affine *Over geodesics* $(\mu_t)_{t\in[0,1]}$*:* $\mathcal{V}(\mu_1) \le \mathcal{V}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu_0), T - \mathrm{id}\rangle_{\mu_0} + \frac{\beta}{2}\|T - \mathrm{id}\|_{\mu_0}^2$ where T is the OT map from μ_0 to μ_1 *Over generalized geodesics* $(\mu_t^\nu)_{t\in[0,1]}$*:* $\mathcal{V}(\mu_1) \le \mathcal{V}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu_0) \circ T_0, T_1 - T_0\rangle_\nu + \frac{\beta}{2}\|T_1 - T_0\|_\nu^2$ where T_0, T_1 are the OT maps from ν to μ_0 to μ_1
α-convexity Equation (2.6) Lemma 2.2.2 Lemma 2.2.4 Remark 2.2.6	$V\colon \mathbb{R}^d \to \mathbb{R}$ diff'able *Equivalent forms:* $\alpha I \preceq \nabla^2 V$ (α-*strongly* convex if $\alpha > 0$) for all $x_0, x_1 \in \mathbb{R}^d$, $V(x_1) \ge V(x_0) + \langle \nabla V(x_0), x_1 - x_0\rangle + \frac{\alpha}{2}\|x_1 - x_0\|^2$	$\mathcal{V}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ diff'able *Over constant-speed curves:* $\mathcal{V}((\mathrm{id}+h)_\#\mu) \ge \mathcal{V}(\mu) + \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu), h\rangle_\mu + \frac{\alpha}{2}\|h\|_\mu^2$ for all $\mu \in \mathsf{BW}(\mathbb{R}^d)$ and $h \in \mathbb{R}^d \to \mathbb{R}^d$ affine *Over geodesics* $(\mu_t)_{t\in[0,1]}$*:* $\mathcal{V}(\mu_1) \ge \mathcal{V}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu_0), T - \mathrm{id}\rangle_{\mu_0} + \frac{\alpha}{2}\|T - \mathrm{id}\|_{\mu_0}^2$ where T is the OT map from μ_0 to μ_1 *Over generalized geodesics* $(\mu_t^\nu)_{t\in[0,1]}$*:* $\mathcal{V}(\mu_1) \ge \mathcal{V}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu_0) \circ T_0, T_1 - T_0\rangle_\nu + \frac{\alpha}{2}\|T_1 - T_0\|_\nu^2$ where T_0, T_1 are the OT maps from ν to μ_0 to μ_1
gradient Section 2.2.2	diff'able $f\colon \mathbb{R}^d \to \mathbb{R}$ $\nabla f(x)\colon \mathbb{R}^d \to \mathbb{R}^d$	diff'able $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ $\nabla_{\mathsf{BW}} \mathcal{F}(\mu) : x \mapsto (\mathbb{E}_\mu \nabla^2 \delta\mathcal{F})(x - m_\mu) + \mathbb{E}_\mu \nabla \delta\mathcal{F}$
gradient step Equation (2.4)	step size $\eta > 0$, $x_k \in \mathbb{R}^d$, function $f\colon \mathbb{R}^d \to \mathbb{R}$ $x_{k+1} = x_k - \eta \nabla f(x_k)$	step size $\eta > 0$, $\mu_k \in \mathsf{BW}(\mathbb{R}^d)$, functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ $\mu_{k+1} = \exp_{\mu_k}(-\eta \nabla_{\mathsf{BW}} \mathcal{F}(\mu_k))$
proximal operator Equation (2.8) Equation (2.18)	step size $\eta > 0$, $x \in \mathbb{R}^d$, function $f\colon \mathbb{R}^d \to \mathbb{R}$ $\mathrm{prox}_{\eta f}(x) := \arg\min_{y\in\mathbb{R}^d} \left\{ f(y) + \frac{1}{2\eta}\|x-y\|^2 \right\}$	step size $\eta > 0$, $\mu_k \in \mathsf{BW}(\mathbb{R}^d)$, functional $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ $\mathrm{BWJKO}_{\eta\mathcal{F}}(\mu) := \arg\min_{\nu\in\mathsf{BW}(\mathbb{R}^d)} \left\{ \mathcal{F}(\nu) + \frac{1}{2\eta} W_2^2(\mu,\nu) \right\}$

Table 2.1: The dictionary between Euclidean space and Bures–Wasserstein space.

Chapter 3

Algorithms and Guarantees

In this chapter, we will translate classical algorithms and proofs in Euclidean optimization to the BW setting, making extensive use of the Euclidean-BW dictionary laid out in Table 2.1. Our work culminates in the development of a suite of implementable algorithms for Gaussian VI (Algorithms 4 and 5) with state-of-the-art convergence guarantees.

3.1 Algorithms for Euclidean optimization

We begin by considering classical algorithms for optimization of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ over Euclidean space, and present crisp proofs of their convergence guarantees via a unifying analytical tool which we call a *one-step inequality*, which is found in the literature on analyses of stochastic forward-backward methods [34]. For the iterates $\{x_k\}_{k=0}^N$ of our optimization algorithms of interest, this inequality takes the generic form of a guarantee

$$\|x_{k+1} - y\|^2 \leq (1 - \alpha\eta)\,\|x_k - y\|^2 - 2\eta(f(x_{k+1}) - f(y)) + \eta^2\sigma_k^2,$$

where $y \in \mathbb{R}^d$ is arbitrary, $\eta > 0$ denotes a chosen step size, α is a strong convexity parameter, and σ_k^2 denotes a bound on the variance of a gradient estimate at iteration k. As we demonstrate, the one-step inequality is obtainable for our optimization algorithms of interest under appropriate conditions on f. Furthermore, the desirable convergence properties of these algorithms all arise as a consequence of satisfying this inequality. The gift of this analysis is that it allows for translation,

Algorithm 1 Gradient descent (GD) and stochastic gradient descent (SGD)

Require: Step size $\eta > 0$; iteration count N; initial point $x_0 \in \mathbb{R}^d$
  for $k = 0$ **to** $N - 1$ **do**
    if gradient descent **then**
      $v_k \leftarrow \nabla f(x_k)$
    else if stochastic gradient descent **then**
      query unbiased gradient oracle for an estimate $\hat{g}_k$ of $\nabla f(x_k)$
      $v_k \leftarrow \hat{g}_k$
    end if
    $x_{k+1} \leftarrow x_k - \eta v_k$
  end for
  output x_N

via Table 2.1, to the setting of the BW space, hence presenting a direct link between the analysis of Euclidean and BW optimization.

3.1.1 (Stochastic) gradient descent

We revisit the setting of Equation (2.3), where we are given a β-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ (meaning that $-\beta I \preceq \nabla^2 f \preceq \beta I$) and seek to optimize the objective

$$\min_{x \in \mathbb{R}^d} \{f(x)\}. \tag{3.1}$$

In this case, we can apply the method of *gradient descent*, which is detailed in Algorithm 1.

We denote by $x_\star$ a minimizer of the objective function, and we let $\mathscr{F}_k$ denote the σ-algebra generated up to iteration k.

We can study the convergence of gradient descent under a variety of different settings, depending on the properties of f. Additional assumptions on f strengthen the guarantees obtained. Given $\alpha \in \mathbb{R}$, f is α-convex if $\alpha I \preceq \nabla^2 f$. If $\alpha = 0$, then f is said to be convex, and if $\alpha > 0$, then f is said to be α-strongly convex.

For either algorithm (GD or SGD), define the (random) error term e_k as equal to $e_k = v_k - \nabla f(x_k)$, where v_k is defined in Algorithm 1, and denote its expected squared norm by $\sigma_k^2 := \mathbb{E}[\|e_k\|^2 \mid \mathscr{F}_k]$. The expectation is taken over the randomness of the gradient oracle (which is exact in the case of GD).

Since the gradient oracle is unbiased by assumption, we have $\mathbb{E}[e_k \mid \mathscr{F}_k] = 0$, so σ_k^2 rep-

resents the conditional variance of the gradient estimate at iteration k. For GD, we have that $e_k = \nabla f(x_k) - \nabla f(x_k) = 0$, and hence $\sigma_k = 0$.

Our analysis of (S)GD hinges on the following one-step inequality:

Lemma 3.1.1 (One-step inequality for (S)GD)**.** *Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ is α-convex and β-smooth, so that $\alpha I \preceq \nabla^2 f \preceq \beta I$. Let $(x_k)_{k\in\mathbb{N}}$ be the iterates of gradient descent or stochastic gradient descent (Algorithm 1). Let η be such that $0 < \eta \leq \frac{1}{\beta}$. Then, for all $y \in \mathbb{R}^d$, we have that*

$$\mathbb{E}\left\|x_{k+1} - y\right\|^2 \leq (1 - \alpha\eta)\,\mathbb{E}\left\|x_k - y\right\|^2 - 2\eta\,\mathbb{E}[f(x_{k+1}) - f(y)] + 2\eta^2\,\mathbb{E}\sigma_k^2\,. \tag{3.2}$$

The key idea of this proof is to decompose the difference $f(x_{k+1}) - f(y)$ as the sum of two terms:

$$f(x_{k+1}) - f(y) = [f(x_k) - f(y)] + [f(x_{k+1}) - f(x_k)].$$

These individual terms may then be controlled using the α-convexity and β-smoothness of f, respectively.

Proof. Using the above decomposition, we have that

$$\begin{aligned}
\mathbb{E}[f(x_{k+1}) - f(y)] &= \mathbb{E}[f(x_k) - f(y)] + \mathbb{E}[f(x_{k+1}) - f(x_k)] \\
&\leq \mathbb{E}[\langle \nabla f(x_k), x_k - y\rangle] - \frac{\alpha}{2}\mathbb{E}\left\|x_k - y\right\|^2 && \text{(by } \alpha\text{-convexity)} \\
&\quad + \mathbb{E}[\langle \nabla f(x_k), x_{k+1} - x_k\rangle] + \frac{\beta}{2}\mathbb{E}\left\|x_{k+1} - x_k\right\|^2 && \text{(by } \beta\text{-smoothness)} \\
&= -\frac{\alpha}{2}\mathbb{E}\left\|x_k - y\right\|^2 + \mathbb{E}\left\langle \nabla f(x_k), x_{k+1} - y\right\rangle + \frac{\beta}{2}\mathbb{E}\left\|x_{k+1} - x_k\right\|^2 \\
&= -\frac{\alpha}{2}\mathbb{E}\left\|x_k - y\right\|^2 - \mathbb{E}\left\langle e_k, x_{k+1} - y\right\rangle - \frac{1}{\eta}\mathbb{E}\left\langle x_{k+1} - x_k, x_{k+1} - y\right\rangle \\
&\quad + \frac{\beta}{2}\mathbb{E}\left\|x_{k+1} - x_k\right\|^2 && \text{(by Algorithm 1 and defn of } e_k\text{)} \\
&= \frac{1}{2\eta}(1 - \alpha\eta)\mathbb{E}\left\|x_k - y\right\|^2 + \eta\mathbb{E}\left\|e_k\right\|^2 && \text{(rearranging, defn of } e_k\text{, } \mathbb{E}e_k = 0\text{)} \\
&\quad + \frac{1}{2\eta}\mathbb{E}\left[\beta\eta\left\|x_{k+1} - x_k\right\|^2 - \left\|x_k - y\right\|^2 - 2\left\langle x_{k+1} - x_k, x_{k+1} - y\right\rangle\right] \\
&\leq \frac{1}{2\eta}(1 - \alpha\eta)\mathbb{E}\left\|x_k - y\right\|^2 + \eta\mathbb{E}\sigma_k^2 + \frac{1}{2\eta}\mathbb{E}\left\|x_k - y\right\|^2. && (\beta\eta \leq 1\text{, defn of } \sigma_k)
\end{aligned}$$

Multiplying both sides of the inequality by 2η and rearranging, we obtain the desired result. □

The one-step inequality is a powerful analytical tool which unifies convergence analysis. Through this single inequality, we can already deduce the key convergence properties of (S)GD under different sets of assumptions on f, as we do in the sequel.

3.1.2 Convergence guarantees for (S)GD

Guarantees for GD

By setting $\sigma_k = 0$ in Lemma 3.1.1, we obtain a number of convergence guarantees for GD. We first consider the case where f is also convex.

Theorem 3.1.2 (Convex case, GD). *Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth and that $0 < \eta \leq \frac{1}{\beta}$. Then for any $N \in \mathbb{N}$, we have that*

$$f(x_N) - f(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{2N\eta}.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta\|x_0 - x_\star\|^2}{\varepsilon^2}$, we obtain the guarantee

$$f(x_N) - f(x_\star) \leq \varepsilon^2.$$

Proof. Since f is convex, Equation (3.2) holds with the choice $\alpha = 0$. Furthermore, we may drop the expectations, since in this case the algorithm is deterministic. Applying the resulting inequality with $y = x_k$, we obtain that

$$f(x_{k+1}) - f(x_k) \leq -\frac{\|x_{k+1} - x_k\|^2}{2\eta} \leq 0. \tag{3.3}$$

On the other hand, choosing $y = x_\star$, we have

$$f(x_{k+1}) - f(x_\star) \leq \frac{\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2}{2\eta}.$$

Telescoping this inequality, we deduce that

$$f(x_N) - f(x_\star) \le \frac{1}{N}\sum_{k=1}^{N}[f(x_k) - f(x_\star)] \le \frac{1}{2\eta N}\sum_{k=0}^{N-1}[\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2] \le \frac{\|x_0 - x_\star\|^2}{2\eta N},$$

where the first inequality holds by Inequality 3.3. Hence, with the choice

$$\eta = \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{\beta\|x_0 - x_\star\|^2}{\varepsilon^2},$$

we obtain the guarantee $f(x_N) - f(x_\star) \le \varepsilon^2$, proving our desired result. □

When the objective is α-strongly convex, we in fact obtain a *linear* rate of convergence:

Theorem 3.1.3 (Strongly convex case, GD)**.** *Suppose that f is α-strongly convex and β-smooth, and that $0 < \eta \le \frac{1}{\beta}$. Then,*

$$\|x_N - x_\star\|^2 \le \exp(-N\alpha\eta)\|x_0 - x_\star\|^2.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta}{\alpha}\log\frac{\alpha\|x_0 - x_\star\|^2}{\varepsilon^2}$, we obtain the guarantees

$$\alpha\|x_N - x_\star\|^2 \le \varepsilon^2, \quad \textit{and} \quad f(x_{2N}) - f(x_\star) \le \varepsilon^2.$$

Proof. Since $f(x_\star) \le f(x_{k+1})$ as $x_\star$ is the minimizer of f, we may iterate Equation (3.2) to obtain

$$\|x_N - x_\star\|^2 \le \exp(-N\alpha\eta)\|x_0 - x_\star\|^2.$$

Hence, with the choice

$$\eta = \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{1}{\alpha\eta}\log\frac{\alpha\|x_0 - x_\star\|^2}{\varepsilon^2} \asymp \frac{\beta}{\alpha}\log\frac{\alpha\|x_0 - x_\star\|^2}{\varepsilon^2},$$

we obtain the guarantee $\alpha\|x_N - x_\star\|^2 \le \varepsilon^2$.

Now, for the guarantee in objective gap, we "reinitialize" the algorithm at x_N and apply the result of Theorem 3.1.2. This argument is inspired by Durmus, Majewski, and Miasojedow [32]. With the same choice of N and η and assuming ε is sufficiently small, we can apply Theorem 3.1.2

to obtain the guarantee

$$f(x_{2N}) - f(x_\star) \leq \frac{\|x_N - x_\star\|^2}{2\eta N} \leq \frac{\varepsilon^2}{2\alpha\eta N} \lesssim \frac{\varepsilon^2}{\log \frac{\alpha\|x_0 - x_\star\|^2}{\varepsilon^2}} \lesssim \varepsilon^2 \,,$$

proving our desired result. □

When f is non-convex, we cannot guarantee convergence to the global minimizer $x_\star$ in general. However, from the one-step inequality, we can obtain a *stationary point* guarantee which states that the norm of $\nabla f(x_N)$ grows small with the number of iterations. This is the canonical "convergence" metric for optimization of non-convex functions [65].

Theorem 3.1.4 (Non-convex case, GD)**.** *Suppose that f is β-smooth and that $0 < \eta \leq \frac{1}{\beta}$. Let $\Delta := f(x_0) - f(x_\star)$. Then,*

$$\min_{k \in \{0,\ldots,N-1\}} \|\nabla f(x_k)\|^2 \leq \frac{2\Delta}{\eta N} \,.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta\Delta}{\varepsilon^2}$, we obtain the guarantee

$$\min_{k \in \{0,\ldots,N-1\}} \|\nabla f(x_k)\|^2 \leq \varepsilon^2 \,. \tag{3.4}$$

Proof. Applying Lemma 3.1.1 with the choice $\alpha = -\beta$ and $y = x_k$, we obtain that

$$\|x_{k+1} - x_k\|^2 \leq -2\eta[f(x_{k+1}) - f(x_k)].$$

Telescoping this inequality, we find that

$$\begin{aligned}
\min_{k \in \{0,\ldots,N-1\}} \|x_{k+1} - x_k\|^2 &\leq \frac{1}{N} \sum_{k=0}^{N-1} \|x_{k+1} - x_k\|^2 \\
&\leq -\frac{2\eta}{N} \sum_{k=0}^{N-1} [f(x_{k+1}) - f(x_k)] \\
&= -\frac{2\eta}{N} [f(x_N) - f(x_0)] \\
&\leq \frac{2\eta\Delta}{N}.
\end{aligned}$$

Hence, we conclude that

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla f(x_k)\|^2 = \min_{k\in\{0,\dots,N-1\}} \frac{1}{\eta^2}\|x_{k+1}-x_k\|^2 \le \frac{2\Delta}{N\eta}.$$

Finally, taking $\eta = \frac{1}{\beta}$ and $N \ge \frac{2\beta\Delta}{\varepsilon^2}$, we obtain that

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla f(x_k)\|^2 \le \varepsilon^2,$$

as desired. □

3.1.3 Guarantees for SGD

In the setting of SGD, the additional noise term σ_k^2 arising from the variance of the gradient estimate significantly affects the convergence analysis. We must assume some bound on the noise, say $\sigma_k^2 \le \sigma^2$, in order to have any hope of bounding the distance to the minimizer or the objective gap. This bound need not be uniform in general, but to simplify the presentation we consider the case where σ^2 uniformly bounds σ_k^2 for all k. Given this noise bound, we can obtain guarantees *in expectation* for the iterates. First, we consider the convex case.

Theorem 3.1.5 (Convex case, SGD)**.** *Suppose that f is convex and β-smooth and that $0 < \eta \le \frac{1}{\beta}$. If $\sigma_k^2 \le \sigma^2$ for all k, then*

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} f(x_k)\right] - f(x_\star) \le \frac{2\|x_0 - x_\star\|^2}{N\eta} + \eta\sigma^2.$$

In particular, with

$$\eta \asymp \frac{\varepsilon^2}{\sigma^2} \wedge \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{\|x_0 - x_\star\|^2}{\varepsilon^2}\left(\frac{\sigma^2}{\varepsilon^2} \vee \beta\right),$$

we obtain the guarantee

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} f(x_k)\right] - f(x_\star) \le \varepsilon^2.$$

Proof. Since f is convex, Equation (3.2) holds with the choice $\alpha = 0$. Assuming also that we have

the uniform bound $\sigma_k^2 \le \sigma^2$, we may apply the resulting inequality with $y = x_\star$ to obtain that

$$\mathbb{E}[f(x_{k+1})] - f(x_\star) \le \frac{\mathbb{E}[\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2] + 2\eta^2\sigma^2}{2\eta}.$$

Telescoping this inequality, we deduce that

$$\begin{aligned}
\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} f(x_k)\right] - f(x_\star) &\le \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}[f(x_k) - f(x_\star)] \\
&\le \frac{1}{2\eta N}\sum_{k=0}^{N-1} [\mathbb{E}[\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2] + 2\eta^2\sigma^2] \\
&\le \frac{\|x_0 - x_\star\|^2}{2\eta N} + \eta\sigma^2.
\end{aligned}$$

Hence, with the choice

$$\eta = \frac{\varepsilon^2}{2\sigma^2} \wedge \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{\|x_0 - x_\star\|^2}{\eta\varepsilon^2} \asymp \frac{\|x_0 - x_\star\|^2}{\varepsilon^2}\left(\frac{\sigma^2}{\varepsilon^2} \vee \beta\right),$$

we obtain the guarantee

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} f(x_k)\right] - f(x_\star) \le \varepsilon^2,$$

proving our desired result. □

Similarly to GD, we obtain faster rates when f is α-strongly convex:

Theorem 3.1.6 (Strongly convex case, SGD)**.** *Suppose that f is α-strongly convex and β-smooth, and that $0 < \eta \le \frac{1}{\beta}$. If $\sigma_k^2 \le \sigma^2$ for all k, then*

$$\mathbb{E}\,\|x_N - x_\star\|^2 \le \exp(-N\alpha\eta)\,\|x_0 - x_\star\|^2 + 2\frac{\eta\sigma^2}{\alpha}.$$

In particular, with

$$\eta \asymp \frac{\varepsilon^2}{\sigma^2} \wedge \frac{1}{\beta}, \quad \textit{and} \quad N \gtrsim \frac{\sigma^2}{\alpha\varepsilon^2}\log\frac{\alpha\,\|x_0 - x_\star\|^2}{\varepsilon^2},$$

we obtain the guarantees

$$\alpha \mathbb{E} \left\| x_N - x_\star \right\|^2 \leq \varepsilon^2, \quad \textit{and} \quad \mathbb{E} \left[\min_{k \in \{1,\dots,2N\}} f(x_k) \right] - f(x_\star) \leq \varepsilon^2 .$$

Proof. Since $f(x_\star) \leq f(x_{k+1})$ as $x_\star$ is the minimizer of f, we may iterate Equation (3.2) to obtain

$$\mathbb{E} \left\| x_N - x_\star \right\|^2 \leq \exp(-N\alpha\eta) \left\| x_0 - x_\star \right\|^2 + 2\frac{\eta\sigma^2}{\alpha} .$$

Hence, with the choice

$$\eta = \frac{\varepsilon^2}{4\sigma^2} \wedge \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{1}{\alpha\eta} \log \frac{\alpha \left\| x_0 - x_\star \right\|^2}{\varepsilon^2} \asymp \frac{\sigma^2}{\alpha\varepsilon^2} \log \frac{\alpha \left\| x_0 - x_\star \right\|^2}{\varepsilon^2} ,$$

we obtain the guarantee $\alpha \mathbb{E} \left\| x_N - x_\star \right\|^2 \leq \varepsilon^2$.

Now, for the guarantee in objective gap, we "reinitialize" the algorithm at x_N and apply the result of Theorem 3.1.5. With the same choice of N and η and assuming ε is sufficiently small, we can apply Theorem 3.1.5 to obtain the guarantee

$$\begin{aligned} \mathbb{E} \left[\min_{k \in \{1,\dots,2N\}} f(x_k) \right] - f(x_\star) &\leq \mathbb{E} \left[\min_{k \in \{N+1,\dots,2N\}} f(x_k) \right] - f(x_\star) \\ &\leq \frac{\mathbb{E} \left\| x_N - x_\star \right\|^2}{2\eta N} + \eta\sigma^2 \\ &\lesssim \varepsilon^2 , \end{aligned}$$

proving our desired result. □

Gradient descent and its stochastic variant are only able to handle the case when f is β-smooth. Intuitively, if f is not β-smooth, then in some regions of space the gradient may fluctuate too wildly for a gradient evaluation at a single point to give any information about the objective function in a non-vanishing neighborhood. However, as we alluded to in Section 2.1, the *(stochastic) proximal gradient algorithm* provides a remedy in the case where the objective has a *composite* structure. We detail the algorithm and its guarantees below.

Algorithm 2 Proximal gradient (PG) and stochastic proximal gradient (SPG)

Require: Step size $\eta > 0$; iteration count N; initial point $x_0 \in \mathbb{R}^d$
  for $k = 0$ **to** $N - 1$ **do**
    if proximal gradient **then**
      $v_k \leftarrow \nabla V(x_k)$
    else if stochastic proximal gradient **then**
      query unbiased gradient oracle for an estimate $\hat{g}_k$ of $\nabla V(x_k)$
      $v_k \leftarrow \hat{g}_k$
    end if
    $x_{k+\frac{1}{2}} \leftarrow x_k - \eta v_k$
    $x_{k+1} \leftarrow \mathrm{prox}_{\eta H}(x_{k+\frac{1}{2}})$
  end for
  output x_N

3.1.4 (Stochastic) proximal gradient

We revisit the setting of Equation (2.7), where we are given a *composite* objective function $f = V + H$, where $V \colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth and $H \colon \mathbb{R}^d \to \mathbb{R}$ is convex but potentially non-smooth:

$$\min_{x \in \mathbb{R}^d} \{f(x)\} = \min_{x \in \mathbb{R}^d} \{V(x) + H(x)\}. \tag{3.5}$$

Suppose we have access to the *proximal operator* of ηH, defined in Equation (2.8) by

$$\mathrm{prox}_{\eta H}(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ H(y) + \frac{1}{2\eta} \|x - y\|^2 \right\}.$$

Given access to this operator along with a gradient oracle for V, we can apply the *proximal gradient* algorithm, which is detailed in Algorithm 2.

Analogously to the setting of (S)GD, we denote by $x_\star$ a minimizer of the objective function f, and we let $\mathscr{F}_k$ denote the σ-algebra generated up to iteration k. We can study the convergence of gradient descent under a variety of different settings, this time depending on the properties of V.

For either algorithm (PG and SPG), define the (random) error term e_k as equal to $e_k = v_k - \nabla V(x_k)$, where v_k is defined in Algorithm 2, and denote its expected squared norm by $\sigma_k^2 := \mathbb{E}[\|e_k\|^2 \mid \mathscr{F}_k]$. The expectation is taken over the randomness of the gradient oracle for V (which is exact in the case of PG). As in the setting of (S)GD, we have $\mathbb{E}[e_k \mid \mathscr{F}_k] = 0$, so σ_k^2 represents the conditional variance of the gradient estimate at iteration k. For PG, we have that $e_k = \nabla V(x_k) -$

$\nabla V(x_k) = 0$, and hence $\sigma_k = 0$.

Once again, our analysis of (S)PG hinges on proving a one-step inequality:

Lemma 3.1.7 (One-step inequality for (S)PG)**.** *Suppose that $f = V + H$ where $V\colon \mathbb{R}^d \to \mathbb{R}$ is α-convex and β-smooth, so that $\alpha I \preceq \nabla^2 V \preceq \beta I$, and where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex (but not necessarily smooth). Let $(x_k)_{k\in\mathbb{N}}$ be the iterates of proximal gradient or stochastic proximal gradient (Algorithm 1). Let η be such that $0 < \eta \leq \frac{1}{\beta}$. Then, for all $y \in \mathbb{R}^d$, we have that*

$$\mathbb{E}\,\|x_{k+1} - y\|^2 \leq (1 - \alpha\eta)\,\mathbb{E}\,\|x_k - y\|^2 - 2\eta\,\mathbb{E}[f(x_{k+1}) - f(y)] + 2\eta^2\,\mathbb{E}\sigma_k^2\,. \tag{3.6}$$

The key idea of this proof is analogous to that of Lemma 3.1.1, except we now decompose the difference $f(x_{k+1}) - f(y)$ as the sum of *three* terms:

$$f(x_{k+1}) - f(y) = [V(x_k) - V(y)] + [V(x_{k+1}) - V(x_k)] + [H(x_{k+1}) - H(y)].$$

These individual terms may then be controlled using the α-convexity and β-smoothness of V and the convexity of H, respectively.

Proof. First, we consider the term $V(x_{k+1}) - V(y)$. We have that

$$\begin{aligned}
\mathbb{E}[V(x_{k+1}) - V(y)] &= \mathbb{E}[V(x_k) - V(y)] + \mathbb{E}[V(x_{k+1}) - V(x_k)] \\
&= \mathbb{E}[\langle \nabla V(x_k), x_k - y\rangle] - \frac{\alpha}{2}\mathbb{E}\,\|x_k - y\|^2 && \text{(by } \alpha\text{-convexity)} \\
&\quad + \mathbb{E}[\langle \nabla V(x_k), x_{k+1} - x_k\rangle] + \frac{\beta}{2}\mathbb{E}\,\|x_{k+1} - x_k\|^2 && \text{(by } \beta\text{-smoothness)} \\
&= -\frac{\alpha}{2}\mathbb{E}\,\|x_k - y\|^2 + \mathbb{E}\,\langle \nabla V(x_k), x_{k+1} - y\rangle + \frac{\beta}{2}\mathbb{E}\,\|x_{k+1} - x_k\|^2 \\
&= -\frac{\alpha}{2}\mathbb{E}\,\|x_k - y\|^2 - \mathbb{E}\,\langle e_k, x_{k+1} - y\rangle - \frac{1}{\eta}\mathbb{E}\left\langle x_{k+\frac{1}{2}} - x_k, x_{k+1} - y\right\rangle \\
&\quad + \frac{\beta}{2}\mathbb{E}\,\|x_{k+1} - x_k\|^2 && \text{(by Algorithm 2 and defn of } e_k\text{)} \\
&= \frac{1}{2\eta}(1 - \alpha\eta)\mathbb{E}\,\|x_k - y\|^2 - \mathbb{E}\,\langle e_k, x_{k+1} - y\rangle && \text{(rearranging)} \\
&\quad + \frac{1}{2\eta}\mathbb{E}\left[\beta\eta\,\|x_{k+1} - x_k\|^2 - \|x_k - y\|^2 - 2\left\langle x_{k+\frac{1}{2}} - x_k, x_{k+1} - y\right\rangle\right] \\
&\leq \frac{1}{2\eta}(1 - \alpha\eta)\mathbb{E}\,\|x_k - y\|^2 - \mathbb{E}\,\langle e_k, x_{k+1} - y\rangle
\end{aligned}$$

$$+\frac{1}{2\eta}\mathbb{E}\left[\|x_{k+1}-x_k\|^2-\|x_k-y\|^2-2\left\langle x_{k+\frac{1}{2}}-x_k,x_{k+1}-y\right\rangle\right].$$

($\beta\eta \le 1$, defn of σ_k)

Now we bound the difference in H. We have that

$$\begin{aligned}
\mathbb{E}[H(x_{k+1})-H(y)] &\le \mathbb{E}\left\langle \nabla H(x_{k+1}), x_{k+1}-y\right\rangle && \text{(convexity of } H\text{)}\\
&= -\frac{1}{\eta}\mathbb{E}\left\langle x_{k+1}-x_{k+\frac{1}{2}}, x_{k+1}-y\right\rangle && \text{(by proximal step in Algorithm 2)}\\
&= \frac{1}{2\eta}\mathbb{E}\left[\left\|x_{k+\frac{1}{2}}-y\right\|^2-\left\|x_{k+1}-x_{k+\frac{1}{2}}\right\|^2-\|x_{k+1}-y\|^2\right].
\end{aligned}$$

Now, we sum the above inequalities to obtain our desired bound on $\mathbb{E}[f(x_{k+1})-f(y)]$. We obtain that

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})-f(y)] &= \mathbb{E}[V(x_{k+1})-V(y)]+\mathbb{E}[H(x_{k+1})-H(y)]\\
&\le \frac{1}{2\eta}\mathbb{E}[(1-\alpha\eta)\|x_k-y\|^2-\|x_{k+1}-y\|^2]\\
&\quad+\frac{1}{2\eta}\mathbb{E}[\|x_{k+1}-x_k\|^2-\|x_k-y\|^2+\left\|x_{k+\frac{1}{2}}-y\right\|^2-\left\|x_{k+1}-x_{k+\frac{1}{2}}\right\|^2]\\
&\quad-\frac{1}{\eta}\mathbb{E}\left\langle x_{k+\frac{1}{2}}-x_k, x_{k+1}-y\right\rangle-\mathbb{E}\left\langle e_k, x_{k+1}-y\right\rangle\\
&= \frac{1}{2\eta}\mathbb{E}[(1-\alpha\eta)\|x_k-y\|^2-\|x_{k+1}-y\|^2]-\mathbb{E}\left\langle e_k, x_{k+1}-y\right\rangle.
\end{aligned}$$

Finally, it remains to bound the error term on the last line. To do this, we define auxiliary variables $\overline{x}_{k+\frac{1}{2}}, \overline{x}_{k+1}$, which keep track of the "ideal" iterates generated using a noiseless gradient oracle:

$$\overline{x}_{k+\frac{1}{2}} := x_k-\eta\nabla V(x_k), \qquad \overline{x}_{k+1} := \mathrm{prox}_{\eta H}(\overline{x}_{k+\frac{1}{2}}).$$

A classical fact property of the proximal operator is that it is *nonexpansive* [71], meaning that for any $u, v \in \mathbb{R}^d$, we have

$$\|u-v\|^2 \ge \left\langle u-v, \mathrm{prox}_{\eta H}(u)-\mathrm{prox}_{\eta H}(v)\right\rangle.$$

Hence, combining this with the fact that

$$\eta e_k = \eta(v_k - \nabla V(x_k)) = \overline{x}_{k+\frac{1}{2}} - x_{k+\frac{1}{2}},$$

we deduce that

$$\eta \left\|e_k\right\|^2 = \frac{1}{\eta}\left\|\overline{x}_{k+\frac{1}{2}} - x_{k+\frac{1}{2}}\right\|^2 \geq \frac{1}{\eta}\left\langle x_{k+\frac{1}{2}} - \overline{x}_{k+\frac{1}{2}}, x_{k+1} - \overline{x}_{k+1}\right\rangle = -\left\langle e_k, x_{k+1} - \overline{x}_{k+1}\right\rangle.$$

Now, this implies that

$$\begin{aligned}
\mathbb{E}\left\langle e_k, x_{k+1} - y\right\rangle &= \mathbb{E}\left[\mathbb{E}\left[\left\langle e_k, x_{k+1} - y\right\rangle \mid x_k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle e_k, x_{k+1} - \overline{x}_{k+1}\right\rangle \mid x_k\right]\right] && (\text{since } e_k \perp\!\!\!\perp (\overline{x}_{k+1}, y) \mid x_k) \\
&= \mathbb{E}\left\langle e_k, x_{k+1} - \overline{x}_{k+1}\right\rangle \\
&\geq -\eta\mathbb{E}\left\|e_k\right\|^2 \\
&\geq -\eta\sigma_k^2.
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}
\mathbb{E}[f(x_{k+1}) - f(y)] &\leq \frac{1}{2\eta}\mathbb{E}[(1-\alpha\eta)\left\|x_k - y\right\|^2 - \left\|x_{k+1} - y\right\|^2] - \mathbb{E}\left\langle e_k, x_{k+1} - y\right\rangle \\
&\leq \frac{1}{2\eta}\mathbb{E}[(1-\alpha\eta)\left\|x_k - y\right\|^2 - \left\|x_{k+1} - y\right\|^2] + \eta\sigma^2,
\end{aligned}$$

and rearranging proves the desired result. □

Having shown a one-step inequality for f along the iterates of Algorithm 2, we automatically obtain a variety of convergence guarantees depending on the properties of V. Since nothing beyond the one-step inequality was assumed for the proofs of Theorem 3.1.2 and Theorem 3.1.3 (for GD), Theorem 3.1.5 and Theorem 3.1.6 (for SGD), we immediately attain analogous convergence guarantees for (S)PG with no additional work. And with only a slight change in the analysis, we will also be able to prove a stationary point guarantee in terms of bounding the squared norm of the gradient for PG, as we do in the sequel.

3.1.5 Convergence guarantees for (S)PG

Guarantees for PG

By setting $\sigma_k = 0$ in Lemma 3.1.7, we obtain convergence guarantees for proximal gradient analogous to those for gradient descent.

Lemma 3.1.8 (Convex case, PG). *Suppose that $f = V + H$ where $V\colon \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, and where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex (but not necessarily smooth). Suppose that $0 < \eta \leq \frac{1}{\beta}$. Then for any $N \in \mathbb{N}$, we have that*

$$f(x_N) - f(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{2N\eta}.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta\|x_0 - x_\star\|^2}{\varepsilon^2}$, we obtain the guarantee

$$f(x_N) - f(x_\star) \leq \varepsilon^2.$$

Proof. Given that the one step inequality Lemma 3.1.7 holds with $\alpha = 0$, the proof is entirely identical to the one for gradient descent (Theorem 3.1.2). □

When V is α-strongly convex, we once again obtain a linear rate of convergence:

Theorem 3.1.9 (Strongly convex case, PG). *Suppose that $f = V + H$ where $V\colon \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex and β-smooth, and where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex (but not necessarily smooth). Suppose that $0 < \eta \leq \frac{1}{\beta}$. Then for any $N \in \mathbb{N}$, we have that*

$$\|x_N - x_\star\|^2 \leq \exp(-N\alpha\eta)\,\|x_0 - x_\star\|^2.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta}{\alpha}\log\frac{\|x_0 - x_\star\|}{\varepsilon}$, we obtain the guarantees

$$\alpha\,\|x_N - x_\star\|^2 \leq \varepsilon^2, \quad \text{and} \quad f(x_{2N}) - f(x_\star) \leq \varepsilon^2.$$

Proof. Given that the one step inequality Lemma 3.1.7 holds with convexity parameter α, the proof is entirely identical to the one for gradient descent (Theorem 3.1.3). □

When V is non-convex, we once again cannot guarantee convergence to the global minimizer $x_\star$. But we can hope to still obtain a "stationary point" guarantee: if the step size η is sufficiently small, Algorithm 2 ought to converge to *something*, so intuitively $\|x_{k+1} - x_k\|^2$ should converge to 0 as the number of iterations k is sent to infinity. However, unlike in the case of gradient descent where $x_{k+1} - x_k$ is directly interpretable in terms of $\nabla f(x_k)$, such a relation is no longer true for Algorithm 2 due to the additional proximal step. In general, if H is non-smooth, we cannot hope to bound $\|\nabla f(x_k)\|$ in terms of $\|x_{k+1} - x_k\|$ in Algorithm 2. However, if we additionally assume that H is smooth, then we can indeed obtain a stationary point guarantee for PG from the one-step inequality.

Theorem 3.1.10 (Non-convex case, PG). *Suppose that $f = V + H$ where $V\colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth, and where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex and also β-smooth. Suppose that $0 < \eta \leq \frac{1}{\beta}$, and define $\Delta := f(x_0) - f(x_\star)$. Then for any $N \in \mathbb{N}$, we have that*

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla f(x_k)\|^2 \leq \frac{8\Delta}{\eta N}\,.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta\Delta}{\varepsilon^2}$, we obtain the guarantee

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla f(x_k)\|^2 \leq \varepsilon^2\,. \tag{3.7}$$

Proof. We proceed identically to the analysis in Theorem 3.1.4, obtaining the inequality

$$\min_{k\in\{0,\dots,N-1\}} \frac{1}{\eta^2}\|x_{k+1} - x_k\|^2 \leq \frac{2\Delta}{N\eta}\,.$$

Now, we have that

$$\begin{aligned}
\frac{1}{2}\|\nabla f(x_k)\|^2 &= \frac{1}{2}\|\nabla V(x_k) + \nabla H(x_k)\|^2 \\
&\leq \|\nabla V(x_k) + \nabla H(x_{k+1})\|^2 + \|\nabla H(x_{k+1}) - \nabla H(x_k)\|^2 \\
&\qquad\qquad \text{(by triangle inequality and Cauchy-Schwarz)} \\
&\leq \|\nabla V(x_k) + \nabla H(x_{k+1})\|^2 + \beta^2\|x_{k+1} - x_k\|^2 \\
&\qquad\qquad \text{(since } \beta\text{-smoothness implies } \nabla H \text{ is } \beta\text{-Lipschitz)}
\end{aligned}$$

$$= \left(\frac{1}{\eta^2} + \beta^2\right) \|x_{k+1} - x_k\|^2 . \quad \text{(since } x_k - x_{k+1} = \eta(\nabla V(x_k) + \nabla H(x_{k+1})) \text{ for PG)}$$

Hence, we conclude that

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla f(x_k)\|^2 \le \min_{k\in\{0,\dots,N-1\}} \frac{4}{\eta^2} \|x_{k+1} - x_k\|^2 \le \frac{8\Delta}{N\eta} .$$

Finally, taking $\eta = \frac{1}{\beta}$ and $N \ge \frac{8\beta\Delta}{\varepsilon^2}$, we obtain that

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla f(x_k)\|^2 \le \varepsilon^2 ,$$

as desired. □

Guarantees for SPG

Just as in the setting of SGD, given some bound on the noise, say $\sigma_k^2 \le \sigma^2$, we can obtain guarantees *in expectation* for the iterates. First, we consider the convex case.

Theorem 3.1.11 (Convex case, SPG). *Suppose that $f = V + H$ where $V\colon \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, and where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex (but not necessarily smooth). Suppose that $0 < \eta \le \frac{1}{\beta}$. If $\sigma_k^2 \le \sigma^2$ for all k, then for any $N \in \mathbb{N}$, we have that*

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} f(x_k)\right] - f(x_\star) \le \frac{2\|x_0 - x_\star\|^2}{N\eta} + 2\eta\sigma^2.$$

In particular, with

$$\eta \asymp \frac{\varepsilon^2}{\sigma^2} \wedge \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{\|x_0 - x_\star\|^2}{\varepsilon^2}\left(\frac{\sigma^2}{\varepsilon^2} \vee \beta\right),$$

we obtain the guarantee

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} f(x_k)\right] - f(x_\star) \le \varepsilon^2.$$

Proof. Given that the one step inequality Lemma 3.1.7 holds with $\alpha = 0$, the proof is entirely identical to the one for stochastic gradient descent (Theorem 3.1.5). □

Similarly, we obtain faster rates when V is α-strongly convex:

Theorem 3.1.12 (Strongly convex case, SPG). *Suppose that $f = V + H$ where $V\colon \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex and β-smooth, and where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex (but not necessarily smooth). Suppose that $0 < \eta \leq \frac{1}{\beta}$. If $\sigma_k^2 \leq \sigma^2$ for all k, then*

$$\mathbb{E}\left\|x_N - x_\star\right\|^2 \leq \exp\left(-N\alpha\eta\right)\left\|x_0 - x_\star\right\|^2 + 2\frac{\eta\sigma^2}{\alpha}.$$

In particular, with

$$\eta \asymp \frac{\varepsilon^2}{\sigma^2} \wedge \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{\sigma^2}{\alpha}\log\frac{\alpha\left\|x_0 - x_\star\right\|}{\varepsilon},$$

we obtain the guarantees

$$\alpha\,\mathbb{E}\left\|x_N - x_\star\right\|^2 \leq \varepsilon^2, \quad \text{and} \quad \mathbb{E}\left[\min_{k\in\{1,\dots,2N\}} f(x_k)\right] - f(x_\star) \leq \varepsilon^2.$$

Proof. Given that the one step inequality Lemma 3.1.7 holds with convexity parameter α, the proof is entirely identical to the one for gradient descent (Theorem 3.1.6). □

Finally, we consider an extension of Algorithms 1 and 2 to the setting of *variance reduction*, as introduced in Section 2.1. In this case, f (resp. V) admits a representation as the average of many component functions $f_i\colon \mathbb{R}^d \to \mathbb{R}$ (resp. V_i), which are each strongly convex and smooth. If m is large, the update equations of gradient descent (2.4) and proximal gradient (2.9) may be computationally expensive, as computing the gradient of f requires evaluating ∇f_i for each $i = 1, \dots, m$. On the other hand, taking the unbiased gradient estimate to equal ∇f_i with $i \in \mathrm{Unif}[m]$ results in cheap iteration complexity but high variance of the gradient estimate, resulting in a slow rate of convergence. The approach of Stochastic Variance Reduced Gradient (SVRG) [42] and its proximal analogue Prox-SVRG [95] performs a slight modification of Algorithms 1 and 2 using a *centering sequence* that is updated infrequently and yet reduces variance of the gradient estimate, hence leveraging the best of both worlds and obtaining an improved complexity guarantee. To simplify the presentation, we will restrict our attention to Prox-SVRG. SVRG can be thought of as a special case of Prox-SVRG by taking $H \equiv 0$.

3.1.6 Prox-SVRG

We once again revisit the setting of Equation (2.7), where we are given a *composite* objective function $f = V + H$, where $V\colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth and $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex but potentially non-smooth:

$$\min_{x\in\mathbb{R}^d}\{f(x)\} = \min_{x\in\mathbb{R}^d}\{V(x) + H(x)\}. \tag{3.8}$$

In addition, we assume that V is α-strongly convex and can be written as the mean of m convex component functions $V_i\colon \mathbb{R}^d \to \mathbb{R}$, so that

$$V(x) = \frac{1}{m}\sum_{i=1}^{m} V_i(x).$$

Given access to the proximal operator for H along with gradient oracles for each component function V_i, we can apply the *Proximal Stochastic Variance Reduced Gradient* (Prox-SVRG) algorithm [95], which is detailed in Algorithm 3. The centering procedure gives rise to an "inner-outer" loop structure of the algorithm. In this paradigm, a centering sequence is computed in the outer loop and its gradient is evaluated exactly, requiring m calls to the gradient for each iteration of the outer loop. This computed gradient is then reshared throughout the iterations of the inner loop, where at each iteration a cheap unbiased gradient estimate is computed. Overall, the total number of calls to a gradient oracle for a component function V_i is equal to $(2N + m)M$, where N is the number of iterations of the inner loop and M is the number of iterations of the outer loop.

3.1.7 Convergence guarantees for Prox-SVRG

Once again, we denote by $x_\star$ a minimizer of the objective function f, and we let $\mathscr{F}_k^{(j)}$ denote the σ-algebra generated up to iteration k of the inner loop on iteration j of the outer loop. We also define the (random) error term $e_k^{(j)}$ as equal to $e_k^{(j)} = v_k^{(j)} - \nabla V(x_k^{(j)})$, where $v_k^{(j)}$ is defined in Algorithm 3, and denote its expected squared norm by $\sigma_{k,j}^2 := \mathbb{E}[\|e_k^{(j)}\|^2 \mid \mathscr{F}_k^{(j)}]$. The expectation is taken over the randomness of the chosen indices $i \in \mathrm{Unif}[m]$. Our gradient estimate is unbiased by construction, so we have $\mathbb{E}[e_k^{(j)} \mid \mathscr{F}_k^{(j)}] = 0$, so $\sigma_{k,j}^2$ represents the conditional variance of the gradient estimate at inner iteration k of outer iteration j. For PG, we have that $e_k = \nabla V(x_k) - \nabla V(x_k) = 0$, and hence $\sigma_k = 0$.

Algorithm 3 Proximal Stochastic Variance Reduced Gradient (Prox-SVRG)

Require: Step size $\eta > 0$; inner loop iteration count N; outer loop iteration count M; initial point $x_0 \in \mathbb{R}^d$

$x_0^{(0)} \leftarrow x_0$
for $j = 0$ **to** $M - 1$ **do**
  compute $\nabla V(x_0^{(j)})$
  for $k = 0$ **to** $N - 1$ **do**
    randomly draw $i \sim \text{Unif}[m]$
    $v_k^{(j)} \leftarrow \nabla V_i(x_k^{(j)}) - \nabla V_i(x_0^{(j)}) + \nabla V(x_0^{(j)})$
    $x_{k+\frac{1}{2}}^{(j)} \leftarrow x_k^{(j)} - \eta v_k^{(j)}$
    $x_{k+1}^{(j)} \leftarrow \text{prox}_{\eta H}(x_{k+\frac{1}{2}}^{(j)})$
  end for
  randomly draw $\ell \sim \text{Unif}[N]$
  $x_0^{(j+1)} \leftarrow x_\ell^{(j)}$
end for
output $x_0^{(M)}$

We can once again prove a one-step inequality for Algorithm 3:

Lemma 3.1.13 (One-step inequality for Prox-SVRG)**.** *Suppose that $f = V + H$ where $V \colon \mathbb{R}^d \to \mathbb{R}$ is α-convex and β-smooth, so that $\alpha I \preceq \nabla^2 V \preceq \beta I$, and where $H \colon \mathbb{R}^d \to \mathbb{R}$ is convex (but not necessarily smooth). Let $(x_k^{(j)})_{j \in \mathbb{N}, k \in \mathbb{N}}$ be the iterates of Prox-SVRG (Algorithm 1). Let η be such that $0 < \eta \leq \frac{1}{\beta}$. Then, for all $y \in \mathbb{R}^d$, we have that*

$$\mathbb{E}\|x_{k+1}^{(j)} - y\|^2 \leq (1 - \alpha\eta)\,\mathbb{E}\|x_k^{(j)} - y\|^2 - 2\eta\,\mathbb{E}[f(x_{k+1}^{(j)}) - f(y)] + 2\eta^2\,\mathbb{E}\sigma_{k,j}^2. \tag{3.9}$$

Proof. For notational convenience, we will drop the dependence on j. In this setting, note that Prox-SVRG is simply an instantiation of stochastic proximal gradient (Algorithm 2) with unbiased stochastic gradient estimator given by

$$\hat{g}_k = \nabla V_i(x_k) - \nabla V_i(x_0) + \nabla V(x_0), \quad i \sim \text{Unif}[m].$$

Hence, the desired result follows as a direct consequence of Lemma 3.1.7; the only difference is the addition of a superscript $x_k^{(j)}$. □

As we just noted, Prox-SVRG is simply a special case of SPG with a particular choice of gradient oracle. The fundamental reason why Prox-SVRG is a useful algorithm is because we can

get a bound on the noise $\sigma_{k,j}^2$ in terms of an objective gap. As $x_k^{(j)}$ approaches $x_\star$, this objective gap shrinks and hence so does the variance of the gradient estimate. The particular form of the variance bound allows us to iterate the one-step inequality Lemma 3.1.13 in a particular way to obtain linear convergence.

We proceed to demonstrate the variance bound and convergence guarantees for Prox-SVRG. First, we introduce a key lemma:

Lemma 3.1.14. *Suppose that $f = V + H$ where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex, and $V\colon \mathbb{R}^d \to \mathbb{R}$ can be written as the average of convex and β-smooth component functions $V_i\colon \mathbb{R}^d \to \mathbb{R}$ for $i \in [m]$, so that*

$$V(x) = \frac{1}{m}\sum_{i=1}^{m} V_i(x).$$

Let $x_\star$ be a minimizer of f. Then

$$\frac{1}{m}\sum_{i=1}^{m} \|\nabla V_i(x) - \nabla V_i(x_\star)\|^2 \le 2\beta[f(x) - f(x_\star)].$$

Proof. We follow the proof of Xiao and Zhang [95, Lemma 1]. For each i, we can consider the V_i-Bregman divergence with respect to $x_\star$, which we define as the function

$$\varphi_i(x) := V_i(x) - V_i(x_\star) - \langle \nabla V_i(x_\star), x - x_\star \rangle .$$

This is a convex function, and since $\nabla\varphi_i(x_\star) = 0$, we deduce that $x_\star$ is a minimizer of φ_i. Also, since V_i is β-smooth and φ_i is equal to an affine shift of V_i, φ_i is also β-smooth, implying that

$$\frac{1}{2\beta}\|\nabla\varphi_i(x)\|^2 \le \varphi_i(x) - \varphi_i\left(x - \frac{1}{\beta}\nabla\varphi_i(x)\right) \le \varphi_i(x) - \varphi_i(x_\star) = \varphi_i(x).$$

Hence, we obtain that

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m} \|\nabla V_i(x) - \nabla V_i(x_\star)\|^2 &= \frac{1}{m}\sum_{i=1}^{m} \|\nabla\varphi_i(x)\|^2 \\
&\le \frac{1}{m}\sum_{i=1}^{m} 2\beta[V_i(x) - V_i(x_\star) - \langle \nabla V_i(x_\star), x - x_\star\rangle] \\
&= 2\beta[V(x) - V(x_\star) - \langle \nabla V(x_\star), x - x_\star\rangle]
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\beta[V(x) - V(x_\star) + \langle \nabla H(x_\star), x - x_\star\rangle] \\
&\qquad\qquad \text{(since } 0 = \nabla f(x_\star) = \nabla V(x_\star) + \nabla H(x_\star)\text{)} \\
&\leq 2\beta[V(x) - V(x_\star) + H(x) - H(x_\star)] \qquad \text{(by convexity of } H\text{)} \\
&= 2\beta[f(x) - f(x_\star)],
\end{aligned}
$$

as desired. □

With this key lemma in mind, we are ready to prove the desired variance bound for the iterates.

Lemma 3.1.15 (Variance bound for Prox-SVRG)**.** *For the iterates of Prox-SVRG (Algorithm 3), we have that*

$$\sigma_{k,j}^2 \leq 4\beta[f(x_k^{(j)}) - f(x_\star) + f(x_0^{(j)}) - f(x_\star)].$$

Proof. We will drop the dependence on j. We have that

$$
\begin{aligned}
\mathbb{E}\,\|e_k\|^2 &= \mathbb{E}\,\|v_k - \nabla V(x_k)\|^2 \\
&= \mathbb{E}\,\|\nabla V_i(x_k) - \nabla V_i(x_0) + \nabla V(x_0) - \nabla V(x_k)\|^2 \qquad \text{(with } i \sim \mathrm{Unif}[m]\text{)} \\
&\leq \mathbb{E}\,\|\nabla V_i(x_k) - \nabla V_i(x_0)\|^2 \qquad \text{(since } \mathbb{E}_i[\nabla V_i(x_k) - \nabla V_i(x_0)] = \nabla V(x_k) - \nabla V(x_0)\text{)} \\
&\leq 2\mathbb{E}\left[\|\nabla V_i(x_k) - \nabla V_i(x_\star)\|^2 + \|\nabla V_i(x_0) - \nabla V_i(x_\star)\|^2\right] \\
&\qquad\qquad \text{(by triangle ineq and Cauchy-Schwarz)} \\
&\leq 4\beta[f(x_k) - f(x_\star) + f(x_0) - f(x_\star)], \qquad \text{(by Lemma 3.1.14)}
\end{aligned}
$$

proving the desired result. □

By combining the variance bound of Lemma 3.1.15 with the one-step inequality Lemma 3.1.13 and telescoping, we obtain the following convergence guarantee:

Theorem 3.1.16 (Strongly convex guarantee, Prox-SVRG)**.** *Suppose that $f = V + H$ where $H\colon \mathbb{R}^d \to \mathbb{R}$ is convex, and $V\colon \mathbb{R}^d \to \mathbb{R}$ is α-convex and can be written as the average of convex and β-smooth*

component functions $V_i \colon \mathbb{R}^d \to \mathbb{R}$ *for* $i \in [m]$*, so that*

$$V(x) = \frac{1}{m} \sum_{i=1}^{m} V_i(x).$$

Then Algorithm 3 run with step size $\eta = \frac{1}{32\beta}$ *and inner loop iteration count* $N = \frac{256\beta}{\alpha}$ *satisfies*

$$\mathbb{E}[f(x_0^{(M)}) - f(x_\star)] \le \exp\left(-\frac{M}{4}\right)[f(x_0) - f(x_\star)].$$

In particular, with $M \gtrsim \log \frac{f(x_0)-f(x_\star)}{\varepsilon^2}$*, we obtain the guarantee*

$$\mathbb{E}[f(x_0^{(M)}) - f(x_\star)] \le \varepsilon^2.$$

Proof. First, we consider a fixed value of the outer loop iteration number j, so we drop the dependence on j in the superscript of the iterates. For notational convenience, let $F(x) \coloneqq f(x) - f(x_\star)$. Combining the variance bound of Lemma 3.1.15 with the one-step inequality Lemma 3.1.13 and taking $y = x_\star$, we obtain that

$$\mathbb{E}\|x_{k+1} - x_\star\|^2 \le \mathbb{E}\|x_k - x_\star\|^2 - 2\eta\mathbb{E}[F(x_{k+1})] + 8\beta\eta^2\mathbb{E}[F(x_k) + F(x_0)].$$

Let $\bar{x}$ be chosen from among $\{x_1, \ldots, x_N\}$ uniformly at random. Summing the above inequality for $k = \{0, \ldots, N-1\}$ and rearranging, we obtain

$$\begin{aligned} 2N\eta(1 - 4\beta\eta)\mathbb{E}[F(\bar{x})] + 8\eta^2\beta\mathbb{E}[F(x_N)] + \mathbb{E}\|x_N - x_\star\|^2 &\le \mathbb{E}\|x_0 - x_\star\|^2 + 8(N+1)\beta\eta^2\mathbb{E}F(x_0) \\ &\le \mathbb{E}\|x_0 - x_\star\|^2 + 16N\beta\eta^2\mathbb{E}F(x_0). \end{aligned}$$

Since $F(x_N) \ge 0$ and $\|x_N - x_\star\|^2 \ge 0$, we find that

$$2\eta(1 - 4\beta\eta)\mathbb{E}[F(\bar{x})] \le \frac{1}{N}\mathbb{E}\|x_0 - x_\star\|^2 + 16\beta\eta^2\mathbb{E}F(x_0) \le \left(\frac{2}{N\alpha} + 16\beta\eta^2\right)\mathbb{E}F(x_0).$$

Hence, with the choice $\eta = \frac{1}{32\beta}$ and $N = 256\frac{\beta}{\alpha}$, we obtain that

$$\frac{1}{32\beta}\mathbb{E}[F(\overline{x})] \leq 2\eta(1-4\beta\eta)\mathbb{E}[F(\overline{x})] = \left(\frac{2}{N\alpha} + \frac{1}{64\beta}\right)\mathbb{E}[F(x_0)] = \left(\frac{3}{128\beta}\right)\mathbb{E}[F(x_0)].$$

Translating this back into the language of the original algorithm, we obtain that

$$\mathbb{E}[f(x_0^{(j+1)}) - f(x_\star)] \leq \frac{3}{4}\mathbb{E}[f(x_0^{(j)}) - f(x_\star)] \leq e^{-1/4}\mathbb{E}[f(x_0^{(j)}) - f(x_\star)]$$

Iterating this, we obtain that

$$\mathbb{E}[f(x_0^{(M)}) - f(x_\star)] \leq e^{-M/4}[f(x_0) - f(x_\star)].$$

Finally, this implies that with $M \gtrsim \log \frac{f(x_0)-f(x_\star)}{\varepsilon^2}$, we have the guarantee

$$\mathbb{E}[f(x_0^{(M)}) - f(x_\star)] \leq \varepsilon^2,$$

as desired. □

Hence, Prox-SVRG significantly improves the total number of gradient calls made to attain an objective gap of $\leq \varepsilon^2$:

- Proximal gradient requires m calls to a gradient oracle to evaluate each ∇V_i for every iteration. As the total number of iterations required to obtain an objective gap of $\leq \varepsilon^2$ is $O(\kappa \log(1/\varepsilon))$, the total query complexity of the algorithm is $O(m\kappa \log(1/\varepsilon))$.
- On the other hand, Prox-SVRG makes m gradient calls per outer iteration and 2 gradient calls per inner iteration, resulting in a total of $O((m + 2\kappa)\log(1/\varepsilon)) = O((m+\kappa)\log(1/\varepsilon))$ gradient calls throughout the algorithm, hence providing a strict improvement.

With these algorithms and their corresponding analyses in mind, we are now ready to return to our main goal: developing Euclidean optimization-inspired algorithms for Gaussian VI.

3.2 (Stochastic) Forward-backward Gaussian variational inference

Using Table 2.1 as a roadmap, we are able to translate both Euclidean *algorithms* and *guarantees* to the BW setting, culminating in the development of Forward-Backward Gaussian Variational Inference (FB–GVI), an algorithm for Gaussian VI with state-of-the-art convergence guarantees.

3.2.1 Revisiting Gaussian VI

We return to the setting of Problem (1.1), and state it more formally. We assume that the target distribution π admits a positive density w.r.t. Lebesgue measure, denoted π as well in an abuse of notation. We write π in the form $\pi \propto \exp(-V)$. Moreover, we assume that the function $V : \mathbb{R}^d \to \mathbb{R}$ is β-smooth. Recall that the KL divergence is defined for every $\mu \in \mathsf{BW}(\mathbb{R}^d)$ as

$$\mathsf{KL}\left(\mu \,\|\, \pi\right) = \int \log \frac{\mu(x)}{\pi(x)} \, \mathrm{d}\mu(x) \,. \tag{3.10}$$

We denote $\mathcal{F} := \mathcal{V} + \mathcal{H}$ as the sum of the potential (associated to the function V) and the entropy. Then, a quick calculation reveals that $\mathcal{F}(\mu) - \mathcal{F}(\pi) = \mathsf{KL}\left(\mu \,\|\, \pi\right)$. Since $\mathcal{F}(\pi)$ is a constant (*i.e.*, does not depend on μ), Problem (1.1) is equivalent to

$$\min_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \left\{ \mathcal{V}(\mu) + \mathcal{H}(\mu) \right\} . \tag{3.11}$$

3.2.2 Proposed algorithm

Recall that the potential $\mathcal{V}$ is "smooth" over the BW space and that the BW gradient of $\mathcal{V}$ admits a closed form (Lemma 2.2.1). Recall also that the entropy $\mathcal{H}$ is "convex" over the BW space and that the BW JKO of $\mathcal{H}$ admits a closed form (Lemma 2.2.2). Hence, the objective in problem (3.11) admits a *composite* structure as the sum of a smooth term and a non-smooth term, and as we have seen in Section 3.1.4, the canonical algorithm for solving such a problem in Euclidean space is the (stochastic) proximal gradient algorithm. Thus, we seek to adapt the proximal gradient (or "forward-backward") algorithm to the BW space. Inspired by the Euclidean-BW dictionary we previously defined in Table 2.1, this leads to the following Forward-Backward Gaussian Varia-

tional Inference (FB–GVI) algorithm (note the analogy with (2.9)):

$$p_{k+\frac{1}{2}} = (\mathrm{id} - \eta \, \nabla_{\mathsf{BW}} \mathcal{V}(p_k))_{\#} p_k \,, \tag{3.12}$$

$$p_{k+1} = \mathrm{JKO}_{\eta \mathcal{H}}(p_{k+\frac{1}{2}}) \,. \tag{3.13}$$

The backward step (3.13) is tractable using (2.21). Although the forward step (3.12) also admits a closed form, the forward step involves computing integrals of ∇V and $\nabla^2 V$ with respect to p_k (see (2.16)). These integrals can be intractable. In order to make the algorithm implementable, we also propose an unbiased stochastic gradient estimator $\hat{g}_k$ of $\nabla_{\mathsf{BW}} \mathcal{V}(p_k)$, computed by drawing a random sample from p_k. The resulting algorithm is called Stochastic FB–GVI, and can be written as:

$$\begin{aligned} p_{k+\frac{1}{2}} &= (\mathrm{id} - \eta \, \hat{g}_k)_{\#} p_k \,, \\ p_{k+1} &= \mathrm{JKO}_{\eta \mathcal{H}}(p_{k+\frac{1}{2}}) \,, \end{aligned} \tag{3.14}$$

where $\hat{g}_k$ is the random affine function defined by

$$\hat{g}_k : x \mapsto \nabla V(\hat{X}_k) + \nabla^2 V(\hat{X}_k) \, (x - m_k) \,, \tag{3.15}$$

where $\hat{X}_k \sim p_k$ and $m_k = \int x \, dp_k(x)$ denotes the mean of p_k.

(Stochastic) FB–GVI is precisely an analogue over BW space of the Euclidean (stochastic) proximal gradient algorithm (Algorithm 2). In this setting, the iterates $(p_k)_{k \in \mathbb{N}}$ defined by (3.14) are a sequence of *random Gaussian distributions*, i.e. random variables taking values in $\mathsf{BW}(\mathbb{R}^d)$. We denote the mean (resp. covariance matrix) of p_k by m_k (resp. Σ_k). FB–GVI and Stochastic FB–GVI can be implemented by keeping track of the means and the covariance matrices of the iterates p_k. The iterations of FB–GVI and Stochastic FB–GVI in terms of m_k and Σ_k are given in Algorithm 4. Efficient algorithms developed for computing the matrix square-root (see, for example, Pleiss et al. [73] and Song, Sebe, and Wang [84]) can be leveraged to improve the per-iteration complexity.

Algorithm 4 FB–GVI and Stochastic FB–GVI

Require: Step size $\eta > 0$; Iteration count N; Initial distribution $p_0 = \mathcal{N}(m_0, \Sigma_0)$
  for $k = 0$ **to** $N - 1$ **do**
    if FB–GVI **then**
      $b_k \leftarrow \mathbb{E}_{p_k} \nabla V$, $S_k \leftarrow \mathbb{E}_{p_k} \nabla^2 V$
    else if Stochastic FB–GVI **then**
      draw $\hat{X}_k \sim \mathcal{N}(m_k, \Sigma_k)$
      $b_k \leftarrow \nabla V(\hat{X}_k)$, $S_k \leftarrow \nabla^2 V(\hat{X}_k)$
    end if
    $m_{k+1} \leftarrow m_k - \eta\, b_k$
    $M_{k+1} \leftarrow I - \eta\, S_k$
    $\Sigma_{k+\frac{1}{2}} \leftarrow M_{k+1} \Sigma_k M_{k+1}$
    $\Sigma_{k+1} \leftarrow \frac{1}{2}(\Sigma_{k+\frac{1}{2}} + 2\eta I + [\Sigma_{k+\frac{1}{2}}(\Sigma_{k+\frac{1}{2}} + 4\eta I)]^{1/2})$
  end for
  output $p_N = \mathcal{N}(m_N, \Sigma_N)$

3.2.3 Variance reduction

Just as stochastic proximal gradient admits a variance-reduced extension (Prox-SVRG, Algorithm 3) in the setting when V can be written as the average of many smooth, convex component functions, we can also develop a variant of Algorithm 4 for the analogous setting in BW space.

Specifically, we consider the setting where $\pi \propto \exp(-V)$, where V is an α-strongly convex function that can be written as the average of m convex, β-smooth component functions $V_i \colon \mathbb{R}^d \to \mathbb{R}$. In this setting, Equation (3.11) is equivalent to

$$\min_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \{\mathcal{V}(\mu) + \mathcal{H}(\mu)\} = \min_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \left\{ \frac{1}{m} \sum_{i=1}^{m} \mathcal{V}_i(\mu) + \mathcal{H}(\mu) \right\},$$

where the functional $\mathcal{V}_i \colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ is defined by $\mathcal{V}_i(\mu) := \mathbb{E}_\mu V_i$. By adapting Algorithm 3 for the setting of FB–GVI, we obtain Variance-Reduced FB–GVI (VRFB–GVI), detailed in Algorithm 5. Just like in the setting of Prox-SVRG, (stochastic) VRFB–GVI employs an inner loop/outer loop structure wherein a "centering sequence" is updated less frequently than at every iteration. As such, we will denote the kth inner-loop iterate of the jth outer-loop iteration by $p_k^{(j)}$. Through this centering procedure, VRFB–GVI attains lower gradient oracle query complexity than FB–GVI in the setting where the number m of component functions is large. We make precise statements of the complexity guarantees of VRFB–GVI in Section 3.3.3.

Algorithm 5 Variance-Reduced FB–GVI and Stochastic Variance-Reduced FB–GVI

Require: Step size $\eta > 0$; Inner loop iteration count N; Outer loop iteration count M; Initial distribution $p_0 = \mathcal{N}(m_0, \Sigma_0)$

$p_0^{(0)} \leftarrow p_0$
for $j = 0$ **to** $M - 1$ **do**
  if Variance-Reduced FB–GVI **then**
    precompute $\mathbb{E}_{p_0^{(j)}}[\nabla V]$, $\mathbb{E}_{p_0^{(j)}}[\nabla^2 V]$
  else if Variance-Reduced Stochastic FB–GVI **then**
    draw $\hat{X}_0^{(j)} \sim p_0^{(j)}$
    precompute $\nabla V(\hat{X}_0^{(j)})$, $\nabla^2(\hat{X}_0^{(j)})$
  end if
  for $k = 0$ **to** $N - 1$ **do**
    randomly draw $i \sim \mathrm{Unif}[m]$
    if Variance-Reduced FB–GVI **then**
      $b_k^{(j)} \leftarrow \mathbb{E}_{p_k^{(j)}} \nabla V_i - \mathbb{E}_{p_0^{(j)}} \nabla V_i + \mathbb{E}_{p_0^{(j)}} \nabla V$
      $S_k^{(j)} \leftarrow \mathbb{E}_{p_k^{(j)}} \nabla^2 V_i - \mathbb{E}_{p_0^{(j)}} \nabla^2 V_i + \mathbb{E}_{p_0^{(j)}} \nabla^2 V$
    else if Stochastic FB–GVI **then**
      draw $\hat{X}_k^{(j)} \sim p_k^{(j)}$
      $b_k^{(j)} \leftarrow \nabla V_i(\hat{X}_k^{(j)}) - \nabla V_i(\hat{X}_0^{(j)}) + \nabla V(\hat{X}_0^{(j)})$
      $S_k^{(j)} \leftarrow \nabla^2 V_i(\hat{X}_k^{(j)}) - \nabla^2 V_i(\hat{X}_0^{(j)}) + \nabla^2 V(\hat{X}_0^{(j)})$
    end if
    $m_{k+1}^{(j)} \leftarrow m_k^{(j)} - \eta\, b_k^{(j)}$
    $M_{k+1}^{(j)} \leftarrow I - \eta\, S_k^{(j)}$
    $\Sigma_{k+\frac{1}{2}}^{(j)} \leftarrow M_{k+1}^{(j)} \Sigma_k^{(j)} M_{k+1}^{(j)}$
    $\Sigma_{k+1}^{(j)} \leftarrow \frac{1}{2}(\Sigma_{k+\frac{1}{2}}^{(j)} + 2\eta I + [\Sigma_{k+\frac{1}{2}}^{(j)}(\Sigma_{k+\frac{1}{2}}^{(j)} + 4\eta I)]^{1/2})$
  end for
  $p_0^{(j+1)} \leftarrow p_N^{(j)}$
end for
output $p_0^{(M)} = \mathcal{N}(m_0^{(M)}, \Sigma_0^{(M)})$

3.3 Convergence theory

In this section, we study the convergence of FB–GVI and Stochastic FB–GVI using their equivalent forms (3.12)–(3.13) and (3.14). We also denote by $\hat{\pi} = \mathcal{N}(\hat{m}, \hat{\Sigma})$ a solution of Problem (1.1) (*i.e.*, a minimizer of the KL objective), and we let $\mathscr{F}_k$ denote the σ-algebra generated up to iteration k (but not including the random sample $\hat{X}_k \sim p_k$ in Stochastic FB–GVI).

We consider several assumptions on V. Given $\alpha \in \mathbb{R}$, V is α-convex if $\alpha I \preceq \nabla^2 V$. If $\alpha = 0$, V is said to be convex, and if $\alpha > 0$, V is said to be (α-)strongly convex. For either algorithm, define

the (random) error function (see the definitions of b_k and S_k in Algorithm 4) as

$$e_k \colon x \mapsto (S_k - \mathbb{E}_{p_k} \nabla^2 V)(x - m_k) + (b_k - \mathbb{E}_{p_k} \nabla V)\,,$$

and denote its expected $L^2(p_k)$ norm by $\sigma_k^2 \coloneqq \mathbb{E}[\|e_k\|_{p_k}^2 \mid \mathscr{F}_k]$. The expectation is taken over the possible randomness of (b_k, S_k) (i.e., over the randomness of $\hat{X}_k$). For Stochastic FB–GVI, $e_k = \hat{g}_k - \nabla_{\mathsf{BW}} \mathcal{V}(p_k)$, where $\hat{g}_k$ is defined by (3.15). Since $\mathbb{E}[e_k \mid \mathscr{F}_k] = 0$ (*i.e.*, the BW stochastic gradient is unbiased), σ_k^2 is the conditional variance of the BW stochastic gradient at iteration k. For FB–GVI, $e_k = \nabla_{\mathsf{BW}} \mathcal{V}(p_k) - \nabla_{\mathsf{BW}} \mathcal{V}(p_k) = 0$, hence $\sigma_k = 0$. Our analysis of FB–GVI and Stochastic FB–GVI relies on the following unified one-step-inequality for the iterates $(p_k)_{k \in \mathbb{N}}$ of both (3.12)–(3.13) and (3.14).

Lemma 3.3.1 (One-step inequality for FB–GVI)**.** *Suppose that V is α-convex and β-smooth. Let $(p_k)_{k \in \mathbb{N}}$ be the iterates of FB–GVI (3.12)–(3.13) or Stochastic FB–GVI (3.14). Let $\eta > 0$ be such that*

$$\eta \leq \begin{cases} \frac{1}{\beta} & \text{if } \sigma_k = 0 \ \text{(FB–GVI)}\,, \\ \frac{1}{2\beta} & \text{else}\,. \end{cases}$$

Then, for all $\nu \in \mathsf{BW}(\mathbb{R}^d)$,

$$\mathbb{E} W_2^2(p_{k+1}, \nu) \leq (1 - \alpha\eta)\, \mathbb{E} W_2^2(p_k, \nu) - 2\eta\, \mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\, \mathbb{E}\sigma_k^2\,. \tag{3.16}$$

Proof. The proof is given in Appendix A.1. □

This one-step inequality is precisely the analogue of Lemma 3.1.1, obtained by translating squared Euclidean distance to squared Wasserstein distance, as per Table 2.1. Our proof of Lemma 3.3.1 heavily employs the differential and geometric structure of the BW space presented in Section 2.2.

3.3.1 Convergence of FB–GVI

In this section, $(p_k)_{k \in \mathbb{N}}$ is the sequence of iterates defined by FB–GVI ((3.12)–(3.13)). We obtain corollaries of Lemma 3.3.1 by setting $\sigma_k = 0$ in (3.16), when V is convex or strongly convex.

Theorem 3.3.2 (Convex case, FB–GVI)**.** *Suppose that V is convex and β-smooth and that $0 < \eta \leq \frac{1}{\beta}$. Then,*

$$\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \frac{W_2^2(p_0, \hat{\pi})}{2N\eta}\,.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta W_2^2(p_0,\hat{\pi})}{\varepsilon^2}$, we obtain the guarantee

$$\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2\,.$$

Proof. The proof is given in Appendix A.3.1. □

Theorem 3.3.3 (Strongly convex case, FB–GVI)**.** *Suppose that V is α-strongly convex and β-smooth, and that $0 < \eta \leq \frac{1}{\beta}$. Then,*

$$W_2^2(p_N, \hat{\pi}) \leq \exp\left(-N\alpha\eta\right)\, W_2^2(p_0, \hat{\pi})\,.$$

In particular, when $\eta = \frac{1}{\beta}$ and $N \gtrsim \frac{\beta}{\alpha} \log \frac{W_2(p_0,\hat{\pi})}{\varepsilon}$, we obtain the guarantees

$$\alpha\, W_2^2(\mu_N, \hat{\pi}) \leq \varepsilon^2\,, \quad \textit{and} \quad \mathcal{F}(p_{2N}) - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2\,.$$

Proof. The proof is given in Appendix A.3.2. □

Theorem 3.3.2 demonstrates a sublinear rate of convergence of FB–GVI for a convex V (in terms of objective gap) and Theorem 3.3.3 demonstrates a *linear* rate of convergence of FB–GVI for a strongly convex V. The convergence rates we obtain are of the same order as the convergence rates of the proximal gradient algorithm [34] (see also Section 3.1.4); hence, the convergence properties of FB–GVI can be seen as inherited from those of the Euclidean proximal gradient algorithm applied to V.

Finally, we also extend our results to the non-convex case, where we obtain a stationary point guarantee, analogously to Theorem 3.1.10.

Theorem 3.3.4 (Non-convex case, FB–GVI)**.** *Suppose that V is β-smooth, and that $0 < \eta \leq \frac{1}{\beta}$. Let*

$\Delta := \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})$. *Then,*

$$\min_{k \in \{0,\dots,N-1\}} \|\nabla_{\mathsf{BW}} \mathcal{F}(p_k)\|_{p_k}^2 \leq \frac{150\Delta}{\eta N}\,.$$

In particular, when $\eta = \frac{1}{\beta}$ *and* $N \gtrsim \frac{\beta\Delta}{\varepsilon^2}$, *we obtain the guarantee*

$$\min_{k \in \{0,\dots,N-1\}} \|\nabla_{\mathsf{BW}} \mathcal{F}(p_k)\|_{p_k}^2 \leq \varepsilon^2\,. \tag{3.17}$$

Proof. The proof is given in Appendix A.3.3. □

To the best of our knowledge, this is the first stationary point guarantee for Gaussian VI. The relevance of this result is that according to Katsevich and Rigollet [45], the favorable statistical properties of Gaussian VI arise, not due to the global minimization of the objective in (1.1), but rather from the first-order optimality (3.17). Hence, Theorem 3.3.4 can be viewed as an algorithmic result for posterior approximation, even in the non-log-concave setting.

We also emphasize that although we assume that V is smooth, it does *not* follow that the objective $\mathcal{F}$ is smooth over the Bures–Wasserstein space, due to the presence of the entropy term $\mathcal{H}$. In fact, the entropy term is only smooth when constrained to a set of Gaussians with covariance matrices having lower-bounded eigenvalues. Hence, the proof of Theorem 3.3.4 requires careful control of the eigenvalues of the iterates of FB–GVI.

3.3.2 Convergence of Stochastic FB–GVI

In this section, $(p_k)_{k \in \mathbb{N}}$ is the sequence of iterates defined by (3.14). To use Lemma 3.3.1, we first prove a bound on σ_k^2, the variance of the BW stochastic gradient.

Lemma 3.3.5. *If* V *is convex and* β*-smooth, then*

$$\sigma_k^2 \leq 6\beta d + 12\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(p_k, \hat{\pi})\,.$$

Moreover, if V *is* α*-strongly convex, the bound above becomes*

$$\sigma_k^2 \leq 6\beta d + \frac{12\beta^3}{\alpha}\, W_2^2(p_k, \hat{\pi})\,.$$

Proof. See Appendix A.4.1. □

The bound on σ_k^2 is reminiscent of the common assumption made in the literature on stochastic gradient algorithms over $\mathbb{R}^d$, that the stochastic gradient has sublinear growth [49, 15]. We emphasize that we do not assume this sublinear growth. Instead, Lemma 3.3.5 proves the sublinear growth for the BW stochastic gradient used in Stochastic FB–GVI. Next, we obtain corollaries of Lemma 3.3.1 for Stochastic FB–GVI by controlling σ_k^2 with Lemma 3.3.5.

Theorem 3.3.6 (Convex case, Stochastic FB–GVI)**.** *Suppose that V is convex and β-smooth and that $0 < \eta \leq \frac{1}{2\beta}$. Define $c := 24\beta^3 \lambda_{\max}(\hat{\Sigma})$. Then,*

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} \mathcal{F}(p_k)\right] - \mathcal{F}(\hat{\pi}) \leq \frac{2W_2^2(p_0,\hat{\pi})}{N\eta} + 2c\eta\, W_2^2(p_0,\hat{\pi}) + 12\beta\eta d\,.$$

In particular, for sufficiently small values of ε^2/d and with

$$\eta \asymp \frac{\varepsilon^2}{cW_2^2(p_0,\hat{\pi}) \vee \beta d}\,, \quad \text{and} \quad N \gtrsim \frac{W_2^2(p_0,\hat{\pi})}{\varepsilon^4}\left(cW_2^2(p_0,\hat{\pi}) \vee \beta d\right)\,,$$

we obtain the guarantee

$$\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} \mathcal{F}(p_k)\right] - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2\,.$$

Proof. See Appendix A.4.3. □

Theorem 3.3.7 (Strongly convex case, Stochastic FB–GVI)**.** *Suppose that V is α-strongly convex and β-smooth, and that $\eta \leq \frac{\alpha^2}{48\beta^3}$. Then,*

$$\mathbb{E}W_2^2(p_N,\hat{\pi}) \leq \exp\Big(-\frac{N\alpha\eta}{2}\Big)\, W_2^2(p_0,\hat{\pi}) + \frac{24\beta\eta d}{\alpha}\,.$$

In particular, for sufficiently small values of ε^2/d and with

$$\eta \asymp \frac{\varepsilon^2}{\beta d}\,, \quad \text{and} \quad N \gtrsim \frac{\beta d}{\alpha\varepsilon^2}\log\frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2}\,,$$

we obtain the guarantees

$$\alpha\,\mathbb{E}W_2^2(p_N,\hat{\pi}) \le \varepsilon^2\,, \quad \textit{and} \quad \mathbb{E}\left[\min_{k\in\{1,\dots,2N\}} \mathcal{F}(p_k)\right] - \mathcal{F}(\hat{\pi}) \le \varepsilon^2\,.$$

Proof. See Appendix A.4.4. □

To our knowledge, Theorem 3.3.6 is the first result to provide a complexity result in terms of the objective gap in Problem (1.1), for log-smooth log-concave target distributions. Moreover, Theorem 3.3.7 improves upon the state-of-the-art obtained in [51] for strongly log-concave target distributions. In particular, ignoring logarithmic factors, their iteration complexity (when written in a scale-invariant way) reads $\widetilde{O}(\frac{\beta^2 d}{\alpha^2\varepsilon^2})$, whereas ours reads $\widetilde{O}(\frac{\beta d}{\alpha\varepsilon^2})$. Note that the linear dependence on the condition number β/α is to be expected for gradient descent methods. We remark that our analysis crucially makes use of the proximal operator (the BW JKO) on the non-smooth entropy in order to obtain our improved rates.

3.3.3 Convergence of Variance-Reduced FB–GVI

In this section, we denote by $(p_k^{(j)})_{j,k\in\mathbb{N}}$ the sequence of iterates defined by Algorithm 5. We define $e_k^{(j)}, \mathscr{F}_k^{(j)}$ analogously to Section 3.3, and define $\sigma_{k,j}^2 := \mathbb{E}[\|e_k^{(j)}\|^2 \mid \mathscr{F}_k^{(j)}]$. Then Equation (3.16) holds identically for Algorithm 5 for all $j,k \in \mathbb{N}$ and $\nu \in \mathsf{BW}(\mathbb{R}^d)$ with the following relabelling:

$$\mathbb{E}W_2^2(p_{k+1}^{(j)},\nu) \le (1-\alpha\eta)\mathbb{E}W_2^2(p_k^{(j)},\nu) - 2\eta\mathbb{E}[\mathcal{F}(p_{k+1}^{(j)}) - \mathcal{F}(\nu)] + 2\eta^2\mathbb{E}\sigma_{k,j}^2. \tag{3.18}$$

In order to prove rates of convergence for (stochastic) VRFB–GVI, we require the following bounds on the variance of the gradient estimate:

Lemma 3.3.8 (Variance bound for VRFB–GVI)**.** *Suppose that each V_i is convex and β-smooth. Then for VRFB–GVI, we have that*

$$\sigma_{k,j}^2 \le 6\beta[\mathcal{F}(p_k^{(j)}) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0^{(j)}) - \mathcal{F}(\hat{\pi})] + 3\beta^2 W_2^2(p_k^{(j)}, p_0^{(j)}).$$

Proof. The proof is deferred to Appendix A.5.1. □

Lemma 3.3.9 (Variance bound for stochastic VRFB–GVI). *Suppose that each V_i is convex and β-smooth. Then for VRFB–GVI, we have that*

$$\sigma_{k,j}^2 \le 72\beta d + \frac{120\beta^3}{\alpha}[W_2^2(p_k^{(j)}, \hat{\pi}) + W_2^2(p_0^{(j)}, \hat{\pi})] + 6\beta[\mathcal{F}(p_k^{(j)}) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0^{(j)}) - \mathcal{F}(\hat{\pi})].$$

Proof. The proof is deferred to Appendix A.5.2. □

Using these variance bounds in conjunction with the one-step inequality, we are able to obtain rates of convergence for (stochastic) VRFB–GVI.

Theorem 3.3.10 (Strongly convex guarantee, VRFB–GVI). *Suppose $V\colon \mathbb{R}^d \to \mathbb{R}$ is α-convex and can be written as the average of convex and β-smooth component functions $V_i\colon \mathbb{R}^d \to \mathbb{R}$ for $i \in [m]$, so that*

$$V(x) = \frac{1}{m}\sum_{i=1}^{m} V_i(x).$$

Suppose that VRFB–GVI (Algorithm 5) is initialized at p_0 with $\Sigma_0^{-1} \preceq 2\beta I$, and run with step size $\eta = \frac{1}{288\beta\kappa}$ and inner loop iteration count $N = \frac{\kappa^2}{144}$. Then

$$\mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \le \exp\left(-\frac{M}{4}\right) W_2^2(p_0, \hat{\pi}).$$

In particular, with $M \gtrsim \log \frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2}$, we obtain the guarantee

$$\alpha\mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \le \varepsilon^2.$$

Proof. The proof is deferred to Appendix A.5.3. □

An analogous result holds for Stochastic VRFB–GVI.

Theorem 3.3.11 (Strongly convex guarantee, Stochastic VRFB–GVI). *Suppose $V\colon \mathbb{R}^d \to \mathbb{R}$ is α-convex and can be written as the average of convex and β-smooth component functions $V_i\colon \mathbb{R}^d \to \mathbb{R}$ for $i \in [m]$, so that*

$$V(x) = \frac{1}{m}\sum_{i=1}^{m} V_i(x).$$

Suppose that stochastic VRFB–GVI (Algorithm 5) is initialized at p_0 with $\Sigma_0^{-1} \preceq 2\beta I$, and run with step size $\eta = \frac{1}{2400}(\frac{\varepsilon^2}{\beta d} \wedge \frac{1}{\beta\kappa^2})$ and inner loop iteration count $N = \frac{2}{\alpha\eta} \asymp \kappa(\frac{d}{\varepsilon^2} \vee \kappa^2)$. Then

$$\mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \leq \exp\left(-\frac{M}{4}\right) W_2^2(p_0, \hat{\pi}) + \frac{\varepsilon^2}{2\alpha}.$$

In particular, with $M \gtrsim \log \frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2}$, we obtain the guarantee

$$\alpha\mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \leq \varepsilon^2.$$

Proof. The proof is deferred to Appendix A.5.4. □

Putting the pieces together, VRFB–GVI attains a αW_2^2 distance to the minimizer of order ε^2 with $O((m + \kappa^2) \log \frac{W_2^2(p_0,\hat{\pi})}{\varepsilon})$ calls to a gradient oracle for $\nabla_{\mathsf{BW}}\mathcal{V}_i$, whereas FB–GVI requires a total of $O(m\kappa \log \frac{W_2^2(p_0,\hat{\pi})}{\varepsilon})$ calls since it must query the gradient oracle for $\mathcal{V}_i$ for all $i \in [m]$ at each iteration. Unfortunately, the guarantee given in Theorem 3.3.10 is not ideal, as the number of inner loop iterations is given by $O(\kappa^2)$ rather than $O(\kappa)$ as in Theorem 3.1.16. The reason for this discrepancy is because our bound on the variance of the stochastic gradient introduces a dependence on $W_2^2(p_0^{(j)}, \hat{\pi})$, whereas no corresponding dependence on $\|x_0^{(j)} - x_\star\|^2$ exists in the analysis of the error term in Prox-SVRG. We hypothesize that this dependence on $W_2^2(p_0^{(j)}, \hat{\pi})$ prevents iteration of the one-step inequality unless the step size is of order $O(\frac{1}{\beta\kappa})$ (rather than $O(\frac{1}{\beta})$ for Prox-SVRG), which in turn results in an inner-loop iteration complexity of $O(\kappa^2)$ rather than $O(\kappa)$. However, in the setting where $m \gg \kappa$, VRFB–GVI outperforms FB–GVI in terms of gradient oracle query complexity.

On the other hand, Stochastic VRFB–GVI *does* obtain a better oracle complexity guarantee than Stochastic FB–GVI. Overall,

- Stochastic VRFB–GVI requires $O((m + \frac{\kappa d}{\varepsilon^2} + \kappa^3) \log \frac{W_2(p_0,\hat{\pi})}{\varepsilon})$ queries to a gradient oracle, while
- Stochastic FB–GVI requires $O(m(\frac{\kappa d}{\varepsilon^2} + \kappa^3) \log \frac{W_2(p_0,\hat{\pi})}{\varepsilon})$ such calls in total.

Here, the dependence on κ^3 arises from the constraint $\eta \lesssim \frac{1}{\beta\kappa^2}$ for both Stochastic VRFB–GVI and Stochastic FB–GVI.

Since Stochastic VRFB–GVI is implementable, this constitutes a genuine practical improvement over the naive implementation of Stochastic FB–GVI in the setting where the potential $\mathcal{V}$ is the average of many component functionals!

Chapter 4

Conclusion

We proposed a novel optimization algorithm, (Stochastic) FB–GVI, for solving the Gaussian VI problem in (1.1). In the setting where the potential admits a representation as the average of many other component potentials, we provide a variance-reduced extension to FB–GVI with improved complexity guarantees.

We view FB–GVI as performing optimization over the Bures–Wasserstein space, echoing a stream of successful works on optimization-inspired design and analysis of sampling and variational inference algorithms. Using this perspective, we also provided new or state-of-the-art convergence rates for solving (1.1), depending on the regularity assumptions on π. As immediate future work, it is intriguing to study the statistical properties (consistency, normal approximation bounds, moment estimation bounds, and robustness properties) of the proposed (Stochastic) FG–GVI algorithm on various specific practical problems of interest. From a technical standpoint, it would also be of interest to either sharpen the analysis of Variance-Reduced FB–GVI to avoid quadratic complexity dependence on the condition number κ, or to devise a new variance reduction technique altogether which is tailored to the Bures–Wasserstein space.

At a broader level, our work opens the door to the following question: Can we develop a rigorous algorithmic framework for general VI, *i.e.*, Problem (1.1) where $\mathsf{BW}(\mathbb{R}^d)$ is replaced by a different or larger set of distributions (for example, mixtures of Gaussians)? We believe that this paper provides a concrete step toward this general goal.

Acknowledgements

This thesis is based largely on the joint work [29] done with Krishnakumar Balasubramanian, Sinho Chewi, and Adil Salim. I am indebted to their help and guidance, without which this project would have never been possible.

Appendix A

Technical Proofs

A.1 Proof of the one-step inequality (Lemma 3.3.1)

The key idea of this proof is to decompose the difference $\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)$ as the sum of three terms,

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu) = [\mathcal{V}(p_{k+1}) - \mathcal{V}(p_k)] + [\mathcal{V}(p_k) - \mathcal{V}(\nu)] + [\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)]\,,$$

where each individual term may be controlled using the inequalities in Lemmas 2.2.2 and 2.2.4. Recalling that Lemma 2.2.2 applies only to *generalized geodesics*, we must take care in defining couplings between $p_k, p_{k+\frac{1}{2}}, p_{k+1}$ and ν. We detail the argument in the following proof.

Proof of Lemma 3.3.1. Recall from Section 3.3 that we defined $\mathscr{F}_k$ as the σ-algebra generated up to iteration k (but not including the random sample $\hat{X}_k \sim p_k$ in Stochastic FB–GVI)). We also have

$$e_k \colon x \mapsto (S_k - \mathbb{E}_{p_k} \nabla^2 V)(x - m_k) + (b_k - \mathbb{E}_{p_k} \nabla V)$$

to be defined as the (random) error of the gradient estimate at iteration k of (stochastic) FB–GVI, for which $\mathbb{E}[e_k \mid \mathscr{F}_k] = 0$. Conditioned on the filtration $\mathscr{F}_k$, we construct the following random variables $X_k, X_{k+\frac{1}{2}}, X_{k+1}, Y_{\mathcal{V}}$ and $Y_{\mathcal{H}}$.

Let $(X_k, Y_{\mathcal{V}}) \sim (p_k, \nu)$ be optimally coupled for the W_2 distance, and let $(X_k, Y_{\mathcal{V}}) \perp\!\!\!\perp e_k$. Since

$\eta \leq \frac{1}{\beta}$ by assumption, we have that

$$I - \eta S_k \succeq (1 - \eta\beta)\, I \succeq 0\,.$$

Recall that by Brenier's theorem [91, Theorem 2.12], if $Y = \nabla\varphi(X)$ for a convex, proper, and lower-semicontinuous function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, then (X, Y) is an optimal coupling for the 2-Wasserstein distance. The condition $I - \eta S_k \succeq 0$ above therefore ensures that $(X_k, X_{k+\frac{1}{2}}) \sim (p_k, p_{k+\frac{1}{2}})$ is an optimal coupling for the W_2 distance, where we define

$$X_{k+\frac{1}{2}} \coloneqq (I - \eta S_k)(X_k - m_k) + m_k - \eta b_k\,.$$

On the other hand, defining X_{k+1} such that

$$\begin{aligned} X_{k+1} &\coloneqq X_{k+\frac{1}{2}} - \eta\, \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}] \\ &= (I - \eta\, \Sigma_{k+1}^{-1})^{-1}(X_{k+\frac{1}{2}} - m_{k+1}) + m_{k+1}\,, \end{aligned}$$

we also get that $(X_{k+\frac{1}{2}}, X_{k+1}) \sim (p_{k+\frac{1}{2}}, p_{k+1})$ are optimally coupled. Finally, we construct the random variable $Y_{\mathcal{H}} \sim \nu$ for which $(X_{k+\frac{1}{2}}, Y_{\mathcal{H}})$ are optimally coupled for the W_2 distance.

First, we bound the difference in energy. From Brenier's theorem, we know that $Y_{\mathcal{H}}$ and X_{k+1} can both be expressed as an affine functions of X_k, thereby enabling the application of Lemma 2.2.4. Doing so, we obtain that

$$\begin{aligned} \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(\nu)] &= \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(p_k)] + \mathbb{E}[\mathcal{V}(p_k) - \mathcal{V}(\nu)] \\ &\leq \mathbb{E}\left\langle \nabla_{\mathsf{BW}}\mathcal{V}(p_k)(X_k), X_k - Y_{\mathcal{V}}\right\rangle - \frac{\alpha}{2}\,\mathbb{E}\left\|X_k - Y_{\mathcal{V}}\right\|^2 \\ &\quad + \mathbb{E}\left\langle \nabla_{\mathsf{BW}}\mathcal{V}(p_k)(X_k), X_{k+1} - X_k\right\rangle + \frac{\beta}{2}\,\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \quad \text{(by Lemma 2.2.4)} \\ &= -\frac{\alpha}{2}\,\mathbb{E}\left\|X_k - Y_{\mathcal{V}}\right\|^2 + \mathbb{E}\left\langle \nabla_{\mathsf{BW}}\mathcal{V}(p_k)(X_k), X_{k+1} - Y_{\mathcal{V}}\right\rangle \\ &\quad + \frac{1}{2\eta}\,\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 - \left(\frac{1}{2\eta} - \frac{\beta}{2}\right)\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \\ &= -\frac{\alpha}{2}\,\mathbb{E}\left\|X_k - Y_{\mathcal{V}}\right\|^2 - \mathbb{E}\left\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\right\rangle - \frac{1}{\eta}\,\mathbb{E}\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}}\rangle \\ &\quad + \frac{1}{2\eta}\,\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 - \left(\frac{1}{2\eta} - \frac{\beta}{2}\right)\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \end{aligned}$$

$$= \frac{1}{2\eta}\,(1-\alpha\eta)\,\mathbb{E}\,\|X_k - Y_{\mathcal{V}}\|^2 - \mathbb{E}\,\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\,\|X_{k+1} - X_k\|^2$$
$$+ \frac{1}{2\eta}\,\mathbb{E}\big[\|X_{k+1} - X_k\|^2 - \|X_k - Y_{\mathcal{V}}\|^2 - 2\,\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}}\rangle\big]\,.$$

Now we bound the difference in entropy. Since $Y_{\mathcal{H}}$ and X_{k+1} are both optimally coupled with $X_{k+\frac{1}{2}}$, we know that $(Y_{\mathcal{H}}, X_{k+1})$ are coupled along a generalized geodesic. Hence, we can apply Lemma 2.2.2 to obtain that

$$\begin{aligned}
\mathbb{E}[\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)] &\le \mathbb{E}\,\langle \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}], X_{k+1} - Y_{\mathcal{H}}\rangle \\
&= -\frac{1}{\eta}\,\mathbb{E}\langle X_{k+1} - X_{k+\frac{1}{2}}, X_{k+1} - Y_{\mathcal{H}}\rangle \\
&= \frac{1}{2\eta}\,\mathbb{E}\big[\|X_{k+\frac{1}{2}} - Y_{\mathcal{H}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] \\
&\le \frac{1}{2\eta}\,\mathbb{E}\big[\|X_{k+\frac{1}{2}} - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big]\,.
\end{aligned}$$

(since $(X_{k+\frac{1}{2}}, Y_{\mathcal{H}})$ are optimally coupled)

Now, we sum the above inequalities to obtain our desired bound on $\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)]$. We obtain that

$$\begin{aligned}
\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &= \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(\nu)] + \mathbb{E}[\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)] \\
&\le \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] \\
&\quad + \frac{1}{2\eta}\,\mathbb{E}\big[\|X_{k+1} - X_k\|^2 - \|X_k - Y_{\mathcal{V}}\|^2 + \|X_{k+\frac{1}{2}} - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2\big] \\
&\quad - \frac{1}{2\eta}\,\mathbb{E}\big[2\,\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}}\rangle\big] \\
&\quad - \mathbb{E}\,\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\,\|X_{k+1} - X_k\|^2 \\
&= \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] \\
&\quad - \mathbb{E}\,\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\,\|X_{k+1} - X_k\|^2\,. \qquad \text{(A.1)}
\end{aligned}$$

Finally, it remains to bound the error term on the last line. For this, we consider two cases based on whether or not the error term e_k is identically zero:

- In the case of FB–GVI where we have access to the exact gradient $\nabla_{\mathsf{BW}}\mathcal{V}(p_k)$, we have that

$e_k \equiv 0$, so

$$-\mathbb{E}\left\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\right\rangle = 0\,.$$

Combining this with Inequality A.1, we obtain that with $\eta \le \frac{1}{\beta}$,

$$\begin{aligned}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &\le \frac{1}{2\eta}\,\mathbb{E}\left[(1-\alpha\eta)\,\left\|X_k - Y_{\mathcal{V}}\right\|^2 - \left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2\right] - \left(\frac{1}{2\eta} - \frac{\beta}{2}\right)\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \\ &\le \frac{1}{2\eta}\,\mathbb{E}\left[(1-\alpha\eta)\,\left\|X_k - Y_{\mathcal{V}}\right\|^2 - \left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2\right].\end{aligned}$$

Rearranging, we conclude that if $e_k \equiv 0$ and $\eta \le \frac{1}{\beta}$,

$$\begin{aligned}\mathbb{E}W_2^2(p_{k+1}, \nu) &\le \mathbb{E}\left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2 && \text{(A.2)}\\ &\le (1-\alpha\eta)\,\mathbb{E}\left\|X_k - Y_{\mathcal{V}}\right\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] && \text{(A.3)}\\ &= (1-\alpha\eta)\,\mathbb{E}W_2^2(p_k, \nu) - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)]\,.\end{aligned}$$

(since conditioned on $\mathscr{F}_k$, $(X_k, Y_{\mathcal{V}})$ are optimally coupled)

- Otherwise, if e_k is not necessarily identically 0, we can still compute

$$\begin{aligned}-\mathbb{E}\left\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\right\rangle &= -\mathbb{E}\left\langle e_k(X_k), X_{k+1} - X_k\right\rangle \quad \text{(since } e_k \perp\!\!\!\perp (X_k, Y_{\mathcal{V}}) \text{ by construction)}\\ &\le \eta\,\mathbb{E}\left\|e_k(X_k)\right\|^2 + \frac{1}{4\eta}\,\mathbb{E}\left\|X_{k+1} - X_k\right\|^2\,.\end{aligned}$$

(Cauchy–Schwarz and Young's inequality)

Hence, combining this with Inequality A.1, we obtain that for $\eta \le \frac{1}{2\beta}$,

$$\begin{aligned}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &\le \frac{1}{2\eta}\,\mathbb{E}\left[(1-\alpha\eta)\,\left\|X_k - Y_{\mathcal{V}}\right\|^2 - \left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2\right] + \eta\,\mathbb{E}\left\|e_k(X_k)\right\|^2 \\ &\quad - \left(\frac{1}{4\eta} - \frac{\beta}{2}\right)\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \\ &\le \frac{1}{2\eta}\,\mathbb{E}\left[(1-\alpha\eta)\,\left\|X_k - Y_{\mathcal{V}}\right\|^2 - \left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2\right] + \eta\,\mathbb{E}\sigma_k^2\,.\end{aligned}$$

Rearranging, we conclude that as long as $\eta \leq \frac{1}{2\beta}$,

$$\begin{aligned}
\mathbb{E}W_2^2(p_{k+1}, \nu) &\leq \mathbb{E}\left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2 \\
&\leq (1 - \alpha\eta)\,\mathbb{E}\left\|X_k - Y_{\mathcal{V}}\right\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\,\mathbb{E}\sigma_k^2 \\
&= (1 - \alpha\eta)\,\mathbb{E}W_2^2(p_k, \nu) - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\,\mathbb{E}\sigma_k^2\,. \\
&\qquad\qquad \text{(since } (X_k, Y_{\mathcal{V}}) \text{ are optimally coupled)}
\end{aligned}$$

Combining these two cases, we have demonstrated our desired inequality. □

Remark A.1.1. Consider specializing the above proof to the case where $\nu = p_k$, for which $Y_{\mathcal{V}} = Y_{\mathcal{H}} = X_k$, so that $(X_k, X_{k+\frac{1}{2}}) \sim (p_k, p_{k+\frac{1}{2}})$ and $(X_{k+\frac{1}{2}}, X_{k+1}) \sim (p_{k+\frac{1}{2}}, p_{k+1})$ are optimally coupled for the W_2 distance. Then from Inequality A.3, we obtain that

$$\begin{aligned}
&\mathbb{E}\left\|X_{k+1} - Y_{\mathcal{H}}\right\|^2 \leq (1 - \alpha\eta)\,\mathbb{E}\left\|X_k - Y_{\mathcal{V}}\right\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] && \text{(Inequality A.3)} \\
\implies &\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \leq -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)]\,. && \text{(since } \nu = p_k \text{ and } Y_{\mathcal{V}} = Y_{\mathcal{H}} = X_k)
\end{aligned}$$

As a corollary, we obtain the following lemma, which will be useful in subsequent analysis.

Lemma A.1.2. *Suppose that V is β-smooth. Let $(p_k)_{k\in\mathbb{N}}$ be the iterates of FB–GVI (3.12)–(3.13). Let $\eta > 0$ be such that $\eta \leq \frac{1}{\beta}$. Let $(X_k, X_{k+\frac{1}{2}}) \sim (p_k, p_{k+\frac{1}{2}})$ and $(X_{k+\frac{1}{2}}, X_{k+1}) \sim (p_{k+\frac{1}{2}}, p_k)$ be optimally coupled for the W_2 distance. Then,*

$$\mathbb{E}\left\|X_{k+1} - X_k\right\|^2 \leq -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)]\,.$$

A.2 Eigenvalue control of the iterates

We will show the following eigenvalue bound result:

Lemma A.2.1. *At the k-th iteration of Algorithm 4, suppose that we have $\gamma_0 I \preceq \Sigma_k^{-1} \preceq \gamma_1 I$. As long as $0 \leq \eta \leq \frac{1}{\gamma_1}$ and $\gamma_0 I \preceq S_k \preceq \gamma_1 I$, we then have that*

$$\gamma_1^{-1} I \preceq \Sigma_{k+1} \preceq \gamma_0^{-1} I\,.$$

Proof. Define the monotonically increasing function $f_\eta \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that

$$f_\eta(x) = \frac{1}{2}\left(x + 2\eta + \sqrt{x\,(x+4\eta)}\right).$$

First, we make note of the following algebraic identity. Define $x_\gamma \coloneqq (1-\eta\gamma)^2/\gamma$. Then we have that

$$\begin{aligned} f_\eta(x_\gamma) &= \frac{1}{2}\left(\frac{(1-\eta\gamma)^2}{\gamma} + 2\eta + \sqrt{\left(\frac{(1-\eta\gamma)^2}{\gamma}\right)\left(\frac{(1-\eta\gamma)^2}{\gamma} + 4\eta\right)}\right) \\ &= \frac{1}{2\gamma}\left(1 + \eta^2\gamma^2 + \sqrt{(1-\eta\gamma)^2\,(1+\eta\gamma)^2}\right) \\ &= \frac{1}{2\gamma}\left(1 + \eta^2\gamma^2 + (1-\eta\gamma)\,(1+\eta\gamma)\right) \\ &= \frac{1}{\gamma}. \end{aligned} \tag{A.4}$$

Now, let $\lambda_{\min}(M), \lambda_{\max}(M)$ denote the minimum and maximum eigenvalues of a matrix $M \in \mathbf{S}^d$. The conditions $\eta \leq \gamma_1^{-1}$ and $S_k \preceq \gamma_1 I$ then imply that $I - \eta S_k \succeq 0$. Hence, we then have that

$$\begin{aligned} \lambda_{\min}(\Sigma_{k+\frac{1}{2}}) &= \lambda_{\min}\big((I-\eta S_k)\,\Sigma_k\,(I-\eta S_k)\big) \\ &\geq \lambda_{\min}^2(I-\eta S_k)\,\lambda_{\min}(\Sigma_k) \\ &\geq (1-\eta\gamma_1)^2\,\lambda_{\min}(\Sigma_k) \\ &\geq \frac{(1-\eta\gamma_1)^2}{\gamma_1} \\ &= x_{\gamma_1}. \end{aligned}$$

Now, we also note that $\Sigma_{k+\frac{1}{2}}$ and Σ_{k+1} commute by construction, so since f_η is a monotonically increasing function,

$$\lambda_{\min}(\Sigma_{k+1}) = f_\eta\big(\lambda_{\min}(\Sigma_{k+\frac{1}{2}})\big) \geq f_\eta(x_{\gamma_1}) = \frac{1}{\gamma_1},$$

where the last equality follows from Equation (A.4).

Similarly, for the upper bound, we have that

$$\begin{aligned}
\lambda_{\max}(\Sigma_{k+\frac{1}{2}}) &= \lambda_{\max}\big((I - \eta S_k)\,\Sigma_k\,(I - \eta S_k)\big) \\
&\leq \lambda_{\max}^2(I - \eta S_k)\,\lambda_{\max}(\Sigma_k) && \text{(since } I - \eta S_k \succeq 0\text{)} \\
&\leq (1 - \eta\gamma_0)^2\,\lambda_{\max}(\Sigma_k) \\
&\leq \frac{(1 - \eta\gamma_0)^2}{\gamma_0}\,.
\end{aligned}$$

Thus, we similarly obtain

$$\lambda_{\max}(\Sigma_{k+1}) = f_\eta\big(\lambda_{\max}(\Sigma_{k+\frac{1}{2}})\big) \leq f_\eta(x_{\gamma_0}) = \frac{1}{\gamma_0}\,.$$

Combining the above results, this proves that $\gamma_1^{-1} I \preceq \Sigma_{k+1} \preceq \gamma_0^{-1} I$ which is what we set out to show. □

Note that for (stochastic) FB–GVI, we have $\alpha I \preceq S_k \preceq \beta I$, so Lemma A.2.1 holds with $\gamma_0 = \alpha$ and $\gamma_1 = \beta$. Hence, we obtain the following corollary:

Corollary A.2.2. *Suppose that Algorithm 4 is initialized with a matrix Σ_0 such that $\beta^{-1} I \preceq \Sigma_0$, that V is β-smooth, and that the step size satisfies $\eta \leq \frac{1}{\beta}$. Then $\beta^{-1} I \preceq \Sigma_k$ for all k.*

A.3 Proofs of the noiseless algorithm convergence rates

We obtain the desired convergence rates for FB–GVI by rearranging and iterating the one-step inequality of Lemma 3.3.1. First, we derive inequalities that hold for both the convex and strongly convex cases.

For FB–GVI, we can apply Lemma 3.3.1 with $\nu = \hat{\pi}$, $\eta \leq \frac{1}{\beta}$ and $\sigma_k = 0$. Furthermore, FB–GVI is deterministic, so we may remove the expectations in Lemma 3.3.1. In this case, the inequality in Lemma 3.3.1 implies that for all k,

$$\begin{aligned}
W_2^2(p_{k+1}, \hat{\pi}) &\leq (1 - \alpha\eta)\,W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})) && \text{(by Lemma 3.3.1)} \\
&\leq \exp(-\alpha\eta)\;W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}))\,. && \text{(A.5)}
\end{aligned}$$

Rearranging eq. (A.5), we obtain

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}) \leq \frac{\exp(-\alpha\eta)\, W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})}{2\eta}. \tag{A.6}$$

On the other hand, we can also apply Lemma 3.3.1 with $\nu = p_k$, $\eta \leq \frac{1}{\beta}$ and $\sigma_k^2 = 0$ to obtain that

$$W_2^2(p_{k+1}, p_k) \leq (1 - \alpha\eta)\, W_2^2(p_k, p_k) - 2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)) = -2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)).$$

Hence, rearranging this inequality, we obtain that

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k) \leq -\frac{W_2^2(p_{k+1}, p_k)}{2\eta} \leq 0, \tag{A.7}$$

meaning that the objective value decreases with each iteration of the algorithm.

A.3.1 Proof of Theorem 3.3.2

Proof. Since V is convex, Inequality A.6 holds with the choice $\alpha = 0$, from which we obtain that

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}) \leq \frac{W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})}{2\eta}.$$

Telescoping this inequality, we obtain that

$$\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \frac{1}{N}\sum_{k=1}^{N}[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi})] \leq \frac{1}{2\eta N}\sum_{k=0}^{N-1}[W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})] \leq \frac{W_2^2(p_0, \hat{\pi})}{2\eta N},$$

where the first inequality holds by Inequality A.7. Hence, with the choice

$$\eta = \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{\beta W_2^2(p_0, \hat{\pi})}{\varepsilon^2},$$

we obtain the guarantee $\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2$, proving our desired result. □

A.3.2 Proof of Theorem 3.3.3

Proof. Since $\mathcal{F}(\hat{\pi}) \leq \mathcal{F}(p_{k+1})$ as $\hat{\pi}$ achieves the minimum of $\mathcal{F}$ among Gaussians, we may iterate Inequality A.6 to obtain

$$W_2^2(p_N, \hat{\pi}) \leq \exp(-N\alpha\eta)\, W_2^2(p_0, \hat{\pi}).$$

Hence, with the choice

$$\eta = \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{1}{\alpha\eta} \log \frac{\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2} \asymp \frac{\beta}{\alpha} \log \frac{\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2},$$

we obtain the guarantee $\alpha W_2^2(p_N, \hat{\pi}) \leq \varepsilon^2$.

Now, for the guarantee in KL divergence, we "reinitialize" the algorithm with distribution p_N and apply the convex result of Theorem 3.3.2. With the same choice of N and η and assuming ε is sufficiently small, we can apply Theorem 3.3.2 to obtain the guarantee

$$\mathcal{F}(p_{2N}) - \mathcal{F}(\hat{\pi}) \leq \frac{W_2^2(p_N, \hat{\pi})}{2\eta N} \leq \frac{\varepsilon^2}{2\alpha\eta N} \lesssim \frac{\varepsilon^2}{\log \frac{\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2}} \lesssim \varepsilon^2,$$

proving our desired result. □

A.3.3 Proof of Theorem 3.3.4

First, we need a lemma.

Lemma A.3.1. *Let $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ be such that $\Sigma_{\mu_0}, \Sigma_{\mu_1} \succeq \beta^{-1} I$. Then if $(X_0, X_1) \sim (\mu_0, \mu_1)$ are optimally coupled for the W_2 distance, we have that*

$$\mathbb{E} \left\| \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0)[X_0] \right\|^2 \leq 20\beta^2\, W_2^2(\mu_0, \mu_1).$$

The proof proceeds as follows. First, we apply the triangle inequality and the Cauchy–Schwarz inequality to decompose the LHS into two terms which we will control separately. For the first term, we appeal to the Lipschitzness of $\nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)$, which is possible since $\Sigma_{\mu_1}^{-1} \preceq \beta I$. Then for the second term, we will utilize Lemma 2.2.4 and Lemma 2.2.7 to derive a bound in

terms of $\mathsf{KL}(\mu_0 \,\|\, \mu_1)$, which we can then further bound in terms of $W_2^2(\mu_0, \mu_1)$. Combining these bounds, we obtain our desired result.

Proof. Applying the triangle inequality and Cauchy–Schwarz, we obtain that

$$\frac{1}{2}\,\mathbb{E}\,\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 \le \mathbb{E}\,\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 + \mathbb{E}\,\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 \,.$$

For the first term, we note that since $\Sigma_{\mu_1}^{-1} \preceq \beta I$ by assumption, we have that

$$\begin{aligned}
\mathbb{E}\,\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 &= \mathbb{E}\|\Sigma_{\mu_1}^{-1}(X_1 - X_0)\|^2 && \text{(by Equation (2.24))} \\
&\le \beta^2\,\mathbb{E}\,\|X_1 - X_0\|^2 && \text{(since } \Sigma_{\mu_1}^{-1} \preceq \beta I\text{)} \\
&= \beta^2\, W_2^2(\mu_0, \mu_1)\,, && \text{(A.8)}
\end{aligned}$$

where the last equality holds since $(X_0, X_1) \sim (\mu_0, \mu_1)$ are optimally coupled by assumption. Now, we bound the second term. Define the functionals $\mathcal{V}_1, \mathcal{F}_1 : \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ such that

$$\begin{aligned}
\mathcal{V}_1(\mu) &\coloneqq -\int \log \mu_1(x)\,\mathrm{d}\mu(x)\,, \\
\mathcal{F}_1(\mu) &\coloneqq \mathcal{V}_1(\mu) + \mathcal{H}(\mu)\,.
\end{aligned}$$

Note that by Equation (2.24), $\nabla_{\mathsf{BW}}\mathcal{V}_1(\mu) = -\nabla \log \mu_1 = -\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)$, so that

$$\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu) = \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu) + \nabla_{\mathsf{BW}}\mathcal{H}(\mu) = \nabla_{\mathsf{BW}}\mathcal{H}(\mu) - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)\,.$$

Furthermore, we also note that

$$\mathsf{KL}(\mu \,\|\, \mu_1) = \mathcal{F}_1(\mu) - \mathcal{F}_1(\mu_1)\,. \tag{A.9}$$

Therefore, the second term that we want to control above can be interpreted as the squared norm of $\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_0)$. We will show that $\mathcal{F}_1$ is smooth, which will allow us to bound the squared gradient norm by a multiple of $\mathcal{F}_1(\mu_0) - \mathcal{F}_1(\mu_1) = \mathsf{KL}(\mu_0 \,\|\, \mu_1)$ by the descent lemma from optimization.

Let $\gamma := c^{-1}\beta$, where $c \in (0,1)$ is chosen to satisfy $c \leq (1-c)^2$. Define the random variable X_0' as follows:

$$\begin{aligned}
X_0' &:= X_0 - \frac{1}{\gamma}\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_0)[X_0] \\
&= X_0 - \frac{1}{\gamma}\left(\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0) - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)\right)[X_0] \\
&= X_0 - \frac{1}{\gamma}\left(-\Sigma_{\mu_0}^{-1}(X_0 - m_{\mu_0}) + \Sigma_{\mu_1}^{-1}(X_0 - m_{\mu_1})\right) && \text{(by Equation (2.24))} \\
&= \underbrace{\left(I + \frac{1}{\gamma}\Sigma_{\mu_0}^{-1} - \frac{1}{\gamma}\Sigma_{\mu_1}^{-1}\right)}_{:=M_0} X_0 + \frac{1}{\gamma}\left(-\Sigma_{\mu_0}^{-1}m_{\mu_0} + \Sigma_{\mu_1}^{-1}m_{\mu_1}\right).
\end{aligned}$$

Let $\mu_0' := \mathrm{law}(X_0')$. Since we have $0 \preceq \Sigma_{\mu_0}^{-1}, \Sigma_{\mu_1}^{-1} \preceq \beta I = c\gamma I$ by assumption, we have that

$$M_0 = I + \frac{1}{\gamma}\Sigma_{\mu_0}^{-1} - \frac{1}{\gamma}\Sigma_{\mu_1}^{-1} \succeq I - \frac{1}{\gamma}\Sigma_{\mu_1}^{-1} \succeq (1-c)\, I \succeq 0\,,$$

so X_0' is equal to the gradient of a convex function of X_0. Hence, by Brenier's theorem, we conclude that $(X_0, X_0') \sim (\mu_0, \mu_0')$ are optimally coupled for the W_2 distance. Thus, by Lemma 2.2.4 applied to the potential $\mathcal{V}_1$, we find that

$$\begin{aligned}
\mathcal{V}_1(\mu_0') - \mathcal{V}_1(\mu_0) &\leq \mathbb{E}\left\langle \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu_0)[X_0], X_0' - X_0\right\rangle + \frac{\beta}{2}\mathbb{E}\left\|X_0' - X_0\right\|^2 && \text{(since } -\nabla^2 \log \mu_0 \preceq \beta I\text{)} \\
&= -\mathbb{E}\left\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0], X_0' - X_0\right\rangle + \frac{\beta}{2}\mathbb{E}\left\|X_0' - X_0\right\|^2. && \text{(A.10)}
\end{aligned}$$

Additionally, we note that since $\beta = c\gamma \leq (1-c)^2\,\gamma$, we have that

$$\Sigma_{\mu_0'} = M_0\Sigma_{\mu_0}M_0 \succeq (1-c)^2\,\Sigma_{\mu_0} \succeq \frac{(1-c)^2}{\beta}\, I \succeq \frac{1}{\gamma}\, I\,.$$

This implies that $\Sigma_{\mu_0'}^{-1}, \Sigma_{\mu_0}^{-1} \preceq \gamma I$. Hence, we can also apply the geodesic smoothness inequality of Lemma 2.2.7 to obtain

$$\mathcal{H}(\mu_0') - \mathcal{H}(\mu_0) \leq \mathbb{E}\left\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0], X_0' - X_0\right\rangle + \frac{\gamma}{2}\mathbb{E}\left\|X_0' - X_0\right\|^2. \qquad \text{(A.11)}$$

Hence, combining Equation (A.9) with Inequality A.10 and Inequality A.11, we obtain that

$$
\begin{aligned}
-\mathsf{KL}\left(\mu_0 \,\|\, \mu_1\right) &\leq \mathsf{KL}\left(\mu_0' \,\|\, \mu_1\right) - \mathsf{KL}\left(\mu_0 \,\|\, \mu_1\right) && \text{(since } \mathsf{KL}\left(\mu_0' \,\|\, \mu_1\right) \geq 0\text{)} \\
&= \mathcal{F}_1(\mu_0') - \mathcal{F}_1(\mu_0) && \text{(by Equation (A.9))} \\
&= \left[\mathcal{V}_1(\mu_0') - \mathcal{V}_1(\mu_0)\right] + \left[\mathcal{H}(\mu_0') - \mathcal{H}(\mu_0)\right] \\
&\leq -\mathbb{E}\left\langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_0], X_0' - X_0 \right\rangle + \frac{\beta}{2}\,\mathbb{E}\left\|X_0' - X_0\right\|^2 && \text{(by Inequality A.10)} \\
&\quad + \mathbb{E}\left\langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0)[X_0], X_0' - X_0 \right\rangle + \frac{\gamma}{2}\,\mathbb{E}\left\|X_0' - X_0\right\|^2 && \text{(by Inequality A.11)} \\
&= \left(-\frac{1}{\gamma} + \frac{\beta}{2\gamma^2} + \frac{1}{2\gamma}\right) \mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_0]\right\|^2 && \text{(definition of } X_0'\text{)} \\
&= -\frac{1-c}{2\gamma}\,\mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_0]\right\|^2 . && \text{(A.12)}
\end{aligned}
$$

To bound the LHS of this inequality, we again apply Lemma 2.2.4 to the potential $\mathcal{V}_1$ as well as Lemma 2.2.7 to $\mathcal{H}$ to obtain

$$
\begin{aligned}
\mathsf{KL}\left(\mu_0 \,\|\, \mu_1\right) &= \mathcal{F}_1(\mu_0) - \mathcal{F}_1(\mu_1) \\
&= \left[\mathcal{V}_1(\mu_0) - \mathcal{V}_1(\mu_1)\right] + \left[\mathcal{H}(\mu_0) - \mathcal{H}(\mu_1)\right] \\
&\leq \mathbb{E}\left\langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu_1)[X_1], X_0 - X_1 \right\rangle + \frac{\beta}{2}\,\mathbb{E}\left\|X_0 - X_1\right\|^2 \\
&\qquad\qquad \text{(by Lemma 2.2.4 since } -\nabla^2 \log \mu_1 \preceq \beta I\text{)} \\
&\quad + \mathbb{E}\left\langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_1], X_0 - X_1 \right\rangle + \frac{\beta}{2}\,\mathbb{E}\left\|X_0 - X_1\right\|^2 \\
&\qquad\qquad \text{(by Lemma 2.2.7 since } \Sigma_{\mu_0}^{-1}, \Sigma_{\mu_1}^{-1} \preceq \beta I\text{)} \\
&= \beta\,\mathbb{E}\left\|X_0 - X_1\right\|^2 && \text{(since } \nabla_{\mathsf{BW}} \mathcal{V}_1(\mu_1) + \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1) = \nabla_{\mathsf{BW}} \mathcal{F}_1(\mu_1) = 0\text{)} \\
&= \beta\, W_2^2(\mu_0, \mu_1) . && \text{(A.13)}
\end{aligned}
$$

Finally, choosing $c = \frac{1}{3}$ so that $c \leq (1-c)^2$ and combining our above inequalities, we find that

$$
\begin{aligned}
\frac{1}{2}\,\mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0)[X_0]\right\|^2 &\leq \mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_0]\right\|^2 \\
&\quad + \mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}} \mathcal{H}(\mu_1)[X_0]\right\|^2 \\
&\leq \beta^2\, W_2^2(\mu_0, \mu_1) + \frac{2\gamma}{1-c}\,\mathsf{KL}\left(\mu_0 \,\|\, \mu_1\right) \\
&\qquad\qquad \text{(by Inequality A.8 and Inequality A.12)}
\end{aligned}
$$

$$\leq 10\beta^2 W_2^2(\mu_0, \mu_1)\,. \qquad \text{(by Inequality A.13)}$$

Rearranging, we obtain our desired result. □

With this result in mind, we are ready to prove our desired stationary point guarantee.

Proof. Let $(X_k, X_{k+\frac{1}{2}}) \sim (p_k, p_{k+\frac{1}{2}})$ and $(X_{k+\frac{1}{2}}, X_{k+1}) \sim (p_{k+\frac{1}{2}}, p_k)$ be optimally coupled for the W_2 distance, noting as in the proof of Lemma 3.3.1 that by construction,

$$\frac{X_k - X_{k+1}}{\eta} = \nabla_{\mathsf{BW}} \mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}]\,.$$

Applying Lemma A.1.2, we obtain that

$$\mathbb{E}\,\|X_{k+1} - X_k\|^2 \leq -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)]\,.$$

Telescoping this inequality, we find that

$$\begin{aligned}
\min_{k\in\{0,\ldots,N-1\}} \mathbb{E}\,\|X_{k+1} - X_k\|^2 &\leq \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\,\|X_{k+1} - X_k\|^2 \\
&\leq -\frac{2\eta}{N}\sum_{k=0}^{N-1} \mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)] \\
&= -\frac{2\eta}{N}\,\mathbb{E}[\mathcal{F}(p_N) - \mathcal{F}(p_0)] \\
&\leq \frac{2\eta\Delta}{N}\,. \qquad \text{(A.14)}
\end{aligned}$$

Now, let $(X_k, X_{k+1}^\star) \sim (p_k, p_{k+1})$ be optimally coupled for the W_2 distance. By Corollary A.2.2, we have that $\Sigma_k^{-1} \preceq \beta I$ for all k, meaning that we can apply Lemma A.3.1 with $\mu_0 = p_k$ and $\mu_1 = p_{k+1}$ to obtain that

$$\mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(p_k)[X_k] - \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}^\star]\right\|^2 \leq 20\beta^2 W_2^2(p_k, p_{k+1})\,. \qquad \text{(A.15)}$$

Furthermore, we have that

$$\mathbb{E}\left\|\nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}^\star] - \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}]\right\|^2 = \mathbb{E}\left\|\Sigma_{k+1}^{-1}\,(X_{k+1}^\star - X_{k+1})\right\|^2$$

$$
\begin{aligned}
&\leq \beta^2 \, \mathbb{E} \left\| X^\star_{k+1} - X_{k+1} \right\|^2 \\
&\leq 2\beta^2 \, \mathbb{E} \left\| X^\star_{k+1} - X_k \right\|^2 + 2\beta^2 \, \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 \\
&= 2\beta^2 \, W_2^2(p_k, p_{k+1}) + 2\beta^2 \, \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 .
\end{aligned} \tag{A.16}
$$

With these inequalities in mind, we obtain that

$$
\begin{aligned}
\frac{1}{3} \left\| \nabla_{\mathsf{BW}} \mathcal{F}(p_k) \right\|_{p_k}^2 &= \frac{1}{3} \, \mathbb{E} \left\| \nabla_{\mathsf{BW}} \mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}} \mathcal{H}(p_k)[X_k] \right\|^2 \\
&\leq \mathbb{E} \left\| \nabla_{\mathsf{BW}} \mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}] \right\|^2 + \mathbb{E} \left\| \nabla_{\mathsf{BW}} \mathcal{H}(p_k)[X_k] - \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X^\star_{k+1}] \right\|^2 \\
&\quad + \mathbb{E} \left\| \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X^\star_{k+1}] - \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}] \right\|^2 \\
&\qquad\qquad \text{(by triangle inequality)} \\
&\leq \frac{1}{\eta^2} \, \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 + 22 \, \beta^2 W_2^2(p_k, p_{k+1}) + 2\beta^2 \, \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 \\
&\qquad\qquad \text{(by Inequality A.15 and Inequality A.16)} \\
&\leq \left(\frac{1}{\eta^2} + 24\beta^2 \right) \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 \quad \text{(since } (X_k, X_{k+1}) \text{ is a coupling of } (p_k, p_{k+1})\text{)} \\
&\leq \frac{25}{\eta^2} \, \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 . \qquad \text{(since } \beta \leq \eta^{-1}\text{)}
\end{aligned}
$$

Combining the above with Inequality A.14, we obtain that

$$
\min_{k \in \{0, \ldots, N-1\}} \left\| \nabla_{\mathsf{BW}} \mathcal{F}(p_k) \right\|_{p_k}^2 \leq \min_{k \in \{0, \ldots, N-1\}} \frac{75}{\eta^2} \, \mathbb{E} \left\| X_{k+1} - X_k \right\|^2 \leq \frac{150\Delta}{\eta N} .
$$

Finally, taking $\eta = \frac{1}{\beta}$ and $N \geq \frac{150\beta\Delta}{\varepsilon^2}$, we obtain that

$$
\min_{k \in \{0, \ldots, N-1\}} \left\| \nabla_{\mathsf{BW}} \mathcal{F}(p_k) \right\|_{p_k}^2 \leq \varepsilon^2 ,
$$

as desired. □

A.4 Proofs of the noisy algorithm convergence rates

We once again utilize Lemma 3.3.1 to obtain our desired rates of convergence. First, we must prove the bound on σ_k for Stochastic FB–GVI given in Lemma 3.3.5.

A.4.1 Proof of Lemma 3.3.5

Proof. Let $\mu = \mathcal{N}(m, \Sigma)$ be an element of $\mathsf{BW}(\mathbb{R}^d)$. We first note that if $X \sim \mu$, then by integration by parts,

$$\begin{aligned}
\Sigma\, \mathbb{E}\nabla^2 V(X) &= \Sigma \int \nabla^2 V \, \mathrm{d}\mu \\
&= -\Sigma \int \nabla \mu \otimes \nabla V && \text{(integration by parts)} \\
&= -\Sigma \int \nabla \ln \mu \otimes \nabla V \, \mathrm{d}\mu \\
&= \int (x - m) \otimes \nabla V \, \mathrm{d}\mu(x) && \text{(since } -\Sigma \nabla \ln \mu(x) = x - m\text{)} \\
&= \mathbb{E}[(X - m) \otimes \nabla V(X)] \,. && \text{(A.17)}
\end{aligned}$$

Hence,

$$\begin{aligned}
\langle \mathbb{E}\nabla^2 V(X), \Sigma \rangle &= \langle \mathbb{E}[\Sigma^{-1}\,(X - m) \otimes \nabla V(X)], \Sigma \rangle && \text{(by Equation (A.17))} \\
&= \mathbb{E}\langle \Sigma^{-1}\,(X - m) \otimes \nabla V(X), \Sigma \rangle && \text{(linearity of expectation and trace)} \\
&= \mathbb{E}\langle \nabla V(X), X - m \rangle \,. && \text{(cyclicity of trace)}
\end{aligned}$$

Now, let $(X_k, Z) \sim (p_k, \hat{\pi})$ be optimally coupled for the W_2 distance and independent of $\hat{X}_k$. Recall also the Brascamp–Lieb inequality [17]: if μ is a measure on $\mathbb{R}^d$ with density $\mu \propto \exp(-W)$, where W is twice continuously differentiable and strictly convex, then for any smooth test function $f : \mathbb{R}^d \to \mathbb{R}$ it holds that $\mathrm{Var}_\mu(f) \le \mathbb{E}\langle \nabla f, (\nabla^2 W)^{-1}\, \nabla f \rangle$. In particular, if we take $f = \langle \nabla V, e \rangle$ for a unit vector e and $\mu = p_k$, it follows that $\mathrm{Var}_{p_k} \langle \nabla V, e \rangle \le \mathbb{E}_{p_k} \langle e, \nabla^2 V\, \Sigma_k\, \nabla^2 V\, e \rangle$. Summing this inequality as e ranges over an orthonormal basis of $\mathbb{R}^d$, we obtain

$$\mathbb{E}_{p_k} \|\nabla V - \mathbb{E}_{p_k} \nabla V\|^2 \le \mathbb{E}_{p_k} \langle [\nabla^2 V]^2, \Sigma_k \rangle \,.$$

Thus, we get that

$$\begin{aligned}
\frac{1}{2}\,\sigma_k^2 &\le \mathbb{E}\|(\nabla^2 V(\hat{X}_k) - \mathbb{E}_{p_k}\nabla^2 V)(X_k - m_k)\|^2 + \mathbb{E}\|\nabla V(\hat{X}_k) - \mathbb{E}_{p_k}\nabla V\|^2 && \text{(by triangle inequality)} \\
&= \left\langle \mathbb{E}_{p_k}[(\nabla^2 V - \mathbb{E}_{p_k}\nabla^2 V)^2], \Sigma_k \right\rangle + \mathbb{E}_{p_k} \left\|\nabla V - \mathbb{E}_{p_k}\nabla V\right\|^2 && \text{(since } X_k \perp\!\!\!\perp \hat{X}_k\text{)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle - \langle \mathbb{E}_{p_k}[\nabla^2 V]^2, \Sigma_k\rangle + \mathbb{E}_{p_k}\left\|\nabla V - \mathbb{E}_{p_k}\nabla V\right\|^2 \\
&\le \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle + \mathbb{E}_{p_k}\left\|\nabla V - \mathbb{E}_{p_k}\nabla V\right\|^2 && \text{(since } \langle \mathbb{E}_{p_k}[(\nabla^2 V)^2], \Sigma_k\rangle \ge 0\text{)} \\
&\le 2\,\mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle && \text{(by Brascamp–Lieb)} \\
&\le 2\beta\,\mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\rangle && \text{(since } \nabla^2 V \preceq \beta I \text{ and } \nabla^2 V, \Sigma_k \succeq 0\text{)} \\
&= 2\beta\,\mathbb{E}\,\langle \nabla V(X_k), X_k - m_k\rangle && \text{(by Equation (A.17))} \\
&= 2\beta\,\underbrace{\mathbb{E}\,\langle \nabla V(Z), Z - \hat{m}\rangle}_{\mathsf{err}_1} + 2\beta\,\underbrace{\mathbb{E}\,\langle \nabla V(X_k) - \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle}_{\mathsf{err}_2} \\
&\quad + 2\beta\,\underbrace{\mathbb{E}\,\langle \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle}_{\mathsf{err}_3} + 2\beta\,\underbrace{\mathbb{E}\,\langle \nabla V(X_k) - \nabla V(Z), Z - \hat{m}\rangle}_{\mathsf{err}_4}.
\end{aligned}
$$

Now, we have the following:

$$
\begin{aligned}
\mathsf{err}_1 &= \mathbb{E}\,\langle \nabla V(Z), Z - \hat{m}\rangle = \langle \mathbb{E}\nabla^2 V(Z), \hat{\Sigma}\rangle = \mathrm{Tr}(I) \\
&\qquad\qquad \text{(by Equation (A.17) and the stationarity conditions in (2.25))} \\
&= d\,, \\
\mathsf{err}_2 &= \mathbb{E}\,\langle \nabla V(X_k) - \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle \\
&\le \frac{1}{2\beta}\,\mathbb{E}\,\|\nabla V(X_k) - \nabla V(Z)\|^2 + \frac{\beta}{2}\,\mathbb{E}\,\|(X_k - m_k) - (Z - \hat{m})\|^2 && \text{(Young's inequality)} \\
&\le \beta\,\mathbb{E}\,\|X_k - Z\|^2 && \text{(since } \nabla V \text{ is } \beta\text{-Lipschitz)} \\
&= \beta\, W_2^2(\mu_k, \hat{\pi})\,, && \text{(since } (X_k, Z) \text{ are optimally coupled)} \\
\mathsf{err}_3 &= \mathbb{E}\,\langle \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle \\
&\le \frac{1}{4\beta}\,\mathbb{E}\,\|\nabla V(Z)\|^2 + \beta\,\mathbb{E}\,\|(X_k - m_k) - (Z - \hat{m})\|^2 && \text{(Young's inequality)} \\
&\le \frac{1}{4\beta}\,\mathbb{E}\langle \nabla^2 V(Z)^2, \hat{\Sigma}\rangle + \beta\, W_2^2(\mu_k, \hat{\pi}) && \text{(Brascamp–Lieb, optimal coupling of } (X_k, Z)\text{)} \\
&\le \frac{d}{4} + \beta\, W_2^2(\mu_k, \hat{\pi})\,, && \text{(since } \mathbb{E}_{\hat{\pi}}\nabla^2 V = \hat{\Sigma}^{-1} \text{ by Equation (2.25) and } \nabla^2 V \preceq \beta I\text{)} \\
\mathsf{err}_4 &= \mathbb{E}\,\langle \nabla V(X_k) - \nabla V(Z), Z - \hat{m}\rangle \\
&\le \frac{\mathrm{Tr}(\hat{\Sigma})}{d}\,\mathbb{E}\,\|\nabla V(X_k) - \nabla V(Z)\|^2 + \frac{d}{4\,\mathrm{Tr}(\hat{\Sigma})}\,\mathbb{E}\,\|Z - \hat{m}\|^2 && \text{(Young's inequality)} \\
&\le \frac{\beta^2\,\mathrm{Tr}(\hat{\Sigma})}{d}\,\mathbb{E}\,\|X_k - Z\|^2 + \frac{d}{4\,\mathrm{Tr}(\hat{\Sigma})}\,\mathrm{Tr}(\hat{\Sigma}) && \text{(since } \nabla V \text{ is } \beta\text{-Lipschitz)}
\end{aligned}
$$

$$\leq \frac{\beta^2 \operatorname{Tr}(\hat{\Sigma})}{d} W_2^2(\mu_k, \hat{\pi}) + \frac{d}{4}.$$

Combining these, we obtain that

$$\sigma_k^2 \leq 4\beta \sum_{i=1}^{4} \mathsf{err}_i \leq 6\beta d + \left(8\beta^2 + \frac{4\beta^3 \operatorname{Tr}(\hat{\Sigma})}{d}\right) W_2^2(\mu_k, \hat{\pi}) \leq 6\beta d + 12\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(\mu_k, \hat{\pi}).$$

(since $\hat{\Sigma}^{-1} = \mathbb{E}_{\hat{\pi}} \nabla^2 V \preceq \beta I$ so $\lambda_{\max}(\hat{\Sigma}) \geq 1/\beta$)

Note that in the strongly convex case, by Equation (2.25), we obtain that

$$\lambda_{\max}(\hat{\Sigma}) = \lambda_{\max}(\mathbb{E}_{\hat{\pi}}[\nabla^2 V]^{-1}) \leq \frac{1}{\alpha},$$

so this bound simplifies to

$$\sigma_k^2 \leq 6\beta d + \frac{12\beta^3}{\alpha} W_2^2(\mu_k, \hat{\pi}).$$

This concludes our proof. □

A.4.2 One-step inequality using the bound on σ_k

We apply the error bound in Lemma 3.3.5 along with the one-step inequality of Lemma 3.3.1 with $\nu = \hat{\pi}$ and $\eta \leq \frac{1}{2\beta}$. This gives us the inequality

$$\begin{aligned}
\mathbb{E} W_2^2(p_{k+1}, \hat{\pi}) &\leq (1 - \alpha\eta)\mathbb{E} W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 2\eta^2\, \mathbb{E}\sigma_k^2 \\
&\leq \left(1 - \alpha\eta + 24\beta^3\eta^2\lambda_{\max}(\hat{\Sigma})\right) \mathbb{E} W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d \\
&\leq \exp\left(-\alpha\eta + 24\beta^3\eta^2\lambda_{\max}(\hat{\Sigma})\right) \mathbb{E} W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d.
\end{aligned} \tag{A.18}$$

A.4.3 Proof of Theorem 3.3.6

Proof. Define $c := 24\beta^3\lambda_{\max}(\hat{\Sigma})$. Since V is convex by assumption, we may take $\alpha = 0$ in Inequality A.18 to obtain that

$$2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) \leq e^{c\eta^2}\, \mathbb{E} W_2^2(p_k, \hat{\pi}) - \mathbb{E} W_2^2(p_{k+1}, \hat{\pi}) + 12\beta\eta^2 d.$$

Define $S_N(\eta) := \sum_{k=1}^{N} e^{-kc\eta^2}$. We then find that

$$\begin{aligned}\sum_{k=0}^{N-1} 2\eta\, e^{-(k+1)c\eta^2} \left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) &\leq \sum_{k=0}^{N-1} e^{-(k+1)c\eta^2} \left(e^{c\eta^2}\, \mathbb{E}W_2^2(p_k, \hat{\pi}) - \mathbb{E}W_2^2(p_{k+1}, \hat{\pi}) + 12\beta\eta^2 d\right) \\ &= W_2^2(p_0, \hat{\pi}) - e^{-Nc\eta^2}\, \mathbb{E}W_2^2(p_N, \hat{\pi}) + 12\beta\eta^2 d \sum_{k=0}^{N-1} e^{-(k+1)c\eta^2} \\ &\leq W_2^2(p_0, \hat{\pi}) + 12\beta\eta^2 d S_N(\eta)\,.\end{aligned}$$

Let $\overline{p}$ be drawn randomly from among $\{p_k\}_{k=1}^{N}$, with probability of choosing p_k proportional to $e^{-kc\eta^2}$. Then we have that

$$\begin{aligned}\mathbb{E}\mathcal{F}(\overline{p}) - \mathcal{F}(\hat{\pi}) &= \frac{1}{2\eta S_N(\eta)} \sum_{k=0}^{N-1} 2\eta\, e^{-(k+1)c\eta^2} \left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) \\ &\leq \frac{1}{2\eta S_N(\eta)} \left(W_2^2(p_0, \hat{\pi}) + 12\beta\eta^2 d S_N(\eta)\right) \\ &= \frac{W_2^2(p_0, \hat{\pi})}{2\eta S_N(\eta)} + 6\beta\eta d\,.\end{aligned}$$

Now, we note that

$$S_N(\eta) = \sum_{k=1}^{N} e^{-kc\eta^2} \geq \sum_{k=1}^{N\wedge(c\eta^2)^{-1}} e^{-kc\eta^2} \geq \sum_{k=1}^{N\wedge(c\eta^2)^{-1}} e^{-1} \geq \frac{N \wedge \lfloor (c\eta^2)^{-1} \rfloor}{e}\,.$$

Thus, we obtain the inequality

$$\begin{aligned}\mathbb{E}\big[\min_{k\in\{1,\ldots,N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) &\leq \mathbb{E}\mathcal{F}(\overline{p}) - \mathcal{F}(\hat{\pi}) \\ &\leq \frac{W_2^2(p_0, \hat{\pi})}{2\eta S_N(\eta)} + 6\beta\eta d \\ &\leq \frac{2W_2^2(p_0, \hat{\pi})}{\eta\left(N \wedge \lfloor (c\eta^2)^{-1} \rfloor\right)} + 6\beta\eta d \\ &\lesssim \frac{W_2^2(p_0, \hat{\pi})}{\eta N} + c\eta W_2^2(p_0, \hat{\pi}) + \beta\eta d\,.\end{aligned}$$

Hence, taking

$$\eta \asymp \frac{\varepsilon^2}{cW_2^2(p_0, \hat{\pi}) \vee \beta d} \asymp \frac{\varepsilon^2}{\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(p_0, \hat{\pi}) \vee \beta d}\,,$$

$$N \gtrsim \frac{W_2^2(p_0, \hat{\pi})}{\eta\varepsilon^2} \asymp \frac{W_2^2(p_0, \hat{\pi})}{\varepsilon^4} \left(\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(p_0, \hat{\pi}) \vee \beta d\right),$$

we get the guarantee $\mathbb{E}\left[\min_{k\in\{1,\dots,N\}} \mathcal{F}(p_k)\right] - \mathcal{F}(\hat{\pi}) \le \varepsilon^2$. □

A.4.4 Proof of Theorem 3.3.7

Proof. In the strongly convex case where $0 \prec \alpha I \preceq \nabla^2 V$, we have the eigenvalue guarantee $\lambda_{\max}(\hat{\Sigma}) \le \frac{1}{\alpha} I$, since $\mathbb{E}_\pi \nabla^2 V = \hat{\Sigma}^{-1}$ by (2.25). Hence, under the assumption that $\eta \le \frac{\alpha^2}{48\beta^3}$, Inequality A.18 implies that

$$\begin{aligned}\mathbb{E}W_2^2(p_{k+1}, \hat{\pi}) &\le \exp\left(-\alpha\eta + \frac{24\beta^3\eta^2}{\alpha}\right) \mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d \\ &\le \exp\left(-\frac{\alpha\eta}{2}\right) \mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d\,.\end{aligned}$$

Since $\mathcal{F}(\hat{\pi}) \le \mathcal{F}(p_{k+1})$, we may iterate this inequality to obtain that

$$\mathbb{E}W_2^2(p_N, \hat{\pi}) \le \exp\left(-\frac{N\alpha\eta}{2}\right) W_2^2(p_0, \hat{\pi}) + \frac{24\beta\eta d}{\alpha}\,.$$

Hence, with the choice

$$\eta \asymp \frac{\varepsilon^2}{\beta d}, \quad \text{and} \quad N \gtrsim \frac{1}{\alpha\eta} \log \frac{\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2} \asymp \frac{\beta d}{\alpha\varepsilon^2} \log \frac{\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2},$$

we obtain the guarantee $\alpha\, \mathbb{E}W_2^2(p_N, \hat{\pi}) \le \varepsilon^2$. Now, for the guarantee in KL divergence, we "reinitialize" the algorithm with distribution p_N and apply the convex result of Theorem 3.3.6. Assuming ε is sufficiently small, we get that

$$c\, \mathbb{E}W_2^2(p_N, \hat{\pi}) \le \frac{c\varepsilon^2}{\alpha} \le \beta d\,,$$

meaning that for the above choice of η, we have

$$\eta \asymp \frac{\varepsilon^2}{\beta d} \asymp \frac{\varepsilon^2}{c\, \mathbb{E}W_2^2(p_N, \hat{\pi}) \vee \beta d}\,.$$

Furthermore, for our choice of N, we have that

$$\frac{\mathbb{E}W_2^2(p_N, \hat{\pi})}{\varepsilon^4}\left(cW_2^2(p_0, \hat{\pi}) \vee \beta d\right) \leq \frac{\beta d}{\alpha \varepsilon^2} \lesssim N\,.$$

Thus, applying Theorem 3.3.6 with our choice of step size η and iteration count N, we obtain that

$$\begin{aligned}\mathbb{E}\big[\min_{k\in\{1,\dots,2N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) &\leq \mathbb{E}\big[\min_{k\in\{N+1,\dots,2N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \\ &\lesssim \mathbb{E}\Big[\frac{W_2^2(p_N, \hat{\pi})}{\eta N} + c\eta\, W_2^2(p_N, \hat{\pi}) + \beta\eta d\Big] \\ &\lesssim \varepsilon^2\,,\end{aligned}$$

proving our desired result. □

A.5 Proofs for VRFB–GVI

A.5.1 Proof of Lemma 3.3.8

Recall that in the setting of Prox-SVRG, the inequality Lemma 3.1.14 was key to deriving a variance bound. Hence, our first goal is to translate this inequality to the setting of the BW space.

Lemma A.5.1. *Suppose that $V_i\colon \mathbb{R}^d \to \mathbb{R}$ are convex and β-smooth for all $i \in [m]$. Let $\mu \in \mathsf{BW}(\mathbb{R}^d)$, and suppose that $(X, Z) \sim (\mu, \hat{\pi})$ are coupled along a generalized geodesic, so that $Z = R(X)$ where $R\colon \mathbb{R}^d \to \mathbb{R}^d$ is the composition of at most two optimal transport maps. Then*

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)[X] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 \leq 2\beta[\mathcal{F}(\mu) - \mathcal{F}(\hat{\pi})].$$

Proof. To begin, we show that

$$\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)[X] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 \leq \mathbb{E}\left\|\nabla_{W_2}\mathcal{V}_i(\mu)[X] - \nabla_{W_2}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2. \tag{A.19}$$

Since $\nabla_{W_2}\mathcal{V}_i(\mu) = \nabla V$ independently of μ, the right-hand side can then be bounded analogously to the Euclidean case (Lemma 3.1.14).

For shorthand, define the function perp_i such that

$$\mathsf{perp}_i(\mu) := \nabla_{W_2}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\mu).$$

Note that by the definition of the BW gradient in Section 2.2.3, we have that for any affine function $h\colon \mathbb{R}^d \to \mathbb{R}^d$,

$$\langle \mathsf{perp}_i(\mu), h\rangle_\mu = \langle \nabla_{W_2}\mathcal{V}_i(\mu), h\rangle_\mu - \langle \nabla_{\mathsf{BW}}\mathcal{V}_i(\mu), h\rangle_\mu = 0.$$

Now, we have that

$$\begin{aligned}
\mathbb{E}\,\|\nabla_{W_2}\mathcal{V}_i(\mu)[X] - \nabla_{W_2}\mathcal{V}_i(\hat{\pi})[Z]\|^2 &= \|\nabla_{W_2}\mathcal{V}_i(\mu) - \nabla_{W_2}\mathcal{V}_i(\hat{\pi})\circ R\|_\mu^2 && \text{(defn of } R\text{)}\\
&= \|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})\circ R\|_\mu^2 + \|\mathsf{perp}_i(\mu) - \mathsf{perp}_i(\hat{\pi})\circ R\|_\mu^2\\
&\quad + 2\,\langle \nabla_{\mathsf{BW}}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})\circ R, \mathsf{perp}_i(\mu) - \mathsf{perp}_i(\hat{\pi})\circ R\rangle_\mu\,.
\end{aligned}$$

Now, since by assumption (X, Z) are coupled along a generalized geodesic, R is an affine function, meaning that $\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)\circ R$ is an affine function as well. Hence, we find that

$$\langle \nabla_{\mathsf{BW}}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})\circ R, \mathsf{perp}_i(\mu)\rangle_\mu = 0.$$

Similarly, we have that R^{-1} is an affine function, meaning that $\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)\circ R^{-1} - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})$ is affine as well. Thus, we also find that

$$\langle \nabla_{\mathsf{BW}}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})\circ R, \mathsf{perp}_i(\hat{\pi})\circ R\rangle_\mu = \left\langle \nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)\circ R^{-1} - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi}), \mathsf{perp}_i(\hat{\pi})\right\rangle_{\hat{\pi}} = 0.$$

Thus, we deduce that

$$\begin{aligned}
\mathbb{E}\,\|\nabla_{W_2}\mathcal{V}_i(\mu)[X] - \nabla_{W_2}\mathcal{V}_i(\hat{\pi})[Z]\|^2 &= \|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})\circ R\|_\mu^2 + \|\mathsf{perp}_i(\mu) - \mathsf{perp}_i(\hat{\pi})\circ R\|_\mu^2\\
&\geq \|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu) - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})\circ R\|_\mu^2\\
&= \mathbb{E}\,\|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)[X] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\|^2\,,
\end{aligned}$$

proving Inequality A.19.

With Inequality A.19 in mind, we proceed to prove the desired claim. We have that

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla_{W_2}\mathcal{V}_i(\mu)[X]-\nabla_{W_2}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla V_i(X)-\nabla V_i(Z)\right\|^2 \\
&\leq \frac{1}{m}\sum_{i=1}^{m}2\beta\mathbb{E}[V_i(X)-V_i(Z)-\langle\nabla V_i(Z),X-Z\rangle] \\
&\qquad\qquad \text{(by Lemma 3.1.14)} \\
&= 2\beta[\mathcal{V}(\mu)-\mathcal{V}(\hat{\pi})-\langle\nabla_{W_2}\mathcal{V}(\hat{\pi}),R^{-1}-\mathrm{id}\rangle_{\hat{\pi}}] \\
&= 2\beta[\mathcal{V}(\mu)-\mathcal{V}(\hat{\pi})-\langle\nabla_{\mathsf{BW}}\mathcal{V}(\hat{\pi}),R^{-1}-\mathrm{id}\rangle_{\hat{\pi}}] \\
&\qquad\qquad \text{(since } R^{-1}-\mathrm{id} \text{ is affine)} \\
&= 2\beta[\mathcal{V}(\mu)-\mathcal{V}(\hat{\pi})+\langle\nabla_{\mathsf{BW}}\mathcal{H}(\hat{\pi}),R^{-1}-\mathrm{id}\rangle_{\hat{\pi}}] \\
&\qquad\qquad \text{(since } \nabla_{\mathsf{BW}}\mathcal{V}(\hat{\pi})+\nabla_{\mathsf{BW}}\mathcal{H}(\hat{\pi})=0) \\
&\leq 2\beta[\mathcal{V}(\mu)-\mathcal{V}(\hat{\pi})+\mathcal{H}(\mu)-\mathcal{H}(\hat{\pi})] \quad \text{(by Lemma 2.2.2)} \\
&= 2\beta[\mathcal{F}(\mu)-\mathcal{F}(\hat{\pi})].
\end{aligned}$$

Note that the last inequality was where we crucially used the fact that $(X,Z)\sim(\mu,\hat{\pi})$ are coupled along a generalized geodesic: it is only in this case that we can invoke the convexity of entropy. Combining this inequality with Inequality A.19, we obtain that

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(\mu)[X]-\nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla_{W_2}\mathcal{V}_i(\mu)[X]-\nabla_{W_2}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 \\
&= 2\beta[\mathcal{F}(\mu)-\mathcal{F}(\hat{\pi})],
\end{aligned}$$

as desired. □

With this result in mind, we are ready to prove our desired variance bound.

Proof. For notational convenience, we drop the dependence on j. Let $(X_k,X_0)\sim(p_k,p_0)$ and $(X_k,Z)\sim(p_k,\hat{\pi})$ be optimally coupled for the W_2 distance. Importantly, we note that (X_0,Z) are coupled along a generalized geodesic. We have that

$$\frac{1}{3}\sigma_k^2=\frac{1}{3}\mathbb{E}\left\|e_k\right\|_{p_k}^2$$

$$
\begin{aligned}
&= \frac{1}{3}\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_k) - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0) + \nabla_{\mathsf{BW}}\mathcal{V}(p_0) - \nabla_{\mathsf{BW}}\mathcal{V}(p_k)\right\|_{p_k}^2 \\
&\leq \frac{1}{3}\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_k) - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)\right\|_{p_k}^2 && \text{(since unbiased estimate)} \\
&\leq \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_k)[X_k] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 \\
&\quad + \frac{1}{2}\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)[X_k] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 && \text{(by triangle ineq and Cauchy-Schwarz)} \\
&\leq \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_k)[X_k] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 \\
&\quad + \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)[X_k] - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)[X_0]\right\|^2 + \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{V}_i(\hat{\pi})[Z]\right\|^2 \\
&&& \text{(triangle ineq and C-S again)} \\
&\leq 2\beta[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi})] && \text{(by Lemma A.5.1)} \\
&\quad + \beta^2\mathbb{E}\left\|X_k - X_0\right\|^2 + 2\beta[\mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})] && \text{(since } \nabla^2 V_i \preceq \beta I \text{ and by Lemma A.5.1)} \\
&\leq 2\beta[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})] + \beta^2 W_2^2(p_k, p_0). && \text{(since } (X_k, X_0) \text{ are opt. coupled)}
\end{aligned}
$$

Rescaling both sides of the inequality and adding back in the dependence on j, we obtain our desired result. □

A.5.2 Proof of Lemma 3.3.8

The proof of this variance bound is essentially a combination of the arguments laid out in Appendix A.5.1 and Appendix A.4.1.

Proof. For notational convenience, we again drop the dependence on j. Let $i \in \text{Unif}[m]$ be the random index chosen at iteration k. Then we have that

$$
\mathbb{E}[e_k \mid i] = \nabla_{\mathsf{BW}}\mathcal{V}_i(p_k) - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0) + \nabla_{\mathsf{BW}}\mathcal{V}(p_0) - \nabla_{\mathsf{BW}}\mathcal{V}(p_k).
$$

Hence, letting $\mathbb{E}_i$ denote expectation with respect to i, we have that

$$
\begin{aligned}
\sigma_k^2 &= \mathbb{E}\left\|e_k\right\|_{p_k}^2 \\
&= \mathbb{E}_i\mathbb{E}\left[\left\|e_k - \mathbb{E}[e_k \mid i]\right\|_{p_k}^2 \mid i\right] + \mathbb{E}_i\left[\left\|\mathbb{E}[e_k \mid i]\right\|_{p_k}^2\right] \\
&\leq \mathbb{E}\left\|e_k - \mathbb{E}[e_k \mid i]\right\|_{p_k}^2 \\
&\quad + 6\beta[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})] + 3\beta^2 W_2^2(p_k, p_0). && \text{(by Lemma 3.3.8)}
\end{aligned}
$$

Hence, it remains to bound $\mathbb{E}\left\|e_k - \mathbb{E}[e_k \mid i]\right\|_{p_k}^2$. Letting $\hat{X}_0, \hat{X}_k$ be defined as in Algorithm 5 and writing

$$\begin{aligned}
\hat{g}_i^{(k)} &: x \mapsto (\nabla^2 V_i(\hat{X}_k))(x - m_k) + \nabla V_i(\hat{X}_k) \\
\hat{g}_i^{(0)} &: x \mapsto (\nabla^2 V_i(\hat{X}_0))(x - m_0) + \nabla V_i(\hat{X}_0) \\
\hat{g}^{(0)} &: x \mapsto (\nabla^2 V(\hat{X}_0))(x - m_0) + \nabla V(\hat{X}_0),
\end{aligned}$$

we have that

$$\begin{aligned}
\mathbb{E}_i[e_k - \mathbb{E}[e_k \mid i]] &= \mathbb{E}_i[(\hat{g}_i^{(k)} - \hat{g}_i^{(0)} + \hat{g}^{(0)}) - (\nabla_{\mathsf{BW}} \mathcal{V}_i(p_k) - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_0) + \nabla_{\mathsf{BW}} \mathcal{V}(p_0))] \\
&= (\hat{g}_i^{(k)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_k) + (\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_0)).
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
\mathbb{E}\left\|e_k - \mathbb{E}[e_k \mid i]\right\|_{p_k}^2 &\le \mathbb{E}\left\|\hat{g}_i^{(k)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_k) + \hat{g}_i^{(0)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_0)\right\|_{p_k}^2 \\
&\le \underbrace{2\,\mathbb{E}\left\|\hat{g}_i^{(k)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_k)\right\|_{p_k}^2}_{\mathsf{var}_1} + \underbrace{2\,\mathbb{E}\left\|\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_0)\right\|_{p_k}^2}_{\mathsf{var}_2}.
\end{aligned}$$

It remains to bound var_1 and var_2, which can be done using nearly identical techniques as the proof of Lemma 3.3.5 in Appendix A.4.1. For var_1, we can follow the proof of Lemma 3.3.5 identically to obtain that

$$\begin{aligned}
\mathsf{var}_1 &= \mathbb{E}\left\|\hat{g}_i^{(k)} - \nabla_{\mathsf{BW}} \mathcal{V}_i(p_k)\right\|_{p_k}^2 \\
&\le 4\mathbb{E}_i \left\langle \nabla^2 V_i, \Sigma_k \nabla^2 V_i \right\rangle \\
&\le 4\beta \mathbb{E}_i \mathbb{E}_{p_k} \left\langle \nabla^2 V_i, \Sigma_k \right\rangle \\
&\le 4\beta \mathbb{E}_{p_k} \left\langle \nabla^2 V, \Sigma_k \right\rangle \\
&\le 12\beta d + \frac{24\beta^3}{\alpha} W_2^2(p_k, \hat{\pi}).
\end{aligned}$$

On the other hand, let $(X_0, X_k) \sim (p_0, p_k)$ be optimally coupled. We have that

$$\begin{aligned}
\mathsf{var}_2 &= \mathbb{E}\left\|\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)\right\|_{p_k}^2 \\
&\leq 2\mathbb{E}\left\|\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)\right\|_{p_0}^2 + 2\mathbb{E}\left\|(\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0))[X_k] - (\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0))[X_0]\right\|^2 \\
&\qquad\qquad \text{(triangle ineq and Cauchy-Schwarz)} \\
&\leq 2\mathbb{E}\left\|\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)\right\|_{p_0}^2 + 2\mathbb{E}\left\|\hat{g}_i^{(0)}[X_k] - \hat{g}_i^{(0)}[X_0]\right\|^2 \\
&\leq 2\mathbb{E}\left\|\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)\right\|_{p_0}^2 + 2\beta^2\mathbb{E}\left\|X_k - X_0\right\|^2 \qquad \text{(since } \nabla^2 V_i(\hat{X}_0) \preceq \beta I\text{)} \\
&= 2\mathbb{E}\left\|\hat{g}_i^{(0)} - \nabla_{\mathsf{BW}}\mathcal{V}_i(p_0)\right\|_{p_0}^2 + 2\beta^2 W_2^2(p_k, p_0) \\
&\leq 24\beta d + \frac{48\beta^3}{\alpha}W_2^2(p_0, \hat{\pi}) + 2\beta^2 W_2^2(p_k, p_0). \qquad \text{(analogously to } \mathsf{var}_1\text{)}
\end{aligned}$$

Combining all of the results above, we conclude that

$$\begin{aligned}
\sigma_k^2 &\leq 72\beta d + \frac{48\beta^3}{\alpha}W_2^2(p_k, \hat{\pi}) + \frac{96\beta^3}{\alpha}W_2^2(p_0, \hat{\pi}) + 4\beta^2 W_2^2(p_k, p_0) \\
&\quad + 6\beta[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})] + 3\beta^2 W_2^2(p_k, p_0) \\
&\leq 72\beta d + \frac{120\beta^3}{\alpha}[W_2^2(p_k, \hat{\pi}) + W_2^2(p_0, \hat{\pi})] + 6\beta[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})]. \\
&\qquad\qquad \text{(triangle ineq and Cauchy-Schwarz)}
\end{aligned}$$

Reintroducing the dependence on j, we obtain the desired result. □

A.5.3 Proof of Theorem 3.3.10

Once again, the idea is to combine the variance bound obtained from Lemma 3.3.8 with the one-step inequality of Inequality 3.18, following the structure of the corresponding proof in Euclidean space for Theorem 3.1.16.

Proof. First, we consider a fixed value of the outer loop iteration number j, so we drop the dependence on j in the superscript of the iterates. Combining the variance bound of Lemma 3.3.8 with the one-step inequality Inequality 3.18, we obtain that

$$\mathbb{E}W_2^2(p_{k+1}, \nu) \leq (1 - \alpha\eta)\mathbb{E}W_2^2(p_k, \nu) - 2\eta\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)]$$

$$+ 6\beta\eta^2 \left(2[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})] + \beta W_2^2(p_k, p_0)\right).$$

First, we note that

$$S_k = \mathbb{E}_{p_k} \nabla^2 V_i - \mathbb{E}_{p_0} \nabla^2 V_i + \mathbb{E}_{p_0} \nabla^2 V \preceq 2\beta I,$$

so by Lemma A.2.1, we have that $\Sigma_k^{-1} \preceq 2\beta I$. Hence, by Lemma 2.2.1 and Lemma 2.2.7, we have that

$$\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) = [\mathcal{V}(p_k) - \mathcal{V}(\hat{\pi})] + [\mathcal{H}(p_k) - \mathcal{H}(\hat{\pi})] \leq \frac{\beta}{2} W_2^2(p_k, \hat{\pi}) + \beta W_2^2(p_k, \hat{\pi}) < 2\beta W_2^2(p_k, \hat{\pi}).$$

Thus, we find that

$$\begin{aligned}
\mathbb{E}W_2^2(p_{k+1}, \nu) &\leq (1 - \alpha\eta)\mathbb{E}W_2^2(p_k, \nu) - 2\eta\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] \\
&\quad + 6\beta^2\eta^2 \left(4W_2^2(p_k, \hat{\pi}) + 4W_2^2(p_0, \hat{\pi}) + W_2^2(p_0, p_k)\right) \\
&\leq (1 - \alpha\eta)\mathbb{E}W_2^2(p_k, \nu) - 2\eta\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] \\
&\quad + 36\beta^2\eta^2 \left(W_2^2(p_k, \hat{\pi}) + W_2^2(p_0, \hat{\pi})\right) \qquad \text{(triangle ineq and Cauchy-Schwarz)}
\end{aligned}$$

Taking $\nu = \hat{\pi}$ and noting that $36\beta^2\eta^2 \leq \frac{\alpha\eta}{2}$ by assumption, we find that

$$\mathbb{E}W_2^2(p_k, \hat{\pi}) \leq \left(1 - \frac{\alpha\eta}{2}\right) \mathbb{E}W_2^2(p_k, \nu) + 36\beta^2\eta^2 W_2^2(p_0, \hat{\pi}).$$

Iterating this inequality, we find that

$$\mathbb{E}W_2^2(p_k, \hat{\pi}) \leq \exp\left(-\frac{N\alpha\eta}{2}\right) W_2^2(p_0, \hat{\pi}) + 72\beta\kappa\eta W_2^2(p_0, \hat{\pi}).$$

Hence, with the choice

$$\eta = \frac{1}{288\beta\kappa}, \quad \text{and} \quad N = \frac{2}{\alpha\eta} = \frac{\kappa^2}{144},$$

we obtain the guarantee

$$\mathbb{E}W_2^2(p_N, \hat{\pi}) \leq \left(\frac{1}{e} + \frac{1}{4}\right) \mathbb{E}W_2^2(p_0, \hat{\pi}) \leq \frac{3}{4}\mathbb{E}W_2^2(p_0, \hat{\pi})$$

Hence, reintroducing the dependence on M, we obtain that

$$\mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \leq \exp\left(-\frac{M}{4}\right) W_2^2(p_0, \hat{\pi}).$$

Finally, with $M \geq 4 \log \frac{\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2}$, we obtain the guarantee

$$\alpha \mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \leq \varepsilon^2,$$

as desired. □

A.5.4 Proof of Theorem 3.3.11

Once again, the blueprint is the same as what we have encountered before: combine the variance bound of Lemma 3.3.9 with the one-step inequality of Inequality 3.18.

Proof. First, we consider a fixed value of the outer loop iteration number j, so we drop the dependence on j in the superscript of the iterates. Combining the variance bound of Lemma 3.3.9 with the one-step inequality Inequality 3.18, we obtain that

$$\begin{aligned}\mathbb{E}W_2^2(p_{k+1}, \nu) \leq (1 - \alpha\eta)\mathbb{E}W_2^2(p_k, \nu) - 2\eta\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] \\ + 12\beta\eta^2 \left(12d + 20\beta\kappa[W_2^2(p_k, \hat{\pi}) + W_2^2(p_0, \hat{\pi})] + \beta[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) + \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})]\right).\end{aligned}$$

Just as in the proof of Theorem 3.3.10, we note that

$$S_k = \mathbb{E}_{p_k}\nabla^2 V_i - \mathbb{E}_{p_0}\nabla^2 V_i + \mathbb{E}_{p_0}\nabla^2 V \preceq 2\beta I,$$

so by Lemma A.2.1, we have that $\Sigma_k^{-1} \preceq 2\beta I$. Hence, by Lemma 2.2.1 and Lemma 2.2.7, we again

have that

$$\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi}) \leq 2\beta W_2^2(p_k, \hat{\pi}).$$

Thus, we find that

$$\begin{aligned}\mathbb{E}W_2^2(p_{k+1}, \nu) \leq (1 - \alpha\eta)\mathbb{E}W_2^2(p_k, \nu) - 2\eta\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] \\ + 144\beta\eta^2 \left(d + 2\beta\kappa[W_2^2(p_k, \hat{\pi}) + W_2^2(p_0, \hat{\pi})]\right).\end{aligned}$$

Taking $\nu = \hat{\pi}$ and noting that $288\beta^2\kappa\eta^2 \leq \frac{\alpha\eta}{2}$ by assumption, we obtain that

$$\mathbb{E}W_2^2(p_k, \hat{\pi}) \leq \left(1 - \frac{\alpha\eta}{2}\right)\mathbb{E}W_2^2(p_k, \nu) + 288\beta^2\kappa\eta^2 W_2^2(p_0, \hat{\pi}) + 144\beta d\eta^2.$$

Iterating this inequality, we find that

$$\begin{aligned}\mathbb{E}W_2^2(p_k, \hat{\pi}) &\leq \exp\left(-\frac{N\alpha\eta}{2}\right) W_2^2(p_0, \hat{\pi}) + 576\beta\kappa^2\eta W_2^2(p_0, \hat{\pi}) + 288\kappa d\eta \\ &\leq \exp\left(-\frac{N\alpha\eta}{2}\right) W_2^2(p_0, \hat{\pi}) + 600\beta\kappa^2\eta W_2^2(p_0, \hat{\pi}) + 300\kappa d\eta.\end{aligned}$$

Hence, with the choice

$$\eta = \frac{1}{2400}\left(\frac{\varepsilon^2}{\beta d} \wedge \frac{1}{\beta\kappa^2}\right), \quad \text{and} \quad N = \frac{2}{\alpha\eta} = \frac{4800}{\alpha}\left(\frac{\beta d}{\varepsilon^2} \vee \beta\kappa^2\right),$$

we obtain the guarantee

$$\mathbb{E}W_2^2(p_N, \hat{\pi}) \leq \left(\frac{1}{e} + \frac{1}{4}\right)\mathbb{E}W_2^2(p_0, \hat{\pi}) + \frac{\varepsilon^2}{8\alpha} \leq \frac{3}{4}\mathbb{E}W_2^2(p_0, \hat{\pi}) + \frac{\varepsilon^2}{8\alpha}.$$

Hence, reintroducing the dependence on M, and iterating this inequality, we obtain that

$$\mathbb{E}W_2^2(p_0^{(M)}, \hat{\pi}) \leq \left(\frac{3}{4}\right)^M \mathbb{E}W_2^2(p_0, \hat{\pi}) + \frac{\varepsilon^2}{2\alpha} \leq \exp\left(-\frac{M}{4}\right)\mathbb{E}W_2^2(p_0, \hat{\pi}) + \frac{\varepsilon^2}{2\alpha}.$$

Finally, with $M \geq 4 \log \frac{2\alpha W_2^2(p_0, \hat{\pi})}{\varepsilon^2}$, we obtain the guarantee

$$\alpha \mathbb{E} W_2^2(p_0^{(M)}, \hat{\pi}) \leq \varepsilon^2,$$

as desired. □

Appendix B

Simulations

In this section, via elementary simulations[1], we demonstrate that FBGVI is implementable, practical and competitive with the Bures–Wasserstein gradient descent (BWGD) method of [51]. We consider two examples:

1. **Gaussian targets.** For the first experiment, we consider a scenario where the target density is

$$\pi(x) \propto \exp\left(-\frac{1}{2}\left\langle (x-\mu), \Sigma^{-1}(x-\mu)\right\rangle\right),$$

where $\mu \sim \mathrm{Unif}([0,1]^{10})$ and $\Sigma^{-1} = U \,\mathrm{diag}\begin{bmatrix}10^{-9} & 10^{-8} & \cdots & 1\end{bmatrix} U^{\mathsf{T}}$, with $U \in \mathbb{R}^{10\times 10}$ chosen as a uniformly random orthogonal matrix. In this case, we have that $\pi \in \mathsf{BW}(\mathbb{R}^{10})$, so the solution to Problem (1.1) is precisely π, and furthermore we have that π is 10^{-9}-strongly log-concave and 1-log-smooth.

We run FB–GVI and stochastic FB–GVI with target potential $\pi \propto \exp(-V)$ initialized at $p_0 = \mathcal{N}(0, I_{10})$, where I_{10} is the 10×10 identity matrix. The step size η is varied, and the resulting plots of $\log \mathsf{KL}\left(p_k \,\|\, \pi\right)$ for different choices of η are displayed in Figure B-1.

2. **Bayesian logistic regression.** We consider the following generative model: given a parameter $\theta \in \mathbb{R}^d$, we draw i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n \in (\mathbb{R}^d \times \{0,1\})^n$ with

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \quad Y_i \mid X_i \sim \mathrm{Bern}(e^{\langle \theta, X_i\rangle}).$$

[1] Code for our experiments can be found at https://github.com/mzydiao/FBGVI/blob/main/FBGVI-Experiments.ipynb.

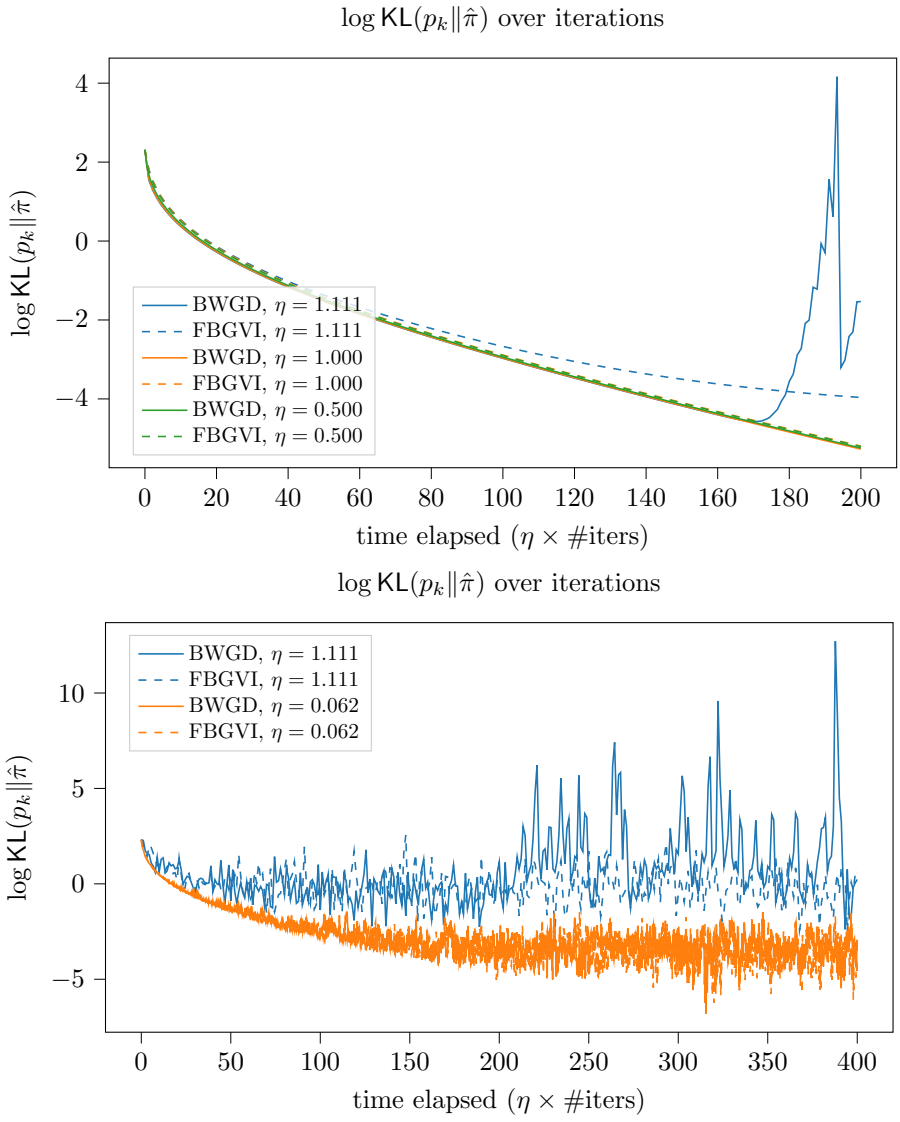

Figure B-1: Gaussian target experiment: results for FB–GVI (top) and stochastic FB–GVI (bottom).

Given these samples $\{(X_i, Y_i)\}_{i=1}^n$ and a uniform (improper) prior on θ, the posterior on θ is given by

$$V(\theta) = \sum_{i=1}^{n} \left[\ln(1 + e^{\langle \theta, X_i \rangle}) - Y_i \ \langle \theta, X_i \rangle\right].$$

We run stochastic FB–GVI with $\pi \propto \exp(-V)$ initialized at $p_0 = \mathcal{N}(0, I_d)$ with varying step sizes η. Since in this scenario we do not know the true minimizer $\hat{\pi}$ nor the normalization constant of π, we cannot directly compute $\mathsf{KL}(p_k \,\|\, \pi)$ nor $W_2^2(p_k, \hat{\pi})$. However, we can still estimate the objective function $\mathcal{F}(p_k)$ as well as the squared BW gradient norm $\mathbb{E}_{p_k} \|\nabla_{\mathsf{BW}} \mathcal{F}(p_k)\|^2$ empirically by drawing samples from p_k. For each choice of step size η, we plot our empirical estimates of $\mathcal{F}(p_k)$ and $\mathbb{E}_{p_k} \|\nabla_{\mathsf{BW}} \mathcal{F}(p_k)\|^2$ over iterations in Figure B-2.

Our results are provided in Figure B-1 and Figure B-2. Based on the plots, we make the following observations:

1. (Stochastic) FB–GVI performs as well as BWGD, if not better. In addition, for sufficiently small η, both FB–GVI and BWGD attain lower objective than the Laplace approximation as seen in Figure B-2. This observation was also made in [45] for BWGD.

2. FB–GVI is stable up to much larger step sizes than BWGD, mirroring the comparative stability of proximal gradient methods versus gradient descent methods in Euclidean space, especially when the step size is large (see, *e.g.*, [88, 87]).

Our empirical results, combined with our theoretical guarantees, lend convincing evidence in support of using FB–GVI for Gaussian variational inference.

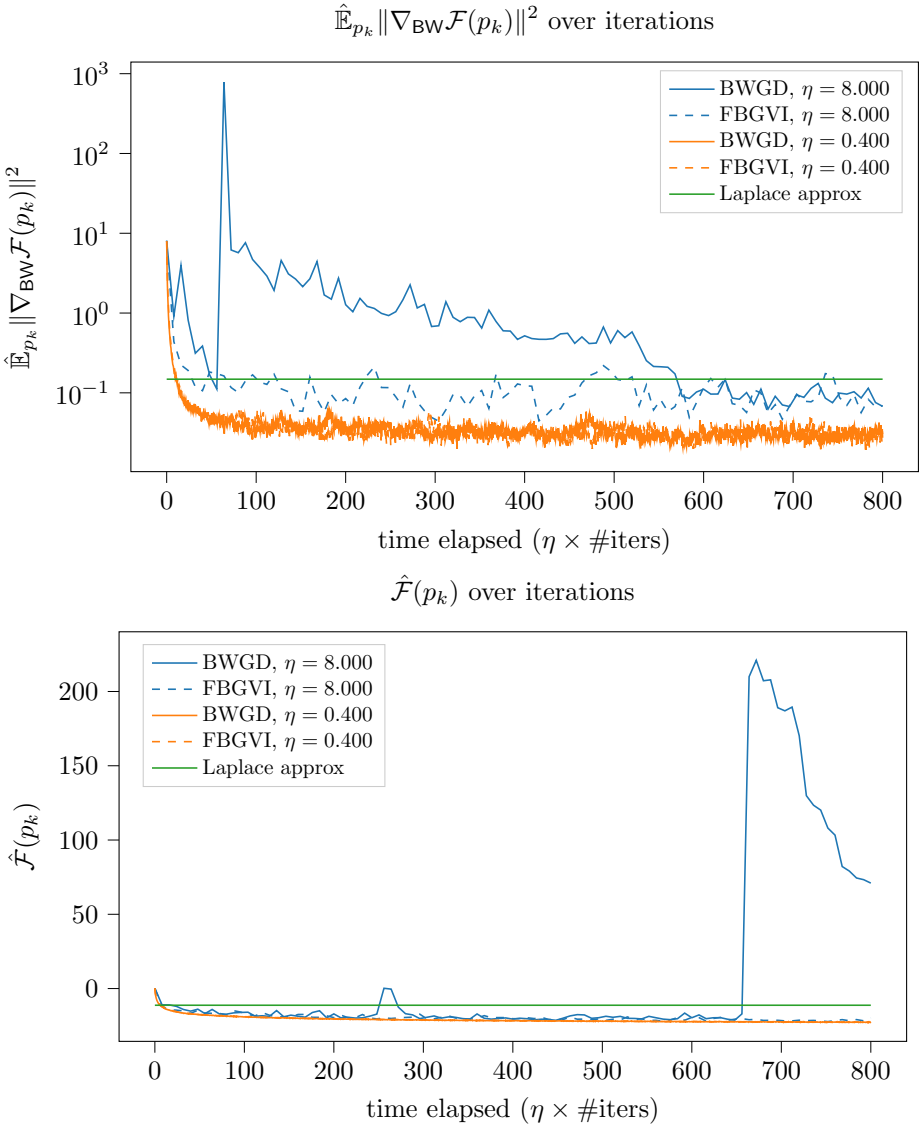

Figure B-2: Bayesian logistic regression experiment: plots of $\log \hat{\mathbb{E}}_{p_k} \|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|^2$ (top) and of $\hat{\mathcal{F}}(p_k)$ (bottom) for stochastic FB–GVI.

Bibliography

[1] Kwangjun Ahn and Sinho Chewi. "Efficient constrained sampling via the mirror-Langevin algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 28405–28418.

[2] Pierre Alquier and James Ridgway. "Concentration of tempered posteriors and of their variational approximations". In: *The Annals of Statistics* 48.3 (2020), pp. 1475–1497.

[3] Jason Altschuler et al. "Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 22132–22145.

[4] David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. "Optimizing functionals on the space of probabilities with input convex neural networks". In: *Transactions on Machine Learning Research* (2022).

[5] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[6] Yves F Atchadé, Gersende Fort, and Eric Moulines. "On perturbed proximal gradient algorithms". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 310–342.

[7] Julio Backhoff-Veraguas et al. "Stochastic gradient descent in Wasserstein space". In: *arXiv preprint arXiv:2201.04232* (2022).

[8] Krishna Balasubramanian et al. "Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo". In: *Conference on Learning Theory*. PMLR. 2022, pp. 2896–2923.

[9] David Barber and Christopher Bishop. "Ensemble learning for multi-layer networks". In: *Advances in Neural Information Processing Systems* 10 (1997).

[10] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Vol. 408. Springer, 2011.

[11] Espen Bernton. "Langevin Monte Carlo and JKO splitting". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1777–1798.

[12] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. "On the Bures–Wasserstein distance between positive definite matrices". In: *Expo. Math.* 37.2 (2019), pp. 165–191.

[13] Pascal Bianchi, Walid Hachem, and Adil Salim. "A constant step forward-backward algorithm involving random maximal monotone operators". In: *J. Convex Anal.* 26.2 (2019), pp. 397–436.

[14] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: a review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[15] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM Review* 60.2 (2018), pp. 223–311.

[16] G. E. P. Box and Mervin E. Muller. "A Note on the Generation of Random Normal Deviates". In: *The Annals of Mathematical Statistics* 29.2 (1958), pp. 610 –611.

[17] Herm Jan Brascamp and Elliott H. Lieb. "On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation". In: *J. Functional Analysis* 22.4 (1976), pp. 366–389.

[18] José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. "A blob method for diffusion". In: *Calculus of Variations and Partial Differential Equations* 58 (2019), pp. 1–53.

[19] Anthony Caterini et al. "Variational inference with continuously-indexed normalizing flows". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 44–53.

[20] Edward Challis and David Barber. "Gaussian Kullback–Leibler approximate inference". In: *Journal of Machine Learning Research* 14.8 (2013).

[21] Shixiang Chen et al. "Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning". In: *IEEE Transactions on Signal Processing* 69 (2021), pp. 4759–4773.

[22] Shixiang Chen et al. "Proximal gradient method for nonsmooth optimization over the Stiefel manifold". In: *SIAM Journal on Optimization* 30.1 (2020), pp. 210–239.

[23] Yongxin Chen et al. "Improved analysis for a proximal algorithm for sampling". In: *Conference on Learning Theory*. PMLR. 2022, pp. 2984–3014.

[24] Badr-Eddine Chérief-Abdellatif, Pierre Alquier, and Mohammad Emtiyaz Khan. "A generalization bound for online variational inference". In: *Asian Conference on Machine Learning*. PMLR. 2019, pp. 662–677.

[25] Sinho Chewi. *Log-concave sampling*. Available at https://chewisinho.github.io/. 2023.

[26] Sinho Chewi et al. "Gradient descent algorithms for Bures–Wasserstein barycenters". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1276–1304.

[27] Sinho Chewi et al. "SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2098–2109.

[28] Arnak S Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[29] Michael Diao et al. *Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein Space*. 2023. arXiv: 2304.05398 [math.ST].

[30] Justin Domke. "Provable smoothness guarantees for black-box variational inference". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2587–2596.

[31] Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. "On the geometry of Stein variational gradient descent". In: *arXiv preprint arXiv:1912.00894* (2019).

[32] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization". In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 2666–2711.

[33] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 978-3-540-35908-1.

[34] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. "A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 680–690.

[35] Andi Han et al. "On Riemannian optimization over positive definite matrices with the Bures–Wasserstein geometry". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[36] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[37] Ye He et al. "Regularized Stein Variational Gradient Flow". In: *arXiv preprint arXiv:2211.07861* (2022).

[38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[39] Antti Honkela and Harri Valpola. "Unsupervised variational Bayesian learning of nonlinear models". In: *Advances in Neural Information Processing Systems* 17 (2004).

[40] Xiaoyin Hu et al. "A Constraint Dissolving Approach for Nonsmooth Optimization over the Stiefel Manifold". In: *arXiv preprint arXiv:2205.10500* (2022).

[41] Wen Huang and Ke Wei. "Riemannian proximal gradient methods". In: *Mathematical Programming* 194.1-2 (2022), pp. 371–413.

[42] Rie Johnson and Tong Zhang. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013.

[43] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[44] Mikołaj J Kasprzak, Ryan Giordano, and Tamara Broderick. "How good is your Gaussian approximation of the posterior? Finite-sample computable error bounds for a variety of useful divergences". In: *arXiv preprint arXiv:2209.14992* (2022).

[45] Anya Katsevich and Philippe Rigollet. "On the approximation accuracy of Gaussian variational inference". In: *arXiv preprint arXiv:2301.02168* (2023).

[46] Durk P Kingma et al. "Improved variational inference with inverse autoregressive flow". In: *Advances in Neural Information Processing Systems* 29 (2016).

[47] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. "An optimization-centric view on Bayes' rule: reviewing and generalizing variational inference". In: *Journal of Machine Learning Research* 23.132 (2022), pp. 1–109.

[48] Anna Korba et al. "A non-asymptotic analysis for Stein variational gradient descent". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4672–4682.

[49] H Kushner and G Yin. "Stochastic approximation and recursive algorithms". In: *Stochastic Modelling and Applied Probability*. Vol. 35. Springer-Verlag NY, 2003.

[50] Marc Lambert, Silvere Bonnabel, and Francis Bach. "The recursive variational Gaussian approximation (R-VGA)". In: *Statistics and Computing* 32.1 (2022), p. 10.

[51] Marc Lambert et al. “Variational inference via Wasserstein gradient flows”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022.

[52] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. “Stochastic zeroth-order Riemannian derivative estimation and optimization”. In: *Mathematics of Operations Research* (2022).

[53] Xiao Li et al. “Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods”. In: *SIAM Journal on Optimization* 31.3 (2021), pp. 1605–1634.

[54] Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. “Fast and simple natural-gradient variational inference with mixture of exponential-family approximations”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3992–4002.

[55] Wu Lin, Mark Schmidt, and Mohammad Emtiyaz Khan. “Handling the positive-definite constraint in the Bayesian learning rule”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6116–6126.

[56] Qiang Liu. “Stein variational gradient descent as gradient flow”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[57] Qiang Liu and Dilin Wang. “Stein variational gradient descent: a general purpose Bayesian inference algorithm”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016.

[58] Jianfeng Lu, Yulong Lu, and James Nolen. “Scaling limit of the Stein variational gradient descent: the mean field regime”. In: *SIAM Journal on Mathematical Analysis* 51.2 (2019), pp. 648–671.

[59] Yuetian Luo and Nicolas Garcia Trillos. “Nonconvex matrix factorization is geodesically convex: global landscape analysis for fixed-rank matrix optimization from a Riemannian perspective”. In: *arXiv preprint arXiv:2209.15130* (2022).

[60] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. “Wasserstein Riemannian geometry of Gaussian densities”. In: *Inf. Geom.* 1.2 (2018), pp. 137–179.

[61] Tyler Maunu, Thibaut Le Gouic, and Philippe Rigollet. “Bures–Wasserstein barycenters and low-rank matrix recovery”. In: *arXiv preprint arXiv:2210.14671* (2022).

[62] Nicholas Metropolis and S. Ulam. “The Monte Carlo Method”. In: *Journal of the American Statistical Association* 44.247 (1949), pp. 335–341.

[63] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[64] Klas Modin. “Geometry of matrix decompositions seen through optimal transport and information geometry”. In: *J. Geom. Mech.* 9.3 (2017), pp. 335–390.

[65] Yurii Nesterov. *Lectures on Convex Optimization*. 2nd. Springer Publishing Company, Incorporated, 2018. ISBN: 3319915770.

[66] Richard Nickl and Sven Wang. “On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms”. In: *Journal of the European Mathematical Society* (2022).

[67] I. Olkin and F. Pukelsheim. “The distance between two random vectors with given dispersion matrices”. In: *Linear Algebra Appl.* 48 (1982), pp. 257–263.

[68] Manfred Opper and Cédric Archambeau. “The variational Gaussian approximation revisited”. In: *Neural Computation* 21.3 (2009), pp. 786–792.

[69] Felix Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: *Comm. Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[70] John W. Paisley, David M. Blei, and Michael I. Jordan. "Variational Bayesian inference with stochastic search". In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. Omnipress, 2012.

[71] Neal Parikh, Stephen Boyd, et al. "Proximal algorithms". In: *Foundations and trends® in Optimization* 1.3 (2014), pp. 127–239.

[72] Zheng Peng et al. "Riemannian Smoothing Gradient Type Algorithms for Nonsmooth Optimization Problem on Manifolds". In: *arXiv preprint arXiv:2212.03526* (2022).

[73] Geoff Pleiss et al. "Fast matrix square roots with applications to Gaussian processes and Bayesian optimization". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22268–22281.

[74] Matias Quiroz, David J Nott, and Robert Kohn. "Gaussian Variational Approximations for High-dimensional State Space Models". In: *Bayesian Analysis* 1.1 (2022), pp. 1–28.

[75] Rajesh Ranganath, Sean Gerrish, and David Blei. "Black–box variational inference". In: *Artificial Intelligence and Statistics*. PMLR. 2014, pp. 814–822.

[76] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1530–1538.

[77] Adil Salim, Anna Korba, and Giulia Luise. "The Wasserstein proximal gradient algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12356–12366.

[78] Adil Salim, Lukang Sun, and Peter Richtarik. "A convergence theory for SVGD in the population limit under Talagrand's inequality T1". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 19139–19152.

[79] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Vol. 87. Progress in Nonlinear Differential Equations and their Applications. Calculus of variations, PDEs, and modeling. Birkhäuser/Springer, Cham, 2015, pp. xxvii+353.

[80] Matthias Seeger. "Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers". In: *Advances in Neural Information Processing Systems* 12 (1999).

[81] Jiaxin Shi and Lester Mackey. "A finite-particle convergence rate for Stein variational gradient descent". In: *arXiv preprint arXiv:2211.09721* (2022).

[82] Jascha Sohl-Dickstein et al. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 2256–2265.

[83] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems* 32 (2019).

[84] Yue Song, Nicu Sebe, and Wei Wang. "Fast differentiable matrix square root and inverse square root". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[85] Vladimir Spokoiny. "Dimension free non-asymptotic bounds on the accuracy of high dimensional Laplace approximation". In: *arXiv preprint arXiv:2204.11038* (2022).

[86] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[87] Panos Toulis, Thibaut Horel, and Edoardo M Airoldi. "The proximal Robbins–Monro method". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.1 (2021), pp. 188–212.

[88] Panos Toulis, Dustin Tran, and Edo Airoldi. "Towards stability and optimality in stochastic gradient descent". In: *Artificial Intelligence and Statistics*. PMLR. 2016, pp. 1290–1298.

[89] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge University Press, 2000.

[90] Cédric Villani. *Optimal transport: Old and new*. Vol. 338. Springer, 2009.

[91] Cédric Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[92] Yifei Wang, Peng Chen, and Wuchen Li. "Projected Wasserstein gradient descent for high-dimensional Bayesian inference". In: *SIAM/ASA Journal on Uncertainty Quantification* 10.4 (2022), pp. 1513–1532.

[93] Zhongruo Wang et al. "A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis". In: *INFORMS Journal on Optimization* 4.2 (2022), pp. 200–214.

[94] Andre Wibisono. "Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2093–3027.

[95] Lin Xiao and Tong Zhang. "A proximal stochastic gradient method with progressive variance reduction". In: *SIAM Journal on Optimization* 24.4 (2014), pp. 2057–2075.

[96] Rentian Yao and Yun Yang. "Mean field variational inference via Wasserstein gradient flow". In: *arXiv preprint arXiv:2207.08074* (2022).

[97] Chao Zhang, Xiaojun Chen, and Shiqian Ma. "A Riemannian smoothing steepest descent method for non-Lipschitz optimization on submanifolds". In: *arXiv preprint arXiv:2104.04199* (2021).

[98] Dewei Zhang and Sam Davanloo Tajbakhsh. "Riemannian Stochastic Variance-Reduced Cubic Regularized Newton Method for Submanifold Optimization". In: *Journal of Optimization Theory and Applications* (2022), pp. 1–38.

[99] Guodong Zhang et al. "Noisy natural gradient as variational inference". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5852–5861.