OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Data Generation without Function Estimation

Hadi Daneshmand DHADI@VIRGINIA.EDU
Department of Computer Science, University of Virginia
Ashkan Soleymani ASHKANSO@MIT.EDU
Department of Electrical Engineering and Computer Science, MIT

Abstract

Estimating the score function—or other population-density-dependent functions —is a fundamental component of most generative models. However, such function estimation is computationally and statistically challenging. *Can we avoid function estimation for data generation?* We propose an **estimation-free** generative method: *A set of points* whose locations are *deterministically* updated with (inverse) *gradient descent* can transport a uniform distribution to arbitrary data distribution, in the mean field regime, **without function estimation, training neural networks, and even noise injection**. The proposed method is built upon recent advances in the physics of interacting particles. Leveraging recent advances in mathematical physics, we prove that the proposed method samples from the true underlying data distribution in the asymptotic regime, without making any strong structural assumptions on the distribution.

1. Introduction

Given i.i.d. samples from an unknown distribution, how to generate a new sample from the distribution? Existing generative models often rely on estimating functions depending on population density such as the score function. Such an estimation is both statistically and computationally challenging. Computationally, estimating the score function even for a simple Gaussian mixture model with maximum likelihood estimation of parameters is NP-hard [8]. Statistically, score estimation suffers the curse of dimensionality [45]. This raises a fundamental question: can we avoid function estimation for data generation? Ultimately, what is typically available during training is an empirical distribution over i.i.d. samples, for which the score function is not even defined. Can we generate new samples directly from this discrete empirical distribution, avoiding (score) function estimation?

We demonstrate that data generation can be achieved without function estimation. Our proposed approach is a novel generative method operating entirely on the empirical distribution of a finite set of training samples. We construct an interactive system that iteratively updates the positions of these data points in two phases: (1) applying standard gradient descent to shape data points (2) performing an inverse gradient descent step to generate new samples. This method builds upon recent advances in the theoretical study of systems of interacting particles [18, 36]. Leveraging this rich literature, we prove that the proposed estimation-free method can transport a uniform measure over a finite ball/sphere to an arbitrary data distribution in the mean-field regime, where the number of samples tends to infinity. Finally, we experimentally validate estimation-free data generation using a finite number of points.

2. Method

Given the support of an empirical distribution over points $x_1, \ldots, x_n \in \mathbb{R}^d$, consider the following optimization problem:

$$x_1^*, \ldots, x_n^* := \arg \min_{x_1,\ldots,x_n \in \mathbb{R}^d} \left(E_n^{(\epsilon)}(x_1, \ldots, x_n) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} W_\epsilon^{(s)}(x_i - x_j) \right), \quad (1)$$

where $W_\epsilon^{(s)}(x-y) = \frac{\|x-y\|^2}{2} + \frac{1}{s(\|x-y\|^2+\epsilon)^{s/2}}$ is called a "attractive-repulsive power-law interaction potential" [9, 37]. The squared-norm term encourages attraction between particles, while the inverse-norm term induces repulsion, preventing collapse[1]. These competing forces lead to a striking structure in the distribution of the optimized points $x_1^*, \ldots, x_n^*$ in the asymptotic regime as $n \to \infty$, where the limiting distribution is characterized by

$$\arg\min_{p \in \Omega} \left(E(p) := \int W^{(s)}(x-y)\, p(x)\, p(y)\, dx\, dy \right), \Omega := \{\text{probability measures over } \mathbb{R}^d\}, \quad (2)$$

where $W^{(s)} := W_0^{(s)}$. The minimizer is unique up to translation and is uniform over

$$\begin{cases} \text{a ball [14]} & \text{if } s = d-2, \\ \text{a sphere [19]} & \text{if } 2 < d \text{ and } -2 \leq s < d-4, \end{cases}$$

with a radius that is finite and computable in closed form. By optimizing the locations $x_1, \ldots, x_n$, one can asymptotically transport any empirical distribution to a simple uniform distribution over a compact ball or sphere. This optimization can be performed via standard gradient descent (GD):

$$\forall i: \quad x_i^{(k+1)} = x_i^{(k)} - \gamma \nabla_{x_i} E_n^{(\epsilon)}(x_1^{(k)}, \ldots, x_n^{(k)}). \quad (3)$$

Figure 1 plots the time evaluation of points $\{x_i^{(k)}\}$ indexed by k starting from the initial points $x_i^{(0)}$ drawn i.i.d. from Gaussian mixture. We observe that the distribution becomes uniform as k increases. Although the distribution is uniform, it contains very interesting information about the initial Gaussian mixture distribution. For example, the ratio of area of different colors are proportional to the number of points in each cluster. This structure indicates that it may be possible to recover original data distribution from the final state. Whereas in the initial state $x_i^{(0)}$ the data points are i.i.d., evolving according to (3) induces strong correlations and entanglement among them under the gradient descent dynamics. Thus, in the limit $k \to \infty$, although the marginal distribution of each data point becomes uniform, their joint distribution still conveys potentially useful information about the original distribution.

Even more interesting is the inverse of gradient descent: assuming GD converges to the minimizer of E_n, inverse GD maps points uniformly distributed on a sphere (or disk) to an arbitrary empirical distribution supported on a finite set of points. The Proposition 9 in the Appendix guarantees that this inverse exists and is computable under mild assumptions. Indeed, the inverse process can be efficiently computed using gradient descent on a convex function, without requiring the score function or any other statistical information about the population density. Combining gradient descent (GD)

1. For $s = 0$, the repulsive term is replaced by the logarithm of the norm, as given by $s \to 0$ [37].

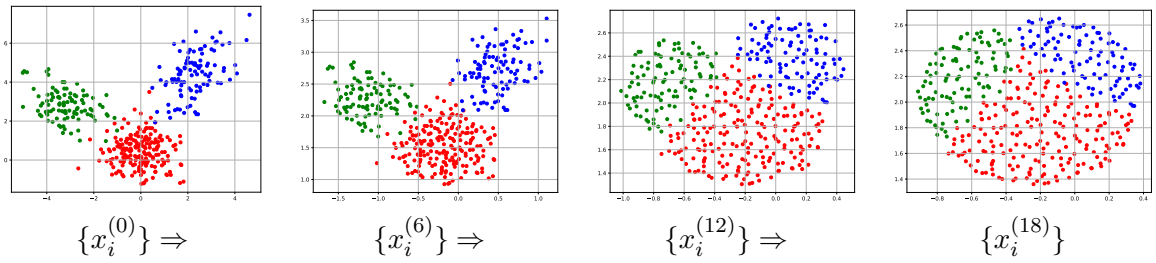


Figure 1: **Convergence of the empirical distribution with optimization.** Scatter plot of gradient descent iterates $x_i^{(k)}$, defined in (3), initialized at $x_i^{(0)}$ drawn from a Gaussian mixture distribution. Different mixture components (modes) are distinguished by color. As k evolves, the points become uniformly distributed on the circle.

and its inverse can deterministically transport an empirical distribution to a uniform distribution and vice versa. This insight enables a three-step sampling method that operates on a finite set of points: (1) apply GD to the n available samples from the target distribution; (2) add a new point drawn uniformly at random from a sphere or ball; and (3) apply inverse GD. We call this three step method **estimation-free sampling (EFS)**.

(1) Forward optimization: Apply gradient descent to $E_n^{(\epsilon)}$ (see (3)) starting from i.i.d. samples $x_1^{(0)}, \ldots, x_n^{(0)}$ drawn from the target distribution. In the next section, we show that this optimization acts as a transport map, pushing the empirical distribution toward a uniform distribution in the asymptotic regime.

(2) Augmentation: Add a new point $y^{(k)}$, sampled from a uniform distribution, to the set $x_0^{(k)}, \ldots, x_n^{(k)}$. To generate $y^{(k)}$, we first estimate the center and radius of the ball or sphere enclosing the points $x_i^{(k)}$. $y^{(k)}$ can also be generated by interpolating between two points in the set, i.e., $y^{(k)} = (1-t)x_i^{(k)} + tx_j^{(k)}$ for some $i \neq j$ and $t \in (0,1)$.

(3) Backward optimization: Compute the inverse GD for the newly generated point $y^{(k)}$ by optimizing a convex function (defined in Proposition 9), using efficient gradient descent with a constant step size. While the forward process involves n points, the backward step involves $n+1$ points, consisting of the n points from the forward step and the newly generated point from step (2). Find Algorithm details in Appendix A.

3. Theory

We demonstrate that the proposed estimation-free sampling (EFS) method can generate new samples from a distribution in the asymptotic regime. As established in Proposition 9, backward step is convex problem with well-established theoretical guarantees. In contrast, the forward optimization is non-convex. Nevertheless, its global convergence can be analyzed using recent advances in Wasserstein gradient flows on interactive-repulsive energy [14, 36]. As the potential function is shift-invariant, *all statements hold up to translation*, and we refrain from repeating "up to translation" for simplicity.

3.1. Forward Transport to Uniform Distribution

While our ultimate goal is to analyze the backward step of EFS, which generates new samples, the forward step also plays a critical role in data generation. Here, we show that the forward step of EFS *transports* the data distribution to a uniform distribution[2].

Consider a continuous-time model where the samples $x_i^{(k)}$ updated by gradient descent, modeled by the following ordinary differential equation (ODE):

$$\frac{dx_i}{dt} = -\frac{1}{n}\sum_{j\neq i} \nabla W^{(s)}(x_i - x_j), \tag{4}$$

ODEs have been widely used to describe the limiting behavior of gradient descent with an infinitesimally small step size [15, 43, 48, 49]. Analyzing the above system becomes challenging due to the coupling between the variables x_i, as the complexity increases with the number of particles n. A common approach to address this challenge is to analyze the evolution of the empirical distribution of the particles at a macroscopic level, rather than tracking their individual trajectories. Define the empirical distribution as $\mu_t^{(n)} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, where δ_x denotes the Dirac measure centered at x. Consider the following PDE:

$$\frac{d\mu}{dt} = \operatorname{div}\left(\mu_t(x)\int \nabla W^{(s)}(x-y)\mu_t(y)\,dy\right), \tag{5}$$

where div denotes the divergence operator acting on the associated vector field. Our analysis adopts the following two regularity assumptions from related work in mathematical physics:

$$\sup_{x,t}|\mu_t(x)| < \infty, \quad \sup_{y,t}\left|\int \nabla^2 W^{(s)}(y-x)\mu_t(x)dx\right| < \infty \qquad \text{(regularity assumptions)}$$

The assumptions above have been widely used in the literature and extensively studied in mathematical physics [36]. In particular, [28] showed that these assumptions hold when the initial distribution μ_0 is sufficiently regular. The following theorem, established in [36], is a key result that connects recent advances in PDE analysis with the EFS framework.

Theorem 1 (Informal Statement of Theorem 1. in [36]) *Suppose that $\mu_0^{(n)}$ converges to a sufficiently regular measure μ_0 in Wasserstein distance. There is a function $f:\mathbb{R}_+\to\mathbb{R}_+$ such that if $n \geq f(t)$, then $\mu_t^{(n)}$ converges weakly to μ_t as $t\to\infty$ and as long as $d-2\leq s<d$ and regularity assumptions hold.*

The macroscopic measure μ_t is significantly easier to analyze than the microscopic trajectories $x_i(t)$. While the energy E_n is generally non-convex in $x_i(t)$, the limiting energy E obeys linear interpolation convexity in μ [37]. This convexity can be exploited to characterize the asymptotic behavior of μ_t as $t\to\infty$ [37].

Theorem 2 ([14, 18]) *The steady state of μ_t is the global minimizer of E (up to translation) which is the uniform measure over:* $\begin{cases} \text{a ball [14]} & \text{if } d-2\leq s<d, \\ \text{a sphere [19]} & \text{if } 2<d \text{ and } -2\leq s<d-4, \end{cases}$ *with finite radius.*

2. A map f transports measure μ to ν if the distribution of $f(x)$ is ν when $x\sim\mu$ [44].

Notably, proving that a steady state (i.e., a local minimizer) is in fact a global minimizer required decades of mathematical development, beginning with the foundational work of Frostman [20], and has only recently been fully resolved in certain parameter regimes [18]. These advances allow us to analyze the asymptotic dynamics of the forward step in EFS.

3.2. Backward Transport to Target Distribution

It is straightforward to verify that the backward optimization can exactly recover the original training data $x_k^{(i)}$ by initializing EFS with $y_k = x_k^{(i)}$. However, our goal is not to reconstruct existing data points, but to generate new samples from the underlying data distribution. We show that this form of generalization is achievable in the asymptotic regime.

Consider the continuous-time formulation of the backward optimization process:

$$\frac{dy^{(n)}}{d\bar{t}} = \frac{1}{n}\sum_{i=1}^{n} \nabla W^{(s)}\left(y_t^{(n)} - x_i(t)\right), \quad \text{where } x_i(t) \text{ is defined in (4).} \tag{6}$$

Here, the differential $d\bar{t}$ indicates that the process is reversed time [5]. This dynamics corresponds to EFS in the limit of an infinitesimally small step size γ and a large terminal time $T \to \infty$. We prove this dynamics *transport* μ_t back to μ_0 in the asymptotic regime as $n \to \infty$. In other words, the distribution of $y_0^{(n)}$ converges to μ_0 as $n \to \infty$.

Theorem 3 *Suppose that $\mu_0^{(n)}$ converges to a sufficiently regular measure μ_0 in the Wasserstein-2 distance, such that regularity assumptions hold. Furthermore, assume that μ_t is not exactly the steady-state solution of the PDE in (5). Let $y_t^{(n)}$ be a random variable with law μ_t, where μ_t is the solution to the continuity equation (5). There is a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that if $n \geq f(t)$, then the distribution of $y_0^{(n)}$—obtained from the reversed-time ODE (6)—converges to μ_0 as $t \to \infty$, for continuous measures μ_0 and μ_t, $s = d - 2$.*

To generate new samples, we assume access to i.i.d. samples from the target distribution. Since the empirical measure $\mu_0^{(n)}$ converges to μ_0 in Wasserstein distance [44] which ensures the assumption holds in the last theorem. As discussed, the measure μ_t converges to a uniform distribution as $t \to \infty$. Consequently, sampling from a nearly uniform distribution allows for effective recovery of the initial data distribution μ_0.

Thus, a set of points whose locations are updated via *(inverse) gradient descent*, can provably generate new samples from an arbitrary distribution μ_0—without function estimation, training neural networks, or injecting noise. Avoiding explicit function estimation enables us to establish asymptotic guarantees for EFS without any structural assumptions on the data distribution. In comparison, methods such as Langevin dynamics depend on the log-concavity of the distribution and its explicit form $\mu_0 = \frac{e^{-U(x)}}{\int e^{-U(x)}\,dx}$. Instead, EFS requires only i.i.d. samples. For more details on related works, please see Section B in the Appendix.

A Note on Proofs and Experimental Results

Due to space constraints, we defer the proofs, experimental results, related work, and discussion of limitations to the Appendix.

References

[1] Fabian Altekrüger, Johannes Hertrich, and Gabriele Steidl. Neural wasserstein gradient flows for discrepancies with riesz kernels. In *International Conference on Machine Learning*, pages 664–690. PMLR, 2023.

[2] David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural network. In *Annual Conference on Neural Information Processing Systems*, 2021.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[4] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International conference on machine learning*, pages 146–155. PMLR, 2017.

[5] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

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

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 2017.

[8] Sanjeev Arora, Ravi Kannan, and Ankur Moitra. Learning mixtures of gaussians is np-hard. In *Proceedings 34th Annual ACM Symposium on Theory of Computing (STOC)*. ACM, 2001.

[9] Daniel Balagué, José A Carrillo, Thomas Laurent, and Gaël Raoul. Nonlocal interactions by repulsive–attractive potentials: radial ins/stability. *Physica D: Nonlinear Phenomena*, 260: 5–25, 2013.

[10] Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1-58):26, 1997.

[11] Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. Efficient gradient flows in sliced-wasserstein space. *Transactions on Machine Learning Research Journal*, 2022.

[12] George EP Box and Mervin E Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 1958.

[13] Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.

[14] José A Carrillo and Ruiwen Shu. From radial symmetry to fractal behavior of aggregation equilibria for repulsive–attractive potentials. *Calculus of Variations and Partial Differential Equations*, 62(1):28, 2023.

[15] 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.

[16] Hadi Daneshmand, Jason D Lee, and Chi Jin. Efficient displacement convex optimization with particle gradient descent. In *International Conference on Machine Learning*, pages 6836–6854. PMLR, 2023.

[17] Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. In *International Conference on Machine Learning*, pages 6185–6215. PMLR, 2022.

[18] Rupert L Frank and Ryan W Matzke. Minimizers for an aggregation model with attractive–repulsive interaction. *Archive for Rational Mechanics and Analysis*, 249(2):15, 2025.

[19] Rupert L Frank, Rupert L Frank, Rupert L Frank, and Ryan W Matzke. Minimizers for an aggregation model with attractive–repulsive interaction. *Archive for Rational Mechanics and Analysis*, 249(2):15, 2025.

[20] O Frostman. Potentiel d'equilibre et theorie des ensembles. *thesis*, 1935.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[22] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[25] Alexander Kolesov, Manukhov Stepan, Vladimir V Palyulin, and Alexander Korotin. Field matching: an electrostatic paradigm to generate and transfer data. *arXiv preprint arXiv:2502.02367*, 2025.

[26] Andrei Kolmogoroff. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104:415–458, 1931.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Fanghua Lin and Ping Zhang. On the hydrodynamic limit of ginzburg-landau vortices. *Discrete and Continuous Dynamical Systems*, 6(1):121–142, 2000.

[29] Jonathan Masci, Ueli Meier, Dan Ciresan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer, 2011.

[30] Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

[31] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2976–2985. PMLR, 2019.

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

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019. URL https://papers.nips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[34] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[35] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 2017.

[36] Sylvia Serfaty. Mean field limit for coulomb-type flows. *Duke Mathematical Journal*, 169(15): 2887–2935, 2020. Appendix by Mitia Duerinckx.

[37] Ruiwen Shu. A family of explicit minimizers for interaction energies. *arXiv preprint arXiv:2501.14666*, 2025.

[38] Ruiwen Shu and Jiuya Wang. Generalized erdős-turán inequalities and stability of energy minimizers. *arXiv e-prints*, pages arXiv–2110, 2021.

[39] Ruiwen Shu and Jiuya Wang. The sharp erdős-turán inequality. *arXiv preprint arXiv:2109.11006*, 2021.

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[41] Min Jae Song. Cryptographic hardness of score estimation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[43] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[44] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[45] Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4958–4991. PMLR, 2024.

[46] Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. *Advances in Neural Information Processing Systems*, 35:16782–16795, 2022.

[47] Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. Pfgm++: Unlocking the potential of physics-inspired generative models. In *International Conference on Machine Learning*, pages 38566–38591. PMLR, 2023.

[48] Peiyuan Zhang, Antonio Orvieto, and Hadi Daneshmand. Rethinking the variational interpretation of accelerated optimization methods. *Advances in Neural Information Processing Systems*, 34:14396–14406, 2021.

[49] Peiyuan Zhang, Antonio Orvieto, Hadi Daneshmand, Thomas Hofmann, and Roy S Smith. Revisiting the role of euler numerical integration on acceleration and stability in convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3979–3987. PMLR, 2021.

Appendix

Appendix A. Algorithm

For completeness, we present the EFS algorithm in Algorithm 1 and its forward and backward optimization subroutines in Algorithm 2 and Algorithm 3, respectively. Remarkably, even when the distribution is uniform over a high-dimensional ball, the samples are effectively drawn uniformly from the unit sphere, which is denoted by $\mathbb{S}^{d-1}$, as almost all the volume concentrates near the boundary [10].

Algorithm 1: Estimation Free Sampling (*EFS*)

Input: I.I.D samples $\{x_1^{(0)}, \ldots, x_n^{(0)}\}$

Parameters: Step size γ, number of forward iterations k, learning rate of backward (proximal step) β, number of iterations for each backward proximal step T, potential parameters s and ϵ

`/* Forward the training data to the sphere by gradient descent */`

Set $\{x_0^{(j)}, \ldots, x_n^{(j)}\}_{j=1}^k \leftarrow \texttt{Forward}(\{x_1^{(0)}, \ldots, x_n^{(0)}\}, \gamma, k, s, \epsilon)$

`/* Draw a new random point uniformly from the sphere */`

Set $c = \frac{1}{n}\sum_{i=1}^n x_i^{(k)}, r = \frac{1}{n}\sum_{i=1}^n \|c - x_i^{(k)}\|$

Draw $\nu \sim \mathbb{S}^{d-1}(c, r)$ `// Sphere with center` c `and radius` r

`/* Backward the new data point to the original space */`

Set $y^0 \leftarrow \texttt{Backward}(\{x_1^{(j)}, \ldots, x_n^{(j)}\}_{j=1}^k, y^{(k)}, \gamma, k, \beta, T, s, \epsilon)$

Output: Generated sample $y^{(0)}$

Algorithm 2: Forward Optimization (*Forward*)

Input: Training data $\{x_1^{(0)}, \ldots, x_n^{(0)}\}$

Parameters: step size γ, number of forward iterations k, potential parameters s and ϵ

for $j \leftarrow 0$ **to** $k-1$ **do**
  for $i \leftarrow 1$ **to** n **do**
    $\Delta_i = \frac{1}{n-1}\sum_{a\in[n], a\neq i} \nabla W_\epsilon^{(s)}(x_i^{(j)} - x_a^{(j)})$
    $x_i^{(j+1)} = x_i^{(j)} - \gamma\Delta_i$
  end
end

Output: $\{x_0^{(j)}, \ldots, x_n^{(j)}\}_{j=1}^k$

Algorithm 3: Backward Optimization (*Backward*)

Input: Data $\{x_1^{(j)}, \ldots, x_n^{(j)}\}_{j=1}^k$, new sample $y^{(k)}$
Parameters: Step size γ, number of forward iterations k, learning rate of backward (proximal step) β, number of iterations for each backward proximal step T, potential parameters s and ϵ
Set: $j = k$
while $j \geq 0$ **do**
  Set: $v_0 = y^{(j)}$
  for $t \leftarrow 0$ **to** T **do**
    $\nabla = \frac{\gamma}{n} \sum_{i \in [n]} \nabla W_\epsilon^{(s)}(v_t - x_i^{(j)})$
    $\Delta = v_t - y^{(j)} - \nabla$
    $v_{t+1} = v_t - \beta\Delta$
  end
  $j = j - 1$ and $y^{(j)} = v_T$
end
Output: $y^{(0)}$

Appendix B. Related works

Data generation with estimation. Generative models traditionally rely on transporting simple to complex distributions. A classical example is the Box–Muller transform: given independent random variables $\theta \sim \text{Uniform}[0, 2\pi]$ and $r \sim \text{Exponential}$, the random vector $v = r(\cos\theta, \sin\theta)$ follows a standard two-dimensional Normal distribution [12]. This illustrates a fundamental idea: sampling from a complex distribution can be achieved via a nonlinear transformation of samples drawn from a simpler distribution.

Modern generative models adopt this principle using function approximation: Given latent variable x sampled from a Gaussian distribution, the goal is to find a parametric function f_θ such that the distribution of $f_\theta(x)$ approximates the data distribution. In generative adversarial networks (GANs), f_θ is implemented by a neural network and trained via adversarial objectives [21]. In normalizing flows, f_θ is a sequence of invertible and differentiable transformations optimized via maximum likelihood [34]. Score-based diffusion models, gradually transform data using estimated score functions of diffusion density [42]. In contrast to these methods, our proposed approach computes a non-linear transformation using a set of interacting points, thereby avoiding the estimation of transport map f_θ.

Varational perspective towards generative models. Despite its distinct mechanism, EFS shares conceptual parallels with diffusion models. Both can be interpreted as discretizations of different *"Wasserstein gradient flows"* [3]. This connection arises from a common variational principle: many simple distributions from which i.i.d. samples are easily drawn, such as the Gaussian distribution or the uniform distribution on a sphere, can be viewed as equilibrium states of physical systems with minimum energy. For example, Gaussian distribution is the minimizer of the negative entropy functional over the space of probability measures [24], while the uniform distribution on the unit sphere minimizes a Riesz-type interaction energy [18]. More generally, such distributions can be

described as

$$p^* = \arg\min_{p \in \Omega} F(p), \quad \Omega := \{\text{probability measures over } \mathbb{R}^d\}, \tag{7}$$

where F is an appropriate energy functional.

This variational formulation provides a principled mechanism to transport any distribution to the minimum energy state by gradient descent on F in the space of probability densities. Diffusion models, optimize $F(p) = \int \log(p(x))p(x)dx$ using Wasserstein gradient flow [24]. More interesting observation is that the reverse of gradient flow which can transport Gaussian distribution to any distribution. For entropy, inverse Wasserstein gradient flow is implemented by the Kolmogorov backward equation [26], enabling transformation from a Gaussian to arbitrary target distributions. This key result forms the basis of modern generative diffusion models [23, 40, 42], where the reverse-time stochastic process includes a drift term proportional to the score function of intermediate densities [5]. However, estimating this score function for arbitrary distributions remains statistically and computationally challenging. In this vein, Wibisono et al. [45] recently proved that estimating the score function of sub-Gaussian distributions with Lipschitz-continuous scores suffers from a curse of dimensionality in the required sample complexity. Song [41] further demonstrated that, under lattice-based cryptographic hardness assumptions, score estimation remains computationally intractable even when the sample complexity is polynomial in the relevant parameters.To circumvent these challenges, our method replaces the entropy-based potential with a Riesz-type interaction energy and optimizes it directly over discrete empirical measures.

Inspired by Formulation (7), one may attempt to optimize functionals of the probability measure p that quantify its distance to a target distribution p^*. For example, the celebrated work of Arjovsky et al. [7] introduce the Wasserstein GAN, which optimizes the Wasserstein-1 distance between the model distribution p and the target distribution p^* by solving its Kantorovich–Rubinstein dual via an adversarial training objective over both the generator (which parameterizes p) and the critic (which approximates the dual). More closely related to gradient flows on the space of probability measures, Arbel et al. [6] consider the Wasserstein gradient flow of the Maximum Mean Discrepancy (MMD) functional and provide particle-based implementations. Additionally, Mroueh et al. [31] propose Sobolev Descent, deriving a particle algorithm from the Sobolev integral probability metric that transports samples along smooth descent paths. More recently, there is been a surge in using the seminal proximal point method of Jordan, Kinderlehrer, and Otto (JKO) [24], to discretisize the backward Wasserstein gradient flows over the chosen energy function F [1, 2, 11, 13, 17, 30]. Central to almost all of these works, is the estimation of the transport map in JKO proximal step with input convex neural networks [4]. In contrast to these works, our approach does not require parameterizing the transport map, the functionals that solve the variational formulations, or the underlying ODE/PDE induced by the continuity equation. Instead, it generates data via direct (inverse) gradient descent on a finite set of points without function estimation.

Particle gradient descent. [32] introduce "particle gradient descent" for optimizing a given energy function F over sparse measures as

$$\min_{x_1,\ldots,x_n} F\left(\frac{1}{n}\sum_{i=1}^{n}\delta(x_i)\right), \quad \delta(x) : \text{the Dirac measure at } x. \tag{8}$$

Particle gradient descent is the gradient descent optimizing the location of $x_1, \ldots, x_n$. Single layer neural networks can be viewed as particle gradient descent on a specific energy functions. Motivated

by this connection, [15] study the connection between gradient descent on particles and gradient flow in the space of probability measure equipped with Wasserstein-2 metric in asymptotic regime $n \to \infty$. [16] establishes a non-asymptotic convergence analysis for particle gradient descent on displacement convex functions. While primary focus of these studies is on analyzing single-layer neural networks, we leverage (inverse) particle gradient descent to develop an estimation-free generative method.

Physics-inspired methods. Xu et al. [46] developed a generative model that maps samples from a uniform distribution over an infinite-radius hemisphere to an arbitrary target distribution. Their model relies on estimating the *Poisson vector field* parameterized by neural networks. This idea, primarily inspired by physical systems—especially electrostatic theory—led to further developments in subsequent works [25, 47].

To cope with the challenges of sampling from an infinite-radius hemisphere, Xu et al. [46] simulate the corresponding ODE by perturbing the training data and then estimate the negative normalized field from these perturbed samples. This contrasts with our approach, in which the primitive distribution is uniform on a finite-dimensional compact manifold (the sphere), and is therefore easy to sample from. Although we also transport samples from a uniform distribution to an arbitrary data distribution, our transport mechanism does not rely on function estimation; rather, it is purely based on (inverse) gradient descent over a finite set of points.

Appendix C. Discussions

We demonstrate, both theoretically and experimentally, that it is possible to transport a uniform distribution to a target distribution without estimating a score function—using only i.i.d. samples from the target. Our method, Estimation-Free Sampling (EFS), avoids both noise injection and function estimation. Instead, it introduces interactions between samples by optimizing an energy functional, thereby shaping the empirical distribution of samples. EFS lies at the intersection of three foundational fields—mathematical optimization, potential theory, and generative AI —offering rich opportunities for interdisciplinary research.

Optimization for Sampling. While most generative models are built upon stochastic differential equations, EFS is purely a *deterministic optimization method*, opening new avenues for the optimization community to use powerful optimization techniques—such as accelerated methods, higher-order optimization methods, and efficient stochastic techniques—to generate data.

Notably, EFS involves both convex and non-convex optimization. Linking the non-convex gradient descent dynamics to a well-studied Wasserstein gradient flow, we establish asymptotic convergence guarantees. However, understanding the behavior of gradient descent in non-asymptotic regimes remains an open challenge. We believe the introduced asymptotic results provides a solid foundation for future non-asymptotic analyses.

Potential Theory. EFS induces attractive-repulsive interactions between training samples using a well-known potential function studied extensively in fractional potential theory [19, 36, 38]. By linking generative modeling to this rich theoretical framework, we gain access to powerful analytical tools for understanding and designing generative models. In particular, our analysis leverages results on Wasserstein gradient flows of attractive-repulsive energies [36], as well as variational analysis of these energy functionals [14, 18]. This bridge between theory and practice opens new directions for developing principled and reliable generative models with theoretical guarantees.

Potential theory can design new interaction potentials tailored to data generation and practical applications. As noted in our experiments, one major challenge was the choice of the power s in the potential function, which led to numerical instability in high dimensions. Investigating alternative potentials that avoid such issues—particularly those that do not involve dimension-dependent blow-up—may help resolve computational challenges of EFS.

Generative AI. EFS opens up several promising directions for future research, particularly in scaling to large-scale machine learning applications. A key practical challenge lies in the quadratic time complexity of the forward optimization step with respect to the number of training samples. We conjecture that stochastic optimization techniques could mitigate this issue and significantly reduce computational cost.

Appendix D. Preliminaries

Notations. $\mathbb{S}^{d-1}(c,r)$ denotes sphere with center $c \in \mathbb{R}^d$ and radius r. Ω denotes the set of probability measure over $\mathbb{R}^d$ with Wasserstein-2 metric, which is denoted by $\mathcal{W}_2(\mu,\nu)$ where $\mu,\nu \in \Omega$. Given vector function $v : \mathbb{R}^d \to \mathbb{R}^d$, its *divergence* is defined as $\mathrm{div}(v) = \sum_{i=1}^d \frac{dv}{dx_i}$. Suppose $E : \Omega \to \mathbb{R}$. Then, the *first variation* of E with respect to μ is denoted by $\frac{dE}{d\mu}$ [35]. Function $f : \mathbb{R}^d \to \mathbb{R}^d$ transport $\mu \in \Omega$ to $\nu \in \Omega$ if for a random vector $x \in \mathbb{R}^d$ drawn from μ, $f(x)$ has density ν. Recall $W_\epsilon^{(s)}$ is the potential function defined as

$$W_\epsilon^{(s)}(x-y) = \frac{\|x-y\|^2}{2} + \frac{1}{s(\|x-y\|^2+\epsilon)^{s/2}}. \tag{9}$$

For simplicity, we use the shorthand notation $W^{(s)} = W_0^{(s)}$ and $W^{(s)}$ by W. The sign $*$ denotes the standard convolution as

$$f * g(y) = \int f(y-x)g(x)dx \tag{10}$$

$\|f\|_{L^2}$ is the L_2 functional norm defined as

$$\|f\|_{L_2}^2 = \int \|f(x)\|^2 dx \tag{11}$$

Weak/Strong convergence. A sequence of measures $\mu^{(n)} \in \Omega$ is said to converge weakly to μ if, for all bounded continuous test functions $f : \mathbb{R}^d \to \mathbb{R}$, the integrals $\int f(x)\mu^{(n)}(x)\,dx$ converge to $\int f(x)\mu(x)\,dx$. A stronger notion is convergence in the $L^2(\nu)$ norm, which requires that

$$\lim_{n\to\infty} \int \|\mu^{(n)}(x) - \mu(x)\|^2 \nu(x) dx = 0.$$

MMD. Define $\mathrm{MMD} : \Omega \times \Omega \to \mathbb{R}_+$ as

$$\mathrm{MMD}^2(\mu,\nu) = \int K(x-y)(\mu(x)-\nu(x))dx(\mu(y)-\nu(y))dy, \quad \text{with } K(\Delta) \coloneqq \frac{1}{s\|\Delta\|^s} \tag{12}$$

where MMD denotes the Maximum Mean Discrepancy between two measures μ and ν.

Suppose that $\widehat{K}$ and $\widehat{\Delta}$ denote the Fourier transforms of the functions $K(x)$ and $\Delta(x) := \mu(x) - \nu(x)$, respectively. Viewing MMD as a convolution and Plancherel's theorem allow us to write MMD as [39]

$$\text{MMD}^2(\mu, \nu) = \int \widehat{K}(w)|\widehat{\Delta}(w)|^2 dw, \tag{13}$$

where $\widehat{K}(w) = C\|w\|^{-d+s}$ with a constant C depending only on d and s [18]. Substituting the Fourier transform into the equation above establishes that $\text{MMD}(\mu, \nu) = 0$ implies $\mu = \nu$. Notably, it is sufficient for the Fourier transform $\widehat{K}$ to be strictly positive almost everywhere to guarantee that K is a *universal kernel* [22].

To avoid potential confusion, note that the MMD associated with the kernel K defined above is not a metric. While machine learning often relies on positive semi-definite kernels to ensure that MMD defines a valid metric, the kernel K is not positive semi-definite. As a result, the corresponding MMD does not obey the triangle inequality.

Gradient dominance of E. We establish an important property of the energy function E. Define functional $F : \Omega \to L_2$ as

$$F(\mu)(y) = \nabla \frac{dE}{d\mu} = \int W(y - x)\mu(x)dx. \tag{14}$$

Remarkably, μ is a steady state of the Wasserstein gradient flow on E if $F(\mu) = 0$. The next theorem represent MMD using F.

Theorem 4 *Suppose that $\mu \in \Omega$ and $\nu \in \Omega$ have the same first-moment, then*

$$\|F(\mu) - F(\nu)\|_{L_2}^2 = c \times \textit{MMD}^2(\mu, \nu)$$

holds for $s = d - 2$ and an absolute constant $c > 0$.

An application of the above theorem recovers the result of Carrillo and Shu [14]: all steady states μ satisfying $F(\mu) = 0$ are global minimizers of E (up to translation). More precisely, the theorem implies that for all μ, ν such that $F(\mu) = F(\nu) = 0$, the following holds:

$$0 = \|F(\mu) - F(\nu)\|_{L_2}^2 = \text{MMD}(\mu, \nu). \tag{15}$$

Beyond recovering existing results, the theorem will also be used to analyze the backward optimization for EFS in Section E.

Proof Let $\widehat{f}(w)$ denote the Fourier transform of function $f(x)$. Define $\widehat{\Delta} := \widehat{\mu} - \widehat{\nu}$ which is equivalent to the Fourier transform of $\Delta := \mu - \nu$. As discussed in Section D, MMD can be written as

$$\text{MMD}^2(\nu, \mu) = \int \widehat{K}(w)|\widehat{\Delta}(w)|^2 dw \tag{16}$$

According to the definition,

$$\nabla W(z) = \nabla K(z) + \text{iden}(z), \quad \text{iden}(z) := z \tag{17}$$

Replacing the above formula into the definition of F obtains

$$\|F(\mu) - F(\nu)\|_{L_2}^2 = \|(\nabla K + \text{iden}) * \mu - (\nabla K + \text{iden}) * \nu\|_{L^2}^2, \tag{18}$$

where $*$ denotes convolution defined in (10) and L_2 is the norm 2 for functions defined in (11). It is easy to check that

$$(\text{iden} * \mu)(y) = y \underbrace{\int \mu(x)dx}_{=1} - \int x\mu(x)dx \tag{19}$$

$$= y \underbrace{\int \nu(x)dx}_{=1} - \underbrace{\int x\mu(x)dx}_{\mathbb{E}_\mu[x]} \tag{20}$$

$$= y \int \nu(x)dx - \underbrace{\int x\nu(x)dx}_{\mathbb{E}_\nu[x]} = (\text{iden} * \nu)(y). \tag{21}$$

The last equation holds because the first moments of μ and ν are equal. This observation allows us to significantly simplify the expression for the quantity of interest as follows:

$$\|F(\mu) - F(\nu)\|_{L_2}^2 = \|\nabla K * \mu - \nabla K * \nu\|_{L_2}^2 \tag{22}$$

Recall two fundamental properties of Fourier transform as: $\begin{cases} \widehat{f * g} = \widehat{f}\widehat{g} \\ \widehat{\partial f} = iw\widehat{f} \end{cases}$. These properties together with Parseval's theorem yield

$$\|F(\mu) - F(\nu)\|_{L_2}^2 = \int \|w\|^2 \left(\widehat{K}(w)\right)^2 |\widehat{\Delta}(w)|^2 dw \tag{23}$$

For $s = d - 2$, it is easy to check that

$$\|w\|^2(\widehat{K}(w))^2 = \widehat{K}(w) \tag{24}$$

holds given $\widehat{K}(w) = c\|w\|^{-2}$ [18]. Combining all the results completes the proof:

$$\|F(\mu) - F(\nu)\|_{L_2}^2 \overset{(23)}{=} c \int \|w\|^2 \left(\widehat{W}(w)\right)^2 |\widehat{\Delta}(w)|^2 dw \tag{25}$$

$$\overset{(24)}{=} c \int \widehat{W}(w)|\widehat{\Delta}(w)|^2 dw \tag{26}$$

$$\overset{(16)}{=} c \times \text{MMD}^2(\mu, \nu) \tag{27}$$

■

A convergence result for ODEs. Consider the following two differential equations

$$\begin{cases} \frac{dy}{dt} = F_t(y_t) \\ \frac{dy^{(n)}}{dt} = F_t^{(n)}(y^{(n)}) \end{cases} \quad , \quad y_0 = y_0^{(n)} \tag{28}$$

where $F_t, F_t^{(n)} : \mathbb{R}^d \to \mathbb{R}^d$ and $y_t, y_t^{(n)} \in \mathbb{R}^d$. If $F_t^{(n)}$ converges to F_t as $n \to \infty$, then $y_t^{(n)}$ converges to y_t in interval $t \in [0, T)$.

Lemma 5 *Define*

$$\epsilon_n \coloneqq \int \|F_t(y) - F_t^{(n)}(y)\|^2 dy.$$

Suppose $F_t(y)$ is almost surely L-Lipschitz in y. If $\lim_{n\to\infty} \epsilon_n = 0$, then $\int_0^T \|y_t - y_t^{(n)}\|^2 dt$ converges to 0 as $n \to \infty$.

Proof We start by writing down ODEs in the integral form and consider their difference,

$$\begin{aligned} \|y^{(n)}(t) - y(t)\| &= \left\| \int_0^t \left[F_s^{(n)}(y^{(n)}) - F_s(y) \right] ds \right\| \\ &= \left\| \int_0^t \left[F_s^{(n)}(y^{(n)}) - F_s(y^{(n)}) \right] ds + \int_0^t \left[F_s(y^{(n)}) - F_s(y) \right] ds \right\| \\ &\leq t \sup_y \left\| F_s^{(n)}(y) - F_s(y) \right\| + L \int_0^t \left\| y^{(n)}(s) - y(s) \right\| ds \\ &\leq t\epsilon_n + L \int_0^t \left\| y^{(n)}(s) - y(s) \right\| ds \end{aligned} \tag{29}$$

The last inequality was followed by convergence of $F_s^{(n)}$ to F_s, and boundedness of L^∞ norm by L^2, the first term is converging to zero, and bounded by $\epsilon_n \xrightarrow{n\to\infty} 0$.

Define $\Delta_n(t) \coloneqq \sup_{s\leq t} \left\| y^{(n)}(s) - y(s) \right\|$. By taking supremum from (29),

$$\Delta_n(t) \leq t\epsilon_n + L \int_0^t \Delta_n(s) ds.$$

In turn, applying Grönwall's inequality yields,

$$\Delta_n(t) \leq t\epsilon_n e^{Lt} \xrightarrow{n\to\infty} 0.$$

Thus,

$$\left\| y^{(n)}(s) - y(s) \right\| \xrightarrow{n\to\infty} 0,$$

uniformly for all $s \leq t$ and proof is completed. ■

Continuity equation and transporting distributions. Let $v_t : \mathbb{R}^d \to \mathbb{R}^d$ be a general vector field. $\mu_t \in \Omega$ is a solution of continuity equation associated with v_t if it obeys

$$\frac{d\mu_t}{dt} = -\text{div}(\mu_t v_t) \tag{30}$$

Given the vector field, we define the following ODE

$$\frac{dy_t}{dt} = v_t(y_t) \tag{31}$$

The above ODE transports μ_0 to μ_t as stated in the following lemma.

Lemma 6 (Lemma 8.1.6. of [3]) *Suppose that the vector field v_t obeys the following 3 conditions*

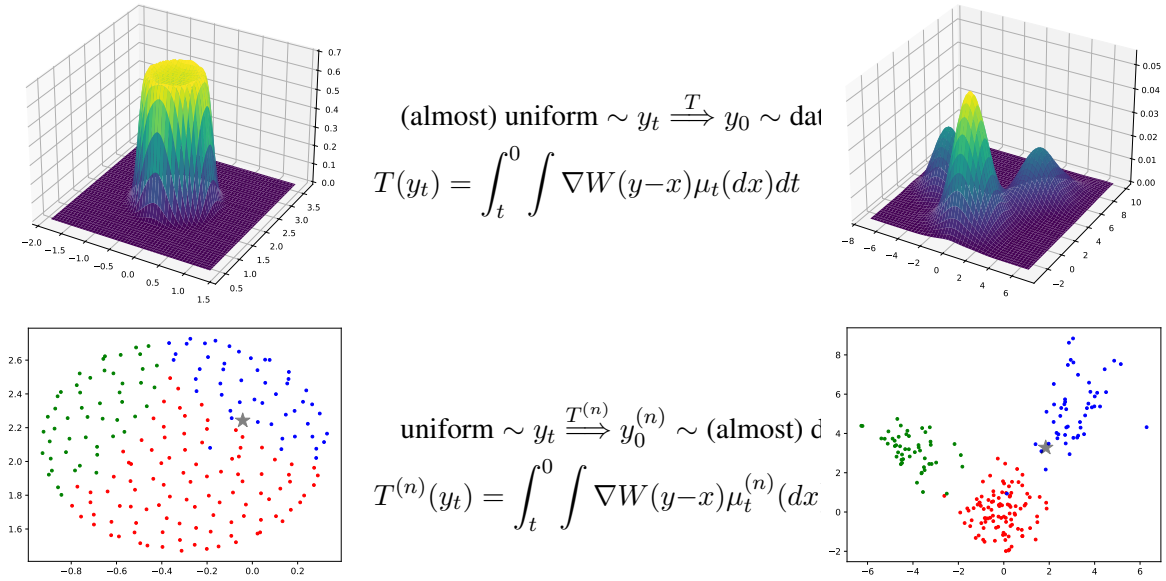
(1) $\int |v_t(x)| d\mu_t(x) dx < \infty$.

(2) For every compact subset $B \subset \mathbb{R}^d$, $\sup_{x \in B} |v_t(x)| < \infty$.

(3) $v_t(x)$ *is Lipschitz*

If y_0 is a random variable with distribution μ_0, then y_t is a random variable with distribution μ_t for a finite t.

Appendix E. Proof of Theorem 3



Proof sketch for Theorem 3. Given the solution to the continuity equation (5), we define the map T above, which provably transports μ_t to data distribution, as stated in Lemma 7. Recall that Theorem 2 states that μ_t converges to the uniform distribution, from which new samples can be drawn. T is not implementable as it requires μ_t, the solution of the continuity equation (5). We show that the empirical distribution $\mu_t^{(n)}$, defined over the point set $\{x_1(t), \ldots, x_n(t)\}$, yields a transport map $T^{(n)}$ (as defined above) that converges to T as $n \to \infty$. The $\star$ symbol in the second row indicates y_t (left) and $y_0^{(n)}$ (right). Colored points represent $\{x_1(t), \ldots, x_n(t)\}$ on the left and $\{x_1(0), \ldots, x_n(0)\}$ on the right.

Recap. Recall that $\mu_t^{(n)}$ defined as $\mu_t^{(n)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ where δ is the Dirac measure and $x_i(t)$ obeys

$$\frac{dx_i}{dt} = -\frac{1}{n}\sum_{j\neq i} \nabla W^{(s)}(x_i - x_j). \tag{32}$$

Using $x_i(t)$, we define $y_t^{(n)}$ which obeys

$$\frac{dy^{(n)}}{d\bar{t}} = F_t^{(n)}(y_t^{(n)}), \quad F_t^{(n)}(y) := \int \nabla W\,(y-x)\,\mu_t^{(n)}(x)dx, \tag{33}$$

where $d\bar{t}$ indicates that the process is reversed time [5].

Reversed-time transport. Recall μ_t as the solution of continuity equation for Wasserstein-2 gradient flow:

$$\frac{d\mu}{dt} = \operatorname{div}\left(\mu(x)\int \nabla W^{(s)}(x-y)\mu_t(y)\,dy\right), \tag{34}$$

Notably, $F_t(y) \coloneqq \int \nabla W(y-x)\mu_t(x)dx$ represents the vector field associated with the above dynamics. Given this vector field, define the following ODE

$$\frac{dy}{dt} = -F_t(y_t), \quad \text{where } F_t(y) \coloneqq \int \nabla W(y-x)\mu_t(x)dx \tag{35}$$

The above ODE transports μ_t backward to μ_0 as stated in the following corollary as a consequence of Lemma 6.

Corollary 7 *Let y_0 be a random vector drawn from μ_0. Then, the distribution of y_t is μ_t.*

The above result together with a simple change of variables obtains the inverse transport from μ_t to μ_0.

Corollary 8 *Consider the following reversed-time dynamics:*

$$\frac{dy_t}{d\bar{t}} = F_t(y_t) \tag{36}$$

The above dynamics transports μ_t to μ_0 as long as μ_t is not a steady state of (34).

Proof According to definition

$$\lim_{\epsilon\to 0}\frac{y_{t+\epsilon}-y_t}{\epsilon} = -F_t(y_t) \implies \frac{dy}{d\bar{t}} = \lim_{\epsilon\to 0}\frac{y_{t-\epsilon}-y_t}{\epsilon} = \lim_{\epsilon\to 0} F_{t-\epsilon}(y_{t-\epsilon}) \tag{37}$$

where $F_{t-\epsilon}(y) = \int \nabla W(y-x)\mu_{t-\epsilon}(x)dx$. F_t is Lipschitz according to regularity assumptions. Thus, invoking Lemma 5 concludes the statement. ■

While the flow y_t transports μ_t to the data distribution, its construction relies on access to μ_t, which is defined as the solution to a PDE. Since μ_t is not feasible to compute, we cannot implement the reverse-time ODE (35) to recover y_0. However, Theorem 3 restated bellow ensures that EFS can reconstruct the backward dynamic as $n \to \infty$.

Theorem 3 *Suppose that $\mu_0^{(n)}$ converges to a sufficiently regular measure μ_0 in the Wasserstein-2 distance, such that regularity assumptions hold. Furthermore, assume that μ_t is not exactly the steady-state solution of the PDE in (5). Let $y_t^{(n)}$ be a random variable with law μ_t, where μ_t is the solution to the continuity equation (5). There is a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that if $n \geq f(t)$, then the distribution of $y_0^{(n)}$—obtained from the reversed-time ODE (6)—converges to μ_0 as $t \to \infty$, for continuous measures μ_0 and μ_t, $s = d-2$.*

Proof [Proof of Theorem 3] We prove that $y_0^{(n)}$ converges to y_0 starting from the same state. To establish this convergence, we use Theorem 1 of [36], which shows that $\mu_t^{(n)}$ converges to μ_t in MMD as $n \to \infty$, namely

$$\lim_{t\to\infty} \text{MMD}^2(\mu_t, \mu_t^{(n_t)}) = 0, \quad \text{for } \beta > 0. \tag{38}$$

Theorem 4 ensures that the above converges leads to the converges of $F_t^{(n)}$ to F_t in functional norm 2, namely

$$\lim_{t\to\infty}\int \|F_t(y) - F_t^{(n_t)}(y)\|^2 dy \overset{\text{Thm 4}}{=} c \lim_{t\to\infty} \text{MMD}^2(\mu_t, \mu_t^{(n_t)}) \overset{[36]}{=} 0 \tag{39}$$

According to regularity assumptions, $F_t(y)$ is Lipschitz. This Lipschitz continuity, together with the result above, ensures that the conditions in Lemma 5 are satisfied. Thus, $y_0^{(n)}$ converges to y_0, as guaranteed by the lemma. Invoking Corollary 8 ensures that the distribution of y_0 coincides with the underlying data distribution μ_0. ■

Appendix F. Proof of Theorem 9

Proposition 9 *Consider the following proximal optimization problem*

$$(y_1^*, y_2^*, \cdots, y_n^*) = \arg \min_{y_1, y_2, \cdots, y_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{d} \|y_i - x_i^{(k)}\|^2 - \gamma E_n^{(\epsilon)}(y_1, y_2, \cdots, y_n) \tag{40}$$

The above optimization is convex with solution $y_i^* = x_i^{(k-1)}$ for all $i \in [n]$, as long as the learning rate γ is sufficiently small.

Proof Let the objective of the proximal step (40) defined as,

$$H(y_1, \cdots, y_n) \coloneqq \frac{1}{2} \sum_{i=1}^{n} \|y_i - x_i^{(k)}\|^2 - \gamma E_n^{(\epsilon)}(y_1, \cdots, y_n). \tag{41}$$

And for ease of notation, let $Y = (y_1, \cdots, y_n) \in \left(\mathbb{R}^d\right)^n$. It is easy to observe that,

$$\nabla^2 H(Y) = I_{n \times d} - \gamma \nabla^2 E_n^{(\epsilon)}(Y).$$

Thus, the optimization problem (9) is convex as long as we can prove an upper bound on the spectral norm of $\nabla^2 E_n^{(\epsilon)}(Y)$ and choose a small learning rate for the inverse gradient map γ. From this point onward, we drop the subscript/superscript ϵ to streamline the notation. Thus W is defined as

$$W(z) = -\frac{m}{(\|z\|^2 + \epsilon)^{m/2}} + \frac{1}{2}\|z\|^2$$

with direct calculations we have,

$$\nabla W(z) = \left(1 + m^2 (r^2 + \epsilon)^{-\frac{m}{2} - 1}\right) z,$$

where $r \coloneqq \|z\|$. Again with differentiation,

$$\nabla^2 W(z) = \left(1 + m^2 (r^2 + \epsilon)^{-\frac{m}{2} - 1}\right) I_d - m^2 (m+2)(r^2 + \epsilon)^{-\frac{m}{2} - 2} z z^\top.$$

Hence, for every unit vector $\nu \in \mathbb{S}^d$,

$$\begin{aligned} \nu^\top \nabla^2 W(z) \nu &\leq 1 + m^2 (r^2 + \epsilon)^{-\frac{m}{2} - 1} \\ &\leq 1 + m^2 \epsilon^{-\frac{m}{2} - 1}. \end{aligned} \tag{42}$$

Recall that,

$$E_n(Y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} W(y_i - y_j).$$

Taking derivatives, for each i we have,

$$\nabla_{y_i} E_n(Y) = \frac{1}{n(n-1)} \sum_{j \neq i} \nabla W(y_i - y_j).$$

Next, for $j \neq i$,

$$\nabla^2_{y_i, y_j} E_n(Y) = \frac{-1}{n(n-1)} \nabla^2 W(y_i - y_j)$$

and,

$$\nabla^2_{y_i} E_n(Y) = \frac{1}{n(n-1)} \sum_{j \neq i} \nabla^2 W(y_i - y_j).$$

Taking an arbitrary block vector $\nu \in \left(\mathbb{R}^d\right)^n$,

$$\begin{aligned}
\langle \nu, \nabla^2 E_n(Y) \nu \rangle &= \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} \nu_i^\top \left(\sum_{j \neq i} \nabla^2 W(y_i - y_j) \right) \nu_i + \sum_{i \neq j} \nu_i^\top \left(-\nabla^2 W(y_i - y_j) \right) \nu_i \right) \\
&= \frac{1}{n(n-1)} \left(\frac{1}{2} \sum_{i \neq j} (\nu_i - \nu_j)^\top \nabla^2 W(y_i - y_j)(\nu_i - \nu_j) \right).
\end{aligned}$$

In turn, by spectral bound on each block in (42),

$$\begin{aligned}
\langle \nu, \nabla^2 E_n(Y) \nu \rangle &\leq \frac{1}{2n(n-1)} \left(\sum_{i \neq j} \left(\left(1 + m^2 (\|y_i - y_j\|^2 + \epsilon)^{-\frac{m}{2}-1} \right) \right) \|\nu_i - \nu_j\|^2 \right) \\
&\leq \frac{1}{2n(n-1)} \left(\sum_{i \neq j} \left(1 + m^2 \epsilon^{-\frac{m}{2}-1} \right) \|\nu_i - \nu_j\|^2 \right) \qquad (43) \\
&= \frac{\left(1 + m^2 \epsilon^{-\frac{m}{2}-1} \right)}{2n(n-1)} \left(\sum_{i \neq j} \|\nu_i - \nu_j\|^2 \right) \\
&\leq \frac{\left(1 + m^2 \epsilon^{-\frac{m}{2}-1} \right)}{2n(n-1)} \left(2n \sum_i \|\nu_i\|^2 \right) \\
&= \frac{1 + m^2 \epsilon^{-\frac{m}{2}-1}}{n-1} \|\nu\|^2,
\end{aligned}$$

where in (43) we used the (42.

Hence, for the operator norm of $\nabla^2 E_n(Y)$ we infer that,

$$\|\nabla^2 E_n(Y)\|_2 \leq \frac{1 + m^2 \epsilon^{-\frac{m}{2}-1}}{n-1}.$$

Thus, by choosing any sufficiently small $\gamma \in (0, \frac{n-1}{1+m^2\epsilon^{-\frac{m}{2}-1}})$,

$$\nabla^2 H(Y) \succ 0,$$

and objective of the proximal step (41) is strongly convex and hence is uniquely minimized.

Considering the optimal solution $(y_1^*, y_2^*, \cdots, y_n^*)$ to (41), by first order optimality condition,

$$\nabla_{y_i} H(y_1^*, y_2^*, \cdots, y_n^*) = y_i^* - x_i^{(k)} - \gamma \nabla_{y_i} E_n(y_1^*, y_2^*, \cdots, y_n^*) = 0.$$

Thus,

$$x_i^{(k)} = y_i^* - \gamma \nabla_{y_i} E_n(y_1^*, y_2^*, \cdots, y_n^*),$$

and hence for all $i \in [n]$,

$$y_i^* = x_i^{(k-1)},$$

and proof is concluded. ■

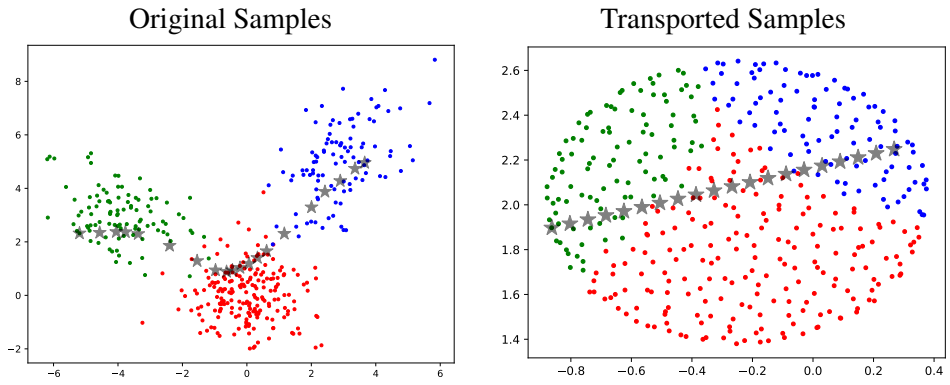


Figure 2: **Interpolation in Latent Space.** *Colored points* are samples drawn from a Gaussian mixture model (left) and their positions after forward optimization (right). The $\star$s are the generated sample y in EFS (right), and their location after applying Algorithm 3 (left). The straight line in the latent space (right) is transformed into a curved trajectory (left) that aligns with the data distribution. Colors correspond to different mixture components.

Appendix G. Experiments

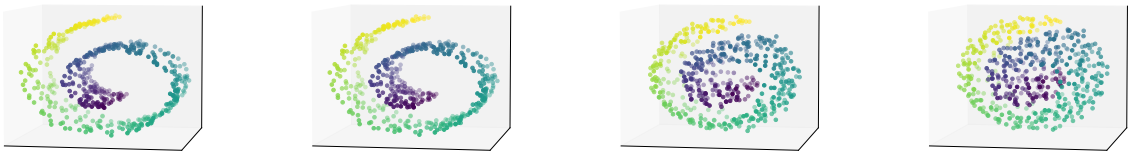
While we have proven EFS can transport an almost uniform distribution to an arbitrary target distribution, our result is asymptotic—it holds in the limit as $n \to \infty$. This asymptotic regime greatly simplifies the theoretical analysis. However, practical applications are inherently constrained to finite n. We bridge this gap with experiments.

As the first estimation-free generative algorithm, EFS needs further research for a comprehensive benchmarking. Our experiments underscore both the practical challenges and the promising potential of this novel generative method. The code is implemented in PyTorch [33] and was executed on a single NVIDIA RTX 6000 GPU with 48GB memory. The total execution time is under 15 minutes.

Mixture of Gaussians: Capturing Data Geometry. EFS effectively recovers the underlying data distribution even after transforming into a set of points uniformly distributed on a circle. Figure 2 displays the original samples drawn from a Gaussian mixture (left) and their transformed positions after the EFS forward optimization step (right), where they are mapped onto the circle. New samples are then generated along a straight line on circle. Backward optimization maps these linear samples onto a curved trajectory that closely follows the geometry of the original data distribution. The backward process adaptively distorts pairwise distances to avoid low-density regions. Furthermore, we observe that EFS creatively generate unseen new samples.

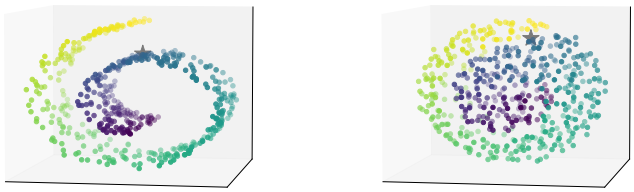
Generating samples from Swiss roll. We repeat the experiments on the Swiss roll dataset, similar to those on the Gaussian mixture in Figures 1, and 2. Table 1 summarizes the parameter choices for this dataset. We evaluate: (i) convergence to a uniform distribution via forward optimization, (ii) generation of new samples via backward optimization, and (iii) preservation of data geometry through interpolation.

(i) *Forward optimization:*In Figure 3, we observe that the distribution of data points becomes uniform over the unit ball. Figure 3 shows that this asymptotic result provides a good approximation even when $n = 500$.



$\{x_i^{(0)}\} \Rightarrow$ $\{x_i^{(36)}\} \Rightarrow$ $\{x_i^{(72)}\} \Rightarrow$ $\{x_i^{(108)}\}$

Figure 3: **Forward convergence for Swiss roll.** Scatter plot of gradient descent iterates $x_i^{(k)}$, defined in (3), initialized at $x_i^{(0)}$ drawn from a Swiss roll with noise level 0.2. As k evolves, the points become uniformly distributed akin to 2d example of Gaussian mixture in Figure 1.



$\{x_i^{(0)}\} \cup \{y^{(0)} = \star\}$ $\{x_i^{(k)}\} \cup \{y^{(k)} = \star\}$

Figure 4: **Sampling:** The black $\star$ is the generated sample $y^{(0)}$; Left: $x_i^{(0)} \overset{i.i.d.}{\sim}$ Swiss roll with noise level 0.2; Right: $x_i^{(k)}$ obtained by forward optimization. Colors: mixture components. The location of the new sample is updated by backward algorithm 3

(ii) *Backward optimization:*We assert that EFS can generate new samples from the Swiss roll dataset using a given set of i.i.d. samples. While Theorem 3 establishes that EFS draws samples from the data distribution in the mean-field regime, Figure 4 supports this result in the non-asymptotic setting.

(iii) *Interpolation:* Figure 5 illustrates how backward optimization can generate new samples from the Swiss roll dataset via interpolation. Similar to the Gaussian mixture case shown in Figure 2, we observe that EFS respects the data density and avoids generating samples in low-density regions.

MNIST: Generating New Samples. EFS can generate new MNIST images [27]. Theorem 3 requires high exponents in the range $d - 2 \leq m < d$. For MNIST with $d = 784$, this exponent leads to numerical overflow due to limited floating-point precision. To avoid this issue, we reduce the data dimension before applying EFS. Specifically, we use a convolutional autoencoder [29] to compress the data into a 15-dimensional latent space (see Appendix for architectural and training details). We then apply encoding $\rightarrow$ EFS $\rightarrow$ decoding to generate new samples. Figure 6 shows

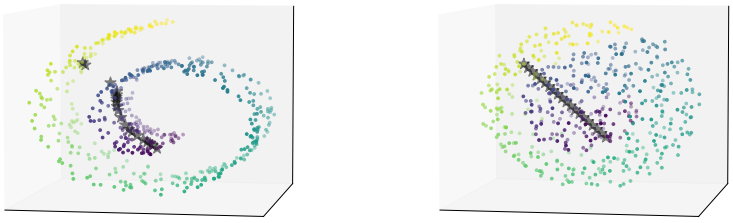


Figure 5: **Interpolation in Latent Space.** *Colored points* are samples drawn from a Swiss roll (left) and their positions after forward optimization (right). The $\star$s are the generated sample $y^{(0)}$ in EFS (right), and their location after applying Algorithm 3 (left). The straight line in the latent space (right) is transformed into a curved trajectory (left) that aligns with the data distribution. We observe that the algorithm avoids generating data from no density regions.

generated samples alongside their nearest neighbors (in Euclidean distance) from decoded original samples $x_1^{(0)}, \ldots, x_n^{(0)}$. The generated digits exhibit stylistic variations distinct from their closest training examples, with minimum Euclidean distance of 0.7508. To emphasize subtle differences, we include an animated image that alternates between the generated sample and its nearest neighbor (open with Adobe Reader). See Table 1 for details.

Generated Images	
Closest Image	
Animation (see with Adobe)	

Figure 6: **Generative MNIST.** *First row:* Images generated by EFS. These are obtained by applying forward and backward optimization in the latent space of an autoencoder, followed by decoding into the image space. *Second row:* The closest decoded training samples $\{x_i^{(0)}\}_{i=1}^n$ for each generated image, determined by Euclidean distance. *Third row:* An animation illustrating subtle differences between the generated images and their nearest neighbors (use Adobe Reader to see). *Note:* The minimum Euclidean distance between each generated image and its nearest neighbor is 0.75.

MNIST: Interpolation Comparison. Recall that in the previous experiment, we used an autoencoder for dimensionality reduction. Notably, autoencoders can generate samples with interpolation in the latent space. To highlight differences, we compare the quality of the autoencoder samples with those produced by EFS.

For EFS, we generate new samples by linearly interpolating between uniformly distributed samples after forward optimization as $y^{(k)} = (1-t)x_i^{(k)} + tx_j^{(k)}$. Then, we apply the backward optimization to $y^{(k)}$. Similarly, autoencoder samples are obtain by decoding $(1-t)x_i^{(0)} + tx_j^{(0)}$

where $x_i^{(0)}$ are encoded samples in the latent space. Figure 7 shows generate samples with these two different strategies. Observe interpolation in the autoencoder's latent space often leads to unrealistic outputs that do not resemble valid digits. As illustrated in the cartoon of Figure 7, such interpolation may cross regions of low data density. In stark contrast, EFS effectively avoids generating samples from these low-density regions, as similarly observed for Gaussian mixtures in Figure 6.

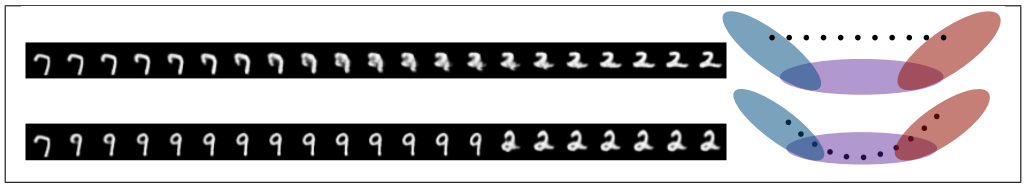


Figure 7: **Interpolation Comparison.** In the interpolation between images 7 and 2, the intermediate autoencoder outputs do not resemble valid digits (1st row), whereas our method produces a transition via $7 \to 9 \to 2$ (2nd row). On the right, we visualize the interpolation paths cross low density regions for autoencoder, while moves along the underlying digit clusters for EFS. Moreover, for EFS, it can be observed that different images are generated for each digit, illustrating that interpolation paths evolve smoothly within clusters.

	γ	k	T	β	ϵ	s	n
Gaussian mixtures	0.1	31	300	0.1	0.001	1	400
MNIST	0.05	120	300	0.01	0.001	$d-2$	15000
Swiss roll	0.05	120	300	0.1	0.001	$d-2$	500

Table 1: Parameters of EFS

G.1. Autoencoder details

As previously mentioned, the experiments in Section G employ an autoencoder. In this section, we provide further details on the architecture and training procedure of the autoencoder. The model is a standard convolutional autoencoder consisting of two encoding layers followed by two decoding layers. The structure of the encoder is summarized in Listing 1. Importantly, the encoder serves as a dimensionality reduction mechanism. Without this reduction, computing $W^{(s)}$, as required in EFS, becomes numerically infeasible.

G.2. Training Protocol

We trained the autoencoder on the MNIST dataset using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 250, and for 120 epochs. The training loss, measured by mean squared error (MSE), decreased to 0.008 after 120 epochs. Training was conducted on an NVIDIA RTX 6000 GPU with 48 GB of memory. The implementation was done in PyTorch [33] (see the notebook `neurips-figures4-5.ipynb` for details). No preprocessing was applied to the data.

```
Autoencoder(
  (encoder): Sequential(
    (0): Conv2d(1, 16, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1)
       )
    (1): ReLU(inplace=True)
    (2): Conv2d(16, 32, kernel_size=(3, 3), stride=(2, 2), padding=(1,
       1))
    (3): ReLU(inplace=True)
    (4): Conv2d(32, 15, kernel_size=(7, 7), stride=(1, 1))
  )
  (decoder): Sequential(
    (0): ConvTranspose2d(15, 32, kernel_size=(7, 7), stride=(1, 1))
    (1): ReLU(inplace=True)
    (2): ConvTranspose2d(32, 16, kernel_size=(3, 3), stride=(2, 2),
       padding=(1, 1), output_padding=(1, 1))
    (3): ReLU(inplace=True)
    (4): ConvTranspose2d(16, 1, kernel_size=(3, 3), stride=(2, 2),
       padding=(1, 1), output_padding=(1, 1))
    (5): Sigmoid()
  )
)
```

Listing 1: Autoencoder