Forward-Backward Gaussian Variational Inference via JKO in the Bures–Wasserstein Space

Anonymous Authors[1]

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 optimization algorithms to solve Gaussian VI based on viewing the KL divergence as a 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. Our proposed algorithm, which we call (Stochastic) Forward-Backward Gaussian Variational Inference (FB–GVI), achieves state-of-the-art convergence rates when π is strongly log-concave. We also obtain the first convergence results when π is a general log-concave distribution.

1. Introduction

Variational inference (VI) (Blei et al., 2017; Knoblauch et al., 2022) has emerged as a tractable alternative to computationally demanding Monte Carlo Markov Chain (MCMC) methods. Of particular interest is the problem of Gaussian VI, in which we approximate a given distribution $\pi \propto \exp(-V)$, where V is a smooth function, by the solution to

$$\underset{\mu \in \mathsf{BW}(\mathbb{R}^d)}{\arg\min}\ \mathsf{KL}(\mu \,\|\, \pi)\,, \tag{1}$$

where KL denotes the Kullback–Leibler divergence and $\mathsf{BW}(\mathbb{R}^d)$ the set of Gaussian distributions over $\mathbb{R}^d$ (see Section 4.1 for formal definitions). Indeed, Gaussian VI has shown superior performance in practice, especially in the presence of large datasets, see, for example, Barber & Bishop (1997); Seeger (1999); Honkela & Valpola (2004); Opper & Archambeau (2009); Quiroz et al. (2022).

In the literature on Gaussian VI, strong *statistical* properties have been shown for the solutions to Problem (1), see, for example, Chérief-Abdellatif et al. (2019); Alquier & Ridgway (2020); Katsevich & Rigollet (2023). For instance, Katsevich & Rigollet (2023) showed that Gaussian VI outperforms Laplace approximation for the estimation of the mean of π. Besides, consider the case where π is the posterior distribution of a sufficiently regular Bayesian model. Then, the Bernstein–von Mises theorem (see Van der Vaart (2000, Chapter 10) and recent non-asymptotic results (Kasprzak et al., 2022; Spokoiny, 2022) state that π is well-approximated by a Gaussian distribution, with the mean given by any asymptotically efficient estimator of the true parameter, and covariance matrix given by the inverse Fisher information matrix. These results collectively provide abundant motivation for efficiently computing the best Gaussian approximation of the target π (Problem (1)).

We hence focus on the *optimization* aspect of Gaussian VI, *i.e.*, solving Problem (1). Several approaches have been proposed which we summarize in the related works (Section 6). In particular, Lambert et al. (2022b) recently proposed an algorithm for Gaussian VI that can be seen as an analog of stochastic gradient descent for Problem (1) over the space $\mathsf{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance, called the Bures–Wasserstein (BW) space. This viewpoint was inspired from the theory of gradient flows over the Wasserstein space, *i.e.*, the space of distributions endowed with the Wasserstein distance (Jordan et al., 1998; Ambrosio et al., 2008). Moreover, this viewpoint has been instrumental for many problems in probabilistic inference (see the related works in Section 6). However, from an optimization standpoint, the approach of Lambert et al. (2022b) relying on the BW gradient of the objective is not the most natural one. Indeed, over the BW space, the objective functional $\mathsf{KL}(\cdot \,\|\, \pi)$ 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.

The composite nature of the KL divergence has inspired more than two decades of research on forward-backward methods on the Wasserstein space (see, for example, Jordan

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

et al., 1998; Bernton, 2018; Wibisono, 2018; Salim et al., 2020). Unfortunately, this line of work is obstructed by the computational intractability of the so-called JKO operator (Jordan et al., 1998) (*i.e.*, the analog of the proximal operator over the Wasserstein space) of the entropy.

In this paper, we introduce a novel algorithm called (Stochastic) Forward-Backward Gaussian Variational Inference (FB–GVI). Similarly to Lambert et al. (2022b), the rich differential and geometric structure of the BW space comprises the linchpin of our approach. However, (Stochastic) FB–GVI additionally incorporates celebrated ideas from the literature on composite and non-smooth optimization. A key insight in this work is that the JKO operator for the entropy, when restricted to the BW space, admits a closed form (Wibisono, 2018), and hence leads to an 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.

We summarize our **contributions** below.

- We propose a new (stochastic) forward-backward algorithm, (Stochastic) FB–GVI, to solve Problem (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 (Ambrosio et al., 2008).

The rest of our paper is organized as follows. In Section 2, we clarify our notation and provide some background material on stochastic and non-smooth composite optimization. In Section 3, we describe the geometric and differential structure of the BW space, which is key to performing optimization over $\mathsf{BW}(\mathbb{R}^d)$. Then, in Section 4 we focus on Problem (1) and propose our algorithms: FB–GVI and Stochastic FB–GVI. The convergence of our algorithms is studied in Section 5. Finally, we discuss related works in Section 6 and conclude in Section 7.

2. Background

In this section, we clarify our notation and provide background on (stochastic) forward-backward algorithms.

2.1. Notation

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

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

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 $\Sigma \in \mathbf{S}^d_{++}$.

2.2. Stochastic and non-smooth convex optimization over $\mathbb{R}^d$

Before introducing optimization concepts on the space of probability measures, we review some details of stochastic and convex non-smooth optimization over $\mathbb{R}^d$. First, a function $V : \mathbb{R}^d \to \mathbb{R}$ 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, V satisfies the Taylor inequality

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

Consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \{V(x) + H(x)\}\,, \tag{3}$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and $H : \mathbb{R}^d \to \mathbb{R}$ is convex but potentially non-smooth. To simplify the presentation, we also assume that V is convex.

Because of the non-smoothness of H, the (sub)gradient descent algorithm applied to Problem (3) may not converge to a minimizer of $V + H$. However, in many situations of interest, the user has closed-form expressions[1] for the proximal operator of H defined by

$$\operatorname{prox}_H(x) \coloneqq \operatorname*{arg\,min}_{y\in\mathbb{R}^d} \left\{H(y) + \frac{1}{2}\|x-y\|^2\right\}. \tag{4}$$

[1] See proximity-operator.net.

Given access to the proximal operator of H and the gradient of V, the forward-backward algorithm (Bauschke et al., 2011) is one of the most natural and efficient techniques to solve Problem (3). The forward-backward algorithm is written as

$$x_{k+1} = \operatorname{prox}_{\eta H}(x_k - \eta \nabla V(x_k))\,. \tag{5}$$

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 situation, the stochastic forward-backward algorithm has been proven to be an efficient alternative to the forward-backward algorithm (Atchadé et al., 2017; Bianchi et al., 2019; Gorbunov et al., 2020). In the stochastic forward-backward algorithm, $\nabla V(x_k)$ is replaced by $\hat{g}_k$ as follows:

$$x_{k+1} = \operatorname{prox}_{\eta H}(x_k - \eta\, \hat{g}_k)\,. \tag{6}$$

We will now adapt the (stochastic) forward-backward algorithm (6) to the BW space. In particular, the iterate at step k will no longer be a random variable x_k, but rather a random Gaussian distribution p_k. Despite the fact that the BW space is not a Euclidean space, its inherent structure still allows us to perform optimization, as explained next.

3. The Bures–Wasserstein space

A detailed presentation of the Wasserstein space and its geometry, which in turn enables optimization over that space, can be found in Ambrosio et al. (2008). In this section, we quickly review the BW space and its geometry, hence providing the requisite tools to perform optimization over the BW space and solve Problem (1). We start with formal definitions of the Wasserstein and BW spaces.

3.1. 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

$$W_2^2(\mu,\nu) = \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \|x - y\|^2 \, \mathrm{d}\gamma(x,y)\,, \tag{7}$$

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, *i.e.*, a map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\# \mu = \nu$ and

$$W_2^2(\mu,\nu) = \int \|x - T(x)\|^2 \, \mathrm{d}\mu(x)\,. \tag{8}$$

In other words, the coupling $(\mathrm{id}, T)_\# \mu$ belongs to $\mathcal{C}(\mu,\nu)$ and attains the infimum in (7).

Moreover, since μ and ν are Gaussian, T is an affine map with symmetric linear part, *i.e.*, can be written as $T(x) = Sx + b$ where $S \in \mathbf{S}^d$ and $b \in \mathbb{R}^d$ (Olkin & Pukelsheim, 1982). In particular, the BW space is a 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 that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T_\# \mu \in \mathcal{P}_2(\mathbb{R}^d)$ 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 $L^2(\mu)$.

3.2. 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 Appendix A.

3.2.1. A “SMOOTH” FUNCTIONAL

Consider a functional $\mathcal{F} : \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$. 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,

$$\mathcal{F}((\mathrm{id} + th)_\# \mu) = \mathcal{F}(\mu) + t \, \langle g_\mu, h \rangle_\mu + o(t)\,. \tag{9}$$

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

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

$$\mathcal{V}(\mu) := \int V \, \mathrm{d}\mu \tag{10}$$

for every $\mu \in \mathsf{BW}(\mathbb{R}^d)$, is a useful example of a differentiable functional over the BW space.

The next result states 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. (2022b, Section C.1) (see also Appendix A.3). Besides, the differentiability of $\mathcal{V}$ is well-established in the literature on the Wasserstein space (Ambrosio et al., 2008, Theorem 10.4.13); we adapt this to the BW space.

Lemma 3.1 (BW gradient of the potential)**.** *Consider the potential functional $\mathcal{V}$ in (10) where V is β-smooth. Then, $\mathcal{V}$ is differentiable at μ and the following Taylor inequality holds: for h affine,*

$$|\mathcal{V}((\mathrm{id} + h)_\# \mu) - \mathcal{V}(\mu) - \langle \nabla_{\mathsf{BW}} \mathcal{V}(\mu), h \rangle_\mu| \le \frac{\beta}{2} \, \|h\|_\mu^2\,. \tag{11}$$

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

$$\nabla_{\mathsf{BW}} \mathcal{V}(\mu) : x \mapsto \mathbb{E}_\mu \nabla V + (\mathbb{E}_\mu \nabla^2 V)(x - m_\mu)\,, \tag{12}$$

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

Proof. The proof of Equation (12) can be found in Lambert et al. (2022b, Section C.1) (see also Appendix A.3), and the proof of Inequality (11) can be found in Appendix B.1. □

Inequality (11) is stronger than differentiability and can be interpreted as the potential $\mathcal{V}$ being β-smooth over the BW space (note the analogy with Inequality (2)). Inequality (11) is a consequence of the smoothness of V.

As in optimization over $\mathbb{R}^d$, when dealing with a convex but potentially non-smooth functional, the user may prefer to handle it through its proximal operator.

3.2.2. A "CONVEX" FUNCTIONAL

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

$$\mathcal{F}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{F}(\mu_0), T - \mathrm{id} \rangle_{\mu_0} \le \mathcal{F}(\mu_1) \,, \qquad (13)$$

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 Bures–Wasserstein JKO operator of $\mathcal{F}$ (Jordan et al., 1998)[2], and defined by (note the analogy with (4))

$$\mathrm{JKO}_{\mathcal{F}}(\mu) \coloneqq \operatorname*{arg\,min}_{\nu \in \mathsf{BW}(\mathbb{R}^d)} \left\{ \mathcal{F}(\nu) + \frac{1}{2} W_2^2(\mu, \nu) \right\} . \qquad (14)$$

We want to emphasize that, in our definition (14), the BW JKO operator is defined over the BW space.

The entropy $\mathcal{H}$ is a useful example of a geodesically convex functional over the BW space. More precisely, the entropy is defined by

$$\mathcal{H}(\mu) = \int \log \mu(x) \, \mathrm{d}\mu(x) \,, \qquad (15)$$

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

The next lemma states that the entropy is geodesically convex and gives a formula for its BW JKO operator. The formula for the BW JKO operator can be obtained by a straightforward adaptation of Wibisono (2018, Example 7). Besides, the geodesic convexity of $\mathcal{H}$ is well-established in the literature on the Wasserstein space (Ambrosio et al., 2008, Remark 9.3.10); we adapt this to the BW space.

Lemma 3.2 (BW JKO of the entropy)**.** *Consider the entropy functional $\mathcal{H}$ defined in (15). Then, $\mathcal{H}$ is geodesically convex and the following stronger inequality holds: for all $\nu, \mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$,*

$$\mathcal{H}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu \le \mathcal{H}(\mu_1) \,, \quad (16)$$

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

Moreover, the BW JKO operator of $\mathcal{H}$ is known in closed form: $\mathrm{JKO}_{\eta\mathcal{H}}(\mu)$ (where $\eta > 0$) is a Gaussian distribution with same mean as μ and with covariance matrix Σ_1 where

$$\Sigma_1 = \frac{1}{2} \left(\Sigma + 2\eta I + [\Sigma(\Sigma + 4\eta I)]^{1/2} \right) , \qquad (17)$$

where Σ is the covariance matrix of μ.

Proof. The proof of Equation (17) can be found in Wibisono (2018, Example 7), and the proof of Inequality (16) can be found in Appendix B.2. □

Inequality (16) is stronger than geodesic convexity and is a consequence of the generalized geodesic convexity of the entropy (Ambrosio et al., 2008, Remark 9.3.10).

Finally, (17) which gives the BW JKO of the entropy in closed form is remarkable, and is at the core of our approach. As a comparison, Salim et al. (2020) proposed an algorithm relying on the JKO of the entropy over the whole Wasserstein space, but the latter JKO is not implementable.

4. (Stochastic) Forward-backward Gaussian variational inference

4.1. Revisiting Gaussian VI

We now restate Problem (1) 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

$$\mathsf{KL}(\mu \,\|\, \pi) = \int \log \frac{\mu(x)}{\pi(x)} \, \mathrm{d}\mu(x) \,. \qquad (18)$$

Recall that our goal is to solve Problem (1). We denote $\mathcal{F} \coloneqq \mathcal{V} + \mathcal{H}$ as the sum of the potential (associated to the function V) and the entropy. Then, a quick calculation reveals that $\mathcal{F}(\mu) - \mathcal{F}(\pi) = \mathsf{KL}(\mu \,\|\, \pi)$. Since $\mathcal{F}(\pi)$ is a constant (*i.e.*, does not depend on μ), Problem (1) is equivalent to

$$\min_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \{ \mathcal{V}(\mu) + \mathcal{H}(\mu) \} \,. \qquad (19)$$

[2] In Jordan et al. (1998) the authors define the JKO operator as an analog of the proximal operator over the Wasserstein space. Inspired by their definition, we define the JKO operator over the BW space, and we call it the BW JKO operator.

4.2. 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 3.1). 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 3.2). To solve the equivalent problem (19), a natural idea is to adapt the forward-backward algorithm to the BW space. This leads to the following Forward-Backward Gaussian Variational Inference (FB–GVI) algorithm (note the analogy with (5)):

$$p_{k+\frac{1}{2}} = (\mathrm{id} - \eta\,\nabla_{\mathsf{BW}}\mathcal{V}(p_k))_{\#}p_k\,, \tag{20}$$

$$p_{k+1} = \mathrm{JKO}_{\eta\mathcal{H}}(p_{k+\frac{1}{2}})\,. \tag{21}$$

The backward step (21) is tractable using (17). Although the forward step (20) also admits a closed form, the forward step involves computing integrals of ∇V and $\nabla^2 V$ w.r.t. p_k, see (12). These integrals can be intractable. Therefore, we propose to build a stochastic estimate $\hat{g}_k$ of $\nabla_{\mathsf{BW}}\mathcal{V}(p_k)$, *i.e.*, a stochastic gradient, by drawing a random sample from p_k. The resulting algorithm is called Stochastic FB–GVI, and can be written as

$$\begin{aligned} p_{k+\frac{1}{2}} &= (\mathrm{id} - \eta\,\hat{g}_k)_{\#}p_k\,,\\ p_{k+1} &= \mathrm{JKO}_{\eta\mathcal{H}}(p_{k+\frac{1}{2}})\,, \end{aligned} \tag{22}$$

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

$$\hat{g}_k : x \mapsto \nabla V(\hat{X}_k) + \nabla^2 V(\hat{X}_k)\,(x - m_k)\,, \tag{23}$$

where $\hat{X}_k$ is sampled from p_k (*i.e.*, $X_k \sim p_k$) and $m_k = \int x\, dp_k(x)$ is the mean of p_k.

Stochastic FB–GVI is an analog, over the BW space, of the stochastic forward-backward algorithm (note the analogy with (6)). In particular, $(p_k)_{k\in\mathbb{N}}$ defined by (22) is a sequence of random Gaussian distributions, *i.e.*, random variables with 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 in terms 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 1. Efficient algorithms developed for computing the matrix square-root (see, for example, Pleiss et al. (2020); Song et al. (2022)) can be leveraged to improve the per-iteration complexity.

5. Convergence theory

In this section, we study the convergence of FB–GVI and Stochastic FB–GVI using their equivalent forms (20)–(21) and (22). We will make use of standard complexity notations, such as $\gtrsim, \asymp$. We also denote by $\hat{\pi} = \mathcal{N}(\hat{m}, \hat{\Sigma})$ a solution of Problem (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).

Algorithm 1 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 $p_N = \mathcal{N}(m_N, \Sigma_N)$
```

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 (weakly) 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 1) as

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

and denote its expected $L^2(p_k)$ norm by

$$\sigma_k^2 := \mathbb{E}[\|e_k\|_{p_k}^2 \mid \mathscr{F}_k]\,.$$

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 (23). 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 (20)–(21) and (22).

Lemma 5.1 (one-step inequality)**.** *Suppose that V is α-convex and β-smooth. Let $(p_k)_{k\in\mathbb{N}}$ be the iterates of FB–GVI* (20)–(21) *or Stochastic FB–GVI* (22)*. Let $\eta > 0$ be such that*

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

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

$$\mathbb{E}W_2^2(p_{k+1}, \nu) \le (1 - \alpha\eta)\,\mathbb{E}W_2^2(p_k, \nu) \tag{24}$$

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

Proof. The proof is given in Appendix C. □

This one-step inequality is similar, by replacing the Wasserstein distance by the Euclidean distance, to the one-step inequality classically used in the analysis of the stochastic forward-backward algorithm (Gorbunov et al., 2020).

The proof of Lemma 5.1 heavily employs the differential and geometric structure of the BW space presented in Section 3.

5.1. Convergence of FB–GVI

In this section, $(p_k)_{k\in\mathbb{N}}$ is the sequence of iterates defined by FB–GVI ((20)–(21)). We obtain corollaries of Lemma 5.1 by setting $\sigma_k = 0$ in (24), in the case where V is weakly convex and in the case where V is strongly convex.

Theorem 5.2 (Weakly convex case, FB–GVI)**.** *Suppose that V is convex and β-smooth and that $0 < \eta \le \frac{1}{\beta}$. Then,*

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

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

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

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

Theorem 5.3 (Strongly convex case, FB–GVI)**.** *Suppose that V is α-strongly convex and β-smooth, and that $0 < \eta \le \frac{1}{\beta}$. Then,*

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

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

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

Proof. The proof is given in Appendix E.2. □

Theorem 5.2 states the sublinear convergence of FB–GVI for a weakly convex (in terms of objective gap) V and Theorem 5.3 states the linear convergence of FB–GVI for a strongly convex V. The convergence rates are of the same order as the convergence rates of the deterministic forward-backward algorithm (Gorbunov et al., 2020) over $\mathbb{R}^d$.

Finally, we also extend our results to the non-convex case, in which case we obtain a stationary point guarantee.

Theorem 5.4 (Smooth case, FB–GVI)**.** *Suppose that V is β-smooth, and that $0 < \eta \le \frac{1}{\beta}$. Let $\Delta := \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})$. Then,*

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

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

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

Proof. The proof is given in Appendix E.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 & Rigollet (2023), the favorable statistical properties of Gaussian VI arise, not due to the global minimization of the objective in (1), but rather from the fixed-point equations characterizing first-order optimality. Hence, Theorem 5.4 can be viewed as an algorithmic result for posterior approximation, even in the non-log-concave setting.

5.2. Convergence of Stochastic FB–GVI

In this section, $(p_k)_{k\in\mathbb{N}}$ is the sequence of iterates defined by (22). To use Lemma 5.1, we first prove a bound on σ_k^2, the variance of the BW stochastic gradient.

Lemma 5.5. *If V is convex and β-smooth, then*

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

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

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

Proof. See Appendix F.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 (Kushner & Yin, 2003; Bottou et al., 2018). We emphasize that we do not assume this sublinear growth. Instead, Lemma 5.5 proves the sublinear growth for the BW stochastic gradient used in Stochastic FB–GVI.

Next, we obtain corollaries of Lemma 5.1 for Stochastic FB–GVI by controlling σ_k^2 with Lemma 5.5.

Theorem 5.6 (Weakly convex case, Stochastic FB–GVI)**.** *Suppose that V is convex and β-smooth and that $0 < \eta \le \frac{1}{2\beta}$. Define $c := 24\beta^3 \lambda_{\max}(\hat{\Sigma})$. Then,*

$$\mathbb{E}\bigl[\min_{k\in\{1,\ldots,N\}} \mathcal{F}(p_k)\bigr] - \mathcal{F}(\hat{\pi})$$

$$\leq \frac{2W_2^2(p_0,\hat{\pi})}{N\eta} + 2c\eta\, W_2^2(p_0,\hat{\pi}) + 12\beta\eta d\,.$$

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

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

we obtain the guarantee

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

Proof. See Appendix F.3. □

Theorem 5.7 (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,*

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

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

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

we obtain the guarantees

$$\alpha\,\mathbb{E}W_2^2(p_N,\hat{\pi}) \leq \varepsilon^2\,,$$
$$\mathbb{E}\big[\min_{k\in\{1,\dots,2N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2\,.$$

Proof. See Appendix F.4. □

To our knowledge, Theorem 5.6 is the first result to provide a complexity result in terms of the objective gap in Problem (1), for log-smooth log-concave target distributions. Moreover, Theorem 5.7 improves upon the state-of-the-art obtained in (Lambert et al., 2022b) 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.

6. Related works

We now discuss streams of work that are closely related to ours, and place our work in the context of the larger literature on sampling and variational inference.

Optimization algorithms for Gaussian VI. Algorithms for solving Gaussian VI have been considered in Paisley et al. (2012); Ranganath et al. (2014); Lambert et al. (2022a). The general approach is to parametrize the set of Gaussian distributions and to apply Euclidean optimization. In particular, Alquier & Ridgway (2020) noticed that when π is the posterior distribution in a Bayesian logistic regression, Problem (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 & Barber (2013). Algorithms based on natural gradient methods (Zhang et al., 2018; Lin et al., 2019; 2020) and normalizing flows (Rezende & Mohamed, 2015; Kingma et al., 2016; Caterini et al., 2021) 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. (2022b), who similarly proposed an optimization algorithm over the BW space, called Bures–Wasserstein Stochastic Gradient Descent (BW–SGD), to solve Problem (1). Their algorithm BW–SGD 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 weakly 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 algorithm (Dalalyan, 2017) is an MCMC algorithm to sample from the target distribution π. The theory of Wasserstein gradient flows (Ambrosio et al., 2008) provides the mathematical tools to view the Langevin algorithm (and its many variants) as an optimization algorithm over Wasserstein space. In the case of the Langevin algorithm, the objective to minimize is $\mathsf{KL}(\cdot \,\|\, \pi)$. Therefore, one can use optimization analysis (over the Wasserstein space) to show convergence bounds for the Langevin algorithm (Wibisono, 2018; Durmus et al., 2019; Balasubramanian et al., 2022; Chen et al., 2022; Chewi, 2023).

Stein Variational Gradient Descent (Liu & Wang, 2016; Liu, 2017) is another method that can be seen as an optimization algorithm for minimizing $\mathsf{KL}(\cdot \,\|\, \pi)$. 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 (Duncan et al., 2019; Lu et al., 2019; Chewi et al., 2020a; Korba et al., 2020; Salim et al., 2022; He et al., 2022; Shi & Mackey, 2022). However, there is an interesting connection between SVGD and BW–GD (Lambert et al., 2022b): 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.

Another closely related work to ours is that of Salim et al. (2020). In the same vein as our work, they view the objective $\mathsf{KL}(\cdot \,\|\, \pi)$ 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 discuss the implementation of the JKO of the entropy. Therefore, to our knowledge, their algorithm 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 morally related to recent works developing efficient algorithms for solving non-smooth optimization problems over certain manifolds. For example, in the deterministic setting, Li et al. (2021) analyzed the sub-gradient method, Chen et al. (2020); Huang & Wei (2022) analyzed the proximal gradient method, Chen et al. (2021) analyzed the proximal point method, and Wang et al. (2022) analyzed the proximal linear method. Stochastic versions were considered in Li et al. (2022); Wang et al. (2022). We also refer to Zhang et al. (2021); Hu et al. (2022); Peng et al. (2022); Zhang & Davanloo Tajbakhsh (2022) 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 (2017); Malagò et al. (2018); Bhatia et al. (2019), and optimization over this space has proven to be fruitful for various applications, see, for example, Chewi et al. (2020b); Altschuler et al. (2021); Han et al. (2021); Lambert et al. (2022b); Luo & Trillos (2022); Maunu et al. (2022).

7. Conclusion

We proposed a novel optimization algorithm, (Stochastic) FB–GVI, for solving the Gaussian VI problem in (1). We view this algorithm 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), 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) FW–GVI algorithm on various specific practical problems of interest.

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) 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.

References

Ahn, K. and Chewi, S. Efficient constrained sampling via the mirror-Langevin algorithm. In Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P. S., Vaughan, J. W., and Dauphin, Y. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28405–28418. Curran Associates, Inc., 2021.

Alquier, P. and Ridgway, J. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497, 2020.

Altschuler, J., Chewi, S., Gerber, P. R., and Stromme, A. Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22132–22145. Curran Associates, Inc., 2021.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Atchadé, Y. F., Fort, G., and Moulines, E. On perturbed proximal gradient algorithms. *The Journal of Machine Learning Research*, 18(1):310–342, 2017.

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pp. 2896–2923. PMLR, 2022.

Barber, D. and Bishop, C. Ensemble learning for multi-layer networks. *Advances in Neural Information Processing Systems*, 10, 1997.

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Bernton, E. Langevin Monte Carlo and JKO splitting. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1777–1798. PMLR, 06–09 Jul 2018.

Bhatia, R., Jain, T., and Lim, Y. On the Bures–Wasserstein distance between positive definite matrices. *Expo. Math.*, 37(2):165–191, 2019.

Bianchi, P., Hachem, W., and Salim, A. A constant step forward-backward algorithm involving random maximal monotone operators. *J. Convex Anal.*, 26(2):397–436, 2019.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Brascamp, H. J. and Lieb, E. H. On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Functional Analysis*, 22(4):366–389, 1976.

Caterini, A., Cornish, R., Sejdinovic, D., and Doucet, A. Variational inference with continuously-indexed normalizing flows. In *Uncertainty in Artificial Intelligence*, pp. 44–53. PMLR, 2021.

Challis, E. and Barber, D. Gaussian Kullback–Leibler approximate inference. *Journal of Machine Learning Research*, 14(8), 2013.

Chen, S., Ma, S., Man-Cho So, A., and Zhang, T. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1): 210–239, 2020.

Chen, S., Deng, Z., Ma, S., and So, A. M.-C. Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning. *IEEE Transactions on Signal Processing*, 69:4759–4773, 2021.

Chen, Y., Chewi, S., Salim, A., and Wibisono, A. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 2984–3014. PMLR, 2022.

Chérief-Abdellatif, B.-E., Alquier, P., and Khan, M. E. A generalization bound for online variational inference. In *Asian Conference on Machine Learning*, pp. 662–677. PMLR, 2019.

Chewi, S. *Log-concave sampling*. 2023. Available at https://chewisinho.github.io/.

Chewi, S., Le Gouic, T., Lu, C., Maunu, T., and Rigollet, P. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020a.

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. Gradient descent algorithms for Bures–Wasserstein barycenters. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1276–1304. PMLR, 09–12 Jul 2020b.

Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Duncan, A., Nüsken, N., and Szpruch, L. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Durmus, A., Majewski, S., and Miasojedow, B. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Gorbunov, E., Hanzely, F., and Richtárik, P. A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 680–690. PMLR, 2020.

Han, A., Mishra, B., Jawanpuria, P., and Gao, J. On Riemannian optimization over positive definite matrices with the Bures–Wasserstein geometry. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

He, Y., Balasubramanian, K., Sriperumbudur, B. K., and Lu, J. Regularized Stein variational gradient flow. *arXiv preprint arXiv:2211.07861*, 2022.

Honkela, A. and Valpola, H. Unsupervised variational Bayesian learning of nonlinear models. *Advances in Neural Information Processing Systems*, 17, 2004.

Hu, X., Xiao, N., Liu, X., and Toh, K.-C. A constraint dissolving approach for nonsmooth optimization over the Stiefel manifold. *arXiv preprint arXiv:2205.10500*, 2022.

Huang, W. and Wei, K. Riemannian proximal gradient methods. *Mathematical Programming*, 194(1-2):371–413, 2022.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Kasprzak, M. J., Giordano, R., and Broderick, T. How good is your Gaussian approximation of the posterior? Finite-sample computable error bounds for a variety of useful divergences. *arXiv preprint arXiv:2209.14992*, 2022.

Katsevich, A. and Rigollet, P. On the approximation accuracy of Gaussian variational inference. *arXiv preprint arXiv:2301.02168*, 2023.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29, 2016.

Knoblauch, J., Jewson, J., and Damoulas, T. An optimization-centric view on Bayes' rule: reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

Kushner, H. and Yin, G. Stochastic approximation and recursive algorithms. In *Stochastic Modelling and Applied Probability*, volume 35. Springer-Verlag NY, 2003.

Lambert, M., Bonnabel, S., and Bach, F. The recursive variational Gaussian approximation (R-VGA). *Statistics and Computing*, 32(1):10, 2022a.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. Variational inference via Wasserstein gradient flows. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b.

Li, J., Balasubramanian, K., and Ma, S. Stochastic zeroth-order Riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 2022.

Li, X., Chen, S., Deng, Z., Qu, Q., Zhu, Z., and Man-Cho So, A. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

Lin, W., Khan, M. E., and Schmidt, M. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning*, pp. 3992–4002. PMLR, 2019.

Lin, W., Schmidt, M., and Khan, M. E. Handling the positive-definite constraint in the Bayesian learning rule. In *International Conference on Machine Learning*, pp. 6116–6126. PMLR, 2020.

Liu, Q. Stein variational gradient descent as gradient flow. *Advances in Neural Information Processing Systems*, 30, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: a general purpose Bayesian inference algorithm. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Lu, J., Lu, Y., and Nolen, J. Scaling limit of the Stein variational gradient descent: the mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

Luo, Y. and Trillos, N. G. Nonconvex matrix factorization is geodesically convex: global landscape analysis for fixed-rank matrix optimization from a Riemannian perspective. *arXiv preprint arXiv:2209.15130*, 2022.

Malagò, L., Montrucchio, L., and Pistone, G. Wasserstein Riemannian geometry of Gaussian densities. *Inf. Geom.*, 1(2):137–179, 2018.

Maunu, T., Le Gouic, T., and Rigollet, P. Bures–Wasserstein barycenters and low-rank matrix recovery. *arXiv preprint arXiv:2210.14671*, 2022.

Modin, K. Geometry of matrix decompositions seen through optimal transport and information geometry. *J. Geom. Mech.*, 9(3):335–390, 2017.

Olkin, I. and Pukelsheim, F. The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.*, 48:257–263, 1982.

Opper, M. and Archambeau, C. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

Paisley, J. W., Blei, D. M., and Jordan, M. I. 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.

Peng, Z., Wu, W.-H., Hu, J., and Deng, K.-K. Riemannian smoothing gradient type algorithms for nonsmooth optimization problem on manifolds. *arXiv preprint arXiv:2212.03526*, 2022.

Pleiss, G., Jankowiak, M., Eriksson, D., Damle, A., and Gardner, J. Fast matrix square roots with applications to Gaussian processes and Bayesian optimization. *Advances in Neural Information Processing Systems*, 33:22268–22281, 2020.

Quiroz, M., Nott, D. J., and Kohn, R. Gaussian variational approximations for high-dimensional state space models. *Bayesian Analysis*, 1(1):1–28, 2022.

Ranganath, R., Gerrish, S., and Blei, D. Black–box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822. PMLR, 2014.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

Salim, A., Korba, A., and Luise, G. The Wasserstein proximal gradient algorithm. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12356–12366. Curran Associates, Inc., 2020.

Salim, A., Sun, L., and Richtarik, P. A convergence theory for SVGD in the population limit under Talagrand's inequality T1. In *International Conference on Machine Learning*, pp. 19139–19152. PMLR, 2022.

Santambrogio, F. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

Seeger, M. Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. *Advances in Neural Information Processing Systems*, 12, 1999.

Shi, J. and Mackey, L. A finite-particle convergence rate for Stein variational gradient descent. *arXiv preprint arXiv:2211.09721*, 2022.

Song, Y., Sebe, N., and Wang, W. Fast differentiable matrix square root and inverse square root. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Spokoiny, V. Dimension free non-asymptotic bounds on the accuracy of high dimensional Laplace approximation. *arXiv preprint arXiv:2204.11038*, 2022.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Villani, C. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

Wang, Z., Liu, B., Chen, S., Ma, S., Xue, L., and Zhao, H. A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis. *INFORMS Journal on Optimization*, 4(2): 200–214, 2022.

Wibisono, A. Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2093–3027. PMLR, 06–09 Jul 2018.

Zhang, C., Chen, X., and Ma, S. A Riemannian smoothing steepest descent method for non-Lipschitz optimization on submanifolds. *arXiv preprint arXiv:2104.04199*, 2021.

Zhang, D. and Davanloo Tajbakhsh, S. Riemannian stochastic variance-reduced cubic regularized Newton method for submanifold optimization. *Journal of Optimization Theory and Applications*, pp. 1–38, 2022.

Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pp. 5852–5861. PMLR, 2018.

A. Differential calculus over the BW space

In this section, we derive a formula for the BW gradient of a generic functional, giving us the tools necessary to define the updates of our forward-backward algorithm. 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 helpful for our proofs of convexity and smoothness inequalities later on. Our derivation relies on specializing the computation rules of Otto calculus, which deals with the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, to the BW space $\mathsf{BW}(\mathbb{R}^d)$.

A.1. Background on Otto calculus

We first give an informal overview of the computation rules of Otto calculus (Otto, 2001), which endows the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ with a formal Riemannian structure. We refer to Ambrosio et al. (2008) 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

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0 . \tag{25}$$

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 Santambrogio, 2015, 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 (25) 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

$$\nabla_{W_2}\mathcal{F}(\mu) \equiv \nabla\delta\mathcal{F}(\mu) . \tag{26}$$

Now we turn our attention to the BW space. The BW space $\mathsf{BW}(\mathbb{R}^d)$ is a submanifold of $\mathcal{P}_2(\mathbb{R}^d)$ (Otto, 2001; Lambert et al., 2022b), 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 (25). Using Equation (26) and integration by parts, we compute an expression for the BW gradient of $\mathcal{F}$ in the following subsection.

A.2. 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. (2022b, Section C.1). We present the 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 a 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 A.1. *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 (25). 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

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

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

$$\begin{aligned}
\partial_t|_{t=0}\mathcal{F}(\mu_t) &= \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu && \text{(definition of } \nabla_{W_2}\mathcal{F}\text{)} \\
&= \langle \nabla\delta\mathcal{F}(\mu), v_0\rangle_\mu && \text{(by Equation (26))} \\
&= \mathbb{E}\langle \nabla\delta\mathcal{F}(X), a + M\,(X - \mathbb{E}X)\rangle \\
&= \mathbb{E}\langle \nabla\delta\mathcal{F}(X), a\rangle + \mathbb{E}\langle \Sigma_\mu M^\mathsf{T}\nabla\delta\mathcal{F}(X), \Sigma_\mu^{-1}\,(X - \mathbb{E}X)\rangle \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle - \int \langle \Sigma_\mu M^\mathsf{T}\nabla\delta\mathcal{F}, \nabla\mu\rangle && \text{(since } \nabla\mu(x) = -\mu(x)\,\Sigma_\mu\,(x - \mathbb{E}X)\text{)} \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle + \mathbb{E}[\operatorname{div}(\Sigma_\mu M^\mathsf{T}\nabla\delta\mathcal{F})(X)] && \text{(integration by parts)} \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle + \langle \mathbb{E}_\mu\nabla^2\delta\mathcal{F}(X), \Sigma_\mu M^\mathsf{T}\rangle\,.
\end{aligned}$$

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. □

A.3. Examples of BW gradients and stationary condition for Problem (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) is equivalent to minimizing $\mathcal{F}$ over the BW space, *i.e.*, solving Problem (19). Using Lambert et al. (2022b, Section C.1), we have the following formulas for the BW gradients of $\mathcal{V}$ and $\mathcal{H}$.

$$\begin{aligned}
\nabla_{\mathsf{BW}}\mathcal{V}(\mu) &: x \mapsto \mathbb{E}_\mu\nabla V + (\mathbb{E}_\mu\nabla^2 V)(x - m_\mu)\,, \\
\nabla_{\mathsf{BW}}\mathcal{H}(\mu) &: x \mapsto -\Sigma_\mu^{-1}\,(x - m_\mu)\,.
\end{aligned} \tag{27}$$

We can also derive the above formulas from Lemma A.1.

Moreover, by the proof of Lemma A.1, 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), the first-order stationary condition $\nabla_{\mathsf{BW}}\mathcal{F}(\hat\pi) = 0$ for Problem (1) reads

$$\mathbb{E}_{\hat\pi} \nabla V = 0 \qquad \text{and} \qquad \mathbb{E}_{\hat\pi} \nabla^2 V = \hat\Sigma^{-1}\,, \tag{28}$$

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

B. 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 Appendix A, we may now proceed to prove Lemma 3.1 (for the potential) and Lemma 3.2 (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 (25), utilizing our calculation rules laid out in Appendix A. In particular, we will use that

$$\begin{aligned}
\mathcal{F}(\mu_1) - \mathcal{F}(\mu_0) &= \int_0^1 \partial_t \mathcal{F}(\mu_t)\,\mathrm{d}t \\
&= \partial_t|_{t=0}\mathcal{F}(\mu_t) + \int_0^1\int_0^t \partial_s^2 F(\mu_s)\,\mathrm{d}s\,\mathrm{d}t \\
&= \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu_0), v_0\rangle_{\mu_0} + \int_0^1 (1-t)\,\partial_t^2 \mathcal{F}(\mu_t)\,\mathrm{d}t\,,
\end{aligned} \tag{29}$$

for both the entropy and the potential.

B.1. Proof of Lemma 3.1

We prove the following result for the potential. This result is stronger than Lemma 3.1, and will be useful in our subsequent analysis.

Lemma B.1. *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:*

$$\begin{aligned}
\mathcal{V}((\mathrm{id}+h)_\#\mu) - \mathcal{V}(\mu) &\ge \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_\mu + \frac{\alpha}{2}\,\|h\|_\mu^2\,, \\
\mathcal{V}((\mathrm{id}+h)_\#\mu) - \mathcal{V}(\mu) &\le \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_\mu + \frac{\beta}{2}\,\|h\|_\mu^2\,.
\end{aligned}$$

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

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

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

Lemma 3.1 then follows as a corollary of the above lemma.

Proof of Lemma 3.1. 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 B.1 with $\alpha = -\beta$, we obtain that

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

Moreover, we have shown in Appendix A.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. □

B.2. Proof of Lemma 3.2

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:

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 Ambrosio et al. (2008) to study gradient flows in the Wasserstein space, and have since been useful for various applications of this theory, e.g., Chewi et al. (2020b); Ahn & Chewi (2021); Altschuler et al. (2021).

Proof of Lemma 3.2. The JKO operator of $\mathcal{H}$ over the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ is derived in Wibisono (2018, Example 7) for a Gaussian measure $\mu = \mathcal{N}(\mu, \Sigma)$, and takes the form $\mu' = \mathcal{N}(\mu, \Sigma_1)$ where Σ_1 is defined in the same manner as Equation (17). 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.

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., Ambrosio et al., 2008, §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

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

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 (Villani, 2003, 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 $\big\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \big\rangle \geq 0$. Hence, using Equation (29) applied to $\mathcal{H}$, we obtain that

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

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

Remark B.2. 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

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

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

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

$$\begin{aligned}
\|T_1 - T_0\|_\nu^2 &= \mathrm{Tr}(\mathrm{Cov}_\nu[T_1 - T_0, T_1 - T_0]) + \|\mathbb{E}_\nu[T_1 - T_0]\|^2 && \text{(by bias-variance decomposition)} \\
&= \mathrm{Tr}(\mathrm{Cov}[(\nabla T_1 - \nabla T_0)(Y), (\nabla T_1 - \nabla T_0)(Y)]) + \|m_1 - m_0\|^2 && \text{(since } T_0, T_1 \text{ are affine)} \\
&= \big\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \big\rangle + \|m_1 - m_0\|^2. && (30)
\end{aligned}$$

In addition, from Chewi et al. (2020b, 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 that

$$\begin{aligned}
\big\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \big\rangle &\le \frac{1}{\|\Sigma_{\mu_t^\nu}\|_{\mathrm{op}}} \big\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \big\rangle \\
&= \frac{1}{\|\Sigma_{\mu_t^\nu}\|_{\mathrm{op}}} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2\right) && \text{(by Equation (30))} \\
&\ge \frac{1}{\lambda} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2\right). && \text{(by Chewi et al. (2020b, Lemma 10))}
\end{aligned}$$

Hence, using Equation (29) applied to $\mathcal{H}$, we obtain that

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

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

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

Once again using Equation (29) applied to $\mathcal{H}$, we obtain that

$$\begin{aligned}
\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 (1-t) \big\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \big\rangle \, \mathrm{d}t \\
&\le \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 \frac{1-t}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \|T_1 - T_0\|_\nu^2 \, \mathrm{d}t. && (31)
\end{aligned}$$

As a corollary of Inequality 31, we obtain a smoothness inequality for the entropy along geodesics, which will be useful for our subsequent analysis.

Lemma B.3 (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*

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

Proof. We apply Inequality 31 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. (2021, Appendix B), we know that $\lambda_{\min}$ is concave along geodesics, so $\lambda_{\min}(\Sigma_{\mu_t}) \ge \gamma^{-1} I$ for all t. Hence, we obtain that

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

proving the desired result. □

C. Proof of the one-step inequality (Lemma 5.1)

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,

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu) = [\mathcal{V}(p_{k+1}) - \mathcal{V}(p_{k+\frac{1}{2}})] + [\mathcal{V}(p_{k+\frac{1}{2}}) - \mathcal{V}(\nu)] + [\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)] ,$$

where each individual term may be controlled using the inequalities in Lemmas 3.2 and B.1. Recalling that Lemma 3.2 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 5.1. Recall from Section 5 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 $\eta \le \frac{1}{\beta}$ by assumption, we have that

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

Recall that by Brenier's theorem (Villani, 2003, 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

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

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

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

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 B.1. Doing so, we obtain that

$$\begin{aligned}
\mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(\nu)] &= \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(p_k)] + \mathbb{E}[\mathcal{V}(p_k) - \mathcal{V}(\nu)] \\
&\le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}(p_k)(X_k), X_k - Y_{\mathcal{V}}\rangle - \frac{\alpha}{2}\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 \\
&\quad + \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}(p_k)(X_k), X_{k+1} - X_k\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_{k+1} - X_k\|^2 \qquad \text{(by Lemma B.1)} \\
&= -\frac{\alpha}{2}\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 + \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}(p_k)(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle \\
&\quad + \frac{1}{2\eta}\,\mathbb{E}\|X_{k+1} - X_k\|^2 - \Bigl(\frac{1}{2\eta} - \frac{\beta}{2}\Bigr)\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
&= -\frac{\alpha}{2}\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \frac{1}{\eta}\,\mathbb{E}\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}}\rangle \\
&\quad + \frac{1}{2\eta}\,\mathbb{E}\|X_{k+1} - X_k\|^2 - \Bigl(\frac{1}{2\eta} - \frac{\beta}{2}\Bigr)\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
&= \frac{1}{2\eta}\,(1 - \alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Bigl(\frac{1}{2\eta} - \frac{\beta}{2}\Bigr)\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
&\quad + \frac{1}{2\eta}\,\mathbb{E}\bigl[\|X_{k+1} - X_k\|^2 - \|X_k - Y_{\mathcal{V}}\|^2 - 2\,\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}}\rangle\bigr]\,.
\end{aligned}$$

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 3.2 to obtain that

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

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

$$\begin{aligned}
\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &= \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(\nu)] + \mathbb{E}[\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)] \\
&\le \frac{1}{2\eta}\,\mathbb{E}\bigl[(1 - \alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\bigr] \\
&\quad + \frac{1}{2\eta}\,\mathbb{E}\bigl[\|X_{k+1} - X_k\|^2 - \|X_k - Y_{\mathcal{V}}\|^2 + \|X_{k+\frac{1}{2}} - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2\bigr] \\
&\quad - \frac{1}{2\eta}\,\mathbb{E}\bigl[2\,\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}}\rangle\bigr] \\
&\quad - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Bigl(\frac{1}{2\eta} - \frac{\beta}{2}\Bigr)\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
&= \frac{1}{2\eta}\,\mathbb{E}\bigl[(1 - \alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\bigr] \\
&\quad - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Bigl(\frac{1}{2\eta} - \frac{\beta}{2}\Bigr)\,\mathbb{E}\|X_{k+1} - X_k\|^2\,. \qquad (32)
\end{aligned}$$

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
$$-\mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle = 0\,.$$
Combining this with Inequality 32, we obtain that with $\eta \leq \frac{1}{\beta}$,
$$\begin{aligned}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &\leq \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\|X_{k+1} - X_k\|^2\\ &\leq \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big]\,.\end{aligned}$$
Rearranging, we conclude that if $e_k \equiv 0$ and $\eta \leq \frac{1}{\beta}$,
$$\begin{aligned}\mathbb{E}W_2^2(p_{k+1},\nu) &\leq \mathbb{E}\|X_{k+1} - Y_{\mathcal{H}}\|^2\\ &\leq (1-\alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] \qquad (33)\\ &= (1-\alpha\eta)\,\mathbb{E}W_2^2(p_k,\nu) - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)]\,.\end{aligned}$$
(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
$$\begin{aligned}-\mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle &= -\mathbb{E}\langle e_k(X_k), X_{k+1} - X_k\rangle && \text{(since } e_k \perp\!\!\!\perp (X_k, Y_{\mathcal{V}}) \text{ by construction)}\\ &\leq \eta\,\mathbb{E}\|e_k(X_k)\|^2 + \frac{1}{4\eta}\,\mathbb{E}\|X_{k+1} - X_k\|^2\,. && \text{(Cauchy–Schwarz and Young's inequality)}\end{aligned}$$
Hence, combining this with Inequality 32, we obtain that for $\eta \leq \frac{1}{2\beta}$,
$$\begin{aligned}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &\leq \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] + \eta\,\mathbb{E}\|e_k(X_k)\|^2\\ &\qquad - \Big(\frac{1}{4\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\|X_{k+1} - X_k\|^2\\ &\leq \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] + \eta\,\mathbb{E}\sigma_k^2\,.\end{aligned}$$
Rearranging, we conclude that as long as $\eta \leq \frac{1}{2\beta}$,
$$\begin{aligned}\mathbb{E}W_2^2(p_{k+1},\nu) &\leq \mathbb{E}\|X_{k+1} - Y_{\mathcal{H}}\|^2\\ &\leq (1-\alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\,\mathbb{E}\sigma_k^2\\ &= (1-\alpha\eta)\,\mathbb{E}W_2^2(p_k,\nu) - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\,\mathbb{E}\sigma_k^2\,.\end{aligned}$$
(since $(X_k, Y_{\mathcal{V}})$ are optimally coupled)

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

Remark C.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 33, we obtain that
$$\begin{aligned}&\mathbb{E}\|X_{k+1} - Y_{\mathcal{H}}\|^2 \leq (1-\alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] && \text{(Inequality 33)}\\ \Longrightarrow\;&\mathbb{E}\|X_{k+1} - X_k\|^2 \leq -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)]\,. && \text{(since } \nu = p_k \text{ and } Y_{\mathcal{V}} = Y_{\mathcal{H}} = X_k)\end{aligned}$$

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

Lemma C.2. *Suppose that V is β-smooth. Let $(p_k)_{k\in\mathbb{N}}$ be the iterates of FB–GVI (20)–(21). 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,*
$$\mathbb{E}\|X_{k+1} - X_k\|^2 \leq -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)]\,.$$

D. Eigenvalue control of the iterates

We will show the following eigenvalue bound result:

Lemma D.1. *At the k-th iteration of Algorithm 1, 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

$$\begin{aligned} f_\eta(x_\gamma) &= \frac{1}{2}\left(\frac{(1-\eta\gamma)^2}{\gamma} + 2\eta + \sqrt{\left(\frac{(1-\eta\gamma)^2}{\gamma}\right)\left(\frac{(1-\eta\gamma)^2}{\gamma} + 4\eta\right)}\right) \\ &= \frac{1}{2\gamma}\left(1 + \eta^2\gamma^2 + \sqrt{(1-\eta\gamma)^2\,(1+\eta\gamma)^2}\right) \\ &= \frac{1}{2\gamma}\left(1 + \eta^2\gamma^2 + (1-\eta\gamma)\,(1+\eta\gamma)\right) \\ &= \frac{1}{\gamma}\,. \end{aligned} \tag{34}$$

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 (34).

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 D.1 holds with $\gamma_0 = \alpha$ and $\gamma_1 = \beta$. Hence, we obtain the following corollary:

Corollary D.2. *Suppose that Algorithm 1 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 $\eta \le \frac{1}{\beta}$. Then $\beta^{-1} I \preceq \Sigma_k$ for all k.*

E. 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 5.1. First, we derive inequalities that hold for both the weakly and strongly convex cases.

For FB–GVI, we can apply Lemma 5.1 with $\nu = \hat{\pi}$, $\eta \le \frac{1}{\beta}$ and $\sigma_k = 0$. Furthermore, FB–GVI is deterministic, so we may remove the expectations in Lemma 5.1. In this case, the inequality in Lemma 5.1 implies that for all k,

$$\begin{aligned} W_2^2(p_{k+1}, \hat{\pi}) &\le (1 - \alpha\eta)\, W_2^2(p_k, \hat{\pi}) - 2\eta\, (\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})) && \text{(by Lemma 5.1)} \\ &\le \exp(-\alpha\eta)\, W_2^2(p_k, \hat{\pi}) - 2\eta\, (\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}))\,. && (35) \end{aligned}$$

Rearranging Equation (35), we obtain

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}) \le \frac{\exp(-\alpha\eta)\, W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})}{2\eta}\,. \tag{36}$$

On the other hand, we can also apply Lemma 5.1 with $\nu = p_k$, $\eta \le \frac{1}{\beta}$ and $\sigma_k^2 = 0$ to obtain that

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

Hence, rearranging this inequality, we obtain that

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k) \le -\frac{W_2^2(p_{k+1}, p_k)}{2\eta} \le 0\,, \tag{37}$$

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

E.1. Proof of Theorem 5.2

Proof. Since V is convex, Inequality 36 holds with the choice $\alpha = 0$, from which we obtain that

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

Telescoping this inequality, we obtain that

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

where the first inequality holds by Inequality 37. Hence, with the choice

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

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

E.2. Proof of Theorem 5.3

Proof. Since $\mathcal{F}(\hat{\pi}) \le \mathcal{F}(p_{k+1})$ as $\hat{\pi}$ achieves the minimum of $\mathcal{F}$ among Gaussians, we may iterate Inequality 36 to obtain

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

Hence, with the choice

$$\begin{aligned}
\eta &= \frac{1}{\beta}\,, \\
N &\gtrsim \frac{1}{\alpha\eta} \log \frac{\alpha W_2^2(p_0, \hat\pi)}{\varepsilon^2} \asymp \frac{\beta}{\alpha} \log \frac{\alpha W_2^2(p_0, \hat\pi)}{\varepsilon^2}\,,
\end{aligned}$$

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

Now, for the guarantee in KL divergence, we "reinitialize" the algorithm with distribution p_N and apply the weakly convex result of Theorem 5.2. With the same choice of N and η and assuming ε is sufficiently small, we can apply Theorem 5.2 to obtain the guarantee

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

proving our desired result. □

E.3. Proof of Theorem 5.4

First, we need a lemma.

Lemma E.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*

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

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 B.1 and Lemma B.3 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

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

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

$$\begin{aligned}
\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 &= \mathbb{E}\|\Sigma_{\mu_1}^{-1}(X_1 - X_0)\|^2 && \text{(by Equation (27))} \\
&\le \beta^2\, \mathbb{E}\|X_1 - X_0\|^2 && \text{(since } \Sigma_{\mu_1}^{-1} \preceq \beta I\text{)} \\
&= \beta^2\, W_2^2(\mu_0, \mu_1)\,, && (38)
\end{aligned}$$

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

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

Note that by Equation (27), $\nabla_{\mathsf{BW}}\mathcal{V}_1(\mu) = -\nabla \log \mu_1 = -\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)$, 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

$$\mathsf{KL}(\mu \,\|\, \mu_1) = \mathcal{F}_1(\mu) - \mathcal{F}_1(\mu_1)\,. \tag{39}$$

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 $c \le (1-c)^2$. Define the random variable X_0' as follows:

$$\begin{aligned}
X_0' &:= X_0 - \frac{1}{\gamma}\,\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_0)[X_0] \\
&= X_0 - \frac{1}{\gamma}\,(\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0) - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1))[X_0] \\
&= X_0 - \frac{1}{\gamma}\,\big(-\Sigma_{\mu_0}^{-1}(X_0 - m_{\mu_0}) + \Sigma_{\mu_1}^{-1}(X_0 - m_{\mu_1})\big) && \text{(by Equation (27))} \\
&= \underbrace{\Big(I + \frac{1}{\gamma}\,\Sigma_{\mu_0}^{-1} - \frac{1}{\gamma}\,\Sigma_{\mu_1}^{-1}\Big)}_{:=M_0} X_0 + \frac{1}{\gamma}\,(-\Sigma_{\mu_0}^{-1}m_{\mu_0} + \Sigma_{\mu_1}^{-1}m_{\mu_1})\,.
\end{aligned}$$

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 B.1 applied to the potential $\mathcal{V}_1$, we find that

$$\begin{aligned}
\mathcal{V}_1(\mu_0') - \mathcal{V}_1(\mu_0) &\le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu_0)[X_0], X_0' - X_0\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0' - X_0\|^2 && (\text{since } -\nabla^2 \log\mu_0 \preceq \beta I) \\
&= -\mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0], X_0' - X_0\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0' - X_0\|^2\,. && (40)
\end{aligned}$$

Additionally, we note that since $\beta = c\gamma \le (1-c)^2\,\gamma$, 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 B.3 to obtain

$$\mathcal{H}(\mu_0') - \mathcal{H}(\mu_0) \le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0], X_0' - X_0\rangle + \frac{\gamma}{2}\,\mathbb{E}\|X_0' - X_0\|^2\,. \tag{41}$$

Hence, combining Equation (39) with Inequality 40 and Inequality 41, we obtain that

$$\begin{aligned}
-\mathsf{KL}(\mu_0 \,\|\, \mu_1) &\le \mathsf{KL}(\mu_0' \,\|\, \mu_1) - \mathsf{KL}(\mu_0 \,\|\, \mu_1) && (\text{since } \mathsf{KL}(\mu_0' \,\|\, \mu_1) \ge 0) \\
&= \mathcal{F}_1(\mu_0') - \mathcal{F}_1(\mu_0) && \text{(by Equation (39))} \\
&= [\mathcal{V}_1(\mu_0') - \mathcal{V}_1(\mu_0)] + [\mathcal{H}(\mu_0') - \mathcal{H}(\mu_0)] \\
&\le -\mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0], X_0' - X_0\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0' - X_0\|^2 && \text{(by Inequality 40)} \\
&\qquad + \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0], X_0' - X_0\rangle + \frac{\gamma}{2}\,\mathbb{E}\|X_0' - X_0\|^2 && \text{(by Inequality 41)} \\
&= \Big(-\frac{1}{\gamma} + \frac{\beta}{2\gamma^2} + \frac{1}{2\gamma}\Big)\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 && (\text{definition of } X_0') \\
&= -\frac{1-c}{2\gamma}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2\,. && (42)
\end{aligned}$$

To bound the LHS of this inequality, we again apply Lemma B.1 to the potential $\mathcal{V}_1$ as well as Lemma B.3 to $\mathcal{H}$ to obtain

$$\mathsf{KL}(\mu_0 \,\|\, \mu_1) = \mathcal{F}_1(\mu_0) - \mathcal{F}_1(\mu_1)$$

$$
\begin{aligned}
&= [\mathcal{V}_1(\mu_0) - \mathcal{V}_1(\mu_1)] + [\mathcal{H}(\mu_0) - \mathcal{H}(\mu_1)] \\
&\le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu_1)[X_1], X_0 - X_1\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0 - X_1\|^2 && \text{(by Lemma B.1 since } -\nabla^2 \log \mu_1 \preceq \beta I\text{)} \\
&\qquad + \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1], X_0 - X_1\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0 - X_1\|^2 && \text{(by Lemma B.3 since } \Sigma_{\mu_0}^{-1}, \Sigma_{\mu_1}^{-1} \preceq \beta I\text{)} \\
&= \beta\,\mathbb{E}\|X_0 - X_1\|^2 && \text{(since } \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu_1) + \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1) = \nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_1) = 0\text{)} \\
&= \beta\, W_2^2(\mu_0, \mu_1)\,. && (43)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{1}{2}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 &\le \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 \\
&\qquad + \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 \\
&\le \beta^2\, W_2^2(\mu_0, \mu_1) + \frac{2\gamma}{1-c}\,\mathsf{KL}(\mu_0 \,\|\, \mu_1) \\
&\qquad\qquad \text{(by Inequality 38 and Inequality 42)} \\
&\le 10\beta^2\, W_2^2(\mu_0, \mu_1)\,. \qquad \text{(by Inequality 43)}
\end{aligned}
$$

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 5.1 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 C.2, we obtain that

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

Telescoping this inequality, we find that

$$
\begin{aligned}
\min_{k \in \{0,\dots,N-1\}} \mathbb{E}\|X_{k+1} - X_k\|^2 &\le \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\|X_{k+1} - X_k\|^2 \\
&\le -\frac{2\eta}{N}\sum_{k=0}^{N-1} \mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)] \\
&= -\frac{2\eta}{N}\,\mathbb{E}[\mathcal{F}(p_N) - \mathcal{F}(p_0)] \\
&\le \frac{2\eta\Delta}{N}\,. \qquad (44)
\end{aligned}
$$

Now, let $(X_k, X_{k+1}^\star) \sim (p_k, p_{k+1})$ be optimally coupled for the W_2 distance. By Corollary D.2, we have that $\Sigma_k^{-1} \preceq \beta I$ for all k, meaning that we can apply Lemma E.1 with $\mu_0 = p_k$ and $\mu_1 = p_{k+1}$ to obtain that

$$
\mathbb{E}\big\|\nabla_{\mathsf{BW}}\mathcal{H}(p_k)[X_k] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}^\star]\big\|^2 \le 20\beta^2\, W_2^2(p_k, p_{k+1})\,. \qquad (45)
$$

Furthermore, we have that

$$
\begin{aligned}
\mathbb{E}\big\|\nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}^\star] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\big\|^2 &= \mathbb{E}\big\|\Sigma_{k+1}^{-1}\,(X_{k+1}^\star - X_{k+1})\big\|^2 \\
&\le \beta^2\,\mathbb{E}\big\|X_{k+1}^\star - X_{k+1}\big\|^2
\end{aligned}
$$

$$\begin{aligned}
&\le 2\beta^2\,\mathbb{E}\big\|X_{k+1}^\star - X_k\big\|^2 + 2\beta^2\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
&= 2\beta^2\,W_2^2(p_k, p_{k+1}) + 2\beta^2\,\mathbb{E}\|X_{k+1} - X_k\|^2\,. \qquad (46)
\end{aligned}$$

With these inequalities in mind, we obtain that

$$\begin{aligned}
\frac{1}{3}\,\|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|_{p_k}^2 &= \frac{1}{3}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}}\mathcal{H}(p_k)[X_k]\|^2 \\
&\le \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\|^2 + \mathbb{E}\big\|\nabla_{\mathsf{BW}}\mathcal{H}(p_k)[X_k] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}^\star]\big\|^2 \\
&\qquad + \mathbb{E}\big\|\nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}^\star] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\big\|^2 && \text{(by triangle inequality)} \\
&\le \frac{1}{\eta^2}\,\mathbb{E}\|X_{k+1} - X_k\|^2 + 22\,\beta^2 W_2^2(p_k, p_{k+1}) + 2\beta^2\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
& && \text{(by Inequality 45 and Inequality 46)} \\
&\le \Big(\frac{1}{\eta^2} + 24\beta^2\Big)\,\mathbb{E}\|X_{k+1} - X_k\|^2 && \text{(since (X_k, X_{k+1}) is a coupling of (p_k, p_{k+1}))} \\
&\le \frac{25}{\eta^2}\,\mathbb{E}\|X_{k+1} - X_k\|^2\,. && \text{(since $\beta \le \eta^{-1}$)}
\end{aligned}$$

Combining the above with Inequality 44, we obtain that

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

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

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

as desired. □

F. Proofs of the noisy algorithm convergence rates

We once again utilize Lemma 5.1 to obtain our desired rates of convergence. First, we must prove the bound on σ_k for Stochastic FB–GVI given in Lemma 5.5.

F.1. Proof of Lemma 5.5

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,

$$\begin{aligned}
\Sigma\,\mathbb{E}\nabla^2 V(X) &= \Sigma \int \nabla^2 V\,\mathrm{d}\mu \\
&= -\Sigma \int \nabla\mu \otimes \nabla V && \text{(integration by parts)} \\
&= -\Sigma \int \nabla \ln\mu \otimes \nabla V\,\mathrm{d}\mu \\
&= \int (x - m) \otimes \nabla V\,\mathrm{d}\mu(x) && \text{(since $-\Sigma\,\nabla \ln\mu(x) = x - m$)} \\
&= \mathbb{E}[(X - m) \otimes \nabla V(X)]\,. && (47)
\end{aligned}$$

Hence,

$$\begin{aligned}
\langle \mathbb{E}\nabla^2 V(X), \Sigma\rangle &= \langle \mathbb{E}[\Sigma^{-1}\,(X - m) \otimes \nabla V(X)], \Sigma\rangle && \text{(by Equation (47))} \\
&= \mathbb{E}\langle \Sigma^{-1}\,(X - m) \otimes \nabla V(X), \Sigma\rangle && \text{(linearity of expectation and trace)} \\
&= \mathbb{E}\langle \nabla V(X), X - m\rangle\,. && \text{(cyclicity of trace)}
\end{aligned}$$

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 (Brascamp & Lieb, 1976): 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

$$\begin{aligned}
\frac{1}{2}\,\sigma_k^2 &\le \mathbb{E}\|(\nabla^2 V(\hat{X}_k) - \mathbb{E}_{p_k}\nabla^2 V)(X_k - m_k)\|^2 + \mathbb{E}\|\nabla V(\hat{X}_k) - \mathbb{E}_{p_k}\nabla V\|^2 && \text{(by triangle inequality)}\\
&= \big\langle \mathbb{E}_{p_k}[(\nabla^2 V - \mathbb{E}_{p_k}\nabla^2 V)^2], \Sigma_k\big\rangle + \mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 && \text{(since } X_k \perp\!\!\!\perp \hat{X}_k\text{)}\\
&= \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle - \langle \mathbb{E}_{p_k}[\nabla^2 V]^2, \Sigma_k\rangle + \mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 &&\\
&\le \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle + \mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 && \text{(since } \langle \mathbb{E}_{p_k}[(\nabla^2 V)^2], \Sigma_k\rangle \ge 0\text{)}\\
&\le 2\,\mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle && \text{(by Brascamp–Lieb)}\\
&\le 2\beta\, \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\rangle && \text{(since } \nabla^2 V \preceq \beta I \text{ and } \nabla^2 V, \Sigma_k \succeq 0\text{)}\\
&= 2\beta\, \mathbb{E}\langle \nabla V(X_k), X_k - m_k\rangle && \text{(by Equation (47))}\\
&= 2\beta\, \underbrace{\mathbb{E}\langle \nabla V(Z), Z - \hat{m}\rangle}_{\mathsf{err}_1} + 2\beta\, \underbrace{\mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle}_{\mathsf{err}_2}\\
&\qquad + 2\beta\, \underbrace{\mathbb{E}\langle \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle}_{\mathsf{err}_3} + 2\beta\, \underbrace{\mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), Z - \hat{m}\rangle}_{\mathsf{err}_4}\,.
\end{aligned}$$

Now, we have the following:

$$\begin{aligned}
\mathsf{err}_1 &= \mathbb{E}\langle \nabla V(Z), Z - \hat{m}\rangle = \langle \mathbb{E}\nabla^2 V(Z), \hat{\Sigma}\rangle = \mathrm{Tr}(I) && \text{(by Equation (47) and the stationarity conditions in (28))}\\
&= d\,,\\
\mathsf{err}_2 &= \mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle\\
&\le \frac{1}{2\beta}\,\mathbb{E}\|\nabla V(X_k) - \nabla V(Z)\|^2 + \frac{\beta}{2}\,\mathbb{E}\|(X_k - m_k) - (Z - \hat{m})\|^2 && \text{(Young's inequality)}\\
&\le \beta\,\mathbb{E}\|X_k - Z\|^2 && \text{(since } \nabla V \text{ is } \beta\text{-Lipschitz)}\\
&= \beta\, W_2^2(\mu_k, \hat{\pi})\,, && \text{(since } (X_k, Z) \text{ are optimally coupled)}\\
\mathsf{err}_3 &= \mathbb{E}\langle \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle\\
&\le \frac{1}{4\beta}\,\mathbb{E}\|\nabla V(Z)\|^2 + \beta\,\mathbb{E}\|(X_k - m_k) - (Z - \hat{m})\|^2 && \text{(Young's inequality)}\\
&\le \frac{1}{4\beta}\,\mathbb{E}\langle \nabla^2 V(Z)^2, \hat{\Sigma}\rangle + \beta\, W_2^2(\mu_k, \hat{\pi}) && \text{(Brascamp–Lieb, optimal coupling of } (X_k, Z)\text{)}\\
&\le \frac{d}{4} + \beta\, W_2^2(\mu_k, \hat{\pi})\,, && \text{(since } \mathbb{E}_{\hat{\pi}}\nabla^2 V = \hat{\Sigma}^{-1} \text{ by Equation (28) and } \nabla^2 V \preceq \beta I\text{)}\\
\mathsf{err}_4 &= \mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), Z - \hat{m}\rangle\\
&\le \frac{\mathrm{Tr}(\hat{\Sigma})}{d}\,\mathbb{E}\|\nabla V(X_k) - \nabla V(Z)\|^2 + \frac{d}{4\,\mathrm{Tr}(\hat{\Sigma})}\,\mathbb{E}\|Z - \hat{m}\|^2 && \text{(Young's inequality)}\\
&\le \frac{\beta^2\,\mathrm{Tr}(\hat{\Sigma})}{d}\,\mathbb{E}\|X_k - Z\|^2 + \frac{d}{4\,\mathrm{Tr}(\hat{\Sigma})}\,\mathrm{Tr}(\hat{\Sigma}) && \text{(since } \nabla V \text{ is } \beta\text{-Lipschitz)}\\
&\le \frac{\beta^2\,\mathrm{Tr}(\hat{\Sigma})}{d}\,W_2^2(\mu_k, \hat{\pi}) + \frac{d}{4}\,.
\end{aligned}$$

Combining these, we obtain that

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

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

Note that in the strongly convex case, by Equation (28), we obtain that

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

so this bound simplifies to

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

This concludes our proof. □

F.2. One-step inequality using the bound on σ_k

We apply the error bound in Lemma 5.5 along with the one-step inequality of Lemma 5.1 with $\nu = \hat{\pi}$ and $\eta \le \frac{1}{2\beta}$. This gives us the inequality

$$\begin{aligned}
\mathbb{E}W_2^2(p_{k+1}, \hat{\pi}) &\le (1 - \alpha\eta)\mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})) + 2\eta^2\,\mathbb{E}\sigma_k^2 \\
&\le \bigl(1 - \alpha\eta + 24\beta^3\eta^2\lambda_{\max}(\hat{\Sigma})\bigr)\,\mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})) + 12\beta\eta^2 d \\
&\le \exp\bigl(-\alpha\eta + 24\beta^3\eta^2\lambda_{\max}(\hat{\Sigma})\,\bigr)\mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})) + 12\beta\eta^2 d\,. \qquad (48)
\end{aligned}$$

F.3. Proof of Theorem 5.6

Proof. Define $c \coloneqq 24\beta^3\lambda_{\max}(\hat{\Sigma})$. Since V is convex by assumption, we may take $\alpha = 0$ in Inequality 48 to obtain that

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

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

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

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

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

Now, we note that

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

Thus, we obtain the inequality

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

Hence, taking

$$\begin{aligned}\eta&\asymp\frac{\varepsilon^2}{cW_2^2(p_0,\hat\pi)\vee\beta d}\\&\asymp\frac{\varepsilon^2}{\beta^3\lambda_{\max}(\hat\Sigma)\,W_2^2(p_0,\hat\pi)\vee\beta d}\,,\\N&\gtrsim\frac{W_2^2(p_0,\hat\pi)}{\eta\varepsilon^2}\\&\asymp\frac{W_2^2(p_0,\hat\pi)}{\varepsilon^4}\,\big(\beta^3\lambda_{\max}(\hat\Sigma)\,W_2^2(p_0,\hat\pi)\vee\beta d\big)\,,\end{aligned}$$

we get the guarantee

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

□

F.4. Proof of Theorem 5.7

Proof. In the strongly convex case where $0\prec\alpha I\preceq\nabla^2V$, we have the eigenvalue guarantee $\lambda_{\max}(\hat\Sigma)\le\frac{1}{\alpha}I$, since $\mathbb{E}_\pi\nabla^2V=\hat\Sigma^{-1}$ by (28). Hence, under the assumption that $\eta\le\frac{\alpha^2}{48\beta^3}$, Inequality 48 implies that

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

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

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

Hence, with the choice

$$\begin{aligned}\eta&\asymp\frac{\varepsilon^2}{\beta d}\,,\\N&\gtrsim\frac{1}{\alpha\eta}\log\frac{\alpha W_2^2(p_0,\hat\pi)}{\varepsilon^2}\\&\asymp\frac{\beta d}{\alpha\varepsilon^2}\log\frac{\alpha W_2^2(p_0,\hat\pi)}{\varepsilon^2}\,,\end{aligned}$$

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 weakly convex result of Theorem 5.6. This argument is inspired by Durmus et al. (2019). 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

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

Thus, applying Theorem 5.6 with our choice of step size η and iteration count N, we obtain that

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

proving our desired result. □