Hessian-guided Perturbed Wasserstein Gradient Flows for Escaping Saddle Points

Naoya Yamamoto
The University of Tokyo
yamamoto-naoya251@g.ecc.u-tokyo.ac.jp

Juno Kim*
UC Berkeley
junokim@berkeley.edu

Taiji Suzuki
The University of Tokyo, RIKEN AIP
taiji@mist.i.u-tokyo.ac.jp

Abstract

Wasserstein gradient flow (WGF) is a common method to perform optimization over the space of probability measures. While WGF is guaranteed to converge to a first-order stationary point, for nonconvex functionals the converged solution does not necessarily satisfy the second-order optimality condition; i.e., it could converge to a saddle point. In this work, we propose a new algorithm for probability measure optimization, *perturbed Wasserstein gradient flow* (PWGF), that achieves second-order optimality for general nonconvex objectives. PWGF enhances WGF by injecting noisy perturbations near saddle points via a Gaussian process-based scheme. By pushing the measure forward along a random vector field generated from a Gaussian process, PWGF helps the solution escape saddle points efficiently by perturbing the solution towards the smallest eigenvalue direction of the Wasserstein Hessian. We theoretically derive the computational complexity for PWGF to achieve a second-order stationary point. Furthermore, we prove that PWGF converges to a global optimum in polynomial time for strictly benign objectives.

1 Introduction

We consider the general problem of optimizing a probability measure: $\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} F(\mu)$, where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of Borel probability measures on $\mathbb{R}^d$ with finite second moment and $F:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$ is a given functional which is not necessarily convex. Optimization of measures appears extensively in machine learning and statistics either explicitly or implicitly, such as in sampling and variational inference (Jordan et al., 1998; Liu & Wang, 2016), generative models (Arbel et al., 2019; Chu et al., 2019) and training neural networks (Mei et al., 2018; Nitanda et al., 2022), and has garnered significant attention both theoretically and practically. To solve such problems, the **Wasserstein gradient flow** (WGF) is extensively employed:

$$\partial_t\mu_t + \nabla\cdot\left(-\nabla\frac{\delta F}{\delta\mu}(\mu_t)\mu_t\right) = 0, \quad (\mu_t)\subset\mathcal{P}_2(\mathbb{R}^d). \tag{1}$$

WGF is the continuity equation of the velocity field $\nabla\frac{\delta F}{\delta\mu}$ and can be interpreted as gradient descent with respect to the 2-Wasserstein metric (Jordan et al., 1998). Moreover, (1) is equivalent to the continuous-time and infinite-particle limit of first-order optimization algorithms such as gradient descent, and serves as the foundation for numerous machine learning methods; see Appendix A for a discussion of related works and applications. It is thus of paramount importance to understand the convergence of WGF, and to develop algorithms which guarantee well-behaved solutions.

*This work was primarily conducted while the author was at the University of Tokyo and RIKEN AIP.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Convergence of WGF. A prominent line of work on the convergence of WGF is the study of neural network optimization using *mean-field theory* (Mei et al., 2018). Mean-field theory models the evolution of an interacting particle system as the number of particles tends to infinity, such as the training dynamics of an infinite-width neural network, through the WGF of the limiting distribution. For simple problem settings such as regression with two-layer networks, the corresponding loss functional is shown to be (linearly) convex, allowing for global convergence analysis of WGF (Chizat & Bach, 2018; Mei et al., 2018). Furthermore, Nitanda et al. (2022); Chizat (2022); Suzuki et al. (2023) obtained linear convergence under the log-Sobolev inequality (LSI) condition using mean-field Langevin dynamics. This adds isotropic noise via Brownian motion, corresponding to an additional entropy regularization which effectively makes the objective strongly convex.

Non-convex objectives. While these works primarily rely on convexity, the vast majority of objectives arising in deep learning (such as the loss function for neural networks with three or more layers) are non-convex, even when lifted to the space of measures. However, due to the inherent difficulty of infinite-dimensional non-convex optimization, convergence guarantees for WGF in this regime has been extremely limited. Recently, Kim & Suzuki (2024) studied a transformer model combining a mean-field neural network with linear attention for in-context learning, resulting in a non-convex optimization problem. Through landscape analysis, they showed that the objective possesses a desirable benignity property: all second-order optima are either saddle points or global optima. A similar result has been demonstrated for energy kernels such as the MMD functional (Boufadene & Vialard, 2024). These findings suggest the necessity of probability measure optimization algorithms that account for second-order optimality. While the first-order optimality of WGF has been shown by Lanzetti et al. (2025), there are few studies that rigorously consider second-order conditions.

Perturbation methods. The main difficulty of non-convex optimization is due to the existence of saddle points. A first-order method with naive initialization may end up converging to a saddle point. Moreover, Du et al. (2017) showed that gradient descent can be significantly delayed near saddle points, taking exponential time to converge. For optimization in finite-dimensional Euclidean space, Ge et al. (2015); Jin et al. (2017); Li (2019) proposed methods to overcome this issue by introducing perturbations in the vicinity of saddle points to 'fall off' the saddle and escape efficiently. These methods, often referred to as **perturbed gradient descent** (PGD), are particularly useful as they achieve second-order optimality in polynomial time while primarily relying on first-order techniques. In particular, PGD guarantees global convergence for problems satisfying a strict benignity property, such as matrix factorization (Jin et al., 2017) and tensor decomposition (Ge et al., 2015).

With these issues in mind, we ask the following question:

Can we develop a perturbative version of Wasserstein gradient flow for non-convex objectives which converges efficiently to second-order optimal points?

Our Contributions. In this study, we propose a perturbative modification to WGF that efficiently avoids saddle points. Considering the tangent bundle structure of $\mathcal{P}_2(\mathbb{R}^d)$ induced by the Wasserstein distance, it is natural to extend the notion of perturbation in Euclidean space to measure space by defining perturbations of the drift function through randomly generated vector fields. Such a perturbative method was conjectured to improve convergence by Kim & Suzuki (2024), but without any theoretical guarantees. Our contributions are summarized as follows:

- We propose a new implementable algorithm, **perturbed Wasserstein gradient flow (PWGF)**, that guarantees second-order optimality for general smooth and non-convex functionals. Unlike the method proposed by Kim & Suzuki (2024), which injects isotropic noise near saddles into the WGF, we guarantee improvement by *directing* the noise using the (Wasserstein) Hessian. Specifically, PWGF pushes the measure along a random velocity field generated from a Gaussian process whose covariance is constructed from the Hessian of the objective.
- We prove that PWGF effectively avoids saddle points and reaches second-order optimal points in time that depends polynomially on precision parameters, enabling the optimization of non-convex distributional objectives. Compared to the finite-dimensional setting (Li, 2019), the analysis requires a careful treatment of an infinite dimensional objective. For this purpose, we utilize techniques from Wasserstein geometry, optimal transport and the theory of Gaussian processes.

Organization. The paper is organized as follows. Section 2 provides theoretical preliminaries, supplemented in Appendix B. Second-order optimality conditions are presented in Section 3. In Section 4, we introduce the proposed PWGF algorithm. Section 5 presents the convergence analysis, along with a rough sketch of the proof. Numerical experiments are provided in Appendix F.

2 Preliminaries

In this paper, we consider the optimization problem $\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} F(\mu)$ over the space of probability distributions, where $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is a real-valued lower-bounded functional defined on $\mathcal{P}_2(\mathbb{R}^d)$. This section introduces the problem of probability measure optimization and reviews key concepts in optimal transport and Wasserstein geometry. See Appendix B for further details.

Notation. Let Id be the identity map on $\mathbb{R}^d$. The canonical projection to the ith coordinate is denoted by p_i. The Euclidean inner product and operator norm are $\langle\cdot,\cdot\rangle, \|\cdot\|$. The Frobenius norm is $\|\cdot\|_{\mathrm{F}}$. The set of real-valued functions on $\mathbb{R}^d$ that are infinitely differentiable with compact support is $C_0^\infty(\mathbb{R}^d)$. The inner product and norm or operator norm in $L^2(\mu)^d$ is $\langle\cdot,\cdot\rangle_{L^2(\mu)}, \|\cdot\|_{L^2(\mu)}$. The trace norm of an operator is $\|\cdot\|_{\mathrm{Tr},L^2(\mu)}$, and the Hilbert-Schmidt norm is $\|\cdot\|_{\mathrm{HS},L^2(\mu)}$. $T \succeq O$ indicates that an operator T is positive semi-definite. The smallest eigenvalue of a compact operator T is denoted by $\lambda_{\min}T$. The exponential of an operator T is denoted by e^T or $\exp T$. We use $\tilde{O}$ to denote big O notation ignoring logarithmic factors.

The set of all Borel probability measures on $\mathbb{R}^d$ with finite second moments is denoted by $\mathcal{P}_2(\mathbb{R}^d)$, and the subset of measures absolutely continuous with respect to Lebesgue measure is denoted by $\mathcal{P}_2^a(\mathbb{R}^d)$. The Dirac measure on $x \in \mathbb{R}^d$ is $\delta_x \in \mathcal{P}_2(\mathbb{R}^d)$. $f\#\mu$ denotes the pushforward of $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ by a measurable map $f : \mathbb{R}^d \to \mathbb{R}^d$.

2.1 Wasserstein Geometry

Definition 2.1 (Wasserstein metric)**.** The 2-Wasserstein metric between $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$W_2(\mu,\nu)^2 \coloneqq \min_{\gamma\in\Gamma(\mu,\nu)} \int \|x-y\|^2 \gamma(\mathrm{d}x\mathrm{d}y), \tag{2}$$

where $\Gamma(\mu,\nu)$ represents the set of all transport plans (or the set of couplings) of μ,ν, that is, all joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginal distributions are μ and ν. We denote the set of all optimal transport plans by $\Gamma_o(\mu,\nu)$ and the optimal transport map by $\mathcal{T}_\mu^\nu$.

A fundamental dynamics in Wasserstein space is the continuity equation with velocity field v_t,

$$\partial_t\mu_t + \nabla\cdot(v_t\mu_t) = 0, \quad \mu_t \in \mathcal{P}_2(\mathbb{R}^d),\ t \in I. \tag{3}$$

Intuitively, (3) describes how a particle distribution μ_t evolves along a vector field v_t. In particular, WGF (1) moves particles according to the vector field $v_t = -\nabla\frac{\delta F}{\delta\mu}(\mu_t)$ (see Section 2.3). Moreover, the continuity equation with velocity field v_t such that

$$v_t \in \mathrm{Tan}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d) \coloneqq \overline{\{\nabla\varphi \mid \varphi \in C_0^\infty(\mathbb{R}^d)\}}^{L^2(\mu_t)}.$$

can be locally approximated by a pushforward along v_t (Proposition B.9). Therefore the continuity equation is computationally approximated by the pushforward representation:

$$\mu_{t+\Delta t} \leftarrow (\mathrm{Id} + \Delta t v_t)\#\mu_t. \tag{4}$$

The absolute continuity of the curve μ_t with respect to the Wasserstein distance is equivalent to satisfying (3) for some $v_t \in L^2(\mu_t)$ (Ambrosio et al. (2008), Theorem 8.3.1). In this sense, the continuity equation provides a concept of differentiation consistent with the Wasserstein metric. For further background on optimal transport theory, see Appendix B.

2.2 Wasserstein Gradient

The Wasserstein gradient is the fundamental quantity for first-order analysis in Wasserstein space.

Definition 2.2 (Wasserstein gradient)**.** The Wasserstein gradient of F at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is a vector field $\nabla_\mu F : \mathbb{R}^d \to \mathbb{R}^d$ such that for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Gamma_o(\mu, \nu)$,

$$F(\nu) - F(\mu) = \int \nabla_\mu F(x)^\top (y - x)\gamma(\mathrm{d}x\mathrm{d}y) + O\big(W_2(\mu, \nu)^2\big).$$

The first variation also frequently appears in the context of probability measure optimization (cf. proximal Gibbs measure (Nitanda et al., 2022)).

Definition 2.3 (First variation)**.** The first variation $\frac{\delta F}{\delta \mu} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ is defined as a functional satisfying for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\frac{\mathrm{d}}{\mathrm{d}h}\bigg|_{h=0} F(\mu + h(\nu - \mu)) = \int \frac{\delta F}{\delta \mu}(\mu, x)(\mu - \nu)(\mathrm{d}x). \tag{5}$$

A naive computation might suggest that $\nabla_\mu F = \nabla \frac{\delta F}{\delta \mu}$; however, this is not generally true without additional conditions. Nevertheless, we do not distinguish between the two, see Appendix B.2.

First-order optimality. The Wasserstein gradient allows us to construct first-order approximations of functionals. Furthermore, Lanzetti et al. (2025) demonstrated two analogies to finite-dimensional optimization. The first is that $\nabla_\mu F = 0$ serves as a necessary condition for local optimality. The second is that $\nabla_\mu F = 0$ becomes a sufficient condition for global optimality in the convex case. Based on these considerations, we define the following.

Definition 2.4 (First-order stationary point)**.** Supoose that a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ satisfies sufficient smoothness. We say that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is a *first-order stationary point*, if μ satisfies $\nabla_\mu F = 0$ μ-a.e.

2.3 Wasserstein Gradient Flow (WGF)

As a counterpart of gradient descent in Euclidean space, the WGF in Wasserstein space is defined as

$$\partial_t \mu_t + \nabla \cdot (-\nabla_\mu F(\mu_t)\mu_t) = 0, \quad (\mu_t) \subset \mathcal{P}_2(\mathbb{R}^d). \tag{6}$$

From previous observations, the direction $v_t = -\nabla_\mu F(\mu_t)$ yields the steepest descent direction of the objective F. Furthermore, by the chain rule (Proposition C.3) it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\|\nabla_\mu F(\mu_t)\|^2_{L^2(\mu_t)} \leq 0$$

This indicates that the WGF monotonically decreases the objective function. Indeed, the WGF can be interpreted as a gradient descent method in the space of probability measures (Jordan et al., 1998). Moreover, WGF becomes stationary iff the solution is at a first-order stationary point (Definition 2.4).

3 Second Order Optimality on Probability Space

In order to study second-order behavior, we define a suitable class of sufficiently regular functionals.

Definition 3.1 (Sufficient smoothness)**.** A functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is *sufficiently smooth* if

- F admits a $L^2(\mu)$-integrable Wasserstein gradient $\nabla_\mu F$ at all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.
- $\nabla_\mu F(\mu, x)$ further admits Wasserstein gradient $\nabla^2_\mu F : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and is differentiable with respect to the second coordinate x for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and μ-a.e. x. Furthermore, $\nabla^2_\mu F(\mu)$ is $L^2(\mu \otimes \mu)$ integrable and μ-ess sup$\|\nabla \nabla_\mu F(\mu, x)\| < \infty$.

Assumption 1. *The objective $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is a sufficiently smooth functional.*

Building upon the discussion of first-order optimality, we extend the analysis to second-order conditions. For simplicity of notation, we define the following operators H_μ, H'_μ for $f \in L^2(\mathbb{R}^d)^d$:

$$H_\mu f(x) = \int \nabla^2_\mu F(\mu, x, y) f(y) \mu(\mathrm{d}y), \quad H'_\mu f(x) = \nabla \nabla_\mu F(\mu, x) f(x).$$

We establish the following proposition.

Proposition 3.2. *For $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ a sufficiently smooth functional, for all $v \in L^2(\mu)^d$,*

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\bigg|_{h=0} F((\mathrm{Id} + hv)\#\mu) = \langle v, (H_\mu + H'_\mu)v \rangle_{L^2(\mu)}. \tag{7}$$

This demonstrates that the second order term of a vector field perturbation is characterized by the operator $H_\mu + H'_\mu$. For a more general version, see Proposition C.7. In particular, when examining stability at first-order stationary points, we have $\nabla_\mu F(\mu) = 0$, which implies $\nabla\nabla_\mu F(\mu) = 0$, that is, $H'_\mu = 0$. Consequently, the change in F due to perturbations along vector fields is determined solely by the integral operator H_μ up to second order. Note that this does not necessarily hold for $\mu \notin \mathcal{P}_2^a(\mathbb{R}^d)$; for further details, refer to Appendix C.

Next, paralleling the work by Lanzetti et al. (2025), we demonstrate that $H_\mu \succeq O$ serves as a sufficient condition for local stability, establishing the analogy with second-order conditions in Euclidean space.

Proposition 3.3 (second-order necessary condition)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be sufficiently smooth. If $\mu^* \in \mathcal{P}_2^a(\mathbb{R}^d)$ is a local minimum of F, then it holds that $H_{\mu^*} \succeq O$.*

Based on this observation, we define the second-order optimality condition for probability measures.

Definition 3.4 (second-order stationary point)**.** For $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ a sufficiently smooth functional and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

- We say that μ is a *second-order stationary point* if μ is a first-order stationary point and $H_\mu \succeq O$.
- We say that μ is a *saddle point* if μ is a first order stationary point and satisfies $H_\mu \not\succeq O$, i.e., the smallest eigenvalue of H_μ is strictly negative : $\lambda_{\min} H_\mu < 0$.

Since we seek approximate solutions for F, we also define *approximate* second order stationary points and saddle points.

Definition 3.5 (approximate second-order stationary point)**.** Suppose that $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is sufficiently smooth and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

- We say that μ is an (ε, δ)-stationary point, if $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} \le \varepsilon$ and $\lambda_{\min} H_\mu \ge -\delta$.
- We say that μ is an (ε, δ)-saddle point, if $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} \le \varepsilon$ and $\lambda_{\min} H_\mu < -\delta$.

While this definition is useful for convergence analysis, it is not fully appropriate in light of the second-order expansion (7). This is because the condition $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} \le \varepsilon$ only implies that $\nabla_\mu F(\mu)$ is small in the sense of L^2 norm and does not indicate that $\nabla\nabla_\mu F(\mu)$ is close to zero. Therefore, in that case, second order optimality should be determined by $H_\mu + H'_\mu$ rather than H_μ. Consequently, this paper assumes that the L^2 norm of Wasserstein gradients being small implies that the supremum norm of their gradients is also small.

Assumption 2. *For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, it holds that*

$$\mu\text{-ess}\sup_{x \in \mathbb{R}^d} \|\nabla\nabla_\mu F(\mu, x)\| \le R_2 \|\nabla_\mu F(\mu)\|_{L^2(\mu)}. \tag{8}$$

Under this assumption, a small $\|\nabla_\mu F(\mu)\|_{L^2(\mu)}$ implies that $\mu\text{-ess}\sup \|\nabla\nabla_\mu F(\mu, x)\| = \|H'_\mu\|_{L^2(\mu)}$ is also small, justifying the definition of (ε, δ)-stationary points.

3.1 Global Convergence for Strictly Benign Objectives

It is known that certain non-convex optimization problems possess a desirable property called *benignity*, i.e. all local minima must be global minima. In Euclidean spaces, examples include tensor decomposition (Ge et al., 2015; Jin et al., 2017). With our definitions of (approximate) second-order optimality, a similar property can be considered for the Wasserstein space.

Definition 3.6 (Strict benignity)**.** The functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is said to be $(\varepsilon, \delta, \alpha)$-strictly benign if at least one of the following conditions holds for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$:

1. $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} > \varepsilon$.

2. $\lambda_{\min} H_\mu < -\delta$.

3. $W_2(\mu, \mu^o) \leq \alpha$ for some global optima μ^o.

We provide examples of non-convex objective functions that exhibit strict benignity. For details on the properties of each objective function, refer to the appendix and the cited papers.

Example 1 (Matrix decomposition)**.** This example is inspired by the fact that finite-dimensional tensor decomposition exhibits strict benignity. We use a mean-field two-layer neural network

$$h_\mu(z) = \int a\sigma(w^\top z)\mu(\mathrm{d}a\mathrm{d}w)$$

to learn a rank-one matrix induced by the target measure μ^o, where z is a data input that follows a certain distribution, the parameter is $x = (a, w) \in \mathbb{R}^{k+l}$, and σ is an activation function, which we set as the sigmoid function. The objective functional can be expressed as follows:

$$F(\mu) = \mathrm{E}_z[\|h_{\mu^o}(z)h_{\mu^o}(z)^\top - h_\mu(z)h_\mu(z)^\top\|_{\mathrm{F}}^2]$$

We defer a detailed analysis of this objective to Appendix G.

Example 2 (3-layer neural network)**.** Consider a three-layer neural network model consisting of a mean-field two-layer network followed by a linear layer. Assuming realizability, the L^2 loss with respect to a teacher network $T^* h_{\mu^*}(z)$ can be written as follows:

$$\tilde{F}(\mu, T) = \mathrm{E}_z\left[\|T^* h_{\mu^*}(z) - T h_\mu(z)\|^2\right] \tag{9}$$

In this case, the optimization problem with respect to T can be explicitly solved. Defining $\mu^o = (T^* \times \mathrm{Id}_{\mathbb{R}^l})\#\mu^*$ and $\Sigma_{\mu,\nu} = \mathrm{E}_z\left[h_\mu(z)h_\nu(z)^\top\right]$, the optimal T satisfies $T = \Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1}$ and (9) reduces to the following probability measure optimization problem:

$$F(\mu) = \mathrm{E}_z\left[\|h_{\mu^o}(z) - \Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1}h_\mu(z)\|^2\right]. \tag{10}$$

Kim & Suzuki (2024) analyze this optimization problem and essentially show strict benignity, assuming $\Sigma_{\mu,\mu}$ is bounded away from degeneracy.

Example 3 (Coulomb MMD)**.** The maximum mean discrepancy (MMD) with Coulomb kernel can be expressed as:

$$F(\mu) = \int \frac{(\mu - \mu^o)^{\otimes 2}(\mathrm{d}x\mathrm{d}y)}{\|x - y\|^{d-2}}.$$

This energy functional has been studied by Boufadene & Vialard (2024), who showed that any absolutely continuous stationary point must be a global optimum. Therefore, F becomes benign in regions that do not involve singular distributions.

4 Perturbed Wasserstein Gradient Flow

In this section, we introduce our proposed algorithm that incorporates perturbations in measure space along random velocity fields to escape saddle points.

4.1 Perturbations in Wasserstein Space

In Euclidean spaces, perturbed gradient descent (PGD) is a first-order optimization method capable of escaping saddle points efficiently by adding perturbations (Ge et al., 2015; Jin et al., 2017; Li, 2019). Despite relying solely on first order information, this method effectively achieves second order optimality, making it highly practical especially for problems with known benign structure. A typical perturbation technique in Euclidean spaces involves adding vectors sampled uniformly from a ball of small radius. Intuitively, such perturbations can uniformly explore all directions, making it likely to include the unstable direction corresponding to the smallest eigenvalue of the Hessian, thereby quickly 'falling off' the saddle.

To extend this idea to the space of probability measures, two issues must be addressed: **(1)** how to induce perturbations in Wasserstein space, and **(2)** whether the perturbation includes the unstable direction that maximally reduces the objective. For **(1)**, Kim & Suzuki (2024) proposed introducing

Algorithm 1 PWGF (continuous-time)

set hyperparameter $\eta_p = \tilde{O}\left(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\right)$, $T_{\text{thres}} = \tilde{O}\left(\frac{1}{\delta}\right)$, and $F_{\text{thres}} = \tilde{O}(\delta^3)$
initialize $\mu^{(0)}$ and $t_p = -T_{\text{thres}}$
for $t > 0$ **do**
  if $\|\nabla_\mu F(\mu_t)\|_{L^2(\mu_t)} \le \varepsilon$ and $t - t_{\text{p}} > t_{\text{thres}}$ **then**
    $\xi \sim \text{GP}(0, K_{\mu_t})$, $\mu_t \leftarrow (\text{Id} + \eta_p \xi)\sharp\mu_t$, $t_{\text{p}} \leftarrow t$
  end if
  if $t = t_{\text{p}} + T_{\text{thres}}$ and $F(\mu_{t_{\text{p}}}) - F(\mu_t) \le F_{\text{thres}}$ **then**
    return $\mu_{t_{\text{p}}}$
  end if
  $\partial_t \mu_t + \nabla \cdot (-\nabla_\mu F(\mu_t)\mu_t) = 0$
end for

perturbations via a multivariate Gaussian process $\xi \sim \text{GP}(0, K)$ with a fixed positive semi-definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$, transforming μ into $(\text{Id}+\eta\xi)\#\mu$. Leveraging vector fields to represent infinitesimal changes in probability measures is both natural and practical. However, they did not provide a theoretical justification for the effectiveness of this method.

We resolve this issue and guarantee **(2)** by constructing a *measure-dependent* kernel $K = K_\mu$ based on the Wasserstein Hessian of the objective:

$$K_\mu(x, y) = \int \nabla_\mu^2 F(\mu, x, z) \nabla_\mu^2 F(\mu, z, y) \mu(\mathrm{d}z). \tag{11}$$

The integral operator defined by this kernel coincides with the squared Hessian operator H_μ^2. Specifically, it holds for $f \in L^2(\mu)^d$ that

$$\begin{aligned} H_\mu^2 f(x) &:= \int \nabla_\mu^2 F(\mu, x, z) \left(\int \nabla_\mu^2 F(\mu, z, y) f(y) \mu(\mathrm{d}y) \right) \mu(\mathrm{d}z) \\ &= \int \left(\int \nabla_\mu^2 F(\mu, x, z) \nabla_\mu^2 F(\mu, z, y) \mu(\mathrm{d}z) \right) f(y) \mu(\mathrm{d}y) \\ &= \int K_\mu(x, y) f(y) \mu(\mathrm{d}y). \end{aligned}$$

The Hessian-based kernel K_μ is symmetric, positive semi-definite, and meets the trace-class condition due to the integrability of $\nabla_\mu F(\mu)$. This means that K_μ satisfies the requirements for a Gaussian process kernel. More importantly, the Gaussian process ξ is 'directed' by H_μ, ensuring that the perturbation is likely to include the direction corresponding to the smallest eigenvalue of H_μ, thereby achieving a reduction in the objective (Proposition 5.4). Details are provided in Appendix D.

4.2 Algorithm of PWGF

Based on the above considerations, we propose **perturbed Wasserstein gradient flow (PWGF)** as a method for solving non-convex optimization problems of probability measures. PWGF alternates between perturbing near saddle points using a Gaussian process with the kernel defined in (11) and evolving via WGF when not near saddle points.

We present the specific algorithm in Algorithm 1, including the saddle detection mechanism. In the space of probability measures, the Hessian H_μ is an operator on $L^2(\mu)^d$, and determining whether its smallest eigenvalue is below a certain threshold is computationally challenging. Drawing on Jin et al. (2017), we propose a practical criterion: Proposition 5.4 says that if a perturbation is introduced at an (ε, δ)-saddle point, the objective function decreases with high probability after a certain period of WGF. Consequently, by always introducing perturbations at first-order stationary points, we can determine whether a given stationary point is a saddle point based on the decrease on the objective within a fixed time threshold.

The time-discretized version of PWGF is provided in Algorithm 2. To implement PWGF numerically, the optimal probability measure is approximated by the ensemble average of N particles as $\mu = \frac{1}{N}\sum_{j=1}^N \delta_{x_j}$ and the pushforward and descent steps are directly applied to each particle as

$$\begin{aligned} \text{(perturbation)} \quad & x_j \leftarrow x_j + \eta_p \cdot \xi(x_j), \\ \text{(gradient descent)} \quad & x_j \leftarrow x_j - \eta \nabla F(\mu^{(k)}, x_j). \end{aligned}$$

Algorithm 2 PWGF (discrete-time)

set hyperparameter $\eta_p = \tilde{O}\left(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\right)$, $\eta = O(1)$, $k_{\mathrm{thres}} = \tilde{O}\left(\frac{1}{\delta}\right)$, and $F_{\mathrm{thres}} = \tilde{O}(\delta^3)$
initialize $\mu^{(0)}$ and $k_p = k_{\mathrm{thres}}$
for $k = 0, 1, \ldots$ **do**
  if $\left\|\nabla_\mu F(\mu^{(k)})\right\|_{L^2(\mu^{(k)})} \leq \varepsilon$ and $k - k_{\mathrm{p}} > k_{\mathrm{thres}}$ **then**
    $\xi \sim \mathrm{GP}(0, K_{\mu^{(k)}})$, $\mu^{(k)} \leftarrow (\mathrm{Id} + \eta_p \xi)\#\mu^{(k)}$
    $k_p \leftarrow k$
  end if
  if $k = k_{\mathrm{p}} + k_{\mathrm{thres}}$ and $F(\mu^{(k_{\mathrm{p}})}) - F(\mu^{(k)}) \leq F_{\mathrm{thres}}$ **then**
    return $\mu^{(k_{\mathrm{p}})}$
  end if
  $\mu^{(k+1)} \leftarrow (\mathrm{Id} - \eta \nabla_\mu F(\mu^{(k)}))\#\mu^{(k)}$
end for

5 Convergence Analysis

In this section, we provide the theoretical guarantee for the PWGF algorithm. In addition to the assumptions introduced in the previous sections, we impose the following Lipschitz continuity of the Wasserstein gradient and Hessian.

Assumption 3. *F satisfies the following smoothness:*

- *$\nabla_\mu F$ is L_1-Lipschitz, i.e. for any $\gamma \in \Gamma(\mu, \nu)$,*

$$\int \|\nabla_\mu F(\mu, x) - \nabla_\mu F(\nu, y)\|^2 \gamma(\mathrm{d}x\mathrm{d}y) \leq L_1^2 \int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y). \tag{12}$$

- *$\nabla_\mu^2 F$ is L_2-Lipschitz, i.e. for any $\gamma \in \Gamma(\mu, \nu)$,*

$$\int \|\nabla_\mu^2 F(\mu, x_1, x_2) - \nabla_\mu^2 F(\nu, y_1, y_2)\|^2 \gamma^{\otimes 2}(\mathrm{d}x\mathrm{d}y) \leq L_2^2 \int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y). \tag{13}$$

- *$\nabla \nabla_\mu F$ is L_3-Lipschitz, i.e. for any $\gamma \in \Gamma(\mu, \nu)$,*

$$\gamma\text{-esssup}_{x \in \mathbb{R}^d} \|\nabla \nabla_\mu F(\mu, x) - \nabla \nabla_\mu F(\nu, y)\|^2 \leq L_3^2 \int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y). \tag{14}$$

Remark 5.1. Similar gradient and Hessian Lipschitz continuity assumptions are made in the convergence analysis of perturbed gradient descent methods in Euclidean spaces (Ge et al., 2015; Jin et al., 2017; Li, 2019). Additionally, the first-order Lipschitz assumption is similar to works on convex functionals (Chizat, 2022; Suzuki et al., 2023).

5.1 Convergence Results for Continuous-time PWGF

The following is the main theorem of this paper, asserting that PWGF terminates in polynomial time and reaches an (ε, δ)-stationary point with high probability.

Theorem 5.2. *Let $\varepsilon,\ \delta,\ \zeta > 0$ be chosen such that $(L_2 + L_3)\,\varepsilon \leq \delta^2$ and $\varepsilon, \delta \leq \tilde{O}(1)$.[2] Let the initial point be $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\Delta F = F(\mu_0) - \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$. For hyperparameters $\eta_p = \tilde{O}\left(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\right)$, $T_{\mathrm{thres}} = \tilde{O}\left(\frac{1}{\delta}\right)$, and $F_{\mathrm{thres}} = \tilde{O}(\delta^3)$, PWGF halts after*

$$t = \tilde{O}\left(\Delta F \left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4}\right)\right)$$

time steps and reaches an (ε, δ)-second-order stationary point with probability $1 - \zeta$.

To prove Theorem 5.2, we present several supporting results. Detailed proofs are deferred to Appendix E. Lemma 5.3 is a fundamental property of WGF, providing a lower bound on the decrease in the objective for non-stationary points.

[2] In the analysis of PGD in Euclidean spaces, $\rho\varepsilon \leq \delta^2$ is often assumed, where ρ is the Lipschitz constant of the Hessian (Jin et al., 2017; Li, 2019). We adopt this assumption to the probabilistic measure space setting, where the Hessian Lipschitz constant becomes $\rho = L_2 + L_3$.

Lemma 5.3. *For a curve of probability measures μ_t following the WGF, the following holds:*

$$F(\mu_0) - F(\mu_t) = \int_0^t \|\nabla_\mu F(\mu_\tau)\|_{\mu_\tau}^2 \mathrm{d}\tau.$$

Proposition 5.4 is the crucial step of our analysis, showing that perturbing an (ε, δ)-saddle point using a Gaussian process with kernel (11) enables WGF to move along the unstable direction and decrease the objective function. In other words, it allows us to **efficiently escape the saddle point**.

Proposition 5.4. *Set $\eta = O(1)$ and let $\varepsilon,\ \delta,\ \eta_p,\ T_{\mathrm{thres}},\ F_{\mathrm{thres}}$ be chosen as in Theorem 5.2. Suppose $\mu^\dagger \in \mathcal{P}_2^a(\mathbb{R}^d)$ satisfies $\left\|\nabla_\mu F(\mu^\dagger)\right\|_{L^2(\mu^\dagger)} < \varepsilon$ and $\lambda_0 := \lambda_{\min} H_{\mu^\dagger} \leq -\delta$. Generating $\xi \sim \mathrm{GP}(0, k_\mu)$ and setting $\mu_0 = (\mathrm{Id} + \eta_p \xi)\sharp\mu^\dagger$ as the initial point of the WGF, we have with probability $1 - \zeta'$:*

$$F(\mu^\dagger) - F(\mu_{T_{\mathrm{thres}}}) \geq F_{\mathrm{thres}}.$$

Combining these results establishes convergence.

Proof of Theorem 5.2. Let $\varepsilon,\ \delta,\ \zeta > 0$ be chosen arbitrarily chosen such that $(L_2 + L_3)\,\varepsilon \leq \delta^2$, and set $\zeta' > 0$ such that ζ' is polynomial in $\frac{1}{\delta}$ and ζ up to logarithmic factors.

From Proposition 5.4, perturbations occur at most $m := \lceil \frac{\Delta F}{F_{\mathrm{thres}}} \rceil$ times. Thus, the probability of failure after m perturbations is at most $1 - (1-\zeta')^m \leq m\zeta'$. Setting $\zeta' = \frac{\zeta}{m}$ ensures that the algorithm reaches an (ε, δ)-second order stationary point with probability at least $1 - \zeta$.

PWGF consists of a gradient descent phase and an evaluation phase, where the decrease in the objective is assessed after applying a perturbation. Then we define the gradient descent phase as *State 0*, and the evaluation phase as *State 1*. Let T_0 denote the total time in *State 0*, where $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} \geq \varepsilon$. By Lemma 5.3, the decrease in the objective during *State 0* is at least $\varepsilon^2 T_0$, implying $T_0 \leq \frac{\Delta F}{\varepsilon^2}$. Moreover, the total time T_1 in *State 1* is bounded as $T_1 \leq m T_{\mathrm{thres}} = \frac{\Delta F T_{\mathrm{thres}}}{F_{\mathrm{thres}}} = \tilde{O}\left(\frac{\Delta F}{\delta^4}\right)$. Hence, PWGF halts in $T_0 + T_1 = \tilde{O}\left(\Delta F\left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4}\right)\right)$. □

5.2 Convergence Analysis for Discrete-time PWGF

We also prove convergence to a second-order stationary point for the time-discretized PWGF (Algorithm 2).

Theorem 5.5. *Let the initial point be $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and denote $\Delta F = F(\mu_0) - \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$. Set $\eta = O(1)$ and let $\varepsilon,\ \delta,\ \eta_p,\ T_{\mathrm{thres}},\ F_{\mathrm{thres}}$ be chosen as in Theorem 5.2. Then, discrete time PWGF halts after*

$$k = \tilde{O}\left(\Delta F\left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4}\right)\right)$$

steps and reaches an (ε, δ)-second-order stationary point with probability $1 - \zeta$.

The proof is similar to the continuous-time case; details are deferred to Appendix E.2. Since second-order stationary points are global optima for a strictly benign objective, the convergence of PWGF to a global solution is also guaranteed.

Corollary 5.6. *Under the same setting as Theorem 5.5, discrete-time PWGF for $(\varepsilon, \delta, \alpha)$-strictly benign objective F halts after $\tilde{O}\left(\Delta F\left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4}\right)\right)$ steps and reaches α-close to some global optima μ^o ; $W_2(\mu, \mu^o) \leq \alpha$ with probability $1 - \zeta$.*

6 Conclusion

We proposed a new method for non-convex probability optimization, perturbed Wasserstein gradient flows (PWGF), which alternates between perturbing near saddle points using a Hessian-guided Gaussian process and evolving via WGF. We have established that PWGF efficiently achieves second-order optimality with high probability. A potential avenue for future work is to reduce computational cost by using a stochastic approximation of the Hessian as the kernel of the Gaussian process, analogous to stochastic gradients. Another direction is to provide a method for analyzing benignity of non-convex distributional objectives, thereby broadening the range of applications of PWGF.

Acknowledgments

NY and JK were partially supported by JST CREST (JPMJCR2015). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2115). This research is supported by the National Research Foundation, Singapore, Infocomm Media Development Authority under its Trust Tech Funding Initiative, and the Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Infocomm Media Development Authority, and the Ministry of Digital Development and Information.

References

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

Arbel, M., Korba, A., SALIM, A., and Gretton, A. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, 2019.

Bonnet, B. A Pontryagin Maximum Principle in Wasserstein spaces for constrained optimal control problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:52, 2019.

Boufadene, S. and Vialard, F.-X. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *arXiv preprint arXiv:2312.00800*, 2024.

Chen, Z., Fan, J., and Wang, K. Multivariate Gaussian processes: definitions, examples and applications. *Metron*, 81(2):181–191, 2023.

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, 2020.

Chizat, L. Mean-field Langevin Dynamics: Exponential Convergence and Annealing. *arXiv preprint arXiv:2202.01009*, 2022.

Chizat, L. and Bach, F. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Chu, C., Blanchet, J. H., and Glynn, P. W. Probability functional descent: A unifying perspective on GANs, variational inference, and reinforcement learning. In *International Conference on Machine Learning*, 2019.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. A Stein variational Newton method. *Advances in Neural Information Processing Systems*, 31, 2018.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient Descent Can Take Exponential Time to Escape Saddle Points. *Advances in Neural Information Processing Systems*, 30, 2017.

Duncan, A., Nüsken, N., and Szpruch, L. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping From Saddle Points—Online Stochastic Gradient for Tensor Decomposition. In *Conference on Learning Theory*, pp. 797–842. PMLR, 2015.

Ge, R., Jin, C., and Zheng, Y. No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis. In *International conference on machine learning*, pp. 1233–1242. PMLR, 2017.

Guo, W., Hur, Y., Liang, T., and Ryan, C. Online Learning to Transport via the Minimal Selection Principle. In *Conference on Learning Theory*, pp. 4085–4109. PMLR, 2022.

Han, J. Wasserstein gradient descent for online learning. *Available at SSRN 4946060*, 2024.

He, Y., Balasubramanian, K., Sriperumbudur, B. K., and Lu, J. Regularized Stein Variational Gradient Flow. *Foundations of Computational Mathematics*, pp. 1–59, 2024.

Hutchinson, M., Terenin, A., Borovitskiy, V., Takao, S., Teh, Y., and Deisenroth, M. Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Independent Projected Kernels. *Advances in Neural Information Processing Systems*, 34:17160–17169, 2021.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to Escape Saddle Points Efficiently. In *International Conference on Machine Learning*, 2017.

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.

Kent, C., Li, J., Blanchet, J., and Glynn, P. W. Modified Frank Wolfe in Probability Space. *Advances in Neural Information Processing Systems*, 34:14448–14462, 2021.

Kim, J. and Suzuki, T. Transformers Learn Nonlinear Features In Context. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Lanzetti, N., Bolognani, S., and Dörfler, F. First-Order Conditions for Optimization in the Wasserstein Space. *SIAM Journal on Mathematics of Data Science*, 7(1):274–300, 2025.

Li, Z. SSRGD: Simple Stochastic Recursive Gradient Descent for Escaping Saddle Points. *Advances in Neural Information Processing Systems*, 32, 2019.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., Zhu, J., and Carin, L. Accelerated First-order Methods on the Wasserstein Space for Bayesian Inference. *stat*, 1050:4, 2018.

Liu, Q. and Wang, D. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Nitanda, A., Wu, D., and Suzuki, T. Convex Analysis of the Mean Field Langevin Dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 9741–9757. PMLR, 2022.

Richemond, P. H. and Maginnis, B. On Wasserstein Reinforcement Learning and the Fokker-Planck Equation. *arXiv preprint arXiv:1712.07185*, 2017.

Santambrogio, F. Optimal Transport for Applied Mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Suzuki, T., Wu, D., and Nitanda, A. Mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. *Advances in Neural Information Processing Systems*, 2023.

Taghvaei, A. and Mehta, P. Accelerated Flow for Probability Distributions. In *International Conference on Machine Learning*, 2019.

Wang, Y. and Li, W. Information Newton's flow: second-order optimization method in probability space. *arXiv preprint arXiv:2001.04341*, 2020.

Wang, Y. and Li, W. Accelerated Information Gradient Flow. *Journal of Scientific Computing*, 90: 1–47, 2022.

Zhang, R., Chen, C., Li, C., and Carin, L. Policy Optimization as Wasserstein Gradient Flows. In *International Conference on Machine Learning*, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The proposed PWGF algorithm is detailed in Section 4. The convergence analysis is given in in Section 5. Proof details are provided throughout the appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the conclusion, as well as Appendix B.2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Assumptions 1,2,3. All complete proofs are provided throughout the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental details are provided in Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The conducted experiments are toy simulations.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conducted 5 experiments under the same conditions, plotting the mean and standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experimental details are provided in Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have checked that the research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is primarily theoretical and no immediate societal impact is expected.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper is primarily theoretical.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Applications of Wasserstein Gradient Flow

A.1 Mean-Field Analysis

Mean-field analysis can be applied to optimization problems where the objective function is expressed as a function of the ensemble average of some underlying functions. We consider the optimization problem :

$$\text{minimize} \quad G\left(\frac{1}{N}\sum_{j=1}^{N} h(x_j)\right) \quad \text{s.t.} \quad x_1, \cdots, x_N \in \mathbb{R}^d, \tag{15}$$

where $h : \mathbb{R}^d \to \mathbb{R}^{d'}$, $G : \mathbb{R}^{d'} \to \mathbb{R}$, and $d' = 1$ for simplicity. The gradient direction of the variable x_j is computed as follows :

$$\nabla_{x_j} G\left(\frac{1}{N}\sum_{j=1}^{N} h(x_j)\right) = \frac{1}{N} G'\left(\frac{1}{N}\sum_{j=1}^{N} h(x_j)\right) \nabla h(x_j). \tag{16}$$

Taking the mean field limit $N \to \infty$ of (15), we lift this problem to the space of measures:

$$\text{minimize} \quad G\left(\int h(x)\mu(\mathrm{d}x)\right) \quad \text{s.t.} \quad \mu \in \mathcal{P}_2(\mathbb{R}^d). \tag{17}$$

The Wasserstein gradient is computed as follows :

$$\nabla_\mu G(\mu) = \nabla \frac{\delta F}{\delta \mu}(\mu, x) = G'\left(\int h(x)\mu(\mathrm{d}x)\right) \nabla h(x). \tag{18}$$

Consequently, from (16) and (18), the update rules for gradient descent applied to the original problem (15) and WGF applied to the lifted problem (17) are equivalent, up to a constant scaling factor. However, properties differ significantly between the two formulations. For example, when F is the identity function, the original problem (15) is not necessarily linear, while the mean-field problem (17) is linear with respect to μ. Similarly, for commonly encountered loss functions with convex losses, (15) is not necessarily convex with respect to the variables $x_1, \ldots, x_N$, but in the mean-field setting (17) becomes convex with respect to μ.

The properties of functions defined on the space of probability measures facilitate the design of WGF-based algorithms with improved performance. The works of Nitanda et al. (2022) and Chizat (2022) have explored the ability of optimization of mean-field Langevin dynamics (MFLD), an approach that augments WGF with Brownian noise, based on the convexity of the objective function. Similarly, Kim & Suzuki (2024) have focused on the benignity of objective functions regarding in-context learning of certain transformer models and, leveraging this insight, have proposed a birth-death modification and also an isotropic perturbation scheme. Our proposed algorithm contributes to this line of research by providing the first convergence guarantees for strictly benign problems on the space of probability measures.

A.2 Additional Related Works

WGF has important applications in Bayesian inference, where posterior distributions are usually estimated from data via variational inference (VI). VI formulates posterior estimation as an optimization problem of the KL divergence. In particular, particle-based VI is founded on the idea of evolving empirical measures formed by particles using WGF. This idea originates from the work of Jordan et al. (1998), which established a connection between diffusion processes and gradient descent in 2-Wasserstein space with entropy regularization.

Stein variational gradient descent (Liu & Wang, 2016; He et al., 2024) circumvents computational difficulties in calculating descent directions through kernel methods. This approach can be seen as a version of WGF where an integral operator acts on the descent direction (Chewi et al., 2020; Duncan et al., 2023). In this context, derivative methods such as Newton's method on the space of probability measures (Detommaso et al., 2018; Wang & Li, 2020) and accelerated methods (Liu et al., 2018; Taghvaei & Mehta, 2019; Wang & Li, 2022) have also been proposed. Furthermore, optimization of measures also appears in contexts such as online optimization (Guo et al., 2022; Han, 2024) and reinforcement learning (Richemond & Maginnis, 2017; Zhang et al., 2018).

B Wasserstein Geometry

B.1 Wasserstein Space

In this section, we provide basic aspects of Wasserstein geometry and propositions used in this paper. We refer to Ambrosio et al. (2008) for a comprehensive review.

We assume that all curves $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ that appear in this section are absolutely continuous and satisfy the continuity equation with respect to a vector field $v_t \in \mathrm{Tan}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$. The existence condition for solutions to the continuity equation corresponding to a vector field v_t is provided, for example, by Ambrosio et al. (2008), Section 8.2.

Definition B.1 (Wasserstein distance)**.** The 2-Wasserstein distance between two points $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$W_2(\mu,\nu)^2 \coloneqq \inf_{\gamma\in\Gamma(\mu,\nu)} \int \|x-y\|^2 \gamma(\mathrm{d}x\mathrm{d}y). \tag{19}$$

Here, $\Gamma(\mu,\nu)$ represents the set of all transport plans:

$$\Gamma(\mu,\nu) = \left\{\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, p_1\#\gamma = \mu,\ p_2\#\gamma = \nu\right\}.$$

Moreover, $\gamma \in \Gamma(\mu,\nu)$ is called the optimal transport plan, if the infmum in (19) is attained by γ. The set of all optimal transport plans is denoted by $\Gamma_o(\mu,\nu)$. Furthermore, if a measurable map $f : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $(\mathrm{Id}\times f)\sharp\mu \in \Gamma_o(\mu,\nu)$, then f is called the optimal transport map between μ and ν. In this case, we denote the optimal transport map f as $\mathcal{T}_\mu^\nu$.

$W_2 : \mathcal{P}_2(\mathbb{R}^d)\times\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defines a metric structure on $\mathcal{P}_2(\mathbb{R}^d)$. Therefore, we consider $\mathcal{P}_2(\mathbb{R}^d)$ as a metric space equipped with the W_2 metric. Convergence in the W_2 metric is equivalent to the weak convergence of measures plus uniform integrability of second moments (Ambrosio et al., 2008).

An optimal transport plan is guaranteed to exist for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, as stated in the following proposition.

Proposition B.2. *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists an optimal transport plan between μ and ν. That is, there exists $\gamma \in \Gamma(\mu,\nu)$ such that*

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

Proof. Refer to Ambrosio et al. (2008), Chapter 6. □

An optimal transport map, not a plan, does not necessarily exist. Specifically, there are cases where even transport maps themselves do not exist. For example, when $d = 1$, $\mu = \delta_0$, and $\nu = \frac{\delta_{-1}+\delta_1}{2}$, no transport map exists, as $T\#\mu \neq \nu$ for any $T : \mathbb{R} \to \mathbb{R}$. However, The following proposition establishes that under certain conditions, the existence and uniqueness of an optimal transport map are guaranteed.

Proposition B.3 (Brenier's theorem)**.** *For any $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists an optimal transport map $\mathcal{T}_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$. Specifically, $\mathcal{T}_\mu^\nu$ satisfies*

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

Furthermore, the following hold:

- *$\Gamma_o(\mu,\nu) = \{(\mathrm{Id}\times\mathcal{T}_\mu^\nu)\sharp\mu\}$, that is, the unique optimal transport plan from μ to ν is induced by $\mathcal{T}_\mu^\nu$.*
- *The optimal transport map can be expressed as $\mathcal{T}_\mu^\nu = \nabla\varphi$, where φ is a convex function defined μ-almost everywhere.*
- *If $\nu \in \mathcal{P}_2^a(\mathbb{R}^d)$ as well, then $\mathcal{T}_\nu^\mu \circ \mathcal{T}_\mu^\nu = \mathrm{Id}$ μ-a.e. and $\mathcal{T}_\mu^\nu \circ \mathcal{T}_\nu^\mu = \mathrm{Id}$ ν-a.e.*

Proof. See Ambrosio et al. (2008), Chapter 6. □

$\mathcal{P}_2(\mathbb{R}^d)$ forms a geodesic metric space due to the existence of optimal transport plans and the pushforward property:

Proposition B.4. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and $\gamma \in \Gamma_0(\mu, \nu)$. Then $\mu_t = ((1-t)p_1 + tp_2)\sharp\gamma$ defines a geodesic between μ and ν, i.e.,*

$$W_2(\mu, \mu_t) = tW_2(\mu, \nu) \quad (\forall t \in [0, 1]).$$

In particular, in case where γ is induced by an optimal transport map $\mathcal{T}_\mu^\nu$, $\mu_t = ((1-t)\mathrm{Id} + t\mathcal{T}_\mu^\nu)\#\mu$ and $\mathcal{T}_\mu^{\mu_t} = (1-t)\mathrm{Id} + t\mathcal{T}_\mu^\nu$ hold.

Proof. See Ambrosio et al. (2008), Lemma 7.2.1. □

The following continuity equation describes how a particle distribution μ_t evolves along a time-dependent vector field v_t.

Definition B.5 (Continuity equation, Ambrosio et al. (2008))**.** Let $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ $(t \in I)$ be a curve in Wasserstein space, and let $v_t \in L(\mu_t)^d$ $(t \in I)$ be the corresponding vector field. The curve μ_t is said to satisfy the continuity equation with respect to the vector field v_t if the distribution equation:

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0 \tag{20}$$

holds. Equation (20) means that for all $\varphi \in C_0^\infty(\mathbb{R}^d)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(x) \mu_t(\mathrm{d}x) = \int \nabla \varphi(x)^\top v_t(x) \mu_t(\mathrm{d}x).$$

The absolute continuity of the curve μ_t with respect to the Wasserstein distance is equivalent to the satisfaction of the continuity equation for some $v_t \in L^2(\mu_t)$ (Ambrosio et al. (2008) Theorem 8.3.1). In this sense, the continuity equation provides a concept of differentiation consistent with the distance structure induced by the Wasserstein distance.

Proposition B.6 (Ambrosio et al. (2008), Theorem 8.3.1)**.** *Let $I \subset \mathbb{R}$ be an open interval, and let $\mu_t : I \to \mathcal{P}_2(\mathbb{R}^d)$ be a continuous curve. μ_t is absolutely continuous, if and if only μ_t satisfies continuity equation (3) for some vector field $v_t \in L^2(\mu_t)$.*

For the vector field that gives the continuity equation (3),

$$\nabla \cdot ((v_t - w_t)\mu_t) = 0 \iff v_t - w_t \in \left\{\nabla \varphi \;\middle|\; \varphi \in C_0^\infty(\mathbb{R}^d)\right\}^{\perp L^2(\mu_t)} =: X_{\mu_t}$$

is equivalent to the fact that the continuity equation gives the same curve μ_t. Noting the orthogonal decomposition $L^2(\mu_t)^d = X_{\mu_t} \oplus X_{\mu_t}^\perp$, the vector field with the minimal L^2 norm among those that give the same curve μ_t must have no component in the subspace X_{μ_t}, and it follows that $v_t \in X_{\mu_t}^\perp$. From these observations, we define the tangent space representing infinitesimal changes in the space of probability measures $\mathcal{P}_2(\mathbb{R}^d)$ as follows :

Definition B.7 (Tangent bundle, Ambrosio et al. (2008))**.** We define the tangent space $\mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu)$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d) \coloneqq X_\mu^\perp = \overline{\{\nabla \varphi \mid \varphi \in C_0^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}.$$

The following proposition, referred to as the Benamou-Brenier formula, demonstrates that the Wasserstein distance is characterized by the minimal action among absolutely continuous curves connecting two given probability measures.

Proposition B.8 (Benamou-Brenier formula)**.** *For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $T > 0$, the following holds:*

$$W_2(\mu, \nu)^2 = \inf \left\{ T \int_0^T \|v_t\|_{L^2(\mu_t)}^2 \,\mathrm{d}t \;\middle|\; \partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0 \; (t \in (0, T)),\; \mu_0 = \mu, \mu_T = \nu \right\}.$$

Proof. From Ambrosio et al. (2008), Theorem 8.3.1,

$$W_2(\mu, \nu)^2 = \inf \left\{ \int_0^1 \|v_t\|_{L^2(\mu_t)}^2 \,\mathrm{d}t \;\middle|\; \partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0 \; (t \in (0, 1)),\; \mu_0 = \mu, \mu_1 = \nu \right\}.$$

By changing the variable $t \mapsto \frac{t}{T}$, the claim follows. □

The following proposition establishes that the infinitesimal behavior of an absolutely continuous curve can be expressed by the pushforward $(\mathrm{Id} + hv_t)\#\mu_t$.

Proposition B.9. *Let μ_t be an absolutely continuous curve satisfying continuous equation with vector field $v_t \in \mathrm{Tan}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$\lim_{h\to 0} \frac{W_2(\mu_{t+h}, (\mathrm{Id} + hv_t)\#\mu_t)}{h} = 0.$$

Proof. See Ambrosio et al. (2008), Proposition 8.4.6. □

The following propositions provide sufficient conditions for a map to be an optimal transport map.

Proposition B.10 (Santambrogio (2015), Theorem 1.48)**.** *Suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and that $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a convex and μ-a.e. differentiable function with $\nabla\varphi \in L^2(\mu)^d$. Then, the map $\nabla\varphi$ provides the optimal transport map from μ to $(\nabla\varphi)\sharp\mu$.*

Proof. Let φ^* be a Legendre-Fenchel transformation of φ. Then, it holds that

$$\begin{aligned}\varphi(x) + \varphi^*(y) &\geq \langle x, y\rangle \quad \forall x, y \in \mathbb{R},\\ \varphi(x) + \varphi^*(y) &= \langle x, y\rangle \quad (\mathrm{Id}\times\nabla\varphi)\#\mu\text{-a.e. } (x, y).\end{aligned}$$

For any transport plan $\gamma \in \Gamma(\mu, \nabla\varphi\#\mu)$, it holds that

$$\begin{aligned}-2\int \langle x, y\rangle\, \gamma(\mathrm{d}x\mathrm{d}y) &\geq -2\int (\varphi(x) + \varphi^*(y))\gamma(\mathrm{d}x\mathrm{d}y)\\ &= -2\int \varphi(x)\mu(\mathrm{d}x) - \int \varphi^*(\nabla\varphi(x))\mu(\mathrm{d}x)\\ &= -2\int \langle x, y\rangle\, (\mathrm{Id}\times\nabla\varphi)\#\mu.\end{aligned}$$

By adding $\int(\|x\|^2+\|y\|^2)\gamma(\mathrm{d}x\mathrm{d}y) = \int(\|x\|^2+\|y\|^2)(\mathrm{Id}\times\nabla\varphi)\#\mu(\mathrm{d}x)$ both sides of the inequality, we obtain

$$\int \|x - y\|^2\gamma(\mathrm{d}x\mathrm{d}y) \geq \int \|x - y\|^2(\mathrm{Id}\times\nabla\varphi)\#\mu(\mathrm{d}x).$$

Then $\nabla\varphi$ is a optimal transport map. □

Proposition B.11 (Lanzetti et al. (2025), Lemma 2.4)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varphi \in C_c^\infty(\mathbb{R}^d)$. Then, there exists $\bar{h} > 0$ such that $\mathrm{Id} + h\nabla\psi$ is an optimal transport map from μ to $(\mathrm{Id} + h\nabla\psi)\#\mu$ for $h \in [-\bar{h}, \bar{h}]$. Furthermore,*

$$W_2(\mu, (\mathrm{Id} + h\nabla\psi)\#\mu) = h\|\nabla\psi\|_{L^2(\mu)}.$$

Proof. From $\psi \in C_c^\infty(\mathbb{R}^d)$, we have $\sup_{x\in\mathbb{R}^d}\|\nabla\psi\|(x) < \infty$, $\sup_{x\in\mathbb{R}^d}\left\|\nabla^2\psi(x)\right\| < \infty$. Then, by taking $\bar{h} = \sup_{x\in\mathbb{R}^d}\left\|\nabla^2\psi(x)\right\|$, we verify that $\frac{1}{2}\|x\|^2 + h\psi(x)$ is convex and $\nabla\left(\frac{1}{2}\|x\|^2 + h\psi(x)\right) = (\mathrm{Id} + h\nabla\psi)(x)$ is $L^2(\mu)$ integrable for all $h \in [-\bar{h}.\bar{h}]$. Noting that $\nabla\left(\frac{1}{2}\|x\|^2 + h\psi(x)\right) = (\mathrm{Id} + h\nabla\psi)(x)$ and applying Proposition B.10, we obtain that $\mathrm{Id} + h\nabla\psi$ yields an optimal transport map. Consequently, the Wasserstein distance between μ and $(\mathrm{Id} + h\nabla\psi)\#\mu$ is computed as:

$$W_2(\mu, (\mathrm{Id} + h\nabla\psi)\#\mu) = \left(\int \|x + h\nabla\psi(x) - x\|^2\mu(\mathrm{d}x)\right)^{\frac{1}{2}} = h\|\nabla\psi\|_{L^2(\mu)}.$$

□

Corollary B.12. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varphi \in C^2(\mathbb{R}^d)$ satisfy $\|\nabla\varphi\|_{L^2(\mu)} < \infty$, $\mu-\mathrm{esssup}_{x\in\mathbb{R}^d}\left\|\nabla^2\varphi(x)\right\| < \infty$. Then, there exists $\bar{h} > 0$ such that $\mathrm{Id}+h\nabla\varphi$ is an optimal transport map from μ to $(\mathrm{Id}+h\nabla\varphi)\#\mu$ for $h \in [-\bar{h}, \bar{h}]$. Furthermore, $W_2(\mu, (\mathrm{Id}+h\nabla\varphi)\#\mu) = h\|\nabla\varphi\|_{L^2(\mu)}$ holds.*

B.2 Wasserstein Gradient

This section provides additional discussion on the relationship between the Wasserstein gradient and the first variation.

Definition B.13 (Wasserstein gradient, Lanzetti et al. (2025); Bonnet (2019))**.** The Wasserstein gradient $\nabla_\mu F : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as a vector-field such that for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Gamma_0(\mu, \nu)$, it holds that

$$F(\nu) - F(\mu) = \int \nabla_\mu F(\mu, x)^\top (y - x) \gamma(\mathrm{d}x\mathrm{d}y) + O\big(W_2(\mu, \nu)^2\big). \tag{21}$$

In particular, there exists a unique element that satisfies $\nabla_\mu F(\mu) \in \mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d)$ (Lanzetti et al. (2025) Proposition 2.5), and we take this as the Wasserstein gradient.

The first variation $\frac{\delta F}{\delta \mu}$ appearing in the WGF equation (1) is defined as follows.

Definition B.14 (First variation)**.** The first variation $\frac{\delta F}{\delta \mu} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ is defined as a functional satisfying for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\frac{\mathrm{d}}{\mathrm{d}h}\bigg|_{h=0} F(\mu + h(\nu - \mu)) = \int \frac{\delta F}{\delta \mu}(\mu, x)(\mu - \nu)(\mathrm{d}x). \tag{22}$$

The first variation, if it exists, is unique up to a constant difference.

Recall the definition of the total variation:

Definition B.15 (Total variation)**.** Suppose that $\mu,\ \nu \in \mathcal{P}_2(\mathbb{R}^d)$. The total variation between μ and ν is:

$$\mathrm{TV}(\mu, \nu) \coloneqq \sup_{B \in \mathcal{B}(\mathbb{R}^d)} |\mu(B) - \nu(B)| = \frac{1}{2} \sup_{\mathcal{C} \subset \mathcal{B}(\mathbb{R}^d), \cup \mathcal{C} = \mathbb{R}^d} \sum_{B \in \mathcal{C}} |\mu(B) - \nu(B)|.$$

As is evident from the definition, the mixture $(1 - h)\mu + h\nu$, defines the constant-speed geodesic between μ, ν in the sense of total variation. Thus, the first variation can be interpreted as the coefficient of differentiation along the geodesic with respect to the total variation distance. On the other hand, as observed in the previous section, the geodesic of W_2 is represented as $((1 - h)p_1 + hp_2)\#\gamma$ for $\gamma \in \Gamma_0(\mu, \nu)$. Therefore, the Wasserstein gradient can be understood as the coefficient of differentiation along the geodesic with respect to the Wasserstein distance.[3]

Through a naive but mathematically non-rigorous calculation, we have

$$\begin{aligned}
F(\nu) - F(\mu) &\approx \int \frac{\delta F}{\delta \mu}(\mu, x)(\nu - \mu)(\mathrm{d}x) \\
&= \int \left(\frac{\delta F}{\delta \mu}(\mu, y) - \frac{\delta F}{\delta \mu}(\mu, x) \right) \gamma(\mathrm{d}x\mathrm{d}y) \\
&\approx \int \left(\left\langle \nabla \frac{\delta F}{\delta \mu}(\mu, x), y - x \right\rangle + \frac{1}{2} \left\langle y - x, \nabla^2 \frac{\delta F}{\delta \mu}(\mu, x)(y - x) \right\rangle \right) \gamma(\mathrm{d}x\mathrm{d}y) \\
&= \int \left\langle \nabla \frac{\delta F}{\delta \mu}(\mu, x), y - x \right\rangle \gamma(\mathrm{d}x\mathrm{d}y) + O\big(W_2(\mu, \nu)^2\big).
\end{aligned}$$

This points to $\nabla_\mu F = \nabla \frac{\delta F}{\delta \mu}(\mu)$. This equation does not hold without certain conditions, but it is valid and implicitly assumed to hold in many practical cases. Kent et al. (2021) make a brief mention of this frustration. The equivalence $\nabla_\mu \cdot = \nabla \frac{\delta \cdot}{\delta \mu}$ specifically provides the following correspondence.

$$\begin{aligned}
\nabla_\mu F(\mu, x) &= \nabla \frac{\delta F}{\delta \mu}(\mu, x), \\
\nabla_\mu^2 F(\mu, x, y) &= \nabla_x \nabla_y \frac{\delta^2 F}{\delta \mu^2}(\mu, x, y), \\
\nabla \nabla_\mu F(\mu, x) &= \nabla^2 \frac{\delta F}{\delta \mu}(\mu, x).
\end{aligned}$$

[3] There is no strict dominance or subordination between total variation and Wasserstein distance in $\mathbb{R}^d$.

C Optimality Conditions for Functionals on Probability Space

In this section, we discuss the details of the optimality conditions for functionals on probability measures based on the Wasserstein gradient. The first-order optimality conditions have been studied in Bonnet (2019) and Lanzetti et al. (2025), and we begin by reviewing these works. We then extend these results to second-order conditions.

C.1 First-order Condition

The first order condition for probability measure optimization is studied by Lanzetti et al. (2025), including constrained optimization problem. Here, we review the results of the first-order condition for unconstrained problem. In the next section, we extend these results and obtain the second-order condition.

The following proposition shows that Equation (21) holds even when γ is not optimal.

Proposition C.1. *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional on probability space, and differentiable at μ. Then, for any transport plan $\gamma \in \Gamma(\mu, \nu)$ (not necessarily an optimal one),*

$$F(\nu) - F(\mu) = \int \nabla_\mu F(\mu, x)^\top (y - x)\gamma(\mathrm{d}x\mathrm{d}y) + O\left(\int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y)\right).$$

Proof. See Lanzetti et al. (2025). □

The following two propositions are useful for computing changes in the objective function. The next proposition provides a first-order Taylor expansion for infinitesimal perturbations induced by a vector field.

Proposition C.2. *For a sufficiently smooth functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and a vector field $v \in L^2(\mu)^d$ ($\mu \in \mathcal{P}_2(\mathbb{R}^d)$), it holds that*

$$F((\mathrm{Id} + hv)\#\mu) - F(\mu) = h \left\langle \nabla_\mu F(\mu), v \right\rangle_{L^2(\mu)} + O(h^2).$$

Proof. Consider the situation of Proposition C.1 where $\nu = (\mathrm{Id} + hv)\#\mu$, $\gamma = (\mathrm{Id} \times (\mathrm{Id} + hv))\#\mu$. Then,

$$\begin{aligned}\int \nabla_\mu \left\langle F(\mu, x), y - x \right\rangle \gamma(\mathrm{d}x\mathrm{d}y) &= \int \left\langle \nabla_\mu F(\mu, x), (x + hv(x) - x \right\rangle \mu(\mathrm{d}x) \\ &= h \left\langle \nabla_\mu F(\mu), v \right\rangle_{L^2(\mu)},\end{aligned}$$

and also

$$\int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y) = \int \|(x + hv(x)) - x\|^2 \mu(\mathrm{d}x) = h^2 \|v\|^2_{L^2(\mu)}$$

holds. So, by Proposition C.1, the claim follows. □

The following proposition provides a formula for differentiating the objective along an absolutely continuous curve. This corresponds to the chain rule for differentiable curves in Euclidean space.

Proposition C.3 (Chain rule)**.** *For a sufficiently smooth functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and absolutely continuous curve μ_t satisfying a continuous equation $\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$, $t \mapsto F(\mu_t)$ is differentiable, and the following holds:*

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = \left\langle \nabla_\mu F(\mu_t), v_t \right\rangle_{L^2(\mu_t)}.$$

Proof. By applying the definition of Wasserstein gradient (Definition B.13) for $\mu = \mu_{t+h}$, $\nu = (\mathrm{Id} + hv_t)\#\mu_t$ and $\gamma \in \Gamma_0(\mu_{t+h}, (\mathrm{Id} + hv_t)\#\mu_t)$, it holds that

$$\begin{aligned}\left| \frac{F(\mu_{t+h}) - F((\mathrm{Id} + hv_t)\#\mu_t)}{h} \right| &\le \left| \frac{1}{h} \int \left\langle \nabla_\mu F(\mu_t, x), y - x \right\rangle \gamma(\mathrm{d}x\mathrm{d}y) \right| + o(1) \\ &\le \|\nabla_\mu F(\mu_t)\|_{L^2(\mu_t)} \frac{W_2(\mu_{t+h}, (\mathrm{Id} + hv_t)\#\mu_t)}{h} + o(1) \\ &\to 0,\end{aligned}$$

as $h \to 0$, where Cauchy Schwarz inequality is used in second line and Proposition B.9 is used in third line. Then by Proposition C.2, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) &= \lim_{h\to 0}\frac{F(\mu_{t+h}) - F(\mu_t)}{h} \\ &= \lim_{h\to 0}\frac{F((\mathrm{Id} + hv_t)\#\mu_t) - F(\mu_t)}{h} \\ &= \langle \nabla_\mu F(\mu_t), v_t\rangle_{L^2(\mu_t)}.\end{aligned}$$

□

The above proposition shows that the Wasserstein gradient plays a crucial role for first-order perturbations. In particular, if $\nabla_\mu F(\mu) = 0$, the objective function does not change under any first-order perturbation. This suggests that it is reasonable to define first-order stationary points as points satisfying $\nabla_\mu F(\mu) = 0$. Furthermore, supporting this observation, Lanzetti et al. (2025) established the following proposition.

Proposition C.4 (First-order necessary condition)**.** *Let $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ be a local minimizer of a differentiable functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ i.e. it holds that there exists a constant $r > 0$ such that*

$$W_2(\mu, \mu^*) < r \implies F(\mu) \le F(\mu^*).$$

Then the Wasserstein gradient of F vanishes at μ^*:

$$\nabla_\mu F(\mu^*, x) = 0 \quad \mu^*\text{-a.e.}x,$$

i.e. $\nabla_\mu F(\mu^*) = 0$ in $L^2(\mu^*)^d$.

Proof. See Lanzetti et al. (2025), Theorem 3.1. □

Proposition C.5 (First-order sufficient condition)**.** *Suppose that F is differentiable and α-geodesically convex with $\alpha \ge 0$, i.e. it holds that*

$$F(\nu) - F(\mu) \ge \int \langle \nabla_\mu F(\mu, x), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y) + \frac{\alpha}{2}W_2(\mu,\nu)^2 \quad \forall \gamma \in \Gamma_0(\mu,\nu).$$

Then, $\nabla_\mu F(\mu^*) = 0$ μ-a.e. implies that μ^* is global minimizer of F, i.e. $F(\mu) \ge F(\mu^*)$ holds for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

Proof. See Lanzetti et al. (2025), Theorem 3.3. □

C.2 Second-order Condition

Building on the results from the previous section, we discuss second-order optimality in measure optimization. To obtain second-order terms, we first prove the following lemma. The key point is that we obtain the derivative $F(\mu_t)$ not only at $t = 0$, but also for $0 < t < 1$. By obtaining this, we can compute the second-order coefficient.[4]

Lemma C.6. *Let F be a sufficiently smooth functional, $\mu,\ \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and μ_h be a constant geodesic induced by $\gamma \in \Gamma(\mu,\nu)$. Then, for any $0 \le t < 1$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) = \int \langle \nabla_\mu F(\mu_t, (1-t)x + ty), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y).$$

Proof. By Proposition C.1, we obtain the following statement $(*)$: let $\mu,\ \nu \in \mathcal{P}_2(\mathbb{R}^d),\ \gamma \in \Gamma(\mu,\nu)$ and let $\mu_t := ((1-t)p_1 + tp_2)\#\gamma$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} F(\mu_t) = \int \langle \nabla_\mu F(\mu, x), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y).$$

[4] Similar to this proposition, higher-order terms (third-order and beyond) can be computed in the same manner. It is conjectured that higher-order terms can also be expressed as the action of ∇_μ or ∇ on F.

Since we suppose $t < 1$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) &= \lim_{s\to 0}\frac{F(\mu_{t+s}) - F(\mu_t)}{s}\\ &= \frac{1}{1-t}\lim_{s\to 0}\frac{F(\mu_{t+(1-t)s}) - F(\mu_t)}{s}\\ &= \frac{1}{1-t}\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} F(\mu_{t+(1-t)s}).\end{aligned}$$

Here, it holds that

$$\begin{aligned}\mu_{t+(1-t)s} &= ((1-(t+(1-t)s))p_1 + (t+(1-t)s)p_2)\#\gamma\\ &= ((1-s)((1-t)p_1 + tp_2) + sp_2)\#\gamma\\ &= ((1-s)p_1 + sp_2)\#(((1-t)p_1 + tp_2)\times p_2)\#\gamma.\end{aligned}$$

Then, by $(*)$ for $\mu \leftarrow \mu_t,\ \nu \leftarrow \nu, \gamma \leftarrow ((1-t)p_1 + tp_2 \times \mathrm{Id})\#\gamma$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) &= \frac{1}{1-t}\lim_{s\to 0}\frac{F(\mu_{t+(1-t)s}) - F(\mu_t)}{s}\\ &= \frac{1}{1-t}\int \langle \nabla_\mu F(\mu_t, x), y - x\rangle\, ((1-t)p_1 + tp_2 \times \mathrm{Id})\#\gamma(\mathrm{d}x\mathrm{d}y)\\ &= \frac{1}{1-t}\int \langle \nabla_\mu F(\mu_t, (1-t)x + ty), y - ((1-t)x + ty)\rangle\, \gamma(\mathrm{d}x\mathrm{d}y)\\ &= \int \langle \nabla_\mu F(\mu_t, (1-t)x + ty), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y).\end{aligned}$$

□

Proposition C.7. *Suppose $F : \mathcal{P}_2(\mathbb{R}^d)$ is sufficiently smooth. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and let μ_h be a constant geodesic induced by $\gamma \in \Gamma_0(\mu, \nu)$. Then, as $h \to 0$,*

$$\begin{aligned}F(\mu_h) - F(\mu) = h\int \langle \nabla_\mu F(\mu, x), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y)\\ + \frac{h^2}{2}\int \left\langle y_1 - x_1, \nabla_\mu^2 F(\mu, x_1, x_2)(y_2 - x_2)\right\rangle \gamma(\mathrm{d}x_1\mathrm{d}y_1)\gamma(\mathrm{d}x_2\mathrm{d}y_2)\\ + \frac{h^2}{2}\int \langle y - x, \nabla\nabla_\mu F(\mu, x)(y - x)\rangle\, \gamma(\mathrm{d}x\mathrm{d}y) + o(h^2).\end{aligned}$$

Proof. By Lemma C.6,

$$\begin{aligned}\left.\frac{\mathrm{d}^2}{\mathrm{d}h^2}\right|_{h=0} F(\mu_h) &= \left.\frac{\mathrm{d}}{\mathrm{d}h}\right|_{h=0}\int \langle \nabla_\mu F(\mu_h, (1-h)x + hy), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y)\\ &= \int \left\langle \left.\frac{\mathrm{d}}{\mathrm{d}h}\right|_{h=0} \nabla_\mu F(\mu_h, (1-h)x + hy), y - x\right\rangle \gamma(\mathrm{d}x\mathrm{d}y)\\ &= \int \left\langle \int \nabla_\mu^2 F(\mu, x_1, x_2)(y_1 - x_1)\gamma(\mathrm{d}x_1\mathrm{d}y_1), y_2 - x_2\right\rangle \gamma\mathrm{d}x_2\mathrm{d}y_2\\ &\qquad + \int \langle \nabla\nabla_\mu F(\mu, x)(y - x), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y)\\ &= \iint \left\langle y_1 - x_1, \nabla_\mu^2 F(\mu, x_1, x_2)(y_2 - x_2)\right\rangle \gamma^{\otimes 2}(\mathrm{d}x_1\mathrm{d}y_1\mathrm{d}x_2\mathrm{d}y_2)\\ &\qquad + \int \langle \nabla\nabla_\mu F(\mu, x)(y - x), y - x\rangle\, \gamma(\mathrm{d}x\mathrm{d}y).\end{aligned}$$

□

Remark C.8. Except for the term $\nabla\nabla_\mu F$, this expression can be interpreted in a manner similar to Taylor expansion in Euclidean space. The term $\nabla\nabla_\mu F$ arises from the change in the metric μ of

the tangent space. This phenomenon is similar to what occurs on Riemannian manifolds, where the metric of the tangent spaces is not necessarily constant. It is worth noting that a similar calculation is performed in Appendix 5 of Bonnet (2019). As stated in the main text, at first-order stationary points, we have $\nabla_\mu F = 0$, which implies $\nabla\nabla_\mu F = 0$ Therefore, second-order optimality at stationary points can be understood in terms of the integral operator property of $\nabla^2_\mu F$.

By setting $\nu = (\mathrm{Id} + hv)\#\mu$ in Proposition C.7, we obtain the following:

Proposition C.9. *Suppose $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is sufficiently smooth. Then, it holds that for all $v \in L^2(\mu)^d$,*

$$F((\mathrm{Id} + hv)\#\mu) - F(\mu) = h\left\langle \nabla_\mu F(\mu), v\right\rangle_{L^2(\mu)} + \frac{h^2}{2}\left\langle v, (H_\mu + H'_\mu)v\right\rangle_{L^2(\mu)} + o(h^2) \quad (h \to 0).$$

For the next lemma, we denote the interior and boundary of a set $A \subset \mathbb{R}^d$ as A^o and ∂A, respectively.

Lemma C.10. *Suppose that $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$ and that $f : \mathbb{R}^d \to \mathbb{R}$ is of class $C^1(\mathbb{R}^d)$. Then, it holds that*

$$f = 0 \quad \mu\text{-a.e.} \implies \nabla f = 0 \quad \mu\text{-a.e.}$$

Proof. Let $B \subset \mathbb{R}^d$ be a closed set $B = \left\{x \in \mathbb{R}^d \mid f(x) = 0\right\}$. For all $x \in B^o \subset B$, there exists $r > 0$ such that

$$\|y - x\| < r \implies y \in B^o \implies f(y) = 0.$$

Then we have $\nabla f(x) = 0$. Thus,

$$\begin{aligned}\mu\left(\left\{x \in \mathbb{R}^d | \nabla f(x) = 0\right\}\right) &\geq \mu(B^o)\\ &= \mu(B) - \mu(\partial B)\\ &= 1 - \mu(\partial B).\end{aligned}$$

Since μ is absolutely continuous with respect to Lebesgue measure and B is a Jordan measurable set, $\mu(\partial B) = 0$ holds. Hence, $\nabla f(x) = 0$ μ-a.e. □

Proposition C.11 (Second-order necessary condition, restatement of Proposition 3.3)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be sufficiently smooth. If $\mu^* \in \mathcal{P}_2^a(\mathbb{R}^d)$ be a local minimum of F, i.e. it holds that there exists a constant $r > 0$ such that*

$$W_2(\mu, \mu^*) < r \implies F(\mu) \leq F(\mu^*).$$

Then it holds that $H_{\mu^*} \succeq O$.

Proof. According to Proposition C.4, $\nabla_\mu F(\mu^*) = 0$ μ-a.e. Then it follows from Lemma C.10 that $\nabla\nabla_\mu F(\mu^*) = 0$ a.e.
For any vector field $v \in L^2(\mu^*)$, there exists a constant $\bar{h} > 0$ such that $h \leq \bar{h} \implies W_2(\mu^*, (\mathrm{Id} + hv)\#\mu^*) = h\|v\|_{L^2(\mu^*)} \leq r$. By applying Proposition C.9, we have

$$\langle v, H_{\mu^*} v\rangle_{L^2(\mu^*)} = F((\mathrm{Id} + hv)\#\mu^*) - F(\mu^*) + o(1) \geq o(1).$$

By letting $h \to 0$, we have $\langle v, H_{\mu^*} v\rangle_{L^2(\mu^*)} \geq 0$ i.e. $H_{\mu^*} \succeq O$. □

D Kernels and Gaussian Processes

The purpose of this section is to present a series of propositions regarding positive semi-definite kernels, the integral operators they define, and Gaussian processes. Additionally, we aim to derive an inequality that evaluates the tail probability of the $L^2(\mu)$ norm of a Gaussian process.

In the following propositions, the kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ of interest corresponds to the integral kernel K_μ of the squared Hessian defined in (11). However, we first examine the properties of general kernels and Gaussian processes, applying these results to K_μ defined in (11) at the end of this section.

In this paper, we consider multivariate Gaussian processes. Kim & Suzuki (2024) were the first to propose introducing random perturbations to probability measures using a multivariate Gaussian process. Another application of multivariate Gaussian processes is modeling vector fields on Riemannian manifolds (Hutchinson et al., 2021). For detailed definition and properties, please refer to Chen et al. (2023).

Definition D.1 (Multivariate Gaussian process)**.** The vector-valued function $\xi : \mathbb{R}^d \to \mathbb{R}^d$ is said to follow a multivariate Gaussian process if any finite collection of variables $\xi(x_1), \cdots, \xi(x_N)$ are jointly normally distributed. This process is determined by the vector-valued mean function $m : \mathbb{R}^d \to \mathbb{R}^d$ and the matrix-valued covariance function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$:

$$m(x) = \mathbb{E}[\xi(x)] \quad (x \in \mathbb{R}^d),$$
$$K(x, \tilde{x}) = \mathbb{E}\big[(\xi(x) - m(x))(\xi(x) - m(x))^\top\big] \quad (x, \tilde{x} \in \mathbb{R}^d).$$

In this case, we denote $\xi \sim \mathrm{GP}(m, K)$.

Proposition D.2. *Suppose $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ satisfies $\int \|K(x, y)\| \, \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) < \infty$ and $K(x, y)^\top = K(y, x)$ for all $x, y \in \mathbb{R}^d$. Then, there exists a sequence $\{\kappa_n\}_{n\geq 1} \subset \mathbb{R}$, which is finite or converges to 0, satisfies $\|T_K\|_{\mathrm{Tr},L^2(\mu)} = \sum_{n\geq 1} |\kappa_n| < \infty$, and is non-increasing. Furthermore, there exists an orthonormal basis $\{\psi_n\}_{n\geq 1} \subset L^2(\mathbb{R}^d)^d$ such that*

$$K(x, y) = \sum_{n\geq 1} \kappa_n \psi_n(x)\psi_n(y)^\top,$$

where the infinite sum converges in the $L^2(\mu)^{d\times d}$ norm.

Proof. This follows from the eigenvalue expansion theorem for compact self-adjoint operators on Hilbert spaces. The trace-class property $\sum_{n\geq 1} |\kappa_n| < \infty$ follows from T_k being trace-class. □

Proposition D.3. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\psi \in L^2(\mu)^d$. Suppose $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a positive semi-definite kernel satisfying $\int \|K(x, y)\| \, \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) < \infty$. Then, $\xi \sim \mathrm{GP}(0, K)$ satisfies $\xi \in L^2(\mu)^d$ almost surely.*

Proof. Using the integrability of k and equivalence of matrix norms:

$$\mathbb{E}[\|\xi\|_\mu^2] = \int \mathbb{E}[\xi(x)^\top \xi(x)] \, \mu(\mathrm{d}x) = \int \mathrm{tr}(K(x, x)) \, \mu(\mathrm{d}x) < \infty.$$

Setting $A_n = \{\|\xi\|_\mu^2 \geq n\}$ for $n \in \mathbb{N}$, it holds that

$$P(\|\xi\|_\mu^2 < \infty) = 1 - P\left(\textstyle\bigcap_{n\geq 1} A_n\right) = 1 - \lim_{n\to\infty} P(A_n) \geq 1 - \limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[\|\xi\|_\mu^2] = 1.$$

Thus, $\xi \in L^2(\mu)^d$ almost surely. □

Proposition D.4. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\psi \in L^2(\mu)^d$. Suppose $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a positive semi-definite kernel satisfying $\int \|K(x, y)\| \, \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) < \infty$. Then, for any $\xi \sim \mathrm{GP}(0, K)$, we have $\langle \psi, \xi \rangle_{L^2(\mu)} \sim \mathcal{N}(0, \langle \psi, T_K \psi \rangle_{L^2(\mu)})$.*

Proof. The proof follows Kim & Suzuki (2024), Lemma E.9. First, we show that $\langle \psi, \xi \rangle_{L^2(\mu)}$ is normally distributed.

We define the closed subspace of square-integrable real-valued random variables by $E = \overline{\mathrm{span}\{\psi(x)^\top \xi(x) \mid x \in \mathbb{R}^d\}}$. For any $Z \in E^\perp$, the following holds:

$$\mathbb{E}[Z\langle \psi, \xi \rangle_{L^2(\mu)}] = \int \mathbb{E}[Z\psi(x)^\top \xi(x)]\, \mu(\mathrm{d}x) = 0.$$

Thus, $\langle \psi, \xi \rangle_{L^2(\mu)} \in (E^\perp)^\perp = E$ holds, meaning that $\langle \psi, \xi \rangle_{L^2(\mu)}$ is the $L^2(P)$ limit, and therefore, the law convergence limit of of normally distributed random variables. As will be shown later, the mean of $\langle \psi, \xi \rangle_{L^2(\mu)}$ is 0, and its variance is $\langle \psi, T_K \psi \rangle_{L^2(\mu)}$. Therefore, the characteristic function converges pointwise to the characteristic function of a normal distribution, implying that $\langle \psi, \xi \rangle_{L^2(\mu)}$ follows a normal distribution.

Moreover, the mean and variance are computed as:

$$\begin{aligned}
\mathbb{E}[\langle \psi, \xi \rangle_{L^2(\mu)}] &= \int \psi(x)^\top \mathbb{E}[\xi(x)]\, \mu(\mathrm{d}x) = 0, \\
\mathbb{E}[\langle \psi, \xi \rangle_{L^2(\mu)}^2] &= \int \psi(x)^\top \mathbb{E}[\xi(x)\xi(y)^\top] \psi(y)\, \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) \\
&= \int \psi(x)^\top T_K \psi(x)\, \mu(\mathrm{d}x) = \langle \psi, T_K \psi \rangle_{L^2(\mu)}.
\end{aligned}$$

Here, Fubini's theorem and the definition of Gaussian processes are used. □

Proposition D.5 (Karhunen–Loève expansion)**.** *Suppose $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ be a positive semidefinite kernel and satisfy $\int \|K(x,y)\| \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) < \infty$. Let $\{\kappa_n\}_{n \geq 1} \subset \mathbb{R}_{\geq 0}$ and $\{\psi_n\}_{n \geq 1} \subset L^2(\mu)^d$ be as in Lemma D.2. The sequence of the random variables $X_n \coloneqq \langle \psi_n, \xi \rangle_{L^2(\mu)} \sim \mathcal{N}(0, \kappa_n)$ is mutually independent. Furthermore, the Gaussian process $\xi \sim \mathrm{GP}(0, K)$ is represented as follows:*

$$\xi(x) = \sum_{n \geq 1} X_n \psi_n(x),$$

where the right-hand infinite sum means convergence in $L^2(P \otimes \mu)$.

Proof. First, we show that the sequence of the random variables $X_n = \langle \psi_n, \xi \rangle_{L^2(\mu)}$ $(n \geq 1)$ is mutually independent.

$$\begin{aligned}
\mathrm{E}[X_n X_m] &= \mathrm{E}\left[\int \psi_n(x)^\top \xi(x) \mu(\mathrm{d}x) \int \xi(y)^\top \psi_m(y) \mu(\mathrm{d}y) \right] \\
&= \int \psi_n(x)^\top \mathrm{E}\left[\xi(x)\xi(y)^\top\right] \psi_m(x) \mu \otimes \mu(\mathrm{d}x\mathrm{d}y) \\
&= \int \psi_n(x)^\top \left(\int K(x,y) \psi_m(y) \mu(\mathrm{d}y) \right) \mu(\mathrm{d}x) \\
&= \int \psi_n(x)^\top (\kappa_m \psi_m(x)) \mu(\mathrm{d}x) \\
&= \kappa_m \delta_{n,m},
\end{aligned}$$

where Fubini's theorem is used in the second and fourth line. The fifth line follows from the fact that ψ_m is the eigenvector of the integral operator T_K.

Thus $n \neq m$ implies the covariance of X_n and X_m is equal to 0. This implies that X_n and X_m are mutually independent, because they follow the normal distribution. Hence we obtain the mutual independency of the sequence $X_n = \langle \psi_n, \xi \rangle_{L^2(\mu)}$ $(n \geq 1)$.

The Karhunen-Loève expansion follows from

$$\begin{aligned}\mathrm{E}\left[\left\|\xi-\sum_{n=1}^{N}X_n\psi_n\right\|_{L^2(\mu)}^2\right] =& \mathrm{E}\Big[\|\xi\|_{L^2(\mu)}^2\Big]-2\sum_{n=1}^{N}\mathrm{E}\Big[X_n\left\langle\xi,\psi_n\right\rangle_{L^2(\mu)}\Big] \\ &+\sum_{n=1}^{N}\sum_{m=1}^{N}\left\langle\psi_n,\psi_m\right\rangle_{L^2(\mu)}\mathrm{E}[X_nX_m] \\ =& \mathrm{E}\Big[\|\xi\|_{L^2(\mu)}^2\Big]-\sum_{n=1}^{N}\mathrm{E}\big[X_n^2\big] \\ =& \sum_{n=N+1}^{\infty}\kappa_n \xrightarrow{N\to\infty} 0.\end{aligned}$$

□

Proposition D.6 (extended Markov's inequality)**.** *For any random variable X and non-decreasing positive-valued measurable function $\varphi:\mathbb{R}\to\mathbb{R}_{>0}$, the following holds :*

$$P(X\geq M)\leq\frac{\mathrm{E}[\varphi(X)]}{\varphi(M)}\quad(\forall M\in\mathbb{R}).$$

Proof. By applying Markov's inequality, we have

$$P(X\geq M)\leq\Pr(\varphi(X)\geq\varphi(M))\leq\frac{\mathrm{E}[\varphi(X)]}{\varphi(M)}\quad(\forall M\in\mathbb{R}).$$

□

With the above preparations, we obtain an upper bound for the tail probability of the L^2 norm of the Gaussian process.

Proposition D.7. *Suppose $K:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^{d\times d}$ be a positive semi-definite symmetric kernel and satisfy $\int\|K(x,y)\|\mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y)<\infty$ holds. Let $\{\kappa_n\}_{n\geq1}\subset\mathbb{R}_{\geq0}$ and $\{\psi_n\}_{n\geq1}\subset L^2(\mu)^d$ be as in Proposition D.2. The Gaussian process $\xi\sim\mathrm{GP}(0,K)$ with kernel function K satisfies the following:*

$$P\Big(\|\xi\|_{L^2(\mu)}\geq M\Big)\leq\exp\left(-\frac{e-1}{2e\kappa_1}M^2+\frac{\sum_{n\geq1}\kappa_n}{2\kappa_1}\right)\quad\Big(\forall M>\|T_K\|_{\mathrm{Tr},L^2(\mu)}^{\frac{1}{2}}\Big).$$

Proof. First, we calculate the characteristic function of $\|\xi\|_{L^2(\mu)}^2$.

Since $\xi\sim\mathrm{GP}(0,K)$, by Lemma D.5, we can express $\|\xi\|_{L^2(\mu)}^2=\sum_{n\geq1}X_n^2=:Y\quad(X_n\overset{\text{i.d.}}{\sim}\mathcal{N}(0,\kappa_n)\ \forall n\geq1)$. Let $Y_N=\sum_{n=1}^{N}{X_n}^2$. Then for any $s\in\mathbb{C}$, we have

$$\mathrm{E}[e^{s{X_n}^2}]=\frac{1}{\sqrt{1-2\kappa_n s}},$$

and so, by the independence of $\{X_n\}_{n\geq1}$,

$$\mathrm{E}[e^{sY_N}]=\prod_{n=1}^{N}\frac{1}{\sqrt{1-2\kappa_n s}}.$$

In particular, setting $s=\mathrm{i}u$ $(u\in\mathbb{R})$, the characteristic function of Y_N is

$$\log\mathrm{E}[e^{\mathrm{i}uY_N}]=-\frac{1}{2}\sum_{n=1}^{N}\log(1-2\mathrm{i}\kappa_n u).\tag{23}$$

Let $\theta_n = \mathrm{Arg}(1 - 2\mathrm{i}\kappa_n u)$ $(-\frac{\pi}{2} < \theta_n < \frac{\pi}{2})$. Then for the real and imaginary parts,

$$\begin{aligned}\left|\mathrm{Re}\sum_{n=1}^{N}(\log(1-2\mathrm{i}\kappa_n u))\right| &\le \frac{1}{2}\sum_{n=1}^{N}\log\bigl(1+4\kappa_n^2u^2\bigr) \le \frac{1}{2}\sum_{n=1}^{N}4{\kappa_n}^2u^2 \\ &\le 2u^2\sum_{n\ge1}{\kappa_n}^2 \le 2u^2\left(\sum_{n\ge1}\kappa_n\right)^2 < \infty, \\ \left|\mathrm{Im}\sum_{n=1}^{N}(\log(1-2\mathrm{i}\kappa_n u))\right| &\le \sum_{n=1}^{N}|\theta_n| \le \sum_{n=1}^{N}|\tan\theta_n| = \sum_{n=1}^{N}2\kappa_n|u| \le 2|u|\sum_{n\ge1}\kappa_n < \infty.\end{aligned}$$

Thus, the characteristic function of Y_N converges pointwise as $N \to \infty$. By Levy's continuity theorem, Y_N weakly converges to $Y = \lim_{N\to\infty} Y_N = \sum_{n\ge1} {X_n}^2$, and its characteristic function is $\log \mathrm{E}[e^{\mathrm{i}uY}] = -\frac{1}{2}\sum_{n\ge1}\log(1-2\mathrm{i}\kappa_n u)$.
Now, we will proceed to

- Show that for some $t > 0$, replacing $\mathrm{i}u$ with t makes the series on the right-hand side of (23) converge. This allows us to show the existence of the moment generating function $\mathrm{E}[e^{tY}]$ for such t.
- Use Lemma D.6 with $\varphi(x) \leftarrow e^{tx},\ X \leftarrow Y$ to obtain the upper bound for the tail probability.
- Adjust $t > 0$ (within the range where the moment generating function exists) to obtain the best possible upper bound for the tail probability.

The function $t \mapsto -\frac{1}{2}\log(1-2\kappa_n t)$ is convex, and the tangent at $t = 0$ is $t \mapsto \kappa_n t$. Thus for $a > 1$, the equation for t;

$$a\kappa_n t = -\frac{1}{2}\log(1-2\kappa_n t)$$

has a solution $t = t_n > 0$ for $t > 0$, and it holds that

$$0 \le t \le t_n \implies -\frac{1}{2}\log(1-2\kappa_n t) \le a\kappa_n t.$$

Since $\kappa_n t_n$ is constant for $n \ge 1$ and $\{\kappa_n\}_{n\ge1}$ is non-increasing, the sequence $\{t_n\}_{n\ge1}$ is non-decreasing. Hence, for any $n \ge 1,\ t \in [0, t_1]$, we have $-\frac{1}{2}\log(1-2\kappa_n t) \le a\kappa_n t$, so

$$\log \mathrm{E}[e^{t\|\xi\|_\mu^2}] \le at\sum_{n\ge1}\kappa_n < \infty.$$

Therefore, for $t \in [0, t_1]$, the moment generating function $\mathrm{E}[e^{t\|\xi\|_\mu^2}]$ exists, and $\mathrm{E}[e^{t\|\xi\|_\mu^2}] \le e^{at\sum_{n\ge1}\kappa_n}$. Hence, for the upper bound of the tail probability of the Gaussian process norm, we have

$$\begin{aligned}P\Bigl(\|\xi\|_\mu \ge M\Bigr) &\le \frac{\mathrm{E}[e^{t\|\xi\|_\mu^2}]}{e^{tM^2}} \\ &\le \exp\left(-tM^2 + at\sum_{n\ge1}\kappa_n\right).\end{aligned}$$

By optimizing the right-hand side of this expression with respect to $t,\ a$, we obtain the best upper bound. For $a > 1,\ at\kappa_1 = -\frac{1}{2}\log(1-2\kappa_1 t)$, considering large M, we solve

$$\text{minimize } f(t) = -tM^2 + at\sum_{n\ge1}\kappa_n, \tag{24}$$

$$\text{s.t. } t > 0,\ \ a = -\frac{1}{2\kappa_1 t}\log(1-2\kappa_1 t) > 1. \tag{25}$$

The derivative of (24) is

$$\frac{\mathrm{d}}{\mathrm{d}t}f(t) = -M^2 + \frac{\sum_{n\geq 1}\kappa_n}{1-2\kappa_1 t} = 0, \tag{26}$$

which gives

$$t \coloneqq t^* = \frac{1}{2\kappa_1}\left(1 - \frac{\sum_{n\geq 1}\kappa_n}{M^2}\right). \tag{27}$$

If $M^2 > \sum_{n\geq 1}\kappa_n$, then $t^* > 0, a > 1$, satisfying the condition (25). The upper bound becomes

$$\begin{aligned} P\Big(\|\xi\|_{L^2(\mu)} \geq M\Big) &\leq \exp f(t^*) \\ &= \exp\left(-t^*M^2 - \frac{\sum_{n\geq 1}\kappa_n}{2\kappa_1}\log(1-2\kappa_1 t^*)\right) \\ &= \exp\left(-\frac{M^2}{2\kappa_1} + \frac{\sum_{n\geq 1}\kappa_n}{2\kappa_1} + \frac{\sum_{n\geq 1}\kappa_n}{2\kappa_1}\log\frac{M^2}{\sum_{n\geq 1}\kappa_n}\right), \end{aligned}$$

where in the third line we used (26), (27). Finally, applying the inequality $-x + \log x \leq -\frac{e-1}{e}x$ at $x = \frac{M^2}{\sum_{n\geq 1}\kappa_n}$, we obtain

$$f(t^*) \leq \frac{\sum_{n\geq 1}\kappa_n}{2\kappa_1}\left(1 - \frac{e-1}{e}\frac{M^2}{\sum_{n\geq 1}\kappa_n}\right).$$

So, we conclude

$$P\Big(\|\xi\|_{L^2(\mu)} \geq M\Big) \leq \exp\left(-\frac{e-1}{2e\kappa_1}M^2 + \frac{\sum_{n\geq 1}\kappa_n}{2\kappa_1}\right).$$

□

We apply the results from the previous section.

Lemma D.8. *Let K_μ be the Hessian-based kernel introduced in (11). Then, the Gaussian process $\xi \sim \mathrm{GP}(0, K_\mu)$ satisfies the following :*

- $\langle \psi_n, \xi\rangle_{L^2(\mu^\dagger)} \sim \mathcal{N}(0, {\lambda_n}^2)$.
- *P-almost surely, it occurs that $\xi \in \mathcal{R}(H_\mu) \subset \mathrm{Tan}_\mu\mathcal{P}_2(\mathbb{R}^d)$. In particular, from the assumptions, $\mathrm{Id}+\eta_p\xi$ gives the optimal transport map from μ to $(\mathrm{Id}+\eta_p\xi)\#\mu$ for sufficiently small $\eta_p > 0$ P-a.s.*
- *For any constant $M > R_2 \geq \left(\sum_{n\geq 1}{\lambda_n}^2\right)^{\frac{1}{2}}$, the following holds :*

$$P\Big(\|\xi\|_{L^2(\mu)} \geq M\Big) \leq \exp\left(-\frac{e-1}{2e{\lambda_1}^2}M^2 + \frac{\sum_{n\geq 1}{\lambda_n}^2}{2{\lambda_1}^2}\right).$$

Proof. Given the assumption $\int \left\|\nabla_\mu^2 F(\mu, x, y)\right\| \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) < \infty$, Lemma D.2 can be applied to $K = \nabla_\mu^2 F(\mu)$. Specifically, there exists a sequence of real numbers $\{\lambda_n\}_{n\geq 1} \subset \mathbb{R}\setminus\{0\}$ satisfying $\|H_\mu\|^2_{\mathrm{HS},L^2(\mu)} = \sum_{n\geq 1}{\lambda_n}^2 < \infty$ and an orthonormal basis $\{\psi_n\}_{n\geq 1}$ of $L^2(\mu)$, such that

$$\nabla_\mu^2 F(\mu, x, y) = \nabla_1\nabla_2\frac{\delta^2 F}{\delta\mu^2}(x, y) = \sum_{n\geq 1}\lambda_n\psi_n(x)\psi_n(y)^\top, \quad \mu^{\otimes 2}\text{-a.e. } (x, y). \tag{28}$$

Regarding the kernel of the squared Hessian K_μ,

$$K_\mu(x, y) = \sum_{n\geq 1}\lambda_n^2\psi_n(x)\psi_n(y)^\top, \quad \mu^{\otimes 2}\text{-a.e. } (x, y).$$

Here, K_μ satisfies the assumptions of Lemma D.5, corresponding to the case where $\kappa_n = \lambda_n^2$ in that lemma. Then, from Proposition D.4, $\langle \psi_n, \xi \rangle_{L^2(\mu)} \sim \mathcal{N}(0, \lambda_n^2)$ holds. Finally, it follows from Proposition D.7 that for any $M \geq R_2 \geq \|H_\mu\|_{\mathrm{HS}, L^2(\mu)} = \left\|H_\mu^2\right\|_{\mathrm{Tr}, L^2(\mu)} = \left(\sum_{n \geq 1} {\lambda_n}^2\right)^{\frac{1}{2}}$,

$$P\Big(\|\xi\|_{L^2(\mu)} \geq M\Big) \leq \exp\left(-\frac{e-1}{2e{\lambda_1}^2} M^2 + \frac{\sum_{n \geq 1} {\lambda_n}^2}{2{\lambda_1}^2}\right).$$

□

E Proofs for Convergence Analysis

The lemmas for the main theorem are re-stated and the proofs are provided here.

E.1 Continuous-time Convergence Analysis

E.1.1 Proof Sketch for Lemmas

The proof of Lemma 5.3 is relatively straightforward using the chain rule. In contrast, the proof of Lemma 5.4 is lengthy and relies on two sub-lemmas inspired by the analysis of SSRGD (Li, 2019) in Euclidean spaces.

Lemma E.1: small increase by perturbation. This lemma ensures that the increase in the objective function due to perturbation is upper bounded by F_{thres} with high probability.

Lemma E.2: large decrease by WGF. Put simply, this lemma asserts that a small deviation in the direction of the eigenvector corresponding to the smallest eigenvalue of the Hessian results in exponential decrease in the objective under WGF. This result is derived from the fact that the dynamics of slightly deviated two points evolve approximately under $H_{\mu^\dagger}$, causing their "distance" to grow exponentially over time.

By combining these results, namely that the objective function does not increase significantly and decreases substantially after perturbation, the lemma is proven.

E.1.2 Proof of Lemma 5.3

Lemma 5.3. *For a curve of probability measures μ_t following the WGF, the following holds:*

$$F(\mu_0) - F(\mu_t) = \int_0^t \|\nabla_\mu F(\mu_\tau)\|_{\mu_\tau}^2 \mathrm{d}\tau.$$

Proof. We have

$$F(\mu_0) - F(\mu_t) = \int_0^t -\frac{\mathrm{d}}{\mathrm{d}\tau} F(\mu_\tau)\mathrm{d}\tau = \int_0^t \|\nabla_\mu F(\mu_\tau)\|_{L^2(\mu_\tau)}^2 \mathrm{d}\tau,$$

due to the chain rule. □

E.1.3 Proof of Proposition 5.4

Proposition 5.4. *Set $\eta = O(1)$ and let ε, δ, η_p, T_{thres}, F_{thres} be chosen as in Theorem 5.2. Suppose $\mu^\dagger \in \mathcal{P}_2^a(\mathbb{R}^d)$ satisfies $\left\|\nabla_\mu F(\mu^\dagger)\right\|_{L^2(\mu^\dagger)} < \varepsilon$ and $\lambda_0 := \lambda_{\min} H_{\mu^\dagger} \leq -\delta$. Generating $\xi \sim \mathrm{GP}(0, k_\mu)$ and setting $\mu_0 = (\mathrm{Id} + \eta_p \xi)\sharp\mu^\dagger$ as the initial point of the WGF, we have with probability $1 - \zeta'$:*

$$F(\mu^\dagger) - F(\mu_{T_{\mathrm{thres}}}) \geq F_{\mathrm{thres}}.$$

Intuitively, Proposition 5.4 indicates that when a perturbation is applied near the saddle point, the objective function decreases over a period of time T_{thres} with high probability, which corresponds to escaping the saddle point. As discussed above, the approach is based on the argument by Li (2019) and utilizes Lemma E.1 and Lemma E.2. These two lemmas postulate that the L^2-norm $\|\xi\|_{L^2(\mu)}$ of the Gaussian process ξ is uniformly bounded. This corresponds to the condition that perturbations in Euclidean space are sampled from spheres with a fixed radius, thus the perturbation size was uniformly bounded. In infinite-dimensional spaces, such uniform sampling cannot be used, which is one of the reasons why the Gaussian process is employed. In this case, the norm of the Gaussian process can take arbitrarily large values even though with low probability. Therefore, it is necessary to exploit tail probability estimates of the norm of the Gaussian process, as in the following.

Lemma D.8. *Let K_μ be the Hessian-based kernel introduced in (11). Then, the Gaussian process $\xi \sim \mathrm{GP}(0, K_\mu)$ satisfies the following :*

- $\langle \psi_n, \xi \rangle_{L^2(\mu^\dagger)} \sim \mathcal{N}(0, {\lambda_n}^2)$.
- *P-almost surely, it occurs that* $\xi \in \mathcal{R}(H_\mu) \subset \mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d)$. *In particular, from the assumptions,* $\mathrm{Id}+\eta_p\xi$ *gives the optimal transport map from* μ *to* $(\mathrm{Id}+\eta_p\xi)\#\mu$ *for sufficiently small* $\eta_p > 0$ *P-a.s.*
- *For any constant* $M > R_2 \geq \left(\sum_{n\geq 1} {\lambda_n}^2\right)^{\frac{1}{2}}$, *the following holds :*

$$P\Big(\|\xi\|_{L^2(\mu)} \geq M\Big) \leq \exp\left(-\frac{e-1}{2e{\lambda_1}^2}M^2 + \frac{\sum_{n\geq 1}{\lambda_n}^2}{2{\lambda_1}^2}\right).$$

Let [5]

$$M = \left(\frac{e{R_1}^2}{e-1}\left(1+2\log\frac{2}{\zeta'}\right)\right)^{\frac{1}{2}} \vee 2R_1\left(\log\frac{4\sqrt{2}}{\zeta'}\right)^{\frac{1}{2}} = \tilde{O}(1). \tag{29}$$

From Assumption 1,

$$R_1{}^2 \geq \int \left\|\nabla_\mu^2 F(\mu,x,y)\right\|_{\mathrm{F}}^2 \mu^{\otimes 2}(\mathrm{d}x\mathrm{d}y) = \sum_{n\geq 1}{\lambda_n}^2 \geq {\lambda_1}^2,$$

so that

$$\begin{aligned} M^2 &\geq \frac{e{R_1}^2}{e-1}\left(1+2\log\frac{2}{\zeta'}\right) \\ &\geq \frac{e}{e-1}\sum_{n\geq 1}{\lambda_n}^2 + \frac{2e}{e-1}{\lambda_1}^2\log\frac{2}{\zeta'} \end{aligned}$$

holds. Then it follows that

$$\exp\left(-\frac{e-1}{2e{\lambda_1}^2}M^2 + \frac{\sum_{n\geq 1}{\lambda_n}^2}{2{\lambda_1}^2}\right) \leq \frac{\zeta'}{2}.$$

Therefore, from Lemma D.8, $\|\xi\|_{L^2(\mu)} \leq \tilde{M} = \tilde{O}(1)$ occurs with probability at least $1 - \frac{\zeta'}{2}$.

We now take the hyperparameters $\eta_p,\ F_{\mathrm{thres}},\ T_{\mathrm{thres}}$ as follows:

$$\begin{aligned} T_{\mathrm{thres}} = \frac{2}{\delta}\log\frac{16{L_1}^{\frac{1}{2}}M}{\sqrt{e}\delta^{\frac{1}{2}}r} &= \tilde{O}\left(\frac{1}{\delta}\right) \\ &= O\left(\frac{1}{\delta}\log\frac{\log\frac{1}{\zeta'}}{\delta^{\frac{1}{2}}\zeta'}\right), \\ \eta_p = \frac{2F_{\mathrm{thres}}}{M(\varepsilon+\sqrt{\varepsilon^2+2L_1F_{\mathrm{thres}}})} &= \tilde{O}\left(\frac{\delta^3}{\varepsilon}\wedge\delta^{\frac{3}{2}}\right) \\ &= O\left(\delta^3\varepsilon^{-1}\left(\log\frac{1}{\zeta'}\right)^{-1}\left(\log\frac{\log\frac{1}{\zeta'}}{\delta^{\frac{1}{2}}\zeta'}\right)^{-3} \wedge \delta^{\frac{3}{2}}\left(\log\frac{1}{\zeta'}\right)^{-1}\left(\log\frac{\log\frac{1}{\zeta'}}{\delta^{\frac{1}{2}}\zeta'}\right)^{-\frac{3}{2}}\right), \\ F_{\mathrm{thres}} = \frac{{T_{\mathrm{thres}}}^{-3}}{18(L_2+L_3)^2}\log^2\frac{3}{2} &= \tilde{O}(\delta^3) \\ &= O\left(\delta^3\left(\log\frac{\log\frac{1}{\zeta'}}{\delta^{\frac{1}{2}}\zeta'}\right)^{-3}\right). \end{aligned} \tag{30}$$

It should be noted that $F_{\mathrm{thres}} = \eta_p M\varepsilon + \frac{{L_1}^2}{2}{\eta_p}^2M^2$ holds.

[5] The right-hand side implies the upper bound $\sqrt{2}\exp\left(-\frac{M^2}{4{\lambda_0}^2}\right) \leq \frac{\zeta'}{4}$, where λ_0 is the smallest eigenvalue of H_μ. This result is utilized later in the proof of Proposition 5.4.

Lemma E.1. *Consider the situation of Proposition 5.4, and let $\mu^\dagger$ be an (ε, δ)-saddle point, let the initial point be $\mu_0 \coloneqq (\mathrm{Id} + \eta_p \xi)\#\mu^\dagger$, and the hyperparameters $\eta_p, F_{\mathrm{thres}}$ be taken as in (30). If the L^2-norm of the Gaussian process ξ satisfies $\|\xi\|_{L^2(\mu)} \leq M$, then it holds that*

$$F(\mu_0) - F(\mu^\dagger) \leq F_{\mathrm{thres}}.$$

Proof. Let $\nu_h = (\mathrm{Id} + h\eta_p\xi)\sharp\mu^\dagger$ for $h \in [0,1]$, joining $\mu^\dagger$ and $\mu_0 = (\mathrm{Id} + \eta_p\xi)\#\mu^\dagger$. From Lemma C.6,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}h} F(\mu_h) &= \eta_p \left\langle \nabla_\mu F(\nu_h) \circ (\mathrm{Id} + h\eta_p\xi), \xi \right\rangle_{L^2(\mu^\dagger)} \\
&\leq \eta_p \|\nabla_\mu F(\nu_h) \circ (\mathrm{Id} + h\eta_p\xi)\|_{L^2(\mu^\dagger)} \|\xi\|_{L^2(\mu^\dagger)} \\
&\leq \eta_p \Big(\big\|\nabla_\mu F(\mu^\dagger)\big\|_{L^2(\mu^\dagger)} + \big\|\nabla_\mu F(\nu_h) \circ (\mathrm{Id} + h\eta_p\xi) - \nabla_\mu F(\mu^\dagger)\big\|_{L^2(\mu^\dagger)} \Big) \|\xi\|_{L^2(\mu^\dagger)} \\
&\leq \eta_p \left(\varepsilon + L_1 \left(\int \|x - y\|^2 (\mathrm{Id} \times (\mathrm{Id} + h\eta_p\xi))\#\mu^\dagger(\mathrm{d}x\mathrm{d}y) \right)^{\frac{1}{2}} \right) \|\xi\|_{L^2(\mu^\dagger)} \\
&= \eta_p \|\xi\|_{L^2(\mu^\dagger)} \varepsilon + L_1 \eta_p^2 \|\xi\|^2_{L^2(\mu^\dagger)} h.
\end{aligned}$$

Here, the Cauchy-Schwarz inequality was used in the second line, the conditions $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} < \varepsilon$ and the gradient's Lipschitz continuity were applied in the fourth line, and Proposition B.4 was invoked in the fifth line. Therefore,

$$\begin{aligned}
F(\mu_0) - F(\mu^\dagger) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}h} F(\mu_h)\, \mathrm{d}h \\
&\leq \eta_p \|\xi\|_{L^2(\mu^\dagger)} \varepsilon + \frac{L_1}{2} \eta_p^2 \|\xi\|^2_{L^2(\mu^\dagger)} \\
&\leq \eta_p M \varepsilon + \frac{L_1}{2} \eta_p^2 M^2 \\
&= F_{\mathrm{thres}}.
\end{aligned}$$

The third line follows from $\|\xi\|_{L^2(\mu^\dagger)} \leq M$ and the fourth line follows from the choice of η_p in (30). □

Lemma E.2. *Consider the situation of Proposition 5.4. That is, let $\mu^\dagger$ be an (ε, δ)-saddle point, $\mu_0 \coloneqq (\mathrm{Id} + \eta_p\xi)\#\mu^\dagger$, and the hyperparameters $\eta_p, F_{\mathrm{thres}}$ be taken as (30). Furthermore, we set $\tilde{\mu}_0 = (\mathrm{Id} + \eta_p\tilde{\xi})\sharp\mu^\dagger$, where $\tilde{\xi} = \xi + r\psi_0$, $r \in \mathbb{R}$ is a constant, and ψ_0 is the eigenvector of the Hessian operator H_μ corresponding to the smallest eigenvalue λ_0.*

Letting $\mu_t,\ \tilde{\mu}_t$ be WGF initialized at $\mu_0,\ \tilde{\mu}_0$ respectively, then $\|\xi\|_{L^2(\mu^\dagger)} \leq M$ and $\frac{\sqrt{2\pi}|\lambda_0|\zeta'}{8} \leq |r| \leq 2M$ implies that there exists $t \in [0, T_{\mathrm{thres}}]$ satisfying

$$(F(\mu_0) - F(\mu_t)) \vee (F(\tilde{\mu}_0) - F(\tilde{\mu}_t)) \geq 2F_{\mathrm{thres}}.$$

Proof. We prove this lemma by contradiction. We assume that for any $t \in [0, T_{\mathrm{thres}}]$, it holds that

$$(F(\mu_0) - F(\mu_t)) \vee (F(\tilde{\mu}_0) - F(\tilde{\mu}_t)) < 2F_{\mathrm{thres}}.$$

We denote the characteristics of the vector fields $-\nabla_\mu F(\mu_t)$ and $-\nabla_\mu F(\tilde{\mu}_t)$ as $X_t, \tilde{X}_t$ respectively. That is, $\mu_t = X_t\#\mu_0, \tilde{\mu}_t = \tilde{X}_t\#\tilde{\mu}_0$, and

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = -\nabla_\mu F(\mu_t) \circ X_t, \quad \frac{\mathrm{d}}{\mathrm{d}t} \tilde{X}_t = -\nabla_\mu F(\tilde{\mu}_t) \circ \tilde{X}_t.$$

It is worth noting here that

$$\begin{aligned}\|X_t - \mathrm{Id}\|_{L^2(\mu_0)} &= \left\| \int_0^t (-\nabla_\mu F(\mu_\tau) \circ X_\tau) \mathrm{d}\tau \right\|_{L^2(\mu_0)} \\ &\leq \int_0^t \|\nabla_\mu F(\mu_\tau)\|_{L^2(\mu_\tau)} \mathrm{d}\tau \\ &\leq t^{\frac{1}{2}} \left(\int_0^t \|\nabla_\mu F(\mu_\tau)\|^2_{L^2(\mu_\tau)} \mathrm{d}\tau \right)^{\frac{1}{2}} \\ &\leq (T_{\mathrm{thres}}(F(\mu_0) - F(\mu_t)))^{\frac{1}{2}},\end{aligned}$$

where the first line follows from the fact that X_t is the characteristic of $-\nabla_\mu F(\mu_t)$, the second from the properties of the Bochner integral, the third from Jensen's inequality, and the fourth from Proposition 5.3. Similarly, $\|\tilde{X}_t - \mathrm{Id}\|_{L^2(\tilde{\mu}_0)} \leq (T_{\mathrm{thres}} F(\tilde{\mu}_0) - F(\tilde{\mu}_t))$ holds. Furthurmore, we set $Y_t \coloneqq X_t \circ (\mathrm{Id} + \eta_p \xi)$ and $\tilde{Y}_t = \tilde{X}_t \circ (\mathrm{Id} + \eta_p \xi)$, which satisfy

$$\begin{aligned}\|Y_t - \mathrm{Id}\|_{L^2(\mu^\dagger)} &\leq \|X_t - (\mathrm{Id} + \eta_p \xi)\|_{L^2(\mu_0)} + \eta_p \|\xi\|_{L^2(\mu^\dagger)} \\ &\leq (T_{\mathrm{thres}}(F(\mu_0) - F(\mu_t)))^{\frac{1}{2}} + \eta_p \|\xi\|_{L^2(\mu^\dagger)} \\ &\leq \sqrt{2} T_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}} + \eta_p M, \end{aligned} \tag{31}$$

and

$$\begin{aligned}\|\tilde{Y}_t - \mathrm{Id}\|_{L^2(\mu^\dagger)} &\leq (T_{\mathrm{thres}}(F(\tilde{\mu}_0) - F(\tilde{\mu}_t)))^{\frac{1}{2}} + \eta_p \left\| \tilde{\xi} \right\|_{L^2(\mu^\dagger)} \\ &\leq (T_{\mathrm{thres}}(F(\tilde{\mu}_0) - F(\tilde{\mu}_t)))^{\frac{1}{2}} + \eta_p \|\xi\|_{L^2(\mu^\dagger)} + \eta_p r \\ &\leq \sqrt{2} T_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}} + \eta_p M + \eta_p r. \end{aligned} \tag{32}$$

We analyze the vector $w_t \coloneqq \tilde{Y}_t - Y_t$. The goal is to obtain a contradiction by confirming that $\|w_t\|_{L^2(\mu^\dagger)}$ becomes large. To achieve this, we investigate how w_t evolves with respect to time:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} w_t &= \frac{\mathrm{d}}{\mathrm{d}t} \tilde{Y}_t - \frac{\mathrm{d}}{\mathrm{d}t} Y_t \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \tilde{X}_t \circ (\mathrm{Id} + \eta_p \tilde{\xi}) - \frac{\mathrm{d}}{\mathrm{d}t} X_t \circ (\mathrm{Id} + \eta_p \xi) \\ &= -\nabla_\mu F(\tilde{\mu}_t) \circ \tilde{X}_t \circ (\mathrm{Id} + \eta_p \tilde{\xi}) + \nabla_\mu F(\mu_t) \circ X_t \circ (\mathrm{Id} + \eta_p \xi) \\ &= -\int_0^1 \frac{\mathrm{d}}{\mathrm{d}h} (\nabla_\mu F(\nu_h) \circ Y_h) \mathrm{d}h, \end{aligned} \tag{33}$$

where ν_h is a curve connecting μ_t and $\tilde{\mu}_t$; $\nu_h \coloneqq ((1-h)Y_t + h\tilde{Y}_t) \sharp \mu^\dagger$ $(h \in [0,1])$, $\nu_0 = \mu_t$ and $\nu_1 = \tilde{\mu}_t$ hold. Moreover, its direction vector is always $w_t = \tilde{Y}_t - Y_t$, so the integrand of (33) is obtained by operating w_t. Specifically,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}h} (\nabla_\mu F(\nu_h) \circ Y_h)(x) \\ &\quad = \int \nabla_\mu^2 F(\nu_h, Y_h(x), Y_h(y)) w_t(y) \mu^\dagger(\mathrm{d}y) + \nabla \nabla_\mu F(\nu_h, Y_h(x)) w_t(x)\end{aligned}$$

holds. Let

$$\frac{\mathrm{d}}{\mathrm{d}t} w_t = -H_{\mu^\dagger} w_t - \left(\int_0^1 \Delta_{t,h} \mathrm{d}h \right) w_t = -H_{\mu^\dagger} w_t - \Delta_t w_t, \tag{34}$$

where $\Delta_{t,h}$ is an operator on $L^2(\mu^\dagger)^d$, which satisfies

$$\begin{aligned}
&\Delta_{t,h} f(x) \\
&= \int \big(\nabla_\mu^2 F(\nu_h, Y_h(x), Y_h(y)) - \nabla_\mu^2 F(\mu^\dagger, x, y)\big) f(y) \mu^\dagger(\mathrm{d}y) \\
&\quad + \nabla\nabla_\mu F(\nu_h, Y_h(x)) f(x) \\
&= \int \big(\nabla_\mu^2 F(\nu_h, Y_h(x), Y_h(y)) - \nabla_\mu^2 F(\mu^\dagger, x, y)\big) f(y) \mu^\dagger(\mathrm{d}y) \\
&\quad + \big(\nabla\nabla_\mu F(\nu_h, Y_h(x)) - \nabla\nabla_\mu F(\mu^\dagger, x)\big) f(x) \\
&\quad + \nabla\nabla_\mu F(\mu^\dagger, x) f(x). \qquad (35)
\end{aligned}$$

Moreover, the operator Δ_t is defined as $\Delta_t = \int_0^1 \Delta_{t,h} \mathrm{d}h$, the norm of which is bounded as $\tilde{O}\Big(T_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}} + \eta_p + \varepsilon\Big)$ from (31), (32), and (35). It should be noted that from (34), w_t evolves according to $-H_{\mu^\dagger}$ unless $\|\Delta_t\|_{L^2(\mu^\dagger)}$ is small. Since

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(e^{tH_{\mu^\dagger}} w_t\Big) = e^{tH_{\mu^\dagger}} \left(\frac{\mathrm{d}}{\mathrm{d}t} w_t + H_{\mu^\dagger} w_t\right) = -e^{tH_{\mu^\dagger}} \Delta_t w_t$$

holds from (34), by integrating we have

$$e^{tH_{\mu^\dagger}} w_t - w_0 = -\int_0^t e^{\tau H_{\mu^\dagger}} \Delta_\tau w_\tau \mathrm{d}\tau,$$

and so

$$w_t = e^{-tH_{\mu^\dagger}} w_0 - \int_0^t e^{(\tau-t)H_{\mu^\dagger}} \Delta_\tau w_\tau \mathrm{d}\tau = e^{-\lambda_0 t} \eta_p r \psi_0 - \int_0^t e^{(\tau-t)H_{\mu^\dagger}} \Delta_\tau w_\tau \mathrm{d}\tau,$$

where we used the equation $w_0 = \tilde{Y}_0 - Y_0 = \eta_p(\tilde{\xi} - \xi) = \eta_p r \psi_0$. Then, it holds that

$$\begin{aligned}
\Big|\|w_t\|_{L^2(\mu^\dagger)} - e^{-\lambda_0 t}\eta_p r\Big| &= \Big|\|w_t\|_{L^2(\mu^\dagger)} - \big\|e^{-\lambda_0 t}\eta_p r\psi_0\big\|_{L^2(\mu^\dagger)}\Big| \\
&\le \big\|w_t - e^{-\lambda_0 t}\eta_p r\psi_0\big\|_{L^2(\mu^\dagger)} \\
&\le \int_0^t \Big\|e^{(\tau-t)H_{\mu^\dagger}}\Big\|_{L^2(\mu^\dagger)} \|\Delta_\tau\|_{L^2(\mu^\dagger)} \|w_\tau\|_{L^2(\mu^\dagger)} \mathrm{d}\tau \\
&\le \int_0^t e^{\lambda_0(\tau-t)} \|\Delta_\tau\|_{L^2(\mu^\dagger)} \|w_\tau\|_{L^2(\mu^\dagger)} \mathrm{d}\tau. \qquad (36)
\end{aligned}$$

Here we use $\|e^{(\tau-t)H_{\mu^\dagger}}\|_{L^2(\mu^\dagger)} = e^{\lambda_0(\tau-t)}$, which is implied by $\lambda_0 = \lambda_{\min}(H_{\mu^\dagger})$ and $\tau - t \le 0$. From (35) and the inequality $|H'_{\mu^\dagger}\|_{L^2(\mu^\dagger)} \le R_2 \big\|\nabla_\mu F(\mu^\dagger)\big\|_{L^2(\mu^\dagger)} \le R_2\varepsilon$, it follows from (31) and (32) that

$$\begin{aligned}
\|\Delta_{t,h}\|_{L^2(\mu^\dagger)} &\le (L_2 + L_3)\|Y_h - \mathrm{Id}\|_{L^2(\mu^\dagger)} + R_2\varepsilon \\
&= (1-h)(L_2 + L_3)\|Y_t - \mathrm{Id}\|_{L^2(\mu^\dagger)} \\
&\qquad + h(L_2 + L_3)\Big\|\tilde{Y}_t - \mathrm{Id}\Big\|_{L^2(\mu^\dagger)} + R_2\varepsilon \\
&\le (L_2 + L_3)\Big(\sqrt{2} T_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}} + \eta_p M + h\eta_p r\Big) + R_2\varepsilon.
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
\|\Delta_t\|_{L^2(\mu^\dagger)} &\le \int_0^1 \|\Delta_{t,h}\|_{L^2(\mu^\dagger)} \mathrm{d}h \\
&\le (L_2 + L_3)\left(\sqrt{2} T_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}} + \eta_p M + \frac{1}{2}\eta_p r\right) + R_2\varepsilon. \\
&\le (L_2 + L_3)\Big(\sqrt{2} T_{\mathrm{thres}}{}^{\frac{1}{2}} F_{\mathrm{thres}}{}^{\frac{1}{2}} + 2\eta_p M\Big) + R_2\varepsilon \\
&=: \Delta.
\end{aligned}$$

By manipulating the inequality (36), the following is obtained:

$$\left| e^{\lambda_0 t} \|w_t\|_{L^2(\mu^\dagger)} - \eta_p r \right| \leq \Delta \int_0^t e^{\lambda_0 \tau} \|w_\tau\|_{L^2(\mu^\dagger)} \mathrm{d}\tau.$$

The application of Gronwall's inequality to the upper bound yields $e^{\lambda_0 t}\|w_t\|_{L^2(\mu^\dagger)} \leq \eta_p r e^{\Delta t}$. Then it holds that

$$\begin{aligned}
\|w_t\|_{L^2(\mu^\dagger)} &\geq \eta_p r e^{-\lambda_0 t} - \Delta e^{-\lambda_0 t} \int_0^t e^{\lambda_0 \tau} \|w_t\|_{L^2(\mu^\dagger)} \, \mathrm{d}\tau \\
&\geq \eta_p r e^{-\lambda_0 t} \left(1 - \Delta \int_0^t e^{\Delta \tau} \, \mathrm{d}\tau \right) \\
&= \eta_p r e^{-\lambda_0 t} (2 - e^{\Delta t}). \qquad (37)
\end{aligned}$$

The left side of (37) is upper-bounded as

$$\begin{aligned}
\|w_t\|_{L^2(\mu^\dagger)} &\leq \left\| \tilde{Y}_t - \mathrm{Id} \right\|_{L^2(\mu^\dagger} + \|Y_t - \mathrm{Id}\|_{L^2(\mu^\dagger} \\
&< 2\sqrt{2} {T_{\mathrm{thres}}}^{\frac{1}{2}} {F_{\mathrm{thres}}}^{\frac{1}{2}} + 2\eta_p M + \eta_p r \\
&\leq 4\sqrt{2} {T_{\mathrm{thres}}}^{\frac{1}{2}} {F_{\mathrm{thres}}}^{\frac{1}{2}},
\end{aligned}$$

where in the last line it follows from ${T_{\mathrm{thres}}}^{\frac{1}{2}} {F_{\mathrm{thres}}}^{\frac{1}{2}} = \tilde{O}(\delta)$, $\eta_p M = o(\delta)$, $\eta_p = o(\delta)$.
On the other hand, using $\eta_p M T_{\mathrm{thres}} = o(1)$, $\varepsilon T_{\mathrm{thres}} = o(1)$ and the definition of F_{thres}, we have

$$\begin{aligned}
t\Delta &\leq T_{\mathrm{thres}} \Delta \\
&= ((L_2 + L_3)(\sqrt{2} {T_{\mathrm{thres}}}^{\frac{1}{2}} {F_{\mathrm{thres}}}^{\frac{1}{2}} + 2\eta_p M) + R_2 \varepsilon) T_{\mathrm{thres}} \\
&= \sqrt{2}(L_2 + L_3) {T_{\mathrm{thres}}}^{\frac{3}{2}} {F_{\mathrm{thres}}}^{\frac{1}{2}} + 2(L_2 + L_3)\eta_p M T_{\mathrm{thres}} + R_2 \varepsilon T_{\mathrm{thres}} \\
&= \frac{1}{3} \log \frac{3}{2} + \frac{1}{3} \log \frac{3}{2} + \frac{1}{3} \log \frac{3}{2} \\
&= \log \frac{3}{2}.
\end{aligned}$$

Then the right side of (37) is lower-bounded as

$$\begin{aligned}
\eta_p r e^{-\lambda_0 t} (2 - e^{\Delta t}) &\geq \frac{\eta_p r}{2} e^{\delta t} \\
&\geq \frac{\eta_p r}{2} (e\delta t)^{\frac{1}{2}} e^{\frac{\delta}{2} t},
\end{aligned}$$

where we used in the second line the inequality $e^{\frac{x}{2}} \leq \sqrt{e} x^{\frac{1}{2}}$ as $x = \delta t$. Letting $t = T_{\mathrm{thres}}$ and transforming (37) yields

$$\begin{aligned}
e^{\frac{\delta}{2} T_{\mathrm{thres}}} &< \frac{8\sqrt{2}}{\sqrt{e}} \frac{{F_{\mathrm{thres}}}^{\frac{1}{2}}}{\delta^{\frac{1}{2}} \eta_p r} \\
&\leq \frac{16 {L_1}^{\frac{1}{2}}}{\sqrt{e}} \frac{M}{\delta^{\frac{1}{2}} r}.
\end{aligned}$$

where the second line is implied by

$$\begin{aligned}
\eta_p M &= \frac{2F_{\mathrm{thres}}}{\varepsilon + \sqrt{\varepsilon^2 + 2L_1 F_{\mathrm{thres}}}} \\
&\geq \frac{2F_{\mathrm{thres}}}{2\varepsilon + \sqrt{2L_1 F_{\mathrm{thres}}}} \\
&\geq \frac{1}{2} \sqrt{\frac{2F_{\mathrm{thres}}}{L_1}}.
\end{aligned}$$

Here it follows from the definition of η and the assumption $\delta^2 \geq (L_2 + L_3)\varepsilon$. Since we set $T_{\mathrm{thres}} = \frac{2}{\delta} \log \left(\frac{16 {L_1}^{\frac{1}{2}} M}{\sqrt{e} \delta^{\frac{1}{2}} r} \right)$, this leads to a contradiction. □

Proof of Proposition 5.4. From the discussion provided after Lemma D.8, by setting $M = \left(\frac{eC}{e-1}\left(1+2\log\frac{2}{\zeta'}\right)\right)^{\frac{1}{2}} = \tilde{O}(1)$, it holds that $\|\xi\|_{L^2(\mu)} \leq M = \tilde{O}(1)$ with probability $1-\frac{\zeta'}{2}$. At this point, by choosing the hyperparameters as in (30), we have $\eta_p = \tilde{O}\left(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\right)$, $T_{\mathrm{thres}} = \tilde{O}\left(\frac{1}{\delta}\right)$, $F_{\mathrm{thres}} = \tilde{O}(\delta^3)$, and Lemma E.1, Lemma E.2 can be applied. There exists $t \in [0, T_{\mathrm{thres}}]$ such that

$$
\begin{aligned}
\mathrm{P}\left(F(\mu_0) - F(\mu_t) \leq 2F_{\mathrm{thres}}\right) &\geq \mathrm{P}\left(F(\mu_0) - F(\mu_t) \leq 2F_{\mathrm{thres}}, \left|\langle \psi_0, \xi\rangle_{L^2(\mu^\dagger)}\right| \leq M\right) \\
&\geq \int_{\mathbb{R}\setminus(-\frac{\sqrt{2\pi}|\lambda_0|\zeta'}{4}, \frac{\sqrt{2\pi}|\lambda_0|\zeta'}{4})} \frac{1}{\sqrt{2\pi}|\lambda_0|} \exp\left(-\frac{x^2}{2{\lambda_0}^2}\right) \mathrm{d}x - \mathrm{P}\left(\left|\langle \psi_0, \xi\rangle_{L^2(\mu^\dagger)}\right| \geq M\right) \\
&\geq 1 - \frac{1}{\sqrt{2\pi}|\lambda_0|} \cdot 2 \cdot \frac{\sqrt{2\pi}|\lambda_0|\zeta'}{8} - \sqrt{2} e^{-\frac{M^2}{4R_1^2}} \\
&\geq 1 - \frac{\zeta'}{4} - \frac{\zeta'}{4} \\
&= 1 - \frac{\zeta'}{2},
\end{aligned}
$$

where several previously established results are used. We provide a detailed explanation below:

- The first line is straightforward.
- The second line is followed by Lemma E.2 as shown below. The key point of Lemma E.2 is that when considering two points respectively perturbed but the ψ_0-direction of perturbation differs by a fixed amount, the WGF dynamics lead to at least one point having an decrease of the objective greater than $2F_{\mathrm{thres}}$.

  Let Ω be the sample space associated with the randomness of the Gaussian process ξ, and define a random variable X as $X = \langle \psi_0, \xi\rangle_{L^2(\mu^\dagger)}$.

  Take $\omega_1, \omega_2 \in \Omega$ such that $|X(\omega_1)| \vee |X(\omega_2)| \leq M, \xi(\omega_2) = \xi(\omega_1) + r\psi_0$, where r is a real constant satisfying the assumption in Lemma E.2. We consider applying Lemma E.2 in this setting.

  Since $|X(\omega_1)| \leq M, |X(\omega_2)| \leq M$ implies $|r| = |X(\omega_1) - X(\omega_2)| \leq 2M$, from Lemma E.2, we obtain that if $|r| \geq |r_0|$, $r_0 = \frac{\sqrt{2\pi}|\lambda_0|\zeta'}{8}$, then at least one of the samples ω_1, ω_2 satisfies $F(\mu_0) - F(\mu_t) \geq 2F_{\mathrm{thres}}$. $---(\star)$

  Based on this, consider the following two cases:

  - There exists a point $x_0 \in [-M, M]$ such that $X(\omega) = x_0 \implies F(\mu_0) - F(\mu_t) < 2F_{\mathrm{thres}}$.
  - For all points $x_0 \in [-M, M]$, $X(\omega) = x_0 \implies F(\mu_0) - F(\mu_t) \geq 2F_{\mathrm{thres}}$.

  In the former case, by $(\star)$, $X \in [-M, M] \setminus (x_0 - r_0, x_0 + r_0) \implies F(\mu_0) - F(\mu_t) \geq 2F_{\mathrm{thres}}$ holds. Therefore, in either case, the following holds:

  $$
  \begin{aligned}
  &\mathrm{P}\left(F(\mu_0) - F(\mu_t) \leq 2F_{\mathrm{thres}}, \left|\langle \psi_0, \xi\rangle_{L^2(\mu^\dagger)}\right| \leq M\right) \\
  \geq &\mathrm{P}\left(X \in [-M, M] \setminus (x_0 - r_0, x_0 + r_0)\right) \\
  = &\int_{[-M,M]\setminus(x_0-r_0, x_0+r_0)} \frac{1}{\sqrt{2\pi}\lambda_0} e^{-\frac{x^2}{2\lambda_0}} dx \\
  \geq &\int_{\mathbb{R}\setminus(x_0-r_0, x_0+r_0)} \frac{1}{\sqrt{2\pi}\lambda_0} e^{-\frac{x^2}{2\lambda_0}} \mathrm{d}x - \int_{\mathbb{R}\setminus(-M,M)} \frac{1}{\sqrt{2\pi}\lambda_0} e^{-\frac{x^2}{2\lambda_0}} \mathrm{d}x
  \end{aligned}
  $$

  Here, the first term is minimized at $x_0 = 0$, which justifies the bound used in the second line above.

- The third line : The lower bound of the first term follows from the fact that the Gaussian PDF reaches its maximum $\frac{1}{\sqrt{2\pi}|\lambda_0|}$ at $x = 0$.

The lower bound of the second term comes from the Gaussian tail bound. (This follows from Proposition D.6 by taking $X = \langle \psi_0, \xi \rangle_{L^2(\mu^\dagger)}$, $\varphi(x) = \exp\left(\frac{x^2}{4\lambda_0^2}\right)$.)

- The fourth line follows from the definition of M.

Thus, with probability $1 - \frac{\zeta'}{2}$, we have

$$F(\mu_0) - F(\mu_t) \geq 2F_{\text{thres}}.$$

Finally, combining with Lemma E.1, the following holds:

$$\begin{aligned} F(\mu^\dagger) - F(\mu_{T_{\text{thres}}}) &= F(\mu\dagger) - F(\mu_0) + F(\mu_0) - F(\mu_{T_{\text{thres}}}) \\ &\geq -F_{\text{thres}} + 2F_{\text{thres}} \\ &= F_{\text{thres}}. \end{aligned}$$

This occurs with probability more than $1 - \left(\frac{\zeta'}{2} + \frac{\zeta'}{2}\right) = 1 - \zeta'$. □

E.2 Discrete-time Convergence Analysis

In this section, we prove the convergence of discrete-time PWGF (Theorem 5.5). For the discretization, the following Lipschitz continuity of the Wasserstein gradient is utilized to bridge the gap with continuous time.

Lemma E.3. *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be sufficiently smooth and suppose that the Wasserstein gradient $\nabla_\mu F$ be L_1-Lipschitz continuous.*[6] *Then, it holds that for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in \Gamma(\mu, \nu)$,*

$$F(\nu) - F(\mu) \leq \int \langle \nabla_\mu F(\mu, x), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) + \frac{L_1}{2} \int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y).$$

Proof. Let $\mu_t = ((1-t)p_1 + tp_2)\#\gamma \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma_t = (\mathrm{Id} \times (1-t)p_1 + tp_2))\#\gamma \in \Gamma(\mu, \mu_t)$. The application of Lemma C.6 implies

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) &= \int \langle \nabla_\mu F(\mu_t, (1-t)x + ty), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) \\ &= \int \langle \nabla_\mu F(\mu, x), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) \\ &\qquad + \int \langle \nabla_\mu F(\mu_t, (1-t)x + ty) - \nabla_\mu F(\mu, x), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) \\ &\leq \int \langle \nabla_\mu F(\mu, x), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) \\ &\qquad + L_1 \left(\int \|x - y\|^2 \gamma_t(\mathrm{d}x\mathrm{d}y) \right)^{\frac{1}{2}} \left(\int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y) \right)^{\frac{1}{2}} \\ &= \int \langle \nabla_\mu F(\mu, x), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) + L_1 t \int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y). \end{aligned}$$

In the third line, we use the Cauchy-Schwarz inequality and Lipschitz continuity. Then, we have

$$\begin{aligned} F(\nu) - F(\mu) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) \mathrm{d}t \\ &= \int \langle \nabla_\mu F(\mu, x), y - x \rangle \, \gamma(\mathrm{d}x\mathrm{d}y) + \frac{L_1}{2} \int \|x - y\|^2 \gamma(\mathrm{d}x\mathrm{d}y). \end{aligned}$$

□

The following proposition is a discrete version of Gronwall's inequality, which will be used in subsequent proofs.

[6] See Assumption 3.

Proposition E.4. *Let $\{a_k\}_{k\geq 0}$ be real-valued sequence and $b \neq 0,\ c \in \mathbb{R}$. Then,*

$$a_k - c \leq b\sum_{l=0}^{k-1} a_l \quad (\forall k \geq 0) \quad \Longrightarrow \quad a_k - a_0 \leq c(b+1)^k.$$

Proof. Letting $d_k \coloneqq \sum_{l=0}^{k} a_k$, we have

$$\begin{aligned} d_k + \frac{c}{b} &\leq a_k + d_{k-1} + \frac{c}{b} \leq (b+1)d_{k-1} + c + \frac{c}{b} \\ &= (b+1)(d_{k-1} + \frac{c}{b}) \leq (b+1)^{k+1}\frac{c}{b}. \end{aligned}$$

Then it holds that

$$a_k \leq c + bd_{k-1} \leq c + b\frac{c}{b}\big((b+1)^k - 1\big) = c(b+1)^k.$$

□

The following proposition corresponds to Lemma 5.3 in continuous time and serves to prove two key points: the evaluation of the decrease in F when the gradient is large, and the ability to reduce the objective function near saddle points.

Proposition E.5. *Let $\left\{\mu^{(l)}\right\}_{l=0}^{k}$ be a sequence of probability measures generated by discrete-time PWGF (Algorithm 2) with step size $\eta \leq \frac{1}{L_1}$. Then it holds that*

$$F(\mu^{(0)}) - F(\mu^{(k)}) \geq \frac{\eta}{2}\sum_{l=0}^{k-1}\left\|\nabla_\mu F(\mu^{(l)})\right\|^2_{L^2(\mu^{(l)})}.$$

Proof. By Lemma E.3, it holds that for any $l = 0, \cdots, k-1$,

$$\begin{aligned} F(\mu^{(l+1)}) - F(\mu^{(l)}) &\leq \int \left\langle \nabla_\mu F(\mu^{(l)}, x), y - x \right\rangle \Big(\mathrm{Id} \times (\mathrm{Id} - \eta\nabla_\mu F(\mu^{(l)}))\Big)\#\mu^{(l)}(\mathrm{d}x\mathrm{d}y) \\ &\quad + \frac{L_1}{2}\int \|x-y\|^2\Big(\mathrm{Id} \times (\mathrm{Id} - \eta\nabla_\mu F(\mu^{(l)}))\Big)\#\mu^{(l)}(\mathrm{d}x\mathrm{d}y) \\ &= -\eta\left(1 - \frac{L_1\eta}{2}\right)\left\|\nabla_\mu F(\mu^{(l)})\right\|^2_{L^2(\mu^{(l)})} \\ &\leq -\frac{\eta}{2}\left\|\nabla_\mu F(\mu^{(l)})\right\|^2_{L^2(\mu^{(l)})}. \end{aligned}$$

Then we have

$$\begin{aligned} F(\mu^{(k)}) - F(\mu^{(0)}) &= \sum_{l=0}^{k-1}\Big(F(\mu^{(l+1)}) - F(\mu^{(l)})\Big) \\ &\leq -\frac{\eta}{2}\sum_{l=0}^{k-1}\left\|\nabla_\mu F(\mu^{(l)})\right\|^2_{L^2(\mu^{(l)})}. \end{aligned}$$

□

In discrete-time PWGF, we take hyperparameters $\eta_p,\ F_{\mathrm{thres}},\ k_{\mathrm{thres}},\ \eta$ as follows:

$$\begin{aligned}
M &= \left(\frac{eR_1{}^2}{e-1}\left(1+2\log\frac{2}{\zeta'}\right)\right)^{\frac{1}{2}} \vee 2R_1\left(\log\frac{4\sqrt{2}}{\zeta'}\right)^{\frac{1}{2}} = \tilde{O}(1),\\
\eta &\le \frac{1}{L_1} = O(1),\\
k_{\mathrm{thres}} &= \frac{2}{\log(1+\eta\delta)}\log\left(\frac{16\sqrt{2}L_1^{\frac{1}{2}}\eta^{\frac{1}{2}}M}{\sqrt{e}r\log^{\frac{1}{2}}(1+\eta\delta)}\right) = \tilde{O}\left(\frac{1}{\delta}\right),\\
F_{\mathrm{thres}} &= \frac{\eta^{-3}k_{\mathrm{thres}}^{-3}}{18(L_2+L_3)^2}\log^2\frac{3}{2} = \tilde{O}(\delta^3),\\
\eta_p &= \frac{2F_{\mathrm{thres}}}{M(\varepsilon+\sqrt{\varepsilon^2+2L_1F_{\mathrm{thres}}})} = \tilde{O}\left(\frac{\delta^3}{\varepsilon}\wedge\delta^{\frac{3}{2}}\right).
\end{aligned} \tag{38}$$

Since η_p and M are set in the same way as continuous time (30), Lemma E.1 holds in discrete time as well. It should also be noted that F_{thres} is defined to correspond to T_{thres} and ηk_{thres}.

Next, we present the discrete-time counterpart of Lemma E.2.

Proposition E.6. *Let $\mu^\dagger$ be an (ε,δ)-saddle point, $\mu^{(0)} := (\mathrm{Id}+\eta_p\xi)\#\mu^\dagger$, and the hyperparameters $\eta_p, F_{\mathrm{thres}}$ be taken as in (38). Furthermore, we set $\tilde{\mu}^{(0)} = (\mathrm{Id}+\eta_p\tilde{\xi})\sharp\mu^\dagger$, where $\tilde{\xi} = \xi + r\psi_0$, $r\in\mathbb{R}$ is a constant, and ψ_0 is the eigenvector of the Hessian operator H_μ corresponding to the smallest eigenvalue λ_0.*
Letting $\mu^{(k)}t,\ \tilde{\mu}^{(k)}$ be WGF initialized at $\mu^{(0)},\tilde{\mu}^{(0)}$ respectively, then $\|\xi\|_{L^2(\mu^\dagger)} \le M$ and $\frac{\sqrt{2\pi}|\lambda_0|\zeta'}{4} \le |r| \le 2M$ implies that there exists $k = 0,\cdots,k_{\mathrm{thres}}]$ satisfying

$$(F(\mu^{(0)}) - F(\mu^{(k)})) \vee (F(\tilde{\mu}^{(0)}) - F(\tilde{\mu}^{(k)})) \ge 2F_{\mathrm{thres}}.$$

Proof. We give a proof by contradiction. Assume that for any $0\le k\le k_{\mathrm{thres}}$,

$$(F(\mu^{(0)}) - F(\mu^{(k)})) \vee (F(\tilde{\mu}^{(0)}) - F(\tilde{\mu}^{(k)})) \vee \Delta_F(\tilde{\mu}^{(0)},\tilde{\mu}^{(k)}) < 2F_{\mathrm{thres}}.$$

The following vector fields are used to evaluate how $\Delta_F(\mu^{(0)},\mu^{(k)})$ and $\Delta_F(\tilde{\mu}^{(0)},\tilde{\mu}^{(k)})$ evolve:

$$\begin{aligned}
Y^{(k)} &= (\mathrm{Id}-\eta\nabla_\mu F(\mu^{(k-1)}))\circ\cdots\circ(\mathrm{Id}-\eta\nabla_\mu F(\mu^{(0)}))\circ(\mathrm{Id}+\eta_p\xi),\\
\tilde{Y}^{(k)} &= (\mathrm{Id}-\eta\nabla_\mu F(\tilde{\mu}^{(k-1)}))\circ\cdots\circ(\mathrm{Id}-\eta\nabla_\mu F(\tilde{\mu}^{(0)}))\circ(\mathrm{Id}+\eta_p\tilde{\xi}).
\end{aligned}$$

Here, we note that

$$\begin{aligned}
\left\|Y^{(k)}-\mathrm{Id}\right\|_{L^2(\mu^{(0)})} &\le \eta_p\|\xi\|_{L^2(\mu^\dagger)} + \sum_{l=0}^{k-1}\left\|Y^{(l+1)}-Y^{(l)}\right\|_{L^2(\mu^{(l)})}\\
&\le \eta_p\|\xi\|_{L^2(\mu^\dagger)} + \eta\sum_{l=0}^{k-1}\left\|\nabla_\mu F(\mu^{(l)})\right\|_{L^2(\mu^{(l)})}\\
&= \eta_p\|\xi\|_{L^2(\mu^\dagger)} + \eta k\sum_{l=0}^{k-1}\frac{1}{k}\left\|\nabla_\mu F(\mu^{(l)})\right\|_{L^2(\mu^{(l)})}\\
&\le \eta_p\|\xi\|_{L^2(\mu^\dagger)} + \eta k\left(\sum_{l=0}^{k-1}\frac{1}{k}\left\|\nabla_\mu F(\mu^{(l)})\right\|^2_{L^2(\mu^{(l)})}\right)^{\frac{1}{2}}\\
&= \eta_p\|\xi\|_{L^2(\mu^\dagger)} + \sqrt{2}\eta^{\frac{1}{2}}k^{\frac{1}{2}}\left(\frac{\eta}{2}\sum_{l=0}^{k-1}\left\|\nabla_\mu F(\mu^{(l)})\right\|^2_{L^2(\mu^{(l)})}\right)^{\frac{1}{2}}\\
&\le \eta_p\|\xi\|_{L^2(\mu^\dagger)} + \sqrt{2}\eta^{\frac{1}{2}}k^{\frac{1}{2}}(F(\mu^{(0)}) - F(\mu^{(k)})^{\frac{1}{2}}\\
&\le \eta_p M + 2\eta^{\frac{1}{2}}k_{\mathrm{thres}}^{\frac{1}{2}}F_{\mathrm{thres}}^{\frac{1}{2}}.
\end{aligned} \tag{39}$$

The third line is due to Jensen's inequality and the fifth line is due to Proposition E.5. Similarly,

$$\left\|\tilde{Y}^{(k)} - \mathrm{Id}\right\|_{L^2(\mu^\dagger)} \leq \eta_p M + \eta_p r + 2\eta^{\frac{1}{2}} k_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}}. \tag{40}$$

We demonstrate that either $\mu^{(k)}$ or $\tilde{\mu}^{(k)}$ significantly decreases the objective function. To this end, we define the following vector field as a measure of the "distance" between $\mu^{(k)}$ and $\tilde{\mu}^{(k)}$:

$$w^{(k)} := \tilde{Y}^{(k)} - Y^{(k)}.$$

$w^{(k)}$ follows the recurrence relation given by the following:

$$\begin{aligned} w^{(k+1)} - w^{(k)} &= -\eta \nabla_\mu F(\tilde{\mu}^{(k)}) \circ \tilde{Y}^{(k)} + \eta \nabla_\mu F(\mu^{(k)}) \circ Y^{(k)} \\ &= -\eta \int_0^1 \frac{\mathrm{d}}{\mathrm{d}h} \Big(\nabla_\mu F(\nu_h) \circ Y_h^{(k)} \Big) \mathrm{d}h \\ &= -\eta H_{\mu^\dagger} w^{(k)} - \eta \Delta^{(k)} w^{(k)}, \end{aligned} \tag{41}$$

where we set $Y_h^{(k)} = (1-h)Y^{(k)} + h\tilde{Y}^{(k)}$, $\nu_h = Y_h^{(k)} \# \mu^\dagger = ((1-h)Y^{(k)} + h\tilde{Y}^{(k)}) \# \mu^\dagger$ and

$$\begin{aligned} \Delta^{(k)} &= \int_0^1 \Delta_h^{(k)} \mathrm{d}h, \\ \Delta_h^{(k)} f(x) &= \int \big(\nabla_\mu^2 F(\nu_h, Y_h(x), Y_h(y)) - \nabla_\mu^2 F(\mu^\dagger, x, y) \big) f(y) \mu^\dagger(\mathrm{d}y) \\ &\quad + (\nabla \nabla_\mu F(\nu_h, Y_h(x)) - \nabla \nabla_\mu F(\mu^\dagger, x)) f(x) \\ &\quad + \nabla \nabla_\mu F(\mu^\dagger, x) f(x). \end{aligned}$$

The recurrence formula (41) yields

$$\begin{aligned} w^{(k)} &= (1 - \eta H_{\mu^\dagger})^k w^{(0)} - \eta \sum_{l=0}^{k-1} (1 - \eta H_{\mu^\dagger})^{k-l-1} \Delta^{(l)} w^{(l)} \\ &= (1 - \eta \lambda_0)^k w^{(0)} - \eta \sum_{l=0}^{k-1} (1 - \eta H_{\mu^\dagger})^{k-l-1} \Delta^{(l)} w^{(l)}. \end{aligned}$$

Here, we use the fact that $w^{(0)} = \tilde{Y}^{(0)} - Y^{(0)} = \eta_p r \psi_0$. Then, we have

$$\begin{aligned} \left| \left\| w^{(k)} \right\|_{L^2(\mu^\dagger)} - (1 - \eta\lambda_0)^k \eta r \right| &\leq \left\| w^{(k)} - (1 - \eta\lambda_0)^k w^{(0)} \right\|_{L^2(\mu^\dagger)} \\ &\leq \eta \sum_{l=0}^{k-1} \left\| 1 - \eta H_{\mu^\dagger} \right\|_{L^2(\mu^\dagger)}^{k-l-1} \left\| \Delta^{(l)} \right\|_{L^2(\mu^\dagger)} \left\| w^{(l)} \right\|_{L^2(\mu^\dagger)} \\ &\leq \eta \Delta \sum_{l=0}^{k-1} (1 - \eta\lambda_0)^{k-l-1} \left\| w^{(l)} \right\|_{L^2(\mu^\dagger)}, \\ \therefore \left| (1 - \eta\lambda_0)^{-k} \left\| w^{(k)} \right\|_{L^2(\mu^\dagger)} - \eta_p r \right| &\leq \eta \Delta \sum_{l=0}^{k-1} (1 - \eta\lambda_0)^{-l-1} \left\| w^{(l)} \right\|_{L^2(\mu^\dagger)}, \end{aligned}$$

where the constant Δ upper bounds the norm of $\Delta^{(k)}$ and set in the same manner as in continuous time:

$$\Delta := (L_2 + L_3)(2\eta^{\frac{1}{2}} k_{\mathrm{thres}}^{\frac{1}{2}} F_{\mathrm{thres}}^{\frac{1}{2}} + 2\eta_p M) + R_2 \varepsilon \geq \left\| \Delta^{(k)} \right\|_{L^2(\mu^\dagger)}.$$

Using the discrete version of Gronwall's inequality (Proposition E.4) with $a_k = (1 - \eta\lambda_0)^{-k} \left\| w^{(k)} \right\|_{L^2(\mu^\dagger)}$, $b = \frac{\eta\Delta}{1-\eta\lambda_0}$, $c = \eta_p r$, we obtain the following:

$$(1 - \eta\lambda_0)^{-l} \left\| w^{(l)} \right\|_{L^2(\mu^\dagger)} \leq \eta_p r \left(1 + \frac{\eta\Delta}{1 - \eta\lambda_0} \right)^l.$$

From this, we have

$$\begin{aligned}
(1-\eta\lambda_0)^{-k}\left\|w^{(k)}\right\| &\geq \eta_p r - \eta\Delta\sum_{l=0}^{k-1}(1-\eta\lambda_0)^{-l-1}\left\|w^{(l)}\right\|_{L^2(\mu^\dagger)} \\
&\geq \eta_p r\left(1-\frac{\eta\Delta}{1-\eta\lambda_0}\sum_{l=0}^{k-1}\left(1+\frac{\eta\Delta}{1-\eta\lambda_0}\right)^l\right) \\
&= \eta_p r\left(1-\frac{\eta\Delta}{1-\eta\lambda_0}\frac{\left(1+\frac{\eta\Delta}{1-\eta\lambda_0}\right)^k-1}{\frac{\eta\Delta}{1-\eta\lambda_0}}\right) \\
&= \eta_p r\left(2-\left(1+\frac{\eta\Delta}{1-\eta\lambda_0}\right)^k\right) \\
&\geq \eta_p r\Big(2-(1+\eta\Delta)^k\Big) \\
&\geq \eta_p r(2-\exp(\eta k_{\mathrm{thres}}\Delta)) \\
&\geq \frac{\eta_p r}{2},
\end{aligned}$$

where the last line holds as $\eta k_{\mathrm{thres}}\Delta \leq \log\frac{3}{2}$ in the same manner as in continuous time,

$$\begin{aligned}
\eta k_{\mathrm{thres}}\Delta &\leq \eta k_{\mathrm{thres}}\Big((L_2+L_3)(2\eta^{\frac{1}{2}}k_{\mathrm{thres}}^{\frac{1}{2}}F_{\mathrm{thres}}^{\frac{1}{2}}+2\eta_p M)+R_2\varepsilon\Big) \\
&\leq 2(L_2+L_3)(\eta k_{\mathrm{thres}})^{\frac{3}{2}}F_{\mathrm{thres}}^{\frac{1}{2}}+2(L_2+L_3)\eta_p M\eta k_{\mathrm{thres}}+R_2\eta k_{\mathrm{thres}}\varepsilon \\
&\leq \frac{1}{3}\log\frac{3}{2}\cdot 3 = \log\frac{3}{2}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\left\|w^{(k)}\right\| &\geq \frac{\eta_p r}{2}(1-\eta\lambda_0)^k \\
&\geq \frac{\eta_p r}{2}(1+\eta\delta)^k \\
&\geq \frac{\eta_p r}{2}\sqrt{e}k^{\frac{1}{2}}(1+\eta\delta)^{\frac{k}{2}}\log^{\frac{1}{2}}(1+\eta\delta). \qquad (42)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\left\|w^{(k)}\right\| &\leq \left\|\tilde{Y}^{(k)}-\mathrm{Id}\right\|_{L^2(\mu^\dagger)}+\left\|Y^{(k)}-\mathrm{Id}\right\|_{L^2(\mu^\dagger)} \\
&< 4\eta^{\frac{1}{2}}k_{\mathrm{thres}}^{\frac{1}{2}}F_{\mathrm{thres}}^{\frac{1}{2}}+2\eta_p M+\eta_p r \\
&\leq 8\eta^{\frac{1}{2}}k_{\mathrm{thres}}^{\frac{1}{2}}F_{\mathrm{thres}}^{\frac{1}{2}}, \qquad (43)
\end{aligned}$$

where in the second line (39) and (40) are used, and in the third line $\eta^{\frac{1}{2}}k_{\mathrm{thres}}{}^{\frac{1}{2}}F_{\mathrm{thres}}{}^{\frac{1}{2}} = \tilde{O}(\delta)$, $\eta_p M = o(\delta)$, $\eta_p r = o(\delta)$. Letting $k = k_{\mathrm{thres}}$, it follows from (42), (43) and $\eta_p M \geq \frac{1}{2}\sqrt{\frac{2F_{\mathrm{thres}}}{L_1}}$ that

$$\begin{aligned}
(1+\eta\delta)^{\frac{k_{\mathrm{thres}}}{2}} &< \frac{16\eta^{\frac{1}{2}}F_{\mathrm{thres}}^{\frac{1}{2}}}{\sqrt{e}\eta_p r\log^{\frac{1}{2}}(1+\eta\delta)} \\
&\leq \frac{16\sqrt{2}L_1^{\frac{1}{2}}\eta^{\frac{1}{2}}M}{\sqrt{e}r\log^{\frac{1}{2}}(1+\eta\delta)}.
\end{aligned}$$

This leads to a contradiction, as we had set k_{thres} as in (38). □

The following proposition corresponds to the discrete-time version of Proposition 5.4.

Proposition E.7. *Let $\varepsilon,\ \delta,\ \zeta' > 0$ be chosen such that $(L_2 + L_3)\,\varepsilon \le \delta^2$. Suppose $\mu^\dagger \in \mathcal{P}_2^a(\mathbb{R}^d)$ satisfies $\left\|\nabla_\mu F(\mu^\dagger)\right\|_{L^2(\mu^\dagger)} < \varepsilon$ and $\lambda_0 := \lambda_{\min} H_{\mu^\dagger} \le -\delta$. x Considering $\xi \sim \mathrm{GP}(0, k_\mu)$ and setting $\mu_0 = (\mathrm{Id} + \eta_p \xi)\sharp\mu^\dagger$ as the initial value of the discrete time PWGF flow $\mu^{(k)}$, with parameters $\eta = O(1),\ \eta_p = \tilde{O}\Big(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\Big),\ k_{\mathrm{thres}} = \tilde{O}\big(\frac{1}{\delta}\big)$, and $F_{\mathrm{thres}} = \tilde{O}(\delta^3)$, the following holds with probability $1 - \zeta'$:*

$$F(\mu^\dagger) - F(\mu^{(k_{\mathrm{thres}})}) \ge F_{\mathrm{thres}}.$$

Proof. From the discussion at the beginning of the previous section, by setting $M \le \left(\frac{eC}{e-1}\left(1 + 2\log\frac{2}{\zeta'}\right)\right)^{\frac{1}{2}} = \tilde{O}(1)$, it holds that $\|\xi\|_{L^2(\mu)} \le M = \tilde{O}(1)$ with probability $1 - \frac{\zeta'}{2}$. By choosing the hyperparameters as in (38), we have $\eta = O(1),\ \eta_p = \tilde{O}\Big(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\Big),\ k_{\mathrm{thres}} = \tilde{O}\big(\frac{1}{\delta}\big),\ F_{\mathrm{thres}} = \tilde{O}(\delta^3)$, and Lemma E.1 and Proposition E.6 can be applied. In a similar manner to the proof of Proposirion 5.4 (continuous version), it follows that there exists $0 \le k \le k_{\mathrm{thres}}$ such that

$$\mathrm{P}\left(F(\mu^{(0)}) - F(\mu^{(k)}) \ge 2F_{\mathrm{thres}}\right) \ge 1 - \frac{\zeta'}{2},$$

Thus, with probability $1 - \frac{\zeta'}{2}$, we have

$$F(\mu^{(0)}) - F(\mu^{(k_{\mathrm{thres}})}) \ge F(\mu^{(0)}) - F(\mu^{(k)}) \ge 2F_{\mathrm{thres}}.$$

Combining with Lemma E.1, the following holds:

$$\begin{aligned} F(\mu^\dagger) - F(\mu^{(k_{\mathrm{thres}})}) &= F(\mu\dagger) - F(\mu^{(0)}) + F(\mu^{(0)}) - F(\mu^{(k_{\mathrm{thres}})}) \\ &\ge -F_{\mathrm{thres}} + 2F_{\mathrm{thres}} \\ &= F_{\mathrm{thres}}. \end{aligned}$$

This occurs with probability more than $1 - \left(\frac{\zeta'}{2} + \frac{\zeta'}{2}\right) = 1 - \zeta'$. □

With the above preparations, we finally prove the convergence of discrete-time PWGF to a second-order stationary point.

Theorem 5.5. *Let the initial point be $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and denote $\Delta F = F(\mu_0) - \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$. Set $\eta = O(1)$ and let $\varepsilon,\ \delta,\ \eta_p,\ T_{\mathrm{thres}},\ F_{\mathrm{thres}}$ be chosen as in Theorem 5.2. Then, discrete time PWGF halts after*

$$k = \tilde{O}\left(\Delta F\left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4}\right)\right)$$

steps and reaches an (ε, δ)-second-order stationary point with probability $1 - \zeta$.

Proof. Let $\varepsilon,\ \delta,\ \zeta > 0$ be chosen arbitrarily chosen such that $(L_2 + L_3)\,\varepsilon \le \delta^2$, and set $\zeta' > 0$ such that ζ' is polynomial in $\frac{1}{\delta}$ and ζ up to logarithmic factors, provided later. By the settings of η, η_p, k_{thres}, and F_{thres} as in Proposition E.7, we have $\eta = O(1)$, $\eta_p = \tilde{O}\Big(\delta^{\frac{3}{2}} \wedge \frac{\delta^3}{\varepsilon}\Big)$, $k_{\mathrm{thres}} = \tilde{O}\big(\frac{1}{\delta}\big)$, and $F_{\mathrm{thres}} = \tilde{O}(\delta^3)$.
From Proposition E.7, perturbations occur at most $m := \lceil \frac{\Delta F}{F_{\mathrm{thres}}} \rceil$ times. Thus, the probability of failure after m perturbations is at most $1 - (1 - \zeta')^m \le m\zeta'$. Setting $\zeta' = \frac{\zeta}{m}$ ensures that the algorithm reaches an (ε, δ)-second order stationary point with probability at least $1 - \zeta$.

Discrete time PWGF determines whether the objective decreases by at least F_{thres} after a certain number of iterations k_{thres} following the application of a perturbation. Then, we define the period between the application of a perturbation and this evaluation as *State* 1, and all other times as *State* 0. Let k_0 denote the total time spent in State 0, where $\|\nabla_\mu F(\mu)\|_{L^2(\mu)} \ge \varepsilon$. By Lemma 5.3, the decrease in the objective function during this time is at least $\varepsilon^2 k_0$, implying $k_0 \le \frac{\Delta F}{\varepsilon^2}$. Moreover, the total time k_1 in State 1 is upper bounded by $k_1 \le m T_{\mathrm{thres}} = \frac{\Delta F k_{\mathrm{thres}}}{F_{\mathrm{thres}}} = \tilde{O}\big(\frac{1}{\delta^4}\big)$. Hence, the algorithm halts in $k_0 + k_1 = \tilde{O}\big(\Delta F\big(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4}\big)\big)$ iterations. □

F Numerical Experiments

F.1 ICFL Functional

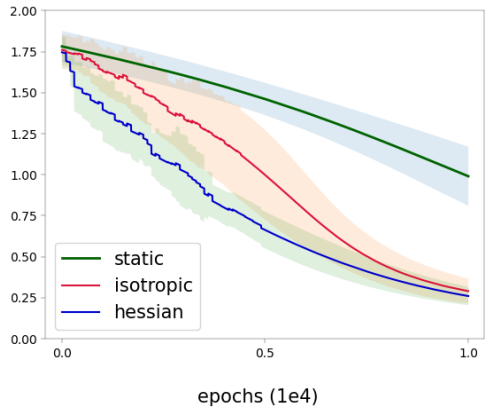

Figure 1: Trajectories of the training loss for no-noise ("static"), isotropic noise ("isotropic") and Hessian guided noise ("hessian") settings.

We conducted numerical experiments using the loss functional of in-context learning of Transformers from Kim & Suzuki (2024) as the objective functional. See below for details.

We compared the dynamics of the loss function under three variants of WGF; WGF without noise (static), WGF with isotropic noise (isotropic), and WGF with noise guided by the Hessian (hessian).

As Figure 1 shows, the loss decreases gradually in the "static" case, whereas the "isotropic" and "hessian" cases exhibit a significant reduction in loss, leading to saturation. Furthermore, the Hessian-based noise demonstrates a more efficient decrease in loss.

Experimental details. We provide a minimal explanation of the loss function used in our numerical experiments for in-context learning in Transformers. For an in-depth exposition on the problem setup, derivation of the loss functional, and an analysis of the loss landscape, we refer to the work by Kim & Suzuki (2024).

We consider a mean field two-layer neural network with a sigmoid activation function, which takes l-dimensional data inputs and k-dimensional data outputs :

$$h_\mu(z) = \int h_x(z)\mu(\mathrm{d}x) = \int a\sigma(w^\top z)\mu(\mathrm{d}x) \quad (x = (a, w) \in \mathbb{R}^k \times \mathbb{R}^l),$$

where $z \in \mathbb{R}^l$ is a given data and follows a certain distribution. We also define the following matrices :

$$\Sigma_{\mu,\nu} = \mathrm{E}_z\left[h_\mu(z)h_\nu(z)^\top\right] \quad (\mu, \nu \in \mathcal{P}_2(\mathbb{R}^{k+l})).$$

We consider performing in-context learning using h_μ as feedforward layer, followed by a reparametrized linear self-attention mechanism which can be described by a single attention matrix $W \in \mathbb{R}^{k\times k}$. The optimal value of W is determined to satisfy $\Sigma_{\mu^o,\mu}W = \Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1}$ with given μ. Thus, the objective with optimal W is derived as follows:

$$F(\mu) = \frac{1}{2}\mathrm{E}\left[\left\|h_{\mu^o}(z) - \Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1}h_\mu(z)\right\|^2\right], \tag{44}$$

where $\mu^o \in \mathcal{P}_2(\mathbb{R}^{k+l})$ is the true feature and $\zeta_{\mu^o,\mu}(z) = h_{\mu^o}(z) - \Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1}h_\mu(z)$. The Wasserstein gradient of the objective (44) is computed as

$$\nabla_\mu F(\mu, a, w) = \begin{pmatrix} -\Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\mathrm{E}_z\left[\sigma(w^\top z)\zeta_{\mu^o,\mu}(z)\right] \\ a^\top\Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\mathrm{E}_z\left[\sigma'(w^\top z)\zeta_{\mu^o,\mu}(z)z^\top\right] \end{pmatrix}.$$

Algorithm 3 PWGD (time/ space discrete)

initialize $x_1^{(0)}, ...x_N^{(0)}$, $\mu^{(0)} \leftarrow \frac{1}{N}\sum_{j=1}^N \delta_{x_j^{(0)}}$
for $k = 0, 1, ...$ **do**
  if $\left\|\nabla_\mu F(\mu^{(k)})\right\|_{L^2(\mu^{(k)})} \leq \varepsilon$ and $k - k_{\mathrm{p}} > k_{\mathrm{thres}}$ **then**
    $\xi \sim \mathrm{GP}(0, k_{\mu^{(k)}})$
    $(\xi_1, ...\xi_N) \leftarrow (\xi(x_1^{(k)}), ..., \xi(x_N^{(k)}))$
    $x_j^{(k)} \leftarrow x_j^{(k)} + \eta_p \xi_j \quad (j = 1, ..., N)$
    $\mu^{(k)} \leftarrow \frac{1}{N}\sum_{j=1}^N \delta_{x_j^{(k)}}$
    $k_p \leftarrow k$
  end if
  if $k = k_{\mathrm{p}} + k_{\mathrm{thres}}$ and $F(\mu^{(k_{\mathrm{p}})}) - F(\mu^{(k)}) \leq F_{\mathrm{thres}}$ **then**
    return $\mu^{(k_{\mathrm{p}})}$
  end if
  $x_j^{(k+1)} \leftarrow x_j^{(k)} - \eta \nabla_\mu F(\mu)(\mu^{(k)}, x_j^{(k)}),\ (j = 1, ..., N)$, $\mu^{(k+1)} \leftarrow \frac{1}{N}\sum_{j=1}^N \delta_{x_j^{(k+1)}}$
end for

Furthermore, Hessian $\nabla_\mu^2 F$ at a first-order optimal point μ is computed as

$$\nabla_\mu^2 F(\mu, a, w, b, v) = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

where

$$\begin{aligned}
&H_{11}(\mu, a, w, b, v) \\
&= \nabla_a \nabla_b^\top \frac{\delta^2 F}{\delta \mu^2}(\mu, a, w, b, v) \\
&= \big(\mathrm{E}_z[\sigma(w^\top z)\sigma(v^\top z)] - \mathrm{E}_z[\sigma(w^\top z)h_\mu(z)]^\top \Sigma_{\mu,\mu}^{-1} \mathrm{E}_z[\sigma(v^\top z)h_\mu(z)]\big)\Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1} \\
&\quad + \mathrm{E}_z[\sigma(w^\top z)\zeta_{\mu^o,\mu}(z)]^\top \mathrm{E}_z[\sigma(v^\top z)\zeta_{\mu^o,\mu}(z)]\Sigma_{\mu,\mu}^{-1}
\end{aligned}$$

and

$$\begin{aligned}
&H_{12}(\mu, a, w, b, v) = H_{21}(\mu, b, v, a, w)^\top \\
&= \nabla_a \nabla_v^\top \frac{\delta^2 F}{\delta \mu^2}(\mu, a, w, b, v) \\
&= \Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1} b \\
&\quad \cdot \Big(\mathrm{E}_z\big[\sigma(w^\top z)\sigma'(v^\top z)z^\top\big] - \mathrm{E}_z\big[\sigma(w^\top z)h_\mu(z)\big]^\top \Sigma_{\mu,\mu}^{-1}\mathrm{E}_z\big[h_\mu(z)\sigma'(v^\top z)z^\top\big]\Big) \\
&\quad - \Sigma_{\mu,\mu}^{-1} b \mathrm{E}_z\big[\sigma(w^\top z)\zeta_{\mu^o,\mu}(z)^\top\big]\mathrm{E}_z\big[\zeta_{\mu^o,\mu}(z)\sigma'(v^\top z)z^\top\big] \\
&\quad + \big(\mathrm{E}_z\big[\sigma(w^\top z)h_\mu(z)^\top\big]\Sigma_{\mu,\mu}^{-1} b\big)\Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\mathrm{E}_z\big[\zeta_{\mu^o,\mu}(z)\sigma'(v^\top z)z^\top\big],
\end{aligned}$$

as well as

$$\begin{aligned}
&H_{22}(\mu, a, w, b, v) \\
&= \big(a^\top \Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1} b\big) \\
&\quad \cdot \big(\mathrm{E}_z\big[z\sigma'(w^\top z)\sigma'(v^\top z)z^\top\big] - \mathrm{E}_z\big[z\sigma'(w^\top z)h_\mu(z)^\top\big]\Sigma_{\mu,\mu}^{-1}\mathrm{E}_z\big[h_\mu(z)\sigma'(v^\top z)z^\top\big]\big) \\
&\quad - \big(a^\top \Sigma_{\mu,\mu}^{-1} b\big)\mathrm{E}_z\big[z\sigma'(w^\top z)\zeta_{\mu^o,\mu}(z)^\top\big]\mathrm{E}_z\big[\zeta_{\mu^o,\mu}(z)\sigma'(w^\top z)z^\top\big] \\
&\quad + \mathrm{E}_z\big[z\sigma'(w^\top z)h_\mu(z)^\top\big]\Sigma_{\mu,\mu}^{-1} b a^\top \Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^o}\mathrm{E}_z\big[\zeta_{\mu^o,\mu}(z)\sigma'(v^\top z)z^\top\big] \\
&\quad + \mathrm{E}_z\big[z\sigma'(w^\top z)\zeta_{\mu^o,\mu}(z)^\top\big]\Sigma_{\mu^o,\mu}\Sigma_{\mu,\mu}^{-1} b a^\top \Sigma_{\mu,\mu}^{-1}\mathrm{E}_z\big[h_\mu(z)\sigma'(v^\top z)z^\top\big].
\end{aligned}$$

The experimental setup is as follows. We compared the dynamics of the loss function under three variants of WGF; WGF without noise (static), WGF with isotropic noise (isotropic), and PWGF

(hessian). To ensure a fair comparison of the three algorithms, no stopping criteria were incorporated into the algorithms.

The input and output dimensions were set to $l = 20,\ k = 5$. We approximated the measure using 400 neurons and generated 800 i.i.d. input data points z from the standard normal distribution $\mathcal{N}(0, 1)$ for each coordinate. The optimization probability measure was randomly initialized and we conducted fiveive experiments under the same conditions, plotting the mean and standard deviation. We used parameters as $\eta_p = 0.015$, $k_{\text{thres}} = 100$. In addition, SGD was used in the optimization process with the learning rate $\eta = 10^{-7}$.

F.2 Matrix-Decomposition Functional

Next, we conducted experiments using the matrix decomposition setting presented in Example 1. Details of the objective function, including analytical expressions for the gradient and Hessian, as well as a proposition suggesting strict benignity of the matrix decomposition objective, are provided in Appendix G.

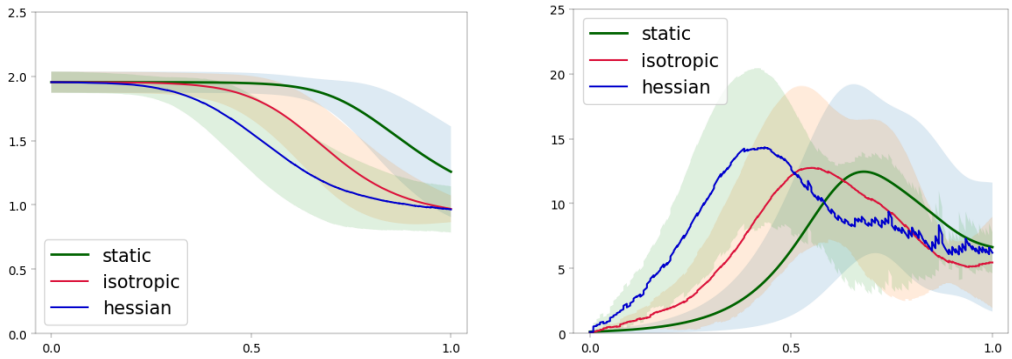

Figure 2: Trajectories of the training loss and the norm of the gradient for no-noise ("static"), isotropic noise ("isotropic") and Hessian guided noise ("hessian") settings.

Due to the stochastic nature of the algorithms, we report the mean and standard deviation over 10 runs, with the corresponding error bars. As shown in Figure 2, the Hessian and isotropic noise injection methods achieve faster objective reduction and exhibit earlier peaks in the gradient norm, demonstrating a more efficient escape from the initial critical point. In contrast, the perturbation-free method tends to stagnate for longer periods. The Hessian method shows the best performance, although the performance of isotropic noise is comparable. The effectiveness of isotropic noise can be attributed to the fact that the infinite-dimensional nature of the problem has not yet manifested due to the number of particles used in the approximation.

In practice, at points where the gradient norm is small, but the point is not a saddle, adding noise may hinder the gradient descent. This issue is particularly pronounced for the method with Hessian noise, where the magnitude of the noise depends on the local curvature. Consequently, whether to inject noise should be determined adaptively in combination with the criteria discussed above.

Experimental details. The input and output dimensions are $l = 15, k = 5$. We approximate the measure using 400 neurons and generate 800 i.i.d. input data points z from the standard normal distribution $\mathcal{N}(0, 1)$ for each coordinate. Similarly to the ICFL case, SGD was used in the optimization process and the learning rate $\eta = 10^{-6}$. We set $k_{\text{thres}} = 100,\ \eta_p = 3 \times 10^{-3},\ F_{\text{thres}} = 10^{-2}$.

G Landscape Analysis of Matrix Decomposition

We analyze matrix factorization (Example 1) as an example of non-convex stochastic optimization:

$$\begin{aligned} F(\mu) &= \frac{1}{2}\mathrm{E}_z[\|h_\mu(z)h_\mu(z)^\top - h_{\mu^o}(z)h_{\mu^o}(z)^\top\|^2] \\ &= \frac{1}{2}E_z\left[\|M_\mu(z) - M_{\mu^o}(z)\|^2\right] \\ &= \frac{1}{2}E_z\left[\|\Delta M_\mu(z)\|^2\right], \end{aligned}$$

where we set

$$\begin{aligned} h_\mu(z) &= \int h_{a,w}(z)\mu(\mathrm{d}a\mathrm{d}w) \\ h_{a,w}(z) &= h_{a,w}(z) \\ &= a\sigma(w^\top z) \\ M_\mu(z) &= h_\mu(z)h_\mu(z)^\top \\ \Delta M_\mu(z) &= M_\mu(z) - M_{\mu^o}(z) \end{aligned}$$

The Wasserstein gradient is computed as:

$$\begin{aligned} \nabla_\mu F(\mu, a_1, a_2, w) &= 2\mathrm{E}_z\left[\nabla_{a,w}h_{a,w}(z)\Delta M_\mu(z)h_\mu(z)\right] \\ &= \begin{pmatrix} 2\mathrm{E}_z\left[\sigma(w^\top z)\Delta M_\mu(z)h_\mu(z)\right] \\ 2\mathrm{E}_z\left[z\sigma'(w^\top z)h_\mu(z)^\top\Delta M_\mu(z)\right]a_1 \end{pmatrix} \end{aligned} \tag{45}$$

The Hessian is computed as:

$$\nabla_\mu^2 F(\mu, x, y) = \nabla_\mu^2 F(\mu, a, w, b, v) \tag{46}$$

$$= \mathrm{E}_z\left[\nabla_x h_x(z)\left(2M_\mu(z) + \|h_\mu(z)\|^2 I_k - M_{\mu^o}(z)\right)\left(\nabla_y h_y(z)\right)^\top\right] \tag{47}$$

From (45) and (47),

- For any $\mu \in \mathcal{P}_2(\mathbb{R}^l)$, $\mu = \delta_0 \otimes \tilde{\mu}$ is a strict saddle point.
- $\mu = (\pm\mathrm{Id}_{\mathbb{R}^k}) \times (\mathrm{Id}_{\mathbb{R}^l})\#\mu^o$ is a global optima.

Furthermore, by a similar argument to Ge et al. (2017), we can deduce the following proposition. This proposition asserts that for an ε-stationary point μ which is not a global minimizer, the objective function can be strictly decreased. This suggests that F possesses strict benignity.

Proposition G.1. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be ε-stationary and not a global optima, i.e; satisfy $\|\nabla_\mu F(\mu)\| \leq \varepsilon$ and $F(\mu) \neq 0$. If $h_\mu(z) \geq 0$ a.s. [7], Assumption 2 and $W_2(\mu, \mu^o) \leq C$ hold, then there exists a curve μ_t s.t. $\mu_0 = \mu$ and at $t = 0$,*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t) \leq -\mathrm{E}_z[\left\|\left(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\right)(z)\right\|_F^2] + O(\varepsilon).$$

Proof. We define a curve μ_t by $\mu_t = (1-t)\mu + t\mu^o$. Then we obtain at $t = 0$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) &= \mathrm{E}_z\left[\mathrm{tr}\left(\left(\frac{\mathrm{d}}{\mathrm{d}t}(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top)\left(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\right)\right)(z)\right)\right] \\ &= \mathrm{E}_z\left[\mathrm{tr}\left(\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}h_{\mu_t}^\top + h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right)\left(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\right)(z)\right)\right], \end{aligned} \tag{48}$$

[7] This condition can be regarded as an extension of non-negative matrix factorization.

$$
\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t) &= \mathrm{E}_z\Big[\mathrm{tr}\Big(\Big(\frac{\mathrm{d}^2}{\mathrm{d}t^2}h_{\mu_t}h_{\mu_t}^\top + h_{\mu_t}\frac{\mathrm{d}^2}{\mathrm{d}t^2}h_{\mu_t}^\top\Big)\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\Big)\Big] \\
&\quad + \mathrm{E}_z\left[\mathrm{tr}\left(\left(2\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right)^\top\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\right)\right] \\
&\quad + \mathrm{E}_z\left[\mathrm{tr}\left(\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}h_{\mu_t}^\top + h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right)^\top\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}h_{\mu_t}^\top + h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right)(z)\right)\right] \\
&= \mathrm{E}_z\Big[(4-3)\mathrm{tr}\Big(\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)^\top\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\Big)\Big] \\
&\quad + 4\mathrm{E}_z\left[\mathrm{tr}\left(\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right)^\top\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\right)\right] \\
&\quad + \mathrm{E}_z\left[\mathrm{tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top(z)\right)\right] \\
&= \mathrm{E}_z\left[\mathrm{tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top(z)\right)\right] \\
&\quad - 3\mathrm{E}_z\Big[\mathrm{tr}\Big(\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)^\top\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\Big)\Big] \\
&\quad - 4\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t),
\end{aligned}
\tag{49}
$$

where we used, in the sixth line, the equations;

$$
\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2}\bigg|_{t=0} h_{\mu_t}(z) &= 0, \\
\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} h_{\mu_t}h_{\mu_t}^\top(z) + h_{\mu_t}\ \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} h_{\mu_t}^\top(z) &= (h_\mu - h_{\mu^o})h_\mu^\top(z) + h_\mu(h_\mu - h_{\mu^o})^\top(z) \\
&= -(h_\mu h_\mu^\top - h_{\mu^o}h_{\mu^o})(z) - (h_\mu - h_{\mu^o})(h_\mu - h_{\mu^o})^\top(z) \\
&= -(h_\mu h_\mu^\top - h_{\mu^o}h_{\mu^o})(z) - \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} h_{\mu_t}\ \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} h_{\mu_t}^\top(z),
\end{aligned}
$$

and in the ninth line;

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t)\bigg|_{t=0} &= -\mathrm{E}_z\big[\mathrm{tr}\big(\big(h_\mu h_\mu^\top - h_{\mu^o}h_{\mu^o}^\top\big)\big(h_\mu h_\mu^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\big)\big] \\
&\quad -\mathrm{E}_z\left[\mathrm{tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\big(h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big)(z)\right)\right]\bigg|_{t=0},
\end{aligned}
$$

which are derived from the definition of μ_t.
Noting that

$$
\left\|\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right\|_F^2(z)\bigg|_{t=0} \le 2\big\|h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big\|_F^2(z),
\tag{50}
$$

which is obtained by straightforward calculation and the assumption $h_\mu \ge 0$ a.s.;

$$
\begin{aligned}
&2\big\|h_{\mu_t}h_{\mu_t}^\top - h_{\mu^o}h_{\mu^o}^\top\big\|_F^2(z) - \left\|\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}\frac{\mathrm{d}}{\mathrm{d}t}h_{\mu_t}^\top\right\|_F^2(z)\bigg|_{t=0} \\
&= 2\big\|h_\mu h_\mu^\top - h_{\mu^o}h_{\mu^o}^\top\big\|_F^2(z) - \big\|(h_{\mu^o} - h_\mu)(h_{\mu^o} - h_\mu)^\top\big\|_F^2(z) \\
&= \|h_\mu\|^4(z) + \|h_{\mu^o}\|^4(z) - 2|h_\mu^\top h_{\mu^o}|^2(z) - \|h_\mu - h_{\mu^o}\|^4(z) \\
&= \Big(\|h_\mu\|^2(z) - \|h_{\mu^o}\|^2(z)\Big)^2 + 4h_\mu^\top h_{\mu^o}\|h_\mu - h_{\mu^o}\|^2(z) \\
&\ge 0
\end{aligned}
$$

Then we obtain, from (49) and (50),

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} F(\mu_t) \le -\mathrm{E}\Big[\big\|h_\mu h_\mu^\top - h_{\mu^o} h_{\mu^o}^\top\big\|_F^2(z)\Big] - 4\,\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t). \tag{51}$$

Finally we will show that

$$\begin{aligned}\Big|\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t)\Big| &= 2\big|\mathrm{E}_z\big[(h_{\mu^o} - h_\mu)^\top \delta M_\mu(z) h_\mu(z)\big]\big| \\ &\le \tilde{C}\|\nabla_\mu F(\mu)\|_{L^2(\mu)} = O(\varepsilon)\end{aligned} \tag{52}$$

for some $\tilde{C} > 0$. For any $\gamma \in \Gamma_o(\mu, \mu^o)$,

$$\begin{aligned}&\Big|\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t)\Big| \\ &= 2\big|\mathrm{E}_z\big[(h_{\mu^o} - h_\mu)^\top \Delta M_\mu(z)) h_\mu(z)\big]\big| \\ &= 2\int \mathrm{E}_z[(h_x(z) - h_y(z))^\top \Delta M_\mu(z) h_\mu(z)]\gamma(\mathrm{d}x\mathrm{d}y) \\ &= 2\int (x-y)^\top \mathrm{E}_z[\nabla h_{x+\theta(y-x)}(z)^\top \Delta M_\mu(z) h_\mu(z)]\gamma(\mathrm{d}x\mathrm{d}y) \\ &= 2\int \Big((x-y)^\top \nabla_\mu F(\mu, x) + (x-y)^\top \nabla\nabla_\mu F(\mu, x + \tilde{\theta}(y-x))(x-y)\Big)\gamma(\mathrm{d}x\mathrm{d}y) \\ &\le 2W_2(\mu,\mu^o)\|\nabla_\mu F(\mu)\|_{L^2(\mu)} + W_2(\mu,\mu^o)^2 \sup_x \|\nabla\nabla_\mu F(\mu, x)\| \\ &\le 2(C + R_2 C^2)\|\nabla_\mu F(\mu)\|_{L^2(\mu)}\end{aligned}$$

where Taylor's expansion is used and $\theta, \tilde{\theta} \in [0, 1]$ in the third and fourth lines, the Cauchy–Schwarz inequality and the definition of the 2-Wasserstein distance in the fifth line, and the assumptions $W_2(\mu, \mu^o) \le C$ and Assumption 2 in the sixth line. Setting $\tilde{C} = 2(C + R_2 C^2)$, we obtain (52) and hence, from (51),

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} F(\mu_t) \le -\mathrm{E}\Big[\big\|h_\mu h_\mu^\top - h_{\mu^o} h_{\mu^o}^\top\big\|_F^2(z)\Big] + O(\varepsilon).$$

□