GRADIENT-DESCENT BASED CONVERGENCE AND ERROR ANALYSIS OF FLOW MATCHING IN DEEP RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching has proven to be a highly efficient technique for generative modeling of complex data, and works by learning the constructed velocity fields of a continuous normalizing flow via deep least-squares regression. However, most existing theoretical investigations on the accuracy of flow models, while significant, assume that the velocity function has been approximated to a certain accuracy, and then use this a priori bound to control the error of generation. We provide a quantitative understanding of the whole generation process, from training to sampling. More precisely, we provide a non-asymptotic convergence analysis of flow matching under gradient descent for a general class of deep ReLU neural networks, which elucidates how to design the training and sampling process for effective generation. For instance, our theory indicates the explicit effect of increasing the depth and width of the network, as well as the gradient-descent hyperparameters, on the final target accuracy. By propagating the convergence error across the generation process, we derive end-to-end error bounds on the Wasserstein-2 distance between the generated and target distributions.

1 INTRODUCTION

Flow matching has emerged as a powerful alternative to diffusion models for generative modeling, offering simpler training objectives and more stable sample generation (Lipman et al., 2023; Liu et al., 2022a). Given samples from a source distribution π_0 (typically Gaussian) and a target distribution π_1, flow matching learns a time-dependent velocity field $\nu(x,t)$ that transports π_0 to π_1 via the ODE $\mathrm{d}Z_t = \nu(Z_t,t)\mathrm{d}t$.

While empirical success has been substantial (Wang et al., 2026), theoretical analyses of flow matching have so far compartmentalized the pipeline: prior work characterizes the optimal velocity field under regularity assumptions (Benton et al., 2024; He et al., 2026; Gaur et al., 2025), the approximation power of neural networks for representing it, and the discretization error of ODE solvers given an accurate velocity, but these analyses uniformly *assume* that training has already produced such an accurate velocity field (Zhou and Liu, 2025). While we defer a more rigorous study of the related work to Appendix A, we point to a gap in the literature that is central in practice:

How does gradient-based training of the neural velocity field affect end-to-end generation quality?

Closing this gap is non-trivial for reasons specific to flow matching. The velocity field must be learned jointly across all $t \in [0,1]$, so optimization couples a continuum of regression problems; the loss is a nested expectation over data, noise, and time; the training landscape is non-convex; and any residual training error compounds along the ODE trajectory at sampling time. Recent work by Wang et al. (2024) proves exponential convergence of gradient descent for over-parameterized score networks in diffusion models, but their analysis hinges on the score-matching structure and does not transfer to flow matching, whose objective regresses against the displacement $X_1 - X_0$ rather than a score. Zhou and Liu (2025) give an end-to-end statistical guarantee for flow matching by exhibiting a hypothesis class containing a velocity within $(\sqrt{d}+1)\epsilon$ of the optimum and assuming this best-in-class element is recovered exactly; the optimization step is left as a black box.

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

1.1 CONTRIBUTIONS

Our main contribution is an end-to-end optimization-aware analysis of flow matching with deep ReLU velocity networks. While Zhou and Liu (2025) establish statistical approximation and generalization guarantees for flow matching under an empirical risk minimization framework, their analysis does not characterize how gradient-based neural-network training reaches the required velocity-field accuracy. We close this gap by proving non-asymptotic convergence guarantees for projected gradient descent on the empirical flow-matching objective in an overparameterized regime, and by incorporating the resulting optimization error into the downstream sampling analysis. Combining this with approximation, statistical estimation, Monte Carlo estimation, and Euler discretization bounds, we obtain explicit training-to-generation guarantees in Wasserstein distance, as well as KL guarantees for Gaussian-smoothed terminal laws.

2 PRELIMINARIES

Notation. We denote $[N] \coloneqq \{1, \dots, N\}$. For matrices $\mathbf{A}$ and $\mathbf{B}$, we say $\mathbf{A} \preceq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive semidefinite. We denote the identity matrix in $\mathbb{R}^{d\times d}$ by $\mathbf{I}_d$. We use $\|\cdot\|_1, \|\cdot\|_2$, and $\|\cdot\|_\infty$ to denote the standard ℓ_1, ℓ_2, and ℓ_∞ norms (respectively). We define the operator norm of a matrix $\mathbf{A}$ as $\|\mathbf{A}\|_{\mathrm{op}} \coloneqq \sup_{\|x\|\le 1} \|\mathbf{A}x\|$. For a twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we define the $L^2(\pi)$-norm of f as $\|f\|_{L^2(\pi)} \coloneqq (\int (f(x))^2 \pi(x) \mathrm{d}x)^{1/2}$. Moreover, the $L^\infty(K)$-norm is given by $\|f\|_{L^\infty(K)} \coloneqq \sup_{x\in K} |f(x)|$. For a vector function $v : \mathbb{R}^d \to \mathbb{R}^d$, we define its $L^2(\pi)$-norm as $\|v\|_{L^2(\pi)} \coloneqq \|\|v\|\|_{L^2(\pi)}$ and $L^\infty(K)$-norm as $\|v\|_{L^\infty(K)} \coloneqq \|\|v\|\|_{L^\infty(K)}$. We use O to hide constants, and $\tilde{O}$ to hide all polylogarithmic factors. Given distributions μ and ν, the Wasserstein-2 distance is $W_2(\mu, \nu) \coloneqq \inf_{x\in\Pi(\mu,\nu)} \mathbb{E}_{(x,y)\sim\pi}[\|x - y\|^2]^{1/2}$, where $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν (specifically, it is a distribution on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ and ν on its first and second factors, respectively). Finally, the KL-divergence is given by $\mathrm{KL}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.

Continuous normalizing flows (CNF). Given $\mathbf{x} \in \mathbb{R}^d$, consider a probability density path $p : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{>0}$ where $\int p_t(x)\mathrm{d}x = 1$, and a time dependent vector field $v : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$. v_t forms a diffeomorphic map (the *flow*) $\phi_t : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ such that $\frac{\mathrm{d}\phi_t(x)}{\mathrm{d}t} = v_t(\phi_t(x))$ and $\phi_0(x) = x$. If $v_t \coloneqq v_\theta(x, t)$ is a neural network parameterized by θ, then ϕ_t is said to be a CNF. Then, $p_t = [\phi_t]_\star p_0$ is the push forward of ϕ_t on p_0, where $[\phi_t]_\star p_0(x) = p_0(\phi_t^{-1}(x)) \cdot \det[\frac{\partial\phi^{-1}(x)}{\partial x}]$.

Flow matching. Given independent empirical observations of $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$, we want to find an ODE on time $t \in [0, 1]$, $\mathrm{d}Z_t = \mathbf{v}(Z_t, t)\mathrm{d}t$, which ensures that if $Z_0 \sim \pi_0$, then $Z_1 \sim \pi_1$. Lipman et al. (2023) shows that the vector field can be found by solving the regression problem

$$v^* = \arg\min_v \mathcal{L}_0(v) \coloneqq \arg\min_v \int_0^1 \mathbb{E}_{X_0,X_1}[\|(X_1 - X_0) - \mathbf{v}(X_t, t)\|^2]\mathrm{d}t, \tag{1}$$

where $X_t = tX_1 + (1 - t)X_0$ is the linear interpolant between X_0 and X_1, for $X_0 \sim \pi_0, X_1 \sim \pi_1$. This framework generalizes continuous normalizing flows (Chen et al., 2018; Grathwohl et al., 2018; Papamakarios et al., 2021; Dinh et al., 2016; Kingma and Dhariwal, 2018) and provides a principled approach to generative modeling (Chen and Lipman, 2024; Liu et al., 2022a; Rezende and Mohamed, 2015; Kingma and Welling, 2013; Goodfellow et al., 2014). Unlike diffusion models (Ho et al., 2020; Song et al., 2021; Sohl-Dickstein et al., 2015; Wang et al., 2026) which define a forward noising process and learns score functions, flow matching uses neither. Then, the exact minimum is achieved by $\mathbf{v}^*(x, t) = \mathbb{E}[X_1 - X_0|X_t = \mathbf{x}]$. The velocity field $\mathbf{v}^*$ is approximated by neural networks. To avoid instability, we often clip the integral interval $[0, 1]$ with T, and consider the loss function

$$\mathcal{L}(v) = \frac{1}{T}\int_0^T \mathbb{E}_{X_0,X_1}[\|(X_1 - X_0) - \mathbf{v}(X_t, t)\|^2]\mathrm{d}t. \tag{2}$$

Then, given a family of neural networks NN, we consider the following approximation error

$$\inf_{\mathbf{v}\in\mathsf{NN}} \int_0^T \|\mathbf{v}(\cdot, t) - \mathbf{v}^*(\cdot, t)\|_{L^2_{\pi_t}}^2 \,\mathrm{d}t = \inf_{\mathbf{v}\in\mathsf{NN}} \mathcal{L}(\mathbf{v}) - \mathcal{L}(\mathbf{v}^*), \tag{3}$$

where π_t is the distribution of X_t. We denote the best approximator $\tilde{v} \in \arg\min_{\mathbf{v}\in\mathsf{NN}} \mathcal{L}(\mathbf{v})$.

Assumption 2.1 (Bounded support)**.** The target distribution π_1 is supported on $[0,1]^d$.

Assumption 2.2 (Lipschitz score)**.** Let $\pi_1(\mathrm{d}x) = e^{-V(x)}\mathrm{d}x$. Moreover, the potential $V(x)$ is twice continuously differentiable and satisfies $-\alpha I \preceq \nabla^2 V(x) \preceq \alpha I$ with $\alpha > 1$.

Lemma 2.3 (Lemmas 1.3 and 1.4 in Zhou and Liu (2025))**.** *Suppose Assumption 2.1 holds. Then, $v^*(x,t)$ is ξ-Lipschitz continuous with respect to x on $\mathbb{R}^d \times [0,T]$, where $\xi \le \max\left\{\frac{1}{1-T}, \frac{Td}{(1-T)^3}\right\}$. Further, if $\frac{1}{2} < T < 1$, we have that v^* is $\frac{d}{(1-T)^3}$-Lipschitz continuous with respect to x. If we additionally assume Assumption 2.2, then $v^*(x,t)$ is $\zeta(\alpha,d)$-Lipschitz continuous on $\mathbb{R}^d \times [0,1]$ with respect to x, where $\zeta(\alpha,d) = d(\alpha + \sqrt{\alpha + 2/d})^2/2$.*

For self-containedness, we provide the proof of Lemma 2.3 from Zhou and Liu (2025) in Appendix G.

3 Learning the Velocity Field

Training objective. Let π_0 be the base distribution $\mathcal{N}(0, I_d)$ and π_1 be the target/data distribution. Let $\pi = \pi_0 \otimes \pi_1$ be the product coupling on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals π_0 and π_1. Then, draw n i.i.d. samples $\{(x_i^{(0)}, x_i^{(1)})\}_{i=1}^n$ where $x_i^{(0)} \sim \pi_0$ and $x_i^{(1)} \sim \pi_1$. For $t \in [0,1]$, let $X_t = (1-t)x^{(0)} + tx^{(1)}$ and $U \coloneqq \frac{\mathrm{d}X_t}{\mathrm{d}t} = X_1 - X_0$. We discretize the time integral using N deterministic collocation points $0 = t_0 < t_1 < \cdots < t_N = T < 1$. For every $i \in [n]$ and discretization time t_j, define $X_{ij} = (1-t_j)x_i^{(0)} + t_j X_i^{(1)}$ and $U_i \coloneqq x_i^{(1)} - x_i^{(0)}$. Define quadrature weights $\beta_j = \frac{t_j - t_{j-1}}{T}$ satisfying $\sum_{j=1}^N \beta_j = 1$. Moreover, let $\beta_{\min} \coloneqq \min_j \beta_j$ and $\beta_{\max} \coloneqq \max_j \beta_j$. Then, the empirical training objective is given by

$$\mathcal{L}_{\mathrm{emp}}(v_\theta) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^N \beta_j \|v_\theta(X_{ij}, t_j) - (x_i^{(1)} - x_i^{(0)})\|^2. \tag{4}$$

In practice, the true velocity field is approximated by neural networks (Shou et al., 2025). To ensure effective learning, the network class should be expressive enough to approximate the velocity field.

Architecture. We first describe the deep-ReLU velocity network, which models the simple shallow MLPs used in practice (Glorot and Bengio, 2010; He et al., 2016; Krizhevsky et al., 2012; LeCun et al., 1998; Mei et al., 2018; Chizat and Bach, 2018; Rotskoff and Vanden-Eijnden, 2018). Fix depth $L \ge 1$ and width m. Let $\sigma(\cdot)$ denote the ReLU activation. Absorbing time into the input by augmenting the dimension, we define the augmented input $\bar{X}_{ij} \in \mathbb{R}^{d+1}$ by $\bar{X}_{ij} \coloneqq [X_{ij} \quad t_j]^\top$. Let $\mathbf{W}_0 \in \mathbb{R}^{m\times(d+1)}, \mathbf{W}_\ell \in \mathbb{R}^{m\times m}$ for $\ell = 1,\dots,L$, and $\mathbf{W}_{L+1} \in \mathbb{R}^{d\times m}$. Then, for $\theta = (\mathbf{W}_0, \dots, \mathbf{W}_{L+1})$, the fully-connected L-ReLU network with bias absorbed is

$$v_\theta(X_{ij}, t_j) = \mathbf{W}_{L+1}\sigma(\mathbf{W}_L \cdots \mathbf{W}_1 \sigma(\mathbf{W}_0 \bar{X}_{ij}) \cdots). \tag{5}$$

Specifically, let $q_{ij,-1} \coloneqq \bar{X}_{ij}$, $r_{ij,0} = \mathbf{W}_0 q_{ij,-1}$, $q_{ij,0} = \sigma(r_{ij,0}), r_{ij,\ell} = \mathbf{W}_\ell q_{ij,\ell-1}, q_{ij,\ell} = \sigma(r_{ij,\ell})$ for $\ell \in [L]$, and then $v_\theta(X_{ij}, t_j) = \mathbf{W}_{L+1} q_{ij,L}$. To ensure v_θ is Lipschitz, we enforce that $\|\mathbf{W}_\ell\|_{\mathrm{op}} \le R_{\mathrm{hid}}$, for $\ell \in [L]$, during the initialization and updates, and $\|\mathbf{W}_{L+1}\|_{\mathrm{op}} \le \Gamma / R_{\mathrm{hid}}^L$. In practice, this is done via a spectral norm regularizer (Yoshida and Miyato, 2017), but for analytic simplicity, we equivalently model this via projected gradient descent (Bartlett et al., 2007). Next, following Chaudhari and Moura (2025); Allen-Zhu et al. (2019); Chen et al. (2022), we define diagonal gate matrices $\mathbf{D}_{ij,\ell} \in \mathbb{R}^{m\times m}$ with $(\mathbf{D}_{ij,\ell})_{kk} = \mathbb{1}\{(r_{ij,\ell})_k > 0\}$ so that $q_{ij,\ell} = \mathbf{D}_{ij,\ell} r_{ij,\ell} = \mathbf{D}_{ij,\ell}\mathbf{W}_\ell q_{ij,\ell-1}$. Splitting $\mathbf{W}_0 = [\mathbf{A} \quad a]$ for $\mathbf{A} \in \mathbb{R}^{m\times d}$ and $a \in \mathbb{R}^m$, we have $\mathbf{W}_0[x;t] = \mathbf{A}x + at$. Since ReLU is 1-Lipschitz and positively homogeneous, and $\|\mathbf{D}_{ij,\ell}\|_2 \le 1$:

$$\mathrm{Lip}_x(v_\theta) \le \|\mathbf{W}_{L+1}\|_2 \|\mathbf{A}\|_2 \cdot \prod_{\ell=1}^L \|\mathbf{W}_\ell\|_2 \le R_{\mathrm{hid}}^L \|\mathbf{A}\|_2 \cdot \frac{\Gamma}{R_{\mathrm{hid}}^L} = \Gamma\|\mathbf{A}\|_2,$$

$$\mathrm{Lip}_t(v_\theta) \le \|\mathbf{W}_{L+1}\|_2 \|a\|_2 \cdot \prod_{\ell=1}^L \|\mathbf{W}_\ell\|_2 \le R_{\mathrm{hid}}^L \|a\|_2 \cdot \frac{\Gamma}{R_{\mathrm{hid}}^L} = \Gamma\|a\|_2.$$

We define the feasible set of parameters as the intersection of convex spectral-norm balls given by

$$\Theta_{\Gamma, R_{\mathrm{hid}}, B_x, B_t} \coloneqq \left\{\theta : \|\mathbf{W}_\ell\|_2 \le R_{\mathrm{hid}}, \forall \ell = 1,\dots,L, \|\mathbf{W}_{L+1}\|_2 \le \frac{\Gamma}{R_{\mathrm{hid}}^L}, \|\mathbf{A}\|_2 \le B_x, \|a\|_2 \le B_t\right\}.$$

Since each constraint is convex (and the intersection of convex sets is convex), Θ is also convex, closed, and projection is separable across the layers of the deep ReLU neural network. For a single layer projection, projecting its weights onto a spectral norm ball of radius R_{hid} is equivalent to projecting its singular values onto the ℓ_∞-norm ball of radius R_{hid}, which is achieved by rescaling the weight entries so that its Frobenius norm (which is efficiently computable) is less than R_{hid}. With this choice, we have $\text{Lip}_x(v_\theta) \le \Gamma B_x$ and $\text{Lip}_t(v_\theta) \le \Gamma B_t$, for all $\theta \in \Theta$.

3.1 Universal Function Approximation

Consider an abstract ReLU network configuration $\mathsf{NN}(L, M, J, K, \kappa, \gamma_1, \gamma_2)$ given by $\{v(x,t) = (\mathbf{W}_{L+1}\sigma(\mathbf{W}_L\sigma(\cdots \mathbf{W}_1\sigma(\mathbf{W}_0[x^\top, t]^\top)\cdots)$:network width bounded by M, $\sup_{x,t}\|v(x,t)\| \le K$, $\|\mathbf{W}_i\|_\infty \le \kappa$ for $i \in [L], \sum_{i=1}^L \|\mathbf{W}_i\|_0 \le J, \|v(x_1,t) - v(x_t,t)\| \le \gamma_1\|x_1 - x_2\|$ for $t \in [0,T]$, and $\|v(x,t_1) - v(x,t_2)\| \le \gamma_2\|t_1 - t_2\|$ for any $x\}$. Zhou and Liu (2025) gives a universal function approximation for the neural network class NN which we state below:

Theorem 3.1 (Theorem 3.1 in Zhou and Liu (2025))**.** *Suppose Assumption 2.1 holds. Given an approximation error $\epsilon > 0$ for any velocity field v^* with Lipschitz constant ζ w.r.t. x on $[0,T]$, we choose NN with $L = O(d + \log\frac{1}{\epsilon}), \gamma_1 = 10d\zeta, \gamma_2 = O(\frac{\sqrt{d^3\log\frac{d}{\epsilon}}}{(1-T)^4}), M = O(\frac{d^{1.5}(\log\frac{d}{\epsilon})^{\frac{d+1}{2}}}{(1-T)^4}\zeta^d\epsilon^{-(d+1)}), J = ML, K = O(\frac{\sqrt{d\log\frac{d}{\epsilon}}}{1-T})$, and $\kappa = O(\max\{\zeta\sqrt{\log\frac{d}{\epsilon}}, \gamma_2\})$. There exists an $\hat{v}_\theta \in \mathsf{NN}$ such that for any $t \in [0,T]$, we have $\|\hat{v}_\theta(\cdot,t) - v^*(\cdot,t)\|_{L^2(\pi_t)} \le (\sqrt{d}+1)\epsilon$.*

For completeness, we include the proof in Appendix C.

Remark 3.2. We set Γ, B_x, B_t such that $\Gamma B_x = 10d\zeta$ and $\Gamma B_t = O(\frac{\sqrt{d^3\log\frac{d}{\epsilon}}}{(1-T)^4})$. For each hidden layer $\ell \in [L]$, let $s_\ell = \max\{1, \|\mathbf{W}_\ell\|_2/R_{\text{hid}}\}$ and define new weights $\mathbf{W}'_\ell = \mathbf{W}_\ell/s_\ell$, $\mathbf{W}'_{L+1} = \mathbf{W}_{L+1}\cdot s_1 s_2 \cdots s_L$, and $\mathbf{W}'_0 = \mathbf{W}_0$. Then, $\|\mathbf{W}'_\ell\|_2 \le R_{\text{hid}}$ for $\ell \in [L]$, and the realized function is unchanged, i.e. $v_\theta = v_{\theta'}$ since ReLU is positively homogeneous as $\sigma(\alpha z) = \alpha\sigma(z)$ and scaling a hidden layer's preactivations by s_ℓ and scaling the next linear map by $1/s_\ell$ does not change the composed function. In doing so, the best neural network in the class that we parameterized in Section 3 is $(\sqrt{d}+1)\epsilon$ away from the optimal velocity function v^*, as measured in $L^2(\pi_t)$-norm.

Remark 3.3. Since the class in Lipschitz in x and t, each residual trajectory is Lipschitz in time. Specifically, if $r_i(t;\theta) = v_\theta(x_i(t),t) - u_i)$, where $u_i = x_i^{(1)} - x_i^{(0)}$ and $x_i(t) = (1-t)x_i^{(0)} + tx_i^{(1)}$,

$$\|r_i(t) - r_i(s)\| \le (\gamma_2 + \gamma_1\|u_i\|)|t-s| = \left(\frac{2\sqrt{d^3\log(d/\epsilon)}}{(1-T)^4)} + 10d\zeta\|u_i\|\right)|t-s|. \tag{6}$$

3.2 Training Approximation

Define residuals $e_{ij} \coloneqq v_\theta(X_{ij}, t_j) - (X_1^{(i)} - X_0^{(i)}) \in \mathbb{R}^d$, and the empirical flow matching objective $\mathcal{L}_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^N w_j\|e_{ij}(\theta)\|^2$. Then, define the per-index block loss $\ell_i(\theta) = \sum_{j=1}^N w_j\|e_{ij}(\theta)\|^2$, which implies $\mathcal{L}_{\text{emp}}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i(\theta)$. We train via full-batch projected gradient descent: $\theta^{(k+1)} = \Pi_\Theta(\theta^{(k)} - h\nabla\mathcal{L}_{\text{emp}}(\theta^{(k)}))$, where $\theta^{(k)} = (\mathbf{W}_0, \mathbf{W}_1^{(k)}, \ldots, \mathbf{W}_L^{(k)}, \mathbf{W}_{L+1})$ are the parameters at time k, $\eta > 0$ is the learning rate, and the gradient decomposes as $\nabla\mathcal{L}_{\text{emp}}(\theta) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^N w_j\nabla f(\theta;i,j)$. Following Allen-Zhu et al. (2019); Wang et al. (2024), we fix $\mathbf{W}_0$ and $\mathbf{W}_{L+1}$ throughout the process and only train on the hidden layers $\mathbf{W}_1, \ldots, \mathbf{W}_L$.

$\mathcal{L}_{\text{emp}}$ is highly non-convex due to neural network composition (Kawaguchi, 2016; Choromanska et al., 2015), which precludes the application of standard convergence analyses. Our approach leverages a semi-smoothness framework (Allen-Zhu et al., 2019; Zou et al., 2020; Oymak and Soltanolkotabi, 2019; Du et al., 2019; Wang et al., 2024); and uses that for large enough m, the network satisfies a universal function approximation theorem (Lee et al., 2019; Jacot et al., 2018; Yang, 2020; Arora et al., 2019), enabling convergence analyses toward the optimal ground-truth velocity function.

We require the following conditions on data and hyperparameters.

Assumption 3.4 (Data and Dimension Scaling)**.** We assume $\|X_0^i\| = \Theta(\sqrt{d})$ and $\|X_1^i\| = \Theta(\sqrt{d})$ for all i, and that $d \ge \Omega(\text{poly}(\log(nN)))$. This is a standard assumption in the literature (Wang et al.,

Algorithm 1 Flow Matching (Lipman et al. (2023))

Require: π_0, π_1, network ν_θ
1: **while** not converged **do**
2: Sample (x_0, x_1, t) and compute $x_t = tx_1 + (1-t)x_0$
3: $\mathcal{L}(\theta) = \|\nu_\theta(x_t, t) - (x_1 - x_0)\|^2$
4: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
5: **return** ν_θ

2024) and is not too restrictive: for instance, it holds with probability at least $1 - O(\exp(-\Omega(d)))$ for standard d-dimensional Gaussian random vectors, and the assumption only focuses on the initial data X_0^i rather than the interpolates X_t^i constructed during the flow matching training process.

Truncated Random initialization. We modify the initialization from Allen-Zhu et al. (2019) as follows: let $R_{\text{hid}} > 2\sqrt{2}$ be a fixed constant and choose a slack $s_{\text{hid}} > 0$ such that $R_{\text{hid}} - s_{\text{hid}} > 2\sqrt{2}$. The input layer is initialized with independent Gaussian entries $\mathbf{W}_\ell^{(0)} \sim \mathcal{N}(0, 2/m)$ for $\ell = 0, 1, \ldots, L$, and we condition on the high-probability events that $\|\mathbf{A}\|_{\text{op}} \le B_x$, $\|a\|_2 \le B_t$, and $\max_{\ell \in [L]} \|\mathbf{W}_\ell^{(0)}\|_{\text{op}} \le R_{\text{hid}} - s_{\text{hid}}$. On this event, the hidden initialization lies strictly inside the projected-gradient feasible set Θ with spectral slack s_{hid}. For the output layer, we draw a random Gaussian matrix $\mathbf{G} \sim \mathcal{N}(0, 1/d)$, and set $\mathbf{W}_{L+1}^{(0)} = cR_{\text{hid}}^{-L} \|\mathbf{G}\|_{\text{op}}^{-1} \mathbf{G}$. Thus, the initialized network satisfies the desired Lipschitz conditions. Notably, normalizing $\mathbf{W}_{L+1}$ by a positive scalar preserves the Gaussian output-column directions used in our eventual gradient lower bound argument.

Theorem 3.5 (Exponential convergence of GD)**.** *Consider the index with the largest-loss block $i^*(k) \in \arg\max_{1 \le i \le n} \ell_i(\theta^{(k)})$ with worst performing time $j^*(k) \in \arg\max_{1 \le j \le N} \beta_j \|v_{\theta^{(k)}}(t_j, X_{i^*(k)_j}) - U_{i^*(k)}\|^2$. Then, there exists $m \ge M_{opt}(\epsilon_{train}) = \Omega(\text{poly}(n, N, d, L, w_{max}/w_{min}, \log(1/\epsilon_{train}))$, and a step-size $h = \Theta(\frac{nN}{m w_{\min}})$ such that with probability at least $1 - \text{poly}(nN)\exp(-\Omega(d))$ over initialization, the iterates satisfy that for all $k \le K$ in the neighborhood,*

$$\mathcal{L}_{em}(\theta^{(k+1)}) \le \left(1 - \frac{c_0 h m d^{\frac{a_0 - 1}{2}} w_{j^*(k)}}{n^3 N^2}\right) \mathcal{L}_{em}(\theta^{(k)}), \tag{7}$$

for $a_0 \in (\frac{1}{2}, 1)$. Therefore, setting $K = \Theta(d^{\frac{1-a_0}{2}} n^2 N \log \frac{d}{\epsilon_{train}})$, we get $\mathcal{L}_{em}(\theta^{(K)}) \le \epsilon_{train}$.

We prove Theorem 3.5 (adapting Allen-Zhu et al. (2019) and Wang et al. (2024)) in Appendix E. The key methodology of the framework is to establish the lower bound of the gradient. Then by integrating it into the semi-smoothness property of the neural network (which we prove via stability arguments), we obtain the exponential rate of convergence of GD. For the lower bound, our proof exploits high-dimensional geometric and probabilistic properties, and shows that the weight projections are inactive during GD, enabling the high-probability bounds to persist throughout the proof.

Remark 3.6 (Discussion of optimality)**.** Theorem 3.5 guarantees geometric decay of the empirical training loss, with iteration complexity depending logarithmically on $1/\epsilon_{\text{train}}$. This matches the classical worst-case accuracy dependence of constant-step gradient descent on smooth strongly convex quadratics, despite the non-convexity and non-smoothness of our neural objective. This comparison concerns the dependence on training accuracy: we do not establish a matching lower bound for the specific flow-matching problem, and the polynomial dependence on sample size, time discretization, and dimension, as well as the required network width, may admit improvements. Theorem 3.5 provides a finite-accuracy guarantee: for each prescribed $\epsilon_{\text{train}} > 0$, the network width must satisfy $m \ge M_{\text{opt}}(\epsilon_{\text{train}})$, and the contraction estimate is justified over the corresponding training horizon while the iterates remain in the controlled neighborhood of initialization. Since the width requirement depends on the target accuracy, taking $\epsilon_{\text{train}} \to 0$ may require increasing the network width, rather than merely training a fixed network longer.

Remark 3.7 (Extension to stochastic optimization.)**.** Our analysis suggests an extension to mini-batch SGD using techniques similar to those of Allen-Zhu et al. (2019). For mini-batches drawn from the fixed empirical flow-matching objective, the gradient lower bound and semi-smoothness estimates could be combined with stochastic-gradient moment bounds and concentration arguments controlling the iterates' distance from initialization. Establishing these additional estimates, with suitable choices

of step size and network width, would provide a route to analogous optimization guarantees, including geometric empirical-loss decay in the interpolating regime. Provided the iterates satisfy the same network constraints, the resulting optimization bound could then be propagated through our statistical and sampling analysis to obtain corresponding end-to-end guarantees. Training with freshly sampled noise and time would additionally require controlling gradient variance that need not vanish at the optimum. We leave a detailed stochastic analysis to future work.

4 GENERALIZATION

We now study how training error propagates through ODE integration to affect generations. Following Zhou and Liu (2025), we relate the loss $\mathcal{L}(v)$ and the L^2-approximation error $\|v(\cdot,t)-v^*(\cdot,t)\|_{L^2(\pi_t)}$.

Lemma 4.1. *The v^* uniquely minimizes $\mathcal{L}(\cdot)$ over all square-integrable vector fields, and*

$$\mathcal{L}(v) = \mathcal{L}(v^*) + \frac{1}{T}\int_0^T \|v(\cdot,t) - v^*(\cdot,t)\|^2_{L^2(\pi_t)}\, dt. \tag{8}$$

Proof. For $Y = X_1 - X_0$, we have $\|Y - v(X_t,t)\|^2 = \|Y - v^*(X_t,t)\|^2 + \|v^*(X_t,t) - v(X_t,t)\|^2 + 2\langle Y - v^*(X_t,t), v^*(X_t,t) - v(X_t,t)\rangle$. Then, conditioning on X_t and using $v^*(X_t,t) = \mathbb{E}[Y \mid X_t]$ yields $\mathbb{E}[Y - v^*(X_t,t) \mid X_t] = 0$, hence the cross term vanishes in expectation. Integrating over $t \in [0,T]$ and dividing by T yields the expression, and uniqueness follows by convexity of $\mathcal{L}(\cdot)$. □

Empirical Evaluation. Let $\ell(x,v) = \frac{1}{T}\int_0^T \int \|x - x_0 - v(tx + (1-t)x_0, t)\|^2 \pi_0(x_0)\mathrm{d}x_0\mathrm{d}t$, where $\pi_0 = \mathcal{N}(0, I_d)$ is the standard Gaussian distribution. Given n i.i.d. samples $\{x_{1,i}\}_{i=1}^n$ from π_1, the empirical loss is given by $\bar{\mathcal{L}}(v) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(x_{1,i}, v)$. We approximate this via an empirical evaluation where we are given $\bar{m}$ many samples $\{(t_j, x_{0,j})\}_{j=1}^{\bar{m}}$ from $\mathrm{Unif}[0,T]$ and π_0 and using

$$\hat{\ell}_{\bar{m}}(x,v) \coloneqq \frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} \|x - x_{0,j} - v(t_j x + (1-t_j)x_{0,j}, t_j\|^2, \tag{9}$$

where $\ell(x,v)$ can be efficiently approximated by $\hat{\ell}_{\bar{m}}(x,v)$ for sufficiently large $\bar{m}$ that we bound later. Then, the empirical risk minimization (ERM) problem is given by $\hat{v} \in \arg\min_{v \in \mathsf{NN}} \bar{\mathcal{L}}(v)$, which the gradient-based training process approximates by

$$\hat{v}_{\text{target}} \in \arg\min_{v \in \mathsf{NN}} \big\{\hat{\mathcal{L}}_{n,\bar{m}}(v) \coloneqq \frac{1}{n}\sum_{i=1}^n \hat{\ell}_{\bar{m}}(x_{1,i}, v)\big\}, \tag{10}$$

and our learned approximation of $\hat{v}_{\text{target}}$ is given by $\hat{v}_{\text{train}}$.

4.1 TOTAL ERROR DECOMPOSITION

We decompose the velocity-field estimation error into training error, estimation error, and approximation error. For any $v \in \mathsf{NN}$, the estimation error decomposes as,

$$\begin{aligned}\mathcal{L}(v) - \hat{\mathcal{L}}_{n,\bar{m}}(v) &= \mathcal{L}(v) - \bar{\mathcal{L}}(v) + \bar{\mathcal{L}}(v) - \hat{\mathcal{L}}_{n,\bar{m}}(v) \\ &= \frac{1}{n}\sum_{i=1}^n (\mathcal{L}(v) - \ell(x_{1:i}, v)) + \frac{1}{n}\sum_{i=1}^n (\ell(x_{1:i}, v) - \hat{\ell}_{\bar{m}}(x_{1:i}, v)).\end{aligned}$$

Therefore, the total error of the velocity field decomposes into:

$$\begin{aligned}&\mathcal{L}(\hat{v}_{\text{train}}) - \mathcal{L}(v^*) = \mathcal{L}(\hat{v}_{\text{train}}) - \inf_{v\in\mathsf{NN}} \mathcal{L}(v) + \inf_{v\in\mathsf{NN}} \mathcal{L}(v) - \mathcal{L}(v^*) \\ &= \mathcal{L}(\hat{v}_{\text{train}}) - \hat{L}_{n,\bar{m}}(\hat{v}_{\text{train}}) + \hat{L}_{n,\bar{m}}(\hat{v}_{\text{train}}) - \hat{L}_{n,\bar{m}}(v_{\text{target}}) + \hat{L}_{n,\bar{m}}(v_{\text{target}}) - \inf_{v\in\mathsf{NN}} L(v) + \inf_{v\in\mathsf{NN}} L(v) - L(v^*) \\ &\le \hat{L}_{n,\bar{m}}(\hat{v}_{\text{train}}) - \hat{L}_{n,\bar{m}}(v_{\text{target}}) + 2\sup_{v\in\mathsf{NN}} |L(v) - \hat{L}_{n,\bar{m}}(v)| + \inf_{v\in\mathsf{NN}} L(v) - L(v^*) \\ &\le \underbrace{\hat{L}_{n,\bar{m}}(\hat{v}_{\text{train}}) - \hat{L}_{n,\bar{m}}(v_{\text{target}})}_{\text{training error}} + \underbrace{2\sup_{v\in\mathsf{NN}} |L(v) - \bar{\mathcal{L}}(v)|}_{\text{outer estimation (data) error}} + \underbrace{2\sup_{v\in\mathsf{NN}} |\bar{\mathcal{L}}(v) - \hat{L}_{n,\bar{m}}(v)|}_{\text{inner Monte Carlo estimation error}} + \underbrace{\inf_{v\in\mathsf{NN}} L(v) - L(v^*)}_{\text{approximation error}}.\end{aligned}$$

The training error is given by Section 3.2. To convert the uniform approximation error in Section 3.1 to the approximation error, we note that

$$\inf_{v\in\mathsf{NN}} L(v) - L(v^*) = \inf_{\theta\in\Theta}\left(\sum_{j=1}^{N} w_j \mathbb{E}[\|v_\theta(X_{t_j},t_j) - v^*(X_{t_j},t_j)\|^2]\right)$$
$$\leq \sum_{j=1}^{N} w_j(\sqrt{d}+1)^2\epsilon^2 = (\sqrt{d}+1)^2\epsilon^2$$

Finally, we bound the estimation errors, $\sup_{v\in\mathsf{NN}}|\mathcal{L}(v)-\bar{\mathcal{L}}(v)|$ and $\sup_{v\in\mathsf{NN}}|\bar{\mathcal{L}}(v)-\hat{\mathcal{L}}_{n,\bar{m}}(v)|$, which we control via covering-number arguments in Section 4.2, building on Zhou and Liu (2025).

4.2 Covering Number Bounds on the Estimation Error Terms

Definition 4.2 (Covering Number)**.** *Let ρ be a pseudometric on a set $\mathcal{M}$, and let $S \subseteq \mathcal{M}$. For any $\delta > 0$, a subset $A \subseteq \mathcal{M}$ is a δ-cover of S if for all $x \in S$ there exists $y \in A$ such that $\rho(x,y) \leq \delta$. The δ-covering number of S, denoted by $\mathcal{N}(\delta, S, \rho)$, is the minimum cardinality of a δ-cover of S.*

Let $K := \sup_{v\in\mathsf{NN}}\|v\|_{L^\infty(\mathbb{R}^d\times[0,T])}$ and $\mathcal{H} := \{\ell(\cdot,v) : v \in \mathsf{NN}\}$. We define the envelope of $\mathcal{H}$ by $B := \sup_{h\in\mathcal{H}}\|h\|_{L^\infty([0,1]^d)}$. By Lemma D.2, this envelope satisfies $B \lesssim d + K^2$. For covering radius $\tau > 0$, define $N_1 := \mathcal{N}\left(\tau, \mathcal{H}, \|\cdot\|_{L^\infty([0,1]^d)}\right)$. By a covering argument and Hoeffding's inequality (Appendix D.1), with probability at least $1-\delta$,

$$\sup_{v\in\mathsf{NN}}\left(\mathcal{L}(v) - \bar{\mathcal{L}}(v)\right) \leq \sqrt{\frac{2B^2\log(N_1/\delta)}{n}} + 2\tau. \tag{11}$$

For the inner Monte Carlo estimation error, fix $x \in [0,1]^d$ and define the truncated integrand class

$$\mathcal{R}_x^{\mathrm{trunc}} := \left\{(x_0,t) \mapsto \|x - x_0 - v(tx+(1-t)x_0,t)\|^2 \mathbf{1}\{\|x_0\|_\infty \leq R\} : v \in \mathsf{NN}\right\}. \tag{12}$$

For covering radius $\tau > 0$, let $N_2 := \sup_{x\in[0,1]^d}\mathcal{N}\left(\tau, \mathcal{R}_x^{\mathrm{trunc}}, \|\cdot\|_{L^\infty([-R,R]^d\times[0,T])}\right)$. On the truncation event $\{\|x_0\|_\infty \leq R\}$, this class has envelope $B_R \lesssim d(R+1)^2 + K^2$, so the truncated inner MC error is controllable by a covering argument and Hoeffding's inequality. The remaining truncation terms is controllable by Gaussian tail bounds. The inner MC estimation error satisfies

$$\sup_{v\in\mathrm{NN}} \bar{\mathcal{L}}(v) - \hat{\mathcal{L}}_{n,\bar{m}}(v) \leq O\left(\frac{\delta^{1/2}(d+1)+K^2}{\sqrt{\bar{m}}} + \frac{\left(d\log(\bar{m}d/\delta)+K^2\right)\sqrt{\log(N_2/\delta)}}{\sqrt{\bar{m}}}\right) \tag{13}$$

with probability $1-\delta$. All derivations are contained within Appendix D.1. Setting $\delta = \frac{1}{4n}$ and $\tau = \frac{1}{n}$ yields the following theorem to bound the estimation plus the approximation error.

Theorem 4.3 (Theorem 4.4 from Zhou and Liu (2025))**.** *Suppose Assumption 2.1 holds. Let v^* be a velocity field with Lipschitz constant ζ with respect to x. Given n samples $\{x_{1,i}\}_{i=1}^n$ drawn from π_1 and $\bar{m}$ samples from π_0 and $\mathrm{Unif}[0,T]$, we choose a neural network as in Theorem 3.1 with $\varepsilon = n^{-\frac{1}{d+5}}$. Then, with probability at least $1-\frac{1}{n}$, it holds that*

$$\frac{1}{T}\int_0^T \|\hat{v}_{\textit{target}}(\cdot,t) - v^*(\cdot,t)\|_{L^2(\pi_t)}^2\, dt = \tilde{\mathcal{O}}\left(\frac{\zeta^{d/2}}{(1-T)^4}\left(n^{-\frac{2}{d+5}} + n^{\frac{d+1}{2(d+5)}}\bar{m}^{-1/2}\right)\right). \tag{14}$$

Here, we omit logarithmic factors in d, $\log n$, $\log\bar{m}$, and $\log(1-T)$. If $\bar{m} = \mathcal{O}(n)$, this yields the convergence rate $\tilde{\mathcal{O}}(\frac{\zeta^{d/2}}{(1-T)^4}\, n^{-\frac{2}{d+5}})$.

For self-containedness, we provide the proof from Zhou and Liu (2025) in Appendix D.

4.3 Bound for Total Error Decomposition

Setting $K = \Theta(d^{\frac{-a_0}{2}} n^2 N \log dn)$ in the gradient-descent convergence bound, we have that the total estimation error of the velocity field decomposes into:

$$\epsilon_{\text{total}} := \mathcal{L}(\hat{v}_{\text{train}}) - \mathcal{L}(v^*) \leq \epsilon_{\text{train}} + \tilde{\mathcal{O}}\left(\frac{\zeta^{d/2}}{(1-T)^4}\, n^{-\frac{2}{d+5}}\right) = \tilde{\mathcal{O}}\left(\frac{\zeta^{d/2}}{(1-T)^4}\, n^{-\frac{2}{d+5}}\right) \tag{15}$$

with probability of at least $1 - 1/n$.

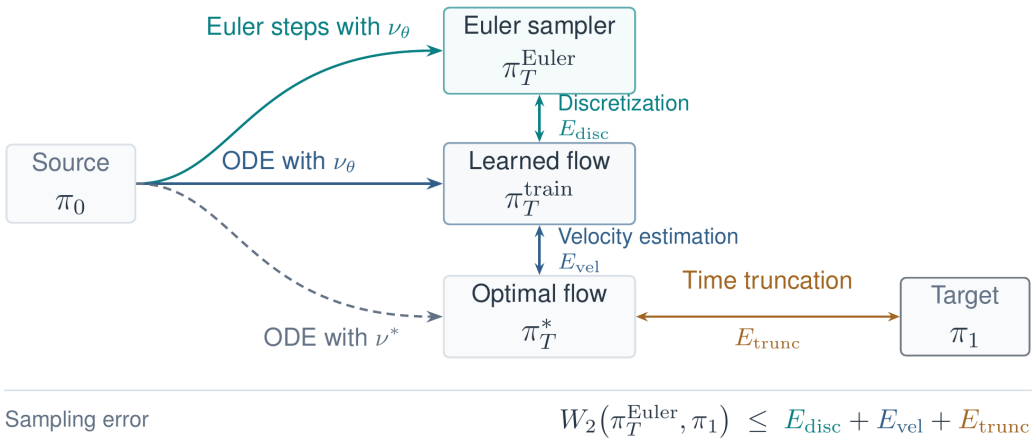

Figure 1: **Error decomposition and sampling discretization framework.** We distinguish between discretization (E_{disc}) arising from the Euler approximation and optimization error (E_{opt}) stemming from the gap between the learned velocity field ν_θ and the ground-truth optimal field ν^*.

5 Sampling

We now examine how the learned velocity field translates into accurate generation. Given a trained network ($\widehat{v}$), we draw an initial sample from (π_0) and approximate the corresponding flow up to time ($T < 1$) using the Euler scheme in Algorithm 2. We decompose the resulting Wasserstein error into contributions from velocity-field estimation, numerical discretization, and terminal-time truncation. Combining these bounds with the preceding training guarantees yields an end-to-end guarantee for the generated distribution.

Algorithm 2 Discretized Sampling with the Learned Flow

1: **Input:** Learned velocity field $\widehat{v}$, source distribution π_0,
2: **Require:** terminal time $T < 1$, grid $0 = t_0 < t_1 < \cdots < t_N = T$
3: **Output:** Sample $\widehat{Z}_T$
4: Sample $\widehat{Z}_0 \sim \pi_0$
5: **for** $j = 0, \ldots, N-1$: $\Delta_j \leftarrow t_{j+1} - t_j, \quad \widehat{Z}_{t_{j+1}} \leftarrow \widehat{Z}_{t_j} + \Delta_j \widehat{v}(\widehat{Z}_{t_j}, t_j)$
6: **return** $\widehat{Z}_T$

5.1 Error Guarantees

We now describe the sampling procedure and the resulting error decomposition. Let π_T^* denote the terminal law at time T under the continuous optimal flow ν^*, and let π_T^{train} denote the terminal law at time T under the continuous learned flow ν_θ. We run the Euler sampler in Algorithm 2 on a grid $0 = t_0 < t_1 < \cdots < t_N = T$, with $\Delta t_j = t_{j+1} - t_j$, and denote its terminal law by π_T^{Euler}. From the derivations in Appendix F, we have, with probability at least $1 - \frac{1}{n}$,

$$W_2(\pi_T^{\text{Euler}}, \pi_1) \le W_2(\pi_T^{\text{Euler}}, \pi_T^{\text{train}}) + W_2(\pi_T^{\text{train}}, \pi_T^*) + W_2(\pi_T^*, \pi_1) \tag{16}$$
$$\lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2) \left(\sum_{j=0}^{N-1} (\Delta t_j)^3 \right)^{1/2} + e^{L_\nu T} \sqrt{T(\epsilon_{\text{total}})} + (1 - T)\sqrt{d}.$$

Theorem 5.1. *Consider the Euler sampler in Algorithm 2 on a uniform grid $0 = t_0 < t_1 < \cdots < t_N = T$, with $\Delta t_j = T/N$. If we choose $T = 1 - n^{-\frac{1}{3(d+5)}}$ and $N \asymp n^{\frac{4}{3(d+5)}}$ and train until $\epsilon_{\text{train}} \lesssim n^{-\frac{2}{3(d+5)}}$, which, by Theorem 3.5, is achieved for*

$$m \ge \Omega\left(\text{poly}\left(n, n^{\frac{4}{3(d+5)}}, d, L, \frac{w_{\max}}{w_{\min}}, \log n \right) \right)$$

and $K = \widetilde{\Theta}\Big(d^{\frac{1-a_0}{2}} n^{2+\frac{4}{3(d+5)}}\Big)$ *projected-gradient steps, then* $W_2(\pi_T^{\mathrm{Euler}}, \pi_1) = \tilde{O}(n^{-\frac{1}{3(d+5)}})$.

We state Theorem 5.1 for the uniform Euler grid used in Algorithm 2. Additionally, Appendix F.6 shows that, under a mild local time-regularity condition on the learned velocity field, a time-adapted discretization preserves the same end-to-end Wasserstein rate while reducing the sufficient number of Euler time steps from $O(n^{\frac{4}{3(d+5)}})$, as implied by the maximum step-size condition in Zhou and Liu (2025), to $O(n^{\frac{1}{d+5}})$, up to logarithmic factors.

Lastly, we establish a KL bound after applying a Gaussian smoothing operator to both the learned and target terminal distributions. For $\sigma > 0$, let φ_σ denote the density of $\mathcal{N}(0, \sigma^2 I_d)$. For any probability measure μ on $\mathbb{R}^d$, define its Gaussian smoothing by $\mathcal{S}_\sigma \mu := \mu * \varphi_\sigma$. We can convert a Wasserstein-2 bound into a Kullback–Leibler bound from applying Proposition F.4 with $\mu = \pi_T^{\mathrm{Euler}}$ and $\nu = \pi_1$, which along with the end-to-end Wasserstein bound from Section 5.1, yield the following.

Corollary 5.2 (End-to-end KL guarantee for smoothed terminal laws)**.** *Suppose the assumptions of the end-to-end Wasserstein theorem in Section 5.1 hold. Then for every* $\sigma > 0$*,*

$$\mathrm{KL}\big(\mathcal{S}_\sigma \pi_T^{\mathrm{Euler}} \,\big\|\, \mathcal{S}_\sigma \pi_1\big) \lesssim \tilde{O}\left(\sigma^{-2} \cdot n^{-\frac{2}{3(d+5)}}\right),$$

We include the proof in Appendix F.5.

Remark 5.3. When the target distribution π_1 is supported on a low-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ with intrinsic dimension $d_{\mathcal{M}} \ll d$, the sample complexity should depend on $d_{\mathcal{M}}$ rather than ambient d. However, our bounds scale polynomially with d. Incorporating manifold assumptions via restricted isometry properties or local PCA could yield dimension-free convergence rates (Narayanan and Mitter, 2010; Weed and Bach, 2017).

Remark 5.4. Our discretization analysis uses first-order Euler integration, yielding $O(T^2/N)$ error. Higher-order methods like Runge-Kutta achieve $O(T^5/N^4)$ local error, potentially reducing the required number of function evaluations. However, the Wasserstein stability analysis becomes more complex. For RK4, each step requires 4 evaluations of ν_θ, and the global error accumulation depends on how these intermediate evaluations propagate errors. Establishing that the improved local convergence rate translates to improved Wasserstein bounds requires generalizing Lemma F.2 to implicit schemes, which is non-trivial when ν_θ is only approximately correct.

6 Conclusion and Future Works

Conclusion. We give an end-to-end convergence and error analysis for flow matching with deep ReLU velocity networks trained by gradient descent. Unlike prior analyses that assume access to an exact velocity field, our results connect the full pipeline: neural-network approximation, gradient-based learning, statistical generalization, Monte Carlo estimation, and numerical sampling. We show that, under standard regularity and bounded-support assumptions, sufficiently overparameterized ReLU networks trained with appropriate step sizes achieve small empirical flow-matching loss, and that this training error propagates through the sampling ODE in a controlled way. Our analysis yields explicit non-asymptotic bounds on the learned velocity-field error and the final generated distribution, including W_2 guarantees for the discretized sampler and KL guarantees after Gaussian smoothing. These results show how the network width, depth, number of training samples, time discretization, terminal-time truncation, and training accuracy jointly affect generation quality. Broadly, our work helps close the gap between optimization theory and sampling guarantees for generative models.

Limitations and Future Work. We identify several avenues of future work. First, our guarantees rely on Assumption 3.4 whereas practical implementations often use more complex architectures and latent contexts (Lin et al., 2023; 2024; Anand et al., 2026). Secondly, extending the analysis to more general probability paths and alternative interpolants would further strengthen the theoretical foundation of flow matching. Thirdly, our theoretical guarantees can be viewed as 'worst-case' guarantees over the target-distribution. It would be interesting to study whether assumptions of a manifold hypothesis or log-concavity in the target distribution can improve the final end-to-end error guarantees.

AI USE STATEMENT

In this work, we used generative AI tools for literature searches, coding implementation, and to aid in the presentation of our experimental results. We also used AI assistance to edit the manuscript, help design scientific figures, and validate our mathematical claims. The authors take full responsibility for verifying the correctness and originality of all the material in the manuscript, including the theoretical claims, experimental results, code, and figures.

ETHICS STATEMENT

This work develops theoretical guarantees for training and sampling in flow-matching models. By identifying how optimization, statistical estimation, and numerical integration affect generation accuracy, our analysis contributes to understanding the reliability and limitations of these models. The guarantees concern distributional approximation under stated mathematical assumptions; they do not establish privacy, fairness, or safety of generated outputs. Applications of flow matching may inherit biases or sensitive information from training data and therefore require evaluation appropriate to their intended use.

REPRODUCIBILITY

We specify the mathematical assumptions, neural-network parameterization, training objective, and sampling procedure underlying our analysis, and provide detailed proofs and supporting lemmas in the appendices. Our numerical experiments examine the dependence of generation error on dimension, training iterations, integration steps, and training-set size. Code for reproducing these experiments is available at `https://anonymous.4open.science/r/flow_matching_full_experiments-3BB6/README.md`. Reproducing the reported figures requires the target-distribution specifications, architecture and optimization settings, sampling configurations, random seeds, and Wasserstein evaluation procedure.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023. URL https://arxiv.org/abs/2303.08797.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL https://arxiv.org/abs/2303.08797.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *International Conference on Machine Learning*, pages 242–252, 2019.

Emile Anand. Towards the pseudorandomness of expander random walks for read-once acc0 circuits, 2025. URL https://arxiv.org/abs/2501.07752.

Emile Anand and Chris Umans. Pseudorandomness of the Sticky Random Walk, 2023. URL https://arxiv.org/abs/2307.11104.

Emile Anand, Ishani Karmarkar, and Guannan Qu. Mean-field sampling for cooperative multi-agent reinforcement learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 8462–8521. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0c8d49f2744d8ee4d5948621d6bb5bc0-Paper-Conference.pdf.

Emile Anand, Abdullah Ateyeh, Xinyuan Cao, and Max Dabagia. Continuous latent contexts enable efficient online learning in transformers, 2026. URL https://arxiv.org/abs/2605.09867.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in neural information processing systems*, volume 32, 2019.

Peter Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. *Advances in neural information processing systems*, 20, 2007.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods, 2024. URL https://arxiv.org/abs/2305.16860.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Herm Jan Brascamp and Elliott H. Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22:366–389, 1976. URL https://api.semanticscholar.org/CorpusID:120558079.

Saptarshi Chakraborty, Quentin Berthet, and Peter L. Bartlett. Generalization properties of score-matching diffusion models for intrinsically low-dimensional data, 2026. URL https://arxiv.org/abs/2603.03700.

Shreyas Chaudhari and José MF Moura. Relu networks as random functions: Their distribution in probability space. *arXiv preprint arXiv:2503.22082*, 2025.

Shreyas Chaudhari, Srinivasa Pranav, Emile Anand, and José M. F. Moura. Peer-to-peer learning dynamics of wide neural networks. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890126.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv preprint arXiv:2002.03938*, 2020.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks : Function approximation and statistical recovery, 2022. URL https://arxiv.org/abs/1908.01842.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023. URL https://arxiv.org/abs/2209.11215.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in wasserstein space. *IEEE Trans. Inf. Theor.*, 70(11): 8087–8106, November 2024. ISSN 0018-9448. doi: 10.1109/TIT.2024.3422412. URL https://doi.org/10.1109/TIT.2024.3422412.

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

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. *Artificial intelligence and statistics*, pages 192–204, 2015.

A. Dembo, T.M. Cover, and J.A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, 1991. doi: 10.1109/18.104312.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting, 2026. URL https://arxiv.org/abs/2602.04770.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Mudit Gaur, Prashant Trivedi, Shuchin Aeron, Amrit Singh Bedi, George K. Atia, and Vaneet Aggarwal. Generative modeling with continuous flows: Sample complexity of flow matching, 2025. URL https://arxiv.org/abs/2512.01286.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling, 2025. URL https://arxiv.org/abs/2505.13447.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Gautam Goel, Mahdi Soltanolkotabi, and Peter Bartlett. Training dynamics of softmax self-attention: Fast global convergence via preconditioning, 2026. URL https://arxiv.org/abs/2603.01514.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yihan He, Qishuo Yin, Yuan Cao, Jianqing Fan, and Han Liu. A theory on flow matching with neural networks, 2026. URL https://arxiv.org/abs/2606.10089.

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

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes, 2025. URL https://arxiv.org/abs/2410.20587.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow using invertible 1x1 convolutions. In *Advances in neural information processing systems*, volume 31, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, volume 32, 2019.

Yongjae Lee and Woo Chang Kim. Concise formulas for the surface area of the intersection of two hyperspherical caps. *KAIST Technical Report*, 2014.

Yiheng Lin, James A Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online Adaptive Policy Selection in Time-Varying Systems: No-Regret via Contractive Perturbations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=hDajsofjRM.

Yiheng Lin, James A. Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Online policy optimization in unknown nonlinear systems. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3475–3522. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/lin24a.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Guan-Horng Liu, Yaron Lipman, Maximilian Nickel, Brian Karrer, Evangelos Theodorou, and Ricky T. Q. Chen. Generalized schrödinger bridge matching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SoismgeX7z.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022b. URL https://arxiv.org/abs/2209.03003.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows, 2026. URL https://arxiv.org/abs/2601.22158.

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

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23, 2010.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *International Conference on Machine Learning*, pages 4951–4960, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape. In *Algorithmic Learning Theory*, pages 671–691. PMLR, 2018.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: a review, 2014. URL https://arxiv.org/abs/1404.5886.

Neta Shaul, Uriel Singer, Itai Gat, and Yaron Lipman. Transition matching: Scalable and flexible generative modeling, 2025. URL https://arxiv.org/abs/2506.23589.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching, 2023. URL https://arxiv.org/abs/2303.16852.

Xiao Shou, Yanna Ding, and Jianxi Gao. Gradient flow matching for learning update dynamics in neural network training, 2025. URL https://arxiv.org/abs/2505.20221.

Uriel Singer and Yaron Lipman. Exploring the design space of transition matching, 2025. URL https://arxiv.org/abs/2512.12465.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2021.

Sophia Tang. Foundations of schrödinger bridges for generative modeling, 2026. URL https://arxiv.org/abs/2603.18992.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Joel A Tropp. *An introduction to matrix concentration inequalities*. arXiv preprint arXiv:1501.01571, 2015.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Archer Wang, Emile Anand, Yilun Du, and Marin Soljačić. Unsupervised decomposition and recombination with discriminator-driven diffusion models, 2026. URL https://arxiv.org/abs/2601.22057.

Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models, 2024. URL https://arxiv.org/abs/2406.12839.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance, 2017. URL https://arxiv.org/abs/1707.00087.

Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. In *International Conference on Machine Learning*, pages 10787–10797. PMLR, 2020.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vES22INUKm.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. In *Machine Learning*, volume 109, pages 467–492. Springer, 2020.

Outline of the Appendix:

- Appendix A lists the related work,
- Appendix B provides the mathematical background and notation used in the paper,
- Appendix C provides the proof for the approximation error of the neural network,
- Appendix D provides the proof of the generalization error,
- Appendix E provides the proof of the rapid convergence of gradient descent,
- Appendix F provides the final end-to-end error analysis,
- Appendix G provides key properties of the velocity function used in the proofs,
- Appendix H contains experiments to verify the bounds established in the paper.

A Related Work

Flow matching and continuous-time transport. Flow matching, introduced by Lipman et al. (2023), and its variants—normalizing flows (Papamakarios et al., 2021), rectified flow (Liu et al., 2022a), OT-coupled flow matching (Tong et al., 2024; Pooladian et al., 2023), geometry-aware variants (Chen and Lipman, 2024), and stochastic interpolants (Albergo and Vanden-Eijnden, 2023; Albergo et al., 2025)—learn a time-dependent velocity transporting a simple source to a target distribution, yielding simulation-free training and ODE-based sampling-alternative to diffusion models. Flow matching belongs to a broader and rapidly growing landscape of generative modeling methods, including generator matching (Holderrieth et al., 2025), transition matching (Shaul et al., 2025; Singer and Lipman, 2025), mean flows (Geng et al., 2025), drifting models (Deng et al., 2026), and pixel mean flows (Lu et al., 2026). Flow matching also sits within a broader line of generative modeling methods built from different choices of interpolants and probability paths, including linear interpolants (Liu et al., 2022b) and Schrödinger bridge-based formulations (Shi et al., 2023; Tang, 2026; Liu et al., 2024). Our work focuses on the canonical formulation of Lipman et al. (2023) under linear interpolants, though our techniques should extend to other interpolant-based variants under analogous regularity assumptions.

Theoretical analyses of flow matching. Recent work has established approximation guarantees for neural networks learning velocity fields (Pooladian et al., 2023; Tong et al., 2024). Benton et al. (2024) bound how velocity-field error propagates to generation error, taking the velocity as given. Albergo et al. (2023) analyzes the Wasserstein distance between generated and target distributions under the assumption that the learned velocity field is exactly K-Lipschitz. Chen et al. (2023) established distribution recovery guarantees under Lipschitz assumptions on the backward score function and provided finite-sample analyses for diffusion models under assumptions of a bounded Hessian. Finally, Zhou and Liu (2025) gave an end-to-end statistical guarantee by combining approximation, generalization, and Wasserstein recovery under bounded-support and Lipschitz-score assumptions; the velocity, however, is obtained via empirical risk minimization rather than an explicit optimization analysis. Our work plugs the missing optimization piece into their pipeline: we prove that projected gradient descent, applied to a network class compatible with their approximation construction, attains the velocity-field accuracy their downstream analysis requires.

Optimization theory for generative models. In score-based diffusion, Wang et al. (2024) establishes exponential convergence of gradient descent on over-parameterized networks; their argument relies on the score-matching loss but does not directly apply to flow matching's displacement-regression objective. Similarly, Chaudhari et al. (2025); Cheng et al. (2024) analyze JKO-type flow models as proximal gradient descent in Wasserstein space at the distribution-level abstraction rather than network parameters. Our convergence proof builds on the over-parameterization techniques of Allen-Zhu et al. (2019), while handling the time-integrated loss, the linear interpolant $X_t = (1-t)X_0 + tX_1$, and the velocity target $X_1 - X_0$, all of which differ structurally from score matching. Chakraborty et al. (2026); Goel et al. (2026)

Statistical guarantees for diffusion and score-based models. Chen et al. (2023) analyzed how accurately diffusion-based samplers can recover the target distribution given an approximate score, under minimal data assumptions, while a separate line of work has bounded distribution-recovery error from finitely many samples under smoothness conditions (Chen et al., 2020). These results parallel our statistical-estimation step but do not address gradient-based training.

B Mathematical Background and Notation

Theorem B.1 (Gershgorin circle theorem)**.** *Let $\mathbf{K} \in \mathbb{C}^{n\times n}$ be an $n \times n$ matrix. For each row $i \in [n]$, define the Gershgorin disc:*

$$D_i = \{z \in \mathbb{C} : |z - \mathbf{K}_{ii}| \leq R_i\},$$

where $R_i = \sum_{j\neq i} |\mathbf{K}_{ij}|$ is the sum of absolute values of off-diagonal entries in row i. Then:

1. *Every eigenvalue λ of K lies in at least one Gershgorin disc: $\lambda \in \bigcup_{i=1}^{n} D_i$.*
2. *If k discs form a connected component disjoint from the other $n-k$ discs, then exactly k eigenvalues (counted with multiplicity) lie in this union.*
3. *For symmetric real matrices, all eigenvalues are real, so they lie on the real line within the disc projections.*

Corollary B.2 (Positive definiteness via Gershgorin)**.** *If $\mathbf{K} \in \mathbb{R}^{n\times n}$ is symmetric and for all $i \in [n]$:*

$$\mathbf{K}_{ii} > R_i = \sum_{j\neq i} |\mathbf{K}_{ij}|,$$

then $\mathbf{K}$ is strictly positive definite (all eigenvalues $\lambda > 0$).

Proof. Since $\mathbf{K}$ is symmetric real, all eigenvalues are real. By Gershgorin, every eigenvalue λ satisfies $|\lambda - \mathbf{K}_{ii}| \leq R_i$ for some i. Hence, $\lambda \geq \mathbf{K}_{ii} - R_i > 0$, where the strict inequality follows by assumption. □

Lemma B.3 (Hanson-Wright inequality)**.** *Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ where X_i are independent sub-gaussian random variables with $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a fixed matrix. Then for all $t \geq 0$:*

$$\Pr\left(|\mathbf{X}^\top \mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{X}^\top \mathbf{A}\mathbf{X}]| \geq t\right) \leq 2\exp\left(-c \cdot \min\left(\frac{t^2}{K^4\|\mathbf{A}\|_F^2}, \frac{t}{K^2\|\mathbf{A}\|_F}\right)\right),$$

where $c > 0$ is an absolute constant, $\|\mathbf{A}\|_F$ is the Frobenius norm, and $\|\mathbf{A}\|$ is the operator norm (Boucheron et al., 2013).

Lemma B.4 (Matrix Bernstein inequality)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{d\times d}$ be independent random symmetric matrices with $\mathbb{E}[\mathbf{X}_i] = 0$ and $\|\mathbf{X}_i\| \leq R$ almost surely for all i. Define the sum $\mathbf{S} = \sum_{i=1}^{n} \mathbf{X}_i$ and the variance parameter:*

$$\sigma^2 = \max\left\{\left\|\sum_{i=1}^{n} \mathbb{E}[\mathbf{X}_i^2]\right\|, \left\|\sum_{i=1}^{n} \mathbb{E}[\mathbf{X}_i^2]\right\|\right\}.$$

Then for all $t \geq 0$:

$$\Pr(\|\mathbf{S}\| \geq t) \leq 2d \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

In particular, when $t \leq \sigma^2/R$, we have $\Pr(\|\mathbf{S}\| \geq t) \leq 2d \cdot \exp(-t^2/(4\sigma^2))$ (Tropp, 2015).

Lemma B.5 (Weyl's inequality)**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric matrices with eigenvalues $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{B}) \geq \lambda_2(\mathbf{B}) \geq \cdots \geq \lambda_n(\mathbf{B})$ respectively. Then for all $i \in [n]$:*

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|,$$

where $\|\cdot\|$ denotes the operator norm (largest singular value).

Lemma B.6 (Gronwall's inequality)**.** *Given $f(t) \in [a, b]$ such that $\frac{\mathrm{d}f}{\mathrm{d}t} \leq \alpha f(t) + g(t)$ on $[a, b]$ and $\alpha \geq 0$, we have that*

$$f(b) \leq e^{\alpha(b-a)} f(a) + \int_a^b e^{\alpha(b-t)} g(t)\mathrm{d}t.$$

Proof. Note that $e^{-\alpha t}\frac{\mathrm{d}f}{\mathrm{d}t} - e^{-\alpha t}\alpha f(t) \leq e^{-\alpha t}g(t)$. Integrating both sides gives

$$e^{-\alpha b}f(b) - e^{-\alpha a}f(a) \leq \int_a^b e^{\alpha(b-t)}g(t)\mathrm{d}t.$$

Then, dividing by $e^{-\alpha b}$ completes the proof. □

B.1 Notation Table

Symbol	Description
π_0, π_1	Source and target distributions
$\nu_\theta(x,t)$	Neural network velocity field with parameters θ
$\nu^*(x,t)$	Optimal velocity field
n	Number of training samples
$\bar{m}$	Number of samples for Monte Carlo estimate of $l(x,v)$
N	Number of time discretization points
d	Ambient dimension
m	Neural network width
L	Number of hidden layers
T	Terminal time ($T < 1$)
δ_0	Minimum pairwise distance between training points
w_j	Time quadrature weights, $w_j = (t_j - t_{j-1})/T$
$w_{\min}, \beta_{\max}$	Minimum and maximum weights
Γ	Neural Network Lipschitz budget
x_{ij}	Interpolated training point: $x_{ij} = t_j x_i^{(1)} + (1-t_j)x_i^{(0)}$
y_i	Velocity target: $y_i = x_i^{(1)} - x_i^{(0)}$
$\mathcal{L}_{\text{emp}}(\theta)$	Empirical training loss
$f(\theta; i, j)$	Per-sample loss: $\|\nu_\theta(x_{ij}, t_j) - y_i\|^2$
$\mathcal{G}^{(k)}$	Good set: indices with above-average loss at iteration k
K	Total number of gradient descent iterations
η	Learning rate
$W_2(\cdot,\cdot)$	Wasserstein-2 distance
L_ν	Lipschitz constant of ν^*, $L_\nu = O((1-T)^{-3})$
L_f	Lipschitz constant of reference drift f_t
M_0, M_1	Bounds on velocity field magnitude and derivatives
L_{grad}	Gradient Lipschitz constant
a_0	Concentration exponent, $a_0 \in (1/2, 1)$
$\epsilon, \epsilon_{\text{train}}$	Target training loss
E_{approx}	Approximation error
E_{opt}	Optimization error
E_{disc}	Discretization error
$Z_t, \hat{Z}_t$	Continuous flow trajectories (optimal and learned)
$\pi_t, \hat{\pi}_t$	Distributions at time t (optimal and learned flows)

C Approximation Error

For completeness, we include the proof of Theorem 3.1 from Zhou and Liu (2025) below.

Proof of Theorem 3.1 (Theorem 3.1 from Zhou and Liu (2025)). Our goal is to construct a neural network $\hat{v} \in \mathsf{NN}$ that approximates the true vector field v^*, ultimately showing that

$$\|v^*(x,t) - \hat{v}(x,t)\|_{L^2(\pi_t)} \leq (\sqrt{d}+1)\epsilon$$

for a given approximation error ϵ. A key challenge is that the domain of $v^*(x,t)$, namely $\mathbb{R}^d \times [0,T]$, is unbounded. To handle this, we decompose $\mathbb{R}^d$ into a compact set $\mathcal{K}$ and its complement $\mathcal{K}^c$. The proof proceeds by constructing $\hat{v}$ to approximate v^* uniformly on $\mathcal{K} \times [0,T]$, while controlling the tail error on $\mathcal{K}^c$. On $\mathcal{K} \times [0,T]$, we construct $\hat{v}$ so that it approximates v^* in the L^∞ norm, whereas in $\mathcal{K}^c$, we define $\hat{v}(x,t) = 0$. The error on $\mathcal{K}^c$ is controlled by Lemma C.1.

C.1 Approximation on the hypercube

We show how to control the approximation error on $\mathcal{K} \times [0,T]$. Let $\mathcal{K} = \{x : \|x\|_\infty \leq R\}$ be a d-dimensional hypercube with edge length $2R > 0$, where R will be specified later. On $\mathcal{K} \times [0,T]$, we approximate each coordinate of the vector field separately. Writing

$$v^*(x,t) = [v_1^*(x,t), \ldots, v_d^*(x,t)]^\top,$$

we construct approximations for each coordinate map $v_k^*(x,t)$. We first rescale the inputs by $x' = \frac{1}{2R}(x + R\mathbf{1})$ and $t' = \frac{t}{T}$, where $\mathbf{1} = [1,\ldots,1]^\top$. This transformation maps the domain to $[0,1]^d \times [0,1]$ and can be implemented exactly using a single ReLU layer. By Lemma G.8, $v^*(x,t)$ is ζ-Lipschitz in x. Define the rescaled function

$$v(x',t') := v^*(2Rx' - R\mathbf{1}, Tt').$$

Under this transformation, v is $2\zeta R$-Lipschitz with respect to x'. Let $T\tau(R)$ denote the Lipschitz constant of $v(x',t')$ with respect to t' for $x' \in [0,1]^d$, where

$$\tau(R) := \sup_{t\in[0,T]} \sup_{x\in[-R,R]^d} \|\partial_t v^*(x,t)\|.$$

Lemma G.3 gives the bound

$$\tau(R) = O\left(\frac{d^{3/2}(R+1)}{(1-T)^4}\right).$$

Thus the problem reduces to approximating v on $[0,1]^d \times [0,1]$.

Next, partition $[0,1]^d$ into non-overlapping hypercubes of side length ϵ_1, and partition the time interval $[0,1]$ into sub-intervals of length ϵ_2. The values of ϵ_1 and ϵ_2 will be chosen according to the desired approximation accuracy. Let $N_1 = \lceil 1/\epsilon_1 \rceil$ and $N_2 = \lceil 1/\epsilon_2 \rceil$. For a multi-index $m = [m_1,\ldots,m_d]^\top \in [N_1]^d$, define

$$\bar{v}_i(x',t') = \sum_{m\in[N_1]^d,\, j\in[N_2]} v_i^*\left(\frac{2Rm}{N_1} - R\mathbf{1}, \frac{Tj}{N_2}\right) \Psi_{m,j}(x',t').$$

Here $\Psi_{m,j} : [0,1]^d \times [0,1]$ is a partition-of-unity function satisfying

$$\sum_{m\in[N_1]^d,\, j\in[N_2]} \Psi_{m,j}(x',t') \equiv 1.$$

We construct $\Psi_{m,j}$ as products of coordinate-wise trapezoidal functions:

$$\Psi_{m,j}(x',t') = \psi\left(3N_2\left(t' - \frac{j}{N_2}\right)\right) \prod_{i=1}^d \psi\left(3N_1\left(x_i' - \frac{m_i}{N_1}\right)\right),$$

where ψ is the trapezoid function

$$\psi(a) = \begin{cases} 1, & |a| < 1, \\ 2 - |a|, & |a| \in [1,2], \\ 0, & |a| > 2. \end{cases}$$

Lemma 10 of Chen et al. (2020) then establishes that the function $\bar{v}_i$ is an approximation to v_i, and that each $\bar{v}_i$ admits a ReLU neural network realization $\widehat{v}_i$ with arbitrarily small approximation error.

C.2 Neural Network Approximation Construction

Uniform approximation. The second assertion guarantees that for any prescribed worst-case L_∞ error $\varepsilon > 0$, there exists a ReLU network $\widehat{\boldsymbol{v}}_\theta$ such that

$$\sup_{\boldsymbol{x}', t' \in [0,1]^d \times [0,1]} \|\widehat{\boldsymbol{v}}_\theta(\boldsymbol{x}', t') - \boldsymbol{v}(\boldsymbol{x}', t')\|_\infty \leq \varepsilon.$$

For this, we choose the spatial and temporal partition sizes $\epsilon_1 = O(\frac{\varepsilon}{\zeta R})$ and $\epsilon_2 = O(\frac{\varepsilon}{\tau(R)})$.

Network construction and complexity. The coordinate-wise argument of Chen et al. (2020)'s Lemma 10 establishes Lipschitz continuity with respect to $\boldsymbol{x}'$ and t'. Each component $\bar{v}_i$ is implemented by a ReLU network $\widehat{v}_i$. Concatenating the component networks yields the vector-valued network $\widehat{\boldsymbol{v}}_\theta = [\widehat{v}_1, \ldots, \widehat{v}_d]^T$. The resulting neural network satisfies

$$L = O\left(\log \frac{1}{\varepsilon} + d\right), \quad M = O\left(\tau(R)(\zeta R)^d \varepsilon^{-(d+1)}\right)$$

$$J = O\left(\tau(R)(\zeta R)^d \varepsilon^{-(d+1)} \left(\log \frac{1}{\varepsilon} + d\right)\right)$$

where L denotes the depth, M the width, and J the sparsity. The weight magnitudes satisfy $\kappa = \max\{1, \zeta R, \tau(R)\}$, and the output magnitude satisfies

$$K = \sqrt{d} \max_i \sup_{(x,t) \in [-R,R]^d \times [0,T]} \|v_i^*(x,t)\| = O\left(\frac{\sqrt{d}R}{1-T}\right).$$

Lipschitz continuity and extension to $\mathbb{R}^d$. Combining the above construction with the input transformation layers $x \mapsto x'$ and $t \mapsto t'$ yields a network that is Lipschitz continuous in both space and time. Specifically, for any $x_1, x_2 \in \mathcal{K}$ and $t \in [0,T]$, we have

$$\|\widehat{v}_\theta(x_1, t) - \widehat{v}_\theta(x_2, t)\|_\infty \leq 10 d \zeta \|x_1 - x_2\|.$$

Similarly, for any $t_1, t_2 \in [0,T]$ and x satisfying $\|x\|_\infty \leq R$, we have that

$$\|\widehat{v}_\theta(x, t_1) - \widehat{v}_\theta(x, t_2)\|_\infty \leq 10 \tau(R) |t_1 - t_2|.$$

Finally, since the partition-of-unity functions $\Psi_{m,j}$ vanish outside $\mathcal{K}$, the network output satisfies $\widehat{v}_\theta(x,t) = 0$ when $\|x\|_\infty > R$. Consequently, the Lipschitz continuity in x extends to the entire space $\mathbb{R}^d$.

C.3 Bounding the approximation error

The L^2 error can be decomposed into two terms,

$$\begin{aligned}\|v^*(x,t) - \hat{v}_\theta(x,t)\|_{L^2(\pi_t)} = &\|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty \leq R\}\|_{L^2(\pi_t)} \\ &+ \|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty > R\}\|_{L^2(\pi_t)}.\end{aligned}$$

The first term can be bounded as

$$\|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty \leq R\}\|_{L^2(\pi_t)} \leq \sqrt{d} \sup_{(x,t) \in \mathcal{K} \times [0,T]} \|v^*(x,t) - \hat{v}_\theta(x,t)\|_\infty \leq \sqrt{d}\,\epsilon.$$

The second term is controlled using Lemma C.1. In particular, choosing $R = O\left(\sqrt{\log \frac{d}{\epsilon}}\right)$ yields

$$\|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty > R\}\|_{L^2(\pi_t)} \leq \epsilon.$$

Combining the two bounds gives $\|v^*(x,t) - \hat{v}_\theta(x,t)\|_{L^2(\pi_t)} \leq (\sqrt{d} + 1)\epsilon$.

Substituting the above choice of R into the network parameters above yields

$$L = O\left(d + \log\frac{1}{\epsilon}\right), \quad M = O\left(\frac{d^{3/2}(\log(d/\epsilon))^{\frac{d+1}{2}}}{(1-T)^4}\zeta^d\epsilon^{-(d+1)}\right),$$
$$J = O\left(\frac{d^{3/2}(\log(d/\epsilon))^{\frac{d+1}{2}}}{(1-T)^4}\zeta^d\epsilon^{-(d+1)}\left(\log\frac{1}{\epsilon} + d\right)\right),$$
$$K = O\left(\frac{\sqrt{d\log(d/\epsilon)}}{1-T}\right), \quad \kappa = O\left(\zeta\sqrt{\log(d/\epsilon)} \vee \frac{\sqrt{d^3\log(d/\epsilon)}}{(1-T)^4}\right),$$
$$\gamma_1 = 10d\zeta, \quad \gamma_2 = O\left(\frac{\sqrt{d^3\log(d/\epsilon)}}{(1-T)^4}\right),$$

which completes the proof. □

C.3.1 Bounding the truncation error

Lemma C.1 (Lemma A.1 in Zhou and Liu (2025)). *Under Assumption 1.1, for any $\epsilon > 0$, choosing $R = O\left(\sqrt{\log\frac{d}{\epsilon}}\right)$ ensures*

$$\|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty > R\}\|_{L^2(\pi_t)} \le \epsilon.$$

Proof. Because $\hat{v}_\theta(x,t) = 0$ for $\|x\|_\infty > R$, we have

$$\|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty > R\}\|_{L^2(\pi_t)} = \|v^*(x,t) \cdot \mathbb{1}\{\|x\|_\infty > R\}\|_{L^2(\pi_t)}.$$

For any $R > 0$, using the representation $v^*(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$, we obtain

$$\begin{aligned}
\int_{\{\|x\|_\infty > R\}} \|v^*(x,t)\|^2 \pi_t(\mathrm{d}x) &= \int_{\{\|x\|_\infty > R\}} \|\mathbb{E}[X_1 - X_0 \mid X_t = x]\|^2 \pi_t(\mathrm{d}x) \\
&\le \int_{\{\|x\|_\infty > R\}} \mathbb{E}[\|X_1 - X_0\|^2 \mid X_t = x]\pi_t(\mathrm{d}x) \\
&= \mathbb{E}\big[\mathbb{E}[\|X_1 - X_0\|^2 \mid X_t] \cdot \mathbb{1}\{\|X_t\|_\infty > R\}\big] \\
&\le \sqrt{\mathbb{E}[\|X_1 - X_0\|^4]} \cdot \sqrt{\Pr(\|X_t\|_\infty > R)},
\end{aligned}$$

where the second line follows from Jensen's inequality, and the last line uses Cauchy-Schwarz.

Next we bound the fourth moment of $\|X_1 - X_0\|$. Using the standard inequality $(a+b)^2 \le 2a^2 + 2b^2$, which follows from $(a-b)^2 \ge 0$, we obtain

$$\begin{aligned}
\mathbb{E}\|X_1 - X_0\|^4 &\le \mathbb{E}\big[(2\|X_1\|^2 + 2\|X_0\|^2)^2\big] \\
&\le \mathbb{E}[4\|X_1\|^4 + 4\|X_0\|^4] \\
&\le 4d^2 + 4\mathbb{E}\left[\left(\sum_{i=1}^d X_{0,i}^2\right)^2\right] \\
&= 4d^2 + 4\mathbb{E}\left[\sum_{k=1}^d X_{0,k}^4 + \sum_{i\ne j} X_{0,i}^2 X_{0,j}^2\right] \\
&= 8d(d+1).
\end{aligned}$$

It remains to control the tail probability of X_t. Then, by applying the union bound, and by noting that π_1 is supported on $[0,1]^d$, we can write

$$\begin{aligned}
\Pr(\|X_t\|_\infty > R) &= \Pr\left(\bigcup_{i=1}^{d} \{|X_{t,i}| > R\}\right) \\
&\le \sum_{i=1}^{d} \Pr(|X_{t,i}| > R) \\
&\le \sum_{i=1}^{d} \Pr(|X_{1,i}| + (1-t)|X_{0,i}| > R) \\
&\le \sum_{i=1}^{d} \Pr\left(|X_{0,i}| > \frac{R-1}{1-t}\right).
\end{aligned}$$

Because $X_{0,i}$ is standard Gaussian, it is sub-Gaussian with parameter 1. Hence

$$\begin{aligned}
\int_{\{\|x\|_\infty > R\}} \|v^*(x,t)\|^2 \pi_t(\mathrm{d}x) &\le \sqrt{8d(d+1)} \cdot \sqrt{\Pr\left(|X_{0,i}| > \frac{R-1}{1-t}\right)} \\
&\le 4\sqrt{d(d+1)} \cdot \exp\left(-\frac{(R-1)^2}{4(1-t)^2}\right).
\end{aligned}$$

Choosing R such that the right-hand side is bounded by ϵ^2 gives

$$R \ge 2(1-t)\left(2\log 2 + \frac{3}{2}\log(d+1) + 2\log\frac{1}{\epsilon}\right)^{1/2} + 1.$$

Therefore, taking $R = O(\sqrt{\log \frac{d}{\epsilon}})$ ensures $\|(v^*(x,t) - \hat{v}_\theta(x,t)) \cdot \mathbb{1}\{\|x\|_\infty > R\}\|_{L^2(\pi_t)} \le \epsilon$, which completes the proof. □

D GENERALIZATION ERROR

Lemma D.1 (Lemma 4.3 from Zhou and Liu (2025)). *The covering number (Def. 4.2) of $\mathcal{H}$ satisfies*

$$\log \mathcal{N}\big(\delta, \mathcal{H}, \|\cdot\|_{L^\infty([0,1]^d)}\big) \lesssim JL \log\left(\frac{(K+\sqrt{d})\, LM_\kappa \sqrt{\log \frac{K^2+d}{\delta}}}{\delta}\right).$$

Proof. First, from Lemma 5.3 of Chen et al. (2022), we have that for the neural network class NN satisfies

$$\log \mathcal{N}(\delta, \mathrm{NN}, \|\cdot\|_{L^\infty([-D,D]^d\times[0,1])}) \lesssim JL \log\left(\frac{LM(D\vee 1)\kappa}{\delta}\right). \tag{17}$$

The key insight for evaluating the covering number of $\mathcal{H}$ is to show that a δ-covering of NN induces a $C(\delta)$-covering of $\mathcal{H}$, where $C(\delta)$ is a function of δ. In this fashion, assume that there are two neural networks v_1 and v_2 satisfying $\|v_1 - v_2\|_{L^\infty([-D,D]^d\times[0,1])} \le \delta$, and the burden of the proof is to show that there exists a function $C(\cdot)$ such that $\|\ell(\cdot, v_1) - \ell(\cdot, v_2)\|_{L^\infty([0,1]^d)} \le C(\delta)$. For this, rewrite $\ell(x, v)$ using

$$\ell(x,v) = \frac{1}{T}\int_0^T \int \|x-x_0\|^2 - 2(x-x_0)^\top v(tx+(1-t)x_0, t) + \|v(tx+(1-t)x_0,t)\|^2 \pi_0(\mathrm{d}x_0)\mathrm{d}t,$$

which admits the following upper bound

$$\begin{aligned}|\ell(x,v_1) - \ell(x,v_2) \le \underbrace{\frac{2}{T}\int_0^T \int \|x-x_0\|\cdot\|v_1-v_2\|\pi_0(\mathrm{d}x_0)\mathrm{d}t}_{(A)} \\ + \underbrace{\frac{1}{T}\int_0^T \int \|v_1-v_2\|\cdot\|v_1+v_2\|\pi_0(\mathrm{d}x_0)\mathrm{d}t}_{(B)},\end{aligned}$$

which follows from Cauchy-Schwarz and the identity $\|x_1 - x_2\|^2 = (x_1 - x_2)^\top(x_1 + x_2)$, and the inputs of v_1 and v_2 are omitted for brevity when there is no ambiguity in the inputs.

To upper bound term (A), the Cauchy-Schwarz inequality implies

$$\begin{aligned}\frac{1}{T}\int_0^T \int \|x-x_0\|\cdot\|v_1-v_2\|\pi_0(\mathrm{d}x_0)\mathrm{d}t \le \sqrt{\frac{1}{T}\int_0^T \int \|x-x_0\|^2\pi_0(\mathrm{d}x_0)\mathrm{d}t} \\ \cdot \sqrt{\frac{1}{T}\int_0^T \int \|v_1-v_2\|^2\pi_0(\mathrm{d}x_0)\mathrm{d}t}.\end{aligned}$$

Since $x \in [0,1]^d$ and x_0 is a standard Gaussian variable, the following inequality holds

$$\sqrt{\frac{1}{T}\int_0^T \int \|x-x_0\|^2\pi_0(\mathrm{d}x_0)\mathrm{d}t} \lesssim \sqrt{d}.$$

Then, via the change of variables $x_t = tx + (1-t)x_0$,

$$\begin{aligned}\frac{1}{T}\int_0^T \int \|v_1(tx+(1-t)x_0,t) - v_2(tx+(1-t)x_0,t)\|^2\pi_0(\mathrm{d}x_0)\mathrm{d}t \\ = \frac{1}{T}\int_0^T \int \|v_1(x_t,t) - v_2(x_t,t)\|^2 \pi_{t|1}(\mathrm{d}x_t|X_1 = x)\mathrm{d}t,\end{aligned}$$

where $\pi_{t|1}$ is the distribution of x_t conditioned on $X_1 = x$. Next, partition $\mathbb{R}^d$ into two subsets $\{x_t \in \mathbb{R}^d : \|x_t\|_\infty \le D\}$ and its complement $\{x_t \in \mathbb{R}^d : \|x_t\|_\infty > D\}$, which gives

$$\begin{aligned}\int \|v_1(x_t,t)-v_2(x_t,t)\|^2 \pi_{t|1}(\mathrm{d}x_t|X_1=x) &= \int_{\{\|x_t\|_\infty \le D\}} \|v_1(x_t,t)-v_2(x_t,t)\|^2 \pi_{t|1}(\mathrm{d}x_t|X_1=x) \\ &\quad + \int_{\{\|x_t\|_\infty > D\}} \|v_1(x_t,t)-v_2(x_t,t)\|^2 \pi_{t|1}(\mathrm{d}x_t|X_1=x) \\ &\lesssim \delta^2 + K^2 \Pr[\|tx + (1-t)X_0\|_\infty > D].\end{aligned}$$

Using the tail bound for Gaussian random variables gives

$$\int \|v_1(x_t,t) - v_2(x_t,t)\|^2 \pi_{t|1}(\mathrm{d}x_t|X_1=x) \lesssim \delta^2 + K^2 d \exp\left(-\frac{(D-1)^2}{2(1-t)^2}\right).$$

Therefore,

$$\begin{aligned}\frac{1}{T}\int_0^T \int \|x - x_0\| \cdot \|v_1 - v_2\| \pi_0(\mathrm{d}x_0)\mathrm{d}t &\lesssim \sqrt{d} \cdot \sqrt{\delta^2 + K^2 d \exp\left(-\frac{(D-1)^2}{2(1-t)^2}\right)} \\ &\lesssim \sqrt{d} \cdot \left(\delta + K\sqrt{d}\exp\left(-\frac{(D-1)^2}{4(1-t)^2}\right)\right) \\ &\lesssim \sqrt{d}\delta + Kd\exp\left(-\frac{(D-1)^2}{4}\right),\end{aligned}$$

where the second inequality uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a \ge 0, b \ge 0$, and the third inequality follows from the fact that $t \in [0,T]$.

Next, to upper bound term (B), using the Cauchy-Schwarz inequality gives

$$\begin{aligned}\frac{1}{T}\int_0^T \int \|v_1 - v_2\| \cdot \|v_1 + v_2\| \pi_0(\mathrm{d}x_0)\mathrm{d}t &\le \sqrt{\frac{1}{T}\int_0^T \int \|v_1 - v_2\|^2 \pi_0(\mathrm{d}x_0)\mathrm{d}t} \\ &\quad \cdot \sqrt{\frac{1}{T}\int_0^T \int \|v_1 + v_2\|^2 \pi_0(\mathrm{d}x_0)\mathrm{d}t} \\ &\lesssim K\left(\delta + K\sqrt{d}\exp\left(-\frac{(D-1)^2}{4}\right)\right),\end{aligned}$$

where the second inequality follows by applying the same argument for the upper bound of term (A).

Together, this implies that

$$\sup_{x\in[0,1]^d} |\ell(x,v_1) - \ell(x,v_2)| \lesssim (K+\sqrt{d})\left(\delta + K\sqrt{d}\exp\left(-\frac{(D-1)^2}{4}\right)\right).$$

Thus, a δ-covering of the neural network class NN with respect to $\|\cdot\|_{L^\infty([-D,D]^d\times[0,1])}$ imduces a $C(K+\sqrt{\delta})(\delta + K\sqrt{d}\exp(-(D-1)^2/4))$-covering of $\mathcal{H}$, where C is a universal constant. Since

$$K\sqrt{d}\exp(-(D-1)^2/4) < \frac{\delta}{2C(K+\sqrt{d})},$$

we have that

$$D \ge 2\sqrt{\log\frac{2CK\sqrt{d}(K+\sqrt{d})}{\delta}} + 1 = D(\delta).$$

Therefore, a $\frac{\delta}{2C(K+\sqrt{d})}$-covering of NN in the $\|\cdot\|_{L^\infty([-D(\delta),D(\delta)]^d\times[0,T])}$-norm induces a δ-covering of $\mathcal{H}$, which implies that

$$\begin{aligned}\log \mathcal{N}(\delta, \mathcal{H}, \|\cdot\|_{L^\infty([0,1]^d)}) &\le \log \mathcal{N}\left(\frac{\delta}{2C(K+\sqrt{d})}, \mathrm{NN}, \|\cdot\|_{L^\infty([-D(\delta),D(\delta)]^d\times[0,T])}\right) \\ &\lesssim JL\log\left(\frac{2C(K+\sqrt{d})LMD(\delta)\kappa}{\delta}\right) \\ &\lesssim JL\log\left(\frac{(K+\sqrt{d})LM\kappa\sqrt{\log((K^2+d)/\delta)}}{\delta}\right),\end{aligned}$$

which completes the proof. □

Lemma D.2 (Uniform bound on the loss; Lemma B.1 of Zhou and Liu (2025))**.** *For any* $v \in \mathrm{NN}(L, M, J, K, \kappa, \gamma_1, \gamma_2)$,

$$\sup_{x\in[0,1]^d} \ell(x, v) \lesssim d + K^2.$$

Proof. Recall that

$$\ell(x, v) = \frac{1}{T}\int_0^T \int \|x - x_0 - v(tx + (1-t)x_0, t)\|^2 \,\pi_0(\mathrm{d}x_0)\,\mathrm{d}t.$$

Using the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned}\ell(x, v) &\lesssim \frac{1}{T}\int_0^T \int \|x - x_0\|^2 \pi_0(\mathrm{d}x_0)\,\mathrm{d}t + \frac{1}{T}\int_0^T \int \|v(tx + (1-t)x_0, t)\|^2 \,\pi_0(\mathrm{d}x_0)\,\mathrm{d}t \\ &\le \frac{1}{T}\int_0^T \int \|x - x_0\|^2 \pi_0(\mathrm{d}x_0)\,\mathrm{d}t + \sup_{z,t}\|v(z,t)\|^2.\end{aligned}$$

Since $x \in [0,1]^d$ and $x_0 \sim \pi_0 = N(0, I_d)$, we have

$$\begin{aligned}\int \|x - x_0\|^2 \pi_0(dx_0) &= \mathbb{E}_{\pi_0}\|x - X_0\|^2 \\ &\lesssim d.\end{aligned}$$

Moreover, by the definition of the network class, $\sup_{z,t}\|v(z,t)\| \le K$. Therefore,

$$\ell(x, v) \lesssim d + K^2.$$

Taking the supremum over $x \in [0,1]^d$ gives the claim. □

D.1 Proof of Theorem 4.3

We reproduce the proof of Theorem 4.3, following Theorem 4.4 from Zhou and Liu (2025), with notation adapted to our setting.

Proof. The proof begins by bounding the data estimation error. For this, we define the loss class

$$\mathcal{H} = \{\ell(\cdot, v) : v \in \mathrm{NN}\}.$$

Since π_1 is supported on $[0,1]^d$, Lemma D.2 gives a uniform bound

$$\sup_{v\in\mathrm{NN}}\sup_{x\in[0,1]^d} \ell(x, v) \lesssim d + K^2.$$

Now, let $N_1 \coloneqq \mathcal{N}\left(\tau, \mathcal{H}, \|\cdot\|_{L^\infty([0,1]^d)}\right)$, and let $\{\ell_k\}_{k=1}^{N_1}$ be a τ-cover of $\mathcal{H}$ under $\|\cdot\|_{L^\infty([0,1]^d)}$. Then for any $v \in \mathrm{NN}$, there exists some ℓ_k such that

$$\|\ell(\cdot, v) - \ell_k\|_{L^\infty([0,1]^d)} \le \tau.$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\left(L(v)-\ell(x_{1,i},v)\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\ell(X_1,v)]-\ell(x_{1,i},v)\right)$$
$$\leq\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\ell_k(X_1)]-\ell_k(x_{1,i})\right)+2\tau.$$

Taking the supremum over $v\in\mathrm{NN}$ gives

$$\sup_{v\in\mathrm{NN}}\frac{1}{n}\sum_{i=1}^{n}\left(L(v)-\ell(x_{1,i},v)\right)\leq\max_{1\leq k\leq N_1}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\ell_k(X_1)]-\ell_k(x_{1,i})\right)+2\tau.$$

For each fixed cover element ℓ_k, define

$$Z_{i,k}:=\mathbb{E}[\ell_k(X_1)]-\ell_k(x_{1,i}).$$

Then $\mathbb{E}[Z_{i,k}]=0$. Moreover, since $\|\ell_k\|_{L^\infty([0,1]^d)}\leq B$, we have $|Z_{i,k}|\leq 2B$. Thus, by Hoeffding's inequality, we have that for any $s>0$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}Z_{i,k}>s\right]\leq\exp\left(-\frac{ns^2}{2B^2}\right),$$

Union bounding over the N_1 cover elements gives

$$\Pr\left[\max_{1\leq k\leq N_1}\frac{1}{n}\sum_{i=1}^{n}Z_{i,k}>s\right]\leq N_1\exp\left(-\frac{ns^2}{2B^2}\right).$$

Choosing $s=\sqrt{\frac{2B^2\log(N_1/\delta)}{n}}$ makes the right-hand side at most δ. Therefore, with probability at least $1-\delta$, we have the bound

$$\max_{1\leq k\leq N_1}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\ell_k(X_1)]-\ell_k(x_{1,i})\right)\leq\sqrt{\frac{2B^2\log(N_1/\delta)}{n}}.$$

Combining this with the covering approximation error gives

$$\sup_{v\in\mathrm{NN}}\frac{1}{n}\sum_{i=1}^{n}\left(L(v)-\ell(x_{1,i},v)\right)\leq\sqrt{\frac{2B^2\log(N_1/\delta)}{n}}+2\tau.$$

Next, for bounding the Monte Carlo error in approximating $\ell(x,v)$, the difficulty is that

$$r((x_0,t),v,x):=\|x-x_0-v(tx+(1-t)x_0,t)\|^2$$

can be unbounded as $x_0\sim N(0,I_d)$. To handle this, we introduce the truncated version

$$r^{\mathrm{tr}}((x_0,t),v,x):=\|x-x_0-v(tx+(1-t)x_0,t)\|^2\cdot\mathbb{1}\{\|x_0\|_\infty\leq R\},$$

and define $\ell^{\mathrm{tr}}(x,v)=\mathbb{E}[r^{\mathrm{tr}}((x_0,t),v,x)]$ and $\hat{\ell}^{\mathrm{tr}}(x,v)=\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}r^{\mathrm{tr}}((x_{0,j},t_j),v,x)$. Then

$$\ell(x,v)-\hat{\ell}(x,v)=\underbrace{\ell(x,v)-\ell^{\mathrm{tr}}(x,v)}_{\text{truncation error I}}+\underbrace{\ell^{\mathrm{tr}}(x,v)-\hat{\ell}^{\mathrm{tr}}(x,v)}_{\text{statistical error}}+\underbrace{\hat{\ell}^{\mathrm{tr}}(x,v)-\hat{\ell}(x,v)}_{\text{truncation error II}}.$$

We first bound *truncation error I*. By Cauchy–Schwarz,

$$\ell(x,v)-\ell^{\mathrm{tr}}(x,v)=\mathbb{E}[r((x_0,t),v,x)\cdot\mathbb{1}\{\|x_0\|_\infty>R\}]$$
$$\leq\sqrt{\mathbb{E}[r((x_0,t),v,x)^2]}\sqrt{\Pr(\|x_0\|_\infty>R)}.$$

We first bound the fourth-moment factor. Since $x\in[0,1]^d$, we have $\|x\|^2\leq d$, and by the definition of the network class, $\sup_{z,t}\|v(z,t)\|\leq K$. Hence

$$\mathbb{E}\big[r((x_0,t),v,x)^2\big]=\mathbb{E}\big[\|x-x_0-v(tx+(1-t)x_0,t)\|^4\big]$$
$$\lesssim\|x\|^4+\mathbb{E}\|x_0\|^4+\sup_{z,t}\|v(z,t)\|^4$$
$$\lesssim d^2+K^4,$$

where we used $x_0 \sim N(0, I_d)$, so $\mathbb{E}\|x_0\|^4 = d^2 + 2d \lesssim d^2$. Therefore,

$$\left(\mathbb{E}\left[r((x_0,t),v,x)^2\right]\right)^{1/2} \lesssim d + K^2.$$

Moreover, since $x_0 \sim N(0, I_d)$, its coordinates $(x_{0,1}, \ldots, x_{0,d})$ are independent standard Gaussian random variables. Hence $\{\|x_0\|_\infty > R\} = \bigcup_{j=1}^d \{|x_{0,j}| > R\}$. Then, by the union bound and the standard Gaussian tail estimate, we have that

$$\begin{aligned}\Pr(\|x_0\|_\infty > R) &= \Pr\left(\bigcup_{j=1}^d \{|x_{0,j}| > R\}\right) \\ &\le \sum_{j=1}^d \Pr(|x_{0,j}| > R) \\ &\le 2d\exp(-R^2/2).\end{aligned}$$

Combining the two estimates gives

$$\ell(x,v) - \ell^{\mathrm{tr}}(x,v) \lesssim (d+K^2)\sqrt{d}\,\exp(-R^2/4).$$

Thus, for any fixed $R > 0$, truncation error (I) is exponentially small in R^2, up to polynomial factors in d and K. We defer the choice of R until after bounding the statistical and truncation error II terms, since they also depend on R.

We then bound *truncation error II*. Note that

$$\widehat{\ell}(x,v) - \widehat{\ell}^{\mathrm{tr}}(x,v) = \frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} r((x_{0,j},t_j),v,x)\cdot \mathbb{1}\{\|x_{0,j}\|_\infty > R\}.$$

In particular, this term is exactly zero on the event

$$\mathcal{E}_R := \left\{\max_{1\le j\le \bar{m}} \|x_{0,j}\|_\infty \le R\right\}.$$

Since $x_{0,j} \sim N(0, I_d)$, the Gaussian union bound gives $\Pr(\|x_{0,j}\|_\infty > R) \le 2d\exp(-R^2/2)$. Applying another union bound over the $\bar{m}$ Monte Carlo samples, we have

$$\begin{aligned}\Pr(\mathcal{E}_R^c) &= \Pr\left(\max_{1\le j\le \bar{m}} \|x_{0,j}\|_\infty > R\right) \\ &\le \sum_{j=1}^{\bar{m}} \Pr(\|x_{0,j}\|_\infty > R) \\ &\le 2\bar{m}d\exp(-R^2/2).\end{aligned}$$

Thus, we have that for all $v \in \mathrm{NN}$,

$$\Pr[\widehat{\ell}(x,v) - \widehat{\ell}^{\mathrm{tr}}(x,v) = 0] \ge 1 - 2\bar{m}d\exp(-R^2/2).$$

Next we bound the *statistical error.* Fix a target sample $x_{1,i}$. We use a standard covering argument, as in Zhou and Liu (2025); Anand et al. (2025), applied to the truncated Monte Carlo losses.

On the event $\|x_0\|_\infty \le R$, since $x_{1,i} \in [0,1]^d$, we have

$$\begin{aligned}\|x_{1,i} - x_0\| &\le \|x_{1,i}\| + \|x_0\| \\ &\le \sqrt{d} + \sqrt{d}R \\ &= \sqrt{d}(R+1).\end{aligned}$$

Moreover, $\sup_{z,t} \|v(z,t)\| \le K$, and hence

$$\begin{aligned} r^{\mathrm{tr}}\big((x_0,t),v,x_{1,i}\big) &:= \|x_{1,i} - x_0 - v(tx_{1,i} + (1-t)x_0, t)\|^2 \cdot \mathbb{1}\{\|x_0\|_\infty \le R\} \\ &\le \Big(\sqrt{d}(R+1) + K\Big)^2 \\ &\lesssim d(R+1)^2 + K^2 =: B_R. \end{aligned}$$

Thus the truncated loss term is uniformly bounded by B_R.

Let $\{v_k\}_{k=1}^{N_R}$ be a $\frac{\tau}{2\sqrt{d}(R+1)+2K}$-cover of NN in the sup-norm over the truncated domain

$$\{(tx_{1,i} + (1-t)x_0, t) : \|x_0\|_\infty \le R,\ t \in [0,T]\}.$$

Then, for any $v \in \mathrm{NN}$, there exists v_k such that

$$\sup_{\|x_0\|_\infty \le R,\ t\in[0,T]} \|v(tx_{1,i} + (1-t)x_0, t) - v_k(tx_{1,i} + (1-t)x_0, t)\| \le \frac{\tau}{2\sqrt{d}(R+1) + 2K}.$$

Using $\big|\|a-b\|^2 - \|a-c\|^2\big| \le \|b-c\|\big(2\|a\| + \|b\| + \|c\|\big)$, with $a = x_{1,i} - x_0$, $b = v_k(tx_{1,i} + (1-t)x_0, t)$, and $c = v(tx_{1,i} + (1-t)x_0, t)$, we obtain

$$\begin{aligned} &\big|r^{\mathrm{tr}}\big((x_0,t),v_k,x_{1,i}\big) - r^{\mathrm{tr}}\big((x_0,t),v,x_{1,i}\big)\big| \\ &\qquad \le \|v_k(tx_{1,i} + (1-t)x_0, t) - v(tx_{1,i} + (1-t)x_0, t)\| \\ &\qquad\qquad \cdot (2\|x_{1,i} - x_0\| + \|v_k\|_\infty + \|v\|_\infty) \\ &\qquad \le \frac{\tau}{2\sqrt{d}(R+1) + 2K}\Big(2\sqrt{d}(R+1) + 2K\Big) \\ &\qquad = \tau. \end{aligned}$$

Therefore, this network cover induces a τ-cover of the truncated squared-loss functions.

For each fixed cover element v_k, the random variables $r^{\mathrm{tr}}\big((x_{0,j},t_j),v_k,x_{1,i}\big)$ for $j = 1,\dots,\bar{m}$, are i.i.d. and bounded by $B_R \lesssim d(R+1)^2 + K^2$. Hence, by Hoeffding's inequality and a union bound over the N_R cover elements (which we prove in Lemma D.3), we have that for any $\varepsilon > 0$,

$$\begin{aligned} &\Pr\left(\max_{1\le k\le N_R}\left|\mathbb{E}_{x_0,t}\big[r^{\mathrm{tr}}\big((x_0,t),v_k,x_{1,i}\big)\big] - \frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} r^{\mathrm{tr}}\big((x_{0,j},t_j),v_k,x_{1,i}\big)\right| > \varepsilon\right) \\ &\qquad\qquad \le N_R \exp\left(-\frac{\bar{m}\varepsilon^2}{8\big(d(R+1)^2 + K^2\big)^2}\right). \end{aligned}$$

Therefore, set $N_R \exp(-\frac{\bar{m}\varepsilon^2}{8(d(R+1)^2+K^2)^2}) = \frac{\delta}{2}$. Equivalently,

$$\varepsilon = \sqrt{\frac{8\big(d(R+1)^2 + K^2\big)^2 \log(2N_R/\delta)}{\bar{m}}}.$$

Therefore, with probability at least $1 - \delta/2$, we have that

$$\begin{aligned} &\max_{1\le k\le N_R}\left|\mathbb{E}_{x_0,t}\Big[r^{\mathrm{tr}}\big((x_0,t),v_k,x_{1,i}\big)\Big] - \frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} r^{\mathrm{tr}}\big((x_{0,j},t_j),v_k,x_{1,i}\big)\right| \\ &\qquad\qquad \le \sqrt{\frac{8\big(d(R+1)^2 + K^2\big)^2 \log(2N_R/\delta)}{\bar{m}}}. \end{aligned}$$

Now fix any $v \in \mathrm{NN}$ and choose its cover element v_k. Since the corresponding truncated squared losses differ by at most τ pointwise, both the expectations and empirical averages differ by at most τ. Therefore, we write that

$$\begin{aligned} \Big|\ell^{\mathrm{tr}}(x_{1,i},v) - \widehat{\ell}^{\mathrm{tr}}_{\bar{m}}(x_{1,i},v)\Big| &\le \Big|\ell^{\mathrm{tr}}(x_{1,i},v_k) - \widehat{\ell}^{\mathrm{tr}}_{\bar{m}}(x_{1,i},v_k)\Big| + 2\tau \\ &\lesssim B_R\sqrt{\frac{\log(N_R/\delta)}{\bar{m}}} + 2\tau. \end{aligned}$$

Taking the supremum over $v \in \mathrm{NN}$, we get

$$\sup_{v\in\mathrm{NN}} \left|\ell^{\mathrm{tr}}(x_{1,i}, v) - \widehat{\ell}^{\mathrm{tr}}_{\bar{m}}(x_{1,i}, v)\right| \lesssim B_R\sqrt{\frac{\log(N_R/\delta)}{\bar{m}}} + 2\tau.$$

Finally, by the standard covering-number bound for ReLU networks on a bounded domain,

$$\log N_R \lesssim JL \log\left(\frac{LM(R\vee 1)\kappa\big(2\sqrt{d}(R+1)+2K\big)}{\tau}\right).$$

Substituting $B_R \lesssim d(R+1)^2 + K^2$ and suppressing logarithmic factors gives

$$\sup_{v\in\mathrm{NN}} \left|\ell^{\mathrm{tr}}(x_{1,i}, v) - \widehat{\ell}^{\mathrm{tr}}_{\bar{m}}(x_{1,i}, v)\right| = \widetilde{O}\Big(\big(d(R+1)^2 + K^2\big)\sqrt{\bar{m}}\Big).$$

Choosing $R = \sqrt{2\log(2\bar{m}d/\delta)}$ ensures, by the Gaussian union bound over the $\bar{m}$ Monte Carlo samples, that $\max_{1\le j\le m}\|x_{0,j}\|_\infty \le R$ with probability at least $1-\delta/2$. On this event, truncation error II vanishes.

Moreover, the same choice of R makes truncation error I exponentially small:

$$(d+K^2)\sqrt{d}\, e^{-R^2/4} \lesssim (d+K^2)\sqrt{d}\cdot\sqrt{\frac{\delta}{2\bar{m}d}}.$$

Combining the two truncation bounds with the truncated statistical error, we obtain that, with probability at least $1-\delta$,

$$\begin{aligned}\sup_{v\in\mathrm{NN}} \left|\ell(x_{1,i}, v) - \widehat{\ell}_m(x_{1,i}, v)\right| &\lesssim (d+K^2)\sqrt{d}\cdot\sqrt{\frac{\delta}{2\bar{m}d}} + 2\tau \\ &\quad + \big(d(R+1)^2 + K^2\big)\sqrt{\frac{\log(N_R/\delta)}{\bar{m}}}.\end{aligned}$$

We now combine the data and Monte Carlo errors. From the previous analysis, with failure probability $\delta/2$ each, and taking a union bound, we obtain, with probability at least $1-2\delta$,

$$\begin{aligned}\sup_{v\in\mathrm{NN}} \left|L(v) - \widehat{L}_{n,\bar{m}}(v)\right| &\le \sup_{v\in\mathrm{NN}} \left|L(v) - \overline{L}(v)\right| + \sup_{v\in\mathrm{NN}} \left|\overline{L}(v) - \widehat{L}_{n,\bar{m}}(v)\right| \\ &\lesssim (d+K^2)\sqrt{\frac{\log(N_1/\delta)}{n}} + (d+K^2)\sqrt{\frac{\delta}{2\bar{m}}} \\ &\quad + \big(d(R+1)^2 + K^2\big)\sqrt{\frac{\log(N_2/\delta)}{\bar{m}}} + \tau.\end{aligned}$$

Next, we incorporate the approximation error and balance the terms appropriately. By the approximation bound already established in the main text, we have that

$$\inf_{v\in\mathsf{NN}} L(v) - L(v^*) \le (\sqrt{d}+1)^2\varepsilon^2.$$

Combining this with the previous analysis gives

$$\begin{aligned}L(\widehat{v}) - L(v^*) &\le L(\widehat{v}) - \inf_{v\in\mathsf{NN}} L(v) + \inf_{v\in\mathsf{NN}} L(v) - L(v^*) \\ &= \widetilde{O}\left(\frac{1}{(1-T)^4}\left(\frac{1}{n} + \frac{1}{\sqrt{\bar{m}n}} + \frac{(1/\varepsilon)^{\frac{d+1}{2}}}{\sqrt{\bar{m}}} + \frac{(1/\varepsilon)^{\frac{d+1}{2}}}{\sqrt{n}} + \varepsilon^2\right)\right),\end{aligned}$$

where we choose the network class as in Theorem 3.1 and suppress factors polynomial in d, ζ and logarithmic in $\bar{m}, n, (1-T)^{-1}$. Here, choosing $\varepsilon = n^{-\frac{1}{d+5}}$ balances the approximation term ε^2 with the outer estimation term $(1/\varepsilon)^{\frac{d+1}{2}}/\sqrt{n}$.

Hence

$$L(\widehat{v}) - L(v^*) = \widetilde{O}\left(\frac{1}{(1-T)^4}\left(n^{-\frac{2}{d+5}} + n^{\frac{d+1}{2(d+5)}} \cdot \frac{1}{\sqrt{\bar{m}}}\right)\right).$$

Therefore,

$$\frac{1}{T}\int_0^T \|\widehat{v}(\cdot,t) - v^*(\cdot,t)\|^2_{L^2(\pi_t)}\mathrm{d}t = \widetilde{O}\left(\frac{1}{(1-T)^4}\left(n^{-\frac{2}{d+5}} + n^{\frac{d+1}{2(d+5)}} \cdot \frac{1}{\sqrt{\bar{m}}}\right)\right).$$

This proves the theorem. □

Lemma D.3 (Lemma B.3 in Zhou and Liu (2025))**.** *Let $\mathcal{G}$ be a bounded function class, i.e. there exists a constant B such that for any $g \in \mathcal{G}$ and any x in its domain, $0 \le g(x) \le B$. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. random variables. For any $\delta \in (0,1)$ and $g \in \mathcal{G}$,*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}[g(X)]) > \epsilon\right] \le \exp\left(-\frac{n\epsilon^2}{2B^2}\right),$$

and

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (\mathbb{E}[g(X)] - g(X_i)) > \epsilon\right] \le \exp\left(-\frac{n\epsilon^2}{2B^2}\right).$$

Proof. The proof begins by computing the moment generating function of $\frac{1}{n}\sum_{i=1}^n (\mathbb{E}[g(X)] - g(X_i))$,

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}[g(X)])\right)\right] = \left(\mathbb{E}\left[\exp(\frac{\lambda}{n}(g(X_1) - \mathbb{E}[g(X)])\right]\right)^n,$$

and the identity follows from the fact that $x_1, \ldots, x_n$ are i.i.d. random variables.

Then, given an independent copy X_1' of X_1, note the following upper bound

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}(g(X_1) - \mathbb{E}[g(X)])\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\lambda}{n}(g(X_1) - \mathbb{E}_{X_1'}[g(X_1')])\right)\right] \\ &\le \mathbb{E}_{X_1, X_1'}\left[\exp\left(\frac{\lambda}{n}(g(X_1) - g(X_1'))\right)\right].\end{aligned}$$

Letting σ be an independent Rademacher random variable, note that $X_1 - X_1' \overset{D}{=} \sigma(X_1 - X_1')$, and hence

$$\begin{aligned}\mathbb{E}_{X_1, X_1'}\left[\exp\left(\frac{\lambda}{n}(g(X_1) - g(X_1'))\right)\right] &= \mathbb{E}_{X_1, X_1'}\left[\mathbb{E}_\sigma\left[\exp\left(\frac{\lambda}{n}\sigma(g(X_1) - g(X_1'))\right)\right]\right] \\ &\le \mathbb{E}_{X_1, X_1'}\left[\exp\left(\frac{\lambda^2(g(X_1) - g(X_1'))^2}{2n^2}\right)\right],\end{aligned}$$

Since $|g(X_1) - g(X_1')| \le B$, we therefore have that

$$\mathbb{E}_{X_1, X_1'}\left[\exp\left(\frac{\lambda^2(g(X_1) - g(X_1'))^2}{2n^2}\right)\right] \le \exp\left(\frac{\lambda^2 B^2}{2n^2}\right).$$

Together, this gives

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}[g(X)])\right)\right] \le \exp\left(\frac{\lambda^2 B^2}{2n}\right).$$

Then, Markov's inequality gives that

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}[g(X)]) > \epsilon\right] \le \exp\left(\frac{\lambda^2 B^2}{2n} - \lambda\epsilon\right).$$

Letting $\lambda = n\epsilon/B^2$ gives the first inequality.

The second inequality follows by the same argument. □

Lemma D.4. *Given a Rademacher random variable σ that takes the values $\{-1, 1\}$ with probability $1/2$. Then, for any $\lambda \in \mathbb{R}$, we have that $\mathbb{E}[e^{\lambda\sigma}] \leq e^{\lambda^2/2}$.*

Proof. By taking expectations and using the power-series expansion for the exponential, we have

$$\begin{aligned}
\mathbb{E}[e^{\lambda\sigma}] &= \frac{1}{2}[e^{-\lambda} + e^{\lambda}] \\
&= \frac{1}{2}\left[\sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty} \frac{(\lambda)^k}{k!}\right] \\
&= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k k!} \\
&= e^{\lambda^2/2},
\end{aligned}$$

which completes the proof. □

E Learning the Velocity Field

This section is devoted to the proof of Theorem 3.5, which we restate in its entirety in Theorem E.2. The key idea of the proof is to establish a semi-smoothness property, adapting the proof of Wang et al. (2024), and a lower bound of the gradient of the loss function, for which we find the largest weighted residual and use it decompose the row-gradient. We then show that there exists a row in which the selected term cannot be canceled using a random-neuron isolation event, and then us the Chernoff bound to yield the gradient lower bound. Owing to the length of the proof, we also provide an overview of the section:

- Appendix E.1 establishes the properties of the weights bestowed by the random initialization,
- Appendix E.2 proves forward-propagation bounds when perturbing the neural network,
- Appendix E.3 establishes the semi-smoothness property of the loss function,
- Appendix E.4 proves the gradient upper bound,
- Appendix E.5 proves that the projection is inactive within the neighborhood $\mathcal{N}_{\omega,\rho}$, and
- Appendix E.6 proves the gradient lower bound

We will rely on the following inequality in Theorem E.2.

Lemma E.1 (Young's inequality)**.** *For any non-negative real numbers a and b, and any conjugate exponents p and q such that $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, we have that $ab \le \frac{a^p}{p} + \frac{b^q}{q}$.*

Theorem E.2. *Let $\mathcal{L}(\theta) = \mathcal{L}_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\beta_j\|v_\theta(X_{ij}, t_j) - U_i\|^2$ where $U_i = x_i^{(1)} - x_i^{(0)}$. Let the projected gradient descent update over the trainable hidden layers be given by $\theta^{(k+1)} = \Pi_{\Theta_{\text{hid}}}(\theta^{(k)} - h\nabla\mathcal{L}(\theta^{(k)}))$, and define the projected-gradient mapping $G_h(\theta) := \frac{1}{h}(\theta - \Pi_{\Theta_{\text{hid}}}(\theta - h\nabla\mathcal{L}(\theta)))$. Now, fix a local neighborhood*

$$\mathcal{N}_{\omega,\rho} = \left\{\theta \in \Theta_{\text{hid}} : \|\theta - \theta^{(0)}\|_F \le \omega, \quad \max_{\ell\in[L]} \|\mathbf{W}_\ell - \mathbf{W}_\ell^{(0)}\|_{2,\infty} \le \frac{\rho}{\sqrt{m}}\right\}.$$

Assume that for every $\theta, \theta' \in \mathcal{N}_{\omega,\rho}$ that the following hold:

1. *Projected-gradient lower bound on the last hidden layer (Theorem E.44): there is a function $\mu(\theta) > 0$ such that $\|G_h(\theta)\|_F^2 \ge \mu(\theta)\mathcal{L}(\theta)$,*
2. *A gradient-mapping upper bound (Lemma E.29): $\|G_h(\theta)\|^2 \le \bar{\mu}\mathcal{L}(\theta)$,*
3. *Semi-smoothness (Corollary E.28): there exist constants $\alpha_{\omega,\rho}, b_{\omega,\rho}, c_{\omega,\rho} \ge 0$ such that*

$$\mathcal{L}(\theta') \le \mathcal{L}(\theta) + \langle\nabla\mathcal{L}(\theta), \theta' - \theta\rangle + \alpha_{\omega,\rho}\sqrt{\mathcal{L}(\theta)}\|\theta' - \theta\|_F + b_{\omega,\rho}\sqrt{\mathcal{L}(\theta)}\|\theta' - \theta\|_F^2 + c_{\omega,\rho}\|\theta' - \theta\|_F^2.$$

Let $\mu_* = \inf_{\theta\in\mathcal{N}_{\omega,\rho}}\mu(\theta)$. Then, if $\alpha_{\omega,\rho}^2 \le \frac{\mu_*}{4}, c_{\omega,\rho} \le \frac{1}{8h}$, and $b_{\omega,\rho}\sqrt{L(\theta^{(0)})} \le \frac{1}{8h}$, then every iterate that stays in $\mathcal{N}_{\omega,\rho}$ satisfies the contraction

$$\mathcal{L}(\theta^{(k+1)}) \le \left(1 - \frac{h\mu_*}{4}\right)\mathcal{L}(\theta^{(k)}).$$

Hence, we have that $\mathcal{L}(\theta^{(K)}) \le \epsilon_{train}$ whenever $K \ge \frac{4}{h\mu_*}\log\frac{\mathcal{L}(\theta^{(0)})}{\epsilon_{train}}$.

Proof. Let $L_k := L(\theta^{(k)})$ and $\Delta_k := \theta^{(k+1)} - \theta^{(k)} = -h\nabla\mathcal{L}(\theta^{(k)}) = -hG_h(\theta^{(k)})$. Using semi-smoothness with $\theta = \theta^{(k)}$ and $\theta' = \theta^{(k+1)}$, and the fact that for projected gradient descent, the

projection's optimality condition gives $\langle \nabla \mathcal{L}(\theta^{(k)}), \Delta_k \rangle \leq -\frac{1}{h}\|\Delta_k\|^2$, we have that

$$\begin{aligned}
L_{k+1} &\leq L_k + \langle \nabla \mathcal{L}(\theta^{(k)}), \Delta_k \rangle + \alpha_{\omega,\rho}\sqrt{L_k}\|\Delta_k\|_F + b_{\omega,\rho}\sqrt{L_k}\|\Delta_k\|_F^2 + c_{\omega,\rho}\|\Delta_k\|_F^2 \\
&\leq L_k - \left(\frac{1}{h} - c_{\omega,\rho} - b_{\omega,\rho}\sqrt{L_k}\right)\|\Delta_k\|_F^2 + \alpha_{\omega,\rho}\sqrt{L_k}\|\Delta_k\|_F \\
&= L_k - \left(\frac{1}{h} - c_{\omega,\rho} - b_{\omega,\rho}\sqrt{L_k}\right)\|\Delta_k\|_F^2 + \alpha_{\omega,\rho}\sqrt{L_k}\sqrt{2h}\cdot\frac{\|\Delta_k\|_F}{\sqrt{2h}} \\
&\leq L_k - \left(\frac{1}{h} - c_{\omega,\rho} - b_{\omega,\rho}\sqrt{L_k}\right)\|\Delta_k\|_F^2 + \frac{1}{4h}\|\Delta_k\|_F^2 + \alpha_{\omega,\rho}^2 h L_k \\
&= L_k - \left(\frac{3}{4h} - c_{\omega,\rho} - b_{\omega,\rho}\sqrt{L_k}\right)\|\Delta_k\|^2 + \alpha_{\omega,\rho}^2 h L_k
\end{aligned}$$

where the last inequality follow by absorbing the linear term by Young's inequality in Lemma E.1. Next, while the contraction holds inductively, we know that $L_k \leq L_0$. Therefore, $b_{\omega,\rho}\sqrt{L_k} \leq b_{\omega,\rho}\sqrt{L_0} \leq \frac{1}{8h}$. So, if $c_{\omega,\rho} \leq \frac{1}{8h}$ and $b_{\omega,\rho}\sqrt{L_k} \leq \frac{1}{8h}$, then we have that $\frac{3}{4h} - c_{\omega,\rho} - b_{\omega,\rho}\sqrt{L_k} \geq \frac{1}{2h}$. Hence,

$$L_{k+1} \leq L_k - \frac{1}{2h}\|\Delta_k\|_F^2 + \alpha_{\omega,\rho}^2 h L_k.$$

Since $\Delta_k = -hG_h(\theta^{(k)})$, we have that $\frac{1}{2h}\|\Delta_k\|^2 = \frac{h}{2}\|G_h(\theta^{(k)})\|^2 \geq \frac{h\mu_*}{2}L_k$, where the last bound follows by the gradient-mapping lower bound.

Together, this gives the contraction property:

$$L_{k+1} \leq \left(1 - \frac{h\mu_*}{2} + \alpha_{\omega,\rho}^2 h\right) L_k.$$

Finally, using $\alpha_{\omega,\rho}^2 \leq \frac{\mu_*}{4}$ yields that $L_{k+1} \leq \left(1 - \frac{h\mu_*}{4}\right) L_k$.

We apply Lemma E.34 after establishing the one-step contraction above and the gradient upper bound of Lemma E.29. Thus the assumptions used in the perturbation and semi-smoothness lemmas remain valid along the entire projected-gradient trajectory.

Finally, the exponential convergence rate follows by unrolling the inequality K times to get

$$L_K \leq \left(1 - \frac{h\mu_*}{4}\right)^K L_0 \leq e^{-\frac{h\mu_* K}{4}} L_0$$

The second part of the theorem follows by noting that the above quantity is bounded by ϵ_{train} when

$$K \geq \frac{4}{h\mu_*}\log\left(\frac{L_0}{\epsilon_{\text{train}}}\right),$$

which completes the proof. □

Lemma E.3 (Projection Inequality). *Let $\theta^+ = \Pi_\Theta(\theta - h\nabla\mathcal{L}(\theta))$. Then,*

$$\langle \nabla \mathcal{L}(\theta), \theta^+ - \theta \rangle \leq -\frac{1}{h}\|\theta^+ - \theta\|_F^2 \tag{18}$$

and

$$\|\theta^+ - \theta\|_F \leq h\|\nabla\mathcal{L}(\theta)\|_F. \tag{19}$$

Proof. Since Θ is convex, we have that for all $Z \in \Theta$,

$$\langle \theta - h\nabla\mathcal{L}(\theta) - \theta^+, Z - \theta^+ \rangle \leq 0.$$

Then, $0 \geq \langle \theta - h\nabla\mathcal{L}(\theta) - \theta^+, \theta - \theta^+ \rangle = \|\theta - \theta^+\|_F^2 - h\langle \nabla\mathcal{L}(\theta), \theta - \theta^+ \rangle$. Rearranging and dividing by h yields Equation 18. Moreover, since $\theta \in \Theta$ implies $\theta = \Pi_\Theta(\theta)$, we have that

$$\|\theta^+ - \theta\|_F = \|\Pi_\Theta(\theta - h\nabla\mathcal{L}(\theta)) - \Pi_\Theta(\theta)\|_F \leq h\|\nabla\mathcal{L}(\theta)\|_F,$$

which proves Equation 19. □

E.1 Random Initialization Lemmas

Let $\mathcal{I} \coloneqq [n] \times [N]$. For $p = (i, j) \in \mathcal{I}$, write $\bar{X}_p \coloneqq \bar{X}_{ij} = [X_{ij}^\top, t_j]^\top \in \mathbb{R}^{d+1}$.

At initialization, we have $q_{p,-1}^{(0)} \coloneqq \bar{X}_p$, $r_{p,0}^{(0)} \coloneqq \mathbf{W}_0^{(0)} \bar{X}_p$, and $q_{p,0}^{(0)} \coloneqq \sigma(r_{p,0}^{(0)})$. Also, for all $\ell = [L]$, we have $r_{p,\ell}^{(0)} \coloneqq \mathbf{W}_\ell^{(0)} q_{p,\ell-1}^{(0)}$, $q_{p,\ell}^{(0)} \coloneqq \sigma(r_{p,\ell}^{(0)})$. Next, define $H_p \coloneqq \|q_{p,0}^{(0)}\|_2$, and let $C > 0$ denote universal constants below. This subsection is devoted to proving the joint initialization event in Corollary E.12 and Remark E.13 which shows that a number of desirable properties hold true with high probability.

Lemma E.4 (Interpolant norm event)**.** *Suppose $X_i^{(0)} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $X_i^{(1)} \in [0,1]^d$, where $X_i^{(0)} \perp X_i^{(1)}$ is derived independently from the product coupling, and let $t_j \le T < 1$. Then, for every $(i, j) \in \mathcal{I}$, we have*

$$\Pr[\|\bar{X}_{ij}\|_2 = \Theta(\sqrt{d})] \ge 1 - 2nN\exp(-\Omega(d)),$$

as well as $\Pr[\|U_i\|_2 \coloneqq \|X_i^{(1)} - X_i^{(0)}\|_2 \le O(\sqrt{d}] \ge 1 - 2nN\exp(-\Omega(d))$. We denote the latter corresponding event by $\mathcal{E}_{data}$.

Proof. Write $X_{ij} = (1 - t_j)X_i^{(0)} + t_j X_i^{(1)}$. Let $\sigma_j \coloneqq 1 - t_j \ge 1 - T > 0$ with $y_{ij} = t_j X_i^{(1)}$. Conditional on $X_i^{(1)}$, we have $X_{ij} = \sigma_j G_i + y_{ij}$ where $G_i \sim \mathcal{N}(0, \mathbf{I}_d)$.

For the upper bound, we have $\|X_{ij}\|_2 \le \sigma_j \|G_i\|_2 + \|y_{ij}\|_2$. Since $X_i^{(1)} \in [0,1]^d$, we have $\|y_{ij}\|_2 \le \sqrt{d}$. Also, we have that $\Pr[\|G_i\|_2 > C\sqrt{d}] \le e^{-cd}$ for a sufficiently large universal C. Therefore, $\|X_{ij}\|_2 \le O(\sqrt{d})$ uniformly over i, j with probability at least $1 - nNe^{-\Omega(d)}$. Since $t_j \le 1$, we also have

$$\begin{aligned}\|\bar{X}_{ij}\|_2^2 &= \|X_{ij}\|_2^2 + t_j^2 \\ &\le O(d) + 1 \le O(d).\end{aligned}$$

Moreover, with the same probability, $\|U_i\|_2 \le \|X_i^{(1)}\|_2 + \|G_i\|_2 \le O(\sqrt{d})$.

It now remains to prove the lower bound for $\|X_{ij}\|_2$: first, suppose $t_j \le 1/2$. Let $\mathbf{P}_{ij}$ be the orthogonal projection onto the orthogonal complement of $X_i^{(1)}$. If $X_i^{(1)} = 0$, we can simply take $\mathbf{P}_{ij}$ to be any orthogonal projection of rank $d - 1$. Then, $\mathbf{P}_{ij} X_{ij} = (1 - t_j)\mathbf{P}_{ij} G_i$. Since $1 - t_j \ge 1/2$ and $P_{ij} G_i$ is a standard Gaussian in a subspace of dimension at least $d - 1$, we have that for all $i \in [n], j \in [N]$ that with probability at least $1 - nNe^{-\Omega(d)}$ that

$$\|X_{ij}\|_2 \ge \frac{1}{2}\|\mathbf{P}_{ij} G_i\|_2 \ge \Omega(\sqrt{d}).$$

Now, support $t_j > 1/2$. Let $v_i \coloneqq \frac{X_i^{(1)}}{\|X_i^{(1)}\|_2}$. Then, $\langle v_i, X_{ij}\rangle = (1 - t_j)\langle v_i, G_i\rangle + t_j \|X_i^{(1)}\|_2$. Since $\langle v_i, G_i\rangle \sim \mathcal{N}(0, 1)$, we have that $\Pr[|\langle v_i, G_i\rangle| \le \frac{C}{4}\sqrt{d}] \ge 1 - nNe^{-\Omega(d)}$. On this event, we can bound $\|X_{ij}\|_2$ using

$$\begin{aligned}\|X_{ij}\|_2 &\ge |\langle v_i, X_{ij}\rangle| \\ &\ge t_j c_1\sqrt{d}) - (1 - t_j)c_1\sqrt{d} = \left(\frac{c_1}{2} - \frac{c_1}{8}\right)\sqrt{d} \\ &\ge \Omega(\sqrt{d}),\end{aligned}$$

where the second line uses Assumption 3.4, completing the proof of the lemma. □

Lemma E.5 (First-layer feature norm preservation)**.** *Let $\mathbf{W}_0^{(0)} \in \mathbb{R}^{m \times (d+1)}$ have independent entries $(\mathbf{W}_0^{(0)})_{ab} \sim \mathcal{N}(0, 2/m)$. If $m \ge \Omega(\log \frac{nN}{\delta_{\text{init}}})$, then with probability at least $1 - \delta_{\text{init}}$, for every $(i, j) \in \mathcal{I}$, we have that*

$$\frac{1}{2}\|\bar{X}_{i,j}\|_2^2 \le H_{ij}^2 \le \frac{3}{2}\|\bar{X}_{ij}\|_2^2,$$

and therefore $H_{ij} = \Theta(\sqrt{d})$ for all $(i, j) \in \mathcal{I}$. Moreover, we denote this event by $\mathcal{E}_0$.

Proof. Fix $(i,j) \in \mathcal{I}$. Then, conditional on $\bar{X}_{i,j}$, each coordinate of $r_{ij,0}^{(0)} = \mathbf{W}_0^{(0)} \bar{X}_{i,j}$ is Gaussian with variance $\frac{2}{m}\|\bar{X}_{i,j}\|_2^2$. Thus, we have that $(r_{ij,0}^{(0)})_k = \sqrt{\frac{2}{m}}\|\bar{X}_{i,j}\|_2 g_k$ with $g_k \sim \mathcal{N}(0,1)$ independently over k. Therefore, we write

$$H_{ij}^2 = \|q_{ij,0}^{(0)}\|_2^2 = \sum_{k=1}^{m} \left(\max\{(r_{ij,0}^{(0)})_k, 0\}\right)^2 = \frac{2\|\bar{X}_{i,j}\|^2}{m} \sum_{k=1}^{m} g_k^2 \mathbb{1}\{g_k > 0\}.$$

Now, let $Y_k \coloneqq 2g_k^2\mathbb{1}\{g_k > 0\}$. Then, Y_k are i.i.d. sub-exponential random variables with $\mathbb{E}[Y_k] = 1$. Thus, by Bernstein's inequality, we have that

$$\Pr\left[\left|\frac{1}{m}\sum_{k=1}^{m} Y_k - 1\right| > \frac{1}{2}\right] \le 2e^{-\Omega(m)},$$

and hence

$$\Pr\left[\frac{1}{2}\|\bar{X}_{i,j}\|_2^2 \le H_{ij}^2 \le \frac{3}{2}\|\bar{X}_{i,j}\|_2^2\right] \ge 1 - 2e^{-\Omega(m)}.$$

Thus, union-bounding over $i \in [n]$ and $j \in [N]$ yields the claim when $m \ge \Omega(\log(nN/\delta_{\text{init}}))$. □

Lemma E.6 (Hidden spectral slack)**.** *Let $(\mathbf{W}_\ell^{(0)})_{ab} \sim \mathcal{N}(0, 2/m)$ for $\ell = 1, \dots, L$, independently. Suppose R_{hid} and s_{hid} satisfy $R_{\text{hid}} - s_{\text{hid}} > 2\sqrt{2}$. Then, there exists $c_{op} > 0$, depending on $R_{\text{hid}} - s_{\text{hid}} - 2\sqrt{2}$ such that*

$$\Pr\left[\max_{\ell \in [L]} \|\mathbf{W}_\ell^{(0)}\|_{op} \le R_{\text{hid}} - s_{\text{hid}}\right] \ge 1 - Le^{-c_{op}m},$$

and we denote this event by $\mathcal{E}_{\text{hid,op}}$.

Proof. Let $\mathbf{W}_\ell^{(0)} = \sqrt{\frac{2}{m}}\mathbf{G}_\ell$, where $\mathbf{G}_\ell \in \mathbb{R}^{m\times m}$ has i.i.d. $\mathcal{N}(0,1)$ entries. The standard Gaussian operator norm bound gives $\Pr[\|\mathbf{G}_\ell\|_{\text{op}} > 2\sqrt{m} + t] \le e^{-t^2/2}$. Therefore,

$$\Pr\left[\|\mathbf{W}_\ell^{(0)}\|_{\text{op}} > 2\sqrt{2} + t\sqrt{\frac{2}{m}}\right] \le e^{-\frac{t^2}{2}}.$$

Thus, let $\Delta_{\text{op}} \coloneqq R_{\text{hid}} - s_{\text{hid}} - 2\sqrt{2} > 0$, and choose $t = \sqrt{\frac{m}{2}}\Delta_{op}$. So, $2\sqrt{2} + t\sqrt{\frac{2}{m}} = 2\sqrt{2} + \Delta_{\text{op}} = R_{\text{hid}} - s_{\text{hid}}$. Hence,

$$\Pr\left[\|\mathbf{W}_\ell^{(0)}\|_{\text{op}} > R_{\text{hid}} - s_{\text{hid}}\right] \le e^{-\frac{m\Delta_{\text{op}}^2}{4}}.$$

Finally, union bounding over $\ell \in [L]$ yields the claim. □

Lemma E.7 (Hidden feature norm preservation)**.** *Assume the event $\mathcal{E}_0$. Then, if $m \ge CL^2\log(nNL/\delta_{init})$, we have that with probability at least $1 - \delta_{init}$, for every $(i,j) \in \mathcal{I}$ and every $\ell = 0, 1, \dots, L$, that $\|q_{ij,\ell}^{(0)}\|_2 = \Theta(H_{ij})$, and we denote this event by $\mathcal{E}_{feat}$.*

Proof. We proceed on a case-by-case basis. First, note that the case $\ell = 0$ is trivial since $H_{ij} = \|q_{ij,0}^{(0)}\|_2$. Then, fix $(i,j) \in \mathcal{I}$ and let $\ell \ge 1$. Conditional on $q_{ij,\ell-1}^{(0)}$, the coordinates of $r_{ij,\ell}^{(0)} = \mathbf{W}_\ell^{(0)} q_{ij,\ell-1}^{(0)}$ are independent Gaussians with variance $\frac{2}{m}\|q_{ij,\ell-1}^{(0)}\|_2^2$. Thus,

$$\|q_{ij,\ell}^{(0)}\|_2^2 = \frac{2}{m}\|q_{ij,\ell-1}^{(0)}\|_2^2 \sum_{k=1}^{m} g_k^2 \mathbb{1}\{g_k > 0\},$$

with $g_k \sim \mathcal{N}(0,1)$ i.i.d. Let $Y_k \coloneqq 2g_k^2 \mathbb{1}\{g_k > 0\}$, where Y_k's are i.i.d. sub-exponential on $\mathbb{E}Y_k = 1$. Then, for every $\epsilon \in (0,1)$, we have

$$\Pr\left[\left|\frac{1}{m}\sum_{k=1}^{m} Y_k - 1\right| > \epsilon \middle| q_{ij,\ell-1}^{(0)}\right] \leq 2e^{-\Omega(m\epsilon^2)}.$$

Thus, we now choose $\epsilon = 1/8L$. Then, for every fixed ij, ℓ, we have that

$$(1-\epsilon)\|q_{ij,\ell-1}^{(0)}\|_2^2 \leq \|q_{ij,\ell}^{(0)}\|_2^2 \leq (1+\epsilon)\|q_{ij,\ell-1}^{(0)}\|_2^2$$

with conditional probability at least $1 - 2e^{-\Omega(m/L^2)}$.

Since this holds uniformly over the past, we union bound over all nNL choices of (ij,ℓ). Thus, if $m \geq \Omega(L^2 \log(nNL/\delta_{init}))$, we then have with probability at least $1-\delta_{init}$ that $(1-\epsilon)\|q_{ij,\ell-1}^{(0)}\|_2^2 \leq \|q_{ij,\ell}^{(0)}\|_2^2 \leq (1+\epsilon)\|q_{ij,\ell-1}^{(0)}\|_2^2$, for all i,j,ℓ.

Then, iterating from 0 to $\ell \leq L$, we have that

$$(1-\epsilon)^L H_{ij}^2 \leq \|q_{ij,\ell}^{(0)}\|_2^2 \leq (1+\epsilon)^L H_{ij}^2.$$

Since $\epsilon = 1/8L$, we have that $(1-\epsilon)^L \geq e^{-1/4}$ and $(1+\epsilon)^L \leq e^{1/4}$. Taking square roots of this proves the lemma. □

Lemma E.8 (Hidden-layer's small margin bound at random initialization)**.** *Assume $\mathcal{E}_{\text{feat}}$. Then, for $(i,j) \in \mathcal{I}, \ell = 1,\dots,L$, and $u \geq 0$, define the initialization's small-margin count*

$$s_{ij,\ell}^{(0)}(u) \coloneqq |\{k \in [m] : |(r_{ij,\ell}^{(0)})_k| \leq u\}|.$$

Next, if $m \geq \Omega(L^2\Lambda)$, then with probability at least $1-\delta_{init}$ over the Gaussian initialization,

$$s_{ij,\ell}^{(0)}(u) \leq O\left(\frac{m^{3/2}u}{H_{ij}} + \Lambda\right)$$

for all $(i,j) \in \mathcal{I}$, all $\ell = 1,\dots,L$, and all $u \geq 0$. So, if $\Lambda \leq \frac{m^{3/2}u}{H_{ij}}$, we get $s_{ij,\ell}^{(0)}(u) \leq O(\frac{m^{3/2}u}{H_{ij}})$. We denote this event by $\mathcal{E}_{sm}$.

Proof. Fix $(i,j) \in \mathcal{I}$ and $\ell \in [L]$. Conditional on all previous layers, $q_{ij,\ell-1}^{(0)}$ is fixed and the coordinates of $r_{ij,\ell}^{(0)} = \mathbf{W}_\ell^{(0)} q_{ij,\ell-1}^{(0)}$ are i.i.d. Gaussian with $(r_{ij,\ell}^{(0)})_k \sim \mathcal{N}(0, \sigma_{ij,\ell}^2)$ with $\sigma_{ij,\ell}^2 = \frac{2}{m}\|q_{ij,\ell-1}^{(0)}\|_2^2$. Let $Z_k \coloneqq |(r_{ij,\ell}^{(0)})_k|$. Then, for all $u \geq 0$, we have that

$$F_{ij,\ell}(u) \coloneqq \Pr[Z_k \leq u | q_{ij,\ell-1}^{(0)}].$$

The density of $\mathcal{N}(0,\sigma_{ij,\ell}^2)$ is bounded by $\frac{1}{\sqrt{2\pi}\sigma_{ij,\ell}}$. Hence, we have that

$$\begin{aligned} F_{ij,\ell}(u) &\leq \frac{2u}{\sqrt{2\pi}\sigma_{ij,\ell}} \\ &= O\left(\frac{u}{\sigma_{ij,\ell}}\right) \\ &= O\left(\frac{\sqrt{m}u}{\|q_{ij,\ell-1}^{(0)}\|_2}\right). \end{aligned}$$

Recall that on the event $\mathcal{E}_{feat}$, $\|q_{ij,\ell-1}^{(0)}\|_2 \geq c_q H_{ij}$ for some constant $c_q > 0$.

So, $F_{ij,\ell}(u) \leq O\left(\frac{\sqrt{m}u}{H_{ij}}\right)$. Then, by applying the empirical small-ball bound lemma (Lemma E.17) conditionally on the previous layers, we have that with probability at least $1 - 2e^{-c\Lambda}$,

$$s_{ij,\ell}^{(0)}(u) \leq 8mF_{ij,\ell}(u) + 24\Lambda \leq O\left(\frac{m^{3/2}u}{H_{ij}} + \Lambda\right)$$

simultaneously for all $u \geq 0$.

Union bounding over all nNL choices of (i, j, ℓ) and using $\Lambda = \log \frac{O(nNL)}{\delta_{init}}$ proves the claim. □

Lemma E.9 (Gate balance at initialization)**.** *For $(i, j) \in \mathcal{I}$ and $\ell = 1, \ldots, L$, define $A^{(0)}_{ij,\ell} \coloneqq \{k \in [m] : (r^{(0)}_{ij,\ell})_k > 0\}$. Then, with probability at least $1 - nNLe^{-\Omega(m)}$, we have, for every $(i, j) \in \mathcal{I}$ and $\ell = 1, \ldots, L$, that $|A^{(0)}_{ij,\ell}| \geq \frac{m}{4}$. We denote this event by $\mathcal{E}_{gate}$.*

Proof. Fix i, j, ℓ. Then, conditional on $q^{(0)}_{ij,\ell-1}$, the entries of $r^{(0)}_{ij,\ell}$ are independent centered Gaussians. Therefore, $\mathbb{1}\{(r^{(0)}_{p,\ell})_k > 0\}$ are i.i.d. Bernoulli$(1/2)$ random variables. Hence, $|A^{(0)}_{ij,\ell}| \sim \mathrm{Bin}(m, 1/2)$. Finally, by Chernoff's inequality, we have that

$$\Pr\left[|A^{(0)}_{ij,\ell}| < \frac{m}{4} | q^{(0)}_{ij,\ell-1}\right] \leq e^{-\Omega(m)}.$$

Finally, union bounding over nNL choices proves the claim. □

For a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$, define $\|\mathbf{A}\|_{2,\infty} \coloneqq \max_{k \in [m]} \|e_k^\top \mathbf{A}\|_2$. Let $\theta^0 = (\mathbf{W}^0_1, \ldots, \mathbf{W}^0_L)$ denote the random initialization, where for every hidden layer $\ell \in [L]$, the entries of $\mathbf{W}^0_\ell$ are independent $\mathcal{N}(0, 2/m)$ random variables.

Lemma E.10. *For the finite sets $\mathcal{Z}_\ell = \{z_{\ell,a} : a \in \mathcal{A}_\ell\} \subset \mathbb{R}^m$ with $\ell = [L]$ and size $S = \sum_{\ell=1}^L |\mathcal{A}_\ell|$, where the random vectors $\mathcal{Z}_\ell$ are independent of $\mathbf{W}^{(0)}_\ell$, let $\Lambda_{\mathrm{row}} \coloneqq \log \frac{2mS}{\delta_{init}}$. Then, with probability at least $1 - \delta_{init}$, for every ℓ, for every $a \in \mathcal{A}_\ell$, and every row $k \in [m]$, we have that*

$$|e_k^\top \mathbf{W}^{(0)}_\ell z_{\ell,a}| \leq 2\sqrt{\frac{\Lambda_{row}}{m}} \|z_{\ell,a}\|_2.$$

Equivalently, $\|\mathbf{W}^{(0)}_\ell z_{\ell,a}\|_\infty \leq 2\sqrt{\frac{\Lambda_{\mathrm{row}}}{m}} \|z_{\ell,a}\|_2$. Moreover, on the same event, if the current weight matrices satisfy $\mathbf{W}_\ell = \mathbf{W}^{(0)}_\ell + \mathbf{E}_\ell$ with $\|\mathbf{E}_\ell\|_{2,\infty} \leq \frac{R_{\mathrm{row}}}{\sqrt{m}}$, we have

$$\|\mathbf{W}_\ell z_{\ell,a}\|_\infty \leq \frac{2\sqrt{\Lambda_{\mathrm{row}}} + R_{\mathrm{row}}}{\sqrt{m}} \|z_{\ell,a}\|_2,$$

and we denote this event by $\mathcal{E}_{\mathrm{row}}$.

Proof. Fix a layer $\ell \in [L]$, an index $a \in \mathcal{A}$, and row $k \in [m]$. Then, conditional on $z_{\ell,a}$, the random variables $e_k^\top \mathbf{W}^{(0)}_\ell z_{\ell,a}$ is Gaussian with mean zero and variance $\frac{2}{m}\|z_{\ell,a}\|_2^2$. If $z_{\ell,a} = 0$, the claim follows immediately. Otherwise,

$$\Pr\left[|e_k^\top \mathbf{W}^{(0)}_\ell z_{\ell,a}| > 2\sqrt{\frac{\Lambda_{\mathrm{row}}}{m}} \|z_{\ell,a}\|_2 \middle| z_{\ell,a}\right] = \Pr[|G| > \sqrt{2\Lambda_{\mathrm{row}}}] \leq 2e^{-\Lambda_{\mathrm{row}}}.$$

Union bounding over at most mS triples (ℓ, a, k) gives failure probability at most $2mSe^{-\Lambda_{row}} = \delta_{init}$. Then, for the perturbed matrix, we have that

$$\begin{aligned}
\|\mathbf{W}_\ell z_{\ell,a}\|_\infty &\leq \|\mathbf{W}^{(0)}_\ell z_{\ell,a}\|_\infty + \|E_\ell z_{\ell,a}\|_\infty \\
&\leq \|\mathbf{W}^{(0)}_\ell z_{\ell,a}\|_\infty + \|E_\ell\|_{2,\infty} \|z_{\ell,a}\|_2 \\
&\leq \|\mathbf{W}^{(0)}_\ell z_{\ell,a}\|_\infty + \frac{R_{\mathrm{row}}}{\sqrt{m}} \|z_{\ell,a}\|_2.
\end{aligned}$$

Finally, combining these bounds proves the claim. □

Lemma E.11 (Output normalization and effective output width)**.** *Let* $\tilde{\mathbf{W}}_{\text{out}} \in \mathbb{R}^{d\times m}$ *with* $(\tilde{\mathbf{W}}_{\text{out}})_{ak} \sim \mathcal{N}(0, 1/d)$*, and define* $\mathbf{W}_{\text{out}}^{(0)} = \frac{\Gamma}{R_{\text{hid}}^L \|\tilde{\mathbf{W}}_{\text{out}}\|}\tilde{\mathbf{W}}_{\text{out}}$*. Additionally, define the effective dimension* $m_{\text{eff}} = \frac{m}{\|\tilde{\mathbf{W}}_{\text{out}}\|_{\text{op}}^2}$*. Then, with probability at least* $1 - 2e^{-\Omega(d+m)} - \delta_{\text{init}}$*, we have*

$$\|\tilde{\mathbf{W}}_{\text{out}}\|_{\text{op}} = \Theta\left(1 + \sqrt{\frac{m}{d}}\right),$$

which in turn implies $m_{\text{eff}} = \Theta\left(\frac{m}{(1+\sqrt{m/d})^2}\right)$.

Also, if $d \geq C \log \frac{m}{\delta_{\text{init}}}$*, then every raw output column* $\tilde{w}_k$ *of* $\tilde{\mathbf{W}}_{\text{out}}$ *satisfies* $\frac{1}{2} \leq \|\tilde{w}_k\|_2 \leq 2$*. Moreover, the normalized output layer satisfies* $\|\mathbf{W}_{\text{out}}^{(0)}\|_{\text{op}} = \frac{\Gamma}{R_{\text{hid}}^L}$*. Finally, the column directions* $u_k \coloneqq \frac{\tilde{w}_k}{\|\tilde{w}_k\|_2}$ *are i.i.d. uniform on* $\mathbb{S}^{d-1}$*, independent of their own column norms and hidden-layer initialization. We denote this event by* $\mathcal{E}_{\text{out}}$.

Proof. Let $\mathbf{G} \coloneqq \sqrt{d}\tilde{\mathbf{W}}_{\text{out}}$. Then, $\mathbf{G} \in \mathbb{R}^{d\times m}$ has i.i.d. $\mathcal{N}(0,1)$ entries. The standard Gaussian operator-norm bound (Vershynin, 2018) gives that with probability at least $1 - e^{-t^2/2}$, $\|\mathbf{G}\|_{\text{op}} \leq \sqrt{d} + \sqrt{m} + t$. Taking $t = \sqrt{d} + \sqrt{m}$ gives $\|\mathbf{G}\|_{\text{op}} \leq 2(\sqrt{d} + \sqrt{m})$ with probability at least $1 - e^{-\Omega(d+m)}$. Hence, $\|\tilde{\mathbf{W}}_{\text{out}}\|_{\text{op}} \leq 2(1 + \sqrt{\frac{m}{d}})$.

For the lower bound, note that with probability at least $1 - e^{-\Omega(dm)}$,

$$\begin{aligned}
\|\mathbf{G}\|_{\text{op}} &\geq \frac{\|\mathbf{G}\|_F}{\sqrt{\text{rank}(\mathbf{G})}} \\
&\geq \frac{\|\mathbf{G}\|_F}{\sqrt{\min\{d,m\}}} \\
&\geq \Omega\left(\frac{\sqrt{dm}}{\sqrt{\min\{d,m\}}}\right) \\
&= \Omega(\sqrt{\max\{d,m\}}),
\end{aligned}$$

where the second inequality uses a loose dimensionality bound, and the third inequality follows from noting that $\|\mathbf{G}\|_F^2 \sim \chi^2_{dm}$ with probability at least $1 - e^{-\Omega(dm)}$. Thus, we also have $\|\tilde{\mathbf{W}}_{\text{out}}\|_{\text{op}} \geq \Omega(\max\{1, \sqrt{\frac{m}{d}}\})$, and hence $\|\tilde{\mathbf{W}}_{\text{out}}\|_{\text{op}} \asymp 1 + \sqrt{\frac{m}{d}}$. Therefore, we have that,

$$m_{\text{eff}} = \frac{m}{\|\tilde{\mathbf{W}_{\text{out}}}\|_{\text{op}}^2} \asymp \frac{m}{(1+\sqrt{m/d})^2}.$$

Note that if $m \leq d$, then $m_{\text{eff}} \asymp m$, while if $m \geq d$ then $m_{\text{eff}} \asymp d$. Hence, $m_{\text{eff}} \asymp m \wedge d$.

Next, for an output column $\tilde{w}_k$, we have $\|\tilde{w}_k\|_2^2 = \frac{1}{d}\chi_d^2$, and standard chi-square concentration gives

$$\Pr\left[\frac{1}{2} \leq \|\tilde{w}_k\|_2 \leq 2\right] \geq 1 - 2e^{-\Omega(d)}.$$

Here, union bounding over $k = 1, \ldots, m$ gives the column-norm bound where $d \geq \Omega(\log(m/\delta_{init}))$. Therefore, $\|\mathbf{W}_{\text{out}}^{(0)}\|_{\text{op}} = \frac{\Gamma}{R_{\text{hid}}^L}$ follows directly. Finally, since each $\tilde{w}_k$ is an isotropic Gaussian vector in $\mathbb{R}^d$, we have that $\frac{\tilde{w}_k}{\|\tilde{w}_k\|_2} \sim \text{Unif}(\mathbb{S}^{d-1})$, where the direction is independent of the norm. Together, these complete the proof. □

Corollary E.12 (Joint initialization event)**.** *Fix* $\delta_{\text{init}} \in (0,1)$ *and set* $\Lambda = \log \frac{CnNLm)}{\delta_{\text{init}}}$*. Assume* $m \geq CL^2\Lambda$ *and* $d \geq C \log \frac{nNm}{\delta_{\text{init}}}$*. Moreover, choose* $R_{\text{hid}}, s_{\text{hid}}$ *such that* $R_{\text{hid}} - s_{\text{hid}} > 2\sqrt{2}$*. Then, conditional on the data event* $\mathcal{E}_{\text{data}}$*, the joint initialization event*

$$\mathcal{E}_{\text{init}} \coloneqq \mathcal{E}_0 \cap \mathcal{E}_{\text{hid,op}} \cap \mathcal{E}_{\text{feat}} \cap \mathcal{E}_{\text{sm}} \cap \mathcal{E}_{\text{gate}} \cap \mathcal{E}_{\text{out}}$$

holds with probability at least $1 - \delta_{\text{init}} - Le^{-c_{\text{op}}m} - nNLe^{-cm}$.

Proof. Recall that on $\mathcal{E}_{\rm init}$, we have that $H_{ij} = \Theta(\sqrt{d})$ for all $(i,j) \in \mathcal{I}$, which implies $\|q_{ij,\ell}^{(0)}\|_2 = \Theta(H_{ij})$ for $\ell = 0, 1, \ldots, L$. Moreover, $\max_{\ell \in [L]} \|\mathbf{W}_\ell^{(0)}\|_{\rm op} \le R_{\rm hid} - s_{\rm hid}$. Additionally, we have that $s_{ij,\ell}^{(0)}(u) \le O(\frac{m^{3/2}u}{H_{ij}} + \Lambda)$, as well as $|\{k : (r_{ij,\ell}^{(0)})_k > 0\}| \ge m/4$. Finally, we also have the output normalization $\|\mathbf{W}_{\rm out}^{(0)}\|_{\rm op} = \frac{\Gamma}{R_{\rm hid}^L}$, along with $m_{\rm eff} \asymp m \wedge d$, and the output column directions $\frac{\tilde{w}_k}{\|\tilde{w}_k\|_2} \sim {\rm Unif}(\mathbb{S}^{d-1})$. Note here that each component follow from the prior lemmas, and that the stated probability follows by taking a union bound. □

Remark E.13. We explicitly list the consequences of Corollary E.12 below for future reference:

$$H_{ij}^2 \asymp d \tag{20}$$

$$\|q_{ij,\ell}^{(0)}\|^2 \asymp d \tag{21}$$

$$\max_\ell \|\mathbf{W}_\ell^{(0)}\|_{\rm op} \le R_{\rm hid} - s_{\rm hid} \tag{22}$$

$$s_{ij,\ell}^{(0)} \le O\left(\frac{m^{3/2}u}{H_{ij}} + \Lambda_{\rm init}\right) \tag{23}$$

$$|A_{ij,\ell}^{(0)}| \ge m/4 \tag{24}$$

$$\|\mathbf{W}_{\rm out}^{(0)}\|_{\rm op} = \frac{\Gamma}{R_{\rm hid}^L}, \quad m_{\rm eff} \asymp m \wedge d \tag{25}$$

In addition, we also state the following corollaries which we will find very useful later.

Corollary E.14. *Recall $\bar{H}_2^2 := \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^N \beta_j H_{ij}^2$, and that on the event $\mathcal{E}_0 \cap \mathcal{E}_{data}$, we have $H_{ij}^2 = \Theta(d)$. Further, since $\sum_{j=1}^N \beta_j = 1$, we also have $\bar{H}_2^2 = \Theta(d)$.*

Corollary E.15. *On the event $\mathcal{E}_{\rm init} \cap \mathcal{E}_{\rm data}$, we have that $\mathcal{L}(\theta^{(0)}) \le L_{\max} \le O(d(1+\Gamma^2))$.*

To see this, recall that for every $(i,j) \in [n] \times [N]$, we have $e_{ij}(\theta^{(0)}) = \mathbf{W}_{\rm out}^{(0)} q_{ij,L}^{(0)} - U_i$. Then, using $\|a-b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we have

$$\begin{aligned}
\mathcal{L}(\theta^{(0)}) &\le \frac{2}{n}\sum_{i,j} \beta_j \|\mathbf{W}_{\rm out}^{(0)} q_{ij,L}^{(0)}\|^2 + \frac{2}{n}\sum_{i,j}\beta_j \|U_i\|^2 \\
&\le \frac{2}{n}\sum_{i,j} \beta_j \|\mathbf{W}_{\rm out}^{(0)}\|_{\rm op}^2 \|q_{ij,L^{(0)}}\|^2 + \frac{2}{n}\sum_{i,j}\beta_j \|X_i^{(1)} - X_i^{(0)}\|^2 \\
&\le \frac{2}{n}\sum_{i,j} \beta_j \cdot \left(\frac{\Gamma}{R_{\rm hid}^L}\right)^2 \cdot (O(H_{ij}))^2 + \frac{2}{n}\sum_{i,j}\beta_j \cdot O(d) \\
&\le \sum_j \beta_j \cdot O(\Gamma^2 d) + \frac{2}{n}\sum_j \beta_j \cdot \sum_i O(d) \\
&= O(\Gamma^2 d) + O(d).
\end{aligned}$$

Therefore, we have that $L(\theta^{(0)}) \le O(d(1+\Gamma^2))$.

Corollary E.16. *Let $\tilde{w}_k$ be the k'th output column and $w_k^{out} = \frac{\Gamma}{R_{\rm hid}^L \|\tilde{W}_{\rm out}\|_{\rm op}} \tilde{w}_k$. On the event $\mathcal{E}_{\rm out}$, if $d \ge \Omega(\log(m/\delta_{\rm init}))$, we have that $\|w_k^{out}\|_2 \asymp \frac{\Gamma}{R_{\rm hid}^L \|\tilde{\mathbf{W}}_{\rm out}\|_{\rm op}}$ for all k. Therefore,*

$$\begin{aligned}
\sum_{k=1}^m \|w_k^{out}\|_2^2 &\asymp \frac{\Gamma^2}{R_{\rm hid}^{2L}} \cdot \frac{m}{\|\tilde{\mathbf{W}}_{\rm out}\|_{\rm op}^2} \\
&= \frac{\Gamma^2}{R_{\rm hid}^{2L}} m_{\rm eff}.
\end{aligned}$$

E.1.1 High-Dimensional Probability Lemmas

Lemma E.17 (Empirical small-ball bound)**.** *Let $Z_1, \ldots, Z_m$ be i.i.d. non-negative random variables with continuous CDF F. Define $N(u) = |\{k : Z_k \le u\}|$. Then, for every $\Lambda \ge 1$, with probability at least $1 - 2e^{-c\Lambda}$, we have that $N(u) \le 8mF(u) + 24\Lambda$, for all $u \ge 0$.*

Proof. The proof follows via a balls-and-bins argument. Let $Z_{(1)} \leq \cdots \leq Z_{(m)}$ be the order statistics. Then, in the event that the event fails, there must exist u such that $N(u) > 8mF(u) + 24\Lambda$. Let $k \coloneqq N(u) > 24\Lambda$ and we have that $F(u) < \frac{k}{8m}$. Define $v_k \coloneqq F^{-1}(\frac{k}{8m})$. Since F is continuous and strictly increasing, $u < v_k$. Thus, since $N(u) = k$, at least k of the Z_r's are at most $u < v_k$. Therefore, $|\{r : Z_r \leq v_k\}| \geq k$; however, $|\{r : z_r \leq v_k\}| \sim \mathrm{Bin}(m, \frac{k}{8m})$. Let $\mu = \frac{k}{8}$. Then,

$$\Pr\left[\mathrm{Bin}\left(m, \frac{k}{8m}\right) \geq k\right] \leq \left(\frac{e\mu}{k}\right)^k = \left(\frac{e}{8}\right)^k.$$

So, by union bound, the event fails with probability at most

$$\sum_{k \geq 24\Lambda} \Pr\left[\mathrm{Bin}\left(m, \frac{k}{8m}\right) \geq k\right] \leq \sum_{k \geq 24\Lambda} \left(\frac{e}{8}\right)^k \leq 2e^{-c\Lambda},$$

which proves the claim. □

Proposition E.18 (Proposition 2 in Wang et al. (2024))**.** *Consider $w \sim \mathcal{N}(0, \mathbf{I})$ where $w \in \mathbb{R}^n$. Then, $\|w\|$ and $w/\|w\|$ are independent random variables and $w/\|w\| \sim \mathrm{Unif}(\mathbb{S}^{n-1})$.*

Lemma E.19 (Lemma 2 in Wang et al. (2024))**.** *Let $w \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ where $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$. Then, for two vectors $v_1, v_2 \in \mathbb{R}^n$, we have that*

$$\Pr[w^\top v_1 \geq 0, w^\top v_2 \geq 0] = \frac{\pi - \angle(v_1, v_2)}{2\pi}.$$

Proof. Since $w \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, it suffices to consider the area of the event. Since the set $\{w \in \mathbb{S}^{n-1} | w^\top v_i\}$ is a semi-hypersphere, the probability is given by the intersection of the two semi-hyperspheres. Therefore,

$$\begin{aligned}\Pr[w^\top v_1 \geq 0, w^\top v_2 \geq 0] &= \frac{\text{area of } \{w \in \mathbb{S}^{n-1} | w^\top v_1 \geq 0\} \cap \text{ area of } \{w \in \mathbb{S}^{n-1} | w^\top v_2 \geq 0\}}{\text{area of the hypersphere}} \\ &= \frac{\pi - \angle(v_1, v_2)}{2\pi},\end{aligned}$$

which proves the lemma. □

Next, by Lee and Kim (2014), consider the unit hypersphere in $\mathbb{R}^d$ given by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. Then, the area of $\mathbb{S}^{d-1}$ is $A_d(1) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$.

Lemma E.20 (Lemma 3 in Wang et al. (2024))**.** *Fix $\xi \in \mathbb{S}^{d-1}$ and let $\xi_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. Then, with probability at least $1 - \exp(-\Omega(d))$, we have that $\angle(\xi_1, \xi_2) \leq \frac{3\pi}{4}$.*

Proof. Fix ξ_1. Then, $\{\xi_2 : \angle(\xi_1, \xi_2) \geq \pi - \theta\}$ are on a hyperspherical cap. Then,

$$\begin{aligned}\Pr[\angle(\xi_1, \xi_2) \geq \pi - \theta] &= \frac{A_d^\theta(1)}{A_d(1)} \\ &= \frac{1}{2} I_{\sin^2\theta}\left(\frac{d-1}{2}, \frac{1}{2}\right) \\ &\propto \frac{1}{2} \frac{\theta^{d-1}}{\sqrt{\pi}\sqrt{\frac{d-1}{2}}},\end{aligned}$$

where the second equality follows by a property from Lee and Kim (2014) that the area of the hyperspherical cap is

$$A_d^\theta(1) = \frac{1}{2} A_d(1) I_{\sin^2\theta}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

Let $\theta = \frac{\pi}{4} < 1$. Then, with probability at least $1 - \exp(-\Omega(d))$, we have that $\angle(\xi_1, \xi_2) \leq \frac{3\pi}{4}$. □

E.2 Perturbation Lemmas

Recall that the input layer W_0 and output layer $W_{\mathrm{out}} = W_{L+1}$ are fixed during training. The trainable parameters are $\theta = (W_1, \dots, W_L)$. For two feasible parameters θ, θ', write $\Delta W_\ell = W'_\ell - W_\ell$, $\Delta = \theta' - \theta$, $\varepsilon_\ell = \|\Delta W_\ell\|_{\mathrm{op}}$, $\delta_\ell = \sum_{k=1}^{\ell} \varepsilon_k$, and $\delta = \delta_L = \delta_{\mathrm{op}}(\theta, \theta')$, with $\delta_0 = 0$. Thus $\delta \le \sqrt{L}\,\|\Delta\|_F$, where $\|\Delta\|_F^2 = \sum_{\ell=1}^{L} \|\Delta W_\ell\|_F^2$. Throughout, $R = R_{\mathrm{hid}} > 1$, $\|W_\ell\|_{\mathrm{op}}, \|W'_\ell\|_{\mathrm{op}} \le R$, and $\|W_{\mathrm{out}}\|_{\mathrm{op}} \le \Gamma R^{-L}$.

For a training input (i, j), write $H_{ij} = \|q_{ij,0}\|_2$, which is fixed during training, and use primes for quantities evaluated at θ'. We assume $H_{ij} > 0$ when using small-margin bounds. The gates are $D_{ij,\ell} = \mathrm{diag}(\mathbf{1}\{(r_{ij,\ell})_k > 0\})$, with the selected convention $\sigma'(0) = 0$. Define $\Delta q_{ij,\ell} = q'_{ij,\ell} - q_{ij,\ell}$ and $\Delta r_{ij,\ell} = r'_{ij,\ell} - r_{ij,\ell}$. The selected linearized feature variation is given recursively by

$$p_{ij,0} = 0, \qquad p_{ij,\ell} = D_{ij,\ell}\big(W_\ell p_{ij,\ell-1} + \Delta W_\ell q_{ij,\ell-1}\big). \tag{26}$$

This is a linear function of Δ with the gates at θ held fixed; no directional differentiability at zero preactivations is assumed.

Lemma E.21 (ReLU secants and forward perturbation bounds)**.** *For $r, r' \in \mathbb{R}^m$, there is a diagonal matrix $E(r, r')$ such that*

$$\sigma(r') - \sigma(r) = (D(r) + E(r, r'))(r' - r), \qquad 0 \preceq D(r) + E(r, r') \preceq I, \tag{27}$$

where $\|E(r, r')\|_{\mathrm{op}} \le 1$ and its support is contained in the coordinates on which $D(r')$ and $D(r)$ differ. For every feasible θ, θ', every (i, j), and $1 \le \ell \le L$,

$$\max\{\|q_{ij,\ell}\|_2, \|q'_{ij,\ell}\|_2, \|r_{ij,\ell}\|_2, \|r'_{ij,\ell}\|_2\} \le R^\ell H_{ij}, \tag{28}$$

$$\max\{\|\Delta q_{ij,\ell}\|_2, \|\Delta r_{ij,\ell}\|_2, \|p_{ij,\ell}\|_2\} \le H_{ij} R^{\ell-1} \delta_\ell. \tag{29}$$

In particular, with $\overline{H}_2^2 = n^{-1} \sum_{i,j} \beta_j H_{ij}^2$, we have

$$\|v_{\theta'}(X_{ij}, t_j) - v_\theta(X_{ij}, t_j)\|_2 \le \frac{\Gamma H_{ij}}{R}\,\delta,$$

$$\left(\frac{1}{n}\sum_{i,j} \beta_j \|v_{\theta'}(X_{ij}, t_j) - v_\theta(X_{ij}, t_j)\|_2^2\right)^{1/2} \le \frac{\Gamma \overline{H}_2}{R}\,\delta \le \frac{\Gamma \overline{H}_2 \sqrt{L}}{R}\,\|\Delta\|_F. \tag{30}$$

Proof. For $r'_k \ne r_k$, define $s_k = (\sigma(r'_k) - \sigma(r_k))/(r'_k - r_k)$, and for equal arguments set $s_k = \mathbf{1}\{r_k > 0\}$. ReLU is monotone and 1-Lipschitz, so $0 \le s_k \le 1$. Taking $E_{kk} = s_k - \mathbf{1}\{r_k > 0\}$ proves (27) and the operator-norm bounds. If the two gates agree, the endpoints are either both positive or both nonpositive; the corresponding secant slope equals the selected gate. Thus $E_{kk} = 0$ on every unchanged coordinate, including when an endpoint is zero and the two gates are both zero.

Since $q_{ij,\ell} = \sigma(W_\ell q_{ij,\ell-1})$ and $r_{ij,\ell} = W_\ell q_{ij,\ell-1}$, $\|q_{ij,\ell}\|_2 \le \|r_{ij,\ell}\|_2 \le R\|q_{ij,\ell-1}\|_2$. Induction from H_{ij} gives (28) for both parameter choices. For their difference, the exact identity

$$\Delta r_{ij,\ell} = W'_\ell \Delta q_{ij,\ell-1} + \Delta W_\ell q_{ij,\ell-1} \tag{31}$$

and the Lipschitz property of ReLU give $\|\Delta q_{ij,\ell}\|_2 \le \|\Delta r_{ij,\ell}\|_2 \le R\|\Delta q_{ij,\ell-1}\|_2 + \varepsilon_\ell R^{\ell-1} H_{ij}$. Because $\Delta q_{ij,0} = 0$, induction proves the claimed difference bound. The recursion (26) gives the same norm inequality for $p_{ij,\ell}$, starting from $p_{ij,0} = 0$. Finally, multiply the last-layer difference bound by $\|W_{\mathrm{out}}\|_{\mathrm{op}} \le \Gamma R^{-L}$ and take the weighted root-mean-square to prove (30). □

Lemma E.22 (Exact remainder and its propagation)**.** *Let $\rho_{ij,\ell} = \Delta q_{ij,\ell} - p_{ij,\ell}$ and $g_{ij,\ell} = (D'_{ij,\ell} - D_{ij,\ell}) r'_{ij,\ell}$. Then $\rho_{ij,0} = 0$, and for every $1 \le \ell \le L$,*

$$\rho_{ij,\ell} = D_{ij,\ell} W_\ell \rho_{ij,\ell-1} + D_{ij,\ell} \Delta W_\ell \Delta q_{ij,\ell-1} + g_{ij,\ell}. \tag{32}$$

The vector $g_{ij,\ell}$ is supported on changed gates and satisfies $\|g_{ij,\ell}\|_2 \le \|\Delta r_{ij,\ell}\|_2$. Define $\Phi_{ij,\ell} = D_{ij,L} W_L \cdots D_{ij,\ell+1} W_{\ell+1}$, with $\Phi_{ij,L} = I$, and $\mathsf{B}_{ij,\ell} = W_{\mathrm{out}} \Phi_{ij,\ell}$. Then the exact output remainder is

$$W_{\mathrm{out}} \rho_{ij,L} = \sum_{\ell=1}^{L} \mathsf{B}_{ij,\ell}\big(D_{ij,\ell} \Delta W_\ell \Delta q_{ij,\ell-1} + g_{ij,\ell}\big). \tag{33}$$

Consequently,

$$\|\rho_{ij,L}\|_2 \le H_{ij} R^{L-2} \sum_{k<\ell} \varepsilon_k \varepsilon_\ell + \sum_{\ell=1}^{L} R^{L-\ell} \|g_{ij,\ell}\|_2, \tag{34}$$

$$\begin{aligned} \|W_{\mathrm{out}} \rho_{ij,L}\|_2 &\le \frac{\Gamma H_{ij}}{R^2} \sum_{k<\ell} \varepsilon_k \varepsilon_\ell + \sum_{\ell=1}^{L} \|\mathsf{B}_{ij,\ell} g_{ij,\ell}\|_2 \\ &\le \frac{\Gamma H_{ij}}{2R^2}\, \delta^2 + \Gamma \sum_{\ell=1}^{L} R^{-\ell} \|g_{ij,\ell}\|_2. \end{aligned} \tag{35}$$

An empty sum is zero, including the quadratic sum when $L = 1$.

Proof. The identities $q_{ij,\ell} = D_{ij,\ell} r_{ij,\ell}$ and $q'_{ij,\ell} = D'_{ij,\ell} r'_{ij,\ell}$ imply $\Delta q_{ij,\ell} - D_{ij,\ell} \Delta r_{ij,\ell} = g_{ij,\ell}$. Expanding both perturbed factors gives

$$\Delta r_{ij,\ell} = W_\ell \Delta q_{ij,\ell-1} + \Delta W_\ell q_{ij,\ell-1} + \Delta W_\ell \Delta q_{ij,\ell-1}. \tag{36}$$

Subtract (26) to obtain (32). On a changed coordinate the two preactivations are on opposite sides of the threshold, allowing equality with zero on the nonpositive side. Hence $|(r'_{ij,\ell})_k| \le |(r'_{ij,\ell} - r_{ij,\ell})_k|$ there. This proves the support and norm statements for $g_{ij,\ell}$.

Unrolling the remainder recursion from $\rho_{ij,0} = 0$ gives (33). Since $\|\Phi_{ij,\ell}\|_{\mathrm{op}} \le R^{L-\ell}$ and $\|\mathsf{B}_{ij,\ell}\|_{\mathrm{op}} \le \Gamma R^{-\ell}$, the gate contributions have the stated bounds. The mixed contribution at $\ell = 1$ vanishes because $\Delta q_{ij,0} = 0$. For $\ell \ge 2$, (29) gives $\|\Delta q_{ij,\ell-1}\|_2 \le H_{ij} R^{\ell-2} \delta_{\ell-1}$. Its contribution to the hidden remainder is therefore at most $H_{ij} R^{L-2} \varepsilon_\ell \delta_{\ell-1}$, and its output contribution is at most $\Gamma H_{ij} R^{-2} \varepsilon_\ell \delta_{\ell-1}$. Summing and using $\sum_\ell \varepsilon_\ell \delta_{\ell-1} = \sum_{k<\ell} \varepsilon_k \varepsilon_\ell \le \delta^2/2$ proves both inequalities. □

Unrolling (26) also gives $p_{ij,L} = \sum_{\ell=1}^{L} \Phi_{ij,\ell} D_{ij,\ell} \Delta W_\ell q_{ij,\ell-1}$, which is the selected first-order variation used in Section E.3. The bound (35) retains the individual propagated gate terms; Section E.3 estimates these terms using the restricted backward-product bound.

Lemma E.23 (Robust small-margin transfer)**.** *Assume the initialization small-margin event from Lemma E.8 holds in the form*

$$s^{(0)}_{ij,\ell}(u) := \#\{k : |(r^{(0)}_{ij,\ell})_k| \le u\} \le C_{\mathrm{sm}} \left(\frac{m^{3/2} u}{H_{ij}} + \Lambda \right), \qquad u \ge 0, \tag{37}$$

simultaneously over (i, j, ℓ), for fixed $C_{\mathrm{sm}} > 0$ and $\Lambda \ge 1$. For a current parameter θ, define $s^{\theta}_{ij,\ell}(u) = \#\{k : |(r_{ij,\ell}(\theta))_k| \le u\}$. If $r_{ij,\ell}(\theta) - r^{(0)}_{ij,\ell} = a + b$ with $\|a\|_2 \le A$ and $\|b\|_\infty \le B$, then for every $u \ge 0$ and $\tau > 0$,

$$s^{\theta}_{ij,\ell}(u) \le C_{\mathrm{sm}} \left(\frac{m^{3/2}(u + B + \tau)}{H_{ij}} + \Lambda \right) + \frac{A^2}{\tau^2}, \tag{38}$$

$$s^{\theta}_{ij,\ell}(u) \le C_{\mathrm{sm}} \left(\frac{m^{3/2}(u + B)}{H_{ij}} + \Lambda \right) + (C_{\mathrm{sm}} + 1) m (A/H_{ij})^{2/3}. \tag{39}$$

These are deterministic statements on the initialization event, including when θ, a, b depend on all initialized weights.

In particular, suppose $\theta^{(0)}$ is feasible and set $\mathcal{N}_\omega = \{\theta \in \Theta_{\mathrm{hid}} : \|\theta - \theta^{(0)}\|_F \le \omega\}$ for $\omega \ge 0$. Define $\eta_\omega = R^{L-1}\sqrt{L}\,\omega$ and

$$s_0 := \min \left\{ m, \left\lceil (C_{\mathrm{sm}} + 1) m \eta_\omega^{2/3} + C_{\mathrm{sm}} \Lambda \right\rceil \right\}. \tag{40}$$

Uniformly over $\theta, \theta' \in \mathcal{N}_\omega$, we have

$$s^{\theta}_{ij,\ell}(u) \le C_{\mathrm{sm}} \left(\frac{m^{3/2} u}{H_{ij}} + \Lambda \right) + (C_{\mathrm{sm}} + 1) m \eta_\omega^{2/3}, \tag{41}$$

$$\|D_{ij,\ell}(\theta) - D^{(0)}_{ij,\ell}\|_0 \le s_0, \qquad \|D_{ij,\ell}(\theta') - D_{ij,\ell}(\theta)\|_0 \le \min\{m, 2s_0\}. \tag{42}$$

Here $\|\cdot\|_0$ counts nonzero diagonal entries.

Proof. If $|(r_{ij,\ell}(\theta))_k| \le u$ and $|a_k| \le \tau$, then $|(r^{(0)}_{ij,\ell})_k| \le u + B + \tau$. It follows that

$$s^{\theta}_{ij,\ell}(u) \le s^{(0)}_{ij,\ell}(u + B + \tau) + \#\{k : |a_k| > \tau\} \le s^{(0)}_{ij,\ell}(u + B + \tau) + A^2/\tau^2. \tag{43}$$

This proves (38). For $A > 0$, choose $\tau = (A^2 H_{ij}/m^{3/2})^{1/3}$. The two terms involving τ sum to $(C_{\mathrm{sm}} + 1)m(A/H_{ij})^{2/3}$. For $A = 0$, let τ decrease to zero in (38). Thus (39) holds in both cases.

For $\theta \in \mathcal{N}_\omega$, the forward perturbation bound applied relative to initialization gives

$$\|r_{ij,\ell}(\theta) - r^{(0)}_{ij,\ell}\|_2 \le H_{ij} R^{\ell-1} \sum_{k=1}^{\ell} \|W_k - W_k^{(0)}\|_{\mathrm{op}} \le H_{ij}\eta_\omega. \tag{44}$$

Use the valid decomposition $a = r_{ij,\ell}(\theta) - r^{(0)}_{ij,\ell}$, $b = 0$ in (39) to obtain (41).

To count changed gates, fix a threshold $\tau > 0$. A gate differing from initialization has either $|(r^{(0)}_{ij,\ell})_k| \le \tau$ or $|(r_{ij,\ell}(\theta) - r^{(0)}_{ij,\ell})_k| > \tau$. Hence the number of changed gates is at most

$$C_{\mathrm{sm}} \left(\frac{m^{3/2}\tau}{H_{ij}} + \Lambda \right) + \frac{H_{ij}^2 \eta_\omega^2}{\tau^2}. \tag{45}$$

For $\eta_\omega > 0$, choose $\tau = H_{ij}\eta_\omega^{2/3}/\sqrt{m}$ and cap the count by m to get the first bound in (42). When $\eta_\omega = 0$, the parameters equal initialization and no gate changes. Finally, the support of the difference of two gate matrices is contained in the union of their differences from initialization. This proves the second bound. □

The local bounds above hold simultaneously over the entire Frobenius neighborhood and therefore over its subset $\mathcal{N}_{\omega,\rho}$. They use the full Euclidean preactivation perturbation and the initialization small-margin event. The resulting s_0 is the gate-count quantity used in the restricted backward-product argument of Section E.3.

E.3 Establishing a Semi-Smoothness Property

Lemma E.24 (First-order expansion)**.** *Let $\theta' = (\mathbf{W}'_1, \dots, \mathbf{W}'_L)$, and define $\Delta\mathbf{W}_\ell \coloneqq \mathbf{W}'_\ell - \mathbf{W}_\ell$ and $\Delta \coloneqq \theta' - \theta$. For all (i, j), let $\mathbf{D}_{ij,\ell}$ be the diagonal ReLU gate matrix at θ, such that $(\mathbf{D}_{ij,\ell})_{kk} = \mathbb{1}\{(r_{ij,\ell})_k > 0\}$. Define the first-order hidden-feature variation*

$$\mathbf{\Gamma}_{ij}(\Delta) \coloneqq \sum_{\ell=1}^{L} \mathbf{D}_{ij,L}\mathbf{W}_L\mathbf{D}_{ij,L-1}\mathbf{W}_{L-1}\dots\mathbf{D}_{ij,\ell+1}\mathbf{W}_{\ell+1}\mathbf{D}_{ij,\ell} \cdot \Delta\mathbf{W}_\ell q_{ij,\ell-1},$$

and the nonlinear remainder $\rho_{ij} \coloneqq q'_{ij,L} - q_{ij,L} - \mathbf{\Gamma}_{ij}(\Delta)$. Then, we have that

$$\begin{aligned}\mathcal{L}_{emp}(\theta') \le \mathcal{L}_{emp}(\theta) &+ \langle \nabla\mathcal{L}_{emp}(\theta), \Delta\rangle \\ &+ 2\sqrt{\mathcal{L}_{emp}(\theta)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\beta_j\|\mathbf{W}_{L+1}\rho_{ij}\|^2} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\beta_j\|\mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\|^2.\end{aligned}$$

Proof. By definition, we have

$$\begin{aligned}\|e_{ij}(\theta')\|^2 &= \|\mathbf{W}_{L+1}q'_{ij,L} - U_i\|^2 \\ &= \|e_{ij}(\theta) + \mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\|^2 \\ &= \|e_{ij}(\theta)\|^2 + \|\mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\|^2 + 2\langle e_{ij}(\theta), \mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\rangle.\end{aligned}$$

Summing with weights β_j and averaging over i, we have

$$\begin{aligned}
&\mathcal{L}_{emp}(\theta') - \mathcal{L}_{emp}(\theta) \\
&\quad= \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\beta_j\langle e_{ij}(\theta), \mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\rangle + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\beta_j\|\mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\|^2 \\
&\quad= \frac{2}{n}\sum_{i,j}\beta_j\langle e_{ij}(\theta), \mathbf{W}_{L+1}\mathbf{\Gamma}_{ij}(\Delta)\rangle + \frac{2}{n}\sum_{i,j}\beta_j\langle e_{ij}(\theta), \mathbf{W}_{L+1}\rho_{ij}\rangle \\
&\qquad\qquad + \frac{1}{n}\sum_{i,j}\beta_j\|\mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\|^2,
\end{aligned}$$

where the second equality follows by noting $q'_{ij,L} - q_{ij,L} = \mathbf{\Gamma}_{ij}(\Delta) + \rho_{ij}$.

Since the input and output layers are fixed, the backpropagation rule with the selected gates $D_{ij,\ell}$ gives $\langle \nabla\mathcal{L}_{\text{emp}}(\theta), \Delta\rangle = \frac{2}{n}\sum_{i,j}\beta_j\langle e_{ij}(\theta), W_{L+1}\mathbf{\Gamma}_{ij}(\Delta)\rangle$. Here $\nabla\mathcal{L}_{\text{emp}}$ denotes this selected backpropagation gradient, including at zero preactivations. The identity follows from the linear recursion for $p_{ij,\ell}$ and $p_{ij,L} = \mathbf{\Gamma}_{ij}(\Delta)$; it does not require directional differentiability at the ReLU threshold. Hence, we write that

$$\begin{aligned}
\mathcal{L}_{emp}(\theta') - \mathcal{L}_{emp}(\theta) - \langle \nabla\mathcal{L}_{emp}(\theta), \Delta\rangle &= \frac{2}{n}\sum_{i,j}\beta_j\langle e_{ij}(\theta), \mathbf{W}_{L+1}\rho_{ij}\rangle \\
&\quad + \frac{1}{n}\sum_{i,j}\beta_j\|\mathbf{W}_{L+1}(q'_{ij,L} - q_{ij,L})\|^2.
\end{aligned}$$

Finally, we bound the first term by Cauchy Schwarz,

$$\begin{aligned}
\frac{2}{n}\sum_{i,j}\beta_j\langle e_{ij}(\theta), \mathbf{W}_{L+1}\rho_{ij}\rangle &\le 2\sqrt{\frac{1}{n}\sum_{i,j}\beta_j\|e_{ij}(\theta)\|^2}\sqrt{\frac{1}{n}\sum_{i,j}\beta_j\|\mathbf{W}_{L+1}\rho_{ij}\|^2} \\
&= 2\sqrt{\mathcal{L}_{emp}(\theta)}\cdot\sqrt{\frac{1}{n}\sum_{i,j}\beta_j\|\mathbf{W}_{L+1}\rho_{ij}\|^2},
\end{aligned}$$

which proves the claim. □

Lemma E.25 (Gate-stability bound)**.** *Let $r, r' \in \mathbb{R}^m$. Define $D \coloneqq \operatorname{diag}(\mathbb{1}\{r_k > 0\})$ and $D' \coloneqq \operatorname{diag}(\mathbb{1}\{r'_k > 0\})$. Let $\Delta r = r' - r$, and suppose it admits a decomposition $\Delta r = a + b$ where $\|a\|_2 \le A$ and $\|b\|_\infty \le B$. Define the small-margin set*

$$S(2\tau) \coloneqq \{k \in [m] : |r_k| \le 2\tau\}$$

with $s(2\tau) \coloneqq |S(2\tau)|$. Then, if $\tau \ge 2B$, we have that

$$\|(D' - D)r'\|_2 \le 4\tau\sqrt{s(2\tau)} + 4A. \tag{46}$$

Proof. Let $F \coloneqq \{k \in [m] : D'_{kk} \ne D_{kk}\}$ be the gate-flip set. If $k \in F$ and $|r_k| > 2\tau$, then (since the sign-gate changed) r_k and r'_k lie on opposite sides of the ReLU threshold 0, by the convention that the gate is 1 only for strictly positive preactivations. Hence,

$$|r'_k - r_k| \ge |r_k| > 2\tau.$$

Since $r'_k - r_k = a_k + b_k$, and since $|b_k| \le B \le \tau/2$, we have that

$$|a_k| \ge |a_k + b_k| - |b_k| > 2\tau - B \ge \frac{3}{2}\tau > \tau.$$

Therefore, we have that $F \subseteq S(2\tau) \cup \{k : |a_k| > \tau\} = \{k \in [m] : |r_k| \le 2\tau\} \cup \{k : |a_k| > \tau\}$. Since $D' - D$ is diagonal and supported on F, we have $\|(D' - D)r'\|_2 = \|r'_F\|_2$. Split $F = G \cup H$,

where $G := F \cap S(2\tau) \cap \{k : |a_k| \le \tau\}$ and $H := F \setminus G$. Then, for $k \in G$, we have that $|r_k| \le 2\tau, |a_k| \le \tau$, and $|b_k| \le B \le \tau/2$. Together, this gives that

$$|r'_k| \le |r_k| + |a_k| + |b_k| \le 2\tau + \tau + \tau/2 \le 4\tau.$$

Therefore, $\|r'_G\|_2 \le 4\tau\sqrt{|G|} \le 4\tau\sqrt{s(2\tau)}$. It remains to control r'_H. By construction, if $K \in H$, then either $k \notin S(2\tau)$, in which case the argument above implies $|a_k| > \tau$ or $k \in S(2\tau)$, but $|a_k| > \tau$. In each of these cases, we have $H \subseteq \{k : |a_k| > \tau\}$. For $k \in H$, we bound $|r'_k| \le 4|a_k|$.

Indeed, if $k \notin S(2\tau)$, then $|r_k| > 2\tau$ and since the gate flipped, we have

$$\begin{aligned} |r'_k| &\le |r'_k - r_k| \\ &= |a_k + b_k| \\ &\le |a_k| + B \\ &\le |a_k| + \frac{\tau}{2} \\ &\le \frac{3}{2}|A_k| \le 4|a_k|. \end{aligned}$$

If $k \in S(2\tau)$ and $|a_k| > \tau$, then similarly we have that

$$\begin{aligned} |r'_k| &\le |r_k| + |a_k| + |b_k| \\ &\le 2\tau + |a_k| + \frac{\tau}{2} \end{aligned}$$

Therefore, we have that $\|r'_H\|_2 \le 4\|a_H\|_2 \le 4\|a\|_2 \le 4A$. Combining the two pieces, we have that

$$\begin{aligned} \|(D' - D)r'\|_2 = \|r'_F\|_2 &\le \|r'_G\|_2 + \|r'_H\|_2 \\ &\le 4\tau\sqrt{s(2\tau)} + 4A, \end{aligned}$$

which proves Equation (46). □

We bound the effect of changed gates after propagation to the output. Write $R = R_{\mathrm{hid}} > 1$, and keep the input and output layers fixed. For the hidden parameters, let $\mathcal{N}_\omega := \{\theta \in \Theta_{\mathrm{hid}} : \|\theta - \theta^{(0)}\|_F \le \omega\}$, where $\omega > 0$ is fixed before drawing the output layer. The bounds below also hold on the smaller neighborhood $\mathcal{N}_{\omega,\rho}$. Set $H_{ij} := \|q_{ij,0}\|_2 > 0$ and $\overline{H}_2^2 := n^{-1}\sum_{i,j}\beta_j H_{ij}^2$. Assume the hidden initialization satisfies $\|W_\ell^{(0)}\|_{\mathrm{op}} \le R$ and the initialization small-margin event from Lemma E.8, written with a fixed constant C_{sm} as

$$\#\{k : |(r_{ij,\ell}^{(0)})_k| \le u\} \le C_{\mathrm{sm}}\left(\frac{m^{3/2}u}{H_{ij}} + \Lambda\right) \quad \text{for every } u \ge 0, \tag{47}$$

simultaneously over all (i,j) and $\ell \in [L]$, where $\Lambda \ge 1$. The output layer is $W_{\mathrm{out}} = \Gamma R^{-L} G/\|G\|_{\mathrm{op}}$, where the entries of $G \in \mathbb{R}^{d\times m}$ are independent $\mathcal{N}(0, 1/d)$ and G is independent of the data and hidden initialization. Thus $\|W_{\mathrm{out}}\|_{\mathrm{op}} = \Gamma R^{-L}$.

For $S \subseteq [m]$, let P_S denote the coordinate projection onto S. Define the backward matrix $\mathsf{B}_{ij,\ell}(\theta) := W_{\mathrm{out}} D_{ij,L} W_L \cdots D_{ij,\ell+1} W_{\ell+1}$, where $\mathsf{B}_{ij,L} = W_{\mathrm{out}}$. All parameter and failure-probability choices below are made independently of G; the trained parameters may depend on G.

Lemma E.26 (Changed gates and restricted backward products)**.** *Fix $\delta_{\mathrm{out}} \in (0,1)$, and suppose $m \ge d$ and $dm \ge 16\log(2/\delta_{\mathrm{out}})$. Put $r_\omega := \sqrt{L}\,\omega$, $\eta_\omega := R^{L-1} r_\omega$, and define*

$$\begin{aligned} s_0 &:= \min\Big\{m, \Big\lceil (C_{\mathrm{sm}}+1)m\eta_\omega^{2/3} + C_{\mathrm{sm}}\Lambda \Big\rceil\Big\}, \qquad s_* := \min\{m, 2s_0\}, \\ A_{m,s_0} &:= \sum_{k=0}^{s_0}\binom{m}{k}, \qquad t_* := \log\frac{2nNL}{\delta_{\mathrm{out}}} + L\log A_{m,s_0} + \log\binom{m}{s_*}, \\ \kappa_* &:= \min\left\{1, \sqrt{\frac{2}{m}}\big(\sqrt{d} + \sqrt{s_*} + \sqrt{2t_*}\big)\right\}, \qquad a_\omega := \min\{1, \kappa_* + r_\omega/R\}. \end{aligned} \tag{48}$$

Conditional on any data and hidden initialization satisfying (47)*, with probability at least* $1-\delta_{\mathrm{out}}$ *over* G*, the following statements hold simultaneously for all* $\theta, \theta' \in \mathcal{N}_\omega$*, all* (i,j)*, and all* $\ell \in [L]$*:*

$$\|D_{ij,\ell}(\theta) - D_{ij,\ell}^{(0)}\|_0 \le s_0, \qquad \|D_{ij,\ell}(\theta') - D_{ij,\ell}(\theta)\|_0 \le s_*, \tag{49}$$

$$\sup_{|S|\le s_*} \|\mathsf{B}_{ij,\ell}(\theta) P_S\|_{\mathrm{op}} \le \Gamma R^{-\ell} a_\omega. \tag{50}$$

In particular, writing $\Delta W_k = W_k' - W_k$ *and* $\delta_\ell := \sum_{k=1}^{\ell} \|\Delta W_k\|_{\mathrm{op}}$*, we have*

$$\|\mathsf{B}_{ij,\ell}(\theta)(D'_{ij,\ell} - D_{ij,\ell}) r'_{ij,\ell}\|_2 \le \frac{\Gamma H_{ij}}{R}\, a_\omega \delta_\ell. \tag{51}$$

Proof. We first obtain deterministic forward bounds. For any two feasible parameters, write $\varepsilon_k = \|\Delta W_k\|_{\mathrm{op}}$. ReLU is 1-Lipschitz, and every hidden matrix has operator norm at most R. Induction over layers therefore gives

$$\|q_{ij,\ell}(\theta)\|_2 \le R^\ell H_{ij}, \qquad \|q'_{ij,\ell} - q_{ij,\ell}\|_2 \le H_{ij} R^{\ell-1} \sum_{k=1}^{\ell} \varepsilon_k, \tag{52}$$

and the same difference bound holds for $r_{ij,\ell}$. For example, the induction step follows from $\|\Delta q_{ij,\ell}\|_2 \le R\|\Delta q_{ij,\ell-1}\|_2 + \varepsilon_\ell R^{\ell-1} H_{ij}$. Comparing $\theta \in \mathcal{N}_\omega$ to initialization and using $\sum_k \|W_k - W_k^{(0)}\|_{\mathrm{op}} \le \sqrt{L}\,\omega$ gives $\|r_{ij,\ell}(\theta) - r_{ij,\ell}^{(0)}\|_2 \le H_{ij}\eta_\omega$.

If a gate differs from initialization, then either its initial preactivation has magnitude at most u, or the corresponding preactivation perturbation has magnitude greater than u. The number of changed gates is consequently at most

$$C_{\mathrm{sm}} \left(\frac{m^{3/2} u}{H_{ij}} + \Lambda \right) + \frac{H_{ij}^2 \eta_\omega^2}{u^2}. \tag{53}$$

Choose $u = H_{ij}\eta_\omega^{2/3}/\sqrt{m}$. This proves the first bound in (49). The support of the difference of two gate matrices is contained in the union of their differences from initialization, proving the second.

We next control backward products uniformly over possible gate patterns. Condition on the data and the entire input and hidden initialization. For each (i,j), consider all diagonal zero-one matrices $\widetilde{D}_k$ differing from $D_{ij,k}^{(0)}$ in at most s_0 coordinates. There are at most A_{m,s_0} possibilities at each layer. For a fixed choice of these masks, a layer ℓ, and a set S of size exactly s_*, form $A = \widetilde{D}_L W_L^{(0)} \cdots \widetilde{D}_{\ell+1} W_{\ell+1}^{(0)} P_S$. This matrix has rank at most s_* and operator norm at most $R^{L-\ell}$, and it is independent of G. The total number of matrices under consideration is at most $nNLA_{m,s_0}^L \binom{m}{s_*}$.

For each fixed A, Gaussian operator-norm concentration, applied to an orthonormal basis for its range, yields

$$\mathbb{P}\left\{ \|GA\|_{\mathrm{op}} > \frac{R^{L-\ell}}{\sqrt{d}} \left(\sqrt{d} + \sqrt{s_*} + \sqrt{2t_*} \right) \right\} \le e^{-t_*}. \tag{54}$$

Indeed, multiplication of G by that fixed basis gives a Gaussian matrix with at most s_* columns, to which the standard bound $\mathbb{P}\{\|Z\|_{\mathrm{op}} > \sqrt{d} + \sqrt{s_*} + u\} \le e^{-u^2/2}$ applies for standard Gaussian entries. A union bound makes (54) simultaneous with failure probability at most $\delta_{\mathrm{out}}/2$. Also, the chi-square lower-tail inequality gives $\mathbb{P}\{\|G\|_F^2 < m/2\} \le e^{-dm/16} \le \delta_{\mathrm{out}}/2$. On its complement, $\|G\|_{\mathrm{op}} \ge \|G\|_F/\sqrt{d} \ge \sqrt{m/(2d)}$. Dividing (54) by this lower bound, and using the deterministic operator-norm bound when it is smaller, gives

$$\|W_{\mathrm{out}} \widetilde{D}_L W_L^{(0)} \cdots \widetilde{D}_{\ell+1} W_{\ell+1}^{(0)} P_S\|_{\mathrm{op}} \le \Gamma R^{-\ell} \kappa_*. \tag{55}$$

Smaller supports are contained in a support of size s_*, so the same estimate applies to them. The probability calculation uses all possible masks before the trained parameters are chosen. It therefore remains valid when the actual masks depend on G.

For the actual masks at θ, replace the initialized matrices in (55) by W_k one at a time, keeping these masks fixed. Every intermediate product obeys the spectral bound. The resulting difference is at most

$$\Gamma R^{-\ell-1} \sum_{k=\ell+1}^{L} \|W_k - W_k^{(0)}\|_{\mathrm{op}} \le \Gamma R^{-\ell} r_\omega / R. \tag{56}$$

Combining this with (55) and the full bound $\|\mathsf{B}_{ij,\ell}(\theta)\|_{\mathrm{op}} \le \Gamma R^{-\ell}$ proves (50).

Finally, set $F = \mathrm{supp}(D'_{ij,\ell} - D_{ij,\ell})$. On every changed coordinate, $|(r'_{ij,\ell})_k| \le |(r'_{ij,\ell} - r_{ij,\ell})_k|$; this also holds at zero preactivations with the convention $D_{kk} = \mathbf{1}\{r_k > 0\}$. Hence $g = (D'_{ij,\ell} - D_{ij,\ell})r'_{ij,\ell}$ is supported on at most s_* coordinates and satisfies $\|g\|_2 \le H_{ij} R^{\ell-1}\delta_\ell$ by (52). Apply (50) to this vector to obtain (51). □

Lemma E.27 (Remainder bound)**.** *On the event of Lemma E.26, let $\theta, \theta' \in \mathcal{N}_\omega$, and write $\varepsilon_\ell := \|\Delta W_\ell\|_{\mathrm{op}}$, $\delta := \sum_{\ell=1}^{L} \varepsilon_\ell$, and $\delta_\ell := \sum_{k=1}^{\ell} \varepsilon_k$. Let $\Gamma_{ij}(\Delta)$ be the selected first-order hidden-feature variation defined in Lemma E.24, and set $\rho_{ij} = q'_{ij,L} - q_{ij,L} - \Gamma_{ij}(\Delta)$. Then*

$$\left(\frac{1}{n}\sum_{i,j}\beta_j\|W_{\mathrm{out}}\rho_{ij}\|_2^2\right)^{1/2} \le \Gamma\overline{H}_2\left(\frac{1}{R^2}\sum_{k<\ell}\varepsilon_k\varepsilon_\ell + \frac{a_\omega}{R}\sum_{\ell=1}^{L}\delta_\ell\right) \le \Gamma\overline{H}_2\left(\frac{\delta^2}{2R^2} + \frac{La_\omega\delta}{R}\right). \tag{57}$$

Moreover, the output difference satisfies

$$\left(\frac{1}{n}\sum_{i,j}\beta_j\|v_{\theta'}(X_{ij},t_j) - v_\theta(X_{ij},t_j)\|_2^2\right)^{1/2} \le \frac{\Gamma\overline{H}_2}{R}\,\delta. \tag{58}$$

Proof. Let $p_{ij,0} = 0$ and recursively define the selected linearization by $p_{ij,\ell} = D_{ij,\ell}(W_\ell p_{ij,\ell-1} + \Delta W_\ell q_{ij,\ell-1})$. Writing $\rho_{ij,\ell} = \Delta q_{ij,\ell} - p_{ij,\ell}$, direct expansion gives the exact identity

$$\rho_{ij,\ell} = D_{ij,\ell}W_\ell\rho_{ij,\ell-1} + D_{ij,\ell}\Delta W_\ell\Delta q_{ij,\ell-1} + (D'_{ij,\ell} - D_{ij,\ell})r'_{ij,\ell}. \tag{59}$$

Here $\rho_{ij,0} = 0$ and $p_{ij,L} = \Gamma_{ij}(\Delta)$. Unrolling and multiplying by W_{out} yields

$$W_{\mathrm{out}}\rho_{ij} = \sum_{\ell=1}^{L}\mathsf{B}_{ij,\ell}(\theta)\left[D_{ij,\ell}\Delta W_\ell\Delta q_{ij,\ell-1} + (D'_{ij,\ell} - D_{ij,\ell})r'_{ij,\ell}\right]. \tag{60}$$

For $\ell \ge 2$, the first contribution is bounded, using (52), by $\Gamma R^{-\ell}\varepsilon_\ell H_{ij} R^{\ell-2}\sum_{k<\ell}\varepsilon_k$. It vanishes for $\ell = 1$ because the input layer is fixed. The gate contribution is bounded by (51). Thus

$$\|W_{\mathrm{out}}\rho_{ij}\|_2 \le \Gamma H_{ij}\left(\frac{1}{R^2}\sum_{k<\ell}\varepsilon_k\varepsilon_\ell + \frac{a_\omega}{R}\sum_{\ell=1}^{L}\delta_\ell\right). \tag{61}$$

Taking the weighted root-mean-square proves the first inequality in (57); the second uses $\sum_{k<\ell}\varepsilon_k\varepsilon_\ell \le \delta^2/2$ and $\sum_\ell \delta_\ell \le L\delta$. Finally, multiply the last-layer instance of (52) by $\|W_{\mathrm{out}}\|_{\mathrm{op}} = \Gamma R^{-L}$ and average to obtain (58). □

Corollary E.28 (Semi-smoothness with explicit constants)**.** *On the same event, for every $\theta, \theta' \in \mathcal{N}_\omega$, with $\Delta = \theta' - \theta$, we have*

$$\mathcal{L}(\theta') \le \mathcal{L}(\theta) + \langle\nabla\mathcal{L}(\theta),\Delta\rangle + \alpha_\omega\sqrt{\mathcal{L}(\theta)}\,\|\Delta\|_F + b_\omega\sqrt{\mathcal{L}(\theta)}\,\|\Delta\|_F^2 + c_\omega\|\Delta\|_F^2, \tag{62}$$

where $\nabla\mathcal{L}$ denotes the selected backpropagation gradient with $\sigma'(0) = 0$, and one may take

$$\alpha_\omega = \frac{2\Gamma\overline{H}_2 a_\omega}{R}\sqrt{\frac{L(L+1)(2L+1)}{6}}, \qquad b_\omega = \frac{\Gamma\overline{H}_2(L-1)}{R^2}, \qquad c_\omega = \frac{\Gamma^2\overline{H}_2^2 L}{R^2}. \tag{63}$$

These constants apply to $\mathcal{N}_{\omega,\rho}$ as well.

Proof. The exact loss expansion and Cauchy–Schwarz, as in Lemma E.24, give

$$\mathcal{L}(\theta') \le \mathcal{L}(\theta) + \langle\nabla\mathcal{L}(\theta),\Delta\rangle + 2\sqrt{\mathcal{L}(\theta)}\left(\frac{1}{n}\sum_{i,j}\beta_j\|W_{\mathrm{out}}\rho_{ij}\|_2^2\right)^{1/2} + \frac{1}{n}\sum_{i,j}\beta_j\|v_{\theta'}(X_{ij},t_j) - v_\theta(X_{ij},t_j)\|_2^2. \tag{64}$$

This is an algebraic expansion with a remainder and does not require differentiability at zero preactivations. We have $\sum_\ell \varepsilon_\ell^2 \le \|\Delta\|_F^2$, and hence

$$\sum_{k<\ell} \varepsilon_k \varepsilon_\ell \le \frac{L-1}{2}\|\Delta\|_F^2, \qquad \delta^2 \le L\|\Delta\|_F^2,$$
$$\sum_{\ell=1}^{L} \delta_\ell = \sum_{k=1}^{L} (L-k+1)\varepsilon_k \le \sqrt{\frac{L(L+1)(2L+1)}{6}}\,\|\Delta\|_F. \tag{65}$$

Substitute (57) and (58) in the loss expansion to obtain (62) with the stated constants. □

The Gaussian probability statement above is conditional on the input and hidden initialization satisfying the stated margin and spectral events. If their joint failure probability is at most δ_{init} under the initialization law used by training, the combined failure probability is at most $\delta_{\mathrm{init}} + \delta_{\mathrm{out}}$. No independence between the output matrix and the trained parameters is required. To apply Theorem E.2, it remains to verify $\alpha_\omega^2 \le \mu_*/4$, $hc_\omega \le 1/8$, and $hb_\omega\sqrt{\mathcal{L}(\theta^{(0)})} \le 1/8$, together with its projected-gradient and neighborhood conditions.

E.4 Gradient Upper Bound

We use the same empirical loss $\mathcal{L}(\theta) = n^{-1}\sum_{i,j}\beta_j\|e_{ij}(\theta)\|_2^2$, where $e_{ij}(\theta) = W_{\mathrm{out}}q_{ij,L}(\theta) - U_i$ and $\sum_j \beta_j = 1$. The input and output layers are fixed, and the trainable parameters lie in $\Theta_{\mathrm{hid}} = \{(W_1,\dots,W_L) : \|W_\ell\|_{\mathrm{op}} \le R,\ \ell \in [L]\}$, where $R = R_{\mathrm{hid}} > 1$. Throughout, $\|W_{\mathrm{out}}\|_{\mathrm{op}} \le \Gamma R^{-L}$, $H_{ij} = \|q_{ij,0}\|_2$, and $\overline{H}_2^2 = n^{-1}\sum_{i,j}\beta_j H_{ij}^2$. It is convenient to write

$$C_{\mathrm{par}} := \frac{\Gamma\overline{H}_2}{R}, \qquad J := \sqrt{L}\,C_{\mathrm{par}}, \qquad \overline{\mu} := 4J^2. \tag{66}$$

Within these two subsections, J denotes the parameter-to-prediction Lipschitz constant defined in (66). We denote the selected backpropagation gradient by $g(\theta) = \nabla\mathcal{L}(\theta)$, with $\sigma'(0) = 0$. Let $\Pi_{\Theta_{\mathrm{hid}}}$ be Euclidean projection in the product Frobenius norm and define $G_h(\theta) = h^{-1}(\theta - \Pi_{\Theta_{\mathrm{hid}}}(\theta - hg(\theta)))$ for $h > 0$.

Lemma E.29 (Gradient upper bound)**.** *For every $\theta \in \Theta_{\mathrm{hid}}$ and every hidden layer ℓ,*

$$\|g_\ell(\theta)\|_F \le 2C_{\mathrm{par}}\sqrt{\mathcal{L}(\theta)}, \qquad \|G_h(\theta)\|_F^2 \le \|g(\theta)\|_F^2 \le \overline{\mu}\mathcal{L}(\theta). \tag{67}$$

Proof. For each (i,j), set $\psi_{ij,L} = D_{ij,L}W_{\mathrm{out}}^\top e_{ij}(\theta)$ and define $\psi_{ij,\ell} = D_{ij,\ell}W_{\ell+1}^\top\psi_{ij,\ell+1}$ for $\ell < L$. The selected backpropagation rule gives

$$g_\ell(\theta) = \frac{2}{n}\sum_{i,j}\beta_j\psi_{ij,\ell}q_{ij,\ell-1}^\top. \tag{68}$$

Since each gate has operator norm at most one, $\|\psi_{ij,\ell}\|_2 \le \Gamma R^{-\ell}\|e_{ij}(\theta)\|_2$. The forward bound gives $\|q_{ij,\ell-1}\|_2 \le R^{\ell-1}H_{ij}$. Taking Frobenius norms and applying the weighted Cauchy–Schwarz inequality therefore yields

$$\|g_\ell(\theta)\|_F \le \frac{2\Gamma}{nR}\sum_{i,j}\beta_j H_{ij}\|e_{ij}(\theta)\|_2 \le \frac{2\Gamma\overline{H}_2}{R}\sqrt{\mathcal{L}(\theta)}. \tag{69}$$

Summing the squared layerwise bounds gives $\|g(\theta)\|_F^2 \le 4LC_{\mathrm{par}}^2\mathcal{L}(\theta)$. The feasible set is closed and convex, so its Euclidean projection is nonexpansive. Since $\Pi_{\Theta_{\mathrm{hid}}}(\theta) = \theta$,

$$\|G_h(\theta)\|_F = \frac{1}{h}\|\Pi_{\Theta_{\mathrm{hid}}}(\theta) - \Pi_{\Theta_{\mathrm{hid}}}(\theta - hg(\theta))\|_F \le \|g(\theta)\|_F. \tag{70}$$

This proves (67), including at zero preactivations under the selected backpropagation convention. □

Corollary E.30 (Data-scaled bounds)**.** *Suppose $\|X_{0,i}\|_2, \|X_{1,i}\|_2 \le C_X\sqrt{d}$ for every i, and set $B_0 = \|W_0\|_{\mathrm{op}}$ and $\overline{U}^2 = n^{-1}\sum_i\|U_i\|_2^2$. Then*

$$\overline{H}_2^2 \le B_0^2(C_X^2 d + T^2), \qquad \overline{\mu} \le \frac{4L\Gamma^2B_0^2(C_X^2d + T^2)}{R^2}. \tag{71}$$

For every feasible parameter, including initialization,

$$\sqrt{\mathcal{L}(\theta)} \le \Gamma\overline{H}_2 + \overline{U} \le \Gamma B_0\sqrt{C_X^2d + T^2} + 2C_X\sqrt{d}. \tag{72}$$

Proof. The interpolant satisfies $\|X_{ij}\|_2 \le C_X\sqrt{d}$, so its augmented input $z_{ij} = (X_{ij}^\top, t_j)^\top$ obeys $\|z_{ij}\|_2^2 \le C_X^2d + T^2$. Consequently $H_{ij} = \|\sigma(W_0z_{ij})\|_2 \le B_0\sqrt{C_X^2d + T^2}$. Averaging and using (66) proves (71). The spectral bounds also give $\|v_\theta(X_{ij},t_j)\|_2 \le \Gamma R^{-L}R^LH_{ij} = \Gamma H_{ij}$. The triangle inequality in the weighted empirical norm yields the first inequality in (72). Finally $\|U_i\|_2 = \|X_{1,i} - X_{0,i}\|_2 \le 2C_X\sqrt{d}$. □

Lemma E.31 (Variation of the square-root loss)**.** *For all feasible θ, θ',*

$$|\sqrt{\mathcal{L}(\theta')} - \sqrt{\mathcal{L}(\theta)}| \le J\|\theta' - \theta\|_F. \tag{73}$$

Thus, if $\mathcal{L}_0 = \mathcal{L}(\theta^{(0)})$ and $\mathcal{N}_\omega = \{\theta \in \Theta_{\mathrm{hid}} : \|\theta - \theta^{(0)}\|_F \le \omega\}$, then $\sqrt{\mathcal{L}(\theta)} \le \sqrt{\mathcal{L}_0} + J\omega$ throughout $\mathcal{N}_\omega$.

Proof. Apply the reverse triangle inequality to the residual arrays in the weighted norm $\|e\|_{\mathrm{emp}}^2 = n^{-1}\sum_{i,j}\beta_j\|e_{ij}\|_2^2$. Their difference is the corresponding prediction difference. The forward perturbation bound from Section E.2 gives

$$\|v_{\theta'} - v_\theta\|_{\mathrm{emp}} \le C_{\mathrm{par}} \sum_{\ell=1}^{L} \|W'_\ell - W_\ell\|_{\mathrm{op}} \le C_{\mathrm{par}}\sqrt{L}\,\|\theta' - \theta\|_F. \tag{74}$$

This proves (73) and its local consequence. □

E.5 Containment and Inactive Projection

We work on the Frobenius neighborhood $\mathcal{N}_\omega$, on which the repaired semi-smoothness and ordinary-gradient lower bounds hold. Suppose initialization has spectral slack $s_{\mathrm{hid}} > 0$, meaning $\max_{\ell\in[L]}\|W_\ell^{(0)}\|_{\mathrm{op}} \le R - s_{\mathrm{hid}}$. We first give a condition for projection to be inactive throughout the neighborhood. We then prove containment and inactivity along the actual trajectory under a separate path-length condition.

Lemma E.32 (Inactive projection throughout the neighborhood)**.** *If*

$$\omega + 2hC_{\mathrm{par}}(\sqrt{\mathcal{L}_0} + J\omega) \le s_{\mathrm{hid}}, \tag{75}$$

then $\theta - hg(\theta) \in \Theta_{\mathrm{hid}}$ and $G_h(\theta) = g(\theta)$ for every $\theta \in \mathcal{N}_\omega$.

Proof. For each layer and each $\theta \in \mathcal{N}_\omega$, $\|W_\ell\|_{\mathrm{op}} \le R - s_{\mathrm{hid}} + \omega$. The layerwise gradient bound and (73) imply

$$\begin{aligned}\|W_\ell - hg_\ell(\theta)\|_{\mathrm{op}} &\le R - s_{\mathrm{hid}} + \omega + h\|g_\ell(\theta)\|_F \\ &\le R - s_{\mathrm{hid}} + \omega + 2hC_{\mathrm{par}}(\sqrt{\mathcal{L}_0} + J\omega) \le R.\end{aligned} \tag{76}$$

Thus the unprojected proposal is feasible, and Euclidean projection fixes it. The gradient-mapping identity follows. □

For the remaining two statements, assume $J > 0$ and fix $\mu_* \in (0, \overline{\mu}]$ such that

$$\|g(\theta)\|_F^2 \ge \mu_* \mathcal{L}(\theta) \qquad \text{for every } \theta \in \mathcal{N}_\omega. \tag{77}$$

Assume also that for every $\theta, \theta' \in \mathcal{N}_\omega$, with $\Delta = \theta' - \theta$, the semi-smoothness estimate is

$$\begin{aligned}\mathcal{L}(\theta') \le \mathcal{L}(\theta) + \langle g(\theta), \Delta\rangle + \alpha_\omega\sqrt{\mathcal{L}(\theta)}\|\Delta\|_F \\ + \big(b_\omega\sqrt{\mathcal{L}(\theta)} + c_\omega\big)\|\Delta\|_F^2, \qquad c_\omega = J^2,\end{aligned} \tag{78}$$

where $\alpha_\omega, b_\omega \ge 0$. These are the ordinary-gradient conclusion of Section E.6 and the semi-smoothness conclusion of Section E.3. Require

$$\alpha_\omega^2 \le \mu_*/4, \qquad hc_\omega \le 1/8, \qquad hb_\omega\sqrt{\mathcal{L}_0} \le 1/8. \tag{79}$$

The last condition is automatic when $b_\omega = 0$.

Lemma E.33 (One-step descent and path length)**.** *Under* (77)–(79)*, let $\theta \in \mathcal{N}_\omega$ satisfy $\mathcal{L}(\theta) \le \mathcal{L}_0$. If the unprojected proposal $\theta^+ = \theta - hg(\theta)$ also lies in $\mathcal{N}_\omega$, then projection is inactive and*

$$\mathcal{L}(\theta^+) \le \mathcal{L}(\theta) - \frac{h}{4}\|g(\theta)\|_F^2 \le (1 - h\mu_*/4)\mathcal{L}(\theta), \tag{80}$$

$$h\|g(\theta)\|_F \le \frac{8}{\sqrt{\mu_*}}\big(\sqrt{\mathcal{L}(\theta)} - \sqrt{\mathcal{L}(\theta^+)}\big). \tag{81}$$

Proof. The proposal is feasible by hypothesis, so projection fixes it. Substituting $\Delta = -hg(\theta)$ into (78) gives

$$\mathcal{L}(\theta^+) \le \mathcal{L}(\theta) - h\big(1 - hb_\omega\sqrt{\mathcal{L}(\theta)} - hc_\omega\big)\|g(\theta)\|_F^2 + h\alpha_\omega\sqrt{\mathcal{L}(\theta)}\|g(\theta)\|_F. \tag{82}$$

The step conditions make the coefficient in parentheses at least $3/4$. Furthermore, the gradient lower bound and $\alpha_\omega \le \sqrt{\mu_*}/2$ imply $\alpha_\omega\sqrt{\mathcal{L}(\theta)} \le \|g(\theta)\|_F/2$. This proves (80). If $\mathcal{L}(\theta) = 0$, the gradient upper bound gives $g(\theta) = 0$, and (81) is immediate. Otherwise,

$$\begin{aligned}\sqrt{\mathcal{L}(\theta)} - \sqrt{\mathcal{L}(\theta^+)} &= \frac{\mathcal{L}(\theta) - \mathcal{L}(\theta^+)}{\sqrt{\mathcal{L}(\theta)} + \sqrt{\mathcal{L}(\theta^+)}} \\ &\ge \frac{h\|g(\theta)\|_F^2}{8\sqrt{\mathcal{L}(\theta)}} \ge \frac{h\sqrt{\mu_*}}{8}\|g(\theta)\|_F, \end{aligned} \tag{83}$$

which proves the path-length inequality. □

Theorem E.34 (Containment and convergence)**.** *Suppose* (77)–(79) *hold on $\mathcal{N}_\omega$ and initialization has the stated spectral slack. Assume in addition that*

$$8\sqrt{\frac{\mathcal{L}_0}{\mu_*}} \le \min\{\omega, s_{\mathrm{hid}}\}. \tag{84}$$

Then the projected GD iterates $\theta^{(k+1)} = \Pi_{\Theta_{\mathrm{hid}}}(\theta^{(k)} - hg(\theta^{(k)}))$ remain in $\mathcal{N}_\omega$ for every $k \ge 0$, and every projection along this trajectory is inactive. Writing $\mathcal{L}_k = \mathcal{L}(\theta^{(k)})$, we have

$$\mathcal{L}_k \le (1 - h\mu_*/4)^k \mathcal{L}_0, \tag{85}$$

$$\sum_{r=0}^{k-1} \|\theta^{(r+1)} - \theta^{(r)}\|_F \le \frac{8}{\sqrt{\mu_*}}(\sqrt{\mathcal{L}_0} - \sqrt{\mathcal{L}_k}), \tag{86}$$

and the infinite sum of step lengths is at most $8\sqrt{\mathcal{L}_0/\mu_*}$. For $0 < \epsilon_{\mathrm{train}} < \mathcal{L}_0$, it suffices to take

$$K \ge \left\lceil \frac{4}{h\mu_*} \log \frac{\mathcal{L}_0}{\epsilon_{\mathrm{train}}} \right\rceil \tag{87}$$

to obtain $\mathcal{L}_K \le \epsilon_{\mathrm{train}}$. If $\epsilon_{\mathrm{train}} \ge \mathcal{L}_0$, no step is needed.

Proof. If $\mathcal{L}_0 = 0$, the gradient upper bound gives $g(\theta^{(0)}) = 0$, and all iterates equal initialization. Hence assume $\mathcal{L}_0 > 0$. Put $A_* = 8/\sqrt{\mu_*}$ and $S_k = \sum_{r=0}^{k-1} \|\theta^{(r+1)} - \theta^{(r)}\|_F$. We prove by induction that all steps so far have inactive projection, that the loss contracts, and that $S_k \le A_*(\sqrt{\mathcal{L}_0} - \sqrt{\mathcal{L}_k})$. These statements hold at $k = 0$.

Suppose they hold through k. First consider the unprojected proposal $\widetilde{\theta}^{(k+1)} = \theta^{(k)} - hg(\theta^{(k)})$. We establish its location before invoking semi-smoothness. The gradient upper bound gives $h\|g(\theta^{(k)})\|_F \le 2hJ\sqrt{\mathcal{L}_k}$. Since $\mu_* \le 4J^2$ and $hJ^2 \le 1/8$, we have $2hJ\sqrt{\mu_*} \le 4hJ^2 \le 1/2$, and in particular $2hJ \le A_*$. Therefore

$$\begin{aligned}\|\widetilde{\theta}^{(k+1)} - \theta^{(0)}\|_F &\le S_k + h\|g(\theta^{(k)})\|_F \\ &\le A_*\sqrt{\mathcal{L}_0} - (A_* - 2hJ)\sqrt{\mathcal{L}_k} \le A_*\sqrt{\mathcal{L}_0} \le \min\{\omega, s_{\mathrm{hid}}\}. \end{aligned} \tag{88}$$

For each proposed hidden matrix, this also gives

$$\|\widetilde{W}_\ell^{(k+1)}\|_{\mathrm{op}} \le \|W_\ell^{(0)}\|_{\mathrm{op}} + \|\widetilde{\theta}^{(k+1)} - \theta^{(0)}\|_F \le R - s_{\mathrm{hid}} + s_{\mathrm{hid}} = R. \tag{89}$$

Thus the proposal is feasible and belongs to $\mathcal{N}_\omega$. Projection fixes it, so $\theta^{(k+1)} = \widetilde{\theta}^{(k+1)}$.

Both parameters are now known to be in the neighborhood, and $\mathcal{L}_k \le \mathcal{L}_0$ by the induction hypothesis. Lemma E.33 therefore yields the next contraction and the step-length bound $\|\theta^{(k+1)} - \theta^{(k)}\|_F \le A_*(\sqrt{\mathcal{L}_k} - \sqrt{\mathcal{L}_{k+1}})$. Adding this to the bound for S_k closes the induction. Notice that condition (75) was not needed: inactivity was established directly for each proposed iterate.

Since $0 < h\mu_*/4 \le 1/8$, iterating the contraction gives (85). The partial path bound proves (86), and its nonnegative partial sums are bounded by $A_*\sqrt{\mathcal{L}_0}$, proving the infinite-sum statement. Finally $(1 - h\mu_*/4)^K \le e^{-h\mu_* K/4}$ gives (87). □

These statements are deterministic on the joint event supporting Sections E.3 and E.6. On that event one may use $\mu_* := \min\{\underline{\mu}, \overline{\mu}\}$, where $\underline{\mu}$ is the certified ordinary-gradient coefficient in Section E.6, provided the semi-smoothness conditions hold with this choice. If $J = 0$, the gradient upper bound gives $g \equiv 0$; a positive ordinary-gradient lower bound then forces $\mathcal{L}_0 = 0$, which is the stationary case already described. The additional containment requirement is equivalently $\mathcal{L}_0 \leq \mu_* \min\{\omega, s_{\mathrm{hid}}\}^2/64$. To obtain a convergence guarantee under random initialization, this inequality must be verified together with the common neighborhood and step-size conditions; the existence of a positive radius for the local lemmas alone does not establish it.

E.6 Gradient Lower Bound

We prove a lower bound for the last hidden-layer gradient, uniformly over the local neighborhood and all residual vectors. The uniformity is needed because both the residuals and the trained features depend on initialization. We first establish separation of the initialized features, then use a Gaussian row-isolation argument and control its change under perturbations.

Write $P = nN$, index the training inputs by $p = (i, j)$, and let $z_p = (X_{ij}^\top, t_j)^\top$. We use $\beta_j = (t_j - t_{j-1})/T$, so that $\sum_j \beta_j = 1$. For this subsection, set $q_p(\theta) = q_{ij,L-1}(\theta)$, $u_p(\theta) = q_p(\theta)/\|q_p(\theta)\|_2$, and $D_p(\theta) = \mathrm{diag}(\mathbf{1}\{(W_L q_p(\theta))_k > 0\})$. The feature norms below ensure that u_p is well defined. As in the semi-smoothness bound, write $R = R_{\mathrm{hid}}$ and $H_p = \|q_{ij,0}\|_2$. The frozen output layer is $W_{\mathrm{out}} = s_{\mathrm{out}} B$, where $B_{ak} \sim \mathcal{N}(0, 1/d)$ independently and $s_{\mathrm{out}} = \Gamma R^{-L}/\|B\|_{\mathrm{op}}$.

Before conditioning, the entries of $W_0^{(0)}, \dots, W_L^{(0)}$ are independent $\mathcal{N}(0, 2/m)$ variables, and B is independent of all these matrices and the data. All Gaussian probability calculations below are made before imposing the spectral conditioning in Section 3. Let $\mathcal{E}_{\mathrm{acc}}$ denote that acceptance event and $p_{\mathrm{acc}} = \mathbb{P}(\mathcal{E}_{\mathrm{acc}}) > 0$. At the end we transfer the joint event to the conditional initialization law. Fix $\eta \in (0, 1/4)$ and a deterministic radius $\omega > 0$ before drawing the weights. We work on $\mathcal{N}_\omega = \{\theta \in \Theta_{\mathrm{hid}} : \|\theta - \theta^{(0)}\|_F \le \omega\}$; hence every conclusion also holds on the smaller neighborhood $\mathcal{N}_{\omega,\rho}$.

Input and feature separation. The following data event uses independent Gaussian source samples, as specified by the flow-matching sampling scheme. Norm bounds alone do not imply separation between distinct trajectories.

Lemma E.35 (Separation of augmented flow inputs)**.** *Suppose $d \ge 2$, $X_{0,i} \sim \mathcal{N}(0, I_d)$ independently of each other and of all target samples, and condition on targets satisfying $c_1\sqrt{d} \le \|X_{1,i}\|_2 \le \sqrt{d}$ for a fixed $c_1 > 0$. Set $c_z = \frac{1}{2}\min\{\frac{1}{2}, c_1\}$ and*

$$\delta_{\mathrm{in}} := \min\left\{\frac{T\beta_{\min}}{10\sqrt{d}}, \frac{1-T}{2\sqrt{10}}\right\}. \tag{90}$$

With probability at least $1 - \delta_{\mathrm{data}}$, where $\delta_{\mathrm{data}} = 4ne^{-d/32} + n(n-1)N^2 e^{-(d-1)/5}$, we have $c_z\sqrt{d} \le \|z_p\|_2 \le \sqrt{5d}$ and $\sin\angle(z_p, z_q) \ge \delta_{\mathrm{in}}$ for every $p \ne q$.

Proof. Gaussian and chi-square tail bounds give, except with probability $4ne^{-d/32}$, simultaneously for all i, $\sqrt{d}/2 \le \|X_{0,i}\|_2 \le 2\sqrt{d}$ and $|\langle X_{0,i}, X_{1,i}/\|X_{1,i}\|_2\rangle| \le \sqrt{d}/4$. In particular the cosine of the angle between these two vectors is at least $-1/2$. For every $t \in [0, T]$, it follows that

$$\|(1-t)X_{0,i} + tX_{1,i}\|_2^2 \ge \tfrac{1}{2}\big((1-t)^2\|X_{0,i}\|_2^2 + t^2\|X_{1,i}\|_2^2\big) \ge c_z^2 d. \tag{91}$$

The upper bound follows from the triangle inequality and $t \le 1$. For two points on the same trajectory, writing $z_i(t) = (X_{0,i}, 0) + t(X_{1,i} - X_{0,i}, 1)$ and expanding the Gram determinant gives

$$\sin\angle(z_i(s), z_i(t)) \ge \frac{|t-s|\,\|X_{0,i}\|_2}{\|z_i(s)\|_2\|z_i(t)\|_2} \ge \frac{|t-s|}{10\sqrt{d}}. \tag{92}$$

Distinct training times differ by at least $T\beta_{\min}$.

For $i \ne i'$, condition on all targets and on $z_{i'}(s)$. Project the spatial coordinates onto a $(d-1)$-dimensional subspace orthogonal to the spatial part of $z_{i'}(s)$; its embedding in $\mathbb{R}^{d+1}$ is orthogonal to $z_{i'}(s)$. The projected $z_i(t)$ has law $\mu + (1-t)g$, with $g \sim \mathcal{N}(0, I_{d-1})$. For $k = d-1$, the Gaussian Laplace transform and Markov's inequality give, for any μ and $\sigma > 0$,

$$\mathbb{P}\{\|\mu + \sigma g\|_2^2 \le \sigma^2 k/4\} \le e^{k/8} 2^{-k/2} \le e^{-k/5}. \tag{93}$$

Thus the orthogonal component has norm at least $(1-T)\sqrt{d-1}/2$ except with that probability. Dividing by $\|z_i(t)\|_2 \le \sqrt{5d}$ and using $d \ge 2$ gives the second term in (90). A union bound proves the claim. □

Lemma E.36 (Penultimate direction separation at initialization)**.** *Condition on inputs satisfying the preceding lemma. Put* $\varepsilon_{\mathrm{f}} = \delta_{\mathrm{in}}^2/[8L(L+1)^2]$ *and* $\delta_h = \delta_{\mathrm{in}}/[2(L+1)]$*. If*

$$m \geq 272\varepsilon_{\mathrm{f}}^{-2} \log \frac{2LP^2}{\eta}, \tag{94}$$

then, under independent Gaussian initialization of $W_0, \ldots, W_{L-1}$*, with probability at least* $1-\eta$*, all intermediate features obey*

$$(1-\varepsilon_{\mathrm{f}})^{\ell+1}\|z_p\|_2^2 \leq \|q_{p,\ell}^{(0)}\|_2^2 \leq (1+\varepsilon_{\mathrm{f}})^{\ell+1}\|z_p\|_2^2, \qquad 0 \leq \ell \leq L-1, \tag{95}$$

$$\min_{p \neq q} \|(I - u_p^{(0)} u_p^{(0)\top}) u_q^{(0)}\|_2 \geq \delta_h. \tag{96}$$

In particular $\|q_p^{(0)}\|_2 \geq H_p/\sqrt{2}$ *and* $H_p \geq \|z_p\|_2/\sqrt{2}$*.*

Proof. For standard jointly Gaussian G, H of correlation c, direct two-dimensional Gaussian integration gives $2\mathbb{E}[G_+H_+] = f(c)$, where $f(c) = (\sqrt{1-c^2} + (\pi - \arccos c)c)/\pi$. Also $2\mathbb{E}G_+^2 = 1$. The function f is convex, increasing, and 1-Lipschitz on $[-1,1]$. Convexity and its endpoint values give $1 - f(c) \geq (1-c)/2$. For $c \geq 0$, put $\Delta = 1-c$ and $g = 1 - f(c)$. Writing $\vartheta = \arccos c \in [0, \pi/2]$ gives

$$0 \leq \Delta - g = \frac{\sin\vartheta - \vartheta\cos\vartheta}{\pi} \leq \frac{\vartheta^3}{3\pi} \leq \frac{\pi^2}{24}\Delta^{3/2} \leq \tfrac{1}{2}\Delta^{3/2}. \tag{97}$$

Together with $g \geq \Delta/2$, this implies $g^{-1/2} - \Delta^{-1/2} \leq 1$. The input sine separation implies $1 - \langle z_p/\|z_p\|_2, z_q/\|z_q\|_2\rangle \geq \delta_{\mathrm{in}}^2/2$. After the first exact kernel iteration the gap is at least $\delta_{\mathrm{in}}^2/4$, and all subsequent correlations are nonnegative. The reciprocal-square-root bound therefore shows that after L iterations the gap is at least $\delta_{\mathrm{in}}^2/(L+1)^2$.

We next compare the finite network to these iterations. Conditional on the preceding layers, each normalized squared norm or inner product is an average of m independent copies of $2G_+^2$ or $2G_+H_+$. For either variable Z, Gaussian moments give $\mathbb{E}|Z - \mathbb{E}Z|^k \leq 8^k k!$ for $k \geq 2$. For completeness, $2G_+H_+ \leq G^2 + H^2$ and $\mathbb{E}G^{2k} \leq 2^k k!$ give the uncentered bound $4^k k!$; centering costs at most a factor 2^k. Expanding the exponential series yields the Bernstein bound

$$\mathbb{P}\left\{\left|m^{-1}\sum_{r=1}^{m} Z_r - \mathbb{E}Z\right| > \varepsilon_{\mathrm{f}}\right\} \leq 2\exp(-m\varepsilon_{\mathrm{f}}^2/272). \tag{98}$$

A conditional union bound over at most LP^2 statistics gives an event of probability at least $1-\eta$. On this event the norm ratios at each layer lie in $[1-\varepsilon_{\mathrm{f}}, 1+\varepsilon_{\mathrm{f}}]$. After normalizing both vectors, the correlation error at one layer is at most $2\varepsilon_{\mathrm{f}}/(1-\varepsilon_{\mathrm{f}}) \leq 4\varepsilon_{\mathrm{f}}$. Since f is 1-Lipschitz, the accumulated error is at most $4L\varepsilon_{\mathrm{f}}$. The final correlation gap is consequently at least $\delta_{\mathrm{in}}^2/[2(L+1)^2]$. ReLU features have nonnegative correlations, so their squared sine is at least their correlation gap, proving (96). Multiplying the norm ratios proves (95). The established per-layer norm ratios also give $\|q_p^{(0)}\|_2^2/H_p^2 \geq (1-\varepsilon_{\mathrm{f}})^{L-1} \geq 1/2$ and $H_p^2/\|z_p\|_2^2 \geq 1 - \varepsilon_{\mathrm{f}} \geq 1/2$, proving the stated comparisons with H_p. □

Uniform perturbation bounds. The next two lemmas quantify the neighborhood conditions used below. They do not require the perturbations to be independent of the weights.

Lemma E.37 (Uniform normalized-feature stability)**.** *Define* $D_\omega = R^{L-2}\sqrt{L-1}\,\omega$ *for* $L \geq 2$*, and* $D_\omega = 0$ *for* $L = 1$*. On the feature event above and on* $\mathcal{E}_{\mathrm{acc}}$*, if* $D_\omega \leq 1/(2\sqrt{2})$*, then uniformly over* $\theta \in \mathcal{N}_\omega$*,*

$$\|q_p(\theta)\|_2 \geq \frac{H_p}{2\sqrt{2}}, \qquad \|u_p(\theta) - u_p^{(0)}\|_2 \leq \epsilon_h, \qquad \epsilon_h := 2\sqrt{2}\,D_\omega. \tag{99}$$

In particular $\min_p \|q_p(\theta)\|_2^2 \geq c_z^2 d/16$*.*

Proof. ReLU is 1-Lipschitz. Telescoping the first $L-1$ trainable layers, whose operator norms are bounded by R, gives $\|q_p(\theta) - q_p^{(0)}\|_2 \leq H_p R^{L-2}\sum_{\ell=1}^{L-1}\|W_\ell - W_\ell^{(0)}\|_{\mathrm{op}} \leq H_p D_\omega$. For $L = 1$ the penultimate feature is frozen, so this difference is zero. The lower norm bound follows from $\|q_p^{(0)}\|_2 \geq H_p/\sqrt{2}$. For nonzero vectors x, y, $\|x/\|x\|_2 - y/\|y\|_2\|_2 \leq 2\|x-y\|_2/\|y\|_2$. Apply this with $y = q_p^{(0)}$ to get (99). The last assertion uses $H_p^2 \geq \|z_p\|_2^2/2 \geq c_z^2 d/2$. □

Lemma E.38 (Uniform gate-change sparsity)**.** *Set* $\zeta_\omega = \sqrt{2L}R^{L-1}\omega$ *and*

$$s := \min\left\{m, \left\lceil 3m\zeta_\omega^{2/3} + 2\log(P/\eta)\right\rceil\right\}. \tag{100}$$

Conditional on any penultimate initialization satisfying the feature event, with probability at least $1-\eta$ *over the unconditioned draw of* $W_L^{(0)}$*, on* $\mathcal{E}_{\mathrm{acc}}$ *we have* $\|D_p(\theta) - D_p^{(0)}\|_0 \le s$ *for every* p *and every* $\theta \in \mathcal{N}_\omega$*.*

Proof. The same telescoping argument, now including layer L, gives $\|r_{p,L}(\theta) - r_{p,L}^{(0)}\|_2 \le H_p R^{L-1}\sqrt{L}\,\omega \le \|q_p^{(0)}\|_2\zeta_\omega$. Write $Q_p = \|q_p^{(0)}\|_2$ and $\tau = \zeta_\omega^{2/3} > 0$. Each initial preactivation is $\mathcal{N}(0, 2Q_p^2/m)$, independently over its coordinates. Thus the probability of magnitude at most $Q_p\tau/\sqrt{m}$ is at most τ. A binomial exponential-moment bound shows that their number is at most $2m\tau + 2\log(P/\eta)$, simultaneously over p, except with probability η. Indeed, for a sum S of independent Bernoulli variables of mean at most $m\tau$, $\mathbb{E}2^S \le e^{m\tau}$; Markov's inequality gives this bound with failure at most η/P. Outside those coordinates, a gate change requires a perturbation larger than $Q_p\tau/\sqrt{m}$. The squared Euclidean perturbation bound limits their number to $m\zeta_\omega^2/\tau^2 = m\tau$. This proves (100) simultaneously for all perturbations. □

Lemma E.39 (Gaussian restricted-output norm)**.** *For the deterministic integer* s *in* (100)*, define*

$$K_B := \frac{\sqrt{m} + \sqrt{d} + \sqrt{2\log(2/\eta)}}{\sqrt{d}},$$
$$K_s := \frac{\sqrt{d} + \sqrt{s} + \sqrt{2s\log(em/s) + 2\log(2/\eta)}}{\sqrt{d}}. \tag{101}$$

With probability at least $1-\eta$*,* $\|B\|_{\mathrm{op}} \le K_B$ *and* $\sup_{|S|\le s}\|B_{[:,S]}\|_{\mathrm{op}} \le K_s$*.*

Proof. For a $d \times r$ standard Gaussian matrix G, the Gaussian operator norm bound is $\mathbb{P}\{\|G\|_{\mathrm{op}} > \sqrt{d} + \sqrt{r} + t\} \le e^{-t^2/2}$. Apply it to $\sqrt{d}B$ with failure $\eta/2$. For submatrices of size s, use the same inequality and $\binom{m}{s} \le (em/s)^s$, allocating total failure $\eta/2$ to their union. Every smaller support is contained in one of size s. □

A lower bound uniform over residual vectors. For $v = (v_p)_{p=1}^P \in (\mathbb{R}^d)^P$, write $\|v\|_{2,\infty} = \max_p \|v_p\|_2$ and define the linear map $\mathsf{T}_\theta(v) = \sum_p D_p(\theta)B^\top v_p u_p(\theta)^\top$. Set

$$A_* := \sqrt{\frac{m\delta_h}{2048dP}}, \qquad \epsilon_* := \min\left\{\frac{1}{4}, \frac{A_*}{4PK_B}\right\}. \tag{102}$$

Lemma E.40 (Uniform residual lower bound at initialization)**.** *Conditional on any penultimate features satisfying* (96)*, suppose*

$$\frac{m\delta_h}{512P} \ge dP\log(1 + 2/\epsilon_*) + \log(1/\eta). \tag{103}$$

Apart from an event of probability at most η *over the independent draws of* $W_L^{(0)}$ *and* B*, the output-norm event implies* $\|\mathsf{T}_{\theta^{(0)}}(v)\|_F \ge (A_*/2)\|v\|_{2,\infty}$ *for every* v*.*

Proof. First fix a nonzero v independently of these two draws and choose $p_* \in \arg\max_p \|v_p\|_2$. Abbreviate $u_* = u_{p_*}^{(0)}$. For a standardized Gaussian row of $W_L^{(0)}$, write $g = zu_* + g_\perp$, where $z \sim \mathcal{N}(0,1)$ is independent of $g_\perp \sim \mathcal{N}(0, I - u_*u_*^\top)$. Put $t = \delta_h/(4P)$. Then $\mathbb{P}\{|z| \le t\} \ge \delta_h/(10P)$. For every $p \ne p_*$, the standard deviation of $\langle g_\perp, u_p^{(0)}\rangle$ is at least δ_h. The Gaussian density bound and a union bound therefore imply

$$\mathbb{P}\{\min_{p\ne p_*}|\langle g_\perp, u_p^{(0)}\rangle| > t\} \ge 1 - \frac{2Pt}{\sqrt{2\pi}\delta_h} \ge \frac{3}{4}. \tag{104}$$

On these two events, changing the sign of z preserves every nonselected gate and switches the selected gate. Independently, the corresponding output column b satisfies $|\langle b, v_{p_*}\rangle| \ge \|v_{p_*}\|_2/(2\sqrt{d})$

with probability at least $1/2$. Conditional on $|z|$, $g_\perp$, and b, the two possible row vectors are R_* and $R_* + \langle b, v_{p_*}\rangle u_*$, each with probability $1/2$. The triangle inequality shows that one has norm at least $\|v\|_{2,\infty}/(4\sqrt{d})$. Thus each row succeeds with probability at least $\delta_h/(64P)$. The rows are independent. A Chernoff bound gives at least $m\delta_h/(128P)$ successful rows except with probability $\exp[-m\delta_h/(512P)]$. Consequently

$$\mathbb{P}\{\|\mathsf{T}_{\theta^{(0)}}(v)\|_F < A_*\|v\|_{2,\infty}\} \le \exp[-m\delta_h/(512P)]. \tag{105}$$

Take a deterministic ϵ_*-net of the Euclidean unit ball in each residual block. Their product has at most $(1 + 2/\epsilon_*)^{dP}$ points, so (103) makes (105) hold on every net point with failure at most η. We intersect this event with $\|B\|_{\mathrm{op}} \le K_B$ after the probability calculation; no conditioning on B was used in the row-independence argument. For any v with $\|v\|_{2,\infty} = 1$, choose a net point $\bar{v}$ with $\|v - \bar{v}\|_{2,\infty} \le \epsilon_*$. Then $\|\bar{v}\|_{2,\infty} \ge 3/4$, and

$$\|\mathsf{T}_{\theta^{(0)}}(v - \bar{v})\|_F \le \sum_p \|B^\top(v_p - \bar{v}_p)\|_2 \le PK_B\epsilon_* \le A_*/4. \tag{106}$$

The triangle inequality proves the claim for unit block norm, and homogeneity proves it for every v. □

Lemma E.41 (Perturbation condition and uniform residual lower bound)**.** *Assume* (94)*,* (103)*,* $D_\omega \le 1/(2\sqrt{2})$*, and*

$$P(K_s + K_B\epsilon_h) \le A_*/4. \tag{107}$$

Conditional on the separated-data event and under the accepted initialization law, with probability at least $1 - 4\eta/p_{\mathrm{acc}}$*, simultaneously for every* $\theta \in \mathcal{N}_\omega$ *and every* v*,*

$$\|\mathsf{T}_\theta(v)\|_F^2 \ge \frac{m\delta_h}{32768dP}\|v\|_{2,\infty}^2. \tag{108}$$

Proof. On the joint feature, gate, output, and initialization-net events, expand the difference of the two linear maps as

$$\mathsf{T}_\theta(v) - \mathsf{T}_{\theta^{(0)}}(v) = \sum_p (D_p - D_p^{(0)})B^\top v_p u_p^\top + \sum_p D_p^{(0)} B^\top v_p (u_p - u_p^{(0)})^\top. \tag{109}$$

Each changed-gate vector is supported on at most s coordinates. The preceding bounds therefore give $\|\mathsf{T}_\theta(v) - \mathsf{T}_{\theta^{(0)}}(v)\|_F \le P(K_s + K_B\epsilon_h)\|v\|_{2,\infty}$. Subtract this from the initialization lower bound $A_*/2$ and use (107); the resulting bound is $\|\mathsf{T}_\theta(v)\|_F \ge (A_*/4)\|v\|_{2,\infty}$. Squaring proves (108). All bounds are uniform, so the same event permits parameters and residual vectors chosen after seeing initialization.

Before spectral conditioning, the four failure probabilities sum to at most 4η. For any failure event F, $\mathbb{P}(F \mid \mathcal{E}_{\mathrm{acc}}) \le \mathbb{P}(F)/p_{\mathrm{acc}}$. Applying this to their union gives the stated conditional probability. □

Corollary E.42 (A sufficient width and positive neighborhood radius)**.** *The conditions in the preceding lemma can be satisfied explicitly. Set* $\lambda_* := \sqrt{\delta_h/(32768P^3)}$ *and* $\sigma_* := \lambda_*^2/[64\log(64e/\lambda_*^2)]$*. In addition to* (94)*, it suffices to take*

$$m \ge \max\left\{\frac{64[d + 2\log(2/\eta)]}{\lambda_*^2},\quad \frac{2[2\log(P/\eta) + 1]}{\sigma_*},\right.$$
$$\left.\frac{512P}{\delta_h}\left[dP\log\left(1 + \frac{2048P^{3/2}}{\sqrt{\delta_h}}\right) + \log(1/\eta)\right]\right\}, \tag{110}$$

and any positive radius satisfying

$$\omega \le \frac{(\sigma_*/6)^{3/2}}{\sqrt{2L}R^{L-1}},\qquad \omega \le \frac{\lambda_*}{16\sqrt{2}R^{L-2}\sqrt{L-1}}\quad (L \ge 2). \tag{111}$$

The second radius condition is omitted when $L = 1$*.*

Proof. Write $q = s/m$, $a = \sqrt{d/m}$, and $b = \sqrt{2\log(2/\eta)/m}$. The first width restriction gives $a + b \le \lambda_*/4$. The first radius restriction and the second width restriction give

$$q \le 3\zeta_\omega^{2/3} + \frac{2\log(P/\eta) + 1}{m} \le \sigma_*. \tag{112}$$

The term $1/m$ accounts for the ceiling in (100). If $\ell_* := \log(64e/\lambda_*^2)$, then $\log(e/\sigma_*) = \ell_* + \log \ell_* \le 2\ell_*$. Since $q \mapsto q\log(e/q)$ increases on $(0, 1]$, we obtain $\sqrt{q} \le \lambda_*/8$ and $\sqrt{2q\log(e/q)} \le \lambda_*/4$. The second radius restriction ensures $\epsilon_h \le \lambda_*/8$; this is automatic for $L = 1$. It also implies $D_\omega \le 1/(2\sqrt{2})$. Consequently

$$\begin{aligned}\frac{P(K_s + K_B\epsilon_h)}{P\sqrt{m/d}} &\le a + b + \sqrt{q} + \sqrt{2q\log(e/q)} + (1 + a + b)\epsilon_h \\ &\le \lambda_*/4 + \lambda_*/8 + \lambda_*/4 + \lambda_*/4 \le \lambda_*.\end{aligned} \tag{113}$$

Because $P\sqrt{m/d}\lambda_* = A_*/4$, this proves (107). The first width restriction also implies $m \ge \max\{d, 2\log(2/\eta)\}$ and hence $K_B \le 3\sqrt{m/d}$. It follows that $\epsilon_* \ge \sqrt{\delta_h}/(1024P^{3/2})$. The last width restriction therefore implies (103). All radius upper bounds are strictly positive, and all width thresholds are finite. This proves the claimed sufficient conditions. □

The loss-gradient lower bound. Let $e_{ij}(\theta) = v_\theta(X_{ij}, t_j) - U_i$, $\mathcal{L}(\theta) = n^{-1}\sum_{i,j}\beta_j\|e_{ij}(\theta)\|_2^2$, and $\ell_i(\theta) = \sum_j \beta_j\|e_{ij}(\theta)\|_2^2$. Choose $i_* \in \arg\max_i \ell_i(\theta)$ and $j_* \in \arg\max_j \beta_j\|e_{i_*j}(\theta)\|_2^2$. Then $\ell_{i_*} \ge \mathcal{L}$ and $\beta_{j_*}\|e_{i_*j_*}\|_2^2 \ge \mathcal{L}/N$. At zero preactivations, gradients below mean the selected backpropagation gradient with $\sigma'(0) = 0$, consistently with the semi-smoothness bound.

Lemma E.43 (Lower bound on the last hidden-layer gradient)**.** *On the event of* (108)*, for every $\theta \in \mathcal{N}_\omega$ and every chosen pair (i_*, j_*),*

$$\|\nabla_{W_L}\mathcal{L}(\theta)\|_F^2 \ge \frac{c_z^2 s_{\text{out}}^2 m\delta_h}{131072\, n^3 N}\beta_{j_*}^2\|e_{i_*j_*}(\theta)\|_2^2. \tag{114}$$

Proof. The exact last-layer formula, including the frozen output scale, is

$$\nabla_{W_L}\mathcal{L}(\theta) = \frac{2s_{\text{out}}}{n}\sum_{i,j} D_{ij}(\theta)B^\top\big(\beta_j\|q_{ij}(\theta)\|_2 e_{ij}(\theta)\big)u_{ij}(\theta)^\top. \tag{115}$$

Apply (108) to $v_{ij} = \beta_j\|q_{ij}(\theta)\|_2 e_{ij}(\theta)$ and multiply by $4s_{\text{out}}^2/n^2$. Using $P = nN$ and $\|q_{ij}(\theta)\|_2^2 \ge c_z^2 d/16$ gives (114). No independence of these residuals from B or the hidden weights is required. □

Theorem E.44 (Lower bound on gradient)**.** *Under the conditions of the perturbation lemma, on its single success event every $\theta \in \mathcal{N}_\omega$ satisfies*

$$\|\nabla_{W_L}\mathcal{L}(\theta)\|_F^2 \ge \mu(\theta)\mathcal{L}(\theta), \qquad \mu(\theta) := \frac{c_z^2 s_{\text{out}}^2 m\delta_h}{131072\, n^3 N^2}\beta_{j_*(\theta)}. \tag{116}$$

If the projection of the gradient step at θ is inactive, the same lower bound holds for $\|G_h(\theta)\|_F^2$. A deterministic uniform lower-bound coefficient on the success event is

$$\underline{\mu} := \frac{c_z^2\Gamma^2 R^{-2L} m\delta_h\beta_{\min}}{131072\, n^3 N^2 K_B^2}. \tag{117}$$

Proof. For the block-first selected pair, $\beta_{j_*}^2\|e_{i_*j_*}\|_2^2 \ge (\beta_{j_*}/N)\mathcal{L}(\theta)$. Insert this into (114). When projection is inactive, $G_h(\theta) = \nabla\mathcal{L}(\theta)$, whose squared norm is at least that of its last-layer component. Finally $s_{\text{out}}^2 = \Gamma^2 R^{-2L}/\|B\|_{\text{op}}^2 \ge \Gamma^2 R^{-2L}/K_B^2$ and $\beta_{j_*} \ge \beta_{\min}$, which proves (117). □

In particular, if $m \ge \max\{d, 2\log(2/\eta)\}$, then $K_B \le 3\sqrt{m/d}$ and

$$\underline{\mu} \ge \frac{c_z^2\Gamma^2 R^{-2L} d\,\delta_{\text{in}}\beta_{\min}}{2359296\,(L+1)n^3N^2}. \tag{118}$$

Thus the output normalization replaces the unnormalized factor m by $m/\|B\|_{\mathrm{op}}^2$. The input-separation factor is the minimum in (90); its time-spacing branch cannot be substituted unless it is the smaller branch. To invoke Theorem E.2, the iterates must remain in the stated neighborhood, the projection must be inactive wherever the ordinary gradient bound is used, and the semi-smoothness conditions must hold with the certified choice $\mu_* := \underline{\mu}$. Those trajectory and step-size conditions are separate from the uniform lower bound proved here.

A common neighborhood and an explicit convergence rate. We now give simultaneous parameter choices for the preceding local bounds and the containment theorem. The output scale must satisfy an additional lower bound; increasing width alone does not establish containment. All parameter choices below are deterministic and made before initialization.

Let $P = nN$, $C_* = 2359296$, and retain c_z, δ_{in}, $\delta_h = \delta_{\mathrm{in}}/[2(L+1)]$, and $\varepsilon_{\mathrm{f}} = \delta_{\mathrm{in}}^2/[8L(L+1)^2]$ from Section E.6. Define

$$\nu := \frac{c_z^2 R^{-2L}\delta_{\mathrm{in}}\beta_{\min}}{C_*(L+1)n^3N^2}, \qquad \mu_0 := \Gamma^2 d\nu, \qquad Q_L := \frac{L(L+1)(2L+1)}{6}. \tag{119}$$

The coefficient μ_0 is the deterministic lower bound in (118). Fix $\delta \in (0,1)$, put $\eta = \delta/20$, $\delta_{\mathrm{out}} = \delta/4$, and $\eta_v = \delta/4$, and set $\Lambda = \max\{1, \log(PLm/\eta)\}$ and $\tau = \log(2PL/\delta_{\mathrm{out}})$. We may use $C_{\mathrm{sm}} = 8$ for the initialization small-margin event, as shown below.

For the lower-bound radius, use $\lambda_* = [\delta_h/(32768P^3)]^{1/2}$ and $\sigma_* = \lambda_*^2/[64\log(64e/\lambda_*^2)]$. For semi-smoothness, define

$$\chi := \min\left\{\frac{1}{2}, \frac{R\sqrt{\nu}}{4\sqrt{10Q_L}}\right\}, \qquad \ell_a := \log\frac{1024e(L+2)}{\chi^2}, \qquad \sigma_a := \frac{\chi^2}{1024(L+2)\ell_a}. \tag{120}$$

Fix spectral slack $s_{\mathrm{hid}} > 0$ with $R - s_{\mathrm{hid}} > 2\sqrt{2}$. A common radius is

$$\omega_0 := \min\left\{s_{\mathrm{hid}}, \quad \frac{(\sigma_*/6)^{3/2}}{\sqrt{2L}R^{L-1}}, \quad \frac{\lambda_*}{16\sqrt{2}R^{L-2}\sqrt{L-1}}, \right.$$
$$\left. \frac{(\sigma_a/18)^{3/2}}{R^{L-1}\sqrt{L}}, \quad \frac{R\chi}{2\sqrt{L}}\right\}. \tag{121}$$

The term containing $\sqrt{L-1}$ is omitted when $L = 1$. In particular, $\omega_0 > 0$, and its definition is independent of Γ, m, and the realized weights.

Theorem E.45 (A sufficient regime for convergence from random initialization)**.** *Assume $d \geq 2$ and the source independence and target-norm hypotheses of the input-separation lemma in Section E.6. Use the accepted Gaussian initialization of Section 3, with output scale*

$$\Gamma \geq \max\left\{2, \frac{32}{\sqrt{\nu}\,\omega_0}\right\}. \tag{122}$$

Require that the acceptance probability satisfy $p_{\mathrm{acc}} \geq 1/2$. Choose m satisfying the feature-width and explicit-width conditions (94) *and* (110)*, with $\eta = \delta/20$, and also*

$$m \geq \max\left\{d, \quad \frac{16\log(2/\delta_{\mathrm{out}})}{d}, \quad \frac{128d}{\chi^2}, \quad \frac{256\tau}{\chi^2}, \quad \frac{2(8\Lambda+1)}{\sigma_a}, \right.$$
$$\left. 20\Gamma^2R^{2-2L}d\left(1 + \sqrt{\frac{2\log(P/\eta_v)}{d}}\right)^2\right\}. \tag{123}$$

The dependence of Λ on m is logarithmic, so these width conditions admit a finite solution. Take

$$h := \frac{R^2}{80\Gamma^2 dL}, \qquad \mathcal{N}_{\omega_0} := \{\theta \in \Theta_{\mathrm{hid}} : \|\theta - \theta^{(0)}\|_F \leq \omega_0\}. \tag{124}$$

Conditional on targets satisfying the stated norm bounds, with probability at least $1 - \delta_{\mathrm{data}} - \delta$ over the source samples and accepted initialization, every projected GD iterate lies in $\mathcal{N}_{\omega_0}$, every projection is inactive, and

$$\mathcal{L}(\theta^{(k)}) \leq 16d\left(1 - \frac{R^2\nu}{320L}\right)^k, \qquad k \geq 0, \tag{125}$$

where $\delta_{\mathrm{data}} = 4ne^{-d/32} + n(n-1)N^2e^{-(d-1)/5}$. *Consequently, for* $0 < \epsilon_{\mathrm{train}} < 16d$, *it suffices to take*

$$K = \left\lceil \frac{320C_*R^{2L-2}L(L+1)n^3N^2}{c_z^2\delta_{\mathrm{in}}\beta_{\min}} \log \frac{16d}{\epsilon_{\mathrm{train}}} \right\rceil. \tag{126}$$

Proof. We first construct a common event. The four events in Section E.6 are feature separation and norm control, last-layer gate sparsity, the output norm bounds, and the uniform initialization residual bound. Their total failure probability before spectral acceptance is at most 4η. We next add the all-layer small-margin event.

Conditional on the previous layers, write $Q = \|q_{p,\ell-1}^{(0)}\|_2 > 0$ and let F be the distribution function of an absolute $\mathcal{N}(0, 2Q^2/m)$ variable. For the independent coordinates of $r_{p,\ell}^{(0)}$, the variables $V_k = F(|(r_{p,\ell}^{(0)})_k|)$ are independent and uniform on $[0,1]$. Write $N(u) = \#\{k : |(r_{p,\ell}^{(0)})_k| \le u\}$. If $N(u) > 2mF(u) + 8\Lambda$ for some u, then for $k = N(u) > 8\Lambda$ the kth order statistic obeys $V_{(k)} < k/(2m)$. The binomial exponential-moment bound gives

$$\mathbb{P}\{V_{(k)} < k/(2m)\} \le \exp[-(\log 2 - 1/2)k]. \tag{127}$$

A union bound over $k \le m$ bounds the probability of any such violation by $me^{-\Lambda}$. A further conditional union bound over the PL preactivation vectors gives failure at most η. On the feature event, $Q \ge H_p/\sqrt{2}$, and the Gaussian density bound gives $F(u) \le \sqrt{m}\,u/H_p$. Therefore, simultaneously for all $u \ge 0$, p, and ℓ,

$$N(u) \le 8\left(\frac{m^{3/2}u}{H_p} + \Lambda\right). \tag{128}$$

The union bound is taken before intersecting with the full feature event; no Gaussian independence is asserted after that intersection. Thus the combined raw failure probability so far is at most 5η.

We verify the semi-smoothness restriction. In the notation of (48), put $q = s_0/m$. The radius bound involving σ_a and the width bound involving Λ in (123) give $q \le \sigma_a$. Since $q < 1/2$, the binomial entropy bounds give

$$\log A_{m,s_0} \le mq\log(e/q), \qquad \log\binom{m}{s_*} \le 2mq\log(e/q). \tag{129}$$

Here $s_* = 2s_0$. Hence

$$\kappa_* \le \sqrt{2d/m} + 2\sqrt{q} + 2\sqrt{\tau/m} + 2\sqrt{(L+2)q\log(e/q)} \le \chi/2. \tag{130}$$

For the last inequality, the first and third terms are each at most $\chi/8$. Also $\log(e/\sigma_a) = \ell_a + \log\ell_a \le 2\ell_a$, so $2\sqrt{q} \le \chi/16$ and the final term is at most $\chi/\sqrt{128}$. Their sum is less than $\chi/2$. The final radius term gives $\sqrt{L}\,\omega_0/R \le \chi/2$; therefore $a_{\omega_0} \le \chi$.

The feature event implies $c_z^2 d/2 \le \overline{H}_2^2 \le 10d$. It follows from the repaired semi-smoothness constants that

$$\alpha_{\omega_0}^2 \le \frac{40\Gamma^2 dQ_L}{R^2}\chi^2 \le \mu_0/4, \qquad c_{\omega_0} \le \frac{10\Gamma^2 dL}{R^2}, \qquad b_{\omega_0} \le \frac{\Gamma\sqrt{10d}(L-1)}{R^2}. \tag{131}$$

The radius bounds involving σ_* and λ_* are exactly those of Corollary E.42, with the λ_* bound omitted when $L = 1$. Thus the ordinary-gradient lower bound holds throughout the same neighborhood with coefficient μ_0. Its definition and the lower feature-norm bound also give $0 < \mu_0 \le 4J^2$.

Conditional on suitable input and hidden weights, Lemma E.26 provides the uniform output-transport event with failure at most δ_{out}. We retain the denominator event used in its proof, $\|G\|_{\mathrm{op}} \ge \sqrt{m/(2d)}$; its failure is already included in δ_{out}. Conditional on all data and input and hidden weights, the Gaussian vector norm bound and a union bound give, with failure at most η_v,

$$\|Gq_{p,L}^{(0)}\|_2 \le \left(1 + \sqrt{\frac{2\log(P/\eta_v)}{d}}\right)\|q_{p,L}^{(0)}\|_2 \quad \text{for all } p. \tag{132}$$

The feature and accepted spectral events imply $\|q_{p,L}^{(0)}\|_2 \le R\sqrt{10d}$. Combining this with the denominator event and the last width condition gives $\|v_{\theta^{(0)}}(z_p)\|_2 \le \sqrt{d}$ for every p. On the data event, $\|U_i\|_2 \le 3\sqrt{d}$, so

$$\mathcal{L}_0 \le 16d, \qquad 8\sqrt{\mathcal{L}_0/\mu_0} \le \frac{32}{\Gamma\sqrt{\nu}} \le \omega_0 \le s_{\mathrm{hid}}. \tag{133}$$

This verifies containment's initial-loss requirement rather than assuming it as an additional random event.

For the stated step size, $hc_{\omega_0} \le 1/8$ and

$$hb_{\omega_0}\sqrt{\mathcal{L}_0} \le \frac{\sqrt{10}}{20\Gamma}\frac{L-1}{L} \le 1/8, \tag{134}$$

where $\Gamma \ge 2$; for $L = 1$ this term is zero. All hypotheses of Theorem E.34 now hold. That theorem gives (125) and (126).

Finally, spectral acceptance depends only on input and hidden weights. It changes the first five failure probabilities to at most $5\eta/p_{\mathrm{acc}}$ in total. The two conditional output probabilities are unaffected. Thus the conditional failure probability on the data event is at most

$$\frac{5\eta}{p_{\mathrm{acc}}} + \delta_{\mathrm{out}} + \eta_v \le 10\eta + \delta/4 + \delta/4 = \delta. \tag{135}$$

Adding δ_{data} proves the claimed probability. All bounds hold uniformly on the neighborhood, so no union bound over GD iterations is required. □

Acceptance probability and the role of the output scale. The condition $p_{\mathrm{acc}} \ge 1/2$ can be checked explicitly. Set $\Delta_h = R - s_{\mathrm{hid}} - 2\sqrt{2}$, $\Delta_x = B_x - \sqrt{2}(1 + \sqrt{d/m})$, and $\Delta_t = B_t - \sqrt{2}$. If all three are positive, Gaussian norm bounds and a union bound give

$$p_{\mathrm{acc}} \ge 1 - Le^{-m\Delta_h^2/4} - e^{-m\Delta_x^2/4} - e^{-m\Delta_t^2/4}. \tag{136}$$

It suffices to make each of the three subtracted terms at most $1/6$. Since the actual initialization is conditioned on acceptance, its spectral-slack failure probability is not added a second time.

The additional output-scale requirement (122) must be compatible with the specified Lipschitz budgets ΓB_x and ΓB_t. If those budgets fix a smaller Γ, the preceding theorem does not apply as stated. The structural event and common radius remain valid, but one must separately establish $\mathcal{L}_0 \le \mu_0\omega_0^2/64$ and choose $h \le [8\max\{J^2, b_{\omega_0}\sqrt{\mathcal{L}_0}\}]^{-1}$ to invoke the containment theorem.

For a uniform training grid, $\beta_{\min} = 1/N$ and the iteration bound has the form

$$K = O\left(\frac{R^{2L-2}L(L+1)n^3N^3}{c_z^2}\max\left\{\frac{N\sqrt{d}}{T}, \frac{1}{1-T}\right\}\log\frac{16d}{\epsilon_{\mathrm{train}}}\right). \tag{137}$$

For example, choosing $\delta = e^{-d/32}$ gives failure probability $\mathrm{poly}(nN)e^{-\Omega(d)}$, conditional on the target-norm hypothesis. If that hypothesis itself holds only probabilistically, its failure probability must also be included.

F End-to-End Guarantees

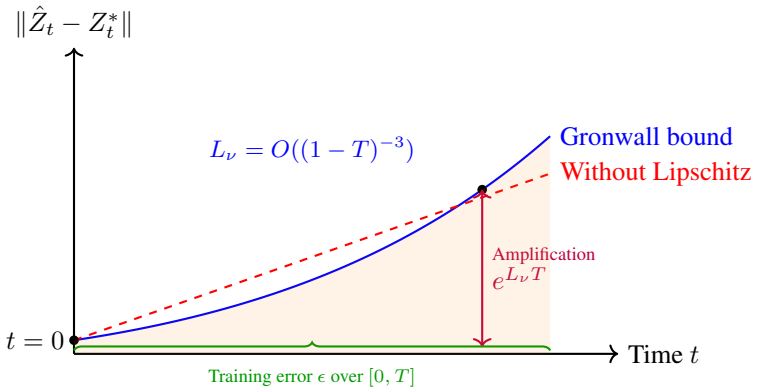

Figure 2: Error propagation via Gronwall's inequality: training error ϵ amplifies exponentially with factor $e^{L_\nu T}$ due to Lipschitz constant L_ν of the optimal velocity field. The bound grows faster near $T \to 1$ as L_ν increases.

F.1 Total Estimation Error Contribution

We now show how training error propagates through the ODE flow.

Lemma F.1. *Let Z_t^* denote the solution of the optimal ODE*

$$\frac{d}{dt} Z_t^* = \nu^*(Z_t^*, t), \qquad Z_0^* \sim \pi_0,$$

and let $\widehat{Z}_t$ denote the solution of the learned ODE

$$\frac{d}{dt} \widehat{Z}_t = \widehat{\nu}_{\text{train}}(\widehat{Z}_t, t), \qquad \widehat{Z}_0 \sim \pi_0.$$

Couple the two flows synchronously so that $\widehat{Z}_0 = Z_0^* \sim \pi_0$, and assume that the optimal velocity $\nu^*(\cdot, t)$ is L_ν-Lipschitz in space uniformly over $t \in [0, T]$, i.e. for all $x, y \in \mathbb{R}^d$ and $t \in [0, T]$, we have $\|\nu^*(x, t) - \nu^*(y, t)\| \le L_\nu \|x - y\|$. Define the trajectory-level velocity estimation error

$$\epsilon := \frac{1}{T} \int_0^T \mathbb{E}\left[\left\| \widehat{\nu}_{\text{train}}(\widehat{Z}_t, t) - \nu^*(\widehat{Z}_t, t) \right\|^2 \right] dt.$$

Then, we have that

$$W_2\Big(\text{Law}(\widehat{Z}_T), \text{Law}(Z_T^*)\Big) \le T e^{L_\nu T} \sqrt{\epsilon}.$$

Proof. Consider the optimal flow Z_t^* and the approximate flow $\hat{Z}_t$ with the same initial condition $Z_0 = \hat{Z}_0 \sim \pi_0$. Their difference satisfies:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(\hat{Z}_t - Z_t^*) &= \nu_\theta(\hat{Z}_t, t) - \nu^*(Z_t^*, t) \\ &= [\nu_\theta(\hat{Z}_t, t) - \nu^*(\hat{Z}_t, t)] + [\nu^*(\hat{Z}_t, t) - \nu^*(Z_t^*, t)]. \end{aligned}$$

By the Lipschitz continuity of ν^*:

$$\|\nu^*(\hat{Z}_t, t) - \nu^*(Z_t^*, t)\| \le L_\nu \|\hat{Z}_t - Z_t^*\|.$$

Integrating from 0 to T:

$$\|\hat{Z}_t - Z_t^*\| \le \int_0^t \|\nu_\theta(\hat{Z}_s, s) - \nu^*(\hat{Z}_s, s)\| \mathrm{d}s + L_\nu \int_0^t \|\hat{Z}_s - Z_s^*\| \mathrm{d}s.$$

By Gronwall's inequality:

$$\|\hat{Z}_t - Z_t^*\| \le e^{L_\nu t} \int_0^t \|\nu_\theta(\hat{Z}_s, s) - \nu^*(\hat{Z}_s, s)\| \mathrm{d}s.$$

Taking expectations and using Jensen's inequality:

$$\begin{aligned}\mathbb{E}[\|\hat{Z}_T - Z_T^*\|^2] &\le e^{2L_\nu T} \mathbb{E}\left[\left(\int_0^T \|\nu_\theta(\hat{Z}_s, s) - \nu^*(\hat{Z}_s, s)\| \mathrm{d}s\right)^2\right] \\ &\le e^{2L_\nu T} \cdot T \int_0^T \mathbb{E}[\|\nu_\theta(\hat{Z}_s, s) - \nu^*(\hat{Z}_s, s)\|^2] \mathrm{d}s.\end{aligned}$$

The integral term involves an expectation over the law of $\hat{Z}_s$, while the training loss ϵ is an expectation over the law of Z_s^*. We bridge this gap via a Wasserstein contraction argument. Let

$$w(t) = W_2(\mathrm{Law}(\hat{Z}_t), \mathrm{Law}(Z_t^*)).$$

By the triangle inequality for Wasserstein distance and the fact that both flows have the same Lipschitz constant:

$$\begin{aligned}w(t + dt) &\le w(t) + \mathrm{d}t \cdot W_2(\mathrm{Law}(\nu_\theta(\hat{Z}_t, t)), \mathrm{Law}(\nu^*(Z_t^*, t))) \\ &\le w(t) + \mathrm{d}t \cdot [L_\nu w(t) + W_2(\mathrm{Law}(\nu_\theta(Z_t^*, t)), \mathrm{Law}(\nu^*(Z_t^*, t)))].\end{aligned}$$

Taking the limit $\mathrm{d}t \to 0$ and using the coupling that induces the Wasserstein distance:

$$\frac{\mathrm{d}}{\mathrm{d}t} w(t) \le L_\nu w(t) + \sqrt{\mathbb{E}_{Z_t^*}[\|\nu_\theta(Z_t^*, t) - \nu^*(Z_t^*, t)\|^2]}.$$

Applying Gronwall's inequality with $w(0) = 0$:

$$w(t) \le e^{L_\nu t} \int_0^t \sqrt{\mathbb{E}_{Z_s^*}[\|\nu_\theta(Z_s^*, s) - \nu^*(Z_s^*, s)\|^2]} \mathrm{d}s.$$

Squaring both sides and using Cauchy-Schwarz:

$$\begin{aligned}w(T)^2 &\le e^{2L_\nu T} \left(\int_0^T \sqrt{\mathbb{E}_{Z_s^*}[\|\nu_\theta - \nu^*\|^2]} \mathrm{d}s\right)^2 \\ &\le T e^{2L_\nu T} \int_0^T \mathbb{E}_{Z_s^*}[\|\nu_\theta(Z_s^*, s) - \nu^*(Z_s^*, s)\|^2] \mathrm{d}s.\end{aligned}$$

The integral is the population training loss, which is bounded by ϵ. Therefore:

$$W_2(\mathrm{Law}(\hat{Z}_T), \mathrm{Law}(Z_T^*))^2 \le T^2 e^{2L_\nu T} \epsilon.$$

Taking the square root gives the final result:

$$W_2(\mathrm{Law}(\hat{Z}_T), \mathrm{Law}(Z_T^*)) \le \sqrt{\epsilon} \cdot T \cdot e^{L_\nu T},$$

which proves the lemma. □

F.2 Discretization Error Contribution

Lemma F.2 (Euler discretization error). *Let $0 = t_0 < t_1 < \cdots < t_N = T$, and write $h_j = t_{j+1} - t_j$. Let $\widehat{v} \in \mathrm{NN}(L, M, J, K, \kappa, \gamma_1, \gamma_2)$, so that*

$$\|\widehat{v}(x, t)\| \le K, \qquad \|\widehat{v}(x, t) - \widehat{v}(y, t)\| \le \gamma_1 \|x - y\|,$$

and $\|\widehat{v}(x, t) - \widehat{v}(x, s)\| \le \gamma_2 |t - s|$. Let Z_t denote the continuous flow

$$dZ_t = \widehat{v}(Z_t, t)\, dt, \qquad Z_0 \sim \pi_0,$$

and let $\widetilde{Z}_t$ denote the continuous-time interpolation of the explicit Euler scheme

$$\widetilde{Z}_{t_{j+1}} = \widetilde{Z}_{t_j} + h_j \widehat{v}(\widetilde{Z}_{t_j}, t_j), \qquad \widetilde{Z}_0 = Z_0.$$

Let $\pi_T^{\mathrm{cont}} = \mathrm{Law}(Z_T)$ and $\pi_T^{\mathrm{Euler}} = \mathrm{Law}(\widetilde{Z}_T)$. Then, we have that

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2)\sqrt{\sum_{j=0}^{N-1} h_j^3}.$$

In particular, for a uniform grid $h_j = T/N$, we have that

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2)\frac{T^{3/2}}{N}.$$

Proof. We couple the two processes by using the same initialization $Z_0 = \widetilde{Z}_0 \sim \pi_0$. By definition of the Wasserstein distance, we have that $W_2^2(\pi_t^{\mathrm{Euler}}, \pi_t^{\mathrm{cont}}) \le \mathbb{E}\|Z_t - \widetilde{Z}_t\|^2$. Define $R_t := \mathbb{E}\|Z_t - \widetilde{Z}_t\|^2$. Then, for $t \in [t_j, t_{j+1})$, the Euler interpolation satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{Z}_t = \widehat{v}(\widetilde{Z}_{t_j}, t_j).$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}R_t = 2\mathbb{E}\Big[\big\langle \widehat{v}(Z_t, t) - \widehat{v}(\widetilde{Z}_{t_j}, t_j), Z_t - \widetilde{Z}_t \big\rangle\Big].$$

We decompose

$$\widehat{v}(Z_t, t) - \widehat{v}(\widetilde{Z}_{t_j}, t_j) = \underbrace{\widehat{v}(Z_t, t) - \widehat{v}(\widetilde{Z}_t, t)}_{A_t} + \underbrace{\widehat{v}(\widetilde{Z}_t, t) - \widehat{v}(\widetilde{Z}_{t_j}, t)}_{B_t} + \underbrace{\widehat{v}(\widetilde{Z}_{t_j}, t) - \widehat{v}(\widetilde{Z}_{t_j}, t_j)}_{C_t}.$$

By spatial Lipschitzness, we have that $2\mathbb{E}\langle A_t, Z_t - \widetilde{Z}_t\rangle \le 2\gamma_1 R_t$. Moreover, since $\widetilde{Z}_t - \widetilde{Z}_{t_j} = (t - t_j)\widehat{v}(\widetilde{Z}_{t_j}, t_j)$, and $\|\widehat{v}\| \le K$, we have $\|B_t\| \le \gamma_1 K(t - t_j)$. Similarly, by time Lipschitzness, we have $\|C_t\| \le \gamma_2(t - t_j)$. Using $2\langle a, b\rangle \le \|a\|^2 + \|b\|^2$, we obtain

$$\frac{d}{dt}R_t \lesssim \gamma_1 R_t + (\gamma_1 K + \gamma_2)^2 (t - t_j)^2.$$

Applying Gronwall's inequality on $[0, T]$, and using $R_0 = 0$, gives

$$\begin{aligned} R_T &\lesssim e^{2\gamma_1 T}(\gamma_1 K + \gamma_2)^2 \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} (t - t_j)^2 \, dt \\ &= e^{2\gamma_1 T}(\gamma_1 K + \gamma_2)^2 \sum_{j=0}^{N-1} \frac{h_j^2}{3} \end{aligned}$$

Taking square roots yields

$$\begin{aligned} W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) &\le R_T^{1/2} \\ &\lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2)\sqrt{\sum_{j=0}^{N-1} h_j^3} \end{aligned}$$

For a uniform grid, $h_j = T/N$, so

$$\sqrt{\sum_{j=0}^{N-1} h_j^3} = \sqrt{N\frac{T^3}{N^3}} = \frac{T^{3/2}}{N}.$$

This proves the claim. □

F.3 Time-Clipping Optimal Flow Contribution

Lemma F.3 (Time-clipping error for the optimal flow)**.** *Suppose that $\pi_0 = \mathcal{N}(0, I_d)$ and that π_1 is supported on $[0,1]^d$. Let π_t^* denote the probability path induced by the optimal flow-matching velocity v^*, so that*

$$\pi_t^* = \mathrm{Law}\big((1-t)X_0 + tX_1\big), \qquad X_0 \sim \pi_0, \quad X_1 \sim \pi_1,$$

with X_0 and X_1 independent. Then, for every $T \in [0,1]$,

$$\begin{aligned} W_2(\pi_T^*, \pi_1) &\le (1-T) \cdot \sqrt{\mathbb{E}\|X_0 - X_1\|^2} \\ &\le \sqrt{2d}\,(1-T). \end{aligned}$$

In particular,

$$W_2(\pi_T^*, \pi_1) \lesssim (1-T)\sqrt{d}.$$

Proof. Let

$$X_T := (1-T)X_0 + TX_1.$$

By the definition of the optimal flow-matching probability path, $X_T \sim \pi_T^*$, while $X_1 \sim \pi_1$. Therefore, the joint law of (X_T, X_1) is an admissible coupling between π_T^* and π_1. Hence

$$\begin{aligned} W_2^2(\pi_T^*, \pi_1) &\le \mathbb{E}\|X_T - X_1\|^2 \\ &= (1-T)^2 \mathbb{E}\|X_0 - X_1\|^2. \end{aligned}$$

Since $X_0 \sim \mathcal{N}(0, I_d)$, we have $\mathbb{E}\|X_0\|^2 = d$. Since π_1 is supported on $[0,1]^d$, we have $\mathbb{E}\|X_1\|^2 \le d$. Moreover, X_0 is independent of X_1 and has mean zero, so

$$\begin{aligned} \mathbb{E}\|X_0 - X_1\|^2 &= \mathbb{E}\|X_0\|^2 + \mathbb{E}\|X_1\|^2 - 2\mathbb{E}\langle X_0, X_1\rangle \\ &\le 2d. \end{aligned}$$

Combining the previous equations gives

$$W_2(\pi_T^*, \pi_1) \le \sqrt{2d}\,(1-T),$$

as claimed. □

F.4 End-to-End Error Bound

Proof of Theorem 5.1. We decompose the total error into three parts from the triangle inequality applied to W_2 distance:

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_1) \le W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) + W_2(\pi_T^{\mathrm{cont}}, \pi_T^*) + W_2(\pi_T^*, \pi_1),$$

where π_T^{Euler} is the distribution obtained by Euler integration with learned ν_θ at early stopping time T, π_T^{cont} is the distribution obtained by continuous-time integration with learned ν_θ, π_T^* is the distribution obtained by continuous-time integration with optimal ν^*, and π_1 is the target distribution.

First, by Lemma F.2, since $\widehat{v} \in \mathrm{NN}(L, M, J, K, \kappa, \gamma_1, \gamma_2)$,

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2)\sqrt{\sum_{j=0}^{N-1} h_j^3}.$$

For a uniform grid $h_j = T/N$, this becomes

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2)\frac{T^{3/2}}{N}.$$

By Lemma F.1:

$$W_2(\pi_T^{\mathrm{cont}}, \pi_T^*) \le \sqrt{\epsilon}_{\mathrm{total}} \cdot T \cdot e^{L_\nu T},$$

where $\epsilon_{\mathrm{total}}$ (from Equation (15)) is the error of the estimated vector field after training.

Finally, we control the time-clipping error. Since the sampler is stopped at time $T < 1$, the optimal terminal law π_T^* does not exactly equal the target law π_1. However, this error is small because the optimal flow-matching path interpolates between π_0 and π_1. By Lemma F.3,

$$W_2(\pi_T^*, \pi_1) \lesssim (1-T)\sqrt{d}.$$

Combining all of the above bounds yield

$$\begin{aligned} W_2(\pi_T^{\mathrm{Euler}}, \pi_1) &\le W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) + W_2(\pi_T^{\mathrm{cont}}, \pi_T^*) + W_2(\pi_T^*, \pi_1) \\ &\lesssim e^{\gamma_1 T}(\gamma_1 K + \gamma_2)\frac{T^{3/2}}{N} + e^{L_\nu T}\sqrt{T\,\epsilon_{\mathrm{total}}} + (1-T)\sqrt{d}. \end{aligned}$$

To get $\epsilon_{\mathrm{train}} \lesssim n^{-\frac{2}{3(d+5)}}$, which, by Theorem 3.5, is achieved for

$$m \ge \Omega\Big(\mathrm{poly}\Big(n, n^{\frac{4}{3(d+5)}}, d, L, \frac{w_{\max}}{w_{\min}}, \log n\Big)\Big)$$

and

$$K = \widetilde{\Theta}\Big(d^{\frac{(1-a_0)}{2}} n^{2+\frac{4}{3(d+5)}}\Big)$$

projected-gradient steps. By choosing the mesh size so that

$$\max_{k=0,\dots,N-1} |t_{k+1} - t_k| = O\Big(n^{-\frac{4}{3(d+5)}}\Big),$$

and setting $T(n) = 1 - n^{-\frac{1}{3(d+5)}}$, the desired rate follows. □

F.5 KL Bound for Gaussian-Smoothed Terminal Laws

Proposition F.4. *Let μ and ν be probability measures on $\mathbb{R}^d$ with finite second moments. Let φ_σ denote the density of $\mathcal{N}(0, \sigma^2 I_d)$, and define*

$$\mathcal{S}_\sigma \mu := \mu * \varphi_\sigma, \qquad \mathcal{S}_\sigma \nu := \nu * \varphi_\sigma.$$

Then for every $\sigma > 0$,

$$\mathrm{KL}(\mathcal{S}_\sigma \mu \,\|\, \mathcal{S}_\sigma \nu) \le \frac{1}{2\sigma^2} W_2^2(\mu, \nu).$$

Proof. Fix any coupling $\gamma \in \Pi(\mu, \nu)$, and let $(X, Y) \sim \gamma$. Let $G \sim \mathcal{N}(0, I_d)$ be independent of (X, Y). Then the laws of $X + \sigma G$ and $Y + \sigma G$ are precisely $\mathcal{S}_\sigma \mu$ and $\mathcal{S}_\sigma \nu$, respectively.

Conditioned on $(X, Y) = (x, y)$, the conditional laws are Gaussian:

$$X + \sigma G \mid (X, Y) = (x, y) \sim \mathcal{N}(x, \sigma^2 I_d), \qquad Y + \sigma G \mid (X, Y) = (x, y) \sim \mathcal{N}(y, \sigma^2 I_d).$$

By convexity of relative entropy under mixing (as in Anand and Umans (2023); Anand (2025)),

$$\mathrm{KL}(\mathcal{S}_\sigma \mu \,\|\, \mathcal{S}_\sigma \nu) \le \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathrm{KL}\big(\mathcal{N}(x, \sigma^2 I_d) \,\big\|\, \mathcal{N}(y, \sigma^2 I_d)\big)\, d\gamma(x, y).$$

Since two Gaussians with common covariance $\sigma^2 I_d$ satisfy

$$\mathrm{KL}\big(\mathcal{N}(x, \sigma^2 I_d) \,\big\|\, \mathcal{N}(y, \sigma^2 I_d)\big) = \frac{\|x - y\|^2}{2\sigma^2},$$

it follows that

$$\mathrm{KL}(\mathcal{S}_\sigma \mu \,\|\, \mathcal{S}_\sigma \nu) \le \frac{1}{2\sigma^2} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\gamma(x, y).$$

Finally, taking the infimum over all couplings $\gamma \in \Pi(\mu, \nu)$ yields

$$\mathrm{KL}(\mathcal{S}_\sigma \mu \,\|\, \mathcal{S}_\sigma \nu) \le \frac{1}{2\sigma^2} W_2^2(\mu, \nu).$$

□

Corollary F.5 (Restatement of Corollary 5.2)**.** *Let π_T^{Euler} be the terminal law generated by the learned flow with discretization, and let π_1 denote the target law. Then for every $\sigma > 0$,*

$$\mathrm{KL}\big(\mathcal{S}_\sigma \pi_T^{\mathrm{Euler}} \,\big\|\, \mathcal{S}_\sigma \pi_1\big) \le \frac{1}{2\sigma^2} W_2^2(\pi_T^{\mathrm{Euler}}, \pi_1).$$

In particular,

$$\mathrm{KL}\big(\mathcal{S}_\sigma \pi_T^{\mathrm{Euler}} \,\big\|\, \mathcal{S}_\sigma \pi_1\big) \lesssim \tilde{O}\left(\frac{n^{-2/3(d+5)}}{\sigma^2}\right).$$

Proof. The first claim is immediate from Proposition F.4 with $\mu = \pi_T^{\mathrm{Euler}}$ and $\nu = \pi_1$. The second follows by substituting the Wasserstein bound and using $(a+b)^2 \le 2(a^2+b^2)$. □

F.6 Reduced step count via an adaptive Euler schedule

Let $\delta := 1 - T$. Suppose the learned velocity satisfies $\|v_\theta(x,t)\| \le K_{\mathrm{out}}$ and $\|v_\theta(x,t) - v_\theta(y,t)\| \le \gamma_1 \|x - y\|$, and the local time-modulus condition

$$\|v_\theta(x,t) - v_\theta(x,s)\| \le \gamma_U \left| \int_s^t (1-r)^{-4} \,\mathrm{d}r \right|.$$

Define $q(t) := \gamma_1 K_{\mathrm{out}} + \gamma_U (1-t)^{-4}$, and choose the Euler mesh by

$$\frac{\int_0^{t_j} q(r)^{1/2} \,\mathrm{d}r}{\int_0^T q(r)^{1/2} \,\mathrm{d}r} = \frac{j}{N_{\mathrm{Eul}}}, \qquad j = 0, \dots, N_{\mathrm{Eul}}.$$

Then

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim \frac{e^{\gamma_1 T}}{N_{\mathrm{Eul}}} \left(\int_0^T q(t)^{1/2} \,\mathrm{d}t \right)^2.$$

Consequently,

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim \frac{e^{\gamma_1 T}}{N_{\mathrm{Eul}}} \left(\gamma_1 K_{\mathrm{out}} + \gamma_U (1-T)^{-2} + 2(\gamma_1 K_{\mathrm{out}} \gamma_U)^{1/2} (1-T)^{-1} \right).$$

In particular, if $\gamma_U = \widetilde{O}(1)$ and $\gamma_1 K_{\mathrm{out}} \lesssim \widetilde{O}((1-T)^{-2})$ then

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim \widetilde{O}\left(\frac{(1-T)^{-2}}{N_{\mathrm{Eul}}} \right).$$

Thus, to make the Euler discretization error no larger than the clipping scale $1 - T$, it suffices to take $N_{\mathrm{Eul}} \gtrsim \widetilde{O}((1-T)^{-3})$. With the end-to-end choice $1 - T = n^{-1/(3(d+5))}$, this becomes $N_{\mathrm{Eul}} \gtrsim \widetilde{O}\big(n^{1/(d+5)}\big)$. When the time-singular term dominates $q(t)$, the mesh has the explicit form

$$t_j = 1 - \left[1 + \frac{j}{N_{\mathrm{Eul}}} \left((1-T)^{-1} - 1 \right) \right]^{-1}.$$

Proof. Couple the continuous learned ODE Z_t and the Euler interpolation $\widehat{Z}_t$ by sharing the initial condition $Z_0 = \widehat{Z}_0 \sim \pi_0$, and let $t(s) := t_j$ for $s \in [t_j, t_{j+1})$ denote the latest mesh point.

On $[t_j, t_{j+1}]$ the Euler step gives $\|\widehat{Z}_t - \widehat{Z}_{t_j}\| \le K_{\mathrm{out}}(t - t_j)$, so combining spatial Lipschitzness of v_θ with the time-modulus assumption,

$$\begin{aligned} \|v_\theta(\widehat{Z}_t, t) - v_\theta(\widehat{Z}_{t_j}, t_j)\| &\le \gamma_1 K_{\mathrm{out}}(t - t_j) + \gamma_U \int_{t_j}^t (1-r)^{-4} \mathrm{d}r \\ &= \int_{t_j}^t q(r) \mathrm{d}r. \end{aligned}$$

For $s \in [t_j, t_{j+1})$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\|Z_s - \widehat{Z}_s\| &\le \|v_\theta(Z_s, s) - v_\theta(\widehat{Z}_s, s)\| + \|v_\theta(\widehat{Z}_s, s) - v_\theta(\widehat{Z}_{t(s)}, t(s))\| \\ &\le \gamma_1 \|Z_s - \widehat{Z}_s\| + \int_{t(s)}^{s} q(r)\mathrm{d}r.\end{aligned}$$

Gronwall's inequality, together with $Z_0 = \widehat{Z}_0$, gives

$$\begin{aligned}\|Z_T - \widehat{Z}_T\| &\le e^{\gamma_1 T} \int_0^T \int_{t(s)}^{s} q(r)\mathrm{d}r\mathrm{d}s \\ &= e^{\gamma_1 T} \sum_{j=0}^{N_{\mathrm{Eul}}-1} \int_{t_j}^{t_{j+1}} \int_{t_j}^{t} q(r)\mathrm{d}r\mathrm{d}t.\end{aligned}$$

Switching the order of integration on each subinterval converts the inner double integral into $\int_{t_j}^{t_{j+1}} (t_{j+1} - r)\, q(r)\, \mathrm{d}r$. Taking expectations under the synchronous coupling and using $W_2^2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \le \mathbb{E}\,\|Z_T - \widehat{Z}_T\|^2$ yields

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim e^{\gamma_1 T} \sum_{j=0}^{N_{\mathrm{Eul}}-1} \int_{t_j}^{t_{j+1}} (t_{j+1} - r)\, q(r)\, \mathrm{d}r.$$

Let $\rho : [0, T] \to (0, \infty)$ be a smooth density with $\int_0^T \rho(t)\, dt = 1$, and generate the mesh by $\int_0^{t_j} \rho(r)\, dr = j/N_{\mathrm{Eul}}$. Then $t_{j+1} - t_j$ is of order $1/(N_{\mathrm{Eul}}\rho(t))$ uniformly for $t \in [t_j, t_{j+1}]$, so

$$\int_{t_j}^{t_{j+1}} (t_{j+1} - r)\, q(r)\, \mathrm{d}r \le (t_{j+1} - t_j) \int_{t_j}^{t_{j+1}} q(r)\, \mathrm{d}r \lesssim \frac{1}{N_{\mathrm{Eul}}} \int_{t_j}^{t_{j+1}} \frac{q(t)}{\rho(t)}\, \mathrm{d}t.$$

Summing yields the asymptotic majorant

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim \frac{e^{\gamma_1 T}}{N_{\mathrm{Eul}}} \int_0^T \frac{q(t)}{\rho(t)}\, \mathrm{d}t.$$

Among all densities ρ with $\int_0^T \rho = 1$, we have that

$$\begin{aligned}\left(\int_0^T \sqrt{q(t)}\, \mathrm{d}t\right)^2 &= \left(\int_0^T \frac{\sqrt{q(t)}}{\sqrt{\rho(t)}} \sqrt{\rho(t)}\mathrm{d}t\right)^2 \\ &\le \left(\int_0^T \frac{q(t)}{\rho(t)}\mathrm{d}t\right)\left(\int_0^T \rho(t)\mathrm{d}t\right) \\ &= \int_0^T \frac{q(t)}{\rho(t)}\mathrm{d}t,\end{aligned}$$

with equality iff $\rho(t) \propto \sqrt{q(t)}$.

Imposing the unit-mass constraint pins down $\rho^\star(t) = q(t)^{1/2} / \int_0^T \sqrt{q(r)}\mathrm{d}r$, and at this optimum

$$W_2(\pi_T^{\mathrm{Euler}}, \pi_T^{\mathrm{cont}}) \lesssim \frac{e^{\gamma_1 T}}{N_{\mathrm{Eul}}} \left(\int_0^T \sqrt{q(t)}\mathrm{d}t\right)^2.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ together with $T \le 1$ and $(1-T)^{-1} - 1 \le (1-T)^{-1}$ gives that for

$$\begin{aligned}\int_0^T q(t)^{1/2}\mathrm{d}t &\le T\sqrt{\gamma_1 K_{\mathrm{out}}} + \sqrt{\gamma_U}\int_0^T (1-t)^{-2}\mathrm{d}t \\ &\le \sqrt{\gamma_1 K_{\mathrm{out}}} + \sqrt{\gamma_U}(1-T)^{-1}.\end{aligned}$$

Squaring and applying $(\sqrt{a}+\sqrt{b})^2 = a+b+2\sqrt{ab}$ with $a = \gamma_1 K_{\text{out}}$ and $b = \gamma_U(1-T)^{-2}$ gives

$$\left(\int_0^T q(t)^{1/2}\,dt\right)^2 \le \gamma_1 K_{\text{out}} + \gamma_U(1-T)^{-2} + 2\sqrt{\gamma_1 K_{\text{out}}\gamma_U}(1-T)^{-1},$$

which is the given three-term bound.

When $\gamma_U(1-t)^{-4}$ dominates inside q, we have $\sqrt{q(t)} \asymp \sqrt{\gamma_U}(1-t)^{-2}$, so $\int_0^t \sqrt{q(r)}dr \asymp \sqrt{\gamma_U}\big[(1-t)^{-1}-1\big]$. The prescription $\int_0^{t_j}\sqrt{q}/\int_0^T\sqrt{q} = j/N_{\text{Eul}}$ reduces to

$$(1-t_j)^{-1} - 1 = \frac{j}{N_{\text{Eul}}}\big[(1-T)^{-1}-1\big],$$

and solving for t_j yields the closed form

$$t_j = 1 - \left[1 + \frac{j}{N_{\text{Eul}}}\big((1-T)^{-1}-1\big)\right]^{-1},$$

which clusters mesh points near $t=1$ at the rate dictated by the singularity of q.

The remaining claims of the statement follow by direct substitution: under $\gamma_U = \widetilde{O}(1)$ and $\gamma_1 K_{\text{out}} \lesssim \widetilde{O}((1-T)^{-2})$, each of the three terms is of order $(1-T)^{-2}$, so

$$W_2(\pi_T^{\text{Euler}}, \pi_T^{\text{cont}}) \lesssim \widetilde{O}\big((1-T)^{-2}/N_{\text{Eul}}\big).$$

Requiring this to be no larger than $1-T$ forces $N_{\text{Eul}} \gtrsim \widetilde{O}((1-T)^{-3})$, and substituting $1-T = n^{-\frac{1}{3(d+5)}}$ gives $N_{\text{Eul}} \gtrsim \widetilde{O}(n^{\frac{1}{d+5}})$. □

G Properties of the true velocity field

This section provides an explicit form of the optimal velocity field v^*, properties of its partial derivative with respect to time, and the Lipschitz continuity proofs of Lemma 2.3. The lemmas and proofs in this section closely mirror those of Zhou and Liu (2025). We begin by computing the velocity field.

Lemma G.1 (Lemma D.1 in Zhou and Liu (2025)). *If π_t is the density of $X_t := (1-t)X_0 + tX_1$, the true velocity field v^* obeys*

$$v^*(x,t) = \frac{1-t}{t}\nabla \log \pi_t(x) + \frac{1}{t}x.$$

Proof. Let $\pi_{t|1}$ be the density of X_t conditioned on X_1. Then, by algebraic manipulation,

$$\begin{aligned}
v^*(x,t) &= \mathbb{E}[X_1 - X_0 | X_t = x] \\
&= \mathbb{E}\left[X_1 - \frac{1}{1-t}((1-t)X_0 + tX_1)|X_t = x\right] \\
&= \frac{1}{1-t}\mathbb{E}[X_1|X_t = x] - \frac{1}{1-t}x \\
&= \frac{1}{1-t}\int \frac{x_1 \pi_{t|1}(x|x_1)\pi_1(x_1)}{\pi_t(x)}\mathrm{d}x_1 - \frac{1}{1-t}x \\
&= \frac{1}{1-t}\int \frac{1}{\sqrt{(2\pi)^d(1-t)^{2d}}}\frac{x_1 \exp(-\frac{\|x-tx_1\|^2}{2(1-t)^2})\pi_1(x_1)}{\pi_t(x)}\mathrm{d}x_1 - \frac{1}{1-t}x \\
&= \frac{1-t}{t}\int \frac{1}{\sqrt{(2\pi)^d(1-t)^{2d}}}\frac{\left(\frac{tx_1-x}{(1-t)^2} + \frac{x}{(1-t)^2}\right)\exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\pi_1(x_1)}{\pi_t(x)}\mathrm{d}x_1 - \frac{1}{1-t}x \\
&= \frac{1-t}{t}\int \frac{1}{\sqrt{(2\pi)^d(1-t)^{2d}}}\frac{\nabla_x \exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\pi_1(x_1)}{\pi_t(x)}\mathrm{d}x_1 + \left(\frac{1}{t(1-t)} - \frac{1}{1-t}x\right) \\
&= \frac{1-t}{t}\nabla_x \log \pi_t(x) + \frac{1}{t}x,
\end{aligned}$$

which completes the proof. □

Lemma G.2 (Lemma D.2 in Zhou and Liu (2025)). *We have that*

$$\frac{\partial v^*(x,t)}{\partial t} = \frac{\mathbb{E}[X_1|X_t=x]-x}{(1-t)^2} + \frac{(1+t)\mathrm{Cov}[X_1|X_t=x]x - t(\mathbb{E}[X_1\|X_1\|^2|X_t=x) - \mathbb{E}[X_1|X_t=x]\mathbb{E}[\|X_1\|^2|X_t=x])}{(1-t)^4},$$

where $\mathrm{Cov}[X_1|X_t=x]$ *is the covariance matrix of* X_1 *conditioned on* $X_t=x$.

Proof. Define $\phi_t(x) \coloneqq \int \exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\pi_1(\mathrm{d}x_1)$, which is the unnormalized version of $\pi_t(x)$, which satisfies $\nabla \log \phi_t(x) = \nabla \log \pi_t(x)$. Using the product rule of the derivatives, this implies

$$\begin{aligned}
\partial_t v^*(x,t) &= -\frac{1}{t^2}\nabla \log \pi_t(x) + \frac{1-t}{t}\partial_t \nabla \log \pi_t(x) - \frac{1}{t^2}x \\
&= -\frac{1}{t(1-t)^2}\mathbb{E}[X_1|X_t=x] + \frac{1}{t^2(1-t)^2}x + \frac{1-t}{t}\partial_t\left(\frac{\nabla\phi_t(x)}{\phi_t(x)}\right) - \frac{1}{t^2}x \\
&= \frac{2-t}{t(1-t)^2}x - \frac{1}{t(1-t)^2}\mathbb{E}[X_1|X_t=x] + \frac{1-t}{t}\left(\frac{\partial_t\nabla\phi_t(x)}{\phi_t(x)} - \frac{\partial_t\phi_t(x)\nabla\phi_t(x)}{(\phi_t(x))^2}\right)
\end{aligned}$$

Here, note that

$$\begin{aligned}
\frac{\partial_t\nabla\phi_t(x)}{\phi_t(x)} &= \frac{1}{\phi_t(x)}\partial_t\int \frac{tx_1-x}{(1-t)^2}\exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\pi_1(\mathrm{d}x_1) \\
&= \frac{1}{\phi_t(x)}\int\left(\frac{(1-t)^2x_1 - 2(tx_1-x)(t-1)}{(1-t)^4}\exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right) - \right. \\
&\quad \left.\frac{tx_1-x}{(1-t)^2}\exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\frac{(t\|x_1\|^2 - x_1^\top x)(1-t)^2 - (t-1)\|x-tx_1\|^2}{(1-t)^4}\right)\pi_1(\mathrm{d}x_1) \\
&= \frac{1+t}{(1-t)^3}\mathbb{E}[X_1|X_t=x] - \frac{2}{(1-t)^3}x - \frac{t^2}{(1-t)^5}\mathbb{E}[X_1\|X_1\|^2|X_t=x] \\
&\quad + \frac{t(1+t)}{(1-t)^5}\mathbb{E}[X_1X_1^\top|X_t=x]x - \frac{t}{(1-t)^5}\mathbb{E}[X_1|X_t=x]\|x\|^2 \\
&\quad + \frac{t}{(1-t)^5}\mathbb{E}[\|X_1\|^2|X_t=x]x - \frac{1+t}{(1-t)^5}\mathbb{E}[X_1^\top x|X_t=x] + \frac{\|x\|^2x}{(1-t)^5}.
\end{aligned}$$

Then, via some calculus,

$$\begin{aligned}
&\frac{\partial_t\phi_t(x)}{\phi_t(x)} \\
&\quad = \frac{1}{\phi_t(x)}\int -\frac{(t\|x_1\|^2 - x_1^\top x)(1-t)^2 + \|x-tx_1\|^2(1-t)}{(1-t)^4}\exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\pi_1(\mathrm{d}x_1) \\
&\quad = -\frac{t}{(1-t)^3}\mathbb{E}[\|X_1\|^2|X_t=x] + \frac{1+t}{(1-t)^3}\mathbb{E}[X_1^\top x|X_t=x] - \frac{\|x\|^2}{(1-t)^3},
\end{aligned}$$

and we also have that

$$\begin{aligned}
\frac{\nabla\phi_t(x)}{\phi_t(x)} &= \frac{1}{\phi_t(x)}\int\frac{tx_1-x}{(1-t)^2}\exp\left(-\frac{\|x-tx_1\|^2}{2(1-t)^2}\right)\pi_1(\mathrm{d}x_1) \\
&= -\frac{x}{(1-t)^2} + \frac{t}{(1-t)^2}\mathbb{E}[X_1|X_t=x].
\end{aligned}$$

Combining these calculations gives that

$$\begin{aligned}
\partial_t v^*(x,t) = &-\frac{1}{(1-t)^2}x + \frac{1}{(1-t)^2}\mathbb{E}[X_1|X_t=x] + \frac{1+t}{(1-t)^4}\mathrm{Cov}[X_1|X_t=x]x \\
&- \frac{t}{(1-t)^4}(\mathbb{E}[X_1\|X_1\|^2|X_t=x] - \mathbb{E}[X_1|X_t=x]\mathbb{E}[\|X_1\|^2|X_t=x]),
\end{aligned}$$

which completes the proof. □

Lemma G.3 (Lemmas D.3 and D.4 in Zhou and Liu (2025)). *We have that*

$$\sup_{t\in[0,T]}\sup_{x\in[-R,R]^d}|v_i^*(x,t)|\le\frac{1+R}{1-T},$$

and

$$\sup_{t\in[0,T]}\sup_{x\in[-R,R]^d}|\partial_t v_i^*(x,t)|=O\left(\frac{d^{3/2}(1+R)}{(1-T)^4}\right).$$

Proof. For the first identity, note that the i th coordinate of the ground-truth velocity function v^* satisfies

$$v_i^*=\frac{1}{1-t}\mathbb{E}[X_1^{(i)}|X_t=x]-\frac{1}{1-t}x_i,$$

where $X_1^{(i)}$ denotes the i'th coordinate of X_1. Note that π_1 is supported on $[-1,1]^d$. Then,

$$\sup_{t\in[0,t]}\sup_{x\in[-R,R]^d}|v_i^*(x,t)|\le\frac{1+R}{1-T}.$$

For the second inequality, note from Lemma G.2 that

$$\begin{aligned}\|\partial_t v^*(x,t)\|\le{}&\frac{1}{(1-t)^2}\|x\|+\frac{1}{(1-t)^2}\|\mathbb{E}[X_1|X_t=x]\|+\frac{1+t}{(1-t)^4}\|\mathrm{Cov}(X_1|X_t=x)\|_{op}\|x\|\\&+\frac{t}{(1-t)^4}\left(\|\mathbb{E}[X_1\|X_1\|^2|X_t=x]\|+\|\mathbb{E}[X_1|X_t=x]\|\cdot\|\mathbb{E}[\|X_1\|^2|X_t=x]\|\right).\end{aligned}$$

Since, by assumption, π_1 is supported on $[0,1]^d$, it holds that

$$\|\mathbb{E}[X_1|X_t=x]\|\le\sqrt{\mathbb{E}[\|X_1\|^2|X_t=x]}\le\sqrt{d},$$

and

$$\|\mathbb{E}[X_1\|X_1\|^2|X_t=x]\|\le\sqrt{\mathbb{E}[\|X_1\|^6|X_t=x]}\le d^{3/2}.$$

The bound for $\|\mathrm{Cov}[X_1|X_t=x]\|_{op}$, uses the following inequality for any $u\in\mathbb{R}^d$,

$$\begin{aligned}u^\top\mathrm{Cov}[X_1|X_t=x]u&=\mathbb{E}[u^\top X_1X_1^\top u|X_t=x]-\mathbb{E}[u^\top X_1|X_t=x]\mathbb{E}[X_1^\top u|X_t=x]\\&=\mathbb{E}[(u^\top X_1)^2|X_t=x]-\mathbb{E}[u^\top X_1|X_t=x]^2\\&\le 2d\|u\|^2.\end{aligned}$$

Therefore, $\|\mathrm{Cov}[X_1|X_t=x]\|_{op}\le 2d$. Together, these imply

$$\sup_{t\in[0,T]}\sup_{x\in[-R,R]^d}\|\partial_t v^*(x,t)\|\le\frac{R\sqrt{d}}{(1-T)^2}+\frac{\sqrt{d}}{(1-T)^2}+\frac{1+T}{(1-T)^4}2d^{3/2}R+\frac{2Td^{3/2}}{(1-T)^4}.$$

Finally, noting that $T<1$ allows the above inequality to simplify to

$$\sup_{t\in[0,T]}\sup_{x\in[-R,R]^d}\|\partial_t v^*(x,t)\|=O\left(\frac{d^{3/2}(1+R)}{(1-T)^4}\right),$$

which completes the proof. □

Lemma G.4 (Lemma D.5 in Zhou and Liu (2025)). *The Jacobian matrix of the velocity field and the conditional covariance matrix enjoy the relationship*

$$\nabla v^*(x,t)=\frac{t}{(1-t)^3}\mathrm{Cov}[X_1|X_t=x]-\frac{1}{1-t}\mathbf{I}_d.$$

Proof. Since $\nabla v^*(x,t) = \frac{1-t}{t}\nabla^2 \log \pi_t(x) + \frac{1}{t}\mathbf{I}_d$, the Hessian $\nabla^2 \log \pi_t(x)$ can be computed as

$$\begin{aligned}
\nabla^2 \log \pi_t(x) &= \nabla \left(\frac{\int_{\mathbb{R}^d} \frac{tx_1 - x}{(1-t)^2} \exp(-\|x - tx_1\|^2/(1-t)^2)\pi_1(\mathrm{d}x_1)}{\int_{\mathbb{R}^d} \exp(-\|x - tx_1\|^2/(1-t)^2)\pi_1(\mathrm{d}x_1)} \right) \\
&= -\frac{1}{(1-t)^2}\mathbf{I}_d + \frac{\int_{\mathbb{R}^d} \left(\frac{tx_1 - x}{(1-t)^2}\right)^{\otimes 2} \exp(-\|x - tx_1\|^2/(1-t)^2)\pi_1(\mathrm{d}x_1)}{\int_{\mathbb{R}^d} \exp(-\|x - tx_1\|^2/(1-t)^2)\pi_1(\mathrm{d}x_1)} \\
&\quad - \left(\frac{\int_{\mathbb{R}^d} \frac{tx_1 - x}{(1-t)^2} \exp(-\|x - tx_1\|^2/(1-t)^2\pi_1(\mathrm{d}x_1))}{\int_{\mathbb{R}^d} \exp(-\|x - tx_1\|^2/(1-t)^2)\pi_1(\mathrm{d}x_1)} \right)^{\otimes 2} \\
&= -\frac{1}{(1-t)^2}\mathbf{I}_d + \frac{t^2}{(1-t)^4}\mathrm{Cov}[X_1 | X_t = x].
\end{aligned}$$

Combining these identities yields the desired result. □

Lemma G.5 (Brascamp-Leib inequality, (Brascamp and Lieb, 1976))**.** *Let $\mu(\mathrm{d}x) = \exp(-U(x))\mathrm{d}x$ be a probability measure on a convex set $\Omega \subseteq \mathbb{R}^d$ whose potential $U : \Omega \to \mathbb{R}$ is twice continuously differentiable and strictly convex. Then, $\mathrm{Cov}_\mu(X) \preceq \mathbb{E}_\mu[(\nabla^2 U(X))^{-1}]$, with equality if $X \sim \mathcal{N}(m, \Sigma)$ with Σ positive definite.*

Lemma G.6 (Cramér-Rao inequality, (Saumard and Wellner, 2014; Dembo et al., 1991))**.** *Let $\mu(\mathrm{d}x) = \exp(-U(x))\mathrm{d}x$ be a probability measure on a convex set $\Omega \subseteq \mathbb{R}^d$ whose potential $U : \Omega \to \mathbb{R}$ is twice continuously differentiable. Then, $\mathrm{Cov}_\mu(X) \succeq \mathbb{E}_\mu[(\nabla^2 U(X))^{-1}]$, with equality if $X \sim \mathcal{N}(m, \Sigma)$ with Σ positive definite.*

Lemma G.7 (Lemma D.6 in Zhou and Liu (2025))**.** *Suppose Assumption 2.1. Then, $v^*(x,t)$ is ξ-Lipschitz continuous with respect to x on $\mathbb{R}^d \times [0,T]$, where $\xi \le \max\left\{\frac{1}{1-T}, \frac{Td}{(1-T)^3}\right\}$. Further, if $\frac{1}{2} < T < 1$, then v^* is $\frac{d}{(1-T)^3}$-Lipschitz continuous with respect to x.*

Proof. Since the target distribution π_1 is supported on $[0,1]^d$, the covariance matrix can be evaluated via $0 \preceq \mathrm{Cov}[X_1 | X_t = x] \preceq d\mathbf{I}_d$. This then implies

$$-\frac{1}{1-t}\mathbf{I}_d \preceq \nabla v^*(x,t) \preceq \left(\frac{td}{(1-t)^3} - \frac{1}{1-t}\right)\mathbf{I}_d,$$

which implies the Lipschitz constant of v^* with respect to x. □

Lemma G.8 (Lemma D.9 in Zhou and Liu (2025))**.** *Suppose Assumption 2.1 and Assumption 2.2. Then, $v^*(x,t)$ is $\zeta(\alpha, d)$-Lipschitz continuous on $\mathbb{R}^d \times [0,1]$ in x, where*

$$\zeta(\alpha, d) = \frac{d}{2}\left(\alpha + \sqrt{\alpha + \frac{2}{d}}\right)^2.$$

Proof. Note that

$$-\nabla^2_{x_1} \log \pi_{1|t}(x_1|x) = -\nabla^2_{x_1} \log \pi_1(x_1) - \nabla^2_{x_1} \log \pi_{t|1}(x|x_1),$$

where $\pi_{1|t}$ is the conditional density of X_1 conditioned on $X_t = x$ and $\pi_{t|1}$ is the conditional density of X_t conditioned on $X_1 = x_1$.

Since X_t can be viewed as tX_1 perturbed by a Gaussian noise, we have

$$\pi_{t|1}(x|x_1) \propto \exp\left(-\frac{\|x - tx_1\|^2}{(1-t)^2}\right).$$

Thus,

$$-\nabla_{x_1}^2 \log \pi_{1|t}(x_1|x) = -\nabla_{x_1}^2 \log \pi_1(x_1) + \frac{t^2}{(1-t)^2}\mathbf{I}_d.$$

Then, Assumption 2.2 implies

$$\left(-\alpha + \frac{t^2}{(1-t)^2}\right)\mathbf{I}_d \preceq -\nabla_{x_1}^2 \log \pi_{1|t}(x_1|x) \preceq \left(\alpha + \frac{t^2}{(1-t)^2}\right)\mathbf{I}_d.$$

By the Cramér-Rao inequality,

$$\mathrm{Cov}[X_1|X_t = x] \succeq \left(\alpha + \frac{t^2}{(1-t)^2}\right)^{-1}\mathbf{I}_d.$$

When $t \in \{t \in (0,1) : -\alpha + \frac{t^2}{(1-t)^2} > 0\}$, Brascamp-Lieb's inequality gives

$$\mathrm{Cov}[X_1|X_t = x] \preceq \left(-\alpha + \frac{t^2}{(1-t)^2}\right)^{-1}\mathbf{I}_d.$$

This gives rise to the property

$$\nabla v^*(x,t) \succeq \frac{t - \alpha(1-t)}{\alpha(1-t)^2 + t^2}\mathbf{I}_d,$$

and for $t \in \{t \in (0,1) : -\alpha + \frac{t^2}{(1-t)^2} > 0\}$,

$$\nabla v^*(x,t) \preceq \frac{t + \alpha(1-t)}{-\alpha(1-t)^2 + t^2}\mathbf{I}_d.$$

Therefore,

$$-\frac{1}{1-t}\mathbf{I}_d \preceq \nabla v^*(x,t) \preceq \left(\frac{td}{(1-t)^3} - \frac{1}{1-t}\right)\mathbf{I}_d.$$

Then, using the property $-\frac{1}{1-t} \leq \frac{t-\alpha(1-t)}{\alpha(1-t)^2+t^2}$ yields the bound

$$\nabla v^*(x,t) \succeq \frac{t - \alpha(1-t)}{\alpha(1-t)^2 + t^2}\mathbf{I}_d \succeq -\frac{\alpha}{\alpha/(1+\alpha)}\mathbf{I}_d = -(1+\alpha)\mathbf{I}_d,$$

where the second inequality follows from the observation that $t - \alpha(1-t) \geq -\alpha$ on $t \in (0,1)$ and $\alpha(1-t)^2 + t^2 \geq \frac{\alpha}{1+\alpha}$.

Next, letting $\frac{td}{(1-t)^3} - \frac{1}{1-t} = \frac{1}{1-t} \cdot \frac{t}{-\alpha(1-t)^2+t^2} = \frac{1}{1-t}$ yields

$$\frac{2}{d} + \alpha = \frac{t^2}{(1-t)^2},$$

which has a $(0,1)$-root in $\frac{\sqrt{\alpha+2/d}}{1+\sqrt{\alpha+2/d}}$.

By the monotonicity of $\frac{x}{1+x}$ on $(0,1)$,

$$\frac{\sqrt{\alpha+2/d}}{1+\sqrt{\alpha+2/d}} > \frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}.$$

Hence, $\nabla v^*(x,t) \preceq g(t)\mathbf{I}_d$, where

$$g(t) = \begin{cases} \left(\frac{td}{(1-t)^3} - \frac{1}{1-t}\right), & t \in \left(0, \frac{\sqrt{\alpha+2/d}}{1+\sqrt{\alpha+2/d}}\right) \\ \frac{t+\alpha(1-t)}{-\alpha(1-t)^2+t^2}, & t \in \left(\frac{\sqrt{\alpha+2/d}}{1+\sqrt{\alpha+2/d}}, 1\right). \end{cases}$$

Taking the derivative of $\frac{td}{(1-t)^3} - \frac{1}{1-t}$, one can observe that it is increasing on $(0, \frac{\sqrt{\alpha+2/d}}{1+\sqrt{\alpha+2/d}})$.

Using the same argument, $\frac{t+\alpha(1-t)}{-\alpha(1-t)^2+t^2}$ is decreasing on $(\frac{\sqrt{\alpha+2/d}}{1+\sqrt{\alpha+2/d}}, 1)$.

Therefore,

$$\nabla v^*(x,t) \preceq \frac{d}{2}\left(\alpha + \sqrt{\alpha + \frac{2}{d}}\right)^2 \mathbf{I}_d,$$

which completes the proof. □

H Numerical Simulations

We complement our theoretical analysis with numerical experiments[1] examining how training and sampling choices affect the accuracy of the generated distribution. Our experiments vary the ambient dimension, the number of optimization iterations, the numerical integration budget, and the training-set size. We evaluate distributional accuracy using the plotted Wasserstein-2 error $W_2(\widehat{\rho}_n^{\mathrm{FM}}, \mu)$, where $\mu = \pi_1$ denotes the target distribution and $\widehat{\rho}_n^{\mathrm{FM}}$ denotes the distribution produced by the learned flow. These experiments illustrate the qualitative dependencies identified by our analysis and the interaction between different sources of error. They are intended to complement the theoretical guarantees rather than establish the sharpness of their asymptotic rates.

Dependence on dimension. Figure 3 examines the effect of ambient dimension with $n = 4096$ training samples, using six random seeds for each dimension. Over the tested range $d \in \{2, \dots, 7\}$, the reported Wasserstein error increases from approximately 0.04 to 0.13. Thus, the same sample budget yields substantially different distributional accuracy across dimensions. This behavior is qualitatively consistent with the dimension dependence of our guarantees and motivates accounting for dimension when allocating training resources. The experiment measures the overall difficulty of the tested problems; it does not separately identify the contributions of statistical estimation, function approximation, and optimization.

Effect of additional training. Figure 4 examines how terminal distributional accuracy changes with the number of training iterations for $d = 2$ and $n = 2048$. The curve reports the median over six seeds, with the shaded region showing the interquartile range. Generation error decreases substantially during the initial training phase, followed by smaller improvements beyond approximately 4000 iterations. This behavior is consistent with the structure of our error decomposition: additional optimization can improve the learned velocity field, while other sources of error may eventually limit its effect on generation quality. Because the plotted quantity is terminal Wasserstein error, this experiment measures the downstream benefit of training rather than directly estimating the convergence rate of the empirical training loss.

Evolution of generated samples during training. Figure 5 provides a qualitative view of the training dependence by comparing generated samples after 10, 100, 1000, and 10000 optimization steps. The panels distinguish samples from the target distribution, the training observations, and samples produced by the learned flow. Early in training, the generated samples are dispersed and do not reproduce the target's separated clusters. As training proceeds, the generated distribution increasingly captures their locations and concentration. In the displayed run, the reported Wasserstein error decreases from 1.064 after 10 steps to 0.029 after 10000 steps. These visualizations connect improvements in the distributional metric to changes in the geometry of the generated samples.

Effect of numerical integration. Figure 6 compares Euler integration and fourth-order Runge–Kutta integration as the number of ODE steps increases, for $d = 2$, $n = 8000$, and five random seeds. Euler integration benefits substantially from refining the initially coarse time grid, before approaching a common error level with RK4. RK4 reaches this regime with fewer integration steps. The resulting plateau is consistent with a setting in which further reductions in numerical integration error have little effect on the remaining generation error. The dashed inverse-step reference illustrates the first-order scaling relevant to Euler integration; the plotted curves themselves measure total generation error, including errors in the learned velocity field. This comparison is expressed in integration steps:

[1] Our code is provided in `https://anonymous.4open.science/r/flow_matching_full_experiments-3BB6/README.md`

because a standard RK4 step requires four velocity evaluations, it should not be interpreted as a comparison at equal computational cost.

Dependence on training-set size. Figure 7 examines distributional accuracy as the number of training samples increases from 128 to 8192, separately for $d = 2$ and $d = 3$. Each panel reports the median over eight seeds and the corresponding interquartile range. Although the curves are not monotone at every sample size, both exhibit an overall decrease in error. The fitted log–log slopes are approximately -0.08 and -0.10, respectively, indicating a measurable benefit from additional training data over the tested range. These slopes describe the finite experimental regime and should not be identified with the asymptotic exponents in our upper bounds.

Figure 7 also includes the empirical-distribution baseline $W_2(\widehat{\mu}_n, \mu)$, where $\widehat{\mu}_n$ is the empirical measure of the training observations. This baseline provides a reference for the discrepancy between the finite dataset and the target distribution. At larger sample sizes, the learned flow remains above this baseline, indicating that its observed error cannot be explained solely by the discrepancy of the training sample. This separation is consistent with additional errors introduced by representing, learning, and numerically integrating the velocity field, although the experiment does not isolate those contributions individually. The empirical measure is used here as a comparator, not as a universal lower bound on the accuracy of a learned generative model.

Qualitative effect of additional data. Figure 8 illustrates the sample-size dependence for a three-dimensional target through the three pairwise coordinate projections. The rows correspond to $n = 128$, 1024, and 8192, with target, training, and generated samples shown together. The generated samples capture the principal cluster structure even at the smallest sample size, while the larger datasets yield a closer visual match in the concentration and separation of the clusters. The reported Wasserstein errors decrease from 0.078 to 0.057 and 0.050, respectively. These projections provide a geometric interpretation of the quantitative improvement, while agreement in pairwise projections alone does not establish agreement of the full joint distributions.

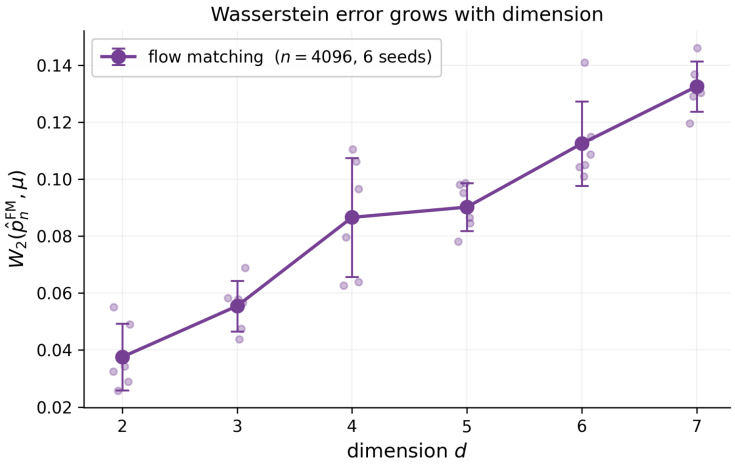

Figure 3: **Dimension scaling.** Wasserstein error as a function of ambient dimension d.



Figure 4: **Gradient-descent iteration scaling.** Wasserstein error as a function of the number of training steps.

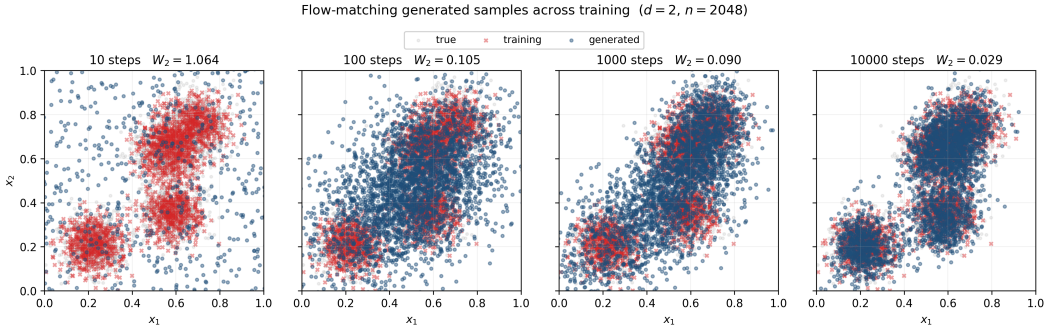

Figure 5: **Evolution of generated samples during training.** Generated samples at increasing numbers of gradient-descent iterations.

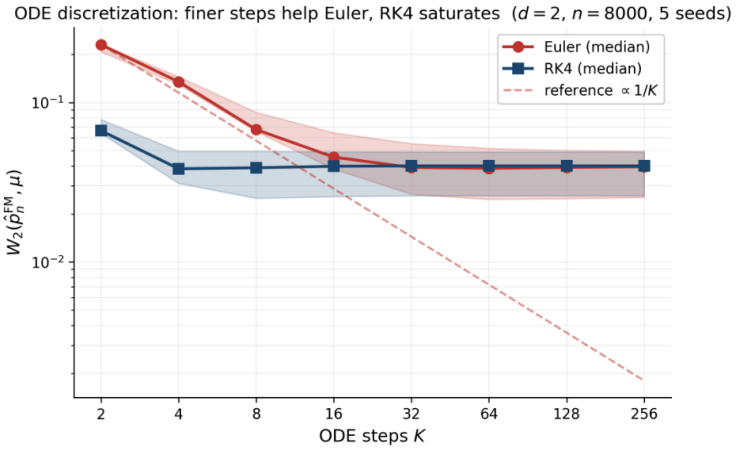

Figure 6: **ODE discretization error.** Wasserstein error as a function of the number of ODE steps. Finer discretization initially reduces error before the total error becomes dominated by other sources.

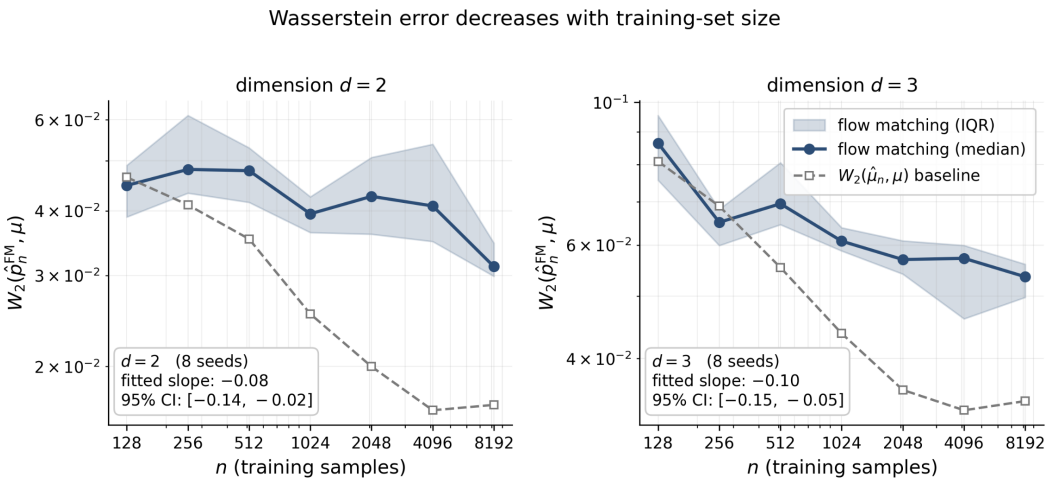

Figure 7: **Sample-size scaling.** Wasserstein error as a function of the number of training samples for $d = 2$ and $d = 3$.

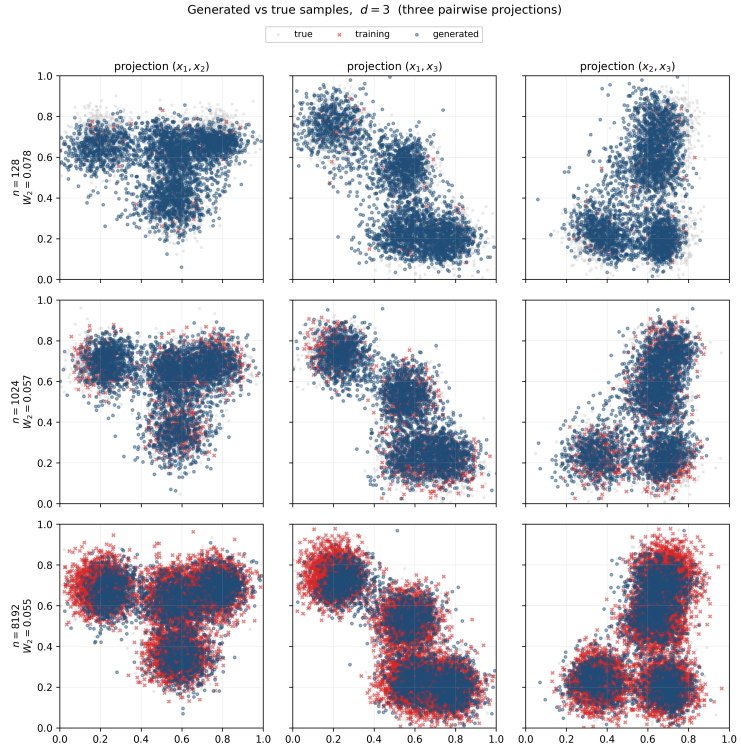

Figure 8: **Generation quality with increasing sample size.** Pairwise projections of target and generated samples in $d = 3$ for increasing training-set sizes n, illustrating improved recovery of the target distribution as n increases.