WASSERSTEIN SADDLE-FREE NEWTON: SADDLE ESCAPE AND LOCAL SPECTRAL CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Extending saddle-free Newton methods to Wasserstein space requires controlling Hessian dependent operators along nonlinear transport trajectories and accounting for a Hessian spectrum that may accumulate at zero. We address these challenges through Wasserstein Saddle-Free Newton (WSFN), a second-order method for nonconvex optimization over Wasserstein space that preconditions the Wasserstein gradient by a regularized inverse squared Hessian operator. The resulting dynamics exploit curvature magnitude without reversing the repulsive effect of negative curvature, while retaining attraction along positive curvature directions, thereby avoiding a fundamental limitation of standard Wasserstein Newton dynamics. Whereas previous Wasserstein saddle escape results rely on perturbed first-order dynamics, WSFN uses curvature both to modify the deterministic transport and to construct the perturbation mechanism. Our analysis combines stability estimates for Hessian operators with a perturbative saddle escape argument to establish polynomial time convergence to approximate second-order stationarity under regularity of landscape assumptions. The contribution of the saddle curvature parameter δ improves from $\tilde{O}(\delta^{-4})$ for prior perturbed first-order Wasserstein methods to $\tilde{O}(\delta^{-3})$. For benign landscapes, this further implies proximity to the set of global minimizers. The cubic dependence corresponds to the scale of the objective decrease that can be guaranteed from negative curvature under Lipschitz Hessian regularity. Near a minimizer, the possible absence of a uniform spectral gap prevents a uniform contraction result and instead calls for a spectral characterization. We prove geometric contraction of the linearized dynamics on positive curvature spectral bands bounded away from zero and polynomial decay under spectral source conditions. Together, these results connect quantitative saddle escape with the local spectral behavior of second-order Wasserstein dynamics.

1 INTRODUCTION

We study the minimization of a non-convex functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ over Wasserstein space, where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the probability measures on $\mathbb{R}^d$ with finite second moment, equipped with the L^2-Wasserstein distance W_2. Throughout, we assume that $\inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu) > -\infty$.

1.1 NONCONVEX OPTIMIZATION ON WASSERSTEIN SPACE

Optimization over probability measures arises in variational inference (Yao & Yang, 2025; Lambert et al., 2022; Wang et al., 2022; Blei et al., 2017; Wibisono, 2018), trajectory inference in path space (Chizat et al., 2022; 2026a), generative modeling (Arbel et al., 2019; Chu et al., 2019; Huang & Malik, 2026), and mean-field neural network training (Hu et al., 2021; Nitanda et al., 2022; Chizat, 2022; Chizat & Bach, 2018; Mei et al., 2018; Sirignano & Spiliopoulos, 2020; Rotskoff & Vanden-Eijnden, 2022; Chizat et al., 2026b). In the context of these applications, much of the convergence theory for Wasserstein gradient flows and their discretizations exploits linear or geodesic convexity. For nonconvex objectives, however, ε-small gradient norm cannot distinguish minimizers from saddles or maxima. Some learning and discrepancy objectives have benign landscapes (Lu et al., 2020; Kim & Suzuki, 2024; Boufadene & Vialard, 2025), in which every local minimizer is global. This motivates methods combining descent with saddle escape, although benignity alone provides neither an escape time bound nor a local convergence rate. Our global analysis first addresses saddle

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

point escape to achieve approximate second-order stationarity. An additional benignity landscape condition then gives proximity to global minimizers. Motivating examples, convexity background, and related optimization methods are discussed in Appendices B, G, and K, respectively.

1.2 From first-order saddle escape to second-order transport

Yamamoto et al. (2025) study perturbed Wasserstein gradient flow (PWGF), using Hessian-guided Gaussian perturbations to escape saddle regions. Under suitable regularity and benignity assumptions, their method reaches a neighborhood of the global minimizer set in polynomial time. The Hessian shapes the perturbations, while the deterministic transport follows the unpreconditioned Wasserstein gradient. This leaves open the possibility of improving escape by using curvature information as preconditioner throughout the transport trajectory.

The choice of the preconditioner is crucial. Wasserstein gradient dynamics expand slowly along directions of weak negative curvature. Where Newton's inverse Hessian is well defined, its linearization instead cancels the curvature sign, so a saddle can become attractive. Table 1 compares these effects. Therefore, a suitable second-order method should rescale curvature while preserving the instability of negative curvature directions. Moreover, because the eigenvalues of an infinite-dimensional Hessian can accumulate at zero, there is no clean spectral gap to guarantee a single, uniform convergence speed. Together, these considerations lead to the following question:

Can curvature dependent Wasserstein transport improve saddle escape while admitting local spectral guarantees in the absence of a uniform positive spectral gap near minimizers?

To address this question, we introduce Wasserstein Saddle-Free Newton (WSFN), with transport direction $-(\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\nabla_\mu F(\mu,\cdot)$, where H_μ is the Wasserstein Hessian, $\nabla_\mu F$ is the Wasserstein gradient, and $\beta > 0$ is the regularization, which makes this direction well defined for bounded self-adjoint Hessians without requiring their invertibility. For suitable stepsizes, the linearization of this transport direction retains attraction along positive curvature directions and repulsion along negative ones. In a special regime of β, it also enables curvature normalization.

WSFN combines this transport with Gaussian perturbations of covariance K_μ^2, where $\mathrm{H}_\mu = \mathrm{M}_\mu + \mathrm{K}_\mu$ and K_μ is the integral kernel operator component. The covariance construction follows PWGF. Our contribution is its coupling to full Hessian preconditioning and the escape analysis of the resulting nonlinear transport. We also characterize the local linearized dynamics spectrally, distinguishing the geometric contraction on strictly positive curvature bands from the polynomial contraction on bands with arbitrarily small positive curvature.

1.3 Technical challenges

The first challenge is that saddle escape is governed by a nonlinear, state dependent transformation of the Wasserstein Hessian. Since Hessians at different iterates act on different L_μ^2 spaces, they must first be represented on a common reference space. Moreover, the spectral maps defining the WSFN preconditioner (see Subsection 3.2) become increasingly sensitive as $\beta \to 0$. Lemma L.16 quantifies this sensitivity, and, together with pulled-back Hessian estimates on the reference space, gives a first-order expansion of the preconditioned gradient (Lemma L.17), which is crucial for the saddle escape analysis.

Perturbative instability of a fixed linearization at a saddle does not guarantee finite-time escape of the nonlinear WSFN trajectory. The key step is Lemma M.3. Unless sufficient objective decrease has already occurred, it controls the displacement of the transport maps emerging from the saddle, thereby limiting the accumulated nonlinear error. Proposition M.5 then turns the resulting spectral instability into a deterministic escape guarantee, while Proposition M.7 combines it with a Gaussian perturbation estimate (Lemma D.6) to obtain a high probability net decrease. By setting $\beta = \Theta(\delta^2)$, curvature normalization reduces the number of escape episodes to $\tilde{O}(1)$ – down from $\tilde{O}(\delta^{-1})$ for PWGF – while guaranteeing a $\tilde{\Omega}(\delta^3)$ decrease per episode. This gives the final $\tilde{O}(\delta^{-3})$ saddle contribution in Theorem 4.6.

Another challenge arises near minimizers because the Wasserstein Hessian need not have a uniform positive spectral gap. Positive spectral values may accumulate at zero, so contraction of every

positive curvature mode does not imply contraction in operator norm. Theorem 4.9 makes this obstruction explicit. Although the linearized iterates converge strongly to the kernel projection, their operator norm remains equal to one. Geometric contraction is recovered only on positive spectral bands away from zero, while source conditions give polynomial decay for arbitrarily small curvature modes. Thus, standard finite-dimensional local convergence arguments cannot be applied directly. Moreover, these are linearized rates, and, as emphasized in Remark 4.10, transferring them to the nonlinear WSFN iterates requires additional strong structural control of the quadratic remainder.

1.4 CONTRIBUTIONS

Our contributions are as follows:

- **A saddle-free second-order method on Wasserstein space.** We formulate WSFN, combining regularized full Hessian preconditioning with Hessian-guided perturbations, and prove the operator stability estimates needed to analyze its transport direction and linearization around critical points (Section 3 and Appendix L).
- **Improved saddle escape contribution.** Under the regularity and admissibility conditions of Theorem 4.6, WSFN reaches approximate second-order stationarity in polynomial time with high probability. With $\beta = \Theta(\delta^2)$, the saddle contribution improves from $\tilde{O}(\delta^{-4})$ for PWGF to $\tilde{O}(\delta^{-3})$. Under additional benignity on the initial sublevel set of F, the second-order stationarity guarantee implies proximity to global minimizers (Corollary 4.8) with the same hitting time bound. Under the standard regime $\varepsilon \lesssim \delta^2$ common in saddle escape analyses, the overall hitting time remains dominated by ε^{-2} (Remark 4.7).
- **Local spectral guarantees.** Near a global minimizer, we prove geometric contraction on positive spectral bands bounded away from zero, and polynomial decay under spectral source conditions on bands with arbitrarily small positive curvature for the linearized WSFN dynamics. (Theorem 4.9). Nonlinear rate transfer requires additional strong assumptions (Remark 4.10).

2 WASSERSTEIN GEOMETRY FOR SECOND-ORDER OPTIMIZATION

We introduce the Wasserstein gradient and Hessian operator used in the analysis of WSFN. Background on operator theory, optimal transport, and Wasserstein geometry is collected in Appendices A, E, and F, respectively.

2.1 NOTATION AND WASSERSTEIN GRADIENT

For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $L^2_\mu := L^2(\mu; \mathbb{R}^d)$, with norm $\|\cdot\|_{L^2_\mu}$ and inner product $\langle \cdot, \cdot \rangle_{L^2_\mu}$. We write Id for the identity map, $\mathrm{T}_\# \mu$ for the pushforward of μ by T, and $\|\cdot\|_{\mu,\infty}$ for the μ-essential supremum norm. Let $\Pi(\mu, \nu)$ and $\Pi_o(\mu, \nu)$ denote, respectively, the couplings and optimal couplings of μ and ν, so that $W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\gamma(x, y)$. We write $B_2(r, \hat{\mu}) := \{\mu : W_2(\mu, \hat{\mu}) < r\}$. The notation $\tilde{O}$, $\tilde{\Theta}$, and $\tilde{\Omega}$ suppresses logarithmic factors.

We use the following definition of Wasserstein differentiability (Ambrosio et al., 2008, Chapter 10).

Definition 2.1 (Wasserstein differentiability). *A functional F is Wasserstein differentiable at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if there exists $\nabla_\mu F(\mu, \cdot) \in L^2_\mu$ such that, for every $\gamma \in \Pi_o(\mu, \nu)$,*

$$F(\nu) = F(\mu) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x \rangle \mathrm{d}\gamma(x, y) + o\left(W_2(\mu, \nu)\right), \text{ as } \nu \to \mu \text{ in } W_2.$$

Throughout, $\nabla_\mu F$ denotes the unique representative belonging to the tangent space $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ (Lanzetti et al., 2025, Proposition 2.11); see Appendix F.1. Under the additional assumptions in Appendix F.4, it satisfies $\nabla_\mu F = \nabla \frac{\delta F}{\delta \mu}$, where $\frac{\delta F}{\delta \mu}$ denotes the first variation of F.

Definition 2.2 (Wasserstein regularity). *We say that F is Wasserstein regular if it is Wasserstein differentiable at every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, and its gradient is Wasserstein differentiable in the measure variable and differentiable in the spatial variable, with derivatives $\nabla^2_\mu F$ and $\nabla \nabla_\mu F$, respectively. For every μ, these derivatives satisfy $\nabla^2_\mu F(\mu, x, y) = \nabla^2_\mu F(\mu, y, x)^\top$, $\mu \otimes \mu$-a.e., $\nabla \nabla_\mu F(\mu, x) = \nabla \nabla_\mu F(\mu, x)^\top$, μ-a.e., together with $\nabla^2_\mu F(\mu, \cdot, \cdot) \in L^2_{\mu \otimes \mu}$ and $\|\nabla \nabla_\mu F(\mu, \cdot)\|_{\mu,\infty} < \infty$.*

2.2 WASSERSTEIN HESSIAN ALONG TRANSPORT CURVES

Our analysis uses transport curves $\mu_t = (\pi_t)_{\#}\mu$, where $\pi_t = \mathrm{Id} + tv$ and $v \in L^2_\mu$, which need not be Wasserstein geodesics. For a fixed transport direction v, define the operators on the space L^2_μ, with $w \in L^2_\mu$, by $\widetilde{\mathrm{M}}_{\mu,t}[w](x) := \nabla\nabla_\mu F(\mu_t, \pi_t(x))w(x)$, and $\widetilde{\mathrm{K}}_{\mu,t}[w](x) := \langle \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(\cdot)), w\rangle_{L^2_\mu}$. Their dependence on the fixed direction v is suppressed in the notation. Set $\widetilde{\mathrm{H}}_{\mu,t} := \widetilde{\mathrm{M}}_{\mu,t} + \widetilde{\mathrm{K}}_{\mu,t}$, and $\mathrm{H}_\mu := \widetilde{\mathrm{H}}_{\mu,0} = \mathrm{M}_\mu + \mathrm{K}_\mu$, where $\mathrm{M}_\mu := \widetilde{\mathrm{M}}_{\mu,0}$ and $\mathrm{K}_\mu := \widetilde{\mathrm{K}}_{\mu,0}$.

Proposition 2.3 (Hessian along transport curves)**.** *Assume that F is Wasserstein regular. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$, and $\mu_t = (\mathrm{Id} + tv)_{\#}\mu$, we have $\frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t) = \langle \widetilde{\mathrm{H}}_{\mu,t} v, v\rangle_{L^2_\mu}$, $t \in [0,1]$, where $\widetilde{\mathrm{H}}_{\mu,t}$ is a bounded self-adjoint operator on L^2_μ.*

The operator is evaluated along the evolving measure μ_t but acts on the fixed space L^2_μ. For absolutely continuous μ and an optimal transport displacement v, the identity at $t = 0$ recovers the usual geodesic Hessian definition in (Villani, 2003, Chapter 8) (see Definition F.7). The proof is given in Appendix L.2; see also Bonet et al. (2024); Wang & Li (2020) for related formulas.

The regularity assumptions imply that M_μ and K_μ are bounded and self-adjoint, with $\| \mathrm{M}_\mu \|_{\mathrm{op}} = \|\nabla\nabla_\mu F(\mu, \cdot)\|_{\mu,\infty}$, and $\| \mathrm{K}_\mu \|_{\mathrm{op}} \le \| \mathrm{K}_\mu \|_{\mathrm{HS}} = \|\nabla^2_\mu F(\mu, \cdot, \cdot)\|_{L^2_{\mu\otimes\mu}}$; see (Arveson, 2001, Theorem 2.1.3) and (Reed & Simon, 1980, Theorem VI.23). In particular, K_μ is Hilbert–Schmidt and therefore compact, so its nonzero spectral values are eigenvalues of finite multiplicity, with zero the only possible accumulation point (Reed & Simon, 1980, Theorems VI.15 and VI.22). Thus, we may write $\underline{\lambda}(\mathrm{K}_\mu) := \inf \sigma(\mathrm{K}_\mu) = \inf_{\|w\|_{L^2_\mu}=1} \langle \mathrm{K}_\mu w, w\rangle_{L^2_\mu}$. Expressions for the Hessian components in terms of the first and second variations of F are given in Proposition F.12.

3 WASSERSTEIN SADDLE-FREE NEWTON

A second-order method for non-convex optimization should improve curvature dynamics without destroying the repulsion needed for saddle escape. To explain this distinction, let μ^* be a critical point of F, and assume that H_{μ^*} is a bounded self-adjoint operator on $L^2_{\mu^*}$.

3.1 CURVATURE NORMALIZATION

At a critical point μ^*, the linearized transport acts on Hessian eigendirections through the spectral multipliers in Table 1. Their detailed derivation is given in Appendix L.5.

Gradient descent expands weakly along directions of small negative curvature. Newton preconditioning removes the curvature dependence, but its multiplier $1 - \tau$ also makes negative curvature directions contract for $\tau \in (0, 2)$. Ideal saddle-free Newton replaces the inverse Hessian by its inverse absolute value, preserving negative curvature repulsion while normalizing its magnitude. The absolute value Hessian principle is inspired by finite-dimensional saddle-free Newton methods (Dauphin et al., 2014). In Wasserstein space, however, $|\mathrm{H}_\mu|$ may lack a bounded inverse, and the absolute value map is not generally operator Lipschitz; see (Reed & Simon, 1980, VI: Bounded Operators, Problem 17). WSFN addresses these difficulties through regularization. The underlying variational construction is given in Appendix L.4.

Method	Preconditioner at μ^*	Multiplier $m_\tau(\lambda)$
Gradient descent	$\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$	$1 - \tau\lambda$
Newton	$\mathrm{H}^{-1}_{\mu^*}$	$1 - \tau$
Ideal saddle-free Newton	$\lvert \mathrm{H}_{\mu^*} \rvert^{-1}$	$1 - \tau\lambda\lvert\lambda\rvert^{-1}$
WSFN	$(\mathrm{H}^2_{\mu^*} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}$	$1 - \tau\lambda(\lambda^2 + \beta)^{-1/2}$

Table 1: Spectral multipliers at a critical point. Newton and ideal saddle-free Newton require boundedly invertible Hessians, whereas WSFN is well defined for $\beta > 0$.

3.2 THE REGULARIZED WSFN UPDATE

For $\beta > 0$, define $P_\beta(A) := (A^2 + \beta\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}$ and $S_\beta(A) := AP_\beta(A)$. The WSFN update is

$$\mu^{n+1} = (\mathrm{Id} - \tau P_\beta(\mathrm{H}_{\mu^n})\nabla_\mu F(\mu^n, \cdot))_{\#}\, \mu^n. \tag{1}$$

Under Assumption 4.4, these iterates are well defined. Proposition L.19 establishes their variational characterization. Regularization also controls the variation of the transport direction. For bounded self-adjoint operators A, B on a common Hilbert space, Lemma L.16 gives $\|P_\beta(A)\|_{\mathrm{op}} \le \beta^{-1/2}$, $\|S_\beta(A)\|_{\mathrm{op}} \le 1$, and

$$\|P_\beta(A) - P_\beta(B)\|_{\mathrm{op}} \le \frac{2}{\pi\beta}\|A - B\|_{\mathrm{op}}, \quad \|S_\beta(A) - S_\beta(B)\|_{\mathrm{op}} \le \beta^{-1/2}\|A - B\|_{\mathrm{op}}.$$

After pulling the Hessians back to a common reference space, these estimates give the first-order expansion of the preconditioned gradient of WSFN in Lemma L.17. Their dependence on β is essential for the saddle escape analysis.

For $0 < \tau \le 1$, the WSFN multiplier in Table 1 lies in $(0, 1)$ for $\lambda > 0$ and exceeds one for $\lambda < 0$. If $\lambda < -\delta$ and $\beta = b\delta^2$ with fixed $b > 0$, then $m_\tau(\lambda) > 1 + \tau(1 + b)^{-1/2}$. Thus, a fixed stepsize gives an expansion factor bounded away from one independently of δ. Turning this linearized instability into nonlinear escape requires the localization argument of Proposition M.5. At zero curvature the multiplier equals one. In this case, local spectral guarantees are stated in Theorem 4.9.

3.3 PERTURBATION EPISODES AND STOPPING

Algorithm 1 combines WSFN steps with perturbation episodes triggered by a small gradient. The perturbation covariance is K_μ^2, following Yamamoto et al. (2025), while the deterministic transport uses the full Hessian $\mathrm{H}_\mu = \mathrm{M}_\mu + \mathrm{K}_\mu$.

Large gradients. When $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} > \varepsilon$, perform (1) with parameters $\beta_{\mathrm{grad}} > 0$ and $0 < \tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$. Lemma M.2 guarantees objective decrease in this phase.

Perturbation episodes. When $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} \le \varepsilon$, store $\bar\mu := \mu$ and initialize $\nu^0 = (\mathrm{Id} + \eta\xi)_{\#}\bar\mu$, with $\xi \sim \mathcal{N}_{L^2_{\bar\mu}}\left(0, \mathrm{K}^2_{\bar\mu}\right)$, using a fresh Gaussian perturbation conditionally on the past. Starting from ν^0, perform at most n_{out} updates of the form (1), with $\beta_{\mathrm{sad}} = \Theta(\delta^2)$ and $\tau_{\mathrm{sad}} \in (0, 1/2]$ fixed independently of δ, subject to the parameter conditions of Theorem 4.6. Keeping τ_{sad} independent of δ is essential to the episode length improvement. Applying the global descent Lemma M.2 with $\beta_{\mathrm{sad}} = \Theta(\delta^2)$ would instead force $\tau_{\mathrm{sad}} = O(\delta)$, losing the uniform $n_{\mathrm{out}} = \tilde O(1)$ episode length guarantee. For more details, see also Remark 4.7.

Acceptance and termination. Immediately after the perturbation and after each subsequent update, test whether $F(\bar\mu) - F(\nu^\ell) \ge F_0$. At the first successful test, accept $\mu := \nu^\ell$ and return to the gradient test. If no $\ell \in \{0, \dots, n_{\mathrm{out}}\}$ satisfies the criterion, terminate and return the stored measure $\widehat\mu := \bar\mu$.

The local descent estimate of Lemma M.3 and the localization argument of Proposition M.5 justify the δ independent saddle stepsize τ_{sad}. The returned measure has small gradient by construction, and its second-order guarantee follows with high probability from Theorem 4.6. Section 5 gives a numerical illustration. Appendix C describes the particle implementation and additional experiments.

Why this covariance is well defined. An L^2_μ-valued Gaussian requires a trace class covariance. Under our assumptions, K_μ is Hilbert–Schmidt, and hence $\mathbb{E}\|\xi\|^2_{L^2_\mu} = \mathrm{Tr}(\mathrm{K}^2_\mu) = \|\,\mathrm{K}_\mu\,\|^2_{\mathrm{HS}} < \infty$. The identity operator is not trace class when L^2_μ is infinite dimensional, and the full Hessian square H^2_μ need not be trace class because of the multiplication component M_μ. Thus, K^2_μ supplies an admissible covariance with finite expected squared perturbation norm, while the full Hessian remains available for deterministic preconditioning. For additional details, see Appendix D.

4 THEORETICAL GUARANTEES

In this section, we state our main results and start with the assumptions under which WSFN enjoys global saddle escape and local spectral convergence guarantees.

4.1 SECOND-ORDER OPTIMALITY AND LANDSCAPE ASSUMPTIONS

At a critical point $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$, we have $\mathrm{M}_{\mu^*} = 0$, so that $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$. Thus, $\mathrm{K}_{\mu^*} \succeq 0$ is a necessary condition for local minimality; see Proposition H.7 and Yamamoto et al. (2025). This motivates the following approximate notions.

Definition 4.1 ((ε, δ)-second-order critical point and saddle point)**.** *Let $\varepsilon, \delta > 0$. A measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ satisfying $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} \le \varepsilon$ is an (ε, δ)-second-order critical point if $\underline{\lambda}(\mathrm{K}_\mu) \ge -\delta$, and an (ε, δ)-saddle point if $\underline{\lambda}(\mathrm{K}_\mu) < -\delta$.*

Small gradient norm alone does not control the multiplication operator of the Hessian. The escape analysis requires this control at the measures initiating perturbation episodes. Thus, we impose a local version of the gradient and Hessian relation in (Yamamoto et al., 2025, Assumption 2).

Assumption 4.2 (Uniform gradient to Hessian control on a set)**.** *Let $S \subseteq \mathcal{P}_2(\mathbb{R}^d)$. There exist $\varepsilon_0, R_F > 0$ such that, for every $\mu \in S$ satisfying $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} \le \varepsilon_0$, we have $\| \mathrm{M}_\mu \|_{\mathrm{op}} \le R_F \|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu}$.*

For the global convergence analysis, it is sufficient to impose this assumption on the initial sublevel set $\mathcal{L}_0 := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : F(\mu) \le F(\mu^0)\}$, and choose the threshold $0 < \varepsilon \le \varepsilon_0$. The constants R_F and ε_0 are fixed independently of the tolerances ε, δ and the episode index. In particular, for every $\mu \in S$ with $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} \le \varepsilon \le \varepsilon_0$, we have $\| \mathrm{H}_\mu - \mathrm{K}_\mu \|_{\mathrm{op}} = \| \mathrm{M}_\mu \|_{\mathrm{op}} \le R_F \varepsilon$. Hence an (ε, δ)-second-order critical point in S also satisfies $\inf \sigma(\mathrm{H}_\mu) \ge -\delta - R_F \varepsilon$. Thus, the stationarity criterion also controls the full Wasserstein Hessian. Under the condition $\varepsilon = \tilde{O}(\delta^2)$, the additional curvature tolerance $R_F \varepsilon$ is lower order than δ, with R_F fixed.

To connect approximate stationarity with global optimality, write $\mathcal{M}^* := \operatorname{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$ and, for $\alpha \ge 0$, define $\mathcal{N}_\alpha(\mathcal{M}^*) := \left\{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) \le \alpha\right\}$. The following quantitative benignity condition is analogous to the strict saddle conditions used in Euclidean saddle escape analyses Jin et al. (2017); Ge et al. (2015). It roughly says that every measure not close to the set of global minimizers either has a large gradient or is localized near a saddle.

Assumption 4.3 (Benignity (saddle property) on a set)**.** *Let $\mathcal{S} \subset \mathcal{P}_2(\mathbb{R}^d)$. Assume that F satisfies the $(\varepsilon, \delta, \alpha_{\mathrm{ben}})$-saddle property on $\mathcal{S}$, namely, $\mathcal{M}^* \neq \varnothing$, and every $\mu \in \mathcal{S} \setminus \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$ satisfies $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} > \varepsilon$ or $\underline{\lambda}(\mathrm{K}_\mu) < -\delta$.*

This places every (ε, δ)-second-order critical point in $\mathcal{S}$ inside $\mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$ (Lemma H.11). Unlike the global formulation in Yamamoto et al. (2025), it suffices to take $\mathcal{S} = \mathcal{L}_0$, because the outer iterates of Algorithm 1, i.e., at which the gradient $\nabla_\mu F$ is evaluated, remain in this sublevel set (Lemma M.9). Examples motivating this property are discussed in Appendix B.

Finally, the analysis requires uniform boundedness and coupling Lipschitz continuity of the Hessian.

Assumption 4.4 (Uniform boundedness of the Wasserstein Hessian)**.** *Assume that F is Wasserstein regular and that there exist $C_{\mathrm{M}}, C_{\mathrm{K}} > 0$ such that, for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\| \mathrm{M}_\mu \|_{\mathrm{op}} \le C_{\mathrm{M}}, \quad \| \mathrm{K}_\mu \|_{\mathrm{HS}} \le C_{\mathrm{K}}.$$

Assumption 4.5 (Coupling Lipschitzness of the Wasserstein Hessian)**.** *Assume that F is Wasserstein regular and that there exist $L_{\mathrm{M}}, L_{\mathrm{K}} > 0$ such that, for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Pi(\mu, \nu)$,*

$$\|\nabla \nabla_\mu F(\nu, \cdot) - \nabla \nabla_\mu F(\mu, \cdot)\|_{\gamma, \infty} \le L_{\mathrm{M}} d_\gamma, \ \left\|\nabla^2_\mu F(\nu, \cdot, \cdot) - \nabla^2_\mu F(\mu, \cdot, \cdot)\right\|_{L^2_{\gamma \otimes \gamma}} \le L_{\mathrm{K}} d_\gamma,$$

where $d^2_\gamma := \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - x\|^2 \mathrm{d}\gamma(x, y)$.

These assumptions give uniform boundedness and Lipschitz control of the Hessian on a common space, with constants $C_{\mathrm{H}} := C_{\mathrm{M}} + C_{\mathrm{K}}$ and $L_{\mathrm{H}} := L_{\mathrm{M}} + L_{\mathrm{K}}$, respectively; see Lemmas L.11,

L.12, L.13, and L.14. Analogous Hessian Lipschitz continuity assumptions have been employed in Euclidean space (Jin et al., 2017; Ge et al., 2015).

Assumption 4.5 imposes coupling regularity, as in (Yamamoto et al., 2025, Assumption 3). Its multiplication operator bound is stronger than ordinary spatial Lipschitz continuity of the Hessian. This condition nevertheless provides the operator norm control needed to compare the pulled-back Hessians on a common reference space and their regularized preconditioners along transport trajectories. The resulting regularity class includes nonconvex objectives. Propositions B.6, N.1 give explicit examples.

4.2 Approximate stationarity and proximity to global minimizers

We first establish approximate second-order stationarity without assuming benignity, and then use benignity to obtain proximity to the global minimizer set.

Theorem 4.6 (Polynomial time approximate second-order stationarity)**.** *Let Assumptions 4.4 and 4.5 hold. Fix $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, and suppose Assumption 4.2 holds on $\mathcal{L}_0$. Choose $0 < \varepsilon \le \varepsilon_0$, $\delta > 0$, and $\zeta \in (0,1)$. Assume $L_{\mathrm{H}}\varepsilon \lesssim \delta^2$. Under the parameter conditions and choices $\beta_{\mathrm{grad}} > 0$, $\tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$, $\beta_{\mathrm{sad}} = \Theta(\delta^2)$, $\tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$, $\kappa = \tilde{O}(1)$, $n_{\mathrm{out}} = \tilde{O}(1)$, $F_0 = \tilde{\Omega}(\delta^3)$, $\eta = \tilde{\Theta}(\frac{\delta^3}{\varepsilon+\delta^{3/2}})$ of Theorem M.8, Algorithm 1, initialized at $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, returns an (ε, δ)-second-order critical point with probability at least $1-\zeta$, using*

$$N_{\mathrm{SOSP}} = \tilde{O}\left(F_{\min}\left(\varepsilon^{-2} + \delta^{-3}\right)\right) \tag{2}$$

WSFN transport steps, where $F_{\min} := F(\mu^0) - \inf_\mu F(\mu)$.

Proof overview. Each large gradient step decreases the objective by order ε^2. For saddle episodes, the local descent estimate of Lemma M.3 converts insufficient objective decrease into a transport displacement bound of order $\beta_{\mathrm{sad}}^{-1/4}\sqrt{F_0}$, up to logarithmic factors and the controlled initial perturbation. The nonlinear escape argument requires this displacement to remain within a radius of order $L_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{sad}}}$, again up to logarithmic factors. Balancing these scales selects a decrease threshold at the cubic scale, $F_0 = \tilde{\Omega}(L_{\mathrm{H}}^{-2}\beta_{\mathrm{sad}}^{3/2}) = \tilde{\Omega}(L_{\mathrm{H}}^{-2}\delta^3)$. Proposition M.7 guarantees this net decrease with high probability within at most $n_{\mathrm{out}} = \tilde{O}(1)$ steps. Counting large gradient steps and successful episodes against the objective gap $F_{\min}$ gives (2). A conditional union bound controls the probability of terminating at a saddle. A detailed proof roadmap is provided in Appendix M.

Remark 4.7 (Improved saddle dependence)**.** *With $\beta_{\mathrm{sad}} = b\delta^2$, $b > 0$, and a saddle episode stepsize τ_{sad} independent of δ, WSFN improves the saddle contribution from $\tilde{O}(\delta^{-4})$ for PWGF (Yamamoto et al., 2025, Theorem 5.5) to $\tilde{O}(\delta^{-3})$, with the other problem constants fixed. The gain comes from shortening each episode from $\tilde{O}(\delta^{-1})$ to $\tilde{O}(1)$ iterations, while retaining a decrease of order $\tilde{\Omega}(\delta^3)$. Indeed, the WSFN preconditioner normalizes the unstable growth rate, but Lipschitz Hessian regularity only guarantees negative curvature of order δ over a radius of order $L_{\mathrm{H}}^{-1}\delta$. Hence, the associated objective depth scale is $L_{\mathrm{H}}^{-2}\delta^3$. Thus, WSFN removes the dynamical slowdown, but not the depth of saddles. Appendix N explains this decomposition and proves sharpness of the cubic decrease scale. We emphasize that this is not a worst case iteration complexity lower bound.*

The restriction $L_{\mathrm{H}}\varepsilon \lesssim \delta^2$ is also required by PWGF (Yamamoto et al., 2025, Theorem 5.5). Therefore, the improvement only concerns the saddle contribution. The large gradient phase admits $\Omega(\varepsilon^{-2})$ examples with uniform regularity constants and bounded initial objective gap $F_{\min}$ (Proposition N.2). Thus, at $\varepsilon = \tilde{\Theta}(\delta^2)$, the leading total bound remains $\tilde{O}(\delta^{-4})$. A uniform $\tilde{O}(\varepsilon^{-3/2})$ large gradient guarantee – and thus an $\tilde{O}(\delta^{-3})$ total bound – typically requires cubic Taylor regularization instead of WSFN's quadratic model (see Appendix L.4). Crucially, cubic regularization leads to an implicit update, sacrificing explicit iterations like (1).

In Euclidean optimization, cubic regularization achieves $O(\varepsilon^{-3/2})$ iteration complexity for first-order stationarity under Lipschitz Hessian assumptions (Nesterov & Polyak, 2006; Cartis et al., 2011); see Grapiglia et al. (2022) for finite-difference Hessian variants. Developing an analogous Wasserstein guarantee through a modified large gradient phase is left for future work.

Under benignity on the initial sublevel set, this second-order stationarity guarantee implies proximity to the global minimizer set without additional iterations.

Corollary 4.8 (Proximity under sublevel benignity)**.** *Under the hypotheses and parameter choices of Theorem 4.6, suppose additionally that Assumption 4.3 holds on $\mathcal{L}_0$. Then, with probability at least $1-\zeta$, the returned measure satisfies $\widehat{\mu} \in \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$, with the same bound (2).*

4.3 Local spectral behavior near a global minimizer

Fix $\tau \in (0,1]$ and $\beta > 0$ as parameters of the deterministic WSFN update. The following result characterizes the linearization of this fixed parameter update. To compare WSFN updates on a fixed reference space, let $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$, set $\mu_v := (\mathrm{Id}+v)_{\#}\mu^*$, and define

$$\Psi_{\tau,\beta}(v) := v - \tau\left[(\mathrm{H}^2_{\mu_v} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\nabla_\mu F(\mu_v,\cdot)\right]\circ(\mathrm{Id}+v).$$

One WSFN step sends μ_v to $(\mathrm{Id}+\Psi_{\tau,\beta}(v))_{\#}\mu^*$.

Theorem 4.9 (Local spectral behavior of WSFN)**.** *Let Assumptions 4.4 and 4.5 hold, and suppose that μ^* is a global minimizer. Fix $\tau \in (0,1]$, $\beta > 0$, and define $\mathcal{L}_{\tau,\beta} := \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^*} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\,\mathrm{K}_{\mu^*}$. There exists $C = C(L_{\mathrm{H}}, C_{\mathrm{K}}, \beta) > 0$ such that, for every $v \in L^2_{\mu^*}$,*

$$\Psi_{\tau,\beta}(v) = \mathcal{L}_{\tau,\beta}v + \mathcal{R}_{\tau,\beta}(v), \quad \|\mathcal{R}_{\tau,\beta}(v)\|_{L^2_{\mu^*}} \le C\|v\|_*^2.$$

The linearized iterates satisfy $\mathcal{L}^n_{\tau,\beta}v \to P_{\ker \mathrm{K}_{\mu^*}}v$ strongly for every v, although $\|\mathcal{L}^n_{\tau,\beta}\|_{\mathrm{op}} = 1$ for every $n \ge 1$. For $m > 0$, let $P_m := \mathbb{1}_{[m,\infty)}(\mathrm{K}_{\mu^*})$. Then, for every $n \ge 1$,

$$\|\mathcal{L}^n_{\tau,\beta}P_m v\|_{L^2_{\mu^*}} \le \left(1 - \tau\frac{m}{\sqrt{m^2+\beta}}\right)^n \|P_m v\|_{L^2_{\mu^*}}.$$

If $(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v = \mathrm{K}^s_{\mu^*}\,w$ for some $s > 0$ and $w \in L^2_{\mu^*}$, then

$$\|\mathcal{L}^n_{\tau,\beta}v - P_{\ker \mathrm{K}_{\mu^*}}v\|_{L^2_{\mu^*}} \le \left(\frac{s\sqrt{C_{\mathrm{K}}^2+\beta}}{e\tau n}\right)^s \|w\|_{L^2_{\mu^*}}, \quad n \ge 1.$$

Spectral interpretation. At an absolutely continuous global minimizer, $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*} \succeq 0$. The proof combines linearization around K_{μ^*} with functional calculus for the multiplier $1 - \tau\lambda(\lambda^2+\beta)^{-1/2}$; see Appendix M.3. Curvature much larger than $\sqrt{\beta}$ is approximately normalized, whereas nearly flat directions remain weakly contracted and kernel components are unchanged. Therefore, the theorem provides geometric contraction on spectral bands separated from zero and polynomial decay under the source condition $(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v = \mathrm{K}^s_{\mu^*}\,w$. This condition suppresses small curvature components through the factor λ^s and is standard in spectral regularization and inverse problems (Bauer et al., 2007; Lin et al., 2020; Blanchard & Mücke, 2018). For more details, we refer to Remark M.10.

Remark 4.10 (Limitations of nonlinear rate transfer)**.** *Theorem 4.9 gives rates for the linearized iterates. Nonlinear rate transfer requires further structural control, such as invariance of a positive spectral band or suitable source norm bounds on the remainder. These are strong requirements, which do not follow from the present assumptions. Accordingly, the neighborhood entry guarantee of Corollary 4.8 and the spectral theorem above do not together imply convergence rates for the nonlinear iteration. Perturbative estimates for a deterministic continuation from the returned measure by Algorithm 1 are given in Theorem M.11. The additional requirements for nonlinear rate transfer are discussed in more detail in Remark M.12.*

5 Numerical illustration

Particle implementation. We represent the measure by $\mu^N = N^{-1}\sum_{i=1}^N \delta_{x_i}$. For the regularized empirical objective below, we evaluate the particle gradient and the action of the full particle Hessian analytically. WSFN applies a Lanczos approximation of $(\mathrm{H}_N^2 + \beta\,\mathrm{I}_{\mathrm{Nd}\times\mathrm{Nd}})^{-1/2}$ to the gradient, without forming or diagonalizing the full $Nd \times Nd$ Hessian. Perturbations use the integral Hessian component K_N, with their magnitudes normalized as specified in Appendix C. This experiment uses prescribed perturbation times instead of stopping and episode rules. Quantitative control of kernel regularization, particle, and Lanczos approximation errors remains outside the present analysis.

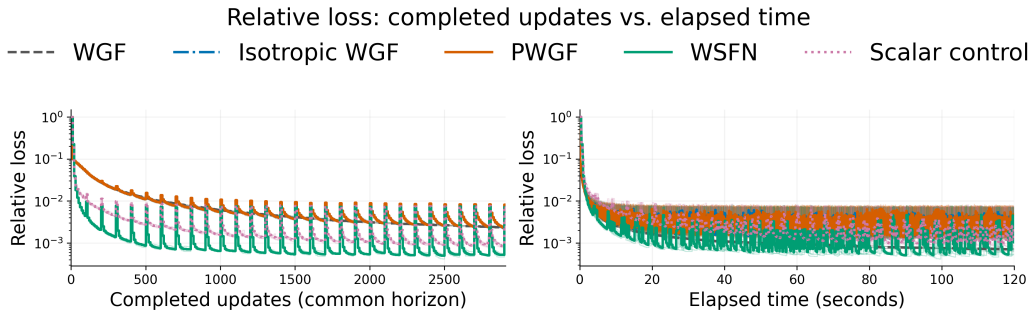


Figure 1: Regularized Coulomb MMD. Relative loss $\frac{F_{\rho,N,M}(X_t;Y)}{F_{\rho,N,M}(X_0;Y)}$ versus completed updates over a common update horizon (left) and elapsed time (right). Thick curves show means of the normalized losses over five paired trials. Thin curves show individual trials. Both vertical axes are logarithmic.

Regularized Coulomb MMD and directional ablation. We consider a smooth regularization of the Coulomb MMD objective in dimension $d = 3$. Given target samples $Y = (y_a)_{a=1}^M$, define $\kappa_\rho(x, y) = (\|x - y\|^2 + \rho^2)^{-1/2}$ and minimize

$$F_{\rho,N,M}(X;Y) = \frac{1}{N^2}\sum_{i,j=1}^{N}\kappa_\rho(x_i,x_j) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{a=1}^{M}\kappa_\rho(x_i,y_a) + \frac{1}{M^2}\sum_{a,b=1}^{M}\kappa_\rho(y_a,y_b). \quad (3)$$

This is the squared MMD between the two empirical measures for the regularized kernel. We use $N = 500$ particles, $M = 400$ target samples, and $\rho = 0.05$.

We compare WGF, WGF with isotropic perturbations, PWGF with Hessian-shaped perturbations, WSFN, and a scalar control. The control replaces the preconditioned direction by a gradient direction of the same norm, computed at the control's own current state. The four methods' learning rates are selected independently on pilot seeds. The control inherits WSFN's selected rate and preconditioner settings. Evaluation uses five paired trials with shared target samples and initial particles, with 120 seconds per method and trial. Except for WGF, methods receive perturbations every 100 updates, using Gaussian draws at corresponding perturbation events.

Figure 1 shows that WSFN reduces the loss more rapidly per update and reaches low loss levels earlier in elapsed time. Between prescribed perturbations, its loss is lower than that of the scalar control, supporting a benefit from the preconditioned direction beyond its locally matched magnitude at the selected WSFN learning rate. The control is not independently tuned, so this comparison does not imply superiority over a tuned scalar adaptive method. Repeated perturbations produce transient loss increases, and WSFN does not dominate the baselines at every time. The gradient diagnostics in Appendix C also show that lower loss does not imply a uniformly smaller gradient norm.

6 Conclusion

We introduced WSFN, combining regularized Hessian preconditioning with Hessian-guided perturbations. Under the stated conditions, WSFN reaches approximate second-order stationarity, improving the saddle contribution from $\tilde{O}(\delta^{-4})$ to $\tilde{O}(\delta^{-3})$. The guarantees rely on Hessian regularity and uniform control at episode starting measures. Verifying or relaxing these conditions for learning objectives remains an important direction. The local spectral rates concern linearized dynamics, and nonlinear rate guarantees require additional control of the remainder. Moreover, the complexity counts exact measure-valued transport steps, leaving particle approximation errors and computational cost outside the current guarantees. Future work should establish nonlinear rates under verifiable structural conditions and quantify how particle, preconditioner, and covariance approximations affect descent and escape. Relating these errors to β and δ would clarify when the improved transport step bound gives computational savings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the numerical implementation of the proposed theoretical methods, refine hypotheses, and support the writing of proofs. We also used these tools to suggest experimental parameters, draft and edit parts of the manuscript, improve readability and organization, summarize and identify relevant literature and suggest the overall structure of the paper.

Generative AI tools were not used to develop the underlying theoretical models or conceptual frameworks, to formulate mathematical claims, nor to provide critical ingredients for establishing the main mathematical results. Uses related to generating synthetic datasets, translation, cleaning or reformatting datasets, qualitative or thematic data analysis, and interpretation of results were not applicable to this work.

All AI assisted outputs were reviewed by the authors. In particular, mathematical claims and proofs were independently checked for correctness, AI assisted code was inspected and tested, experimental suggestions were evaluated by the authors before use, and literature references and summaries were verified against the original sources. The authors take full responsibility for the final content of this work, including all text, mathematical claims, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions underlying our theoretical results are explicitly outlined in Section 4, with complete proofs provided in Appendix M. Auxiliary results and supporting proofs appear in Appendices D and L, while the remaining appendices offer additional motivation, background material, and theoretical context. Finally, complete setup details and hyperparameters for our numerical experiments are described in Appendix C.

REFERENCES

L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

W. Arveson. *A Short Course on Spectral Theory*. Graduate Texts in Mathematics. Springer New York, 2001.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

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

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in Wasserstein space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

Siwan Boufadene and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4):4556–4587, 2025.

P. Cardaliaguet, F. Delarue, J.M. Lasry, and P.L. Lions. *The Master Equation and the Convergence Problem in Mean Field Games*. Annals of Mathematics Studies. Princeton University Press, 2019.

René A. Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*. Springer International Publishing, 2018.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130:295–319, 2011.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. A concise second-order complexity analysis for unconstrained optimization using high-order regularized models. *Optimization Methods and Software*, 35(2):243–256, 2020.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham, 2025. École d'Été de Probabilités de Saint-Flour XLIX – 2019.

Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

Lénaïc Chizat and Francis R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS*, 2018.

Lénaïc Chizat, Stephen Zhang, Matthieu Heitz, and Geoffrey Schiebinger. Trajectory inference via mean-field Langevin in path space. In *Advances in Neural Information Processing Systems*, 2022.

Lénaïc Chizat, Maria Colombo, and Xavier Fernández-Real. Convergence of drift-diffusion PDEs arising as Wasserstein gradient flows of convex functions. *Archive for Rational Mechanics and Analysis*, 250(48), 2026a.

Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative convergence of Wasserstein gradient flows of Kernel Mean Discrepancies, 2026b. arXiv:2603.01977.

Casey Chu, Jose Blanchet, and Peter Glynn. Probability functional descent: A unifying perspective on GANs, Variational Inference, and Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1213–1222. PMLR, 09–15 Jun 2019.

J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1994.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, 2014.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. EMS Press, Berlin, 2023.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 797–842. PMLR, 03–06 Jul 2015.

G. N. Grapiglia, M. L. N. Gonçalves, and G. N. Silva. A cubic regularization of Newton's method with finite difference Hessian approximations. *Numerical Algorithms*, 90:607–630, 2022.

S.J. Gustafson and I.M. Sigal. *Mathematical Concepts of Quantum Mechanics*. Universitext. Springer International Publishing, 2020.

E. Hille and R.S. Phillips. *Functional analysis and semi-groups*. American Mathematical Society. American Mathematical Soc., 1957.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043–2065, 2021.

Yu-Jui Huang and Zachariah Malik. Generative modeling by minimizing the Wasserstein-2 loss. *SIAM Journal on Mathematics of Data Science*, 8(3):973–999, 2026.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1724–1732. PMLR, 06–11 Aug 2017.

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

T. Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer Berlin Heidelberg, 1995.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

Marc Lambert, Sinho Chewi, Francis R. Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14434–14447, 2022.

Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-order conditions for optimization in the Wasserstein space. *SIAM Journal on Mathematics of Data Science*, 7(1):274–300, 2025.

Razvan-Andrei Lascu, Mateusz B. Majka, David Šiška, and Łukasz Szpruch. Linear convergence of proximal descent schemes on the Wasserstein space, 2024. arXiv:2411.15067.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Yiping Lu, Chao Ma, Jianfeng Lu, and Lexing Ying. A mean-field analysis of Deep ResNet and Beyond: Towards Provable Optimization Via Overparameterization From Depth. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6426–6436. PMLR, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E7665 – E7671, 2018.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex Analysis of the Mean Field Langevin Dynamics. *arXiv e-prints*, art. arXiv:2201.10469, January 2022.

F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics. I: Functional Analysis*. Academic Press, New York, 1980.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75 (9):1889–1935, 2022.

Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12356–12366. Curran Associates, Inc., 2020.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

C. Villani. *Topics in optimal transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

Yifei Wang and Wuchen Li. Information Newton's flow: second-order optimization method in probability space, 2020. arXiv:2001.04341.

Yifei Wang, Peng Chen, and Wuchen Li. Projected Wasserstein gradient descent for high-dimensional Bayesian inference. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1513–1532, 2022.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2093–3027. PMLR, 06–09 Jul 2018.

Naoya Yamamoto, Juno Kim, and Taiji Suzuki. Hessian-guided perturbed Wasserstein gradient flows for escaping saddle points. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Rentian Yao and Yun Yang. Mean-field variational inference via Wasserstein gradient flow. *Journal of Machine Learning Research*, 26(309):1–46, 2025.

Rentian Yao, Xiaohui Chen, and Yun Yang. Wasserstein proximal coordinate gradient algorithms. *Journal of Machine Learning Research*, 25(269):1–66, 2024.

Shuailong Zhu and Xiaohui Chen. Convergence analysis of the Wasserstein proximal algorithm beyond geodesic convexity, 2025. arXiv:2501.14993.

7 APPENDIX

The appendices are organized to guide the reader from motivating examples and implementation to background material and, finally, the technical analysis. Appendix A collects operator theoretic notation on L^2_μ. Appendix B presents non-convex objectives on Wasserstein space that motivate the benignity assumption, and Appendix C gives the particle-based implementation of WSFN. Appendix D introduces the L^2_μ-valued Gaussian process machinery used for the perturbation step. Appendix E recalls the optimal transport background, Appendix F develops the differential calculus on Wasserstein space, and Appendix G reviews the main convexity notions in Wasserstein space. Appendix H gathers first and second-order optimality conditions, Appendix I recalls the corresponding first and second-order dynamics, Appendix J provides calculus along curves together with flow approximation results linking discrete dynamics to their continuous-time limits, and Appendix K places WSFN alongside related methods in a common Wasserstein Proximal Taylor framework. Finally, Appendix L collects the auxiliary lemmas used throughout the analysis, and additional local convergence results. Appendix M contains the proofs of the main results. Appendix N explains the improvement of the saddle contribution from $\tilde{O}(\delta^{-4})$ to $\tilde{O}(\delta^{-3})$.

CONTENTS

1 Introduction **1**

1.1 Nonconvex optimization on Wasserstein space 1

1.2 From first-order saddle escape to second-order transport 2

1.3 Technical challenges 2

1.4 Contributions 3

2 Wasserstein geometry for second-order optimization **3**

2.1 Notation and Wasserstein gradient 3

2.2 Wasserstein Hessian along transport curves 4

3 Wasserstein Saddle-Free Newton **4**

3.1 Curvature normalization 4

3.2 The regularized WSFN update 5

3.3 Perturbation episodes and stopping 5

4 Theoretical guarantees **6**

4.1 Second-order optimality and landscape assumptions 6

4.2 Approximate stationarity and proximity to global minimizers 7

4.3 Local spectral behavior near a global minimizer 8

5 Numerical illustration **8**

6 Conclusion **9**

7 Appendix **13**

A Additional notation for operators on L^2_μ **16**

B Motivating examples of qualitative benign landscapes **16**

B.1 Landscape analysis of a mean-field residual flow 18

C Particle implementation of WSFN **20**

C.1 Numerical experiments 22

C.2 Regularized Coulomb MMD: implementation details 22

C.3 In-context feature learning: implementation details 24

C.4 Mean-field deep residual networks: implementation details. 26

C.5 Matrix decomposition: implementation details 29

D L^2_μ-valued Gaussian Processes **31**

E Background on Optimal Transport **35**

F Differential calculus on Wasserstein space **36**
F.1 Wasserstein subdifferentials and differentiability 36
F.2 Examples of Wasserstein differentiable functionals 37
F.3 Wasserstein Hessian along geodesics 38
F.4 First and second variation 38

G Convexity in Wasserstein space **40**
G.1 Geodesic convexity 40
G.2 Linear convexity 42

H Optimality conditions in the Wasserstein space **43**
H.1 First-order optimality conditions 43
H.2 Second-order optimality conditions 43
H.3 Consequences under the saddle property 46

I First and second-order dynamics on Wasserstein space **47**
I.1 Wasserstein gradient flow and Wasserstein gradient descent 47
I.2 Wasserstein Newton flow and Wasserstein Newton method 48

J Calculus along curves and flow approximations **48**
J.1 Chain rule along absolutely continuous curves 49
J.2 Convergence of Wasserstein gradient descent to the flow 49
J.3 Convergence of the Wasserstein Newton scheme to the flow 51

K The Wasserstein Proximal Taylor (WPT-k) method **53**
K.1 Iterates of the WPT-1 method 54
K.2 Iterates of the WPT-2 method 54
K.3 Wasserstein Saddle-Free Newton as a proximal Taylor scheme 55

L Auxiliary results **55**
L.1 Calculus along transport curves 55
L.2 Hessian identities and second-order expansions 57
L.3 Operator bounds and preconditioner identities 61
L.4 Well-posedness and characterization of the optimizer 72
L.5 Linearization of preconditioned Wasserstein updates 75
L.6 Relation with Newton's method 76

M Proofs of main results **78**
M.1 Auxiliary lemmas for the WSFN dynamics 79
M.2 Escape from saddle regions 83
M.3 Proofs of the convergence theorems 93

M.4 Local spectral analysis and nonlinear limitations 100

N Scope of the improved saddle dependence **103**

N.1 Episode duration and objective decrease counting 103

N.2 Curvature normalization removes slow amplification 103

N.3 Why the objective decrease scale is cubic 104

N.4 Sharpness of the cubic decrease scale 105

N.5 Why smaller regularization does not improve this balance 105

N.6 What is sharp and what might improve further 106

A Additional notation for operators on L^2_μ

In this section, we collect several standard notions and conventions for bounded linear operators on L^2_μ that will be used throughout the paper. For an operator $A_\mu : L^2_\mu \to L^2_\mu$, the operator norm is defined as $\|A_\mu\|_{\mathrm{op}} = \sup_{v\neq 0} \frac{\|A_\mu v\|_{L^2_\mu}}{\|v\|_{L^2_\mu}}$. The operator A_μ is called *positive* (respectively, *non-negative*) and denoted by $A_\mu \succ 0$ (respectively, $A_\mu \succeq 0$) if $\langle A_\mu v, v\rangle_{L^2_\mu} > 0$ (respectively, $\langle A_\mu v, v\rangle_{L^2_\mu} \geq 0$), for all $v \in L^2_\mu$. The operator A_μ is called *uniformly positive* and denoted by $A_\mu \succeq \lambda \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$, where $\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ is the identity operator, if there exists $\lambda > 0$ such that $\langle A_\mu v, v\rangle_{L^2_\mu} \geq \lambda \|v\|^2_{L^2_\mu}$ for all $v \in L^2_\mu$. The spectrum of the operator A_μ is denoted by $\sigma(A_\mu)$. For a bounded linear operator $A_\mu : L^2_\mu \to L^2_\mu$, its adjoint, denoted by A^*_μ, is the bounded linear operator $A^*_\mu : L^2_\mu \to L^2_\mu$ satisfying $\langle A_\mu v, w\rangle_{L^2_\mu} = \langle v, A^*_\mu w\rangle_{L^2_\mu}$, for all $v, w \in L^2_\mu$. A bounded linear operator A_μ is called self-adjoint if $A_\mu = A^*_\mu$. A Hilbert–Schmidt (HS) operator $A_\mu : L^2_\mu \to L^2_\mu$ is a bounded operator that has finite HS norm $\|A_\mu\|^2_{\mathrm{HS}} := \sum_{n=1}^\infty \|A_\mu e_n\|^2_{L^2_\mu}$, where $\{e_n\}_{n=1}^\infty$ is an orthonormal basis. We denote by $\mathcal{LS}(L^2_\mu)$ and $\mathcal{B}(L^2_\mu)$ the spaces of linear symmetric and linear bounded operators from L^2_μ to L^2_μ, respectively. If $A_\mu \in \mathcal{B}(L^2_\mu)$, then $\|A_\mu\|_{\mathrm{op}} \leq \|A_\mu\|_{\mathrm{HS}}$.

B Motivating examples of qualitative benign landscapes

The examples below present different forms of qualitative benignity, such as the absence of spurious local minima or the fact that non-global critical points are saddles. These qualitative statements motivate the quantitative benignity condition used in Corollary 4.8, but do not provide uniform constants $(\varepsilon, \delta, \alpha_{\mathrm{ben}})$ on the whole Wasserstein space. Proposition B.5 gives sufficient compactness and continuity conditions under which the quantitative property follows.

Example B.1 (Coulomb MMD)**.** *Let $d \geq 3$ and $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ be a target probability measure. The Maximum Mean Discrepancy (MMD) associated with the Coulomb kernel is given by*

$$F(\mu) = \frac{1}{d-2} \iint_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{\|x-y\|^{d-2}} \mathrm{d}(\mu-\mu^*)(x)\mathrm{d}(\mu-\mu^*)(y).$$

Boufadene & Vialard (2025) showed that, although this functional is not geodesically convex in Wasserstein space, it has no local minima other than the global minimizer μ^*. Hence, every local minimizer of F is global. Thus, the Coulomb MMD provides a natural example of a non-convex functional with a benign optimization landscape. Moreover, they proved that any critical point coincides with μ^* on the interior of its support. If μ is a critical point of F, then $\mu|_{\mathrm{Int}(\mathrm{supp}\,\mu)} = \mu^*|_{\mathrm{Int}(\mathrm{supp}\,\mu)}$. In particular, if for some open set $U \subset \mathrm{Int}(\mathrm{supp}\,\mu)$ the measures μ and μ^* are absolutely continuous, i.e., admit densities ρ and ρ^* on U, then $\rho = \rho^*$ a.e. on U. Thus any discrepancy between μ and μ^* could only remain in the atomic part or on the boundary of the support.

Example B.2 (Three-layer mean-field network)**.** *Consider a three-layer teacher-student network in which the first two layers are expressed as a mean-field two-layer NN*

$$h_\mu(x) = \int_\Theta h_\theta(x)\mathrm{d}\mu(\theta),$$

and the third layer is a linear map $T \in \mathbb{R}^{k\times k}$. Here, $h_\theta(x) := a\sigma(w^\top x) \in \mathbb{R}^k$ for $\theta = (a, w) \in \mathbb{R}^k \times \mathbb{R}^d$, input data point $x \sim \mathcal{D}$, and activation function σ, whereas the integer k denotes the feature dimension of the hidden representation. The L^2 loss with respect to a teacher network $x \mapsto \hat{T}h_{\hat{\mu}}(x)$ is

$$L_{\mathrm{NN}}(\mu, T) = \mathbb{E}_{x\sim\mathcal{D}}\left[\|\hat{T}h_{\hat{\mu}}(x) - Th_\mu(x)\|^2\right].$$

Define $\Sigma_{\mu,\nu} := \mathbb{E}_{x\sim\mathcal{D}}[h_\mu(x)h_\nu(x)^\top]$ and $\mu^* := \hat{T}_\#\hat{\mu}$, so that $h_{\mu^*}(x) = \hat{T}h_{\hat{\mu}}(x)$. In the two-timescale regime where the linear layer updates infinitely fast, the optimal third-layer parameter is $T = \Sigma_{\mu^*,\mu}\Sigma_{\mu,\mu}^{-1}$, and the problem reduces optimizing the functional

$$F(\mu) = \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}\left[\left\|h_{\mu^*}(x) - \Sigma_{\mu^*,\mu}\Sigma_{\mu,\mu}^{-1}h_\mu(x)\right\|^2\right].$$

Kim & Suzuki (2024) show that, although functional F is non-convex, its landscape is benign on the non-degenerate region where $\Sigma_{\mu,\mu}$ is invertible, i.e., every local minimizer is global, and every non-global critical point is a saddle.

Example B.3 (Mean-field deep ResNet)**.** *Let ρ be the distribution of the input data x, and let $y(x)$ denote the target function. Consider the mean-field continuum model of a deep residual network*

$$\frac{\mathrm{d}}{\mathrm{d}t}X_\mu(x,t) = \int_\Omega f(X_\mu(x,t),\theta)\mu(\theta,t)\mathrm{d}\theta, \quad X_\mu(x,0) = \langle w_2, x\rangle.$$

Here $X_\mu(x,t)$ is the hidden feature at depth $t \in [0,1]$ for input x, $\theta \in \Omega \subset \mathbb{R}^{d_2\times d_2}$ is the parameter of the residual block, where Ω is the parameter space, $\rho(\theta,t)$ is a probability distribution over residual block parameters and depth, $w_2 \in \mathbb{R}^{d_1\times d_2}$ represents the first convolution layer in the ResNet, which extracts features before sending them to the residual blocks, $w_1 \in \mathbb{R}^{d_2}$ is a fixed linear pooling operator that transfers the final feature $X_\mu(x,1)$ to the classification result, and $f(\cdot,\theta)$ is the residual block with parameter θ that aims to learn a feature transformation from $\mathbb{R}^{d_2}$ to $\mathbb{R}^{d_2}$. For instance, $f(z,\theta) = \sigma(\theta z)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function applied component-wise. The associated L^2 loss is

$$F(\mu) = \mathbb{E}_{x\sim\rho}\left[\frac{1}{2}\left(\langle w_1, X_\mu(x,1)\rangle - y(x)\right)^2\right].$$

Lu et al. (2020) prove that whenever $F(\mu) > 0$ there exists $\nu \in \mathcal{P}_2(\Omega)$ such that

$$\int_\Omega \frac{\delta F}{\delta\mu}(\mu,\theta,t)\mathrm{d}(\nu-\mu)(\theta) < 0.$$

Thus, every local minimizer of F must satisfy $F(\mu) = 0$. Since the loss F is non-negative, every local minimizer is therefore a global minimizer.

Example B.4 (Matrix decomposition)**.** *Inspired by the strict benignity of finite-dimensional tensor decomposition Ge et al. (2015), Yamamoto et al. (2025) study the following mean-field version. Let*

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

be a mean-field two-layer neural network, where $z \in \mathbb{R}^l$ is sampled from a given input distribution $\mathcal{D}$, $x = (a,w) \in \mathbb{R}^{k+l}$ denotes the parameter of the network, and σ is an activation function. Given a target measure μ^*, define

$$F(\mu) = \mathbb{E}_{z\sim\mathcal{D}}\left[\left\|h_{\mu^*}(z)h_{\mu^*}(z)^\top - h_\mu(z)h_\mu(z)^\top\right\|_{\mathrm{F}}^2\right],$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. This functional corresponds to learning the rank-one matrix generated by the target representation $h_{\mu^*}(z)h_{\mu^*}(z)^\top$. (Yamamoto et al., 2025, Proposition G.1) proves that for an ε-stationary point μ^* which is not a global minimizer, the objective F can be strictly decreased. This suggests that F possesses benignity.

Proposition B.5 (From qualitative to quantitative saddles)**.** *Let $\mathcal{S} \subset \mathcal{P}_2(\mathbb{R}^d)$ be W_2-compact and suppose that $\mathcal{M}^*$ is nonempty. Assume that the maps $\mu \mapsto \|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu}$ and $\mu \mapsto \underline{\lambda}(\mathrm{K}_\mu)$ are continuous on $\mathcal{S}$. Assume further that every second-order critical point in $\mathcal{S}$ is globally minimizing, i.e., $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} = 0$, and $\underline{\lambda}(\mathrm{K}_\mu) \geq 0$ imply $\mu \in \mathcal{M}^*$. Then, for every $\alpha_{\mathrm{ben}} > 0$, there exists $c_{\alpha_{\mathrm{ben}}} > 0$ such that $\mu \in \mathcal{S}$, $\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) := \inf_{\nu^* \in \mathcal{M}^*} W_2(\mu,\nu^*) \geq \alpha_{\mathrm{ben}}$ implies $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} \geq c_{\alpha_{\mathrm{ben}}}$ or $\underline{\lambda}(\mathrm{K}_\mu) \leq -c_{\alpha_{\mathrm{ben}}}$. Hence, F satisfies the saddle property on $\mathcal{S}$ for any $0 < \varepsilon, \delta < c_{\alpha_{\mathrm{ben}}}$.*

Proof. Define $g(\mu) := \|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu}$, $\chi(\mu) := (-\underline{\lambda}(\mathrm{K}_\mu))_+$, where $a_+ := \max\{a, 0\}$, and set $\Phi(\mu) := \max\{g(\mu), \chi(\mu)\}$. By assumption, g and $\underline{\lambda}$ are continuous on $\mathcal{S}$. Hence, χ and Φ are also continuous on $\mathcal{S}$.

Fix $\alpha_{\mathrm{ben}} > 0$, and define $\mathcal{C}_{\alpha_{\mathrm{ben}}} := \{\mu \in \mathcal{S} : \mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) \geq \alpha_{\mathrm{ben}}\}$. The map $\mu \mapsto \mathrm{dist}_{W_2}(\mu, \mathcal{M}^*)$ is 1-Lipschitz with respect to W_2. Indeed, for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$|\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) - \mathrm{dist}_{W_2}(\nu, \mathcal{M}^*)| \leq W_2(\mu, \nu),$$

which follows directly from the triangle inequality. Therefore, $\mathcal{C}_{\alpha_{\mathrm{ben}}}$ is a closed subset of the compact set $\mathcal{S}$, and is therefore compact. If $\mathcal{C}_{\alpha_{\mathrm{ben}}} = \varnothing$, then the asserted implication is vacuous, and we may choose any $c_{\alpha_{\mathrm{ben}}} > 0$. Hence, suppose that $\mathcal{C}_{\alpha_{\mathrm{ben}}} \neq \varnothing$.

We claim that $\Phi(\mu) > 0$, $\mu \in \mathcal{C}_{\alpha_{\mathrm{ben}}}$. Indeed, suppose for a contradiction that there exists $\mu \in \mathcal{C}_{\alpha_{\mathrm{ben}}}$ such that $\Phi(\mu) = 0$. Since both g and χ are non-negative, this implies $g(\mu) = 0$, and $\chi(\mu) = 0$, i.e., $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} = 0$, $\underline{\lambda}(\mathrm{K}_\mu) \geq 0$. By assumption, it follows that $\mu \in \mathcal{M}^*$.

However, this would imply $\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) = 0$, contradicting $\mu \in \mathcal{C}_{\alpha_{\mathrm{ben}}}$, and $\alpha_{\mathrm{ben}} > 0$. Thus, $\Phi(\mu) > 0$ throughout $\mathcal{C}_{\alpha_{\mathrm{ben}}}$. Since Φ is continuous and $\mathcal{C}_{\alpha_{\mathrm{ben}}}$ is nonempty and compact, Φ attains its minimum on this set. Define

$$c_{\alpha_{\mathrm{ben}}} := \min_{\mu \in \mathcal{C}_{\alpha_{\mathrm{ben}}}} \Phi(\mu).$$

The strict positivity proved above implies $c_{\alpha_{\mathrm{ben}}} > 0$.

Now let $\mu \in \mathcal{S}$ satisfy $\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) \geq \alpha_{\mathrm{ben}}$. Then $\mu \in \mathcal{C}_{\alpha_{\mathrm{ben}}}$, and hence

$$\max\left\{\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu}, (-\underline{\lambda}(\mathrm{K}_\mu))_+\right\} \geq c_{\alpha_{\mathrm{ben}}}.$$

It follows that either $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} \geq c_{\alpha_{\mathrm{ben}}}$, or $(-\underline{\lambda}(\mathrm{K}_\mu))_+ \geq c_{\alpha_{\mathrm{ben}}}$. In the latter case, since $c_{\alpha_{\mathrm{ben}}} > 0$, we necessarily have $-\underline{\lambda}(\mathrm{K}_\mu) \geq c_{\alpha_{\mathrm{ben}}}$, and therefore $\underline{\lambda}(\mathrm{K}_\mu) \leq -c_{\alpha_{\mathrm{ben}}}$. This proves the first assertion.

Finally, let $0 < \varepsilon, \delta < c_{\alpha_{\mathrm{ben}}}$. For every $\mu \in \mathcal{S}$ satisfying $\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) \geq \alpha_{\mathrm{ben}}$, the preceding dichotomy gives either $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} \geq c_{\alpha_{\mathrm{ben}}} > \varepsilon$, or $\underline{\lambda}(\mathrm{K}_\mu) \leq -c_{\alpha_{\mathrm{ben}}} < -\delta$. Hence, for every $\mu \in \mathcal{S}$, at least one of the following holds: $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} > \varepsilon$, $\underline{\lambda}(\mathrm{K}_\mu) < -\delta$, or $\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) < \alpha_{\mathrm{ben}}$. Thus, F satisfies Assumption 4.3 on $\mathcal{S}$ for every $0 < \varepsilon, \delta < c_{\alpha_{\mathrm{ben}}}$. □

B.1 Landscape analysis of a mean-field residual flow

We analyze the specific residual network architecture with rotational weights from Example B.3, which gives a nonconvex objective that globally satisfies all regularity assumptions, possesses critical points with negative curvature, and satisfies the quantitative benignity condition.

Proposition B.6 (A globally regular rotational residual flow)**.** *Consider a mean-field residual flow with scalar parameter $\theta \in \mathbb{R}$, state in $\mathbb{R}^2$, and residual vector field*

$$f(z, \theta) = \theta J z, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Let the input distribution D satisfy $\mathbb{E}_{x \sim D}[xx^\top] = I_2$. With readout e_1 and target $y(x) = x_1$, define

$$F(\mu) := \frac{1}{2}\mathbb{E}_{x \sim D}\left[\left(e_1^\top X_\mu(x, 1) - x_1\right)^2\right].$$

Hence, Assumptions 4.2, 4.4, and 4.5 hold globally. One admissible choice of positive constants is $R_F = \varepsilon_0 = C_{\mathrm{M}} = C_{\mathrm{K}} = L_{\mathrm{M}} = L_{\mathrm{K}} = 1$. *Moreover, every nonglobal critical point has a negative eigenvalue equal to* -1 *in its integral Hessian. Finally, for every* $0 < \alpha_{\mathrm{ben}} < \frac{\pi}{2}$, $0 < \delta < 1$, $0 < \varepsilon < \min\left\{\sin\alpha_{\mathrm{ben}}, \sqrt{1-\delta^2}\right\}$, *Assumption 4.3 holds on all of* $\mathcal{P}_2(\mathbb{R})$.

Proof. Write $m_\mu := \int_{\mathbb{R}} \theta \mathrm{d}\mu(\theta)$. Since $\mu \in \mathcal{P}_2(\mathbb{R})$, its mean m_μ is finite. The parameter distribution $\mu \in \mathcal{P}_2(\mathbb{R})$ is the same at every depth, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} X_\mu(x,t) = \int_{\mathbb{R}} \theta J X_\mu(x,t) \mathrm{d}\mu(\theta), \quad X_\mu(x,0) = x, \quad t \in [0,1].$$

The state equation reduces to

$$\frac{\mathrm{d}}{\mathrm{d}t} X_\mu(x,t) = m_\mu J X_\mu(x,t).$$

Its solution is

$$X_\mu(x,t) = R_{t m_\mu} x, \quad R_s = \begin{pmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{pmatrix}.$$

Therefore, using $\mathbb{E}_{x\sim D}[xx^\top] = I_2$,

$$F(\mu) = \frac{1}{2}\mathbb{E}_{x\sim D}\left[\left((\cos m_\mu - 1)x_1 - \sin m_\mu x_2\right)^2\right] = \frac{1}{2}\left[(\cos m_\mu - 1)^2 + \sin^2 m_\mu\right] = 1 - \cos m_\mu.$$

This proves $0 \le F \le 2$.

For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ and $\gamma \in \Pi(\mu,\nu)$, we have $m_\nu - m_\mu = \int_{\mathbb{R}^2} (\vartheta - \theta)\mathrm{d}\gamma(\theta,\vartheta)$, so $|m_\nu - m_\mu| \le d_\gamma$. In particular, $|m_\nu - m_\mu| \le W_2(\mu,\nu)$. Taylor expansion of $s \mapsto 1 - \cos s$ gives, for every optimal coupling $\gamma \in \Pi_o(\mu,\nu)$,

$$F(\nu) - F(\mu) = \int_{\mathbb{R}^2} \sin m_\mu (\vartheta - \theta) \mathrm{d}\gamma(\theta,\vartheta) + O\left(W_2^2(\mu,\nu)\right).$$

Hence $\nabla_\mu F(\mu,\theta) = \sin m_\mu$. Differentiating this expression in the spatial and measure variables gives $\nabla\nabla_\mu F(\mu,\theta) = 0$, $\nabla_\mu^2 F(\mu,\theta,\vartheta) = \cos m_\mu$. These derivatives have the symmetry and integrability required by Wasserstein regularity.

Let $q_\mu(\theta) \equiv 1$ and denote by P_μ the orthogonal projection onto the constant functions in L^2_μ, i.e., $P_\mu v = \left(\int_{\mathbb{R}} v \,\mathrm{d}\mu\right) q_\mu$. Since $\|q_\mu\|_{L^2_\mu} = 1$, the Hessian components are $\mathrm{M}_\mu = 0$, $\mathrm{K}_\mu = \cos m_\mu P_\mu$. Thus Assumption 4.2 holds for any positive R_F and ε_0. Moreover, $\|\,\mathrm{M}_\mu\,\|_{\mathrm{op}} = 0$, $\|\,\mathrm{K}_\mu\,\|_{\mathrm{HS}} = |\cos m_\mu| \le 1$, which proves Assumption 4.4.

The multiplication component inequality in Assumption 4.5 holds because its left-hand side vanishes. For the integral component, its kernel is constant in both spatial arguments, so

$$\left\|\nabla_\mu^2 F(\nu,\cdot,\cdot) - \nabla_\mu^2 F(\mu,\cdot,\cdot)\right\|_{L^2_{\gamma\otimes\gamma}} = |\cos m_\nu - \cos m_\mu| \le |m_\nu - m_\mu| \le d_\gamma.$$

This proves Assumption 4.5 with the stated constants.

The characterization of $\mathcal{M}^*$ follows from $F(\mu) = 1 - \cos m_\mu$. The global minimizer set is $\mathcal{M}^* = \{\mu \in \mathcal{P}_2(\mathbb{R}) : m_\mu \in 2\pi\mathbb{Z}\}$, and $\mathrm{dist}_{W_2}(\mu, \mathcal{M}^*) = \mathrm{dist}(m_\mu, 2\pi\mathbb{Z})$. For any $\nu \in \mathcal{M}^*$, we get

$$W_2(\mu,\nu) \ge |m_\mu - m_\nu| \ge \mathrm{dist}(m_\mu, 2\pi\mathbb{Z}).$$

Conversely, choose $a \in 2\pi\mathbb{Z}$ nearest to m_μ and set $\nu := (\mathrm{Id} + a - m_\mu)_{\#}\,\mu$. Then $m_\nu = a$, so $\nu \in \mathcal{M}^*$, and the translation coupling gives $W_2(\mu,\nu) \le |a - m_\mu|$. Together, these inequalities prove the asserted distance identity.

The critical points satisfy $\sin m_\mu = 0$. At a nonglobal critical point, $m_\mu = (2k+1)\pi$ for some $k \in \mathbb{Z}$, and therefore $\mathrm{K}_\mu\, q_\mu = -q_\mu$. In particular, these critical points have negative curvature. The objective is nonconvex even along translation geodesics, since at such a measure $F\left((\mathrm{Id} + s)_{\#}\mu\right) = 1 + \cos s$.

Finally, let $r := \mathrm{dist}(m_\mu, 2\pi\mathbb{Z}) \in [0,\pi]$. By the distance identity, $r = \mathrm{dist}_{W_2}(\mu, \mathcal{M}^*)$, and $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} = \sin r$, $\mathrm{K}_\mu\, q_\mu = \cos r q_\mu$. Suppose $r > \alpha_{\mathrm{ben}}$. If $\cos r < -\delta$, then $\lambda(\mathrm{K}_\mu) < -\delta$. Otherwise, $\alpha_{\mathrm{ben}} < r \le \arccos(-\delta)$, and hence

$$\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} = \sin r \ge \min\left\{\sin\alpha_{\mathrm{ben}}, \sqrt{1-\delta^2}\right\} > \varepsilon.$$

This proves Assumption 4.3 holds on $\mathcal{P}_2(\mathbb{R})$. □

Algorithm 1: WSFN for approximate second-order stationarity

Input: Initial measure $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$; thresholds $\varepsilon, \delta > 0$; regularization $\beta_{\mathrm{grad}} > 0$, $\beta_{\mathrm{sad}} = \Theta(\delta^2)$; stepsize $\tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$, $\tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$; perturbation amplitude $\eta > 0$; maximum episode length $n_{\mathrm{out}} \in \mathbb{N}$; required decrease $F_0 > 0$

Output: An approximate second-order critical point $\widehat{\mu}$

$\mu := \mu^0$, $N_{\mathrm{step}} := 0$

while true **do**

  if $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} > \varepsilon$ **then**

    $\mu = \left(\mathrm{Id} - \tau_{\mathrm{grad}}(\mathrm{H}^2_\mu + \beta_{\mathrm{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\nabla_\mu F(\mu, \cdot)\right)_{\#}\mu$

    $N_{\mathrm{step}} = N_{\mathrm{step}} + 1$

  else

    Store the pre-perturbation iterate $\bar{\mu} = \mu$, $\bar{F} = F(\bar{\mu})$.

    Draw a fresh perturbation $\xi \sim \mathcal{N}_{L^2_{\bar{\mu}}}(0, \mathrm{K}^2_{\bar{\mu}})$, and initialize the episode by

$$\nu^0 = (\mathrm{Id} + \eta\xi)_{\#}\,\bar{\mu}.$$

    $\ell = 0$

    while $\ell < n_{\mathrm{out}}$ and $\bar{F} - F(\nu^\ell) < F_0$ **do**

      $\nu^{\ell+1} = \left(\mathrm{Id} - \tau_{\mathrm{sad}}\left(\mathrm{H}^2_{\nu^\ell} + \beta_{\mathrm{sad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}\nabla_\mu F(\nu^\ell, \cdot)\right)_{\#}\nu^\ell$

      $\ell = \ell + 1$

      $N_{\mathrm{step}} = N_{\mathrm{step}} + 1$

    end

    if $\bar{F} - F(\nu^\ell) \ge F_0$ **then**

      Accept the episode output $\mu = \nu^\ell$

    else

      $\widehat{\mu} = \bar{\mu}$

      return $(\widehat{\mu}, N_{\mathrm{step}})$

    end

  end

end

C PARTICLE IMPLEMENTATION OF WSFN

In this appendix, we describe a particle-based realization of the WSFN scheme (1). The goal is to translate the measure-valued Algorithm 1 into an algorithm acting on finitely many particles. This provides a practical approximation of the dynamics analyzed in the paper.

Algorithm 2 implements WSFN using the empirical measure $\mu = N^{-1}\sum_{j=1}^N \delta_{x_j}$. When the empirical Wasserstein gradient norm exceeds ε, all particles are updated using the full Hessian preconditioned direction with parameters $(\tau_{\mathrm{grad}}, \beta_{\mathrm{grad}})$. Otherwise, the algorithm stores the current particle configuration and initiates a perturbed saddle episode, without explicitly testing for negative curvature.

The perturbations are sampled jointly across particles using the covariance kernel C_μ. Writing $k_\mu(x, y) := \nabla^2_\mu F(\mu, x, y)$, this kernel is related to K^2_μ by

$$C_\mu(x, y) = \int_{\mathbb{R}^d} k_\mu(x, z)k_\mu(y, z)^\top \mathrm{d}\mu(z), \quad (\mathrm{K}^2_\mu\, v)(x) = \int_{\mathbb{R}^d} C_\mu(x, y)v(y)\mathrm{d}\mu(y),$$

where the second identity uses the symmetry of k_μ. Thus, C_μ is the matrix-valued covariance kernel whose integral operator is K^2_μ, as in Lemma D.4. For the empirical measure, the sampled Gaussian vector has block covariance

$$[\mathbf{C}_\mu]_{ij} = \frac{1}{N}\sum_{k=1}^N k_\mu(x_i, x_k)k_\mu(x_j, x_k)^\top.$$

The perturbation uses the integral component of the Hessian, whereas the subsequent transport uses the full Hessian H_μ.

After perturbing the stored particles, the algorithm performs at most n_{out} WSFN steps with $\beta_{\mathrm{sad}} = \Theta(\delta^2)$ and $\tau_{\mathrm{sad}} \in (0, 1/2]$. It checks the objective after the perturbation and after every subsequent update, accepting the first configuration whose objective is at least F_0 below the stored pre-perturbation value. If the criterion is never met within the episode budget, the algorithm terminates and returns the stored particles. The counter N_{step} records the transport steps actually performed, including those in the terminal unsuccessful episode.

Algorithm 2: Particle approximation of WSFN

Input: Particles $\{x_j^0\}_{j=1}^N$; thresholds $\varepsilon, \delta > 0$; regularization $\beta_{\mathrm{grad}} > 0$, $\beta_{\mathrm{sad}} = \Theta(\delta^2)$; step size $\tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$, $\tau_{\mathrm{sad}} \in (0, 1/2]$; perturbation amplitude $\eta > 0$; maximum episode length $n_{\mathrm{out}} \in \mathbb{N}$; required decrease $F_0 > 0$

Output: An empirical approximate second-order critical point $\widehat{\mu}$

$x_j := x_j^0$ for $j = 1, \dots, N$, $\mu := \frac{1}{N}\sum_{j=1}^N \delta_{x_j}$, $N_{\mathrm{step}} := 0$

while true **do**

  if $\left(\frac{1}{N}\sum_{j=1}^N \|\nabla_\mu F(\mu, x_j)\|^2\right)^{1/2} > \varepsilon$ **then**

    for $j = 1, \dots, N$ **do**

      $x_j \leftarrow x_j - \tau_{\mathrm{grad}}\left[\left(\mathrm{H}_\mu^2 + \beta_{\mathrm{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}\nabla_\mu F(\mu, \cdot)\right](x_j)$

    end

    $\mu = \frac{1}{N}\sum_{j=1}^N \delta_{x_j}$.

    $N_{\mathrm{step}} \leftarrow N_{\mathrm{step}} + 1$

  else

    Store the pre-perturbation particles $\bar{x}_j = x_j$, $\bar{\mu} = \frac{1}{N}\sum_{j=1}^N \delta_{\bar{x}_j}$, $\bar{F} = F(\bar{\mu})$.

    Draw the jointly Gaussian vector $(\xi(\bar{x}_1), \dots, \xi(\bar{x}_N)) \sim \mathcal{N}(0, \mathbf{C}_{\bar{\mu}})$, where the (i, j) block of $\mathbf{C}_{\bar{\mu}}$ is $[\mathbf{C}_{\bar{\mu}}]_{ij} = C_{\bar{\mu}}(\bar{x}_i, \bar{x}_j)$.

    Initialize the temporary episode particles by

$$y_j^0 = \bar{x}_j + \eta\xi(\bar{x}_j), \quad j = 1, \dots, N,$$

    and set $\nu^0 = \frac{1}{N}\sum_{j=1}^N \delta_{y_j^0}$.

    $\ell = 0$

    while $\ell < n_{\mathrm{out}}$ and $\bar{F} - F(\nu^\ell) < F_0$ **do**

      for $j = 1, \dots, N$ **do**

        $y_j^{\ell+1} = y_j^\ell - \tau_{\mathrm{sad}}\left[\left(\mathrm{H}_{\nu^\ell}^2 + \beta_{\mathrm{sad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}\nabla_\mu F(\nu^\ell, \cdot)\right](y_j^\ell)$

      end

      $\nu^{\ell+1} = \dfrac{1}{N}\displaystyle\sum_{j=1}^N \delta_{y_j^{\ell+1}}$

      $\ell = \ell + 1$

      $N_{\mathrm{step}} = N_{\mathrm{step}} + 1$

    end

    if $\bar{F} - F(\nu^\ell) \ge F_0$ **then**

      $x_j = y_j^\ell$ for $j = 1, \dots, N$

      $\mu = \nu^\ell$

    else

      $\widehat{\mu} = \bar{\mu}$

      return $(\widehat{\mu}, N_{\mathrm{step}})$

    end

  end

end

C.1 NUMERICAL EXPERIMENTS

We now illustrate the particle implementation of WSFN on a set of numerical experiments. These are not intended as a comprehensive empirical evaluation or as evidence of state-of-the-art practical performance. A detailed study of scalable approximations to the WSFN preconditioner is left for future work. The hardware and software used for the reported experiments are an Apple M4 laptop with 10-core CPU and PyTorch 2.11.

Computational cost and future directions. Across the reported experiments, the implemented WSFN variants attain lower or comparable losses at the common update horizons, although this behavior does not consistently translate into an advantage in elapsed time. This pattern is consistent with the additional cost of constructing and applying curvature operators and approximating the action of the regularized inverse square root preconditioner. The numerical illustrations suggest a benefit from preconditioning in terms of progress per update in several examples, while highlighting a computational limitation of the current implementations. Although the primary contribution of this work is theoretical, these observations motivate further investigation of efficient curvature representations and preconditioner applications, together with control of their approximation errors. Showing when the resulting savings translate into improved optimization performance under a common runtime budget remains an important direction for future work.

C.2 REGULARIZED COULOMB MMD: IMPLEMENTATION DETAILS

For the experiment in Section 5, the $M = 400$ target points are sampled independently from the equally weighted mixture $\frac{1}{4}\sum_{\ell=1}^{4}\mathcal{N}(c_\ell, 0.25^2 I_3)$, with $c_1 = (2,0,1)$, $c_2 = (-2,0,-1)$, $c_3 = (0,2,-1)$, $c_4 = (0,-2,1)$.

The $N = 500$ initial particles are independent draws from $\mathcal{N}(0, 0.05^2 I_3)$. Each trial uses a new target sample and initial particle configuration, shared by all methods within that trial. We evaluate (3) with $\rho = 0.05$, including the student and target diagonal terms. The target self-interaction term is constant during optimization and is computed once per trial. The numerical objective uses a smooth regularized kernel instead of the singular Coulomb kernel.

Gradient and full Hessian actions. Equip particle vectors with the norm $\|v\|_N = (N^{-1}\sum_i \|v_i\|^2)^{1/2}$. Writing $F = F_{\rho,N,M}(\cdot; Y)$, the particle gradient is $g = N\nabla_X F$, and the Hessian for this metric is $H_N = N\nabla_X^2 F$. For $u \in \mathbb{R}^3$, define

$$s_\rho(u) = \|u\|^2 + \rho^2, \quad a_\rho(u) = \frac{u}{s_\rho(u)^{3/2}}, \quad B_\rho(u) = s_\rho(u)^{-3/2}\left(I_3 - \frac{3uu^\top}{s_\rho(u)}\right).$$

Then

$$g_i = -\frac{2}{N}\sum_j a_\rho(x_i - x_j) + \frac{2}{M}\sum_a a_\rho(x_i - y_a),$$

and the integral component and full Hessian act as

$$(K_N v)_i = \frac{2}{N}\sum_j B_\rho(x_i - x_j)v_j,$$

$$(H_N v)_i = -\frac{2}{N}\sum_j B_\rho(x_i - x_j)(v_i - v_j) + \frac{2}{M}\sum_a B_\rho(x_i - y_a)v_i. \tag{4}$$

The difference form in (4) cancels student self-interactions exactly. The implementation caches pairwise terms and evaluates these actions without assembling a dense particle Hessian.

Transport directions and scalar control. WGF, isotropic WGF, and PWGF use the direction g. WSFN uses $p_J \approx (H_N^2 + \beta I_{3N})^{-1/2} g$, computed by Lanczos iterations on $H_N^2 + \beta\, \mathrm{I}_{3\mathrm{N}\times 3\mathrm{N}}$ with full reorthogonalization, a maximum of J iterations, and an absolute Lanczos residual stopping tolerance ϵ_{L}. Computations check for nonfinite quantities and violations of the known lower spectral bound

Setting	Value in the experiment
Selected learning rates	WGF, IsoWGF, PWGF: 0.01; WSFN: 0.3
Preconditioner $(\beta, J, \epsilon_{\mathrm{L}})$	$\beta = 10^{-5}$, $J = 12$, $\epsilon_L = 10^{-6}$
Kick scales $(\eta_{\mathrm{iso}}, \eta_{\mathrm{H}})$	$\eta_{\mathrm{iso}} = \eta_{\mathrm{H}} = 0.1$

Table 2: Specific settings for the regularized Coulomb MMD experiment.

β. At its own current particle configuration, the scalar control computes the same p_J but takes the direction

$$p_{\mathrm{sc}} = \begin{cases} \dfrac{\|p_J\|_N}{\|g\|_N} g, & \|g\|_N > 0, \\ 0, & \|g\|_N = 0. \end{cases}$$

It shares WSFN's learning rate, β, Lanczos settings, and perturbation rule. Thus the control matches the local direction norm and incurs the curvature computation needed for that match. It is not a separately tuned inexpensive scalar method.

Perturbations. At updates $100, 200, \ldots$, a vector z with independent $\mathcal{N}(0, I_3)$ particle blocks generates the kick

$$r_{\mathrm{iso}} = \eta_{\mathrm{iso}} z, \quad r_{\mathrm{H}} = \eta_{\mathrm{H}} \|z\|_N \frac{K_N z}{\|K_N z\|_N}.$$

We set $r_{\mathrm{H}} = 0$ when $K_N z = 0$. Isotropic WGF uses r_{iso}, and PWGF, WSFN, and the scalar control use r_{H}. WGF receives no kicks. Within each paired trial, the same Gaussian vector is used at the same kick index, while K_N is evaluated at each method's own state. Both the transport direction and the kick are computed before the update $X^+ = X - \tau p + r$. Away from prescribed events, $r = 0$.

For comparison, the unnormalized theoretical perturbation has the particle realization $\xi_i = \frac{1}{\sqrt{N}} \sum_j \nabla^2_\mu F_\rho(\mu^N, x_i, x_j) z_j$, with $\xi = \sqrt{N} K_N z$, whose block covariance is the kernel associated with $\mathrm{K}^2_{\mu^N}$. The implemented norm normalization changes this distribution, so the kicks above do not have that exact covariance in general. Together with the prescribed timing, this makes the experiment an ablation of the numerical directions and perturbations, instead of implementing the full stopping and episode rules.

Tuning, timing, and reporting. The learning rates of WGF, isotropic WGF, PWGF, and WSFN are selected separately by minimizing the mean terminal relative loss on pilot trials. The pilot search uses equal numbers of candidate rates and a 20-second budget per candidate and pilot seed. The selected rates are fixed for the five paired evaluation trials, each with a 120-second budget per method and no simultaneous update cap. Run order is randomized. The scalar control inherits WSFN's selected rate. The final selected values are recorded in Table 2. Timing includes gradient and curvature calculations, Lanczos iterations, kicks, numerical checks, and diagnostic evaluations. Data generation, the target self-interaction calculation, warm-up iterations, and plotting are excluded. The update panels use a common observed horizon, while the time panels use the common 120-second budget. Gradient panels report $\|g\|_N$, not the norm of the preconditioned direction.

Interpretation. Figure 2 shows that WSFN reaches low relative losses in fewer updates, with faster recovery after the prescribed kicks, and this advantage is also visible in the early elapsed time loss curves. Its lower loss between perturbations relative to the scalar control supports a directional benefit under the shared WSFN settings. This is a conditional ablation since the control has not received an independent learning rate search.

The recurring loss and gradient peaks in the perturbed methods align with the prescribed kick schedule. Faster methods encounter more such events within the same time budget, producing dense fluctuations in the time panels. These excursions exclude a claim of uniform pointwise dominance or monotone loss decrease. WGF has no prescribed kicks and maintains smaller gradient norms for much of the late elapsed time interval, despite WSFN's favorable loss reduction. Hence, the experiment illustrates an advantage in reaching low loss under these settings, without showing uniformly faster convergence to stationarity or empirically verifying the theoretical saddle escape complexity bounds.

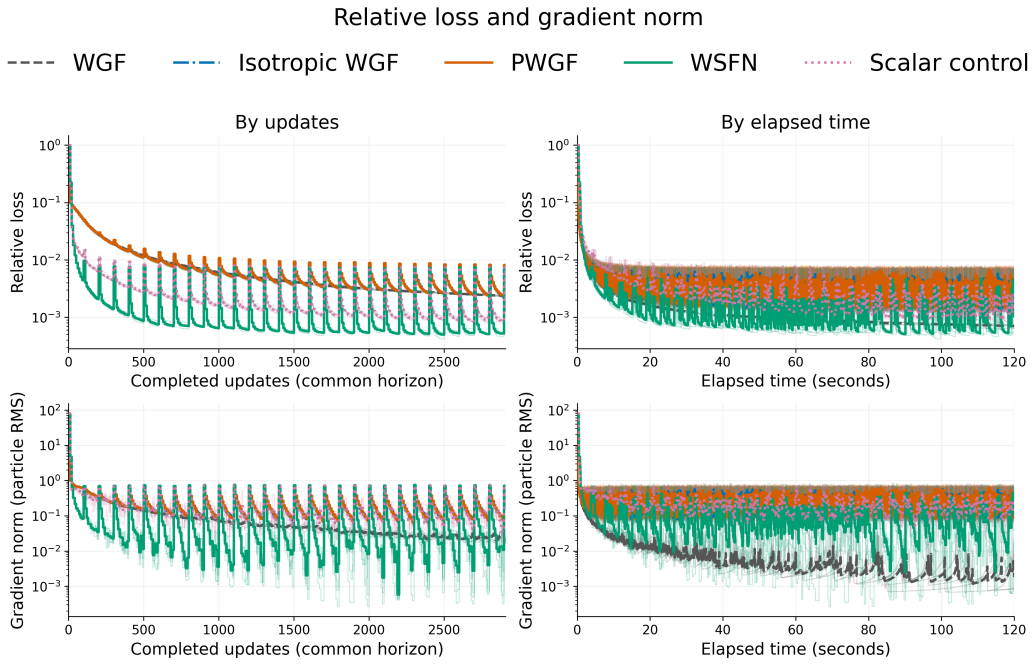


Figure 2: Regularized Coulomb MMD: relative loss (top) and particle RMS gradient norm $\|g\|_N$ (bottom), plotted against completed updates over a common horizon (left) and elapsed time (right). Thick curves show means over five paired trials. Thin curves show individual trials. All vertical axes are logarithmic. The upper panels contain the same loss data as Figure 1.

C.3 In-context feature learning: implementation details

We consider the in-context feature learning objective from Example B.2,

$$F(\mu) = \frac{1}{2}\operatorname{Tr}(\Sigma_{\mu^*,\mu^*}) - \frac{1}{2}\operatorname{Tr}\left(\Sigma_{\mu^*,\mu}\Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\mu^*}\right).$$

The teacher network is fixed, and the student is represented by an empirical measure with $N = 400$ particles. We use input dimension $d = 15$, feature dimension $k = 5$, and $n_{\mathrm{p}} = 300$ training prompts, on which the covariance matrices are estimated. Each evaluation seed specifies a data set and student initialization shared by all methods. The five evaluation seeds are disjoint from the two pilot seeds used for learning rate selection. The teacher seed is 1729, the pilot seeds are $1101, 1102$, and the evaluation seeds are $2101, \ldots, 2105$.

Implemented curvature operator and scalar control. The implemented WSFN variant uses the symmetrized block curvature surrogate $\widehat{H}_N$, which is the full Wasserstein Hessian. With $\beta \in \{10^{-5}, 10^{-4}, 10^{-3}\}$, learned jointly with the learning rate τ, its direction is computed by $p_N \approx (\widehat{H}_N^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2} g_N$, where g_N is the gradient of the implemented loss in stored particle coordinates. We use at most 10 Lanczos iterations with tolerance 10^{-6}. At its own current state, the scalar control replaces p_N by

$$d_N^{\mathrm{sc}} = \frac{\|p_N\|_N}{\|g_N\|_N} g_N, \qquad \|v\|_N^2 = \frac{1}{N}\sum_{i=1}^{N} \|v_i\|^2,$$

with $d_N^{\mathrm{sc}} = 0$ when $g_N = 0$. It inherits WSFN's selected learning rate and preconditioner settings, including the cost of computing p_N, and serves as a diagnostic of the contribution of the preconditioned direction.

Perturbations. For this controlled ablation, perturbed methods receive prescribed kicks every 5 updates. In particle coordinates, the displacements are $\frac{1}{2N} Z$ for isotropic WGF and $0.0015 \widehat{H}_N Z$

Table 3: Learning rates selected from pilot runs for the reported evaluation.

Method	Selected learning rate	Selected β
WGF	0.03	N/a
Isotropic WGF	0.03	N/a
PWGF	0.01	N/a
WSFN	0.000095	$\beta = 0.00001$
Scalar control	Same as WSFN	Same as WSFN

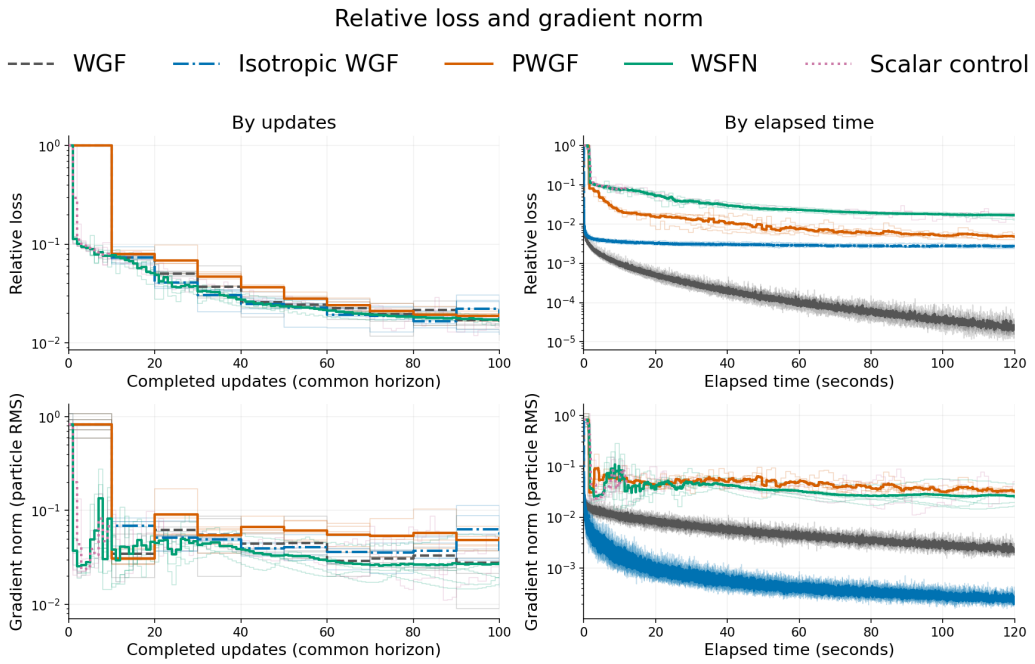


Figure 3: In-context feature learning with $N = 400$ particles over five paired evaluation trials. Top: loss relative to its initial value. Bottom: particle RMS norm of the implemented gradient. Left: completed updates, restricted to the common observed horizon. Right: elapsed time, with a 120-second budget per method and trial. Thick curves show arithmetic means across all five trials. Thin curves show individual trials.

for PWGF, WSFN, and the scalar control, where Z is standard Gaussian. Matching kick indices use common Gaussian draws, with each curvature operator evaluated at its own pre-update state. WGF receives no perturbations.

Evaluation and timing. Let $S, T \in \mathbb{R}^{n_{\mathrm{p}} \times k}$ contain the student and teacher features, respectively. When S has full column rank, a reduced QR factorization $S = QR$ gives the implemented loss

$$L_N = \frac{1}{2n_{\mathrm{p}}} \|T - QQ^\top T\|_{\mathrm{F}}^2.$$

Gradients are obtained by differentiating this residual formulation.

We compare WGF, WGF with isotropic Gaussian perturbations, PWGF, WSFN, and a scalar control. Learning rates are tuned separately for the first four methods using equal sized candidate grids and 20-second pilot runs on two seeds. Selection minimizes the mean relative loss across the pilot seeds. The selected rates (Table 3) are then frozen for five disjoint evaluation trials, each with a 120-second budget per method. Within each trial, all methods share the same prompts and initial student parameters. Run order is randomized. Elapsed time includes optimization, curvature calculations, perturbations. Model and data setup, and disposable warm-up runs are excluded. Post deadline loss improvements receive no credit in the fixed-time comparison.

Interpretation. The ICL experiment shows no clear advantage of the implemented WSFN variant at the common update horizon. WSFN achieves somewhat lower losses than the first-order baselines over parts of the trajectory, but the curves approach one another as the update count increases. The scalar control remains comparable to WSFN and attains a slightly lower mean loss near the endpoint. Together with the variation across individual trials, this comparison provides no clear evidence of an additional benefit from directional preconditioning at the selected settings.

The elapsed time comparison strongly favors the first-order methods. WGF attains a loss almost three orders of magnitude smaller than WSFN within the same runtime. In this implementation, the cost of curvature construction and preconditioning is not compensated by a sufficient improvement per update.

The gradient norm panels are consistent with this comparison. At matched update counts, WSFN reaches a somewhat smaller reported particle RMS gradient norm near the endpoint, but its loss remains comparable to the other methods. Within the elapsed time budget, WGF and isotropic WGF reach substantially smaller gradient norms than WSFN, PWGF, and the scalar control. The perturbed trajectories exhibit continuing fluctuations, so individual peaks should be distinguished from the underlying decrease.

These results identify a limitation of the implemented curvature surrogate method on this ICL instance. Its computational overhead is substantial, and the scalar ablation does not reveal a clear directional advantage. The experiment does not prove a general ordering of the methods or provide direct evidence concerning saddle escape rates or second-order stationarity.

C.4 Mean-field deep residual networks: implementation details.

We consider the mean-field deep ResNet of Example B.3 with residual block $f(z, \Theta) = \tanh(\Theta z)$, where the activation is applied componentwise. We use input dimension $d_1 = 15$, hidden dimension $d_2 = 4$, $L = 8$ depth slices, $N = 24$ matrix-valued particles per slice, and $n = 300$ training samples. Writing $\mu_\ell^N = N^{-1} \sum_{j=1}^N \delta_{\Theta_{\ell j}}$, we discretize the depth dynamics by the explicit Euler scheme

$$X_{r,0} = w_2^\top x_r, \quad X_{r,\ell+1} = X_{r,\ell} + \frac{1}{LN} \sum_{j=1}^N \tanh(\Theta_{\ell j} X_{r,\ell}), \quad \ell = 0, \dots, L-1.$$

The input map $w_2 \in \mathbb{R}^{d_1 \times d_2}$ and pooling vector $w_1 \in \mathbb{R}^{d_2}$ are fixed. Only the residual block matrices $\Theta_{\ell j} \in \mathbb{R}^{d_2 \times d_2}$ are trained, giving $LNd_2^2 = 3072$ scalar parameters. The full-batch empirical objective is

$$F_{L,N}(\Theta) = \frac{1}{2n} \sum_{r=1}^n \left(w_1^\top X_{r,L} - y_r \right)^2. \tag{5}$$

We evaluate this loss directly through its squared residuals.

Data, teacher, and initialization. Inputs are independent samples from $\mathcal{N}(0, I_{d_1})$. A fixed teacher uses the same architecture and Euler grid to generate noiseless targets y_r. We obtain w_2 from the reduced QR decomposition of a standard Gaussian matrix and obtain w_1 by normalizing a standard Gaussian vector. For independent matrices A_0, A_1 with entries distributed as $\mathcal{N}(0, d_2^{-1})$, the teacher particles are

$$\Theta_{\ell j}^* = 0.8 \left[A_0 + 0.4 \sin\left(\frac{2\pi(\ell + 1/2)}{L} \right) A_1 \right] + \frac{0.25}{\sqrt{d_2}} E_{\ell j}^*,$$

where the entries of $E_{\ell j}^*$ are independent standard Gaussians. The teacher and the fixed input and output maps are generated with seed 1729 and held fixed throughout tuning and evaluation. Student particles are initialized as $\Theta_{\ell j}^0 = 0.25 \frac{E_{\ell j}}{\sqrt{d_2}}$, with independent standard Gaussian entries. Each trial draws new inputs and initial particles. All methods within a trial share these draws and targets.

Particle geometry and curvature. We use the discretized tangent metric

$$\langle U, V \rangle_{L,N} = \frac{1}{LN} \sum_{\ell,j} \mathrm{Tr}(U_{\ell j}^\top V_{\ell j}). \tag{6}$$

Hence, the gradient and full Hessian in stored particle coordinates are $G = (LN)\nabla_\Theta F_{L,N}$, $H[V] = (LN)D^2_\Theta F_{L,N}[V]$. Hessian-vector products are computed by automatic differentiation of (5). This differentiates the full discrete objective, including interactions across depth slices.

For the Hessian-guided perturbations, we compute the integral component as $K = H - M$. Let $a_{r,\ell}$ be the discrete adjoint of the unaveraged per-sample loss, with terminal value $a_{r,L} = (w_1^\top X_{r,L} - y_r)w_1$. Holding the states and adjoints fixed when differentiating a block parameter gives

$$(M[V])_{\ell j} = \frac{1}{n}\sum_{r=1}^{n}\left[a_{r,\ell+1} \odot \tanh''(\Theta_{\ell j}X_{r,\ell}) \odot (V_{\ell j}X_{r,\ell})\right] X_{r,\ell}^\top,$$

where $\odot$ denotes componentwise multiplication. Thus K retains both within-depth and cross-depth interactions, including its diagonal particle blocks. Neither a dense Hessian nor a matrix of particle pair curvature blocks is formed during optimization.

Perturbations. We compare WGF, isotropic WGF, PWGF, WSFN, and a scalar control. The first three methods use descent direction G. WSFN uses the approximation $P \approx (H^2 + \beta\,\mathrm{I})^{-1/2}G$. At its own current state, the scalar control computes the same approximate direction P and replaces it by $P_{\mathrm{sc}} = \frac{\|P\|_{L,N}}{\|G\|_{L,N}}G$, with $P_{\mathrm{sc}} = 0$ when $G = 0$. It inherits WSFN's selected learning rate, damping, and perturbation settings and incurs the cost of computing P. This compares direction at a locally matched length.

Perturbations are prescribed every 100 completed updates, beginning at update 100. For a tensor Z with independent standard Gaussian entries, isotropic WGF adds $\eta_{\mathrm{iso}}\sqrt{LN}\,Z$, while PWGF, WSFN, and the scalar control add $\eta_K\sqrt{LN}\,K[Z]$, with $\eta_{\mathrm{iso}} = 10^{-4}$ and $\eta_K = 10^{-3}$. WGF has no perturbations. These increments are added to $\Theta - \tau D$, where D is the method's descent direction. The perturbation amplitude is not multiplied by τ. Both direction and perturbation are evaluated at the pre-update state. The factor $\sqrt{LN}$ makes the covariance operator of $\sqrt{LN}\,K[Z]$ equal to K^2 in the metric (6). Gaussian draws are shared by event index within each paired trial. Perturbations are not normalized to a common length, and their amplitudes remain fixed during tuning.

Lanczos approximation. We apply at most 12 Lanczos steps to $H^2 + \beta\,\mathrm{I}$, using two Hessian-vector products per operator action and two passes of full reorthogonalization. Breakdown is detected with absolute tolerance 10^{-12} and relative tolerance 10^{-9}. The small projected matrix is diagonalized after scaling. Projected eigenvalues violating its lower bound β beyond relative tolerance 10^{-7} trigger a numerical failure.

Parameter selection and evaluation. We tune on pilot seeds 1101 and 1102. Each of WGF, isotropic WGF, and PWGF independently searches the 12 learning rates

$$\tau \in \{10^j,\ 3\cdot 10^j : j = -3, -2, -1, 0, 1, 2\}.$$

WSFN searches 12 joint configurations defined by

$$\beta \in \{10^{-4}, 10^{-3}, 10^{-2}\}, \quad a \in \{0.3, 1, 3, 10\}, \quad \tau = a\sqrt{\beta}.$$

Each configuration receives 20 seconds per pilot seed. Up to three nonfailing finalists per method are then restarted from the original initializations for 120 seconds per seed. The finalist count is the same for every method. Selection minimizes the mean terminal relative loss over these longer pilot runs. The scalar control inherits the selected WSFN configuration and is not independently tuned.

After freezing the configurations, we evaluate on five fresh paired seeds, $2101, \ldots, 2105$, with 120 seconds per method and seed. Method order is randomized within each evaluation seed. Timing includes gradient evaluation, curvature operations, Lanczos iterations, perturbations, and post-update diagnostics. It excludes data construction, initial diagnostics, disposable warm-up, printing, and plotting.

Reported diagnostics. Figure 4 reports relative empirical loss and the particle gradient norm

$$R_s = \frac{F_{L,N}(\Theta_s)}{F_{L,N}(\Theta_0)}, \quad \|G_s\|_{L,N} = \left(\frac{1}{LN}\sum_{\ell,j}\|(G_s)_{\ell j}\|_{\mathrm{F}}^2\right)^{1/2}.$$

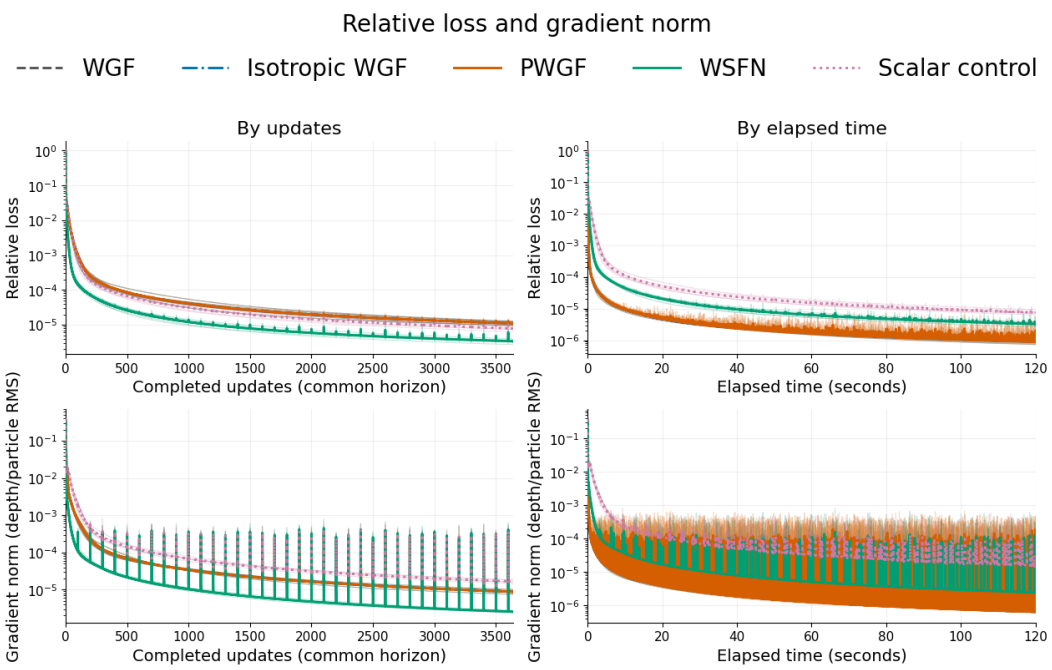


Figure 4: Mean-field deep ResNet regression. Relative empirical loss (top) and gradient norm in the particle metric (bottom), versus completed updates at a common recorded horizon (left) and elapsed time (right). Thick curves show means over five paired evaluation trials. Thin curves show individual trials

The update plots use a common recorded horizon among successful runs, while the time plots use the common 120-second budget. Since perturbations are scheduled by update count, faster methods can experience more perturbations within the same time budget.

This experiment studies the finite-sample, finite-particle, finite-depth implementation with prescribed perturbation times. It does not implement the theoretical stopping and episode rules. The scalar ablation measures the effect of direction at WSFN's selected settings and does not imply superiority over an independently tuned scalar adaptive method. Quantitative control of the discretization and Lanczos errors remains outside the present analysis.

Interpretation. Figure 4 shows that WSFN makes more progress per update than the first-order baselines. WSFN exhibits this advantage early and maintains it throughout the common update horizon. However, the additional cost of curvature computations offsets this improvement in elapsed time. Within 120 seconds, the first-order baselines reach relative losses around 10^{-6}, whereas WSFN remains near 4×10^{-6}. The first-order loss curves largely overlap in this comparison.

The locally norm matched scalar control remains stable and attains an intermediate loss at the common update horizon. Its improvement over the first-order baselines is consistent with scalar rescaling contributing to the observed progress. WSFN achieves a further reduction and outperforms the control both per update and within the elapsed time budget, supporting a benefit from changing the update direction under the selected settings. This comparison is conditional on the control inheriting WSFN's learning rate and damping and matching the preconditioned direction's norm at its own current state.

The gradient norm panels display regular pulses aligned with the prescribed perturbations every 100 updates. Between perturbations, WSFN attains smaller gradient norms at comparable update counts. The corresponding loss increases are small, and descent resumes after each perturbation. Faster methods experience more perturbation events within the same time budget, contributing to the dense fluctuations in the elapsed time panels.

All methods reduce the empirical training loss by several orders of magnitude. These results support improved progress of WSFN per update and a benefit from directional preconditioning at the selected settings, while the elapsed time comparison favors the first-order baselines. The trajectories do not identify a saddle escape episode, and the loss and gradient norm diagnostics alone do not prove faster saddle escape or convergence to a second-order stationary point.

C.5 Matrix decomposition: implementation details

We consider the matrix decomposition problem of Example B.4, in which a student mean-field network learns the rank-one feature matrices of a fixed teacher by minimizing

$$F(\mu) = \frac{1}{n}\sum_{r=1}^{n} \left\| h_\mu(x_r)h_\mu(x_r)^\top - h_{\mu^*}(x_r)h_{\mu^*}(x_r)^\top \right\|_{\mathrm{F}}^2 .$$

Here x_r are the input samples and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

We use input dimension $d = 15$, feature dimension $k = 5$, $n = 300$ samples, and $N = 300$ particles. The particle parameter dimension is $q = d + k = 20$. Inputs are sampled from $\mathcal{N}(0, 9I_d)$ and the teacher is fixed across trials. The four methods' learning rates are selected independently on pilot seeds. Each evaluation seed specifies a data set and student initialization shared by all methods. The five evaluation seeds are disjoint from the two pilot seeds used for learning rate selection. The teacher seed is 1729, the pilot seeds are $1101, 1102$, and the evaluation seeds are $2101, \ldots, 2105$.

Implemented curvature operator. The network uses the sigmoid activation. In the stored coordinates $\theta_i = (a_i, w_i)$, the output weight normalization is absorbed into a_i, so that $h_\theta(x) = \sum_{i=1}^{N} a_i \sigma(w_i^\top x)$. Write $h_r = h_\theta(x_r)$, $h_r^* = h_{\mu^*}(x_r)$, and $J_{ri} = D_{\theta_i} h_\theta(x_r) \in \mathbb{R}^{k\times q}$. The curvature action supplied to Lanczos is

$$(\widehat{H}_N v)_i = \frac{2}{nN}\sum_{r=1}^{n}\sum_{j=1}^{N} J_{ri}^\top B_r J_{rj} v_j, \quad B_r = 2h_r h_r^\top + \|h_r\|^2 I_k - h_r^*(h_r^*)^\top .$$

This kernel type surrogate does not assemble the separate multiplication component involving second derivatives of the particle feature map. Accordingly, the reported experiment concerns this implemented curvature approximation instead of the exact full Hessian discretization. WSFN uses at most 12 Lanczos iterations, with stopping tolerance 10^{-8} and full reorthogonalization.

Scalar control. Let g_N denote the gradient of G in the stored parameter coordinates, and let p_N be the computed Lanczos approximation to $(\widehat{H}_N^2 + \beta\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2} g_N$. With $\|v\|_N^2 = N^{-1}\sum_{i=1}^{N}\|v_i\|^2$, the control uses

$$d_N^{\mathrm{sc}} = \begin{cases} \dfrac{\|p_N\|_N}{\|g_N\|_N}\, g_N, & \|g_N\|_N > 0, \\ 0, & \|g_N\|_N = 0. \end{cases}$$

Thus $\|d_N^{\mathrm{sc}}\|_N = \|p_N\|_N$ at the same state. The control evaluates both vectors at its own current state and inherits WSFN's selected learning rate, regularization, and Lanczos settings. It incurs the preconditioner cost and is used to examine the effect of directional rescaling. Its learning rate is not tuned independently, and the two evolving trajectories need not have identical step lengths.

Perturbations. WGF is unperturbed. The other methods receive a perturbation after every 100 updates, irrespective of loss decrease. At each scheduled event, the methods use common underlying Gaussian blocks $Z_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_q)$. Isotropic WGF adds $3\times 10^{-4} Z$, while PWGF, WSFN, and the scalar control add $3\times 10^{-3}\widehat{H}_N Z$. The transport direction and perturbation are computed from the same pre-update state. The curvature shaped perturbations depend on each method's own state and need not coincide as vectors. This common prescribed schedule serves the ablation and differs from stagnation triggered perturbations and the theoretical episode rules.

Table 4: Learning rates and β selected from pilot runs for the reported evaluation.

Method	Selected learning rate	Selected β
WGF	0.0015	N/a
Isotropic WGF	0.0015	N/a
PWGF	0.0015	N/a
WSFN	0.001	0.0001
Scalar control	Same as WSFN	Same as WSFN

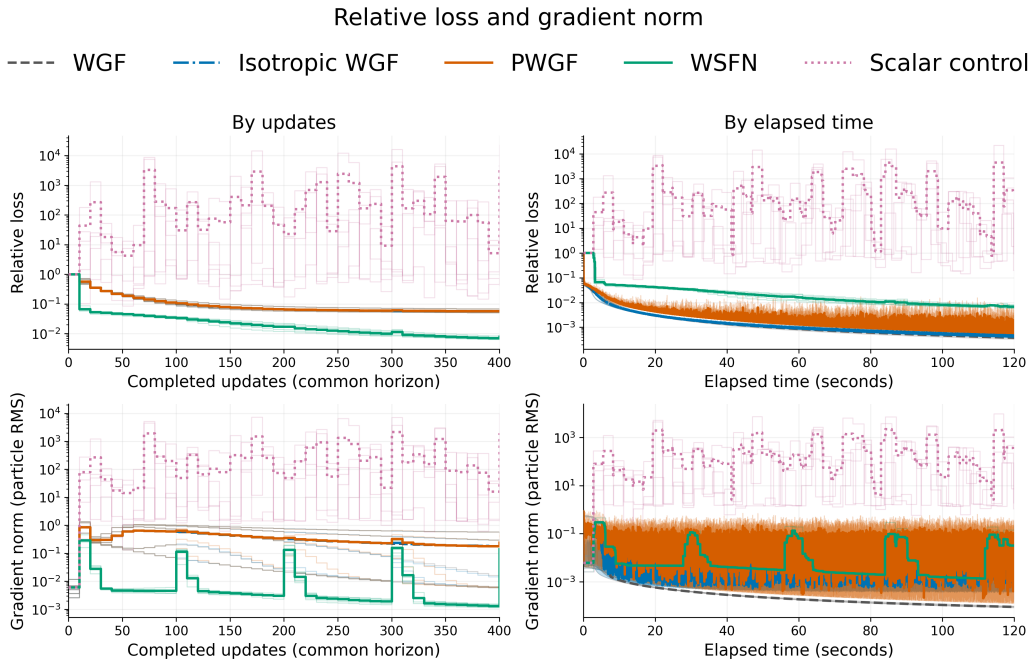


Figure 5: Matrix decomposition ablation with $N = 300$ particles over five paired evaluation trials. Top: loss relative to its initial value. Bottom: particle RMS norm of the implemented gradient. Left: completed updates, restricted to the common observed horizon. Right: elapsed time, with a budget of 120 seconds per method and trial. Thick curves show means. Thin curves show individual trials.

Regularization and learning rate selection. Each of WGF, isotropic WGF, PWGF, and WSFN is tuned independently by minimizing mean relative loss over two pilot seeds, with a 20-second budget per candidate and seed. The initial grids are

$$\mathcal{T}_{\text{first}} = 5 \times 10^{-6}\{0.1, 0.3, 1, 3, 10, 30, 60, 100, 200, 300, 600, 10000\},$$
$$\mathcal{T}_{\text{WSFN}} = 5 \times 10^{-6}\{10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\},$$
$$\beta_{\text{WSFN}} = 10^{-3}\{0.1, 0.3, 1\}$$

The first grid is searched separately for each first-order method. If any selected rate lies at a grid boundary, the search adds one rate below and above every method's grid by a factor of 10, with at most one such expansion. This preserves equal candidate counts across the four methods. The selected rates (Table 4) are frozen before evaluation.

Evaluation and timing. Each method is evaluated for 120 seconds on each of five fresh seeds. Run order is randomized. Elapsed time includes forward and backward passes, curvature actions, Lanczos iterations, perturbations, and diagnostics recorded every 10 updates. Data and model construction, disposable warm-up runs, and plotting and printing are excluded. Only checkpoints completed within the time budget contribute to fixed-time comparisons, and the last observed value is held constant between checkpoints. The iteration plots are restricted to the common observed update horizon.

Interpretation. Figure 5 shows that WSFN makes substantially more progress per update than the first-order baselines. At the common horizon of approximately 400 updates, WSFN decreases the loss rapidly during the initial phase and continues improving after the first-order curves have largely flattened over this update range. However, this advantage does not translate into better performance within the elapsed time budget. PWGF also attains lower losses than WSFN within this budget, although its trajectory exhibits persistent fluctuations. Thus, the additional cost of curvature operations outweighs WSFN's improved progress per update in this implementation.

The locally norm matched scalar control exhibits repeated large excursions in both loss and gradient norm throughout the experiment. The individual trajectories also show substantial variation, with extreme excursions strongly influencing the mean. This comparison supports a contribution from the preconditioned direction to the observed stability. However, the scalar control inherits WSFN's learning rate and damping and matches the preconditioned direction's norm at its own current state. Hence, the comparison does not show superiority over an independently tuned scalar adaptive method or identify negative curvature exploitation as the specific mechanism responsible for the improvement.

The gradient norm panels further illustrate the effect of the prescribed perturbations. WSFN's pulses near updates 100, 200, and 300 coincide with the perturbation schedule. Between later perturbations, its gradient norm decreases to the order of 10^{-3}, substantially below the first-order mean curves at comparable update counts. The accompanying loss increases are modest relative to those of the scalar control. In elapsed time, unperturbed WGF reaches the smallest gradient norm, approximately 10^{-4}. Faster perturbed methods undergo more perturbation events within the same time budget, contributing to the dense fluctuations in their elapsed time curves.

These results support improved progress per update and greater stability than the locally norm matched control at the selected settings, while the runtime comparison favors the first-order baselines. The loss and gradient norm trajectories alone do not imply faster saddle escape or convergence to a second-order stationary point.

D L^2_μ-VALUED GAUSSIAN PROCESSES

In this section, we recall the definition of Gaussian processes (GPs) and the notion of the trace of an operator, which together allow us to define L^2_μ-valued Gaussian processes. Related material can also be found in (Yamamoto et al., 2025, Appendix D). We adopt the following definition of a Gaussian process from (Rasmussen & Williams, 2005, Chapter 2).

Definition D.1 (Gaussian process)**.** *Let $I = \{x_1, x_2, \dots, x_n\}$ be a finite collection of points in $\mathbb{R}^d$. The random variable $\xi : \mathbb{R}^d \to \mathbb{R}^d$ is said to be a Gaussian process if the collection of random variables $\{\xi(x_i)\}_{1\le i\le n}$ is jointly normally distributed. A Gaussian process is completely determined by its mean function m and covariance function K, which are defined by*

$$m(x) = \mathbb{E}[\xi(x)], \quad K(x,y) = \mathbb{E}[(\xi(x) - m(x))(\xi(y) - m(y))^\top].$$

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

We next recall a standard result, adapted from (Reed & Simon, 1980, Theorem VI.18).

Theorem D.2. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\{\psi_n\}_{n=1}^\infty$ be an orthonormal basis in L^2_μ. Then for any nonnegative operator $A_\mu \in \mathcal{B}(L^2_\mu)$ we define $\mathrm{Tr}A = \sum_{n=1}^\infty \langle \psi_n, A_\mu \psi_n\rangle_{L^2_\mu}$. The number $\mathrm{Tr}A_\mu$ is called the trace of A_μ and is independent of the orthonormal basis chosen.*

Definition D.3. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. An operator $A_\mu \in \mathcal{B}(L^2_\mu)$ is called trace-class if and only if $\mathrm{Tr}\,|A_\mu| < \infty$.*

The next result shows that there exists a centered L^2_μ-valued Gaussian random element with covariance operator K^2_μ.

Lemma D.4 (L^2_μ-valued GP)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and assume that $\nabla^2_\mu F(\mu,\cdot,\cdot) \in L^2_{\mu\otimes\mu}$ and that $\nabla^2_\mu F(\mu,x,y) = \nabla^2_\mu F(\mu,y,x)^\top$ for $\mu\otimes\mu$-a.e. (x,y). Then K_μ is self-adjoint and Hilbert–Schmidt, and K^2_μ is nonnegative, symmetric and trace-class. There exists a centered L^2_μ-valued Gaussian*

random element $\xi \sim \mathcal{N}_{L^2_\mu}(0, \mathrm{K}^2_\mu)$. *Its covariance operator has kernel*

$$C_\mu(x,y) = \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu,x,z)\nabla^2_\mu F(\mu,z,y)\mathrm{d}\mu(z)$$

$\mu \otimes \mu$*-a.e.* (x,y) *and*

$$\mathrm{K}^2_\mu[v](x) = \int_{\mathbb{R}^d} C_\mu(x,y)v(y)\mathrm{d}\mu(y),$$

for all $v \in L^2_\mu$. *Moreover,*

$$\mathbb{E}\|\xi\|^2_{L^2_\mu} = \mathrm{Tr}(\mathrm{K}^2_\mu) = \|\nabla^2_\mu F(\mu,\cdot,\cdot)\|^2_{L^2_{\mu\otimes\mu}}.$$

For every $t > 0$,

$$\mathbb{P}\left(\|\xi\|^2_{L^2_\mu} > \mathrm{Tr}(\mathrm{K}^2_\mu) + 2\sqrt{\mathrm{Tr}(\mathrm{K}^4_\mu)t} + 2\|\,\mathrm{K}_\mu\,\|^2_{\mathrm{op}}t\right) \le e^{-t}.$$

Proof. By (Reed & Simon, 1980, Theorem VI.23), since $\nabla^2_\mu F(\mu,\cdot,\cdot) \in L^2_{\mu\otimes\mu}$, the integral operator K_μ is Hilbert–Schmidt with $\|\,\mathrm{K}_\mu\,\|_{\mathrm{HS}} = \|\nabla^2_\mu F(\mu,\cdot,\cdot)\|_{L^2_{\mu\otimes\mu}}$. We also have that $\|\,\mathrm{K}_\mu\,\|_{\mathrm{op}} \le \|\,\mathrm{K}_\mu\,\|_{\mathrm{HS}} = \|\nabla^2_\mu F(\mu,\cdot,\cdot)\|_{L^2_{\mu\otimes\mu}} < \infty$, so K_μ is bounded for every $\mu \in L^2_\mu$. Moreover, K_μ is symmetric for every $\mu \in L^2_\mu$ since $\nabla^2_\mu F(\mu,x,y) = \nabla^2_\mu F(\mu,y,x)^\top$ for $\mu\otimes\mu$-a.e. (x,y). Thus K_μ is self-adjoint for every $\mu \in L^2_\mu$.

Since K_μ is HS, belongs to $\mathcal{B}(L^2_\mu)$, and is self-adjoint, it follows from (Reed & Simon, 1980, Definition, p. 210) that $\mathrm{Tr}(\mathrm{K}^2_\mu) < \infty$. To show that $\mathrm{K}^2_\mu \ge 0$, let $v \in L^2_\mu$. Then

$$\langle v, \mathrm{K}^2_\mu\, v\rangle_{L^2_\mu} = \int_{\mathbb{R}^d} \left\langle v(x), \mathrm{K}^2_\mu\, v(x)\right\rangle \mathrm{d}\mu(x) = \|\,\mathrm{K}_\mu\, v\|^2_{L^2_\mu} \ge 0,$$

which implies that K^2_μ defines a non-negative operator on L^2_μ. Thus we have $|\,\mathrm{K}^2_\mu\,| = \mathrm{K}^2_\mu$, hence by Definition D.3, combined with $\mathrm{Tr}(\mathrm{K}^2_\mu) < \infty$, K^2_μ is trace-class. The symmetry of K^2_μ follows from the symmetry of K_μ.

For the existence of $\xi \sim \mathcal{N}_{L^2_\mu}(0, \mathrm{K}^2_\mu)$, it suffices (cf. (Dudley, 2002, Theorem 12.1.3.)) to show that $C_\mu(x,y) = C_\mu(y,x)^\top$, $\mu\otimes\mu$-a.e., and C_μ is positive semi-definite in the matrix sense, i.e., for any $n \in \mathbb{N}$, any $x_1, ..., x_n \in \mathbb{R}^d$, and any $u_1, ..., u_n \in \mathbb{R}^d$,

$$\sum_{i,j=1}^n u_i^\top C_\mu(x_i,x_j)u_j \ge 0.$$

Using the rule $(AB)^\top = B^\top A^\top$,

$$C_\mu(y,x)^\top = \left(\int_{\mathbb{R}^d} \nabla^2_\mu F(\mu,y,z)\nabla^2_\mu F(\mu,z,x)\mathrm{d}\mu(z)\right)^\top = \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu,z,x)^\top \nabla^2_\mu F(\mu,y,z)^\top \mathrm{d}\mu(z).$$

By regularity of F, we have $\nabla^2_\mu F(\mu,z,x)^\top = \nabla^2_\mu F(\mu,x,z)$ and $\nabla^2_\mu F(\mu,y,z)^\top = \nabla^2_\mu F(\mu,z,y)$, μ-a.e. Hence

$$C_\mu(y,x)^\top = \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu,x,z)\nabla^2_\mu F(\mu,z,y)\mathrm{d}\mu(z) = C_\mu(x,y),$$

which proves $C_\mu(x,y) = C_\mu(y,x)^\top$.

Fix $n \in \mathbb{N}$, $x_1, ..., x_n \in \mathbb{R}^d$ and $u_1, ..., u_n \in \mathbb{R}^d$. Then

$$\begin{aligned}\sum_{i,j=1}^n u_i^\top C_\mu(x_i,x_j)u_j &= \sum_{i,j=1}^n \int_{\mathbb{R}^d} u_i^\top \nabla^2_\mu F(\mu,x_i,z)\nabla^2_\mu F(\mu,z,x_j)u_j \mathrm{d}\mu(z) \\ &= \int_{\mathbb{R}^d} \sum_{i,j=1}^n u_i^\top \nabla^2_\mu F(\mu,x_i,z)\nabla^2_\mu F(\mu,z,x_j)u_j \mathrm{d}\mu(z).\end{aligned}$$

Using the symmetry of $\nabla^2_\mu F$, we can rewrite the integrand as

$$\begin{aligned} u_i^\top \nabla^2_\mu F(\mu, x_i, z) \nabla^2_\mu F(\mu, z, x_j) u_j &= u_i^\top \nabla^2_\mu F(\mu, z, x_i)^\top \nabla^2_\mu F(\mu, z, x_j) u_j \\ &= \big(\nabla^2_\mu F(\mu, z, x_i) u_i\big)^\top \big(\nabla^2_\mu F(\mu, z, x_j) u_j\big). \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{i,j=1}^n u_i^\top C_\mu(x_i, x_j) u_j &= \int_{\mathbb{R}^d} \sum_{i,j=1}^n \big(\nabla^2_\mu F(\mu, z, x_i) u_i\big)^\top \big(\nabla^2_\mu F(\mu, z, x_j) u_j\big) \mathrm{d}\mu(z) \\ &= \int_{\mathbb{R}^d} \left\| \sum_{i=1}^n \nabla^2_\mu F(\mu, z, x_i) u_i \right\|^2 \mathrm{d}\mu(z) \geq 0, \end{aligned}$$

which proves positive semi-definiteness.

It remains to show that $\xi \in L^2_\mu$ almost surely. Using the definition of C_μ and Tonelli,

$$\begin{aligned} \operatorname{Tr} C_\mu(x, x) &= \operatorname{Tr}\left(\int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, z) \nabla^2_\mu F(\mu, z, x) \mathrm{d}\mu(z) \right) \\ &= \int_{\mathbb{R}^d} \operatorname{Tr}\big(\nabla^2_\mu F(\mu, x, z) \nabla^2_\mu F(\mu, z, x)\big) \mathrm{d}\mu(z). \end{aligned}$$

By symmetry $\nabla^2_\mu F(\mu, z, x) = \nabla^2_\mu F(\mu, x, z)^\top$, hence

$$\operatorname{Tr}\big(\nabla^2_\mu F(\mu, x, z) \nabla^2_\mu F(\mu, z, x)\big) = \operatorname{Tr}\big(\nabla^2_\mu F(\mu, x, z) \nabla^2_\mu F(\mu, x, z)^\top\big) = \|\nabla^2_\mu F(\mu, x, z)\|^2.$$

Therefore,

$$\int_{\mathbb{R}^d} \operatorname{Tr} C_\mu(x, x) \mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\nabla^2_\mu F(\mu, x, z)\|^2 \mathrm{d}\mu(z) \mathrm{d}\mu(x) < \infty.$$

For each fixed $x \in \mathbb{R}^d$, $\xi(x)$ is a centered Gaussian vector in $\mathbb{R}^d$ with covariance matrix $C_\mu(x, y)$, hence

$$\mathbb{E}\|\xi(x)\|^2 = \operatorname{Tr}\big(\mathbb{E}[\xi(x)\xi(x)^\top]\big) = \operatorname{Tr} C_\mu(x, x).$$

By Tonelli's theorem,

$$\begin{aligned} \mathbb{E}\|\xi\|^2_{L^2_\mu} &= \mathbb{E} \int_{\mathbb{R}^d} \|\xi(x)\|^2 \mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \mathbb{E}\|\xi(x)\|^2 \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^d} \operatorname{Tr} C_\mu(x, x) \mathrm{d}\mu(x) = \|\nabla^2_\mu F(\mu, \cdot, \cdot)\|^2_{L^2_{\mu\otimes\mu}}. \end{aligned} \tag{7}$$

Hence $\mathbb{E}\|\xi\|^2_{L^2_\mu} < \infty$. Since $\|\xi\|^2_{L^2_\mu} \geq 0$, by Markov's inequality, for any $a > 0$

$$\mathbb{P}(\|\xi\|^2_{L^2_\mu} \geq a) \leq \frac{\mathbb{E}\|\xi\|^2_{L^2_\mu}}{a}.$$

Sending $a \to \infty$ implies $\mathbb{P}(\|\xi\|^2_{L^2_\mu} < \infty) = 1$, so $\|\xi\|_{L^2_\mu} < \infty$ a.s. and $\xi \in L^2_\mu$ a.s.

To verify the relation between K^2_μ and C_μ, let $v \in L^2_\mu$. Then

$$\begin{aligned} \mathrm{K}^2_\mu[v](x) = \mathrm{K}_\mu[\mathrm{K}_\mu[v]](x) &= \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, z)\, \mathrm{K}_\mu[v](z) \mathrm{d}\mu(z) \\ &= \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, z) \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, z, y) v(y) \mathrm{d}\mu(y) \mathrm{d}\mu(z) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, z) \nabla^2_\mu F(\mu, z, y) v(y) \mathrm{d}\mu(y) \mathrm{d}\mu(z) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, z) \nabla^2_\mu F(\mu, z, y) v(y) \mathrm{d}\mu(z) \mathrm{d}\mu(y) \\ &= \int_{\mathbb{R}^d} C_\mu(x, y) v(y) \mathrm{d}\mu(y), \end{aligned}$$

where the penultimate equality follows from Fubini's theorem.

Furthermore, let $\{\psi_n\}_{n=1}^{\infty}$ be an orthonormal basis in L^2_μ. Then

$$\mathrm{Tr}(\mathrm{K}_\mu^2) = \sum_{n=1}^{\infty} \langle \psi_n, \mathrm{K}_\mu^2\, \psi_n \rangle_{L^2_\mu} = \sum_{n=1}^{\infty} \| \mathrm{K}_\mu\, \psi_n \|^2_{L^2_\mu} = \| \mathrm{K}_\mu \|^2_{\mathrm{HS}} = \| \nabla^2_\mu F(\mu, \cdot, \cdot) \|^2_{L^2_{\mu\otimes\mu}},$$

where the last equality follows from the definition of the HS norm. Hence by (7),

$$\mathbb{E}\|\xi\|^2_{L^2_\mu} = \mathrm{Tr}(\mathrm{K}_\mu^2) = \| \nabla^2_\mu F(\mu, \cdot, \cdot) \|^2_{L^2_{\mu\otimes\mu}}.$$

Lastly, to prove the exponential bound we use the Karhunen–Loève representation. If $\mathrm{K}_\mu\, e_j = \lambda_j e_j$ is a spectral decomposition and $Z_j \sim \mathcal{N}(0,1)$ i.i.d., set

$$\xi = \sum_{j\geq 1} \lambda_j Z_j e_j.$$

Since $\sum_{j\geq 1} \lambda_j^2 = \| \mathrm{K}_\mu \|^2_{\mathrm{HS}} < \infty$, due to (Reed & Simon, 1980, Theorem VI.22 (e)), this series converges in L^2_μ and almost surely. Furthermore,

$$\|\xi\|^2_{L^2_\mu} = \sum_{j\geq 1} \lambda_j^2 Z_j^2,$$

so the squared norm is a weighted chi-square random variable. Applying the standard weighted chi-square inequality to finite truncations and passing to the limit gives the displayed tail bound; see (Hsu et al., 2012, Proposition 1). □

Lemma D.5 (Linear functionals of a Gaussian process are Gaussian)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and assume that $\nabla^2_\mu F(\mu, \cdot, \cdot) \in L^2_{\mu\otimes\mu}$ and that $\nabla^2_\mu F(\mu, x, y) = \nabla^2_\mu F(\mu, y, x)^\top$ for $\mu \otimes \mu$-a.e. (x, y). Let $\xi \sim \mathcal{N}_{L^2_\mu}(0, \mathrm{K}_\mu^2)$ be a centered L^2_μ-valued Gaussian random element. Then for any $\psi \in L^2_\mu$ the random variable $\langle \psi, \xi \rangle_{L^2_\mu} \sim \mathcal{N}\big(0, \langle \psi, \mathrm{K}_\mu^2\, \psi \rangle_{L^2_\mu}\big)$.*

Proof. See the proof of (Kim & Suzuki, 2024, Lemma E.9). □

Lemma D.6. *Let $\mu_{\mathrm{sad}} \in \mathcal{P}_2(\mathbb{R}^d)$ be an (ε, δ)-saddle point, i.e., $\|\nabla_\mu F(\mu_{\mathrm{sad}})\|_{L^2_{\mu_{\mathrm{sad}}}} \leq \varepsilon$ and $\lambda_0 := \underline{\lambda}(\mathrm{K}_{\mu_{\mathrm{sad}}}) < -\delta$. Let $\xi \sim \mathcal{N}_{L^2_{\mu_{\mathrm{sad}}}}(0, \mathrm{K}^2_{\mu_{\mathrm{sad}}})$ and $q_0 \in L^2_{\mu_{\mathrm{sad}}}$ be a unit eigenvector of the operator $\mathrm{K}_{\mu_{\mathrm{sad}}}$ corresponding to the eigenvalue λ_0. Define the orthogonal projection of ξ onto the orthogonal complement of the span of q_0 by $\xi_\perp := \xi - \langle q_0, \xi \rangle_{L^2_{\mu_{\mathrm{sad}}}} q_0$. Then $\langle q_0, \xi \rangle_{L^2_{\mu_{\mathrm{sad}}}} \sim \mathcal{N}(0, \lambda_0^2)$, and $\langle q_0, \xi \rangle_{L^2_{\mu_{\mathrm{sad}}}}$ and $\xi_\perp$ are independent.*

Proof. By Lemma D.5, for every $u \in L^2_{\mu_{\mathrm{sad}}}$, the scalar random variable $\langle \xi, u \rangle_{L^2_{\mu_{\mathrm{sad}}}}$ is $\mathcal{N}(0, \langle u, \mathrm{K}^2_{\mu_{\mathrm{sad}}}\, u \rangle_{L^2_{\mu_{\mathrm{sad}}}})$. In particular, by

$$\mathbb{E}\big[\langle \xi, q_0 \rangle^2_{L^2_{\mu_{\mathrm{sad}}}}\big] = \langle \mathrm{K}^2_{\mu_{\mathrm{sad}}}\, q_0, q_0 \rangle_{L^2_{\mu_{\mathrm{sad}}}} = \lambda_0^2 \langle q_0, q_0 \rangle_{L^2_{\mu_{\mathrm{sad}}}} = \lambda_0^2,$$

since $\|q_0\|_{L^2_{\mu_{\mathrm{sad}}}} = 1$. Therefore,

$$\langle \xi, q_0 \rangle_{L^2_{\mu_{\mathrm{sad}}}} \sim \mathcal{N}(0, \lambda_0^2).$$

Next, we show that

$$\mathbb{E}\big[\langle \xi, u \rangle_{L^2_{\mu_{\mathrm{sad}}}} \langle \xi, v \rangle_{L^2_{\mu_{\mathrm{sad}}}}\big] = \langle \mathrm{K}^2_{\mu_{\mathrm{sad}}}\, u, v \rangle_{L^2_{\mu_{\mathrm{sad}}}},$$

for every $u, v \in L^2_{\mu_{\mathrm{sad}}}$. Indeed,

$$\begin{aligned}
\mathbb{E}\big[\langle \xi, u \rangle_{L^2_{\mu_{\mathrm{sad}}}} \langle \xi, v \rangle_{L^2_{\mu_{\mathrm{sad}}}}\big] &= \mathbb{E}\left[\left(\int_{\mathbb{R}^d} \xi(x) u(x) \mathrm{d}\mu_{\mathrm{sad}}(x)\right)\left(\int_{\mathbb{R}^d} \xi(y) v(y) \mathrm{d}\mu_{\mathrm{sad}}(y)\right)\right] \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{E}[\xi(x)\xi(y)] u(x) v(y) \mathrm{d}\mu_{\mathrm{sad}}(x) \mathrm{d}\mu_{\mathrm{sad}}(y) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C_{\mu_{\mathrm{sad}}}(x, y) u(x) v(y) \mathrm{d}\mu_{\mathrm{sad}}(x) \mathrm{d}\mu_{\mathrm{sad}}(y).
\end{aligned}$$

On the other hand, by Lemma D.4,

$$\mathrm{K}^2_{\mu_{\mathrm{sad}}}[u](x) = \int_{\mathbb{R}^d} C_{\mu_{\mathrm{sad}}}(x,y)u(y)\mathrm{d}\mu_{\mathrm{sad}}(y).$$

Therefore,

$$\begin{aligned}\langle \mathrm{K}^2_{\mu_{\mathrm{sad}}} u, v\rangle_{L^2_{\mu_{\mathrm{sad}}}} &= \int_{\mathbb{R}^d} \mathrm{K}^2_{\mu_{\mathrm{sad}}}[u](x)v(x)\mathrm{d}\mu_{\mathrm{sad}}(x)\\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} C_{\mu_{\mathrm{sad}}}(x,y)u(y)\mathrm{d}\mu_{\mathrm{sad}}(y)\right)v(x)\mathrm{d}\mu_{\mathrm{sad}}(x)\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C_{\mu_{\mathrm{sad}}}(x,y)u(y)v(x)\mathrm{d}\mu_{\mathrm{sad}}(y)\mathrm{d}\mu_{\mathrm{sad}}(x).\end{aligned}$$

Renaming the dummy variables x and y, we obtain

$$\langle \mathrm{K}^2_{\mu_{\mathrm{sad}}} u, v\rangle_{L^2_{\mu_{\mathrm{sad}}}} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C_{\mu_{\mathrm{sad}}}(x,y)u(x)v(y)\mathrm{d}\mu_{\mathrm{sad}}(x)\mathrm{d}\mu_{\mathrm{sad}}(y).$$

Hence

$$\mathbb{E}\big[\langle \xi, u\rangle_{L^2_{\mu_{\mathrm{sad}}}}\langle \xi, v\rangle_{L^2_{\mu_{\mathrm{sad}}}}\big] = \langle \mathrm{K}^2_{\mu_{\mathrm{sad}}} u, v\rangle_{L^2_{\mu_{\mathrm{sad}}}}.$$

Let $P_\perp : L^2_{\mu_{\mathrm{sad}}} \to L^2_{\mu_{\mathrm{sad}}}$ denote the orthogonal projection onto the orthogonal complement of the span of q_0, that is,

$$P_\perp v := v - \langle q_0, v\rangle_{L^2_{\mu_{\mathrm{sad}}}} q_0, \quad v \in L^2_{\mu_{\mathrm{sad}}}.$$

Fix $v \in L^2_{\mu_{\mathrm{sad}}}$. Since $\xi_\perp = P_\perp \xi$, we have

$$\langle \xi_\perp, v\rangle_{L^2_{\mu_{\mathrm{sad}}}} = \langle \xi, P_\perp v\rangle_{L^2_{\mu_{\mathrm{sad}}}},$$

due to the fact that orthogonal projections are symmetric operators. Hence

$$\mathrm{Cov}\big(\langle \xi, q_0\rangle_{L^2_{\mu_{\mathrm{sad}}}}, \langle \xi, P_\perp v\rangle_{L^2_{\mu_{\mathrm{sad}}}}\big) = \langle \mathrm{K}^2_{\mu_{\mathrm{sad}}} q_0, P_\perp v\rangle_{L^2_{\mu_{\mathrm{sad}}}} = \lambda_0^2\langle q_0, P_\perp v\rangle_{L^2_{\mu_{\mathrm{sad}}}} = 0,$$

because $P_\perp v$ and q_0 are orthogonal. Now let $(e_k)_{k\geq 1}$ be an orthonormal basis of $L^2_{\mu_{\mathrm{sad}}}$. For every $m \geq 1$, the random vector

$$\big(\langle \xi, q_0\rangle_{L^2_{\mu_{\mathrm{sad}}}}, \langle \xi_\perp, e_1\rangle_{L^2_{\mu_{\mathrm{sad}}}}, ..., \langle \xi_\perp, e_m\rangle_{L^2_{\mu_{\mathrm{sad}}}}\big)$$

is jointly Gaussian, since each component is a continuous linear functional of the Gaussian random element ξ. The computation above shows that

$$\mathrm{Cov}\big(\langle \xi, q_0\rangle_{L^2_{\mu_{\mathrm{sad}}}}, \langle \xi_\perp, e_j\rangle_{L^2_{\mu_{\mathrm{sad}}}}\big) = 0, \quad j = 1, \ldots, m.$$

Therefore, by the characterization of Gaussian vectors, $\langle \xi, q_0\rangle_{L^2_{\mu_{\mathrm{sad}}}}$ is independent of

$$\big(\langle \xi_\perp, e_1\rangle_{L^2_{\mu_{\mathrm{sad}}}}, ..., \langle \xi_\perp, e_m\rangle_{L^2_{\mu_{\mathrm{sad}}}}\big)$$

for every $m \geq 1$. Since $\xi_\perp$ is completely determined by its coordinates $\big(\langle \xi_\perp, e_k\rangle_{L^2_{\mu_{\mathrm{sad}}}}\big)_{k\geq 1}$, it follows that $\langle \xi, q_0\rangle_{L^2_{\mu_{\mathrm{sad}}}}$ is independent of $\xi_\perp$. This proves the claim. □

E Background on Optimal Transport

In this appendix, we recall several standard results from optimal transport that are used throughout the paper. We begin with notation and then state Brenier's theorem, which ensures the existence and structure of optimal transport maps when the source measure is absolutely continuous.

Let $\mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ be the subset of probability measures that are absolutely continuous with respect to the Lebesgue measure. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, define $W_2^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\gamma(x,y)$, where $\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_0)_\#\gamma = \mu, (\pi_1)_\#\gamma = \nu\}$ denotes the set of couplings between μ and ν, with $\pi_0(x,y) = x$ and $\pi_1(x,y) = y$. We denote by $\Pi_o(\mu,\nu)$ the set of optimal couplings. When the optimal coupling is of the form $\gamma = (\mathrm{Id}, \mathrm{T}_\mu^\nu)_\#\mu$ with $\mathrm{T}_\mu^\nu \in L^2_\mu$ satisfying $\mathrm{T}_{\mu\,\#}^\nu\mu = \nu$, we call T_μ^ν the optimal transport map from μ to ν.

Theorem E.1 (Brenier's theorem; (Ambrosio et al., 2008, Theorem 6.2.4)). *Let $\mu \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then:*

- *there exists a unique optimal coupling $\gamma^* := \left(\mathrm{Id}, \mathrm{T}_\mu^\nu\right)_\# \mu$ such that*

$$W_2^2(\mu,\nu) = \int_{\mathbb{R}^d} \left|x - \mathrm{T}_\mu^\nu(x)\right|^2 \mathrm{d}\mu(x),$$

  where $\mathrm{T}_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ is the unique μ-almost everywhere (a.e.) optimal transport map from μ to ν.

- *Moreover, $\mathrm{T}_\mu^\nu(x) = \nabla\varphi(x)$ μ-a.e. for a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ with $\varphi(x) := \frac{1}{2}|x|^2 - \psi(x)$, where ψ is a c-concave function.*[1]

- *If $\nu \in \mathcal{P}_2^\lambda(\mathbb{R}^d)$, then $\gamma^* = (\mathrm{T}_\nu^\mu, \mathrm{Id})_\# \nu$, where $\mathrm{T}_\nu^\mu : \mathbb{R}^d \to \mathbb{R}^d$ is the unique ν-a.e. optimal transport map such that*

$$\mathrm{T}_\nu^\mu \circ \mathrm{T}_\mu^\nu = \mathrm{Id}, \quad \mu\text{-a.e. and } \mathrm{T}_\mu^\nu \circ \mathrm{T}_\nu^\mu = \mathrm{Id}, \quad \nu\text{-a.e.}$$

The next result gives a useful sufficient condition ensuring that a map is optimal.

Theorem E.2 ((Santambrogio, 2015, Theorem 1.48)). *Suppose $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $u : \mathbb{R}^d \to \mathbb{R}$ is a convex and differentiable μ-a.e. and $\int_{\mathbb{R}^d} |\nabla u(x)|^2 \mu(\mathrm{d}x) < \infty$. Then ∇u is an optimal transport map from μ to $(\nabla u)_\# \mu$.*

We conclude with a standard characterization of the tangent velocity of an absolutely continuous curve in Wasserstein space in terms of the infinitesimal behavior of optimal transport.

Proposition E.3 ((Ambrosio et al., 2008, Proposition 8.4.6)). *Let $T > 0$ and let $\mu : [0,T] \to \mathcal{P}_2(\mathbb{R}^d)$ be an absolutely continuous curve satisfying $\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$, $t \in (0,T)$, $\mu|_{t=0} := \mu_0$, with vector field $v_t \in \mathcal{T}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$. Then, for a.e. $t \in (0,T)$, we have*

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

F Differential calculus on Wasserstein space

In this appendix, we collect the differential notions on Wasserstein space used throughout the paper. We begin with Wasserstein subdifferentials, Wasserstein differentiability, and the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$. We then recall how these first-order objects can be expressed in terms of first variations, and how the corresponding second-order structure is described through the Wasserstein Hessian and the second variation of the functional.

F.1 Wasserstein subdifferentials and differentiability

We recall the notions of Wasserstein sub-differentiability and differentiability; see (Ambrosio et al., 2008, Chapter 10), (Carmona & Delarue, 2018, Chapter 5), and Bonnet (2019).

Definition F.1 (Wasserstein sub- and super-differential). *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

- *a map $\xi \in L^2_\mu$ belongs to the sub-differential $\partial^- F(\mu)$ of F at μ if for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$F(\nu) \geq F(\mu) + \sup_{\gamma \in \Pi_o(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \xi(x), y - x \rangle \, \mathrm{d}\gamma(x,y) + o(W_2(\mu,\nu)).$$

  If $\partial^- F(\mu) \neq \emptyset$, we say the function F is Wasserstein sub-differentiable at μ.

- *A map $\xi \in L^2_\mu$ belongs to the super-differential $\partial^+ F(\mu)$ of F at μ if $-\xi \in \partial^-(-F)(\mu)$. If $\partial^+ F(\mu) \neq \emptyset$, we say the function F is Wasserstein super-differentiable at μ.*

[1] A function $\xi : \mathbb{R}^d \to \mathbb{R}$ is c-concave if and only if $x \mapsto \frac{1}{2}|x|^2 - \xi(x)$ is convex and lower semi-continuous.

A functional is said to be Wasserstein differentiable when its sub- and super-differentials intersect.

Definition F.2 (Wasserstein differentiability)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional and let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. We say F is Wasserstein differentiable at μ if $\partial^- F(\mu) \cap \partial^+ F(\mu) \neq \emptyset$. In this case, we say $\nabla_\mu F(\mu) \in \partial^- F(\mu) \cap \partial^+ F(\mu)$ is a Wasserstein gradient of F at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, satisfying for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, and $\gamma \in \Pi_o(\mu, \nu)$,*

$$F(\nu) = F(\mu) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x \rangle \, \mathrm{d}\gamma(x, y) + o(W_2(\mu, \nu)).$$

Let $C_c^\infty(\mathbb{R}^d)$ denote the space of smooth functions with compact support in $\mathbb{R}^d$. The tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla \psi : \psi \in C_c^\infty(\mathbb{R}^d)\}} \subset L^2_\mu,$$

where the closure is taken in L^2_μ (Ambrosio et al., 2008, Definition 8.4.1), (Carmona & Delarue, 2018, Section 5.4.3). Moreover, (Lanzetti et al., 2025, Proposition 2.11) shows that whenever F is Wasserstein differentiable, there exists a unique Wasserstein gradient in $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$, and we may therefore restrict attention to this canonical representative.

The next result, due to (Lanzetti et al., 2025, Proposition 2.12), shows that Wasserstein gradients provide a first-order expansion even along perturbations that are not induced by optimal transport plans.

Proposition F.3 ((Lanzetti et al., 2025, Proposition 2.12))**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in \Pi(\mu, \nu)$ and let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be Wasserstein differentiable at μ with Wasserstein gradient $\nabla_\mu F(\mu) \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$F(\nu) = F(\mu) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x \rangle \mathrm{d}\gamma(x, y) + o\left(\sqrt{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y)} \right).$$

Together, these notions provide the first-order differential structure used throughout the paper.

F.2 Examples of Wasserstein differentiable functionals

We now give examples of Wasserstein (sub)-differentiable functionals.

Example F.4 (The L^2-Wasserstein distance; (Lanzetti et al., 2025, Proposition 2.17), (Ambrosio et al., 2008, Section 9.3, 10.4))**.** *Fix $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, and for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ consider the functional $\mathcal{W}(\mu) = \frac{1}{2} W_2^2(\mu, \nu)$. If $\mu \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$, the functional $\mathcal{W}$ is Wasserstein differentiable at μ with Wasserstein gradient given by $\nabla_\mu \mathcal{W}(\mu) = \mathrm{Id} - \mathrm{T}_\mu^\nu$, where T_μ^ν is the (unique) optimal transport map from μ to ν.*

Example F.5 (Potential and interaction energies; (Lanzetti et al., 2025, Proposition 2.20 and 2.21))**.** *The potential energy $\mathcal{V}(\mu) = \int_{\mathbb{R}^d} V(x) \mathrm{d}\mu(x)$ and interaction energy $\mathcal{U}(\mu) = \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} U(x - y) \mathrm{d}\mu(x) \mathrm{d}\mu(y)$ with $V : \mathbb{R}^d \to \mathbb{R}$ and $U : \mathbb{R}^d \to \mathbb{R}$ in $C^2(\mathbb{R}^d)$ and with uniformly bounded Hessian are Wasserstein differentiable. In particular, their Wasserstein gradients are $\nabla_\mu \mathcal{V}(\mu) = \nabla V$ and $\nabla_\mu \mathcal{U}(\mu) = \nabla U \star \mu := \int_{\mathbb{R}^d} U(\cdot - y) \mathrm{d}\mu(y)$, provided that $\nabla U \star \mu \in L^2_\mu$. The Maximum mean Discrepancy (MMD) functional, which may expressed as a sum of the potential energy $\mathcal{V}$ and interaction energy $\mathcal{U}$ is also Wasserstein differentiable Arbel et al. (2019).*

Example F.6 (The entropy functional; (Lanzetti et al., 2025, Proposition 2.25))**.** *The entropy functional defined as $\mathcal{H}(\mu) = \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}x}(x) \log \frac{\mathrm{d}\mu}{\mathrm{d}x}(x) \mathrm{d}x$ for $\mu \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ is not Wasserstein differentiable but sub-differentiable (Lanzetti et al., 2025, Proposition 2.25). To guarantee that the Wasserstein sub-gradient of $\mathcal{H}$ is non-empty, $\frac{\mathrm{d}\mu}{\mathrm{d}x}$ is required to satisfy Sobolev regularity cf. (Ambrosio et al., 2008, Theorem 10.4.13). Then, if $\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}x} \in L^2_\mu$, the only sub-gradient of $\mathcal{H}$ in the tangent space is $\nabla_\mu \mathcal{H}(\mu) = \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}x}$; see (Ambrosio et al., 2008, Theorem 10.4.17). Thus, the relative entropy $\mathrm{KL}(\cdot|\rho)$ with respect to a fixed target measure $\rho \propto e^{-V}$, which is the sum of the potential energy $\mathcal{V}$ and entropy term $\mathcal{H}$, is Wasserstein sub-differentiable (Lanzetti et al., 2025, Example 2.29).*

F.3 Wasserstein Hessian along geodesics

We next recall the Wasserstein Hessian along geodesics (Villani, 2003, Chapter 8), which plays the role of the second-order differential of F in Wasserstein space and characterizes the curvature of F along geodesic curves.

Definition F.7 (Wasserstein Hessian along geodesics)**.** *The Wasserstein Hessian of F is an operator over $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d) \subset L^2_\mu$ defined by*

$$\mathrm{H}_\mu[v_0](x) := \nabla\nabla_\mu F(\mu, x)v_0(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, \bar{x})v_0(\bar{x})\mathrm{d}\mu(\bar{x}),$$

verifying $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\big|_{t=0} F(\mu_t) = \langle \mathrm{H}_\mu\, v_0, v_0\rangle_{L^2_\mu}$ if $(\mu_t, v_t)_{t\in[0,1]}$ is a Wasserstein geodesic starting at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, with $v_t \in \mathcal{T}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$ satisfying the flow equation $\partial_t\mu_t + \nabla\cdot(v_t\mu_t) = 0$.

Definition F.7 and Proposition 2.3 can be used to compute the Wasserstein Hessian of several standard functionals, including the Kullback–Leibler divergence Otto & Villani (2000), the maximum mean discrepancy Arbel et al. (2019), and the kernel Stein discrepancy Korba et al. (2021). Additional examples are provided in Example F.13 and Example F.14.

F.4 First and second variation

Under suitable regularity assumptions, the Wasserstein gradient and Hessian of a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ can be expressed in terms of its first and second variations. In this subsection, we recall these notions and the assumptions under which the Wasserstein differential structure of F can be identified with the Euclidean derivatives of $\frac{\delta F}{\delta \mu}$ and $\frac{\delta^2 F}{\delta \mu^2}$. We begin with the first variation (Santambrogio, 2015, Definition 7.12).

Definition F.8 (First variation)**.** *We say $F \in \mathcal{C}^1$, i.e., it admits a first variation at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, if there exists a jointly continuous function $\frac{\delta F}{\delta \mu} : \mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}$, unique up to an additive constant, called the first variation of F, for which there exists $\kappa_1 > 0$ such that, for all $(\mu, x) \in \mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d$, $\|\frac{\delta F}{\delta \mu}(\mu, x)\| \le \kappa_1\left(1 + \|x\|^2\right)$, and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\lim_{t\searrow 0} \frac{F(\mu + t(\nu - \mu)) - F(\mu)}{t} = \int_{\mathbb{R}^d} \frac{\delta F}{\delta \mu}(\mu, x)\mathrm{d}(\nu - \mu)(x),$$

and $\int_{\mathbb{R}^d} \frac{\delta F}{\delta \mu}(\mu, x)\mathrm{d}\mu(x) = 0$.

The next assumption ensures that the Wasserstein gradient of F can be identified with the Euclidean gradient of its first variation.

Assumption F.9. *Assume that $F \in \mathcal{C}^1$, the map $\mathbb{R}^d \ni x \mapsto \frac{\delta F}{\delta \mu}(\mu, x) \in \mathbb{R}$ is differentiable, there exists $\kappa_3 > 0$ such that $\|\nabla_x \frac{\delta F}{\delta \mu}(\mu, x)\| \le \kappa_3(1 + \|x\|)$, for all (μ, x), and the map $\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d \ni (\mu, x) \mapsto \nabla \frac{\delta F}{\delta \mu}(\mu, x) \in \mathbb{R}^d$ is jointly continuous.*

Under Assumption F.9, it follows from (Carmona & Delarue, 2018, Proposition 5.48; Theorem 5.64) that F is Wasserstein differentiable, with $\nabla_\mu F(\mu, \cdot) = \nabla \frac{\delta F}{\delta \mu}(\mu, \cdot)$. See also (Cardaliaguet et al., 2019, Proposition 2.2.3) and (Chewi et al., 2025, Proposition 5.10).

The same idea extends to second order. If, for each fixed $x \in \mathbb{R}^d$, the map $\mu \mapsto \frac{\delta F}{\delta \mu}(\mu, x)$ admits a first variation, then one may define the second variation of F (Cardaliaguet et al., 2019, Section 2.2.2).

Definition F.10 (Second variation)**.** *We say $F \in \mathcal{C}^2$, i.e., it admits a second variation $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, if there exists a jointly continuous function $\frac{\delta^2 F}{\delta \mu^2} : \mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$, unique up to an additive term $\phi(x) \in \mathbb{R}$, called the second variation F, for which there exists $\kappa_2(x) > 0$ such that, for all $(\mu, x') \in \mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d$, $\left|\frac{\delta F}{\delta \mu}(\mu, x, x')\right| \le \kappa_2(x)\left(1 + \|x'\|^2\right)$, and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\lim_{t\searrow 0} \frac{1}{t}\left(\frac{\delta F}{\delta \mu}(\mu + t(\nu - \mu), x) - \frac{\delta F}{\delta \mu}(\mu, x)\right) = \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta \mu^2}(\mu, x, x')\mathrm{d}(\nu - \mu)(x'),$$

The next assumption guarantees that the Wasserstein Hessian of F can be computed from the Euclidean derivatives of its first and second variations. Let C^2 denote the space of twice continuously differentiable functions.

Assumption F.11. *Assume that $F \in \mathcal{C}^2$, the maps $\mathbb{R}^d \ni x \mapsto \frac{\delta F}{\delta \mu}(\mu, x)$, $\mathbb{R}^d \times \mathbb{R}^d \ni (x, x') \mapsto \frac{\delta^2 F}{\delta \mu^2}(\mu, x, x')$ are C^2, $\|\nabla^2 \frac{\delta F}{\delta \mu}(\mu, \cdot)\|_{\mu,\infty} < \infty$, $\nabla_{x'} \nabla_x \frac{\delta^2 F}{\delta \mu^2}(\mu, \cdot, \cdot) \in L^2_{\mu \otimes \mu}$, and, for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\frac{\delta}{\delta \mu}\left(\nabla \frac{\delta F}{\delta \mu}(\mu, \cdot)\right)(\cdot) = \nabla \frac{\delta^2 F}{\delta \mu^2}(\mu, \cdot, \cdot)$.*

The identity $\frac{\delta}{\delta \mu}\left(\nabla \frac{\delta F}{\delta \mu}(\mu, \cdot)\right)(\cdot) = \nabla \frac{\delta^2 F}{\delta \mu^2}(\mu, \cdot, \cdot)$ is used when differentiating the Wasserstein gradient $\nabla_\mu F(\mu_t, \cdot) = \nabla_x \frac{\delta F}{\delta \mu}(\mu_t, \cdot)$ along a transport curve $\mu_t = (\mathrm{Id} + tv)_{\#}\mu$. Under the C^2 regularity assumptions above, Schwarz's theorem also implies that the resulting Hessian operator is symmetric on L^2_μ.

Proposition F.12 (Wasserstein Hessian via first and second variations of F)**.** *Let Assumption F.9, F.11 hold. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$ and for $t \in [0,1]$ define the curve $\mu_t = (\pi_t)_{\#}\mu$, with $\pi_t := \mathrm{Id} + tv$. Then, for any $t \in [0,1]$,*

$$\frac{\mathrm{d}}{\mathrm{d}t} \nabla_z \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)) = \underbrace{\nabla_z^2 \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)) v(x)}_{= \widetilde{\mathrm{M}}_{\mu,t}[v](x)} + \underbrace{\int_{\mathbb{R}^d} \nabla_{\bar{z}} \nabla_z \frac{\delta^2 F}{\delta \mu^2}(\mu_t, \pi_t(x), \pi_t(\bar{x})) v(\bar{x}) \mathrm{d}\mu(\bar{x})}_{= \widetilde{\mathrm{K}}_{\mu,t}[v](x)}.$$

Hence,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t) = \left\langle (\widetilde{\mathrm{M}}_{\mu,t} + \widetilde{\mathrm{K}}_{\mu,t}) v, v \right\rangle_{L^2_\mu} = \left\langle \widetilde{\mathrm{H}}_{\mu,t} v, v \right\rangle_{L^2_\mu}.$$

Proof. Applying Lemma L.2 to the map $\mu \mapsto \partial_{z_i} \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x))$ for $1 \le i \le d$ with fixed $\pi_t(x)$ gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\partial_{z_i} \frac{\delta F}{\delta \mu}(\mu, \pi_t(x))\right)\Big|_{\mu=\mu_t} &= \int_{\mathbb{R}^d} \left\langle \nabla_\mu \left(\partial_{z_i} \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x))\right)(\pi_t(\bar{x})), v(\bar{x}) \right\rangle \mathrm{d}\mu(\bar{x}) \\
&= \int_{\mathbb{R}^d} \left\langle \nabla_{\bar{z}} \frac{\delta}{\delta \mu}\left(\partial_{z_i} \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x))\right)(\pi_t(\bar{x})), v(\bar{x}) \right\rangle \mathrm{d}\mu(\bar{x}) \\
&= \int_{\mathbb{R}^d} \left\langle \nabla_{\bar{z}}\, \partial_{z_i} \frac{\delta^2 F}{\delta \mu^2}(\mu_t, \pi_t(x), \pi_t(\bar{x})), v(\bar{x}) \right\rangle \mathrm{d}\mu(\bar{x}).
\end{aligned}$$

Then, noting that $\partial_t \pi_t(x) = v(x)$ and using the chain rule,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \partial_{z_i} \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\partial_{z_i} \frac{\delta F}{\delta \mu}(\mu_t, u(x))\right)\Big|_{u=\pi_t} + \frac{\mathrm{d}}{\mathrm{d}t}\left(\partial_{z_i} \frac{\delta F}{\delta \mu}(\mu, \pi_t(x))\right)\Big|_{\mu=\mu_t} \\
&= \left\langle \nabla_z \partial_{z_i} \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)), v(x) \right\rangle \\
&\quad + \int_{\mathbb{R}^d} \left\langle \nabla_{\bar{z}}\, \partial_{z_i} \frac{\delta^2 F}{\delta \mu^2}(\mu_t, \pi_t(x), \pi_t(\bar{x}), v(\bar{x}) \right\rangle \mathrm{d}\mu(\bar{x}).
\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t} \nabla_z \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)) = \nabla_z^2 \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)) v(x) + \int_{\mathbb{R}^d} \nabla_{\bar{z}} \nabla_z \frac{\delta^2 F}{\delta \mu^2}(\mu_t, \pi_t(x), \pi_t(\bar{x})) v(\bar{x}) \mathrm{d}\mu(\bar{x}).$$

Therefore, by Lemma L.2,

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\left\langle \nabla_z \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)), v(x)\right\rangle \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^d}\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\nabla_z \frac{\delta F}{\delta \mu}(\mu_t, \pi_t(x)), v(x)\right\rangle \mathrm{d}\mu(x) \\ &= \left\langle ((\widetilde{\mathrm{M}}_{\mu,t} + \widetilde{\mathrm{K}}_{\mu,t})v, v\right\rangle_{L^2_\mu} = \left\langle \widetilde{\mathrm{H}}_{\mu,t}v, v\right\rangle_{L^2_\mu}.\end{aligned}$$

□

We conclude with two examples from Bonet et al. (2024) for which the assumptions above hold and Proposition F.12 yields an explicit expression for the Wasserstein Hessian.

Example F.13 (Potential energy; Bonet et al. (2024))**.** *Let $\mathcal{V}(\mu) = \int_{\mathbb{R}^d} V(x)\mathrm{d}\mu(x)$ with $V \in C^2$ with uniformly bounded Hessian. Then, we have $\frac{\delta \mathcal{V}}{\delta \mu}(\mu, x) = V(x)$ and $\frac{\delta^2 \mathcal{V}}{\delta \mu^2}(\mu, x, x') = 0$ (using Definition F.8, F.10). Thus, applying Proposition F.12, we recover for $\mu_t = (\pi_t)_{\#}\mu$,*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{V}(\mu_t) = \int_{\mathbb{R}^d}\left\langle \nabla^2 V(\pi_t(x))v(x), v(x)\right\rangle \mathrm{d}\mu(x).$$

Example F.14 (Interaction energy; Bonet et al. (2024))**.** *Let $\mathcal{U}(\mu) = \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} U(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y)$ with $U \in C^2$ symmetric and with uniformly bounded Hessian. Then, we have for all $x, y \in \mathbb{R}^d$, $\frac{\delta \mathcal{U}}{\delta \mu}(\mu, x) = (U \star \mu)(x)$ and $\frac{\delta^2 \mathcal{U}}{\delta \mu^2}(\mu, x, y) = U(x-y)$. Therefore,*

$$\nabla_x^2 \frac{\delta \mathcal{U}}{\delta \mu}(\mu, x) = \nabla_x^2 (U \star \mu)(x) = \int_{\mathbb{R}^d} \nabla_x^2 U(x-y)\mathrm{d}\mu(y),$$

$$\nabla_y \nabla_x \frac{\delta^2 \mathcal{U}}{\delta \mu^2}(\mu, x, y) = -\nabla_x^2 U(x-y).$$

Thus, applying Proposition F.12, for $\mu_t = (\pi_t)_{\#}\mu$, the Hessian operator is

$$\widetilde{\mathrm{H}}_{\mu,t}[v](x) = \int_{\mathbb{R}^d} \nabla^2 U(\pi_t(x) - \pi_t(\bar{x}))(v(x) - v(\bar{x}))\mathrm{d}\mu(\bar{x}),$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{U}(\mu_t) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left\langle \nabla^2 U(\pi_t(x) - \pi_t(\bar{x}))(v(x) - v(\bar{x})), v(x)\right\rangle \mathrm{d}\mu(\bar{x})\mathrm{d}\mu(x).$$

G Convexity in Wasserstein space

In this appendix, we recall the two notions of convexity on Wasserstein space that are most relevant to the paper: (i) convexity along Wasserstein geodesics and (ii) convexity along linear interpolations of measures. Although both notions play an important role in optimization over probability measures, they are generally distinct and lead to different classes of objective functionals.

G.1 Geodesic convexity

We begin with the notion of convexity along Wasserstein geodesics, which is the natural analogue of convexity induced by the geometry of the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ (Ambrosio et al., 2008, Definition 9.1.1).

Definition G.1. *We say a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is α-convex along geodesics, for $\alpha \geq 0$, if for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists an optimal transport plan $\gamma^* \in \Pi_o(\mu, \nu)$ such that*

$$F(\mu_t) \leq (1-t)F(\mu) + tF(\nu) - \alpha\frac{t(1-t)}{2}W_2^2(\mu, \nu), \tag{8}$$

for all $t \in [0, 1]$, where $\mu_t = (\pi_t)_{\#}\gamma^* := ((1-t)x + ty)_{\#}\gamma^*$ is a Wasserstein constant-speed geodesic between μ and ν.

A canonical example of geodesic convexity is provided by the relative entropy with respect to a log-concave reference measure.

Example G.2 (The relative entropy)**.** *Suppose $\rho(\mathrm{d}x) \propto e^{-U(x)}\mathrm{d}x \in \mathcal{P}_2(\mathbb{R}^d)$, with a convex potential $U : \mathbb{R}^d \to \mathbb{R}$. In this case, we say ρ is a log-concave measure. Thus by (Ambrosio et al., 2008, Theorem 9.4.10) the relative entropy $\mathrm{KL}(\cdot|\rho)$ is geodesically convex, and hence*

$$\mathrm{KL}\left(\mu_t|\rho\right) \leq (1-t)\,\mathrm{KL}(\mu|\rho) + t\,\mathrm{KL}(\nu|\rho).$$

The following proposition, adapted from (Villani, 2008, Proposition 16.2), collects several equivalent ways of characterizing geodesic convexity.

Proposition G.3. *Assume $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma^* \in \Pi_o(\mu,\nu)$ and $\mu_t = (\pi_t)_{\#}\gamma^*$, for all $t \in [0,1]$, where $\pi_t = (1-t)x + ty$. Then, the following statements are equivalent:*

1. *For all $t \in [0,1]$, $F(\mu_t) \leq (1-t)F(\mu) + tF(\nu)$, i.e., F is convex along the geodesic $t \mapsto \mu_t$.*
2. *For all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we have*
$$F(\nu) \geq F(\mu) + \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x\rangle \,\mathrm{d}\gamma^*(x,y).$$
3. *For all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we have*
$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_\mu F(\nu, y) - \nabla_\mu F(\mu, x), y - x\rangle \,\mathrm{d}\gamma^*(x,y) \geq 0.$$
4. *For all $s \in [0,1]$, $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=s} F(\mu_t) \geq 0$.*

Proof. (1) $\implies$ (2). Let $t \in (0,1)$. Then

$$F(\mu_t) \leq (1-t)F(\mu) + tF(\nu) \iff \frac{F(\mu_t) - F(\mu)}{t} \leq F(\nu) - F(\mu).$$

Passing to the limit $t \to 0$ and using Lemma L.1, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} F(\mu_t) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x\rangle \,\mathrm{d}\gamma^*(x,y) \leq F(\nu) - F(\mu).$$

(2) $\implies$ (3). Replace the geodesic μ_t with the geodesic μ_{1-t} in (2) to get

$$F(\mu) \geq F(\nu) + \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_\mu F(\nu, y), x - y\rangle \,\mathrm{d}\gamma^*(x,y).$$

By adding up the inequality in (2), we obtain (3).

(3) $\implies$ (4). First, we have

$$\begin{aligned}\int_0^1 \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t)\mathrm{d}t &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=1} F(\mu_t) - \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} F(\mu_t) \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_\mu F(\nu, y) - \nabla_\mu F(\mu, x), y - x\rangle \,\mathrm{d}\gamma^*(x,y) \geq 0.\end{aligned}$$

Fix $s \in (0,1)$ and let $\varepsilon > 0$ be such that $s + \varepsilon \leq 1$. Define the rescaled curve $\nu_t^\varepsilon := \mu_{s+\varepsilon t}$, for $t \in [0,1]$. Then by the chain rule

$$0 \leq \frac{1}{\varepsilon^2}\int_0^1 \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\nu_t^\varepsilon)\mathrm{d}t = \int_0^1 \frac{\mathrm{d}^2}{\mathrm{d}t^2}\bigg|_{r=s+\varepsilon t} F(\mu_r)\mathrm{d}t. \tag{9}$$

Because $t \mapsto \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t)$ is continuous on $[0,1]$, it is bounded in a neighborhood of s, thus there exists $M > 0$ such that $\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\big|_{r=s+\varepsilon t} F(\mu_r)\right| \le M$ for all $t \in [0,1]$ and all sufficiently small $\varepsilon > 0$. Moreover, for each fixed $t \in [0,1]$ we have $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\big|_{r=s+\varepsilon t} F(\mu_r) \to \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_s)$ as $\varepsilon \to 0$. Hence, by dominated convergence applied to (9),

$$0 \le \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \int_0^1 \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\nu_t^\varepsilon) \mathrm{d}t = \int_0^1 \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_s) \mathrm{d}t = \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_s), \tag{10}$$

as claimed.

(4) $\implies$ (1). Let $\varphi(t) = F(\mu_t)$ for all $t \in [0,1]$. From (Villani, 2008, Equation 16.5), for all $t \in [0,1]$,

$$\varphi(t) = (1-t)\varphi(0) + t\varphi(1) - \int_0^1 G(s,t) \frac{\mathrm{d}^2}{\mathrm{d}t^2} \varphi(s) \mathrm{d}s,$$

where G is the Green function defined as $G(s,t) = s(1-t)\mathbb{1}_{\{s \le t\}} + t(1-s)\mathbb{1}_{\{t \le s\}} \ge 0$ in (Villani, 2008, Equation 16.6). Then, $\frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_s) \ge 0$ implies that $\int_0^1 G(s,t) \frac{\mathrm{d}^2}{\mathrm{d}t^2} \varphi(s) \mathrm{d}s \ge 0$, and thus

$$\varphi(t) = F(\mu_t) \le (1-t)\varphi(0) + t\varphi(1) = (1-t)F(\mu) + tF(\nu).$$

□

Geodesic convexity is intrinsic to the Wasserstein geometry, but it is not the only convexity notion that appears in optimization over measures. In many mean-field and learning problems, one instead works with convexity along linear interpolations of probability measures.

G.2 Linear convexity

We now turn to convexity along linear interpolations of measures, which is widely used in the mean-field optimization literature and is generally different from geodesic convexity.

Definition G.4 (Linear convexity of F)**.** *We say a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is linearly convex, if for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and any $t \in [0,1]$,*

$$F((1-t)\mu + t\nu) \le (1-t)F(\mu) + tF(\nu). \tag{11}$$

This notion is standard in the mean-field optimization literature; see, for instance, Hu et al. (2021); Nitanda et al. (2022); Chizat (2022); Lascu et al. (2024). Moreover, if $F \in \mathcal{C}^1$, then linear convexity implies the first-order inequality

$$F(\nu) - F(\mu) \ge \int_{\mathbb{R}^d} \frac{\delta F}{\delta \mu}(\mu, x) \, \mathrm{d}(\nu - \mu)(x),$$

see (Hu et al., 2021, Lemma 4.1).

The following examples illustrate that linear convexity and geodesic convexity need not coincide. In particular, some objectives of interest in generative modeling are linearly convex without being geodesically convex.

Example G.5 (L^2-Wasserstein and MMD distances)**.** *Examples of optimization problems that are linearly convex but not geodesically convex include minimizing distances with respect to a fixed target distribution, a setting common in generative modeling Huang & Malik (2026); Arbel et al. (2019). One example is the minimization of the squared L^2-Wasserstein distance $F(\mu) = \frac{1}{2}\mathcal{W}_2^2(\mu, \mu^*)$, studied in Huang & Malik (2026), which can be interpreted as training a Generative Adversarial Network (GAN) generator to match a target measure μ^* in $\mathcal{W}_2$. While F in this case fails to be geodesically convex (see, e.g., (Ambrosio et al., 2008, Example 9.1.5)), it is linearly convex. Another example, given in Arbel et al. (2019), is the minimization of the Maximum Mean Discrepancy (MMD) $F(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \mu^*)$. Here, again, F is linearly convex but not geodesically convex (Arbel et al., 2019, Section 3.1).*

H Optimality conditions in the Wasserstein space

In this appendix, we collect the first and second-order optimality notions used in the paper. We begin with the standard first-order characterization of minimizers in terms of the Wasserstein gradient, and then turn to second-order conditions involving the operator K_μ. We conclude by recording the consequences of the saddle-property assumption, which links second-order criticality to global minimality.

H.1 First-order optimality conditions

We start with the first-order notions of minimality and stationarity. These are the Wasserstein counterparts of the familiar first-order optimality conditions from finite-dimensional optimization.

Definition H.1 (Local, global and strict minimizers)**.** *A probability measure $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ is a local minimizer of F if there exists $r > 0$ such that $F(\mu) \geq F(\mu^*)$ for all $\mu \in B_2(r, \mu^*)$. If $F(\mu) > F(\mu^*)$ for all $\mu \in B_2(r, \mu^*)$, $\mu \neq \mu^*$, then μ^* is the strict local minimizer of F. If this holds for any $r > 0$, then μ^* is a (strict) global minimizer.*

Definition H.2 (First-order stationary point)**.** *We say $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ is a first-order stationary point of F if $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$.*

Definition H.3 (ε-first-order stationary point)**.** *We say $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ is an ε-first-order stationary point of F if $\|\nabla_\mu F(\mu^*, \cdot)\|_{L^2_{\mu^*}} \leq \varepsilon$.*

The following theorem, due to Lanzetti et al. (2025), gives the basic first-order necessary condition for local optimality.

Theorem H.4 (First-order necessary optimality condition; (Lanzetti et al., 2025, Theorem 3.2))**.** *Let μ^* be a local minimizer of F and assume F is Wasserstein differentiable at μ^*. Then, the Wasserstein gradient of F vanishes at μ^*, i.e., $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$.*

Under an additional geodesic convexity assumption, first-order stationarity also becomes sufficient for global optimality.

Theorem H.5 (First-order sufficient optimality condition; (Lanzetti et al., 2025, Theorem 3.3))**.** *Assume that F is α-geodesically convex with $\alpha \geq 0$. Suppose there exists $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ such that F is Wasserstein differentiable at μ^* and the Wasserstein gradient of F vanishes at μ^*, i.e., $\nabla_\mu F(\mu^*, \cdot) = 0$, μ^*-a.e. Then, μ^* is a global minimizer of F. If $\alpha > 0$, μ^* is the unique global minimizer of F.*

H.2 Second-order optimality conditions

We next introduce the second-order optimality notions used in the paper. As explained in the main text, when μ^* is a critical point, the second-order behavior is determined by the operator K_{μ^*}, which motivates the definitions below.

Definition H.6 (Second-order critical and saddle point)**.** *We say $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ satisfying $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$ is (i) a second-order critical point of F if $\langle \mathrm{K}_{\mu^*} v, v \rangle_{L^2_{\mu^*}} \geq 0$, for all $v \in L^2_{\mu^*}$, and (ii) a saddle point of F if $\underline{\lambda}(\mathrm{K}_{\mu^*}) < 0$.*

The next result provides the second-order necessary condition satisfied by local minimizers.

Proposition H.7 (Second-order necessary optimality condition; (Yamamoto et al., 2025, Proposition 3.3))**.** *Assume $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular (cf. Definition 2.2). Suppose $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ is a local minimizer of F, i.e., there exists $r > 0$ such that $F(\mu) \geq F(\mu^*)$, for all $\mu \in B_2(r, \mu^*)$. Assume further that the maps $(\mu, x) \mapsto \nabla \nabla_\mu F(\mu, x)$ and $(\mu, x, \bar{x}) \mapsto \nabla^2_\mu F(\mu, x, \bar{x})$ are jointly continuous. Then $\mathrm{K}_{\mu^*} \succeq 0$.*

Proof. By Theorem H.4, $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$. Since μ^* is absolutely continuous, $\mathrm{M}_{\mu^*} = 0$ μ^*-a.e. (Yamamoto et al., 2025, Lemma C.10.), and hence $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$. Fix $v \in L^2_{\mu^*}$ and set $\mu_t := (\mathrm{Id} + tv)_{\#}\mu^*$. The coupling induced by this map gives $W_2(\mu_t, \mu^*) \leq |t| \|v\|_{L^2_{\mu^*}}$. Thus μ_t belongs

to the neighborhood of local minimality for sufficiently small $|t|$. The function $f(t) := F(\mu_t)$ therefore has a local minimum at $t = 0$. By Proposition 2.3,

$$0 \leq f''(0) = \langle \mathrm{H}_{\mu^*} v, v \rangle_{L^2_{\mu^*}} = \langle \mathrm{K}_{\mu^*} v, v \rangle_{L^2_{\mu^*}}.$$

Hence $\mathrm{K}_{\mu^*} \succeq 0$. □

We continue with the proof of a second-order sufficient optimality condition, which relies only on the local second-order geometry of the objective.

Proposition H.8 (Second-order sufficient condition under local radial convexity)**.** *Assume that $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular. Let $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ satisfy $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$. Assume that K_{μ^*} is strictly positive on the Wasserstein tangent space, i.e.,*

$$\langle \mathrm{K}_{\mu^*} v, v \rangle_{L^2_{\mu^*}} > 0, \tag{12}$$

for every $v \in \mathcal{T}_{\mu^*}\mathcal{P}_2(\mathbb{R}^d) \setminus \{0\}$. Suppose that there exists $r > 0$ such that the following holds. For every $\mu \in B_2(r, \mu^*)$, let $\mathrm{T}^{\mu}_{\mu^*}$ be the optimal transport map from μ^* to μ and set $v_\mu := \mathrm{T}^{\mu}_{\mu^*} - \mathrm{Id}$, $\mu_t := (\mathrm{Id} + t v_\mu)_{\#} \mu^*$, $t \in [0, 1]$. Assume that

$$\left\langle \widetilde{\mathrm{H}}_{\mu^*, t} v_\mu, v_\mu \right\rangle_{L^2_{\mu^*}} \geq 0, \tag{13}$$

for every $t \in [0, 1]$, and that the map $t \mapsto \langle \widetilde{\mathrm{H}}_{\mu^*, t} v_\mu, v_\mu \rangle_{L^2_{\mu^*}}$ is continuous at $t = 0$. Then μ^* is a strict local minimizer of F, i.e., $F(\mu) > F(\mu^*)$, for every $\mu \in B_2(r, \mu^*) \setminus \{\mu^*\}$.

Proof. Fix $\mu \in B_2(r, \mu^*) \setminus \{\mu^*\}$. Since μ^* is absolutely continuous, $\|v_\mu\|_{L^2_{\mu^*}} = W_2(\mu, \mu^*) > 0$. Moreover, $v_\mu \in \mathcal{T}_{\mu^*}\mathcal{P}_2(\mathbb{R}^d)$. Define $f(t) := F(\mu_t)$. By Proposition 2.3,

$$f''(t) = \left\langle \widetilde{\mathrm{H}}_{\mu^*, t} v_\mu, v_\mu \right\rangle_{L^2_{\mu^*}}.$$

Since μ^* is stationary, Definition H.2 implies

$$f'(0) = \langle \nabla_\mu F(\mu^*, \cdot), v_\mu \rangle_{L^2_{\mu^*}} = 0.$$

Furthermore, stationarity and absolute continuity give $\mathrm{M}_{\mu^*} = 0$ (Yamamoto et al., 2025, Lemma C.10.), and hence $\widetilde{\mathrm{H}}_{\mu^*, 0} = \mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$. Therefore, by (12),

$$f''(0) = \langle \mathrm{K}_{\mu^*} v_\mu, v_\mu \rangle_{L^2_{\mu^*}} > 0.$$

By the continuity assumption, there exists $s_\mu \in (0, 1]$ such that

$$f''(t) \geq \frac{1}{2} \langle \mathrm{K}_{\mu^*} v_\mu, v_\mu \rangle_{L^2_{\mu^*}} > 0,$$

for all $t \in [0, s_\mu]$. On the other hand, local radial convexity gives $f''(t) \geq 0$ on all of $[0, 1]$. Taylor's formula with integral remainder therefore implies

$$\begin{aligned} F(\mu) - F(\mu^*) = \int_0^1 (1 - t) f''(t) \mathrm{d}t &\geq \frac{1}{2} \langle \mathrm{K}_{\mu^*} v_\mu, v_\mu \rangle_{L^2_{\mu^*}} \int_0^{s_\mu} (1 - t) \mathrm{d}t \\ &= \frac{1}{2} \left(s_\mu - \frac{s_\mu^2}{2} \right) \langle \mathrm{K}_{\mu^*} v_\mu, v_\mu \rangle_{L^2_{\mu^*}} > 0. \end{aligned}$$

Thus $F(\mu) > F(\mu^*)$ for every $\mu \in B_2(r, \mu^*) \setminus \{\mu^*\}$. □

Remark H.9 (Local radial convexity)**.** *Condition (13) requires convexity only along geodesics starting from μ^*, which is weaker than geodesic convexity over all $B_2(r, \mu^*)$. Because K_{μ^*} is compact, its positive eigenvalues can accumulate at zero, preventing a uniform lower bound $\mathrm{K}_{\mu^*} \succeq m \, \mathrm{I}_{\mathrm{d} \times \mathrm{d}}$, $m > 0$. Standard Hessian continuity is insufficient to guarantee preserved positivity in directions with arbitrarily small curvature. Strict positivity of K_{μ^*} ensures positive initial curvature along these geodesics, continuity maintains this positivity over an initial segment, and radial convexity prevents it from dropping below zero later. Together, this gives strict local minimality.*

Spectrally projected transport neighborhoods. Let $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$. Under Wasserstein regularity, K_{μ^*} is compact and self-adjoint on $L^2_{\mu^*}$. For $m > 0$, define

$$P_m := \mathbb{1}_{[m,\infty)}(\mathrm{K}_{\mu^*}), \quad V_m := \operatorname{Ran} P_m. \tag{14}$$

We consider cutoffs for which $V_m \neq \{0\}$. Compactness implies that K_{μ^*} has only finitely many eigenvalues, counted with multiplicity, in $[m, \infty)$. Thus V_m is finite-dimensional and is spanned by the eigenvectors whose eigenvalues are at least m. By the spectral calculus,

$$\langle \mathrm{K}_{\mu^*} v, v\rangle_{L^2_{\mu^*}} \geq m\|v\|^2_{L^2_{\mu^*}}, \quad v \in V_m. \tag{15}$$

Because μ^* is absolutely continuous, for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ there exists a unique μ^*-a.e. optimal transport map $\mathrm{T}^{\mu}_{\mu^*}$ from μ^* to μ. Define the projected transport family

$$\mathcal{M}_m(\mu^*) := \left\{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathrm{T}^{\mu}_{\mu^*} - \mathrm{Id} \in V_m\right\},$$

which represents the collection of measures reachable from μ^* through optimal displacement fields belonging to V_m, and its restricted Wasserstein neighborhood

$$B_{2,m}(r, \mu^*) := B_2(r, \mu^*) \cap \mathcal{M}_m(\mu^*).$$

Now we can prove that μ^* is a strict local minimizer of F relative to the projected transport family $\mathcal{M}_m(\mu^*)$.

Proposition H.10 (Projected second-order sufficient optimality condition)**.** *Assume that $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular. Let $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ satisfy $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$. Fix $m > 0$ such that $V_m \neq \{0\}$, where V_m is defined in (14). For $\mu \in \mathcal{M}_m(\mu^*)$, set $v_\mu := \mathrm{T}^{\mu}_{\mu^*} - \mathrm{Id}$, $\mu_t := (\mathrm{Id} + tv_\mu)_{\#}\mu^*$, $t \in [0, 1]$. Assume that there exists $r_0 > 0$ and a nondecreasing function $\omega_m : [0, \infty) \to [0, \infty)$, satisfying $\lim_{\rho \downarrow 0} \omega_m(\rho) = 0$, such that, for every $\mu \in B_{2,m}(r_0, \mu^*)$ and every $t \in [0, 1]$,*

$$\left\| P_m \left(\widetilde{\mathrm{H}}_{\mu^*,t} - \mathrm{K}_{\mu^*}\right) P_m \right\|_{\mathrm{op}} \leq \omega_m \left(W_2(\mu, \mu^*)\right). \tag{16}$$

Then there exists $r_m \in (0, r_0)$ such that

$$F(\mu) \geq F(\mu^*) + \frac{m}{4} W_2^2(\mu, \mu^*),$$

for all $\mu \in B_{2,m}(r_m, \mu^*)$. In particular, μ^* is a strict local minimizer relative to $\mathcal{M}_m(\mu^*)$.

Proof. Choose $r_m \in (0, r_0)$ sufficiently small that

$$\omega_m(r_m) \leq \frac{m}{2}. \tag{17}$$

Fix $\mu \in B_{2,m}(r_m, \mu^*)$. By the definition of $\mathcal{M}_m(\mu^*)$, we have $v_\mu \in V_m$, and hence $P_m v_\mu = v_\mu$. Since $\mathrm{T}^{\mu}_{\mu^*}$ is optimal,

$$\|v_\mu\|_{L^2_{\mu^*}} = W_2(\mu, \mu^*). \tag{18}$$

Define $f(t) := F(\mu_t)$. Proposition 2.3 gives

$$f''(t) = \left\langle \widetilde{\mathrm{H}}_{\mu^*,t} v_\mu, v_\mu \right\rangle_{L^2_{\mu^*}}.$$

Since $v_\mu = P_m v_\mu$, we obtain

$$\begin{aligned} f''(t) &= \langle \mathrm{K}_{\mu^*} v_\mu, v_\mu\rangle_{L^2_{\mu^*}} + \left\langle \left(\widetilde{\mathrm{H}}_{\mu^*,t} - \mathrm{K}_{\mu^*}\right) v_\mu, v_\mu \right\rangle_{L^2_{\mu^*}} \\ &= \langle \mathrm{K}_{\mu^*} v_\mu, v_\mu\rangle_{L^2_{\mu^*}} + \left\langle P_m \left(\widetilde{\mathrm{H}}_{\mu^*,t} - \mathrm{K}_{\mu^*}\right) P_m v_\mu, v_\mu \right\rangle_{L^2_{\mu^*}}. \end{aligned}$$

Using (15), (16), and (17), it follows that

$$f''(t) \geq m\|v_\mu\|^2_{L^2_{\mu^*}} - \omega_m \left(W_2(\mu, \mu^*)\right) \|v_\mu\|^2_{L^2_{\mu^*}} \geq \frac{m}{2}\|v_\mu\|^2_{L^2_{\mu^*}}, \quad t \in [0, 1].$$

The stationarity condition gives

$$f'(0) = \langle \nabla_\mu F(\mu^*, \cdot), v_\mu \rangle_{L^2_{\mu^*}} = 0.$$

Therefore, Taylor's formula with integral remainder gives

$$F(\mu) - F(\mu^*) = \int_0^1 (1-t) f''(t) \mathrm{d}t \geq \frac{m}{2} \|v_\mu\|^2_{L^2_{\mu^*}} \int_0^1 (1-t)\mathrm{d}t = \frac{m}{4} \|v_\mu\|^2_{L^2_{\mu^*}} = \frac{m}{4} W_2^2(\mu, \mu^*),$$

where the final identity follows from (18). For $\mu \neq \mu^*$, the right-hand side is strictly positive, proving the relative strict local minimality. □

The conclusion concerns strict local minimality relative to $\mathcal{M}_m(\mu^*)$. The operator K_{μ^*} may have a nontrivial kernel or negative spectrum on $V_m^\perp$. The proof requires coercivity only on V_m, where the spectral cutoff supplies the explicit bound $\mathrm{K}_{\mu^*}|_{V_m} \succeq m I_{V_m}$. Hence, the proposition does not assert local minimality with respect to arbitrary Wasserstein perturbations.

Finite-dimensional restriction. Under the present compact Hessian setting, a uniform positive quadratic growth bound cannot hold on an entire Wasserstein neighborhood of an absolutely continuous stationary measure. Indeed, suppose that there exist $c, r > 0$ such that

$$F(\mu) \geq F(\mu^*) + c W_2^2(\mu, \mu^*), \quad \mu \in B_2(r, \mu^*). \tag{19}$$

Take $v = \nabla\phi$ with $\phi \in C_c^\infty(\mathbb{R}^d)$ and set $\mu_s = (\mathrm{Id} + sv)_\# \mu^*$. For sufficiently small $s > 0$, the function $x \mapsto \frac{1}{2}\|x\|^2 + s\phi(x)$ is convex. Hence, by E.2, $\mathrm{Id} + s\nabla\phi$ is an optimal transport map, and $W_2(\mu_s, \mu^*) = s\|v\|_{L^2_{\mu^*}}$. Stationarity, $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$, and the second-order expansion in Lemma L.6 give

$$F(\mu_s) - F(\mu^*) = \frac{s^2}{2} \langle \mathrm{K}_{\mu^*} v, v \rangle_{L^2_{\mu^*}} + o(s^2).$$

Combining the two bounds, dividing by s^2, and letting $s \downarrow 0$ gives

$$\langle \mathrm{K}_{\mu^*} v, v \rangle_{L^2_{\mu^*}} \geq 2c \|v\|^2_{L^2_{\mu^*}}. \tag{20}$$

This holds for every $v = \nabla\phi$ with $\phi \in C_c^\infty(\mathbb{R}^d)$. By density and boundedness of K_{μ^*}, it extends to $\mathcal{T}_{\mu^*}\mathcal{P}_2(\mathbb{R}^d)$. Choose an orthonormal sequence $(v_j)_{j\geq 1}$ in $\mathcal{T}_{\mu^*}\mathcal{P}_2(\mathbb{R}^d)$. Then $v_j \rightharpoonup 0$ in $L^2_{\mu^*}$, and compactness implies $\mathrm{K}_{\mu^*} v_j \to 0$ strongly. Thus $\langle \mathrm{K}_{\mu^*} v_j, v_j \rangle_{L^2_{\mu^*}} \to 0$, contradicting the preceding lower bound. More generally, a compact operator cannot satisfy a uniform positive quadratic lower bound on an infinite-dimensional linear subspace. A fixed positive spectral band provides a finite-dimensional family on which such a bound is available.

The spectral cutoff is a natural projection. The spectral cutoff is canonical and selects exactly the eigendirections whose curvature is at least m. Moreover, P_m commutes with K_{μ^*} and with every bounded spectral function of K_{μ^*}, including $(\mathrm{K}^2_{\mu^*} + \beta\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}$. If the positive eigenvalues accumulate at zero, decreasing m retains progressively more positive spectral directions while weakening the available uniform curvature bound. If there are only finitely many positive eigenvalues, V_m stabilizes once m falls below the smallest positive eigenvalue.

H.3 Consequences under the saddle property

We now combine the optimality conditions above with the saddle property assumption from the main text. This allows us to relate approximate second-order critical points to the global minimizer set.

Lemma H.11 (Approximate second-order critical points are close to the global minimizer set)**.** *Suppose Assumption 4.3 holds on $\mathcal{S} \subseteq \mathcal{P}_2(\mathbb{R}^d)$. If $\mu \in \mathcal{S}$ satisfies $\|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} \leq \varepsilon$, and $\underline{\lambda}(\mathrm{K}_\mu) \geq -\delta$, then $\mu \in \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$.*

Proof. Otherwise, Assumption 4.3 would imply either $\|\nabla_\mu F(\mu,\cdot)\|_{L^2_\mu} > \varepsilon$ or $\underline{\lambda}(\mathrm{K}_\mu) < -\delta$, contradicting the hypotheses. □

If Assumption 4.3 holds with $\alpha_{\mathrm{ben}} = 0$, and $\mathcal{M}^*$ is closed, then every second-order critical point in $\mathcal{S}$ is a global minimizer.

I First and second-order dynamics on Wasserstein space

In this appendix, we briefly recall the first and second-order transport schemes associated with the minimization of F over Wasserstein space.

I.1 Wasserstein gradient flow and Wasserstein gradient descent

We begin with the Benamou–Brenier formula, which characterizes the L^2-Wasserstein distance in dynamic form; see, e.g., (Figalli & Glaudo, 2023, Theorem 4.1.3). Given $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu_0,\mu_1) = \inf\left\{ \int_0^1 \int_{\mathbb{R}^d} |v_s(x)|^2 \mu_s(\mathrm{d}x)\mathrm{d}s : \partial_s\mu_s + \nabla\cdot(v_s\mu_s) = 0 \right\},$$

where the infimum is taken over all curves $[0,1] \ni t \mapsto (\mu_t, v_t) \in \mathcal{P}_2(\mathbb{R}^d) \times L^2_{\mu_t}$ such that $t \mapsto \mu_t$ is continuous with respect to weak convergence topology.

This suggests defining the steepest-descent dynamics for F by choosing the velocity field $v_t = -\nabla_\mu F(\mu_t,\cdot)$, where $\nabla_\mu F$ belongs of the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at μ_t. Define the first-order Taylor model by

$$\mathfrak{T}_1^F(\mu)[v] := F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu}.$$

By Lemma L.3,

$$F((\mathrm{Id}+v)_\#\mu) = \mathfrak{T}_1^F(\mu)[v] + o(\|v\|_{L^2_\mu}), \quad \|v\|_{L^2_\mu} \to 0.$$

Adding a quadratic penalty defines the gradient descent direction

$$v_{\mathrm{GD}}(\mu^n) := \operatorname*{argmin}_{v\in L^2_{\mu^n}} \left\{ \mathfrak{T}_1^F(\mu^n)[v] + \frac{1}{2}\|v\|^2_{L^2_{\mu^n}} \right\}.$$

The objective is strongly convex, and its unique minimizer satisfies

$$v_{\mathrm{GD}}(\mu^n) = -\nabla_\mu F(\mu^n,\cdot).$$

Thus, for a stepsize $\tau > 0$, the update is

$$\mu^{n+1} = (\mathrm{Id} + \tau v_{\mathrm{GD}}(\mu^n))_\#\mu^n = (\mathrm{Id} - \tau\nabla_\mu F(\mu^n,\cdot))_\#\mu^n, \quad \mu^0 \in \mathcal{P}_2(\mathbb{R}^d).$$

If $(\mu,x) \mapsto \nabla_\mu F(\mu,x)$ is L_F-Lipschitz (cf. Proposition J.2), then Lemma L.4 implies that, for $\tau \le L_F^{-1}$,

$$F(\mu^{n+1}) \le F(\mu^n) - \frac{\tau}{2}\|\nabla_\mu F(\mu^n,\cdot)\|^2_{L^2_{\mu^n}} \le F(\mu^n).$$

Hence Wasserstein gradient descent is a descent method and becomes stationary exactly at first-order critical points, that is, when $\nabla_\mu F(\mu,\cdot) = 0$ in L^2_μ (cf. Definition H.2). By Proposition J.2, passing to the limit $\tau \to 0$ yields the Wasserstein gradient flow

$$\partial_t\mu_t = \nabla\cdot(\nabla_\mu F(\mu_t,\cdot)\mu_t)\,, \text{ for } t > 0, \quad \mu|_{t=0} := \mu^0. \tag{21}$$

By Proposition J.1, F dissipates along the flow, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) = -\|\nabla_\mu F(\mu_t,\cdot)\|^2_{L^2_{\mu_t}} \le 0, \tag{22}$$

for all $t > 0$. Thus, Wasserstein gradient flow is the natural first-order descent dynamics for F on $\mathcal{P}_2(\mathbb{R}^d)$ and becomes stationary precisely at first-order critical points.

I.2 Wasserstein Newton flow and Wasserstein Newton method

We work under the hypotheses of Lemma L.6 and define the second-order Taylor model by

$$\mathfrak{T}_2^F(\mu)[w] := F(\mu) + \langle \nabla_\mu F(\mu, \cdot), w \rangle_{L^2_\mu} + \frac{1}{2} \langle \mathrm{H}_\mu\, w, w \rangle_{L^2_\mu}.$$

For every fixed $v \in L^2_\mu$, Lemma L.6 gives

$$F((\mathrm{Id} + tv)_{\#}\mu) = \mathfrak{T}_2^F(\mu)[tv] + o(t^2), \quad t \to 0.$$

This suggests defining the Newton direction by minimizing the quadratic model. Given $n \in \mathbb{N}_0$, $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\tau > 0$, let

$$v_{\mathrm{N}}(\mu^n) = \operatorname*{arg\,min}_{v \in L^2_{\mu^n}} \mathfrak{T}_2^F(\mu^n)[v], \quad \mu^{n+1} = (\mathrm{Id} + \tau v_{\mathrm{N}}(\mu^n))_{\#}\mu^n.$$

Assume that H_{μ^n} is self-adjoint and uniformly positive on $L^2_{\mu^n}$ with constant $c > 0$, so that $\mathrm{H}_{\mu^n}^{-1}$ exists. Then the minimizer is characterized by

$$\mathrm{H}_{\mu^n}\, v_{\mathrm{N}}(\mu^n) = -\nabla_\mu F(\mu^n, \cdot),$$

that is,

$$v_{\mathrm{N}}(\mu^n) = -\,\mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot).$$

The resulting Wasserstein Newton scheme is therefore

$$\mu^{n+1} = \Big(\mathrm{Id} - \tau\, \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot)\Big)_{\#}\mu^n, \quad \mu^0 \in \mathcal{P}_2(\mathbb{R}^d).$$

Under Assumption 4.4, let $C_{\mathrm{H}} := C_{\mathrm{M}} + C_{\mathrm{K}}$. Suppose that $\mathrm{H}_{\mu^n} \succeq c\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$, $c > 0$, and choose $0 < \tau \le \frac{c}{C_{\mathrm{H}}}$. Applying the first estimate of Lemma L.14 gives

$$\begin{aligned}
F(\mu^{n+1}) &\le F(\mu^n) - \tau \langle \nabla_\mu F(\mu^n, \cdot), \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot) \rangle_{L^2_{\mu^n}} + \frac{C_{\mathrm{H}} \tau^2}{2} \| \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot) \|^2_{L^2_{\mu^n}} \\
&= F(\mu^n) - \tau \langle \mathrm{H}_{\mu^n}\, \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot), \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot) \rangle_{L^2_{\mu^n}} + \frac{C_{\mathrm{H}} \tau^2}{2} \| \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot) \|^2_{L^2_{\mu^n}} \\
&\le F(\mu^n) - \tau \left(c - \frac{C_{\mathrm{H}} \tau}{2} \right) \| \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot) \|^2_{L^2_{\mu^n}} \\
&\le F(\mu^n) - \frac{\tau c}{2} \| \mathrm{H}_{\mu^n}^{-1}\, \nabla_\mu F(\mu^n, \cdot) \|^2_{L^2_{\mu^n}}.
\end{aligned}$$

Thus, under uniform Hessian boundedness and the stated coercivity and stepsize conditions, the damped Newton transport step decreases the objective strictly whenever $\nabla_\mu F(\mu^n, \cdot) \neq 0$.

As in the first-order case, one may view the corresponding continuous-time dynamics as the limit of the explicit transport scheme; see Proposition J.4. This yields the Wasserstein Newton flow

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t\, \mathrm{H}_{\mu_t}^{-1}\, \nabla_\mu F(\mu_t, \cdot) \right), \text{ for } t > 0, \quad \mu|_{t=0} := \mu^0.$$

This coincides with the Information Newton's flow introduced by Wang & Li (2020). By Proposition J.1,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\langle \nabla_\mu\, F(\mu_t, \cdot), \mathrm{H}_{\mu_t}^{-1}\, \nabla_\mu F(\mu_t, \cdot) \rangle_{L^2_{\mu_t}} \le -c \| \mathrm{H}_{\mu_t}^{-1}\, \nabla_\mu F(\mu_t, \cdot) \|^2_{L^2_{\mu_t}} \le 0.$$

The Newton method, however, comes with two important limitations. First, the scheme is only well defined when H_μ^{-1} exists at every iteration, which is a strong requirement. Second, inverse Hessian preconditioning does not distinguish between local minimizers and saddle points. As discussed in the main text, it may destroy repulsion along directions of negative curvature and make saddle points locally attractive.

J Calculus along curves and flow approximations

In this appendix, we collect two types of auxiliary results used in the paper. First, we recall a chain rule formula for differentiating the objective along absolutely continuous curves in Wasserstein space. We then use this calculus to relate the discrete transport schemes introduced in the paper to their corresponding gradient flows.

J.1 Chain rule along absolutely continuous curves

The following result provides the basic differential identity for evaluating the variation of F along an absolutely continuous curve satisfying the continuity equation.

Proposition J.1. *Assume $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein differentiable and let $\mu : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ be an absolutely continuous curve satisfying $\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$, $t \in (0,1]$, $\mu|_{t=0} := \mu_0$, with vector field $v_t \in \mathcal{T}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$.*

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

Proof. By applying Definition 2.1 with $\mu = \mu_{t+h}$, $\nu = (\mathrm{Id} + h v_t)_{\#}\mu_t$, $h > 0$, and $\gamma \in \Pi_o(\mu_{t+h}, (\mathrm{Id} + h v_t)_{\#}\mu_t)$, we have

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

as $h \to 0$, where the second inequality follows from the Cauchy-Schwarz inequality and the optimality of γ, while the last inequality follows from Proposition E.3. For fixed t at which $v_t \in L^2_{\mu_t}$ is defined, Lemma L.3 gives

$$F((\mathrm{Id} + h v_t)_{\#}\mu_t) - F(\mu_t) = h \langle \nabla_\mu F(\mu_t, \cdot), v_t \rangle_{L^2_{\mu_t}} + o(h), \quad h \downarrow 0.$$

Thus,

$$\lim_{h \downarrow 0} \frac{F((\mathrm{Id} + h v_t)_{\#}\mu_t) - F(\mu_t)}{h} = \langle \nabla_\mu F(\mu_t, \cdot), v_t \rangle_{L^2_{\mu_t}}.$$

Finally,

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

□

J.2 Convergence of Wasserstein gradient descent to the flow

We first show that the Wasserstein gradient descent scheme converges to the corresponding Wasserstein gradient flow as the time step tends to zero.

Proposition J.2 (Convergence of Wasserstein gradient descent to the flow)**.** *Let $T > 0$ and let $\mu : [0,T] \to \mathcal{P}_2(\mathbb{R}^d)$ be an absolutely continuous curve satisfying $\partial_t \mu_t = \nabla \cdot (\nabla_\mu F(\mu_t, \cdot)\mu_t)$ in $\mathcal{D}'\left((0,T) \times \mathbb{R}^d\right)$, $\mu|_{t=0} := \mu_0$. Assume that $\int_0^T \|\nabla_\mu F(\mu_t, \cdot)\|^2_{L^2_{\mu_t}} \mathrm{d}t < \infty$. For $\tau > 0$, define $E_\tau(\mu) := (\mathrm{Id} - \tau \nabla_\mu F(\mu, \cdot))_{\#}\mu$. Assume there exists $L_F > 0$ such that, for all $\gamma \in \Pi(\mu, \nu)$,*

$$\|\nabla_\mu F(\nu, \cdot) - \nabla_\mu F(\mu, \cdot)\|_{L^2_\gamma} \le L_F \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - x\|^2 \mathrm{d}\gamma(x,y) \right)^{1/2}. \tag{23}$$

Let $(\mu^n_\tau)_{n \ge 0}$ be defined by $\mu^{n+1}_\tau = E_\tau(\mu^n_\tau)$, with $\mu^0_\tau := \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then,

$$\lim_{\tau \to 0} \max_{n \le \lfloor \frac{T}{\tau} \rfloor} W_2(\mu^n_\tau, \mu_{n\tau}) = 0.$$

Proof. For notational convenience, set $N_\tau := \left\lfloor \frac{T}{\tau} \right\rfloor$, and denote the iterates corresponding to the step size τ by $(\mu^n_\tau)_{n \ge 0}$. We first establish the stability of the Euler map. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and let $\pi \in \Pi_o(\mu, \nu)$. The measure $((\mathrm{Id} - \tau \nabla_\mu F(\mu, \cdot)) \times (\mathrm{Id} - \tau \nabla_\mu F(\nu, \cdot)))_{\#}\pi$ is a coupling of $E_\tau(\mu)$ and $E_\tau(\nu)$. Therefore, by Minkowski's inequality and (23),

$$\begin{aligned} W_2\left(E_\tau(\mu), E_\tau(\nu)\right) &\le \left(\int \left\| (x - y) - \tau \left(\nabla_\mu F(\mu, x) - \nabla_\mu F(\nu, y) \right) \right\|^2 \mathrm{d}\pi(x,y) \right)^{1/2} \\ &\le \left(\int \|x - y\|^2 \mathrm{d}\pi(x,y) \right)^{1/2} + \tau \left(\int \|\nabla_\mu F(\mu, x) - \nabla_\mu F(\nu, y)\|^2 \mathrm{d}\pi(x,y) \right)^{1/2} \\ &\le (1 + L_F \tau) W_2(\mu, \nu). \end{aligned} \tag{24}$$

We next prove a uniform one-step consistency estimate. Set $M_T := \sup_{0\le t\le T} \|\nabla_\mu F(\mu_t,\cdot)\|_{L^2_{\mu_t}}$. This quantity is finite. Indeed, if $\gamma_t \in \Pi_o(\mu_0,\mu_t)$, then

$$\begin{aligned}\|\nabla_\mu F(\mu_t,\cdot)\|_{L^2(\mu_t)} &\le \|\nabla_\mu F(\mu_0,\cdot)\|_{L^2(\mu_0)} + \left(\int \|\nabla_\mu F(\mu_t,y) - \nabla_\mu F(\mu_0,x)\|^2 \,\mathrm{d}\gamma_t(x,y)\right)^{1/2} \\ &\le \|\nabla_\mu F(\mu_0,\cdot)\|_{L^2(\mu_0)} + L_F W_2(\mu_t,\mu_0).\end{aligned}$$

Since $t \mapsto \mu_t$ is continuous in W_2 on the compact interval $[0,T]$, the right-hand side is uniformly bounded. The curve $(\mu_t)_{t\in[0,T]}$ solves the continuity equation $\partial_t\mu_t + \nabla\cdot(v_t\mu_t) = 0$, $v_t = -\nabla_\mu F(\mu_t,\cdot)$. By the superposition principle (Ambrosio et al., 2008, Theorem 8.2.1), there exists a probability measure $\boldsymbol{\eta}$ in Γ_T, where Γ_T is the space of continuous maps from $[0,T]$ into $\mathbb{R}^d$, endowed with the sup norm, such that, writing $X_t(\omega) := \omega(t) \in AC^2([0,T];\mathbb{R}^d)$, we have $(X_t)_\#\boldsymbol{\eta} = \mu_t$ and

$$X_s - X_t = -\int_t^s \nabla_\mu F(\mu_r, X_r)\mathrm{d}r$$

for a.e. $s,t \in (0,T)$. Fix $t \in [0, T-\tau]$. Since $\mathrm{Law}_{\boldsymbol{\eta}}\left(X_t - \tau\nabla_\mu F(\mu_t,X_t)\right) = E_\tau(\mu_t)$, we obtain

$$\begin{aligned}W_2\left(E_\tau(\mu_t),\mu_{t+\tau}\right) &\le \|X_{t+\tau} - X_t + \tau\nabla_\mu F(\mu_t,X_t)\|_{L^2_{\boldsymbol{\eta}}} \\ &= \left\|\int_t^{t+\tau} \left(\nabla_\mu F(\mu_t,X_t) - \nabla_\mu F(\mu_s,X_s)\right)\mathrm{d}s\right\|_{L^2_{\boldsymbol{\eta}}} \\ &\le \int_t^{t+\tau} \|\nabla_\mu F(\mu_t,X_t) - \nabla_\mu F(\mu_s,X_s)\|_{L^2_{\boldsymbol{\eta}}}\,\mathrm{d}s.\end{aligned} \tag{25}$$

For each s, the law of (X_t,X_s) is a coupling of μ_t and μ_s. Hence (23) implies

$$\|\nabla_\mu F(\mu_t,X_t) - \nabla_\mu F(\mu_s,X_s)\|_{L^2_{\boldsymbol{\eta}}} \le L_F\|X_t - X_s\|_{L^2_{\boldsymbol{\eta}}}.$$

Moreover,

$$\|X_s - X_t\|_{L^2_{\boldsymbol{\eta}}} \le \int_t^s \|\nabla_\mu F(\mu_r,X_r)\|_{L^2_{\boldsymbol{\eta}}}\mathrm{d}r \le M_T(s-t).$$

Substituting these estimates into (25) gives the uniform bound

$$\sup_{0\le t\le T-\tau} W_2\left(E_\tau(\mu_t),\mu_{t+\tau}\right) \le \frac{L_F M_T}{2}\tau^2. \tag{26}$$

Now set $e_n := W_2(\mu^n_\tau,\mu_{n\tau})$. For every $0 \le n < N_\tau$, using (24) and (26), we obtain

$$e_{n+1} \le W_2\left(E_\tau(\mu^n_\tau), E_\tau(\mu_{n\tau})\right) + W_2\left(E_\tau(\mu_{n\tau}),\mu_{(n+1)\tau}\right) \le (1+L_F\tau)e_n + \frac{L_F M_T}{2}\tau^2.$$

Since $e_0 = 0$, iteration gives

$$e_n \le \frac{L_F M_T}{2}\tau^2\sum_{k=0}^{n-1}(1+L_F\tau)^k = \frac{M_T}{2}\tau\left((1+L_F\tau)^n - 1\right) \le \frac{M_T}{2}\left(e^{L_F T} - 1\right)\tau.$$

Hence,

$$\max_{0\le n\le N_\tau} W_2(\mu^n_\tau,\mu_{n\tau}) \le \frac{M_T}{2}\left(e^{L_F T} - 1\right)\tau,$$

and hence

$$\lim_{\tau\downarrow 0}\max_{0\le n\le \lfloor T/\tau\rfloor} W_2(\mu^n_\tau,\mu_{n\tau}) = 0.$$

□

J.3 Convergence of the Wasserstein Newton scheme to the flow

We next consider the second-order transport scheme associated with Wasserstein Newton's method and show that, under suitable regularity and uniform positivity assumptions, it converges to the corresponding Newton flow.

For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, write $\mathrm{M}_\mu(x) := \nabla\nabla_\mu F(\mu, x)$, $\mathrm{K}_\mu(x, x') := \nabla^2_\mu F(\mu, x, x')$, so that

$$(\mathrm{H}_\mu\, v)(x) = \mathrm{M}_\mu(x)v(x) + \int_{\mathbb{R}^d} \mathrm{K}_\mu(x, x')v(x')\mathrm{d}\mu(x').$$

Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Pi(\mu, \nu)$, define the first and second coordinate lifts of the Hessians by

$$\left(\widehat{\mathrm{H}}^{\gamma,1}_\mu w\right)(x, y) := \mathrm{M}_\mu(x)w(x, y) + \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathrm{K}_\mu(x, x')w(x', y')\mathrm{d}\gamma(x', y'),$$
$$\left(\widehat{\mathrm{H}}^{\gamma,2}_\nu w\right)(x, y) := \mathrm{M}_\nu(y)w(x, y) + \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathrm{K}_\nu(y, y')w(x', y')\mathrm{d}\gamma(x', y').$$

Assumption J.3 (Uniform coercivity of the coupling lifts)**.** *There exists $\lambda > 0$ such that, for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Pi(\mu, \nu)$, the operators $\widehat{\mathrm{H}}^{\gamma,1}_\mu$ and $\widehat{\mathrm{H}}^{\gamma,2}_\nu$ are bounded and self-adjoint on L^2_γ and satisfy*

$$\left\langle \widehat{\mathrm{H}}^{\gamma,1}_\mu w, w\right\rangle_{L^2_\gamma} \geq \lambda\|w\|^2_{L^2_\gamma}, \quad \left\langle \widehat{\mathrm{H}}^{\gamma,2}_\nu w, w\right\rangle_{L^2_\gamma} \geq \lambda\|w\|^2_{L^2_\gamma}$$

for every $w \in L^2_\gamma$.

Proposition J.4 (Convergence of the Wasserstein Newton scheme to the flow)**.** *Let Assumptions 4.5 and J.3 hold, and write $L_{\mathrm{H}} := L_{\mathrm{M}} + L_{\mathrm{K}}$. Assume that there exists $L_F > 0$ such that, for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Pi(\mu, \nu)$,*

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \|\nabla_\mu F(\nu, y) - \nabla_\mu F(\mu, x)\|^2\mathrm{d}\gamma(x, y) \leq L_F^2 \int_{\mathbb{R}^d\times\mathbb{R}^d} \|y - x\|^2\mathrm{d}\gamma(x, y). \tag{27}$$

Let $T > 0$, and let $\mu : [0, T] \to \mathcal{P}_2(\mathbb{R}^d)$ be an absolutely continuous curve satisfying $\partial_t\mu_t = \nabla \cdot (\mu_t\, \mathrm{H}^{-1}_{\mu_t} \nabla_\mu F(\mu_t, \cdot))$ in $\mathcal{D}'((0, T) \times \mathbb{R}^d)$, $\mu|_{t=0} = \mu^0$. For $\tau > 0$, set $E_\tau(\nu) := (\mathrm{Id} -\tau\, \mathrm{H}^{-1}_\nu \nabla_\mu F(\nu, \cdot))_\#\nu$, $\mu^0_\tau := \mu^0$, $\mu^{n+1}_\tau := E_\tau(\mu^n_\tau)$. Define

$$G_T := \sup_{0\leq t\leq T} \|\nabla_\mu F(\mu_t, \cdot)\|_{L^2_{\mu_t}}, \quad C_{N,T} := \frac{L_F}{\lambda} + \frac{L_{\mathrm{H}}G_T}{\lambda^2},$$

where $\lambda > 0$ is the coercivity constant in Assumption J.3. Then $G_T < \infty$, the discrete iterates are well defined, and

$$\lim_{\tau\downarrow 0} \max_{0\leq n\leq\lfloor T/\tau\rfloor} W_2(\mu^n_\tau, \mu_{n\tau}) = 0.$$

Proof. The proof is divided into the following four steps.

Step 1: finite-horizon bounds and well-posedness. Applying (27) to an optimal coupling $\gamma \in \Pi_o(\mu, \nu)$ and using the reverse triangle inequality gives

$$\left|\|\nabla_\mu F(\nu, \cdot)\|_{L^2_\nu} - \|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu}\right| \leq L_F W_2(\mu, \nu).$$

Hence,

$$G_T \leq \|\nabla_\mu F(\mu^0, \cdot)\|_{L^2_{\mu^0}} + L_F \sup_{0\leq t\leq T} W_2(\mu_t, \mu^0) < \infty,$$

because $t \mapsto \mu_t$ is continuous on $[0, T]$. Applying Assumption J.3 to the diagonal coupling shows that $\mathrm{H}_\nu \succeq \lambda\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ on L^2_ν, and $\|\, \mathrm{H}^{-1}_\nu\, \|_{\mathrm{op}} \leq \lambda^{-1}$, for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Hence

$$\|\, \mathrm{H}^{-1}_\nu\, \nabla_\mu F(\nu, \cdot)\|_{L^2_\nu} \leq \lambda^{-1}\|\nabla_\mu F(\nu, \cdot)\|_{L^2_\nu} < \infty.$$

It follows that $E_\tau(\nu) \in \mathcal{P}_2(\mathbb{R}^d)$, so all discrete iterates are well defined. Along the exact flow,

$$\sup_{0\le t\le T} \| \mathrm{H}_{\mu_t}^{-1} \nabla_\mu F(\mu_t, \cdot)\|_{L^2_{\mu_t}} \le \frac{G_T}{\lambda}. \tag{28}$$

Step 2: comparison with the exact flow. Fix $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \Pi(\mu, \nu)$, and set $d_\gamma^2 := \int \|y - x\|^2 \mathrm{d}\gamma(x, y)$. Write $\widehat{\mathrm{H}}_\mu := \widehat{\mathrm{H}}_\mu^{\gamma,1}$, $\widehat{\mathrm{H}}_\nu := \widehat{\mathrm{H}}_\nu^{\gamma,2}$ for the coupling lifts defined above. Assumption 4.5 gives

$$\|\widehat{\mathrm{H}}_\nu - \widehat{\mathrm{H}}_\mu\|_{\mathrm{op}} \le L_{\mathrm{H}} d_\gamma. \tag{29}$$

Here the multiplication part is bounded using its L^∞_γ norm, and the integral part using the $L^2_{\gamma\otimes\gamma}$ norm of its kernel. Moreover, Assumption J.3 yields

$$\|\widehat{\mathrm{H}}_\mu^{-1}\|_{\mathrm{op}} \le \lambda^{-1}, \quad \|\widehat{\mathrm{H}}_\nu^{-1}\|_{\mathrm{op}} \le \lambda^{-1}. \tag{30}$$

Define the lifted gradients and Newton directions by $\widehat{g}_\mu(x, y) := \nabla_\mu F(\mu, x)$, $\widehat{g}_\nu(x, y) := \nabla_\mu F(\nu, y)$, and $\widehat{N}_\mu(x, y) := \mathrm{H}_\mu^{-1} \nabla_\mu F(\mu, x)$, $\widehat{N}_\nu(x, y) := \mathrm{H}_\nu^{-1} \nabla_\mu F(\nu, y)$, respectively. Since the first marginal of γ is μ,

$$\begin{aligned}
\left(\widehat{\mathrm{H}}_\mu \widehat{N}_\mu\right)(x, y) &= \mathrm{M}_\mu(x)\, \mathrm{H}_\mu^{-1} \nabla_\mu F(\mu, x) + \int \mathrm{K}_\mu(x, x')\, \mathrm{H}_\mu^{-1} \nabla_\mu F(\mu, x') \mathrm{d}\gamma(x', y') \\
&= \mathrm{M}_\mu(x)\, \mathrm{H}_\mu^{-1} \nabla_\mu F(\mu, x) + \int \mathrm{K}_\mu(x, x')\, \mathrm{H}_\mu^{-1} \nabla_\mu F(\mu, x') \mathrm{d}\mu(x') \\
&= \nabla_\mu F(\mu, x) = \widehat{g}_\mu(x, y).
\end{aligned}$$

Hence, $\widehat{N}_\mu = \widehat{\mathrm{H}}_\mu^{-1} \widehat{g}_\mu$. Similarly, $\widehat{N}_\nu = \widehat{\mathrm{H}}_\nu^{-1} \widehat{g}_\nu$. Therefore,

$$\widehat{N}_\nu - \widehat{N}_\mu = \widehat{\mathrm{H}}_\nu^{-1} \left[\widehat{g}_\nu - \widehat{g}_\mu + \left(\widehat{\mathrm{H}}_\mu - \widehat{\mathrm{H}}_\nu\right) \widehat{N}_\mu\right].$$

Using (27), (29) and (30) gives

$$\|\widehat{N}_\nu - \widehat{N}_\mu\|_{L^2_\gamma} \le \frac{L_F}{\lambda} d_\gamma + \frac{L_{\mathrm{H}}}{\lambda} d_\gamma \| \mathrm{H}_\mu^{-1} \nabla_\mu F(\mu, \cdot)\|_{L^2_\mu} \le \left(\frac{L_F}{\lambda} + \frac{L_{\mathrm{H}}}{\lambda^2} \|\nabla_\mu F(\mu, \cdot)\|_{L^2_\mu}\right) d_\gamma.$$

In particular, when $\mu = \mu_t$ belongs to the exact flow,

$$\|N_\nu - N_{\mu_t}\|_{L^2_\gamma} \le C_{N,T} d_\gamma,$$

uniformly over $t \in [0, T]$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, and $\gamma \in \Pi(\mu_t, \nu)$. Taking an optimal coupling and pushing it forward by the two Euler maps, Minkowski's inequality consequently gives

$$W_2(E_\tau(\nu), E_\tau(\mu_t)) \le (1 + C_{N,T}\tau) W_2(\nu, \mu_t).$$

Step 3: uniform consistency. By (28), the Newton velocity $-\mathrm{H}_{\mu_t}^{-1} \nabla_\mu F(\mu_t, \cdot)$ satisfies

$$\int_0^T \| \mathrm{H}_{\mu_t}^{-1} \nabla_\mu F(\mu_t, \cdot)\|^2_{L^2_{\mu_t}} \mathrm{d}t \le T \left(\frac{G_T}{\lambda}\right)^2.$$

The superposition principle (Ambrosio et al., 2008, Theorem 8.2.1) gives a probability measure η in Γ_T, where Γ_T is the space of continuous maps from $[0, T]$ into $\mathbb{R}^d$, endowed with the sup norm, such that, writing $X_t(\omega) := \omega(t) \in AC^2([0, T]; \mathbb{R}^d)$,

$$(X_t)_\# \eta = \mu_t, \quad \dot{X}_t = -\mathrm{H}_{\mu_t}^{-1} \nabla_\mu F(\mu_t, X_t), \tag{31}$$

for η-a.e. X and a.e. $t \in (0, T)$. Thus for a.e. $0 < t \le s < T$

$$\|X_s - X_t\|_{L^2_\eta} \le \frac{G_T}{\lambda}(s - t).$$

Applying Step 2 to the coupling $(X_t, X_s)_\# \eta$ yields

$$\| \mathrm{H}_{\mu_s}^{-1} \nabla_\mu F(\mu_s, X_s) - \mathrm{H}_{\mu_t}^{-1} \nabla_\mu F(\mu_t, X_t)\|_{L^2_\eta} \le C_{N,T} \frac{G_T}{\lambda}(s - t).$$

For $0 \le t \le T-\tau$, the laws of $X_t - \tau\,\mathrm{H}_{\mu_t}^{-1}\,\nabla_\mu F(\mu_t, X_t)$ and $X_{t+\tau}$ are $E_\tau(\mu_t)$ and $\mu_{t+\tau}$, respectively. Moreover, by (31),

$$X_{t+\tau} = X_t - \int_t^{t+\tau} \mathrm{H}_{\mu_s}^{-1}\,\nabla_\mu F(\mu_s, X_s)\mathrm{d}s.$$

Hence

$$W_2(E_\tau(\mu_t), \mu_{t+\tau}) \le \left\| \int_t^{t+\tau} \left(\mathrm{H}_{\mu_s}^{-1}\,\nabla_\mu F(\mu_s, X_s) - \mathrm{H}_{\mu_t}^{-1}\,\nabla_\mu F(\mu_t, X_t)\right) \mathrm{d}s \right\|_{L^2_\eta} \le \frac{C_{N,T} G_T}{2\lambda}\tau^2.$$

Step 4: global error estimate. Set $e_n := W_2(\mu_\tau^n, \mu_{n\tau})$. For $0 \le n < \lfloor T/\tau \rfloor$, Steps 2 and 3 give

$$\begin{aligned} e_{n+1} &\le W_2(E_\tau(\mu_\tau^n), E_\tau(\mu_{n\tau})) + W_2(E_\tau(\mu_{n\tau}), \mu_{(n+1)\tau}) \\ &\le (1 + C_{N,T}\tau) e_n + \frac{C_{N,T} G_T}{2\lambda}\tau^2. \end{aligned}$$

Since $e_0 = 0$, summing the geometric series yields

$$e_n \le \frac{G_T}{2\lambda}\left((1 + C_{N,T}\tau)^n - 1\right)\tau \le \frac{G_T}{2\lambda}\left(e^{C_{N,T}T} - 1\right)\tau.$$

This proves the stated estimate and convergence. □

Remark J.5 (A sufficient condition for coupling lift coercivity)**.** *Assumption J.3 holds if there exist $\lambda_{\mathrm{H}}, \lambda_{\mathrm{M}} > 0$ such that, uniformly over $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{H}_\mu \succeq \lambda_{\mathrm{H}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ on L^2_μ, and $\mathrm{M}_\mu(x) \succeq \lambda_{\mathrm{M}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ for μ-a.e. x. In that case, Assumption J.3 holds with $\lambda = \min\{\lambda_{\mathrm{H}}, \lambda_{\mathrm{M}}$. Indeed, fix $\gamma \in \Pi(\mu,\nu)$ and disintegrate $\gamma(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)\gamma_x(\mathrm{d}y)$. For $w \in L^2_\gamma$, define*

$$\overline{w}(x) := \int_{\mathbb{R}^d} w(x,y)\mathrm{d}\gamma_x(y), \quad r(x,y) := w(x,y) - \overline{w}(x).$$

Then $\int r(x,y)\mathrm{d}\gamma_x(y) = 0$ for μ-a.e. x, and $\|w\|^2_{L^2_\gamma} = \|\overline{w}\|^2_{L^2_\mu} + \|r\|^2_{L^2_\gamma}$. A direct calculation gives

$$\begin{aligned} \left\langle \widehat{\mathrm{H}}_\mu^{\gamma,1} w, w \right\rangle_{L^2_\gamma} &= \langle \mathrm{H}_\mu\,\overline{w}, \overline{w}\rangle_{L^2_\mu} + \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \mathrm{M}_\mu(x) r(x,y), r(x,y)\rangle\,\mathrm{d}\gamma(x,y) \\ &\ge \lambda_{\mathrm{H}}\|\overline{w}\|^2_{L^2_\mu} + \lambda_{\mathrm{M}}\|r\|^2_{L^2_\gamma} \\ &\ge \min\{\lambda_{\mathrm{H}}, \lambda_{\mathrm{M}}\}\|w\|^2_{L^2_\gamma}. \end{aligned}$$

The same argument, using a disintegration with respect to the second marginal, applies to $\widehat{\mathrm{H}}_\nu^{\gamma,2}$.

K THE WASSERSTEIN PROXIMAL TAYLOR (WPT-k) METHOD

In this section, we show that several iterative schemes on Wasserstein space can be viewed within a common a so-called proximal Taylor framework. This includes Wasserstein gradient descent, proximal descent, mirror descent, preconditioned gradient descent, Newton's method, its Levenberg–Marquardt regularization, and the Wasserstein Saddle-Free Newton method. The common principle is to combine a local Taylor model of the objective with a strongly convex penalty that controls the transport step.

Assume that for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein gradient $\nabla_\mu F(\mu)$ and Hessian H_μ of F exist. Let $\mathfrak{T}_k^F(\mu)[v]$ denote the kth-order Taylor expansion of F at μ in the direction $v \in L^2_\mu$, for $k \in \{1,2\}$. Given $n \in \mathbb{N}$, $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\tau > 0$, the Wasserstein Proximal Taylor method is defined by

$$v^{n+1} = \operatorname*{arg\,min}_{v \in L^2_{\mu^n}} \left\{ \mathfrak{T}_k^F(\mu^n)[v] + \mathrm{d}_{\mu^n}(\mathrm{Id} + v, \mathrm{Id}) \right\}, \quad \mu^{n+1} = (\mathrm{Id} + \tau v^{n+1})_\# \mu^n, \tag{32}$$

where $\mathrm{d}_{\mu^n} : L^2_{\mu^n} \times L^2_{\mu^n} \to [0,\infty)$ is a strongly convex penalty. In the cases considered below, we assume the minimizer is well-defined and unique.

K.1 ITERATES OF THE WPT-1 METHOD

Gradient and proximal descent. For $k = 1$, the first-order Taylor model is

$$\mathfrak{T}_1^F(\mu^n)[v] = F(\mu^n) + \langle \nabla_\mu F(\mu^n, \cdot), v \rangle_{L^2_{\mu^n}}.$$

Lemma L.3 gives the approximation

$$F((\mathrm{Id} + v)_\# \mu^n) = \mathfrak{T}_1^F(\mu^n)[v] + o(\|v\|_{L^2_{\mu^n}}), \quad \|v\|_{L^2_{\mu^n}} \to 0.$$

If we choose $\mathrm{d}_{\mu^n}(\mathrm{Id} + v, \mathrm{Id}) = \frac{1}{2}\|v\|^2_{L^2_{\mu^n}}$, then (32) becomes

$$\mu^{n+1} = (\mathrm{Id} - \tau \nabla_\mu F(\mu^n, \cdot))_\# \, \mu^n, \tag{33}$$

which is Wasserstein gradient descent Salim et al. (2020); Lascu et al. (2024).

We may also consider entropy-regularized objective functionals of the form $F = F_0 + \mathrm{KL}(\cdot|\pi)$, where $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ is a fixed reference measure and F_0 is a smooth nonlinear functional on $\mathcal{P}_2(\mathbb{R}^d)$. To handle the non-smoothness of the relative entropy in the Wasserstein geometry, F is typically minimized using a forward-backward splitting scheme. This method, widely known as Wasserstein proximal descent Zhu & Chen (2025); Lascu et al. (2024); Yao et al. (2024); Salim et al. (2020), alternates between a forward gradient step on the smooth component F_0 and a backward proximal (JKO Jordan et al. (1998)) step on the relative entropy:

$$\begin{cases} \mu^{n+\frac{1}{2}} = (\mathrm{Id} - \tau \nabla_\mu F_0(\mu^n, \cdot))_\# \, \mu^n, \\ \mu^{n+1} = \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathrm{KL}(\mu|\pi) + \frac{1}{2\tau} W_2^2(\mu, \mu^{n+\frac{1}{2}}) \right\}. \end{cases}$$

Mirror descent. More generally, let $\phi_{\mu^n} : L^2_{\mu^n} \to \mathbb{R}$ be convex and continuously Gâteaux differentiable, with derivative ϕ'_{μ^n}, and let $\mathrm{d}^{\phi_{\mu^n}}(\mathrm{Id} + v, \mathrm{Id}) = \phi_{\mu^n}(\mathrm{Id} + v) - \phi_{\mu^n}(\mathrm{Id}) - \langle \phi'_{\mu^n}[\mathrm{Id}], v \rangle_{L^2_{\mu^n}}$ be the associated Bregman divergence. Then (32) becomes

$$\mu^{n+1} = \left((\phi^*_{\mu^n})' \left[\phi'_{\mu^n}(\mathrm{Id}) - \tau \nabla_\mu F(\mu^n) \right] \right)_\# \, \mu^n, \tag{34}$$

which is Wasserstein mirror descent Bonet et al. (2024). Here $(\phi^*_{\mu^n})' = (\phi'_{\mu^n})^{-1}$ provided that ϕ_{μ^n} is strictly convex, lower semi-continuous, and superlinear. In the quadratic case $\phi_{\mu^n}(v) = \frac{1}{2}\|v\|^2_{L^2_{\mu^n}}$, (34) reduces to (33).

Preconditioned gradient descent. A further special case yields Wasserstein preconditioned gradient descent. Let $h : \mathbb{R}^d \to \mathbb{R}$ be strictly convex, define $\phi^h_{\mu^n}(v) = \int h \circ v \, \mathrm{d}\mu^n$, $\mathrm{d}_{\mu^n}(\mathrm{Id} + v, \mathrm{Id}) = \tau \phi^h_{\mu^n}(-v\tau^{-1}) = \int \tau h(-v(x)\tau^{-1}) \, \mathrm{d}\mu^n(x)$. Then (32) yields

$$\mu^{n+1} = (\mathrm{Id} - \tau \nabla h^* \circ \nabla_\mu F(\mu^n))_\# \, \mu^n, \tag{35}$$

which is Wasserstein preconditioned gradient descent Bonet et al. (2024), where $\nabla h^* = (\nabla h)^{-1}$. When $h(z) = \frac{1}{2}\|z\|^2$, (35) again coincides with (33).

K.2 ITERATES OF THE WPT-2 METHOD

Newton method. For $k = 2$, under the hypotheses of Lemma L.6, we get the second-order Taylor model

$$\mathfrak{T}_2^F(\mu^n)[v] := F(\mu^n) + \langle \nabla_\mu F(\mu^n, \cdot), v \rangle_{L^2_\mu} + \frac{1}{2} \langle \mathrm{H}_{\mu^n} v, v \rangle_{L^2_\mu}.$$

If $\mathrm{d}_{\mu^n}(\mathrm{Id} + v, \mathrm{Id}) = 0$, then (32) becomes

$$\mu^{n+1} = \left(\mathrm{Id} - \tau \, \mathrm{H}^{-1}_{\mu^n} \nabla_\mu F(\mu^n, \cdot) \right)_\# \mu^n,$$

which is the Wasserstein Newton method.

Levenberg–Marquardt regularization. If instead $\mathrm{d}_{\mu^n}(\mathrm{Id}+v,\mathrm{Id}) = \frac{1}{2\tau}\|v\|^2_{L^2_{\mu^n}}$ and $\mu^{n+1} = (\mathrm{Id}+v^{n+1})_\#\mu^n$, then Proposition L.18 shows that (32) becomes

$$\mu^{n+1} = \Big(\mathrm{Id} - \big(\mathrm{H}_{\mu^n} + \tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\big)^{-1}\,\nabla_\mu F(\mu^n)\Big)_\#\,\mu^n, \tag{36}$$

which is the Wasserstein analogue of the Levenberg–Marquardt regularization of Newton's method.

K.3 Wasserstein Saddle-Free Newton as a proximal Taylor scheme

The saddle-free Newton construction also fits into this framework. Since the quadratic Hessian term measures the curvature correction added to the first-order Taylor approximation in the direction v, a natural choice is

$$\mathrm{d}_{\mu^n}(\mathrm{Id}+v,\mathrm{Id}) = \big|\mathfrak{T}_1^F(\mu^n)[v] - \mathfrak{T}_2^F(\mu^n)[v]\big| = \frac{1}{2}\big|\langle \mathrm{H}_{\mu^n}\, v, v\rangle_{L^2_{\mu^n}}\big|.$$

Using Lemma L.9, this term admits the upper bounds

$$\frac{1}{2}\big|\langle \mathrm{H}_{\mu^n}\, v, v\rangle_{L^2_{\mu^n}}\big| \le \frac{1}{2}\langle |\,\mathrm{H}_{\mu^n}\,|v, v\rangle_{L^2_{\mu^n}} \le \frac{1}{2}\left\langle (\mathrm{H}^2_{\mu^n} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{1/2} v, v\right\rangle_{L^2_{\mu^n}},$$

where the first inequality follows from the proof of Lemma L.9 (see also (Kato, 1995, Problem 2.35)) and the second follows directly from Lemma L.9.

Replacing the trust-region penalty d_{μ^n} by these upper bounds yields two saddle-free variants. The first is the ideal saddle-free Newton scheme

$$v^{n+1} = \operatorname*{arg\,min}_{v\in L^2_{\mu^n}} \left\{ F(\mu^n) + \langle \nabla_\mu F(\mu^n), v\rangle_{L^2_{\mu^n}} + \frac{1}{2}\langle |\,\mathrm{H}_{\mu^n}\,|v, v\rangle_{L^2_{\mu^n}} \right\}, \quad \mu^{n+1} = (\mathrm{Id}+\tau v^{n+1})_\#\mu^n,$$

which gives (78).

The second is the regularized Wasserstein Saddle-Free Newton scheme

$$\begin{aligned} v^{n+1} &= \operatorname*{arg\,min}_{v\in L^2_{\mu^n}} \left\{ F(\mu^n) + \langle \nabla_\mu F(\mu^n), v\rangle_{L^2_{\mu^n}} + \frac{1}{2}\left\langle (\mathrm{H}^2_{\mu^n} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{1/2} v, v\right\rangle_{L^2_{\mu^n}} \right\}, \\ \mu^{n+1} &= (\mathrm{Id}+\tau v^{n+1})_\#\mu^n, \end{aligned}$$

which yields (1).

L Auxiliary results

This appendix collects the technical lemmas used throughout the paper. We first record basic calculus identities for the functional F along transport curves, then derive the second-order identities underlying the Wasserstein Hessian, and finally gather the operator estimates needed for the analysis of the regularized WSFN preconditioner.

L.1 Calculus along transport curves

We begin with first-order identities describing how F evolves along curves of the form $\mu_t = (\mathrm{Id}+tv)_\#\mu$. These results will be used repeatedly in the derivation of both the gradient and Hessian expansions.

Lemma L.1 (I. Derivative of F along W_2–interpolating curves)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be Wasserstein differentiable. Let $\mu, \gamma \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\gamma$ and for $t \in [0,1]$ define the curve $\mu_t = (\pi_t)_\#\gamma$, with $\pi_t := \mathrm{Id}+tv$, such that $\mu_0 = \mu$ and $\mu_1 = (\pi_1)_\#\gamma$. Then, for any $t \in [0,1]$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_\mu F(\mu_t, \pi_t(x,y)), v(x,y)\rangle\,\mathrm{d}\gamma(x,y).$$

Note that for $v(x,y) = y - x$ we recover (Yamamoto et al., 2025, Lemma C.6).

Proof. Fix $t \in [0,1)$ and let $h > 0$ be such that $t + h \in [0,1]$. Consider the coupling

$$\gamma_{t,t+h} := (\pi_t, \pi_{t+h})_{\#}\gamma \in \Pi(\mu_t, \mu_{t+h}).$$

Applying Proposition F.3 with $\mu = \mu_t$, $\nu = \mu_{t+h}$, and $\gamma = \gamma_{t,t+h}$ yields

$$F(\mu_{t+h}) = F(\mu_t) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu_t, x), y - x \rangle \, \mathrm{d}\gamma_{t,t+h}(x,y) + o\left(\sqrt{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - x\|^2 \, \mathrm{d}\gamma_{t,t+h}(x,y)} \right).$$

Since $\gamma_{t,t+h} = (\pi_t, \pi_{t+h})_{\#}\gamma$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu_t, x), y - x \rangle \, \mathrm{d}\gamma_{t,t+h}(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu_t, \pi_t(x,y)), \pi_{t+h}(x,y) - \pi_t(x,y) \rangle \, \mathrm{d}\gamma(x,y),$$

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - x\|^2 \, \mathrm{d}\gamma_{t,t+h}(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\pi_{t+h}(x,y) - \pi_t(x,y)\|^2 \, \mathrm{d}\gamma(x,y).$$

Observing that

$$\pi_{t+h}(x,y) - \pi_t(x,y) = hv(x,y),$$

we obtain

$$F(\mu_{t+h}) - F(\mu_t) = h \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu_t, \pi_t(x,y)), v(x,y) \rangle \, \mathrm{d}\gamma(x,y) + o\left(h \sqrt{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|v(x,y)\|^2 \, \mathrm{d}\gamma(x,y)} \right).$$

Dividing by h, letting $h \to 0$ and noting that $\frac{o(h)}{h} \to 0$, as $h \to 0$, we conclude that

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu_t, \pi_t(x,y)), v(x,y) \rangle \, \mathrm{d}\gamma(x,y).$$

□

Lemma L.2 (II. Derivative of F along W_2–interpolating curves)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be Wasserstein differentiable. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$ and for $t \in [0,1]$ define the curve $\mu_t = (\pi_t)_{\#}\mu$, with $\pi_t := \mathrm{Id} + tv$. Then, for any $t \in [0,1]$,*

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = \int_{\mathbb{R}^d} \langle \nabla_\mu F(\mu_t, \pi_t(x)), v(x) \rangle \, \mathrm{d}\mu(x).$$

Proof. Fix $t \in [0,1)$ and let $h > 0$ be such that $t + h \in [0,1]$. Consider the coupling

$$\gamma_{t,t+h} := (\pi_t, \pi_{t+h})_{\#}\mu \in \Pi(\mu_t, \mu_{t+h}).$$

Applying Proposition F.3 with $\mu = \mu_t$, $\nu = \mu_{t+h}$, $\gamma = \gamma_{t,t+h}$ and following the proof of Lemma L.1 yields the conclusion. □

Lemma L.3 (First-order expansion along L^2_μ perturbations)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be Wasserstein differentiable. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mu_t := (\pi_t)_{\#}\mu$, with $\pi_t := \mathrm{Id} + tv$, for any $v \in L^2_\mu$. Then, we have the first-order expansion*

$$F(\mu_t) - F(\mu) = t \langle \nabla_\mu F(\mu, \cdot), v \rangle_{L^2_\mu} + o(t \|v\|_{L^2_\mu}), \quad t \to 0.$$

Proof. Fix $t \in [0,1)$. Consider the coupling

$$\gamma_t := (\mathrm{Id}, \pi_t)_{\#}\mu \in \Pi(\mu, \mu_t).$$

Applying Proposition F.3 with $\mu = \mu$, $\nu = \mu_t$, and $\gamma = \gamma_t$ yields

$$F(\mu_t) = F(\mu) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x \rangle \, \mathrm{d}\gamma_t(x,y) + o\left(\sqrt{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - x\|^2 \, \mathrm{d}\gamma_t(x,y)} \right).$$

Since $\gamma_t = (\mathrm{Id}, \pi_t)_{\#}\mu$,

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla_\mu F(\mu, x), y - x \rangle \, \mathrm{d}\gamma_t(x,y) &= \int_{\mathbb{R}^d} \langle \nabla_\mu F(\mu, x), \pi_t(x) - x \rangle \, \mathrm{d}\mu(x) \\ &= t \int_{\mathbb{R}^d} \langle \nabla_\mu F(\mu, x), v(x) \rangle \, \mathrm{d}\mu(x), \end{aligned}$$

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \|y-x\|^2 \,\mathrm{d}\gamma_t(x,y) = \int_{\mathbb{R}^d} \|\pi_t(x)-x\|^2 \,\mathrm{d}\mu(x) = t^2\|v\|_{L^2_\mu}^2.$$

We obtain

$$F(\mu_t) - F(\mu) = t\int_{\mathbb{R}^d} \langle \nabla_\mu F(\mu,x), v(x)\rangle \,\mathrm{d}\mu(x) + o(t\|v\|_{L^2_\mu}).$$

□

Lemma L.4 (First-order smoothness of F)**.** *Let $F:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$ be Wasserstein differentiable and assume $\nabla_\mu F$ is L_F-Lipschitz, i.e., there exists $L_F>0$ such that, for all $\gamma\in\Pi(\mu,\nu)$,*

$$\|\nabla_\mu F(\nu,\cdot) - \nabla_\mu F(\mu,\cdot)\|_{L^2_\gamma} \le L_F\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}\|y-x\|^2 \mathrm{d}\gamma(x,y)\right)^{1/2}.$$

Let $\mu\in\mathcal{P}_2(\mathbb{R}^d)$, $v\in L^2_\mu$ and define $\mu_t := (\mathrm{Id}+tv)_{\#}\mu$, for $t\in[0,1]$. Then

$$F\left((\mathrm{Id}+v)_{\#}\mu\right) \le F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{L_F}{2}\|v\|^2_{L^2_\mu}.$$

Proof. By Lemma L.2,

$$F((\mathrm{Id}+v)_{\#}\mu) - F(\mu) = \int_0^1\int_{\mathbb{R}^d} \langle \nabla_\mu F(\mu_t,\pi_t(x)), v(x)\rangle \,\mathrm{d}\mu(x)\mathrm{d}t.$$

Take $\gamma = (\mathrm{Id}, \mathrm{Id}+tv)_{\#}\mu \in \Pi(\mu,\mu_t)$. Subtracting $\langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu}$ from both sides gives

$$\begin{aligned}
F((\mathrm{Id}+v)_{\#}\mu) - F(\mu) - \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} &= \int_0^1 \langle \nabla_\mu F(\mu_t)\circ\pi_t - \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu}\,\mathrm{d}t \\
&\le \int_0^1 \|\nabla_\mu F(\mu_t)\circ\pi_t - \nabla_\mu F(\mu,\cdot)\|_{L^2_\mu}\,\|v\|_{L^2_\mu}\mathrm{d}t \\
&\le L_F\int_0^1 t\|v\|^2_{L^2_\mu}\mathrm{d}t = \frac{L_F}{2}\|v\|^2_{L^2_\mu}.
\end{aligned}$$

□

L.2 Hessian identities and second-order expansions

We now turn to the second-order structure of F along transport curves. The results in this subsection justify the form of the Wasserstein Hessian used in the main text and yield the corresponding second-order expansions of both the objective and its gradient.

Proof of Proposition 2.3. We follow the proofs of (Bonet et al., 2024, Lemma 11, Proposition 12). By Lemma L.2,

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t) &= \frac{\mathrm{d}}{\mathrm{d}t}\langle \nabla_\mu F(\mu_t,\pi_t), v\rangle_{L^2_\mu} = \left\langle \frac{\mathrm{d}}{\mathrm{d}t}\nabla_\mu F(\mu_t,\pi_t), v\right\rangle_{L^2_\mu} \\
&= \left\langle \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{\mu=\mu_t}\nabla_\mu F(\mu,\pi_t), v\right\rangle_{L^2_\mu} + \left\langle \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{u=\pi_t}\nabla_\mu F(\mu_t,u), v\right\rangle_{L^2_\mu} \\
&= \int_{\mathbb{R}^d}\left\langle \int_{\mathbb{R}^d}\nabla^2_\mu F(\mu_t,\pi_t(x),\pi_t(\bar{x}))v(\bar{x})\mathrm{d}\mu(\bar{x}), v(x)\right\rangle \mathrm{d}\mu(x) \\
&+ \int_{\mathbb{R}^d}\langle \nabla\nabla_\mu F(\mu_t,\pi_t(x))\partial_t\pi_t(x), v(x)\rangle\,\mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left\langle \nabla^2_\mu F(\mu_t,\pi_t(x),\pi_t(\bar{x}))v(\bar{x}), v(x)\right\rangle \mathrm{d}\mu(\bar{x})\mathrm{d}\mu(x) \\
&+ \int_{\mathbb{R}^d}\langle \nabla\nabla_\mu F(\mu_t,\pi_t(x))v(x), v(x)\rangle\,\mathrm{d}\mu(x),
\end{aligned}$$

where the last equality follows since $\partial_t \pi_t = v$. Hence,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t) = \int_{\mathbb{R}^d} \left\langle (\widetilde{\mathrm{M}}_{\mu,t} + \widetilde{\mathrm{K}}_{\mu,t})[v](x), v(x) \right\rangle \mathrm{d}\mu(x) = \left\langle \widetilde{\mathrm{H}}_{\mu,t} v, v \right\rangle_{L^2_\mu}.$$

Now we prove that $\widetilde{\mathrm{H}}_{\mu,t} \in \mathcal{LS}(L^2_\mu)$, i.e., for all $v, w \in L^2_\mu$ and $\alpha_1, \alpha_2 \in \mathbb{R}$,

$$\widetilde{\mathrm{H}}_{\mu,t}[\alpha_1 v + \alpha_2 w] = \alpha_1 \widetilde{\mathrm{H}}_{\mu,t}[v] + \alpha_2 \widetilde{\mathrm{H}}_{\mu,t}[w],$$
$$\left\langle \widetilde{\mathrm{H}}_{\mu,t} v, w \right\rangle_{L^2_\mu} = \left\langle v, \widetilde{\mathrm{H}}_{\mu,t} w \right\rangle_{L^2_\mu}.$$

The linearity follows immediately from a direct calculation. By assumption, $\nabla \nabla_\mu F(\mu, x) \in \mathbb{R}^{d \times d}$ is symmetric, i.e.,

$$\nabla \nabla_\mu F(\mu, x) = \nabla \nabla_\mu F(\mu, x)^\top.$$

Hence, using the identity $\langle u, v \rangle = u^\top v$, for all $u, v \in \mathbb{R}^d$,

$$\begin{aligned}
\left\langle \widetilde{\mathrm{M}}_{\mu,t} v, w \right\rangle_{L^2_\mu} &= \int_{\mathbb{R}^d} \langle \nabla \nabla_\mu F(\mu, \pi_t(x)) v(x), w(x) \rangle \ \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \left(\nabla_\mu F(\mu, \pi_t(x)) v(x) \right))^\top w(x) \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} v(x)^\top \nabla_\mu F(\mu, \pi_t(x))^\top w(x) \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \left\langle v(x), \nabla_\mu F(\mu, \pi_t(x))^\top w(x) \right\rangle \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \langle v(x), \nabla_\mu F(\mu, \pi_t(x)) w(x) \rangle \, \mathrm{d}\mu(x) \\
&= \left\langle v, \widetilde{\mathrm{M}}_{\mu,t} w \right\rangle_{L^2_\mu}.
\end{aligned} \tag{37}$$

By regularity of F, $\nabla^2_\mu F \in \mathbb{R}^{d \times d}$ is symmetric, i.e.,

$$\nabla^2_\mu F(\mu, x, \bar{x})^\top = \nabla^2_\mu F(\mu, \bar{x}, x).$$

Hence,

$$\begin{aligned}
\left\langle \widetilde{\mathrm{K}}_{\mu,t} v, w \right\rangle_{L^2_\mu} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left\langle \nabla^2_\mu F(\mu, \pi_t(x), \pi_t(\bar{x})) v(\bar{x}), w(x) \right\rangle \mathrm{d}\mu(\bar{x}) \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\nabla^2_\mu F(\mu, \pi_t(x), \pi_t(\bar{x})) v(\bar{x}) \right)^\top w(x) \mathrm{d}\mu(\bar{x}) \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} v(\bar{x})^\top \nabla^2_\mu F(\mu, \pi_t(x), \pi_t(\bar{x}))^\top w(x) \mathrm{d}\mu(\bar{x}) \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} v(\bar{x})^\top \nabla^2_\mu F(\mu, \pi_t(\bar{x}), \pi_t(x)) w(x) \mathrm{d}\mu(\bar{x}) \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left\langle v(\bar{x}), \nabla^2_\mu F(\mu, \pi_t(\bar{x}), \pi_t(x)) w(x) \right\rangle \mathrm{d}\mu(x) \mathrm{d}\mu(\bar{x}) \\
&= \int_{\mathbb{R}^d} \left\langle v(\bar{x}), \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, \pi_t(\bar{x}), \pi_t(x)) w(x) \mathrm{d}\mu(x) \right\rangle \mathrm{d}\mu(\bar{x}) \\
&= \left\langle v, \widetilde{\mathrm{K}}_{\mu,t} w \right\rangle_{L^2_\mu},
\end{aligned} \tag{38}$$

where the penultimate equality follows by Fubini's theorem. Adding (37) and (38) gives the conclusion. □

Proposition L.5 (Wasserstein Hessian of F along W_2–curves)**.** *Assume $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$ and for $t \in [0,1]$ define the curve $\mu_t = (\pi_t)_\# \mu$, with $\pi_t = \mathrm{Id} + tv$. If π_t is invertible μ-a.e., for all $t \in [0,1]$, setting $v_t = v \circ \pi_t^{-1}$,*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t) = \langle \mathrm{H}_{\mu_t} v_t, v_t \rangle_{L^2_{\mu_t}},$$

where, the Hessian operator $\mathrm{H}_\mu : L^2_\mu \to L^2_\mu$ *is defined by*

$$\mathrm{H}_\mu[v](x) = \underbrace{\nabla\nabla_\mu F(\mu, x)v(x)}_{=:\mathrm{M}_\mu[v](x)} + \underbrace{\int_{\mathbb{R}^d} \nabla^2_\mu F(\mu, x, \bar{x})v(\bar{x})\mathrm{d}\mu(\bar{x})}_{=:\mathrm{K}_\mu[v](x)},$$

and $\mathrm{H}_\mu \in \mathcal{LS}(L^2_\mu)$.

Proof. We follow the calculations presented after of (Bonet et al., 2024, Proposition 12). The relation between $\widetilde{\mathrm{H}}_{\mu,t}$ and H_{μ_t} can be established for example if π_t is invertible for all t. In this case, setting $v_t = v \circ \pi_t^{-1} \in L^2_{\mu_t}$, we can write by Proposition 2.3,

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2} F(\mu_t) &= \langle \widetilde{\mathrm{H}}_{\mu,t} v, v\rangle_{L^2_\mu} \\
&= \int_{\mathbb{R}^d} \langle \widetilde{\mathrm{H}}_{\mu,t}[v](x), v(x)\rangle \, \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \langle \widetilde{\mathrm{H}}_{\mu,t}[v]\big(\pi_t^{-1}(x_t)\big), v_t(x_t)\rangle \, \mathrm{d}\mu_t(x_t) \\
&= \langle \mathrm{H}_{\mu_t} v_t, v_t\rangle_{L^2_{\mu_t}}.
\end{aligned}$$

The last equality follows from Proposition 2.3 and that $\mu_t = (\pi_t)_\# \mu$, as for all $x \in \mathrm{supp}(\mu)$, which give

$$\begin{aligned}
\widetilde{\mathrm{H}}_{\mu,t}[v](x) &= \nabla\nabla_\mu F(\mu_t, \pi_t(x))v(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(\bar{x}))v(\bar{x})\mathrm{d}\mu(\bar{x}) \\
&= \nabla\nabla_\mu F(\mu_t, x_t)v_t(x_t) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, x_t, \bar{x}_t))v_t(\bar{x}_t)\mathrm{d}\mu_t(\bar{x}_t) \\
&= \mathrm{H}_{\mu_t}[v_t](x_t),
\end{aligned}$$

and thus $\widetilde{\mathrm{H}}_{\mu,t}[v](\pi_t^{-1}(x_t)) = \mathrm{H}_{\mu_t}[v_t](x_t)$. The fact that $\mathrm{H}_\mu \in \mathcal{LS}(L^2_\mu)$ follows directly from Proposition 2.3. □

Lemma L.6 (Second-order expansion of F along L^2_μ perturbations)**.** *Assume that Assumption 4.4 holds and that the maps* $(\nu, x) \mapsto \nabla\nabla_\mu F(\nu, x)$, $(\nu, x, y) \mapsto \nabla^2_\mu F(\nu, x, y)$ *are jointly continuous, with the measure variable endowed with the* W_2 *topology. Fix* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $v \in L^2_\mu$*, and define* $\pi_t := \mathrm{Id} + tv$, $\mu_t := (\pi_t)_\# \mu$*. Then*

$$F(\mu_t) = F(\mu) + t\langle \nabla_\mu F(\mu, \cdot), v\rangle_{L^2_\mu} + \frac{t^2}{2}\langle \mathrm{H}_\mu \, v, v\rangle_{L^2_\mu} + o(t^2), \quad t \to 0.$$

Proof. Write

$$m_t(x) := \nabla\nabla_\mu F(\mu_t, \pi_t(x)), \qquad k_t(x, y) := \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y)).$$

Since $W_2(\mu_t, \mu_s) \le |t - s|\|v\|_{L^2_\mu}$, joint continuity implies that, as $t \to s$, $\nabla\nabla_\mu F(\mu_t, \pi_t(x)) \to \nabla\nabla_\mu F(\mu_s, \pi_s(x))$, and $\nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y)) \to \nabla^2_\mu F(\mu_s, \pi_s(x), \pi_s(y))$, for μ-a.e. x and $\mu \otimes \mu$-a.e. (x, y), respectively. Assumption 4.4 and the pushforward identities give

$$\|\nabla\nabla_\mu F(\mu_t, \cdot)\|_{L^\infty_\mu} \le C_{\mathrm{M}}, \quad \|\nabla^2_\mu F(\mu_t, \cdot, \cdot)\|_{L^2_{\mu\otimes\mu}} \le C_{\mathrm{K}}.$$

Hence, dominated convergence implies

$$\int \langle \nabla\nabla_\mu F(\mu_t, \pi_t(x))v(x), v(x)\rangle \mu(x) \to \int \langle \nabla\nabla_\mu F(\mu_s, \pi_s(x))v(x), v(x)\rangle \mathrm{d}\mu(x).$$

For the kernel term, pointwise convergence and uniform $L^2_{\mu\otimes\mu}$ boundedness imply

$$\nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y)) \rightharpoonup \nabla^2_\mu F(\mu_s, \pi_s(x), \pi_s(y)) \text{ in } L^2_{\mu\otimes\mu}.$$

Indeed, the uniform L^2 bound implies uniform integrability in L^1, so pointwise convergence gives L^1 convergence. Convergence against bounded test functions follows, and truncation of an arbitrary

L^2 test function, together with the uniform L^2 bound, gives weak L^2 convergence. Since $(x,y) \mapsto v(x)v(y)^\top$ belongs to $L^2_{\mu\otimes\mu}$, $\|v(x)v(y)^\top\|_{L^2_{\mu\otimes\mu}} = \|v\|^2_{L^2_\mu}$, we obtain

$$\iint \langle \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y))v(y), v(x)\rangle \mathrm{d}\mu(x)\mathrm{d}\mu(y) \to \iint \langle \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y))v(y), v(x)\rangle \mathrm{d}\mu(x)\mathrm{d}\mu(y).$$

Thus $q(t) := \langle \widetilde{\mathrm{H}}_{\mu,t} v, v\rangle_{L^2_\mu}$ is continuous, and $q(0) = \langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu}$.

Set $\phi(t) := F(\mu_t)$. By Lemma L.2 and Proposition 2.3,

$$\phi'(0) = \langle \nabla_\mu F(\mu, \cdot), v\rangle_{L^2_\mu}, \quad \phi''(t) = q(t).$$

Taylor's formula with integral remainder therefore gives

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2} q(0) + r_{\mu,v}(t),$$

where $r_{\mu,v}(t) := t^2 \int_0^1 (1-s)\,(q(st) - q(0))\,\mathrm{d}s$. Hence

$$\frac{|r_{\mu,v}(t)|}{t^2} \le \frac{1}{2} \sup_{|u|\le|t|} |q(u) - q(0)| \to 0,$$

which proves the assertion. □

Lemma L.7 (First-order expansion of the Wasserstein gradient)**.** *Let Assumption 4.5 hold. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $v \in L^2_\mu$. For $t \in [0,1]$, define $\pi_t := \mathrm{Id} + tv$, $\mu_t := (\pi_t)_\#\mu$. Then*

$$\nabla_\mu F(\mu_t, \cdot) \circ \pi_t - \nabla_\mu F(\mu, \cdot) = t\,\mathrm{H}_\mu\, v + o(t) \text{ in } L^2_\mu, \quad t \downarrow 0.$$

Proof. By Proposition 2.3,

$$\widetilde{\mathrm{H}}_{\mu,t} v(x) = \nabla \nabla_\mu F(\mu_t, \pi_t(x))v(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y))v(y)\mu(\mathrm{d}y).$$

Fix an arbitrary $w \in L^2_\mu$. Apply the chain rule componentwise to $\nabla_\mu F(\mu_t, \pi_t(x))$, differentiating both its spatial argument and its measure argument. For the latter, use the coupling $\gamma_{t,t+h} := (\pi_t, \pi_{t+h})_\#\mu \in \Pi(\mu_t, \mu_{t+h})$, whose displacement is hv. The first-order expansion in Proposition F.3 gives the weak derivative identity

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle \nabla_\mu F(\mu_t, \cdot) \circ \pi_t, w\rangle_{L^2_\mu} &= \int_{\mathbb{R}^d} \langle \nabla \nabla_\mu F(\mu_t, \pi_t(x))v(x), w(x)\rangle \mu(\mathrm{d}x) \\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(y))v(y), w(x)\rangle \mu(\mathrm{d}y)\mu(\mathrm{d}x) \\
&= \langle \widetilde{\mathrm{H}}_{\mu,t} v, w\rangle_{L^2_\mu}.
\end{aligned}$$

The terms are integrable as their absolute values are bounded, respectively, by $\|\nabla\nabla_\mu F(\mu_t, \cdot)\|_{L^\infty_\mu}\|v\|_{L^2_\mu}\|w\|_{L^2_\mu}$ and $\|\nabla^2_\mu F(\mu_t, \cdot, \cdot)\|_{L^2_{\mu\otimes\mu}}\|v\|_{L^2_\mu}\|w\|_{L^2_\mu}$. Assumption 4.5 gives

$$\begin{aligned}
\|\nabla\nabla_\mu F(\mu_t, \cdot) - \nabla\nabla_\mu F(\mu_s, \cdot)\|_{L^\infty_\mu} &\le |t-s| L_{\mathrm{M}} \|v\|_{L^2_\mu}, \\
\|\nabla^2_\mu F(\mu_t, \cdot, \cdot) - \nabla^2_\mu F(\mu_s, \cdot, \cdot)\|_{L^2_{\mu\otimes\mu}} &\le |t-s| L_{\mathrm{K}} \|v\|_{L^2_\mu}.
\end{aligned}$$

In particular, these coefficient norms are bounded along the curve, and $t \mapsto \widetilde{\mathrm{H}}_{\mu,t} v$ is continuous in L^2_μ. Integrating the weak identity therefore gives

$$\langle \nabla_\mu F(\mu_t, \cdot) \circ \pi_t - \nabla_\mu F(\mu, \cdot), w\rangle_{L^2_\mu} = \left\langle \int_0^t \widetilde{\mathrm{H}}_{\mu,s} v \mathrm{d}s, w \right\rangle_{L^2_\mu}.$$

Since this holds for every $w \in L^2_\mu$, we obtain

$$\nabla_\mu F(\mu_t, \cdot) \circ \pi_t - \nabla_\mu F(\mu, \cdot) = \int_0^t \widetilde{\mathrm{H}}_{\mu,s} v \mathrm{d}s \text{ in } L^2_\mu.$$

Finally, $\widetilde{\mathrm{H}}_{\mu,0} = \mathrm{H}_\mu$, so

$$\|\nabla_\mu F(\mu_t, \cdot) \circ \pi_t - \nabla_\mu F(\mu, \cdot) - t\,\mathrm{H}_\mu\, v\|_{L^2_\mu} \le \int_0^t \|(\widetilde{\mathrm{H}}_{\mu,s} - \mathrm{H}_\mu)v\|_{L^2_\mu}\mathrm{d}s \le t \sup_{0\le s\le t} \|(\widetilde{\mathrm{H}}_{\mu,s} - \mathrm{H}_\mu)v\|_{L^2_\mu} = o(t).$$

□

Lemma L.8 (L^2_μ-smoothness of the Wasserstein gradient of F)**.** *Let Assumption 4.5 hold. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $v \in L^2_\mu$. For $t \in [0,1]$, define $\pi_t = \mathrm{Id} + tv$, $\mu_t = (\pi_t)_\# \mu$. Then*

$$\|\nabla_\mu F((\mathrm{Id}+v)_\#\mu, \cdot) \circ \pi_1 - \nabla_\mu F(\mu, \cdot) - \mathrm{H}_\mu\, v\|_{L^2_\mu} \le \frac{L_\mathrm{M} + L_\mathrm{K}}{2}\|v\|^2_{L^2_\mu}.$$

Proof. By the integral identity in Lemma L.7,

$$\nabla_\mu F((\mathrm{Id}+v)_\#\mu, \cdot) \circ (\mathrm{Id}+v) - \nabla_\mu F(\mu, \cdot) - \mathrm{H}_\mu\, v = \int_0^1 (\widetilde{\mathrm{H}}_{\mu,t} - \mathrm{H}_\mu)v\mathrm{d}t.$$

Lemma L.12 gives

$$\|\widetilde{\mathrm{H}}_{\mu,t} - \mathrm{H}_\mu\|_{\mathrm{op}} \le t(L_\mathrm{M} + L_\mathrm{K})\|v\|_{L^2_\mu}.$$

Hence,

$$\left\|\int_0^1 (\widetilde{\mathrm{H}}_{\mu,t} - \mathrm{H}_\mu)v\mathrm{d}t\right\|_{L^2_\mu} \le (L_\mathrm{M} + L_\mathrm{K})\int_0^1 t\|v\|^2_{L^2_\mu}\mathrm{d}t = \frac{1}{2}(L_\mathrm{M} + L_\mathrm{K})\|v\|^2_{L^2_\mu}.$$

□

L.3 Operator bounds and preconditioner identities

We conclude with several technical operator estimates. These bounds control the size and regularity of the Hessian along transport curves and are used to justify the regularized saddle-free preconditioner.

Lemma L.9. *Let Assumption 4.4 hold and let $\beta \ge 0$. Then $(\mathrm{H}_\mu)_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}$ is self-adjoint and, for all $v \in L^2_\mu$,*

$$\left|\langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu}\right| \le \left\langle \left(\mathrm{H}^2_\mu + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}} v, v\right\rangle_{L^2_\mu},$$

where H^2_μ denotes the composition $\mathrm{H}^2_\mu[v] := \mathrm{H}_\mu[\mathrm{H}_\mu[v]]$.

Proof. By Proposition L.5, $(\mathrm{H}_\mu)_\mu \subset \mathcal{LS}(L^2_\mu)$, i.e., $(\mathrm{H}_\mu)_\mu$ is linear and symmetric, and since $(\mathrm{H}_\mu)_\mu$ is uniformly bounded by Assumption 4.4, it follows by (Gustafson & Sigal, 2020, Theorem 23.22) that $(\mathrm{H}_\mu)_\mu$ is self-adjoint, i.e., $\mathrm{H}_\mu = \mathrm{H}^*_\mu$.

Since H_μ is self-adjoint and $\mathrm{H}^*_\mu\,\mathrm{H}_\mu = \mathrm{H}^2_\mu \succeq 0$, using (Reed & Simon, 1980, Definition; VI: Bounded Operators, page 196), we may define the absolute value operator by $|\,\mathrm{H}_\mu\,| := \left(\mathrm{H}^2_\mu\right)^{\frac{1}{2}}$. By (Kato, 1995, Problem 2.35),

$$\left|\langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu}\right| \le \langle |\mathrm{H}_\mu|\, v, v\rangle_{L^2_\mu},$$

for any $v \in L^2_\mu$.

Using the property of adjoint, $(AB)^* = B^*A^*$, we get

$$(\mathrm{H}^2_\mu)^* = (\mathrm{H}_\mu\,\mathrm{H}_\mu)^* = \mathrm{H}^*_\mu\,\mathrm{H}^*_\mu = \mathrm{H}_\mu\,\mathrm{H}_\mu = \mathrm{H}^2_\mu,$$

hence H^2_μ is self-adjoint. The identity operator $\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ is trivially self-adjoint and the sum of two self-adjoint operators is self-adjoint. Thus, $\mathrm{H}^2_\mu + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ is self-adjoint. Furthermore, because $\mathrm{H}^2_\mu \succeq 0$,

the operator $\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}} \succeq 0$. Then by (Reed & Simon, 1980, Theorem VI.9 (square root lemma)) the square root operator $\left(\mathrm{H}_\mu^2 + \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}}$ is well-defined.

Let $\sigma(\mathrm{H}_\mu)$ be the spectrum of H_μ and $E(\lambda)$ be the spectral measure associated to the eigenvalue $\lambda \in \sigma(\mathrm{H}_\mu)$. Since H_μ is self-adjoint, we have $\sigma(\mathrm{H}_\mu) \subset \mathbb{R}$. Then by spectral calculus

$$\begin{aligned}
\langle |\mathrm{H}_\mu|\, v, v\rangle_{L^2_\mu} = \int_{\sigma(\mathrm{H}_\mu)} |\lambda| \mathrm{d}\langle v, E(\lambda) v\rangle_{L^2_\mu} &= \int_{\sigma(\mathrm{H}_\mu)} |\lambda| \mathrm{d}\langle v, E(\lambda) v\rangle_{L^2_\mu} \\
&= \int_{\sigma(\mathrm{H}_\mu)} (\lambda^2)^{\frac{1}{2}} \mathrm{d}\langle v, E(\lambda) v\rangle_{L^2_\mu} \\
&\le \int_{\sigma(\mathrm{H}_\mu)} (\lambda^2 + \beta)^{\frac{1}{2}} \mathrm{d}\langle v, E(\lambda) v\rangle_{L^2_\mu} \\
&= \left\langle \left(\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}} v, v \right\rangle_{L^2_\mu}.
\end{aligned}$$

□

Lemma L.10 (Essential supremum under pushforward)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$, let $T : \mathbb{R}^d \to \mathbb{R}^d$ be measurable and set $\nu := T_{\#}\mu$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be measurable. Then*

$$\|f \circ T\|_{\mu,\infty} = \|f\|_{\nu,\infty},$$

Proof. By definition of essential supremum, for any measurable f,

$$\|f\|_{\mu,\infty} := \inf\{M \ge 0 : |f(x)| \le M \text{ for } \mu\text{-a.e. } x\}.$$

Fix $M \ge 0$. We claim that $f \le M$ ν-a.e. $\iff f \circ T \le M$ μ-a.e. Indeed,

$$f \le M \ \nu\text{-a.e.} \iff \nu(\{y \in \mathbb{R}^d : f(y) > M\}) = 0.$$

Since $\nu = T_{\#}\mu$, by definition of pushforward, we have $\nu(\{y : f(y) > M\}) = \mu\big(T^{-1}(\{y : f(y) > M\})\big)$. But $T^{-1}(\{y : f(y) > M\}) = \{x \in \mathbb{R}^d : f(T(x)) > M\} = \{x \in \mathbb{R}^d : (f \circ T)(x) > M\}$. Therefore,

$$\nu(\{y : f(y) > M\}) = 0 \iff \mu(\{x : (f \circ T)(x) > M\}) = 0.$$

Thus the sets $\{M \ge 0 : f \le M\ \nu\text{-a.e.}\}$ and $\{M \ge 0 : f \circ T \le M\ \mu\text{-a.e.}\}$ coincide. Taking infimum over $M \ge 0$ yields the conclusion. □

Lemma L.11 (Uniform operator norm bound for Hessian)**.** *Let Assumption 4.4 hold. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$ and define $\mu_t := (\pi_t)_{\#}\mu$, with $\pi_t = \mathrm{Id} + tv$, for all $t \in [0,1]$. Then, for every $t \in [0,1]$,*

$$\|\widetilde{\mathrm{H}}_{\mu,t}\|_{\mathrm{op}} \le C_{\mathrm{M}} + C_{\mathrm{K}}.$$

Proof. Fix $t \in [0,1]$. Let $v \in L^2_\mu$. By definition, $\widetilde{\mathrm{M}}_{\mu,t}[v](x) = \nabla\nabla_\mu F(\mu_t, \pi_t(x)) v(x)$. Hence

$$\|\widetilde{\mathrm{M}}_{\mu,t}[v](x)\| \le \left\|\nabla\nabla_\mu F(\mu_t, \pi_t(x))\right\| \|v(x)\|.$$

Squaring and integrating with respect to μ yields

$$\|\widetilde{\mathrm{M}}_{\mu,t} v\|^2_{L^2_\mu} \le \left(\mu\text{-}\operatorname*{ess\,sup}_x \left\|\nabla\nabla_\mu F(\mu_t, \pi_t(x))\right\|\right)^2 \|v\|^2_{L^2_\mu}.$$

By applying Lemma L.10 with $T = \pi_t$, $\nu = \mu_t$, and $f(y) = \|\nabla\nabla_\mu F(\mu_t, y)\|$, we get

$$\mu\text{-}\operatorname*{ess\,sup}_x \left\|\nabla\nabla_\mu F(\mu_t, \pi_t(x))\right\| = \|\nabla\nabla_\mu F(\mu_t, \cdot)\|_{\mu_t,\infty}.$$

By Assumption 4.4, $\|\nabla\nabla_\mu F(\mu_t, \cdot)\|_{\mu_t,\infty} \le C_{\mathrm{M}}$. Therefore $\|\widetilde{\mathrm{M}}_{\mu,t} v\|_{L^2_\mu} \le C_{\mathrm{M}} \|v\|_{L^2_\mu}$, and hence

$$\|\widetilde{\mathrm{M}}_{\mu,t}\|_{\mathrm{op}} \le C_{\mathrm{M}}. \tag{39}$$

By definition

$$\widetilde{\mathrm{K}}_{\mu,t}[v](x) = \int_{\mathbb{R}^d} \nabla_\mu^2 F(\mu_t, \pi_t(x), \pi_t(\bar{x}))v(\bar{x})\mathrm{d}\mu(\bar{x}).$$

By Cauchy–Schwarz in $\bar{x}$,

$$\begin{aligned}\|\widetilde{\mathrm{K}}_{\mu,t}[v](x)\|^2 &= \left\|\int_{\mathbb{R}^d} \nabla_\mu^2 F(\mu_t, \pi_t(x), \pi_t(\bar{x}))v(\bar{x})\mathrm{d}\mu(\bar{x})\right\|^2 \\ &\le \left(\int_{\mathbb{R}^d} \|\nabla_\mu^2 F(\mu_t, \pi_t(x), \pi_t(\bar{x}))\|^2 \mathrm{d}\mu(\bar{x})\right)\left(\int_{\mathbb{R}^d} \|v(\bar{x})\|^2 \mathrm{d}\mu(\bar{x})\right).\end{aligned}$$

Integrating in x gives

$$\|\widetilde{\mathrm{K}}_{\mu,t}v\|_{L^2_\mu}^2 \le \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\nabla_\mu^2 F(\mu_t, \pi_t(x), \pi_t(\bar{x}))|^2 \mathrm{d}\mu(x)\mathrm{d}\mu(\bar{x})\right)\|v\|_{L^2_\mu}^2.$$

Thus

$$\|\widetilde{\mathrm{K}}_{\mu,t}\|_{\mathrm{op}} \le \|\nabla_\mu^2 F(\mu_t, \cdot, \cdot)\circ \pi_t\|_{L^2_{\mu\otimes\mu}}. \tag{40}$$

Now use $\mu_t = (\pi_t)_{\#}\mu$ in both variables to get

$$\|\nabla_\mu^2 F(\mu_t, \cdot, \cdot)\circ \pi_t\|_{L^2_{\mu\otimes\mu}} = \|\nabla_\mu^2 F(\mu_t, \cdot, \cdot)\|_{L^2_{\mu_t\otimes\mu_t}} \le C_{\mathrm{K}}.$$

by Assumption 4.4. Combining with (40),

$$\|\widetilde{\mathrm{K}}_{\mu,t}\|_{\mathrm{op}} \le C_{\mathrm{K}}. \tag{41}$$

Since $\widetilde{\mathrm{H}}_{\mu,t} = \widetilde{\mathrm{M}}_{\mu,t} + \widetilde{\mathrm{K}}_{\mu,t}$, we have by the triangle inequality

$$\|\widetilde{\mathrm{H}}_{\mu,t}\|_{\mathrm{op}} \le \|\widetilde{\mathrm{M}}_{\mu,t}\|_{\mathrm{op}} + \|\widetilde{\mathrm{K}}_{\mu,t}\|_{\mathrm{op}} \le C_{\mathrm{M}} + C_{\mathrm{K}},$$

using (39) and (41). □

Lemma L.12 (Coupling Lipschitzness implies Lipschitzness along W_2–curves)**.** *Let Assumption 4.5 hold. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$ and define $\mu_t := (\mathrm{Id} + tv)_{\#}\mu$, for $t \in [0,1]$. Then, for any $t, s \in [0,1]$,*

$$\|\widetilde{\mathrm{H}}_{\mu,t} - \widetilde{\mathrm{H}}_{\mu,s}\|_{\mathrm{op}} \le |t-s|(L_{\mathrm{M}} + L_{\mathrm{K}})\|v\|_{L^2_\mu}.$$

Proof. By Proposition 2.3, the operator $\widetilde{\mathrm{H}}_{\mu,t}$ is defined by

$$\widetilde{\mathrm{H}}_{\mu,t}[v](x) = \nabla\nabla_\mu F(\mu_t, \pi_t(x))v(x) + \int_{\mathbb{R}^d} \nabla_\mu^2 F(\mu_t, \pi_t(x), \pi_t(\bar{x}))v(\bar{x})\mathrm{d}\mu(\bar{x}),$$

where

$$\widetilde{\mathrm{M}}_{\mu,t}[v](x) = \nabla\nabla_\mu F(\mu_t, \pi_t(x))v(x), \quad \widetilde{\mathrm{K}}_{\mu,t}[v](x) = \int_{\mathbb{R}^d} \nabla_\mu^2 F(\mu_t, \pi_t(x), \pi_t(\bar{x}))v(\bar{x})\mathrm{d}\mu(\bar{x}).$$

For any $t, s \in [0,1]$, we have

$$\|\widetilde{\mathrm{H}}_{\mu,t} - \widetilde{\mathrm{H}}_{\mu,s}\|_{\mathrm{op}} \le \|\widetilde{\mathrm{M}}_{\mu,t} - \widetilde{\mathrm{M}}_{\mu,s}\|_{\mathrm{op}} + \|\widetilde{\mathrm{K}}_{\mu,t} - \widetilde{\mathrm{K}}_{\mu,s}\|_{\mathrm{op}}.$$

Taking $\gamma = (\pi_s, \pi_t)_{\#}\mu \in \Pi(\mu_s, \mu_t)$ gives

$$\begin{aligned}\|(\widetilde{\mathrm{M}}_{\mu,t} - \widetilde{\mathrm{M}}_{\mu,s})v\|_{L^2_\mu}^2 &\le \int_{\mathbb{R}^d} \left\|(\widetilde{\mathrm{M}}_{\mu,t} - \widetilde{\mathrm{M}}_{\mu,s})[v](x)\right\|^2 \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^d} \|(\nabla\nabla_\mu F(\mu_t, \pi_t(x)) - \nabla\nabla_\mu F(\mu_s, \pi_s(x)))v(x)\|^2 \mathrm{d}\mu(x) \\ &\le \|\nabla\nabla_\mu F(\mu_t, \cdot) - \nabla\nabla_\mu F(\mu_s, \cdot)\|_{\mu,\infty}^2 \|v\|_{L^2_\mu}^2 \\ &\le (t-s)^2 L_{\mathrm{M}}^2 \|v\|_{L^2_\mu}^2 \|v\|_{L^2_\mu}^2,\end{aligned}$$

for all $v \in L^2_\mu$. Thus, $\|\widetilde{\mathrm{M}}_{\mu,t} - \widetilde{\mathrm{M}}_{\mu,s}\|_{\mathrm{op}} \le |t-s| L_{\mathrm{M}} \|v\|_{L^2_\mu}$. Similarly,

$$\begin{aligned}
\|(\widetilde{\mathrm{K}}_{\mu,t} - \widetilde{\mathrm{K}}_{\mu,s})v\|^2_{L^2_\mu} &\le \int_{\mathbb{R}^d} \left\|(\widetilde{\mathrm{K}}_{\mu,t} - \widetilde{\mathrm{K}}_{\mu,s})[v](x)\right\|^2 \mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \left\| \int_{\mathbb{R}^d} (\nabla^2_\mu F(\mu_t, \pi_t(x), \pi_t(\bar{x})) - \nabla^2_\mu F(\mu_s, \pi_s(x), \pi_s(\bar{x})))v(\bar{x}) \mathrm{d}\mu(\bar{x}) \right\|^2 \mathrm{d}\mu(x) \\
&\le \|\nabla^2_\mu F(\mu_t, \cdot, \cdot) - \nabla^2_\mu F(\mu_s, \cdot, \cdot)\|_{L^2_{\mu\otimes\mu}} \|v\|^2_{L^2_\mu} \\
&\le (t-s)^2 L^2_{\mathrm{K}} \|v\|^2_{L^2_\mu} \|v\|^2_{L^2_\mu},
\end{aligned}$$

where the penultimate inequality follows from Cauchy-Schwarz inequality. Thus, $\|(\widetilde{\mathrm{K}}_{\mu,t} - \widetilde{\mathrm{K}}_{\mu,s})\|_{\mathrm{op}} \le |t-s| L_{\mathrm{K}} \|v\|_{L^2_\mu}$. Finally,

$$\begin{aligned}
\|\widetilde{\mathrm{H}}_{\mu,t} - \widetilde{\mathrm{H}}_{\mu,s}\|_{\mathrm{op}} &\le \|\widetilde{\mathrm{M}}_{\mu,t} - \widetilde{\mathrm{M}}_{\mu,s}\|_{\mathrm{op}} + \|\widetilde{\mathrm{K}}_{\mu,t} - \widetilde{\mathrm{K}}_{\mu,s}\|_{\mathrm{op}} \\
&\le |t-s| L_{\mathrm{M}} \|v\|_{L^2_\mu} + |t-s| L_{\mathrm{K}} \|v\|_{L^2_\mu} \\
&= |t-s| (L_{\mathrm{M}} + L_{\mathrm{K}}) \|v\|_{L^2_\mu}.
\end{aligned}$$

□

Lemma L.13 (Two-map Lipschitz continuity of the pulled-back Hessian)**.** *Let Assumption 4.5 hold. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $u, v \in L^2_\mu$, and define $Z_u := \mathrm{Id} + u$, $Z_v := \mathrm{Id} + v$, together with $\mu_u := (Z_u)_{\#}\mu$, $\mu_v := (Z_v)_{\#}\mu$. Define the corresponding pulled-back Hessians on L^2_μ by*

$$\widetilde{\mathrm{H}}_{\mu,u}[h](x) := \nabla\nabla_\mu F(\mu_u, Z_u(x)) h(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F\left(\mu_u, Z_u(x), Z_u(y)\right) h(y) \mathrm{d}\mu(y)$$

and analogously for $\widetilde{\mathrm{H}}_{\mu,v}[h]$. Then

$$\left\|\widetilde{\mathrm{H}}_{\mu,u} - \widetilde{\mathrm{H}}_{\mu,v}\right\|_{\mathrm{op}} \le L_{\mathrm{H}} \|v - u\|_{L^2_\mu}.$$

Proof. Consider the coupling $\gamma_{u,v} := (Z_u, Z_v)_{\#}\mu \in \Pi(\mu_u, \mu_v)$. Its quadratic transportation cost is

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - x\|^2 \mathrm{d}\gamma_{u,v}(x,y) = \int_{\mathbb{R}^d} \|Z_v(x) - Z_u(x)\|^2 \mathrm{d}\mu(x) = \|v - u\|^2_{L^2_\mu}. \tag{42}$$

Decompose the pulled-back Hessian as $\widetilde{\mathrm{H}}_{\mu,v}[h] = \widetilde{\mathrm{M}}_{\mu,v}[h] + \widetilde{\mathrm{K}}_{\mu,v}[h]$, where $\widetilde{\mathrm{M}}_{\mu,v}[h] = \nabla\nabla_\mu F(\mu_v, Z_v(x)) h(x)$, and

$$\widetilde{\mathrm{K}}_{\mu,v}[h](x) = \int_{\mathbb{R}^d} \nabla^2_\mu F\left(\mu_v, Z_v(x), Z_v(y)\right) h(y) \mathrm{d}\mu(y).$$

Applying the first part of Assumption 4.5 to $\gamma_{u,v}$, and using (42), gives

$$\mu\text{-}\operatorname{ess\,sup}_{x \in \mathbb{R}^d} \|\nabla\nabla_\mu F(\mu_v, Z_v(x)) - \nabla\nabla_\mu F(\mu_u, Z_u(x))\| \le L_{\mathrm{M}} \|v - u\|_{L^2_\mu}.$$

Consequently,

$$\left\|\widetilde{\mathrm{M}}_{\mu,v} - \widetilde{\mathrm{M}}_{\mu,u}\right\|_{\mathrm{op}} \le L_{\mathrm{M}} \|v - u\|_{L^2_\mu}. \tag{43}$$

Similarly, the second part of Assumption 4.5 gives

$$\left(\iint \left\|\nabla^2_\mu F\left(\mu_v, Z_v(x), Z_v(y)\right) - \nabla^2_\mu F\left(\mu_u, Z_u(x), Z_u(y)\right)\right\|^2 \mathrm{d}\mu(x)\mathrm{d}\mu(y) \right)^{1/2} \le L_{\mathrm{K}} \|v - u\|_{L^2_\mu}.$$

The operator norm of an integral operator is bounded by its Hilbert–Schmidt norm. Hence,

$$\left\|\widetilde{\mathrm{K}}_{\mu,v} - \widetilde{\mathrm{K}}_{\mu,u}\right\|_{\mathrm{op}} \le L_{\mathrm{K}} \|v - u\|_{L^2_\mu}. \tag{44}$$

Combining (43) and (44) gives

$$\left\|\widetilde{\mathrm{H}}_{\mu,v} - \widetilde{\mathrm{H}}_{\mu,u}\right\|_{\mathrm{op}} \le (L_{\mathrm{M}} + L_{\mathrm{K}}) \|v - u\|_{L^2_\mu},$$

as claimed. □

Lemma L.14 (Second-order smoothness of F)**.** *Assume $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $v \in L^2_\mu$ and for $t \in [0,1]$ define the curve $\mu_t = (\pi_t)_{\#}\mu$, with $\pi_t = \mathrm{Id} + tv$. If Assumption 4.4 holds, then*

$$F\left((\mathrm{Id}+v)_{\#}\mu\right) \leq F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{C_\mathrm{M}+C_\mathrm{K}}{2}\|v\|^2_{L^2_\mu},$$

If Assumption 4.5 holds,

$$F\left((\mathrm{Id}+v)_{\#}\mu\right) \leq F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{1}{2}\langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu} + \frac{L_\mathrm{M}+L_\mathrm{K}}{6}\|v\|^3_{L^2_\mu}.$$

Proof. We start with the proof of the first statement. By Lemma L.2,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t) = \int_{\mathbb{R}^d} \langle \nabla_\mu F(\mu,x), v(x)\rangle \,\mathrm{d}\mu(x),$$

By the fundamental theorem of calculus and integration by parts,

$$\begin{aligned}
F\left((\mathrm{Id}+v)_{\#}\mu\right) - F(\mu) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t)\mathrm{d}t = \left(t\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t)\right)\Big|_{t=0}^{t=1} - \int_0^1 t\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t)\mathrm{d}t \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=1} F(\mu_t) - \int_0^1 t\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t)\mathrm{d}t \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t) + \int_0^1 \frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t)\mathrm{d}t - \int_0^1 t\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t)\mathrm{d}t \\
&= \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \int_0^1 (1-t)\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t)\mathrm{d}t.
\end{aligned}$$

By Proposition 2.3,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mu_t) = \left\langle \widetilde{\mathrm{H}}_{\mu,t} v, v\right\rangle_{L^2_\mu},$$

and so

$$F\left((\mathrm{Id}+v)_{\#}\mu\right) - F(\mu) = \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \int_0^1 (1-t)\left\langle \widetilde{\mathrm{H}}_{\mu,t} v, v\right\rangle_{L^2_\mu}\mathrm{d}t.$$

Hence, by Lemma L.11,

$$\begin{aligned}
F\left((\mathrm{Id}+v)_{\#}\mu\right) &= F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \int_0^1 (1-t)\left\langle \widetilde{\mathrm{H}}_{\mu,t} v, v\right\rangle_{L^2_\mu}\mathrm{d}t \\
&\leq F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \sup_{t\in[0,1]}\|\widetilde{\mathrm{H}}_{\mu,t}\|_{\mathrm{op}}\int_0^1 (1-t)\|v\|^2_{L^2_\mu}\mathrm{d}t \\
&\leq F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{C_\mathrm{M}+C_\mathrm{K}}{2}\|v\|^2_{L^2_\mu}.
\end{aligned}$$

For the proof of the second statement, we proceed identically until we obtain

$$F\left((\mathrm{Id}+v)_{\#}\mu\right) = F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \int_0^1 (1-t)\left\langle \widetilde{\mathrm{H}}_{\mu,t} v, v\right\rangle_{L^2_\mu}\mathrm{d}t.$$

Hence, by Lemma L.12 with $s = 0$,

$$\begin{aligned}
F\left((\mathrm{Id}+v)_{\#}\mu\right) &= F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{1}{2}\langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu} + \int_0^1 (1-t)\left\langle (\widetilde{\mathrm{H}}_{\mu,t} - \mathrm{H}_\mu) v, v\right\rangle_{L^2_\mu}\mathrm{d}t \\
&\leq F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{1}{2}\langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu} + \int_0^1 (1-t)\|\widetilde{\mathrm{H}}_{\mu,t} - \mathrm{H}_\mu\,\|_{\mathrm{op}}\|v\|^2_{L^2_\mu}\mathrm{d}t \\
&\leq F(\mu) + \langle \nabla_\mu F(\mu,\cdot), v\rangle_{L^2_\mu} + \frac{1}{2}\langle \mathrm{H}_\mu\, v, v\rangle_{L^2_\mu} + \frac{L_\mathrm{M}+L_\mathrm{K}}{6}\|v\|^3_{L^2_\mu}.
\end{aligned}$$

□

Lemma L.15 (Composition identity for the regularized Hessian)**.** *Let Assumption 4.4 hold. Let $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$, let $\mathrm{T}_t : \mathbb{R}^d \to \mathbb{R}^d$ be measurable and set $\mu_t := (\mathrm{T}_t)_{\#}\mu^*$ for $t \in [0,1)$. Define*

$$\widetilde{\mathrm{H}}_{\mu,t}[v](x) := \nabla\nabla_\mu F(\mu_t, \mathrm{T}_t(x))v(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, \mathrm{T}_t(x), \mathrm{T}_t(y))v(y)\mathrm{d}\mu^*(y), \quad v \in L^2_{\mu^*},$$

$$\mathrm{H}_{\mu_t}[w](x) := \nabla\nabla_\mu F(\mu_t, x)w(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, x, y)w(y)\mathrm{d}\mu_t(y), \quad w \in L^2_{\mu_t}.$$

Then

$$(\widetilde{\mathrm{H}}^2_{\mu,t} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}(\nabla_\mu F(\mu_t) \circ \mathrm{T}_t) = \Big((\mathrm{H}^2_{\mu_t} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\nabla_\mu F(\mu_t)\Big) \circ \mathrm{T}_t, \quad \beta > 0.$$

Proof. Define the composition operator $\mathrm{J}_t : L^2_{\mu_t} \to L^2_{\mu^*}$ by $(\mathrm{J}_t\, w)(x) := w(\mathrm{T}_t(x))$. For $w \in L^2_{\mu_t}$,

$$\|\,\mathrm{J}_t\, w\|^2_{L^2_{\mu^*}} = \int_{\mathbb{R}^d} |w(\mathrm{T}_t(x))|^2 \mathrm{d}\mu^*(x) = \int_{\mathbb{R}^d} |w(y)|^2 \mathrm{d}\mu_t(y),$$

because $\mu_t = (\mathrm{T}_t)_{\#}\mu^*$. Thus

$$\|\,\mathrm{J}_t\, w\|_{L^2_{\mu^*}} = \|w\|_{L^2_{\mu_t}}.$$

Then

$$\widetilde{\mathrm{H}}_{\mu,t}[\mathrm{J}_t\, w](x) = \nabla\nabla_\mu F(\mu_t, \mathrm{T}_t(x))w(\mathrm{T}_t(x)) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, \mathrm{T}_t(x), \mathrm{T}_t(\bar{x}))w(\mathrm{T}_t(\bar{x}))\mathrm{d}\mu^*(\bar{x}).$$

For the first term,

$$\nabla\nabla_\mu F(\mu_t, \mathrm{T}_t(x))w(\mathrm{T}_t(x)) = \big(\nabla\nabla_\mu F(\mu_t, \cdot)w(\cdot)\big)(\mathrm{T}_t(x)).$$

For the second term, use again $\mu_t = (\mathrm{T}_t)_{\#}\mu^*$ to get

$$\int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, \mathrm{T}_t(x), \mathrm{T}_t(\bar{x}))w(\mathrm{T}_t(\bar{x}))\mathrm{d}\mu^*(\bar{x}) = \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_t, \mathrm{T}_t(x), \bar{y})w(\bar{y})\mathrm{d}\mu_t(\bar{y}).$$

Therefore

$$\widetilde{\mathrm{H}}_{\mu,t}[\mathrm{J}_t\, w](x) = (\mathrm{H}_{\mu_t}\, w)(\mathrm{T}_t(x)) = \mathrm{J}_t(\mathrm{H}_{\mu_t}\, w)(x).$$

Hence

$$\widetilde{\mathrm{H}}_{\mu,t}\,\mathrm{J}_t = \mathrm{J}_t\,\mathrm{H}_{\mu_t}.$$

We prove that for every integer $k \geq 1$,

$$\widetilde{\mathrm{H}}^k_{\mu,t}\,\mathrm{J}_t = \mathrm{J}_t\,\mathrm{H}^k_{\mu_t}.$$

by induction on k, where H^k denotes the k-times composition of the operator H with itself. The case $k = 1$ is already proved. Assume

$$\widetilde{\mathrm{H}}^k_{\mu,t}\,\mathrm{J}_t = \mathrm{J}_t\,\mathrm{H}^k_{\mu_t},$$

for some $k \geq 1$. Then

$$\widetilde{\mathrm{H}}^{k+1}_{\mu,t}\,\mathrm{J}_t = \widetilde{\mathrm{H}}_{\mu,t}(\widetilde{\mathrm{H}}^k_{\mu,t}\,\mathrm{J}_t) = \widetilde{\mathrm{H}}_{\mu,t}(\mathrm{J}_t\,\mathrm{H}^k_{\mu_t}) = (\widetilde{\mathrm{H}}_{\mu,t}\,\mathrm{J}_t)\,\mathrm{H}^k_{\mu_t} = \mathrm{J}_t\,\mathrm{H}^{k+1}_{\mu_t},$$

which concludes the induction. Let $p_n(\lambda) = \sum_{j=0}^n a_j\lambda^j$, with $\{a_j\}_{j=0}^n \subset \mathbb{R}$. Then

$$p_n(\widetilde{\mathrm{H}}_{\mu,t})\,\mathrm{J}_t = \sum_{j=0}^n a_j\widetilde{\mathrm{H}}^j_{\mu,t}\,\mathrm{J}_t = \sum_{j=0}^n a_j\,\mathrm{J}_t\,\mathrm{H}^j_{\mu_t} = \mathrm{J}_t\sum_{j=0}^n a_j\,\mathrm{H}^j_{\mu_t} = \mathrm{J}_t\,p_n(\mathrm{H}_{\mu_t}).$$

Because H_{μ_t} and $\widetilde{\mathrm{H}}_{\mu,t}$ are bounded self-adjoint due to Assumption 4.4, their spectra are compact subsets of $[-M, M]$, where $M := \max\{\|\,\mathrm{H}_{\mu_t}\,\|_{\mathrm{op}}, \|\widetilde{\mathrm{H}}_{\mu,t}\|_{\mathrm{op}}\}$. Let $f \in C([-M, M])$. By the Weierstrass theorem, there exists a sequence of polynomials $(p_n)_n$ such that $\sup_{\lambda\in[-M,M]} |p_n(\lambda) - f(\lambda)| \to 0$. Hence, by functional calculus,

$$\|p_n(\mathrm{H}_{\mu_t}) - f(\mathrm{H}_{\mu_t})\|_{\mathrm{op}} \to 0, \qquad \|p_n(\widetilde{\mathrm{H}}_{\mu,t}) - f(\widetilde{\mathrm{H}}_{\mu,t})\|_{\mathrm{op}} \to 0, \quad n \to \infty.$$

From earlier,

$$p_n(\widetilde{\mathrm{H}}_{\mu,t})\,\mathrm{J}_t = \mathrm{J}_t\,p_n(\mathrm{H}_{\mu_t}).$$

Passing to the limit in operator norm and using that J_t is bounded since it is an isometry, we obtain

$$f(\widetilde{\mathrm{H}}_{\mu,t})\,\mathrm{J}_t = \mathrm{J}_t\,f(\mathrm{H}_{\mu_t}).$$

Take $f(\lambda) = \frac{1}{\sqrt{\beta+\lambda^2}}$, $\beta > 0$. Then

$$(\widetilde{\mathrm{H}}^2_{\mu,t} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{J}_t = \mathrm{J}_t(\mathrm{H}^2_{\mu_t} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}.$$

Evaluating this identity on $w \in L^2_{\mu_t}$ yields

$$(\widetilde{\mathrm{H}}^2_{\mu,t} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}(w \circ \mathrm{T}_t) = \Big((\mathrm{H}^2_{\mu_t} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} w\Big) \circ \mathrm{T}_t,$$

Taking $w = \nabla_\mu F(\mu_t)$ gives the conclusion. □

Lemma L.16 (Lipschitz estimates for the regularized WSFN spectral maps)**.** *Let $\mathcal{H}$ be a complex or real Hilbert space, let $A, B \in \mathcal{B}(\mathcal{H})$ be bounded self-adjoint operators, and let $\beta > 0$. Define*

$$P_\beta(A) := \left(A^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}, \quad S_\beta(A) := A\left(A^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}. \tag{45}$$

Then

$$\|P_\beta(A) - P_\beta(B)\|_{\mathrm{op}} \le \frac{2}{\pi\beta}\|A - B\|_{\mathrm{op}}, \tag{46}$$

and

$$\|S_\beta(A) - S_\beta(B)\|_{\mathrm{op}} \le \frac{1}{\sqrt{\beta}}\|A - B\|_{\mathrm{op}}. \tag{47}$$

Moreover, $\|P_\beta(A)\|_{\mathrm{op}} \le \frac{1}{\sqrt{\beta}}$, $\|S_\beta(A)\|_{\mathrm{op}} \le 1$. The dependence $\beta^{-1/2}$ in (47) is sharp.

Proof. We first consider a complex Hilbert space. For a real Hilbert space, the same argument can be applied to its complexification. Self-adjointness and the relevant operator norms are preserved.

For $a > 0$, define

$$Q_a(A) := \left(A^2 + a^2\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}, \quad R_a(A) := A\left(A^2 + a^2\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}.$$

For any nonzero vector $v \in \mathcal{H}$, we have

$$\langle (A \pm ia\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})v, v\rangle = \langle Av, v\rangle \pm ia\|v\|^2.$$

Since A is self-adjoint, its eigenvalues are real, and $\langle Av, v\rangle = \langle v, Av\rangle \in \mathbb{R}$. Taking squared norm

$$\|(A \pm ia\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})v\|^2 = \|Av\|^2 + a^2\|v\|^2 \ge a^2\|v\|^2.$$

Taking square root gives

$$\|(A \pm ia\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})v\| \ge a\|v\|,$$

for all nonzero $v \in \mathcal{H}$. Therefore, the operators $A \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ are invertible and

$$\left\|(A \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1}\right\|_{\mathrm{op}} \le \frac{1}{a}.$$

The resolvent identity (Hille & Phillips, 1957, Theorem 4.8.2) gives, for either choice of sign,

$$(A \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1} - (B \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1} = (A \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1}\,(B - A)\,(B \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1}.$$

Hence,

$$\left\|(A \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1} - (B \pm \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1}\right\|_{\mathrm{op}} \le \frac{1}{a^2}\|A - B\|_{\mathrm{op}}. \tag{48}$$

Using the functional calculus and the fact that $A - \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ and $A + \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ commute, we have

$$Q_a(A) = \frac{1}{2\mathrm{i}a}\left[(A - \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1} - (A + \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1}\right] \tag{49}$$

and

$$R_a(A) = \frac{1}{2}\left[(A - \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1} + (A + \mathrm{i}a\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1}\right]. \tag{50}$$

Combining (48) with (49) gives

$$\|Q_a(A) - Q_a(B)\|_{\mathrm{op}} \le \frac{1}{a^3}\|A - B\|_{\mathrm{op}}, \tag{51}$$

whereas (48) and (50) give

$$\|R_a(A) - R_a(B)\|_{\mathrm{op}} \le \frac{1}{a^2}\|A - B\|_{\mathrm{op}}. \tag{52}$$

We now use the scalar identities

$$\frac{1}{\sqrt{x^2+\beta}} = \frac{2}{\pi}\int_0^\infty \frac{1}{x^2+\beta+t^2}\mathrm{d}t$$

and

$$\frac{x}{\sqrt{x^2+\beta}} = \frac{2}{\pi}\int_0^\infty \frac{x}{x^2+\beta+t^2}\mathrm{d}t.$$

Since A and B are bounded, the corresponding integrals converge in operator norm. Indeed, using the sub-multiplicativity of the operator norm

$$\left\|A\left(A^2 + (\beta+t^2)\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\right\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}}\left\|\left(A^2 + (\beta+t^2)\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\right\|_{\mathrm{op}}$$

Since operator A is self-adjoint, its spectrum $\sigma(A) \subset \mathbb{R}$, the spectrum of $A^2 + (\beta+t^2)\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ lies in $[\beta+t^2, \infty)$. Therefore

$$\left\|\left(A^2 + (\beta+t^2)\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\right\|_{\mathrm{op}} = \sup_{\lambda\in\sigma(A)}\frac{1}{\lambda^2+\beta+t^2} \le \frac{1}{\beta+t^2}.$$

Multiplying by $\|A\|_{\mathrm{op}}$ gives

$$\left\|A\left(A^2 + (\beta+t^2)\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\right\|_{\mathrm{op}} \le \frac{\|A\|_{\mathrm{op}}}{\beta+t^2}.$$

Hence

$$\int_0^\infty \left\|A\left(A^2 + (\beta+t^2)I\right)^{-1}\right\|_{\mathrm{op}}\mathrm{d}t \le \|A\|_{\mathrm{op}}\int_0^\infty \frac{1}{\beta+t^2}\mathrm{d}t = \frac{\|A\|_{\mathrm{op}}\,\pi}{2\sqrt{\beta}} < \infty$$

Since the scalar integral converges for $\beta > 0$, the operator integral converges absolutely in the operator norm. Therefore, by continuous functional calculus,

$$P_\beta(A) = \frac{2}{\pi}\int_0^\infty Q_{\sqrt{\beta+t^2}}(A)\mathrm{d}t \tag{53}$$

and

$$S_\beta(A) = \frac{2}{\pi}\int_0^\infty R_{\sqrt{\beta+t^2}}(A)\mathrm{d}t. \tag{54}$$

It follows from (53) and (51) that

$$\|P_\beta(A) - P_\beta(B)\|_{\mathrm{op}} \le \frac{2}{\pi}\int_0^\infty \frac{\|A-B\|_{\mathrm{op}}}{(\beta+t^2)^{3/2}}\mathrm{d}t = \frac{2}{\pi\beta}\|A - B\|_{\mathrm{op}},$$

where we used $\int_0^\infty \frac{1}{(\beta+t^2)^{3/2}}\mathrm{d}t = \frac{1}{\beta}$. This proves (46).

Similarly, by (54) and (52),

$$\|S_\beta(A) - S_\beta(B)\|_{\mathrm{op}} \le \frac{2}{\pi}\int_0^\infty \frac{\|A-B\|_{\mathrm{op}}}{\beta+t^2}\mathrm{d}t = \frac{1}{\sqrt{\beta}}\|A - B\|_{\mathrm{op}},$$

because $\int_0^\infty \frac{1}{\beta+t^2}\mathrm{d}t = \frac{\pi}{2\sqrt{\beta}}$. This proves (47).

The norm bounds follow immediately from the scalar inequalities

$$\frac{1}{\sqrt{x^2+\beta}} \le \frac{1}{\sqrt{\beta}}, \quad \frac{|x|}{\sqrt{x^2+\beta}} \le 1,$$

and the spectral calculus.

Finally, the dependence $\beta^{-1/2}$ is sharp already in dimension one. Indeed, for $s_\beta(x) := \frac{x}{\sqrt{x^2+\beta}}$, we have $s'_\beta(0) = \frac{1}{\sqrt{\beta}}$. Hence no smaller uniform Lipschitz constant is possible. □

Lemma L.17 (First-order expansion of the preconditioned gradient)**.** *Let Assumptions 4.4, 4.5 hold. Let $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ and assume that, for some $R_F > 0$,*

$$\| \mathrm{M}_{\mu^*} \|_{\mathrm{op}} \le R_F \|\nabla_\mu F(\mu^*, \cdot)\|_{L^2_{\mu^*}}. \tag{55}$$

In particular, condition (55) follows from Assumption 4.2 whenever $\mu^* \in S$ and $\|\nabla_\mu F(\mu^*, \cdot)\|_{L^2_{\mu^*}} \le \varepsilon_0$. Let $Z, \widetilde{Z} \in L^2_{\mu^*}(\mathbb{R}^d; \mathbb{R}^d)$ be measurable maps, and define $\mu_Z := Z_\# \mu^*$, $\mu_{\widetilde{Z}} := \widetilde{Z}_\# \mu^*$. For $\beta > 0$, define the pulled-back preconditioned gradient by

$$\mathfrak{G}_\beta(Z) := \left[\left(\mathrm{H}^2_{\mu_Z} + \beta \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \right)^{-1/2} \nabla_\mu F(\mu_Z, \cdot) \right] \circ Z.$$

Then

$$\mathfrak{G}_\beta(\widetilde{Z}) - \mathfrak{G}_\beta(Z) = \left(\mathrm{K}^2_{\mu^*} + \beta \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \right)^{-1/2} \mathrm{K}_{\mu^*} \left(\widetilde{Z} - Z \right) + R_\beta(Z, \widetilde{Z})$$

in $L^2_{\mu^*}$. Then the remainder satisfies

$$\begin{aligned} \|R_\beta(Z, \widetilde{Z})\|_{L^2_{\mu^*}} \le \Bigg[& \frac{2L_H}{\sqrt{\beta}} \left(\|Z - \mathrm{Id}\,\|_{L^2_{\mu^*}} + \|\widetilde{Z} - \mathrm{Id}\,\|_{L^2_{\mu^*}} \right) \\ & + \frac{L_H^2}{2\pi\beta} \left(\|Z - \mathrm{Id}\,\|^2_{L^2_{\mu^*}} + \|\widetilde{Z} - \mathrm{Id}\,\|^2_{L^2_{\mu^*}} \right) \\ & + \left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_H}{\pi\beta} \right) \|\nabla_\mu F(\mu^*, \cdot)\|_{L^2_{\mu^*}} \Bigg] \|\widetilde{Z} - Z\|_{L^2_{\mu^*}}. \end{aligned} \tag{56}$$

If $\|Z - \mathrm{Id}\,\|_{L^2_{\mu^*}} \le r$, $\|\widetilde{Z} - \mathrm{Id}\,\|_{L^2_{\mu^*}} \le r$, then

$$\|R_\beta(Z, \widetilde{Z})\|_{L^2_{\mu^*}} \le \left[\frac{4L_H r}{\sqrt{\beta}} + \frac{L_H^2 r^2}{\pi\beta} + \left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_H}{\pi\beta} \right) \|\nabla_\mu F(\mu^*, \cdot)\|_{L^2_{\mu^*}} \right] \|\widetilde{Z} - Z\|_{L^2_{\mu^*}}. \tag{57}$$

Proof. For simplicity, set $u := Z - \mathrm{Id}$, $v := \widetilde{Z} - \mathrm{Id}$, and $h := v - u = \widetilde{Z} - Z$. Thus, For $w \in L^2_{\mu^*}$, define $Z_w := \mathrm{Id} + w$, $\mu_w := (Z_w)_\# \mu^*$, and introduce the pulled-back Hessian $\widehat{H}_w : L^2_{\mu^*} \to L^2_{\mu^*}$ by

$$\widehat{H}_w[\varphi](x) := \nabla \nabla_\mu F\left(\mu_w, Z_w(x)\right) \varphi(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F\left(\mu_w, Z_w(x), Z_w(y)\right) \varphi(y) \mu^*(\mathrm{d}y).$$

By Lemma L.13,

$$\|\widehat{\mathrm{H}}_w - \widehat{\mathrm{H}}_z\|_{\mathrm{op}} \le L_H \|w - z\|_{L^2_{\mu^*}}, \quad w, z \in L^2_{\mu^*}. \tag{58}$$

Moreover,

$$\widehat{\mathrm{H}}_0 = \mathrm{H}_{\mu^*} = \mathrm{M}_{\mu^*} + \mathrm{K}_{\mu^*}. \tag{59}$$

For $w \in L^2_{\mu^*}$, set

$$g_w := \nabla_\mu F(\mu_w, \cdot) \circ Z_w, \; P_w := \left(\widehat{\mathrm{H}}^2_w + \beta \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \right)^{-1/2}, \; S_w := \widehat{\mathrm{H}}_w \left(\widehat{\mathrm{H}}^2_w + \beta \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \right)^{-1/2}.$$

Since P_w is a spectral function of the self-adjoint operator $\widehat{\mathrm{H}}_w$, it commutes with $\widehat{\mathrm{H}}_w$ (Reed & Simon, 1980, Theorem VII.2 (g)), and therefore

$$S_w = P_w \widehat{\mathrm{H}}_w = \widehat{\mathrm{H}}_w P_w. \tag{60}$$

The composition identity in Lemma L.15 gives

$$\mathfrak{G}_\beta(Z_w) = P_w g_w. \tag{61}$$

We first derive a convenient decomposition of g_w. Along the curve $t \mapsto \mu_{tw} = (\mathrm{Id} + tw)_{\#}\mu^*$, the pulled-back gradient differentiation identity gives

$$\frac{\mathrm{d}}{\mathrm{d}t} g_{tw} = \widehat{\mathrm{H}}_{tw} w \text{ in } L^2_{\mu^*}.$$

Integrating from 0 to 1 gives $g_w = g_0 + \int_0^1 \widehat{\mathrm{H}}_{tw} w \mathrm{d}t$. Define $E(w) := \int_0^1 \left(\widehat{\mathrm{H}}_{tw} - \widehat{\mathrm{H}}_w\right) w \mathrm{d}t$. Then

$$g_w = g_0 + \widehat{\mathrm{H}}_w w + E(w). \tag{62}$$

By (58),

$$\|E(w)\|_{L^2_{\mu^*}} \le \int_0^1 \|\widehat{\mathrm{H}}_{tw} - \widehat{\mathrm{H}}_w\|_{\mathrm{op}} \|w\|_{L^2_{\mu^*}} \mathrm{d}t \le L_H \int_0^1 (1-t)\|w\|^2_{L^2_{\mu^*}} \mathrm{d}t = \frac{L_H}{2}\|w\|^2_{L^2_{\mu^*}}. \tag{63}$$

We also need a two-point estimate for E. Adding and subtracting $(\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_v)u$ gives

$$E(v) - E(u) = \int_0^1 \left(\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_v\right)(v-u)\mathrm{d}t + \int_0^1 \left[\left(\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_v\right) - \left(\widehat{\mathrm{H}}_{tu} - \widehat{\mathrm{H}}_u\right)\right] u \mathrm{d}t.$$

For the first term,

$$\left\| \int_0^1 \left(\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_v\right)(v-u)\mathrm{d}t \right\|_{L^2_{\mu^*}} \le L_H \int_0^1 (1-t)\|v\|_{L^2_{\mu^*}} \|v-u\|_{L^2_{\mu^*}} \mathrm{d}t = \frac{L_H}{2}\|v\|_{L^2_{\mu^*}} \|v-u\|_{L^2_{\mu^*}}. \tag{64}$$

For the second term, again using (58),

$$\left\| \left(\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_v\right) - \left(\widehat{\mathrm{H}}_{tu} - \widehat{\mathrm{H}}_u\right) \right\|_{\mathrm{op}} \le \|\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_{tu}\|_{\mathrm{op}} + \|\widehat{\mathrm{H}}_v - \widehat{\mathrm{H}}_u\|_{\mathrm{op}} \le L_H(1+t)\|v-u\|_{L^2_{\mu^*}}.$$

Hence,

$$\begin{aligned} \left\| \int_0^1 \left[\left(\widehat{\mathrm{H}}_{tv} - \widehat{\mathrm{H}}_v\right) - \left(\widehat{\mathrm{H}}_{tu} - \widehat{\mathrm{H}}_u\right)\right] u \mathrm{d}t \right\|_{L^2_{\mu^*}} &\le L_H \|u\|_{L^2_{\mu^*}} \|v-u\|_{L^2_{\mu^*}} \int_0^1 (1+t)\mathrm{d}t \\ &= \frac{3L_H}{2}\|u\|_{L^2_{\mu^*}} \|v-u\|_{L^2_{\mu^*}}. \end{aligned} \tag{65}$$

Combining (64) and (65), we obtain

$$\|E(v) - E(u)\|_{L^2_{\mu^*}} \le \frac{L_H}{2}\left(\|v\|_{L^2_{\mu^*}} + 3\|u\|_{L^2_{\mu^*}}\right)\|v-u\|_{L^2_{\mu^*}}. \tag{66}$$

We now expand the preconditioned gradient. Using (62), (60), and (61), we have

$$\begin{aligned} \mathfrak{G}_\beta(\widetilde{Z}) - \mathfrak{G}_\beta(Z) &= P_v g_v - P_u g_u \\ &= P_v\left(g_0 + \widehat{\mathrm{H}}_v v + E(v)\right) - P_u\left(g_0 + \widehat{\mathrm{H}}_u u + E(u)\right) \\ &= S_v(v-u) + (S_v - S_u)u \\ &\quad + P_v\left(E(v) - E(u)\right) + (P_v - P_u)E(u) + (P_v - P_u)g_0. \end{aligned}$$

Let

$$S_* := \mathrm{K}_{\mu^*}\left(\mathrm{K}^2_{\mu^*} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}. \tag{67}$$

Because S_* is a spectral function of K_{μ^*}, $S_* = \left(\mathrm{K}^2_{\mu^*} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}\mathrm{K}_{\mu^*}$. Subtracting $S_*(v-u)$ from (67) gives

$$\begin{aligned} R_\beta(Z, \widetilde{Z}) &= (S_v - S_*)(v-u) + (S_v - S_u)u \\ &\quad + P_v\left(E(v) - E(u)\right) + (P_v - P_u)E(u) + (P_v - P_u)g_0. \end{aligned} \tag{68}$$

We estimate the five terms in (68). By (59), (58), and (55),

$$\|\widehat{\mathrm{H}}_v - \mathrm{K}_{\mu^*}\|_{\mathrm{op}} \le \|\widehat{\mathrm{H}}_v - \widehat{\mathrm{H}}_0\|_{\mathrm{op}} + \|\,\mathrm{M}_{\mu^*}\|_{\mathrm{op}} \le L_H\|v\|_{L^2_{\mu^*}} + R_F\|g_0\|_{L^2_{\mu^*}}.$$

Applying the Hessian estimate from Lemma L.16, i.e.,

$$\left\|A(A^2+\beta I)^{-1/2} - B(B^2+\beta I)^{-1/2}\right\|_{\mathrm{op}} \le \frac{1}{\sqrt{\beta}}\|A-B\|_{\mathrm{op}},$$

we obtain

$$\|(S_v - S_*)(v-u)\|_{L^2_{\mu^*}} \le \left(\frac{L_H}{\sqrt{\beta}}\|v\|_{L^2_{\mu^*}} + \frac{R_F}{\sqrt{\beta}}\|g_0\|_{L^2_{\mu^*}}\right)\|v-u\|_{L^2_{\mu^*}}. \tag{69}$$

Similarly,

$$\|(S_v - S_u)u\|_{L^2_{\mu^*}} \le \frac{L_H}{\sqrt{\beta}}\|u\|_{L^2_{\mu^*}}\|v-u\|_{L^2_{\mu^*}}. \tag{70}$$

Since $\|P_v\|_{\mathrm{op}} \le \frac{1}{\sqrt{\beta}}$, (66) gives

$$\|P_v(E(v)-E(u))\|_{L^2_{\mu^*}} \le \frac{L_H}{2\sqrt{\beta}}\left(\|v\|_{L^2_{\mu^*}} + 3\|u\|_{L^2_{\mu^*}}\right)\|v-u\|_{L^2_{\mu^*}}. \tag{71}$$

The other estimate from Lemma L.16 and (58) imply

$$\|P_v - P_u\|_{\mathrm{op}} \le \frac{2L_H}{\pi\beta}\|v-u\|_{L^2_{\mu^*}}. \tag{72}$$

Combining (72) with (63), we get

$$\|(P_v - P_u)E(u)\|_{L^2_{\mu^*}} \le \frac{L_H^2}{\pi\beta}\|u\|^2_{L^2_{\mu^*}}\|v-u\|_{L^2_{\mu^*}}. \tag{73}$$

Finally,

$$\|(P_v - P_u)g_0\|_{L^2_{\mu^*}} \le \frac{2L_H}{\pi\beta}\|g_0\|_{L^2_{\mu^*}}\|v-u\|_{L^2_{\mu^*}}. \tag{74}$$

Combining (69)–(74) gives the asymmetric estimate

$$\begin{aligned}\|R_\beta(Z,\widetilde{Z})\|_{L^2_{\mu^*}} \le \Bigg[&\frac{L_H}{2\sqrt{\beta}}\left(5\|u\|_{L^2_{\mu^*}} + 3\|v\|_{L^2_{\mu^*}}\right)\\ &+ \frac{L_H^2}{\pi\beta}\|u\|^2_{L^2_{\mu^*}} + \left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_H}{\pi\beta}\right)\|g_0\|_{L^2_{\mu^*}}\Bigg]\|v-u\|_{L^2_{\mu^*}}.\end{aligned} \tag{75}$$

Interchanging u and v in the preceding argument gives

$$\begin{aligned}\|R_\beta(Z,\widetilde{Z})\|_{L^2_{\mu^*}} \le \Bigg[&\frac{L_H}{2\sqrt{\beta}}\left(3\|u\|_{L^2_{\mu^*}} + 5\|v\|_{L^2_{\mu^*}}\right) + \frac{L_H^2}{\pi\beta}\|v\|^2_{L^2_{\mu^*}}\\ &+ \left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_H}{\pi\beta}\right)\|g_0\|_{L^2_{\mu^*}}\Bigg]\|v-u\|_{L^2_{\mu^*}}.\end{aligned} \tag{76}$$

Indeed, after interchanging the two maps, the corresponding remainder is $-R_\beta(Z,\widetilde{Z})$. Since the norm of the remainder is bounded by both right-hand sides, it is also bounded by their arithmetic mean. Averaging (75) and (76) gives

$$\begin{aligned}\|R_\beta(Z,\widetilde{Z})\|_{L^2_{\mu^*}} \le \Bigg[&\frac{2L_H}{\sqrt{\beta}}\left(\|u\|_{L^2_{\mu^*}} + \|v\|_{L^2_{\mu^*}}\right) + \frac{L_H^2}{2\pi\beta}\left(\|u\|^2_{L^2_{\mu^*}} + \|v\|^2_{L^2_{\mu^*}}\right)\\ &+ \left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_H}{\pi\beta}\right)\|g_0\|_{L^2_{\mu^*}}\Bigg]\|v-u\|_{L^2_{\mu^*}}.\end{aligned}$$

Substituting $u = Z - \mathrm{Id}$, $v = \widetilde{Z} - \mathrm{Id}$, $g_0 = \nabla_\mu F(\mu^*,\cdot)$ proves (56). If $\|u\|_{L^2_{\mu^*}}, \|v\|_{L^2_{\mu^*}} \le r$, then

$$\|u\|_{L^2_{\mu^*}} + \|v\|_{L^2_{\mu^*}} \le 2r, \quad \|u\|^2_{L^2_{\mu^*}} + \|v\|^2_{L^2_{\mu^*}} \le 2r^2,$$

which gives (57). □

L.4 WELL-POSEDNESS AND CHARACTERIZATION OF THE OPTIMIZER

In this subsection, we verify that the variational problems defining the Newton-type updates are well-posed. More precisely, we show that each update admits a unique minimizer in the corresponding Hilbert space and derive the associated first-order optimality condition. This yields the explicit form of the update whenever the relevant operator is invertible. We begin with the Levenberg–Marquardt regularization of the Wasserstein Newton step.

Levenberg–Marquardt regularized Newton method. It is worth emphasizing that WSFN differs from the Wasserstein analogue of the Levenberg–Marquardt regularized Newton method, namely,

$$\mu^{n+1} = \Big(\mathrm{Id} - \big(\mathrm{H}_{\mu^n} + \tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\big)^{-1}\,\nabla_\mu F(\mu^n,\cdot)\Big)_{\#}\,\mu^n.$$

The latter interpolates between Newton's method (as $\tau \to +\infty$) and gradient descent (as $\tau \to 0$). However, for the regularized Hessian $\mathrm{H}_{\mu^n} + \tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ to be invertible, one must either assume that $\mathrm{H}_{\mu^n} \succeq 0$, or impose a restriction of the form $\tau < \lambda^{-1}$ when $\mathrm{H}_{\mu^n} \succeq -\lambda\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. In other words, the regularization $\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ suppresses negative curvature in non-convex regions, thereby enforcing a descent direction but potentially at the expense of a prohibitively small stepsize. By contrast, WSFN regularizes the non-negative operator $\mathrm{H}^2_{\mu^n}$. As a result, the parameter β can be chosen independently of the eigenvalues of H_{μ^n}, enough to guarantee invertibility of $\mathrm{H}^2_{\mu^n} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$.

Proposition L.18 (Well-posedness and first-order optimality condition of Levenberg–Marquardt algorithm)**.** *Assume $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is Wasserstein regular. Let $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\tau > 0$. Define the sequence of iterates $(v^n)_{n\in\mathbb{N}}$ generated by the update*

$$v^{n+1} = \operatorname*{arg\,min}_{v\in L^2_{\mu^n}} \left\{ \langle \nabla_\mu F(\mu^n,\cdot), v\rangle_{L^2_{\mu^n}} + \frac{1}{2}\langle \mathrm{H}_{\mu^n}\, v, v\rangle_{L^2_{\mu^n}} + \frac{1}{2\tau}\|v\|^2_{L^2_{\mu^n}} \right\},$$
$$\mu^{n+1} = (\mathrm{Id} + v^{n+1})_{\#}\mu^n.$$

Assume that there exists $\delta > 0$ such that $\langle \mathrm{H}_{\mu^n}\, v, v\rangle_{L^2_{\mu_n}} \geq -\delta\|v\|^2_{L^2_{\mu_n}}$, for all $v \in L^2_{\mu_n}$, and $\tau < \delta^{-1}$, or H_{μ^n} is non-negative, for all n. Then the sequence $(v^n)_{n\in\mathbb{N}}$ is well-defined and remains in $L^2_{\mu^n}$, for all n. Moreover, v^{n+1} satisfies

$$v^{n+1} = -\big(\mathrm{H}_{\mu^n} + \tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\big)^{-1}\,\nabla_\mu F(\mu^n,\cdot) \;\text{ in } L^2_{\mu^n}.$$

Proof. For all $v \in L^2_{\mu^n}$ and $\tau > 0$, define

$$J_n(v) = \langle \nabla_\mu F(\mu^n,\cdot), v\rangle_{L^2_{\mu^n}} + \frac{1}{2}\langle \mathrm{H}_{\mu^n}\, v, v\rangle_{L^2_{\mu^n}} + \frac{1}{2\tau}\|v\|^2_{L^2_{\mu^n}}.$$

Note that, given $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, for each $n \in \mathbb{N}$, the functional $J_n(v)$ admits a unique minimizer, denoted v^{n+1}, since

$$J_n(v) \geq -\|\nabla_\mu F(\mu^n,\cdot)\|_{L^2_{\mu^n}}\|v\|_{L^2_{\mu^n}} + \frac{1}{2}\left(\frac{1}{\tau} - \delta\right)\|v\|^2_{L^2_{\mu^n}}.$$

Now, for $\varepsilon \in [0,1]$, and all $v \in L^2_{\mu^n}$, define $v^{n,\varepsilon} := v^{n+1} + \varepsilon(v - v^{n+1}) \in L^2_{\mu^n}$. By linearity of the first term of J_n,

$$\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\left(\langle \nabla_\mu F(\mu^n,\cdot), v^{\varepsilon,n}\rangle_{L^2_{\mu^n}} - \big\langle \nabla_\mu F(\mu^n,\cdot), v^{n+1}\big\rangle_{L^2_{\mu^n}}\right) = \big\langle \nabla_\mu F(\mu^n,\cdot), v - v^{n+1}\big\rangle_{L^2_{\mu^n}}.$$

By symmetry of H_{μ^n}, a straightforward calculation gives

$$\begin{aligned}
&\lim_{\varepsilon\downarrow 0}\frac{1}{2\varepsilon}\left(\langle \mathrm{H}_{\mu^n}\, v^{n,\varepsilon}, v^{n,\varepsilon}\rangle_{L^2_{\mu^n}} - \langle \mathrm{H}_{\mu^n}\, v^{n+1}, v^{n+1}\rangle_{L^2_{\mu^n}}\right)\\
&= \lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\left(\varepsilon\langle \mathrm{H}_{\mu^n}\, v^{n+1}, v - v^{n+1}\rangle_{L^2_{\mu^n}} + \frac{\varepsilon^2}{2}\langle \mathrm{H}_{\mu^n}(v - v^{n+1}), v - v^{n+1}\rangle_{L^2_{\mu^n}}\right)\\
&= \langle \mathrm{H}_{\mu^n}\, v^{n+1}, v - v^{n+1}\rangle_{L^2_{\mu^n}}.
\end{aligned}$$

Similarly,

$$\lim_{\varepsilon\downarrow 0}\frac{1}{2\tau\varepsilon}\left(\|v^{\varepsilon,n}\|^2_{L^2_{\mu^n}}-\|v^{n+1}\|^2_{L^2_{\mu^n}}\right)=\tau^{-1}\langle v^{n+1},v-v^{n+1}\rangle_{L^2_{\mu^n}}.$$

Since v^{n+1} is the unique minimizer of J_n,

$$\begin{aligned}0&\le\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\left(J_n(v^{n,\varepsilon})-J_n(v^{n+1})\right)\\&=\left\langle\nabla_\mu F(\mu^n,\cdot),v-v^{n+1}\right\rangle_{L^2_{\mu^n}}+\langle \mathrm{H}_{\mu^n}v^{n+1},v-v^{n+1}\rangle_{L^2_{\mu^n}}+\tau^{-1}\langle v^{n+1},v-v^{n+1}\rangle_{L^2_{\mu^n}}\\&=\left\langle\nabla_\mu F(\mu^n,\cdot)+\mathrm{H}_{\mu^n}v^{n+1}+\tau^{-1}v^{n+1},v-v^{n+1}\right\rangle_{L^2_{\mu^n}},\end{aligned}$$

for all $v\in L^2_{\mu^n}$. Hence,

$$\nabla_\mu F(\mu^n,\cdot)+\mathrm{H}_{\mu^n}v^{n+1}+\tau^{-1}v^{n+1}=0\ \text{ in } L^2_{\mu^n}.$$

Equivalently,

$$\left(\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)v^{n+1}=-\nabla_\mu F(\mu^n,\cdot)\ \text{ in } L^2_{\mu^n}. \tag{77}$$

Since there exists $\delta>0$ such that $\langle \mathrm{H}_{\mu^n}v,v\rangle_{L^2_{\mu_n}}\ge-\delta\|v\|^2_{L^2_{\mu_n}}$, for all $v\in L^2_{\mu_n}$, and $\tau<\delta^{-1}$, then

$$\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\succeq m\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}},$$

with $m:=\tau^{-1}-\delta>0$. Therefore, $\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ is an invertible operator in $L^2_{\mu^n}$, and by (77)

$$v^{n+1}=-\left(\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\nabla_\mu F(\mu^n,\cdot)\ \text{ in } L^2_{\mu^n}.$$

If H_{μ^n} is non-negative, for all n, then

$$\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\succeq\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}},$$

hence $\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ is again an invertible operator in $L^2_{\mu^n}$ and the same expression for v^{n+1} holds.

Finally, $v^{n+1}\in L^2_{\mu^n}$, since $\nabla_\mu F(\mu^n,\cdot)\in L^2_{\mu^n}$ and either $\|\left(\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\|_{\mathrm{op}}\le m^{-1}$ or $\|\left(\mathrm{H}_{\mu^n}+\tau^{-1}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1}\|_{\mathrm{op}}\le\tau$. □

Wasserstein saddle-free Newton method. Assume that H_μ is bounded and self-adjoint on L^2_μ. Given $v\in L^2_\mu$, the quadratic Taylor model that approximates F around μ is $\mathfrak{T}^F_2(\mu)[v]=F(\mu)+\langle\nabla_\mu F(\mu,\cdot),v\rangle_{L^2_\mu}+\frac{1}{2}\langle \mathrm{H}_\mu v,v\rangle_{L^2_\mu}$. When H_μ is invertible, the Newton direction solves the stationarity equation $\mathrm{H}_\mu v=-\nabla_\mu F(\mu,\cdot)$. For an indefinite Hessian, however, this quadratic model is unbounded below along negative curvature directions, so the Newton direction is not its minimizer.

Instead, define $|\,\mathrm{H}_\mu\,|:=(\mathrm{H}_\mu^2)^{1/2}$ and replace the quadratic model by $\mathcal{Q}^{\mathrm{SFN}}_\mu(v):=\mathfrak{T}^F_1(\mu)[v]+\frac{1}{2}\langle|\,\mathrm{H}_\mu\,|v,v\rangle_{L^2_\mu}$, where $T^F_1(\mu)[v]:=F(\mu)+\langle\nabla_\mu F(\mu,\cdot),v\rangle_{L^2_\mu}$ is the linear Taylor model. Since $\mathrm{H}_\mu\preceq|\,\mathrm{H}_\mu\,|$, we have $\mathfrak{T}^F_2(\mu)[v]\le\mathcal{Q}^{\mathrm{SFN}}_\mu(v)$; see Lemma L.9 and (Kato, 1995, Problem 2.35). This is an upper bound on the quadratic Taylor model. If $|\,\mathrm{H}_\mu\,|\succeq m\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ for some $m>0$, the surrogate $\mathcal{Q}^{\mathrm{SFN}}_\mu$ is strongly convex and has the unique minimizer $v_{\mathrm{SFN}}(\mu)=-|\,\mathrm{H}_\mu\,|^{-1}\nabla_\mu F(\mu,\cdot)$. The absolute value Hessian principle is inspired by finite-dimensional saddle-free Newton methods. However, its use on Wasserstein space must address the following operator level difficulties.

Lack of bounded invertibility. Non-negativity of $|\,\mathrm{H}_\mu\,|$ does not guarantee that its inverse exists as a bounded operator, as the Hessian may have a nontrivial kernel. Even when its kernel is trivial, spectral values may accumulate at zero due to compactness. Thus, injectivity alone does not ensure that $|\,\mathrm{H}_\mu\,|^{-1}\nabla_\mu F(\mu,\cdot)$ is defined for every gradient in L^2_μ or depends continuously on it. Requiring a uniform gap around zero would exclude precisely the small curvature regimes that our local spectral analysis is intended to accommodate.

Lack of operator Lipschitz stability. Although the scalar function $x \mapsto |x|$ is Lipschitz, the map $A \mapsto |A|$ need not be operator Lipschitz in the operator norm for bounded self-adjoint operators; see (Reed & Simon, 1980, VI: Bounded Operators, Problem 17). Thus, a Lipschitz estimate for the Hessian does not necessarily give the corresponding estimate for its absolute value. This matters because quantitative saddle escape requires comparing preconditioners along evolving nonlinear transport trajectories, not only computing their spectra at a fixed critical point.

We address these difficulties by replacing $|\,\mathrm{H}_\mu\,|$ with $(\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{1/2}$, $\beta > 0$. Functional calculus gives $\mathrm{H}_\mu \preceq |\,\mathrm{H}_\mu\,| \preceq (\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{1/2}$, and $(\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{1/2} \succeq \sqrt{\beta}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. Thus, the regularized quadratic model still bounds the second-order Taylor model and is now coercive without a spectral gap assumption on H_μ.

Given $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, $\beta > 0$, and $\tau > 0$, we define WSFN by

$$\begin{aligned} v_{\mathrm{WSFN}}(\mu^n) &= \operatorname*{arg\,min}_{v \in L^2_{\mu^n}} \left\{ \mathfrak{T}_1^F(\mu^n)[v] + \frac{1}{2} \left\langle (\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{1/2} v, v \right\rangle_{L^2_{\mu^n}} \right\}, \\ \mu^{n+1} &= \left(\mathrm{Id} + \tau v_{\mathrm{WSFN}}(\mu^n)\right)_{\#} \mu^n. \end{aligned} \tag{78}$$

The minimizer is unique and satisfies $v_{\mathrm{WSFN}}(\mu) = -P_\beta(\mathrm{H}_\mu)\nabla_\mu F(\mu, \cdot)$. Hence,

$$\mu^{n+1} = \left(\mathrm{Id} - \tau(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2} \nabla_\mu F(\mu^n, \cdot)\right)_{\#} \mu^n.$$

Proposition L.19 (Well-posedness and first-order optimality condition of (78))**.** *Let Assumption 4.4 hold. Let $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\tau, \beta > 0$. Then the sequence of iterates $(v^n)_{n\in\mathbb{N}}$ generated by the update (78) is well-defined and remains in $L^2_{\mu^n}$, for all n. Moreover, v^{n+1} satisfies*

$$v^{n+1} = -\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-\frac{1}{2}} \nabla_\mu F(\mu^n, \cdot) \ \text{ in } L^2_{\mu^n}.$$

Proof. By Lemma L.9, $A_{\mu^n} := \mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ is self-adjoint. Furthermore, because $\mathrm{H}_{\mu^n}^2 \succeq 0$, and $\beta > 0$ the operator $A_{\mu^n} \succeq \beta$. Let $\sigma(A_{\mu^n})$ be the spectrum of A_{μ^n}. Since A_{μ^n} is self-adjoint, we have $\sigma(A_{\mu^n}) \subset \mathbb{R}$, and in particular $\sigma(A_{\mu^n}) \subset [\beta, \infty)$.

Let f be a continuous function over the spectrum $\sigma(A_\mu)$. By (Reed & Simon, 1980, Theorem VII.1 (continuous functional calculus)),

$$f(A_{\mu^n})^* = \bar{f}(A_{\mu^n}),$$

where $\bar{f}$ denotes the complex conjugate of f. Take $f(\lambda) = \lambda^{\frac{1}{2}}$ for $\lambda \in \sigma(A_{\mu^n})$. Since $\sigma(A_{\mu^n}) \subset [\beta, \infty)$, it follows that $f(\lambda) = \bar{f}(\lambda)$. Therefore,

$$\left(\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}}\right)^* = \left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}},$$

i.e., $\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}}$ is self-adjoint.

For all $v \in L^2_{\mu^n}$, define

$$J_n(v) = \langle \nabla_\mu F(\mu^n, \cdot), v \rangle_{L^2_{\mu^n}} + \frac{1}{2} \langle \left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}} v, v \rangle_{L^2_{\mu^n}}.$$

Therefore, given $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$, for each $n \in \mathbb{N}$, the functional $J_n(v)$ admits a unique minimizer, denoted v^{n+1}, since it is the sum of a linear and a quadratic term.

Hence, following the argument in Proposition L.18, since $\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}}$ is self-adjoint, it is symmetric, so we get

$$\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}} v^{n+1} = -\nabla_\mu F(\mu^n, \cdot) \ \text{ in } L^2_{\mu^n}. \tag{79}$$

Note that $\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}} \subset [\sqrt{\beta}, \infty)$. It follows that $\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{\frac{1}{2}}$ is invertible and so (79) becomes

$$v^{n+1} = -\left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-\frac{1}{2}} \nabla_\mu F(\mu^n, \cdot) \ \text{ in } L^2_{\mu^n}.$$

Finally, $v^{n+1} \in L^2_{\mu^n}$, since $\nabla_\mu F(\mu^n, \cdot) \in L^2_{\mu^n}$ and $\left\| \left(\mathrm{H}_{\mu^n}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-\frac{1}{2}} \right\|_{\mathrm{op}} \leq \beta^{-\frac{1}{2}}$. □

These two results justify the explicit formulas for the regularized Newton and WSFN updates used in the main text.

L.5 Linearization of preconditioned Wasserstein updates

This appendix derives the spectral multipliers in Table 1. All comparisons are made on the common reference space $L^2_{\mu^*}$, where μ^* is a critical point, i.e., $\nabla_\mu F(\mu^*, \cdot) = 0$.

Curvature interpretation. The second-order expansion in Lemma L.6 gives, along $\mathrm{H}_{\mu^*} e = \lambda e$,

$$F\left((\mathrm{Id} + ae)_{\#}\mu^*\right) = F(\mu^*) + \frac{1}{2}\lambda a^2 \|e\|^2_{L^2_{\mu^*}} + o(a^2) \quad \text{as } a \to 0.$$

Thus, the sign of λ describes the curvature of the objective along this perturbation curve. The multipliers below describe how each linearized update acts on that direction.

Representation on a common reference space. For $v \in L^2_{\mu^*}$, define $\mathrm{T}_v := \mathrm{Id} + v$, $\mu_v := (\mathrm{T}_v)_{\#}\mu^*$. Composition defines an isometric embedding $\mathcal{C}_v : L^2_{\mu_v} \to L^2_{\mu^*}$, given by $\mathcal{C}_v h := h \circ \mathrm{T}_v$. In particular, this construction does not require T_v to be invertible. The pulled-back Hessian is the operator $\widetilde{\mathrm{H}}_{\mu^*,v} : L^2_{\mu^*} \to L^2_{\mu^*}$ given by

$$\widetilde{\mathrm{H}}_{\mu^*,v}[h](x) := \nabla \nabla_\mu F(\mu_v, \mathrm{T}_v(x)) h(x) + \int_{\mathbb{R}^d} \nabla^2_\mu F(\mu_v, \mathrm{T}_v(x), \mathrm{T}_v(y)) h(y) \mathrm{d}\mu^*(y).$$

Under regularity of F, in particular symmetry, and Assumption 4.4, $\widetilde{\mathrm{H}}_{\mu^*,v}$ is bounded and self-adjoint near $v = 0$, $\widetilde{\mathrm{H}}_{\mu^*,0} = \mathrm{H}_{\mu^*}$, and, under Assumption 4.5, by Lemma L.13,

$$\|\widetilde{\mathrm{H}}_{\mu^*,v} - \widetilde{\mathrm{H}}_{\mu^*,w}\|_{\mathrm{op}} \le L_{\mathrm{H}} \|v - w\|_{L^2_{\mu^*}}.$$

Define $G(v) := \nabla_\mu F(\mu_v, \cdot) \circ \mathrm{T}_v$. Proposition 2.3 gives $G(v) = \int_0^1 \widetilde{\mathrm{H}}_{\mu^*,sv} v \mathrm{d}s$, because $G(0) = 0$. Thus,

$$G(v) = \mathrm{H}_{\mu^*} v + R_G(v), \quad \|R_G(v)\|_{L^2_{\mu^*}} \le \frac{L_{\mathrm{H}}}{2} \|v\|^2_{L^2_{\mu^*}}. \tag{80}$$

Directly from the definitions, $\widetilde{\mathrm{H}}_{\mu^*,v} \mathcal{C}_v = \mathcal{C}_v \mathrm{H}_{\mu_v}$. Since the operators are self-adjoint, this intertwining identity extends by functional calculus to operator functions $f(\widetilde{\mathrm{H}}_{\mu^*,v})\mathcal{C}_v = \mathcal{C}_v f(\mathrm{H}_{\mu_v})$, whenever the indicated functions of the operators are well defined; see (Reed & Simon, 1980, Theorem VII.2 (g)). Thus, preconditioned gradients can be represented on $L^2_{\mu^*}$ without identifying composition with an invertible change of coordinates.

A common linearization formula. Let $\mathrm{B}(v)$ be the common reference preconditioner, and assume that $\mathrm{B}(v)$ is continuous in operator norm at $v = 0$. Define the displacement update on $L^2_{\mu^*}$ by

$$\Psi^{\mathrm{B}}_\tau(v) := v - \tau\, \mathrm{B}(v) G(v).$$

Equivalently, the updated transport map from μ^* is $\mathrm{Id} + \Psi^{\mathrm{B}}_\tau(v)$. Using (80), we have

$$\mathrm{B}(v)G(v) - \mathrm{B}(0)\,\mathrm{H}_{\mu^*} v = \mathrm{B}(v)\,(G(v) - \mathrm{H}_{\mu^*} v) + (\mathrm{B}(v) - \mathrm{B}(0))\,\mathrm{H}_{\mu^*} v.$$

The first term is $O(\|v\|^2_{L^2_{\mu^*}})$ because $\mathrm{B}(v)$ is locally bounded, and the second is $o(\|v\|_{L^2_{\mu^*}})$ by operator norm continuity. Hence

$$\Psi^{\mathrm{B}}_\tau(v) = (\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau\, \mathrm{B}(0)\,\mathrm{H}_{\mu^*})\, v + o(\|v\|_{L^2_{\mu^*}}), \tag{81}$$

and therefore by definition of Fréchet derivative,

$$D\Psi^{\mathrm{B}}_\tau(0) = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau\, \mathrm{B}(0)\,\mathrm{H}_{\mu^*}.$$

No derivative of $\mathrm{B}(v)$ is needed. Continuity suffices because the gradient vanishes at the reference measure μ^*.

Gradient descent. For $\mathrm{B}(v) = \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$, (81) gives $D\Psi^{\mathrm{GD}}_\tau(0) = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau\, \mathrm{H}_{\mu^*}$. Thus, if $\mathrm{H}_{\mu^*} e = \lambda e$, then $D\Psi^{\mathrm{GD}}_\tau(0) e = (1 - \tau\lambda) e$.

Newton. Suppose that H_{μ^*} has a bounded inverse. By continuity of $\widetilde{\mathrm{H}}_{\mu^*,v}$, the operators $\widetilde{\mathrm{H}}_{\mu^*,v}$ are also boundedly invertible for sufficiently small $\|v\|_{L^2_{\mu^*}}$, and $\widetilde{\mathrm{H}}^{-1}_{\mu^*,v} \to \mathrm{H}^{-1}_{\mu^*}$ in operator norm. Taking $\mathrm{B}(v) = \widetilde{\mathrm{H}}^{-1}_{\mu^*,v}$ gives $D\Psi^{\mathrm{N}}_\tau(0) = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau\,\mathrm{H}^{-1}_{\mu^*}\mathrm{H}_{\mu^*} = (1-\tau)\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. The multiplier is therefore independent of the sign of the Hessian eigenvalue.

Ideal saddle-free Newton. Again suppose that H_{μ^*} is boundedly invertible. Its spectrum is then separated from zero. The same holds for $\widetilde{\mathrm{H}}_{\mu^*,v}$ in a sufficiently small neighborhood of $v = 0$, and continuous functional calculus on this neighborhood gives $|\widetilde{\mathrm{H}}_{\mu^*,v}|^{-1} \to |\,\mathrm{H}_{\mu^*}\,|^{-1}$ in operator norm. Taking $\mathrm{B}(v) = |\widetilde{\mathrm{H}}_{\mu^*,v}|^{-1}$ gives $D\Psi^{\mathrm{SFN}}_\tau(0) = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau|\,\mathrm{H}_{\mu^*}\,|^{-1}\mathrm{H}_{\mu^*}$. For $\mathrm{H}_{\mu^*}\,e = \lambda e$, then $D\Psi^{\mathrm{SFN}}_\tau(0)e = \big(1-\tau\lambda|\lambda|^{-1}\big)\,e$.

Thus, for $0 < \tau < 2$, positive curvature directions contract, whereas negative curvature directions expand by $1+\tau$. The continuity used here does not contradict the lack of a general operator Lipschitz bound for $A \mapsto |A|$ since this argument is local to a neighborhood whose spectra are separated from zero.

Regularized WSFN. For $\beta > 0$, take $\mathrm{B}(v) = P_\beta(\widetilde{\mathrm{H}}_{\mu^*,v}) = (\widetilde{\mathrm{H}}^2_{\mu^*,v} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}$. Lemma L.16 ensures operator norm continuity without a spectral gap assumption. Therefore,

$$D\Psi^{\mathrm{WSFN}}_\tau(0) = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau P_\beta(\mathrm{H}_{\mu^*})\,\mathrm{H}_{\mu^*} = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau S_\beta(\mathrm{H}_{\mu^*}).$$

In an eigendirection,

$$D\Psi^{\mathrm{WSFN}}_\tau(0)e = \left(1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}\right)e.$$

The normalized Hessian estimate also gives a quantitative remainder with explicit dependence on β. Indeed, write $G(v) = \widetilde{\mathrm{H}}_{\mu^*,v}v + E(v)$, where

$$E(v) = \int_0^1 \left(\widetilde{\mathrm{H}}_{\mu^*,sv} - \widetilde{\mathrm{H}}_{\mu^*,v}\right) v\,\mathrm{d}s, \quad \|E(v)\|_{L^2_{\mu^*}} \le \frac{L_{\mathrm{H}}}{2}\|v\|^2_{L^2_{\mu^*}}.$$

Then

$$P_\beta(\widetilde{\mathrm{H}}_{\mu^*,v})G(v) - S_\beta(\mathrm{H}_{\mu^*})v = \Big(S_\beta(\widetilde{\mathrm{H}}_{\mu^*,v}) - S_\beta(\mathrm{H}_{\mu^*})\Big)\,v + P_\beta(\widetilde{\mathrm{H}}_{\mu^*,v})E(v).$$

Applying Lemma L.16 to the two terms gives

$$\Big\|P_\beta(\widetilde{\mathrm{H}}_{\mu^*,v})G(v) - S_\beta(\mathrm{H}_{\mu^*})v\Big\|_{L^2_{\mu^*}} \le \frac{3L_{\mathrm{H}}}{2\sqrt{\beta}}\|v\|^2_{L^2_{\mu^*}}.$$

In particular,

$$\big\|\Psi^{\mathrm{WSFN}}_\tau(v) - (\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau S_\beta(\mathrm{H}_{\mu^*}))\,v\big\|_{L^2_{\mu^*}} \le \frac{3\tau L_{\mathrm{H}}}{2\sqrt{\beta}}\|v\|^2_{L^2_{\mu^*}}.$$

L.6 Relation with Newton's method

In this subsection, we record one complementary local result concerning the quadratic convergence rate of the Wasserstein Newton method on positive spectral subspaces. The relation between WSFN and Wasserstein Newton must be understood spectrally. Since K_{μ^*} is compact, its inverse is generally unbounded on $(\ker \mathrm{K}_{\mu^*})^\perp$. Thus, one does not have $(\mathrm{K}^2_{\mu^*} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2} \to \mathrm{K}^{-1}_{\mu^*}$ in operator norm on $L^2_{\mu^*}$ as $\beta \downarrow 0$.

The Newton interpretation is nevertheless valid on every positive spectral subspace. Indeed, since K_{μ^*} is compact, $\mathrm{Ran}(P_m)$ is finite-dimensional, and the restricted operator $\mathrm{K}_m := \mathrm{K}_{\mu^*}|_{\mathrm{Ran}(P_m)}$ is boundedly invertible, with $\|\,\mathrm{K}_m^{-1}\,\|_{\mathrm{op}} \le \frac{1}{m}$. The spectral functional calculus gives

$$\Big\|\Big((\mathrm{K}^2_{\mu^*} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2} - \mathrm{K}_m^{-1}\Big)\,P_m\Big\|_{\mathrm{op}} \le \frac{1}{m} - \frac{1}{\sqrt{m^2+\beta}} \to 0$$

as $\beta \downarrow 0$. Thus, on $\mathrm{Ran}(P_m)$, the WSFN preconditioner converges in operator norm to the Newton preconditioner. In particular, for the unit stepsize $\tau = 1$,

$$\|\mathcal{L}_{1,\beta} P_m\|_{\mathrm{op}} \leq 1 - \frac{m}{\sqrt{m^2 + \beta}} \to 0,$$

as $\beta \downarrow 0$. Hence the first-order term vanishes in the limit $\beta \downarrow 0$ on every fixed positive spectral subspace, which is precisely the cancellation responsible for the quadratic convergence rate of Newton's method.

On the full space, however, one only has the strong convergence $(\mathrm{K}_{\mu^*}^2 + \beta\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\, \mathrm{K}_{\mu^*} \to \mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}}$, as $\beta \downarrow 0$, and so $\mathcal{L}_{1,\beta} \to P_{\ker \mathrm{K}_{\mu^*}}$ strongly. This convergence is not uniform. If the positive eigenvalues of K_{μ^*} accumulate at zero, then $\|\mathcal{L}_{1,\beta}\|_{\mathrm{op}} = 1$ for every $\beta > 0$. Accordingly, the limit $\beta \downarrow 0$ does not imply a full space quadratic convergence theorem for the nonlinear WSFN iteration. The corresponding quadratic Newton result is valid after restricting the dynamics to a fixed positive spectral subspace, as the following theorem shows.

Theorem L.20 (Quadratic convergence of spectrally projected Wasserstein Newton)**.** *Let Assumptions 4.4 and 4.5 hold. Let $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ be a global minimizer of F. Fix $m > 0$, and define $P_m := \mathbb{1}_{[m,\infty)}(\mathrm{K}_{\mu^*})$. Assume that $\mathrm{Ran}(P_m) \neq \{0\}$. For $v \in \mathrm{Ran}(P_m)$, let $Z_v := \mathrm{Id} + v$, and $\mu_v := (Z_v)_{\#}\mu^*$. Its gradient on $\mathrm{Ran}(P_m)$ is $G_m(v) := P_m\left(\nabla_\mu F(\mu_v, \cdot) \circ Z_v\right)$. Let $A_m(v) := P_m \widetilde{\mathrm{H}}_{\mu^*,v}|_{\mathrm{Ran}(P_m)} : \mathrm{Ran}(P_m) \to \mathrm{Ran}(P_m)$ denote the corresponding projected pulled-back Hessian defined by*

$$\widetilde{\mathrm{H}}_{\mu^*,v}[h](x) := \nabla\nabla_\mu F(\mu_v, Z_v(x))h(x) + \int_{\mathbb{R}^d} \nabla_\mu^2 F\left(\mu_v, Z_v(x), Z_v(y)\right) h(y) \mathrm{d}\mu^*(y),$$

for $h \in L^2_{\mu^*}$. Then

$$\|A_m(v) - A_m(u)\|_{\mathrm{op}} \leq L_{\mathrm{H}} \|v - u\|_{L^2_{\mu^*}}, \quad u, v \in \mathrm{Ran}(P_m).$$

Let $0 < \alpha < \frac{m}{2L_{\mathrm{H}}}$, and consider the projected Newton iteration

$$v^{n+1} = v^n - A_m(v^n)^{-1} G_m(v^n), \quad \mu^n := (\mathrm{Id} + v^n)_{\#}\mu^*,$$

initialized at $v^0 \in \mathrm{Ran}(P_m)$ with $\|v^0\|_{L^2_{\mu^*}} < \alpha$. Then $A_m(v^n)$ is invertible, $v^n \in B_{\mathrm{Ran}(P_m)}(\alpha)$ for every $n \geq 0$, and

$$W_2(\mu^n, \mu^*) \leq \frac{m}{L_{\mathrm{H}}} \left(\frac{1}{2}\right)^{2^n}.$$

Proof. Since μ^* is a global minimizer, $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$, and hence $G_m(0) = 0$. Moreover, $\mathrm{M}_{\mu^*} = 0$, so $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$. It follows that $A_m(0) = P_m\, \mathrm{K}_{\mu^*}|_{\mathrm{Ran}(P_m)} = \mathrm{K}_{\mu^*}|_{\mathrm{Ran}(P_m)}$.

By the definition of $\mathrm{Ran}(P_m)$, $\langle A_m(0)z, z\rangle_{L^2_{\mu^*}} \geq m\|z\|^2_{L^2_{\mu^*}}$, for every $z \in \mathrm{Ran}(P_m)$. For $h \in \mathrm{Ran}(P_m)$,

$$\left(A_m(v) - A_m(u)\right) h = P_m\left(\widetilde{\mathrm{H}}_{\mu^*,v} - \widetilde{\mathrm{H}}_{\mu^*,u}\right) h,$$

for every $v, u \in \mathrm{Ran}(P_m)$. Since P_m is orthogonal projection, $\|P_m z\|_{L^2_{\mu^*}} \leq \|z\|_{L^2_{\mu^*}}$, and therefore by Lemma L.13,

$$\|A_m(v) - A_m(u)\|_{\mathrm{op}} \leq L_{\mathrm{H}} \|v - u\|_{L^2_{\mu^*}}. \tag{82}$$

For $v \in B_{\mathrm{Ran}(P_m)}(\alpha)$, this Lipschitz estimate gives

$$\|A_m(v) - A_m(0)\|_{\mathrm{op}} \leq L_{\mathrm{H}} \|v\|_{L^2_{\mu^*}} < \frac{m}{2}.$$

Therefore, $A_m(v) \succeq \frac{m}{2} I_{\mathrm{Ran}(P_m)}$, so $A_m(v)$ is invertible and

$$\|A_m(v)^{-1}\|_{\mathrm{op}} \leq \frac{2}{m}. \tag{83}$$

Because $G_m(0) = 0$, the fundamental theorem of calculus gives

$$G_m(v) = \int_0^1 A_m(tv) v \mathrm{d}t.$$

Consequently, one projected Newton step satisfies

$$v - A_m(v)^{-1}G_m(v) = A_m(v)^{-1}\left(A_m(v)v - G_m(v)\right) = A_m(v)^{-1}\int_0^1 \left(A_m(v) - A_m(tv)\right) v\mathrm{d}t.$$

Using (82),

$$\left\| \int_0^1 \left(A_m(v) - A_m(tv)\right) v\mathrm{d}t \right\|_{L^2_{\mu^*}} \le \int_0^1 L_{\mathrm{H}}(1-t)\|v\|^2_{L^2_{\mu^*}}\,\mathrm{d}t = \frac{L_{\mathrm{H}}}{2}\|v\|^2_{L^2_{\mu^*}}.$$

Combining this with (83) gives

$$\|v^{n+1}\|_{L^2_{\mu^*}} \le \frac{L_{\mathrm{H}}}{m}\|v^n\|^2_{L^2_{\mu^*}}.$$

Since $\frac{L_{\mathrm{H}}}{m}\|v^0\|_{L^2_{\mu^*}} < \frac{1}{2}$, the one-step estimate shows inductively that

$$\|v^{n+1}\|_{L^2_{\mu^*}} \le \frac{1}{2}\|v^n\|_{L^2_{\mu^*}} < \alpha.$$

Thus all the iterates remain in $B_{\mathrm{Ran}(P_m)}(\alpha)$. Setting $q_n := \frac{L_{\mathrm{H}}}{m}\|v^n\|_{L^2_{\mu^*}}$, we have $q_{n+1} \le q_n^2$, and therefore $q_n \le q_0^{2^n}$. This gives

$$\|v^n\|_{L^2_{\mu^*}} \le \frac{m}{L_{\mathrm{H}}}\left(\frac{L_{\mathrm{H}}}{m}\|v^0\|_{L^2_{\mu^*}}\right)^{2^n}.$$

Finally, the coupling $(\mathrm{Id}, \mathrm{Id}+v^n)_{\#}\,\mu^* \in \Pi(\mu^*, \mu^n)$ implies

$$W_2^2(\mu^n, \mu^*) \le \int_{\mathbb{R}^d} \|v^n(x)\|^2 \mathrm{d}\mu^*(x),$$

which proves the Wasserstein estimate. □

M PROOFS OF MAIN RESULTS

In this section, we prove the main theoretical results of the paper.

PROOF ROADMAP

The analysis proceeds through the following steps, using the auxiliary operator estimates of Appendix L.

Operator stability and transport direction. We first represent Hessians and preconditioners on a common reference space. Lemma L.13 controls the variation of the pulled-back Hessian under changes in the transport map, while Lemma L.16 gives operator Lipschitz estimates for the regularized preconditioner and normalized Hessian. Together with the composition identity of Lemma L.15, these estimates give the expansion of the preconditioned gradient in Lemma L.17. This expansion is the main tool for comparing nearby WSFN trajectories. All estimates track the dependence on regularization, which is essential when $\beta_{\mathrm{sad}} = \Theta(\delta^2)$.

Turn insufficient decrease into localization and escape. Lemma M.3 relates objective decrease to transport displacement. Thus, unless a trajectory has already achieved the required decrease, its transport maps remain in a controlled neighborhood of the episode's starting measure. Within this neighborhood, Hessian regularity and Lemma L.17 control the nonlinear errors relative to the frozen linearized dynamics. Proposition M.5 uses this localization to establish a deterministic two-trajectory escape guarantee for initializations differing along an unstable direction.

High probability net decrease. The perturbation covariance is K_μ^2, whereas the subsequent transport uses the full Hessian H_μ. Lemma D.6 decomposes the Gaussian perturbation into its component along a negative eigendirection of K_μ and an independent orthogonal component. This decomposition connects the random initialization to the deterministic comparison in Proposition M.5. Proposition M.7 then combines Gaussian estimates with control of the perturbation cost to obtain a high probability decrease measured relative to the *pre-perturbation* objective.

Count descent steps and saddle episodes. Theorem 4.6 combines the large gradient descent estimate in Lemma M.2 with the uniform saddle episode guarantee. Under the prescribed parameter choices, $\beta_{\mathrm{sad}} = \Theta(\delta^2)$ gives an episode length $n_{\mathrm{out}} = \tilde{O}(1)$ and a required net decrease $F_0 = \tilde{\Omega}(\delta^3)$. Writing $F_{\min} := F(\mu^0) - \inf_\mu F(\mu)$, objective decrease bounds the number of successful episodes by $\frac{F_{\min}}{F_0}$. Including the possible terminal episode, the number of transport steps spent in episodes satisfies

$$N_{\mathrm{sad}} \le n_{\mathrm{out}} \left(\frac{F_{\min}}{F_0} + 1 \right) = \tilde{O}\left(F_{\min} \delta^{-3} + 1 \right).$$

A conditional union bound over fresh perturbations controls the probability of terminating at a saddle.

For PWGF, the corresponding episode duration is $\tilde{O}(\delta^{-1})$ Yamamoto et al. (2025). Thus, shortening the episodes accounts for the improvement of the saddle contribution from $\tilde{O}(\delta^{-4})$ to $\tilde{O}(\delta^{-3})$. The inverse cubic factor remains because of the guaranteed objective decrease per episode. Appendix N explains this balance. Remark 4.7 clarifies why, under the required tolerance coupling, this improvement does not change the leading total complexity bound.

Characterize the local linearized dynamics. The local analysis returns to the operator expansion and applies spectral calculus at a minimizer. Theorem 4.9 separates the kernel component from positive curvature spectral bands, proving geometric contraction on bands bounded away from zero and polynomial decay under spectral source conditions. The source conditions control the contribution of small positive curvature components. These rates concern the linearized iterates. Remark 4.10 explains the additional control of nonlinear error accumulation and spectral transfer needed to extend them to the full iteration.

We now turn to the analysis of the WSFN iteration. The next lemmas provide the basic control needed on perturbed initializations and on the objective decrease along the discrete dynamics.

M.1 Auxiliary lemmas for the WSFN dynamics

In this subsection, we collect the deterministic estimates used in the analysis of the algorithm. These results control the effect of the initial perturbation and the decrease of the objective along the WSFN iterates.

Lemma M.1. *Let Assumption 4.4 hold, with $C_{\mathrm{H}} := C_{\mathrm{M}} + C_{\mathrm{K}}$. Let μ_{sad} be an (ε, δ)-saddle point, let $\xi \sim \mathcal{N}_{L^2_{\mu_{\mathrm{sad}}}}(0, \mathrm{K}^2_{\mu_{\mathrm{sad}}})$. If $\|\xi\|_{L^2_{\mu_{\mathrm{sad}}}} \le \kappa$ and $\eta = \frac{2F_0}{\kappa(\varepsilon + \sqrt{\varepsilon^2 + 2C_{\mathrm{H}}F_0})}$, for some $\kappa > 0$ and $F_0 > 0$, then $\mu^{n_{\mathrm{in}}} = (\mathrm{Id} + \eta\xi)_{\#}\mu_{\mathrm{sad}}$ satisfies*

$$F(\mu^{n_{\mathrm{in}}}) - F(\mu_{\mathrm{sad}}) \le F_0.$$

Proof. Let $\pi_t = (\mathrm{Id} + t\eta\xi)_{\#}\mu_{\mathrm{sad}}$ for $t \in [0,1]$, with $\pi_0 = \mu_{\mathrm{sad}}$ and $\pi_1 = (\mathrm{Id} + \eta\xi)_{\#}\mu_{\mathrm{sad}} = \mu^{n_{\mathrm{in}}}$. Applying Lemma L.14 gives

$$\begin{aligned} F(\mu^{n_{\mathrm{in}}}) &\le F(\mu_{\mathrm{sad}}) + \eta \langle \nabla_\mu F(\mu_{\mathrm{sad}}), \xi \rangle_{L^2_{\mu_{\mathrm{sad}}}} + \frac{C_{\mathrm{H}}}{2} \eta^2 \|\xi\|^2_{L^2_{\mu_{\mathrm{sad}}}} \\ &\le F(\mu_{\mathrm{sad}}) + \eta\kappa\varepsilon + \frac{C_{\mathrm{H}}}{2}\eta^2\kappa^2 \\ &= F(\mu_{\mathrm{sad}}) + F_0, \end{aligned}$$

where the second inequality follows from $\|\xi\|_{L^2_{\mu_{\mathrm{sad}}}} \le \kappa$ and the last equality follows from the choice of η. □

Lemma M.2 (Large gradient descent lemma)**.** *Let Assumption 4.4 hold. Let $\{\mu^k\}_{k=0}^{n-1} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a sequence of iterates generated by scheme (1) with stepsize $\tau_{\mathrm{grad}} \le \frac{\sqrt{\beta_{\mathrm{grad}}}}{C_{\mathrm{H}}}$. Then*

$$F(\mu^{k+1}) \le F(\mu^k) - \frac{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}}{2} \|(\mathrm{H}^2_{\mu^k} + \beta_{\mathrm{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}},$$

for all $k \geq 0$, *and*

$$F(\mu^0) - F(\mu^n) \geq \frac{\tau_{\text{grad}}\sqrt{\beta_{\text{grad}}}}{2} \sum_{k=0}^{n-1} \|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}}.$$

Proof. Applying Lemma L.14 with $\mu = \mu^k$ and $v = -\tau(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)$ gives

$$\begin{aligned} &F\left((\mathrm{Id} - \tau(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k))_\# \mu^k\right) \\ &\leq F(\mu^k) - \tau_{\text{grad}} \left\langle \nabla_\mu F(\mu^k), (\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k) \right\rangle_{L^2_{\mu^k}} \\ &+ \frac{\tau^2_{\text{grad}}(C_{\mathrm{M}} + C_{\mathrm{K}})}{2} \|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}}. \end{aligned}$$

By scheme (1),

$$F\left((\mathrm{Id} - \tau_{\text{grad}}(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2} \nabla_\mu F(\mu^k))_\# \mu^k\right) = F(\mu^{k+1}).$$

Since H_{μ^k} is bounded and symmetric by Assumption 4.4, it is self-adjoint, and hence so is $(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}$. Therefore, using that $\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}} \succeq \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ implies $(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \preceq \beta_{\text{grad}}^{-\frac{1}{2}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$ gives

$$\|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}} \leq \frac{1}{\sqrt{\beta_{\text{grad}}}} \left\langle \nabla_\mu F(\mu^k), (\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k) \right\rangle_{L^2_{\mu^k}},$$

Therefore, using $\tau_{\text{grad}} \leq \sqrt{\beta_{\text{grad}}}(C_{\mathrm{M}} + C_{\mathrm{K}})^{-1}$ gives

$$\begin{aligned} F(\mu^{k+1}) &\leq F(\mu^k) - \tau_{\text{grad}}\sqrt{\beta_{\text{grad}}} \|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}} \\ &\quad + \frac{\tau_{\text{grad}}\sqrt{\beta_{\text{grad}}}}{2} \|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}} \\ &= F(\mu^k) - \frac{\tau_{\text{grad}}\sqrt{\beta_{\text{grad}}}}{2} \|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}}. \end{aligned}$$

Summing from $k = 0$ to $k = n - 1$ yields

$$F(\mu^0) - F(\mu^n) \geq \frac{\tau_{\text{grad}}\sqrt{\beta_{\text{grad}}}}{2} \sum_{k=0}^{n-1} \|(\mathrm{H}^2_{\mu^k} + \beta_{\text{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^k)\|^2_{L^2_{\mu^k}}.$$

□

Lemma M.3 (Local descent of WSFN with an order-one stepsize)**.** *Let Assumptions 4.4 and 4.5 hold. Let* $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ *satisfy* $\|\nabla_\mu F(\mu^*, \cdot)\|_{L^2_{\mu^*}} \leq \varepsilon$. *For* $z \in L^2_{\mu^*}(\mathbb{R}^d; \mathbb{R}^d)$, *define* $Z_z := \mathrm{Id} + z$, $\mu_z := (Z_z)_\# \mu^*$, *and let*

$$d_z := \left(\mathrm{H}^2_{\mu_z} + \beta_{\text{sad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2} \nabla_\mu F(\mu_z, \cdot),$$

for $\beta_{\text{sad}} > 0$. *For* $r > 0$, *set* $D_{\beta_{\text{sad}}}(r, \varepsilon) := \frac{\varepsilon}{\sqrt{\beta_{\text{sad}}}} + r + \frac{L_H r^2}{2\sqrt{\beta_{\text{sad}}}}$. *Then, for every* $\|z\|_{L^2_{\mu^*}} \leq r$ *and every* $\tau_{\text{sad}} \in (0, 1]$,

$$F\left((\mathrm{Id} - \tau_{\text{sad}} d_z)_\# \mu_z\right) \leq F(\mu_z) - \tau \left[1 - \frac{\tau_{\text{sad}}}{2} - \frac{L_H \tau^2_{\text{sad}}}{6\sqrt{\beta_{\text{sad}}}} D_{\beta_{\text{sad}}}(r, \varepsilon)\right] \langle \nabla_\mu F(\mu_z, \cdot), d_z \rangle_{L^2_{\mu_z}}. \quad (84)$$

In particular, if $0 < \tau_{\text{sad}} \leq \frac{1}{2}$, $r \leq \frac{\sqrt{\beta_{\text{sad}}}}{L_H}$, $\varepsilon \leq \frac{\beta_{\text{sad}}}{L_H}$, *then*

$$F\left((\mathrm{Id} - \tau_{\text{sad}} d_z)_\# \mu_z\right) \leq F(\mu_z) - \frac{\tau_{\text{sad}}}{2} \langle \nabla_\mu F(\mu_z, \cdot), d_z \rangle_{L^2_{\mu_z}}, \quad (85)$$

and

$$F\left((\mathrm{Id} - \tau_{\text{sad}} d_z)_\# \mu_z\right) \leq F(\mu_z) - \frac{\tau_{\text{sad}}\sqrt{\beta_{\text{sad}}}}{2} \|d_z\|^2_{L^2_{\mu_z}}. \quad (86)$$

Proof. Denote the pulled-back gradient and Hessian by $\widehat{g}_z := \nabla_\mu F(\mu_z, \cdot) \circ Z_z$ and $\widehat{\mathrm{H}}_z$, respectively. Lemma L.15 gives

$$\widehat{d}_z := d_z \circ Z_z = \left(\widehat{\mathrm{H}}_z^2 + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2} \widehat{g}_z.$$

In particular, $\|\widehat{d}_z\|_{L^2_{\mu^*}} = \|d_z\|_{L^2_{\mu_z}}$. Along the curve $t \mapsto \mu_{tz} = (\mathrm{Id} + tz)_\# \mu^*$, we have $\widehat{g}_z = \widehat{g}_0 + \int_0^1 \widehat{\mathrm{H}}_{tz} z \mathrm{d}t$. Adding and subtracting $\widehat{\mathrm{H}}_z z$, we obtain

$$\widehat{g}_z = \widehat{g}_0 + \widehat{\mathrm{H}}_z z + \int_0^1 \left(\widehat{\mathrm{H}}_{tz} - \widehat{\mathrm{H}}_z\right) z \mathrm{d}t. \tag{87}$$

By Lemma L.13,

$$\left\| \int_0^1 \left(\widehat{\mathrm{H}}_{tz} - \widehat{\mathrm{H}}_z\right) z \mathrm{d}t \right\|_{L^2_{\mu^*}} \le \int_0^1 \|\widehat{\mathrm{H}}_{tz} - \widehat{\mathrm{H}}_z\|_{\mathrm{op}} \|z\|_{L^2_{\mu^*}} \mathrm{d}t \le L_H \int_0^1 (1-t) \|z\|^2_{L^2_{\mu^*}} \mathrm{d}t$$
$$= \frac{L_H}{2} \|z\|^2_{L^2_{\mu^*}}.$$

Let $P_z := \left(\widehat{\mathrm{H}}_z^2 + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}$, $S_z := P_z \widehat{\mathrm{H}}_z$. By spectral calculus, $\|P_z\|_{\mathrm{op}} \le \frac{1}{\sqrt{\beta_{\mathrm{sad}}}}$, $\|S_z\|_{\mathrm{op}} \le 1$. Using (87), we get

$$\widehat{d}_z = P_z \widehat{g}_0 + S_z z + P_z \int_0^1 \left(\widehat{\mathrm{H}}_{tz} - \widehat{\mathrm{H}}_z\right) z \mathrm{d}t.$$

Thus

$$\begin{aligned}
\|d_z\|_{L^2_{\mu_z}} = \|\widehat{d}_z\|_{L^2_{\mu^*}} &\le \frac{1}{\sqrt{\beta_{\mathrm{sad}}}} \|\widehat{g}_0\|_{L^2_{\mu^*}} + \|z\|_{L^2_{\mu^*}} + \frac{1}{\sqrt{\beta_{\mathrm{sad}}}} \left\| \int_0^1 \left(\widehat{\mathrm{H}}_{tz} - \widehat{\mathrm{H}}_z\right) z \mathrm{d}t \right\|_{L^2_{\mu^*}} \\
&\le \frac{\varepsilon}{\sqrt{\beta_{\mathrm{sad}}}} + \|z\|_{L^2_{\mu^*}} + \frac{L_H}{2\sqrt{\beta_{\mathrm{sad}}}} \|z\|^2_{L^2_{\mu^*}} \\
&\le D_{\beta_{\mathrm{sad}}}(r, \varepsilon).
\end{aligned}$$

We now apply the cubic Taylor estimate of Lemma L.14 at μ_z in the direction $-\tau d_z$. This gives

$$\begin{aligned}
F\left((\mathrm{Id} - \tau_{\mathrm{sad}} d_z)_\# \mu_z\right) \le F(\mu_z) &- \tau_{\mathrm{sad}} \left\langle \nabla_\mu F(\mu_z, \cdot), d_z \right\rangle_{L^2_{\mu_z}} \\
&+ \frac{\tau_{\mathrm{sad}}^2}{2} \left\langle \mathrm{H}_{\mu_z} d_z, d_z \right\rangle_{L^2_{\mu_z}} + \frac{L_H \tau_{\mathrm{sad}}^3}{6} \|d_z\|^3_{L^2_{\mu_z}}.
\end{aligned} \tag{88}$$

Since $d_z = \left(\mathrm{H}^2_{\mu_z} + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2} \nabla_\mu F(\mu_z)$, we have

$$\nabla_\mu F(\mu_z, \cdot) = \left(\mathrm{H}^2_{\mu_z} + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{1/2} d_z.$$

Therefore,

$$\left\langle \nabla_\mu F(\mu_z, \cdot), d_z \right\rangle_{L^2_{\mu_z}} = \left\langle \left(\mathrm{H}^2_{\mu_z} + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{1/2} d_z, d_z \right\rangle_{L^2_{\mu_z}}.$$

The scalar inequality $\lambda \le \sqrt{\lambda^2 + \beta_{\mathrm{sad}}}$, $\lambda \in \mathbb{R}$, implies, by spectral calculus, $\mathrm{H}_{\mu_z} \preceq \left(\mathrm{H}^2_{\mu_z} + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{1/2}$. Hence

$$\left\langle \mathrm{H}_{\mu_z} d_z, d_z \right\rangle_{L^2_{\mu_z}} \le \left\langle \nabla_\mu F(\mu_z, \cdot), d_z \right\rangle_{L^2_{\mu_z}}.$$

Moreover, $\left(\mathrm{H}^2_{\mu_z} + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{1/2} \succeq \sqrt{\beta_{\mathrm{sad}}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$, so

$$\left\langle \nabla_\mu F(\mu_z, \cdot), d_z \right\rangle_{L^2_{\mu_z}} \ge \sqrt{\beta_{\mathrm{sad}}} \|d_z\|^2_{L^2_{\mu_z}}. \tag{89}$$

It follows that

$$\|d_z\|^3_{L^2_{\mu_z}} \le \frac{\|d_z\|_{L^2_{\mu_z}}}{\sqrt{\beta_{\mathrm{sad}}}} \left\langle \nabla_\mu F(\mu_z, \cdot), d_z \right\rangle_{L^2_{\mu_z}}.$$

Substituting these estimates into (88) gives

$$\begin{aligned} F\left((\mathrm{Id}-\tau_{\mathrm{sad}} d_z)_{\#}\mu_z\right) &\le F(\mu_z) - \tau_{\mathrm{sad}}\left(1-\frac{\tau_{\mathrm{sad}}}{2}\right)\langle \nabla_\mu F(\mu_z,\cdot), d_z\rangle_{L^2_{\mu_z}} \\ &\quad + \frac{L_H \tau_{\mathrm{sad}}^3}{6\sqrt{\beta_{\mathrm{sad}}}} \|d_z\|_{L^2_{\mu_z}} \langle \nabla_\mu F(\mu_z,\cdot), d_z\rangle_{L^2_{\mu_z}} \\ &\le F(\mu_z) - \tau_{\mathrm{sad}}\left[1-\frac{\tau_{\mathrm{sad}}}{2} - \frac{L_H\tau_{\mathrm{sad}}^2}{6\sqrt{\beta_{\mathrm{sad}}}} D_{\beta_{\mathrm{sad}}}(r,\varepsilon)\right]\langle \nabla_\mu F(\mu_z,\cdot), d_z\rangle_{L^2_{\mu_z}}, \end{aligned}$$

which proves (84). By the assumption on the constants, $\frac{\varepsilon}{\sqrt{\beta_{\mathrm{sad}}}} \le \frac{\sqrt{\beta_{\mathrm{sad}}}}{L_H}$, $r \le \frac{\sqrt{\beta_{\mathrm{sad}}}}{L_H}$, and

$$\frac{L_H r^2}{2\sqrt{\beta_{\mathrm{sad}}}} \le \frac{\sqrt{\beta_{\mathrm{sad}}}}{2L_H}.$$

Therefore,

$$D_{\beta_{\mathrm{sad}}}(r,\varepsilon) \le \frac{5\sqrt{\beta_{\mathrm{sad}}}}{2L_H}.$$

Since $\tau_{\mathrm{sad}} \le \frac{1}{2}$,

$$1 - \frac{\tau_{\mathrm{sad}}}{2} - \frac{L_H\tau_{\mathrm{sad}}^2}{6\sqrt{\beta_{\mathrm{sad}}}} D_{\beta_{\mathrm{sad}}}(r,\varepsilon) \ge 1 - \frac{1}{4} - \frac{5}{12}\frac{1}{4} = \frac{31}{48} > \frac{1}{2}.$$

This proves (85). Finally, (89) gives (86). □

Lemma M.4 (Bootstrap control of a WSFN saddle episode)**.** *Let the assumptions of Lemma M.3 hold, and suppose* $0 < \tau_{\mathrm{sad}} \le \frac{1}{2}$, $\varepsilon \le \frac{\beta_{\mathrm{sad}}}{L_H}$. *Let* $\mu^0 = (Z^0)_{\#}\mu^*$, *and define recursively* $Z^{k+1} = (\mathrm{Id} - \tau_{\mathrm{sad}} d_k)\circ Z^k$, $\mu^k = (Z^k)_{\#}\mu^*$, *where*

$$d_k := \left(\mathrm{H}^2_{\mu^k} + \beta_{\mathrm{sad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2} \nabla_\mu F(\mu^k,\cdot).$$

Fix $N \in \mathbb{N}$ *and assume that*

$$F(\mu^0) - F(\mu^n) \le 2F_0, \quad 0 \le n \le N. \tag{90}$$

Let $r \le \frac{\sqrt{\beta_{\mathrm{sad}}}}{L_H}$. *If* $\|Z^0 - \mathrm{Id}\,\|_{L^2_{\mu^*}} + 2\sqrt{\tau N}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2} < r$, *then* $\|Z^n - \mathrm{Id}\,\|_{L^2_{\mu^*}} < r$, $0 \le n \le N$, *and*

$$\|Z^n - \mathrm{Id}\,\|_{L^2_{\mu^*}} \le \|Z^0 - \mathrm{Id}\,\|_{L^2_{\mu^*}} + 2\sqrt{\tau n}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2}. \tag{91}$$

Proof. Suppose, for a contradiction, that an iterate leaves the ball of radius r. Let

$$m := \min\left\{n \in \{1,\dots,N\} : \|Z^n - \mathrm{Id}\,\|_{L^2_{\mu^*}} \ge r\right\}.$$

Then

$$\|Z^k - \mathrm{Id}\,\|_{L^2_{\mu^*}} < r, \quad 0 \le k < m.$$

Therefore, Lemma M.3 applies at every iteration $k = 0,\dots,m-1$, and gives

$$F(\mu^0) - F(\mu^m) \ge \frac{\tau_{\mathrm{sad}}\sqrt{\beta_{\mathrm{sad}}}}{2} \sum_{k=0}^{m-1} \|d_k\|^2_{L^2_{\mu^k}}.$$

Furthermore, $Z^{k+1} - Z^k = -\tau_{\mathrm{sad}} d_k \circ Z^k$, and hence $\|Z^{k+1} - Z^k\|_{L^2_{\mu^*}} = \tau_{\mathrm{sad}}\|d_k\|_{L^2_{\mu^k}}$. It follows from the triangle inequality and Cauchy–Schwarz that

$$\|Z^m - Z^0\|_{L^2_{\mu^*}} \le \tau_{\mathrm{sad}} \sum_{k=0}^{m-1} \|d_k\|_{L^2_{\mu^k}} \le \tau_{\mathrm{sad}}\sqrt{m}\left(\sum_{k=0}^{m-1}\|d_k\|^2_{L^2_{\mu^k}}\right)^{1/2} \le \left(\frac{2\tau_{\mathrm{sad}} m}{\sqrt{\beta_{\mathrm{sad}}}}\left[F(\mu^0) - F(\mu^m)\right]\right)^{1/2}.$$

Using (90),

$$\|Z^m - Z^0\|_{L^2_{\mu^*}} \le 2\sqrt{\tau_{\mathrm{sad}} m}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2}.$$

Hence,

$$\begin{aligned} \|Z^m - \mathrm{Id}\,\|_{L^2_{\mu^*}} &\le \|Z^0 - \mathrm{Id}\,\|_{L^2_{\mu^*}} + \|Z^m - Z^0\|_{L^2_{\mu^*}} \\ &\le \|Z^0 - \mathrm{Id}\,\|_{L^2_{\mu^*}} + 2\sqrt{\tau_{\mathrm{sad}} N}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2} < r, \end{aligned}$$

contradicting the definition of m. Hence no such first exit time exists. The same calculation with m replaced by any $n \le N$ proves (91). □

M.2 Escape from saddle regions

With these preliminary estimates in hand, we now fix the parameter choices used in the saddle escape analysis. Let us define $\widetilde{\delta} := \frac{\delta}{\sqrt{\delta^2+\beta}} > 0$. Let $\zeta_{\mathrm{ep}} \in (0,1)$ and choose the parameters β, τ, κ, n_{out}, F_0, and η as follows. Fix $b > 0$ and $\tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$, and set

$$\beta := \beta_{\mathrm{sad}} := b\delta^2, \quad \tau := \tau_{\mathrm{sad}}. \tag{92}$$

Take δ sufficiently small that

$$\beta_{\mathrm{sad}} = b\delta^2 \le \min\left\{1, \left(\frac{C_{\mathrm{H}}\tau_{\mathrm{sad}}}{2}\right)^2\right\}. \tag{93}$$

Moreover,

$$\widetilde{\delta} := \frac{\delta}{\sqrt{\delta^2+\beta_{\mathrm{sad}}}} = \frac{1}{\sqrt{1+b}}, \quad \varrho := 1 + \tau_{\mathrm{sad}}\widetilde{\delta} > 1.$$

In particular, $\widetilde{\delta}$ is now a positive constant independent of δ. Define $c_{\mathrm{ep}} := \frac{\sqrt{2\pi}}{4}\delta\zeta_{\mathrm{ep}}$, and take $\kappa > 0$ such that

$$\kappa \ge C_{\mathrm{K}}\left(1 + 2\sqrt{\log\frac{4}{\zeta_{\mathrm{ep}}}} + 2\log\frac{4}{\zeta_{\mathrm{ep}}}\right) = \tilde{O}(1). \tag{94}$$

For sufficiently small δ, define

$$n_{\mathrm{out}} := 1 + \left\lceil \frac{2}{\log\varrho}\log\left(\frac{16\sqrt{2C_{\mathrm{H}}\tau_{\mathrm{sad}}}\kappa}{\sqrt{e}\beta_{\mathrm{sad}}^{1/4}c_{\mathrm{ep}}\log^{1/2}\varrho}\right)\right\rceil. \tag{95}$$

Note that

$$\beta_{\mathrm{sad}}^{1/4}c_{\mathrm{ep}} = b^{1/4}\delta^{1/2}\frac{\sqrt{2\pi}}{4}\delta\zeta_{\mathrm{ep}} = b^{1/4}\delta^{3/2}\frac{\sqrt{2\pi}}{4}\zeta_{\mathrm{ep}}.$$

Therefore

$$n_{\mathrm{out}} = O\left(1 + \log\frac{\kappa}{\delta^{3/2}\zeta_{\mathrm{ep}}}\right) = \tilde{O}(1),$$

where $\tilde{O}$ hides logarithmic dependence on δ^{-1}, ζ_{ep}^{-1}, and the Gaussian norm control parameter κ. Choose $c_F > 0$, independent of δ, sufficiently small as specified below, and set

$$F_0 := c_F\frac{\log^2(3/2)\beta_{\mathrm{sad}}^{3/2}}{L_{\mathrm{H}}^2(\tau_{\mathrm{sad}}n_{\mathrm{out}})^3}. \tag{96}$$

Choose c_F sufficiently small that

$$\sqrt{c_F}\log\frac{3}{2}\left(\frac{2}{\tau_{\mathrm{sad}}} + \frac{3\sqrt{2}}{\sqrt{C_{\mathrm{H}}}\tau_{\mathrm{sad}}^{3/2}}\right) \le \frac{1}{2}, \tag{97}$$

and

$$8\sqrt{c_F} + 8\sqrt{\frac{2c_F}{C_{\mathrm{H}}\tau_{\mathrm{sad}}}} + c_F\log\frac{3}{2}\left(\frac{10}{\pi C_{\mathrm{H}}\tau_{\mathrm{sad}}^2} + \frac{8\sqrt{2}}{\pi\sqrt{C_{\mathrm{H}}}\tau_{\mathrm{sad}}^{3/2}} + \frac{4}{\pi\tau_{\mathrm{sad}}}\right) \le \frac{3}{4}. \tag{98}$$

The definition of F_0 becomes

$$F_0 = c_F\frac{\log^2(3/2)b^{3/2}\delta^3}{L_{\mathrm{H}}^2(\tau_{\mathrm{sad}}n_{\mathrm{out}})^3} = \tilde{\Omega}\left(\delta^3\right).$$

Define the perturbation amplitude

$$\eta := \frac{2F_0}{\kappa\left(\varepsilon + \sqrt{\varepsilon^2 + 2C_{\mathrm{H}}F_0}\right)} = \tilde{\Theta}\left(\frac{\delta^3}{\varepsilon + \delta^{3/2}}\right). \tag{99}$$

Define $Q_\beta := \frac{R_F}{\sqrt{\beta_{\mathrm{sad}}}} + \frac{2L_{\mathrm{H}}}{\pi\beta_{\mathrm{sad}}}$. The small gradient condition is

$$\varepsilon \le \min\left\{\frac{3}{4}\sqrt{2C_{\mathrm{H}}F_0}, \frac{\beta_{\mathrm{sad}}}{L_{\mathrm{H}}}, \frac{\log(3/2)}{4\tau_{\mathrm{sad}}n_{\mathrm{out}}Q_\beta}\right\}, \tag{100}$$

which, up to logarithmic factors, becomes

$$\varepsilon = \tilde{O}\left(\frac{\delta^2}{L_{\mathrm{H}}}\right).$$

We next analyze the behavior of WSFN near an (ε, δ)-saddle. The key idea is that, after a small perturbation, the component of the iterate in a direction of negative curvature is amplified by the saddle-free dynamics, which leads to a decrease in the objective within a finite number of steps.

Proposition M.5 (Deterministic escape from a saddle region)**.** *Let Assumptions 4.4 and 4.5 hold. Let $\mu^{\mathrm{sad}} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy $\|\nabla_\mu F(\mu^{\mathrm{sad}}, \cdot)\|_{L^2_{\mu^{\mathrm{sad}}}} \le \varepsilon$, $\lambda_0 := \underline{\lambda}(\mathrm{K}_{\mu^{\mathrm{sad}}}) < -\delta$. Assume that, for some $R_F > 0$, $\|\mathrm{M}_{\mu^{\mathrm{sad}}}\|_{\mathrm{op}} \le R_F\|\nabla_\mu F(\mu^{\mathrm{sad}}, \cdot)\|_{L^2_{\mu^{\mathrm{sad}}}}$. Let $q_0 \in L^2_{\mu^{\mathrm{sad}}}$ be a unit eigenvector of $\mathrm{K}_{\mu^{\mathrm{sad}}}$ corresponding to λ_0. Fix $\zeta_{\mathrm{ep}} \in (0,1)$, and choose parameters β, τ, κ, n_{out}, F_0 and η as in (92)–(99). Suppose that the small gradient condition (100) holds. Let $\xi \in L^2_{\mu^{\mathrm{sad}}}$ satisfy $\|\xi\|_{L^2_{\mu^{\mathrm{sad}}}} \le \kappa$, and, for some $c \in \mathbb{R}$ satisfying $c_{\mathrm{ep}} \le |c| \le 2\kappa$, set $\tilde{\xi} := \xi + cq_0$. Let $\mu^{n_{\mathrm{in}}+n}$ and $\tilde{\mu}^{n_{\mathrm{in}}+n}$ be the trajectories of the WSFN scheme (1) initialized respectively from $\mu^{n_{\mathrm{in}}} := (\mathrm{Id}+\eta\xi)_{\#}\mu^{\mathrm{sad}}$, $\tilde{\mu}^{n_{\mathrm{in}}} := (\mathrm{Id}+\eta\tilde{\xi})_{\#}\mu^{\mathrm{sad}}$. Assume that δ is sufficiently small, depending on the problem constants and the logarithmic parameters. Then, there exists $n \in \{1, \dots, n_{\mathrm{out}}\}$ such that*

$$\max\left\{F(\mu^{n_{\mathrm{in}}}) - F(\mu^{n_{\mathrm{in}}+n}), F(\tilde{\mu}^{n_{\mathrm{in}}}) - F(\tilde{\mu}^{n_{\mathrm{in}}+n})\right\} \ge 2F_0.$$

Proof. We give a proof by contradiction. Assume that for any $n \in \{1, ..., n_{\mathrm{out}}\}$,

$$\max\left\{F(\mu^{n_{\mathrm{in}}}) - F(\mu^{n_{\mathrm{in}}+n}), F(\tilde{\mu}^{n_{\mathrm{in}}}) - F(\tilde{\mu}^{n_{\mathrm{in}}+n})\right\} \le 2F_0.$$

For every $n \in \{0, ..., n_{\mathrm{out}}\}$, define

$$\Psi_\tau[\mu^{n_{\mathrm{in}}+n}] = \mathrm{Id} - \tau(\mathrm{H}^2_{\mu^{n_{\mathrm{in}}+n}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\nabla_\mu F(\mu^{n_{\mathrm{in}}+n}).$$

Then the n-step maps of the SFN scheme starting from μ^* are

$$Z^{n_{\mathrm{in}}+n} := \underbrace{\Psi_\tau[\mu^{n_{\mathrm{in}}+n-1}] \circ \dots \circ \Psi_\tau[\mu^{n_{\mathrm{in}}}]}_{=:Y^{n_{\mathrm{in}}+n}} \circ (\mathrm{Id}+\eta\xi), \quad n \ge 1, \quad Z^{n_{\mathrm{in}}} := \mathrm{Id}+\eta\xi,$$
$$\tilde{Z}^{n_{\mathrm{in}}+n} = \underbrace{\Psi_\tau[\tilde{\mu}^{n_{\mathrm{in}}+n-1}] \circ \dots \circ \Psi_\tau[\tilde{\mu}^{n_{\mathrm{in}}}]}_{=:\tilde{Y}^{n_{\mathrm{in}}+n}} \circ (\mathrm{Id}+\eta\tilde{\xi}), \quad n \ge 1, \quad \tilde{Z}^{n_{\mathrm{in}}} := \mathrm{Id}+\eta\tilde{\xi}.$$

Note that $Y^{n_{\rm in}} = \tilde{Y}^{n_{\rm in}} = \mathrm{Id}$. We have

$$\begin{aligned}
W_2(\mu^{n_{\rm in}+n}, \mu^{\rm sad}) &\le \|Z^{n_{\rm in}+n} - \mathrm{Id}\,\|_{L^2_{\mu^{\rm sad}}} \\
&\le \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \|Z^{n_{\rm in}+n} - (\mathrm{Id} + \eta\xi)\|_{L^2_{\mu^{\rm sad}}} \\
&= \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \|Y^{n_{\rm in}+n} - \mathrm{Id}\,\|_{L^2_{\mu^{n_{\rm in}}}} \\
&\le \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \sum_{k=0}^{n-1} \|Y^{n_{\rm in}+k+1} - Y^{n_{\rm in}+k}\|_{L^2_{\mu^{n_{\rm in}}}} \\
&= \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \sum_{k=0}^{n-1} \|\Psi_\tau[\mu^{n_{\rm in}+k}] \circ Y^{n_{\rm in}+k} - Y^{n_{\rm in}+k}\|_{L^2_{\mu^{n_{\rm in}}}} \\
&= \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \sum_{k=0}^{n-1} \|\Psi_\tau[\mu^{n_{\rm in}+k}] - \mathrm{Id}\,\|_{L^2_{\mu^k}} \\
&= \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \tau \sum_{k=0}^{n-1} \|(\mathrm{H}^2_{\mu^{n_{\rm in}+k}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^{n_{\rm in}+k})\|_{L^2_{\mu^k}} \\
&\le \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \tau\sqrt{n} \left(\sum_{k=0}^{n-1} \|(\mathrm{H}^2_{\mu^{n_{\rm in}+k}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^{n_{\rm in}+k})\|^2_{L^2_{\mu^k}} \right)^{1/2} \\
&\le \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \tau\sqrt{n} \left(\frac{2}{\tau\sqrt{\beta}} (F(\mu^{n_{\rm in}}) - F(\mu^{n_{\rm in}+n})) \right)^{1/2} \\
&\le \eta\|\xi\|_{L^2_{\mu^{\rm sad}}} + \sqrt{2}\sqrt{\frac{\tau n}{\sqrt{\beta}}} (2F_0)^{1/2} \\
&\le \eta\kappa + 2\sqrt{\tau n_{\rm out}}\beta^{-1/4} F_0^{1/2},
\end{aligned} \tag{101}$$

where the fourth inequality follows from the Cauchy-Schwarz inequality, the fifth inequality follows from Lemma M.3 and the sixth inequality follows from the contradiction hypothesis. Similarly,

$$W_2(\tilde{\mu}^{n_{\rm in}+n}, \mu^{\rm sad}) \le \|\tilde{Z}^{n_{\rm in}+n} - \mathrm{Id}\,\|_{L^2_{\mu^{\rm sad}}} \le \eta\kappa + \eta|c| + 2\sqrt{\tau n_{\rm out}}\beta^{-1/4} F_0^{1/2}. \tag{102}$$

In the following we will show that $\mu^{n_{\rm in}+n}$ and $\tilde{\mu}^{n_{\rm in}+n}$ cannot be too far apart. To this end, since $\mu^{n_{\rm in}+n} = (Z^{n_{\rm in}+n})_\# \mu^{\rm sad}$ and $\tilde{\mu}^{n_{\rm in}+n} = (\tilde{Z}^{n_{\rm in}+n})_\# \mu^{\rm sad}$, we may define

$$w^{n_{\rm in}+n} := \tilde{Z}^{n_{\rm in}+n} - Z^{n_{\rm in}+n} \in L^2_{\mu^{\rm sad}},$$

and prove that $\|w^{n_{\rm in}+n}\|_{L^2_{\mu^{\rm sad}}}$ cannot be too large. In particular, we have

$$\begin{aligned}
&w^{n_{\rm in}+n+1} - w^{n_{\rm in}+n} \\
&= \tilde{Z}^{n_{\rm in}+n+1} - \tilde{Z}^{n_{\rm in}+n} - (Z^{n_{\rm in}+n+1} - Z^{n_{\rm in}+n}) \\
&= -\tau(\mathrm{H}^2_{\tilde{\mu}^{n_{\rm in}+n}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\tilde{\mu}^{n_{\rm in}+n}) \circ \tilde{Z}^{n_{\rm in}+n} + \tau(\mathrm{H}^2_{\mu^{n_{\rm in}+n}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}} \nabla_\mu F(\mu^{n_{\rm in}+n}) \circ Z^{n_{\rm in}+n} \\
&= -\tau(\mathrm{K}^2_{\mu^{\rm sad}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\, \mathrm{K}_{\mu^{\rm sad}} (\tilde{Z}^{n_{\rm in}+n} - Z^{n_{\rm in}+n}) - \tau R_\beta(Z^{n_{\rm in}+n}, \tilde{Z}^{n_{\rm in}+n}) \\
&= -\tau(\mathrm{K}^2_{\mu^{\rm sad}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\, \mathrm{K}_{\mu^{\rm sad}}\, w^{n_{\rm in}+n} - \tau R_\beta(Z^{n_{\rm in}+n}, \tilde{Z}^{n_{\rm in}+n}),
\end{aligned} \tag{103}$$

where the penultimate equality follows from Lemma L.17 and $R_\beta(Z^{n_{\rm in}+n}, \tilde{Z}^{n_{\rm in}+n})$ satisfies

$$\begin{aligned}
\|R_\beta(Z^{n_{\rm in}+n}, \tilde{Z}^{n_{\rm in}+n}))\|_{L^2_{\mu^{\rm sad}}} \le \Bigg[&\frac{2L_H}{\sqrt{\beta}} \left(\|Z^{n_{\rm in}+n} - \mathrm{Id}\,\|_{L^2_{\mu^{\rm sad}}} + \|\tilde{Z}^{n_{\rm in}+n} - \mathrm{Id}\,\|_{L^2_{\mu^{\rm sad}}} \right) \\
&+ \frac{L_H^2}{2\pi\beta} \left(\|Z^{n_{\rm in}+n} - \mathrm{Id}\,\|^2_{L^2_{\mu^{\rm sad}}} + \|\tilde{Z}^{n_{\rm in}+n} - \mathrm{Id}\,\|^2_{L^2_{\mu^{\rm sad}}} \right) \\
&+ \left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_H}{\pi\beta} \right) \|\nabla_\mu F(\mu^{\rm sad}, \cdot)\|_{L^2_{\mu^{\rm sad}}} \Bigg] \|\tilde{Z}^{n_{\rm in}+n} - Z^{n_{\rm in}+n}\|_{L^2_{\mu^{\rm sad}}}.
\end{aligned}$$

The assumed bound on $\mathrm{M}_{\mu^{\mathrm{sad}}}$ verifies condition (55) of Lemma L.17. That lemma applies to arbitrary reference measures in $\mathcal{P}_2(\mathbb{R}^d)$. By induction on n, the recurrence formula (103) implies

$$w^{n_{\mathrm{in}}+n} = \left(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\right)^n w^{n_{\mathrm{in}}} - \tau\sum_{k=0}^{n-1}\left(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\right)^{n-k-1} R_\beta(Z^{n_{\mathrm{in}}+k}, \tilde{Z}^{n_{\mathrm{in}}+k}).$$

We have $\|\nabla^2_\mu F(\mu^{\mathrm{sad}},\cdot,\cdot)\|_{L^2_{\mu^{\mathrm{sad}}\otimes\mu^{\mathrm{sad}}}} < \infty$, so by (Reed & Simon, 1980, Theorem VI.23), the operator $\mathrm{K}_{\mu^{\mathrm{sad}}}$ is Hilbert-Schmidt, hence it is compact by (Reed & Simon, 1980, Theorem VI.22). Moreover, it is self-adjoint since it is bounded and symmetric by Assumption 4.4. By self-adjointness, $\underline{\lambda}(\mathrm{K}_{\mu^{\mathrm{sad}}}) \in \sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})$. Since $\underline{\lambda}(\mathrm{K}_{\mu^{\mathrm{sad}}}) < -\delta < 0$, by compactness, $\lambda_0 = \underline{\lambda}(\mathrm{K}_{\mu^{\mathrm{sad}}})$ is an eigenvalue of $\mathrm{K}_{\mu^{\mathrm{sad}}}$. Thus we may choose a corresponding unit eigenvector $q_0 \in L^2_{\mu^{\mathrm{sad}}}$, i.e.,

$$\mathrm{K}_{\mu^{\mathrm{sad}}}\, q_0 = \lambda_0 q_0, \quad \lambda_0 < -\delta.$$

Since q_0 is an eigenvector of $\mathrm{K}_{\mu^{\mathrm{sad}}}$ with corresponding eigenvalue λ_0, it follows that q_0 is also an eigenvector of $\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}$ with corresponding eigenvalue $1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}$ for $\tau > 0$. Thus, using the fact that $w^{n_{\mathrm{in}}} = \tilde{Z}^{n_{\mathrm{in}}} - Z^{n_{\mathrm{in}}} = \eta c q_0$, we get

$$\left(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\right)^n w^{n_{\mathrm{in}}} = \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^n w^{n_{\mathrm{in}}}.$$

Therefore, using $\|w^{n_{\mathrm{in}}}\|_{L^2_{\mu^{\mathrm{sad}}}} = \eta|c|$, we have

$$\left\|\left(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\right)^n w^{n_{\mathrm{in}}}\right\|_{L^2_{\mu^{\mathrm{sad}}}} = \eta|c|\left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^n.$$

Now we upper bound $\|\left(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\right)\|_{\mathrm{op}}$ using spectral functional calculus. The key observation is that the function $f(x) = \frac{x}{\sqrt{\beta+x^2}}$ is non-decreasing on $\mathbb{R}$.

Since $\mathrm{K}_{\mu^{\mathrm{sad}}}$ is self-adjoint, we have $\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}}) \subset \mathbb{R}$. Let Ω_{K} be the Borel σ-algebra of subsets of $\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})$. By (Conway, 1994, 1.1. Definition, Chapter IX), a spectral measure for $(\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}}), \Omega_{\mathrm{K}}, L^2_{\mu^{\mathrm{sad}}})$ is defined as a function $E : \Omega_{\mathrm{K}} \to \mathcal{B}(L^2_{\mu^{\mathrm{sad}}})$ such that

- for each $\Delta \in \Omega_{\mathrm{K}}$, $E(\Delta)$ is a projection,
- $E(\emptyset) = 0$ and $E(\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})) = 1$,
- $E(\Delta_1 \cap \Delta_2) = E(\Delta_1)E(\Delta_2)$ for $\Delta_1, \Delta_2 \in \Omega_{\mathrm{K}}$,
- if $\{\Delta_n\}_{n\geq 1}$ are pairwise disjoint sets from Ω_{K}, then

$$E\left(\cup_{n=1}^\infty \Delta_n\right) = \sum_{i=1}^\infty E(\Delta_n).$$

By (Conway, 1994, 1.9. Lemma, Chapter IX), if E is a spectral measure for $(\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}}), \Omega_{\mathrm{K}}, L^2_{\mu^{\mathrm{sad}}})$ and $\phi, \psi \in L^2_{\mu^{\mathrm{sad}}}$, then

$$E_{\phi,\psi}(\Delta) := \langle E(\Delta)\phi, \psi\rangle_{L^2_{\mu^{\mathrm{sad}}}}$$

defines a countably additive measure on Ω_{K} with $\|E\|_{\mathrm{TV}} = |E|(\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})) \leq \|\phi\|_{L^2_{\mu^{\mathrm{sad}}}}\|\psi\|_{L^2_{\mu^{\mathrm{sad}}}}$.

Since $\mathrm{K}_{\mu^{\mathrm{sad}}}$ is self-adjoint on $L^2_{\mu^{\mathrm{sad}}}$, it is a normal operator. Hence, by (Conway, 1994, 2.2. The Spectral Theorem, Chapter IX), there is a unique spectral measure E on the Borel subsets of $\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})$ such that

$$\mathrm{K}_{\mu^{\mathrm{sad}}} = \int_{\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})} \lambda \mathrm{d}E(\lambda).$$

Let $f : \sigma(\mathrm{K}_{\mu^{\mathrm{sad}}}) \to \mathbb{R}$ be bounded Ω_{K}-measurable. Then, by (Conway, 1994, 1.10. Proposition, Chapter IX), there exists a unique operator $f(\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})) \in \mathcal{B}(L^2_{\mu^{\mathrm{sad}}})$ such that for all $\phi, \psi \in L^2_{\mu^{\mathrm{sad}}}$,

$$\langle \phi, f(\mathrm{K}_{\mu^{\mathrm{sad}}})\psi \rangle_{L^2_{\mu^{\mathrm{sad}}}} = \int_{\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})} f(\lambda) \mathrm{d}\langle \phi, E(\lambda)\psi \rangle_{L^2_{\mu^{\mathrm{sad}}}}.$$

Choose $\phi = \psi \in L^2_{\mu^{\mathrm{sad}}}$ and $f_\tau(\lambda) = 1 - \tau(\lambda^2 + \beta)^{-\frac{1}{2}}\lambda$, for $\tau > 0$, which is bounded on $\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})$ since $\mathrm{K}_{\mu^{\mathrm{sad}}}$ is bounded. Moreover, f_τ is non-increasing on $\mathbb{R}$. Hence, for any $\lambda \in \sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})$, since $\lambda \geq \lambda_0$, we get

$$1 - \tau(\lambda^2 + \beta)^{-\frac{1}{2}}\lambda = f_\tau(\lambda) \leq f_\tau(\lambda_0) = 1 - \tau(\lambda_0^2 + \beta)^{-\frac{1}{2}}\lambda_0.$$

From $\tau \leq \frac{1}{2} < 1$, it follows that $f_\tau(\lambda) > 0$, which implies

$$\left(1 - \tau(\lambda^2 + \beta)^{-\frac{1}{2}}\lambda\right)^2 \leq \left(1 - \tau(\lambda_0^2 + \beta)^{-\frac{1}{2}}\lambda_0\right)^2.$$

Hence, by (Conway, 1994, 4.7. Theorem, Chapter X),

$$\begin{aligned}
&\|(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}})\psi\|^2_{L^2_{\mu^{\mathrm{sad}}}} = \int_{\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})} \left(1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}\right)^2 \mathrm{d}\langle \psi, E(\lambda)\psi\rangle_{L^2_{\mu^{\mathrm{sad}}}} \\
&\leq \int_{\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}})} \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^2 \mathrm{d}\langle \psi, E(\lambda)\psi\rangle_{L^2_{\mu^{\mathrm{sad}}}} \\
&\leq \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^2 E(\sigma(\mathrm{K}_{\mu^{\mathrm{sad}}}))\|\psi\|^2_{L^2_{\mu^{\mathrm{sad}}}} \\
&= \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^2 \|\psi\|^2_{L^2_{\mu^{\mathrm{sad}}}}.
\end{aligned}$$

Therefore,

$$\|\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\,\|_{\mathrm{op}} \leq 1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}.$$

Putting everything together gives

$$\begin{aligned}
&\left| \|w^{n_{\mathrm{in}}+n}\|_{L^2_{\mu^{\mathrm{sad}}}} - \eta|c| \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^n \right| \\
&\left\| w^{n_{\mathrm{in}}+n} - \left(\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\right)^n w^{n_{\mathrm{in}}} \right\|_{L^2_{\mu^{\mathrm{sad}}}} \\
&\leq \tau \sum_{k=0}^{n-1} \|\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau(\mathrm{K}^2_{\mu^{\mathrm{sad}}} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\,\mathrm{K}_{\mu^{\mathrm{sad}}}\,\|^{n-k-1}_{\mathrm{op}} \|R_\beta(Z^{n_{\mathrm{in}}+k}, \tilde{Z}^{n_{\mathrm{in}}+k})\|_{L^2_{\mu^{\mathrm{sad}}}} \qquad (104) \\
&\leq \tau \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^{n-k-1} \sum_{k=0}^{n-1} \|R_\beta(Z^{n_{\mathrm{in}}+k}, \tilde{Z}^{n_{\mathrm{in}}+k})\|_{L^2_{\mu^{\mathrm{sad}}}} \\
&\leq \frac{\tau\Delta}{1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}} \sum_{k=0}^{n-1} \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^{n-k} \|w^{n_{\mathrm{in}}+k}\|_{L^2_{\mu^{\mathrm{sad}}}},
\end{aligned}$$

where the last inequality follows from the fact that

$$\begin{aligned}
&\|R(Z^{n_{\mathrm{in}}+k}, \tilde{Z}^{n_{\mathrm{in}}+k})\|_{L^2_{\mu^{\mathrm{sad}}}} \\
&\leq \underbrace{\left(\frac{4L_{\mathrm{H}}}{\sqrt{\beta}}(2A+B) + \frac{L^2_{\mathrm{H}}}{\pi\beta}\left(5A^2 + 4AB + B^2\right) + \varepsilon\left(\frac{R_F}{\sqrt{\beta}} + \frac{2L_{\mathrm{H}}}{\pi\beta}\right)\right)}_{:=\Delta>0} \|w^{n_{\mathrm{in}}+k}\|_{L^2_{\mu^{\mathrm{sad}}}},
\end{aligned}$$

due to (101), (102), the assumption $|c| \le 2\kappa$, and $A := \eta\kappa$, $B := 2\sqrt{\tau n_{\mathrm{out}}}\beta^{-1/4}F_0^{1/2}$.

For any $n \in \{0, ..., n_{\mathrm{out}}\}$, set

$$a_{n_{\mathrm{in}}+n} := \left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^{-n} \|w^{n_{\mathrm{in}}+n}\|_{L^2_{\mu^{\mathrm{sad}}}},\ b = \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}},\ d = \eta|c|.$$

Observe that $a_{n_{\mathrm{in}}} = \|w^{n_{\mathrm{in}}}\|_{L^2_{\mu^{\mathrm{sad}}}} = \eta|c| = d$. Thus, since $b > 0$, it can be proved by induction on n that the inequality

$$a_{n_{\mathrm{in}}+n} \le d + b\sum_{k=0}^{n-1} a_{n_{\mathrm{in}}+k},$$

obtained from (104), implies

$$a_{n_{\mathrm{in}}+n} \le d(1+b)^n,$$

for all $n \in \{0, ..., n_{\mathrm{out}}\}$. Therefore,

$$\left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^{-n} \|w^{n_{\mathrm{in}}+n}\|_{L^2_{\mu^{\mathrm{sad}}}} \le \eta|c|\left(1 + \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\right)^n. \tag{105}$$

From this, using (104) and (105) gives

$$\begin{aligned}
\left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^{-n} \|w^{n_{\mathrm{in}}+n}\|_{L^2_{\mu^{\mathrm{sad}}}} &\ge \eta|c| - \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\sum_{k=0}^{n-1}\left(1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}\right)^{-k} \|w^{n_{\mathrm{in}}+k}\|_{L^2_{\mu^{\mathrm{sad}}}} \\
&\ge \eta|c| - \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\eta|c|\sum_{k=0}^{n-1}\left(1 + \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\right)^k \\
&= \eta|c| - \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\eta|c|\frac{\left(1 + \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\right)^n - 1}{\frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}} \\
&= \eta|c|\left(2 - \left(1 + \frac{\tau\Delta}{1-\tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}}}\right)^n\right) \\
&\ge \eta|c|\left(2 - (1+\tau\Delta)^n\right) \\
&\ge \eta|c|\left(2 - \exp\left(n_{\mathrm{out}}\tau\Delta\right)\right),
\end{aligned}$$

where the penultimate inequality follows from the fact that $1 - \tau\frac{\lambda_0}{\sqrt{\lambda_0^2+\beta}} \ge 1$ since $\lambda_0 < -\delta < 0$, and the last inequality follows from the standard inequality $(1+x)^n \le e^{nx}$.

The aim is now to prove that $\tau_{\mathrm{sad}} n_{\mathrm{out}}\Delta \le \log\frac{3}{2}$. We use the coupled scaling $\beta_{\mathrm{sad}} = b\delta^2$ for $b > 0$ and keep all dependence on β_{sad} explicit. Under the contradiction hypothesis, for every $n \in \{0, \dots, n_{\mathrm{out}}\}$,

$$\left\|Z^{n_{\mathrm{in}}+n} - \mathrm{Id}\right\|_{L^2_{\mu^{\mathrm{sad}}}} \le A + B. \tag{106}$$

Similarly, since $|c| \le 2\kappa$,

$$\left\|\widetilde{Z}^{n_{\mathrm{in}}+n} - \mathrm{Id}\right\|_{L^2_{\mu^{\mathrm{sad}}}} \le \eta\kappa + \eta|c| + B \le 3A + B. \tag{107}$$

We first verify that the local bootstrap in Lemma M.4 closes. From the definition of η,

$$A = \eta\kappa = \frac{2F_0}{\varepsilon + \sqrt{\varepsilon^2 + 2C_{\mathrm{H}}F_0}} = \frac{\sqrt{\varepsilon^2 + 2C_{\mathrm{H}}F_0} - \varepsilon}{C_{\mathrm{H}}} \le \sqrt{\frac{2F_0}{C_{\mathrm{H}}}}. \tag{108}$$

Using the definition of F_0, we obtain

$$A \le \sqrt{\frac{2c_F}{C_{\mathrm{H}}}} \frac{\log(3/2)\beta_{\mathrm{sad}}^{3/4}}{L_{\mathrm{H}}(\tau_{\mathrm{sad}} n_{\mathrm{out}})^{3/2}}. \tag{109}$$

Moreover,

$$B = 2\sqrt{c_F} \frac{\log(3/2)\sqrt{\beta_{\mathrm{sad}}}}{L_{\mathrm{H}} \tau_{\mathrm{sad}} n_{\mathrm{out}}}. \tag{110}$$

Hence,

$$3A + B \le \frac{\sqrt{\beta_{\mathrm{sad}}}}{L_{\mathrm{H}}} \sqrt{c_F} \log \frac{3}{2} \left(\frac{2}{\tau_{\mathrm{sad}} n_{\mathrm{out}}} + \frac{3\sqrt{2}\beta_{\mathrm{sad}}^{1/4}}{\sqrt{C_{\mathrm{H}}}(\tau_{\mathrm{sad}} n_{\mathrm{out}})^{3/2}} \right) \le \frac{\sqrt{\beta_{\mathrm{sad}}}}{2L_{\mathrm{H}}},$$

where the final inequality follows from $\tau_{\mathrm{sad}} n_{\mathrm{out}} \ge \tau_{\mathrm{sad}}$, $\beta_{\mathrm{sad}} \le 1$, and (97). Therefore both trajectories remain inside the ball of radius $\sqrt{\beta_{\mathrm{sad}}}/L_{\mathrm{H}}$ throughout the episode. Lemma M.3 hence applies up to time n_{out} by the first exit argument.

Applying Lemma L.17 at the reference measure μ^{sad}, and using $\left\|\nabla_\mu F(\mu^{\mathrm{sad}}, \cdot)\right\|_{L^2_{\mu^{\mathrm{sad}}}} \le \varepsilon$, (106) and (107) imply

$$\left\| R_\beta \left(Z^{n_{\mathrm{in}}+n}, \widetilde{Z}^{n_{\mathrm{in}}+n} \right) \right\|_{L^2_{\mu^{\mathrm{sad}}}} \le \Delta \left\| w^{n_{\mathrm{in}}+n} \right\|_{L^2_{\mu^{\mathrm{sad}}}},$$

where

$$\Delta = \frac{8L_{\mathrm{H}}A}{\sqrt{\beta_{\mathrm{sad}}}} + \frac{4L_{\mathrm{H}}B}{\sqrt{\beta_{\mathrm{sad}}}} + \frac{L_{\mathrm{H}}^2}{2\pi\beta_{\mathrm{sad}}} \left(10A^2 + 8AB + 2B^2\right) + \varepsilon \left(\frac{R_F}{\sqrt{\beta_{\mathrm{sad}}}} + \frac{2L_{\mathrm{H}}}{\pi\beta_{\mathrm{sad}}} \right).$$

We now estimate $\tau_{\mathrm{sad}} n_{\mathrm{out}} \Delta$. By (109) and (110),

$$\begin{aligned}
\tau_{\mathrm{sad}} n_{\mathrm{out}} \Delta \le\ & 8\sqrt{c_F} \log \frac{3}{2} + 8\sqrt{\frac{2c_F}{C_{\mathrm{H}}}} \frac{\log(3/2)\beta_{\mathrm{sad}}^{1/4}}{(\tau_{\mathrm{sad}} n_{\mathrm{out}})^{1/2}} + \frac{10c_F \log^2(3/2)}{\pi C_{\mathrm{H}}} \frac{\sqrt{\beta_{\mathrm{sad}}}}{(\tau_{\mathrm{sad}} n_{\mathrm{out}})^2} \\
& + \frac{8\sqrt{2}c_F \log^2(3/2)}{\pi\sqrt{C_{\mathrm{H}}}} \frac{\beta_{\mathrm{sad}}^{1/4}}{(\tau_{\mathrm{sad}} n_{\mathrm{out}})^{3/2}} + \frac{4c_F \log^2(3/2)}{\pi \tau_{\mathrm{sad}} n_{\mathrm{out}}} + \tau_{\mathrm{sad}} n_{\mathrm{out}} \varepsilon \left(\frac{R_F}{\sqrt{\beta_{\mathrm{sad}}}} + \frac{2L_{\mathrm{H}}}{\pi\beta_{\mathrm{sad}}} \right).
\end{aligned}$$

Since $\beta_{\mathrm{sad}} \le 1$ and $\tau_{\mathrm{sad}} n_{\mathrm{out}} \ge \tau_{\mathrm{sad}}$, the first five terms on the right-hand side are bounded above by

$$\log \frac{3}{2} \left[8\sqrt{c_F} + 8\sqrt{\frac{2c_F}{C_{\mathrm{H}}\tau_{\mathrm{sad}}}} + c_F \log\left(\frac{3}{2}\right) \left(\frac{10}{\pi C_{\mathrm{H}} \tau_{\mathrm{sad}}^2} + \frac{8\sqrt{2}}{\pi\sqrt{C_{\mathrm{H}}}\, \tau_{\mathrm{sad}}^{3/2}} + \frac{4}{\pi\tau_{\mathrm{sad}}} \right) \right] \le \frac{3}{4} \log \frac{3}{2},$$

where we used (98). Moreover, (100) gives

$$\tau_{\mathrm{sad}} n_{\mathrm{out}} \varepsilon \left(\frac{R_F}{\sqrt{\beta_{\mathrm{sad}}}} + \frac{2L_{\mathrm{H}}}{\pi\beta_{\mathrm{sad}}} \right) \le \frac{1}{4} \log \frac{3}{2}.$$

Thus,

$$\tau_{\mathrm{sad}} n_{\mathrm{out}} \Delta \le \log \frac{3}{2}.$$

Let $\rho_0 := 1 - \tau_{\mathrm{sad}} \frac{\lambda_0}{\sqrt{\lambda_0^2 + \beta_{\mathrm{sad}}}}$. Since $\lambda_0 < 0$, we have $\rho_0 > 1$. The discrete perturbation estimate derived from the recurrence for $w^{n_{\mathrm{in}}+n}$ therefore gives, for every $n \le n_{\mathrm{out}}$,

$$\rho_0^{-n} \left\| w^{n_{\mathrm{in}}+n} \right\|_{L^2_{\mu^{\mathrm{sad}}}} \ge \eta|c| \left(2 - \exp(\tau_{\mathrm{sad}} n_{\mathrm{out}} \Delta)\right) \ge \eta|c| \left(2 - \exp\left(\log \frac{3}{2}\right)\right) = \frac{\eta|c|}{2}.$$

Since $\lambda_0 \le -\delta$,

$$\rho_0 = 1 + \tau_{\mathrm{sad}} \frac{|\lambda_0|}{\sqrt{\lambda_0^2 + \beta_{\mathrm{sad}}}} \ge 1 + \tau_{\mathrm{sad}} \frac{\delta}{\sqrt{\delta^2 + \beta_{\mathrm{sad}}}} = \varrho.$$

Therefore,

$$\left\| w^{n_{\mathrm{in}}+n} \right\|_{L^2_{\mu^{\mathrm{sad}}}} \ge \frac{\eta|c|}{2} \varrho^n. \tag{111}$$

On the other hand, the triangle inequality in $L^2_{\mu^{\mathrm{sad}}}$ and (106), (107) give

$$\left\|w^{n_{\mathrm{in}}+n}\right\|_{L^2_{\mu^{\mathrm{sad}}}} \le 4A + 2B = 4\eta\kappa + 4\sqrt{\tau_{\mathrm{sad}} n_{\mathrm{out}}}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2}.$$

Furthermore, by (108), $A \le \sqrt{\frac{2F_0}{C_{\mathrm{H}}}}$. The smallness condition (93), together with $\tau_{\mathrm{sad}} n_{\mathrm{out}} \ge \tau_{\mathrm{sad}}$, implies

$$A \le \sqrt{\tau_{\mathrm{sad}} n_{\mathrm{out}}}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2}.$$

Hence

$$\left\|w^{n_{\mathrm{in}}+n}\right\|_{L^2_{\mu^{\mathrm{sad}}}} \le 8\sqrt{\tau_{\mathrm{sad}} n_{\mathrm{out}}}\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2}. \tag{112}$$

We now evaluate the lower and upper bounds at $n = n_{\mathrm{out}}$. From (111),

$$\left\|w^{n_{\mathrm{in}}+n_{\mathrm{out}}}\right\|_{L^2_{\mu^{\mathrm{sad}}}} \ge \frac{\eta|c|}{2}\varrho^{n_{\mathrm{out}}}.$$

For every $x > 0$ and $m \ge 1$,

$$(1+x)^m \ge \sqrt{em\log(1+x)}(1+x)^{m/2}.$$

Applying this inequality with $x = \tau_{\mathrm{sad}}\widetilde{\delta}$ and $m = n_{\mathrm{out}}$ gives

$$\left\|w^{n_{\mathrm{in}}+n_{\mathrm{out}}}\right\|_{L^2_{\mu^{\mathrm{sad}}}} \ge \frac{\eta|c|}{2}\sqrt{en_{\mathrm{out}}\log\varrho}\,\varrho^{n_{\mathrm{out}}/2}. \tag{113}$$

Combining (113) and (112) gives

$$\varrho^{n_{\mathrm{out}}/2} \le \frac{16\sqrt{\tau_{\mathrm{sad}}}F_0^{1/2}}{\eta|c|\sqrt{e}\beta_{\mathrm{sad}}^{1/4}\log^{1/2}\varrho}. \tag{114}$$

It remains to eliminate $F_0^{1/2}/\eta$. By (100), $\varepsilon \le \frac{3}{4}\sqrt{2C_{\mathrm{H}}F_0}$. Therefore,

$$\eta\kappa = \frac{2F_0}{\varepsilon + \sqrt{\varepsilon^2 + 2C_{\mathrm{H}}F_0}} \ge \frac{1}{2}\sqrt{\frac{2F_0}{C_{\mathrm{H}}}}.$$

Equivalently,

$$\frac{F_0^{1/2}}{\eta} \le \kappa\sqrt{2C_{\mathrm{H}}}.$$

Substituting this estimate into (114), and using $|c| \ge c_{\mathrm{ep}}$, gives

$$\varrho^{n_{\mathrm{out}}/2} \le \frac{16\sqrt{2C_{\mathrm{H}}\tau_{\mathrm{sad}}}\kappa}{\sqrt{e}\beta_{\mathrm{sad}}^{1/4}c_{\mathrm{ep}}\log^{1/2}\varrho}.$$

Taking logarithms, we obtain

$$n_{\mathrm{out}} \le \frac{2}{\log\varrho}\log\left(\frac{16\sqrt{2C_{\mathrm{H}}\tau_{\mathrm{sad}}}\kappa}{\sqrt{e}\beta_{\mathrm{sad}}^{1/4}c_{\mathrm{ep}}\log^{1/2}\varrho}\right).$$

This contradicts the definition (95) of n_{out}. The contradiction proves that at least one of the two trajectories decreases the objective by at least $2F_0$ within n_{out} iterations. □

The previous proposition gives a deterministic mechanism for escaping a saddle once the perturbation has a sufficiently large component in the unstable eigendirection. We now formalize this escape phase as a saddle-point episode and quantify its success probability.

Informally, a saddle-point episode begins when the iterate is near an (ε, δ)-saddle and a perturbation is injected. The episode is said to be successful if, after running the WSFN iteration for the prescribed number of steps, the objective has decreased by at least a fixed amount.

Definition M.6 (Perturbation episode with early acceptance)**.** *Let $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy $\|\nabla_\mu F(\bar{\mu}, \cdot)\|_{L^2_{\bar{\mu}}} \le \varepsilon$. Draw $\xi \sim \mathcal{N}_{L^2_{\bar{\mu}}}(0, \mathrm{K}^2_{\bar{\mu}})$, $\nu^0 := (\mathrm{Id} + \eta\xi)_{\#}\bar{\mu}$. Define the auxiliary trajectory*

$$\nu^{\ell+1} := \left(\mathrm{Id} - \tau_{\mathrm{sad}} \left(\mathrm{H}^2_{\nu^\ell} + \beta_{\mathrm{sad}} \, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2} \nabla_\mu F(\nu^\ell, \cdot)\right)_{\#} \nu^\ell, \quad 0 \le \ell < n_{\mathrm{out}}.$$

Define the episode stopping index by

$$\ell_{\mathrm{ep}} := \min\left(\left\{\ell \in \{0, \dots, n_{\mathrm{out}}\} : F(\bar{\mu}) - F(\nu^\ell) \ge F_0\right\} \cup \{n_{\mathrm{out}}\}\right), \quad \mu_{\mathrm{out}} := \nu^{\ell_{\mathrm{ep}}}.$$

The algorithm computes only the prefix $\nu^0, \dots, \nu^{\ell_{\mathrm{ep}}}$. The episode is successful if $F(\bar{\mu}) - F(\mu_{\mathrm{out}}) \ge F_0$. Upon success, the algorithm accepts μ_{out} and returns to the gradient norm test. Otherwise, it terminates and returns the stored measure $\bar{\mu}$. When $\bar{\mu}$ is an (ε, δ)-saddle point, we call this a saddle point episode. In the notation of Proposition M.5, $\bar{\mu} = \mu^{\mathrm{sad}}$ and $\nu^\ell = \mu^{n_{\mathrm{in}}+\ell}$.

Proposition M.7 (Escape from saddle point episode)**.** *Let Assumptions 4.4 and 4.5 hold. Let $\mu^{\mathrm{sad}} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy $\|\nabla_\mu F(\mu^{\mathrm{sad}}, \cdot)\|_{L^2_{\mu^{\mathrm{sad}}}} \le \varepsilon$, $\lambda_0 := \underline{\lambda}(\mathrm{K}_{\mu^{\mathrm{sad}}}) < -\delta$. Assume that, for some $R_F > 0$, $\| \mathrm{M}_{\mu^{\mathrm{sad}}} \|_{\mathrm{op}} \le R_F \|\nabla_\mu F(\mu^{\mathrm{sad}}, \cdot)\|_{L^2_{\mu^{\mathrm{sad}}}}$. Let $q_0 \in L^2_{\mu^{\mathrm{sad}}}$ be a unit eigenvector of $\mathrm{K}_{\mu^{\mathrm{sad}}}$ corresponding to λ_0. Fix $\zeta_{\mathrm{ep}} \in (0, 1)$, and choose parameters β, τ, κ, n_{out}, F_0 and η as in (92)–(99). Suppose that the small gradient condition 100 holds. Let $\xi \sim \mathcal{N}_{L^2_{\mu^{\mathrm{sad}}}}(0, \mathrm{K}^2_{\mu^{\mathrm{sad}}})$, and define the WSFN trajectory $\mu^{n_{\mathrm{in}}+n}$, $0 \le n \le n_{\mathrm{out}}$, initialized at $\mu^{n_{\mathrm{in}}} := (\mathrm{Id} + \eta\xi)_{\#}\mu^{\mathrm{sad}}$. Let ℓ_{ep} be the stopping index from Definition M.6, with $\bar{\mu} = \mu^{\mathrm{sad}}$ and $\nu^\ell = \mu^{n_{\mathrm{in}}+\ell}$, and set $\mu_{\mathrm{out}} := \mu^{n_{\mathrm{in}}+\ell_{\mathrm{ep}}}$. Then*

$$\mathbb{P}\left(F(\mu^{\mathrm{sad}}) - F(\mu_{\mathrm{out}}) \ge F_0\right) \ge 1 - \frac{3\zeta_{\mathrm{ep}}}{4}.$$

Proof. By Lemma D.4, for every $t > 0$,

$$\mathbb{P}\left(\|\xi\|^2_{L^2_\mu} > \mathrm{Tr}(\mathrm{K}^2_\mu) + 2\sqrt{\mathrm{Tr}(\mathrm{K}^4_\mu)t} + 2\|\,\mathrm{K}_\mu\,\|^2_{\mathrm{op}} t\right) \le e^{-t}.$$

Since $\mathrm{Tr}(\mathrm{K}^4_\mu) \le \|\nabla^2_\mu F(\mu, \cdot, \cdot)\|^4_{L^2_{\mu\otimes\mu}} \le C^4_{\mathrm{K}}$, $\|\,\mathrm{K}_\mu\,\|^2_{\mathrm{op}} \le \|\nabla^2_\mu F(\mu, \cdot, \cdot)\|^2_{L^2_{\mu\otimes\mu}} \le C^2_{\mathrm{K}}$, and setting κ as in (94), i.e.,

$$\kappa \ge C_{\mathrm{K}}\left(1 + 2\sqrt{\log\frac{4}{\zeta_{\mathrm{ep}}}} + 2\log\frac{4}{\zeta_{\mathrm{ep}}}\right)$$

it follows from the choice $t = \log(4/\zeta_{\mathrm{ep}})$ that

$$\mathbb{P}\left(\|\xi\|_{L^2_{\mu^{\mathrm{sad}}}} > \kappa\right) \le \frac{\zeta_{\mathrm{ep}}}{4}.$$

Thus $\|\xi\|_{L^2_{\mu^{\mathrm{sad}}}} \le \kappa$ with probability at least $1 - \frac{\zeta_{\mathrm{ep}}}{4}$.

We apply Proposition M.5 to the trajectory continued up to n_{out} steps. Define

$$\mathcal{E} := \left\{\exists n \in \{1, \dots, n_{\mathrm{out}}\} : F(\mu^{n_{\mathrm{in}}}) - F(\mu^{n_{\mathrm{in}}+n}) \ge 2F_0\right\}.$$

We first prove that $\mathbb{P}(\mathcal{E}) \ge 1 - \frac{\zeta_{\mathrm{ep}}}{2}$. Let $\mathcal{A} = \mathcal{E}^c$ be the event that the scheme does not escape the saddle point, meaning $F(\mu^{n_{\mathrm{in}}}) - F(\mu^{n_{\mathrm{in}}+n}) < 2F_0$ for all $n \in \{1, ..., n_{\mathrm{out}}\}$. Define the bounded norm event by

$$\mathcal{B} := \left\{\|\xi\|_{L^2_{\mu^{\mathrm{sad}}}} \le \kappa\right\}.$$

We have $\mathbb{P}(\mathcal{B}) \ge 1 - \frac{\zeta_{\mathrm{ep}}}{4}$. Using the union bound, we get

$$\mathbb{P}(\mathcal{E}) \ge \mathbb{P}(\mathcal{E} \cap \mathcal{B}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

To lower bound $\mathbb{P}(\mathcal{E})$, we must upper bound $\mathbb{P}(\mathcal{A} \cap \mathcal{B})$. We decompose the perturbation ξ into a component along the unit eigenvector q_0 and an orthogonal component $\xi_\perp$ so that

$$\xi = \langle q_0, \xi\rangle_{L^2_{\mu^{\mathrm{sad}}}} q_0 + \xi_\perp,$$

and $\langle q_0, \xi_\perp \rangle_{L^2_{\mu^{\text{sad}}}} = 0$.

For each fixed orthogonal component $\xi_\perp$, we define the "stuck region" $I_{\xi_\perp}$ as the set scalars for which the perturbation ξ remains bounded and the scheme fails to escape, that is

$$I_{\xi_\perp} := \{x \in \mathbb{R} | x q_0 + \xi_\perp \in \mathcal{A} \cap \mathcal{B}\}.$$

Take $x_1, x_2 \in I_{\xi_\perp}$. Let $\xi_1 = x_1 q_0 + \xi_\perp$ and $\xi_2 = x_2 q_0 + \xi_\perp$. Let $c = x_2 - x_1$, so $\xi_2 = \xi_1 + c q_0$. By definition, both ξ_1 and ξ_2 are in $\mathcal{B}$, meaning $\|\xi_1\|_{L^2_{\mu^{\text{sad}}}} \leq \kappa$ and $\|\xi_2\|_{L^2_{\mu^{\text{sad}}}} \leq \kappa$. The triangle inequality gives

$$|c| = |x_2 - x_1| \leq 2\kappa.$$

According to Proposition M.5, if $\|\xi_1\|_{L^2_{\mu^{\text{sad}}}} \leq \kappa$ and $c_{\text{ep}} \leq |c| \leq 2\kappa$, then at least one of the initializations ($\mu^{n_{\text{in}}}$ or $\tilde{\mu}^{n_{\text{in}}}$) must escape. However, since we assumed both x_1 and x_2 are in the stuck region $\mathcal{A}$, this escape condition cannot hold. Because we already know $|c| \leq 2\kappa$, it must be the case that

$$|x_2 - x_1| < c_{\text{ep}}.$$

This implies that the maximum distance between any two points in $I_{\xi_\perp}$ is bounded. Thus, $I_{\xi_\perp}$ is contained within an interval of length c_{ep}. By Lemma D.6, $\langle q_0, \xi \rangle_{L^2_{\mu^{\text{sad}}}} \sim \mathcal{N}(0, \lambda_0^2)$, with density

$$p(x) = \frac{1}{\sqrt{2\pi}|\lambda_0|} \exp\left(-\frac{x^2}{2\lambda_0^2}\right),$$

and $\langle q_0, \xi \rangle_{L^2_{\mu^{\text{sad}}}}$ is independent from $\xi_\perp$. The density is bounded by its value at $x = 0$, so $p(x) \leq \frac{1}{\sqrt{2\pi}|\lambda_0|}$. Due to independence, we can now bound the conditional probability of $\langle q_0, \xi \rangle_{L^2_{\mu^{\text{sad}}}}$ falling into the stuck region by

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B} | \xi_\perp) = \mathbb{P}(\langle q_0, \xi \rangle_{L^2_{\mu^{\text{sad}}}} \in I_{\xi_\perp} | \xi_\perp) = \int_{I_{\xi_\perp}} p(x) \mathrm{d}x \leq \frac{1}{\sqrt{2\pi}|\lambda_0|} c_{\text{ep}} = \frac{\delta}{|\lambda_0|} \frac{\zeta_{\text{ep}}}{4} < \frac{\zeta_{\text{ep}}}{4},$$

since $|\lambda_0| > \delta$. Taking the expectation over the orthogonal component $\xi_\perp$, the unconditional probability of being in $\mathcal{A} \cap \mathcal{B}$ is

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{E}_{\xi_\perp}\left[\mathbb{P}(\langle q_0, \xi \rangle_{L^2_{\mu^{\text{sad}}}} \in I_{\xi_\perp} | \xi_\perp)\right] \leq \frac{\zeta_{\text{ep}}}{4}.$$

We substitute this back into our initial probability bound

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}) \geq \left(1 - \frac{\zeta_{\text{ep}}}{4}\right) - \frac{\zeta_{\text{ep}}}{4} = 1 - \frac{\zeta_{\text{ep}}}{2}.$$

Define the perturbation cost event $\mathcal{D} := \left\{F(\mu^{n_{\text{in}}}) - F(\mu^{\text{sad}}) \leq F_0\right\}$. By Lemma M.1, $\mathbb{P}(\mathcal{D}) \geq 1 - \frac{\zeta_{\text{ep}}}{4}$. On $\mathcal{E} \cap \mathcal{D}$, there exists $n \in \{1, \ldots, n_{\text{out}}\}$ such that

$$F(\mu^{\text{sad}}) - F(\mu^{n_{\text{in}}+n}) = F(\mu^{\text{sad}}) - F(\mu^{n_{\text{in}}}) + F(\mu^{n_{\text{in}}}) - F(\mu^{n_{\text{in}}+n}) \geq -F_0 + 2F_0 = F_0.$$

Hence the set of indices satisfying the acceptance criterion is nonempty. By the definition of ℓ_{ep}, the algorithm accepts an iterate no later than this index, and its output satisfies

$$F(\mu^{\text{sad}}) - F(\mu_{\text{out}}) \geq F_0.$$

Therefore,

$$\mathbb{P}\left(F(\mu^{\text{sad}}) - F(\mu_{\text{out}}) \geq F_0\right) \geq \mathbb{P}(\mathcal{E} \cap \mathcal{D}) \geq 1 - \mathbb{P}(\mathcal{E}^c) - \mathbb{P}(\mathcal{D}^c) \geq 1 - \frac{3\zeta_{\text{ep}}}{4}.$$

□

M.3 PROOFS OF THE CONVERGENCE THEOREMS

We are now ready to prove the main convergence results. The argument combines two facts: outside the saddle region, large gradient steps decrease the objective by a uniform amount, while inside the saddle region, successful saddle-point episodes also give a decrease. To turn this into a global theorem, the Algorithm 1 uses phase specific parameters $\beta_{\mathrm{sad}} = b\delta^2$, $\tau = \tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$ during saddle episodes, while the large gradient phase retains parameters satisfying its own global descent condition $\tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$.

Theorem M.8 (Polynomial time approximate second-order stationarity)**.** *Let Assumptions 4.4 and 4.5 hold. Fix $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$. Assume that Assumption 4.2 holds on $S = \mathcal{L}_0$, with $\varepsilon_0, R_F > 0$. Fix $0 < \varepsilon \le \varepsilon_0$, $\delta > 0$, $\beta_{\mathrm{grad}} > 0$, and $\zeta \in (0,1)$, and set $F_{\min} := F(\mu^0) - \inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} F(\mu) > 0$. Run Algorithm 1 with $0 < \tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$, and the episode parameters $\beta_{\mathrm{sad}} = \Theta(\delta^2)$, $\tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$, $\kappa = \tilde{O}(1), n_{\mathrm{out}} = \tilde{O}(1), F_0 = \tilde{\Omega}(\delta^3), \eta = \tilde{\Theta}(\frac{\delta^3}{\varepsilon+\delta^{3/2}})$, prescribed in (92)-(99), subject to (100), using the same constant R_F at every episode. Choose these parameters jointly with $M_{\max} \in \mathbb{N}$ and $\zeta_{\mathrm{ep}} \in (0,1)$ such that $M_{\max} \ge \left\lfloor \frac{F_{\min}}{F_0} \right\rfloor + 1$, and $\frac{3}{4} M_{\max}\zeta_{\mathrm{ep}} \le \zeta$. At every initiated episode, draw a fresh Gaussian perturbation conditionally on the preceding history, with covariance $\mathrm{K}_{\bar{\mu}}^2$ on $L^2_{\bar{\mu}}$, where $\bar{\mu}$ is the episode's starting measure. Then the algorithm terminates and, with probability at least $1-\zeta$, returns an (ε,δ)-second-order critical point $\widehat{\mu}$. The number N_{SOSP} of WSFN transport steps satisfies*

$$N_{\mathrm{SOSP}} \le \frac{2(C_{\mathrm{H}}^2 + \beta_{\mathrm{grad}})}{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}} \frac{F_{\min}}{\varepsilon^2} + n_{\mathrm{out}} \left(\frac{F_{\min}}{F_0} + 1 \right). \tag{115}$$

Hence, with the large gradient parameters and other problem constants fixed,

$$N_{\mathrm{SOSP}} = \tilde{O}\left(F_{\min} \left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^3} \right) \right). \tag{116}$$

Proof. We first verify that Proposition M.7 applies uniformly whenever an episode is initiated at an (ε,δ)-saddle. Every accepted outer update decreases the objective. This follows from Lemma M.2 for a large gradient step, and from the acceptance criterion $F(\bar{\mu}) - F(\mu^{\mathrm{out}}) \ge F_0$ for a successful perturbation episode. Hence, every outer iterate, and every episode starting measure, belongs to $\mathcal{L}_0$.

Let $\bar{\mu}_j$ be the starting measure of an initiated j-th episode, and let $\mathcal{F}_j$ denote the history immediately before its perturbation is drawn. By the triggering condition, $\|\nabla_\mu F(\bar{\mu}_j, \cdot)\|_{L^2_{\bar{\mu}_j}} \le \varepsilon \le \varepsilon_0$. Since $\bar{\mu}_j \in \mathcal{L}_0$, Assumption 4.2 gives

$$\| \mathrm{M}_{\bar{\mu}_j} \|_{\mathrm{op}} \le R_F \|\nabla_\mu F(\bar{\mu}_j, \cdot)\|_{L^2_{\bar{\mu}_j}}.$$

The constant R_F is independent of j and of the realized history. Conditionally on $\mathcal{F}_j$, the reference measure $\bar{\mu}_j$ is fixed and the perturbation has the Gaussian law required by Proposition M.7. Therefore, on the event $\underline{\lambda}(\mathrm{K}_{\bar{\mu}_j}) < -\delta$, Proposition M.7 implies

$$\mathbb{P}\left(F(\bar{\mu}_j) - F(\mu_j^{\mathrm{out}}) \ge F_0 | \mathcal{F}_j \right) \ge 1 - \frac{3\zeta_{\mathrm{ep}}}{4}. \tag{117}$$

Further, we divide the proof into five steps.

Step 1: uniform decrease at every large gradient step. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, define

$$P_\mu := \left(\mathrm{H}_\mu^2 + \beta_{\mathrm{grad}} \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \right)^{-1/2}.$$

By Lemma M.2, every WSFN step satisfies

$$F(\mu^+) \le F(\mu) - \frac{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}}{2} \|P_\mu \nabla_\mu F(\mu, \cdot)\|^2_{L^2_\mu}. \tag{118}$$

Assumption 4.4 and Lemma L.11 imply $\| \mathrm{H}_\mu \|_{\mathrm{op}} \le C_{\mathrm{H}}$. Therefore, $\mathrm{H}_\mu^2 + \beta_{\mathrm{grad}} \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \preceq (C_{\mathrm{H}}^2 + \beta_{\mathrm{grad}}) \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. Since both sides are strictly positive, inversion reverses the operator order, so that

$\left(\mathrm{H}_\mu^2 + \beta_{\mathrm{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1} \succeq \frac{1}{C_{\mathrm{H}}^2+\beta_{\mathrm{grad}}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. Consequently,

$$\begin{aligned}\|P_\mu \nabla_\mu F(\mu,\cdot)\|_{L_\mu^2}^2 &= \left\langle \nabla_\mu F(\mu,\cdot), \left(\mathrm{H}_\mu^2 + \beta_{\mathrm{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1} \nabla_\mu F(\mu,\cdot)\right\rangle_{L_\mu^2} \\ &\geq \frac{1}{C_{\mathrm{H}}^2+\beta_{\mathrm{grad}}}\,\|\nabla_\mu F(\mu,\cdot)\|_{L_\mu^2}^2 .\end{aligned}$$

Substituting this estimate into (118) gives

$$F(\mu) - F(\mu^+) \geq \frac{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}}{2(C_{\mathrm{H}}^2+\beta_{\mathrm{grad}})}\,\|\nabla_\mu F(\mu,\cdot)\|_{L_\mu^2}^2 .$$

In particular, every outer iterate satisfying $\|\nabla_\mu F(\mu,\cdot)\|_{L_\mu^2} > \varepsilon$ produces an objective decrease strictly larger than $G_0 := \frac{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}}{2(C_{\mathrm{H}}^2+\beta_{\mathrm{grad}})}\varepsilon^2$. Let N_{grad} denote the total number of large gradient steps before termination. Since the objective is bounded below, $N_{\mathrm{grad}} G_0 \leq F(\mu^0) - \inf_\mu F(\mu) = F_{\min}$. Hence

$$N_{\mathrm{grad}} \leq \frac{F_{\min}}{G_0} = \frac{2(C_{\mathrm{H}}^2+\beta_{\mathrm{grad}})}{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}}\,\frac{F_{\min}}{\varepsilon^2}. \tag{119}$$

Step 2: the number of successful episodes is finite. Call an episode successful if its realized output satisfies $F(\bar{\mu}) - F(\mu^{\mathrm{out}}) \geq F_0$. Whenever an episode is successful, the next outer iterate is μ^{out}, and therefore the outer objective decreases by at least F_0. Let N_{succ} be the total number of successful episodes before termination. Telescoping the decreases over all successful episodes and all large gradient steps gives $N_{\mathrm{grad}} G_0 + N_{\mathrm{succ}} F_0 \leq F_{\min}$. In particular,

$$N_{\mathrm{succ}} \leq \frac{F_{\min}}{F_0}. \tag{120}$$

Every outer stage that does not terminate is either a large gradient step or a successful episode. By (119) and (120), only finitely many such stages can occur. The algorithm must therefore eventually encounter an unsuccessful episode and terminate. There is exactly one unsuccessful episode, namely the final episode. Consequently, the total number N_{ep} of initiated episodes satisfies

$$N_{\mathrm{ep}} = N_{\mathrm{succ}} + 1 \leq \left\lfloor \frac{F_{\min}}{F_0} \right\rfloor + 1 \leq M_{\max}. \tag{121}$$

Let $\ell_{\mathrm{ep},j} \leq n_{\mathrm{out}}$ denote the number of WSFN transport steps performed in the j-th episode. Since each large gradient step uses one transport step,

$$N_{\mathrm{SOSP}} = N_{\mathrm{grad}} + \sum_{j=1}^{N_{\mathrm{ep}}} \ell_{\mathrm{ep},j} \leq N_{\mathrm{grad}} + n_{\mathrm{out}} N_{\mathrm{ep}}.$$

Combining (119) and (121) therefore gives

$$N_{\mathrm{SOSP}} \leq \frac{2(C_{\mathrm{H}}^2+\beta_{\mathrm{grad}})F_{\min}}{\tau_{\mathrm{grad}}\sqrt{\beta_{\mathrm{grad}}}\varepsilon^2} + n_{\mathrm{out}}\left(\frac{F_{\min}}{F_0}+1\right),$$

which is (115).

Step 3: simultaneous high probability control of all true saddle episodes. If fewer than j episodes are initiated, define all events associated with the j-th episode to be empty. Let $\mathcal{B}_j$ be the event that:

1. the j-th episode is initiated at an (ε,δ)-saddle $\bar{\mu}_j$, i.e., $\|\nabla_\mu F(\bar{\mu}_j,\cdot)\|_{L_{\bar{\mu}_j}^2} \leq \varepsilon$, $\underline{\lambda}(\mathrm{K}_{\bar{\mu}_j}) < -\delta$,
2. the episode is unsuccessful, i.e., $F(\bar{\mu}_j) - F(\mu_j^{\mathrm{out}}) < F_0$.

Conditionally on $\mathcal{F}_j$, the point $\bar{\mu}_j$ is fixed and the perturbation used in that episode is fresh. Therefore, on the event that $\bar{\mu}_j$ is an (ε,δ)-saddle, the uniform escape guarantee (117) gives $\mathbb{P}(\mathcal{B}_j|\mathcal{F}_j) \leq$

$\frac{3\zeta_{\mathrm{ep}}}{4}$. Taking expectations yields $\mathbb{P}(\mathcal{B}_j) \le \frac{3\zeta_{\mathrm{ep}}}{4}$. Define the good event $\mathcal{G} := \bigcap_{j=1}^{M_{\max}} \mathcal{B}_j^c$. By the union bound and the choice of ζ_{ep},

$$\mathbb{P}(\mathcal{G}^c) \le \sum_{j=1}^{M_{\max}} \mathbb{P}(\mathcal{B}_j) \le M_{\max} \frac{3\zeta_{\mathrm{ep}}}{4} = M_{\max} \frac{3}{4} \frac{4\zeta}{3M_{\max}} = \zeta.$$

Thus, $\mathbb{P}(\mathcal{G}) \ge 1 - \zeta$. On $\mathcal{G}$, every episode initiated at an (ε, δ)-saddle is successful.

Step 4: the returned iterate is approximately second-order stationary. Let $\widehat{\mu}$ denote the iterate returned by the algorithm. By the stopping rule, the final episode was initiated at $\widehat{\mu}$, and hence

$$\|\nabla_\mu F(\widehat{\mu}, \cdot)\|_{L^2_{\widehat{\mu}}} \le \varepsilon. \tag{122}$$

Moreover, the final episode was unsuccessful:

$$F(\widehat{\mu}) - F(\mu^{\mathrm{out}}) < F_0. \tag{123}$$

We now work on the good event $\mathcal{G}$. Suppose, for contradiction, that $\underline{\lambda}(\mathrm{K}_{\widehat{\mu}}) < -\delta$. Together with (122), this would imply that $\widehat{\mu}$ is an (ε, δ)-saddle. By the definition of $\mathcal{G}$, every episode initiated at such a point is successful, so we would have $F(\widehat{\mu}) - F(\mu^{\mathrm{out}}) \ge F_0$, contradicting (123). Therefore, $\underline{\lambda}(\mathrm{K}_{\widehat{\mu}}) \ge -\delta$. Combining this with (122) proves that $\widehat{\mu}$ is an (ε, δ)-second-order critical point on an event of probability at least $1 - \zeta$.

Step 5: asymptotic form of the complexity bound. By (96) and $\beta_{\mathrm{sad}} = b\delta^2$,

$$n_{\mathrm{out}} \left(\frac{F_{\min}}{F_0} + 1 \right) = \frac{L_{\mathrm{H}}^2 \tau_{\mathrm{sad}}^3}{c_F \log^2(3/2) b^{3/2}} \frac{n_{\mathrm{out}}^4}{\delta^3} F_{\min} + n_{\mathrm{out}}.$$

Since $n_{\mathrm{out}} = \tilde{O}(1)$, from (115),

$$N_{\mathrm{SOSP}} = \tilde{O}\left(\frac{C_{\mathrm{H}}^2 + \beta_{\mathrm{grad}}}{\tau_{\mathrm{grad}} \sqrt{\beta_{\mathrm{grad}}}} \frac{F_{\min}}{\varepsilon^2} + \delta^{-3} F_{\min} + 1 \right).$$

For the fixed initial measure μ^0, we assume that $F_{\min} > 0$ is independent of the tolerances. Since $F_0(\delta) \to 0$ as $\delta \downarrow 0$, there exists $\delta_0 > 0$, possibly depending on $F_{\min}$, such that $F_0(\delta) \le F_{\min}$ whenever $0 < \delta \le \delta_0$. Hence,

$$n_{\mathrm{out}} \left(\frac{F_{\min}}{F_0} + 1 \right) \le 2 n_{\mathrm{out}} \frac{F_{\min}}{F_0},$$

so the terminal episode is absorbed into the episode complexity term, giving (116). □

Lemma M.9 (Sublevel control of the episode iterates)**.** *Let Assumptions 4.4 and 4.5 hold. Choose $\beta_{\mathrm{grad}} > 0$ and $0 < \tau_{\mathrm{grad}} \le C_{\mathrm{H}}^{-1} \sqrt{\beta_{\mathrm{grad}}}$, and choose the episode parameters as in (92)-(99), subject to (100). Consider Algorithm 1 with early acceptance. Let $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$ and let $\bar{\mu}^0 := \mu^0, \bar{\mu}^1, \ldots$ denote the outer iterates of the WSFN algorithm, i.e., the iterates at which the gradient $\nabla_\mu F$ is evaluated. If an episode is initiated at $\bar{\mu}^r$, write $\nu_r^0 := (\mathrm{Id} + \eta \xi_r)_{\#} \bar{\mu}^r$ for its perturbed initialization and*

$$\nu_r^{\ell+1} := \left(\mathrm{Id} - \tau \left(\mathrm{H}^2_{\nu_r^\ell} + \beta \, \mathrm{I}_{\mathrm{d} \times \mathrm{d}} \right)^{-1/2} \nabla_\mu F(\nu_r^\ell, \cdot) \right)_{\#} \nu_r^\ell, \quad 0 \le \ell < \ell_{\mathrm{ep},r},$$

for the temporary iterates within that episode. Define $\mathcal{L}_0 := \left\{ \mu \in \mathcal{P}_2(\mathbb{R}^d) : F(\mu) \le F(\mu^0) \right\}$, and let $\mathcal{E}_\kappa := \bigcap_{r:\ \text{episode } r \text{ initialized}} \left\{ \|\xi_r\|_{L^2_{\bar{\mu}^r}} \le \kappa \right\}$. Then:

1. *Every outer iterate and the terminal returned iterate belong to $\mathcal{L}_0$, i.e., $F(\bar{\mu}^r) \le F(\mu^0)$ for every outer iteration r.*
2. *On the event $\mathcal{E}_\kappa$, every computed temporary iterate in every perturbation episode satisfies $F(\nu_r^\ell) \le F(\mu^0) + F_0$, $0 \le \ell \le \ell_{\mathrm{ep},r}$.*

Hence, both conclusions hold on the intersection of $\mathcal{E}_\kappa$ with the event that all episodes initiated at (ε, δ)-saddles are successful.

Proof. We first prove the bound on the temporary episode iterates and then prove invariance of the initial sublevel set by induction over the outer iterations.

Step 1: control of the perturbation increase. Consider an episode initiated at an outer iterate $\bar{\mu}^r$. By the algorithmic trigger, $\|\nabla_\mu F(\bar{\mu}^r,\cdot)\|_{L^2_{\bar{\mu}^r}} \le \varepsilon$. On the event $\mathcal{E}_\kappa$, the corresponding perturbation satisfies $\|\xi_r\|_{L^2_{\bar{\mu}^r}} \le \kappa$. Applying the second-order smoothness estimate of Lemma L.14 to $\nu_r^0 = (\mathrm{Id}+\eta\xi_r)_{\#}\,\bar{\mu}^r$ gives

$$F(\nu_r^0) \le F(\bar{\mu}^r) + \eta\,\langle \nabla_\mu F(\bar{\mu}^r,\cdot),\xi_r\rangle_{L^2_{\bar{\mu}^r}} + \frac{C_{\mathrm{H}}}{2}\eta^2\|\xi_r\|^2_{L^2_{\bar{\mu}^r}} \le F(\bar{\mu}^r) + \eta\varepsilon\kappa + \frac{C_{\mathrm{H}}}{2}\eta^2\kappa^2.$$

It remains to verify that the last two terms sum to F_0 in (96). Set $s := \sqrt{\varepsilon^2 + 2C_H F_0}$. The choice of η gives $\eta\kappa = \frac{2F_0}{\varepsilon+s} = \frac{s-\varepsilon}{C_H}$, where the second equality follows from $(s-\varepsilon)(s+\varepsilon) = 2C_H F_0$. Therefore,

$$\eta\varepsilon\kappa + \frac{C_H}{2}\eta^2\kappa^2 = \frac{\varepsilon(s-\varepsilon)}{C_H} + \frac{(s-\varepsilon)^2}{2C_H} = \frac{(s-\varepsilon)(s+\varepsilon)}{2C_H} = F_0.$$

Hence

$$F(\nu_r^0) \le F(\bar{\mu}^r) + F_0. \tag{124}$$

Note that this argument uses only the small gradient condition and the bound on the perturbation norm. It does not require $\bar{\mu}^r$ to induce negative curvature.

Step 2: control of the computed episode iterates. Fix an episode r and work on $\mathcal{E}_\kappa$. Set $r_{\mathrm{loc}} := \frac{\sqrt{\beta_{\mathrm{sad}}}}{L_{\mathrm{H}}}$. The parameter choices (93), (96), (97), and (99) give

$$\eta\kappa + 2\sqrt{\tau_{\mathrm{sad}} n_{\mathrm{out}}}\,\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2} \le \frac{r_{\mathrm{loc}}}{2} < r_{\mathrm{loc}}.$$

Indeed, this follows from the same parameter estimate used to bound $3A+B$ in the proof of Proposition M.5, with $A = \eta\kappa$, $B = 2\sqrt{\tau_{\mathrm{sad}} n_{\mathrm{out}}}\,\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2}$. Write $\nu_r^j = (Z_r^j)_{\#}\bar{\mu}_r$, where

$$Z_r^0 = \mathrm{Id}+\eta\xi_r, \quad Z_r^{j+1} = \left(\mathrm{Id} - \tau_{\mathrm{sad}} d_r^j\right)\circ Z_r^j,$$

and $d_r^j := \left(\mathrm{H}^2_{\nu_r^j} + \beta_{\mathrm{sad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2} \nabla_\mu F(\nu_r^j,\cdot)$. For every $k < \ell_{\mathrm{ep},r}$, the acceptance criterion has failed at every $j \le k$. Thus

$$F(\bar{\mu}_r) - F(\nu_r^j) < F_0, \quad 0 \le j \le k.$$

Combining this with the perturbation cost bound (124) gives

$$F(\nu_r^0) - F(\nu_r^j) = F(\nu_r^0) - F(\bar{\mu}_r) + F(\bar{\mu}_r) - F(\nu_r^j) < 2F_0, \quad 0 \le j \le k.$$

For $k \ge 1$, apply Lemma M.4 to this prefix with $N = k$. For $k = 0$, use $\|Z_r^0 - \mathrm{Id}\,\|_{L^2_{\bar{\mu}_r}} \le \eta\kappa$ directly. In either case,

$$\|Z_r^k - \mathrm{Id}\,\|_{L^2_{\bar{\mu}_r}} \le \eta\kappa + 2\sqrt{\tau_{\mathrm{sad}} k}\,\beta_{\mathrm{sad}}^{-1/4} F_0^{1/2} < r_{\mathrm{loc}}.$$

Since (100) implies $\varepsilon \le \frac{\beta_{\mathrm{sad}}}{L_{\mathrm{H}}}$, Lemma M.3 applies at the starting point of every computed update, including the final update of a successful episode. Hence,

$$F(\nu_r^{k+1}) \le F(\nu_r^k), \quad 0 \le k < \ell_{\mathrm{ep},r}.$$

Together with (124), this gives

$$F(\nu_r^\ell) \le F(\nu_r^0) \le F(\bar{\mu}_r) + F_0, \quad 0 \le \ell \le \ell_{\mathrm{ep},r}. \tag{125}$$

Thus, whenever $F(\bar{\mu}_r) \le F(\mu^0)$, all computed temporary iterates satisfy $F(\nu_r^\ell) \le F(\mu^0) + F_0$.

Step 3: invariance of the initial sublevel set. We now prove by induction over the outer iterations that $F(\bar{\mu}^r) \le F(\mu^0)$. The assertion is immediate at $r = 0$, since $\bar{\mu}^0 = \mu^0$. Suppose that $F(\bar{\mu}^r) \le F(\mu^0)$ for some outer iteration r. There are two cases.

Case 1: large gradient step. If $\|\nabla_\mu F(\bar{\mu}^r,\cdot)\|_{L^2_{\bar{\mu}^r}} > \varepsilon$, the algorithm performs one WSFN step and sets

$$\bar{\mu}^{r+1} = \left(\mathrm{Id} - \tau_{\mathrm{grad}}\left(\mathrm{H}^2_{\bar{\mu}^r} + \beta_{\mathrm{grad}}\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-1/2}\nabla_\mu F(\bar{\mu}^r,\cdot)\right)_{\#}\bar{\mu}^r.$$

By Lemma M.2, $F(\bar{\mu}^{r+1}) \le F(\bar{\mu}^r) \le F(\mu^0)$.

Case 2: saddle point episode. Suppose instead that $\|\nabla_\mu F(\bar{\mu}^r, \cdot)\|_{L^2_{\bar{\mu}^r}} \le \varepsilon$, so that an episode is initiated. If the episode is successful, then $F(\bar{\mu}_r) - F(\nu_r^{\ell_{\mathrm{ep},r}}) \ge F_0$. The algorithm accepts $\bar{\mu}_{r+1} = \nu_r^{\ell_{\mathrm{ep},r}}$, and hence $F(\bar{\mu}_{r+1}) \le F(\bar{\mu}_r) - F_0 \le F(\mu^0)$.

If the episode is unsuccessful, the algorithm rejects the temporary trajectory and returns the stored pre-perturbation iterate $\widehat{\mu} = \bar{\mu}^r$. By the induction hypothesis, $F(\widehat{\mu}) = F(\bar{\mu}^r) \le F(\mu^0)$. The algorithm then terminates, so there are no subsequent outer iterates. Thus, in every case, each accepted outer iterate and the terminal returned iterate belong to $\mathcal{L}_0$. Combining this fact with (125) proves both conclusions. □

Proof of Corollary 4.8. Let $\mathcal{G}_{\mathrm{SOSP}}$ denote the event on which the conclusion of Theorem 4.6 holds. Thus, $\mathbb{P}(\mathcal{G}_{\mathrm{SOSP}}) \ge 1-\zeta$, and, on $\mathcal{G}_{\mathrm{SOSP}}$, the algorithm terminates and returns an iterate $\widehat{\mu}$ satisfying

$$\|\nabla_\mu F(\widehat{\mu}, \cdot)\|_{L^2_{\widehat{\mu}}} \le \varepsilon, \quad \underline{\lambda}(\mathrm{K}_{\widehat{\mu}}) \ge -\delta. \tag{126}$$

By Lemma M.9, every outer iterate of the algorithm belongs to $\mathcal{L}_0$. In particular, since the algorithm returns the stored pre-perturbation outer iterate when the final episode is unsuccessful, we have

$$\widehat{\mu} \in \mathcal{L}_0. \tag{127}$$

This conclusion holds on every sample path on which the algorithm terminates.

We now work on the event $\mathcal{G}_{\mathrm{SOSP}}$. Suppose, for a contradiction, that $\widehat{\mu} \notin \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$. Since $\widehat{\mu} \in \mathcal{L}_0$ by (127), the $(\varepsilon, \delta, \alpha_{\mathrm{ben}})$-saddle property on $\mathcal{L}_0$ implies that either $\|\nabla_\mu F(\widehat{\mu}, \cdot)\|_{L^2_{\widehat{\mu}}} > \varepsilon$, or $\underline{\lambda}(\mathrm{K}_{\widehat{\mu}}) < -\delta$. Both alternatives contradict (126). Therefore, $\widehat{\mu} \in \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$ on $\mathcal{G}_{\mathrm{SOSP}}$.

Hence, $\mathbb{P}\left(\widehat{\mu} \in \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)\right) \ge \mathbb{P}(\mathcal{G}_{\mathrm{SOSP}}) \ge 1 - \zeta$. Finally, by the definition of $\mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$, this means that

$$\mathrm{dist}_{W_2}(\widehat{\mu}, \mathcal{M}^*) = \inf_{\mu^* \in \mathcal{M}^*} W_2(\widehat{\mu}, \mu^*) \le \alpha_{\mathrm{ben}}.$$

The complexity bound is inherited directly from Theorem 4.6. □

Proof of Theorem 4.9. Since μ^* is a global minimizer, $\nabla_\mu F(\mu^*, \cdot) = 0$ in $L^2_{\mu^*}$. Moreover, $\mathrm{M}_{\mu^*} = 0$, μ^*-a.e., and hence $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$. The second-order necessary optimality condition in Proposition H.7 gives $\mathrm{K}_{\mu^*} \succeq 0$.

We first identify the derivative of the update map. Define the pulled-back Wasserstein gradient by $G(v) := \nabla_\mu F(\mu_v, \cdot) \circ Z_v$. By Lemma L.8, which follows from the pulled-back gradient differentiation identity associated with Proposition 2.3 and the Hessian estimate of Lemma L.12,

$$\|G(v) - G(0) - \mathrm{H}_{\mu^*} v\|_{L^2_{\mu^*}} \le \frac{L_{\mathrm{H}}}{2} \|v\|^2_{L^2_{\mu^*}}.$$

Since $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ is a global minimizer, hence local, Theorem H.4 gives $G(0) = 0$. Moreover, because $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$, we get

$$G(v) = \mathrm{K}_{\mu^*} v + R_G(v), \tag{128}$$

for some $R_G(v)$ such that $\|R_G(v)\|_{L^2_{\mu^*}} \le \frac{L_{\mathrm{H}}}{2} \|v\|^2_{L^2_{\mu^*}}$, for all $v \in L^2_{\mu^*}$.

For $v \in L^2_{\mu^*}$, let $\widehat{H}(v) := \widetilde{\mathrm{H}}_{\mu^*,1}$ denote the pulled-backed Hessian at μ_v, and set $\widehat{P}(v) := \left(\widehat{H}(v)^2 + \beta\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}}\right)^{-\frac{1}{2}}$. Lemma L.15 shows that $\widehat{P}(v)$ intertwines with the preconditioner H_{μ_v}, so that

$$\widehat{P}(v) J_v = J_v (\mathrm{H}^2_{\mu_v} + \beta\, \mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}},$$

where $J_v : L^2_{\mu_v} \to L^2_{\mu^*}$ is the composition operator given by $J_v w := w \circ (\mathrm{Id} + v)$ for $w \in L^2_{\mu_v}$. Moreover, Lemma L.12 and the identity $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$ give

$$\|\widehat{H}(v) - \mathrm{K}_{\mu^*}\|_{\mathrm{op}} \le L_{\mathrm{H}} \|v\|_{L^2_{\mu^*}}.$$

Applying Lemma L.16 gives

$$\left\|\widehat{P}(v) - (\mathrm{K}_{\mu^*}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\right\|_{\mathrm{op}} \le \frac{2L_{\mathrm{H}}}{\pi\beta}\|v\|_{L^2_{\mu^*}}. \tag{129}$$

Finally, since $(\lambda^2+\beta)^{-1/2} \le \beta^{-1/2}$ for every $\lambda \in \mathbb{R}$, continuous functional calculus gives $\|\widehat{P}(v)\|_{\mathrm{op}} \le \beta^{-1/2}$.

Using (128) and (129), we obtain

$$\begin{aligned}\widehat{P}(v)G(v) &= (\mathrm{K}_{\mu^*}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\,\mathrm{K}_{\mu^*}\,v \\ &\quad + \left(\widehat{P}(v) - (\mathrm{K}_{\mu^*}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\right)\mathrm{K}_{\mu^*}\,v + \widehat{P}(v)R_G(v),\end{aligned}$$

and therefore

$$\begin{aligned}\left\|\widehat{P}(v)G(v) - (\mathrm{K}_{\mu^*}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\,\mathrm{K}_{\mu^*}\,v\right\|_{L^2_{\mu^*}} &\le \left(\frac{2L_{\mathrm{H}}}{\pi\beta}\|\,\mathrm{K}_{\mu^*}\,\|_{\mathrm{op}} + \|\widehat{P}(v)\|_{\mathrm{op}}\frac{L_{\mathrm{H}}}{2}\right)\|v\|^2_{L^2_{\mu^*}} \\ &\le L_{\mathrm{H}}\left(\frac{2C_{\mathrm{K}}}{\pi\beta} + \frac{1}{2\sqrt{\beta}}\right)\|v\|^2_{L^2_{\mu^*}}.\end{aligned}$$

Substitution into $\Psi_{\tau,\beta}$ proves that there exists $C = C(L_{\mathrm{H}}, C_{\mathrm{K}}, \beta) > 0$ such that, for every $v \in L^2_{\mu^*}$,

$$\Psi_{\tau,\beta}(v) = \mathcal{L}_{\tau,\beta}v + \mathcal{R}_{\tau,\beta}(v), \quad \|\mathcal{R}_{\tau,\beta}(v)\|_{L^2_{\mu^*}} \le C\|v\|_*^2.$$

We next analyze the operator $\mathcal{L}_{\tau,\beta}$. Let E be the spectral measure of K_{μ^*}. By the spectral functional calculus,

$$\mathcal{L}_{\tau,\beta} = \int_{\sigma(\mathrm{K}_{\mu^*})}\left(1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}\right)\mathrm{d}E(\lambda).$$

Then, by (Conway, 1994, 1.10. Proposition, Chapter IX), for every $v \in L^2_{\mu^*}$ and every $n \ge 1$,

$$\mathcal{L}^n_{\tau,\beta}v = \int_{\sigma(\mathrm{K}_{\mu^*})}\left(1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}\right)^n\mathrm{d}E(\lambda)v. \tag{130}$$

For $\lambda > 0$, $0 \le 1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}} < 1$, whereas the multiplier equals 1 at $\lambda = 0$. Dominated convergence in (130) gives

$$\lim_{n\to\infty}\mathcal{L}^n_{\tau,\beta}v = E(\{0\})v = P_{\ker\mathrm{K}_{\mu^*}}v,$$

where $P_{\ker\mathrm{K}_{\mu^*}}v$ is the projection of v onto the null space K_{μ^*}. If K_{μ^*} is injective, then $\ker\mathrm{K}_{\mu^*} = \{0\}$ and $\lim_{n\to\infty}\mathcal{L}^n_{\tau,\beta}v = 0$.

Because K_{μ^*} is compact on the infinite dimensional space $L^2_{\mu^*}$, one has $0 \in \sigma(\mathrm{K}_{\mu^*})$ by (Reed & Simon, 1980, Theorem VI. 15). Hence $1 \in \sigma(\mathcal{L}_{\tau,\beta})$, and therefore $\|\mathcal{L}^n_{\tau,\beta}\|_{\mathrm{op}} = 1$.

Now assume that $(\mathrm{Id} - P_{\ker\mathrm{K}_{\mu^*}})v = \mathrm{K}^s_{\mu^*}w$ for some $s > 0$ and $w \in L^2_{\mu^*}$. Recall that K_{μ^*} is bounded, self-adjoint, and positive semidefinite, with $\|\,\mathrm{K}_{\mu^*}\,\|_{\mathrm{op}} \le C_{\mathrm{K}}$. The linearized WSFN operator can be written through the functional calculus of K_{μ^*} as $\mathcal{L}_{\tau,\beta} = \ell_{\tau,\beta}(\mathrm{K}_{\mu^*})$, where $\ell_{\tau,\beta}(\lambda) := 1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}$ for $\lambda \in \sigma(\mathrm{K}_{\mu^*})$.

We first show that $\mathcal{L}^n_{\tau,\beta}P_{\ker\mathrm{K}_{\mu^*}} = P_{\ker\mathrm{K}_{\mu^*}}$. Indeed, for every $u \in \ker\mathrm{K}_{\mu^*}$, one has $\mathrm{K}_{\mu^*}u = 0$, and therefore

$$(\mathrm{K}_{\mu^*}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\,\mathrm{K}_{\mu^*}\,u = 0.$$

It follows that

$$\mathcal{L}_{\tau,\beta}u = u - \tau(\mathrm{K}_{\mu^*}^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\,\mathrm{K}_{\mu^*}\,u = u.$$

Thus, $\mathcal{L}_{\tau,\beta}$ acts as the identity on $\ker\mathrm{K}_{\mu^*}$. Since $P_{\ker\mathrm{K}_{\mu^*}}u \in \ker\mathrm{K}_{\mu^*}$ for every $u \in L^2_{\mu^*}$, we obtain

$$\mathcal{L}^n_{\tau,\beta}P_{\ker\mathrm{K}_{\mu^*}}u = P_{\ker\mathrm{K}_{\mu^*}}u,$$

Alternatively, using the spectral calculus, $P_{\ker \mathrm{K}_{\mu^*}} = \mathbb{1}_{\{0\}}(\mathrm{K}_{\mu^*})$ and hence

$$\mathcal{L}_{\tau,\beta}^n P_{\ker \mathrm{K}_{\mu^*}} = \ell_{\tau,\beta}(\mathrm{K}_{\mu^*})^n \mathbb{1}_{\{0\}}(\mathrm{K}_{\mu^*}) = \ell_{\tau,\beta}(0)^n \mathbb{1}_{\{0\}}(\mathrm{K}_{\mu^*}) = P_{\ker \mathrm{K}_{\mu^*}},$$

because $\ell_{\tau,\beta}(0) = 1$.

We now decompose $v = P_{\ker \mathrm{K}_{\mu^*}} v + (\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v$ for every $v \in L^2_{\mu^*}$. Using $\mathcal{L}_{\tau,\beta}^n P_{\ker \mathrm{K}_{\mu^*}} = P_{\ker \mathrm{K}_{\mu^*}}$ and the source condition $(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v = \mathrm{K}_{\mu^*}^s w$, we obtain

$$\begin{aligned}
\mathcal{L}_{\tau,\beta}^n v - P_{\ker \mathrm{K}_{\mu^*}} v &= \mathcal{L}_{\tau,\beta}^n P_{\ker \mathrm{K}_{\mu^*}} v + \mathcal{L}_{\tau,\beta}^n (\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v - P_{\ker \mathrm{K}_{\mu^*}} v \\
&= \mathcal{L}_{\tau,\beta}^n (\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v \\
&= \mathcal{L}_{\tau,\beta}^n \mathrm{K}_{\mu^*}^s w.
\end{aligned}$$

Because both $\mathcal{L}_{\tau,\beta}$ and $\mathrm{K}_{\mu^*}^s$ are functions of K_{μ^*}, the spectral calculus gives

$$\left\| \mathcal{L}_{\tau,\beta}^n \mathrm{K}_{\mu^*}^s \right\|_{\mathrm{op}} = \sup_{\lambda \in \sigma(\mathrm{K}_{\mu^*})} \lambda^s \left| 1 - \tau \frac{\lambda}{\sqrt{\lambda^2 + \beta}} \right|^n \leq \sup_{0 \leq \lambda \leq C_{\mathrm{K}}} \lambda^s \left(1 - \tau \frac{\lambda}{\sqrt{\lambda^2 + \beta}} \right)^n. \tag{131}$$

Here we used $\sigma(\mathrm{K}_{\mu^*}) \subseteq [0, C_{\mathrm{K}}]$. Moreover, since $\tau \in (0, 1]$, $\beta > 0$, and $\lambda \geq 0$,

$$0 \leq \tau \frac{\lambda}{\sqrt{\lambda^2 + \beta}} < 1.$$

The elementary inequality $1 - x \leq e^{-x}$ implies

$$\left(1 - \tau \frac{\lambda}{\sqrt{\lambda^2 + \beta}} \right)^n \leq \exp\left(-\tau n \frac{\lambda}{\sqrt{\lambda^2 + \beta}} \right).$$

For $0 \leq \lambda \leq C_{\mathrm{K}}$, we also have

$$\sqrt{\lambda^2 + \beta} \leq \sqrt{C_{\mathrm{K}}^2 + \beta},$$

and consequently

$$\frac{\lambda}{\sqrt{\lambda^2 + \beta}} \geq \frac{\lambda}{\sqrt{C_{\mathrm{K}}^2 + \beta}}.$$

It follows that

$$\left(1 - \tau \frac{\lambda}{\sqrt{\lambda^2 + \beta}} \right)^n \leq \exp\left(-\frac{\tau n}{\sqrt{C_{\mathrm{K}}^2 + \beta}} \lambda \right). \tag{132}$$

Combining (131) and (132), we obtain

$$\left\| \mathcal{L}_{\tau,\beta}^n \mathrm{K}_{\mu^*}^s \right\|_{\mathrm{op}} \leq \sup_{\lambda \geq 0} \lambda^s \exp\left(-\frac{\tau n}{\sqrt{C_{\mathrm{K}}^2 + \beta}} \lambda \right).$$

For every $a > 0$ and $s > 0$, the function $\lambda \mapsto \lambda^s e^{-a\lambda}$ attains its maximum over $[0, \infty)$ at $\lambda = \frac{s}{a}$, and

$$\sup_{\lambda \geq 0} \lambda^s e^{-a\lambda} = \left(\frac{s}{ea} \right)^s.$$

Taking $a = \frac{\tau n}{\sqrt{C_{\mathrm{K}}^2 + \beta}}$ gives

$$\left\| \mathcal{L}_{\tau,\beta}^n \mathrm{K}_{\mu^*}^s \right\|_{\mathrm{op}} \leq \left(\frac{s \sqrt{C_{\mathrm{K}}^2 + \beta}}{e \tau n} \right)^s.$$

Therefore,

$$\left\| \mathcal{L}_{\tau,\beta}^n v - P_{\ker \mathrm{K}_{\mu^*}} v \right\|_{L^2_{\mu^*}} = \left\| \mathcal{L}_{\tau,\beta}^n \mathrm{K}_{\mu^*}^s w \right\|_{L^2_{\mu^*}} \leq \left(\frac{s \sqrt{C_K^2 + \beta}}{e \tau n} \right)^s \|w\|_{L^2_{\mu^*}},$$

as claimed.

Finally, on the range of P_m, every spectral value of K_{μ^*} satisfies $\lambda \geq m > 0$. Since $\lambda \mapsto 1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}$ is non-increasing on $[0,\infty)$,

$$\|\mathcal{L}_{\tau,\beta}^n P_m\|_{\mathrm{op}} \leq \left(1 - \tau\frac{m}{\sqrt{m^2+\beta}}\right)^n,$$

as required. □

Remark M.10 (Interpretation of Theorem 4.9)**.** *The multiplier $r_{\tau,\beta}(\lambda) = 1-\tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}$ distinguishes two curvature regimes. When $\lambda \gg \sqrt{\beta}$, it is approximately $1-\tau$. When $0 < \lambda \ll \sqrt{\beta}$, it is approximately $1 - \tau\frac{\lambda}{\sqrt{\beta}}$. Thus, regularized WSFN approximately normalizes sufficiently large positive curvature, while nearly flat directions remain weakly contracted.*

The absence of a uniform contraction on the full space must be distinguished from the behavior on the positive spectral subspace. If K_{μ^*} has infinitely many positive eigenvalues, they accumulate at zero, and no common geometric contraction factor applies to all positive curvature directions. If K_{μ^*} has finite rank, its nonzero spectrum is instead separated from zero, so the positive spectral component contracts geometrically. In either case, the null-space component is unchanged by the linearized dynamics.

The source condition $(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v = \mathrm{K}_{\mu^*}^s w$ is a Hölder-type spectral regularity condition, as used in spectral regularization, inverse problems, and statistical learning; see (Bauer et al., 2007, Section 2.3), (Lin et al., 2020, Assumption 2), and (Blanchard & Mücke, 2018, Exercise 2.12). On a positive eigenmode with eigenvalue λ, it introduces the factor λ^s and thereby suppresses the contribution of nearly flat directions. Larger s imposes stronger spectral regularity and gives the faster linearized rate n^{-s}.

Therefore, the spectral band and source condition estimates provide different quantitative conclusions. The former restricts attention to curvature bounded away from zero, whereas the latter controls the distribution of the initial perturbation across the positive spectrum.

M.4 Local spectral analysis and nonlinear limitations

This appendix derives perturbative estimates for a deterministic continuation of WSFN, and explains the additional requirements for transferring the linearized rates to nonlinear iterates. Throughout, fix a global minimizer $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$, $\tau \in (0,1]$, and $\beta > 0$. For fixed τ and β, write

$$\Phi_{\tau,\beta}(\mu) := \left(\mathrm{Id} - \tau(\mathrm{H}_\mu^2 + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-1/2}\nabla_\mu F(\mu,\cdot)\right)_{\#}\mu.$$

Algorithm 1 terminates when it returns $\widehat{\mu}$. The iterates considered below constitute an additional deterministic continuation from that measure, with a fixed pair $(\tau,\beta) \in (0,1]\times(0,\infty)$, not a further phase already guaranteed by the algorithm. This pair need not coincide with either phase specific pair used before termination.

Theorem M.11 (Post entry perturbative estimates)**.** *Let Assumptions 4.4 and 4.5 hold. Let $\widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ and define $v^0 := \mathrm{T}_{\mu^*}^{\widehat{\mu}} - \mathrm{Id}$, $v^{n+1} := \Psi_{\tau,\beta}(v^n)$, for $n \geq 0$. Then $\rho^n := (\mathrm{Id}+v^n)_{\#}\mu^*$ satisfies $\rho^0 = \widehat{\mu}$, $\rho^{n+1} = \Phi_{\tau,\beta}(\rho^n)$, and there exists $C(L_{\mathrm{H}}, C_{\mathrm{K}}, \beta) > 0$ such that*

$$v^{n+1} = \mathcal{L}_{\tau,\beta}v^n + \mathcal{R}_{\tau,\beta}(v^n), \tag{133}$$

with $\|\mathcal{R}_{\tau,\beta}(v^n)\|_{L^2_{\mu^*}} \leq C\|v^n\|^2_{L^2_{\mu^*}}$. Moreover:

- *If, in addition, $\left(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}}\right)v^0 = \mathrm{K}_{\mu^*}^s w$ for some $w \in L^2_{\mu^*}$ and $s > 0$, then*

$$\left\|v^n - P_{\ker \mathrm{K}_{\mu^*}}v^0\right\|_{L^2_{\mu^*}} \leq \left(\frac{s\sqrt{C_{\mathrm{K}}^2+\beta}}{e\tau n}\right)^s \|w\|_{L^2_{\mu^*}} + C\sum_{j=0}^{n-1}\|v^j\|^2_{L^2_{\mu^*}}. \tag{134}$$

- *For $m > 0$, let $q_m := 1 - \tau\frac{m}{\sqrt{m^2+\beta}}$. Then*

$$\|P_m v^n\|_{L^2_{\mu^*}} \leq q_m^n\|P_m v^0\|_{L^2_{\mu^*}} + C\sum_{j=0}^{n-1} q_m^{n-1-j}\|v^j\|^2_{L^2_{\mu^*}}. \tag{135}$$

Proof. Since μ^* is absolutely continuous, the optimal transport map $\mathrm{T}_{\mu^*}^{\widehat{\mu}}$ exists. Its properties give $\widehat{\mu} = (\mathrm{T}_{\mu^*}^{\widehat{\mu}})_{\#}\mu^* = (\mathrm{Id} + v^0)_{\#}\mu^*$.

Starting from $\bar{\mu}^0 := \widehat{\mu}$, consider the WSFN update

$$\bar{\mu}^{n+1} := \Big(\mathrm{Id} - \tau(\mathrm{H}^2_{\bar{\mu}^n} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\nabla_\mu F(\bar{\mu}^n, \cdot)\Big)_{\#}\bar{\mu}^n, \quad n \geq 0.$$

Let $\Psi_{\tau,\beta}$ denote the pulled-back WSFN map centered at μ^*, and define recursively

$$v^{n+1} := \Psi_{\tau,\beta}(v^n), \quad n \geq 0.$$

Suppose inductively that $\bar{\mu}^n = (\mathrm{Id} + v^n)_{\#}\mu^*$. The case $n = 0$ holds by construction. By the definition of the WSFN update,

$$\begin{aligned}
\bar{\mu}^{n+1} &= \Big(\mathrm{Id} - \tau(\mathrm{H}^2_{\bar{\mu}^n} + \beta\,\mathrm{I}_{\mathrm{d}\times\mathrm{d}})^{-\frac{1}{2}}\nabla_\mu F(\bar{\mu}^n, \cdot)\Big)_{\#}\bar{\mu}^n \\
&= \Big(\Big[\mathrm{Id} - \tau(\mathrm{H}^2_{\bar{\mu}^n} + \beta\,\mathrm{I}_{d\times d})^{-\frac{1}{2}}\nabla_\mu F(\bar{\mu}^n, \cdot)\Big] \circ (\mathrm{Id} + v^n)\Big)_{\#}\mu^* \\
&= (\mathrm{Id} + \Psi_{\tau,\beta}(v^n))_{\#}\,\mu^* = (\mathrm{Id} + v^{n+1})_{\#}\mu^*.
\end{aligned}$$

This proves $\bar{\mu}^n = (\mathrm{Id} + v^n)_{\#}\mu^*$ for all $n \geq 0$ by induction. The identity $\Phi_{\tau,\beta}(\mu_v) = (\mathrm{Id} + \Psi_{\tau,\beta}(v))_{\#}\mu^*$ shows inductively that ρ^n follows the deterministic WSFN update.

Applying Theorem 4.9 at the selected minimizer μ^* gives (133). Iterating this identity yields

$$v^n = \mathcal{L}^n_{\tau,\beta}v^0 + \sum_{j=0}^{n-1}\mathcal{L}^{n-1-j}_{\tau,\beta}\mathcal{R}_{\tau,\beta}(v^j), \tag{136}$$

for every $n \geq 1$. Since $\left\|\mathcal{L}^\ell_{\tau,\beta}\right\|_{\mathrm{op}} \leq 1$, $\ell \geq 0$, the remainder estimate in (133) gives

$$\left\|v^n - \mathcal{L}^n_{\tau,\beta}v^0\right\|_{L^2_{\mu^*}} \leq C\sum_{j=0}^{n-1}\|v^j\|^2_{L^2_{\mu^*}}.$$

Now, if $(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v^0 = \mathrm{K}^s_{\mu^*}\,w$, for some $s > 0$ and $w \in L^2_{\mu^*}$, the source condition estimate in Theorem 4.9 gives

$$\left\|\mathcal{L}^n_{\tau,\beta}v^0 - P_{\ker \mathrm{K}_{\mu^*}}v^0\right\|_{L^2_{\mu^*}} \leq \left(\frac{s\sqrt{C^2_{\mathrm{K}} + \beta}}{e\tau n}\right)^s \|w\|_{L^2_{\mu^*}}.$$

Combining this with (136) gives

$$\begin{aligned}
&\left|\left\|v^n - P_{\ker \mathrm{K}_{\mu^*}}v^0\right\|_{L^2_{\mu^*}} - \left\|\sum_{j=0}^{n-1}\mathcal{L}^{n-1-j}_{\tau,\beta}\mathcal{R}_{\tau,\beta}(v^j)\right\|_{L^2_{\mu^*}}\right| \\
&\leq \left\|v^n - P_{\ker \mathrm{K}_{\mu^*}}v^0 - \sum_{j=0}^{n-1}\mathcal{L}^{n-1-j}_{\tau,\beta}\mathcal{R}_{\tau,\beta}(v^j)\right\|_{L^2_{\mu^*}} \\
&\leq \left(\frac{s\sqrt{C^2_{\mathrm{K}} + \beta}}{e\tau n}\right)^s \|w\|_{L^2_{\mu^*}}.
\end{aligned}$$

Hence

$$\|v^n - P_{\ker \mathrm{K}_{\mu^*}}v^0\|_{L^2_{\mu^*}} \leq \left(\frac{s\sqrt{C^2_{\mathrm{K}} + \beta}}{e\tau n}\right)^s \|w\|_{L^2_{\mu^*}} + \sum_{j=0}^{n-1}\left\|\mathcal{L}^{n-1-j}_{\tau,\beta}\mathcal{R}_{\tau,\beta}(v^j)\right\|_{L^2_{\mu^*}}$$

Since $\|\mathcal{L}^\ell_{\tau,\beta}\|_{\mathrm{op}} \leq 1$, and the quadratic remainder estimate on $\mathcal{R}_{\tau,\beta}(v^j)$ proves (134).

Finally, since $P_m = \mathbb{1}_{[m,\infty)}(\mathrm{K}_{\mu^*})$ and $\mathcal{L}_{\tau,\beta} = \ell_{\tau,\beta}(\mathrm{K}_{\mu^*})$, $\ell_{\tau,\beta}(\lambda) := 1 - \tau\frac{\lambda}{\sqrt{\lambda^2+\beta}}$, $\lambda \in \sigma(\mathrm{K}_{\mu^*})$, both operators are Borel functions of the same self-adjoint bounded operator K_{μ^*}. (Reed & Simon, 1980, Theorem VII.2 (a)) implies $P_m\mathcal{L}_{\tau,\beta} = \mathcal{L}_{\tau,\beta}P_m$. Theorem 4.9 hence gives

$$\left\|\mathcal{L}^\ell_{\tau,\beta}P_m\right\|_{\mathrm{op}} \leq q^\ell_m.$$

Applying P_m to (136) therefore proves (135). □

Connection with the global guarantee. On the event in Corollary 4.8, the returned measure $\widehat{\mu}$ belongs to $\mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$. To obtain a small initial displacement relative to a fixed minimizer, one must additionally select $\mu^* \in \mathcal{M}^* \cap \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ that is close to $\widehat{\mu}$. For example, if $W_2(\widehat{\mu}, \mu^*) \le \alpha_{\mathrm{ben}}$, then $\|v^0\|_{L^2_{\mu^*}} = W_2(\widehat{\mu}, \mu^*) \le \alpha_{\mathrm{ben}}$. The fact that $\widehat{\mu} \in \mathcal{N}_{\alpha_{\mathrm{ben}}}(\mathcal{M}^*)$ alone does not ensure that a nearby minimizer is absolutely continuous or that the distance to the minimizer set is attained. The estimates above are pathwise and do not provide a further nonlinear convergence rate. They also do not establish that the continuation remains close to the selected minimizer. Although $\mathrm{Id} + v^0$ is an optimal transport map, the recursively constructed maps $\mathrm{Id} + v^n$ need not be optimal for $n \ge 1$.

Remark M.12 (Additional requirements for nonlinear rate transfer)**.** Accumulation and spectral transfer of nonlinear errors. *From the proof of Theorem M.11,*

$$\left\|v^n - \mathcal{L}_{\tau,\beta}^n v^0\right\|_{L^2_{\mu^*}} \le C \sum_{j=0}^{n-1} \|v^j\|^2_{L^2_{\mu^*}}.$$

There is no uniform geometric damping in this bound. In particular, small nonlinear errors can generate components in nearly flat directions. The null-space component can also change. Since $P_{\ker \mathrm{K}_{\mu^*}} \mathcal{L}_{\tau,\beta} = P_{\ker \mathrm{K}_{\mu^*}}$,

$$P_{\ker \mathrm{K}_{\mu^*}} v^n = P_{\ker \mathrm{K}_{\mu^*}} v^0 + \sum_{j=0}^{n-1} P_{\ker \mathrm{K}_{\mu^*}} \mathcal{R}_{\tau,\beta}(v^j).$$

Thus, $P_{\ker \mathrm{K}_{\mu^*}} v^0$ *is the limit of the linearized iteration, not an automatically preserved component or prescribed limit of the nonlinear iteration.*

Spectral band invariance. *One sufficient structural route to a nonlinear spectral band rate would be local invariance of* $V_m := \mathrm{Ran}\, P_m$*, that is,* $\Psi_{\tau,\beta}\left(V_m \cap B_{L^2_{\mu^*}}(0,r)\right) \subseteq V_m$*, where* $B_{L^2_{\mu^*}}(0,r) := \{v \in L^2_{\mu^*} : \|v\|_{L^2_{\mu^*}} < r\}$ *for the open ball of radius* $r > 0$ *centered at the zero vector field. Because* $\mathcal{L}_{\tau,\beta}$ *already preserves* V_m*, this is equivalent to* $(\mathrm{Id} - P_m)\mathcal{R}_{\tau,\beta}(v) = 0$*, for* $v \in V_m \cap B_{L^2_{\mu^*}}(0,r)$*. Under this additional condition,*

$$\|\Psi_{\tau,\beta}(v)\|_{L^2_{\mu^*}} \le \left(q_m + C\|v\|_{L^2_{\mu^*}}\right) \|v\|_{L^2_{\mu^*}}.$$

Choosing r *sufficiently small that* $q_m + Cr < 1$ *would therefore give geometric decay for sufficiently small initial data in* V_m*. This invariance is not a consequence of the present assumptions. It rules out nonlinear transfer from the band* $[m, \infty)$ *into the complementary spectral subspace* $[0, m)$ *and requires compatibility between the nonlinear terms and the spectral decomposition of the Hessian frozen at* μ^**.*

Source space control. *A possible route to nonlinear source condition rates would instead require quantitative control in the source space, for example* $(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})v^j = \mathrm{K}^s_{\mu^*} w_j$*, with* $\sup_{j \ge 0} \|w_j\|_{L^2_{\mu^*}} < \infty$*, or an estimate on the positive spectral part of the nonlinear remainder of the form*

$$(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})\mathcal{R}_{\tau,\beta}(v) \in \mathrm{Ran}(\mathrm{K}^s_{\mu^*}), \quad \left\|\mathrm{K}^{-s}_{\mu^*}(\mathrm{Id} - P_{\ker \mathrm{K}_{\mu^*}})\mathcal{R}_{\tau,\beta}(v)\right\|_{L^2_{\mu^*}} \le C_s \|v\|^2_{L^2_{\mu^*}}.$$

Here $\mathrm{K}^{-s}_{\mu^*}$ *is the spectral inverse with domain* $\mathrm{Ran}(\mathrm{K}^s_{\mu^*})$*, a subspace of* $(\ker \mathrm{K}_{\mu^*})^\perp$*. It is not assumed to be bounded or defined on the whole orthogonal complement. Such a weighted estimate is substantially stronger than the available ambient norm bound. Indeed, suppose that* K_{μ^*} *has positive eigenpairs* $\mathrm{K}_{\mu^*} e_j = \lambda_j e_j$*,* $\|e_j\|_{L^2_{\mu^*}} = 1$*,* $\lambda_j \downarrow 0$*. For* $h_j := \lambda_j^{s/2} e_j$*,*

$$\|h_j\|_{L^2_{\mu^*}} = \lambda_j^{s/2} \to 0, \quad \|\,\mathrm{K}^{-s}_{\mu^*} h_j\|_{L^2_{\mu^*}} = \lambda_j^{-s/2} \to \infty.$$

Thus, an error can be arbitrarily small in $L^2_{\mu^*}$ *while arbitrarily large in the source norm. This obstruction concerns the case of positive eigenvalues accumulating at zero, which is true for* K_{μ^*} *since it is Hilbert-Schmidt. Source space preservation or a weighted remainder estimate would still need to be combined with control of the accumulated errors in (136), including the null-space component. Neither displayed condition alone establishes the nonlinear rate* n^{-s}*.*

These invariance and source space conditions are strong structural requirements and can be difficult to verify. They illustrate possible routes to nonlinear rates, but cannot be considered as assumptions established by the present analysis. Accordingly, the unconditional rates in Theorem 4.9 are stated for the linearized dynamics.

N SCOPE OF THE IMPROVED SADDLE DEPENDENCE

This section explains the improvement of the saddle contribution from $\tilde{O}(\delta^{-4})$ to $\tilde{O}(\delta^{-3})$. The distinction is between the time required to escape an individual saddle neighborhood and the objective decrease used to count such episodes. Throughout, the regularity constants are fixed independently of δ.

We consider the saddle episode analysis with $\beta_{\mathrm{sad}} = b\delta^2$, $b > 0$, $\tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$, where b and τ_{sad} are fixed. The small gradient conditions and the local descent and first-exit estimates of that analysis are understood to hold. An episode uses at most n_{out} transport steps and accepts an output only after certifying a net decrease of at least F_0 relative to its pre-perturbation iterate.

N.1 EPISODE DURATION AND OBJECTIVE DECREASE COUNTING

We briefly recall the counting argument that was used in the proof of Theorem 4.6.Let $F_{\min} := F(\mu^0) - \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$. Suppose all accepted non-episode steps are descending. If N_{succ} episodes are successful, telescoping the accepted objective decreases gives

$$N_{\mathrm{succ}} F_0 \le F_{\min}.$$

For an algorithm that terminates at its first unsuccessful episode, there is at most one additional episode. Hence the number N_{sad} of transport steps spent in episodes satisfies

$$N_{\mathrm{sad}} \le n_{\mathrm{out}} \left(\frac{F_{\min}}{F_0} + 1 \right). \tag{137}$$

This counting inequality is deterministic. The high-probability escape guarantee is used separately to justify that termination does not occur at an (ε, δ)-saddle.

For comparison, the discrete-time PWGF analysis of (Yamamoto et al., 2025, Theorem 5.5 and Appendix E.2, equation (38)) uses an episode budget n_{out} of order $\tilde{O}(\delta^{-1})$ and a certified decrease F_0 at the cubic scale. Their resulting saddle term is $\tilde{O}(F_{\min}\delta^{-4})$, with regularity constants suppressed.

For WSFN, the revised parameter choices in (95) and (96) have the form

$$n_{\mathrm{out}} = \tilde{O}(1), \quad F_0 = c_F \frac{\beta_{\mathrm{sad}}^{3/2}}{L_{\mathrm{H}}^2 (\tau_{\mathrm{sad}} n_{\mathrm{out}})^3},$$

where $c_F > 0$ is independent of δ and may absorb fixed numerical factors such as $\log^2(3/2)$. Substituting $\beta_{\mathrm{sad}} = b\delta^2$ into (137) gives

$$N_{\mathrm{sad}} \le \frac{L_{\mathrm{H}}^2 \tau_{\mathrm{sad}}^3}{c_F b^{3/2}} \frac{n_{\mathrm{out}}^4}{\delta^3} F_{\min} + n_{\mathrm{out}} = \tilde{O}\left(\frac{L_{\mathrm{H}}^2 F_{\min}}{\delta^3} + 1 \right).$$

Note that the one-power improvement comes from episode duration, not from increasing the polynomial scale of the certified decrease.

N.2 CURVATURE NORMALIZATION REMOVES SLOW AMPLIFICATION

Consider an exact critical point $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$. Then $\mathrm{H}_{\mu^*} = \mathrm{K}_{\mu^*}$, and consider a unit eigenvector $q \in L^2_{\mu^*}$ with $\mathrm{K}_{\mu^*} q = -\gamma q$, where $\gamma > \delta > 0$. By Table 1, the frozen linearized Wasserstein gradient descent iteration is

$$v^{k+1} = (\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau_{\mathrm{GD}} \mathrm{H}_{\mu^*}) v^k, \quad v^0 \in L^2_{\mu^*},$$

where $\tau_{\mathrm{GD}} > 0$. Here "frozen" means that the Hessian is kept fixed at μ^*, not evaluated along the nonlinear trajectory.

Define the signed amplitude of the displacement in the direction q by $a_k := \langle v^k, q \rangle_{L^2_{\mu^*}}$. Thus, $a_k q$ is the orthogonal projection of v^k onto $\mathrm{span}\{q\}$, and its norm is $|a_k|$. By self-adjointness of H_{μ^*} and the eigenvector identity, we obtain

$$a_{k+1} = \left\langle (\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau_{\mathrm{GD}} \mathrm{H}_{\mu^*}) v^k, q \right\rangle_{L^2_{\mu^*}} = \left\langle v^k, (\mathrm{I}_{\mathrm{d}\times\mathrm{d}} - \tau_{\mathrm{GD}} \mathrm{H}_{\mu^*}) q \right\rangle_{L^2_{\mu^*}} = (1 + \tau_{\mathrm{GD}}\gamma) a_k.$$

Iterating this scalar recurrence gives

$$a_k = (1 + \tau_{\mathrm{GD}}\gamma)^k a_0, \quad |a_k| = (1 + \tau_{\mathrm{GD}}\gamma)^k |a_0|. \tag{138}$$

In particular, if $v^0 = a_0 q$, then $v^k = a_k q$ for every k, so $\|v^k\|_{L^2_{\mu^*}} = |a_k|$.

Fix a target amplitude $r > 0$ and suppose $0 < |a_0| < r$. Define the amplification time by $N_{\mathrm{GD}}(r, a_0) := \min\{k \in \mathbb{N}_0 : |a_k| \geq r\}$. The threshold r is measured in the reference $L^2_{\mu^*}$ norm. When $v^0 = a_0 q$, it is the radius of the displacement ball that the linearized trajectory must leave. For a general initial displacement, it is a threshold for the component along q.

Using (138), $|a_k| \geq r$ if and only if $|a_0|(1 + \tau_{\mathrm{GD}}\gamma)^k \geq r$. Using $\log(1 + \tau_{\mathrm{GD}}\gamma) > 0$, solving for k gives

$$N_{\mathrm{GD}}(r, a_0) = \left\lceil \frac{\log(|a_0|^{-1} r)}{\log(1 + \tau_{\mathrm{GD}}\gamma)} \right\rceil.$$

The numerator is the logarithm of the required amplification factor, whereas the denominator is the logarithmic growth per iteration. If $a_0 = 0$, this component remains zero under the frozen dynamics and never reaches a positive threshold. For $\gamma = \Theta(\delta)$ and a fixed admissible τ_{GD}, we have $\log(1 + \tau_{\mathrm{GD}}\gamma) = \Theta(\tau_{\mathrm{GD}}\delta)$, for sufficiently small δ, and hence

$$N_{\mathrm{GD}}(r, a_0) = O\left(\frac{1}{\tau_{\mathrm{GD}}\delta} \log\left(\frac{r}{|a_0|}\right)\right).$$

This isolates the inverse curvature δ^{-1} slowdown. The formula is exact for the frozen linearized component around μ^*, not for the nonlinear Wasserstein gradient descent trajectory. Transferring this amplification mechanism to the latter requires the tools from Yamamoto et al. (2025).

WSFN instead has the frozen multiplier

$$a_{k+1} = \left(1 + \tau_{\mathrm{sad}} \frac{\gamma}{\sqrt{\gamma^2 + \beta_{\mathrm{sad}}}}\right) a_k.$$

With $\beta_{\mathrm{sad}} = b\delta^2$,

$$1 + \tau_{\mathrm{sad}} \frac{\gamma}{\sqrt{\gamma^2 + \beta_{\mathrm{sad}}}} \geq 1 + \frac{\tau_{\mathrm{sad}}}{\sqrt{1 + b}} > 1. \tag{139}$$

Thus the amplification takes $O(\log(r/|a_0|))$ steps, without a polynomial factor in δ^{-1}. This rescaling that preserves curvature sign is the mechanism behind saddle-free Newton preconditioning. The nonlinear escape proof is still needed to control Hessian variation and the nonzero gradient at an approximate saddle, as the frozen multiplier alone is not an escape theorem. In particular, an order-one episode stepsize must be justified by local descent, instead of being inferred from (139); see Lemma M.3

N.3 Why the objective decrease scale is cubic

Preconditioning changes the update direction but does not change the objective. Suppose first that $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$. is an exact critical point and $q \in L^2_{\mu^*}$ is a unit vector satisfying $\langle \mathrm{H}_{\mu^*} q, q \rangle_{L^2_{\mu^*}} = -\gamma$, where $\gamma > \delta > 0$. Lemma L.14 gives

$$F((\mathrm{Id} + rq)_{\#}\mu^*) - F(\mu^*) \leq -\frac{\gamma}{2} r^2 + \frac{L_{\mathrm{H}}}{6} r^3.$$

The quadratic term and the cubic error balance when $r = \Theta(\frac{\gamma}{L_{\mathrm{H}}})$. For example, taking $r = \frac{\gamma}{L_{\mathrm{H}}}$ gives

$$F(\mu^*) - F((\mathrm{Id} + rq)_{\#}\mu^*) \geq \frac{\gamma^3}{3L_{\mathrm{H}}^2}.$$

Thus, at the weakest relevant curvature scale $\gamma = \Theta(\delta)$, the objective depth scale is asymptotically equivalent to $\frac{\delta^3}{L_{\mathrm{H}}^2}$. The same scale survives at an approximate saddle. If $\mathrm{K}_\mu\, q = \lambda q$ with $\lambda < -\delta$,

$\|q\|_{L^2_\mu} = 1$, and $\|\,\mathrm{M}_\mu\,\|_{\mathrm{op}} \le R_F\varepsilon \le \frac{\delta}{2}$, then $\langle \mathrm{H}_\mu\, q, q\rangle_{L^2_\mu} \le -\frac{\delta}{2}$. Choose the sign of q so that $\langle \nabla_\mu F(\mu), q\rangle \le 0$. The same cubic Taylor estimate, with $r = \frac{\delta}{2L_{\mathrm{H}}}$, gives

$$F(\mu) - F((\mathrm{Id} + rq)_{\#}\mu) > \frac{\delta^3}{24L_{\mathrm{H}}^2}.$$

The additional small gradient restrictions in the WSFN proof also control the sensitivity of the varying preconditioner.

N.4 Sharpness of the cubic decrease scale

The cubic scale is not only a feature of the Taylor upper bound. The following example shows that the entire available objective decrease can have precisely this order.

Proposition N.1 (A rank-one example with cubic objective depth)**.** *Fix $L > 0$ and $\bar\gamma > 0$. For $\gamma \in (0, \bar\gamma]$, define $f_\gamma(x) := \frac{\gamma^3}{L^2}[\cos(\frac{Lx}{\gamma}) - 1]$, and $F_\gamma(\mu) := f_\gamma(m(\mu))$, with $m(\mu) := \int_{\mathbb{R}} x\mu(\mathrm{d}x)$. Then F_γ is bounded below and Wasserstein regular. Its multiplication Hessian part vanishes, its integral Hessian part is rank at most one, Assumption 4.4 holds uniformly in γ, and Assumption 4.5 is satisfied with constant L. At every centered $\mu^* \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R})$, $\nabla_\mu F_\gamma(\mu^*, \cdot) = 0$, $\mathrm{K}_{\mu^*}\, q = -\gamma q$, for $q(x) \equiv 1$, and*

$$F_\gamma(\mu^*) - \inf_{\nu \in \mathcal{P}_2(\mathbb{R})} F_\gamma(\nu) = \frac{2\gamma^3}{L^2}. \tag{140}$$

Proof. Direct differentiation gives $f'_\gamma(x) = -\frac{\gamma^2}{L}\sin\left(\frac{Lx}{\gamma}\right)$, $f''_\gamma(x) = -\gamma\cos\left(\frac{Lx}{\gamma}\right)$, $f'''_\gamma(x) = L\sin\left(\frac{Lx}{\gamma}\right)$. Thus $\|f''_\gamma\|_\infty \le \bar\gamma$ and f''_γ is L-Lipschitz, uniformly in γ. The Wasserstein derivatives are $\nabla_\mu F_\gamma(\mu, x) = f'_\gamma(m(\mu))$, $\nabla\nabla_\mu F_\gamma(\mu, x) = 0$, and $\nabla^2_\mu F_\gamma(\mu, x, y) = f''_\gamma(m(\mu))$. These identities follow by differentiating the mean along transport curves. They give

$$\mathrm{M}_\mu = 0, \quad \mathrm{K}_\mu[v](x) = f''_\gamma(m(\mu)) \int_{\mathbb{R}} v(y)\mu(\mathrm{d}y).$$

In particular,

$$\|\,\mathrm{K}_\mu\,\|_{\mathrm{op}} = |f''_\gamma(m(\mu))| \le \bar\gamma.$$

For any coupling $\pi \in \Pi(\mu, \nu)$,

$$|m(\nu) - m(\mu)| = \left|\int (y - x)\pi(\mathrm{d}x, \mathrm{d}y)\right| \le \left(\int |y - x|^2 \pi(\mathrm{d}x, \mathrm{d}y)\right)^{1/2}.$$

Since the kernel is constant in its spatial variables, its $L^2_{\pi\otimes\pi}$ difference is bounded by L times the right-hand side. This proves the claimed coupling Lipschitz estimate. The multiplication part contributes zero.

At a centered measure, $f'_\gamma(0) = 0$ and $f''_\gamma(0) = -\gamma$. Since $\|q\|_{L^2_{\mu^*}} = 1$ for $q \equiv 1$, the asserted eigenvalue identity follows. Finally, $f_\gamma(0) = 0$ and $\inf_x f_\gamma(x) = -\frac{2\gamma^3}{L^2}$. This infimum is attained by translating μ^* to have mean $\frac{\pi\gamma}{L}$, which proves (140). □

Taking $\gamma = 2\delta$ gives a point with curvature strictly below $-\delta$ whose total available decrease is only $\frac{16\delta^3}{L^2}$. Hence, uniformly over this regularity class, one cannot guarantee a decrease of order $\omega(\delta^3)$ from every such point.

N.5 Why smaller regularization does not improve this balance

The normalized curvature is bounded above by one. Thus, once β is comparable to δ^2, decreasing β further cannot produce another polynomial improvement in the unstable multiplier. At the same time, the estimates in Lemma L.16 deteriorate as β decreases. This trade-off can be seen by formally optimizing the leading expression supplied by the escape proof. Fix $\tau_{\mathrm{sad}} \in (0, \frac{1}{2}]$ independently of

β and δ. Suppressing logarithms, the balances are $\tau_{\mathrm{sad}} n_{\mathrm{out}} = \tilde{O}(\tilde{\delta}^{-1})$, and $F_0 = \tilde{\Omega}(\frac{\beta^{3/2}}{L_{\mathrm{H}}^2 (\tau_{\mathrm{sad}} n_{\mathrm{out}})^3})$, with $\tilde{\delta} = \frac{\delta}{\sqrt{\delta^2+\beta}}$. The episode counting bound from the proof of Theorem 4.6 is

$$N_{\mathrm{sad}} \le F_{\min} \frac{n_{\mathrm{out}}}{F_0} + n_{\mathrm{out}}.$$

Since $F_0 > 0$, we have $\frac{1}{F_0} = \tilde{O}\left(\frac{L_{\mathrm{H}}^2 (\tau_{\mathrm{sad}} n_{\mathrm{out}})^3}{\beta^{3/2}}\right)$. Using n_{out}, we therefore obtain

$$\frac{n_{\mathrm{out}}}{F_0} = \tilde{O}\left(\frac{\tau_{\mathrm{sad}} L_{\mathrm{H}}^2}{\beta^{3/2}} n_{\mathrm{out}}^4\right) = \tilde{O}\left(\frac{L_{\mathrm{H}}^2}{\beta^{3/2}} \tilde{\delta}^{-4}\right).$$

By the definition of $\tilde{\delta}$,

$$\tilde{\delta}^{-4} = \left(\frac{\sqrt{\delta^2+\beta}}{\delta}\right)^4 = \frac{(\delta^2+\beta)^2}{\delta^4}.$$

Thus,

$$\frac{n_{\mathrm{out}}}{F_0} = \tilde{O}\left(L_{\mathrm{H}}^2 \frac{(\delta^2+\beta)^2}{\beta^{3/2}\delta^4}\right).$$

Suppressing logarithmic factors and fixed multiplicative constants, this motivates the cost proxy

$$\mathcal{C}(\beta, \delta) := L_{\mathrm{H}}^2 \frac{(\delta^2+\beta)^2}{\beta^{3/2}\delta^4}. \tag{141}$$

This proxy describes the leading coefficient of $F_{\min}$ in the upper bound. The additive cost of the terminal episode remains bounded by $n_{\mathrm{out}} = \tilde{O}(\tilde{\delta}^{-1})$. To optimize the proxy over β for fixed $\delta > 0$, introduce the dimensionless ratio $z := \frac{\beta}{\delta^2} > 0$, so that $\beta = z\delta^2$. Then $(\delta^2+\beta)^2 = \delta^4(1+z)^2$, $\beta^{3/2} = z^{3/2}\delta^3$. Substituting these identities into (141) gives

$$\mathcal{C}(\beta, \delta) = L_{\mathrm{H}}^2 \frac{\delta^4(1+z)^2}{z^{3/2}\delta^3\delta^4} = \frac{L_{\mathrm{H}}^2}{\delta^3} h(z), \quad h(z) := \frac{(1+z)^2}{z^{3/2}}.$$

Since $h(z) > 0$, logarithmic differentiation gives $\log h(z) = 2\log(1+z) - \frac{3}{2}\log z$, and hence

$$\frac{h'(z)}{h(z)} = \frac{2}{1+z} - \frac{3}{2z} = \frac{4z - 3(1+z)}{2z(1+z)} = \frac{z-3}{2z(1+z)}.$$

The denominator is positive for $z > 0$. Therefore, $h'(z) < 0$ on $(0, 3)$ and $h'(z) > 0$ on $(3, \infty)$, so $z = 3$ is the unique global minimizer. Its value is $h(3) = \frac{16}{3\sqrt{3}}$. Equivalently, $\operatorname{argmin}_{\beta>0} \mathcal{C}(\beta, \delta) = 3\delta^2$, and $\min_{\beta>0} \mathcal{C}(\beta, \delta) = \frac{16}{3\sqrt{3}} \frac{L_{\mathrm{H}}^2}{\delta^3}$. Thus the leading order proxy selects $\beta = \Theta(\delta^2)$ and retains the exponent δ^{-3}. Taking β much smaller or much larger than δ^2 worsens this proxy, since $h(z) \sim z^{-3/2}$ as $z \downarrow 0$, and $h(z) \sim z^{1/2}$, as $z \to \infty$.

The value $\beta = 3\delta^2$ minimizes only the displayed proxy with τ_{sad} held fixed. Logarithmic factors and the remaining admissibility conditions can change the preferred constant. Moreover, minimizing an upper bound proxy does not establish a lower bound on the actual iteration complexity.

N.6 What is sharp and what might improve further

Proposition N.1 establishes sharpness of the uniform objective decrease scale, not an $\Omega(\delta^{-3})$ iteration lower bound for WSFN or for arbitrary second-order methods. In that example, $F_{\min}$ is of order δ^3, so it does not force δ^{-3} successive episodes. A complexity lower bound with a fixed initial objective gap $F_{\min}$ would need a separate construction and an argument excluding faster ways of reaching approximate second-order stationarity.

Stronger global information could improve the episode count without changing the local escape rate. Such information is additional to qualitative benignity since the absence of nonglobal local minima does not itself bound the number of saddle neighborhoods encountered by a trajectory.

Higher-order regularity and higher-order models provide another route. For Euclidean unconstrained optimization, Cartis et al. (2020) show that adaptive regularization with Taylor models of order

$p \geq 2$, for a lower bounded objective with Lipschitz p-th derivative, attains first and second-order tolerances (ε, δ) with evaluation complexity $O\left(\max\left\{\varepsilon^{-(p+1)/p}, \delta^{-(p+1)/(p-1)}\right\}\right)$, with the other problem and algorithm parameters fixed. The second-order exponent is 3 for $p = 2$ and 2 for $p = 3$. These are Euclidean evaluation complexity results under their own model solution requirements, and do not imply consequences of the present Wasserstein analysis. They nevertheless show why the cubic exponent should not be presented as a universal barrier independent of smoothness, oracle information, or algorithm design. Indeed, the example above has $\|f_\gamma^{(4)}\|_\infty = \frac{L^2}{\gamma}$, so its third derivative is not Lipschitz with a constant uniform in γ. It therefore does not satisfy the stronger uniform regularity needed for the $p = 3$ comparison.

Finally, the improved saddle contribution must be distinguished from the leading total complexity. Theorem 4.6 contains a large gradient contribution proportional to $F_{\min}\varepsilon^{-2}$, while condition 100 requires $\varepsilon \leq \frac{b\delta^2}{L_{\mathrm{H}}}$. Hence, the total upper bound is $\tilde{O}(F_{\min}\varepsilon^{-2} + 1)$ throughout the admissible regime. At the largest admissible polynomial scale $\varepsilon = \tilde{\Theta}(\delta^2)$, this becomes $\tilde{O}(F_{\min}\delta^{-4})$ for fixed $F_{\min} > 0$.

Therefore, WSFN improves the certified cost of saddle episodes without improving the leading polynomial order of the present total upper bound. Improving that leading order requires a sharper analysis of the large gradient phase or a weaker restriction linking ε and δ. A further reduction of the episode contribution alone would not change the leading order of the current guarantee.

Proposition N.2 (Limitation of the large gradient phase)**.** *Under the restriction $\tau_{\mathrm{grad}} \leq C_{\mathrm{H}}^{-1}\sqrt{\beta_{\mathrm{grad}}}$, the large gradient phase admits $\Omega(\varepsilon^{-2})$ worst case examples. These examples satisfy Assumptions 4.2, 4.4, and 4.5 with constants independent of ε, and have initial objective gap $F_{\min}$ bounded above and below by positive constants.*

Proof. Choose a nonnegative smooth function $q \in C_c^\infty((-1, 0))$ with $\|q\|_\infty = 1$, and set $a := \int_{\mathbb{R}} q(t)\mathrm{d}t > 0$. For $0 < \varepsilon < \frac{a}{4}$, define f_ε by

$$f_\varepsilon(0) = 0, \quad f_\varepsilon'(x) = 2\varepsilon - \int_x^0 q(t)\mathrm{d}t.$$

Then $f_\varepsilon'' = q$, so the Hessian bound and its Lipschitz constant are independent of ε. Moreover, f_ε is convex and bounded below, with a minimizer in $(-1, 0)$. For $x \geq 0$, we have $f_\varepsilon(x) = 2\varepsilon x$.

Define the functional on $\mathcal{P}_2(\mathbb{R})$ by $F_\varepsilon(\mu) := f_\varepsilon(m_\mu)$, with $m_\mu := \int_{\mathbb{R}} x\mathrm{d}\mu(x)$. Its Wasserstein derivatives satisfy $\nabla_\mu F_\varepsilon(\mu, x) = f_\varepsilon'(m_\mu)$, $\mathrm{M}_\mu = 0$, $(\mathrm{K}_\mu v)(x) = q(m_\mu)\int_{\mathbb{R}} v\mathrm{d}\mu$. Hence $\|\mathrm{K}_\mu\|_{\mathrm{HS}} \leq 1$, and Assumptions 4.2, 4.4 are automatic. Assumption 4.5 follows from

$$|q(m_\nu) - q(m_\mu)| \leq \|q'\|_\infty |m_\nu - m_\mu| \leq \|q'\|_\infty d_\gamma,$$

where $\gamma \in \Pi(\mu, \nu)$. Thus the required regularity constants are uniform in ε.

Initialize with an absolutely continuous measure of mean $m_0 = (2\varepsilon)^{-1}$, for example $\mu^0 = \mathcal{N}(m_0, 1)$. Since $\min f_\varepsilon \in [-2\varepsilon, 0]$,

$$1 \leq F_\varepsilon(\mu^0) - \inf_\mu F_\varepsilon(\mu) \leq 1 + 2\varepsilon.$$

While $m_k \geq 0$, the gradient norm equals $2\varepsilon > \varepsilon$ and $\mathrm{H}_{\mu^k} = 0$. Consequently, only large gradient steps are performed, and

$$m_{k+1} = m_k - \frac{2\tau_{\mathrm{grad}}\varepsilon}{\sqrt{\beta_{\mathrm{grad}}}},$$

so that

$$m_k - m_{k+1} \leq \frac{2\varepsilon}{C_{\mathrm{H}}}.$$

Reaching a negative mean therefore requires at least $\frac{C_{\mathrm{H}}}{4\varepsilon^2}$ steps. The iterates are translations of the initial measure, so absolute continuity is preserved. □