Flow Matching from Viewpoint of Proximal Operators

Kenji Fukumizu[1] **Wei Huang**[2,1] **Han Bao**[1,3,2] **Shuntuo Xu**[4] **Nisha Chandramoorthy**[5]

[1]The Institute of Statistical Mathematics. 10-3 Midoricho, Tachikawa 190-8562 Tokyo, Japan
[2]RIKEN AIP. 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
[3]Tohoku University. 6-3-09 Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579 Japan
[4]East China Normal University. 3663 North Zhongshan Road, Shanghai 200062, China
[5]The University of Chicago. 5747 S. Ellis Avenue Chicago, IL 60637 USA

Abstract

We discuss the dynamics of Optimal Transport Conditional Flow Matching (OT-CFM) under the manifold hypothesis, where the support of the target distribution is a low dimensional manifold. We first derive an exact proximal expression of the vector field, using an extended Brenier potential for the OT; the vector field can be expressed by a denoiser, which is presented as the proximal operator associated with the potential function. Based on this explicit proximal form of the vector field, we analyze the manifold stability for perturbations to initial points, using the Lyapunov exponents of the dynamics. By analyzing the local behavior with the tools in convex analysis, we show that the dynamics contract linearly in the perturbation of the normal directions to the manifold, while perturbations in the tangential directions are preserved; this implies the stability of the manifold structure to the perturbations. To justify the theoretical arguments on the population OT, we also establish the convergence of minibatch OT-CFM with neural network learning to population OT-CFM under regularity and stability assumptions, with rate exponents governed by the intrinsic dimension of the data manifold. Experiments on synthetic data, MNIST, and CIFAR-10 support the predicted geometric behavior.

1 Introduction

Dynamical generative models such as diffusion models [55, 27, 56] and flow matching [37, 38, 1] have recently achieved remarkable success in broad application fields [66], showing strong sampling ability for high-dimensional domains. Prominent examples include text-to-image generation [51, 53], posterior sampling for inverse problems [30], and diffusion-based policies for robots [15]. In scientific domains, diffusion models have also led to strong results in molecular design [16, 64]. In parallel, flow-based models have emerged as competitive alternatives and have been applied to image generation [22] and scientific modeling such as weather forecasting [17] and molecule editing [29].

This paper focuses on flow matching (FM), which learns a vector field $v_t(x)$ whose induced ODE transports a source distribution P_0 (typically standard normal for generation tasks) to a target distribution P_1. The method of Conditional FM (CFM, [37]) generates random conditional paths

$$x_t = \alpha_t x_0 + \beta_t x_1 \qquad (t \in [0,1]) \tag{1}$$

given $(x_0, x_1) \sim \pi$ and schedulers $\alpha_t, \beta_t \in [0,1]$, where π is a joint distribution with marginals P_0 and P_1. A random velocity field along the conditional path is constructed by

$$v_t(x_t; x_0, x_1) = \dot{\alpha}_t x_0 + \dot{\beta}_t x_1, \tag{2}$$

where $\dot{\alpha}_t$ and $\dot{\beta}_t$ denote the time derivatives. Such random velocities are used as teaching data at x_t to train a neural network to estimate the marginal vector field

$$v_t(x_t) \coloneqq \mathbb{E}_{x_0,x_1}\left[v_t(x_t; x_0, x_1) \mid x_t = \alpha_t x_0 + \beta_t x_1\right].$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

With the above CFM construction, many different vector fields can transport P_0 to P_1. The degree of freedom comes from the choice of coupling π between P_0 and P_1, which is implicit in the interpolation path (1). The popular *OT-CFM* constructs the coupling by solving (minibatch) optimal transport (OT) problems in the stochastic gradient descent learning of neural networks [59, 48]. This improves training by reducing the variance of the random teaching data [48] and solves ODE more easily by straighter vector fields, resulting in better empirical performance.

It is often observed that the strong empirical performance of generative models is more evident when the data are distributed on a low-dimensional manifold, a phenomenon called the *manifold hypothesis*; high-resolution images are typical examples. Some recent theoretical studies analyze diffusion models under the manifold hypothesis, including manifold learning [47], minimax estimation rates governed by intrinsic dimension [44, 23, 65], mixtures of low-rank Gaussians [63], a clean subspace/tangent–normal decomposition [13, 14], Lyapunov stability near manifolds [12], and geometric effects/singular behavior of the score around manifolds [36, 39]. On the other hand, similar analysis of FM, especially of OT-CFM under the manifold hypothesis, is scarce, while there are some recent works on FM with independent coupling [33, 46]. Since the learning of OT-CFM is based on a distinct principle from the score, different approaches are required to understand the dynamical behavior around a data manifold.

This paper presents a novel framework for analyzing the vector field and dynamics of OT-CFM, in the general case where the target distribution concentrates on a low-dimensional manifold. We derive an explicit formula for the vector field using the extended Brenier potential and its proximal operator, which are popular tools in convex analysis [10, 50, 45]. The formula expresses the vector field as an affine form of the proximal operator of the potential at the scaled location x_t/t, and provides a denoiser interpretation to OT-CFM in a manner analogous to that in diffusion models.

Based on the proximal expression of the vector field, under the manifold hypothesis, we analyze the stability of the data manifold against the perturbation of initial points. Applying existing results on proximal sensitivity under partly smoothness [25, 19], we derive an explicit tangent–normal decomposition along population OT trajectories. For affine interpolation and the terminal time variable $\tau = -\log(1-t)$, infinitesimal perturbations in the normal direction decay with exponent -1, while tangent perturbations approach finite, nonzero limits and have exponent 0. This implies that small changes in the initial point causes a move in the data manifold.

Our analysis primarily uses a theoretical assumption that the conditional random paths employ the population OT coupling. While this is strong compared to the minibatch OT used in practice, it is not restrictive; we show that the vector field of the minibatch OT with finite data converges to the one given by the population OT for the large sample size limit under some assumptions. We also show that the convergence rate depends only on the manifold dimensionality, which is of independent interest.

Our main contributions are summarized as follows:

- We provide a framework of OT-CFM using an extended Brenier potential, which expresses the OT-CFM vector field with the subdifferential of the potential, covering the manifold hypothesis cases.
- We show that the OT-CFM vector field admits an explicit expression by the proximal operator, which has an interpretation as a denoiser.
- Using convex analysis tools, we prove that perturbations to the initial point shrink in the normal directions to the data manifold and are preserved in the tangential directions, meaning that the manifold structure is stable against perturbation.
- We theoretically and empirically demonstrate that the analysis based on the population OT provides an approximation of the practical minibatch OT-CFM.

2 Formulation of OT-based Flow Matching via Proximal Operators

2.1 OT-CFM

We use the notation of the conditional path (1) and vector field (2). Throughout this paper, we assume that the source distribution P_0 is absolutely continuous with respect to the Lebesgue measure

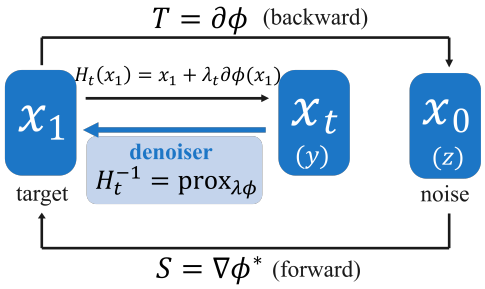


Figure 1: Relation among a target sample x_1, noise sample x_0, and an interpolation x_t. Here ϕ denotes the extended Brenier potential (Sec. 2.2). The target x_1 and noise x_0 are connected via $x_0 \in T(x_1)$, where $T = \partial\phi$ is the (extended) Brenier map. In parallel, x_t is mapped back to x_1 via the denoiser (Sec. 2.2).

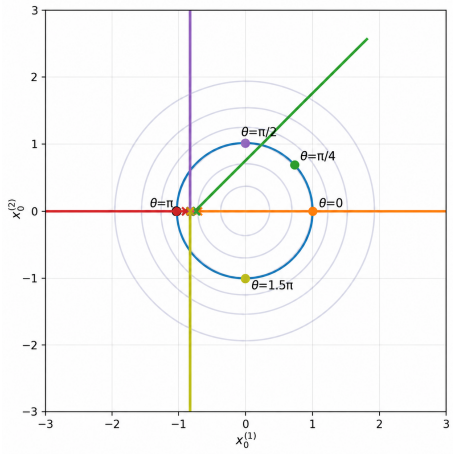


Figure 2: Examples of coupling. $P_0 = N(0, I_2)$ and P_1: von Mises distribution $\rho_{1,0}$. The point on the circle and the half line in the same color are coupled by OT. Note that the direction of $\partial\phi(x_1)$ is normal to the manifold at x_1.

on $\mathbb{R}^d$, and $P_0, P_1 \in \mathcal{P}_2(\mathbb{R}^d)$, which have a finite second moment. The source P_0 is typically the standard normal $N(0, I_d)$ in practice, but except Sec.2.2 we do not use special properties of Gaussian. Also, we do *not* assume that P_1 admits a density function; rather, we are interested in the case where the support of P_1 is a low-dimensional manifold. The extended real numbers are denoted by $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$.

For simplicity of theoretical expositions, this paper only considers the affine paths $x_t = (1-t)x_0 + tx_1$. This is not restrictive; extending to general schedules (α_t, β_t) is easy under appropriate changes. By eliminating x_0 in (2) by $x_0 = (x_t - tx_1)/(1-t)$, we obtain

$$v_t(x_t) = -\frac{x_t - \mathbb{E}[x_1|x_t]}{1-t}. \tag{3}$$

This is essentially the same as the drift term of diffusion models, in which the term $\mathbb{E}[x_1|x_t]$ is often regarded a *denoiser* to estimate clean x_1 from noisy x_t. The form of (3) is general to any CFM methods using affine paths. The only non-trivial part is $\mathbb{E}[x_1|x_t]$, which is also the only part dependent on the coupling (x_0, x_1).

In our theoretical analysis, we primarily consider the OT-CFM *in population*; instead of OT coupling in minibatches, the population OT between P_1 and P_0 is considered to generate random paths (2) and $\mathbb{E}[x_1|x_t]$. In particular, the OT with the quadratic cost, $\inf_{\pi \in \Pi(P_0,P_1)} \int \frac{1}{2}\|x_0 - x_1\|^2 \, d\pi(x_0, x_1)$, is considered, where $\Pi(P_0, P_1)$ denotes the joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with the marginals $P_0, P_1 \in \mathcal{P}_2(\mathbb{R}^d)$. We will discuss the convergence of the minibatch OT with neural net learning theoretically in Sec. 4 and empirically in Sec. 6, demonstrating that the theoretical argument based on population OT provides a reasonable approximation of the practical minibatch OT-CFM.

2.2 Extended Brenier potential and proximal form of OT-CFM vector field

To derive $\mathbb{E}[x_1|x_t]$ in (3), we consider the transport map from P_1 to P_0 to eliminate x_0[1] When the target P_1 does not have a density, such as in the manifold hypothesis, an OT *map* or *Brenier map* from P_1 to P_0 may not exist [10]. Nevertheless, for the quadratic cost, the OT plan always admits an expression by the subdifferential of some convex potential. More concretely, it is known [60, Theorem 12.2; 52] that, for an OT plan $\pi^\star$, there exists a proper lower semicontinuous (lsc) convex function $\phi : \mathbb{R}^d \to \overline{\mathbb{R}}$ such that

$$(x_0, x_1) \in \mathrm{supp}(\pi^\star) \quad \Longrightarrow \quad x_0 \in \partial\phi(x_1), \tag{4}$$

[1]We can also consider the transport map, either from P_0 to P_1 or P_1 to P_0. We use the latter in this paper to see the local behavior around the manifold.

where $\partial\phi$ is the subdifferential of ϕ [7]. Extending the usual terminology to distributions P_1 that need not admit a density, we call such ϕ the *extended Brenier potential* for the transport from P_1 to P_0 in this paper. The extended Brenier potential may not be differentiable; if P_1 has a p.d.f., then ϕ is differentiable P_1-a.e., and (4) reduces to the Brenier map $x_0 = \nabla\phi(x_1)$ [10]. Intuitively, if the support of P_1 is an m-dimensional manifold, a point x_1 couples with a $(d-m)$-dimensional subset in $\partial\phi(x_1)$, induced by $\pi^\star$ (see Sec. 2.2.1 for an example).

We express the vector field of the OT-CFM using the proximal operator (see, e.g., [45, 50]), which is commonly used in convex analysis. Let (x_0, x_1) be a pair sampled by the OT plan $\pi^\star$. Then, using $x_0 \in \partial\phi(x_1)$, we have, for $t \in (0,1)$,

$$x_t = (1-t)x_0 + tx_1 \;\in\; t\Big(I + \frac{1-t}{t}\partial\phi\Big)(x_1), \tag{5}$$

where I is the identity operator. This gives an inclusion expression of x_t in terms of x_1. To obtain $\mathbb{E}[x_1|x_t]$, we wish to have an inverse relation, which can be realized by the proximal operator.

Generally, for a convex function $F : \mathbb{R}^d \to \overline{\mathbb{R}}$ and $\lambda > 0$, the *proximal operator* of F with parameter λ is defined by

$$\mathrm{prox}_{\lambda F}(y) \coloneqq \arg\min_{u\in\mathbb{R}^d}\Big\{F(u) + \frac{1}{2\lambda}\|u-y\|^2\Big\}. \tag{6}$$

This can be regarded as the minimizer of the L^2-regularized objective. For differentiable F, $x = \mathrm{prox}_{\lambda F}(y)$ if and only if $y = x + \lambda\nabla F(x)$, which allows the interpretation as an implicit Euler step.

The basic relation for our purpose is the next lemma, which does *not* require the differentiability of ϕ.

Lemma 2.1. *For (x_0, x_1) sampled from the population OT coupling $\pi^\star$, let $x_t = (1-t)x_0 + tx_1$ and ϕ be defined as above. For $t \in (0,1)$,*

$$x_1 = \mathrm{prox}_{\lambda_t\phi}\Big(\frac{x_t}{t}\Big), \qquad \lambda_t = \frac{1-t}{t}. \tag{7}$$

Consequently, the vector field given by the population OT coupling is represented by

$$u^\star(x_t) = -\frac{x_t - \mathrm{prox}_{\lambda_t\phi}(x_t/t)}{1-t}. \tag{8}$$

Proof. From (5), we have $0 \in \partial(\lambda_t\phi(x_1) + \frac{1}{2}\|x_1 - \frac{x_t}{t}\|^2)$, which means x_1 is the minimizer of $\lambda_t\phi(\cdot) + \frac{1}{2}\|\cdot - \frac{x_t}{t}\|^2$. Thus, $x_1 = \mathrm{prox}_{\lambda_t\phi}(x_t/t)$ by definition. □

Note that in the current setting of P_0 and P_1, x_1 is a deterministic, single-valued map of x_t. Eq. (8) provides a formula for the OT-CFM vector field as an explicit function of location x_t. We use this explicit form to discuss the stability of the manifold structure in Sec. 3.

Denoiser view. The prox term of (8) can be interpreted as a precise *denoiser* in the following "noise-corrupted observation" sense, when $P_0 \sim N(0, I_d)$. It is similar to diffusion models, where $\mathbb{E}[x_1|x_t]$ is regarded as a denoiser. For simplicity, suppose that there exists a Brenier *map* $T = \nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\# P_1 = P_0$, where ϕ is a differentiable Brenier potential and $T_\# P_1$ is the pushforward of P_1 by T. Sample a clean data $x_1 \sim P_1$ and define its *OT code* by $z \coloneqq T(x_1)$, which follows Gaussian distribution as a marginal. Now, define the observed (corrupted) variable by $y \coloneqq x_1 + \lambda z = x_1 + \lambda\nabla\phi(x_1)$. Then the proximal map recovers the clean data x_1 given corrupted y by $\mathrm{prox}_{\lambda\phi}(y) = x_1$ exactly (Eq. (7)). Hence $\mathrm{prox}_{\lambda\phi}$ is a deterministic denoiser for this OT-induced corruption: it inverts the "noising" mechanism $x \mapsto x + \lambda T(x)$. This makes a clear contrast to diffusion models, where independent Gaussian noise is used; hence, its denoising is statistical. See Fig. 1 for the relationship among (x_1, y, z).

2.2.1 An example of OT coupling for low-dimensional manifold

As an example of OT coupling under the manifold hypothesis, we consider the von Mises distribution on the unit circle $\mathbb{S}^1$ as P_1 and $N(0, I_2)$ as P_0. The details are shown in Appendix A. The density of von Mises distribution with mean $\mu \in [0, 2\pi)$ and concentration $\kappa > 0$ is given by

$$\rho_{\kappa,\mu}(\theta) = \frac{\exp\{\kappa\cos(\theta-\mu)\}}{2\pi I_0(\kappa)}, \qquad (\theta \in [0, 2\pi)),$$

where I_0 is the modified Bessel function of the first kind.

Let $e_\theta = (\cos\theta, \sin\theta) \in \mathbb{S}^1$. Using the Kantorovich dual formulation, we can see that the function $a(\theta)$ that attains the minimum of

$$\min_a \int_{\mathbb{R}^2} \sup_{\theta\in[0,2\pi)} \{\langle x, e_\theta\rangle - a(\theta)\} dP_0(x) + \int_0^{2\pi} a(\theta) d\theta,$$

gives the fiber $\partial\phi(e_\theta)$ in the form of

$$\partial\phi_a(e_\theta) = \left\{a'(\theta)e_\theta^\perp + se_\theta : s \geq r_a(\theta)\right\},$$

where $r_a(\theta) \coloneqq \sup_{\eta\neq\theta} \frac{a(\theta)-a(\eta)+a'(\theta)\sin(\eta-\theta)}{1-\cos(\eta-\theta)}$ (see Appendix A for details). Figure 2 shows the $e_\theta = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ and the coupled fibers in the same color. We can see that the fiber coupled with e_θ is normal to the tangent space $T_{e_\theta}\mathbb{S}$. This is generic; see Sec. 3.2.

3 Analysis of Terminal Lyapunov Exponents

This section analyzes the terminal behavior of flow matching (FM) under the low-dimensional manifold hypothesis, based on the proximal formulation developed in Sec. 2. In data generation, it is often observed that small perturbations to the initial condition lead to variations that remain close to the data manifold, suggesting anisotropic stability of the dynamics: contraction in normal directions and neutral behavior along tangential directions. To formalize this observation, we study the local behavior of trajectories via the Jacobian of the vector field, and characterize the behaviors using Lyapunov exponents. The explicit proximal expression (8) allows us to apply active-manifold sensitivity results [25, 19]. For the affine interpolation $x_t = (1-t)x_0 + tx_1$, this yields an explicit Jacobian along each qualifying OT trajectory. We use it to characterize the terminal growth rates of infinitesimal perturbations.

3.1 Lyapunov exponents

We review the definition of Lyapunov exponents. Consider, in general, the ODE with time $\tau \in [0, \infty)$

$$x'(\tau) = u_\tau(x(\tau)), \tag{9}$$

where $x'(\tau) \coloneqq \frac{dx(\tau)}{d\tau}$ denotes the time derivative with respect to τ in the sequel. For $\tau_0 > 0$ and $v \in \mathbb{R}^d \setminus \{0\}$, the *Lyapunov exponent* in the direction v is defined (if it exists) by

$$\lambda_{\tau_0}(v) \coloneqq \lim_{\tau\to+\infty} \frac{1}{\tau - \tau_0} \log\left\|D_x\Phi_{\tau,\tau_0}(x(\tau_0))\, v\right\|, \tag{10}$$

where Φ_{τ,τ_0} denotes the associated flow from τ_0 to τ; $x(\tau) = \Phi_{\tau,\tau_0}(x(\tau_0))$, and $D_x\Phi_{\tau,\tau_0}(x(\tau_0))$ is the space derivative of Φ_{τ,τ_0} at $x(\tau_0)$. See, e.g. [4, 26]. The norm $\|\cdot\|$ is any norm on $\mathbb{R}^d$; as all norms are equivalent, $\lambda_{\tau_0}(v)$ is norm-independent. Negative exponents (resp., positive) correspond to exponential contraction (resp., expansion) of nearby trajectories in τ-time. Let

$$\xi(\tau) \coloneqq D_x\Phi_{\tau,\tau_0}(x(\tau_0))v.$$

From (9), ξ satisfies the ODE

$$\xi'(\tau) = D_x u_\tau(x(\tau))\xi(\tau).$$

We therefore analyze the growth of solutions of this variational equation. For the OT-CFM considered below, the proximal sensitivity formula provides a fixed tangent–normal decomposition along the reference trajectory, allowing the equation to be solved explicitly.

3.2 Proximal sensitivity and tangent–normal splitting

We use existing results of active-manifold sensitivity [25, 19] to describe the local behavior of the proximal denoiser. See Appendix C for details.

Fix an OT pair (x_1, x_0) with $x_0 \in \partial\phi(x_1)$ and define

$$f \coloneqq \phi - \langle x_0, \cdot\rangle, \qquad \Xi_\lambda \coloneqq \mathrm{prox}_{\lambda f} \quad (\lambda > 0). \tag{11}$$

Then f is proper, lsc, and convex, and $0 \in \partial f(x_1)$, so that $\Xi_\lambda(x_1) = x_1$. In the sequel, we make the following assumptions:

(M) $\mathcal{M} \subset \mathbb{R}^d$ is a C^2 submanifold of dimension $m(< d)$, and the support of P_1 is $\mathcal{M}$.
(PS) f is C^2-partly smooth at x_1 relative to $\mathcal{M}$. (See Sec. C.1 and [25] for the definition.)
(ND) $0 \in \mathrm{ri}(\partial f(x_1))$, equivalently $x_0 \in \mathrm{ri}(\partial \phi(x_1))$, where ri stands for relative interior.

For $x_1 \in \mathcal{M}$, the tangent and normal spaces are denoted by $T_{x_1}\mathcal{M}$ and $N_{x_1}\mathcal{M}$, respectively (normal with respect to the inner product of $\mathbb{R}^d$), so that $\mathbb{R}^d \cong T_{x_1}\mathcal{M} \oplus N_{x_1}\mathcal{M}$. The assumption (PS) means that $f|_{\mathcal{M}}$ is C^2 near x_1 that ∂f is nonempty at nearby points of $\mathcal{M}$ and continuous at x_1 relative to $\mathcal{M}$, and that $\mathrm{par}(\partial f(x_1)) = N_{x_1}\mathcal{M}$ [25, Definition 2.3]. Here $\mathrm{par}(C) \coloneqq \{a - b : a, b \in \mathrm{aff}(C)\}$, and continuity is understood in the set-valued sense of inner and outer semicontinuity. These assumptions do not require the differentiability of ϕ in the ambient $\mathbb{R}^d$. A sufficient condition for (ND) to hold almost surely is given in Appendix C.3.

Write $T \coloneqq T_{x_1}\mathcal{M}$ and $N \coloneqq N_{x_1}\mathcal{M}$ for simplicity, and let Π_T and Π_N be the orthogonal projections onto these spaces. Define

$$H \coloneqq \mathrm{Hess}_{\mathcal{M}}(f|_{\mathcal{M}})(x_1) : T \to T, \tag{12}$$

using the Riemannian metric induced by the ambient Euclidean space. Since x_1 minimizes f, its Hessian H at x_1 is self-adjoint and positive semidefinite.

Proposition 3.1. *Under (M), (PS), and (ND), for every fixed $\lambda > 0$, Ξ_λ is C^1 on a neighborhood U_λ of x_1 and $\Xi_\lambda(U_\lambda) \subset \mathcal{M}$. Its Jacobian at x_1 is*

$$D\Xi_\lambda(x_1) = (I_T + \lambda H)^{-1}\Pi_T, \tag{13}$$

where I_T is the identity on T. In particular,

$$\|D\Xi_\lambda(x_1) - \Pi_T\|_{\mathrm{op}} \le \lambda \|H\|_{\mathrm{op}}. \tag{14}$$

The statements of local identification $\Xi_\lambda(U_\lambda) \subset \mathcal{M}$ and smoothness of Ξ_λ follow from Hare and Lewis [25, Theorem 5.3] and the sensitivity argument of Daniilidis et al. [19, Theorems 28–29]. Although the latter reference treats general prox-regular functions and states these results for sufficiently small $\lambda > 0$, in the present convex setting, the argument applies to every $\lambda > 0$. Appendix C.2 proves these claims, and derives (13) by differentiating the manifold-restricted optimality equation.

Application to the OT denoiser. For $0 < t < 1$, use the affine interpolation and denoiser

$$x_t = (1-t)x_0 + tx_1, \qquad \lambda_t = (1-t)/t, \qquad Q_t(x) \coloneqq \mathrm{prox}_{\lambda_t \phi}(x/t).$$

It is easy to see the shift identity $\mathrm{prox}_{\lambda\phi}(y) = \Xi_\lambda(y - \lambda x_0)$, which gives $Q_t(x_t + h) = \Xi_{\lambda_t}(x_1 + h/t)$. Hence, from (13), the relation

$$DQ_t(x_t) = \frac{1}{t} D\Xi_{\lambda_t}(x_1) = \big(tI_T + (1-t)H\big)^{-1}\Pi_T \tag{15}$$

holds. Thus $DQ_t(x_t)$ annihilates normal directions and converges to Π_T as $t \uparrow 1$. Note that the directions T and N are determined at the location $x_1 \in \mathcal{M}$.

3.3 Terminal Lyapunov exponents and manifold stability

To follow the standard terminology of Lyapunov analysis, we use the time interval $[0, \infty)$ by setting $\tau \coloneqq -\log(1-t)$ or $t(\tau) = 1 - e^{-\tau}$. Then, $\tau \in [0, \infty)$ and $dt/d\tau = 1 - t$. The rescaled OT-CFM equation is

$$x'(\tau) = u_\tau(x(\tau)), \qquad u_\tau(x) \coloneqq Q_{t(\tau)}(x) - x,$$

where $x'(\tau)$ denotes the derivative with respect to time τ. Fix the OT trajectory $x(\tau) = (1 - t(\tau))x_0 + t(\tau)x_1$ and use the notation of Sec. 3.2. Put $B_t \coloneqq tI_T + (1-t)H$. Then, Eq. (15) gives the block decomposition

$$D_x u_\tau(x(\tau)) = -\Pi_N + (B_t^{-1} - I_T)\Pi_T, \qquad t = t(\tau). \tag{16}$$

In particular, the fixed subspaces $T_{x_1}\mathcal{M}$ and $N_{x_1}\mathcal{M}$ are preserved over t or τ by the variational equation. Note again that these spaces are defined at the endpoint x_1, not at the intermediate point x_t.

Theorem 3.2. *Fix an OT pair* (x_1, x_0) *and assume* (M), (PS), *and* (ND). *For any* $\tau_0 > 0$, *let* $t_0 = t(\tau_0)$ *and write* $v = v_T + v_N$ *according to* $\mathbb{R}^d = T_{x_1}\mathcal{M} \oplus N_{x_1}\mathcal{M}$. *Then*

$$D_x\Phi_{\tau,\tau_0}(x(\tau_0))v = B_{t(\tau)}B_{t_0}^{-1}v_T + e^{-(\tau-\tau_0)}v_N. \tag{17}$$

Consequently, the Lyapunov exponents defined by (10) *satisfy*

$$\lambda_{\tau_0}(v) = -1 \quad \textit{for } v \in N_{x_1}\mathcal{M} \setminus \{0\}, \qquad \lambda_{\tau_0}(v) = 0 \quad \textit{for } v \in T_{x_1}\mathcal{M} \setminus \{0\}.$$

The same zero exponent holds whenever $v_T \neq 0$.

Proof. For any finite time interval beginning at $\tau_0 > 0$, the local smoothness in Prop. 3.1 gives the variational equation

$$\xi'(\tau) = D_x u_\tau(x(\tau))\xi(\tau), \qquad \xi(\tau_0) = v. \tag{18}$$

By (16), the normal component solves $\xi_N' = -\xi_N$. For the tangent component, observe that $B_t^{-1} - I_T = (1-t)(I_T - H)B_t^{-1}$ and $\frac{d}{d\tau}B_{t(\tau)} = (1-t)(I_T - H)$. It follows from (16) that

$$\xi_T' = B_{t(\tau)}'B_{t(\tau)}^{-1}\xi_T.$$

Therefore, the solution is $\xi_T(\tau) = B_{t(\tau)}B_{t_0}^{-1}v_T$, proving (17). Since $H \succeq 0$, B_{t_0} is invertible for $t_0 > 0$, and $B_t \to I_T$ as $t \uparrow 1$. Thus a nonzero tangent component has the nonzero limit $B_{t_0}^{-1}v_T$, while the normal component decays exactly as $e^{-(\tau-\tau_0)}$. The claimed exponents follow. □

Theorem 3.2 concerns infinitesimal perturbations along a reference OT trajectory. Their normal components contract by the exact factor $e^{-(\tau-\tau_0)} = (1-t)/(1-t_0)$, whereas nonzero tangential components approach finite, nonzero limits. Note that the normal Lyapunov exponent -1 depends on the time scheduling; it is not difficult to see that if we use (1) with $\alpha_t = (1-t)^\gamma$, the exponent is $-\gamma$. The result of the theorem agrees with the stability analysis of the diffusion models of [12], which showed that if the Lyapunov exponents along the tangential directions are dominant, the manifold structure is stable in dynamics. Some recent work [9, 36, 39] showed the Taylor expansion of the score function, demonstrating zero Lyapunov exponents along the manifold and negative along the normal directions. Although the Lyapunov exponents suggest similar behaviors, the trajectories of OT-CFM may not approach to the manifold in the normal direction.

4 Minibatch OT-CFM and Population OT-CFM

Theoretical arguments in the previous sections consider the population OT for constructing the endpoints of the paths in CFM. However, in practical situations where we have only finite data, CFM methods often use the minibatch OT, where a batch of size B is taken from the training data and also from P_0. The samples in the two batches are coupled by an OT method, such as the Hungarian algorithm, resulting in the matched samples $\{(X_{0,i}, X_{1,i}^{\mathrm{match}})\}_{i=1}^B$. A random path $X_t = (1-t)X_{0,i} + tX_{1,i}^{\mathrm{match}}$ is constructed by a random choice of a pair $(X_{0,i}, X_{1,i}^{\mathrm{match}})$, and a neural network model uses the velocity $V(X_t, t) \coloneqq X_1^{\mathrm{match},i} - X_0^i$ as the teaching data at (X_t, t) for stochastic gradient descent (SGD).

This section makes a theoretical argument that the population OT-CFM provides a reasonable surrogate of the practical minibatch OT-CFM by showing the convergence of the latter to the former as the sample size (N) and the batch size (B) go to infinity. This implies that the minibatch OT-CFM and population OT-CFM provide similar behaviors asymptotically. We also derive a convergence rate, in which the exponents depend only on the manifold dimensionality. We give the result informally in this section, and the details are deferred in Section D.

One of the difficulties in discussing the finite-sample convergence is that the vector field given by the batches can be defined only on the finite paths, which may have measure zero. Thus, we need to discuss regression of the vector field, typically by neural network models. This paper considers an idealized regression procedure following [8], but combines it with neural network learning as a new development. Consider independent batches $X_{0,1}, \ldots, X_{0,B} \overset{\mathrm{iid}}{\sim} P_0$ and $X_{1,1}, \ldots, X_{1,B} \overset{\mathrm{iid}}{\sim} P_1$, and

define empirical distributions $\widehat{P}_{0,B} \coloneqq \frac{1}{B}\sum_{i=1}^{B}\delta_{X_{0,i}}$ and $\widehat{P}_{1,B} \coloneqq \frac{1}{B}\sum_{j=1}^{B}\delta_{X_{1,j}}$. Suppose an OT solver selects an optimal permutation $\widehat{\pi}_B$. Then, a neural network u is learned via minimizing

$$\mathbb{E}_{(X_0,X_1)\sim\widehat{\pi}_B,t}\|u(X_t,t)-V(X_0,X_1)\|^2,$$

where t is taken independently. Assume that we generate such random batches L times independently. Then, it is well known that the least-squares learning ideally converges as $L\to\infty$ to the average of the vector field over the random choice of batches:

$$\mathbb{E}_B\mathbb{E}_{(X_0,X_1)\sim\widehat{\pi}_B}[V(X_0,X_1)|X_t] = \arg\min_u \mathbb{E}_B\mathbb{E}_{(X_0,X_1)\sim\widehat{\pi}_B,t}\|u(X_t,t)-V(X_0,X_1)\|^2, \quad (19)$$

where $\mathbb{E}_B$ is the average over the random batches. We consider the *expected batch OT plan* $\bar{\pi} \coloneqq \mathbb{E}_B[\widehat{\pi}_B]$ following [8], the average of the minibatch OT plan $\widehat{\pi}_B$ over the random batches. If we take a random pair with the law $\bar{\pi}$, the resulting expected vector field is equal to (19), so we can analyze $\bar{\pi}$.

Neural network learning. For simplicity, we restrict our discussion in this section to a sparse ReLU network defined on a compact set. Let $K_i \coloneqq \operatorname{supp} P_i$ $(i=0,1)$, and assume that they are compact sets and K_1 equals an m-dimensional manifold M. Although compact K_0 excludes Gaussian distributions, this does not cause much practical harm since the source P_0 for FM can be arbitrary, such as truncated Gaussian. Let $I=[0,c]$ with $0<c\le 1$. Then, the set of possible paths $\mathcal{X}_I \coloneqq \left\{(t,(1-t)x_0+tx_1) : t\in I,\ x_0\in K_0,\ x_1\in K_1\right\}$ is a compact subset of $\mathbb{R}^{d+1}$; we assume $\mathcal{X}_I\subset[0,1]^{d+1}$, w.l.o.g.

Let $\rho(a)\coloneqq\max\{a,0\}$ be the ReLU activation. A network $\Phi:\mathbb{R}^{q_0}\to\mathbb{R}^{q_J}$ with depth J and widths $\mathbf{q}=(q_0,\dots,q_J)$ is written as

$$\Phi = F_J\circ\rho\circ F_{J-1}\circ\cdots\circ\rho\circ F_1, \qquad F_j(z)=W_jz+b_j, \quad (20)$$

where the ReLU is applied componentwise, and the last affine map is not followed by an activation.

Fix constants $C_D, R_f, C_A, \kappa_A > 0$. We consider a network model $\mathcal{F}_P$ with a complexity budget P (sparsity) defined by

$$\begin{aligned} &J\le C_D\log^2(P+1), \quad \max_j q_j \le P+d+1, \quad \textstyle\sum_j\|W_j\|_0+\|b_j\|_0\le P,\\ &\max_j\{\|W_j\|_\infty,\|b_j\|_\infty\}\le C_AP^{\kappa_A}, \qquad \sup_{z\in\mathcal{X}_I}\|\Phi(z)\|\le R_f. \end{aligned}$$

Convergence result. Let $(X_0,X_1)\sim\pi^\star$ be a pair sampled from the population OT, and define

$$X_t^\star = (1-t)X_0+tX_1 \quad (21)$$

and the corresponding ideal vector field $u^\star$ (8). Suppose that $\widehat{u}_{B,L,P}\in\mathcal{F}_P$ is a trained vector field with L independent repetitions of size-B batches. The ODE for generation is given by

$$\frac{d\widehat{X}_t}{dt} = \widehat{u}_{B,L,P}(\widehat{X}_t,t), \qquad \widehat{X}_0 = X_0^\star = X_0. \quad (22)$$

We assume that the ideal vector field $u^\star$ is smooth of degree s: $u^\star\in C^s([0,1]^{d+1};\mathbb{R}^d)$.

Extending the analysis of the vector field and its flow to the terminal time $t=1$ requires additional care because of the potentially singular behavior of (3) as $t\uparrow 1$; see, e.g., Wan et al. [62, Appendix C.2]. We therefore consider $c<1$ (ideally close to 1) for $I=[0,c]$ in the following theorem.

Theorem 4.1 (Informal)**.** *Let $m>4$ and choose the integer parameter budget P appropriately depending on L and m. Under some assumptions listed in Appendix D, for all sufficiently large L and every $\eta\in(0,1)$, with probability at least $1-\eta$ over the training batches,*

$$\left(\mathbb{E}_\star\sup_{t\in I}\|\widehat{X}_t-X_t^\star\|^2\right)^{1/2} \lesssim B^{-1/m} + \left(\frac{\log^5(eL)}{L}\right)^{s/(2s+m+1)} + \sqrt{\frac{\log(2/\eta)}{L}} + \sqrt{\varepsilon_{\mathrm{opt}}}, \quad (23)$$

where $\mathbb{E}_\star$ denotes the expectation with respect to the independent test data and $\varepsilon_{\mathrm{opt}}$ is the optimization error of neural networks. The Wasserstein error at time c satisfies the same upper bound.

This theorem shows that the resulting distributions of the minibatch OT-CFM and the population OT-CFM are close for large batch size B and the number of batches L. The first term comes from the minibatch bias and the second term is the neural approximation-estimation contribution. The

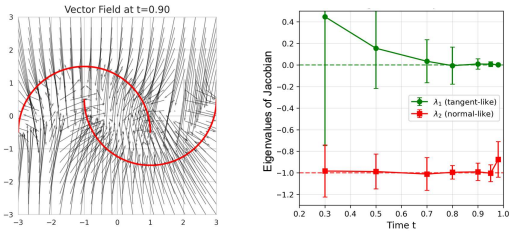


Figure 3: Left: OT-CFM vector field ($t = 0.9$) obtained by a neural network. An attracting force on the manifold can be seen. Right: Mean and STD of the eigenvalues of $(1-t)Dv_t$ over different trajectories. The second eigenvalue corresponds to a Lyapounov exponent close to -1.

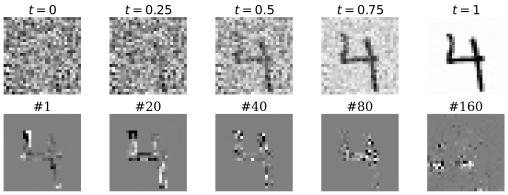


Figure 4: Top: Generation with OT-CFM. Bottom: Eigenvectors of the Jacobian of the vector field $v_t(x)$ with respect to the 1st, 20th, 40th, 80th, and 160th eigenvalues in the descending order.

polynomial rates of convergence on B and L depend only on the manifold dimensionality m. Note that the dimensionality of the input to the learned vector field $\widehat{u}_t(x)$ is $d+1$, and thus we need special arguments for the approximation error of the model to derive the bound depending on m. To achieve this, we basically extend the approximation on M by using the affine structure of the population OT, extending the arguments developed for diffusion models in [23] to the OT-CFM. To the best of our knowledge, this is the first result of the convergence rate of minibatch OT-CFM depending on the intrinsic dimensionality under the manifold hypothesis.

5 Related work

Proximal view to diffusion models. Diffusion models have been discussed from the proximal viewpoint with the Wasserstein metric [68, 35, 11]. These works consider proximal operators in infinite dimensional probability spaces and thus differ from the current paper.

Diffusion models and the manifold hypothesis. Some recent work discusses the behavior of diffusion models under the manifold hypothesis; Pidstrigach [47] derived conditions under which score-based models sample the underlying low-dimensional data manifold. Oko et al. [44] showed that the asymptotic convergence rate of learning the target distribution depends on the dimension of the manifold, not the ambient space. In addition to the work discussed in the end of Sec. 3, the dynamical behavior of diffusion models under low-dimensional subspaces has also been studied, such as the orthogonal decomposition of the space [13] and the provable recovery [63]. Compared with these results, analogous analysis for FM under the manifold hypothesis remains limited.

Flow matching and proximal viewpoints. FM models are often trained with improved couplings of minibatch OT [48, 59]. Our work provides a theoretical basis for this line with a convex-analytic characterization, enabling stability analysis in manifold-supported cases. Our framework does not apply directly to rectified flow [38]. Note that our usage of proximal operators is different from the standard view of finite proximal-steps, which include Schrödinger bridges/entropic OT as KL-proximal iterations [40, 18]; or Wasserstein-proximal step [3]. The most relevant to our work is Optimal Flow Matching [32], which uses convex potentials to carry straight trajectories (P_0 to P_1). Although their flow inversion is equivalent to a proximal operator, they do not explicitly adopt the proximal or denoiser perspective, nor do they provide a geometric analysis.

6 Experiments

Toy examples. We applied the minibatch OT-CFM to the dataset "Two Moons" (Fig. 3), where the target P_1 is uniform on the red curves. See Sec. E.1 for an additional toy example. The vector field is trained by an MLP with batch size 512. The Hungarian algorithm is applied for the minibatch OT. The left figure depicts the trained vector field, and the right figure shows the means and STDs of two eigenvalues of $D_x v_t(x_t)$ for different t over 100 initial points. The values are well separated with $\lambda_1 = 0$ and $\lambda_2 = -\gamma = -1.0$ for $t \to 1$, which agrees with Theorem 3.2. Note that the example has two arcs with boundaries, at which the theoretical analysis in Sec. 3 does not necessarily apply.

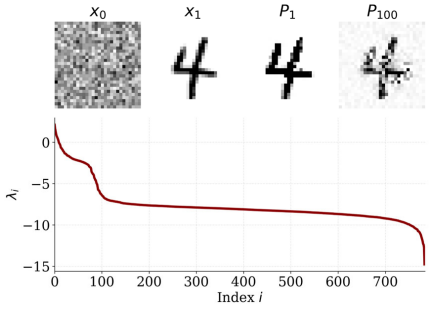


Figure 5: Top: x_1 generated with $x_0 \sim N(0, I_d)$, and perturbed images by adding to x_1 the eigenvectors P_1 and P_{100}. Bottom: sorted eigenvalues of $(1-t)D_x v(x_t)$ in the descending order.

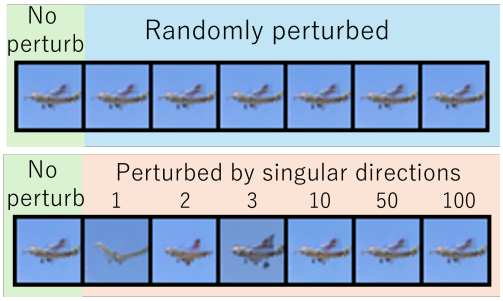


Figure 6: Top: Generated images with randomly perturbed initial points. Bottom: Generated images x_0^k with initial points by adding the right-sigular vector of $D_x\widehat{\Phi}(x_0)$. The number k indicates the k-th largest singular values. The random perturbations give almost the same images as the original, while the perturbations for the 1st, 2nd, and 3rd sigular vectors generate slightly modified images from the original.

MNIST. We used MNIST images to train OT-CFM with minibatch size 256. Fig. 4, bottom, presents the top eigenvectors (up to 160th) of $D_x v_t(x)$. Up to the 80th eigenvalue, they have a strong correlation to the shape "4". Fig. 5 shows the eigenvalues of $(1-t)D_x v(x_t)$ at $t = 0.98$. We can see that there is a clear gap in the eigenspectrum; about 100 eigenvalues are significantly larger than the others. The second right (P_1) and rightmost (P_{100}) images are perturbed from a generated image x_1 with eigenvectors of the 1st and 100th largest eigenvalue, respectively. The results suggest that the eigenspaces with the eigenvalues ≈ 0 correspond to the tangential directions to the manifold that constitutes the MNIST images. See also Sec. E.3. In line with the assertion of Theorem 3.2, We have also carried out experiments by perturbing the initial point and investigated its effects on the generated images. The perturbation causes only the change along the manifold of MNIST images, and we can also analyze the effective dimensionality with PCA. See Sec. E.4 for the details.

CIFAR-10. We conduct experiments on perturbations to initial points with CIFAR-10. Sinkhorn algorithm is applied to approximate OT with minibatch size 1024, and the UNet is used to learn a vector field. After training, perturbations are applied to initial points in two ways: (A) Uniformly random directions, (B) Singular directions: right singular vectors of the singular value decomposition of $D_x\widehat{\Phi}(x_0)$. See the results in Fig. 6. The dimensionality of the data manifold should be much smaller than the ambient dimension of 3072. Most of the components in the random perturbation tend to be normal to the manifold, and they shrink at the terminal points. In contrast, the results show that the top singular vectors tend to be tangential to the manifold, resulting in small changes along the manifold.

Minibatch OT-CFM. Beyond the convergence of the vector field of minibatch OT in Sec. 4, we empirically compare the Lyapunov exponents between the minibatch and population OT-CFM. We describe the summary here and defer the details to Sec. E.2. The empirical eigenvalues of the Jacobian $(1-t)D_x v_t$ obtained by minibatch OT-CFM with uniform data on $\mathbb{S}^1$ in $\mathbb{R}^2$ do not differ significantly from the theoretical ones for population OT-CFM as $t \to 1$, while larger batch sizes converge more smoothly. The results provide empirical support for the convergence of the minibatch OT-CFM in the level of the derivative of vector fields.

7 Conclusions

We developed a proximal-operator framework for OT-based conditional FM. By representing the quadratic-cost OT via an extended Brenier potential, we obtained an exact proximal form of the FM vector field and its denoiser viewpoint, without requiring the target distribution to admit a density. We also established a terminal Lyapunov theory under the manifold hypothesis, which implies stability of the manifold structure against perturbation to the dynamics.

Acknowledgements

KF is partially supported by JSPS KAKENHI 26K02990 and JSPS Grant-in-Aid for Transformative Research Areas (A) 22H05106. WH is supported by JSPS KAKENHI (24K20848) and JST BOOST (JPMJBY24G6). HB is supported by JST PRESTO (JPMJPR24K6). The authors also thank the anonymous reviewers for their constructive comments, particularly for drawing attention to relevant literature on active manifolds and suggesting refinements to the convergence proof.

References

[1] M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhauser Verlag AG, Basel, Switzerland, 2 edition, Dec. 2008.

[3] M. Balcerak, T. Amiranashvili, A. Terpin, S. Shit, L. Bogensperger, S. Kaltenbach, P. Koumoutsakos, and B. Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[4] L. Barreira and Y. Pesin. *Lyapunov Exponents and Smooth Ergodic Theory*. Student Mathematical Library. American Mathematical Society, 2002. ISBN 9780821829219. URL https://books.google.co.jp/books?id=9kLyBwAAQBAJ.

[5] P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006. doi: 10.1007/s00440-005-0462-3.

[6] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[7] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2nd edition, 2017.

[8] S. Boïté, J. Delon, and K. Nadjahi. Expected batch optimal transport plans and consequences for flow matching. *arXiv preprint arXiv:2605.12174*, 2026.

[9] V. D. Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=MhK5aXo3gB.

[10] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[11] K. F. Caluya and A. Halder. Wasserstein proximal algorithms for the schrödinger bridge problem: Density control with nonlinear drift. *IEEE Trans. Automat. Contr.*, 67(3):1163–1178, Mar. 2022.

[12] N. Chandramoorthy and A. De Clercq. When and how can inexact generative models still sample from the data manifold? *Advances in Neural Information Processing Systems*, 2025.

[13] M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023.

[14] S. Chen, H. Zhang, M. Guo, Y. Lu, P. Wang, and Q. Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. *Advances in Neural Information Processing Systems*, pages 27340–27371, 2024.

[15] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025. doi: 10.1177/02783649241273668. URL https://doi.org/10.1177/02783649241273668.

[16] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. S. Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

[17] G. Couairon, R. Singh, A. Charantonis, C. Lessig, and C. Monteleoni. Archesweathergen: Skillful and compute-efficient probabilistic weather forecasting with machine learning. *Science Advances*, 12(17):eadx2372, 2026. doi: 10.1126/sciadv.adx2372. URL https://www.science.org/doi/abs/10.1126/sciadv.adx2372.

[18] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, page 2292–2300, 2013.

[19] A. Daniilidis, W. Hare, and J. Malick. Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. *Optimization*, 55(5–6):481–503, 2006. doi: 10.1080/02331930600815884.

[20] D. Drusvyatskiy and A. S. Lewis. Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential. *SIAM Journal on Optimization*, 23(1):256–267, 2013. doi: 10.1137/120876551.

[21] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021. doi: 10.1109/TIT.2021.3062161.

[22] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024.

[23] G. Fu, T. Suzuki, W. S. Lee, and A. Nitanda. Intrinsic Wasserstein rates for score-based generative models on smooth manifolds. *arXiv:2605.15822*, 2026. URL https://arxiv.org/abs/2605.15822.

[24] K. Fukumizu, T. Suzuki, N. Isobe, K. Oko, and M. Koyama. Flow matching achieves almost minimax optimal convergence. In *The Thirteenth International Conference on Learning Representations*, Oct. 2024.

[25] W. L. Hare and A. S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004. URL https://journalofconvexanalysis.com/articles/jca11017/jca11017.pdf.

[26] M. Hirsch, C. Pugh, and M. Shub. *Invariant Manifolds*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2006. ISBN 9783540373827. URL https://books.google.co.uk/books?id=Py98CwAAQBAJ.

[27] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

[28] S. Hundrieser, T. Staudt, and A. Munk. Empirical optimal transport between different measures adapts to lower complexity. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):824–846, 2024.

[29] K. Ikeda, M. Koyama, J. Zhang, K. Hayashi, and K. Fukumizu. Pairwise optimal transports for training all-to-all flow-based condition transfer model. *Advances in Neural Information Processing Systems*, 2025.

[30] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, pages 23593–23606, 2022.

[31] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. doi: 10.1214/009053606000001019.

[32] N. Kornilov, P. Mokrov, A. Gasnikov, and A. Korotin. Optimal flow matching: Learning straight trajectories in just one step. *Advances in Neural Information Processing Systems*, 37: 104180–104204, 2024.

[33] S. Kumar, Y. Wang, and L. Lin. Flow matching is adaptive to manifold structures. *arXiv [stat.ML]*, Feb. 2026.

[34] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3): 702–725, 2002. doi: 10.1137/S1052623401387623.

[35] W. Li, S. Liu, and S. Osher. A kernel formula for regularized wasserstein proximal operators. *Res. Math. Sci.*, 10(4):43, Dec. 2023.

[36] X. Li, Z. Shen, Y.-P. Hsieh, and N. He. When scores learn geometry: Rate separations under the manifold hypothesis. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=34V0IZytle`.

[37] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

[38] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

[39] Z. Liu, W. Zhang, and T. Li. Improving the Euclidean diffusion generation of manifold data by mitigating score function singularity. *Advances in Neural Information Processing Systems*, 2025.

[40] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014. ISSN 1078-0947. doi: 10.3934/dcds.2014.34.1533. URL `https://www.aimsciences.org/article/id/d5bcf817-901d-4104-b7da-eade7847c53e`.

[41] R. Mifflin and C. Sagastizábal. Proximal points are on the fast track. *J. Convex Anal.*, 9(2): 563–579, 2002.

[42] S. A. Miller and J. Malick. Newton methods for nonsmooth convex minimization: connections among -lagrangian, riemannian newton and SQP methods. *Math. Program.*, 104(2-3):609–633, Nov. 2005.

[43] I. Ohn and Y. Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7):627, 2019. doi: 10.3390/e21070627.

[44] K. Oko, S. Akiyama, and T. Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 26517–26582. PMLR, 4 2023.

[45] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, Jan. 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL `https://doi.org/10.1561/2400000003`.

[46] S. Pi, M. Lu, M. Su, W. Wu, J. Y.-C. Hu, and H. Liu. Learning manifold data with flow matching. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=9mA5d705Gd`.

[47] J. Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, pages 35852–35865, 2022.

[48] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28100–28127. PMLR, July 2023.

[49] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

[50] R. T. Rockafellar and R. J.-B. Wets. *Variatioal Analysis*. Springer, 1998.

[51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[52] L. Rüschendorf and S. T. Rachev. A characterization of random variables with minimum L2-distance. *J. Multivar. Anal.*, 32(1):48–54, Jan. 1990.

[53] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

[54] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875.

[55] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[56] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[57] T. Staudt and S. Hundrieser. Convergence of empirical optimal transport in unbounded settings. *Bernoulli*, 31(3):1929–1954, 2025.

[58] L. Szabó. A simple proof for the Jordan measurability of convex sets. *Elemente der Mathematik*, 52(2):84–86, 1997. doi: 10.1007/s000170050014.

[59] A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW.

[60] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[61] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

[62] Z. Wan, Q. Wang, G. Mishne, and Y. Wang. Elucidating flow matching ODE dynamics via data geometry and denoisers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 62020–62083. PMLR, 2025. URL https://proceedings.mlr.press/v267/wan25e.html.

[63] P. Wang, H. Zhang, Z. Zhang, S. Chen, Y. Ma, and Q. Qu. Diffusion models learn low-dimensional distributions via subspace clustering. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025. URL https://openreview.net/forum?id=R8xgXKmVds.

[64] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. Vazquez Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[65] K. Yakovlev and N. Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In *Proceedings of Machine Learning Research Vol. 291: The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, pages 5824–5891, 2025.

[66] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, Apr. 2024.

[67] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

[68] B. J. Zhang, S. Liu, W. Li, M. A. Katsoulakis, and S. Osher. Wasserstein proximal operators describe score-based generative models and resolve memorization. *SIAM J. Math. Data Sci.*, 8 (3):623–648, Sept. 2026.

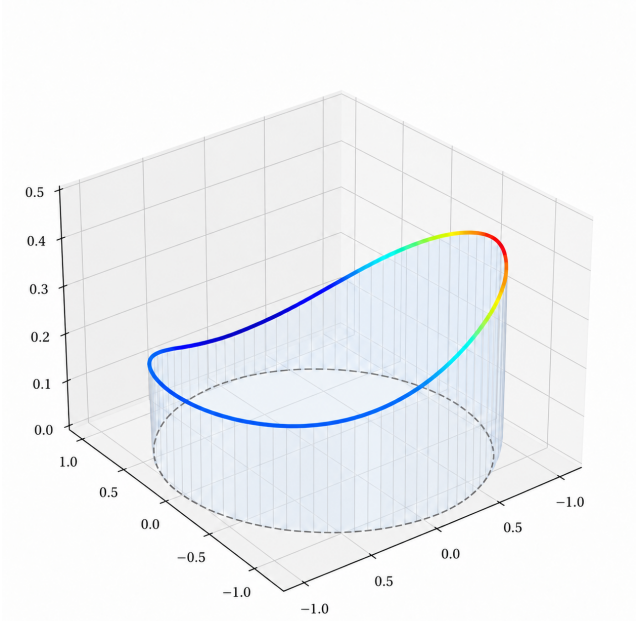


Figure 7: The probability density function of von Mises distribution ($\mu = 0$, $\kappa = 1$). The peak is attained at $(1, 0)$.

A An example of OT under manifold hypothesis

This appendix provides a two-dimensional example illustrating the extended Brenier potential and its subgradient fibers when the target distribution is supported on a one-dimensional manifold. The example demonstrates that an OT fiber is parallel to the normal direction at the target manifold point.

Let $M = \mathbb{S}^1 = \{e_\theta = (\cos\theta, \sin\theta) : \theta \in [0, 2\pi)\}$ is the unit circle in $\mathbb{R}^2$, and define $e_\theta^\perp := (-\sin\theta, \cos\theta)$. We take $P_0 = N(0, I_2)$ and P_1 is a von Mises distribution on $\mathbb{S}^1$:

$$P_1(d\theta) = \rho_{\kappa,\mu}(\theta)\, d\theta, \qquad \rho_{\kappa,\mu}(\theta) = \frac{\exp\{\kappa\cos(\theta - \mu)\}}{2\pi I_0(\kappa)},$$

where $\kappa \geq 0$, $\mu \in [0, 2\pi)$, and I_0 is the modified Bessel function of the first kind. The case $\kappa = 0$ corresponds to the uniform distribution on the circle. See Figure 7 for an example of the density function $\rho_{\kappa,\mu}$.

A.1 Dual representation of the OT map

Consider the Kantorovich dual problem:

$$\sup_{u,v}\left\{\int_{\mathbb{R}^2} u(x)dP_0(x) + \int_{\mathbb{S}^1} v(e)dP_1(e) : u(x) + v(e) \leq \frac{1}{2}\|x - e\|^2\right\},$$

and let $u_\star$, $v_\star$ be the Kantorovich potential for P_0, P_1, respectively. If we introduce

$$a(\theta) := \frac{1}{2}\|e_\theta\|^2 - v(e_\theta) = \frac{1}{2} - v(e_\theta)$$

and

$$\Phi_a(x) := \sup_{\theta\in[0,2\pi)}\left\{\langle x, e_\theta\rangle - a(\theta)\right\} = \sup_{\theta\in[0,2\pi)}\left\{v(e_\theta) - \frac{1}{2}\|e_\theta\|^2 + \langle x, e_\theta\rangle\right\}, \tag{24}$$

we have $u(x) = \frac{1}{2}\|x\|^2 - \Phi_a(x)$, and the Kantorovich dual is equivalent to the convex minimization problem

$$\min_a \mathcal{J}(a), \qquad \mathcal{J}(a) := \int_{\mathbb{R}^2} \Phi_a(x)\, dP_0(x) + \int_0^{2\pi} a(\theta)\rho_{\kappa,\mu}(\theta)\, d\theta. \tag{25}$$

It is also known that the Brenier map S from P_0 to P_1 is given by

$$S(x) \in \arg\min_{e\in\mathbb{S}^1} \left\{ \frac{1}{2}\|x-e\|^2 - v_\star(e) \right\},$$

and the above expression can be rewritten as

$$S(x) \in \arg\max_{\theta\in[0,2\pi)} \langle x, e_\theta\rangle - a_\star(\theta),$$

where $a_\star$ corresponds to $v_\star$ and

$$S(x) = e_{\Theta_{a_\star}(x)}, \qquad \Theta_{a_\star}(x) \in \arg\max_{\theta\in[0,2\pi)} \big\{\langle x, e_\theta\rangle - a_\star(\theta)\big\}.$$

The function $\Phi_{a_\star}$ is convex, and 1-Lipschitz because $|\Phi_a(x) - \Phi_a(x')| \le \sup_\theta |\langle x-x', e_\theta\rangle| \le \|x-x'\|$. From Rademacher's theorem, $\Phi_{a_\star}$ is differentiable almost everywhere in Lebesgue measure, and also differentiable almost everywhere in P_0. Consequently, $\arg\max_{\theta\in[0,2\pi)}\{\langle x, e_\theta\rangle - a(\theta)\}$ is unique P_0-a.e., and

$$S(x) = \nabla\Phi_{a_\star}(x) = e_{\Theta_\star(x)}, \qquad P_0\text{-a.e.}$$

The corresponding extended-Brenier potential for the reverse transport relation $P_1 \to P_0$ can be taken by the convex conjugate

$$\phi_a = \Phi_a^*,$$

or equivalently, from (24),

$$\phi_a = (a + \iota_{\mathbb{S}^1})^{**},$$

where $a + \iota_{\mathbb{S}^1}$ denotes the $(\mathbb{R}\cup\{+\infty\})$-valued function that takes the value $a(\theta)$ at $e_\theta \in \mathbb{S}^1$ and $+\infty$ outside $\mathbb{S}^1$. Thus ϕ_a is the lower-semicontinuous convex envelope of the boundary data a on the circle.

By Fenchel duality,

$$x_0 \in \partial\phi_{a_\star}(e_\theta) \quad\Longleftrightarrow\quad e_\theta \in \partial\Phi_{a_\star}(x_0).$$

Therefore, the fiber $\partial\phi_{a_\star}(e_\theta)$ is exactly the set of source points x_0 for which e_θ is the target point of OT.

A.2 Subgradient fibers on the circle

This subsection calculates the explicit form of the fiber $\partial\phi_a(e_\theta)$. Assume that a is C^2 and that e_θ is an active boundary point of the convex envelope. From the first-order condition of the maximization of (24), we have $0 = \langle x, e_\theta^\perp\rangle - a'(\theta)$, and hence, every point in the fiber has the form

$$x = a'(\theta)e_\theta^\perp + se_\theta$$

for some scalar $s \in \mathbb{R}$. The global maximality condition is then given by

$$\langle x, e_\theta\rangle - a(\theta) \ge \langle x, e_\eta\rangle - a(\eta) \qquad \text{for all } \eta.$$

Substituting $x = a'(\theta)e_\theta^\perp + se_\theta$, we obtain

$$s\{1-\cos(\eta-\theta)\} \ge a(\theta) - a(\eta) + a'(\theta)\sin(\eta-\theta).$$

Therefore, in full generality, the fiber can be written as

$$\partial\phi_a(e_\theta) = \left\{a'(\theta)e_\theta^\perp + se_\theta : s \ge r_a(\theta)\right\},$$

where

$$r_a(\theta) := \sup_{\eta\neq\theta} \frac{a(\theta) - a(\eta) + a'(\theta)\sin(\eta-\theta)}{1-\cos(\eta-\theta)}.$$

This formula makes the tangent–normal structure transparent. Since $T_{e_\theta}\mathbb{S}^1 = \mathbb{R}e_\theta^\perp$ and $N_{e_\theta}\mathbb{S}^1 = \mathbb{R}e_\theta$, the parallel subspace of the fiber is

$$\mathrm{par}\big(\partial\phi_a(e_\theta)\big) = \mathbb{R}e_\theta = N_{e_\theta}\mathbb{S}^1.$$

Thus, the fiber direction is normal to the circle. On the other hand, an individual OT path is

$$x_0 - e_\theta = a'(\theta)e_\theta^\perp + (s-1)e_\theta,$$

which contains a tangential component $a'(\theta)e_\theta^\perp$ and not normal to the manifold in general.

A.3 Numerical computation of fibers

We describe the numerical procedure used to compute the subgradient of the extended Brenier potential for von Mises distribution. The numerical example below uses $\kappa = 1$ and $\mu = 0$.

Fourier approximation of the dual potential. Since $\mu = 0$, both P_0 and P_1 are symmetric across the first coordinate axis. We therefore use the truncated Fourier approximation

$$a_M(\theta) = \sum_{m=1}^{M} c_m \cos(m\theta).$$

We can omit the constant term by fixing the additive normalization of a.

We discretize the circle by $\theta_j = \frac{2\pi j}{J}$ $(j = 0, \dots, J-1)$, and set $e_j := e_{\theta_j}$. The quadrature weights for P_1 are

$$b_j := \frac{\rho_{\kappa,\mu}(\theta_j)}{\sum_{\ell=0}^{J-1} \rho_{\kappa,\mu}(\theta_\ell)}, \qquad \sum_{j=0}^{J-1} b_j = 1.$$

We independently generate $X_1, \dots, X_N \sim N(0, I_2)$ and approximate $\mathcal{J}$ (25) by

$$\widehat{\mathcal{J}}_{N,J}(c) = \frac{1}{N} \sum_{i=1}^{N} \max_{0 \le j < J} \{\langle X_i, e_j \rangle - a_M(\theta_j)\} + \sum_{j=0}^{J-1} b_j a_M(\theta_j). \tag{26}$$

This is a finite-dimensional convex function of $c = (c_1, \dots, c_M)$, but it is non-smooth because of the maximum.

Soft-max approximation and continuation. To optimize (26) numerically, we approximate it by replacing the maximum with the soft-max:

$$\mathrm{smax}_\tau(z_0, \dots, z_{J-1}) := \tau \log \left(\sum_{j=0}^{J-1} e^{z_j/\tau} \right), \qquad \tau > 0. \tag{27}$$

It satisfies

$$\max_j z_j \le \mathrm{smax}_\tau(z) \le \max_j z_j + \tau \log J.$$

Thus the approximation error converges uniformly to zero as $\tau \downarrow 0$. The resulting smooth objective is

$$\widehat{\mathcal{J}}_{N,J,\tau}(c) = \frac{\tau}{N} \sum_{i=1}^{N} \log \left[\sum_{j=0}^{J-1} \exp \left\{ \frac{\langle X_i, e_j \rangle - a_M(\theta_j)}{\tau} \right\} \right] + \sum_{j=0}^{J-1} b_j a_M(\theta_j). \tag{28}$$

We minimize (28) by a quasi-Newton method such as L-BFGS. To reduce the smoothing bias, we use a continuation scheme $\tau_1 > \tau_2 > \cdots > \tau_L > 0$, using the solution obtained at τ_ℓ as the initial value for the optimization at $\tau_{\ell+1}$.

Global computation of the fiber threshold. For a fixed θ, every point in the affine line containing the fiber has the form

$$x_{\theta,s} := a'_M(\theta) e_\theta^\perp + s e_\theta.$$

Let

$$q_\theta(\eta) := \frac{a_M(\theta) - a_M(\eta) + a'_M(\theta) \sin(\eta - \theta)}{1 - \cos(\eta - \theta)},$$

where $q_\theta(\theta)$ is defined by $-a''_M(\theta)$ at $\eta = \theta$, and define

$$r_{a_M}(\theta) := \max_{\eta \in [0, 2\pi)} q_\theta(\eta).$$

The resulting numerical fiber is given by

$$\partial \phi_a(e_\theta) \approx \left\{ a'_M(\theta) e_\theta^\perp + s e_\theta : s \ge r_{a_M}(\theta) \right\}. \tag{29}$$

Figure 2 is made by the above method for

$$\theta \in \left\{ 0,\ \frac{\pi}{4},\ \frac{\pi}{2},\ \pi,\ \frac{3\pi}{2} \right\}.$$

B Semiderivatives and Proximal Operators

We summarize some basic definitions in convex analysis: semiderivatives and proximal operators used in this paper. See Parikh and Boyd [45] and Rockafellar and Wets [50] for more details.

For a mapping $G : \mathbb{R}^d \to \mathbb{R}^d$, the *semiderivative* of G at x in direction h is defined by

$$DG(x; h) \coloneqq \lim_{\substack{\varepsilon \downarrow 0 \\ h' \to h}} \frac{G(x + \varepsilon h') - G(x)}{\varepsilon}, \tag{30}$$

whenever the limit exists. This is a generalization of the directional derivative with a one-sided limit.

For a convex function $F : \mathbb{R}^d \to \overline{\mathbb{R}}$ and $\lambda > 0$, the *proximal operator* of F with parameter λ is defined by

$$\operatorname{prox}_{\lambda F}(y) = \arg\min_x \left\{ F(x) + \frac{1}{2\lambda}\|x - y\|^2 \right\}. \tag{31}$$

The proximal operator is often used for optimization steps. A key identity from this viewpoint is, if F is proper, lsc, convex, and differentiable,

$$x = y - \lambda \nabla F(x) \quad \iff \quad x = \operatorname{prox}_{\lambda F}(y).$$

Thus, $x_{t+1} = \operatorname{prox}_{\lambda F}(x_t)$ can be regarded as the implicit Euler step for the gradient flow $\dot{x} = -\nabla F(x)$.

The *Moreau envelope* (or *Moreau–Yosida regularization*) of F with parameter λ is defined by

$$M_{\lambda F}(y) = \min_x \left\{ F(x) + \frac{1}{2\lambda}\|x - y\|^2 \right\}. \tag{32}$$

If F is convex, $x \mapsto F(x) + \frac{1}{2\lambda}\|x - y\|^2$ is strictly convex, so the proximal operator $\operatorname{prox}_{\lambda F}(y)$ is the unique minimizer to give $M_{\lambda F}(y)$.

The proximal operator of a convex function satisfies the following firm nonexpansiveness.

Proposition B.1 (Firm nonexpansiveness of the proximal operator)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, lower semicontinuous, and convex. Then, for $\lambda > 0$, $\operatorname{prox}_{\lambda f}$ is firmly nonexpansive: for all $x, y \in \mathbb{R}^d$,*

$$\|\operatorname{prox}_{\lambda f}(x) - \operatorname{prox}_{\lambda f}(y)\|^2 \le \langle \operatorname{prox}_{\lambda f}(x) - \operatorname{prox}_{\lambda f}(y), x - y \rangle.$$

In particular,

$$\|\operatorname{prox}_{\lambda f}(x) - \operatorname{prox}_{\lambda f}(y)\| \le \|x - y\|.$$

Proof. Let $P_\lambda(x) := \operatorname{prox}_{\lambda f}(x)$ for simplicity. Fix $x, y \in \mathbb{R}^d$ and write $p = P_\lambda(x)$ and $q = P_\lambda(y)$. The proximal optimality conditions give

$$\frac{x - p}{\lambda} \in \partial f(p), \qquad \frac{y - q}{\lambda} \in \partial f(q).$$

By the monotonicity of the convex subdifferential,

$$0 \le \left\langle p - q, \frac{x - p}{\lambda} - \frac{y - q}{\lambda} \right\rangle = \frac{1}{\lambda}\left(\langle p - q, x - y \rangle - \|p - q\|^2 \right).$$

This proves firm nonexpansiveness. By the Cauchy–Schwarz inequality,

$$\|p - q\|^2 \le \langle p - q, x - y \rangle \le \|p - q\|\,\|x - y\|.$$

If $p \neq q$, division by $\|p - q\|$ yields nonexpansiveness; if $p = q$, the conclusion is immediate. □

We recall two standard regularity notions from variational analysis (see, e.g., Rockafellar and Wets [50, Chs. 1 & 13]). Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper lower-semicontinuous (lsc) function.

Definition B.2 (Prox-boundedness)**.** The function f is *prox-bounded* if there exists $\lambda > 0$ such that $M_{\lambda f}(x) > -\infty$ for at least one $x \in \mathbb{R}^d$ (equivalently, $M_{\lambda f}$ is not identically $-\infty$). The *prox-threshold* of f is defined as

$$\bar{\lambda}_f := \sup\Big\{\lambda > 0 : \exists x \text{ s.t. } M_{\lambda f}(x) > -\infty\Big\} \in (0, +\infty].$$

Intuitively, prox-boundedness means that f does not decrease faster than a quadratic, so the proximal subproblem is well-posed for sufficiently small step sizes $\lambda < \bar{\lambda}_f$.

Definition B.3 (Prox-regularity)**.** Let $\bar{x} \in \operatorname{dom} f$ and $\bar{v} \in \partial f(\bar{x})$. We say that f is *prox-regular at* $\bar{x}$ *for* $\bar{v}$ if there exist $\varepsilon > 0$ and $\rho \geq 0$ such that

$$f(x') \;\geq\; f(x) + \langle v, x' - x\rangle - \frac{\rho}{2}\|x' - x\|^2 \tag{33}$$

holds for all $x, x' \in B(\bar{x}, \varepsilon)$ and all $v \in \partial f(x)$ satisfying $\|v - \bar{v}\| < \varepsilon$ and $|f(x) - f(\bar{x})| < \varepsilon$.

Inequality (33) is a "convexity up to a quadratic" condition: it says that, locally and along nearby subgradients, the function admits a quadratic supporting model. Prox-regularity is a standard assumption guaranteeing good local behavior of the proximal mapping (e.g. local single-valuedness and Lipschitz properties), which is used in second-order variational characterizations of $\operatorname{prox}_{\lambda f}$ and its directional derivatives.

Remark B.4 (Convex case). If f is proper lsc and *convex*, then (33) holds globally with $\rho = 0$ by the subgradient inequality, hence f is prox-regular everywhere. Moreover, convexity implies the existence of an affine minorant (e.g. if $v \in \partial f(\bar{x})$, then $f(y) \geq f(\bar{x}) + \langle v, y - \bar{x}\rangle$ for all y), which in turn implies that $M_{\lambda f}(x) > -\infty$ for every $\lambda > 0$ and $x \in \mathbb{R}^d$ (i.e. $\bar{\lambda}_f = +\infty$). Thus, for a proper, lsc, convex function f, the prox-boundedness and prox-regularity are straightforward.

C Details of the Lyapunov analysis in Section 3

This appendix specializes in the existing results on active-manifold sensitivity to the convex potential used in Section 3.2, verifies the affine flow formula, and retains the relative-interior criterion for nondegeneracy.

C.1 Partly smoothness

We show the definition of partly smoothness. See [25] for details. Assume that M is a C^p submanifold embedded in $\mathbb{R}^d$, and $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is a function.

Definition C.1. Let x be a point of M. A function f is *C^p-partly smooth* at x relative to M if the following conditions hold:

(i) (smoothness) $f|_M$ is a C^p function in a neighborhood of x;
(ii) (regularity) f is regular at all points $x \in M$ in a neighborhood of x, with $\partial f(x) \neq \emptyset$;
(iii) (sharpness) the affine span of $\partial f(x)$ is a translate of the normal space $N_x M$;
(iv) (sub-continuity) $\partial f(x)$ restricted to M is continuous at x.

C.2 Proximal sensitivity in the convex setting

We prove Proposition 3.1 by applying the identification theorem of Hare and Lewis [25, Theorem 5.3] and differentiating the manifold-restricted proximal problem, as in Daniilidis et al. [19, Theorems 28–29]. The latter paper states its general prox-regular result for sufficiently small proximal parameters. The argument below explains why every fixed $\lambda > 0$ is allowed in the present convex setting.

Proof of Proposition 3.1. **Identification for each fixed parameter.** Fix $\lambda > 0$ and recall $\Xi_\lambda = \operatorname{prox}_{\lambda f}$. We first show that for $x_1 \in M$ there is an open neighborhood U_λ of x_1 in $\mathbb{R}^d$ such that $\Xi_\lambda(U_\lambda) \subset M$.

Suppose $y_k \to x_1$ and set $z_k = \Xi_\lambda(y_k)$. Nonexpansiveness of Ξ_λ (Prop. B.1) and the fact $\Xi_\lambda(x_1) = x_1$ yield $\|z_k - x_1\| \leq \|y_k - x_1\|$, hence $z_k \to x_1$. By definition of proximal operator, we have

$f(z_k) + \frac{1}{2\lambda}\|y_k - z_k\|^2 \le f(x_1) + \frac{1}{2\lambda}\|y_k - x_1\|^2$. Also, $f(x_1) \le f(z_k)$ from $0 \in \partial f(x_1)$. Accordingly,

$$f(x_1) \le f(z_k) \le f(x_1) + \frac{1}{2\lambda}\|y_k - x_1\|^2,$$

so that $f(z_k) \to f(x_1)$. Proximal optimality gives $0 \in \partial f(y_k) + \frac{1}{\lambda}(z_k - y_k)$, and thus $w_k := \frac{y_k - z_k}{\lambda} \in \partial f(z_k)$ satisfies $w_k \to 0$. This implies

$$\operatorname{dist}\big(0, \partial f(z_k)\big) \le \operatorname{dist}(0, w_k) \to 0 \quad (k \to \infty).$$

A proper lsc convex function is prox-regular. Thus Hare and Lewis [25, Theorem 5.3], with (PS)–(ND), implies $z_k \in \mathcal{M}$ for all sufficiently large k. Since this applies to every sequence $y_k \to x_1$, there is a neighborhood U_λ of x_1 such that $\Xi_\lambda(U_\lambda) \subset \mathcal{M}$.

Smoothness and the Jacobian. Let $V \subset \mathbb{R}^m$ be an open neighborhood of 0 and $\chi : V \to \mathcal{M}$ be a C^2 local chart with $\chi(0) = x_1$ and $D\chi(0) = R$, where the columns of R form an orthonormal basis of $T = T_{x_1}\mathcal{M}$, i.e., $R^*R = I_m$. Write $g = f \circ \chi$. Since $0 \in \partial f(x_1)$ and $f|_M$ is C^2, g takes a minimum at 0 and

$$\nabla g(0) = 0, \qquad \nabla^2 g(0) = R^* H R \succeq 0$$

hold. For y near x_1, the identified proximal minimizer is a critical point of

$$s \longmapsto \lambda g(s) + \tfrac{1}{2}\|\chi(s) - y\|^2.$$

Its stationarity equation is

$$G(s, y) := \lambda \nabla g(s) + D\chi(s)^*(\chi(s) - y) = 0.$$

At $(s, y) = (0, x_1)$, using the assumption $R^*R = I_m$, we have

$$D_s G(0, x_1) = I_m + \lambda R^* H R \succ 0, \qquad D_y G(0, x_1) = -R^*.$$

The implicit function theorem supplies a unique C^1 local critical branch $s = s(y)$. By identification and continuity of Ξ_λ, after shrinking the neighborhood of x_1, the coordinate $\widehat{s}(y) := \chi^{-1}(\Xi_\lambda(y))$ is well-defined and satisfies $\widehat{s}(y) \to 0$ as $y \to x_1$. Since $\Xi_\lambda(y)$ minimizes the original proximal objective, $\widehat{s}(y)$ minimizes the manifold-restricted objective and hence satisfies $G(\widehat{s}(y), y) = 0$. Local uniqueness in the implicit function theorem then gives $\widehat{s}(y) = s(y)$, and hence $\Xi_\lambda(y) = \chi(s(y))$. Therefore

$$D\Xi_\lambda(x_1) = R(I_m + \lambda R^* H R)^{-1} R^* = (I_T + \lambda H)^{-1} \Pi_T.$$

Since $H \succeq 0$, we have $\|(I_T + \lambda H)^{-1}\|_{\mathrm{op}} \le 1$, and consequently,

$$\|D\Xi_\lambda(x_1) - \Pi_T\|_{\mathrm{op}} = \|\lambda H (I_T + \lambda H)^{-1}\|_{\mathrm{op}} \le \lambda \|H\|_{\mathrm{op}}.$$

This proves the proposition. The same argument, allowing $\lambda_k \to \lambda_* > 0$ and using λ as an additional parameter in G, gives joint local C^1 dependence on (λ, y) around (λ_*, x_1). □

Remark C.2. For finite-valued convex functions satisfying (ND) and admitting a fast track, Mifflin and Sagastizábal [41, Theorem 5.2] establish local identification by the proximal map. Together with the equivalence between partial smoothness and the existence of a fast track in this setting [42, Corollary 2.10], this yields $\Xi_\lambda(U_\lambda) \subset \mathcal{M}$ under (PS)–(ND) for every fixed $\lambda > 0$. The proof above applies the identification theorem for prox-regular functions [25, Theorem 5.3] to the possibly extended-real-valued convex function f considered here.

C.3 Nondegeneracy: a disintegration criterion for relative-interior subgradients

This subsection records a sufficient condition to ensure the assumption (ND). We discuss a convenient measure-theoretic criterion ensuring that, for an optimal coupling concentrated on the graph of a subdifferential, the conditional points lie in the *relative interior* of the corresponding fiber almost surely. The statement is purely measure-theoretic and can be applied to OT couplings in the non-smooth setting.

Let $\phi : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, lower semicontinuous, and convex. Let π be a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$, with first marginal P_1, such that

$$\pi\big(\{(x_1, x_0) : x_0 \in \partial\phi(x_1)\}\big) = 1. \tag{34}$$

By the disintegration theorem [2, Theorem 5.3.1], write

$$\pi(dx_1, dx_0) = P_1(dx_1)\, \pi_{x_1}(dx_0). \tag{35}$$

For P_1-almost every x_1, set

$$C(x_1) := \partial\phi(x_1), \qquad L(x_1) := \operatorname{aff} C(x_1), \qquad k(x_1) := \dim L(x_1).$$

The graph-support condition implies $\pi_{x_1}(C(x_1)) = 1$ for P_1-almost every x_1.

For $E := \{(x_1, x_0) : x_0 \in \operatorname{ri} C(x_1)\}$, disintegration gives

$$\pi(E) = \int \pi_{x_1}\big(\operatorname{ri} C(x_1)\big)\, P_1(dx_1).$$

Thus, (ND) holds π-almost surely if and only if $\pi_{x_1}(\operatorname{ri} C(x_1)) = 1$ for P_1-almost every x_1. The following sufficient condition is an immediate consequence of this identity and the nullity of the relative boundary of a convex set.

Lemma C.3. *Assume that, for P_1-almost every x_1,*

$$\pi_{x_1} \ll \mathcal{H}^{k(x_1)} \restriction_{L(x_1)}. \tag{36}$$

Then $x_0 \in \operatorname{ri} \partial\phi(x_1)$ for π-almost every (x_1, x_0).

Proof. The relative boundary of a nonempty closed convex set C has zero k-dimensional measure in $\operatorname{aff} C$, where $k = \dim \operatorname{aff} C$. For bounded sets this follows from the Jordan measurability theorem of Szabó [58, p. 84]; the general case follows by localization in bounded boxes within the affine hull. If $k = 0$, the relative boundary is empty. Hence (36) and $\pi_{x_1}(C(x_1)) = 1$ imply $\pi_{x_1}(\operatorname{ri} C(x_1)) = 1$ almost everywhere. The preceding disintegration identity concludes the proof. □

Condition (36) is a sufficient condition on the conditional laws: it ensures that these laws do not charge the relative boundaries of the subgradient fibers.

D Convergence of minibatch OT–CFM in Section 4

This appendix gives the detailed convergence analysis for Section 4. The corresponding detailed result to Theorem 4.1 is Corollary D.48.

D.1 Preliminaries

For the theoretical analysis of the minibatch OT-CFM in this paper, we consider idealized neural network learning in combination with the *expected batch OT plan* introduced in [8], where independent (fresh) minibatches of size B taken from the population are repeatedly provided L times, and the network minimizes the squared error for the teaching vector field along the paths constructed by the minibatch OT.

The analysis has the following structure:

Fresh batches of size $B \implies$ Expected batch plan $\overline{\pi}_B \implies$ Expected batch OT vector field v_B,

$B \to \infty \implies v_B \to u^\star$ population vector field,

$L \to \infty$ fresh training batches $\implies \widehat{u}_{B,L,P} \to v_B$ within the neural class,

Spatial Lipschitz/One-sided Lipschitz control $\implies$ Flow convergence.

Assuming that the distribution of data is supported by a low-dimensional manifold, we will prove the convergence of the vector field of the minibatch OT–CFM to that of the population OT–CFM with a convergence rate that depends on manifold dimensionality. The purpose is not to provide a tight statistical generalization bound for the OT–CFM, and we do not claim that the obtained bound is optimal.

D.1.1 Optimal Transport Conditional Flow Matching

Let P_0 and P_1 be in the class $\mathcal{P}_2(\mathbb{R}^d)$, the probability distributions on $\mathbb{R}^d$ with a finite second moment. The *optimal transport (OT) plan* $\pi^\star$ between P_0 and P_1 with quadratic cost is defined by

$$\pi^\star := \arg\min_{\pi \in \Pi(P_0, P_1)} \int \frac{1}{2}\|x_0 - x_1\|^2 d\pi(x_0, x_1), \tag{37}$$

where $\Pi(P_0, P_1) := \{\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \mid \mathrm{Proj}_{0,\#} = P_0, \mathrm{Proj}_{1,\#} = P_1\}$.

In *Conditional Flow Matching* [37, 1, 38], the vector field is constructed by using random paths from $x_0 \sim P_0$ to $x_1 \sim P_1$:

$$X_t := (1-t)X_0 + tX_1. \tag{38}$$

The joint distribution π of (X_0, X_1) is arbitrary for the general class of CFM, while this appendix considers the OT coupling: $(X_0, X_1) \sim \pi^\star$. This appendix focuses on the affine paths (38), but an extension to an arbitrary parameterization $\alpha_t x_0 + \beta_t x_1$ is possible with the necessary changes. Along the path X_t, the velocity vector

$$V(X_0, X_1) := X_1 - X_0$$

is placed at X_t. This creates a random vector field, and a regression model, such as a neural network, learns a vector field using the least squares criterion:

$$\widehat{u}(x,t) = \arg\min_u \mathbb{E}_\pi[\|u(x,t) - V(X_0, X_1)\|^2 \mid X_t = x].$$

The least square solution is then given by the mean of the random vector field:

$$\widehat{u}(x,t) = \mathbb{E}_\pi[V(X_0, X_1) \mid X_t = x]. \tag{39}$$

D.1.2 Minibatch OT-CFM and expected batch OT plan

OT-based CFM methods have been popularly used to make the regression learning of the vector field easier [59, 48]. In practice, the population OT is not feasible to obtain. Alternatively, a resampled batch of size B is taken from the training data and the source distribution P_0, and some method of optimal transport, such as the Hungarian algorithm, is used to create a coupling $(X_{0,i}, X_{1,i}^{match})_{i=1}^B$. The endpoints of the path (38) are then determined by this coupling. The velocities along sampled points on the minibatch-OT paths are provided to neural networks as the minibatch training data for stochastic gradient descent (SGD).

For ease of theoretical analysis, we consider an idealized procedure for this minibath-OT coupling and neural net learning, using the expected batch OT plan. Consider independent batches of size B

$$X_{0,1}, \ldots, X_{0,B} \overset{\text{iid}}{\sim} P_0, \qquad X_{1,1}, \ldots, X_{1,B} \overset{\text{iid}}{\sim} P_1,$$

and define

$$\widehat{P}_{0,B} := \frac{1}{B}\sum_{i=1}^B \delta_{X_{0,i}}, \qquad \widehat{P}_{1,B} := \frac{1}{B}\sum_{j=1}^B \delta_{X_{1,j}}.$$

Fix a measurable solver that selects an empirical optimal coupling

$$\widehat{\pi}_B \in \operatorname*{argmin}_{\pi \in \Pi(\widehat{P}_{0,B}, \widehat{P}_{1,B})} \int \frac{1}{2}\|x_0 - x_1\|^2 \, \mathrm{d}\pi.$$

We assume to select an optimal permutation plan. Then, the neural network $u(x,t)$ minimizes

$$\mathbb{E}_{(X_0,X_1)\sim\widehat{\pi}_B, t}[\|u(X_t,t) - V(X_0,X_1)\|^2 | X_t],$$

where t is taken independently.

Assume that we generate the above random batches and couplings L times independently from the populations P_0 and P_1, and the corresponding training data are provided for neural networks. Then, the least squares learning ideally converges as $L \to \infty$ to the average of the vector field over the random batches:

$$\mathbb{E}_B \mathbb{E}_{(X_0,X_1)\sim\widehat{\pi}_B, t}[\|u(X_t,t) - V(X_0,X_1)\|^2 | X_t], \tag{40}$$

where $\mathbb{E}_B$ is the average over the random choice of the batches. Following Boïté et al. [8], we introduce the average of the minibatch OT plan over the random batches.

Definition D.1. The *expected batch OT plan* $\overline{\pi}_B$ is the unique probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ satisfying

$$\int g \, \mathrm{d}\overline{\pi}_B = \mathbb{E}_B \left[\int g \, \mathrm{d}\widehat{\pi}_B \right] \tag{41}$$

for every bounded Borel function g, where the expectation is taken over the random choice of batches of size B from the populations P_0 and P_1.

Equivalently, sample two batches of size B, solve empirical OT, and then choose one matched pair according to the empirical optimal plan. The law of the selected pair is equal to $\overline{\pi}_B$.

The following properties of the expected batch OT plan are established in [8].

Proposition D.2 (Boïté et al. [8], Section A.1.3 and Proposition 3.2)**.** *For every $B \geq 1$,*

$$\overline{\pi}_B \in \Pi(P_0, P_1).$$

Moreover,

$$\int \frac{1}{2}\|x_0 - x_1\|^2 \, \overline{\pi}_B(\, \mathrm{d}x_0, \, \mathrm{d}x_1) = \frac{1}{2}\mathbb{E}_B\left[W_2^2(\widehat{P}_{0,B}, \widehat{P}_{1,B})\right]. \tag{42}$$

Define the signed expected-batch cost bias Δ_B and the plan W_2-distance d_B by

$$\begin{aligned} \Delta_B &:= \int \frac{1}{2}\|x_0 - x_1\|^2 \, \mathrm{d}\overline{\pi}_B - \int \frac{1}{2}\|x_0 - x_1\|^2 \, \mathrm{d}\pi^\star \\ &= \frac{1}{2}\left\{\mathbb{E}_B W_2^2(\widehat{P}_{0,B}, \widehat{P}_{1,B}) - W_2^2(P_0, P_1)\right\} \geq 0, \end{aligned} \tag{43}$$

$$d_B := W_2(\overline{\pi}_B, \pi^\star). \tag{44}$$

The following proposition shows the convergence of the expected batch OT plan to the population OT as $B \to \infty$.

Proposition D.3. *Assume that $P_0, P_1 \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > 2$. Then,*

$$\Delta_B \longrightarrow 0 \quad \textit{and} \quad d_B \longrightarrow 0 \quad \textit{as } B \to \infty.$$

Proof. The convergence $\Delta_B \to 0$ follows from Boïté et al. [8, Proposition 3.3(1)], while $\overline{\pi}_B \rightharpoonup \pi^\star$ follows from Boïté et al. [8, Proposition 3.4]. Since $\overline{\pi}_B$ and $\pi^\star$ have the same fixed marginals, their second moments on $\mathbb{R}^d \times \mathbb{R}^d$ coincide. The standard characterization of Wasserstein convergence therefore upgrades the weak convergence to $W_2(\pi_B, \pi^\star) \to 0$ (see Villani [61, Theorem 6.9], for example). □

D.1.3 Interpolation laws and Brenier maps

For the affine schedule, define for $t \in [0, 1]$

$$F_t(x_0, x_1) := (1 - t)x_0 + tx_1, \tag{45}$$

and let

$$\mu_t^\star := (F_t)_\# \pi^\star, \qquad \nu_{B,t} := (F_t)_\# \overline{\pi}_B. \tag{46}$$

The vector field of the expected batch OT plan is

$$v_B(t, x) := \mathbb{E}_{(X_0, X_1) \sim \overline{\pi}_B}\left[X_1 - X_0 \mid F_t(X_0, X_1) = x\right], \tag{47}$$

defined $\nu_{B,t}$-almost everywhere.

Let $P_0, P_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and let

$$K_0 := \operatorname{supp} P_0, \qquad K_1 := \operatorname{supp} P_1.$$

We make the following four assumptions throughout this section.

Assumption D.4. The source distribution P_0 is absolutely continuous with respect to Lebesgue measure, and K_0 and K_1 are compact.

Assumption D.5. There exist an open set Ω_0 that contains K_0 and a proper lower-semicontinuous convex function $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ such that $S := \nabla\Psi$ is continuous on Ω_0 and

$$S_{\#}P_0 = P_1, \qquad \pi^\star = (\mathrm{Id}, S)_{\#}P_0. \tag{48}$$

Assumption D.6. There is a constant $L_S > 0$ such that the following Bregman co-coercivity

$$D_\Psi(x, z) := \Psi(x) - \Psi(z) - \langle S(z), x - z\rangle \geq \frac{1}{2L_S}\left\|S(x) - S(z)\right\|^2 \tag{49}$$

holds for all $x, z \in K_0$.

Assumption D.7. Fix $0 < c < 1$ and

$$I = [0, c]. \tag{50}$$

All vector field, regression, and flow bounds concern early stopping at time $t = c$.

Remark D.8. The assumption D.4 does not cover the Gaussian distribution for the source. However, unlike diffusion models, a Gaussian distribution for the source is not necessary for flow matching, and a truncated Gaussian is often sufficient in practice. This simplifies theoretical arguments significantly.

Remark D.9. The standard theory guaranties the existence of Brenier map defined on $\mathbb{R}^d$, if P_0 is absolutely continuous with respect to the Lebesgue measure. In the assumption D.5, we assume the continuous Brenier map S exists only on Ω_0 that contains K_0, whereas the Brenier potential Ψ, which may take ∞, exists on $\mathbb{R}^d$.

Remark D.10 (A checkable sufficient condition of D.6). A sufficient condition for (49) is that Ψ is the continuously differentiable functions on a convex open neighborhood Ω of compact K_0 such that its derivative $S = \nabla\Psi$ is Lipschitz with

$$\mathrm{Lip}(S; \Omega) \leq L_\Omega.$$

In fact, there is $\rho > 0$ such that $K_0 + B(0, \rho) \subset \Omega$, where $B(0, \rho)$ is the closed ball with radius ρ. Set $L_S := L_\Omega \max\{1, \mathrm{Diam}(K_0)/\rho\}$. For any fixed $x, z \in K_0$, define

$$h_z(u) := \Psi(u) - \Psi(z) - \langle S(z), u - z\rangle$$

for $u \in \Omega$. Then, h_z is convex and differentiable on Ω, $h_z(u) \geq 0$ by the convexity of Ψ, and $\nabla h_z(u) = S(u) - S(z)$. If we take $y := x - (S(x) - S(z))/L_S$, then $y \in \Omega$, because $\|y - x\| \leq \rho$ can be easily confirmed. Since the line segment $[x, y]$ is contained in Ω, we have

$$0 \leq h_z(y) \leq h_z(x) + \langle \nabla h_z(x), y - x\rangle + \frac{L_\Omega}{2}\|y - x\|^2.$$

Simple calculation yields

$$0 \leq \Psi(x) - \Psi(z) - \langle S(z), x - z\rangle - \frac{1}{2L_S}\|S(x) - S(z)\|^2,$$

which proves the claim.

Remark D.11. We state the Bregman co-coercivity on the source support K_0 only because it is the property used below and is required only for pairs of source points. Note also that, if we impose the Lipschitz property of Ψ on the whole $\mathbb{R}^d$, the range of the map $S = \nabla\Psi$ must be convex, which excludes general low dimensional manifold structures for K_1. We, therefore, consider a restricted compact domain K_0, and the co-coercivity or Lipschitz property on the restricted domain.

Remark D.12 (Scope of the assumption D.5). Assumption D.5 may not hold in general. For example, for the rotationally symmetric quadratic OT from the uniform distribution on $K_0 := \{x \in \mathbb{R}^d \mid 1/2 \leq \|x\| \leq 1\}$ to the uniform distribution on the unit circle $\mathbb{S}^1$, the forward Brenier map is given by

$$S(x) = \frac{x}{\|x\|}.$$

The potential function Ψ must take the form $\Psi(x) = \|x\|$ to satisfy $\nabla\Psi(x) = S(x)$. So, if we take $\Omega_0 := \{x \in \mathbb{R}^d \mid 1/4 < \|x\| < 2\}$, Assumption D.5 is satisfied. However, if we consider the uniform distribution on the unit disk for P_0, the Bernier map S cannot be continuous at 0.

Proposition D.13. *Under the assumption D.6,*

$$\|S(x) - S(z)\| \le L_S \|x - z\|, \qquad \textit{for any } x, z \in K_0. \tag{51}$$

Proof. Adding (49) with (x, z) and (z, x) yields

$$\langle S(x) - S(z), x - z\rangle \ge \frac{1}{L_S}\|S(x) - S(z)\|^2.$$

Combining Cauchy-Schwartz $\langle S(x) - S(z), x - z\rangle \le \|S(x) - S(z)\|\|x - z\|$ shows the Lipschitz property. □

D.2 OT–CFM vector field and Expected-batch OT

D.2.1 Lipschitz regularity of the population OT vector field

Consider the vector field $u^\star$ given by the population OT plan $\pi^\star$. Introduce $\phi := \Psi^*$, which serves as the extended Brenier potential such that $x_0 \in \partial\phi(x_1)$ whenever $(x_0, x_1) \sim \pi^\star$. As shown in Sec. 2, we have the expression of the population OT vector field with the proximal operator of ϕ:

$$u^\star(t, x) := \frac{Q_t(x) - x}{1 - t}, \tag{52}$$

where $Q_t(x) = \mathrm{prox}_{\lambda_t \phi}\left(\frac{x}{t}\right)$ with $\lambda_t = (1 - t)/t$.

For $x_0 \in K_0$, define the exact interpolation as a function of x_0

$$Z_t(x_0) := (1 - t)x_0 + tS(x_0).$$

Since $x_0 \in \partial\phi(S(x_0))$, the proximal optimality condition gives

$$Q_t(Z_t(x_0)) = S(x_0), \qquad u^\star(t, Z_t(x_0)) = S(x_0) - x_0. \tag{53}$$

We show Lipschitz regularity of the proximal operator and vector field used in the sequel. For a vector field $f(t, x)$, define the *one-sided Lipschitz (OSL) coefficient* by

$$\rho_f(t) := \sup_{x \ne y} \frac{\langle f(t, x) - f(t, y), x - y\rangle}{\|x - y\|^2}. \tag{54}$$

Proposition D.14. *For every $t \in (0, 1)$,*

$$\mathrm{Lip}(Q_t) \le \frac{1}{t}, \tag{55}$$

$$\mathrm{Lip}_x\, u^\star(t, \cdot) \le \max\left\{\frac{1}{t}, \frac{1}{1 - t}\right\} =: \ell_t, \tag{56}$$

$$\rho_{u^\star}(t) \le \frac{1}{t}. \tag{57}$$

Proof. Write: The firm nonexpansiveness of the proximal operator (B.1) yields

$$\|Q_t(x) - Q_t(y)\|^2 \le \frac{1}{t}\langle Q_t(x) - Q_t(y), x - y\rangle. \tag{58}$$

Combined with Cauchy-Schwartz, this proves (55). It follows from (52) that

$$\begin{aligned}(1 - t)^2 \|u^\star(t, x) - u^\star(t, y)\|^2 &= \|Q_t(x) - Q_t(y) - (x - y)\|^2 \\ &= \|x - y\|^2 + \|Q_t(x) - Q_t(y)\|^2 - 2\langle Q_t(x) - Q_t(y), x - y\rangle \\ &\le \|x - y\|^2 + (1 - 2t)\|Q_t(x) - Q_t(y)\|^2,\end{aligned}$$

where the last inequality uses (58). If $t \ge 1/2$, the RHS is at most $\|x - y\|^2$. If $t \le 1/2$, use $\|Q_t(x) - Q_t(y)\| \le t^{-1}\|x - y\|$ to obtain the RHS $\le \frac{(1-t)^2}{t^2}\|x - y\|^2$. This proves (56). Further, using (55),

$$\langle u^\star(t, x) - u^\star(t, y), x - y\rangle = \frac{\langle Q_t(x) - Q_t(y), x - y\rangle - \|x - y\|^2}{1 - t} \le \frac{t^{-1} - 1}{1 - t}\|x - y\|^2 = \frac{1}{t}\|x - y\|^2,$$

which is (57). □

D.2.2 Expected batch regression risk

For a measurable field $f : I \times \mathbb{R}^d \to \mathbb{R}^d$, define the norms by

$$\|f\|_{B,I}^2 := \int_I \int \|f(t,x)\|^2 \ \nu_{B,t}(\,\mathrm{d}x)\ \mathrm{d}t, \tag{59}$$

$$\|f\|_{\star,I}^2 := \int_I \int \|f(t,x)\|^2 \ \mu_t^\star(\,\mathrm{d}x)\ \mathrm{d}t. \tag{60}$$

The *expected-batch regression risk* is defined by

$$\mathcal{R}_B(f) := \int_I \mathbb{E}_{\overline{\pi}_B}\left[\|f(t, F_t(X_0,X_1)) - V(X_0,X_1)\|^2\right] \mathrm{d}t. \tag{61}$$

Lemma D.15. *Any measurable vector field f with finite risk satisfies*

$$\mathcal{R}_B(f) - \mathcal{R}_B(v_B) = \|f - v_B\|_{B,I}^2 . \tag{62}$$

Proof. For each t, write $X_t = F_t(X_0,X_1)$ and $Y = V(X_0,X_1) = X_1 - X_0$. Since $v_B(t,X) = \mathbb{E}_{\overline{\pi}_B}[Y \mid X_t = X]$, the orthogonal projection property of the conditional expectation gives

$$\mathbb{E}_{\overline{\pi}_B} \|f(t,X) - Y\|^2 = \mathbb{E}_{\overline{\pi}_B} \|f(t,X) - v_B(t,X)\|^2 + \mathbb{E}_{\overline{\pi}_B} \|v_B(t,X) - Y\|^2 .$$

Integration over $t \in I$ proves the claim. □

D.3 Bias of the expected-batch OT plan

The difference of expected batch OT plan from the population OT causes a bias of the expected batch OT vector field. We quantify this bias below.

D.3.1 Fenchel slack and cost gap

Define the Fenchel slack associated with $\phi = \Psi^*$ by

$$G_\phi(x_1,x_0) := \phi(x_1) + \Psi(x_0) - \langle x_0, x_1\rangle \geq 0. \tag{63}$$

The equality $G_\phi(x_1,x_0) = 0$ holds whenever $x_0 \in \partial\phi(x_1)$. The following fact is a generalization of the result in Boïté et al. [8, Appendix A.5.3], where they express the excess cost by the expected dual slack for the semidiscrete OT.

Lemma D.16. *For every $\pi \in \Pi(P_0,P_1)$,*

$$\int G_\phi \ \mathrm{d}\pi = \int \frac{1}{2}\|x-y\|^2 \ \mathrm{d}\pi - \int \frac{1}{2}\|x-y\|^2 \ \mathrm{d}\pi^\star. \tag{64}$$

In particular,

$$\int G_\phi \ \mathrm{d}\overline{\pi}_B = \Delta_B. \tag{65}$$

Proof. Because π and $\pi^\star$ have the same marginals,

$$\int \{\phi(x_1) + \Psi(x_0)\} \ \mathrm{d}\pi = \int \{\phi(x_1) + \Psi(x_0)\} \ \mathrm{d}\pi^\star.$$

The latter equals $\int \langle x_0, x_1\rangle \ \mathrm{d}\pi^\star$ by Fenchel equality on the population OT graph. Hence

$$\int G_\phi \ \mathrm{d}\pi = \int \{\phi(x_1) + \Psi(x_0)\} d\pi^\star - \int \langle x_0, x_1\rangle \ \mathrm{d}\pi = \int \langle x_0, x_1\rangle \ \mathrm{d}\pi^\star - \int \langle x_0, x_1\rangle \ \mathrm{d}\pi.$$

The marginal terms in the cost $\frac{1}{2}\|x_0\|^2 + \frac{1}{2}\|x_1\|^2 - \langle x_0, x_1\rangle$ cancel between π and $\pi^\star$. We thus obtain (64). (65) is given by the definition of Δ_B. □

D.3.2 Bias of vector field

The following argument compares an expected-batch interpolation with the population OT interpolation issued from the same source point.

Lemma D.17. *Let $x_0 \in K_0$ and $x_1 \in K_1$ be arbitrary points, and $t \in I$. Set*

$$x_t^B := F_t(x_0, x_1), \qquad x_t^\star := F_t(x_0, S(x_0)) = Z_t(x_0).$$

Then

$$\left\|V(x_0, x_1) - u^\star(t, x_t^B)\right\| \le \left(1 + t\ell_t\right) \|x_1 - S(x_0)\|, \tag{66}$$

where ℓ_t is defined in (56)*.*

Proof. Note first that (x_0, x_1) is an arbitrary pair, while $(x_0, S(x_0))$ is an OT pair. By the exact-path identity (53), $u^\star(t, x_t^\star) = S(x_0) - x_0$ and $V(x_0, x_1) = x_1 - x_0$. Additionally, we have

$$x_t^B - x_t^\star = (1-t)x_0 + tx_1 - \{(1-t)x_0 + tS(x_0)\} = t\{x_1 - S(x_0)\}.$$

Therefore, using Proposition D.14,

$$\begin{aligned}\left\|V(x_0, x_1) - u^\star(t, x_t^B)\right\| &\le \|x_1 - S(x_0)\| + \left\|u^\star(t, x_t^\star) - u^\star(t, x_t^B)\right\| \\ &\le \|x_1 - S(x_0)\| + \ell_t \left\|x_t^\star - x_t^B\right\| \le \left(1 + t\ell_t\right) \|x_1 - S(x_0)\|.\end{aligned}$$

□

Remark D.18. The factor t in the (66) cancels the possible t^{-1} growth of the Lipschitz bound ℓ_t near the initial time.

Define

$$\kappa_I^{\mathrm{cs}} := \int_0^c \left(1 + t\ell_t\right)^2 \mathrm{d}t = \begin{cases} 4c, & 0 < c \le \frac{1}{2}, \\ (1-c)^{-1}, & \frac{1}{2} < c < 1. \end{cases} \tag{67}$$

Also set

$$\varepsilon_B^2 := \int \|x_1 - S(x_0)\|^2 \, \overline{\pi}_B(\,\mathrm{d}x_0,\, \mathrm{d}x_1). \tag{68}$$

The following theorem expresses an upper bound for the bias of the exected-batch OT vector field in terms of $L^2(\nu_{B,t}dt)$

Theorem D.19. *The expected-batch OT vector field satisfies*

$$\|v_B - u^\star\|_{B,I}^2 \le \kappa_I^{\mathrm{cs}} \varepsilon_B^2. \tag{69}$$

Proof. Fix $t \in I$ and let $(X_0, X_1) \sim \overline{\pi}_B$, $X_t := F_t(X_0, X_1)$, and $V := X_1 - X_0$. By definition $v_B(t, X_t) = \mathbb{E}[V \mid X_t]$, the orthogonality of conditional expectation gives

$$\mathbb{E}\|V - u^\star(t, X_t)\|^2 = \mathbb{E}\|V - v_B(t, X_t)\|^2 + \mathbb{E}\|v_B(t, X_t) - u^\star(t, X_t)\|^2.$$

Theorefore,

$$\int \|v_B(t,x) - u^\star(t,x)\|^2 \, \nu_{B,t}(dx) \le \int \|V(x_0,x_1) - u^\star(t, F_t(x_0,x_1))\|^2 \, \overline{\pi}_B(dx_0, dx_1).$$

Applying Lemma D.17, we obtain

$$\int \|v_B(t,x) - u^\star(t,x)\|^2 \, \nu_{B,t}(dx) \le (1 + t\ell_t)^2 \int \|x_1 - S(x_0)\|^2 \, d\overline{\pi}_B$$

Integrating over t proves the claim. □

D.3.3 Graph-restricted detachment and plan discrepancy

Proposition D.20. *Under Assumption D.6, for every $x_0 \in K_0$ and $x_1 \in K_1$,*

$$G_\phi(x_1, x_0) \geq \frac{1}{2L_S} \|x_1 - S(x_0)\|^2 . \tag{70}$$

Consequently, we have the following bounds.

$$\varepsilon_B^2 \leq 2L_S \Delta_B, \tag{71}$$

$$d_B \leq \varepsilon_B \leq \sqrt{2L_S \Delta_B}, \tag{72}$$

$$\|v_B - u^\star\|_{B,I}^2 \leq 2L_S \kappa_I^{\mathrm{cs}} \Delta_B. \tag{73}$$

Proof. Fix $x_0 \in K_0$ and $x_1 \in K_1$. Since $K_1 = S(K_0)$, choose $z_0 \in K_0$ such that $x_1 = S(z_0)$. Fenchel equality at $(z_0, S(z_0))$ gives $\phi(x_1) = \langle z_0, x_1 \rangle - \Psi(z_0)$. Therefore

$$G_\phi(x_1, x_0) = \phi(x_1) + \Psi(x_0) - \langle x_1, x_0 \rangle = \Psi(x_0) - \Psi(z_0) - \langle S(z_0), x_0 - z_0 \rangle = D_\Psi(x_0, z_0),$$

which is the Bregman divergence from z_0 to x_0. Thus, the Assumption D.6 yields

$$G_\phi(x_1, x_0) \geq \frac{1}{2L_S} \|S(x_0) - S(z_0)\|^2 = \frac{1}{2L_S} \|S(x_0) - x_1\|^2 ,$$

which proves (70). Integrating with respect to $\overline{\pi}_B$ and using Lemma D.16 gives (71).

While the proof of $d_B \leq \varepsilon_B$ is basically given in [8, Appendix A.5.3], we show the proof here for completeness. To compare $\overline{\pi}_B$ and $\pi^\star$, draw $(X_0, X_1) \sim \overline{\pi}_B$. Since the first marginal of $\overline{\pi}_B$ is P_0, we see that $(X_0, S(X_0)) \sim (\mathrm{Id}, S)_\# P_0 = \pi^\star$. Hence, the joint law of $((X_0, X_1), (X_0, S(X_0)))$ is an admissible coupling of $\overline{\pi}_B$ and $\pi^\star$, with squared transportation cost $\mathbb{E} \|X_1 - S(X_0)\|^2 = \varepsilon_B^2$. This proves $d_B = W_2(\pi^\star, \overline{\pi}_B) \leq \varepsilon_B$. Finally, combine Theorem D.19 with (71) to obtain (73). □

D.4 Neural regression from fresh batches

In practical methods of OT-CFM, a continuous vector field is learned by a regression model, such as neural networks, so that we can generate a new sample with an arbitrary initial point. This section derives the neural network contribution from standard approximation and regression theory.

For ease of exposing the theory of neural approximation and finite-sample rate, we now restrict ourselves to regression on a compact set in this section. Recall that $K_i = \operatorname{supp} P_i$ $(i = 0, 1)$. Let us introduce the space of possible paths:

$$\mathcal{X}_I := \left\{ \big(t, (1-t)x_0 + tx_1\big) : t \in I,\ x_0 \in K_0,\ x_1 \in K_1 \right\} \subset I \times \mathbb{R}^d. \tag{74}$$

Assumption D.21.

$$\mathcal{X}_I \subset [0,1]^{d+1}.$$

Note that, under the assumption of compact supports, $\mathcal{X}_I \subset [0,1]^{d+1}$ is satisfied by applying an affine rescaling of the space. This rescaling is used only for easiness of stating the bounds below. An unbounded source can be treated by a separate truncation or unbounded-envelope regression analysis, but they are left out of the scope of this appendix.

We also use the notation

$$R_V := \sup_{x_0 \in K_0,\, x_1 \in K_1} \|x_1 - x_0\| < \infty. \tag{75}$$

D.4.1 Neural Network Learning

For each iteration $m = 1, \ldots, L$, sample two fresh endpoint batches of size B independently over m, solve the empirical quadratic OT problem, and draw independent time variables $T_{m,1}, \ldots, T_{m,B} \sim \mathrm{Unif}(I)$. After reindexing according to the selected OT coupling, write

$$U_m := \left\{ \big(X_{0,m,i}, X_{1,m,i}^{\mathrm{match}}, T_{m,i}\big) \right\}_{i=1}^B .$$

The batch objects $U_1, \ldots, U_L$ are i.i.d., while the matched pairs inside a batch are generally dependent.

For a measurable field $f : I \times \mathbb{R}^d \to \mathbb{R}^d$, define

$$\mathcal{L}_f(U_m) := \frac{c}{B} \sum_{i=1}^{B} \left\| f\big(T_{m,i}, F_{T_{m,i}}(X_{0,m,i}, X_{1,m,i}^{\text{match}})\big) - V(X_{0,m,i}, X_{1,m,i}^{\text{match}}) \right\|^2, \tag{76}$$

$$\widehat{R}_{B,L}(f) := \frac{1}{L} \sum_{m=1}^{L} \mathcal{L}_f(U_m), \tag{77}$$

where c is the early stopping time. Then $\mathbb{E}[\mathcal{L}_f(U_m)] = \mathcal{R}_B(f)$.

Throughout the remainder of Sec. D.4, the regression model is a sparse deep ReLU network. Let $\rho(a) := \max\{a, 0\}$ be the ReLU activation function. A network $\Phi : \mathbb{R}^{q_0} \to \mathbb{R}^{q_J}$ with depth J and widths $\mathbf{q} = (q_0, \ldots, q_J)$ is written as

$$\Phi = F_J \circ \rho \circ F_{J-1} \circ \cdots \circ \rho \circ F_1, \qquad F_j(z) = W_j z + b_j, \tag{78}$$

where the last affine map is not followed by an activation. We use

$$\mathcal{L}(\Phi) := J, \quad \mathcal{W}(\Phi) := \max_{0 \le j \le J} q_j, \quad \mathcal{M}(\Phi) := \sum_{j=1}^{J} \big(\|W_j\|_0 + \|b_j\|_0\big),$$
$$\mathcal{B}(\Phi) := \max_j \{\|W_j\|_{\max}, \|b_j\|_\infty\}, \tag{79}$$

for the depth, maximal width, number of nonzero parameters, and maximal parameter magnitude, respectively. In particular, P denotes sparsity $\mathcal{M}(\Phi)$, and is used as a complexity budget to decide the other parameters. Fix positive constants $C_D, C_A, \kappa_A, R_f < \infty$, and let $\mathcal{F}_P$ be the class of vector-valued ReLU networks $\Phi : I \times \mathbb{R}^d \to \mathbb{R}^d$ satisfying

$$\mathcal{L}(\Phi) \le \lceil C_D \log^2(P+1) \rceil, \;\; \mathcal{W}(\Phi) \le P+d+1, \;\; \mathcal{M}(\Phi) \le P, \;\; \mathcal{B}(\Phi) \le A_P, \;\; \sup_{z \in \mathcal{X}_I} \|\Phi(z)\| \le R_f, \tag{80}$$

where $A_P := C_A P^{\kappa_A}$. The width condition is inessential: after deleting inactive nodes, a network with at most P nonzero parameters can be represented with maximal width at most $P + d + 1$. The depth budget is broad enough for the structured construction in Section D.4.3. A fixed output envelope R_f can, if desired, be enforced by coordinatewise ReLU clipping, which has constant additional depth and sparsity.

For training, we assume that an approximate empirical minimizer $\widehat{u}_{B,L,P} \in \mathcal{F}_P$ is obtained:

$$\widehat{R}_{B,L}(\widehat{u}_{B,L,P}) \le \inf_{f \in \mathcal{F}_P} \widehat{R}_{B,L}(f) + \varepsilon_{\text{opt}}, \tag{81}$$

where ε_{opt} is the optimization error.

For $\epsilon > 0$, the *covering number* $N(\epsilon, \mathcal{F}_P, \|\cdot\|_{\infty,\mathcal{X}_I})$ is the smallest integer $K \ge 1$ for which there exist $f_1, \ldots, f_K \in \mathcal{F}_P$ satisfying

$$\sup_{f \in \mathcal{F}_P} \min_{1 \le j \le K} \|f - f_j\|_{\infty,\mathcal{X}_I} \le \epsilon.$$

Lemma D.22. *There is a constant C_{ent}, depending only on C_D and the fixed input and output dimensions, such that, for $0 < \epsilon \le 1$,*

$$\log N(\epsilon, \mathcal{F}_P, \|\cdot\|_{\infty,\mathcal{X}_I}) \le C P \log^2(P+1) \log\left(\frac{C A_P P \log(P+1)}{\epsilon}\right). \tag{82}$$

The last logarithm can be bounded by $C \log(CP/\epsilon)$.

Proof. For scalar output, Proposition 5.1 of Ohn and Kim [43], applied to the 1-Lipschitz ReLU activation, gives

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le 2\mathcal{L}(\Phi)(\mathcal{M}(\Phi)+1) \log\big(\epsilon^{-1}\mathcal{L}(\Phi)(\mathcal{W}(\Phi)+1)(\mathcal{B}(\Phi) \vee 1)\big).$$

The vector-output extension follows by covering the d coordinates and changing only a constant depending on d. Substitution of (80) gives the inequality in (82) Finally, $A_P \le C_A P^{\kappa_A}$ implies the last assertion. □

D.4.2 A finite-sample oracle inequality for random batches

Define

$$K_I^2 := |I|(R_f + R_V)^2 = c(R_f + R_V)^2, \tag{83}$$

where R_f and R_V are defined in (80) and (75), respectively.

Theorem D.23. *Suppose Assumption D.21 holds and let $L \geq 2$. There exist constants $C, C_0 > 0$, independent of B, L, P, such that, for every $\eta \in (0,1)$, with probability at least $1-\eta$ over the random batches, the following inequality holds:*

$$\|\widehat{u}_{B,L,P} - v_B\|_{B,I}^2 \leq C \inf_{f \in \mathcal{F}_P} \|f - v_B\|_{B,I}^2 + CK_I^2 \frac{P \log^2(P+1) \log(C_0 A_P P L) + \log(2/\eta)}{L} + C\varepsilon_{\mathrm{opt}}. \tag{84}$$

Consequently, with probability at least $1-\eta$,

$$\begin{aligned} \|\widehat{u}_{B,L,P} - u^\star\|_{B,I} \leq{}& C\, a_P(B) + C\sqrt{\kappa_{S,I} \Delta_B} \\ &+ CK_I \sqrt{\frac{P \log^2(P+1) \log(C_0 A_P P L) + \log(2/\eta)}{L}} + C\sqrt{\varepsilon_{\mathrm{opt}}} \end{aligned} \tag{85}$$

holds, where

$$a_P(B) := \inf_{f \in \mathcal{F}_P} \|f - u^\star\|_{B,I}.$$

Proof. The following proof is a specialization of the standard bounded empirical-risk minimization under a Bernstein-type variance condition. We include it for completeness.

First, the loss map satisfies

$$\sup_Z |\mathcal{L}_f(Z) - \mathcal{L}_g(Z)| \leq 2c(R_f + R_V)\|f - g\|_{\infty, \mathcal{X}_I}, \tag{86}$$

so that Lemma D.22 controls the entropy of the excess loss class $\{h_f := \mathcal{L}_f - \mathcal{L}_{v_B} \mid f \in \mathcal{F}_P\}$. We first show that this excess-loss class satisfies the Bernstein condition. Let $\mathbb{P}$ denote expectation with respect to one batch object Z, and let

$$\mathbb{P}_L h := \frac{1}{L} \sum_{m=1}^{L} h(U_m).$$

From

$$|\mathcal{L}_f(Z) - \mathcal{L}_{v_B}(Z)| \leq \frac{c}{B} \sum_{i=1}^{B} \|f(T_i, F_i) - v_B(T_i, F_i)\| \left(\|f(T_i, F_i)\| + \|v_B(T_i, F_i)\| + 2\|V_i\| \right),$$

we have

$$|h_f(Z)| \leq 2c(R_f + R_V) \frac{1}{B} \sum_{i=1}^{B} \|f(T_i, F_i) - v_B(T_i, F_i)\|.$$

Jensen's inequality and Lemma D.15 give

$$\begin{aligned} \mathbb{P} h_f^2 &\leq 4c^2 (R_f + R_V)^2 \mathbb{P} \left(\frac{1}{B} \sum_{i=1}^{B} \|f(T_i, F_i) - v_B(T_i, F_i)\| \right)^2 \\ &\leq CK_I^2 \mathbb{P} \left(\frac{1}{B} \sum_{i=1}^{B} \|f(T_i, F_i) - v_B(T_i, F_i)\|^2 \right) \\ &= CK_I^2 \|f - v_B\|_{B,I}^2 \\ &= CK_I^2 \mathbb{P} h_f. \end{aligned} \tag{87}$$

Thus, the excess-loss class satisfies the $(1, CK_I^2)$-Bernstein condition in the sense of Bartlett and Mendelson [5, Definition 2.6].

Consider a finite K_I^2/L-net of the excess-loss class in the uniform norm. The preceding loss-Lipschitz estimate (86) and Lemma D.22 imply that its logarithmic cardinality is bounded by

$$CP\log^2(P+1)\log(C_0A_PPL).$$

Applying the scalar Bernstein inequality [Theorem 1.1, 5] with (87) to each element of the finite net and taking a union bound yields

$$|(\mathbb{P}_L-\mathbb{P})h_f| \le C\sqrt{\frac{K_I^2(\mathbb{P}h_f)\{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)\}}{L}} + CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}$$

with probability larger than $1-\eta$. Using $\sqrt{ab}\le a/8+4b$, we further obtain

$$|(\mathbb{P}_L-\mathbb{P})h_f| \le \frac{1}{4}\mathbb{P}h_f + CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}.$$

Taking a union bound over the net, and then extending the estimate from the net to the whole class by the loss-Lipschitz bound, we obtain, with probability at least $1-\eta$,

$$\sup_{f\in\mathcal{F}_P}\left\{|(\mathbb{P}_L-\mathbb{P})h_f|-\frac{1}{4}\mathbb{P}h_f\right\} \le CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}. \tag{88}$$

This is the standard Bernstein-type uniform deviation inequality for a bounded excess-loss class satisfying a Bernstein condition; see Bartlett et al. [6, Theorem 3.3 and Sections 3.1.4–3.2] and Koltchinskii [31, Theorem 1].

Applying (88) to $\widehat{u} := \widehat{u}_{B,L,P}$ gives

$$\mathbb{P}h_{\widehat{u}} \le \mathbb{P}_Lh_{\widehat{u}} + \frac{1}{4}\mathbb{P}h_{\widehat{u}} + CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}.$$

Then, the ERM property of $\widehat{u}$, i.e., $\mathbb{P}_Lh_{\widehat{u}} \le \mathbb{P}_Lh_f + \varepsilon_{\mathrm{opt}}$ means

$$\frac{3}{4}\mathbb{P}h_{\widehat{u}} \le \mathbb{P}_Lh_f + CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L} + \varepsilon_{\mathrm{opt}}$$

for an arbitrary f. Finally, (88) for f gives

$$\mathbb{P}_Lh_f \le \frac{5}{4}\mathbb{P}h_f + CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}.$$

Combining the above expressions finally yields

$$\|\widehat{u}_{B,L,P}-v_B\|_{B,I}^2 \le C\|f-v_B\|_{B,I}^2 + CK_I^2\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L} + C\varepsilon_{\mathrm{opt}}.$$

Taking the infimum over $f\in\mathcal{F}_P$ proves (84).

To prove (85), by the triangle inequality,

$$\|\widehat{u}_{B,L,P}-u^\star\|_{B,I} \le \|\widehat{u}_{B,L,P}-v_B\|_{B,I} + \|v_B-u^\star\|_{B,I}.$$

Taking square roots in (84) with $\sqrt{a+b+c}\le\sqrt{a}+\sqrt{b}+\sqrt{c}$ gives

$$\|\widehat{u}_{B,L,P}-v_B\|_{B,I} \le C\inf_{f\in\mathcal{F}_P}\|f-v_B\|_{B,I} + CK_I\sqrt{\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}} + C\sqrt{\varepsilon_{\mathrm{opt}}}.$$

For every $f\in\mathcal{F}_P$, another application of the triangle inequality yields

$$\|f-v_B\|_{B,I} \le \|f-u^\star\|_{B,I} + \|u^\star-v_B\|_{B,I}.$$

Hence, by the definition of $a_P(B)$,

$$\inf_{f\in\mathcal{F}_P}\|f-v_B\|_{B,I} \le a_P(B) + \|u^\star-v_B\|_{B,I}.$$

Finally, Proposition D.19 gives

$$\|u^\star-v_B\|_{B,I} \le \sqrt{2\kappa_{S,I}\Delta_B}.$$

Combining the preceding three displays completes the proof. □

D.4.3 Intrinsic-dimensional approximation of the population OT denoiser

Standard smooth-function approximation treats the population OT field as a function of $d+1$ variables [67, 54, 43]; the vector field $u(t,t)$ indeed takes a $d+1$-dimensional variable as its input. In this subsection, we exploit the proximal representation of the population OT field. Under additional smoothness assumptions on the target manifold and on the restriction of the reverse Brenier potential to that manifold, we obtain an approximation exponent depending on $m+1$, where $m = \dim M$.

Our construction first develops the OT manifold proximal map on the exact population OT interpolation graph and then extends it to a fixed ambient tube around the graph. The tube extension is important: it yields a direct approximation bound under the expected-batch interpolation law by perturbing the proximal objective. This construction adapts the approach of the finite-anchor projection network developed by Fu et al. [23] from their Euclidean metric projection to the OT-induced manifold proximal map. More precisely, the squared-distance objective for the nearest-point projection in their diffusion case is replaced by the time-dependent proximal objective $\mathcal{J}_{(t,x)}$ generated by the reverse Brenier potential, and the reach-based projection gap is replaced by an OT-specific Fenchel–Bregman quadratic-growth inequality. The proof architecture is therefore related to that of Fu et al. [23], while the stability mechanism and the selected map are specific to the OT proximal geometry.

D.4.3.1 ReLU arithmetic modules. We use the ReLU-network notation and complexity measures $\mathcal{L}(\Phi)$, $\mathcal{W}(\Phi)$, $\mathcal{M}(\Phi)$, and $\mathcal{B}(\Phi)$ introduced in (78)–(79). The following standard arithmetic modules will be used repeatedly.

Lemma D.24 (ReLU arithmetic networks)**.** *Fix $B, q \in \mathbb{N}$. There are constants $C, c > 0$, depending only on B and q, such that for every $\varepsilon \in (0, 1/2)$ the following hold;*

1. *There is a ReLU network*
$$\mathrm{Mult}_{B,\varepsilon} : [-B, B]^2 \to \mathbb{R}$$
such that
$$\sup_{|a|,|b|\le B} |\mathrm{Mult}_{B,\varepsilon}(a,b) - ab| \le \varepsilon,$$
and
$$\mathrm{Mult}_{B,\varepsilon}(0,b) = \mathrm{Mult}_{B,\varepsilon}(a,0) = 0.$$
Moreover,
$$\mathcal{L}(\mathrm{Mult}_{B,\varepsilon}) + \mathcal{M}(\mathrm{Mult}_{B,\varepsilon}) \le C\log(1/\varepsilon), \qquad \mathcal{B}(\mathrm{Mult}_{B,\varepsilon}) \le \varepsilon^{-c}.$$

2. *Let*
$$p(u,z) = \sum_{|\beta|\le q} \ell_\beta(u) z^\beta,$$
where u ranges over a fixed compact subset of a Euclidean space, each ℓ_β is affine in u, and z ranges over $[-B, B]^m$. There is a ReLU network $\mathrm{Poly}_{p,\varepsilon}$ such that
$$\sup_{u,z} \left|\mathrm{Poly}_{p,\varepsilon}(u,z) - p(u,z)\right| \le \varepsilon,$$
with
$$\mathcal{L}(\mathrm{Poly}_{p,\varepsilon}) + \mathcal{M}(\mathrm{Poly}_{p,\varepsilon}) \le C\log(1/\varepsilon), \qquad \mathcal{B}(\mathrm{Poly}_{p,\varepsilon}) \le \varepsilon^{-c}.$$

3. *For every fixed interval $[a,b] \subset (0,\infty)$, there is a ReLU network*
$$\mathrm{Inv}_{[a,b],\varepsilon} : [a,b] \to \mathbb{R}$$
such that
$$\sup_{r\in[a,b]} \left|\mathrm{Inv}_{[a,b],\varepsilon}(r) - r^{-1}\right| \le \varepsilon,$$
and
$$\mathcal{L}(\mathrm{Inv}_{[a,b],\varepsilon}) + \mathcal{M}(\mathrm{Inv}_{[a,b],\varepsilon}) \le C\log^2(1/\varepsilon), \qquad \mathcal{B}(\mathrm{Inv}_{[a,b],\varepsilon}) \le \varepsilon^{-c}.$$

Proof. The multiplication and monomial constructions follow from the standard sawtooth approximation of the square function and polarization; see, for example, Yarotsky [67] and Schmidt-Hieber [54, Lemmas A.2–A.4].

For part 1, the construction in Schmidt-Hieber [54] additionally gives the zero-preserving property in part 1.

For part 2, use the monomial approximation networks of Schmidt-Hieber [54, Lemmas A.3–A.4]. The affine functions $\ell_\beta(u)$ are represented exactly. Multiplying their outputs by the monomial approximations using part 1 and summing the resulting terms gives the desired network. Since the dimensions and degree are fixed and all intermediate quantities are bounded on the fixed domain, choosing the individual approximation errors as sufficiently small constant multiples of ε gives the stated uniform error with depth and sparsity $O(\log(1/\varepsilon))$.

For part 3, let $g(u) := u^{-1}$ on $[1, b/a]$. For every $n \in \mathbb{N}_0$,

$$\sup_{u\in[1,b/a]} |g^{(n)}(u)| = \sup_{u\in[1,b/a]} \frac{n!}{u^{n+1}} \le n!.$$

Thus, Elbrächter et al. [21, Lemma A.6] provides a ReLU approximation Φ_δ of g with uniform error at most δ, fixed width, depth $O(\log^2(1/\delta))$, and bounded weights. Taking $\delta = \min\{a, 1\}\varepsilon$ and defining

$$\mathrm{Inv}_{[a,b],\varepsilon}(x) := a^{-1}\Phi_\delta(x/a)$$

gives uniform error at most ε. Fixed width implies that sparsity has the same order as depth, and the fixed input and output rescalings preserve these orders. □

D.4.3.2 The manifold proximal objective Recall that $M = K_1 \subset \mathbb{R}^d$ is the support of the target distribution, which is an m-dimensional embedded manifold, and that $\phi := \Psi^*$ is the reverse extended Brenier potential. We write

$$\bar{\phi} := \phi|_M.$$

We first define the minimizer of the proximal objective restricted to M, and approximate it by a neural network model. For $X = (t, x) \in I \times \mathbb{R}^d$, define the manifold proximal objective

$$\mathcal{J}_X(z) := (1-t)\bar{\phi}(z) + \frac{t}{2}\|z\|^2 - \langle x, z\rangle, \qquad z \in M. \tag{89}$$

This corresponds to the proximal objective on $\mathbb{R}^d$ through

$$(1-t)\left\{\phi(z) + \frac{1}{2\lambda_t}\left\|z - \frac{x}{t}\right\|^2\right\} = (1-t)\phi(z) + \frac{t}{2}\|z\|^2 - \langle x, z\rangle + \frac{\|x\|^2}{2t}. \tag{90}$$

Thus, $\mathcal{J}_X$ is essentially the proximal objective restricted on M. However, the form (89) is preferred for the following reasons; it remains well defined at $t = 0$, where the usual variables x/t and λ_t are singular; and for every fixed $z \in M$, the value $X \mapsto \mathcal{J}_X(z)$ is affine in $X = (t, x)$. The last property is crucial for the finite-anchor ReLU construction below.

For $x_0 \in K_0$, set

$$Z_t(x_0) := (1-t)x_0 + tS(x_0),$$

and introduce the population OT interpolation graph by

$$\Gamma_I^\star := \left\{\big(t, Z_t(x_0)\big) : t \in I,\ x_0 \in K_0\right\}. \tag{91}$$

We define the *manifold proximal map* on $\Gamma_I^\star$ by

$$\Pi_{\mathrm{OT}}(X) := \operatorname*{argmin}_{z\in M} \mathcal{J}_X(z), \qquad X \in \Gamma_I^\star. \tag{92}$$

Assumption D.25. Let $s = r + \alpha > 0$, where $r \in \mathbb{N}_0$ and $\alpha \in (0, 1]$. Assume the following.

1. $M \subset \mathbb{R}^d$ is a compact embedded $C^{r+2,\alpha}$ manifold without boundary, of dimension m.
2. The restriction $\bar{\phi} = \phi|_M$ is finite and belongs to $C^{r+2,\alpha}(M)$.
3. The Bregman co-coercivity of Assumption D.6 holds:

$$D_\Psi(x, z) := \Psi(x) - \Psi(z) - \langle S(z), x - z\rangle \ge \frac{1}{2L_S}\|S(x) - S(z)\|^2, \qquad x, z \in K_0. \tag{93}$$

4. The manifold admits uniformly regular local charts: there exist $\rho_{\mathrm{ch}} > 0$ and $C_{\mathrm{ch}} < \infty$ such that, for every $y \in M$, there is a chart $Y_y : B^m_{\rho_{\mathrm{ch}}}(0) \to M$ with $Y_y(0) = y$ satisfying

$$C_{\mathrm{ch}}^{-1} I_m \preceq DY_y(\xi)^\top DY_y(\xi) \preceq C_{\mathrm{ch}} I_m \tag{94}$$

and

$$\|Y_y\|_{C^{r+2,\alpha}} + \|Y_y^{-1}\|_{C^{r+2,\alpha}} \le C_{\mathrm{ch}} \tag{95}$$

on their respective domains.

The uniform-chart condition is a standard bounded-geometry consequence for a compact, sufficiently smooth embedded manifold. It is stated explicitly to guarantee that the constants in the network construction must be uniform over the centers of the local charts.

Lemma D.26. *Suppose Assumption D.25 holds. Let $x_0 \in K_0$, $y = S(x_0)$, $t \in I$, and*

$$X = \big(t, Z_t(x_0)\big) \in \Gamma_I^\star.$$

Then y is the unique minimizer of $\mathcal{J}_X$ on M:

$$y = \operatorname*{argmin}_{z \in M} \mathcal{J}_X(z) = \Pi_{\mathrm{OT}}(X). \tag{96}$$

For $t > 0$, it also holds that

$$y = Q_t(Z_t(x_0)) = \mathrm{prox}_{\lambda_t \phi}\left(\frac{Z_t(x_0)}{t}\right). \tag{97}$$

Proof. Since $x_0 \in \partial\phi(y)$, the proximal optimality condition gives $y = \mathrm{prox}_{\lambda_t\phi}(y + \lambda_t x_0)$. Because $y + \lambda_t x_0 = Z_t(x_0)/t$, this proves (97). The proximal minimizer on $\mathbb{R}^d$ is unique and belongs to M, so it is also the unique minimizer of $\mathcal{J}_X$, the restriction to M. The case $t = 0$ follows directly from $x_0 \in \partial\phi(y)$, which means that y minimizes $z \mapsto \phi(z) - \langle x_0, z\rangle$. □

This lemma implies that, on the exact population OT graph, the manifold proximal map Π_{OT} equals to the denoizer Q_t.

Lemma D.27. *Let $x_0 \in K_0$. Under Assumption D.25, for every $X = (t, Z_t(x_0)) \in \Gamma_I^\star$ and every $z \in M$,*

$$\mathcal{J}_X(z) - \mathcal{J}_X(S(x_0)) \ge \frac{\mu_\star}{2}\|z - S(x_0)\|^2, \tag{98}$$

where $\mu_\star := \min\{1, \frac{1}{L_S}\} > 0$.

Proof. Let $y := S(x_0)$. Since $M = S(K_0)$, choose $z_0 \in K_0$ such that $z = S(z_0)$. Fenchel equality at (x_0, y) and (z_0, z) gives

$$\bar\phi(z) - \bar\phi(y) - \langle x_0, z - y\rangle = \Psi(x_0) - \Psi(z_0) - \langle S(z_0), x_0 - z_0\rangle = D_\Psi(x_0, z_0). \tag{99}$$

Using $x = Z_t(x_0) = (1-t)x_0 + ty$, we obtain

$$\begin{aligned}\mathcal{J}_X(z) - \mathcal{J}_X(y) &= (1-t)\left\{\bar\phi(z) - \bar\phi(y) - \langle x_0, z - y\rangle\right\} + \frac{t}{2}\|z - y\|^2 \\ &= (1-t)D_\Psi(x_0, z_0) + \frac{t}{2}\|z - y\|^2.\end{aligned} \tag{100}$$

By (93), $D_\Psi(x_0, z_0) \ge \frac{1}{2L_S}\|y - z\|^2$, and consequently,

$$\mathcal{J}_X(z) - \mathcal{J}_X(y) \ge \frac{1}{2}\left(t + \frac{1-t}{L_S}\right)\|z - y\|^2 \ge \frac{\mu_\star}{2}\|z - y\|^2.$$

□

Lemma D.27 establishes a uniform quadratic-growth bound for $\mathcal{J}_X$ on M when $X \in \Gamma_I^\star$. To discuss the paths given by minibatch OT, we need to consider a slightly larger set than $\Gamma_I^\star$ for parameter X. In the next lemma, we extend the above bound of quadratic growth to parameter $X^e = (t, Z_t(x_0) + e)$ for sufficiently small $\|e\|$.

Lemma D.28. *Under Assumption D.25, there exist constants $\delta_0 > 0$ and $\mu_{\mathrm{tub}} > 0$ such that the following holds. Define*

$$\Gamma^\star_{I,\delta_0} := \left\{ \big(t, Z_t(x_0) + e\big) : t \in I,\ x_0 \in K_0,\ \|e\| \le \delta_0 \right\}. \tag{101}$$

For

$$X^e = \big(t, Z_t(x_0) + e\big) \in \Gamma^\star_{I,\delta_0}, \qquad y = S(x_0),$$

the restriction $\mathcal{J}_{X^e}|_M$ has a unique minimizer, denoted by

$$y_e := \Pi_{\mathrm{OT}}(X^e) := \operatorname*{argmin}_{z \in M} \mathcal{J}_{X^e}(z).$$

Moreover,

$$\|y_e - y\| \le \frac{2}{\mu_\star}\|e\|, \tag{102}$$

and, for every $z \in M$,

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y_e) \ge \frac{\mu_{\mathrm{tub}}}{2}\|z - y_e\|^2. \tag{103}$$

The constants are uniform over all representations of X^e in (101).

Proof. **localization of every global minimizer.** Fix $t \in I$ and $x_0 \in K_0$, and write $X^0 := (t, Z_t(x_0))$, $X^e := (t, Z_t(x_0) + e)$, and $y := S(x_0)$. Since M is compact and $\mathcal{J}_{X^e}|_M$ is continuous, a global minimizer exists. Let y_e be any such minimizer; uniqueness is not assumed at this stage. The definition (89) gives $\mathcal{J}_{X^e}(z) = \mathcal{J}_{X^0}(z) - \langle e, z\rangle$ for any $z \in M$, and therefore

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y) = \mathcal{J}_{X^0}(z) - \mathcal{J}_{X^0}(y) - \langle e, z - y\rangle.$$

Lemma D.27 with $X = X^0$ and $S(x_0) = y$ yields

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y) \ge \frac{\mu_\star}{2}\|z - y\|^2 - \|e\|\,\|z - y\|, \qquad z \in M. \tag{104}$$

Taking $z = y_e$ and using $\mathcal{J}_{X^e}(y_e) \le \mathcal{J}_{X^e}(y)$ gives

$$\frac{\mu_\star}{2}\|y_e - y\|^2 \le \|e\|\,\|y_e - y\|,$$

which proves (102). This estimate holds for every global minimizer, independently of whether the minimizer is unique.

Stability with parameters. To prove the quadratic growth (103), we use the results of Lewis [34], where they discuss strong minimizers, i.e., quadratic growth in a neighborhood. In particular, to derive the uniform coefficient of the quadratic growth in (103), we use Theorem 5.5 of Lewis [34], which discusses the strong minimizers depending on a parameter. For our purpose, we consider $\mathcal{J}_X(z)$ as a function of z with parameter X.

Define $f_X(z) := \mathcal{J}_X(z)$ for $z \in M$ and $+\infty$ for $z \notin M$, and parametric mapping

$$\tilde{f}(X, z) := f_X(z),$$

where $X = (t, x) \in \mathbb{R}^{d+1}$ is regarded as a parameter for the minimization problem given by $z \mapsto f_X(z)$. Let $\tilde{M} := \mathbb{R}^{d+1} \times M$. The function $\tilde{f}$ is proper and lower semicontinuous. By Assumption D.25(1)–(2), $\tilde{M}$ is a C^2 manifold, and $\tilde{f}$ is a C^2 function on $\tilde{M}$ extended by $+\infty$ outside $\tilde{M}$.

The quadratic growth (98) means that $y = S(x_0)$ is a strong local minimizer of $\tilde{f}(X^0, \cdot)|_M$. Moreover, the transversality condition required for the theorem in [34] is obvious under Assumption D.25(1)–(2). From Lewis [34, Theorem 5.5], there are neighborhoods V of X^0 and U of y such that, for every $X' \in V$, the restriction $\mathcal{J}_{X'}|_{M\cap U}$ has a unique critical point, which is a strong local minimizer and depends on X' in C^1-smooth.

Global minimizer. We next show that every global minimizer lies in U after shrinking V. Choose $a > 0$ such that $B_y(a) := \{y' \in \mathbb{R}^d \mid \|y' - y\| \le a\} \subset U$, and set $\gamma := \mu_\star a^2/2 > 0$. For $z \in M \setminus U$, (98) implies

$$\mathcal{J}_{X^0}(z) - \mathcal{J}_{X^0}(y) \ge \gamma.$$

For $X' = (t', x')$, the definition (89) also gives

$$\mathcal{J}_{X'}(z) - \mathcal{J}_{X^0}(z) = (t' - t)\Big(\frac{1}{2}\|z\|^2 - \bar{\phi}(z)\Big) - \langle x' - Z_t(x_0), z\rangle.$$

Compactness of M therefore provides a finite constant $C_J > 0$ such that

$$\sup_{z\in M} |\mathcal{J}_{X'}(z) - \mathcal{J}_{X^0}(z)| \le C_J\|X' - X^0\|.$$

Shrink V so that $2C_J\|X' - X^0\| < \gamma/2$ on V. For every $z \in M \setminus U$, it then follows that

$$\mathcal{J}_{X'}(z) - \mathcal{J}_{X'}(y) \ge \gamma - 2C_J\|X' - X^0\| > \gamma/2,$$

which shows that no point of $M \setminus U$ can be a global minimizer. Hence, the local minimizer in $M \cap U$ is also the global minimizer. We may therefore denote this point by $\Pi_{\mathrm{OT}}(X')$, which is single-valued and C^1 on V.

Apply this argument at every $X^0 \in \Gamma_I^\star$. On overlapping parameter neighborhoods, the resulting maps agree, because both give the unique global minimizer of the same objective. They therefore define a single C^1 map Π_{OT} on an open neighborhood $\mathcal{V}$ of $\Gamma_I^\star$.

Uniform Lipschitz bound. We will show the Lipschitz property of $e \mapsto y_e = \Pi_{\mathrm{OT}}(X^e)$. By the compactness of $\Gamma_I^\star$, choose $\rho > 0$ such that

$$\Gamma_{I,2\rho}^\star = \{(t, Z_t(x_0) + e) : t \in I,\ x_0 \in K_0,\ \|e\| \le 2\rho\} \subset \mathcal{V}.$$

Since Π_{OT} is C^1 on $\mathcal{V}$, the constant

$$L := \max\Big\{1, \sup_{X\in\Gamma_{I,2\rho}^\star} \|D_x\Pi_{\mathrm{OT}}(X)\|_{\mathrm{op}}\Big\}$$

is finite. For fixed t, x_0 and $\|e_1\|, \|e_2\| \le 2\rho$, put $e_\theta := (1-\theta)e_1 + \theta e_2$ for $\theta \in [0,1]$. The entire segment $(t, Z_t(x_0) + e_\theta)$ remains in this tube. Expressing $y_{e_1} - y_{e_2}$ in terms of the integral over the segment yields

$$\|y_{e_2} - y_{e_1}\| \le L\|e_2 - e_1\|, \qquad \|e_1\|, \|e_2\| \le 2\rho, \tag{105}$$

with the same L for every $X^0 \in \Gamma_I^\star$.

We now choose the radius δ_0 of the smaller tube in the statement of the lemma. Take $0 < \delta_0 \le \min\{\rho/2, \mu_\star\rho/4\}$. Then, (102) gives $\|y_e - y\| \le \rho/2$ whenever $\|e\| \le \delta_0$.

Quadratic growth with a common constant. We derive the uniform quadratic-growth bound for $X^e \in \Gamma_{I,\delta_0}^\star$ from the uniform Lipschitz estimate (105) and the optimality of the perturbed minimizers.

We first consider the case $z \in M$ with $\|z - y\| < \rho$. Fix $\|e\| \le \delta_0$, and introduce the additional perturbation

$$h := \frac{z - y_e}{2L}.$$

Since $\|z - y_e\| \le \|z - y\| + \|y - y_e\| < 3\rho/2$, $\|e\| \le \delta_0 \le \rho/2$, and $L \ge 1$, we have $\|e + h\| \le \|e\| + \frac{\|z-y_e\|}{2L} < \frac{5\rho}{4} < 2\rho$. Thus, $X^{e+h} := (t, Z_t(x_0) + e + h)$ belongs to the larger tube $\Gamma_{I,2\rho}^\star$ so that $y_{e+h} := \Pi_{\mathrm{OT}}(X^{e+h})$ is well defined. (105) gives $\|y_{e+h} - y_e\| \le L\|h\|$. From the identity $\mathcal{J}_{X^{e+h}} = \mathcal{J}_{X^e} - \langle h, \cdot\rangle$, the optimality of y_{e+h} for $\mathcal{J}_{X^{e+h}}$ and the optimality of y_e for $\mathcal{J}_{X^e}$ provide

$$\mathcal{J}_{X^e}(z) - \langle h, z\rangle \ge \mathcal{J}_{X^e}(y_{e+h}) - \langle h, y_{e+h}\rangle \ge J_{X^e}(y_e) - \langle h, y_{e+h}\rangle,$$

from which we obtain

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y_e) \ge \langle h, z - y_{e+h}\rangle = \langle h, z - y_e\rangle - \langle h, y_{e+h} - y_e\rangle.$$

Cauchy-Schwarz and the Lipschitz bound (105) imply $\langle h, y_{e+h} - y_e\rangle \le \|h\|\,\|y_{e+h} - y_e\| \le L\|h\|^2$. The definition of h therefore yields

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y_e) \ge \langle h, z - y_e\rangle - L\|h\|^2 = \frac{1}{4L}\|z - y_e\|^2.$$

Consequently, with $\mu_{\mathrm{loc}} := 1/(2L) > 0$,

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y_e) \ge \frac{\mu_{\mathrm{loc}}}{2}\|z - y_e\|^2 \quad \text{for} \quad z \in M, \quad \|z - y\| < \rho. \tag{106}$$

Both ρ and μ_{loc} are independent of t, x_0, e within the ranges specified above.

It remains to consider $z \in M$ with $\|z - y\| \geq \rho$. The choice of δ_0 implies $\|e\| \leq \frac{\mu_\star \rho}{4} \leq \frac{\mu_\star}{4}\|z - y\|$. Using (104) and $\mathcal{J}_{X^e}(y_e) \leq \mathcal{J}_{X^e}(y)$, we obtain

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y_e) \geq \mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y) \geq \frac{\mu_\star}{4}\|z - y\|^2.$$

Since $\|y_e - y\| \leq \rho/2$, we have $\|z - y_e\| \leq \|z - y\| + \frac{\rho}{2} \leq \frac{3}{2}\|z - y\|$. Thus

$$\mathcal{J}_{X^e}(z) - \mathcal{J}_{X^e}(y_e) \geq \frac{\mu_\star}{9}\|z - y_e\|^2, \qquad \|z - y\| \geq \rho.$$

Together with (106), this proves (103) with $\mu_{\mathrm{tub}} := \min\{\frac{1}{2L}, \frac{2\mu_\star}{9}\}$.

The choices of δ_0 and μ_{tub} depend only on the fixed objective family and its OT interpolation graph, not on a particular t or x_0. □

Remark D.29. The arguments to derive the local quadratic growth bound are consistent with the general equivalence between tilt stability and uniform local quadratic growth established by Drusvyatskiy and Lewis [20, Corollary 3.2]. We derived the bound more directly in the proof.

Observe that neither Lemma D.27 nor Lemma D.28 requires a centered normal fiber $0 \in C_y$, or a smooth global map $y \mapsto s(y) \in \partial\phi(y)$. The entire normal OT fiber is encoded through the input x in the proximal objective $\mathcal{J}_X$.

D.4.3.3 Intrinsic-dimensional ReLU approximation We now state the main approximation theorem.

Theorem D.30. *Suppose Assumption D.25 holds, with $s = r + \alpha > 0$ as specified there. There exist constants $C, h_0, R_\Phi > 0$, depending only on $d, m, s, c, K_0, M, L_S, \|\bar{\phi}\|_{C^{r+2,\alpha}(M)}$, and C_{ch}, such that, for every $h \in (0, h_0)$, there is a ReLU network $\Phi_h : I \times \mathbb{R}^d \to \mathbb{R}^d$ satisfying*

$$\sup_{X \in \Gamma^\star_{I,\delta_0}} \|\Phi_h(X) - \Pi_{\mathrm{OT}}(X)\| \leq Ch^s, \tag{107}$$

and

$$\mathcal{M}(\Phi_h) \leq Ch^{-(m+1)} \log^2(1/h), \qquad \mathcal{L}(\Phi_h) \leq C \log^2(1/h), \qquad \mathcal{B}(\Phi_h) \leq h^{-C}. \tag{108}$$

The network may additionally be chosen with the global output bound

$$\sup_{(t,x) \in I \times \mathbb{R}^d} \|\Phi_h(t, x)\| \leq R_\Phi. \tag{109}$$

Consequently, for every sufficiently large P, there exists a ReLU network Φ_P with

$$\mathcal{M}(\Phi_P) \leq P, \qquad \mathcal{L}(\Phi_P) \leq C \log^2 P, \qquad \mathcal{B}(\Phi_P) \leq P^C, \tag{110}$$

such that

$$\sup_{X \in \Gamma^\star_{I,\delta_0}} \|\Phi_P(X) - \Pi_{\mathrm{OT}}(X)\| \leq C \left(\frac{P}{\log^2 P}\right)^{-s/(m+1)}. \tag{111}$$

Remark D.31. Theorem D.30 shows that the manifold proximal map Π_{OT} can be approximated by a ReLU network in such accuracy that the polynomial exponent with respect to the number of nonzero parameters is $s/(m+1)$, independent of the ambient dimension d. The constants and the polynomial weight-envelope exponent may still depend on d.

Proof. The proof is constructive.

Step 1: h-net of $I \times M$. Equip $\mathcal{B} := I \times M$ with the metric

$$d_{\mathcal{B}}\big((t, y), (t', y')\big) := |t - t'| + d_M(y, y').$$

Choose a maximal h-separated set $\{b_j = (t_j, y_j)\}_{j=1}^{N_h} \subset \mathcal{B}$. It is also an h-net. Since $I \times M$ is a compact $(m+1)$-dimensional manifold, the size of the set is bounded by

$$N_h \leq Ch^{-(m+1)}. \tag{112}$$

For $X = (t, x) \in \Gamma^\star_{I,\delta_0}$, write

$$y := \Pi_{\mathrm{OT}}(X) \in M.$$

Step 2: energy-based localization. For each $j = 1, \dots, N_h$, define the score $\Sigma_j(X)$ to measure the discrepancy between (t, y) and b_j by

$$\Sigma_j(X) := \mathcal{J}_X(y_j) + \lambda h|t - t_j|, \tag{113}$$

where $\lambda > 0$ is a fixed constant, and set

$$\underline{\Sigma}_h(X) := \min_{1 \le j \le N_h} \Sigma_j(X). \tag{114}$$

Notice that $\mathcal{J}_X(y_j)$ is affine in the input $X = (t, x)$ of $\Sigma_j(X)$. Also, $|t - t_j| = \rho(t - t_j) + \rho(t_j - t)$ and $\min\{a, b\} = a - \rho(a - b)$. Hence, all the scores and their minimum can be represented exactly by a ReLU network with $O(N_h)$ nonzero parameters and depth $O(\log N_h)$.

We next establish the localization property. Since y is a minimizer of $\mathcal{J}_X|_M$, its intrinsic gradient vanishes:

$$\nabla_M \mathcal{J}_X(y) = 0.$$

Smoothness of $\bar{\phi}$ and compactness of M give a constant $L_J < \infty$ such that, whenever $d_M(y, z)$ is sufficiently small,

$$0 \le \mathcal{J}_X(z) - \mathcal{J}_X(y) \le L_J d_M(y, z)^2, \tag{115}$$

uniformly in $X \in \Gamma^\star_{I,\delta_0}$.

For (t, y), choose j_0 such that $d_{\mathcal{B}}\big((t, y), b_{j_0}\big) \le h$. Then, it follows from (115) that $\Sigma_{j_0}(X) \le \mathcal{J}_X(y) + C_0 h^2$ for a constant C_0, and therefore,

$$\underline{\Sigma}_h(X) \le \mathcal{J}_X(y) + C_0 h^2. \tag{116}$$

Fix $A > 0$ and define the raw gates

$$\widetilde{\eta}_j(X) := \left[1 - \frac{\Sigma_j(X) - \underline{\Sigma}_h(X)}{Ah^2}\right]_+. \tag{117}$$

The score difference is nonnegative, so $0 \le \widetilde{\eta}_j \le 1$. The index attaining the minimum score satisfies $\widetilde{\eta}_j(X) = 1$.

If $\widetilde{\eta}_j(X) > 0$, then $\Sigma_j(X) < \underline{\Sigma}_h(X) + Ah^2$, which in combination with (116) and Lemma D.28 yields

$$\frac{\mu_{\mathrm{tub}}}{2}\|y_j - y\|^2 + \lambda h|t_j - t| \le (A + C_0)h^2.$$

Consequently,

$$\|y_j - y\| \le Ch, \qquad |t_j - t| \le Ch. \tag{118}$$

for a constant $C > 0$. For sufficiently small h, local equivalence of the Euclidean and geodesic metrics on M gives

$$\widetilde{\eta}_j(X) > 0 \quad \Longrightarrow \quad d_{\mathcal{B}}\big((t, y), b_j\big) \le C_{\mathrm{loc}} h \tag{119}$$

for a constant C_{loc}.

Since the set $\{b_j\}$ is h-separated, the packing estimate on the $(m + 1)$-dimensional manifold $I \times M$ implies that there is a constant $N_{\mathrm{ov}} < \infty$, independent of h, such that

$$\#\{j : \widetilde{\eta}_j(X) > 0\} \le N_{\mathrm{ov}} \tag{120}$$

for every $X \in \Gamma^\star_{I,\delta_0}$. We call such j an active index.

Step 3: local proximal experts. For $X \in \Gamma^\star_{I,\delta_0}$, fix an active index j and use the uniformly regular chart $Y_j := Y_{y_j} : B^m_{\rho_{\mathrm{ch}}}(0) \to M$ in the assumption D.25. From (119), by taking sufficiently small h, the true minimizer y belongs to this chart and can be written as

$$y = Y_j(\xi_j^\star), \qquad \|\xi_j^\star\| \le C_r h. \tag{121}$$

Define

$$A_j(\xi) := \bar{\phi}(Y_j(\xi))$$

and

$$J_{j,X}(\xi) := \mathcal{J}_X(Y_j(\xi)).$$

Its gradient is

$$G_j(X,\xi) := \nabla_\xi J_{j,X}(\xi) = (1-t)\nabla A_j(\xi) + DY_j(\xi)^\top \{tY_j(\xi) - x\}. \tag{122}$$

In particular, for each fixed ξ, the map $X = (t,x) \mapsto G_j(X,\xi)$ is affine.

We first consider $s > 1$. We use the C^1 regularity of Π_{OT} on an open neighborhood of $\Gamma^\star_{I,\delta_0}$, established in the proof of Lemma D.28. In particular, compactness of the tube gives

$$L_\Pi := \max\left\{1, \sup_{X \in \Gamma^\star_{I,\delta_0}} \|D_x \Pi_{\mathrm{OT}}(X)\|_{\mathrm{op}}\right\} < \infty.$$

This constant is independent of h and of the anchor index j. For fixed j, the expression

$$\xi_j^\star(X) = Y_j^{-1}(\Pi_{\mathrm{OT}}(X))$$

defines a C^1 map on a neighborhood of any input for which the minimizer belongs to the chart. Such a neighborhood exists at each active pair (j,X), by (121) and continuity of Π_{OT}. Since $\Pi_{\mathrm{OT}}(X)$ minimizes $\mathcal{J}_X$ on M, we have

$$G_j(X, \xi_j^\star(X)) = 0. \tag{123}$$

Thus, the local solution and its smooth dependence on X are inherited from Lemma D.28.

Step 3: A uniform Hessian bound at the minimizer. Fix an active pair (j,X) and abbreviate

$$B_j := DY_j(\xi_j^\star), \qquad H_j := D^2_\xi J_{j,X}(\xi_j^\star).$$

The matrix H_j is symmetric positive semidefinite, since $\xi_j^\star$ is an interior local minimizer of the C^2 function $J_{j,X}$. By (122), differentiation with respect to x, with ξ held fixed, gives $D_x G_j(X,\xi) = -DY_j(\xi)^\top$. Differentiating the stationarity equation (123) with respect to x therefore yields

$$H_j D_x \xi_j^\star(X) = B_j^\top.$$

The chart metric bound (94) makes B_j full column rank. Hence $B_j^\top$ has rank m, and the last identity forces the $m \times m$ matrix H_j to be invertible. Combined with positive semidefiniteness, this shows that H_j is positive definite. Consequently,

$$D_x \xi_j^\star(X) = H_j^{-1} B_j^\top.$$

Differentiating $\Pi_{\mathrm{OT}}(X) = Y_j(\xi_j^\star(X))$ now gives

$$D_x \Pi_{\mathrm{OT}}(X) = B_j H_j^{-1} B_j^\top.$$

In particular, this derivative is symmetric positive semidefinite, and its operator-norm bound implies $B_j H_j^{-1} B_j^\top \preceq L_\Pi I_d$. Multiply this inequality on the left by $(B_j^\top B_j)^{-1} B_j^\top$ and on the right by its transpose. Using (94), we obtain

$$H_j^{-1} \preceq L_\Pi (B_j^\top B_j)^{-1} \preceq C_{\mathrm{ch}} L_\Pi I_m.$$

Thus, with $\mu_1 := (2C_{\mathrm{ch}} L_\Pi)^{-1} > 0$,

$$D^2_\xi J_{j,X}(\xi_j^\star) \succeq 2\mu_1 I_m, \tag{124}$$

uniformly over all inputs in the stable tube and all active indices.

Step 4: Local approximation of Π_{OT} by a neural network. In this part, we approximate $X \mapsto \Pi_{\mathrm{OT}}(X)$ on a local cube. Because $\Pi_{\mathrm{OT}}(X)$ is defined by a minimization on M, we approximate it by using the iteration of gradient descent with a neural network approximation of the gradient.

We first derive a uniform bound for the second derivative of the objective function $J_{j,X}$ and show that the gradient descent is a contracting map. The smoothness of $\bar{\phi}$, the uniform chart bounds (94)–(95), and compactness of the tube give constants $C_H, L_1 < \infty$, independent of h, j, X, such that, for any ξ with $\|\xi\| < \rho_{\mathrm{ch}}/2$,

$$\|D^2_\xi J_{j,X}(\xi) - D^2_\xi J_{j,X}(\zeta)\|_{\mathrm{op}} \le C_H \|\xi - \zeta\|^\alpha \quad \text{and} \quad \|D^2_\xi J_{j,X}(\xi)\|_{\mathrm{op}} \le L_1. \tag{125}$$

Enlarge L_1 if necessary so that $L_1 > 2\mu_1$. For C_r in (121), choose a fixed R_r such that $R_r > 2C_r + 1$, and decrease $h_0 \in (0,1)$ if necessary so that

$$(\sqrt{m}R_r + C_r)h_0 < \rho_{\mathrm{ch}}/2, \qquad C_H\big((\sqrt{m}R_r + C_r)h_0\big)^{\alpha} \le \mu_1.$$

Then, for any $h < h_0$ and ξ in the local cube $[-R_r h, R_r h]^m$, we have $\|\xi\| \le \sqrt{m}\|\xi\|_\infty \le \sqrt{m}R_r h$ and therefore (121) gives

$$\|\xi - \xi_j^\star\| \le \|\xi\| + \|\xi_j^\star\| \le (\sqrt{m}R_r + C_r)h.$$

By the choice of the constants and (125), we have $\|D_\xi^2 J_{j,X}(\xi) - D_\xi^2 J_{j,X}(\xi_j^\star)\|_{\mathrm{op}} \le \mu_1$. Combining this with (124) yields, for $\|\xi\|_\infty \le R_r h$ and $h < h_0$,

$$\mu_1 I_m \preceq D_\xi^2 J_{j,X}(\xi) \preceq L_1 I_m. \tag{126}$$

Choose

$$\gamma := \frac{2}{L_1 + \mu_1}, \qquad q_{\mathrm{ctr}} := \frac{L_1 - \mu_1}{L_1 + \mu_1} \in (0,1). \tag{127}$$

For ξ, ζ in the local cube $[-R_r h, R_r h]^m$, define the averaged Hessian

$$\overline{H}_{j,X}(\xi,\zeta) := \int_0^1 D_\xi^2 J_{j,X}(\zeta + \theta(\xi - \zeta))\, d\theta.$$

The segment lies in the cube, so $\mu_1 I_m \preceq \overline{H}_{j,X}(\xi,\zeta) \preceq L_1 I_m$. Define the map of the gradient descent of $J_{j,X}$ by

$$T_{j,X}(\xi) := \xi - \gamma G_j(X,\xi).$$

The fundamental theorem of calculus gives

$$T_{j,X}(\xi) - T_{j,X}(\zeta) = \big(I_m - \gamma \overline{H}_{j,X}(\xi,\zeta)\big)(\xi - \zeta).$$

The choice of γ implies $\|I_m - \gamma \overline{H}_{j,X}(\xi,\zeta)\|_{\mathrm{op}} \le q_{\mathrm{ctr}}$; hence

$$\|T_{j,X}(\xi) - T_{j,X}(\zeta)\| \le q_{\mathrm{ctr}}\|\xi - \zeta\|, \tag{128}$$

which means that $T_{j,X}$ satisfies a uniform contraction estimate on the local cube. All constants in these estimates are independent of h, j, and the active input X.

Next, we approximate $\xi_j^* = Y_j^{-1}(\Pi_{\mathrm{OT}}(X))$ using a neural network that approximately realizes the gradient descent optimization. Because $s = r + \alpha$ and the map $\xi \mapsto G_j(X,\xi)$ has a uniformly bounded $C^{r,\alpha}$-norm, its Taylor expansion of order r at the origin,

$$T_{j,X}^{(r)}(\xi) := \sum_{|\beta| \le r} \frac{D_\xi^\beta G_j(X,0)}{\beta!}\xi^\beta, \tag{129}$$

satisfies

$$\sup_{\|\xi\|_\infty \le R_r h} \Big\| G_j(X,\xi) - T_{j,X}^{(r)}(\xi) \Big\| \le C h^s. \tag{130}$$

Each coefficient $D_\xi^\beta G_j(X,0)$ is affine in X. By Lemma D.24, there is a ReLU network $\widehat{G}_{j,h}$ such that it achieves approximation of order h^s, i.e.,

$$\sup_{\substack{X \in \Gamma^\star_{I,\delta_0} \\ \|\xi\|_\infty \le R_r h}} \Big\| \widehat{G}_{j,h}(X,\xi) - G_j(X,\xi) \Big\| \le \delta_h, \qquad \delta_h := c_G h^s, \tag{131}$$

where $c_G > 0$ is chosen below, and

$$\mathcal{L}(\widehat{G}_{j,h}) + \mathcal{M}(\widehat{G}_{j,h}) \le C\log(1/h). \tag{132}$$

Starting from $\widehat{\xi}_j^{(0)} = 0$, define the iteration of the approximate gradient descent by

$$\widehat{\xi}_j^{(\ell+1)} = \widehat{\xi}_j^{(\ell)} - \gamma \widehat{G}_{j,h}\big(X, \widehat{\xi}_j^{(\ell)}\big). \tag{133}$$

Let $e_\ell := \|\widehat{\xi}_j^{(\ell)} - \xi_j^\star\|$. Note that $\xi_j^\star$ is a fixed point of $T_{j,X}$. As long as the iterates remain in the local cube, (128) and (131) imply

$$e_{\ell+1} \le q_{\mathrm{ctr}} e_\ell + \gamma \delta_h.$$

Since $e_0 = \|\xi_j^\star\| \le C_r h$, induction arguments give

$$e_\ell \le q_{\mathrm{ctr}}^\ell C_r h + \frac{\gamma \delta_h}{1 - q_{\mathrm{ctr}}}. \tag{134}$$

By choosing R_r sufficiently large and then c_G in $\delta_h = c_G h^s$ sufficiently small, we can guarantee that all iterates remain in the local cube.

Let

$$K_h := \left\lceil \frac{(s-1)\log(1/h) + C_K}{|\log q_{\mathrm{ctr}}|} \right\rceil \tag{135}$$

with C_K sufficiently large. Then, $q_{\mathrm{ctr}}^{K_h} h \le C h^s$, and therefore

$$\left\| \widehat{\xi}_j^{(K_h)} - \xi_j^\star \right\| \le C h^s. \tag{136}$$

By a similar Taylor expansion of the chart Y_j on the cube, Lemma D.24 guarantees the existence of a ReLU network $\widehat{Y}_{j,h}$ that satisfies

$$\sup_{\|\xi\|_\infty \le R_r h} \left\| \widehat{Y}_{j,h}(\xi) - Y_j(\xi) \right\| \le C h^s. \tag{137}$$

Define the local expert

$$E_{j,h}(X) := \widehat{Y}_{j,h}\big(\widehat{\xi}_j^{(K_h)}(X)\big). \tag{138}$$

Equations (136) and (137) yield

$$\widetilde{\eta}_j(X) > 0 \quad \Longrightarrow \quad \|E_{j,h}(X) - y\| \le C h^s. \tag{139}$$

The number of iterations is $K_h = O(\log(1/h))$, and each approximate gradient evaluation uses $O(\log(1/h))$ nonzero parameters and layers. Consequently,

$$\mathcal{L}(E_{j,h}) + \mathcal{M}(E_{j,h}) \le C \log^2(1/h). \tag{140}$$

When $0 < s \le 1$, no iterative correction is required. We can simply set $E_{j,h}(X) := y_j$. For an active index, (118) gives

$$\|E_{j,h}(X) - y\| \le Ch \le Ch^s.$$

because $h \in (0,1)$.

Finally, for later use, each expert may be clipped coordinatewise to a fixed box $[-B, B]^d$ that contains M. Such clipping is represented exactly using ReLU and ensures a uniform output bound without increasing the approximation error on active patches.

Step 5: Partition of unity and assembly. With $\tilde{\eta}_j$ in (117), define

$$R_h(X) := \sum_{j=1}^{N_h} \widetilde{\eta}_j(X). \tag{141}$$

Since at least one index attains the minimum score, we have $R_h(X) \ge 1$. By (120), we also have $1 \le R_h(X) \le N_{\mathrm{ov}}$ on $\Gamma^\star_{I,\delta_0}$. We construct a partition of unity $\{\eta_j\}_{j=1}^{N_h}$ by

$$\eta_j(X) := \frac{\widetilde{\eta}_j(X)}{R_h(X)}, \tag{142}$$

which satisfies

$$\eta_j(X) \ge 0, \qquad \sum_{j=1}^{N_h} \eta_j(X) = 1.$$

Moreover, if $\eta_j(X) > 0$, then the local expert $E_{j,h}$ satisfies (139). Hence, the exact mixture

$$\overline{\Phi}_h(X) := \sum_{j=1}^{N_h} \eta_j(X) E_{j,h}(X) \tag{143}$$

obeys

$$\|\overline{\Phi}_h(X) - y\| \le \sum_{j=1}^{N_h} \eta_j(X) \|E_{j,h}(X) - y\| \le C h^s. \tag{144}$$

The normalization and products are realized by ReLU networks; by Lemma D.24, approximate $r \mapsto r^{-1}$ on $[1, N_{\mathrm{ov}}]$ to an accuracy of ch^s, and use zero-preserving multiplication networks to form the weights and products $\eta_j E_{j,h}$. The zero-preserving property is important: if $\widetilde{\eta}_j(X) = 0$, the corresponding product is exactly zero. Thus, approximation errors accumulate only over the at most N_{ov} active indices, rather than over all N_h indices.

We choose the arithmetic accuracy to be sufficiently small, which gives a ReLU network Φ_h to approximate $\overline{\Phi}_h(X)$. More explicitly, we choose the reciprocal and multiplication subnetworks with uniform approximation error at most $c_{\mathrm{arith}} h^s$. Since the clipped expert outputs are uniformly bounded, the arithmetic error in each active term $\eta_j(X)E_{j,h}(X)$ is at most $C_{\mathrm{arith}} h^s$, where $C_{\mathrm{arith}} > 0$ is independent of h, j, and X. By zero preservation and the overlap bound, the total assembly error is therefore at most $C_{\mathrm{arith}} N_{\mathrm{ov}} \varepsilon_h$. Taking $0 < c_{\mathrm{arith}} \le \min\{\frac{1}{4}, \frac{1}{C_{\mathrm{arith}} N_{\mathrm{ov}}}\}$ gives a ReLU network Φ_h such that

$$\sup_{X \in \Gamma^\star_{I,\delta_0}} \|\Phi_h(X) - \overline{\Phi}_h(X)\| \le C h^s. \tag{145}$$

Together with (144), this proves (107). Finally, apply a coordinatewise ReLU clipping map to the assembled output, with a fixed box containing a unit neighborhood of M. For sufficiently small h this clipping is inactive on $\Gamma^\star_{I,\delta_0}$, while it gives the global bound (109). This final operation has constant depth and size.

The N_h experts are implemented in parallel. Each requires $O(\log^2(1/h))$ nonzero parameters. The score network, minimum tree, raw gates, summation, reciprocal, and multiplication networks require at most

$$C N_h \log^2(1/h)$$

additional nonzero parameters. By (112),

$$\mathcal{M}(\Phi_h) \le C h^{-(m+1)} \log^2(1/h),$$

which proves the bound of $\mathcal{M}$ in (110). The parallel construction and the local iterations give

$$\mathcal{L}(\Phi_h) \le C \log^2(1/h).$$

The factors h^{-2} in the gates and the accuracy-dependent arithmetic networks give a polynomial parameter envelope, proving (108).

Finally, for sufficiently large P, choose

$$h_P := c_P \left(\frac{P}{\log^2 P} \right)^{-1/(m+1)} \tag{146}$$

with $c_P > 0$ sufficiently small. Then $\mathcal{M}(\Phi_{h_P}) \le P$, and

$$h_P^s \le C \left(\frac{P}{\log^2 P} \right)^{-s/(m+1)}.$$

This proves (111). □

D.4.3.4 Approximation of the population OT vector field On the exact interpolation graph,

$$u^\star(t,x) = \frac{\Pi_{\mathrm{OT}}(t,x) - x}{1-t}. \tag{147}$$

Since $t \in [0,c]$ with $c < 1$, the factor $(1-t)^{-1}$ is uniformly bounded.

Corollary D.32. *Under the assumptions of Theorem D.30, for every sufficiently large P there is a ReLU network*

$$f_P : I \times \mathbb{R}^d \to \mathbb{R}^d$$

with

$$\mathcal{M}(f_P) \le P, \qquad \mathcal{L}(f_P) \le C \log^2 P, \qquad \mathcal{B}(f_P) \le P^C,$$

such that

$$\sup_{(t,x)\in\Gamma_I^\star} \|f_P(t,x) - u^\star(t,x)\| \le C_c \left(\frac{P}{\log^2 P}\right)^{-s/(m+1)}. \tag{148}$$

Consequently,

$$\|f_P - u^\star\|_{\star,I} \le C_c \left(\frac{P}{\log^2 P}\right)^{-s/(m+1)}. \tag{149}$$

Proof. Let Φ_P be given by Theorem D.30. A one-dimensional ReLU network can approximate

$$t \mapsto \frac{1}{1-t}, \qquad t \in [0,c],$$

to the same accuracy using Lemma D.24. Combining this reciprocal network with multiplication networks and the affine map $(t,x) \mapsto \Phi_P(t,x) - x$ yields f_P.

Since $(1-t)^{-1} \le (1-c)^{-1}$, the error in (148) follows from (111). The additional arithmetic networks have only polylogarithmic complexity and therefore do not change the parameter exponent.

Finally, $\mu_t^\star$ is supported on $\{Z_t(x_0) : x_0 \in K_0\}$. Hence

$$\begin{aligned}\|f_P - u^\star\|_{\star,I}^2 &= \int_I \int \|f_P(t,x) - u^\star(t,x)\|^2 \, \mu_t^\star(dx)\, dt \\ &\le |I| \sup_{(t,x)\in\Gamma_I^\star} \|f_P(t,x) - u^\star(t,x)\|^2.\end{aligned}$$

□

Corollary D.33. *Let $a_P(B)$ be defined in Theorem D.23. Under Assumption D.25, for all sufficiently large P,*

$$a_P(B) \le C_c \left(\frac{P}{\log^2 P}\right)^{-s/(m+1)} + C_c \varepsilon_B. \tag{150}$$

Consequently, Proposition D.20 gives

$$a_P(B) \le C_c \left(\frac{P}{\log^2 P}\right)^{-s/(m+1)} + C_{c,L_S} \Delta_B^{1/2}. \tag{151}$$

Proof. Let

$$r_P := \left(\frac{P}{\log^2 P}\right)^{-s/(m+1)},$$

and let Φ_P be the ReLU network that approximates the endpoint in Theorem D.30. Draw $(X_0, X_1) \sim \pi_B$ and set

$$X_t^B := (1-t)X_0 + tX_1, \qquad X_t^\star := (1-t)X_0 + tS(X_0), \qquad D := \|X_1 - S(X_0)\|. \tag{152}$$

Then

$$X_t^B - X_t^\star = t\{X_1 - S(X_0)\}. \tag{153}$$

Recall also that δ_0 is a constant to define $\Gamma_{I,\delta_0}^\star$.

Suppose first that $\|X_t^B - X_t^\star\| = tD \le \delta_0$. Lemma D.28 and Theorem D.30 imply

$$\begin{aligned}\|\Phi_P(t, X_t^B) - S(X_0)\| &\le \|\Phi_P(t, X_t^B) - \Pi_{\mathrm{OT}}(t, X_t^B)\| + \|\Pi_{\mathrm{OT}}(t, X_t^B) - S(X_0)\| \\ &\le Cr_P + CtD.\end{aligned} \tag{154}$$

For the second term, (102) is applied from $\|X_t^B - X_t^\star\| \le \delta_0$. For $t > 0$, Proposition D.14 and the exact identity $Q_t(X_t^\star) = S(X_0)$ give

$$\|Q_t(X_t^B) - S(X_0)\| \le \frac{1}{t}\|X_t^B - X_t^\star\| \le D. \tag{155}$$

At $t = 0$ the corresponding identity is understood through the on-support definition $u^\star(0, X_0) = S(X_0) - X_0$.

Second, Suppose $tD > \delta_0$. The global clipping (109) shows $\|\Phi_P(t, X_t^B)\| \le R_\Phi$. Since (155) always holds, it follows from the compactness of $K_1 = M$ that $\|Q_t(X_t^B)\| = \|Q_t(X_t^B) - S(X_0)\| + \|S(X_0)\|$ is bounded. Therefore, we have a constant $C > 0$ such that $\|\Phi_P(t, X_t^B) - Q_t(X_t^B)\| \le C$. Since $t \le c$ and $tD > \delta_0$ implies $1 < D/c\delta_0$, we can replace C with $\frac{C}{c\delta_0}D$.

Combining the two cases yields, for almost every $t \in I$,

$$\|\Phi_P(t, X_t^B) - Q_t(X_t^B)\| \le Cr_P + CD. \tag{156}$$

Construct f_P using $(\Phi_P - x)/(1-t)$ with arithmetic accuracy $O(r_P)$, and clip its output coordinate-wise to a fixed box containing the range of $u^\star$ on $\mathcal{X}_I$. Since $(1-t)^{-1} \le (1-c)^{-1}$, the arithmetic error and (156) give

$$\|f_P(t, X_t^B) - u^\star(t, X_t^B)\| \le C_c(r_P + D) \tag{157}$$

for almost every t. Integration with respect to $dt\,\pi_B(dX_0, dX_1)$ and Minkowski's inequality give

$$\|f_P - u^\star\|_{B,I} \le C_c r_P + C_c \varepsilon_B.$$

This proves (150); the bound (151) follows from $\varepsilon_B^2 \le 2L_S\Delta_B$. □

Remark D.34 (Regularity assumptions for the learned field). The preceding approximation error bounds do not require Lipschitz constants of the comparison networks to be bounded uniformly in the network size. Moreover, these approximation results do not provide uniform Lipschitz or one-sided Lipschitz (OSL) control of the learned vector field. The regularity conditions needed for the subsequent prediction-error comparison and flow-stability analysis are imposed separately.

Remark D.35 (No second derivative of the ReLU network is used). The Hessian bounds used in the local construction concern the smooth manifold objective $j_{j,X} = J_X \circ Y_j$, not the approximating ReLU networks. The neural approximation argument requires only function-value approximation of the smooth maps involved in the construction; it does not require differentiating their ReLU realizations. Thus the piecewise-affine regularity of ReLU is compatible with the proof, and no second derivatives of the ReLU networks are used.

Remark D.36 (Role of the OT fibers). The construction never explicitly approximates the sets

$$C_y = \big(\partial\phi(y) \cap K_0\big) - \nabla_M(\phi|_M)(y) \subset N_y M.$$

In particular, it does not assume that $0 \in C_y$, nor that the subgradient multifunction admits a smooth global section. The OT fiber enters only through the fact that, for an exact OT point

$$x = (1-t)x_0 + ty, \qquad x_0 \in \partial\phi(y),$$

the point y is the uniformly stable minimizer of $\mathcal{J}_{(t,x)}$.

This is the reason that the normal dimension $d - m$ does not enter the covering number. The only mesh is placed on

$$I \times M,$$

whose dimension is $m + 1$. The ambient input $x \in \mathbb{R}^d$ enters the local stationarity equation

$$(1-t)\nabla A_j(\xi) + DY_j(\xi)^\top\{tY_j(\xi) - x\} = 0. \tag{158}$$

affinely and is processed by the same local expert for all points in the corresponding normal OT fiber.

D.5 Prediction-error comparison between expected-batch and population OT interpolations

The regression bound in Section D.4 measures the prediction error under the expected-batch interpolation distribution $\nu_{B,t}$. The flow analysis in Section D.6, however, requires the error along population OT paths, whose distribution is $\mu_t^\star$. We compare these two errors by coupling the paths through the same source point. The comparison inequalities are independent of the approximation theorem. We then use the intrinsic-dimensional approximation from Section D.4.3 to obtain the neural rate in Corollary D.42.

The quantitative comparison below uses spatial Lipschitz control of the learned field, whereas the subsequent ODE stability estimate uses one-sided Lipschitz control. Neither the regression bound nor the intrinsic approximation result provides uniform control of these learned-field quantities as P increases; the regularity assumptions are therefore stated separately.

Throughout this section, we retain the compact-support and forward-map conditions in Assumptions D.4 and D.5, together with the Bregman co-coercivity condition on the source support in Assumption D.6. Let $I = [0, c]$ as before, where $c < 1$ is fixed, and set

$$\mathcal{X}_t := \{(1-t)x_0 + tx_1 : x_0 \in K_0,\ x_1 \in K_1\}.$$

Recall the definitions in (43) and (44):

$$\Delta_B = \frac{1}{2}\left\{\mathbb{E}_B W_2^2(\widehat{P}_{0,B}, \widehat{P}_{1,B}) - W_2^2(P_0, P_1)\right\}, \qquad d_B = W_2(\overline{\pi}_B, \pi^\star).$$

Also define

$$D_{B,I}^2 := \int_I W_2^2(\nu_{B,t}, \mu_t^\star)\,\mathrm{d}t. \tag{159}$$

Lemma D.37.

$$D_{B,I} \le \min\left\{\sqrt{c}\,d_B, \sqrt{\frac{2L_S c^3}{3}}\,\Delta_B^{1/2}\right\}. \tag{160}$$

Proof. The map $F_t(x_0, x_1) = (1-t)x_0 + tx_1$ has Lipschitz constant $\sqrt{(1-t)^2 + t^2}$ for the Euclidean metric. Because $\nu_{B,t} = F_{t,\#}\overline{\pi}_B$ and $\mu_t^\star = F_{t,\#}\pi^\star$, we have

$$W_2(\nu_{B,t}, \mu_t^\star) \le \sqrt{(1-t)^2 + t^2}\,d_B.$$

Next, consider $(X_0, X_1) \sim \overline{\pi}_B$ and couple $F_t(X_0, X_1)$ with $F_t(X_0, S(X_0))$. The latter has distribution $\mu_t^\star$ because the first marginal of $\overline{\pi}_B$ is P_0. Their difference is $t\{X_1 - S(X_0)\}$, so the squared transportation cost of this coupling is $t^2\varepsilon_B^2$. Then, we have

$$W_2(\nu_{B,t}, \mu_t^\star) \le t\varepsilon_B \le t\sqrt{2L_S\Delta_B}.$$

The integral over $[0, c]$ gives the assertion. □

Lemma D.38. *Let $(X_0, X_1) \sim \overline{\pi}_B$, and define*

$$X_t^B := (1-t)X_0 + tX_1, \qquad X_t^\star := (1-t)X_0 + tS(X_0).$$

For every measurable map $h : I \times \mathbb{R}^d \to \mathbb{R}^q$,

$$\|h\|_{\star,I} \le \|h\|_{B,I} + \left(\int_I \mathbb{E}_{\overline{\pi}_B}\left[\|h(t, X_t^\star) - h(t, X_t^B)\|^2\right]\mathrm{d}t\right)^{1/2}. \tag{161}$$

Suppose, in addition, that a measurable function $L_h : I \to [0, \infty)$ satisfies

$$\|h(t, X_t^\star) - h(t, X_t^B)\| \le L_h(t)\|X_t^\star - X_t^B\| \qquad \overline{\pi}_B\text{-a.s.} \tag{162}$$

for almost every $t \in I$. Then,

$$\|h\|_{\star,I} \le \|h\|_{B,I} + \sqrt{2L_S\Delta_B}\left(\int_I t^2 L_h(t)^2\,\mathrm{d}t\right)^{1/2}. \tag{163}$$

Proof. From $X_t^B \sim \nu_{B,t}$ and $X_t^\star \sim \mu_t^\star$ with $(X_0, X_1) \sim \overline{\pi}_B$, the application of the triangular inequality in $L^2(I \times \overline{\pi}_B)$ to

$$h(t, X_t^\star) = h(t, X_t^B) + \{h(t, X_t^\star) - h(t, X_t^B)\}$$

proves (161). Next, the relation $X_t^\star - X_t^B = t\{S(X_0) - X_1\}$ and (162) give

$$\left(\int_I \mathbb{E}_{\overline{\pi}_B} \|h(t, X_t^\star) - h(t, X_t^B)\|^2 \, \mathrm{d}t\right)^{1/2}$$
$$\leq \left(\int_I t^2 L_h(t)^2 \mathbb{E}_{\overline{\pi}_B} \|X_1 - S(X_0)\|^2 \, \mathrm{d}t\right)^{1/2} = \varepsilon_B \left(\int_I t^2 L_h(t)^2 \, \mathrm{d}t\right)^{1/2}.$$

This proves (163) from (71). □

Theorem D.39. *Suppose Assumption D.21 holds, and let $\widehat{u}_{B,L,P} \in \mathcal{F}_P$ satisfy the approximate empirical minimization condition* (81)*, with $L \geq 2$. Set $h_{B,L,P} := \widehat{u}_{B,L,P} - u^\star$ and define the common-source path-comparison term*

$$\mathfrak{T}_{B,L,P} := \left(\int_I \mathbb{E}_{\overline{\pi}_B}\left[\|h_{B,L,P}(t, X_t^\star) - h_{B,L,P}(t, X_t^B)\|^2\right] \mathrm{d}t\right)^{1/2}. \tag{164}$$

Here the evaluation pair $(X_0, X_1) \sim \overline{\pi}_B$ is independent of the training randomness, and the expectation is taken with the learned network held fixed. For every $\eta \in (0,1)$, with probability at least $1 - \eta$ over the training batches,

$$\begin{aligned}\|\widehat{u}_{B,L,P} - u^\star\|_{\star,I} \leq\ & C a_P(B) + C\sqrt{\kappa_{S,I}\Delta_B} \\ & + CK_I\sqrt{\frac{P\log^2(P+1)\log(C_0 A_P P L) + \log(2/\eta)}{L}} \\ & + C\sqrt{\varepsilon_{\mathrm{opt}}} + \mathfrak{T}_{B,L,P}.\end{aligned} \tag{165}$$

The constants are independent of B, L, P, η.

Proof. For each training realization, apply Lemma D.38 with $h = \widehat{u}_{B,L,P} - u^\star$ to obtain

$$\|\widehat{u}_{B,L,P} - u^\star\|_{\star,I} \leq \|\widehat{u}_{B,L,P} - u^\star\|_{B,I} + \mathfrak{T}_{B,L,P}.$$

Theorem D.23 bounds the first term of the R.H.S. on the same training event of probability at least $1 - \eta$. □

Theorem D.40. *Suppose Assumption D.21 holds. Assume further that there is a deterministic measurable function $L_P : I \to [0, \infty)$ such that, for almost every $t \in I$,*

$$\|\widehat{u}_{B,L,P}(t,x) - \widehat{u}_{B,L,P}(t,y)\| \leq L_P(t)\|x - y\|, \qquad x, y \in \mathcal{X}_t \tag{166}$$

holds for every realization of the training procedure. Define

$$\Lambda_P^{\mathrm{tr}}(I) := \left(\int_0^c t^2\{L_P(t) + \ell_t\}^2 \, \mathrm{d}t\right)^{1/2}, \qquad \ell_t := \max\left\{\frac{1}{t}, \frac{1}{1-t}\right\} \qquad (t > 0). \tag{167}$$

If $\Lambda_P^{\mathrm{tr}}(I) < \infty$, then, for every $\eta \in (0,1)$, with probability at least $1 - \eta$,

$$\begin{aligned}\|\widehat{u}_{B,L,P} - u^\star\|_{\star,I} \leq\ & C a_P(B) + C\sqrt{\kappa_{S,I}\Delta_B} + CK_I\sqrt{\frac{P\log^2(P+1)\log(C_0 A_P P L) + \log(2/\eta)}{L}} \\ & + C\sqrt{\varepsilon_{\mathrm{opt}}} + \sqrt{2L_S\Delta_B}\,\Lambda_P^{\mathrm{tr}}(I).\end{aligned} \tag{168}$$

Proof. For $h = \widehat{u}_{B,L,P} - u^\star$, condition (166) and Proposition D.14 imply

$$\|h(t,x) - h(t,y)\| \leq \{L_P(t) + \ell_t\}\|x - y\|, \qquad x, y \in \mathcal{X}_t,$$

for almost every $t > 0$. Both common-source interpolation points belong to $\mathcal{X}_t$. Therefore Lemma D.38 gives

$$\mathfrak{T}_{B,L,P} \leq \varepsilon_B \Lambda_P^{\mathrm{tr}}(I) \leq \sqrt{2L_S\Delta_B}\,\Lambda_P^{\mathrm{tr}}(I).$$

Substitute this estimate into (165). The value of ℓ_t at $t = 0$ is irrelevant to the integral. In particular, $t\ell_t$ remains bounded on $(0, c]$, so no positive initial-time cutoff is needed. □

Remark D.41. Condition (166) concerns the learned neural network. Bounds on weights or spectral normalization can provide such Lipschitz control. Restricting the network class by a Lipschitz, spectral, or OSL constraint cannot increase its covering entropy. If training is instead restricted to a smaller Lipschitz-constrained class, the oracle inequality can use the approximation error of that smaller class. The intrinsic approximation bound may not be carried over to such a restriction, however. To the best of our knowledge, including the Lipschitz condition to the approximation class remain open. Although [44] and [24] use a partition of the time interval to avoid the effect of Lipschitz constants in bounding the flow errors, the technique requires separate training of many neural networks depending on the time domains, which is not realistic in practical learning.

Corollary D.42. *Suppose that Assumptions D.21, D.25, and the Lipschitz condition* (166) *hold. Then, for every sufficiently large P, every $L \geq 2$, and every $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$\begin{aligned}\|\widehat{u}_{B,L,P} - u^\star\|_{\star,I} \leq C\Delta_B^{1/2} &+ C\left(\frac{P}{\log^2(P+1)}\right)^{-s/(m+1)} \\ &+ CK_I\sqrt{\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}} \\ &+ C\sqrt{\varepsilon_{\mathrm{opt}}} + \sqrt{2L_S\Delta_B}\,\Lambda_P^{\mathrm{tr}}(I).\end{aligned} \tag{169}$$

The constant C may depend on the fixed geometric and model parameters, including c, L_S, d, m, s, but not on B, L, P, η.

Moreover, by the choice $P_L^{\mathrm{int}} \asymp (\log(eL))^2\left(\frac{L}{(\log(eL))^5}\right)^{(m+1)/(2s+m+1)}$, for all sufficiently large L, with probability at least $1-\eta$,

$$\begin{aligned}\|\widehat{u}_{B,L,P_L^{\mathrm{int}}} - u^\star\|_{\star,I} \leq C\Delta_B^{1/2} &+ C\left(\frac{\log^5(eL)}{L}\right)^{s/(2s+m+1)} \\ &+ C\sqrt{\frac{\log(2/\eta)}{L}} + C\sqrt{\varepsilon_{\mathrm{opt}}} + \sqrt{2L_S\Delta_B}\,\Lambda_{P_L^{\mathrm{int}}}^{\mathrm{tr}}(I).\end{aligned} \tag{170}$$

In particular, the neural approximation-estimation contribution is $\widetilde{O}(L^{-s/(2s+m+1)})$.

Proof. Corollary D.33 and (168) prove (169).

Let $H_L := \log(eL)$. For the parameter choice, put $k := m+1$ and $N_L := (L/H_L^5)^{k/(2s+k)}$. Then $P_L^{\mathrm{int}} \asymp H_L^2 N_L$ and $\log(P_L^{\mathrm{int}}+1) \asymp H_L$. Consequently,

$$\left(\frac{P_L^{\mathrm{int}}}{\log^2(P_L^{\mathrm{int}}+1)}\right)^{-s/k} \asymp N_L^{-s/k} = \left(\frac{H_L^5}{L}\right)^{s/(2s+k)}.$$

The polynomial weight envelope and the fact that P_L^{int} is polynomial in L up to logarithmic factors give

$$\log(C_0A_{P_L^{\mathrm{int}}}P_L^{\mathrm{int}}L) \leq CH_L.$$

Thus the non-confidence part of the root estimation error is at most

$$C\left(\frac{P_L^{\mathrm{int}}H_L^3}{L}\right)^{1/2} \asymp C\left(\frac{N_LH_L^5}{L}\right)^{1/2} = C\left(\frac{H_L^5}{L}\right)^{s/(2s+k)}.$$

Splitting off the confidence term proves (170). □

The exponent $m+1$ comes from the structured approximation in Section D.4.3. The logarithmic factor accounts for the $\log^2 P$ approximation overhead in (151). The power 5 is the sufficient logarithmic factor supplied by the present approximation and entropy bounds; no optimality of that factor is claimed. To absorb the last term of (170) into $C\Delta_B^{1/2}$, one still needs a uniform bound on $\Lambda_{P_L^{\mathrm{int}}}^{\mathrm{tr}}(I)$, which is relevant to the Lipschitz condition.

D.6 Flow convergence rate

Let $\mathcal{F}_{\mathrm{tr}}$ be the sigma-field generated by the random batches for training, sampled time variables, and any algorithmic randomness. Independently of $\mathcal{F}_{\mathrm{tr}}$, draw $(X_0, X_1) \sim \pi^\star$ and write $\mathbb{E}_\star$ for expectation only over this evaluation pair. Define

$$X_t^\star = (1-t)X_0 + tX_1,$$

and let the ODE for generation

$$\frac{d\widehat{X}_t}{dt} = \widehat{u}_{B,L,P}(t, \widehat{X}_t), \qquad \widehat{X}_0 = X_0^\star = X_0. \tag{171}$$

D.6.1 From vector field bound to flow bound

Assumption D.43 (One-sided Lipschitz (OSL) condition for the learned field)**.** The map $t \mapsto \rho_{\widehat{u}_{B,L,P}}(t)$ in (54) is measurable and

$$R^{\mathrm{OSL}}_{B,L,P}(I) := \int_I (\rho_{\widehat{u}_{B,L,P}}(t))_+ \,\mathrm{d}t < \infty. \tag{172}$$

Note that, if the global spatial Lipschitz condition on $\mathbb{R}^d$

$$\sup_{t\in I} \mathrm{Lip}_x\, \widehat{u}_{B,L,P}(t,\cdot) \le L_P \tag{173}$$

holds, then the OSL holds:

$$R^{\mathrm{OSL}}_{B,L,P}(I) \le cL_P. \tag{174}$$

Thus, when Sections D.4–D.6 are combined, the OSL condition is not an additional regularity requirement if the ordinary Lipschitz bound is imposed pathwise. It is stated separately because it is strictly weaker and can yield a sharper stability factor. In contrast, (166) only controls the field on $\mathcal{X}_t$ and does not by itself imply the global OSL condition above.

Theorem D.44. *Under Assumption D.43, conditionally on $\mathcal{F}_{\mathrm{tr}}$,*

$$\left(\mathbb{E}_\star \sup_{t\in I} \|\widehat{X}_t - X_t^\star\|^2\right)^{1/2} \le e^{R^{\mathrm{OSL}}_{B,L,P}(I)} \sqrt{c}\, \|\widehat{u}_{B,L,P} - u^\star\|_{\star,I}. \tag{175}$$

Let $\widehat{\mu}_c := \mathrm{Law}_\star(\widehat{X}_c \mid \mathcal{F}_{\mathrm{tr}})$ denote the conditional distribution of the generated state $\widehat{X}_c$ given the training sigma-field $\mathcal{F}_{\mathrm{tr}}$. Then

$$W_2(\widehat{\mu}_c, \mu_c^\star) \le \left(\mathbb{E}_\star \|\widehat{X}_c - X_c^\star\|^2\right)^{1/2} \tag{176}$$

and is bounded by the right-hand side of (175)*.*

Proof. Let $e(t) = \|\widehat{X}_t - X_t^\star\|$. Using the OSL coefficient of the learned field,

$$D^+ e(t) \le \rho_{\widehat{u}_{B,L,P}}(t) e(t) + \|\widehat{u}_{B,L,P}(t, X_t^\star) - u^\star(t, X_t^\star)\|, \qquad e(0) = 0.$$

Here D^+ denotes the upper right Dini derivative. Variation of constants and (172) give

$$e(t) \le e^{R^{\mathrm{OSL}}_{B,L,P}(I)} \int_0^t \|\widehat{u}_{B,L,P}(t, X_t^\star) - u^\star(t, X_t^\star)\| \,\mathrm{d}s.$$

Taking the supremum and applying Cauchy–Schwarz yields

$$\sup_{t\in I} e(t)^2 \le e^{2R^{\mathrm{OSL}}_{B,L,P}(I)} c \int_I \|\widehat{u}_{B,L,P}(t, X_t^\star) - u^\star(t, X_t^\star)\|^2 \,\mathrm{d}s.$$

Taking $\mathbb{E}_\star$ completes the proof. □

Corollary D.45 (Finite-sample end-to-end bound on I)**.** *Suppose the assumptions of Theorems D.40 and D.44 hold and*

$$R^{\mathrm{OSL}}_{B,L,P}(I) \le R \tag{177}$$

for a deterministic $R < \infty$. Then, with probability at least $1-\eta$,

$$\begin{aligned}
&\left(\mathbb{E}_\star \sup_{t\in I} \|\widehat{X}_t - X^\star_t\|^2\right)^{1/2} \\
&\quad \le C_{c,R}\Bigg[a_P(B) + \sqrt{\kappa_{S,I}\Delta_B} + K_I\sqrt{\frac{P\log^2(P+1)\log(C_0A_PPL)+\log(2/\eta)}{L}} + \sqrt{\varepsilon_{\mathrm{opt}}} \\
&\qquad\qquad + \sqrt{2L_S\Delta_B}\,\Lambda^{\mathrm{tr}}_P(I)\Bigg],
\end{aligned} \tag{178}$$

where $C_{c,R} = Ce^R\sqrt{c}$.

D.6.2 Quantitative batch-size rates

In this subsection, we discuss the convergence rate of empirical OT of batch size B to the population one, that is, the rate of Δ_B.

Define

$$\Psi_m(B) := \begin{cases} B^{-1/2}, & m<4, \\ B^{-1/2}\log(B+1), & m=4, \\ B^{-2/m}, & m>4. \end{cases} \tag{179}$$

When the compact-support condition of Assumption D.21 holds, lower-complexity adaptation results imply, under the corresponding metric-entropy hypotheses, that when one marginal is supported on an m-dimensional regular compact manifold,

$$\mathbb{E}\left|W_2^2(\widehat{P}_{0,B},\widehat{P}_{1,B}) - W_2^2(P_0,P_1)\right| \le C\Psi_m(B). \tag{180}$$

See [28]. Hence

$$\Delta_B \le C\Psi_m(B). \tag{181}$$

Under Assumption D.5, the batch contribution to the root field and strong-flow errors on I is therefore

$$O(\Psi_m(B)^{1/2}) = O\left(\begin{cases} B^{-1/4}, & m<4, \\ B^{-1/4}\sqrt{\log(B+1)}, & m=4, \\ B^{-1/m}, & m>4. \end{cases}\right). \tag{182}$$

Remark D.46. Extensions of empirical OT-cost rates to unbounded distributions require explicit moment and tail assumptions and may contain additional tail terms; see [57].

Remark D.47. The rate above is a corollary within the standing globally regular forward-map regime (Assumption D.5). Singular transports such as Gaussian-to-circle (Remark D.12) and Gaussian-to-discrete OT do not satisfy this assumption and are not asserted to obey the present initial-time field or flow rate.

D.6.3 Main result: Explicit intrinsic-dimensional rates of flow

We now combine the structured approximation in Section D.4.3, the prediction-error comparison in Corollary D.42, and the flow stability estimate in Theorem D.44. In this section, $m = \dim M$ and $I = [0,c]$ with $c<1$ fixed. The resulting neural rate depends on the intrinsic space–time dimension $m+1$. As in Section D.6, $\mathbb{E}_\star$ denotes expectation over the independent population evaluation pair, with the training realization held fixed.

Corollary D.48 (Main result)**.** *Suppose that Assumptions D.21, D.25, D.43, and the Lipschitz condition* (166) *hold. Choose the integer parameter budget $P = P_L^{\mathrm{int}}$ as*

$$P_L^{\mathrm{int}} \asymp \log^2(eL)\left(\frac{L}{\log^5(eL)}\right)^{(m+1)/(2s+m+1)}. \tag{183}$$

Suppose further that, uniformly in the asymptotic regime and for every training realization,

$$\Lambda^{\mathrm{tr}}_{P_L^{\mathrm{int}}}(I) \le \Lambda, \qquad R^{\mathrm{OSL}}_{B,L,P_L^{\mathrm{int}}}(I) \le R \tag{184}$$

for deterministic constants $\Lambda, R < \infty$. Then, for all sufficiently large L and every $\eta \in (0,1)$, with probability at least $1-\eta$ over the training batches,

$$\left(\mathbb{E}_\star \sup_{t\in I} \|\widehat{X}_t - X^\star_t\|^2\right)^{1/2} \le C\left[\Psi_m(B)^{1/2} + \left(\frac{\log^5(eL)}{L}\right)^{s/(2s+m+1)} + \sqrt{\frac{\log(2/\eta)}{L}} + \sqrt{\varepsilon_{\mathrm{opt}}}\right], \tag{185}$$

where Ψ_m is defined in (179)*. The constant C may depend on the fixed geometric parameter and network architecture, including d, m, s, c, L_S, and on Λ, R, but not on B, L, η. The Wasserstein error $W_2(\widehat{\mu}_c, \mu^\star_c)$ satisfies the same upper bound. Thus the neural approximation–estimation contribution to the root strong-flow error is $\widetilde{O}(L^{-s/(2s+m+1)})$.*

Proof. The approximation $a_P(B)$, already incorporated into Corollary D.42, is bounded as

$$a_P(B) \le C\left(\frac{P}{\log^2(P+1)}\right)^{-s/(m+1)} + C\Delta_B^{1/2}.$$

For $P = P_L^{\mathrm{int}}$, on a training event of probability at least $1-\eta$, combining the field estimate (170), Theorem D.44, and the assumption (184) proves (185). The Wasserstein conclusion follows from (176). □

In Corollary D.48, the intrinsic manifold dimension m enters through two separate estimates: $\Psi_m(B)$ controls the empirical-OT cost bias, while $s/(2s+m+1)$ comes from structured neural approximation and estimation. Manifold support is not sufficient for the latter; the smooth manifold proximal geometry by OT is essential in the approximation arguments in Section D.4.3.

Remark D.49 (Explicit dependence on the learned-field budgets). Without the uniform bounds in (184), but retaining the learned-field Lipschitz condition and the ODE solvability and finite-OSL assumptions, combine (169) with (175). For every sufficiently large P, this gives, on the same training event of probability at least $1-\eta$,

$$\left(\mathbb{E}_\star \sup_{t\in I} \|\widehat{X}_t - X^\star_t\|^2\right)^{1/2} \le Ce^{R^{\mathrm{OSL}}_{B,L,P}(I)}\sqrt{c}\left[\Psi_m(B)^{1/2} + \left(\frac{P}{\log^2(P+1)}\right)^{-s/(m+1)} + K_I\sqrt{\frac{P\log^2(P+1)\log(C_0A_PPL) + \log(2/\eta)}{L}} + \sqrt{\varepsilon_{\mathrm{opt}}} + \sqrt{2L_S\Delta_B}\,\Lambda^{\mathrm{tr}}_P(I)\right]. \tag{186}$$

Thus growth of $\Lambda^{\mathrm{tr}}_P(I)$ affects the common-source path-comparison correction, and growth of $R^{\mathrm{OSL}}_{B,L,P}(I)$ affects the entire flow bound exponentially. Moreover, imposing a smaller Lipschitz-constrained regression class requires verifying the approximation bound within that class, as discussed in Remark D.41.

All rates in this section concern a fixed $c < 1$ and comparison with $\mu^\star_c$, not a quantitative terminal-time guarantee at $t = 1$. No uniformity of the constants as $c \uparrow 1$ is asserted; the terminal extension requires the additional care discussed in Section D.6.4 because of the divergence of the velocity $u^\star(x_t, t)$.

D.6.4 Conditional extension to the terminal time

The spatial Lipschitz bound in Proposition D.14 grows like $(1-t)^{-1}$ as $t \uparrow 1$. Accordingly, the preceding estimates are stated on $I = [0, c]$ for a fixed $c < 1$, without asserting uniformity of their constants as $c \uparrow 1$. Nevertheless, the one-sided Lipschitz stability argument extends to the whole interval $[0, 1]$ under the additional integrability and terminal-limit assumptions stated below.

Theorem D.50. *Assume*

$$\int_0^1 (\rho_{\widehat{u}}(t))_+ \,\mathrm{d}t \le R, \tag{187}$$

that the learned and exact trajectories admit limits at $t = 1$, and that

$$\int_0^1 \mathbb{E}_\star \|\widehat{u}(t, X_t^\star) - u^\star(t, X_t^\star)\|^2 \, \mathrm{d}t < \infty.$$

Then

$$\left(\mathbb{E}_\star \sup_{0 \le t \le 1} \|\widehat{X}_t - X_t^\star\|^2\right)^{1/2} \le e^R \left(\int_0^1 \mathbb{E}_\star \|\widehat{u}(t, X_t^\star) - u^\star(t, X_t^\star)\|^2 \, \mathrm{d}t\right)^{1/2}. \tag{188}$$

Thus a quantitative rate of the full-interval vector field is inherited by the endpoint map with the same exponent.

Proof. Apply Theorem D.44 on $I = [0, c]$ and let $c \uparrow 1$, using monotone convergence and the assumed terminal limits. □

D.7 Summary of the quantitative conclusions

We summarize the intrinsic-dimensional conclusion obtained by combining Section D.4.3, Corollary D.42, and Theorem D.44. In addition to the standing compact-support, forward-map, and source-support Bregman co-coercivity conditions, this conclusion requires the smooth manifold proximal geometry of Assumption D.25. We also assume the learned-field Lipschitz condition (166) and the ODE solvability and OSL conditions of Theorem D.44. Fix $I = [0, c]$ with $c < 1$, and write

$$\mathcal{E}^{\mathrm{flow}}_{B,L,P}(I) := \left(\mathbb{E}_\star \sup_{t \in I} \|\widehat{X}_t - X_t^\star\|^2\right)^{1/2}.$$

The expectation is over the independent population evaluation pair, conditionally on the training randomness, as in Section D.6.

Intrinsic approximation and finite-sample error. The essential approximation input is (151):

$$a_P(B) \le C \left(\frac{P}{\log^2(P+1)}\right)^{-s/(m+1)} + C\Delta_B^{1/2}, \qquad m = \dim M.$$

This bound is under the expected-batch interpolation distribution; its $\Delta_B^{1/2}$ term accounts for the perturbation away from population OT paths. It does not require a global Lipschitz bound on the comparison network.

Combining (169) with (175) gives, on the same training event of probability at least $1 - \eta$,

$$\begin{aligned} \mathcal{E}^{\mathrm{flow}}_{B,L,P}(I) \le C e^{R^{\mathrm{OSL}}_{B,L,P}(I)} \sqrt{c} \Bigg[& \Delta_B^{1/2} + \left(\frac{P}{\log^2(P+1)}\right)^{-s/(m+1)} \\ & + K_I \sqrt{\frac{P \log^2(P+1) \log(C_0 A_P P L) + \log(2/\eta)}{L}} \\ & + \sqrt{\varepsilon_{\mathrm{opt}}} + \sqrt{2 L_S \Delta_B}\, \Lambda_P^{\mathrm{tr}}(I) \Bigg]. \end{aligned} \tag{189}$$

In particular, if the deterministic bounds

$$\Lambda_P^{\mathrm{tr}}(I) \le \Lambda, \qquad R^{\mathrm{OSL}}_{B,L,P}(I) \le R \tag{190}$$

hold uniformly in the asymptotic regime for every training realization, then the root strong-flow error has the form

$$\begin{aligned} \mathcal{E}^{\mathrm{flow}}_{B,L,P}(I) \lesssim \; & \underbrace{\Delta_B^{1/2}}_{\text{expected-batch OT error}} + \underbrace{\left(P/\log^2(P+1)\right)^{-s/(m+1)}}_{\text{intrinsic ReLU approximation}} \\ & + \underbrace{K_I \sqrt{\frac{P \log^2(P+1) \log(C_0 A_P P L) + \log(2/\eta)}{L}}}_{\text{estimation from } L \text{ independent training batches}} \\ & + \underbrace{\sqrt{\varepsilon_{\mathrm{opt}}}}_{\text{optimization}}. \end{aligned} \tag{191}$$

The hidden constants may depend on the fixed geometry, d, m, s, c, L_S, the sieve constants, Λ, and R, but not on B, L, P, η. Thus the algebraic approximation exponent is governed by the intrinsic space–time dimension $m+1$, although constants may depend on the ambient dimension d. The effective sample count in this argument is L, not LB, because independence is used between training batches.

Optimized neural and batch-size rates. For the intrinsic parameter choice (183), namely

$$P_L^{\mathrm{int}} \asymp \log^2(eL)\left(\frac{L}{\log^5(eL)}\right)^{(m+1)/(2s+m+1)},$$

conditions (190) yield

$$\mathcal{E}^{\mathrm{flow}}_{B,L,P_L^{\mathrm{int}}}(I) \le C\left[\Delta_B^{1/2} + \left(\frac{\log^5(eL)}{L}\right)^{s/(2s+m+1)} + \sqrt{\frac{\log(2/\eta)}{L}} + \sqrt{\varepsilon_{\mathrm{opt}}}\right] \tag{192}$$

with probability at least $1-\eta$. Hence the neural contribution is $\widetilde{O}(L^{-s/(2s+m+1)})$. Under the empirical-cost hypotheses in Section D.6.2, the additional bound (181) gives

$$\mathcal{E}^{\mathrm{flow}}_{B,L,P_L^{\mathrm{int}}}(I) \le C\left[\Psi_m(B)^{1/2} + \left(\frac{\log^5(eL)}{L}\right)^{s/(2s+m+1)} + \sqrt{\frac{\log(2/\eta)}{L}} + \sqrt{\varepsilon_{\mathrm{opt}}}\right], \tag{193}$$

where Ψ_m is defined in (179). Thus intrinsic dimension enters both the empirical-OT contribution and the structured neural approximation contribution, through distinct arguments. The Wasserstein error $W_2(\widehat{\mu}_c, \mu_c^\star)$ obeys the same upper bounds by (176).

Scope of the conclusion. The result above uses Assumption D.25; manifold support alone does not justify replacing d by m in a generic smooth-function approximation theorem. Moreover, neither the approximation theorem nor the polynomial weight envelope supplies the uniform learned-field budgets in (190). If those budgets grow, their dependence must remain explicit as in (189).

A quantitative endpoint rate at $t=1$ requires additional full-interval error estimates and the conditions of Theorem D.50, which are not established by the preceding fixed-c analysis.

E Additional experiments and detailed experimental settings

E.1 Toy examples

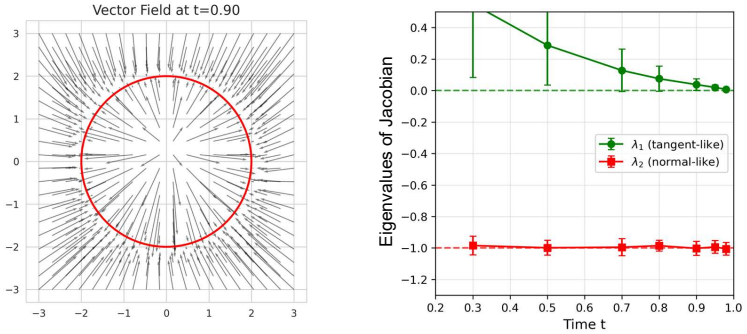
We include an additional toy example of "circle", which consists of the uniform distribution on $\mathbb{S}^1$. The experimental setting is the same as "Two Moons" in Sec. 6. We also replicate the example of "Two Moons" shown there. Figure 8 shows the results. In a similar way to the result for "Two Moons" in Sec. 6, the trained vector field around the target manifold are normal to the manifold. The two eigenvalues of $D_x v_t(x_t)$ for different t are clearly separated with $\lambda_1 = 0$ and $\lambda_2 = -\gamma = -1.0$ for $t \to 1$. The results on the circle date agrees better with Theorem 3.2 than "Two Moons", since the target distribution is a manifold rigorously.

E.2 Experimental results of Minibatch OT

We trained OT-CFM on the circle dataset (Sec. E.1) with minibatch sizes $B \in \{16, 64, 128, 256, 512, 1024\}$ (3 seeds each, 30k iterations) and computed the rescaled eigenvalues $(1-t)\lambda_i$ of the Jacobian $D_x v_t$ along the time trajectory $t \in [0.3, 1.0]$.

Fig. 9 clearly shows that larger batch sizes produce eigenvalues that follow more closely the theoretical values of Theorem 3.2 for the population OT. Interestingly, even with a small batch size, the eigenvalues do not differ significantly from the theoretical ones. In the left panel, the engenvalues converge to $\hat{\lambda}_1 = 0$ as $t \to 1$ for all batch sizes, with larger batches converging more smoothly. In the right panel, larger batches (e.g., $B = 512, 1024$) closely track the theoretical value $-\gamma = -1.0$ across $t \in [0.3, 0.9]$, while smaller batches (e.g., $B = 16$) deviate more, underestimating the normal contraction rate. This provides direct empirical support for the convergence of the minibatch OT-CFM and confirms that our theoretical analysis gives a reasonable approximation for minibatch OT.

(a) Circle



(b) Two Moons

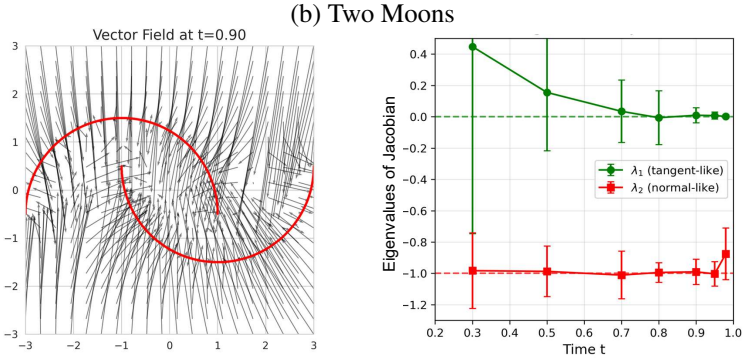


Figure 8: Left: OT-CFM vector field at time $t = 0.9$ obtained by a Neural Network, which appears to be an attracting force on the manifold. Right: Mean and STD of the eigenvalues of the Jacobian $(1-t)Dv_t$ over different trajectories. The second eigenvalue corresponds to a terminal Lyapounov exponent close to -1, as predicted by the analysis in Sec. 3.

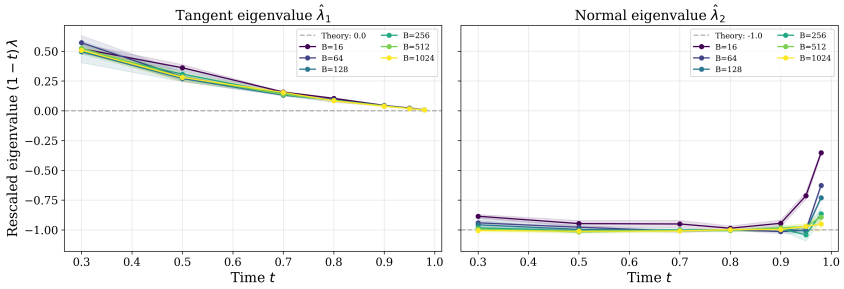


Figure 9: Eigenvalues of Jacobian of the vector field trained with minibatch OT-CFM.

E.3 Additional example: stability of data manifold

This subsection includes another example with the digit “7" in addition to the MNIST perturbation experiment in Sec.6. The results for digit “7" (Figure 10) demonstrate properties similar to those of case “4". The dimension of the larger eigenvalues is around 100 again.

E.4 Robustness of the data manifold to learning errors

The implication of our results in Sec.3 is that any small perturbation of the initial or intermediate point will be reduced to movement tangent to the data manifold at the terminal time. We demonstrate this theoretical argument with the MNIST dataset in Figure 11, where eight random perturbations $x_0^{(k)} = x_0 + \epsilon\delta_k$ ($\epsilon = 2.0$, δ_k unit-norm random direction in $\mathbb{R}^{784}$) are applied at $t = 0$ and evolved to $t = 0.99$ via the learned ODE. All perturbed outputs remain valid MNIST digits, and the displacement maps $\Delta x_1 = x_1' - x_1$ exhibit structured, edge-localized patterns corresponding to tangential deformations, confirming that the normal components of the initial perturbation are

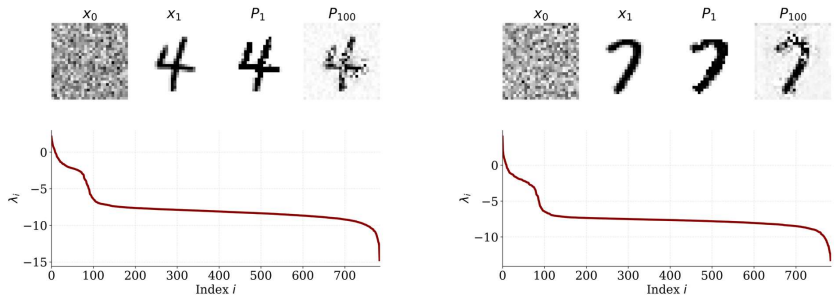


Figure 10: Perturbed MNIST images with eigenvectors

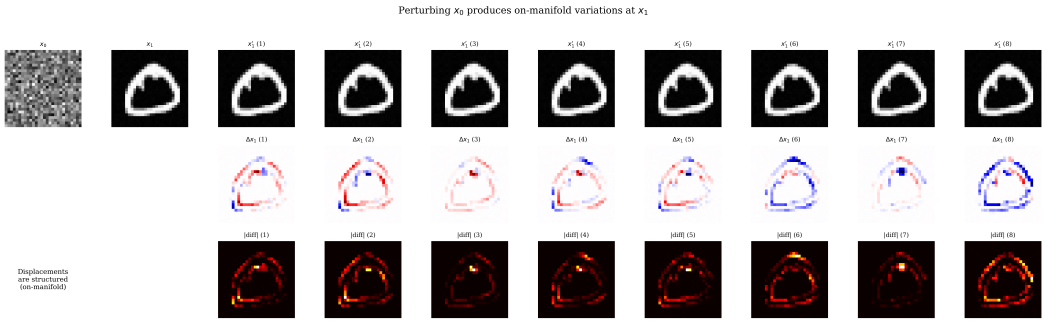


Figure 11: Generated images with perturbed initial points. The top rows show the generated images, and the bottom rows shows the displacement $\Delta x_1 = x_1' - x_1$.

contracted by the flow. Quantitatively, PCA on the displacement matrix (Figure 12) shows that the effective rank of Δx_1 is only 24 (at 95% cumulative variance) out of 784 ambient dimensions when disturbed at $t = 0$. On the other hand, when the perturbation is applied at $t = 0.75$, the effective rank is 181 because less contraction time remains, directly reflecting the exponential normal contraction predicted by Theorem 3.2.

E.5 CIFAR-10 perturbation experiments

This subsection provides further details and additional examples for the CIFAR-10 experiment summarized in Sec. 6. Each CIFAR-10 image is represented as a vector in $\mathbb{R}^{3\times32\times32} \simeq \mathbb{R}^{3072}$, and its pixel values are scaled to $[-1, 1]$. The source distribution is $P_0 = \mathcal{N}(0, I_{3072})$, and we use the affine interpolation $x_t = (1-t)x_0 + tx_1$. At every training iteration, batches of size $B = 1024$ are drawn from the source and target distributions. Their coupling is approximated by the Sinkhorn

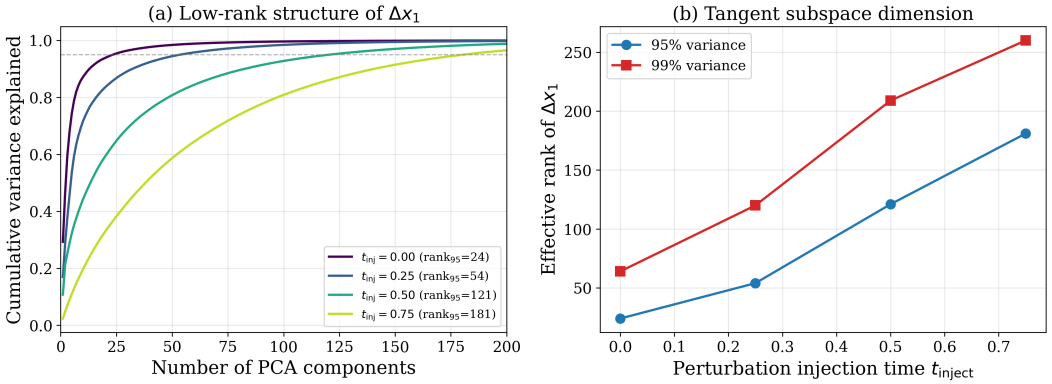


Figure 12: Results of the PCA for the displacement maps.

algorithm with entropic regularization parameter 0.01, 200 Sinkhorn iterations, and sampling from the resulting transport plan.

The vector field is represented by a time-conditioned U-Net with approximately 57.8 million parameters. The base channel width is 128, and residual blocks and self-attention layers are used at the 16×16, 8×8, and 4×4 resolutions. We use dropout 0.1 and a sinusoidal time embedding applied to $1000t$. The model is trained for 2000 epochs using AdamW, with the learning rate decreased from 2×10^{-4} to 10^{-6} by cosine annealing. Training uses BF16 arithmetic and an exponential moving average with decay 0.999. The generated samples below use the EMA parameters and a fixed-step Heun solver up to $t_{\mathrm{end}} = 0.999$.

Endpoint sensitivity with respect to the initial point. Let

$$\widehat{\Phi}(x_0) := \widehat{\Phi}_{t_{\mathrm{end}},0}(x_0)$$

denote the endpoint map induced by the learned vector field and the numerical ODE solver. To directly examine the influence of perturbations to x_0, we consider its Jacobian

$$J_\Phi(x_0) := D_{x_0}\widehat{\Phi}(x_0) \in \mathbb{R}^{3072\times 3072}.$$

Rather than explicitly forming this matrix, we use Jacobian–vector and vector–Jacobian products and approximate the leading singular values and right singular vectors of the endpoint-map Jacobian using randomized SVD:

$$J_\Phi(x_0) \simeq U\,\mathrm{diag}(\sigma_1,\dots,\sigma_r)V^\top, \qquad \sigma_1 \geq \cdots \geq \sigma_r.$$

The columns v_k of V are directions in the initial Gaussian space, while $\sigma_k u_k$ describes the corresponding first-order displacement at the endpoint:

$$\widehat{\Phi}(x_0 + \varepsilon v_k) - \widehat{\Phi}(x_0) = \varepsilon\sigma_k u_k + o(\varepsilon).$$

Thus, the directions shown below are right singular directions of the endpoint flow map.

For both random and singular-direction perturbations, we use a root-mean square perturbation of 0.01 per coordinate, that is, the Euclidean magnitude of the perturbation is $\varepsilon = 0.01\sqrt{3072} \simeq 0.55$

Independent random perturbations. We first draw independent isotropic directions

$$r_j \sim \mathcal{N}(0, I_{3072}), \qquad \bar{r}_j = \frac{r_j}{\|r_j\|},$$

and generate

$$x_1^{(j)} = \widehat{\Phi}(x_0 + \varepsilon\bar{r}_j).$$

The left panel of Figure 13 shows generated images $x_1^{(j)}$ for two different x_0, for which $\widehat{\Phi}(x_0)$ are airplane and automobile. The generated images remain almost unchanged under the independent random perturbations.

This behavior has a natural high-dimensional interpretation. If the local tangent space has dimension m and r is uniform on the unit sphere in $\mathbb{R}^d$, then

$$\mathbb{E}\|\Pi_T r\|^2 = \frac{m}{d}, \qquad \mathbb{E}\|\Pi_N r\|^2 = \frac{d-m}{d}.$$

For CIFAR-10, the ambient dimensionality is $d = 3072$, but the dimensionality of the image space is considered to be much smaller; various works have given different estimates, which are usually less than 100 (see, e.g., [49]). Consequently, most of the energy of a generic random direction tends to lie in the normal component. The almost invariance in The result in Fig. 13 is therefore consistent with the terminal contraction of normal perturbations established in Theorem 3.2.

Perturbations along dominant flow-map directions. We next perturb x_0 along selected right singular vectors:

$$x_1^{(k)} = \widehat{\Phi}(x_0 + \varepsilon v_k).$$

In the right panels of Fig. 13, the generated images with v_k are shown, where v_k corresponds to the leading k-th singular value σ_k. In contrast to isotropic random perturbations, the leading

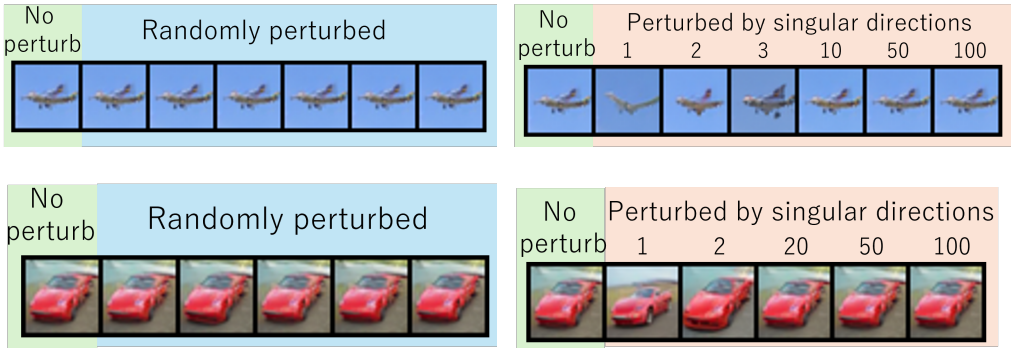


Figure 13: CIFAR-10 generation under perturbations of random directions and sigular vectors at the initial point. In each panel, the leftmost image is generated from the unperturbed initial point x_0, and the remaining images are generated from perturbed initial ones. Left: random perturbations. The generated images remain nearly identical to the reference image, consistently with contraction of the predominantly normal components. Right: Perturbations by the k-th right singular vector v_k of $D_{x_0}\widehat{\Phi}(x_0)$ in descending singular-value order. The number above each image indicates k. Leading directions produce coherent changes of pose, shape, viewpoint, or background while the samples remain visually close to the CIFAR-10 image manifold. Higher-rank directions have visibly weaker effects.

singular directions ($k = 1, 2, 3$) produce visible but structured variations. For the airplane, the first few directions modify the shape of the aircraft while preserving an airplane-like image. For the automobile, the leading directions alter the front or side geometry while retaining a coherent automobile image. The changes become much smaller for directions with larger k.

These observations are consistent with the tangent–normal decomposition in Section 3. Directions that survive the terminal dynamics should appear among the dominant singular directions of $D_{x_0}\widehat{\Phi}$, whereas strongly contracted normal directions correspond to small singular values. Thus, the leading v_k provides empirical tangent-like directions: they induce variations within the image manifold instead of producing unstructured pixel noise.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and instrudction describe the scope and contribution of this work; proximal view to FM, its application to theoretical stability analysis of manifold structure, and experimental verification.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have included the limitations in Conclusion section explicitly.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We have provided full set of assumptions to mathematical results. All the complete proofs are given in Appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We explain the detailed exprimental settings and provide the code in Supplementary matrerials.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We have provided our codes with sufficinet information in Supplementary materials

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We have provided the sufficient information in the paper and detailed settings in the codes.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We have provided the mean and STD in the experiments, if availabe, to illustrate the statistical significance of the resutls.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Our experiments are simple to demonstrate the theretical results, and provided the compute information in Experiments section.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have read the code of ethics, and do not find any issues about this submission.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The current paper is limited to the theoretical views of generative models, and does not have any direct societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our paper proposes a theoretical viewpoint, and do not release any models.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: Out work does not use any previous assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: Our work proposes a theoretical viewpoint and does not introduce any new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our work does not make any crowdsorucing experimetns or with human subjects.

    Guidelines:n

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

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

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

    Answer: [N/A]

    Justification: Our work does not have any experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: We have used LLM only for writing, editing, and proof assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.