On the Mean-field Analysis of Normalized Steepest Descent via Linear Minimization Oracles

Yongtao Wu YONGTAO.WU@EPFL.CH
EPFL

Fanghui Liu FANGHUI.LIU@SJTU.EDU.CN
Shanghai Jiao Tong University

Taiji Suzuki TAIJI@MIST.I.U-TOKYO.AC.JP
RIKEN AIP; The University of Tokyo

Volkan Cevher VOLKAN.CEVHER@EPFL.CH
EPFL

Abstract

In this paper, we develop a mean-field analysis of normalized steepest descent for two-layer neural networks in Wasserstein space. We first study the induced gradient flow in continuous time, establishing global convergence guarantees in both Euclidean and Wasserstein spaces. Our analysis reveals a fundamental distinction between normalized and unnormalized dynamics: while the latter exhibits linear convergence, the normalized flow reaches the optimum in finite time. We further extend the framework to finite-particle and discrete-time settings. We introduce LMO-driven Langevin dynamics and develop adaptive LMO particle schemes, establishing non-asymptotic convergence and stationarity guarantees. Collectively, our results provide a theoretical foundation for normalized steepest descent, a class of optimization methods that has recently become popular for training neural networks.

1. Introduction

The growing cost of training large language models (LLMs) has renewed interest in optimization methods beyond standard Euclidean geometry. Adaptive optimizers such as Adam [23] and AdamW [28] rely on coordinate-wise second-moment estimates under the ℓ_2 norm, which only partially capture the geometry of large-scale neural network losses [48]. This has motivated a family of normalized steepest descent methods based on non-Euclidean geometries, including spectral-norm-based optimizers such as Muon and its variants [2, 22, 26, 35].

Despite their empirical success, the theoretical foundations of normalized steepest descent remain limited. A useful framework for analyzing optimization dynamics is the mean-field limit of two-layer neural networks [13, 14, 31, 32, 38, 41, 42], where training is described by an evolution over probability measures $\mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathscr{F}(\mu) := F(\mu) + \lambda \operatorname{Ent}(\mu), \qquad \lambda \geq 0. \tag{1}$$

For $\lambda > 0$, the associated Wasserstein gradient flow corresponds to mean-field Langevin dynamics (MFLD), whose convergence properties are well understood in both continuous and discrete settings [11, 12, 21, 30, 32, 43]. See more detail on related work in Sec. B.

However, existing analyses exclusively study *unnormalized* steepest descent under ℓ_2 geometry. For normalized steepest descent, it remains unclear how to formulate the corresponding Wasserstein gradient flow and analyze its finite-particle approximation. This raises the following question:

Can we establish a convergence theory for normalized steepest descent in Wasserstein space?

In this work, we study two-layer neural networks in the mean-field regime and develop a convergence analysis of normalized steepest descent via linear minimization oracles (LMOs) [17, 25].

Definition 1 (LMO) *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ with dual norm $\|\cdot\|_*$. We define:*

$$\mathrm{lmo}(\boldsymbol{g};\tau) \;:=\; \operatorname*{arg\,min}_{\boldsymbol{y}:\|\boldsymbol{y}\|\le\|\boldsymbol{g}\|_*^{\tau}} \langle \boldsymbol{g},\boldsymbol{y}\rangle\,, \quad \forall \boldsymbol{g}\in\mathbb{R}^d,\ \tau\in\{0,1\}\,.$$

By LMO, we can unify normalized and unnormalized steepest descent (with different radii) under the same framework. For instance, given a function $f:\mathbb{R}^d\to\mathbb{R}$ and a sequence of step sizes $\{\eta_k\}_{k\ge1}$, the LMO update is: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \eta_k\,\mathrm{lmo}\left(\nabla f(\boldsymbol{x}_k);\tau\right)$. It allows for a general optimization framework, e.g., $\tau=0$ corresponding to normalized steepest descent, and $\tau=1$ corresponding to (unnormalized) steepest descent; see more examples in Tab. 2.

2. Notation

We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of all probability measures on $\mathbb{R}^d$ with finite second moment. The minimizer of $\mathscr{F}$ is denoted by $\mu^\star$. We denote by $\mathscr{F}^\star$ (resp. $F^\star$) the optimal value when $\lambda>0$ (resp. $\lambda=0$). For a functional $F:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$, we denote its first variation with respect to μ by $\frac{\delta F}{\delta\mu}(\mu)$, which is assumed to exist and to be continuously differentiable throughout the paper. We use $f:\mathbb{R}^d\to\mathbb{R}$ as the objective function when optimizing in Euclidean space while F and $\mathscr{F}$ are the functionals $\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$. We denote by $\|\cdot\|$ a general primal norm, and by $\|\cdot\|_*$ its dual norm on $\mathbb{R}^d$. We denote the negative Shannon entropy as $\mathrm{Ent}(\mu) := \int_{\mathbb{R}^d}\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}(x)\right)\mathrm{d}\mu(x)$. For $\mu\ll\nu$, let $H(\mu|\nu)$ denote the Kullback–Leibler divergence. We often identify absolutely continuous probability measures with their densities with respect to the Lebesgue measure for notational simplicity. The 2-Wasserstein distance between $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$ is defined as: $W_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\boldsymbol{x}-\boldsymbol{y}\|^2\,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y})\right)^{1/2}$, where $\Pi(\mu,\nu)$ denotes the set of all couplings of μ and ν. Let $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ and let $f:\mathbb{R}^d\to\mathbb{R}^d$ be measurable, given a norm $\|\cdot\|$, we define the space of p–integrable vector fields as $L^p(\mu;\|\cdot\|) := \left\{f:\mathbb{R}^d\to\mathbb{R}^d \;\middle|\; \|f\|_{L^p(\mu;\|\cdot\|)}<\infty\right\}$, where $\|f\|_{L^p(\mu;\|\cdot\|)} := \left(\int_{\mathbb{R}^d}\|f(x)\|^p\,\mathrm{d}\mu(x)\right)^{1/p}$, $1\le p<\infty$. We denote by $C^1(\mathbb{R}^d)$ the space of continuously differentiable functions on $\mathbb{R}^d$. We use $\delta_{\boldsymbol{x}}$ to denote the Dirac measure at $\boldsymbol{x}$.

3. LMO gradient flow

In this section, we study the analysis of LMO gradient flow to optimization over probability measures based on the objective defined in Eq. (1), which is the extension of our result for optimization over Euclidean space (Thm. 7 in Sec. C of the appendix). We first introduce the following assumptions.

Assumption 1 *The functional F is convex and $\mathscr{F}$ admits a minimizer $\mu^\star$.*

Remark 2 *The same assumption has been made in Chizat [12] for analyzing two-layer networks. Proposition 2.5 of Hu et al. [21] guarantees the existence and uniqueness of the minimizer of $\mathscr{F}$ under mild conditions. Hence Asm 1 can be verified.*

Assumption 2 (Smoothness in Wasserstein space) *We assume that the first variation of F is L-Lipschitz, i.e., there exists a constant $L > 0$ such that for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$:*

$$\left\|\nabla \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) - \nabla \frac{\delta F}{\delta \mu}(\nu)(\boldsymbol{x}')\right\|_* \leq \frac{L}{2}\big(W_2(\mu,\nu) + \|\boldsymbol{x} - \boldsymbol{x}'\|\big),$$

Remark 3 *Note that Asm 2 with the ℓ_2 norm is commonly used in the literature [12, 43]. We extend it to a general norm $\|\cdot\|$ with the associated dual norm $\|\cdot\|_*$. Moreover, for two-layer mean-field neural networks, Asm 2 is satisfied under mild conditions on the neuron and the loss, as shown in Prop. 1 of the appendix.*

In the analysis of MFLD for neural networks, log-Sobolev inequalities play a crucial role in establishing linear convergence rates [12, 32] and it can be verified for two-layer mean-field neural network (Proposition 5.1 in Chizat [12]). In this work, we consider a generalized log-Sobolev assumption formulated with respect to the dual norm of a general norm, i.e., $\|\cdot\|_*$ [1, 5]. When $\|\cdot\|$ is the Euclidean ℓ_2 norm, this assumption reduces to the standard log-Sobolev inequality.

Assumption 3 (Generalized Log-Sobolev inequality) *Fix a reference measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Let ν be the Gibbs measure on $\mathbb{R}^d$ with density $d\nu(x) \propto \exp\Big(-\frac{1}{\lambda}\frac{\delta F(\mu)}{\delta \mu}(x)\Big)\,dx$. Assume there exists $\zeta > 0$ such that for every probability measure $\varphi \ll \nu$, it holds $H(\varphi \,\|\, \nu) \leq \frac{1}{2\zeta}\int \left\|\nabla \log \frac{\mathrm{d}\varphi}{\mathrm{d}\nu}(x)\right\|_*^2 \mathrm{d}\varphi(x)$.*

To deliver our analysis, we need the following assumption to avoid "ill-conditioned" measures.

Assumption 4 *Given probability measures $\{\mu_t\}$ along the trajectory of interest, there exists a constant $\kappa \in (0, 1]$ such that, the associated Wasserstein gradient field $g_{\mu_t}(\boldsymbol{x}) := \nabla \frac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x})$ satisfies the uniform lower bound $\|g_{\mu_t}\|_{L^1(\mu_t;\|\cdot\|_*)} \geq \kappa \|g_{\mu_t}\|_{L^2(\mu_t;\|\cdot\|_*)}$.*

Remark: Asm 4 admits a similar form of reverse Hölder's inequality [34], see Lemma 1 in Sec. D.2 for further discussion. Asm 4 rules out the case where $\|g_\mu(\boldsymbol{x})\|_*$ is highly concentrated on a μ-negligible region (e.g., spiky gradients). Here we give a simple sufficient condition to Asm 4 as below, with proof deferred to Sec. D.1.

Condition 1 (Sufficient condition of Asm 4) Assume that for all admissible probability measures μ along the trajectory, there exists a constant $G_{\max} < \infty$ such that $\|g_\mu(\boldsymbol{x})\|_* \leq G_{\max}$ holds μ-a.s. (classical bounded gradient assumption), and there exist constants $m_1 > 0$, $m_2 \in (0, 1]$ such that $\mu\Big(\Big\{\boldsymbol{x} : \|g_\mu(\boldsymbol{x})\|_* \geq m_1\Big\}\Big) \geq m_2$. Then Asm 4 holds with $\kappa = \frac{m_2 m_1}{G_{\max}}$.

Note that the PDE associated with the MFLD as presented in Eq. (7) can be written as follows:

$$\partial_t \mu_t = \nabla \cdot (\mu_t \boldsymbol{v}_t), \quad \boldsymbol{v}_t := \nabla \tfrac{\delta F}{\delta \mu}(\mu_t) + \lambda \nabla \log \mu_t\,. \tag{2}$$

We are now ready to state our main theorem with an analysis on an LMO analogue of the dynamics in Eq. (2). The proof can be found in Sec. D.4.

Algorithm 1: Adaptive LMO Particle Descent

Input: $N, \eta, S_{\text{init}}, \tau, K$, initial particles $(\boldsymbol{x}_0^i)_{i=1}^N \sim \mu_0$, and $s \in \{\texttt{global}, \texttt{per-particle}\}$
Output: $\mu_K = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_K^i}$
$S_{-1} \leftarrow S_{\text{init}}$
if $s = \texttt{per-particle}$ **then** $S_{-1}^i \leftarrow S_{\text{init}}$ for all $i \in [N]$
for $k = 0$ **to** $K-1$ **do**
  $\mu_k \leftarrow \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_k^i}$
  for $i = 1$ **to** N **do**
    $\boldsymbol{g}_k^i \leftarrow \nabla \frac{\delta F}{\delta \mu}(\mu_k)(\boldsymbol{x}_k^i)$, $\boldsymbol{v}_k^i \leftarrow \mathrm{lmo}(\boldsymbol{g}_k^i; \tau)$
    if $s = \texttt{per-particle}$ **then** $S_k^i \leftarrow S_{k-1}^i + \|\boldsymbol{v}_k^i\|^2$, $\eta_k^i \leftarrow \eta/\sqrt{S_k^i}$
  end
  if $s = \texttt{global}$ **then** $S_k \leftarrow S_{k-1} + \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{v}_k^i\|^2$, $\eta_k \leftarrow \eta/\sqrt{S_k}$, $\eta_k^i \leftarrow \eta_k$ for all $i \in [N]$
  $\boldsymbol{x}_{k+1}^i \leftarrow \boldsymbol{x}_k^i + \eta_k^i \boldsymbol{v}_k^i$ for all $i \in [N]$
end

Theorem 4 *Under Asm. 1 to 4, consider the following LMO-based continuity equation:*

$$\partial_t \mu_t = \nabla \cdot \big(\mu_t \, \boldsymbol{v}_t\big), \qquad \boldsymbol{v}_t := -\,\mathrm{lmo}\left(\nabla \frac{\delta F}{\delta \mu}(\mu_t) + \lambda \nabla \log \mu_t; 0\right). \tag{3}$$

Then for all $t \geq 0$, we can obtain: $\mathscr{F}(\mu_t) - \mathscr{F}(\mu^\star) \leq \left[\max\left\{\sqrt{\mathscr{F}(\mu_0) - \mathscr{F}(\mu^\star)} - \frac{\kappa}{2}\sqrt{2\,\zeta\,\lambda}\,t,\, 0\right\}\right]^2$.
In particular, $\mathscr{F}(\mu_t) = \mathscr{F}(\mu^\star)$ *for all* $t \geq t^\star$, *with* $t^\star = \frac{\sqrt{2}}{\kappa\sqrt{\zeta\,\lambda}}\sqrt{\mathscr{F}(\mu_0) - \mathscr{F}(\mu^\star)}$.

The result above can be considered as the extension of Thm. 7 (Euclidean space) to Wasserstein space, where finite time convergence is achieved instead of linear convergence.

4. Discrete time analysis

The result in the previous section establishes convergence for the continuous-time LMO flow in Wasserstein space with an infinite number of particles. In practice, however, one must work with a finite number of particles and a discrete-time dynamics. In this section, we study the convergence of the algorithm in discrete time with finite particles.

4.1. Discrete time analysis: adaptive LMO particle descent

Here, we work in a general non-convex setting without the requirement of Asm 3 and seek a stationarity guarantee. We consider the case without entropy regularization ($\lambda = 0$) for simplicity. A technical issue is that an optimal step size relies on the knowledge of the smoothness constant L. Motivated by AdaGrad [16], we consider adaptive step sizes that automatically scale with the accumulated gradient, with the following two cases (see Alg. 1): **global** adaptive step size for all particles; **per-particle** adaptive step size for each particle. Below, we provide the analysis for the first choice. The proof and additional result for the second choice can be found in Sec. F.6.

Theorem 5 *Consider Alg. 1 with a global adaptive step size for normalized steepest descent ($\tau = 0$). Let $\{\boldsymbol{g}_k^i\}_{k\geq 0, i\in[N]}$, $\{\mu_k\}_{k\geq 0}$ be produced by Alg. 1. Define $\Delta_k := F(\mu_k) - F^\star$. Under Asm 2:*

$$\min_{0\le k<K}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_* \le \frac{\Delta_0\sqrt{S_{init}}}{K\eta}+\frac{L\eta}{\sqrt{K}}+\frac{2}{\sqrt{K}\eta}\left(\Delta_0+\frac{L}{2}\eta^2\log\Big(\frac{S_{init}+K}{S_{init}}\Big)\right).$$

4.2. Discrete time analysis: MFLD

Note that in the finite-particle setting, the Kullback–Leibler (KL) term is not well-defined, since the negative entropy is not meaningful for a discrete empirical measure. Therefore, we instead consider the joint distribution of N particles. Specifically, let $\mu^{(N)} \in \mathcal{P}^{(N)}$ denote the law of the particle system $\boldsymbol{X} = (\boldsymbol{x}^i)_{i=1}^N$, where $\mathcal{P}^{(N)}$ is the space of probability measures on $(\mathbb{R}^{d\times N}, \mathcal{B}(\mathbb{R}^{d\times N}))$.

The discrete-time Euler–Maruyama scheme is:

$$\boldsymbol{x}_{k+1}^i = \boldsymbol{x}_k^i - \eta_k \boldsymbol{v}_k^i + \sqrt{2\lambda\,\eta_k}\,\boldsymbol{\xi}_k^i, \quad \boldsymbol{\xi}_k^i \sim \mathcal{N}(0, \boldsymbol{I}_d)\,, \tag{4}$$

with step size $\eta_k > 0$. The term $\boldsymbol{v}_k^i$ represents the drift direction. In previous works [43], $\boldsymbol{v}_k^i$ has been analyzed as the full gradient $\nabla\frac{\delta F}{\delta\mu}(\mu_k)(\boldsymbol{x}_k^i)$, as a stochastic approximation computed on random mini-batches, or as a variance-reduced estimator. Here, for the theoretical analysis in discrete time, we will focus on the following:

$$\boldsymbol{v}_k^i = -\operatorname{lmo}\left(\nabla\frac{\delta F}{\delta\mu}(\mu_k)(\boldsymbol{x}_k^i)+\lambda\nabla_i\log\mu_k^{(N)}(\boldsymbol{X}_k);0\right)-\lambda\nabla_i\log\mu_k^{(N)}(\boldsymbol{X}_k)\,. \tag{5}$$

Similarly, when $\tau = 1$ with $\|\cdot\|$ being ℓ_2 norm, $\boldsymbol{v}_k^i$ reduced to the full gradient $\nabla\frac{\delta F}{\delta\mu}(\mu_k)(\boldsymbol{x}_k^i)$. The motivation for such a choice is that we can rewrite the velocity field in the LMO Wasserstein gradient flow Eq. (3) as follows by adding and subtracting the score term:

$$\boldsymbol{v}_t = -\operatorname{lmo}\left(\nabla\tfrac{\delta F}{\delta\mu}(\mu_t)+\lambda\nabla\log\mu_t;\,0\right)-\lambda\nabla\log\mu_t+\lambda\nabla\log\mu_t\,.$$

Therefore, Eq. (5) can be considered as the finite-particle Euler–Maruyama scheme for Eq. (3), with the last term $\lambda\nabla\log\mu_t$ corresponding to the noise in Eq. (4). In Sec. E, we present a convergence analysis of LMO noisy particle descent and establish a discrete-time analogue of the finite-time convergence results derived for the continuous flows in Thm. 4 and 7. The precise statement and proof are provided in Thm. 10 and Sec. E.

5. Conclusion

This paper establishes convergence guarantees for two-layer neural networks using LMO Wasserstein gradient flows. This paper also provides a convergence rate for a practical finite-particle scheme as well as adaptive particle methods in a discrete time setting. Interesting open directions include extending to deep neural networks, weakening the uniform log-Sobolev and the non-highly concentrated gradient assumptions, and extending the theory to mirror update. Furthermore, developing a convergence analysis for adaptive particle-based LMO dynamics under both smoothness and log-Sobolev assumptions remains largely unexplored, particularly for entropy-regularized objectives.

References

[1] Radosław Adamczak, Witold Bednorz, and Paweł Wolff. Moment estimates implied by modified log-sobolev inequalities. *ESAIM: Probability and Statistics*, 21:467–494, 2017.

[2] Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

[3] Amit Attia and Tomer Koren. Sgd with adagrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pages 1147–1171. PMLR, 2023.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pages 560–569. PMLR, 2018.

[5] Sergey G Bobkov and Michel Ledoux. From brunn-minkowski to brascamp-lieb and to logarithmic sobolev inequalities. *Geometric and Functional Analysis*, 10(5):1028–1052, 2000.

[6] Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in wasserstein space. *Advances in Neural Information Processing Systems*, 37:25311–25374, 2024.

[7] Siwan Boufadène 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.

[8] David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic Spectral Descent for Restricted Boltzmann Machines. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

[9] David Carlson, Ya-Ping Hsieh, Edo Collins, Lawrence Carin, and Volkan Cevher. Stochastic spectral descent for discrete graphical models. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):296–311, 2015.

[10] David E Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28, 2015.

[11] Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 61, pages 2357–2404. Institut Henri Poincaré, 2025.

[12] Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[13] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1):487–532, 2022.

[14] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[15] Lénaïc Chizat, Maria Colombo, and Xavier Fernández-Real. Convergence of drift-diffusion pdes arising as wasserstein gradient flows of convex functions. *arXiv preprint arXiv:2507.12385*, 2025.

[16] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[17] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[18] Kaja Gruntkowska, Alexander Gaponov, Zhirayr Tovmasyan, and Peter Richtárik. Error feedback for muon and friends. *arXiv preprint arXiv:2510.00643*, 2025.

[19] Anming Gu and Juno Kim. Mirror mean-field langevin dynamics. *arXiv preprint arXiv:2505.02621*, 2025.

[20] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28, 2015.

[21] Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, pages 2043–2065. Institut Henri Poincaré, 2021.

[22] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[23] Diederik P Kingma. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[24] Marcel Kreuter. Sobolev spaces of vector-valued functions. *Ulm University Faculty of Mathematics and Economies*, 2015.

[25] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

[26] Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

[27] Fanghui Liu, Luca Viano, and Volkan Cevher. Understanding deep neural function approximation in reinforcement learning via epsilon-greedy exploration. *Advances in Neural Information Processing Systems*, 35:5093–5108, 2022.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[29] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[30] Atsushi Nitanda. Improved particle approximation error for mean field neural networks. *Advances in Neural Information Processing Systems*, 37:113823–113845, 2024.

[31] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

[32] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

[33] Atsushi Nitanda, Anzelle Lee, Damian Tan Xing Kai, Mizuki Sakaguchi, and Taiji Suzuki. Propagation of chaos for mean-field langevin dynamics and its application to model ensemble. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=SjT6JK4KZv.

[34] Jorge Alberto Paz Moyado, Yamilet del Carmen Quintana Mato, José Manuel Rodríguez García, and José María Sigarreta Almira. New reverse hölder-type inequalities and applications. 2023.

[35] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Forty-second International Conference on Machine Learning*, 2025.

[36] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[37] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *International Conference on Machine Learning*, 2019.

[38] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.

[39] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[40] Louis Sharrock and Christopher Nemeth. Tuning-free sampling via optimization on the space of probability measures. *arXiv preprint arXiv:2510.25315*, 2025.

[41] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[42] Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[43] Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. *Advances in Neural Information Processing Systems*, 36:15545–15577, 2023.

[44] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 2006.

[45] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

[46] Jan Vondrák. Submodularity and curvature: The optimal algorithm (combinatorial optimization and discrete algorithms). 23:253–266, 2010.

[47] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR, 2023.

[48] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37:131786–131823, 2024.

[49] Shuailong Zhu and Xiaohui Chen. Convergence analysis of the wasserstein proximal algorithm beyond geodesic convexity. *arXiv preprint arXiv:2501.14993*, 2025.

Acknowledgments

We thank the workshop committee and reviewers for their work. This work was supported under project ID # 37 as part of the Swiss AI Initiative, through a grant from the ETH Domain and computational resources provided by the Swiss National Supercomputing Centre (CSCS) under the Alps infrastructure. This work was funded by the Swiss National Science Foundation (SNSF) under grant number 2000-1-240094. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048. TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015). This research is supported by the National Research Foundation, Singapore and the Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and the Ministry of Digital Development and Information.

Contents of the Appendix

The Appendix is organized as follows:

- In Sec. A, we summarize the main symbols and notation used throughout the paper.
- In Sec. B, we provide further detail on related work.
- In Sec. C, we provide theorems and proofs on LMO gradient flow in Euclidean space.
- In Sec. D and E, we provide the theoretical analysis and the detailed proofs for the LMO-based mean-field Langevin dynamics, including the proof for continuous time in Sec. D, and discrete setting in Sec. E.
- In Sec. F, we present the full technical proofs for our results on the optimization in measure space in the noiseless setting ($\lambda = 0$) and an additional theorem on adaptive LMO descent, and the corresponding proof.
- The ituition on the adaptive step size design is given at Sec. G.
- In Sec. H, we study the annealed dynamics in the mean-field Langevin dynamics setting.
- We provide additional experiments with varying dimension, and comparison between global adaptive stepsize and per-particle adaptive stepsize in Sec. I.
- We discuss the broader impact of this work in Sec. J.

Appendix A. Notation

A detailed summary of the notation can be found at Tab. 1.

The definition of the first variation of a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is as follows:

Definition 6 *For a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, its first variation $\frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x})$ is defined by*

$$\lim_{\varepsilon \to 0} \frac{F((1-\varepsilon)\mu + \varepsilon\nu) - F(\mu)}{\varepsilon} = \int \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) \mathrm{d}(\nu - \mu)(\boldsymbol{x}),$$

for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

Appendix B. Further detail on related work

B.1. Steepest descent

Steepest descent was originally studied for unconstrained smooth optimization problems in $\mathbb{R}^d$. Early work by Carlson et al. [8, 9, 10] showed that the smoothness assumption of the objective function implies a local majorization bound that naturally induces a notion of steepest descent in possibly non-Euclidean geometries. This perspective yields unnormalized steepest descent directions. Spectral descent arises as a special case when the underlying norm is the spectral norm. Building on this idea, Jordan et al. [22] proposed Muon, which applies a normalized steepest descent update by orthogonalizing the momentum under a spectral norm constraint. More recently, Pethick et al. [35]

introduced Scion, providing a general framework for training neural networks using norm-constrained LMOs, including guidelines for norm choices, proper scalings across layers, and convergence guarantees. Several further modifications and extensions of Muon have since been proposed [2, 18]. On the other hand, Vondrák [46] analyzed a continuous greedy algorithm based on linear optimization over a matroid polytope, whereas we study normalized descent dynamics induced by norm-based oracles in Euclidean and Wasserstein spaces.

B.2. Wasserstein gradient flows & MFLD

A growing body of work considers the training of overparameterized models at the level of probability measures, by viewing infinite-width two layer neural networks as minimizing a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ in Wasserstein space, see Eq. (1). [14, 31, 42]. A neuron with parameter $\boldsymbol{x}$ produces output $h(\boldsymbol{x}, \boldsymbol{z})$ for input $\boldsymbol{z}$, and an infinitely wide network is represented by a probability distribution μ over parameters, with predicted output $f_\mu(\boldsymbol{z}) = \int h(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\mu(\boldsymbol{x})$. Given a data distribution $\mathcal{D}$ and a loss function ℓ, the energy functional is defined as the population risk $F(\mu) = \mathbb{E}_{(\boldsymbol{z},y)\sim\mathcal{D}}\big[\ell\big(\int h(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\mu(\boldsymbol{x}), y\big)\big]$. A finite-width network with parameters $(\boldsymbol{x}^i)_{i=1}^N$ corresponds to the empirical measure $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_i}$, in which case the predicted output becomes $f_{\mu_N}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^N h(\boldsymbol{x}_i, \boldsymbol{z})$. Thus, μ_N serves as a finite-particle approximation of μ.

Previous works show that gradient descent on the finite network converges to a Wasserstein gradient flow in the mean-field limit, and use this perspective to obtain global convergence guarantees for mean-field neural networks [14, 37, 42]. Quantitative analysis with particle gradient descent is given in Chizat [13]. The extension to the mirror descent algorithm beyond gradient descent is presented in Bonet et al. [6]. In the entropy-regularized setting, the associated Wasserstein gradient flow becomes the MFLD. To minimize $\mathscr{F}$, a common approach is to consider the MFLD:

$$\mathrm{d}\boldsymbol{x}_t = -\nabla \frac{\delta F}{\delta \mu}(\mu_t)(\boldsymbol{x}_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}\boldsymbol{w}_t, \; \mu_t = \mathrm{Law}(\boldsymbol{x}_t)\,, \tag{6}$$

where $\boldsymbol{w}_t$ is a standard Brownian motion in $\mathbb{R}^d$. The law of a solution to MFLD solves the following PDE:

$$\partial_t \mu_t = \nabla \cdot \Big(\mu_t \nabla \tfrac{\delta F}{\delta \mu}(\mu_t)\Big) + \lambda \, \Delta \mu_t \,. \tag{7}$$

The convergence of MFLD to the optimum is proved in Hu et al. [21], Song et al. [42]. Later, under transport smoothness and uniform log-Sobolev conditions, one can obtain linear convergence of μ_t toward the minimizer $\mu^\star$ of $\mathscr{F}$ [12, 32].

To relate this mean-field limit to a practical finite model, one considers noisy particle gradient descent. For a finite particle system, given $\boldsymbol{X} = (\boldsymbol{x}^i)_{i=1}^N \in \mathbb{R}^{Nd}$, denote by its associated empirical measure as $\mu_{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}^i}$, then a single iteration is:

$$\boldsymbol{x}_{k+1}^i = \boldsymbol{x}_k^i - \eta \nabla \frac{\delta F}{\delta \mu}(\mu_{\boldsymbol{X}_k})(\boldsymbol{x}_k^i) + \sqrt{2\lambda\eta}\, \boldsymbol{\xi}_k^i, \; i \in [N], \tag{8}$$

where $\boldsymbol{\xi}_k^i \sim \mathcal{N}(0, I_d)$ are i.i.d. normal random variables. Passing to the limit $\eta \to 0$ yields the stochastic differential equation:

$$\mathrm{d}\boldsymbol{x}_t^i = -\nabla \frac{\delta F}{\delta \mu}(\mu_{\boldsymbol{X}_t})(\boldsymbol{x}_t^i)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}\boldsymbol{w}_t^i, \; i \in [N]\,. \tag{9}$$

When the number of particles grows to infinity, the empirical distribution $\mu_{\boldsymbol{X}_t}$ converges to a deterministic measure μ_t satisfying the MFLD in Eq. (6). In this case, the trajectories of particles are given by i.i.d. samples from Eq. (6), referred to as propagation of chaos [44].

Recent works establish propagation-of-chaos guarantees for mean-field Langevin dynamics, showing that the finite-particle system converges to the ideal mean-field limit with quantitative error bounds [11, 43]. Refined analysis is given by Nitanda [30], Nitanda et al. [33], which removes the dependence on the log-Sobolev inequality constants and shows that the objectives scale as $\mathcal{O}(1/N)$ uniformly in time. Gu and Kim [19] further extends the idea of MFLD to mirror descent dynamics.

However, all previous results rely on the unnormalized Wasserstein gradient-flow structure, which does not cover the recent normalized steepest descent dynamics, so the existing convergence and propagation-of-chaos theories do not directly apply. In addition, Sharrock and Nemeth [40] studies adaptive step size for Wasserstein gradient descent in the geodesically convex setting, whereas our focus is on normalized and unnormalized steepest descent dynamics for nonconvex objectives.

Note that Eq. (9) has an equivalent vector form:

$$\mathrm{d}\boldsymbol{X}_t = -N\nabla F(\mu_{\boldsymbol{X}_t})\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}\boldsymbol{W}_t\,,$$

where $\boldsymbol{W}_t$ is a standard Brownian motion in $\mathbb{R}^{Nd}$. Classical results imply that $\mu_t^{(N)} = \mathrm{Law}(\boldsymbol{X}_t)$ converges to the invariant Gibbs measure $\frac{\mathrm{d}\mu_*^{(N)}}{\mathrm{d}\boldsymbol{X}}(\boldsymbol{X}) \propto \exp\bigl(-\frac{N}{\lambda}F(\mu_{\boldsymbol{X}})\bigr)$. This stationary distribution is the unique minimizer of the finite-particle free energy

$$\mathscr{F}^{(N)}(\mu^{(N)}) = N\,\mathbb{E}_{\boldsymbol{X}\sim\mu^{(N)}}[F(\mu_{\boldsymbol{X}})] + \lambda\,\mathrm{Ent}(\mu^{(N)})\,,$$

where $\mu^{(N)} \in \mathcal{P}_2(\mathbb{R}^{dN})$. A fundamental question is how well $\mu_*^{(N)}$ approximates the ideal law $(\mu^\star)^{\otimes N}$. Recent propagation-of-chaos results for MFLD [11, 43] prove

$$\frac{1}{N}\Bigl(\mathscr{F}^{(N)}(\mu_*^{(N)}) - \mathscr{F}(\mu^\star)\Bigr) = \mathcal{O}\Bigl(\frac{\lambda}{\zeta N}\Bigr)\,,$$

where ζ is the constant in the log-Sobolev inequality. The resulting bounds depend inversely on log-Sobolev constants, which can deteriorate exponentially in the regularization parameter, leading to pessimistic particle complexity estimates. To mitigate this issue, Nitanda [30], Nitanda et al. [33] develop refined propagation-of-chaos and particle-bias bounds for mean-field neural networks, showing that the particle approximation error in the objective scales as $\mathcal{O}(1/N)$ uniformly in time and removing the exponential sensitivity to log-Sobolev constants.

Appendix C. Analysis of LMO gradient flow in Euclidean space

C.1. LMO gradient flow in Euclidean space

We consider the unconstrained optimization problem $\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$, where $f:\mathbb{R}^d\to\mathbb{R}$ satisfies $f\in C^1(\mathbb{R}^d)$. Next, we state the following assumption, which is a generalized version of the strongly convex assumption with ℓ_2 norm in optimization literature [29].

Assumption 5 *We assume the objective $f:\mathbb{R}^d\to\mathbb{R}$ is α-strongly convex w.r.t. the norm $\|\cdot\|$, i.e., there exists $\alpha>0$ such that:*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{\alpha}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\,,\ \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d\,.$$

Note that the norm $\|\cdot\|$ can be ℓ_p ($1 \le p \le \infty$) or, more generally, any norm of interest.

Below, we define the gradient flow with LMO, which generalizes the classical gradient flow by replacing the direction $-\nabla f(\boldsymbol{x}_t)$ with an LMO-induced direction.

$$\mathrm{d}\boldsymbol{x}_t = \mathrm{lmo}\left(\nabla f(\boldsymbol{x}_t); \tau\right) \mathrm{d}t\,, \tag{10}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^d$ is the initialization of the optimization trajectory. In the case of using ℓ_2 norm with $\tau = 1$ in LMO, classical result [39] shows that under Asm 5, one gets: $f(\boldsymbol{x}_t) - f^\star = (f(\boldsymbol{x}_0) - f^\star)e^{-2\alpha t}$. As a comparison, we provide the following theorem in the case of $\tau = 0$ (i.e., normalized steepest descent).

Theorem 7 (Convergence of normalized LMO gradient flow) *Consider the gradient flow in Eq. (10) with any norm and $\tau = 0$ in LMO, under Asm 5 and assume the gradient flow solution exists, then we have the following convergence result:*

$$f(\boldsymbol{x}_t) - f^\star \le \left[\max\left\{\sqrt{f(\boldsymbol{x}_0) - f^\star} - \sqrt{\tfrac{\alpha}{2}}t, 0\right\}\right]^2.$$

The proof can be found in Sec. C.2. This shows that the trajectory reaches the optimum in finite time, $f(\boldsymbol{x}_t) = f^\star,\ \forall\, t \ge t^\star := \sqrt{\frac{2}{\alpha}\left(f(\boldsymbol{x}_0) - f^*\right)}$. Note that the assumption on the existence of the gradient flow solution can be verified by checking the local Lipschitz continuity of the corresponding LMO vector field and then applying the Picard–Lindelöf theorem [45]. Refer to Sec. C.3 for a concrete proof for the normalized gradient flow case.

C.2. Proof of Thm. 7

Proof [Proof of Thm. 7] First, we prove that strong convexity under a general norm $\|\cdot\|$ can still get a PL inequality: Strong convexity at $\boldsymbol{x}$ with $\boldsymbol{y} = \boldsymbol{x}^*$ gives

$$f^* \;\ge\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x}\rangle + \frac{\alpha}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2\,.$$

Rearranging, using the Cauchy–Schwarz inequality for dual norms, and Young's inequality:

$$f(\boldsymbol{x}) - f^* \;\le\; \|\nabla f(\boldsymbol{x})\|_* \|\boldsymbol{x} - \boldsymbol{x}^*\| - \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \le \frac{1}{2\alpha}\|\nabla f(\boldsymbol{x})\|_*^2 + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 - \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2\,.$$

As a result, we have:

$$f(\boldsymbol{x}) - f^* \;\le\; \frac{1}{2\alpha}\|\nabla f(\boldsymbol{x})\|_*^2. \tag{11}$$

Next, we will prove the convergence. Along any differentiable trajectory $\boldsymbol{x}(t)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}(t)) = \langle \nabla f(\boldsymbol{x}(t)), \dot{\boldsymbol{x}}(t)\rangle = \langle \nabla f(\boldsymbol{x}(t)), \mathrm{lmo}(\nabla f(\boldsymbol{x}(t)), 0)\rangle = -\|\nabla f(\boldsymbol{x}(t))\|_*. \tag{12}$$

where we use the equality $\langle \boldsymbol{g}, \mathrm{lmo}(\boldsymbol{g}, 0)\rangle = \min_{\|\boldsymbol{y}\|\le 1}\langle \boldsymbol{g}, \boldsymbol{y}\rangle = -\|\boldsymbol{g}\|_*$.
Define $G(t) := \sqrt{f(\boldsymbol{x}(t)) - f^*} \ge 0$. Combining Eq. (11) and Eq. (12) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}(t)) = \frac{\mathrm{d}}{\mathrm{d}t} G(t)^2 = 2G(t)\frac{\mathrm{d}}{\mathrm{d}t} G(t) \le -\sqrt{2\alpha}\, G(t)\,.$$

Therefore, when $G(t) > 0$, we can divide both sides by $2G(t)$ and obtain $\frac{\mathrm{d}}{\mathrm{d}t}G(t) \leq -\sqrt{\frac{\alpha}{2}}$. Integrating both sides gives

$$G(t) \leq G(0) - \sqrt{\frac{\alpha}{2}}\, t.$$

Hence $G(t)$ decreases linearly until it possibly reaches 0. Once $G(t) = 0$, i.e., $f(\boldsymbol{x}(t)) = f^*$, we have $\nabla f(\boldsymbol{x}(t)) = 0$, and from Eq. (12).

$$\frac{\mathrm{d}}{dt} f(\boldsymbol{x}(t)) = -\|\nabla f(\boldsymbol{x}(t))\|_* = 0\,,$$

so $G(t)$ remains constant at 0 thereafter.

Therefore, we get

$$f(\boldsymbol{x}(t)) - f^\star \leq \left[\max\left\{\sqrt{f(\boldsymbol{x}(0)) - f^\star} - \sqrt{\tfrac{\alpha}{2}}\, t,\ 0\right\}\right]^2.$$

■

C.3. Uniqueness and existence of normalized gradient flow

Theorem 8 (Uniqueness and existence of normalized gradient flow) *Under Asm 5, consider setting $\tau = 0$ with ℓ_2 norm, the LMO gradient flow reduces to the following normalized gradient flow*

$$\dot{\boldsymbol{x}}(t) = -\frac{\nabla f(\boldsymbol{x}(t))}{\|\nabla f(\boldsymbol{x}(t))\|_2}, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0\,, \tag{13}$$

with the convention that $\dot{\boldsymbol{x}}(t) = 0$ whenever $\nabla f(\boldsymbol{x}(t)) = 0$. And we assume $\nabla f(\boldsymbol{x}_0) \neq 0$ for a nontrivial trajectory. Then we have i) Local existence and uniqueness. There exists $T > 0$ and a unique trajectory $\boldsymbol{x} : [0, T] \to \mathbb{R}^d$ satisfying Eq. (13) with $\boldsymbol{x}(0) = \boldsymbol{x}_0$. Moreover, for sufficiently small T, we have $\nabla f(\boldsymbol{x}(t)) \neq 0$ for all $t \in [0, T]$, so uniqueness holds on $[0, T]$.

ii) Global existence. There exists a global absolutely continuous solution $x : [0, \infty) \to \mathbb{R}^d$ to Eq. (13), and we have $\|\boldsymbol{x}(t) - \boldsymbol{x}_0\| \leq t, \quad \forall t \geq 0$. Moreover, if a trajectory reaches $\boldsymbol{x}^\star$ with $\nabla f(\boldsymbol{x}^\star) = 0$, then multiple continuations of the solution may exist beyond that time.

Proof First we show the local Lipschitzness. Define $\phi : \mathbb{R}^d \setminus \{0\} \to \mathbb{R}^d$ by $\phi(\boldsymbol{g}) := \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$. We show that if $\|\boldsymbol{g}\|_2 \geq \varepsilon > 0$ and $\|\boldsymbol{h}\|_2 \geq \varepsilon > 0$, then

$$\|\phi(\boldsymbol{g}) - \phi(\boldsymbol{h})\|_2 \leq \frac{2}{\varepsilon}\|\boldsymbol{g} - \boldsymbol{h}\|_2. \tag{14}$$

Indeed, we have

$$\begin{aligned} \|\phi(\boldsymbol{g}) - \phi(\boldsymbol{h})\|_2 &= \left\| \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} - \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \right\|_2 \leq \left\| \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} - \frac{\boldsymbol{g}}{\|\boldsymbol{h}\|_2} \right\|_2 + \left\| \frac{\boldsymbol{g} - \boldsymbol{h}}{\|\boldsymbol{h}\|_2} \right\|_2 \\ &= \frac{|\|\boldsymbol{h}\|_2 - \|\boldsymbol{g}\|_2|}{\|\boldsymbol{h}\|_2} + \frac{\|\boldsymbol{g} - \boldsymbol{h}\|_2}{\|\boldsymbol{h}\|_2} \leq \frac{\|\boldsymbol{g} - \boldsymbol{h}\|_2}{\|\boldsymbol{h}\|_2} + \frac{\|\boldsymbol{g} - \boldsymbol{h}\|_2}{\|\boldsymbol{h}\|_2} \ \leq\ \frac{2}{\varepsilon}\,\|\boldsymbol{g} - \boldsymbol{h}\|_2\,. \end{aligned} \tag{15}$$

Fix $\hat{\boldsymbol{x}}$ with $\nabla F(\hat{\boldsymbol{x}}) \neq 0$ and set $\varepsilon := \frac{1}{2}\|\nabla F(\hat{\boldsymbol{x}})\|_2 > 0$. By continuity of ∇F, there exists $r > 0$ such that

$$\|\nabla f(\boldsymbol{x})\|_2 \geq \varepsilon, \qquad \forall \boldsymbol{x} \in B(\hat{\boldsymbol{x}}, r)\,. \tag{16}$$

For any $\boldsymbol{x}, \boldsymbol{y} \in B(\hat{\boldsymbol{x}}, r)$, apply Eq. (14) with $g = \nabla f(\boldsymbol{x})$ and $h = \nabla f(\boldsymbol{y})$:

$$\Big\| \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} - \frac{\nabla f(\boldsymbol{y})}{\|\nabla f(\boldsymbol{y})\|} \Big\|_2 \leq \frac{2}{\varepsilon}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \frac{2L}{\varepsilon}\|\boldsymbol{x} - \boldsymbol{y}\|_2,$$

where the last inequality uses the global Lipschitz property of ∇F based on Asm 5.A. This proves that if $\nabla F(\hat{\boldsymbol{x}}) \neq 0$, then there exists a neighborhood U of $\hat{\boldsymbol{x}}$ and a constant $L_U < \infty$ such that

$$\Big\| \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|_2} - \frac{\nabla f(\boldsymbol{y})}{\|\nabla f(\boldsymbol{y})\|_2} \Big\| \;\leq\; L_U\|\boldsymbol{x} - \boldsymbol{y}\|_2, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in U\,. \tag{17}$$

Next, we are ready to show the local existence and uniqueness. Given $\nabla f(\boldsymbol{x}_0) \neq 0$, in its neighborhood, we have shown local Lipschitz property. By the Picard–Lindelöf theorem [45], there exists $T > 0$ and a unique solution $\boldsymbol{x} : [0, T] \to \mathbb{R}^d$ to Eq. (13) with $\boldsymbol{x}(0) = \boldsymbol{x}_0$. For sufficiently small T, the trajectory remains in the region where Eq. (16) holds, so $\nabla f(\boldsymbol{x}(t)) \neq 0$ for all $t \in [0, T]$.

Next, clearly, we have $\| \frac{\nabla f(\boldsymbol{x}(t))}{\|\nabla f(\boldsymbol{x}(t))\|_2}\|_2 = 1$. Hence, any absolutely continuous solution satisfies

$$\|\boldsymbol{x}(t) - \boldsymbol{x}(0)\|_2 \leq \int_0^t \| \frac{\nabla f(\boldsymbol{x}(t))}{\|\nabla f(\boldsymbol{x}(t))\|_2}\|_2 \,\mathrm{d}s = t\,,$$

Because trajectories cannot escape to infinity in finite time, standard continuation arguments imply that any local solution can be extended globally for all $t \geq 0$.

Lastly, we discuss the possible non-uniqueness at stationary points. If $\nabla f(\boldsymbol{x}^\star) = 0$, then by convention we set $\dot{\boldsymbol{x}}(t) = \mathbf{0}$ at $\boldsymbol{x}^\star$. However, the right-hand side of Eq. (13) is not locally Lipschitz in any neighborhood of $\boldsymbol{x}^\star$, since the normalization map $\boldsymbol{g} \mapsto \boldsymbol{g}/\|\boldsymbol{g}\|$ becomes singular at $\boldsymbol{g} = \mathbf{0}$. Consequently, the Picard–Lindelöf theorem cannot be applied at such points, and multiple absolutely continuous solutions may originate from the same initial condition $\boldsymbol{x}(0) = \boldsymbol{x}^\star$. ■

Appendix D. Proof for LMO-based MFLD in continuous time

D.1. Proof of Condition 1

Proof By the mass condition:

$$\|g_\mu\|_{L^1(\mu;\|\cdot\|_*)} = \int \|g_\mu(\boldsymbol{x})\|_* \,\mathrm{d}\mu(\boldsymbol{x}) \;\geq\; \int_{\{\|g_\mu(\boldsymbol{x})\|_* \geq m_1\}} \|g_\mu(\boldsymbol{x})\|_* \,\mathrm{d}\mu(\boldsymbol{x}) \;\geq\; m_1 m_2.$$

Using the uniform bound $\|g_\mu(\boldsymbol{x})\|_* \leq G_{\max}$ μ-a.s.,

$$\|g_\mu\|^2_{L^2(\mu;\|\cdot\|_*)} \;\leq\; \int G^2_{\max} \,\mathrm{d}\mu(\boldsymbol{x}) = G^2_{\max}\,.$$

Therefore, $\|g_\mu\|_{L^1(\mu;\|\cdot\|_*)} \;\geq\; m_1 m_2 \;\geq\; \frac{m_1 m_2}{G_{\max}}\|g_\mu\|_{L^2(\mu;\|\cdot\|_*)}$. This proves Asm 4 with $\kappa = m_1 m_2 / G_{\max}$.

The empirical case follows by replacing integrals with empirical averages: $\int f(\boldsymbol{x})\, d\hat{\mu}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}_i)$. ■

D.2. Connection of Asm 4 to reverse Hölder inequality

Lemma 1 (Reverse form of Hölder's inequality [34]) Let $1 < p \leq 2$ and let $q = \frac{p}{p-1}$ be its conjugate exponent. For any $g \in L^q(\mu)$ and $h \in L^p(\mu)$ with $\|h\|_{L^p(\mu)}, \|g\|_{L^q(\mu)} > 0$, then we have:

$$\|hg\|_{L^1(\mu)} \geq \|h\|_{L^p(\mu)} \|g\|_{L^q(\mu)} \left(1 - \frac{1}{p} \left\| \frac{|h|^{p/2}}{\|h\|_{L^p(\mu)}^{p/2}} - \frac{|g|^{q/2}}{\|g\|_{L^q(\mu)}^{q/2}} \right\|_{L^2(\mu)}^2 \right).$$

Remark 9 *Asm 4 can be viewed as a simplified consequence of the reverse form of Hölder's inequality (Lemma 1). Indeed, in the special case $p = q = 2$ and $h = 1$, Lemma 1 yields the quantitative bound $\|g\|_{L^1(\mu)} \geq \|g\|_{L^2(\mu)} \left(1 - \frac{1}{2} \int \left(1 - \frac{|g(\boldsymbol{x})|}{\|g\|_{L^2(\mu)}}\right)^2 \mathrm{d}\mu(\boldsymbol{x})\right)$. Hence, whenever $\left(1 - \frac{|g(\boldsymbol{x})|}{\|g\|_{L^2(\mu)}}\right)$ is smaller, then Asm 4 holds. This type of non-degeneracy condition is conceptually related to the coverage or exploration inequalities used in reinforcement learning analysis (e.g., Lemma 3 in Liu et al. [27]).*

D.3. Smoothness for two-layer mean-field neural networks

Proposition 1 (Smoothness for two-layer mean-field neural networks) Consider the two-layer mean-field neural network $f_\mu(\boldsymbol{z}) = \int h(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\mu(\boldsymbol{x})$, $F(\mu) = \mathbb{E}_{(\boldsymbol{z},y)\sim\mathcal{D}}\big[\ell(f_\mu(\boldsymbol{z}), y)\big]$. Assume that: (i) the neuron $h(\boldsymbol{x}, \boldsymbol{z})$ is differentiable in $\boldsymbol{x}$; (ii) $\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z})$ is bounded in $\|\cdot\|_*$ and Lipschitz continuous with respect to $\|\cdot\|$ and $\|\cdot\|_*$ (e.g., $h(\boldsymbol{x}, \boldsymbol{z}) = \sigma(\boldsymbol{x}^\top \boldsymbol{z})$ with σ being tanh or sigmoid activation); (iii) the loss $\ell(f, y)$ has Lipschitz continuous first derivative in f (e.g., logistic loss; squared loss under bounded outputs). Then Asm 2 holds.

Proof We first compute the first variation:

$$\frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) = \mathbb{E}_{(\boldsymbol{z},y)}\Big[\partial_f \ell(f_\mu(\boldsymbol{z}), y) \, h(\boldsymbol{x}, \boldsymbol{z})\Big].$$

Taking gradient with respect to $\boldsymbol{x}$ gives

$$\nabla \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) = \mathbb{E}_{(\boldsymbol{z},y)}\Big[\partial_f \ell(f_\mu(\boldsymbol{z}), y) \, \nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z})\Big].$$

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$. Using the convexity of the dual norm,

$$\left\|\nabla \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) - \nabla \frac{\delta F}{\delta \mu}(\nu)(\boldsymbol{x}')\right\|_* \leq \mathbb{E}_{(\boldsymbol{z},y)}\big[A_1(\boldsymbol{z}, y) + A_2(\boldsymbol{z}, y)\big],$$

where we define

$$A_1(\boldsymbol{z}, y) := \big|\partial_f \ell(f_\mu(\boldsymbol{z}), y) - \partial_f \ell(f_\nu(\boldsymbol{z}), y)\big| \cdot \|\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z})\|_*,$$
$$A_2(\boldsymbol{z}, y) := \big|\partial_f \ell(f_\nu(\boldsymbol{z}), y)\big| \cdot \|\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z}) - \nabla_{\boldsymbol{x}} h(\boldsymbol{x}', \boldsymbol{z})\|_*.$$

By assumption, there exist constants $B_1, B_2, H_1, H_2 > 0$ such that for all $\boldsymbol{z}, y, \boldsymbol{x}, \boldsymbol{x}'$:

$$\begin{aligned}
&|\partial_f \ell(f, y)| \le B_1, \\
&|\partial_f \ell(f_1, y) - \partial_f \ell(f_2, y)| \le B_2 |f_1 - f_2|, \\
&\|\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z})\|_* \le H_1, \\
&\|\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z}) - \nabla_{\boldsymbol{x}} h(\boldsymbol{x}', \boldsymbol{z})\|_* \le H_2 \|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned}$$

Therefore,

$$A_1(\boldsymbol{z}, y) \le B_2 H_1 \, |f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})|, \qquad A_2(\boldsymbol{z}, y) \le B_1 H_2 \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

Taking expectation gives

$$\left\| \nabla \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) - \nabla \frac{\delta F}{\delta \mu}(\nu)(\boldsymbol{x}') \right\|_* \le B_2 H_1 \, \mathbb{E}_{\boldsymbol{z}} |f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})| + B_1 H_2 \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

It remains to bound $\mathbb{E}_{\boldsymbol{z}} |f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})|$. For any coupling π between μ and ν,

$$\begin{aligned}
|f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})| &= \left| \int h(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\mu(\boldsymbol{x}) - \int h(\boldsymbol{x}', \boldsymbol{z}) \, \mathrm{d}\nu(\boldsymbol{x}') \right| \\
&\le \int |h(\boldsymbol{x}, \boldsymbol{z}) - h(\boldsymbol{x}', \boldsymbol{z})| \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{x}').
\end{aligned}$$

Since $\|\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{z})\|_* \le H_1$, the map $\boldsymbol{x} \mapsto h(\boldsymbol{x}, \boldsymbol{z})$ is H_1-Lipschitz w.r.t. $\|\cdot\|$, hence

$$|h(\boldsymbol{x}, \boldsymbol{z}) - h(\boldsymbol{x}', \boldsymbol{z})| \le H_1 \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

Therefore,

$$|f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})| \le H_1 \int \|\boldsymbol{x} - \boldsymbol{x}'\| \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{x}'),$$

and taking the infimum over all couplings gives

$$|f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})| \le H_1 W_1(\mu, \nu) \le H_1 W_2(\mu, \nu).$$

Taking expectation over $\boldsymbol{z}$ yields

$$\mathbb{E}_{\boldsymbol{z}} |f_\mu(\boldsymbol{z}) - f_\nu(\boldsymbol{z})| \le H_1 W_2(\mu, \nu).$$

Substituting back, we obtain

$$\begin{aligned}
\left\| \nabla \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}) - \nabla \frac{\delta F}{\delta \mu}(\nu)(\boldsymbol{x}') \right\|_* &\le B_2 H_1^2 W_2(\mu, \nu) + B_1 H_2 \|\boldsymbol{x} - \boldsymbol{x}'\|. \\
&\le \frac{L}{2} \big(W_2(\mu, \nu) + \|\boldsymbol{x} - \boldsymbol{x}'\| \big),
\end{aligned} \tag{18}$$

with

$$L = 2 \max\{B_2 H_1^2, \; B_1 H_2\}.$$

This proves the claim. ∎

D.4. Proof of Thm. 4

Proof We first present the following lemma that is crucial in our analysis.

Lemma 2 (Entropy Sandwich [12]) Under Asm. 2 and 3 and assuming that $F(\mu)$ is convex w.r.t μ, let μ^* be the unique minimizer of $\mathscr{F}$. For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, given the definition of ν in Asm 3, the following inequality holds:

$$\mathscr{F}(\mu) - \mathscr{F}(\mu^*) \;\le\; \lambda\, \mathrm{H}(\mu \,\|\, \nu)\,. \tag{19}$$

Now we start our proof. Define the suboptimality $\mathcal{E}(t) := \mathscr{F}(\mu_t) - \mathscr{F}(\mu^\star) \ge 0$. Since

$$\partial_t \mu_t + \nabla\cdot(\mu_t v_t) = 0,$$

starting from the chain rule in Wasserstein space and the definition of v_t gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) \;&=\; \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}(\mu_t) \\
&=\; \int \big\langle \nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}),\, v_t(\boldsymbol{x}) \big\rangle \,\mathrm{d}\mu_t(\boldsymbol{x}) \\
&=\; \int \big\langle \nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}),\, \mathrm{lmo}\big(\nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}); 0\big) \big\rangle \,\mathrm{d}\mu_t(\boldsymbol{x}) \\
&=\; -\int \Big\| \nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}) \Big\|_* \,\mathrm{d}\mu_t(\boldsymbol{x}) \\
&\le\; -\kappa \left(\int \Big\| \nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}) \Big\|_*^2 \,\mathrm{d}\mu_t(\boldsymbol{x}) \right)^{1/2},
\end{aligned} \tag{20}$$

where in the second last step, we use Lemma 7 and the last step follows from Asm 4.

According to the "entropy sandwich" lemma (Lemma 2) and the definition of v_t,

$$\mathcal{E}(t) \;\le\; \lambda\, H(\mu_t \,\|\, \nu_t), \qquad \nu_t \;\propto\; \exp\Big(-\tfrac{1}{\lambda} \tfrac{\delta F}{\delta \mu}(\mu_t) \Big)\,, \tag{21}$$

we have:

$$\nabla_{\boldsymbol{x}} \log \frac{\mu_t}{\nu_t} \;=\; \nabla_{\boldsymbol{x}} \log \mu_t + \frac{1}{\lambda} \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu_t) \;=\; \frac{1}{\lambda} \nabla_{\boldsymbol{x}} \frac{\delta \mathscr{F}}{\delta \mu}(\mu_t)\,. \tag{22}$$

Next, by the uniform log-Sobolev inequality (Asm 3), we get

$$\frac{\mathcal{E}(t)}{\lambda} \;\le\; H(\mu_t \,\|\, \nu_t) \;\le\; \frac{1}{2\zeta\, \lambda^2} \int \Big\| \nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}) \Big\|_*^2 \,\mathrm{d}\mu_t(\boldsymbol{x})\,. \tag{23}$$

From Eq. (20) and Eq. (23), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) \;\le\; -\kappa \left(\int \Big\| \nabla_{\boldsymbol{x}} \tfrac{\delta \mathscr{F}}{\delta \mu}(\mu_t)(\boldsymbol{x}) \Big\|_*^2 \,\mathrm{d}\mu_t(\boldsymbol{x}) \right)^{1/2} \le -\kappa\, (2\zeta\, \lambda\, \mathcal{E}(t))^{1/2}\,. \tag{24}$$

Let $G(t) := \sqrt{\mathcal{E}(t)}$. Then $\frac{\mathrm{d}}{\mathrm{d}t} G(t) \le -\frac{\kappa}{2}\sqrt{2\zeta\,\lambda}$ so that $G(t) \le \max\{G(0) - \frac{\kappa}{2}\sqrt{2\,\zeta\,\lambda} t,\, 0\}$. Taking the square completes the proof. ∎

Algorithm 2: LMO Noisy Particle Descent (LMO-NPD)

Data: number of particles N; step sizes $\{\eta_k\}$; diffusion level $\lambda > 0$; iterations K
Result: empirical measure $\mu_K^{(N)} = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_K^i}$
Initialize particles $\boldsymbol{x}_0^i \sim \mu_0$ for $i = 1, \dots, N$;
for $k = 0$ **to** $K-1$ **do**
  $\mu_k^{(N)} \leftarrow \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_k^i}$;
  for $i = 1$ **to** N **do**
    Evaluate mean-field force $\boldsymbol{v}_k^i$ according to Eq. (87) or Eq. (5);
    Sample Gaussian noise $\boldsymbol{\xi}_k^i \sim \mathcal{N}(0, I_d)$;
    $\boldsymbol{x}_{k+1}^i \leftarrow \boldsymbol{x}_k^i - \eta_k \boldsymbol{v}_k^i + \sqrt{2\lambda\eta_k}\,\boldsymbol{\xi}_k^i$;
  end
end

Appendix E. Discrete time analysis of LMO noisy particle descent

In this section, we consider the empirical risk minimization problem of mean-field neural networks. Throughout, the norm $\|\cdot\|$ and $\|\cdot\|_*$ on $\mathbb{R}^{dN}$ are understood as the sum of the corresponding norms over N blocks. Let $h(\boldsymbol{x}, \cdot) : \mathcal{Z} \to \mathbb{R}$ be the neuron with parameter $\boldsymbol{x} \in \mathbb{R}^d$, where $\mathcal{Z}$ is the data space, and let $F(\mu) = \frac{1}{n}\sum_{j=1}^N \ell\big(\mathbb{E}_{\boldsymbol{x}\sim\mu}[h(\boldsymbol{x}, \boldsymbol{z}_j)], y_j\big)$, where $(\boldsymbol{z}_i, y_i)_{i=1}^N$ are training data.

We then define the following objective on $\mathcal{P}^{(N)}$:

$$\mathscr{F}^N(\mu^{(N)}) = N\,\mathbb{E}_{\boldsymbol{X}\sim\mu^{(N)}}\big[F(\mu_{\boldsymbol{X}})\big] + \lambda\,\mathrm{Ent}(\mu^{(N)})\,.$$

Below, we introduce the assumption on the loss and the neuron, which has been used in Nitanda [30].

Assumption 6 (From Nitanda [30])

- *The loss $\ell(\cdot, y)$ is convex and $\hat{L}$-smooth.*
- *There exists $\hat{R} > 0$ such that, for all $\boldsymbol{z} \in \mathcal{Z}$, we have $\mathbb{E}_{\boldsymbol{x}\sim\mu^\star}\big[|h(\boldsymbol{x}, \boldsymbol{z})|^2\big] \le \hat{R}^2$.*

Below, we provide an assumption that is similar to Asm 2 and similar to the bounded gradient assumption.

Assumption 7 *There exists a constant $M_1 > 0$ such that $\big\|\boldsymbol{v}_k^i\big\|_2 \le M_1, \forall k \ge 0, i \in [N]$. Fix a norm $\|\cdot\|$ on $\mathbb{R}^d$ with dual $\|\cdot\|_*$. There exists a constant $M_2 > 0$ s.t. $\forall \mu, \nu$ on $(\mathbb{R}^d)^N$ and $\forall \boldsymbol{Y}, \boldsymbol{X} \in (\mathbb{R}^d)^N$:*

$$\big\|\nabla \log \mu(\boldsymbol{Y}) - \nabla \log \nu(\boldsymbol{X})\big\|_* \le M_2\Big(W_2(\mu, \nu) + \|\boldsymbol{Y} - \boldsymbol{X}\|\Big)\,.$$

Now we are ready to provide the convergence theorem for Alg. 2 with proof deferred to Sec. E.1.

Theorem 10 (Convergence of LMO-NPD) *Let $\{\mu_k^{(N)}\}_{k\ge0}$ be the measure generated by Alg. 2 with drift selected by Eq. (87). Define: $C_1 = \sup_{\boldsymbol{a}\in\mathbb{R}^{dN}\setminus\{0\}} \frac{\|\boldsymbol{a}\|^2}{\|\boldsymbol{a}\|_*^2}$. and $C_2 := \frac{8}{\kappa\sqrt{2\zeta\lambda}}\Big(LN +$*

$\lambda M_2 N + \lambda M_2 \sqrt{N}\Big) \sqrt{2\eta\big(M_1^2 + 2\lambda\,\mathbb{E}\|\xi\|^2\big)}$. *Consider a fixed step size* η *that satisfies:*

$$0 < \eta \le \min\left\{1,\ \frac{\kappa^2}{32\,\lambda^2\,M_2^2\,C_1\,N\,\big(M_1^2 + 2\lambda\,\mathbb{E}\|\xi\|^2\big)}\right\}.$$

Then, under Asm. 1 to 4, 6 and 7, for all $k \ge 0$*,*

$$\begin{aligned}&\frac{1}{N}\,\mathscr{F}^{(N)}(\mu_k^{(N)}) - \mathscr{F}(\mu^\star)\\ &\le \frac{\hat{L}\hat{R}^2}{2N} + \frac{1}{N}\left[\max\left\{\sqrt{\mathscr{F}^{(N)}(\mu_0^{(N)}) - \mathscr{F}^{(N)}(\mu_*^{(N)})} - \frac{\kappa}{8}\sqrt{2\zeta\lambda}\,k\eta,\ C_2\right\}\right]^2.\end{aligned}$$

E.1. Proof for Thm. 10

E.1.1. AUXILIARY LEMMAS

Lemma 3 (Theorem 1 in Nitanda [30]) Let $\mu^\star$ be the minimizer of $\mathscr{F}$ and $\mu_*^{(N)}$ the minimizer of $\mathscr{F}^{(N)}$, and denote by $(\mu^\star)^{\otimes N}$ the N-fold product of $\mu^\star$. Under Asm. 1, 2 and 6, we have:

$$\frac{1}{N}\mathscr{F}^{(N)}(\mu_*^{(N)}) - \mathscr{F}(\mu^\star) \ \le\ \frac{\hat{L}\hat{R}^2}{2N}.$$

Lemma 4 (LMO mismatch inequality on the unit ball) Let $(\mathbb{R}^d, \|\cdot\|)$ be a normed space with dual norm $\|\cdot\|_*$. Then for all $\boldsymbol{h}, \boldsymbol{g} \in \mathbb{R}^d$,

$$0 \ \le\ -\big\langle \boldsymbol{h},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle \ \le\ 2\,\|\boldsymbol{h}-\boldsymbol{g}\|_*.$$

Proof By optimality of $\mathrm{lmo}(\boldsymbol{h};0)$ for $\boldsymbol{h}$ over $\{\boldsymbol{s} : \|\boldsymbol{s}\| \le 1\}$,

$$\langle \boldsymbol{h}, \mathrm{lmo}(\boldsymbol{h};0)\rangle \le \langle \boldsymbol{h}, \mathrm{lmo}(\boldsymbol{g};0)\rangle,$$

hence

$$-\big\langle \boldsymbol{h},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle \ge 0.$$

By optimality of $\mathrm{lmo}(\boldsymbol{g};0)$ for $\boldsymbol{g}$ over $\{\boldsymbol{s} : \|\boldsymbol{s}\| \le 1\}$,

$$\langle \boldsymbol{g}, \mathrm{lmo}(\boldsymbol{g};0)\rangle \le \langle \boldsymbol{g}, \mathrm{lmo}(\boldsymbol{h};0)\rangle,$$

so

$$\big\langle \boldsymbol{g},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle \ge 0.$$

Therefore,

$$\begin{aligned}&-\big\langle \boldsymbol{h},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle\\ &= -\big\langle \boldsymbol{h}-\boldsymbol{g},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle - \big\langle \boldsymbol{g},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle\\ &\le -\big\langle \boldsymbol{h}-\boldsymbol{g},\ \mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\rangle\\ &\le \|\boldsymbol{h}-\boldsymbol{g}\|_*\,\big\|\mathrm{lmo}(\boldsymbol{h};0) - \mathrm{lmo}(\boldsymbol{g};0)\big\|,\end{aligned}$$

where the last line uses Hölder's inequality. Finally, by the triangle inequality,

$$\big\|\mathrm{lmo}(\boldsymbol{h};0)-\mathrm{lmo}(\boldsymbol{g};0)\big\| \le \big\|\mathrm{lmo}(\boldsymbol{h};0)\big\| + \big\|\mathrm{lmo}(\boldsymbol{g};0)\big\| \le 2,$$

and hence

$$-\big\langle \boldsymbol{h},\ \mathrm{lmo}(\boldsymbol{h};0)-\mathrm{lmo}(\boldsymbol{g};0)\big\rangle \le 2\,\|\boldsymbol{h}-\boldsymbol{g}\|_* .$$

■

Lemma 5 Under the same setting and assumption as in Thm. 10, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}^{(N)}(\nu_t) = -\int \nu_t(\boldsymbol{Y})\,\|G_t(\boldsymbol{Y})\|_2^2\,\mathrm{d}\boldsymbol{Y} - \int \nu_t(\boldsymbol{Y})\,G_t(\boldsymbol{Y})\delta_t(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}.$$

where

$$\delta_t(\boldsymbol{Y}) := \mathbb{E}[-\mathrm{lmo}(G_0(\boldsymbol{Y}_0);0) - \lambda\nabla\log\nu_0(\boldsymbol{Y}_0)\,|\,\boldsymbol{Y}_t=\boldsymbol{Y}] - N\nabla F(\boldsymbol{Y}). \tag{25}$$

$$G_t(\boldsymbol{Y}) := \nabla\frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y}) = N\nabla F(\boldsymbol{Y}) + \lambda\nabla\log\nu_t(\boldsymbol{Y}) = \lambda\nabla\log\frac{\nu_t}{\mu_*^{(N)}}(\boldsymbol{Y}). \tag{26}$$

Proof The proof follows the idea in Nitanda [30]. We study the discrete update

$$\boldsymbol{x}_{k+1}^i = \boldsymbol{x}_k^i - \eta\,\boldsymbol{v}_k^i + \sqrt{2\lambda\eta}\,\boldsymbol{\xi}_k^i, \qquad i\in\{1,\dots,N\},$$

where $\boldsymbol{\xi}_k^i \sim \mathcal{N}(0,\boldsymbol{I}_d)$ are i.i.d. Gaussian noises. Let $\mathbf{X}_k = (\boldsymbol{x}_k^1,\dots,\boldsymbol{x}_k^N)$ and $\mu_k^{(N)} = \mathrm{Law}(\mathbf{X}_k)$.

To analyze one step, we introduce the one-step interpolation process

$$\mathrm{d}\boldsymbol{y}_t^i = -\boldsymbol{v}_k^i\,\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}\boldsymbol{w}_t^i, \qquad t\in[0,\eta],$$

with $\boldsymbol{Y}_0 = \mathbf{X}_k$. Its explicit solution is

$$\boldsymbol{y}_t^i = \boldsymbol{y}_0^i - t\,\boldsymbol{v}_k^i + \sqrt{2\lambda t}\,\boldsymbol{\xi}^i, \qquad \boldsymbol{\xi}^i\sim\mathcal{N}(0,I_d),$$

so that $\boldsymbol{Y}_\eta = \mathbf{X}_{k+1}$. We denote $\nu_t := \mathrm{Law}(\boldsymbol{Y}_t)$, hence $\nu_0 = \mu_k^{(N)}$ and $\nu_\eta = \mu_{k+1}^{(N)}$.

For notational simplicity we identify measures with their densities with respect to the Lebesgue measure and write $\mu_*^{(N)}(\boldsymbol{Y})$ for the density of the target Gibbs measure. Let $\nu_{0t}(\boldsymbol{Y}_0,\boldsymbol{Y})$ be the joint law of $(\boldsymbol{Y}_0,\boldsymbol{Y}_t)$, and $\nu_{t|0}$, $\nu_{0|t}$ the corresponding conditional distributions. Then

$$\nu_{0t}(\boldsymbol{Y}_0,\boldsymbol{Y}) = \nu_0(\boldsymbol{Y}_0)\,\nu_{t|0}(\boldsymbol{Y}\,|\,\boldsymbol{Y}_0) = \nu_t(\boldsymbol{Y})\,\nu_{0|t}(\boldsymbol{Y}_0\,|\,\boldsymbol{Y}).$$

The conditional density satisfies

$$\frac{\partial}{\partial t}\nu_{t|0}(\boldsymbol{Y}\,|\,\boldsymbol{Y}_0) = \nabla\cdot\big(\nu_{t|0}(\boldsymbol{Y}\,|\,\boldsymbol{Y}_0)\,\boldsymbol{V}_k\big) + \lambda\,\Delta\nu_{t|0}(\boldsymbol{Y}\,|\,\boldsymbol{Y}_0),$$

where $\boldsymbol{V}_k := (\boldsymbol{v}_k^1,\dots,\boldsymbol{v}_k^N)$ is constant in t, i.e.,

$$\begin{aligned}\boldsymbol{V}_k(\boldsymbol{Y}_0) &= -\mathrm{lmo}\big(N\nabla F(\boldsymbol{Y}_0)+\lambda\nabla\log\nu_0(\boldsymbol{Y}_0);0\big) - \lambda\nabla\log\nu_0(\boldsymbol{Y}_0)\\ &= -\mathrm{lmo}(G_0(\boldsymbol{Y}_0);0) - \lambda\nabla\log\nu_0(\boldsymbol{Y}_0)\end{aligned} \tag{27}$$

where the LMO acts blockwise over the N particles. And the last inequality is because that $\mu_*^{(N)}$ is with density proportional to $\exp\big(-\frac{N}{\lambda}F(\mu_{\boldsymbol{Y}})\big)$, we have $\lambda\,\nabla\log\mu_*^{(N)}(\boldsymbol{Y}) \;=\; -N\,\nabla F(\boldsymbol{Y})$, so we have (see Chen et al. [11] for the proof):

$$\nabla\frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y}) = N\nabla F(\boldsymbol{Y}) + \lambda\nabla\log\nu_t(\boldsymbol{Y}) = \lambda\nabla\log\frac{\nu_t}{\mu_*^{(N)}}(\boldsymbol{Y})\,. \tag{28}$$

Integrating against $\nu_0(\boldsymbol{Y}_0)$ and using $\nu_{0t} = \nu_0\nu_{t|0} = \nu_t\nu_{0|t}$, we get

$$\begin{aligned}
\frac{\partial}{\partial t}\,\nu_t(\boldsymbol{Y}) &= \int \frac{\partial}{\partial t}\,\nu_{t|0}(\boldsymbol{Y}\,|\,\boldsymbol{Y}_0)\,\nu_0(\boldsymbol{Y}_0)\,\mathrm{d}\boldsymbol{Y}_0 \\
&= \int \Big(\nabla\cdot\big(\nu_{0t}(\boldsymbol{Y}_0,\boldsymbol{Y})\,\boldsymbol{V}_k\big) \;+\; \lambda\,\Delta\nu_{0t}(\boldsymbol{Y}_0,\boldsymbol{Y})\Big)\,\mathrm{d}\boldsymbol{Y}_0 \\
&= \nabla\cdot\Big(\nu_t(\boldsymbol{Y})\,\underbrace{\int \nu_{0|t}(\boldsymbol{Y}_0\,|\,\boldsymbol{Y})\,\boldsymbol{V}_k(\boldsymbol{Y}_0)\,\mathrm{d}\boldsymbol{Y}_0}_{=\,\mathbb{E}[\boldsymbol{V}_k(\boldsymbol{Y}_0)\,|\,\boldsymbol{Y}_t=\boldsymbol{Y}]}\Big) \;+\; \lambda\,\Delta\nu_t(\boldsymbol{Y}).
\end{aligned}$$

By adding and subtracting yields

$$\frac{\partial}{\partial t}\,\nu_t(\boldsymbol{Y}) \;=\; \lambda\,\nabla\cdot\Big(\nu_t(\boldsymbol{Y})\,\nabla\log\frac{\nu_t}{\mu_*^{(N)}}(\boldsymbol{Y})\Big) \;+\; \nabla\cdot\Big(\nu_t(\boldsymbol{Y})\,\delta_t(\boldsymbol{Y})\Big),$$

where we define

$$\begin{aligned}
\delta_t(\boldsymbol{Y}) :&= \mathbb{E}[\boldsymbol{V}_k(\boldsymbol{Y}_0)\,|\,\boldsymbol{Y}_t=\boldsymbol{Y}] \;-\; N\,\nabla F(\boldsymbol{Y}) \\
&= \mathbb{E}[-\,\mathrm{lmo}(G_0(\boldsymbol{Y}_0);0) - \;\lambda\,\nabla\log\nu_0(\boldsymbol{Y}_0)\,|\,\boldsymbol{Y}_t=\boldsymbol{Y}] \;-\; N\,\nabla F(\boldsymbol{Y})
\end{aligned} \tag{29}$$

Differentiating along $t\mapsto\nu_t$ and integrating by parts, we obtain:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\,\mathscr{F}^{(N)}(\nu_t) \\
&= \int \frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y})\frac{\partial\nu_t}{\partial t}(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} \\
&= \lambda\int \frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y})\nabla\cdot\Big(\nu_t(\boldsymbol{Y})\,\nabla\log\tfrac{\nu_t}{\mu_*^{(N)}}(\boldsymbol{Y})\Big)\,\mathrm{d}\boldsymbol{Y} + \int \frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y})\nabla\cdot\Big(\nu_t(\boldsymbol{Y})\,\delta_t(\boldsymbol{Y})\Big)\,\mathrm{d}\boldsymbol{Y} \\
&= -\lambda\int \nu_t(\boldsymbol{Y})\,\nabla\frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y})^\top\nabla\log\frac{\nu_t}{\mu_*^{(N)}}(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} - \int \nu_t(\boldsymbol{Y})\,\nabla\frac{\delta\mathscr{F}^{(N)}(\nu_t)}{\delta\mu^{(N)}}(\boldsymbol{Y})^\top\delta_t(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}.
\end{aligned}$$

And the proof is finished by using Eq. (25). ■

Lemma 6 Under the same setting and assumptions as in Thm. 10, fix a norm $\|\cdot\|$ on $\mathbb{R}^d$. For all $t\in[0,\eta]$ we have

$$\mathbb{E}_{\nu_{0t}}\left[\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|^2\right] \;\le\; 2Nt^2M_1^2 \;+\; 4\lambda t\,N\,\mathbb{E}\|\boldsymbol{\xi}\|^2, \tag{30}$$

where $\boldsymbol{\xi}\sim\mathcal{N}(0,I_d)$. Moreover,

$$\mathbb{E}_{\nu_{0t}}\left[\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|\right] \;\le\; N\sqrt{2t^2M_1^2 + 4\lambda t\,\mathbb{E}\|\boldsymbol{\xi}\|^2}. \tag{31}$$

Proof Recall the interpolation SDE for $t \in [0, \eta]$:

$$\mathrm{d}\boldsymbol{y}_t^i = -\boldsymbol{v}_k^i \,\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}\boldsymbol{w}_t^i, \qquad \boldsymbol{y}_0^i = \boldsymbol{x}_k^i,$$

whose explicit solution is

$$\boldsymbol{y}_t^i = \boldsymbol{y}_0^i - t\boldsymbol{v}_k^i + \sqrt{2\lambda t}\,\boldsymbol{\xi}^i, \qquad \boldsymbol{\xi}^i \sim \mathcal{N}(0, \boldsymbol{I}_d) \text{ i.i.d.}$$

Hence

$$\boldsymbol{y}_t^i - \boldsymbol{y}_0^i = -\,t\boldsymbol{v}_k^i + \sqrt{2\lambda t}\,\boldsymbol{\xi}^i.$$

Using the elementary inequality (valid for any norm)

$$\|\boldsymbol{a} + \boldsymbol{b}\|^2 \le 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2,$$

we obtain

$$\begin{aligned}\|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|^2 &\le 2t^2\|\boldsymbol{v}_k^i\|^2 \;+\; 2\big\|\sqrt{2\lambda t}\,\boldsymbol{\xi}^i\big\|^2 \\ &= 2t^2\|\boldsymbol{v}_k^i\|^2 \;+\; 4\lambda t\,\|\boldsymbol{\xi}^i\|^2.\end{aligned}$$

Summing over $i = 1, \dots, N$,

$$\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|^2 \;\le\; 2t^2 \sum_{i=1}^N \|\boldsymbol{v}_k^i\|^2 \;+\; 4\lambda t \sum_{i=1}^N \|\boldsymbol{\xi}^i\|^2.$$

Taking expectation and using Asm 7 (i.e., $\|\boldsymbol{v}_k^i\| \le M_1$), as well as i.i.d. of $\boldsymbol{\xi}^i$, we get

$$\begin{aligned}\mathbb{E}_{\nu_{0t}}\left[\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|^2\right] &\le 2t^2\,\mathbb{E}\Big[\sum_{i=1}^N \|\boldsymbol{v}_k^i\|^2\Big] \;+\; 4\lambda t\,\mathbb{E}\Big[\sum_{i=1}^N \|\boldsymbol{\xi}^i\|^2\Big] \\ &\le 2Nt^2M_1^2 + 4\lambda t\,N\,\mathbb{E}\|\boldsymbol{\xi}\|^2,\end{aligned}$$

which proves Eq. (30).

For the first-moment bound, by Jensen's inequality and Cauchy–Schwarz $(\sum_{i=1}^N a_i)^2 \le N\sum_{i=1}^N a_i^2$ for $a_i \ge 0$, we have

$$\begin{aligned}\mathbb{E}_{\nu_{0t}}\left[\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|\right] &\le \left(\mathbb{E}_{\nu_{0t}}\left[\Big(\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|\Big)^2\right]\right)^{1/2} \\ &\le \left(N\,\mathbb{E}_{\nu_{0t}}\left[\sum_{i=1}^N \|\boldsymbol{y}_t^i - \boldsymbol{y}_0^i\|^2\right]\right)^{1/2} \\ &\le \Big(N\,(2Nt^2M_1^2 + 4\lambda t\,N\,\mathbb{E}\|\boldsymbol{\xi}\|^2)\Big)^{1/2} \\ &= N\sqrt{2t^2M_1^2 + 4\lambda t\,\mathbb{E}\|\boldsymbol{\xi}\|^2},\end{aligned}$$

where we used the previous bound in the second-to-last inequality. This proves Eq. (31). ■

E.1.2. PROOF OF THE MAIN THEOREM

Proof We use the interpolation scheme and notation of Lemma 5. Specifically, Lemma 5 gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}^{(N)}(\nu_t) = -\int \nu_t(\boldsymbol{Y})\,\|G_t(\boldsymbol{Y})\|_2^2\,\mathrm{d}\boldsymbol{Y} - \int \nu_t(\boldsymbol{Y})\,G_t(\boldsymbol{Y})\delta_t(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}. \tag{32}$$

where

$$\delta_t(\boldsymbol{Y}) := \mathbb{E}_{\boldsymbol{Y}_0|}[-\operatorname{lmo}(G_0(\boldsymbol{Y}_0)) - \lambda\nabla\log\nu_0(\boldsymbol{Y}_0)\,|\,\boldsymbol{Y}_t=\boldsymbol{Y}] - N\nabla F(\boldsymbol{Y}).$$

Define

$$\begin{aligned}
&r_t(\boldsymbol{Y}) := r_t^{\mathrm{lmo}}(\boldsymbol{Y}) + r_t^{\mathrm{sc}}(\boldsymbol{Y}),\\
&r_t^{\mathrm{lmo}}(\boldsymbol{Y}) := \operatorname{lmo}\big(G_t(\boldsymbol{Y});0\big) - \mathbb{E}_{\boldsymbol{Y}_0|}\big[\operatorname{lmo}(G_0(\boldsymbol{Y}_0);0)\mid \boldsymbol{Y}_t=\boldsymbol{Y}\big],\\
&r_t^{\mathrm{sc}}(\boldsymbol{Y}) := \lambda\nabla\log\nu_t(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{Y}_0|}\big[\lambda\nabla\log\nu_0(\boldsymbol{Y}_0)\mid \boldsymbol{Y}_t=\boldsymbol{Y}\big],
\end{aligned}$$

Continuing from Eq. (32), we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}^{(N)}(\nu_t) &= -\int \nu_t\|G_t\|_2^2 - \int \nu_t\,G_t^\top\delta_t\\
&= -\int \nu_t\|G_t\|_2^2 - \int \nu_t G_t^\top\big(-\operatorname{lmo}(G_t;0) - G_t + r_t\big)\\
&= -\int \nu_t\|G_t\|_2^2 + \int \nu_t\big\langle G_t, \operatorname{lmo}(G_t;0)\big\rangle + \int \nu_t\|G_t\|_2^2 - \int \nu_t\langle G_t, r_t\rangle\\
&= \int \nu_t\big\langle G_t, \operatorname{lmo}(G_t;0)\big\rangle - \int \nu_t\langle G_t, r_t^{\mathrm{lmo}}\rangle - \int \nu_t\langle G_t, r_t^{\mathrm{sc}}\rangle\\
&= -\int \nu_t\|G_t\|_* - \int \nu_t\langle G_t, r_t^{\mathrm{lmo}}\rangle - \int \nu_t\langle G_t, r_t^{\mathrm{sc}}\rangle,
\end{aligned}$$

where we abbreviate $\nu_t(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}$ by ν_t and use Lemma 7. We now bound the term $-\int \nu_t\langle G_t, r_t^{\mathrm{lmo}}\rangle$. Using the joint distribution ν_{0t} of $(\boldsymbol{Y}_0,\boldsymbol{Y})$ and Lemma 4, we obtain

$$\begin{aligned}
-\int \nu_t\langle G_t, r_t^{\mathrm{lmo}}\rangle &= -\mathbb{E}_{\boldsymbol{Y}\sim\nu_t}\Big[\big\langle G_t(\boldsymbol{Y}), \operatorname{lmo}(G_t(\boldsymbol{Y});0) - \mathbb{E}_{\boldsymbol{Y}_0|\boldsymbol{Y}}\operatorname{lmo}(G_0(\boldsymbol{Y}_0);0)\big\rangle\Big]\\
&= -\mathbb{E}_{(\boldsymbol{Y}_0,\boldsymbol{Y})\sim\nu_{0t}}\Big[\big\langle G_t(\boldsymbol{Y}), \operatorname{lmo}(G_t(\boldsymbol{Y});0) - \operatorname{lmo}(G_0(\boldsymbol{Y}_0);0)\big\rangle\Big]\\
&\le 2\,\mathbb{E}_{(\boldsymbol{Y}_0,\boldsymbol{Y})\sim\nu_{0t}}\big[\|G_t(\boldsymbol{Y}) - G_0(\boldsymbol{Y}_0)\|_*\big].
\end{aligned} \tag{33}$$

We now bound $\|G_t(\boldsymbol{Y}) - G_0(\boldsymbol{Y}_0)\|_*$. Recall

$$\begin{aligned}
G_t(\boldsymbol{Y}) &= N\nabla F(\boldsymbol{Y}) + \lambda\nabla\log\nu_t(\boldsymbol{Y}),\\
G_0(\boldsymbol{Y}_0) &= N\nabla F(\boldsymbol{Y}_0) + \lambda\nabla\log\nu_0(\boldsymbol{Y}_0).
\end{aligned} \tag{34}$$

By the triangle inequality in the dual norm,

$$\|G_t(\boldsymbol{Y}) - G_0(\boldsymbol{Y}_0)\|_* \le N\|\nabla F(\boldsymbol{Y}) - \nabla F(\boldsymbol{Y}_0)\|_* + \lambda\|\nabla\log\nu_t(\boldsymbol{Y}) - \nabla\log\nu_0(\boldsymbol{Y}_0)\|_*. \tag{35}$$

Since $\nabla_{\boldsymbol{y}^i} F(\boldsymbol{Y}) = \frac{1}{N}\nabla \frac{\delta F}{\delta \mu}(\mu_{\boldsymbol{Y}})(\boldsymbol{y}^i)$ (Eq. (28)) and Asm 2, we get

$$\begin{aligned} N\|\nabla F(\boldsymbol{Y}) - \nabla F(\boldsymbol{Y}_0)\|_* &\le N \cdot \frac{L}{2N} \sum_{i=1}^N \big(W_2(\mu_{\boldsymbol{Y}}, \mu_{\boldsymbol{Y}_0}) + \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|\big) \\ &\le N \cdot \frac{L}{2} \frac{1}{\sqrt{N}} \Big(\sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2 \Big)^{1/2} + \frac{L}{2} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|, \end{aligned} \tag{36}$$

where in the last step we used the empirical coupling $\pi := \frac{1}{N}\sum_{i=1}^N \delta_{(\boldsymbol{y}_0^i, \boldsymbol{y}^i)} \in \Pi(\mu_{\boldsymbol{Y}_0}, \mu_{\boldsymbol{Y}})$.

By Asm 7, we also have:

$$\begin{aligned} \|\nabla \log \nu_t(\boldsymbol{Y}) - \nabla \log \nu_0(\boldsymbol{Y}_0)\|_* &\le M_2 \bigg(W_2(\nu_t, \nu_0) + \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \bigg) \\ &\le M_2 \bigg(\big(\mathbb{E}_{\nu_{0t}} \|\boldsymbol{Y} - \boldsymbol{Y}_0\|^2\big)^{1/2} + \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \bigg) \\ &= M_2 \bigg((\mathbb{E}_{\nu_{0t}} \Big[\sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2 \Big])^{1/2} + \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \bigg) \end{aligned} \tag{37}$$

where we choose the coupling $\nu_{0t} = \mathrm{Law}(\boldsymbol{Y}_0, \boldsymbol{Y})$ in the second to the last step and in the last step is by the definition mentioned in the notation section that the norm $\|\cdot\|$ on $\mathbb{R}^{dN}$ is understood as the sum of the corresponding norms over the N blocks.

Substituting Eq. (36) and Eq. (37) into Eq. (35) and taking expectation with respect to ν_{0t}, we obtain

$$\begin{aligned} &\mathbb{E}_{\nu_{0t}} \|G_t(\boldsymbol{Y}) - G_0(\boldsymbol{Y}_0)\|_* \\ &\le\ N \cdot \frac{L}{2} \cdot \frac{1}{\sqrt{N}}\, \mathbb{E}_{\nu_{0t}} \Big(\sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2 \Big)^{1/2} \ +\ \frac{L}{2}\, \mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \\ &\quad +\ \lambda M_2\, \mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \ +\ \lambda M_2 \Big(\mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2 \Big)^{1/2} \\ &\le\ N \cdot \frac{L}{2} \cdot \frac{1}{\sqrt{N}} \Big(\mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2 \Big)^{1/2} \ +\ \frac{L}{2}\, \mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \\ &\quad +\ \lambda M_2\, \mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\| \ +\ \lambda M_2 \Big(\mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2 \Big)^{1/2} \\ &\le\ N \cdot \frac{L}{2} \cdot \frac{1}{\sqrt{N}} \Big(2Nt^2 M_1^2 + 4\lambda t\, N\, \mathbb{E}\|\boldsymbol{\xi}\|^2 \Big)^{1/2} \ +\ \frac{L}{2}\, N \sqrt{2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2} \\ &\quad +\ \lambda M_2\, N \sqrt{2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2} \ +\ \lambda M_2 \Big(2Nt^2 M_1^2 + 4\lambda t\, N\, \mathbb{E}\|\boldsymbol{\xi}\|^2 \Big)^{1/2} \\ &=\ \Big(LN + \lambda M_2 N + \lambda M_2 \sqrt{N} \Big) \sqrt{2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2}. \end{aligned} \tag{38}$$

where we used Jensen's inequality to bound $\mathbb{E}_{\nu_{0t}} (\sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2)^{1/2} \le (\mathbb{E}_{\nu_{0t}} \sum_{i=1}^N \|\boldsymbol{y}^i - \boldsymbol{y}_0^i\|^2)^{1/2}$, and Lemma 6 in the last inequality.

We now bound the term involving r_t^{sc}. By the tower property and Jensen's inequality, we have

$$\begin{aligned}
&\int \nu_t(\boldsymbol{Y})\, \|r_t^{\mathrm{sc}}(\boldsymbol{Y})\|_*^2 \,\mathrm{d}\boldsymbol{Y} \\
&= \mathbb{E}_{\nu_t}\left\| \mathbb{E}\big[\lambda(\nabla \log \nu_t(\boldsymbol{Y}) - \nabla \log \nu_0(\boldsymbol{Y}_0)) \mid \boldsymbol{Y}\big] \right\|_*^2 \\
&\le \lambda^2 \mathbb{E}_{(\boldsymbol{Y}_0,\boldsymbol{Y})\sim\nu_{0t}} \Big[\|\nabla \log \nu_t(\boldsymbol{Y}) - \nabla \log \nu_0(\boldsymbol{Y}_0)\|_*^2 \Big] \qquad \text{(by Jensen)} \\
&\le 2\lambda^2 M_2^2\, \mathbb{E}_{\nu_{0t}} \|\boldsymbol{Y} - \boldsymbol{Y}_0\|^2 + 2\lambda^2 M_2^2\, W_2^2(\nu_t, \nu_0) \qquad \text{(by Asm 7 and } (a+b)^2 \le 2a^2 + 2b^2) \\
&\le 4\lambda^2 M_2^2\, \mathbb{E}_{\nu_{0t}} \|\boldsymbol{Y} - \boldsymbol{Y}_0\|^2 \qquad \text{(using the same coupling as before)} \\
&\le 4\lambda^2 M_2^2 \big(2Nt^2 M_1^2 + 4\lambda t\, N\, \mathbb{E}\|\boldsymbol{\xi}\|^2\big) \qquad \text{(by Lemma 6)}.
\end{aligned} \tag{39}$$

Next, using Hölder's inequality (Proposition 2.10 in Kreuter [24]), we obtain

$$\begin{aligned}
&-\int \nu_t \langle G_t, r_t^{\mathrm{sc}} \rangle \\
&\le \Big(\int \nu_t \|G_t\|_*^2 \Big)^{1/2} \Big(\int \nu_t \|r_t^{\mathrm{sc}}\|^2 \Big)^{1/2} \\
&\le \sqrt{C} \Big(\int \nu_t \|G_t\|_*^2 \Big)^{1/2} \Big(\int \nu_t \|r_t^{\mathrm{sc}}\|_*^2 \Big)^{1/2} \qquad \text{(by norm equivalence } \|\boldsymbol{a}\| \le \sqrt{C}\|\boldsymbol{a}\|_*) \\
&\le \Big(\int \nu_t \|G_t\|_*^2 \Big)^{1/2} \cdot \sqrt{C} \cdot 2\lambda M_2 \big(2Nt^2 M_1^2 + 4\lambda t\, N\, \mathbb{E}\|\boldsymbol{\xi}\|^2\big)^{1/2} \qquad \text{(by Eq. (39))}.
\end{aligned} \tag{40}$$

Thus, define

$$\begin{aligned}
\beta_t &:= 2\lambda M_2 \sqrt{CN}\, \sqrt{2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2}. \\
a_t &:= 2\Big(LN + \lambda M_2 N + \lambda M_2 \sqrt{N}\Big) \sqrt{2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2}.
\end{aligned} \tag{41}$$

Putting together the bounds for the three terms in Eq. (32) we obtain, for all $t \in [0, \eta]$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathscr{F}^{(N)}(\nu_t) - \mathscr{F}^{(N)}(\mu_*^{(N)}) \;&\le\; -\int \nu_t \|G_t\|_* \;+\; \beta_t \left(\int \nu_t \|G_t\|_*^2 \right)^{1/2} \;+\; a_t \\
&\le -(\kappa - \beta_t) \left(\int \nu_t \|G_t\|_*^2 \right)^{1/2} \;+\; a_t \\
&\le -(\kappa - \beta_t) \sqrt{2\zeta\lambda}\, \sqrt{\mathscr{F}^{(N)}(\nu_t) - \mathscr{F}^{(N)}(\mu_*^{(N)})} \;+\; a_t.
\end{aligned} \tag{42}$$

where we use Asm 4 in the second step, log-sobolev inequality Asm 3 and Lemma 2 in the third step when $\beta_t \le \kappa/2$ (which will be satisfied by the following argument).

Next, we choose η so that $\beta_t \le \kappa/2$ on $t \in [0, \eta]$. Since $t \mapsto 2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2$ is increasing on $[0, \infty)$, we have $\sup_{t\in[0,\eta]} \beta_t = \beta_\eta$. For $0 < \eta \le 1$, we bound

$$2t^2 M_1^2 + 4\lambda t\, \mathbb{E}\|\boldsymbol{\xi}\|^2 \;\le\; 2\eta M_1^2 + 4\lambda\eta\, \mathbb{E}\|\boldsymbol{\xi}\|^2 = 2\eta\big(M_1^2 + 2\lambda\, \mathbb{E}\|\boldsymbol{\xi}\|^2\big),$$

and therefore, by Eq. (41),

$$\beta_\eta \;\le\; 2\lambda M_2\sqrt{CN}\,\sqrt{2\eta\big(M_1^2+2\lambda\,\mathbb{E}\|\boldsymbol{\xi}\|^2\big)}. \tag{43}$$

Hence it suffices to choose $\eta>0$ such that

$$2\lambda M_2\sqrt{CN}\,\sqrt{2\eta\big(M_1^2+2\lambda\,\mathbb{E}\|\boldsymbol{\xi}\|^2\big)} \;\le\; \frac{\kappa}{2},$$

Define

$$\eta_0 := \min\left\{1,\; \frac{\kappa^2}{32\,\lambda^2\,M_2^2\,C\,N\,\big(M_1^2+2\lambda\,\mathbb{E}\|\boldsymbol{\xi}\|^2\big)}\right\}. \tag{44}$$

Then for all $t\in[0,\eta]$ whenever $0<\eta\le\eta_0$, we have $\beta_t\le\kappa/2$ and thus $\kappa-\beta_t\ge\kappa/2$. Consequently, Eq. (42) simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathscr{F}^{(N)}(\nu_t)-\mathscr{F}^{(N)}(\mu_*^{(N)}) \;\le\; -a\sqrt{\mathscr{F}^{(N)}(\nu_t)-\mathscr{F}^{(N)}(\mu_*^{(N)})}\;+\;a_t,\quad a\;:=\;\frac{\kappa}{2}\sqrt{2\zeta\lambda} \tag{45}$$

Lastly, we convert this differential inequality into a per-iteration recursion. Fix an iteration k and consider the interpolation on $t\in[0,\eta]$. For notational convenience, set

$$\mathcal{E}(t) := \mathscr{F}^{(N)}(\nu_t)-\mathscr{F}^{(N)}(\mu_*^{(N)}),\qquad \bar{a}_\eta := \sup_{t\in[0,\eta]} a_t.$$

From the previous step, we have for all $t\in[0,\eta]$,

$$\mathcal{E}'(t) \;\le\; -a\sqrt{\mathcal{E}(t)}\;+\;a_t,\qquad a\;:=\;\frac{\kappa}{2}\sqrt{2\zeta\lambda}. \tag{46}$$

Define $y(t):=\sqrt{\mathcal{E}(t)}$. Whenever $y(t)>0$,

$$2y(t)y'(t)=\mathcal{E}'(t)\;\le\;-ay(t)+a_t\;\le\;-ay(t)+\bar{a}_\eta.$$

Hence, for all t with $y(t)>0$,

$$y'(t)\;\le\;-\frac{a}{2}\;+\;\frac{\bar{a}_\eta}{2y(t)}. \tag{47}$$

Set the threshold $R\;:=\;\frac{2\bar{a}_\eta}{a}$. We distinguish two cases.

Case 1: $y(t)\ge R$ for all $t\in[0,\eta]$. Then Eq. (47) gives

$$y'(t)\le-\frac{a}{2}+\frac{a}{4}=-\frac{a}{4},\qquad t\in[0,\eta].$$

Integrating over $[0,\eta]$ yields

$$y(\eta)\le y(0)-\frac{a}{4}\eta.$$

Case 2: $y(t_0)<R$ for some $t_0\in[0,\eta]$. Let t_0 be the first such time (if $y(0)<R$, take $t_0=0$). We claim that $y(t)\le R$ for all $t\in[t_0,\eta]$. Indeed, suppose for contradiction that there exists $t_1\in(t_0,\eta]$ with $y(t_1)>R$. By continuity, define

$$s:=\max\{t\in[t_0,t_1]:y(t)=R\}.$$

Then $y(s) = R$ and $y(t) > R$ for all $t \in (s, t_1]$. But whenever $y(t) \geq R$, Eq. (47) and the definition of R imply

$$y'(t) \leq -\frac{a}{2} + \frac{\bar{a}_\eta}{2y(t)} \leq -\frac{a}{2} + \frac{\bar{a}_\eta}{2R} = -\frac{a}{2} + \frac{a}{4} = -\frac{a}{4} < 0,$$

so y is strictly decreasing on $(s, t_1]$, contradicting $y(t) > R$ there. Therefore $y(t) \leq R$ on $[t_0, \eta]$, and in particular $y(\eta) \leq R$.

Combining the two cases, we obtain the one-step bound

$$y(\eta) \;\leq\; \max\Big\{y(0) - \frac{a}{4}\eta,\; \frac{2\bar{a}_\eta}{a}\Big\}.$$

Recalling $y(t) = \sqrt{\mathcal{E}(t)}$, we get

$$\sqrt{\mathscr{F}^{(N)}(\mu_{k+1}^{(N)}) - \mathscr{F}^{(N)}(\mu_*^{(N)})} \;\leq\; \max\Big\{\sqrt{\mathscr{F}^{(N)}(\mu_k^{(N)}) - \mathscr{F}^{(N)}(\mu_*^{(N)})} - \frac{a}{4}\eta,\; \frac{2\bar{a}_\eta}{a}\Big\}. \tag{48}$$

Iterating Eq. (48) yields, for all $k \geq 0$,

$$\mathscr{F}^{(N)}(\mu_k^{(N)}) - \mathscr{F}^{(N)}(\mu_*^{(N)}) \;\leq\; \left[\max\Big\{\sqrt{\mathcal{E}_0^{(N)}} - \frac{a}{4}\,k\eta,\; \frac{2\bar{a}_\eta}{a}\Big\}\right]^2. \tag{49}$$

Finally, add and subtract $\mathscr{F}^{(N)}(\mu_*^{(N)})$:

$$\frac{1}{N}\mathscr{F}^{(N)}(\mu_k^{(N)}) - \mathscr{F}(\mu^\star) = \frac{1}{N}\Big(\mathscr{F}^{(N)}(\mu_k^{(N)}) - \mathscr{F}^{(N)}(\mu_*^{(N)})\Big) + \Big(\frac{1}{N}\mathscr{F}^{(N)}(\mu_*^{(N)}) - \mathscr{F}(\mu^\star)\Big). \tag{50}$$

The finite-particle bias term is controlled by Lemma 3:

$$\frac{1}{N}\mathscr{F}^{(N)}(\mu_*^{(N)}) - \mathscr{F}(\mu^\star) \;\leq\; \frac{\hat{L}\hat{R}^2}{2N}. \tag{51}$$

Combining Eq. (49)–Eq. (51) gives, for all $k \geq 0$,

$$\begin{aligned}&\frac{1}{N}\mathscr{F}^{(N)}(\mu_k^{(N)}) - \mathscr{F}(\mu^\star)\\ &\leq\; \frac{1}{N}\left[\max\Big\{\sqrt{\mathscr{F}^{(N)}(\mu_0^{(N)}) - \mathscr{F}^{(N)}(\mu_*^{(N)})} - \frac{a}{4}\,k\eta,\; \frac{2\bar{a}_\eta}{a}\Big\}\right]^2 + \frac{\hat{L}\hat{R}^2}{2N}, \quad a = \frac{\kappa}{2}\sqrt{2\zeta\lambda}.\end{aligned} \tag{52}$$

■

Appendix F. LMO gradient flow in measure space in noiseless setting ($\lambda = 0$)

Theorem 11 (Convergence of normalized LMO Wasserstein flow) *Consider the following LMO Wasserstein flow:*

$$\partial_t \mu_t = \nabla\cdot(\mu_t v_t), \quad v_t := -\operatorname{lmo}\Big(\nabla\tfrac{\delta F}{\delta \mu}(\mu_t)(\boldsymbol{x});\, 0\Big),$$

we assume the solution $(\mu_t)_{t\geq 0}$ exists. Then under Asm. 4 and 8, for all $t \geq 0$,

$$F(\mu_t) - F^\star \leq \left[\max\Big\{\sqrt{F(\mu_0) - F^\star} - \kappa\sqrt{\tfrac{\alpha}{2}}\,t,\; 0\Big\}\right]^2.$$

In particular, $F(\mu_t) = F^\star$ for all $t \geq t^\star$, where $t^\star = \frac{1}{\kappa}\sqrt{\frac{2}{\alpha}}\,\sqrt{F(\mu_0) - F^\star}$.

F.1. Proof of Thm. 11

Proof Let $\Delta(t) := F(\mu_t) - F^\star \geq 0$, $g_{\mu_t}(x) := \nabla_x \frac{\delta F}{\delta \mu}(\mu_t)(x)$. By the chain rule for first variations along continuity equations,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = \int \big\langle g_{\mu_t}(x), v_t(x) \big\rangle \, \mathrm{d}\mu_t(x).$$

Plugging $v_t(x)$ and using Lemma 7, we get

$$\Delta'(t) = \frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\|g_{\mu_t}\|_{L^1(\mu_t;\|\cdot\|_*)}. \tag{53}$$

By the non-degeneracy assumption Asm 4,

$$\|g_{\mu_t}\|_{L^1(\mu_t;\|\cdot\|_*)} \;\geq\; \kappa \, \|g_{\mu_t}\|_{L^2(\mu_t;\|\cdot\|_*)}. \tag{54}$$

Finally, by the transport PL inequality (Asm 8),

$$\|g_{\mu_t}\|_{L^2(\mu_t;\|\cdot\|_*)} \geq \sqrt{2\alpha \, \Delta(t)}. \tag{55}$$

Combining Eq. (53)–Eq. (55) yields the differential inequality

$$\Delta'(t) \leq -\kappa\sqrt{2\alpha} \, \sqrt{\Delta(t)}.$$

Let $y(t) := \sqrt{\Delta(t)}$. Wherever $y(t) > 0$, we have $y'(t) = \Delta'(t)/(2\sqrt{\Delta(t)}) \leq -\kappa\sqrt{\alpha/2}$. Integrating gives

$$y(t) \leq \max\Big\{ y(0) - \kappa\sqrt{\alpha/2}\, t, \; 0 \Big\},$$

and squaring yields the convergence rate given in Thm. 11. The expression for $t^\star$ followed by setting the bracketed term to zero. ∎

F.2. Proof of Lemma 12

Proof Fix μ and T, and define the displacement interpolation

$$\mu_t := (\mathrm{id} + tv)_{\#}\mu, \qquad t \in [0,1]\,,$$

where $v(\boldsymbol{x}) := T(\boldsymbol{x}) - \boldsymbol{x}$. Let $h(t) := F(\mu_t)$. Lemma A.2 in Chizat [12] provides the derivative formula

$$h'(t) = \int_{\mathbb{R}^d} \Big\langle \nabla_{\boldsymbol{x}} \tfrac{\delta F}{\delta \mu}(\mu_t)\big(x + tv(\boldsymbol{x})\big), \, v(\boldsymbol{x}) \Big\rangle \, \mathrm{d}\mu(\boldsymbol{x}). \tag{56}$$

Fix $s, t \in [0,1]$ and define $G_{t,s}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu_t)\big(\boldsymbol{x} + tv(\boldsymbol{x})\big) - \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu_s)\big(\boldsymbol{x} + sv(\boldsymbol{x})\big)$. Then, subtracting Eq. (56) at times t and s, we obtain

$$h'(t) - h'(s) = \int \big\langle G_{t,s}(\boldsymbol{x}), \, v(\boldsymbol{x}) \big\rangle \, \mathrm{d}\mu(\boldsymbol{x}).$$

By $|\langle u, v\rangle| \leq \|u\|_* \, \|v\|$ and Hölder's inequality applied to the scalar functions $\|G_{t,s}(\cdot)\|_*$ and $\|v(\cdot)\|$ (Proposition 2.10 in Kreuter [24]), we obtain

$$|h'(t) - h'(s)| \;\leq\; \|G_{t,s}\|_{L^2(\mu;\|\cdot\|_*)} \, \|v\|_{L^2(\mu;\|\cdot\|)}. \tag{57}$$

Next we bound $\|G_{t,s}\|_{L^2(\mu;\|\cdot\|_*)}$. By Asm 2:

$$\begin{aligned}\|G_{t,s}(\boldsymbol{x})\|_* &= \big\|\nabla_{\boldsymbol{x}}\tfrac{\delta F}{\delta \mu}(\mu_t)(\boldsymbol{x}+tv(\boldsymbol{x})) - \nabla_{\boldsymbol{x}}\tfrac{\delta F}{\delta \mu}(\mu_s)(\boldsymbol{x}+sv(\boldsymbol{x}))\big\|_* \\ &\le \frac{L}{2}\big(W_2(\mu_t,\mu_s) + |t-s|\,\|v(\boldsymbol{x})\|\big).\end{aligned} \tag{58}$$

Recall that W_2 is defined using the same norm $\|\cdot\|$: i.e.,

$$W_2^2(\mu_t,\mu_s) = \inf_{\pi\in\Pi(\mu_t,\mu_s)}\int \|\boldsymbol{x}-\boldsymbol{y}\|^2\,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}).$$

Consider the measurable map $\Phi_{t,s}(x) := (x+tv(x),\, x+sv(x))$ and define $\pi_{t,s} := (\Phi_{t,s})_{\#}\mu$. Then, for any Borel set $A\subset\mathbb{R}^d$, we have

$$\pi_{t,s}(A\times\mathbb{R}^d) = \mu\big(\{x:\ x+tv(x)\in A\}\big) = (\mathrm{id}+tv)_{\#}\mu(A) = \mu_t(A),$$

and similarly,

$$\pi_{t,s}(\mathbb{R}^d\times A) = \mu\big(\{x:\ x+sv(x)\in A\}\big) = (\mathrm{id}+sv)_{\#}\mu(A) = \mu_s(A).$$

Hence $\pi_{t,s}\in\Pi(\mu_t,\mu_s)$.

By the definition of the Wasserstein distance:

$$W_2^2(\mu_t,\mu_s) \;\le\; \int \big\|\big(\boldsymbol{x}+tv(\boldsymbol{x})\big)-\big(\boldsymbol{x}+sv(\boldsymbol{x})\big)\big\|^2\,\mathrm{d}\mu(\boldsymbol{x}) = |t-s|^2\int\|v(\boldsymbol{x})\|^2\,\mathrm{d}\mu(\boldsymbol{x}).$$

Therefore, we get:

$$\|G_{t,s}(\boldsymbol{x})\|_* \le \frac{L}{2}|t-s|\big(\|v\|_{L^2(\mu;\|\cdot\|)} + \|v(\boldsymbol{x})\|\big).$$

Squaring and using $(a+b)^2\le 2a^2+2b^2$:

$$\|G_{t,s}(\boldsymbol{x})\|_*^2 \le \frac{L^2}{2}|t-s|^2\Big(\|v\|^2_{L^2(\mu;\|\cdot\|)} + \|v(\boldsymbol{x})\|^2\Big).$$

Integrating over μ yields:

$$\begin{aligned}\|G_{t,s}\|^2_{L^2(\mu;\|\cdot\|_*)} &= \int\|G_{t,s}(\boldsymbol{x})\|_*^2\,\mathrm{d}\mu(\boldsymbol{x}) \\ &\le \frac{L^2}{2}|t-s|^2\Big(\|v\|^2_{L^2(\mu;\|\cdot\|)} + \int\|v(\boldsymbol{x})\|^2\,\mathrm{d}\mu(\boldsymbol{x})\Big) \\ &= L^2|t-s|^2\|v\|^2_{L^2(\mu;\|\cdot\|)}.\end{aligned}$$

Plugging back Eq. (57), we obtain:

$$|h'(t)-h'(s)| \;\le\; L|t-s|\,\|v\|^2_{L^2(\mu;\|\cdot\|)}. \tag{59}$$

Taking $s=0$ in Eq. (59) and integrating over $t\in[0,1]$:

$$h(1) = h(0)+\int_0^1 h'(t)\,\mathrm{d}t \;\le\; h(0)+h'(0)+\int_0^1|h'(t)-h'(0)|\,\mathrm{d}t \;\le\; h(0)+h'(0)+\frac{L}{2}\|v\|^2_{L^2(\mu;\|\cdot\|)}.$$

Finally, by definition $h(0) = F(\mu)$ and $h(1) = F(T_{\#}\mu)$, and Eq. (56), substituting these identities into the previous inequality gives exactly Eq. (85), completing the proof. ∎

F.3. Auxiliary lemma

Lemma 7 (LMO identities) Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ with dual norm $\|\cdot\|_*$, and let $\mathrm{lmo}(\cdot;\tau)$ be defined as in Thm. 1. Then, for every $\boldsymbol{g} \in \mathbb{R}^d$,

$$\big\langle \boldsymbol{g}, \mathrm{lmo}(\boldsymbol{g};\tau)\big\rangle \;=\; -\,\|\boldsymbol{g}\|_*^{\tau+1}, \qquad \big\|\,\mathrm{lmo}(\boldsymbol{g};\tau)\big\| \;=\; \|\boldsymbol{g}\|_*^{\tau}.$$

Proof If $\boldsymbol{g} = 0$, then $\|\boldsymbol{g}\|_* = 0$ and any $\boldsymbol{y}$ with $\|\boldsymbol{y}\| \le 0$ must be $\boldsymbol{y} = 0$, so $\mathrm{lmo}(0;\tau) = 0$ and both identities hold trivially. Hence we may assume $\boldsymbol{g} \neq 0$.

By Definition 1, we have

$$\mathrm{lmo}(\boldsymbol{g};\tau) = -\,\|\boldsymbol{g}\|_*^{\tau}\, s(\boldsymbol{g}), \quad \text{where } s(\boldsymbol{g}) \in \arg\max_{\|\boldsymbol{u}\|\le 1} \langle \boldsymbol{g}, \boldsymbol{u}\rangle \;\text{ and }\; \langle \boldsymbol{g}, s(\boldsymbol{g})\rangle = \|\boldsymbol{g}\|_*.$$

Since the maximum of a linear functional over the unit ball is attained on the unit sphere, we can choose $s(\boldsymbol{g})$ with $\|s(\boldsymbol{g})\| = 1$.

Then

$$\big\langle \boldsymbol{g}, \mathrm{lmo}(\boldsymbol{g};\tau)\big\rangle = \big\langle \boldsymbol{g},\, -\|\boldsymbol{g}\|_*^{\tau}\, s(\boldsymbol{g})\big\rangle = -\,\|\boldsymbol{g}\|_*^{\tau}\,\langle \boldsymbol{g}, s(\boldsymbol{g})\rangle = -\,\|\boldsymbol{g}\|_*^{\tau}\,\|\boldsymbol{g}\|_* = -\,\|\boldsymbol{g}\|_*^{\tau+1},$$

and

$$\big\|\,\mathrm{lmo}(\boldsymbol{g};\tau)\big\| = \big\|-\,\|\boldsymbol{g}\|_*^{\tau}\, s(\boldsymbol{g})\big\| = \|\boldsymbol{g}\|_*^{\tau}\,\|s(\boldsymbol{g})\| = \|\boldsymbol{g}\|_*^{\tau}.$$

This proves the claim. ∎

Lemma 8 (Upper bound for $\widehat{\Delta}_K$ in Eq. (73)) The term $\widehat{\Delta}_K$ defined in Eq. (73) can be bounded as follows:

$$\widehat{\Delta}_K \;=\; \max_{0\le t<K} \Delta_t \;\le\; \Delta_0 + \frac{L}{2}\,\eta^2 \log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big). \tag{60}$$

Proof From the smoothness inequality (Eq. (70), dropping the negative term $-\eta_k \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_k^i\|_*^{1+\tau}$) we have

$$F(\mu_{k+1}) - F(\mu_k) \;\le\; \frac{L}{2}\eta_k^2 \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_k^i\|_*^{2\tau}.$$

Summing over $s = 0, \dots, k-1$ yields

$$F(\mu_k) - F(\mu_0) \;\le\; \frac{L}{2}\sum_{s=0}^{k-1}\eta_s^2 \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_s^i\|_*^{2\tau} \;\le\; \frac{L}{2}\,\eta^2 \log\Big(\frac{S_{k-1}}{S_{\text{init}}}\Big) \;\le\; \frac{L}{2}\,\eta^2 \log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big),$$

for all $k < K$, where we use Lemma 9. Hence

$$\Delta_k = F(\mu_k) - F^\star \;\le\; \Delta_0 + \frac{L}{2}\,\eta^2 \log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big), \qquad 0 \le k < K,$$

and therefore

$$\widehat{\Delta}_K \;=\; \max_{0\le k<K} \Delta_t \;\le\; \Delta_0 + \frac{L}{2}\,\eta^2 \log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big). \tag{61}$$

∎

Lemma 9 (From Wang et al. [47] (Lemma 10)) Let $\{a_k\}_{k=0}^{\infty}$ be a sequence of non-negative real numbers with $a_0 > 0$. Then the following inequality holds:

$$\sum_{k=1}^{K} \frac{a_k}{\sum_{s=0}^{k} a_s} \leq \ln\left(\sum_{k=0}^{K} a_k\right) - \ln a_0.$$

Lemma 10 (From Attia and Koren [3] (Lemma 4)) Let $\{a_k\}_{k=0}^{\infty}$ be a sequence of non-negative real numbers. Then the following inequality holds:

$$\sum_{k=1}^{K} \frac{a_k}{\sqrt{\sum_{s=0}^{k} a_s}} \leq 2\sqrt{\sum_{k=1}^{K} a_k}$$

Lemma 11 (Polynomial growth of the accumulator S_{K-1} for $\tau = 1$) Under the same setting and assumption as in Thm. 5, We have:

$$\log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big) \;\leq\; \log\left(1 + \frac{4KL(F(\mu_0) - F^\star) + \frac{1}{2}(L(\sqrt{N}+1)\,\eta)^2 K^3}{S_{\text{init}}}\right). \tag{62}$$

Proof We work in the $\tau = 1$ case. Recall that

$$\boldsymbol{g}_k^i := \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu_k)(X_k^i), \qquad \boldsymbol{v}_k^i := \operatorname{lmo}(\boldsymbol{g}_k^i; 1), \qquad \|\boldsymbol{v}_k^i\| = \|\boldsymbol{g}_k^i\|_*.$$

By using Asm 2, we have

$$\|\boldsymbol{g}_k^i - \boldsymbol{g}_{k-1}^i\|_* \leq \frac{L}{2}\Big(W_2(\mu_k, \mu_{k-1}) + \|\boldsymbol{x}_k^i - \boldsymbol{x}_{k-1}^i\|\Big).$$

Using the update $\boldsymbol{x}_k^i = \boldsymbol{x}_{k-1}^i + \eta_{k-1}\boldsymbol{v}_{k-1}^i$ and the coupling that pairs each $\boldsymbol{x}_{k-1}^j$ with $\boldsymbol{x}_k^j$, we obtain

$$\|\boldsymbol{x}_k^i - \boldsymbol{x}_{k-1}^i\| = \eta_{k-1}\|\boldsymbol{v}_{k-1}^i\|, \qquad W_2(\mu_k, \mu_{k-1}) \leq \eta_{k-1}\sqrt{B_{k-1}},$$

where $B_{k-1} = \frac{1}{N}\sum_{j=1}^{N} \|\boldsymbol{v}_{k-1}^j\|^2$. Hence

$$\begin{aligned}
\|\boldsymbol{g}_k^i\|_* &\leq \|\boldsymbol{g}_{k-1}^i\|_* + \frac{L}{2}\eta_{k-1}\big(\sqrt{B_{k-1}} + \|\boldsymbol{v}_{k-1}^i\|\big) \\
&\leq \|\boldsymbol{g}_{k-1}^i\|_* + \frac{L}{2}\eta_{k-1}\Big(\sqrt{B_{k-1}} + \sqrt{\sum_{j=1}^{N}\|\boldsymbol{v}_{k-1}^j\|^2}\Big) \\
&= \|\boldsymbol{g}_{k-1}^i\|_* + \frac{L}{2}\eta_{k-1}(\sqrt{N}+1)\sqrt{B_{k-1}}.
\end{aligned} \tag{63}$$

By the definition of the global adaptive stepsizes,

$$\eta_{k-1} = \frac{\eta}{\sqrt{S_{k-1}}}, \qquad S_{k-1} = S_{\text{init}} + \sum_{s=0}^{k-1} B_s,$$

so $\eta_{k-1}\sqrt{B_{k-1}} \le \eta$. Consequently,

$$\|\boldsymbol{g}_k^i\|_* \le \|\boldsymbol{g}_{k-1}^i\|_* + \frac{L}{2}(\sqrt{N}+1)\,\eta, \tag{64}$$

Iterating Eq. (64) gives, for all $k \ge 1$,

$$\|\boldsymbol{g}_k^i\|_* \le \|\boldsymbol{g}_0^i\|_* + k\,\frac{L}{2}(\sqrt{N}+1)\,\eta. \tag{65}$$

Squaring Eq. (65) and using $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$\|\boldsymbol{g}_k^i\|_*^2 \le 2\|\boldsymbol{g}_0^i\|_*^2 + \frac{1}{2}(L(\sqrt{N}+1)\,\eta)^2 k^2.$$

Thus, we have

$$B_k = \frac{1}{N}\sum_{j=1}^{N}\|v_k^j\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^2 \le 2B_0 + \frac{1}{2}(L(\sqrt{N}+1)\,\eta)^2 k^2, \qquad B_0 := \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_0^i\|_*^2.$$

Therefore

$$\begin{aligned} S_{K-1} &= S_{\text{init}} + \sum_{k=0}^{K-1} B_k \\ &\le S_{\text{init}} + \sum_{k=0}^{K-1}\left(2B_0 + \frac{1}{2}(L(\sqrt{N}+1)\,\eta)^2 k^2\right) \\ &= S_{\text{init}} + 2KB_0 + \frac{1}{2}(L(\sqrt{N}+1)\,\eta)^2 \sum_{k=0}^{K-1} k^2 \\ &\le S_{\text{init}} + 2KB_0 + \frac{1}{2}(L(\sqrt{N}+1)\,\eta)^2 K^3\,. \end{aligned} \tag{66}$$

Next, we will bound the term $B_0 = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_0^i\|_*^2$. Continuing from Eq. (70) but with a fix step size $\hat{\eta}$, and $k = 0$, we have

$$F(\mu_1) - F(\mu_0) \;\le\; -\hat{\eta}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_0^i\|_*^2 \;+\; \frac{L}{2}\,\hat{\eta}^2\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_0^i\|_*^2. \tag{67}$$

By setting $\hat{\eta} = \frac{1}{L}$, we get

$$B_0 \le 2L(F(\mu_0) - F^\star)$$

Finally, dividing Eq. (66) by S_{init}, applying the monotonicity of $\log$ yields Eq. (62). ■

F.4. Global adaptive algorithm with unnormalized LMO particle descent

Theorem 12 *Consider the same setting as in Thm. 5 with the case $\tau = 1$, we have the following convergence guarantee:*

$$\min_{0\le k<K} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_* \le \frac{\sqrt{2\Delta_0}\sqrt[4]{S_{init}}}{\sqrt{K\eta}} + \frac{\sqrt{2}L\eta}{\sqrt{K}} + \frac{2\sqrt{2}}{\eta\sqrt{K}}\Delta_0 + \frac{2\sqrt{2}}{\eta\sqrt{K}}\left(\frac{L}{2}\eta^2 \log\left(1 + \frac{4KL\Delta_0 + \frac{1}{2}(L(\sqrt{N}+1)\eta)^2K^3}{S_{init}}\right)\right).$$

Remark: The proof can be found at Sec. F.5 As a comparison: i) When $\tau = 1$, the right-hand side contains a logarithmic dependence on the number of particles N, whereas this dependence disappears when $\tau = 0$. ii) When $\tau = 0$, we have $\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{2\tau} = 1$ by Lemma 7. Thus, the step size $\eta_k = \frac{\eta}{\sqrt{k}}$ does not depend on the past gradient. iii) The first term in the right-hand side of the bound is $\mathcal{O}(K^{-1/2})$ for $\tau = 1$ and $\mathcal{O}(K^{-1})$ for $\tau = 0$.

F.5. Proof of Thm. 5 and Thm. 12

F.5.1. PROOF

Proof [Proof of Thm. 5] At iteration k, the LMO-APGD update can be written as a pushforward

$$\mu_{k+1} = (T_k)_{\#}\mu_k, \qquad T_k(\boldsymbol{x}) = \boldsymbol{x} + \eta_k\, \boldsymbol{v}_k(\boldsymbol{x}),$$

where $\boldsymbol{v}_k(\boldsymbol{x}) = \boldsymbol{v}_k^i$ whenever $\boldsymbol{x} = \boldsymbol{x}_k^i$. Define $g_k(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu_k)(\boldsymbol{x})$ and $g_k(\boldsymbol{x}) = \boldsymbol{g}_k^i$ whenever $\boldsymbol{x} = \boldsymbol{x}_k^i$. By Asm 2, we have Eq. (85), by applying with T_k, we obtain

$$F(\mu_{k+1}) - F(\mu_k) \le \int \left\langle g_k(\boldsymbol{x}),\, T_k(\boldsymbol{x}) - \boldsymbol{x} \right\rangle \mathrm{d}\mu_k(\boldsymbol{x}) \;+\; \frac{L}{2}\int \left\|T_k(\boldsymbol{x}) - \boldsymbol{x}\right\|^2 \mathrm{d}\mu_k(\boldsymbol{x}) \tag{68}$$

$$= \int \langle g_k(\boldsymbol{x}),\, \eta_k\, \boldsymbol{v}_k(\boldsymbol{x})\rangle \,\mathrm{d}\mu_k(\boldsymbol{x}) \;+\; \frac{L}{2}\int \|\eta_k\, \boldsymbol{v}_k(\boldsymbol{x})\|^2 \,\mathrm{d}\mu_k(\boldsymbol{x}). \tag{69}$$

Since μ_k is the empirical measure $\mu_k = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{x}_k^i}$, we have

$$\int \langle g_k(\boldsymbol{x}), \eta_k\, \boldsymbol{v}_k(\boldsymbol{x})\rangle \,\mathrm{d}\mu_k(\boldsymbol{x}) = \eta_k \frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{g}_k^i, \boldsymbol{v}_k^i\rangle,$$

$$\int \|\eta_k\, \boldsymbol{v}_k(\boldsymbol{x})\|^2 \,\mathrm{d}\mu_k(\boldsymbol{x}) = \eta_k^2 \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{v}_k^i\|^2 .$$

By the LMO identities (Lemma 7) for $\tau \in \{0, 1\}$, we have

$$\langle \boldsymbol{g}_k^i, \boldsymbol{v}_k^i\rangle = -\|\boldsymbol{g}_k^i\|_*^{1+\tau}, \qquad \|\boldsymbol{v}_k^i\|^2 = \|\boldsymbol{g}_k^i\|_*^{2\tau},$$

and therefore

$$F(\mu_{k+1}) - F(\mu_k) \;\le\; -\eta_k \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{1+\tau} \;+\; \frac{L}{2}\eta_k^2 \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{2\tau}. \tag{70}$$

Rearranging yields

$$\eta_k \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{1+\tau} \;\le\; F(\mu_k) - F(\mu_{k+1}) \;+\; \frac{L}{2}\eta_k^2 \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{2\tau}. \tag{71}$$

Divide both sides of Eq. (71) by $K\eta_k$ and then summing up over $k = 0, \dots, K-1$:

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{1+\tau} \le \underbrace{\frac{1}{K}\sum_{k=0}^{K-1}\frac{F(\mu_k) - F(\mu_{k+1})}{\eta_k}}_{(\star)} + \frac{L}{2K}\sum_{k=0}^{K-1}\eta_k \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{2\tau}. \tag{72}$$

We now bound the term $(\star)$. Define

$$\Delta_k := F(\mu_k) - F^\star, \qquad \widehat{\Delta}_K := \max_{0\le j<K} \Delta_j, \tag{73}$$

We set $\Delta_{-1} := \Delta_0$ and $\eta_{-1} := \frac{\eta}{\sqrt{S_{\text{init}}}}$ for convenience. Then

$$\begin{aligned}
(\star) &= \frac{1}{K}\sum_{k=0}^{K-1}\frac{\Delta_k - \Delta_{k+1}}{\eta_k} \\
&= \frac{1}{K}\sum_{k=-1}^{K-1}\frac{\Delta_k - \Delta_{k+1}}{\eta_k} \quad \text{(the } k=-1 \text{ term vanishes since } \Delta_{-1} = \Delta_0\text{)} \\
&= \frac{1}{K}\left(\frac{\Delta_{-1}}{\eta_{-1}} + \sum_{k=0}^{K-1}\Delta_k\Big(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\Big) - \frac{\Delta_K}{\eta_{K-1}}\right) \\
&\le \frac{\Delta_0}{K\eta_{-1}} + \frac{\widehat{\Delta}_K}{K}\sum_{k=0}^{K-1}\Big(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\Big),
\end{aligned}$$

where we used $\Delta_k \le \widehat{\Delta}_K$, and η_k is a non-increasing sequence, and dropped the negative term.

Recall that $\eta_k = \eta/\sqrt{S_k}$ and by Lemma 7 we have

$$S_k = S_{k-1} + B_k, \qquad B_k := \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{v}_k^i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^{2\tau}.$$

Then

$$\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} = \frac{\sqrt{S_k} - \sqrt{S_{k-1}}}{\eta} = \frac{S_k - S_{k-1}}{\eta\big(\sqrt{S_k} + \sqrt{S_{k-1}}\big)} \le \frac{S_k - S_{k-1}}{\eta\sqrt{S_k}} = \frac{B_k}{\eta\sqrt{S_k}} = \frac{\eta_k B_k}{\eta^2}.$$

Hence

$$(\star) \le \frac{\Delta_0}{K\eta_{-1}} + \frac{\widehat{\Delta}_K}{K\eta^2}\sum_{k=0}^{K-1}\eta_k B_k.$$

Substituting into Eq. (72), we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{1+\tau} \;\le\; \frac{\Delta_0}{K\eta_{-1}} + \frac{\widehat{\Delta}_K}{K\eta^2}\sum_{k=0}^{K-1}\eta_k B_k + \frac{L}{2K}\sum_{k=0}^{K-1}\eta_k B_k. \tag{74}$$

By Lemma 10, we have

$$\sum_{k=0}^{K-1} \eta_k B_k = \sum_{k=0}^{K-1} \frac{\eta B_k}{\sqrt{S_k}} = \sum_{k=0}^{K-1} \frac{\eta B_k}{\sqrt{\sum_{s=0}^{k} B_s + S_{\text{init}}}} \;\le\; 2\eta \sqrt{\sum_{k=0}^{K-1} B_k}. \tag{75}$$

Substituting Eq. (75) into Eq. (74), we obtain

$$\min_{0\le k<K} \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^{1+\tau} \;\le\; \frac{\Delta_0}{K\eta_{-1}} + \Big(\frac{2\widehat{\Delta}_K}{K\eta} + \frac{L\eta}{K}\Big) \sqrt{\sum_{k=0}^{K-1} B_k}. \tag{76}$$

Case $\tau = 0$. In this case, for each i, $\|\boldsymbol{v}_k^i\| \in \{0, 1\}$. Indeed, when $\tau = 0$ the LMO solves

$$\boldsymbol{v}_k^i \;=\; \mathrm{lmo}(\boldsymbol{g}_k^i; 0) \;=\; \arg\min_{\|\boldsymbol{y}\|\le 1} \langle \boldsymbol{g}_k^i, \boldsymbol{y}\rangle,$$

which returns a unit vector in the direction of $-\boldsymbol{g}_k^i$ whenever $\boldsymbol{g}_k^i \neq 0$, and returns the zero vector only in the degenerate case $\boldsymbol{g}_k^i = 0$. Consequently, $0 \le B_k = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{v}_k^i\|^2 \le 1$.

Thus, the right of Eq. (76) can be further bounded by combining with the bound for $\widehat{\Delta}_K$ in Lemma 8 (using $S_K \le S_{\text{init}} + K$):

$$\min_{0\le k<K} \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_* \le \frac{\Delta_0}{K\eta_{-1}} + \frac{2}{K\eta}\left(\Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{\text{init}} + K}{S_{\text{init}}}\Big)\right)\sqrt{K} + \frac{L\eta}{K}\sqrt{K},$$

Case $\tau = 1$. We now bound the $\sqrt{\sum_{k=0}^{K-1} B_k}$ terms in Eq. (76) using Young's inequality ($ab \le \frac{a^2}{4} + b^2$). Specifically,

$$\frac{2\widehat{\Delta}_K}{K\eta}\sqrt{\sum_{k=0}^{K-1} B_k} \le \frac{\sum_{k=0}^{K-1} B_k}{4K} + \frac{4\widehat{\Delta}_K^2}{\eta^2 K},$$

$$\frac{L\eta}{K}\sqrt{\sum_{k=0}^{K-1} B_k} \le \frac{\sum_{k=0}^{K-1} B_k}{4K} + \frac{L^2\eta^2}{K},$$

and thus

$$\frac{1}{K}\sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^2 \le \frac{\Delta_0}{K\eta_{-1}} + \frac{\sum_{k=0}^{K-1} B_k}{2K} + \frac{4\widehat{\Delta}_K^2}{\eta^2 K} + \frac{L^2\eta^2}{K}. \tag{77}$$

Substituting the bound for $\widehat{\Delta}_K$ in Lemma 8 gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^2 \tag{78}$$

$$\le \frac{\Delta_0}{K\eta_{-1}} + \frac{\sum_{k=0}^{K-1} B_k}{2K} + \frac{4}{\eta^2 K}\left(\Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big)\right)^2 + \frac{L^2\eta^2}{K}$$

$$= \frac{\Delta_0}{K\eta_{-1}} + \frac{1}{2}\frac{1}{K}\sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^2 + \frac{4}{\eta^2 K}\left(\Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{K-1}}{S_{\text{init}}}\Big)\right)^2 + \frac{L^2\eta^2}{K}. \tag{79}$$

Lastly, by rearranging the term and using the bound for $\log\left(\frac{S_{K-1}}{S_{\text{init}}}\right)$ in Lemma 11, using

$$\min_{0\le k<K} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_* \;\le\; \sqrt{\min_{0\le k<K} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_*^2}\,.$$

and

$$\sqrt{a+b+c} \;\le\; \sqrt{a} \;+\; \sqrt{b} \;+\; \sqrt{c}, \quad \text{for } a,b,c \ge 0,$$

we finishes the proof. ■

F.6. Per-particle adaptive LMO particle descent

In this subsection, we consider a per-particle AdaGrad variant of the LMO update, where each particle maintains its own accumulator and stepsize. We give a non-convex convergence bound in the Wasserstein setting for the case $\tau = 0$. The framework of the proof generally follows the idea in Attia and Koren [3].

Theorem 13 (Non-convex analysis of per-particle AdaGrad LMO descent, $\tau = 0$) *Under Asm 2, let $(\mu_k)_{k\ge 0}$ be the sequence of empirical measures produced by Algorithm 1 with $\tau = 0$, i.e.,*

$$\boldsymbol{v}_k^i \;=\; \mathrm{lmo}(\boldsymbol{g}_k^i; 0), \qquad \boldsymbol{g}_k^i := \nabla_{\boldsymbol{x}} \tfrac{\delta F}{\delta \mu}(\mu_k)(X_k^i),$$

and per-particle stepsizes $\eta_k^i = \eta/\sqrt{S_k^i}$ where $S_k^i = S_{init} + \sum_{s=0}^{k} \|v_s^i\|^2$ and $S_{init} > 0$. Define the suboptimality $\Delta_k := F(\mu_k) - F^\star$. Then, for any $K \ge 1$,

$$\min_{0\le k<K} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{g}_k^i\|_* \;\le\; \frac{\sqrt{S_{init}+K}}{K\eta}\left(\Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{init}+K}{S_{init}}\Big)\right). \tag{80}$$

Proof At iteration k, the update can be written as a pushforward

$$\mu_{k+1} = (T_k)_{\#}\mu_k, \qquad T_k(\boldsymbol{x}) = \boldsymbol{x} + \eta_k(\boldsymbol{x})\,\boldsymbol{v}_k(\boldsymbol{x}),$$

where $\boldsymbol{v}_k(\boldsymbol{x}) = \boldsymbol{v}_k^i$ and $\eta_k(\boldsymbol{x}) = \eta_k^i$ whenever $\boldsymbol{x} = \boldsymbol{x}_k^i$, and

$$g_k(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \tfrac{\delta F}{\delta \mu}(\mu_k)(\boldsymbol{x}).$$

By Asm 2 (Eq. (85)), applied with T_k, we obtain

$$\begin{aligned} F(\mu_{k+1}) - F(\mu_k) &\le \int \big\langle g_k(\boldsymbol{x}),\, T_k(\boldsymbol{x}) - \boldsymbol{x}\big\rangle \,\mathrm{d}\mu_k(\boldsymbol{x}) \;+\; \frac{L}{2}\int \big\|T_k(\boldsymbol{x}) - x\big\|^2 \,\mathrm{d}\mu_k(\boldsymbol{x}) \\ &= \int \big\langle g_k(\boldsymbol{x}),\, \eta_k(\boldsymbol{x})\,\boldsymbol{v}_k(\boldsymbol{x})\big\rangle \,\mathrm{d}\mu_k(\boldsymbol{x}) \;+\; \frac{L}{2}\int \|\eta_k(\boldsymbol{x})\,\boldsymbol{v}_k(\boldsymbol{x})\|^2 \,\mathrm{d}\mu_k(\boldsymbol{x}). \end{aligned}$$

Since μ_k is empirical, $\mu_k = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{x}_k^i}$, we can write

$$F(\mu_{k+1}) - F(\mu_k) \;\le\; \frac{1}{N}\sum_{i=1}^{N}\Big(\eta_k^i \langle \boldsymbol{g}_k^i, \boldsymbol{v}_k^i\rangle + \frac{L}{2}(\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2\Big).$$

For $\tau = 0$, the LMO identities (Lemma 7) give

$$\langle \boldsymbol{g}_k^i, \boldsymbol{v}_k^i \rangle = -\|\boldsymbol{g}_k^i\|_*, \qquad \|\boldsymbol{v}_k^i\|^2 \in \{0, 1\},$$

where $\|\boldsymbol{v}_k^i\|^2 = 1$ whenever $\boldsymbol{g}_k^i \neq 0$, and $\|\boldsymbol{v}_k^i\|^2 = 0$ when $\boldsymbol{g}_k^i = 0$. Hence,

$$F(\mu_{k+1}) - F(\mu_k) \;\leq\; -\frac{1}{N}\sum_{i=1}^{N} \eta_k^i \|\boldsymbol{g}_k^i\|_* \;+\; \frac{L}{2N}\sum_{i=1}^{N} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2. \tag{81}$$

Rearranging and using $\Delta_k = F(\mu_k) - F^\star$ gives

$$\frac{1}{N}\sum_{i=1}^{N} \eta_k^i \|\boldsymbol{g}_k^i\|_* \;\leq\; \Delta_k - \Delta_{k+1} \;+\; \frac{L}{2N}\sum_{i=1}^{N} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2. \tag{82}$$

Summing Eq. (82) over $k = 0, \ldots, K-1$ yields

$$\begin{aligned} \sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^{N} \eta_k^i \|\boldsymbol{g}_k^i\|_* \;&\leq\; \Delta_0 - \Delta_K \;+\; \frac{L}{2N}\sum_{k=0}^{K-1}\sum_{i=1}^{N} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2 \\ &\leq\; \Delta_0 \;+\; \frac{L}{2N}\sum_{k=0}^{K-1}\sum_{i=1}^{N} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2 \,. \end{aligned} \tag{83}$$

We now bound the double sum on the right-hand side. Fix a particle i and set

$$a_k^i := \|\boldsymbol{v}_k^i\|^2, \qquad S_k^i = S_{\text{init}} + \sum_{s=0}^{k} a_s^i, \qquad \eta_k^i = \frac{\eta}{\sqrt{S_k^i}}.$$

Then

$$\sum_{k=0}^{K-1} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2 = \eta^2 \sum_{k=0}^{K-1} \frac{a_k^i}{S_k^i}.$$

By Lemma 9 applied to the sequence $\{a_k^i\}_{k\geq 0}$ with initial offset S_{init}, we have

$$\sum_{k=0}^{K-1} \frac{a_k^i}{S_{\text{init}} + \sum_{s=0}^{k} a_s^i} \;\leq\; \log\Big(S_{\text{init}} + \sum_{s=0}^{K-1} a_s^i\Big) - \log(S_{\text{init}}) = \log\Big(\frac{S_K^i}{S_{\text{init}}}\Big).$$

Thus

$$\sum_{k=0}^{K-1} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2 \;\leq\; \eta^2 \log\Big(\frac{S_K^i}{S_{\text{init}}}\Big).$$

Summing over i and using Jensen's inequality ($\frac{1}{N}\sum_i \log u_i \leq \log(\frac{1}{N}\sum_i u_i)$), we obtain

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2 &\leq \eta^2 \cdot \frac{1}{N}\sum_{i=1}^{N} \log\Big(\frac{S_K^i}{S_{\text{init}}}\Big) \\ &\leq \eta^2 \log\Big(\frac{1}{N}\sum_{i=1}^{N} \frac{S_K^i}{S_{\text{init}}}\Big) = \eta^2 \log\Big(\frac{\overline{S}_K}{S_{\text{init}}}\Big), \end{aligned}$$

where $\overline{S}_K := \frac{1}{N}\sum_{i=1}^N S_K^i$. Moreover, since $a_k^i = \|\boldsymbol{v}_k^i\|^2 \le 1$ for all k, we have

$$S_K^i = S_{\text{init}} + \sum_{s=0}^{K-1} a_s^i \;\le\; S_{\text{init}} + K, \qquad \text{so} \qquad \overline{S}_K \le S_{\text{init}} + K.$$

Therefore

$$\frac{1}{N}\sum_{i=1}^N \sum_{k=0}^{K-1} (\eta_k^i)^2 \|\boldsymbol{v}_k^i\|^2 \;\le\; \eta^2 \log\Big(\frac{S_{\text{init}} + K}{S_{\text{init}}}\Big).$$

Plugging this into Eq. (83) gives

$$\sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^N \eta_k^i \|\boldsymbol{g}_k^i\|_* \;\le\; \Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{\text{init}} + K}{S_{\text{init}}}\Big). \tag{84}$$

Next, we lower bound the left-hand side by using the pointwise lower bound on the stepsizes. For every i and k,

$$S_k^i \le S_{\text{init}} + \sum_{s=0}^{K-1} a_s^i \le S_{\text{init}} + K, \qquad \Rightarrow \qquad \eta_k^i = \frac{\eta}{\sqrt{S_k^i}} \;\ge\; \frac{\eta}{\sqrt{S_{\text{init}} + K}}.$$

Hence

$$\frac{\eta}{\sqrt{S_{\text{init}} + K}} \sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_k^i\|_* \;\le\; \sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^N \eta_k^i \|\boldsymbol{g}_k^i\|_*.$$

Combining this with Eq. (84), we obtain

$$\sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_k^i\|_* \;\le\; \frac{\sqrt{S_{\text{init}} + K}}{\eta}\left(\Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{\text{init}} + K}{S_{\text{init}}}\Big)\right).$$

Dividing both sides by K gives

$$\min_{0\le k<K} \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_k^i\|_* \le \frac{1}{K}\sum_{k=0}^{K-1} \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{g}_k^i\|_* \;\le\; \frac{\sqrt{S_{\text{init}} + K}}{K\eta}\left(\Delta_0 + \frac{L}{2}\eta^2 \log\Big(\frac{S_{\text{init}} + K}{S_{\text{init}}}\Big)\right).$$

This completes the proof. ∎

Appendix G. Intuition on the adaptive step size design

Given Asm 2, we can derive the following lemma.

Lemma 12 (Upper bound of F by Asm 2) Under Asm 2, for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and every measurable transport map $T \in L^2(\mu; \|\cdot\|)$, we have:

$$F(T_\#\mu) \;\le F(\mu) + \int \Big\langle \nabla \tfrac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x}),\, T(\boldsymbol{x}) - \boldsymbol{x} \Big\rangle \,\mathrm{d}\mu(\boldsymbol{x}) + \frac{L}{2}\int \|T(\boldsymbol{x}) - \boldsymbol{x}\|^2 \,\mathrm{d}\mu(\boldsymbol{x}). \tag{85}$$

Remark: Eq. (85) is an analogue of the descent lemma in Euclidean space, and the proof can be found in Sec. F.2.

To understand why the adaptive step size appears naturally, let us start by freezing the step sizes and examining the basic descent mechanism. The analysis above is based on the Wasserstein-type descent lemma Eq. (85), plugging the LMO update $T_k(\boldsymbol{x}) = \boldsymbol{x} + \eta \boldsymbol{v}_k(\boldsymbol{x})$ gives the one-step inequality:

$$F(\mu_{k+1}) - F(\mu_k) \leq \frac{\eta}{N}\sum_{i=1}^{N}\langle \boldsymbol{g}_k^i, \boldsymbol{v}_k^i\rangle + \frac{L\eta^2}{2N}\sum_{i=1}^{N}\|\boldsymbol{v}_k^i\|^2 = -\frac{\eta}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^{1+\tau} + \frac{L\eta^2}{2N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^{2\tau},$$

where the last inequality is by Lemma 7. Summing $k = 0, \ldots, K-1$, rearranging, and dividing by η yields

$$\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^{1+\tau} \leq \frac{\Delta_0}{\eta} + \frac{L\eta}{2}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^{2\tau}. \tag{86}$$

The right-hand side is minimized at the optimal stepsize:

$$\eta^\star = \sqrt{\frac{2\Delta_0}{L\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_*^{2\tau}}}\,.$$

Plugging $\eta^\star$ back into Eq. (86) gives the following rates for $\tau = 0$: $\min_k \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{g}_k^i\|_* \leq \sqrt{\frac{2\Delta_0 L}{K}}$. Thus, under the standard quadratic Wasserstein-type descent lemma Eq. (85), we can achieve the $\mathcal{O}(K^{-1/2})$ stationarity rate. The adaptive step size designed in Alg. 1 is precisely a computable surrogate for the ideal but non-implementable $\eta^\star$.

Appendix H. Annealing dynamics

We consider the following LMO noisy particle descent rule, which augments the standard noisy particle gradient descent in Eq. (8) with an LMO oracle:

$$\boldsymbol{x}_{k+1}^i = \boldsymbol{x}_k^i + \eta\ \mathrm{lmo}\Big(\nabla\tfrac{\delta F}{\delta\mu}(\mu_{\boldsymbol{X}_k})(\boldsymbol{x}_k^i);\ 0\Big) + \sqrt{2\lambda\eta}\,\boldsymbol{\xi}_k^i\,, \tag{87}$$

for $i \in [N]$, where $\boldsymbol{\xi}_k^i \sim \mathcal{N}(0, I_d)$ are i.i.d. Gaussian random variables. Passing formally to the limits $\eta \to 0$ and $N \to \infty$ motivates the study of the following nonlinear Fokker–Planck equation:

$$\begin{aligned} &\partial_t \mu_t = \nabla\cdot\big(\mu_t\,\boldsymbol{v}_t\big), \\ &\boldsymbol{v}_t := -\,\mathrm{lmo}\Big(\nabla\tfrac{\delta F}{\delta\mu}(\mu_t); 0\Big) + \lambda_t\nabla\log\mu_t\,. \end{aligned} \tag{88}$$

Compared with Eq. (3), the key difference is that the first variation term is now passed solely through an LMO operator and the entropy regularization is allowed to vary over time. We emphasize that this is an annealing dynamics, where the regularization parameter λ_t decays to zero as $t \to \infty$, following the annealing framework for MFLD introduced in Chizat [12]. For our analysis, we introduce a Polyak-Lojasiewicz(PL)-type assumptions in Wasserstein space.

Assumption 8 (Wasserstein PL inequality) *Let $\mu^\star \in \mathcal{P}_2(\mathbb{R}^d)$ be a minimizer of F, and denote $F^\star := F(\mu^\star)$. Define $g_\mu(x) := \nabla_x\frac{\delta F}{\delta\mu}(\mu)(x)$. We assume that there exists $\alpha > 0$ such that for every μ, $\|g_\mu\|\_L^2(\mu;\|\cdot\|_*)^2 \geq 2\alpha\big(F(\mu) - F^\star\big)$.*

Remark 14 *Note that such assumption have been studied in prior work [7, 15].*

We also need the following regularity assumption along the trajectory.

Assumption 9 (Divergence control along the trajectory) *Consider the dynamics Eq. (88), denote by* $g_\mu(\boldsymbol{x}) := \nabla \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x})$. *we assume there exists* $C_{\mathrm{div}} < \infty$ *such that* $\forall t \geq 0$, $\left| \int (\nabla \cdot g_{\mu_t}) \, \mathrm{d}\mu_t \right| \leq C_{\mathrm{div}}$.

Now we have the following convergence guarantee for such annealing LMO dynamics.

Theorem 15 *Under Asm. 4, 8 and 9, consider the flow Eq. (88) with* $\lambda_t = \frac{\gamma}{\log(e+t)}$ *for some* $\gamma > 0$, *and assume the solution exists, denoted by* $\{\mu_t\}_{t \geq 0}$. *Then there exists* $t_0 \geq 0$ *such that for all* $t \geq t_0$, *we have:*

$$F(\mu_t) - F(\mu^\star) = O\big(\log^{-2} t\big) \, .$$

Proof [Proof of Thm. 15] Denote by $\mathcal{E}(t) := F(\mu_t) - F(\mu^\star)$. Since μ_t solves the continuity equation $\partial_t \mu_t = -\nabla \cdot (\mu_t v_t)$ with $v_t = s_t - \lambda_t \nabla \log \mu_t$ and $s_t(\boldsymbol{x}) := \mathrm{lmo}(g_{\mu_t}(\boldsymbol{x}); 0)$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = \int_{\mathbb{R}^d} \big\langle g_t(\boldsymbol{x}), v_t(\boldsymbol{x}) \big\rangle \, \mathrm{d}\mu_t(\boldsymbol{x}), \qquad g_t := g_{\mu_t}. \tag{89}$$

By the LMO property on the unit ball, $\langle g_t(\boldsymbol{x}), s_t(\boldsymbol{x}) \rangle = -\|g_t(\boldsymbol{x})\|_*$ for μ_t-a.e. $\boldsymbol{x}$, hence

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\int_{\mathbb{R}^d} \|g_t(\boldsymbol{x})\|_* \, \mathrm{d}\mu_t(\boldsymbol{x}) \; - \; \lambda_t \int_{\mathbb{R}^d} \big\langle g_t(\boldsymbol{x}), \nabla \log \mu_t(\boldsymbol{x}) \big\rangle \, \mathrm{d}\mu_t(\boldsymbol{x}). \tag{90}$$

Next, we use integration by parts for the cross term. Write $\mathrm{d}\mu_t(\boldsymbol{x}) = \rho_t(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$. Then $\nabla \log \mu_t = \nabla \rho_t / \rho_t$ and

$$\int \langle g_t, \nabla \log \mu_t \rangle \, \mathrm{d}\mu_t = \int \langle g_t, \nabla \rho_t \rangle \, \mathrm{d}\boldsymbol{x} = -\int (\nabla \cdot g_t) \, \rho_t \, \mathrm{d}\boldsymbol{x} = -\int (\nabla \cdot g_t) \, \mathrm{d}\mu_t,$$

Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) = -\|g_t\|_{L^1(\mu_t; \|\cdot\|_*)} \; + \; \lambda_t \int_{\mathbb{R}^d} (\nabla \cdot g_t)(\boldsymbol{x}) \, \mathrm{d}\mu_t(\boldsymbol{x}). \tag{91}$$

By Assumption 9,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) \leq -\|g_t\|_{L^1(\mu_t; \|\cdot\|_*)} \; + \; \lambda_t C_{\mathrm{div}}. \tag{92}$$

Now,by Asm 4, Asm 8 :

$$\|g_t\|_{L^1(\mu_t; \|\cdot\|_*)} \geq \kappa \|g_t\|_{L^2(\mu_t; \|\cdot\|_*)} \geq \kappa \sqrt{2\alpha \, \mathcal{E}(t)} \, .$$

Plugging the previous bound to Eq. (92) gives

$$\mathcal{E}'(t) \leq -\kappa \sqrt{2\alpha \, \mathcal{E}(t)} + \lambda_t C_{\mathrm{div}},$$

Now specialize to the schedule $\lambda_t = \gamma / \log(e + t)$ and define

$$b(t) := \frac{C_{\mathrm{div}} \gamma}{\log(e + t)}, \qquad y(t) := \sqrt{\mathcal{E}(t)}.$$

Since $\mathcal{E}$ is nonnegative and absolutely continuous, y is absolutely continuous and differentiable a.e. on $\{y > 0\}$, with $\mathcal{E}'(t) = 2y(t)y'(t)$. Thus, for a.e. t with $y(t) > 0$,

$$2y(t)y'(t) \le -\kappa\sqrt{2\alpha}\, y(t) + b(t). \tag{93}$$

Dividing by $2y(t)$ yields

$$y'(t) \le -\frac{\kappa\sqrt{2\alpha}}{2} + \frac{b(t)}{2y(t)}. \tag{94}$$

Define the barrier

$$h(t) := \frac{2}{\kappa\sqrt{2\alpha}}\, b(t) = \frac{2C_{\mathrm{div}}\gamma}{\kappa\sqrt{2\alpha}\, \log(e+t)}.$$

Whenever $y(t) \ge h(t)$, we have $b(t)/(2y(t)) \le \kappa\sqrt{2\alpha}/4$, and Eq. (94) implies

$$y'(t) \le -\frac{\kappa\sqrt{2\alpha}}{4} \qquad \text{for a.e. such } t. \tag{95}$$

Fix any $T \ge 0$. If $y(t) \ge h(t)$ for all $t \ge T$, then integrating Eq. (95) yields

$$y(t) \le y(T) - \frac{\kappa\sqrt{2\alpha}}{4}(t - T) \xrightarrow[t\to\infty]{} -\infty,$$

a contradiction. Hence, for every $T \ge 0$ there exists $t \ge T$ such that

$$y(t) \le h(t). \tag{96}$$

Since h is C^1 and strictly decreasing,

$$h'(t) = -\frac{2C_{\mathrm{div}}\gamma}{\kappa\sqrt{2\alpha}\,(e+t)\,\log^2(e+t)}.$$

As $h'(t) \uparrow 0$ when $t \to \infty$, there exists $t_* \ge 0$ such that

$$h'(t) \ge -\frac{\kappa\sqrt{2\alpha}}{4} \qquad \forall\, t \ge t_*. \tag{97}$$

Choose $t_0 \ge t_*$ such that $y(t_0) \le h(t_0)$. We claim that

$$y(t) \le h(t) \qquad \forall\, t \ge t_0.$$

Otherwise, define

$$\hat{t} := \inf\{\, t \ge t_0 :\ y(t) > h(t)\,\}.$$

Then $y(\hat{t}) = h(\hat{t})$ and for $t > \hat{t}$ sufficiently close to $\hat{t}$, $y(t) \ge h(t)$. Integrating Eq. (95) and using Eq. (97) gives

$$y(t) - y(\hat{t}) \le -\frac{\kappa\sqrt{2\alpha}}{4}(t - \hat{t}), \qquad h(t) - h(\hat{t}) \ge -\frac{\kappa\sqrt{2\alpha}}{4}(t - \hat{t}),$$

which implies $y(t) - h(t) \le 0$, a contradiction. Hence the barrier is invariant.

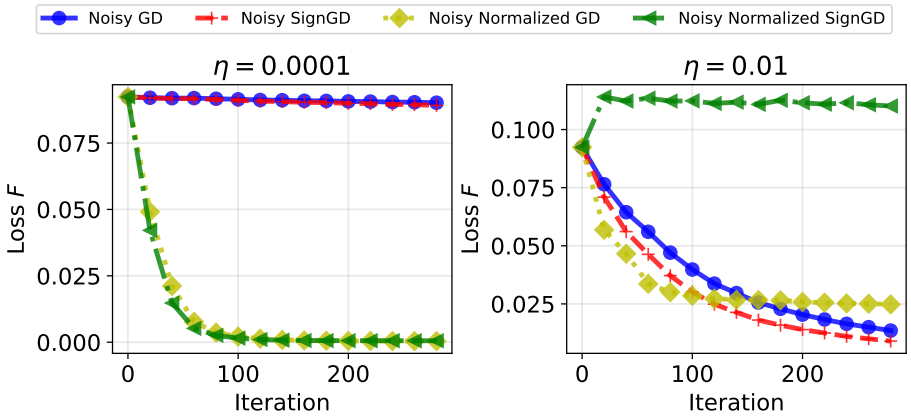

Figure 1: Results on the teacher–student regression task for the mean field Langevin dynamics with LMO noisy particle descent.

Therefore, for all $t \geq t_0$,

$$\sqrt{\mathcal{E}(t)} = y(t) \leq h(t) = \frac{2C_{\mathrm{div}}\gamma}{\kappa\sqrt{2\alpha}\,\log(e+t)},$$

and squaring yields

$$\mathcal{E}(t) \leq \frac{4C_{\mathrm{div}}^2}{2\alpha\,\kappa^2}\,\lambda_t^2 \qquad \forall\, t \geq t_0.$$

Thus, $F(\mu_t) - F(\mu^\star) = O\big(\log^{-2} t\big)$. ∎

Appendix I. Experiment

I.1. Main experiment

We consider a teacher–student regression task following the setup in Zhu and Chen [49]. Given M input samples $\{\boldsymbol{z}_j\}_{j=1}^M \subset \mathbb{R}^d$ drawn i.i.d. from the standard Gaussian distribution, the labels are generated by a teacher model $y_j = \sin(\hat{\boldsymbol{x}}^\top \boldsymbol{z}_j), \quad j = 1, \ldots, M$, where $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ is a fixed random vector drawn from the standard Gaussian distribution. The student model is a two-layer neural network with N particles with tanh activation and mean-field scaling: $h(\boldsymbol{x}_i, \boldsymbol{z}) = \tanh(\boldsymbol{x}_i^\top \boldsymbol{z})$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes the weight vector of the i-th neuron and $\boldsymbol{X} = (\boldsymbol{x}_i)_{i=1}^N \in \mathbb{R}^{N\times d}$. We consider the entropy regularized objective in Eq. (1), with $F(\mu_{\boldsymbol{X}}) = \frac{1}{M}\sum_{j=1}^M \big(\frac{1}{N}\sum_{i=1}^N h(\boldsymbol{x}_i, \boldsymbol{z}_j) - y_j\big)^2$ and $\lambda = 0.01$. We compare different LMO noisy particle descent methods based on Eq. (87), including noisy GD, noisy SignGD, noisy normalized GD, and noisy normalized SignGD. We choose both a smaller and a larger step size: $\eta = 10^{-4}$ and $\eta = 10^{-2}$. The convergence results are shown in Fig. 1, where each curve represents the mean over three random seeds. For a very small step size, *noisy GD and noisy SignGD essentially do not make progress, while normalized methods converge rapidly*. For a larger step size, unnormalized methods converge faster compared with normalized methods. This behavior is consistent with our previous theoretical findings, which show that the normalized LMO

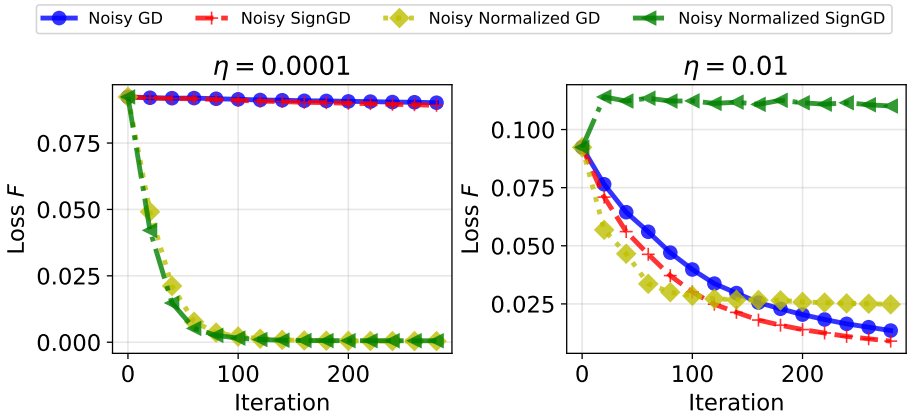

Figure 2: Results on the teacher–student regression task for the mean field Langevin dynamics with LMO noisy particle descent. The dimension of the neuron $d = 100$.

with $\tau = 0$ enjoys faster convergence than the unnormalized one in the gradient flow setting, i.e., with small step size.

Additional experiments on adaptive step-size schemes, comparing global and per-particle variants under different initialization settings, are provided in Sec. I of the appendix. These results are consistent with our main findings, showing that normalized steepest descent converges faster in the small step-size regime.

I.2. Additional experiment

In Fig. 1 of the previous section, the result is under $d = 2$. In this section, we set $d = 100$ under the same teacher–student regression task, the additional result is given in Fig. 2. We can still see that for a very small step size, noisy GD and noisy SignGD essentially make much less progress compared to normalized methods. And such a gap becomes smaller as the step size increases.

Next, we run experiments regarding the adaptive stepsize scheme developed in Sec. 4.1. Under the same teacher-student regression setting as in Sec. I.1, we conducted additional experiments comparing global adaptive and per-particle adaptive step sizes for both normalized and unnormalized methods. The result is provided in Fig. 3. We have the following observations: Firstly, the results are consistent with our original findings: normalized methods converge faster in the small step-size regime than the unnormalized method.

The difference between global adaptive and per-particle adaptive step sizes is extremely small in this setting. Since all particles are initialized in the same way, their trajectories tend to remain similar, leading to comparable per-particle quantities. Combined with a relatively large number of particles ($m = 100$) in this experiment, this can result in very similar accumulators and hence nearly identical effective step sizes and convergence behavior. Our original paper only shows that both achieve the same convergence rate instead of distinguishing between these two variants.

To further investigate potential differences, we design an additional experiment to amplify particle difference. Specifically, we reduce the number of particles from $N = 100$ to $N = 3$, and replace the original standard Gaussian initialization with a log-normal scaling initialization of the

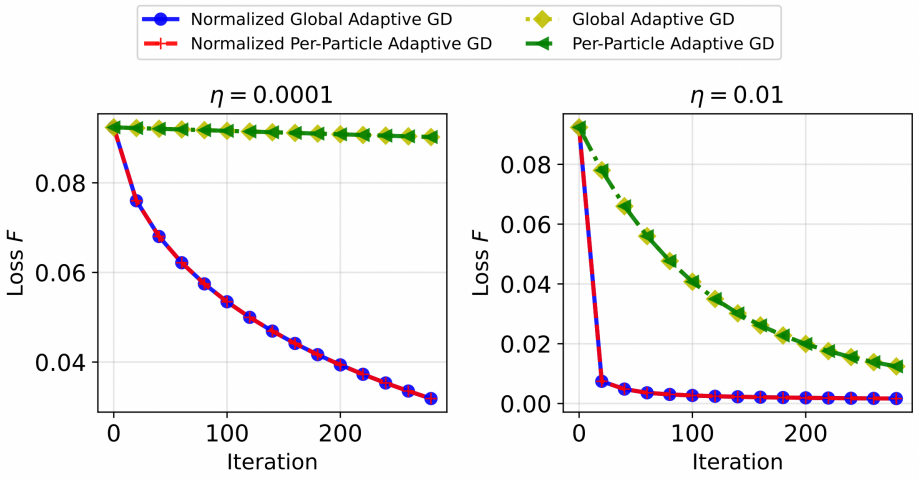

Figure 3: Results on the teacher–student regression task for comparing global adaptive stepsize and per-particle adaptive stepsize.

form $x_i^0 = \alpha_i \widetilde{x}_i, \quad \widetilde{x}_i \sim \mathcal{N}(0, I_d), \quad \alpha_i = \exp(\xi_i),\ \xi_i \sim \mathcal{N}(0,1)$, independently across particles $i = 1, \ldots, N$. This increases the variability of their magnitudes.

The results are shown in Fig. 4. Under the normalized scheme, both global and per-particle adaptive methods remain essentially identical as in previous experiment, since the step size does not depend on the magnitude of past gradients.

However, for the unnormalized methods, we observe a noticeable (though still moderate) difference, where the global adaptive method performs slightly better than the per-particle adaptive one. While we do not provide a formal theoretical analysis for this phenomenon, an explanation is that when N is small and particle-wise gradient magnitudes differ significantly under different initialization, the global adaptive scheme averages these variations, leading to a more stable step size, whereas the per-particle scheme adapts each particle independently and may introduce higher variance in the updates.

Appendix J. Broader Impact

This paper provides a theoretical analysis of optimization methods for language model training, with a focus on normalized steepest descent dynamics in Wasserstein space and their particle-based implementations. The primary goal of this work is to improve the theoretical understanding of large-scale non-convex model training, which may indirectly contribute to reducing computational costs and energy consumption in modern machine learning systems. From a methodological perspective, our results provide principled guidance for the analysis of geometry-aware optimization algorithms in Wasserstein space, and do not constitute or evaluate applications in any specific domain. We do not foresee direct negative societal or ethical consequences arising uniquely from this work, and this work does not raise new ethical concerns.

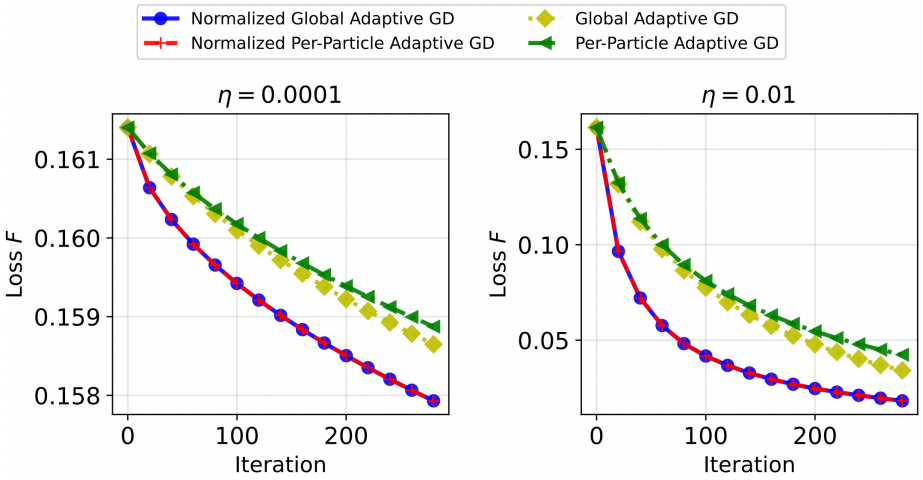

Figure 4: Results on the teacher–student regression task for comparing global adaptive stepsize and per-particle adaptive stepsize with log-normal scaling initialization.

Table 1: Core symbols and constants used in this paper.

Symbol	Type	Description
		Optimization on $\mathbb{R}^d$
$\boldsymbol{x}_k,\ \boldsymbol{x}(t)$	$\mathbb{R}^d$	Parameter at iteration k or time t
$f(\boldsymbol{x})$	$\mathbb{R}^d \to \mathbb{R}$	Objective in Euclidean space
$f^\star$	$\mathbb{R}$	Optimal value $f^\star := \min_{\boldsymbol{x}} f(\boldsymbol{x})$
$\nabla f(\boldsymbol{x})$	$\mathbb{R}^d$	Gradient of f at $\boldsymbol{x}$
$\|\cdot\|,\ \|\cdot\|_*$	$-$	A norm on $\mathbb{R}^d$ and its dual: $\|\boldsymbol{y}\|_* := \sup_{\|\boldsymbol{x}\|\le 1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle$
$\mathrm{lmo}(\boldsymbol{g};\tau)$	$\mathbb{R}^d$	LMO direction (any minimizer), Def. 1
η_k	>0	Stepsize at iteration k
Δ_k	$\mathbb{R}_{\ge 0}$	Suboptimality: $\Delta_k := f(\boldsymbol{x}_k) - F^\star$
		Optimization in measure space and particle systems
$\mathcal{P}_2(\mathbb{R}^d)$	$-$	Probability measures on $\mathbb{R}^d$ with finite second moment
$\mu,\ \mu_t$	$\mathcal{P}_2(\mathbb{R}^d)$	Measure (generic / continuous time)
$F(\mu)$	$\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$	Energy functional on measures
$\frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x})$	$\mathbb{R}$	First variation of F at μ evaluated at $\boldsymbol{x}$
$g_\mu(\boldsymbol{x})$	$\mathbb{R}^d$	Wasserstein gradient field: $g_\mu(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu)(\boldsymbol{x})$
$\mathscr{F}(\mu)$	$\mathbb{R}$	Objective in measure space: $\mathscr{F}(\mu) := F(\mu) + \lambda\,\mathrm{Ent}(\mu)$
$\mathrm{Ent}(\mu)$	$\mathbb{R}$	(Negative) Shannon entropy: if μ has density ρ, $\mathrm{Ent}(\mu) = \int \rho(\boldsymbol{x}) \log \rho(\boldsymbol{x})\, d\boldsymbol{x}$
$W_2(\mu,\nu)$	$\mathbb{R}_{\ge 0}$	2-Wasserstein distance between μ, ν
$\boldsymbol{X}_k = (\boldsymbol{x}_k^i)_{i=1}^N$	$(\mathbb{R}^d)^N$	Collection of particles at iteration k
$\mu_{\boldsymbol{X}_k}$	$\mathcal{P}_2(\mathbb{R}^d)$	Empirical measure: $\mu_{\boldsymbol{X}_k} := \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_k^i}$ (and often $\mu_k := \mu_{\boldsymbol{X}_k}$)
$\mu^{(N)},\ \mu_t^{(N)},\ \mu_k^{(N)}$	$\mathcal{P}_2(\mathbb{R}^{dN})$	Law of the N-particle system on $\mathbb{R}^{dN}$ (continuous / discrete time)
$\mathscr{F}^{(N)}(\mu^{(N)})$	$\mathbb{R}$	Finite-particle free energy on $\mathcal{P}_2(\mathbb{R}^{dN})$
$\mu_*^{(N)}$	$\mathcal{P}_2(\mathbb{R}^{dN})$	Minimizer of $\mathscr{F}^{(N)}$ (finite-particle optimum)
$(\mu^\star)^{\otimes N}$	$\mathcal{P}_2(\mathbb{R}^{dN})$	N-fold product of the mean-field minimizer $\mu^\star$
$\boldsymbol{\xi},\ \boldsymbol{w}$	$\mathbb{R}^d$	Gaussian noise / Brownian motion
N	$\mathbb{N}$	Number of particles
$\boldsymbol{g}_k^i$	$\mathbb{R}^d$	Mean-field gradient at particle: $\boldsymbol{g}_k^i := \nabla_{\boldsymbol{x}} \frac{\delta F}{\delta \mu}(\mu_{\boldsymbol{X}_k})(\boldsymbol{x}_k^i)$
$\boldsymbol{v}_k^i$	$\mathbb{R}^d$	Drift / update direction at particle i (cf. Eq. Eq. (5))
		Constants in assumptions and rates
L	>0	Smoothness / transport-smoothness constant (Assumps. 5.A, 2)
α	>0	Strong convexity constant (Assumps. 5.B)
κ	$(0,1]$	Non-degeneracy constant (Assump. 4)
ζ	>0	Log-Sobolev constant (Assump. 3)
C_1	>0	Norm-comparison constant on $\mathbb{R}^{dN}$ used in Thm. 10: $C_1 := \sup_{\boldsymbol{a}\neq 0} \frac{\|\boldsymbol{a}\|^2}{\|\boldsymbol{a}\|_*^2}$
M_1	>0	Uniform drift bound (Assump. 7)
M_2	>0	Score Lipschitz constant (Assump. 7)

Table 2: Examples of LMO and the updates in Euclidean space.[1]

Instance	Norm	τ	$\mathrm{lmo}(\boldsymbol{g};\tau)$	Update
Unnormalized steepest descent				
GD [36] / Adam [23]	ℓ_2	1	$-\boldsymbol{g}$	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f(\boldsymbol{x}_k)$
SignGD [9]	ℓ_∞	1	$-\lVert\boldsymbol{g}\rVert_1 \operatorname{sign}(\boldsymbol{g})$	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \lVert\nabla f(\boldsymbol{x}_k)\rVert_1 \operatorname{sign}(\nabla f(\boldsymbol{x}_k))$
Spectral descent [9]	Spectral	1	$-\sum_i \sigma_i \boldsymbol{u}\boldsymbol{v}^\top$	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \sum_i \sigma_i \boldsymbol{u}\boldsymbol{v}^\top$
Normalized steepest descent				
Normalized GD [20]	ℓ_2	0	$-\dfrac{\boldsymbol{g}}{\lVert\boldsymbol{g}\rVert_2}$	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \dfrac{\nabla f(\boldsymbol{x}_k)}{\lVert\nabla f(\boldsymbol{x}_k)\rVert_2}$
Normalized SignGD [4]	ℓ_∞	0	$-\operatorname{sign}(\boldsymbol{g})$	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \operatorname{sign}(\nabla f(\boldsymbol{x}_k))$
Muon [22] / Scion [35]	Spectral	0	$-\boldsymbol{u}\boldsymbol{v}^\top$	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \boldsymbol{u}\boldsymbol{v}^\top$

[1] In the spectral norm case, the gradient $\boldsymbol{g} = \nabla f(\boldsymbol{x}_k)$ is a matrix with singular value decomposition $\nabla f(\boldsymbol{x}_k) = \boldsymbol{u}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)\,\boldsymbol{v}^\top$, and the dual norm is the nuclear norm $\lVert\nabla f(\boldsymbol{x}_k)\rVert_* = \sum_i \sigma_i$.