NON-ASYMPTOTIC BOUNDS FOR BREGMAN SCORE MATCHING WITH MIXED RELU-REQU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Score matching provides a tractable approach to learning unnormalized statistical models, while finite-sample guarantees for deep estimators trained with general Bregman objectives remain limited. We study deep Bregman score matching, which extends the classical quadratic formulation to losses that depend jointly on the learned score field and its input Jacobian. We analyze empirical risk minimization over mixed ReLU–ReQU networks, consisting of a ReLU backbone followed by a shallow terminal ReQU block. For distributions on the full Euclidean space with bounded, Sobolev-smooth score functions, we establish non-asymptotic bounds for excess Bregman risk and population mean-squared score error under explicit regularity and tail conditions. Experiments evaluate score recovery, activation design, and the choice of Bregman divergence. On mixture-distribution benchmarks, the proposed estimator achieves lower score errors than representative kernel- and Stein-based estimators across the tested sample sizes and dimensions. Under the same quadratic score-matching objective and matched parameter counts, the mixed architecture achieves lower score and Jacobian errors than pure ReLU and pure ReQU networks. Further experiments show that different choices of Bregman divergence can improve accuracy for different score-estimation tasks, such as recovering selected coordinates or small-magnitude score components.

1 INTRODUCTION

Score estimation targets the input gradient of a log-density without evaluating its normalizing constant, making it useful when likelihood-based inference is intractable. Applications include sampling with learned unnormalized models (Strathmann et al., 2015), inverse problems (Song et al., 2022), and generative modeling (Song & Ermon, 2019; Ho et al., 2020).

Score matching (Hyvärinen, 2005) yields a training objective that can be evaluated without the unknown density. Its classical quadratic loss treats score errors uniformly across coordinates and magnitudes. Some estimation tasks give priority to selected coordinates, whereas others require accurate recovery of small score components. The Bregman formulation of Gutmann & Hirayama (2011) accommodates these priorities through a convex generating function while retaining the same population target score. Anisotropic quadratics provide fixed directional weights; hypentropy gives weights that decrease with component magnitude. Under finite samples and restricted network capacity, these choices can lead to different allocations of estimation error. Our contribution is the finite-sample analysis of this established criterion over a mixed neural class and the evaluation of these task-dependent choices.

Score estimation has also been studied through denoising (Vincent, 2011; Block et al., 2020), sliced score matching (Song et al., 2020), and kernel or Stein estimators (Sriperumbudur et al., 2017; Sutherland et al., 2018; Li & Turner, 2018; Zhou et al., 2020). Wibisono et al. (2024) obtained minimax-optimal rates up to logarithmic factors for subgaussian distributions with Lipschitz scores using a regularized Gaussian-kernel estimator. Under a Gaussian-noise model with low-dimensional latent structure, Yakovlev et al. (2025) established intrinsic-dimension-dependent rates for implicit and denoising score matching, together with log-density Hessian guarantees. Our analysis treats general Bregman losses over mixed ReLU–ReQU networks without assuming low intrinsic dimension. Deep regression results control neural outputs (Anthony & Bartlett, 1999; Bartlett et al., 2017; Yarotsky, 2017; Petersen & Voigtlaender, 2018; Shen et al., 2020; Lu et al., 2021; Bauer & Kohler,

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.

2019; Schmidt-Hieber, 2020; Jiao et al., 2023); the Jacobian-dependent Bregman loss requires joint control of output and derivative classes.

Differentiable rectified-power networks can approximate functions and their derivatives simultaneously (Ali & Nouy, 2021; Belomestny et al., 2023; Siegel & Xu, 2022; Hon & Yang, 2022). Shen et al. (2023) used fully RePU-activated networks to study deep score matching under the quadratic objective. We use the mixed architecture introduced by Yang et al. (2025), with ReLU layers in the main network body and ReQU layers confined to a shallow terminal block. General parameter choices need not define a globally C^1 function; the analysis uses weak Jacobians and simultaneous approximation of function values and first derivatives.

For a nonquadratic generator, the Hessian multiplying the network Jacobian varies with the predicted score. Both the coefficient and the Jacobian therefore change when the network changes. We control this loss through compatible covers of network outputs and Jacobians, and construct a single approximating network satisfying the width, depth, and global envelope constraints. This yields non-asymptotic score-error bounds for bounded smooth scores on $\mathbb{R}^d$. For fixed quadratic objectives, local smoothness and finite Fisher information also permit unbounded scores and yield rates for finite Gaussian mixtures.

Experiments compare the mixed-network estimator with kernel and Stein methods on mixture distributions. A separate study compares pure ReLU, pure ReQU, and mixed networks under the same quadratic loss and matched parameter counts. Further experiments examine accuracy on selected score coordinates and small-magnitude components using different Bregman-based estimators. A Wasserstein-autoencoder application evaluates latent-score estimation.

The remainder of the paper is organized as follows. Section 2 introduces the notation, score-matching objective, and network estimator. Section 3 states the assumptions and presents the mean-squared score bound and convergence rate. Section 4 reports the experiments, and Section 5 discusses their scope and limitations. The appendices contain the proofs and additional experimental details.

2 Problem setup

2.1 Notation

Let $\mathcal{D} = (X_1, \dots, X_n)$ be independent observations with common density p on $\mathbb{R}^d$, and let $X \sim p$ be an independent draw. Expectations over X and $\mathcal{D}$ are denoted by $\mathbb{E}$ and $\mathbb{E}_{\mathcal{D}}$, respectively. We write s^* for the true score, f for a candidate network, and $\widehat{s}$ for the fitted score.

For vectors, $\langle u, v\rangle = u^\top v$, $\|u\|_2^2 = \sum_j u_j^2$, and $\|u\|_\infty = \max_j |u_j|$. For matrices, $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$ and $\|A\|_{\mathrm{op}} = \sup_{\|u\|_2=1} \|Au\|_2$. The symbols I_d, Tr, and $\preceq$ denote the identity matrix, trace, and positive-semidefinite order. For a vector field $f : \mathbb{R}^d \to \mathbb{R}^d$, the weak input Jacobian is $J_f = (\partial_j f_i)_{i,j=1}^d$, and

$$\|f\|_\infty = \operatorname*{ess\,sup}_{x\in\mathbb{R}^d} \|f(x)\|_\infty, \qquad \|f\|_{L^2(p)}^2 = \int_{\mathbb{R}^d} \|f(x)\|_2^2 \, p(x)\, dx, \tag{1}$$

$$\|J_f\|_{F,\infty} = \operatorname*{ess\,sup}_{x\in\mathbb{R}^d} \|J_f(x)\|_F, \qquad \|f\|_{W^{m,\infty}} = \max_{|\alpha|\le m} \|D^\alpha f\|_\infty. \tag{2}$$

Here $\alpha \in \mathbb{N}_0^d$ is a multi-index with $|\alpha| = \sum_j \alpha_j$, and D^α is the componentwise weak derivative. The space $W^{m,\infty}(\mathbb{R}^d;\mathbb{R}^d)$ consists of vector fields with finite norm in (2). The local space $W^{m,\infty}_{\mathrm{loc}}(\mathbb{R}^d;\mathbb{R}^d)$ requires this norm to be finite on every bounded cube, with the essential suprema restricted to that cube. For nonnegative quantities, $a \lesssim b$ means $a \le cb$ for a positive constant c; $a \gtrsim b$ means $b \lesssim a$, and $a \asymp b$ means both. The relation $\lesssim$ hides no logarithmic factors; $\widetilde{O}$ suppresses logarithmic factors in n. We use calligraphic $\mathcal{L}$ for loss functionals and plain L for network depth.

2.2 Score and Bregman loss

The target of estimation is

$$s^*(x) = \nabla \log p(x), \qquad x \in \mathbb{R}^d. \tag{3}$$

The convex function defining a Bregman divergence is called its generator. For a twice differentiable generator ϕ, the divergence is

$$D_\phi(u, v) = \phi(u) - \phi(v) - \langle \nabla\phi(v), u - v \rangle . \tag{4}$$

Following Gutmann & Hirayama (2011, Section 3.4), expand the population discrepancy as

$$\begin{aligned}\mathbb{E}D_\phi(s^*(X), f(X)) = \mathbb{E}\phi(s^*(X)) + \mathbb{E}[-\phi(f(X)) + \langle \nabla\phi(f(X)), f(X)\rangle] \\ - \mathbb{E}\langle \nabla\phi(f(X)), s^*(X)\rangle .\end{aligned}$$

Under the density, score, and generator conditions stated below, integration by parts replaces the last inner-product expectation by $-\mathbb{E}\,\mathrm{Tr}(\nabla^2\phi(f(X))J_f(X))$. The resulting sample-observable loss is

$$\ell_\phi(f; x) = -\phi(f(x)) + \langle \nabla\phi(f(x)), f(x)\rangle + \mathrm{Tr}\big(\nabla^2\phi(f(x))J_f(x)\big) . \tag{5}$$

The population and empirical losses are therefore

$$\begin{aligned}\mathcal{L}_\phi(f) &:= \mathbb{E}\ell_\phi(f; X) = \mathbb{E}D_\phi(s^*(X), f(X)) - \mathbb{E}\phi(s^*(X)), \\ \widehat{\mathcal{L}}_\phi(f) &:= \frac{1}{n}\sum_{i=1}^{n} \ell_\phi(f; X_i).\end{aligned} \tag{6}$$

The term $\mathbb{E}\phi(s^*(X))$ is independent of f. For $\phi(u) = \|u\|_2^2/2$, the classical population score-matching loss is $\mathcal{L}_\phi(f) = \mathbb{E}[\|f(X)\|_2^2/2 + \mathrm{div}\, f(X)]$ (Hyvärinen, 2005).

Choice of Bregman divergence. The generator specifies how the discrepancy responds to score errors. Differentiating (4) with respect to its second argument gives

$$\nabla_v D_\phi(u, v) = \nabla^2\phi(v)(v - u). \tag{7}$$

Thus the Hessian at the candidate score weights the error response. The same Hessian enters the observable Jacobian term in (5), so this choice is part of training. Under the integration-by-parts conditions below, all strictly convex choices retain s^* as the unrestricted population target.

Table 1 gives the three choices studied here. Anisotropic quadratics induce Mahalanobis-type discrepancies (Banerjee et al., 2005), with a positive-definite matrix specifying coordinate or directional priorities. The hypentropy generator of Ghai et al. (2020) is defined on signed vectors by

$$\phi_\beta(u) = \sum_{j=1}^{d} h_\beta(u_j), \qquad h_\beta(t) = t\,\mathrm{arsinh}(t/\beta) - \sqrt{t^2 + \beta^2}, \qquad \beta > 0. \tag{8}$$

Its curvature decreases with component magnitude. Small components therefore receive greater relative weight than large ones; β controls the scale of this variation. These curvature choices specify estimation priorities, without guaranteeing smaller Euclidean error in every coordinate or magnitude range.

Table 1: Three Bregman generators used in the experiments. The Hessian specifies the local error response in (7). Here $A = A^\top \succ 0$ and $\beta > 0$; h_β is defined in (8).

Generator and potential	Hessian $\nabla^2\phi(u)$	Estimation priority
Quadratic $\phi(u) = \frac{1}{2}\|u\|_2^2$	I_d	Uniform coordinate weighting.
Anisotropic quadratic / Mahalanobis $\phi_A(u) = \frac{1}{2}u^\top Au$	A	Fixed weights on selected coordinates or directions.
Hypentropy $\phi_\beta(u) = \sum_{j=1}^{d} h_\beta(u_j)$	Diagonal, with entries $(u_j^2 + \beta^2)^{-1/2}$	Greater relative weight for small-magnitude components.

2.3 Network class and estimator

The loss in (5) involves both network values and input derivatives, so the approximation and complexity analysis must control them jointly. RePU networks provide simultaneous approximation of smooth functions and their derivatives (Shen et al., 2023; Belomestny et al., 2023). Following Yang et al. (2025), we retain a ReLU backbone and use a shallow terminal ReQU block for polynomial and product operations. The resulting ordered architecture is defined below.

Definition 1 (Bounded mixed-network class)**.** *The depth of a network is the number of its hidden layers, and its width is the maximum total number of neurons in any hidden layer. For $W \geq d$ and $L \geq 2$, let $\mathcal{N}^{\mathrm{mix}}_{L,W}$ denote the class of vector-valued networks $f : \mathbb{R}^d \to \mathbb{R}^d$ with depth at most L and width at most W.*

Each network consists of a ReLU block f_1 with L_1 hidden layers, followed by a ReQU block f_2 with L_2 hidden layers and a d-dimensional affine output. The blocks have compatible dimensions and satisfy

$$f = f_2 \circ f_1, \qquad L_1 + L_2 \leq L, \qquad L_2 \leq c_{\mathrm{term}} \log(eL), \tag{9}$$

where $c_{\mathrm{term}} > 0$ is a fixed architectural constant. The activation functions

$$\mathrm{ReLU}(z) = \max\{z, 0\}, \qquad \mathrm{ReQU}(z) = \max\{z, 0\}^2$$

act componentwise, with all ReLU hidden layers preceding all ReQU hidden layers.

For deterministic capacities $W_n \geq d$, $L_n \geq 2$ and output and Jacobian bounds $M_n, B_{J,n} > 0$, define

$$\mathcal{F}_n = \left\{ f \in \mathcal{N}^{\mathrm{mix}}_{L_n,W_n} : \|f\|_\infty \leq M_n, \quad \|J_f\|_{F,\infty} \leq B_{J,n} \right\}. \tag{10}$$

The empirical loss uses a fixed, bounded, sample-independent measurable representative of each weak Jacobian. To isolate statistical estimation error, we assume that empirical minimization is attained and select a measurable estimator

$$\widehat{s} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \widehat{\mathcal{L}}_\phi(f). \tag{11}$$

Optimization error is set to zero throughout the analysis under this exact-ERM idealization; convergence of the numerical training algorithm is not studied.

3 Main theoretical results

We study the mean-squared population error $\mathbb{E}_{\mathcal{D}} \|\widehat{s} - s^*\|^2_{L^2(p)}$ in two settings.

3.1 Bregman score estimation with bounded smooth scores

Throughout this subsection, the output and Jacobian envelopes in (10) are fixed independently of n, with $M_n = M$ and $B_{J,n} = B_J$.

Assumption 1 (Density positivity and differentiability)**.**

$$p \in C^1(\mathbb{R}^d), \qquad p(x) > 0 \quad \textit{for every } x \in \mathbb{R}^d.$$

Assumption 2 (Score regularity)**.** *For a fixed integer $m \geq 2$ and a constant $H_0 > 0$,*

$$s^* \in W^{m,\infty}(\mathbb{R}^d; \mathbb{R}^d), \qquad \|s^*\|_{W^{m,\infty}} \leq H_0.$$

Remark 1. Under the norm convention in (2), Assumption 2 requires $\|D^\alpha s^*\|_\infty \leq H_0$ for every multi-index with $|\alpha| \leq m$. Thus each score coordinate is uniformly bounded, as are all pure and mixed weak partial derivatives through order m. In particular, $\|s^*\|_\infty \leq H_0$ and $\|J_{s^*}\|_{F,\infty} \leq dH_0$. Classical continuity of the derivatives of order m is not required. The smoothness order m and bound H_0 are fixed independently of n.

Assumption 3 (Bregman generator)**.** *The function ϕ is twice continuously differentiable on an open neighborhood of $[-M, M]^d$. There are constants $0 < \lambda \leq \Lambda < \infty$ and $L_\phi < \infty$ such that, for all $u, v \in [-M, M]^d$,*

$$\lambda I_d \preceq \nabla^2\phi(u) \preceq \Lambda I_d, \qquad \left\|\nabla^2\phi(u) - \nabla^2\phi(v)\right\|_{\mathrm{op}} \leq L_\phi \|u - v\|_2. \tag{12}$$

Remark 2. The cube in (12) is the range of score values. Strong convexity relates Bregman risk to squared score error; the remaining conditions control the derivative-dependent loss.

Let U and V be arbitrary positive integers. For a sample of size n, consider the mixed-network class $\mathcal{F}_n$ with width at most W_n and at most L_n hidden layers. A concrete choice of these bounds is

$$W_n = 2^{d+6} d\, m^{d+1} (U+d) \lceil \log_2(8U) \rceil,$$
$$L_n = 9(V+1) \lceil \log_2(4V) \rceil + 1.$$

The following bound applies when the width and depth bounds satisfy (13).

The terminal-layer coefficient in Definition 1 is fixed at a sufficiently large value depending only on d, m. The output and Jacobian envelopes are fixed with $M > H_0$ and $B_J > dH_0$. The comparison network approximates the score on $[-\varrho, \varrho]^d$, leaving a tail contribution outside this region. The radius ϱ is an analysis parameter and does not modify the estimator.

Theorem 1 (Mean-squared score error bound)**.** *Under Assumptions 1–3, there is a constant $c_{\mathrm{cap}} > 0$, depending only on d, m, H_0, M, B_J. When $n \geq 2$, $\varrho \geq 1$, and the width and depth bounds satisfy*

$$\min\left\{ \frac{W_n}{\log(eW_n)}, \frac{L_n}{\log(eL_n)} \right\} \geq c_{\mathrm{cap}} \varrho^{d/4}, \tag{13}$$

every estimator $\widehat{s} \in \mathcal{F}_n$ satisfying (11) *obeys*

$$\mathbb{E}_{\mathcal{D}} \left\| \widehat{s} - s^* \right\|_{L^2(p)}^2 \lesssim W_n L_n \sqrt{\frac{\log(eW_n L_n)}{n}} + \varrho^{2m} \left(\frac{W_n L_n}{\log(eW_n)\log(eL_n)} \right)^{-4m/d} + \mathbb{P}(\|X\|_\infty > \varrho). \tag{14}$$

The hidden constant depends only on $d, m, H_0, M, B_J, \lambda, \Lambda, L_\phi$, and is independent of n, ϱ, W_n, L_n and p within the stated class.

The sampling term follows from a Rademacher bound for the derivative-dependent loss class. Width and depth enter through the joint complexity of network outputs and Jacobians; the loss-class multiplier depends only on $d, M, B_J, \Lambda, L_\phi$, while strong convexity contributes the separate factor $2/\lambda$.

In (14), increasing width and depth reduces the approximation term while increasing the sampling term. The tail term accounts for the part of the distribution outside the approximation region. First-derivative approximation is used to keep the comparison network within the Jacobian envelope; the squared output error determines the approximation term in the population bound.

Assumption 4 (Moment generating function)**.** *The moment generating function $t \mapsto \mathbb{E} e^{\langle t, X \rangle}$ is uniformly bounded on a fixed neighborhood of the origin in $\mathbb{R}^d$. The neighborhood and the bound do not depend on n.*

Remark 3. A uniform result requires a common neighborhood and bound across the density family; see Appendix B.4 for an example.

A logarithmically growing radius controls the tail term in (14). Balancing the sampling and approximation terms gives the following network-size choice.

Corollary 1 (Convergence rate)**.** *Suppose Assumptions 1–4 hold. Choose*

$$W_n = L_n = \left\lceil n^{\frac{d}{4d+16m}} \right\rceil. \tag{15}$$

There exists n_0 such that, for $n \geq n_0$, every estimator satisfying (11) *obeys*

$$\mathbb{E}_{\mathcal{D}} \left\| \widehat{s} - s^* \right\|_{L^2(p)}^2 \lesssim n^{-\frac{2m}{d+4m}} [\log(en)]^{2m+8m/d}. \tag{16}$$

The threshold n_0 and hidden constant depend only on $d, m, H_0, M, B_J, \lambda, \Lambda, L_\phi$, and the neighborhood and bound in Assumption 4.

Corollary 2 (Excess Bregman risk bound)**.** *Under the assumptions and network-size choice of Corollary 1, for all $n \geq n_0$,*

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}_\phi(\widehat{s}) - \mathcal{L}_\phi(s^*)] = \mathbb{E}_{\mathcal{D}} \mathbb{E} D_\phi(s^*(X), \widehat{s}(X)) \lesssim n^{-\frac{2m}{d+4m}} [\log(en)]^{2m+8m/d}. \tag{17}$$

The hidden constant depends only on $d, m, H_0, M, B_J, \Lambda, L_\phi$ and the neighborhood and bound in Assumption 4.

Remark 4. Corollary 1 has the same polynomial rate $n^{-2m/(d+4m)}$ as Shen et al. (2023, Theorem 13) for quadratic score matching with RePU networks and C^m-smooth scores. Here the result covers mixed networks and general Bregman losses; the assumptions and logarithmic factors differ. For fixed parameters, the generators in Table 1 share this exponent. Their constants involve λ through risk conversion and Λ, L_ϕ through empirical-loss control, without ranking Euclidean errors.

3.2 QUADRATIC SCORE MATCHING BEYOND BOUNDED SCORES

For quadratic generators, the population identity of Hyvärinen (2005) extends to unbounded scores with finite Fisher information under the bounded-candidate conditions used here. To obtain an explicit approximation bound, we replace Assumption 2 by the following condition.

Assumption 5 (Finite Fisher information and local score regularity)**.** *For a fixed integer $m \geq 2$, the true score satisfies $s^* \in W^{m,\infty}_{\mathrm{loc}}(\mathbb{R}^d;\mathbb{R}^d)$ and*

$$I(p) := \mathbb{E}\,\|s^*(X)\|_2^2 < \infty. \tag{18}$$

For $\varrho \geq 1$, write $H_\varrho = \max\{1, \|s^*\|_{W^{m,\infty}([-\varrho,\varrho]^d)}\}$. This local bound may grow with ϱ. We use the same class $\mathcal{F}_n$ and estimator $\widehat{s}$ from (10)–(11), now with $M_n, B_{J,n} \geq 1$ allowed to grow with n and $\phi = \phi_A$.

Theorem 2 (Score bound under finite Fisher information and local smoothness)**.** *Under Assumptions 1 and 5, fix $A = A^\top \succ 0$ and $\phi_A(u) = u^\top Au/2$. There is a constant $c_{\mathrm{size}} > 0$, depending only on d, m. When $n \geq 2$, $\varrho \geq 1$, $W_n \geq W_0$, $L_n \geq L_0$, and*

$$M_n \geq 2H_\varrho, \qquad B_{J,n} \geq 2dH_\varrho, \qquad \frac{W_n L_n}{\log(eW_n)\log(eL_n)} \geq c_{\mathrm{size}}\varrho^{d/2}, \tag{19}$$

every estimator $\widehat{s} \in \mathcal{F}_n$ obeys

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\,\|\widehat{s} - s^*\|^2_{L^2(p)} \lesssim\ & (M_n^2 + B_{J,n})W_n L_n \sqrt{\frac{\log(eW_nL_n)}{n}} \\ & + H_\varrho^2 \varrho^{2m}\left(\frac{W_nL_n}{\log(eW_n)\log(eL_n)}\right)^{-4m/d} \\ & + \mathbb{E}\Big[(M_n^2 + \|s^*(X)\|_2^2)\mathbf{1}\{\|X\|_\infty > \varrho\}\Big].\end{aligned} \tag{20}$$

Here $W_0, L_0 > 0$ are fixed architecture thresholds depending only on d, m; the hidden constant depends only on d, m, A. The dependence on the local score size, envelopes, and tail is explicit.

Finite Fisher information alone gives no quantitative control of H_ϱ or this tail term. For Gaussian mixtures, their growth and decay yield an explicit rate.

Corollary 3 (Rate for finite Gaussian mixtures)**.** *Let p be a fixed finite mixture of nonsingular Gaussian densities and let $m \geq 2$ be fixed. For the quadratic estimator above, choose*

$$W_n = L_n = \left\lceil n^{\frac{d}{4d+16m}} \right\rceil, \qquad M_n \asymp B_{J,n} \asymp [\log(en)]^{(m+1)/2}, \tag{21}$$

with sufficiently large fixed envelope multipliers. For all sufficiently large n,

$$\mathbb{E}_{\mathcal{D}}\,\|\widehat{s} - s^*\|^2_{L^2(p)} \lesssim n^{-\frac{2m}{d+4m}}[\log(en)]^{\,2m+1+8m/d}. \tag{22}$$

The hidden constant and the threshold on n depend only on d, m, A and the fixed mixture parameters. Thus the error is $\widetilde{O}(n^{-2m/(d+4m)})$.

Remark 5 (Gaussian and heavy-tailed examples). For a single $N(\mu, \Sigma)$, the score permits $W_n = 4d$, $L_n = 2$, a sufficiently large $M_n \asymp \sqrt{\log(en)}$, and fixed $B_{J,n} \geq \max\{1, \|\Sigma^{-1}\|_F\}$, giving the sharper order $\log(en)/\sqrt{n}$ through a clipped affine comparator. Student t scores have finite energy and local smoothness even when the input second moment is infinite.

4 EXPERIMENTS

We examine the coordinate and magnitude priorities of the three generators in Table 1. Quadratic-loss experiments compare kernel and Stein estimators, activation architectures, and a Wasserstein-autoencoder application. Appendix G.5 reports an additional network-capacity study at fixed sample size. For an independent test set, global score error is measured by

$$\mathrm{NormL2} = \frac{1}{d\,n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}} \|\widehat{s}(x_i) - s^*(x_i)\|_2^2. \tag{23}$$

The coordinate- and magnitude-specific errors supplement this global metric. Training parameterizations and validation criteria are stated for each study.

4.1 TASK-DEPENDENT SCORE ESTIMATION

This diagnostic considers recovery of selected coordinates and small-magnitude score components. All three generators use the same cosine density on $[0,1]^8$,

$$p(x) \propto \exp\left\{\sum_{j=1}^{8} \kappa_j \cos(2\pi m_j x_j)\right\}, \qquad (m_1,\dots,m_8) = (3,3,1,1,1,1,1,1), \qquad \kappa_j = \frac{0.5}{m_j}. \tag{24}$$

Its score is $s_j^*(x) = -\pi \sin(2\pi m_j x_j)$. The first two coordinates have higher spatial frequency but the same amplitude bound as the others. The coordinate task therefore prioritizes more rapidly varying components without increasing their target amplitude. Each of five paired runs uses 500 training, $1{,}000$ validation, and $5{,}000$ test observations. The shared backbone has affine dimensions $8 \to 128 \to 128 \to 128 \to 8$ and activation sequence ReLU–ReLU–ReQU. Model selection uses the observable validation score-matching loss.

Coordinate priorities. The anisotropic quadratic specializes Table 1 to

$$\phi_{A_a}(u) = \tfrac{1}{2} u^\top A_a u, \qquad A_a = \mathrm{diag}(a,a,1,1,1,1,1,1), \tag{25}$$

with $a \in \{4,8,16,32\}$ and the quadratic baseline corresponding to $a = 1$. Table 2 puts the prioritized-coordinate error first and retains the other-coordinate and global errors. Increasing a from 4 to 16 lowers the mean error on coordinates 1–2 from 5.1213 to 4.9984, below the baseline 5.1593. At $a = 16$, the other-coordinate error is higher than the baseline, whereas moderate weighting at $a = 4$ gives the lowest global error in the sweep. The $a = 32$ stress test further lowers prioritized-coordinate error but raises the other-coordinate and global errors relative to the baseline. These are descriptive trends over the reported settings.

Table 2: Coordinate and global errors for the three generator choices under dual parameterization. Entries are means ± sample standard deviations over five paired runs. The first metric measures the prioritized coordinates; lower is better throughout.

Generator	$\mathrm{MSE}_{1:2}$	$\mathrm{MSE}_{3:8}$	NormL2
Quadratic	5.1593 ± 0.1112	0.5906 ± 0.0723	1.7328 ± 0.0785
Anisotropic, $a = 4$	5.1213 ± 0.1206	$\mathbf{0.5224 \pm 0.0989}$	$\mathbf{1.6721 \pm 0.0979}$
Anisotropic, $a = 8$	5.0526 ± 0.0659	0.5867 ± 0.1494	1.7032 ± 0.1221
Anisotropic, $a = 16$	4.9984 ± 0.0491	0.6277 ± 0.1595	1.7204 ± 0.1256
Anisotropic, $a = 32$	$\mathbf{4.9718 \pm 0.0515}$	0.6606 ± 0.1663	1.7384 ± 0.1297
Hypentropy, $\beta = 1$	5.1701 ± 0.1391	0.7965 ± 0.1659	1.8899 ± 0.1426

Score-magnitude priorities. For hypentropy in (8), we use $\beta = 1$. Its curvature decreases with the predicted component magnitude. Evaluation bins instead use the 50th and 90th percentiles of pooled *true* score-component magnitudes; these bins are used for reporting errors, not choosing the training weights. Table 3 shows lower mean error for hypentropy in the low-magnitude bin, together with higher errors in the other bins and global NormL2. This is the small-component emphasis motivated in Section 2.2, with an observed cost in other error measures. Neither experiment establishes a uniformly superior generator.

Table 3: Score-component MSE by true-score magnitude, together with global NormL2. Low, medium, and high bins use the 50th and 90th percentiles of pooled sample–coordinate magnitudes. Entries are means ± sample standard deviations over five runs; lower is better.

Generator	Low	Medium	High	NormL2
Quadratic	0.749 ± 0.035	$\mathbf{2.567 \pm 0.120}$	$\mathbf{3.314 \pm 0.159}$	$\mathbf{1.7328 \pm 0.0785}$
Hypentropy, $\beta = 1$	$\mathbf{0.686 \pm 0.044}$	2.917 ± 0.232	3.804 ± 0.335	1.8899 ± 0.1426

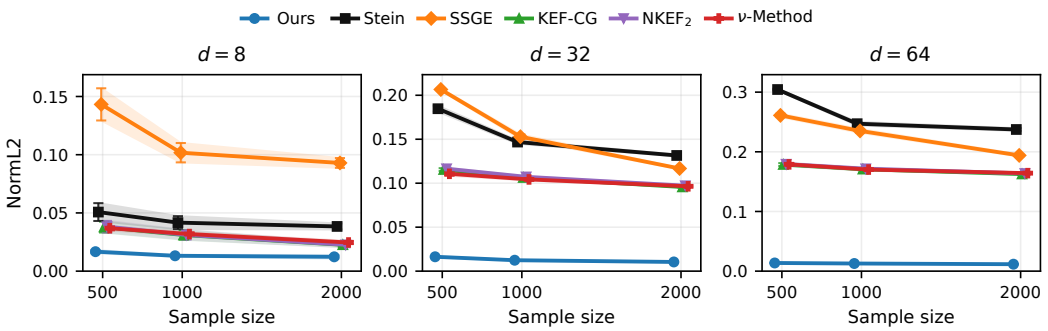

Figure 1: Normalized score-estimation error on the grid-mixture benchmark. "Ours" denotes the mixed ReLU–ReQU estimator. Numerical means and standard deviations are reported in Table 6.

4.2 Score recovery against kernel and Stein estimators

The nonparametric comparators are the Stein gradient estimator (Stein) (Li & Turner, 2018), the spectral Stein gradient estimator (SSGE) (Shi et al., 2018), the Nyström kernel exponential-family estimator NKEF$_2$ (Sutherland et al., 2018), and the KEF-CG and iterative curl-free ν-Method estimators (Zhou et al., 2020).

We use the grid Gaussian-mixture benchmark of Sutherland et al. (2018); Zhou et al. (2020), with $d \in \{8, 32, 64\}$ and $n \in \{500, 1000, 2000\}$. The mixed network is trained using the quadratic objective. Score error is evaluated by (23). Each configuration uses $1{,}024$ test observations and ten repeated trials. Figure 1 shows the mean errors over repeated trials. The mixed estimator has the lowest reported mean score error in every tested configuration.

4.3 Matched activation ablation

We compare pure ReLU, pure ReQU, and mixed ReLU–ReQU networks under the same quadratic score-matching objective. The target is an eight-dimensional correlated Gaussian with mean coordinates equally spaced from -1 to 1 and covariance $\Sigma_{ij} = 0.4^{|i-j|}$. Its score and Jacobian are available analytically. Each of five paired runs uses $5{,}000$ training, $1{,}000$ validation, and $2{,}000$ test observations.

All models have three hidden layers of width 128 and $35{,}208$ parameters. The mixed model uses two ReLU hidden layers followed by one ReQU hidden layer. Learning rates are selected separately for the three architectures. Epoch selection uses the true score on the independent validation split, so this is an oracle-assisted architecture comparison.

Metrics. All metrics are computed in the training-standardized coordinates. Let z_i denote the standardized test observations, and let s_z^* and $J_{s_z^*}$ denote the corresponding analytic score and its constant Jacobian. We measure score error by $\mathrm{NormL2} = \frac{1}{d\, n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \|\widehat{s}(z_i) - s_z^*(z_i)\|_2^2$, directional agreement by $\mathrm{Cosine} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \frac{\langle \widehat{s}(z_i), s_z^*(z_i) \rangle}{\|\widehat{s}(z_i)\|_2 \|s_z^*(z_i)\|_2}$, and relative Jacobian error on 256 test observations by $\text{Jacobian-RelF} = \frac{\left(\frac{1}{256} \sum_{i=1}^{256} \|J_{\widehat{s}}(z_i) - J_{s_z^*}\|_F^2 \right)^{1/2}}{\|J_{s_z^*}\|_F}$. Lower NormL2 and Jacobian-RelF indicate smaller estimation errors, whereas higher cosine similarity indicates better directional agreement. Table 4 reports lower score and Jacobian errors and higher cosine similarity for the mixed architecture than for either homogeneous architecture.

4.4 Latent-score estimation in Wasserstein autoencoders

Background. WAE (Tolstikhin et al., 2018) learn latent-variable generative models by minimizing a reconstruction loss together with a penalty that matches the aggregated posterior $q_\eta(z) = \int q_\eta(z \mid x) p_X(x)\, dx$ to a prior $p(z)$. With a deterministic decoder G, the objective takes the form $\mathbb{E}_{p_X(x)} \mathbb{E}_{q_\eta(z|x)} \|x - G(z)\|_2^2 + \lambda_{\mathrm{WAE}}\, \mathrm{D}_{\mathrm{WAE}}(q_\eta(z), p(z))$. When $\mathrm{D}_{\mathrm{WAE}}$ is chosen as the KL diver-

Table 4: Matched activation ablation on the correlated Gaussian. Entries are means ± sample standard deviations over five runs. Lower NormL2 and Jacobian-RelF are better; higher cosine similarity is better.

Architecture	Learning rate	NormL2	Cosine	Jacobian-RelF
Pure ReLU	3×10^{-3}	0.8139 ± 0.0204	0.8419 ± 0.0133	0.8093 ± 0.0125
Pure ReQU	3×10^{-3}	0.4258 ± 0.1187	0.9676 ± 0.0037	0.9017 ± 0.1341
Mixed	3×10^{-4}	$\mathbf{0.1197 \pm 0.0124}$	$\mathbf{0.9800 \pm 0.0039}$	$\mathbf{0.4198 \pm 0.0306}$

Table 5: AIS-based NLL estimates on MNIST (mean ± std). "Ours" denotes the mixed ReLU–ReQU estimator. Lower is better.

Method	$d_z = 8$	$d_z = 32$	$d_z = 64$	$d_z = 128$
Stein	130.49 ± 3.40	130.52 ± 0.93	134.66 ± 1.35	150.86 ± 3.84
KEF-CG	126.06 ± 3.16	$\mathbf{126.72 \pm 1.21}$	128.76 ± 2.47	127.59 ± 1.13
ν-Method	128.57 ± 1.32	131.26 ± 0.90	133.33 ± 2.22	126.43 ± 0.79
SSGE	122.44 ± 0.32	127.44 ± 0.61	130.12 ± 0.35	139.77 ± 0.72
NKEF_2	128.88 ± 0.42	128.58 ± 0.84	127.11 ± 0.83	$\mathbf{125.43 \pm 0.55}$
Ours	$\mathbf{117.85 \pm 0.53}$	$\mathbf{126.72 \pm 0.32}$	$\mathbf{126.70 \pm 0.29}$	132.44 ± 0.93

gence, the entropy-gradient term of $q_\eta(z)$ is generally intractable and can be approximated using a score estimator for $\nabla_z \log q_\eta(z)$ (Song et al., 2020). Thus, WAE training provides a downstream empirical setting to examine whether learned latent-space score estimates can serve as a useful regularization component in generative modeling.

Setup and metric. We train WAEs on MNIST with $d_z \in \{8, 32, 64, 128\}$, comparing the mixed estimator with the same nonparametric baselines defined above. Each configuration is repeated over ten independent trials, and we report mean and standard deviation. The latent score estimator uses the quadratic score-matching objective. Generative performance is evaluated by Annealed Importance Sampling (AIS)-based test negative log-likelihood (NLL) (Neal, 2001), where lower is better.

Results. Table 5 reports AIS-based NLL estimates on MNIST. The mixed estimator attains the lowest reported mean at $d_z = 8$ and $d_z = 64$, and has the same displayed mean as KEF-CG at $d_z = 32$. At $d_z = 128$, NKEF_2 has the lowest mean. The small difference at $d_z = 64$ is not a claim of statistical significance. The mixed-network implementation uses latent-buffer and score-refitting stabilizers; the comparison concerns the full implemented methods and does not isolate the effect of each stabilizer.

5 Conclusion and Limitations

We derived mean-squared score bounds for mixed ReLU–ReQU estimators, separating sampling, approximation, and tail errors. Anisotropic quadratic and hypentropy objectives favor selected coordinates and small components in the reported diagnostics, with trade-offs in other metrics. Quadratic grid-mixture and activation comparisons separately assess the estimator and architecture; WAE provides a latent-space application.

Jacobian errors are evaluated empirically; the theory does not give a Jacobian or Hessian estimation rate. Exact trace evaluation remains part of the theoretical criterion. There is no minimax lower bound or sampling theorem. The practical experiments do not certify the global function-class constraints, and the original grid-baseline selection record and some training details require further documentation.

AI use statement

In this work, we used generative AI tools to assist in the formulation of some mathematical claims and in proof development, including suggesting intermediate arguments and bounds used in establishing the stochastic error bound in our theoretical analysis. All such AI-assisted mathematical content was

independently reconstructed, checked, and substantially revised by the authors before inclusion in the manuscript.

We did not use generative AI tools to generate synthetic datasets, develop the theoretical model or conceptual framework of the paper, design the experimental methodology, implement the proposed methods, or interpret the experimental results. Other required disclosure categories are not applicable to this work.

We additionally used generative AI tools for suggesting aspects of the paper structure, title, and keywords; editing the manuscript for readability; and assisting with literature search and reference checking.

All AI-assisted work was reviewed and verified by the authors. In particular, mathematical arguments and bounds were manually checked for correctness, references and bibliographic information were verified against the original sources, and AI-edited text was reviewed to ensure that it accurately reflects the authors' intended claims. The authors take full responsibility for the final content of the paper, including all mathematical results, experimental claims, text, and references.

Reproducibility statement

The estimator is specified in Section 2.3, and the assumptions are stated in Section 3. Proofs and network constructions are provided in Appendices B–E, with constant dependencies stated alongside the bounds. Appendix F proves the extension under finite Fisher information and verifies its distributional examples. Appendix G describes the data-generating distributions, evaluation metrics, and reported training and selection procedures. The reporting scope, oracle-assisted selection, and unavailable implementation details are identified there.

References

Mazen Ali and Anthony Nouy. Approximation of smoothness classes by deep rectifier networks. *SIAM Journal on Numerical Analysis*, 59(6):3032–3051, 2021.

Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Denis Belomestny, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161:242–253, 2023.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, 2011.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 386–407. PMLR, 2020.

Michael U Gutmann and Jun-ichiro Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 283–290. AUAI Press, 2011.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Sean Hon and Haizhao Yang. Simultaneous neural network approximation for smooth functions. *Neural Networks*, 154:152–164, 2022.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Differentiable neural networks with RePU activation: With applications to score estimation and isotonic regression. *arXiv preprint arXiv:2305.00608*, 2023.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4644–4653. PMLR, 2018.

Jonathan W Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 574–584. PMLR, 2020.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Danica J. Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with Nyström kernel exponential families. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 652–660. PMLR, 2018.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical Bayes smoothing. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4958–4991. PMLR, 2024.

Konstantin Yakovlev, Anna Markovich, and Nikita Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density Hessian estimation. *arXiv preprint arXiv:2512.24378*, 2025.

Yahong Yang, Yue Wu, Haizhao Yang, and Yang Xiang. Nearly optimal approximation rates for deep super ReLU networks on Sobolev spaces. *arXiv preprint arXiv:2310.10766v5*, 2025.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11513–11522. PMLR, 2020.

A PROOF NOTATION AND APPENDIX ORGANIZATION

For scalar integrands write $Pg = \mathbb{E}g(X)$ and $P_n g = n^{-1}\sum_i g(X_i)$. The excess Bregman risk and the two quantities in the ERM decomposition are

$$\mathcal{E}_\phi(f) = \mathcal{L}_\phi(f) - \mathcal{L}_\phi(s^*). \tag{A.1}$$

$$\mathfrak{S}_n = \mathbb{E}_{\mathcal{D}} \sup_{f\in\mathcal{F}_n} |\widehat{\mathcal{L}}_\phi(f) - \mathcal{L}_\phi(f)|, \qquad \mathfrak{A}_n = \inf_{f\in\mathcal{F}_n} \mathcal{E}_\phi(f). \tag{A.2}$$

The infimum defining $\mathfrak{A}_n$ need not be attained. The norm conventions in (1)–(2) apply on restricted domains as well. Suprema are interpreted through outer expectations and probabilities when required for measurability. The independent sample is specified in Section 2.1.

Appendix B establishes the population score-matching identity. Appendix C gives the output and derivative complexity bounds and controls the stochastic error. Appendix D constructs the comparison networks and verifies their global bounds. Appendix E proves Theorem 1 and Corollaries 1 and 2. Appendix F proves the finite-Fisher-information results, including the Gaussian-mixture rate and distributional examples. Appendix G contains the experimental settings, metrics, and additional results.

B POPULATION CALCULUS ON THE FULL SPACE

Throughout Appendices B–E, the class $\mathcal{F}_n$ has fixed envelopes $M_n = M$ and $B_{J,n} = B_J$, as in Section 3.1.

B.1 BASIC REGULARITY AND AFFINE NORMALIZATION

Lemma B.1 (Regularity supplied by the envelopes)**.** *Every $f \in \mathcal{F}_n$ has a continuous representative in $W^{1,\infty}(\mathbb{R}^d;\mathbb{R}^d)$. Moreover, Assumptions 1 and 2 imply*

$$\nabla p = p\, s^*, \qquad \int_{\mathbb{R}^d} \|\nabla p(x)\|_2 \, dx \le \sqrt{d}\, H_0. \tag{B.1}$$

In particular $p \in W^{1,1}(\mathbb{R}^d)$.

Proof. Mixed networks are locally Lipschitz, since each affine map and activation is locally Lipschitz. The local weak-derivative calculus and the $W^{1,\infty}$ characterization (Brezis, 2011, Chapter 9), together with the envelopes in (10), give a global $W^{1,\infty}$ representative. Continuity turns the essential output bound into a pointwise bound: a strict violation would persist on an open set of positive measure.

The ordinary chain rule gives $\nabla p = p\nabla \log p = ps^*$. Thus $\int \|\nabla p(x)\|_2 \, dx \le \sqrt{d}H_0 \int p(x)\,dx = \sqrt{d}H_0$, which is (B.1). Since $p \in L^1$, this also proves $p \in W^{1,1}$. □

Lemma B.2 (Affine invariance)**.** *Replacing $\phi(u)$ by $\widetilde{\phi}(u) = \phi(u) - \phi(0) - \langle \nabla\phi(0), u\rangle$ leaves D_ϕ, $\mathcal{E}_\phi$, $\widehat{\mathcal{L}}_\phi(f) - \mathcal{L}_\phi(f)$, and the empirical minimizers unchanged. The single-sample loss changes only by the constant $\phi(0)$.*

Proof. For an affine function $a(u) = c + \langle b, u\rangle$, one has $a(u) - a(v) - \langle b, u - v\rangle = 0$. Substitution into (4) proves invariance of the divergence. The Hessian is unchanged, while

$$-\widetilde{\phi}(u) + \left\langle \nabla\widetilde{\phi}(u), u \right\rangle = -\phi(u) + \langle \nabla\phi(u), u\rangle + \phi(0).$$

Therefore $\ell_{\widetilde{\phi}}(f;x) = \ell_\phi(f;x) + \phi(0)$. A constant cancels in population differences, empirical differences, and empirical-minus-population differences. This proves all statements. In subsequent envelope and chaining calculations we use the normalized generator and keep the notation ϕ. □

B.2 Integration by parts without a boundary condition

Lemma B.3 (Full-space integration by parts). *Under Assumptions 1–3, for $f \in \mathcal{F}_n \cup \{s^*\}$,*

$$\mathbb{E}\,\langle \nabla\phi(f(X)), s^*(X)\rangle = -\mathbb{E}\,\mathrm{Tr}\big(\nabla^2\phi(f(X))J_f(X)\big)\,. \tag{B.2}$$

Both expectations are absolutely integrable. Assumption 4 is not used.

Proof. We verify integrability using the envelopes, then remove a compactly supported cutoff. Weak chain and product rules are used in their standard Sobolev form (Brezis, 2011, Chapter 9).

Step 1: Check the transformed field and integrability.

Set $v_f = \nabla\phi \circ f$ and apply Lemma B.2. On the convex score range,

$$v_f(x) = \int_0^1 \nabla^2\phi(tf(x))f(x)\,dt, \qquad \|v_f(x)\|_2 \le \Lambda\sqrt{d}M.$$

The weak chain rule gives $J_{v_f} = \nabla^2\phi(f)J_f$ almost everywhere. Hence

$$\operatorname{div} v_f = \mathrm{Tr}(\nabla^2\phi(f)J_f), \qquad |\operatorname{div} v_f| \le \Lambda\sqrt{d}B_J. \tag{B.3}$$

Here $|\,\mathrm{Tr}(AB)| \le \|A\|_F \|B\|_F$ and $\left\|\nabla^2\phi\right\|_F \le \sqrt{d}\Lambda$ were used. For $f = s^*$, the same estimate follows from $\|J_{s^*}\|_{F,\infty} \le dH_0 < B_J$. Equation (B.1) implies

$$\int |\,\langle v_f, \nabla p\rangle\,| \le \Lambda d M H_0, \qquad \int p|\operatorname{div} v_f| \le \Lambda\sqrt{d}B_J.$$

The entries of J_{v_f} are also bounded, so $pv_f \in W^{1,1}$ and $\operatorname{div}(pv_f) = p\operatorname{div} v_f + \langle \nabla p, v_f\rangle$.

Step 2: Remove the cutoff.

Choose $0 \le \chi \le 1$ in $C_c^\infty(\mathbb{R}^d)$, equal to one on the unit ball and zero outside the ball of radius two; let $\chi_T(x) = \chi(x/T)$. Testing the weak divergence gives

$$\int \chi_T p\,(\operatorname{div} v_f + \langle s^*, v_f\rangle) = -\int p\,\langle v_f, \nabla\chi_T\rangle\,. \tag{B.4}$$

The right side is at most $\Lambda\sqrt{d}M\,\|\nabla\chi\|_\infty / T$ in absolute value. The left integrand is dominated by $p(\Lambda\sqrt{d}B_J + \Lambda d M H_0)$, which is integrable. Dominated convergence as $T \to \infty$ proves (B.2). The cutoff T is only used to justify integration by parts. □

B.3 Population risk identity

The next identity specializes the Bregman score-matching calculation of Gutmann & Hirayama (2011, Section 3.4) to our full-space weak-derivative setting. Lemma B.3 supplies its integration-by-parts justification under the present assumptions.

Proposition B.1 (Population Bregman identity). *Under Assumptions 1–3, every $f \in \mathcal{F}_n$ satisfies*

$$\mathcal{E}_\phi(f) = \mathcal{L}_\phi(f) - \mathcal{L}_\phi(s^*) = \mathbb{E}D_\phi(s^*(X), f(X)) \tag{B.5}$$

and

$$\frac{\lambda}{2}\,\|f - s^*\|^2_{L^2(p)} \le \mathcal{E}_\phi(f) \le \frac{\Lambda}{2}\,\|f - s^*\|^2_{L^2(p)}\,. \tag{B.6}$$

Proof of Proposition B.1. Expand the Bregman discrepancy, apply the full-space integration identity, and then use the Hessian bounds to compare it with squared score error.

Step 1: Expand the expected divergence.

All score values lie in the compact cube $[-M, M]^d$, so ϕ and $\nabla\phi$ are bounded on their values. Expanding (4) gives

$$\begin{aligned}\mathbb{E}D_\phi(s^*(X), f(X)) = \mathbb{E}\phi(s^*(X)) \\ + \mathbb{E}[-\phi(f(X)) + \langle\nabla\phi(f(X)), f(X)\rangle - \langle\nabla\phi(f(X)), s^*(X)\rangle]\,.\end{aligned} \tag{B.7}$$

Applying Lemma B.3 to its last term yields

$$\begin{aligned}\mathbb{E}D_\phi(s^*, f) &= \mathbb{E}\phi(s^*) + \mathbb{E}\{-\phi(f) + \langle\nabla\phi(f), f\rangle - \langle\nabla\phi(f), s^*\rangle\} \\ &= \mathbb{E}\phi(s^*) + \mathcal{L}_\phi(f).\end{aligned} \tag{B.8}$$

Apply the same identity to $f = s^*$; then $D_\phi(s^*, s^*) = 0$, so $\mathcal{L}_\phi(s^*) = -\mathbb{E}\phi(s^*)$. Subtraction in (B.8) proves the first equality in (B.5).

Step 2: Convert Bregman discrepancy to squared error.

For $u, v \in [-M, M]^d$, apply the one-dimensional fundamental theorem of calculus twice to $t \mapsto \phi(v + t(u - v))$:

$$D_\phi(u, v) = \int_0^1 (1 - t)(u - v)^\top \nabla^2\phi(v + t(u - v))(u - v)\, dt. \tag{B.9}$$

The segment is contained in the cube. Assumption 3 and $\int_0^1 (1 - t)dt = 1/2$ give

$$\frac{\lambda}{2}\|u - v\|_2^2 \le D_\phi(u, v) \le \frac{\Lambda}{2}\|u - v\|_2^2\,.$$

Substitute $u = s^*(X)$ and $v = f(X)$ and integrate to obtain the inequalities in (B.6). □

B.4 A FULL-SPACE BOUNDED-SCORE EXAMPLE

The density, score, and tail assumptions allow smooth distributions with unbounded support. For $\kappa_{\text{ex}} > 0$, consider

$$p_{\text{ex}}(x) = Z_{\text{ex}}^{-1} \exp\left(-\kappa_{\text{ex}}\sqrt{1 + \|x\|_2^2}\right), \qquad x \in \mathbb{R}^d,$$

where Z_{ex} is the normalizing constant. The inequality $0 < Z_{\text{ex}} \le \int_{\mathbb{R}^d} e^{-\kappa_{\text{ex}}\|x\|_2}\, dx < \infty$ shows that this is a positive C^∞ density. Direct differentiation gives

$$s_{\text{ex}}(x) = -\frac{\kappa_{\text{ex}} x}{\sqrt{1 + \|x\|_2^2}}, \qquad \|s_{\text{ex}}(x)\|_2 \le \kappa_{\text{ex}}.$$

For a multi-index α of order q, each coordinate of $D^\alpha s_{\text{ex}}$ is a finite sum of terms $P(x)(1 + \|x\|_2^2)^{-k-1/2}$, where $k \ge 0$ and $\deg P \le 2k + 1 - q$; terms with negative degree bounds vanish. This follows inductively: differentiating a numerator lowers its degree, while differentiating a denominator raises k by one and multiplies the numerator by a coordinate. Each term is bounded on $\mathbb{R}^d$. Thus, for every fixed integer $m \ge 2$, a finite constant $H_0 = H_0(d, m, \kappa_{\text{ex}})$ satisfies $\|s_{\text{ex}}\|_{W^{m,\infty}} \le H_0$.

For $0 < a < \kappa_{\text{ex}}$, the radial exponential bound gives

$$\mathbb{E}_{p_{\text{ex}}} e^{a\|X\|_2} \le Z_{\text{ex}}^{-1} \int_{\mathbb{R}^d} e^{-(\kappa_{\text{ex}} - a)\|x\|_2}\, dx < \infty.$$

Since $\langle t, X\rangle \le a\|X\|_2$ whenever $\|t\|_2 \le a$, the moment generating function is uniformly bounded on this ball. The example therefore satisfies Assumptions 1, 2, and 4. Together with any generator satisfying Assumption 3 and fixed envelopes $M > H_0$, $B_J > dH_0$, it provides a full-space model for Theorem 1 and Corollary 1.

C EMPIRICAL-PROCESS TOOLS AND THE STOCHASTIC BOUND

C.1 A FIXED BOUNDED JACOBIAN REPRESENTATIVE

For a specified network realization, compute its formal input Jacobian $\widetilde{J}_f$ using $\mathrm{ReLU}'(z) = \mathbf{1}\{z > 0\}$ and $\mathrm{ReQU}'(z) = 2\,\mathrm{ReLU}(z)$. The zero derivative convention at a ReLU knot is fixed and sample-independent. Use

$$J_f(x) = \widetilde{J}_f(x)\,\mathbf{1}\{\left\|\widetilde{J}_f(x)\right\|_F \le B_J\} \tag{C.1}$$

in the empirical loss. This guard implements the bounded Jacobian representative specified in Section 2.3. It does not change the function f or its weak Jacobian class as equivalence classes modulo Lebesgue null sets.

Lemma C.1 (Validity of the representative convention)**.** *For every $f \in \mathcal{F}_n$,* (C.1) *is a Borel measurable representative of the weak input Jacobian and satisfies $\|J_f(x)\|_F \le B_J$ for every x. All capacity estimates below include the guard. If the raw convention already obeys the pointwise envelope for every candidate, the guard can be omitted.*

Proof. Every intermediate neuron is locally Lipschitz. The weak chain rule for the positive-part function (Brezis, 2011, Chapter 9) states that $\partial_j \mathrm{ReLU}(h) = \mathbf{1}\{h > 0\}\partial_j h$ almost everywhere. On the level set $\{h = 0\}$ the weak gradient of the locally Lipschitz h is zero almost everywhere; thus the chosen zero derivative is compatible with this chain rule even if the level set has positive measure. The ReQU identity follows by applying the ordinary product rule to $\mathrm{ReLU}(h)^2$. Induction over finitely many neurons gives $\widetilde{J}_f = J_f^{\mathrm{weak}}$ outside a Lebesgue null set for each fixed network.

By the essential bound in (10), the set where $\left\|\widetilde{J}_f\right\|_F > B_J$ is another null set. Multiplication by the indicator in (C.1) therefore leaves the weak derivative unchanged almost everywhere. Borel measurability follows from finite arithmetic operations and threshold comparisons, and the pointwise envelope follows from the definition. This argument is for each fixed candidate; it does not claim that the union of exceptional sets over all candidates is null. The next lemma controls the selected representatives on arbitrary finite samples and hence also covers data-dependent choices of $\widehat{s}$. □

C.2 CAPACITY OF OUTPUTS AND SELECTED FIRST DERIVATIVES

Define the coordinate output and Jacobian classes by

$$\mathcal{S}_{n,k} = \{x \mapsto f_k(x) : f \in \mathcal{F}_n\}, \qquad \mathcal{J}_{n,ij} = \{x \mapsto [J_f(x)]_{ij} : f \in \mathcal{F}_n\}. \tag{C.2}$$

The pseudo-dimension of a real-valued class is the VC dimension of its subgraphs: it is the largest integer q for which fixed $(x_r, t_r)_{r=1}^q$ admit all 2^q patterns of comparisons $g(x_r) > t_r$. This definition allows arbitrary thresholds at each point.

Classical polynomial sign inequality. For r polynomials of degree at most $D \ge 1$ in $P_{\mathrm{net}} \ge 1$ real variables, a zero-inclusive sign-pattern bound is

$$\#\{(\operatorname{sgn} q_1(\theta), \dots, \operatorname{sgn} q_r(\theta)) : \theta \in \mathbb{R}^{P_{\mathrm{net}}}\} \le [16eD(r + P_{\mathrm{net}} + 1)]^{P_{\mathrm{net}}}. \tag{C.3}$$

For completeness, the zero-inclusive form follows from the strict-sign inequality of Warren (1968). Choose a representative point for each realized ternary pattern and one sufficiently small positive η for this finite collection. The signs of $q_i - \eta, q_i + \eta$ at these points encode negative, zero, and positive values injectively. Apply the strict-sign bound to the $2r$ perturbed polynomials. When $2r \ge P_{\mathrm{net}}$ it is at most $(8eDr/P_{\mathrm{net}})^{P_{\mathrm{net}}}$, and when $2r < P_{\mathrm{net}}$ it is at most $2^{2r} \le 4^{P_{\mathrm{net}}}$. Both are bounded by (C.3). The strict-sign theorem is the classical external input; its use here does not impose a parameter-size bound.

Let $V_n \ge 1$ be a common upper bound for the pseudo-dimensions of the classes in (C.2).

The proof follows the layerwise sign-counting method for neural networks (Anthony & Bartlett, 1999; Warren, 1968). We track the terminal ReQU depth and the representative guard explicitly, since they affect the polynomial degrees.

Lemma C.2 (Architectural pseudo-dimension bound)**.** *For $W \geq d$, $L \geq 2$, each class in* (C.2) *has pseudo-dimension at most*

$$C_{\dim} W^2 L^2 \log(eWL), \tag{C.4}$$

where $C_{\dim}$ depends only on d, c_{term}. This remains true when the output and derivative envelopes restrict the parameter set.

Proof. We count activation and guard sign patterns on a finite sample. On each resulting cell, outputs and derivatives are polynomials of controlled degree; their comparison patterns then bound pseudo-dimension.

Step 1: Count parameters and bound polynomial degrees.

Pad each hidden layer to width W with zero neurons, without adding layers. For each choice of depth and of the ReLU/ReQU split, the number of affine parameters is at most

$$P_{\mathrm{net}} = (L+1)(W+d+1)^2.$$

Let $T_L = \min\{L, \lfloor c_{\mathrm{term}} \log(eL) \rfloor\}$. On a region of parameter space with all previous activation comparisons fixed, a preactivation is a polynomial in the parameters. An affine layer increases its degree by at most one; a ReLU leaves the polynomial degree unchanged on each branch, and a ReQU multiplies it by at most two. Induction therefore bounds the degree of every output polynomial by

$$D_L = (L+2)2^{T_L}.$$

On each such branch, differentiating the polynomial with respect to an input coordinate does not increase its degree in the parameters. Thus D_L also bounds the raw derivative degrees. At zero ReLU preactivations the inactive branch agrees with the convention used in Lemma C.1.

Step 2: Count evaluation patterns on an arbitrary sample.

Fix $q \geq 1$ input points and fixed subgraph thresholds. At each hidden layer at most qW polynomial comparisons are needed, each of degree at most D_L. The norm guard adds q comparisons of degree at most $2D_L$, since it tests $\sum_{i,j} \widetilde{J}_{ij}^2 \leq B_J^2$. Final output or derivative threshold comparisons add at most one more stage of q polynomials of degree at most D_L on the refined branches.

Using (C.3) at each stage, and restricting its global sign bound to each previously created parameter region, bounds the number of patterns for a fixed architecture by

$$[32eD_L q(W + P_{\mathrm{net}} + d + 1)]^{P_{\mathrm{net}}(L+2)}.$$

There are at most $(L+1)^2$ choices of depth and split. With

$$h = P_{\mathrm{net}}(L+2), \qquad K_L = 32eD_L(W + P_{\mathrm{net}} + d + 1)(L+1)^2, \tag{C.5}$$

the growth function of either scalar class is consequently at most $(K_L q)^h$. No independence between output coordinates is used. Restriction to $\mathcal{F}_n$ cannot create new patterns.

Step 3: Convert the growth bound to pseudo-dimension.

If q points are pseudo-shattered, then $2^q \leq (K_L q)^h$. Writing $x = q/h$ gives

$$x \leq \log_2(K_L h) + \log_2 x.$$

For $x \geq 4$, $\log_2 x \leq x/2$. If $x > 4\log_2(4K_L h)$, this inequality implies $x - \log_2 x \geq x/2 > \log_2(K_L h)$, a contradiction. The range $x < 4$ is included by enlarging the same upper bound. Hence

$$V_n = \lceil 4h \log_2(4K_L h) \rceil \tag{C.6}$$

is a permissible common upper bound.

Step 4: Simplify in terms of actual width and depth.

Since $W \geq d$ and $L \geq 2$, $P_{\mathrm{net}} \lesssim W^2 L$ and $h \lesssim W^2 L^2$, with absolute hidden constants. Also

$$\log D_L \leq \log(L+2) + c_{\mathrm{term}} \log 2 \, \log(eL).$$

Insert these inequalities into (C.5) and (C.6). Both $\log K_L$ and $\log h$ are $\lesssim \log(eWL)$, with hidden constants depending only on d, c_{term}. This proves (C.4). □

C.3 Loss envelope and Lipschitz control

Lemma C.3 (Bounds for the derivative-dependent loss)**.** *After the invariant affine normalization in Lemma B.2, the loss satisfies*

$$|\ell_\phi(f;x)| \le B_\ell := \frac{\Lambda}{2} dM^2 + \Lambda\sqrt{d}B_J, \quad f \in \mathcal{F}_n,\ x \in \mathbb{R}^d. \tag{C.7}$$

Moreover, with

$$A_\ell = \sqrt{d}(\Lambda M + L_\phi B_J), \qquad B'_\ell = \Lambda\sqrt{d}, \tag{C.8}$$

any $f, g \in \mathcal{F}_n$ obey

$$|\ell_\phi(f;x) - \ell_\phi(g;x)| \le A_\ell \left\|f(x) - g(x)\right\|_2 + B'_\ell \left\|J_f(x) - J_g(x)\right\|_F. \tag{C.9}$$

Proof. We bound the value-dependent and Jacobian-dependent parts separately. The same split gives the envelope and a Lipschitz estimate for constructing covers.

Step 1: Bound the nondifferential part.

Put $h_\phi(u) = \langle \nabla\phi(u), u\rangle - \phi(u)$. Direct differentiation gives $\nabla h_\phi(u) = \nabla^2\phi(u)u$, without needing a third derivative of ϕ. Since $h_\phi(0) = 0$,

$$h_\phi(u) = \int_0^1 t\, u^\top \nabla^2\phi(tu)u\, dt, \qquad 0 \le h_\phi(u) \le \frac{\Lambda}{2}\left\|u\right\|_2^2.$$

For $u \in [-M, M]^d$, $\left\|u\right\|_2 \le \sqrt{d}M$. Furthermore, integrating ∇h_ϕ along the segment from v to u yields

$$|h_\phi(u) - h_\phi(v)| \le \Lambda\sqrt{d}M \left\|u - v\right\|_2.$$

Step 2: Bound the trace and its increment.

The Frobenius Cauchy–Schwarz inequality and (C.1) give

$$|\operatorname{Tr}(\nabla^2\phi(u)J)| \le \Lambda\sqrt{d}B_J.$$

For $\left\|J\right\|_F, \left\|K\right\|_F \le B_J$, add and subtract $\operatorname{Tr}(\nabla^2\phi(v)J)$ to obtain

$$\begin{aligned}
&|\operatorname{Tr}(\nabla^2\phi(u)J) - \operatorname{Tr}(\nabla^2\phi(v)K)| \\
&\quad\le \left\|\nabla^2\phi(u) - \nabla^2\phi(v)\right\|_F \left\|J\right\|_F + \left\|\nabla^2\phi(v)\right\|_F \left\|J - K\right\|_F \\
&\quad\le \sqrt{d}L_\phi B_J \left\|u - v\right\|_2 + \sqrt{d}\Lambda \left\|J - K\right\|_F.
\end{aligned}$$

Combine the two steps, substitute $u = f(x)$ and $v = g(x)$, and use (5). This proves both bounds. The argument uses no density or moment condition. □

Remark C.1 (Curvature constants for the three generators). On the score range $[-M, M]^d$, the generators in Table 1 satisfy Assumption 3 with the following choices:

$$\begin{aligned}
\text{Quadratic:} \quad & (\lambda, \Lambda, L_\phi) = (1, 1, 0), \\
\text{Anisotropic quadratic:} \quad & (\lambda, \Lambda, L_\phi) = \big(\lambda_{\min}(A), \lambda_{\max}(A), 0\big), \\
\text{Hypentropy:} \quad & (\lambda, \Lambda, L_\phi) = \left((M^2 + \beta^2)^{-1/2},\ \beta^{-1},\ \frac{2}{3\sqrt{3}\,\beta^2}\right).
\end{aligned}$$

The quadratic Hessians are constant matrices I_d and A, so their Hessian Lipschitz constants are zero. For the hypentropy generator in (8), direct differentiation gives

$$h'_\beta(t) = \operatorname{arsinh}(t/\beta), \qquad h''_\beta(t) = (t^2 + \beta^2)^{-1/2}, \qquad h'''_\beta(t) = -\frac{t}{(t^2 + \beta^2)^{3/2}}.$$

The range $|t| \le M$ gives the stated lower and upper curvature bounds. Writing $r = |t|/\beta$ yields $|h'''_\beta(t)| = \beta^{-2} r(1 + r^2)^{-3/2}$. The derivative of $r(1 + r^2)^{-3/2}$ is $(1 - 2r^2)(1 + r^2)^{-5/2}$, so its maximum for $r \ge 0$ is $2/(3\sqrt{3})$, attained at $r = 1/\sqrt{2}$. The mean value theorem and the diagonal Hessian then give

$$\left\|\nabla^2\phi_\beta(u) - \nabla^2\phi_\beta(v)\right\|_{\mathrm{op}} \le \frac{2}{3\sqrt{3}\,\beta^2}\left\|u - v\right\|_\infty \le \frac{2}{3\sqrt{3}\,\beta^2}\left\|u - v\right\|_2.$$

This is a valid global Lipschitz bound and need not be the smallest one on the specified score range. For fixed M, decreasing β increases these upper curvature and Hessian-variation bounds, which enter the loss estimates (C.7)–(C.9).

C.4 From pseudo-dimension to distribution-free covering numbers

For a class $\mathcal{G}$ and a probability measure Q, write $N(\varepsilon, \mathcal{G}, L^2(Q))$ for its covering number with centers in the class. The metric is the scalar $L^2(Q)$ norm. The scalar entropy estimate is the standard VC-subgraph covering argument (Anthony & Bartlett, 1999, Chapter 12). We give its short quantitative specialization because its constants enter the loss-class bound.

Lemma C.4 (Scalar VC-type entropy). *If $|g| \le b$ for every $g \in \mathcal{G}$, $b > 0$, and $\mathrm{Pdim}(\mathcal{G}) \le V$ with $V \ge 1$, then for every probability measure Q and $0 < \varepsilon \le 2b$,*

$$\log N(\varepsilon, \mathcal{G}, L^2(Q)) \le 4V \log\left(\frac{16eb}{\varepsilon}\right). \tag{C.10}$$

Proof. We apply a VC packing estimate to subgraphs, then convert L^1 separation into L^2 separation.

Step 1: Use the VC counting bound.

For an η-separated collection of $N \ge 2$ indicators, take $k = \lceil 3 \log N/\eta \rceil$. A pair agrees on all k independent observations with probability at most $e^{-k\eta}$. A union bound shows that some sample distinguishes all N indicators. The Sauer–Shelah counting bound (Anthony & Bartlett, 1999) therefore gives $N \le \sum_{j=0}^{V} \binom{k}{j}$. If $k < V$, then $\log N < V \log 2$. Otherwise, $N \le (ek/V)^V$ and $k \le 5 \log N/\eta$, so $x = \log N/V$ satisfies $x \le \log(5ex/\eta)$. Comparing $x - \log x$ at $2\log(12e/\eta)$ yields, in both cases,

$$\log N \le 2V \log(12e/\eta). \tag{C.11}$$

Step 2: Normalize and cover the subgraphs.

For $g_b = (g + b)/(2b)$, the indicators $\mathbf{1}\{t < g_b(x)\}$ under $Q \otimes \mathrm{Unif}[0, 1]$ have VC dimension at most V. Their disagreement probability is

$$\frac{\|g - h\|_{L^1(Q)}}{2b} \ge \frac{\|g - h\|^2_{L^2(Q)}}{4b^2},$$

using $|g-h|^2 \le 2b|g-h|$. Thus an $L^2(Q)$ ε-packing induces an indicator packing with $\eta = \varepsilon^2/(4b^2)$. By (C.11), its logarithmic size is at most $2V \log(48eb^2/\varepsilon^2) \le 4V \log(16eb/\varepsilon)$. A maximal packing is a cover with centers in the class, proving (C.10). □

Lemma C.5 (Entropy of the Bregman loss class). *Let $q_d = d + d^2$ and define the constants*

$$\begin{aligned} b &= \max\{1, M, B_J\}, & K_\ell &= \max\{1, \sqrt{d}A_\ell + dB'_\ell\}, \\ A^{\mathrm{ent}}_\ell &= \max\{B_\ell, 32ebK_\ell\}, & H_\ell &= 4q_d + 2. \end{aligned} \tag{C.12}$$

For $\mathcal{G}_n = \{\ell_\phi(f; \cdot) : f \in \mathcal{F}_n\}$ and $\mathcal{G}^{\pm}_n = \mathcal{G}_n \cup (-\mathcal{G}_n) \cup \{0\}$, every probability measure Q and $0 < \varepsilon \le B_\ell$ satisfy

$$\log N(\varepsilon, \mathcal{G}^{\pm}_n, L^2(Q)) \le H_\ell V_n \log(1 + A^{\mathrm{ent}}_\ell/\varepsilon). \tag{C.13}$$

Proof. Scalar covers are combined through actual network representatives. The loss Lipschitz bound then converts their product cover into a cover of the derivative-dependent loss class.

Step 1: Cover the scalar output and Jacobian classes.

All q_d classes in (C.2) have envelope at most b. Choose scalar covers of radius $\eta = \varepsilon/(2K_\ell)$. The definitions imply

$$B_\ell \le b(\sqrt{d}A_\ell + dB'_\ell) \le bK_\ell,$$

so $\eta \le b/2$ in the required range. Lemma C.4 applies to every cover.

Step 2: Use compatible representatives.

Assign a candidate f to the indices of its scalar cover centers. For each nonempty combination of indices, choose one actual network g assigned to that combination. Then each coordinate of $f-g$ and each entry of $J_f - J_g$ has $L^2(Q)$ norm at most 2η. Consequently

$$\big\|\|f-g\|_2\big\|_{L^2(Q)} \le 2\sqrt{d}\eta, \qquad \big\|\|J_f - J_g\|_F\big\|_{L^2(Q)} \le 2d\eta.$$

Minkowski's inequality and (C.9) therefore give $\|\ell_\phi(f) - \ell_\phi(g)\|_{L^2(Q)} \le 2(\sqrt{d}A_\ell + dB'_\ell)\eta \le \varepsilon$. This construction is important: independent scalar centers need not assemble into a valid network or Jacobian. The centers used for the loss cover are actual compatible candidates.

Step 3: Multiply cover sizes and symmetrize.

The number of occupied combinations is at most the product of the q_d scalar covering numbers. Thus

$$\log N(\varepsilon, \mathcal{G}_n, L^2(Q)) \le 4q_d V_n \log(32ebK_\ell/\varepsilon).$$

Adding the negative class and zero multiplies the size by at most three. Since $A_\ell^{\mathrm{ent}} \ge B_\ell \ge \varepsilon$ and $V_n \ge 1$, the added $\log 3$ is bounded by $2V_n \log(1 + A_\ell^{\mathrm{ent}}/\varepsilon)$. Increasing the logarithm in the preceding display to $\log(1 + A_\ell^{\mathrm{ent}}/\varepsilon)$ proves (C.13). □

C.5 Chaining with an integrable small-scale entropy

Lemma C.6 (Explicit conditional chaining bound)**.** *For independent Rademacher signs $\sigma_1, \dots, \sigma_n$ and any fixed sample $\mathcal{D}$,*

$$\mathbb{E}_\sigma \sup_{g \in \mathcal{G}_n} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right| \le C_{\mathrm{rad}} \sqrt{V_n/n}, \tag{C.14}$$

where a permissible constant is

$$C_{\mathrm{rad}} = 3\sqrt{2}B_\ell \left[\sqrt{\log 2 + 2H_\ell \log(1 + A_\ell^{\mathrm{ent}}/B_\ell)} + 2\sqrt{2H_\ell \log 2} \right]. \tag{C.15}$$

Proof. We use the dyadic chaining argument for sub-Gaussian increments (van der Vaart & Wellner, 1996, Section 2.2), retaining the constants needed in (C.15).

Step 1: Apply the conditional maximal inequality.

For $Z_g = n^{-1} \sum_i \sigma_i g(X_i)$, independence of the signs gives

$$\mathbb{E}_\sigma e^{t(Z_g - Z_h)} \le \exp\left\{ \frac{t^2}{2n} \|g-h\|^2_{L^2(P_n)} \right\}.$$

The finite sub-Gaussian maximal inequality (van der Vaart & Wellner, 1996, Section 2.2) then implies, for N pairs at distance at most r,

$$\mathbb{E}_\sigma \max_{j \le N} |Z_{g_j} - Z_{h_j}| \le \frac{r}{\sqrt{n}} \sqrt{2\log(2N)}. \tag{C.16}$$

Step 2: Sum the dyadic increments.

In $\mathcal{G}_n^{\pm}$, take covers T_k of radius $\varepsilon_k = B_\ell 2^{-k}$, with $T_0 = \{0\}$, and select projections $\pi_k g \in T_k$. Adjacent projections have distance at most $3B_\ell 2^{-k}$. The residual after K levels is bounded by ε_K by Cauchy–Schwarz. Writing $a_0 = \log(1 + A_\ell^{\mathrm{ent}}/B_\ell)$, Lemma C.5 gives $\log|T_k| \le H_\ell V_n (a_0 + k\log 2)$. Apply (C.16) to the at most $|T_k||T_{k-1}|$ pairs at each level and let $K \to \infty$. The result is bounded by

$$3\sqrt{2}B_\ell \sqrt{V_n/n} \sum_{k=1}^\infty 2^{-k} \sqrt{\log 2 + 2H_\ell a_0 + 2H_\ell k \log 2}.$$

Finally, use $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$, $\sum_{k \ge 1} 2^{-k} = 1$, and $\sum_{k \ge 1} 2^{-k}\sqrt{k} \le 2$. This yields exactly the constant in (C.15), with no sample-dependent small-scale cutoff. □

Proposition C.1 (Uniform stochastic error). *Under Assumption 3 and Definition 1, for every $n \geq 1$,*

$$\mathfrak{S}_n \leq \frac{C_{\mathrm{sto}}}{2}\sqrt{\frac{V_n}{n}}. \tag{C.17}$$

The constant C_{sto} depends only on $d, M, B_J, \Lambda, L_\phi$. It is independent of n, p, λ and the approximation radius. An admissible expression is given in (C.18).

Proof of Proposition C.1. The standard symmetrization inequality (van der Vaart & Wellner, 1996, Section 2.3), followed by Lemma C.6, gives

$$\mathfrak{S}_n \leq 2\mathbb{E}_{\mathcal{D}}\mathbb{E}_\sigma \sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right| \leq 2C_{\mathrm{rad}}\sqrt{V_n/n}.$$

Set

$$C_{\mathrm{sto}} = 4C_{\mathrm{rad}}. \tag{C.18}$$

The definitions in (C.15), (C.7), (C.8), and (C.12) give the stated constant dependencies. Symmetrization is applied with outer expectations under the convention in Appendix A. □

Lemma C.7 (Uniform concentration about the mean). *Let $Z(\mathcal{D}) = \sup_{f\in\mathcal{F}_n}|\widehat{\mathcal{L}}_\phi(f) - \mathcal{L}_\phi(f)|$. Then, with probability at least $1-\delta$,*

$$Z(\mathcal{D}) \leq \mathfrak{S}_n + B_\ell\sqrt{2\log(1/\delta)/n}. \tag{C.19}$$

Proof. Changing one observation changes $Z(\mathcal{D})$ by at most $2B_\ell/n$. The bounded-differences inequality of McDiarmid (1989), with coordinate bounds $2B_\ell/n$, gives

$$\mathbb{P}\{Z - \mathbb{E}Z > t\} \leq \exp\{-nt^2/(2B_\ell^2)\}.$$

Substituting $t = B_\ell\sqrt{2\log(1/\delta)/n}$ and $\mathbb{E}Z = \mathfrak{S}_n$ proves (C.19), with the outer-probability convention when needed. □

D Simultaneous approximation and global admissibility

D.1 The unit-cube construction and its lower-order bounds

We retain the following scalar width and depth budgets for the local construction and its assembled network:

$$\begin{aligned}\bar{W}_{m,d}(U) &= 2^{d+6}m^{d+1}(U+d)\lceil\log_2(8U)\rceil,\\ \bar{L}(V) &= 9(V+1)\lceil\log_2(4V)\rceil.\end{aligned} \tag{D.1}$$

Here U, V are auxiliary approximation indices; they are not actual widths, depths, or sample sizes. Choose $U \geq U_0(m,d)$ and $V \geq V_0(m,d)$ so that the scalar construction applies. One can take $U_0 \geq \max\{2, d, (md)^d\}$ and $V_0 \geq \max\{2, m+1, d\}$, increasing these fixed thresholds for integer rounding and the fixed-depth product operations if necessary. This records, in particular, the width condition $\log U \geq d(\log m + \log d)$ and the depth required for polynomial representation in the cited construction.

The weights and the local approximants are constructed on one mesh. The following realization uses only ReLU folding before a fixed-depth ReQU block. All supports in this subsection are relative to $[0,1]^d$.

Lemma D.1 (Same-grid partition weights). *Let $U, V \geq 2$ be integers and set*

$$a = \lfloor U^{1/d}\rfloor, \qquad b = \lfloor V^{2/d}\rfloor, \qquad K_{U,V} = a^2 b, \qquad r = \lceil\log_2(4b)\rceil. \tag{D.2}$$

For $K = K_{U,V}$, define the two families of intervals

$$\begin{aligned} I_{1,K} &= \bigcup_{i=0}^{K-1}\left[\frac{i}{K}, \frac{i+3/4}{K}\right],\\ I_{2,K} &= [0,1]\cap\bigcup_{i=0}^{K}\left[\frac{i-1/2}{K}, \frac{i+1/4}{K}\right], \qquad \Omega_\nu = \prod_{j=1}^d I_{\nu_j,K}, \quad \nu\in\{1,2\}^d. \end{aligned} \tag{D.3}$$

There are C^1 weights ρ_ν on the unit cube satisfying

$$0 \le \rho_\nu \le 1, \qquad \sum_\nu \rho_\nu = 1, \qquad \operatorname{supp}_{[0,1]^d} \rho_\nu \subseteq \Omega_\nu, \qquad \|\partial_j \rho_\nu\|_\infty \le 8K. \tag{D.4}$$

All 2^d weights can be realized jointly by a ReLU block of depth $r+2$ and width at most $d\max\{a,2\}$, followed by a ReQU block of depth at most $1+\lceil \log_2 d \rceil$ and width at most $2^{d+3}(d+1)$. The output layer is affine. In particular, the ReLU width is $O_d(U)$ and its depth is $O_d(\log(eV))$.

Proof. We realize the fine mesh by composing ReLU folding maps, then apply a fixed ReQU gate. Tensor products give the multivariate weights with the required supports and derivative bounds.

Step 1: Realize the fine-grid triangular wave.

Let $\tau(t) = \operatorname{dist}(t, 2\mathbb{Z})$ and $T_j(t) = \tau(jt)$ for $t \in [0,1]$ and an integer $j \ge 1$. The slopes of T_j alternate between j and $-j$ at k/j, so

$$T_j(t) = j\,\mathrm{ReLU}(t) + 2j \sum_{k=1}^{j-1} (-1)^k \,\mathrm{ReLU}\left(t - \frac{k}{j}\right), \qquad 0 \le t \le 1. \tag{D.5}$$

This is a one-hidden-layer ReLU realization with j neurons; for $j = 1$ the sum is empty.

Every $t \in \mathbb{R}$ can be written as $2k + \tau(t)$ or $2k - \tau(t)$ for an integer k. Evenness and 2-periodicity of τ then give

$$\tau(j\tau(t)) = \tau(jt), \qquad j \in \mathbb{N}_+, \quad t \in \mathbb{R}. \tag{D.6}$$

For $x \in [0,1]$, set

$$\psi_K(x) = T_a \circ T_a \circ T_2^{\circ r}\left(\frac{2bx + (4a^2)^{-1}}{2^r}\right) = \tau\left(2Kx + \frac{1}{4}\right). \tag{D.7}$$

The initial affine argument belongs to $[0,1]$, since $2^r \ge 4b$ and $2b + (4a^2)^{-1} \le 4b$. Every folding map also takes values in $[0,1]$. Thus (D.5) applies at each stage, and repeated use of (D.6) proves the equality in (D.7). Adjacent affine layers are absorbed into the next affine transformation. There are r binary folds and two a-folds, giving $r+2$ ReLU hidden layers of width at most $\max\{a,2\}$. The two a-folds generate the factor a^2 without an a^2-width layer.

Step 2: Turn the fold into two smooth weights.

Define a fixed ReQU gate by

$$\eta(u) = 2\,\mathrm{ReQU}\left(2u - \frac{1}{2}\right) - 4\,\mathrm{ReQU}(2u-1) + 2\,\mathrm{ReQU}\left(2u - \frac{3}{2}\right). \tag{D.8}$$

Expanding the three squares on their activation intervals yields

$$\eta(u) = \begin{cases} 0, & u \le 1/4, \\ 8(u-1/4)^2, & 1/4 \le u \le 1/2, \\ 1 - 8(3/4-u)^2, & 1/2 \le u \le 3/4, \\ 1, & u \ge 3/4. \end{cases} \tag{D.9}$$

The values and first derivatives agree at the three junctions. Consequently $0 \le \eta \le 1$ and $\|\eta'\|_\infty \le 4$. Put

$$\rho_1(x) = \eta(\psi_K(x)), \qquad \rho_2(x) = 1 - \rho_1(x), \qquad x \in [0,1]. \tag{D.10}$$

Then $\rho_1 + \rho_2 = 1$ and $\|\rho_i'\|_\infty \le 4(2K) = 8K$ by the chain rule. At a corner of ψ_K, its value is either 0 or 1; the gate is constant in a neighborhood of either value. Therefore the compositions in (D.10) are C^1 even at the folding corners.

On one period, $\eta(\tau(t)) > 0$ precisely when $1/4 < t < 7/4$ modulo 2, whereas $1 - \eta(\tau(t)) > 0$ when $-3/4 < t < 3/4$ modulo 2. Substituting $t = 2Kx + 1/4$ shows that the relative supports of ρ_1 and ρ_2 lie in $I_{1,K}$ and $I_{2,K}$, respectively. This also covers the two endpoints of the unit interval.

Step 3: Form the tensor weights in the terminal block.

Set

$$\rho_\nu(x) = \prod_{j=1}^{d} \rho_{\nu_j}(x_j), \qquad \nu \in \{1,2\}^d. \tag{D.11}$$

The sum factorizes as $\sum_\nu \rho_\nu = \prod_j (\rho_1(x_j) + \rho_2(x_j)) = 1$. The support inclusion follows coordinatewise. For each j,

$$\partial_j \rho_\nu(x) = \rho'_{\nu_j}(x_j) \prod_{k \neq j} \rho_{\nu_k}(x_k),$$

which proves the derivative bound in (D.4).

For the network realization, let

$$\begin{aligned} S(z) &= \mathrm{ReQU}(z) + \mathrm{ReQU}(-z) = z^2, \\ \mathsf{P}(u,v) &= \frac{S(u+v) - S(u-v)}{4} = uv, \qquad \mathsf{I}(u) = \frac{S(u+1) - S(u-1)}{4} = u. \end{aligned} \tag{D.12}$$

Each of P and I uses one ReQU hidden layer with four neurons. After the d coordinate folding networks, one ReQU layer with $3d$ neurons computes the gates. The two affine combinations η and $1-\eta$ share these gate outputs. A balanced product tree then forms each product in (D.11) using at most $\lceil \log_2 d \rceil$ further ReQU layers. An unpaired factor is carried by I. At each such layer, no more than d factors per multi-index are present, so four neurons per factor, together with the shared gate layer, fit within the stated width $2^{d+3}(d+1)$. All new operations after the folds are ReQU operations. Finally, $a \le U$ and $b \le V^{2/d}$ give the asserted width and depth orders. □

Quoted localized construction. Specialize Yang et al. (2025), Section 4.3, to the unit cube, regularity order m, and integrability index ∞ (their regularity index n is our m; their N, L are our U, V). For $g \in W^{m,\infty}([0,1]^d)$ with norm at most G, use the local networks q_ν, $\nu \in \{1,2\}^d$, on the corresponding translated regions Ω_ν in (D.3). Combine their local error estimates with the weights constructed in Lemma D.1. On the common mesh this gives

$$\begin{aligned} &\sum_\nu \rho_\nu = 1, \qquad \|\partial_j \rho_\nu\|_\infty \le 8K_{U,V}, \qquad K_{U,V} = \lfloor U^{1/d} \rfloor^2 \lfloor V^{2/d} \rfloor \le (UV)^{2/d}, \\ &\|g - q_\nu\|_{L^\infty(\Omega_\nu)} \le C_{\mathrm{loc}} G (UV)^{-2m/d}, \\ &\max_j \|\partial_j (g - q_\nu)\|_{L^\infty(\Omega_\nu)} \le C_{\mathrm{loc}} G (UV)^{-2(m-1)/d}, \end{aligned} \tag{D.13}$$

The constant C_{loc} depends only on m, d. The two local error estimates and local branch architecture are the input from Yang et al. (2025), Section 4.3, equation (23); the partition properties are proved above. Their supports refer to the same regions used for the local error estimates. The next lemma verifies the capacity of the assembled network using the explicit weights.

Lemma D.2 (Assembly within the width–depth budget)**.** *Use the local networks q_ν in* (D.13) *and the weights of Lemma D.1. For the fixed thresholds U_0, V_0 specified above, the single network*

$$q = \sum_{\nu \in \{1,2\}^d} \rho_\nu q_\nu \tag{D.14}$$

can be realized with scalar width and total hidden depth bounded by (D.1)*. All ReLU layers precede all ReQU layers, and the terminal-block length is at most $c_{\mathrm{app}} \log(e\bar{L}(V))$ for a fixed c_{app} depending only on d, m.*

Proof. We run the local approximation branches and the same-grid weights in parallel, align their activation blocks, and count the final product operations within the prescribed budgets.

Step 1: Align the local branches and the folding branch.

Write

$$Q = m^{d+1}(U+d)\lceil \log_2(8U)\rceil, \qquad D_{\mathrm{loc}} = 9(V+1)\lceil \log_2 V\rceil.$$

The local realization in Yang et al. (2025), Section 4.3, has width at most $28Q$ and depth at most D_{loc} for each q_ν. The translated local constructions use the same layer pattern; their coefficients and affine input maps change with ν. Thus their ReLU blocks can be run in parallel with a common length, followed by their ReQU blocks with a common length. These local block lengths and their logarithmic terminal-block bound are the local-network input used here.

Run the d folding networks from (D.7) alongside the local ReLU blocks. If either branch finishes first, propagate its nonnegative ReLU features through identity ReLU layers. Delay the gate (D.8) until all ReLU computations finish. The local ReQU blocks and the gate/product blocks then run in parallel; I in (D.12) carries completed scalar outputs when their terminal lengths differ. A final ReQU multiplication layer computes $\rho_\nu q_\nu$, and an affine output sums over ν. This ordering introduces no ReLU layer after a ReQU layer.

Step 2: Count the maximum total width.

The 2^d local branches require at most $28 \cdot 2^d Q$ neurons at any layer. The folding channels, the tensor gates, and the scalar identity and final-product channels can be bounded together by

$$2d\max\{a,2\} + 2^{d+4}(d+1).$$

The first term covers both active folds and carried coordinate values. For the second, the product trees have at most d factors for each of 2^d multi-indices, each product or identity uses four neurons, and completed local outputs add at most four neurons per multi-index. The shared $3d$ gate neurons also fit within this bound.

For $m \geq 2$ and $U \geq \max\{2,d\}$, the definition of Q gives $Q \geq d\max\{a,2\}$ and $Q \geq d+1$. Therefore

$$\mathrm{width}(q) \leq 28 \cdot 2^d Q + 2Q + 16 \cdot 2^d Q \leq 64 \cdot 2^d Q = \bar{W}_{m,d}(U). \tag{D.15}$$

The factor 2^d for the local branches has been included in the total width.

Step 3: Count depth and the terminal layers.

Let $t = 1 + \lceil \log_2 d\rceil$, the depth of the weights' ReQU block. Parallel alignment adds at most $r+2$ layers before the ReQU transition and at most $t+1$ layers after it. Hence

$$\mathrm{depth}(q) \leq D_{\mathrm{loc}} + r + 4 + \lceil \log_2 d\rceil \leq D_{\mathrm{loc}} + 18(V+1) = \bar{L}(V). \tag{D.16}$$

For the second inequality, use $r \leq 3 + 2\log_2 V$, $V \geq d$, $\lceil \log_2 d\rceil \leq V$, and $\log_2 V \leq V$. These yield $r + 4 + \lceil \log_2 d\rceil \leq 7 + 3V \leq 18(V+1)$. The last equality uses $\lceil \log_2(4V)\rceil = \lceil \log_2 V\rceil + 2$.

Only the gate, tensor products, padding, and final multiplication add ReQU layers. If the local terminal coefficient is $c_{\mathrm{loc,term}}$, their combined terminal length is at most $c_{\mathrm{loc,term}} \log(e\bar{L}(V)) + t + 1$. Thus $c_{\mathrm{app}} = c_{\mathrm{loc,term}} + t + 1$ suffices. The constants and the fixed thresholds may depend on d, m, but no additional power of U or V enters either capacity bound. □

Lemma D.3 (Simultaneous output and first-derivative approximation)**.** *For the* single assembled network q *above,*

$$\|q - g\|_{L^\infty([0,1]^d)} \leq c_0 G(UV)^{-2m/d}, \tag{D.17}$$

$$\max_j \|\partial_j(q-g)\|_{L^\infty([0,1]^d)} \leq c_1 G(UV)^{-2(m-1)/d}, \tag{D.18}$$

where $c_0 = C_{\mathrm{loc}}$ *and* $c_1 = (1 + 8 \cdot 2^d)C_{\mathrm{loc}}$. *The architecture has bounds* (D.1)*, as verified in Lemma D.2. Its terminal-block coefficient* c_{app} *includes the fixed gate and product layers and depends only on* m, d*, independently of* U, V, G.

Proof. The cited local estimates are assembled with the weights just constructed. The partition identity controls the output error, while the product rule controls the first derivatives.

Step 1: Bound the output of the assembled approximant.

Since $\sum_\nu \rho_\nu = 1$, $q - g = \sum_\nu \rho_\nu (q_\nu - g)$. At each point, only terms on the support of their partition functions contribute. Use their local bounds in (D.13) and $\rho_\nu \geq 0$ to obtain

$$|q - g| \leq \sum_\nu \rho_\nu |q_\nu - g| \leq C_{\mathrm{loc}} G (UV)^{-2m/d} \sum_\nu \rho_\nu = C_{\mathrm{loc}} G (UV)^{-2m/d}.$$

Taking the essential supremum proves (D.17).

Step 2: Differentiate the same assembled approximant.

By the product rule for locally Lipschitz functions, almost everywhere,

$$\partial_j (q - g) = \sum_\nu (\partial_j \rho_\nu)(q_\nu - g) + \sum_\nu \rho_\nu \partial_j (q_\nu - g).$$

The partition functions vanish off their supports, and their weak derivatives vanish almost everywhere off those supports. Thus only local errors controlled by (D.13) enter. There are 2^d terms in the first sum, whereas nonnegativity and unit sum bound the second by its maximal local error. It follows that

$$\begin{aligned} |\partial_j (q - g)| &\leq 8 \cdot 2^d K_{U,V} C_{\mathrm{loc}} G (UV)^{-2m/d} + C_{\mathrm{loc}} G (UV)^{-2(m-1)/d} \\ &\leq (1 + 8 \cdot 2^d) C_{\mathrm{loc}} G (UV)^{-2(m-1)/d}, \end{aligned}$$

where the last inequality uses $K_{U,V} \leq (UV)^{2/d}$. This proves (D.18).

Step 3: Use the verified assembly budget.

Both estimates concern the single network in (D.14). Lemma D.1 supplies its weights at $K_{U,V}$, and Lemma D.2 verifies its width, depth, and terminal-block bounds. The error calculation adds no network operations. For $m = 2$, the local first-order exponent remains positive; the proof uses the local first-order estimate directly. □

D.2 Rescaling, vector outputs, and exact input saturation

Lemma D.4 (Local rescaling to a growing cube)**.** *Under Assumption 2, for $\varrho \geq 1$ and permissible U, V, there is a vector-output mixed network g_ϱ with width at most $d\bar{W}_{m,d}(U)$ and depth at most $\bar{L}(V)$ satisfying on $Q_\varrho = [-\varrho, \varrho]^d$*

$$\|g_\varrho - s^*\|_{\infty, Q_\varrho} \leq b_0 := c_0 H_0 (2\varrho)^m (UV)^{-2m/d}, \tag{D.19}$$

$$\max_{i,j} \|\partial_j (g_{\varrho,i} - s_i^*)\|_{L^\infty(Q_\varrho)} \leq b_1 := c_1 H_0 (2\varrho)^{m-1} (UV)^{-2(m-1)/d}. \tag{D.20}$$

No constant depending on ϱ is suppressed in c_0, c_1.

Proof. An affine change of variables puts the growing cube on the unit cube. Applying the scalar approximation to each coordinate and scaling derivatives back gives the stated powers of the radius.

Step 1: Compute the rescaled Sobolev norm.

For each coordinate i, define $h_i(z) = s_i^*(2\varrho z - \varrho \mathbf{1})$ on $[0,1]^d$. The weak affine chain rule gives

$$D_z^\alpha h_i(z) = (2\varrho)^{|\alpha|} (D_x^\alpha s_i^*)(2\varrho z - \varrho \mathbf{1}).$$

Since $2\varrho \geq 1$ and Assumption 2 bounds every derivative through order m, $\|h_i\|_{W^{m,\infty}([0,1]^d)} \leq H_0 (2\varrho)^m$. Set $G = H_0 (2\varrho)^m$ in Lemma D.3 and choose its network q_i.

Step 2: Transform back and track the derivative factor.

Define $g_{\varrho,i}(x) = q_i((x + \varrho \mathbf{1})/(2\varrho))$. The output norm is unaffected by the change of variables, yielding (D.19). Each first derivative receives the factor $(2\varrho)^{-1}$, so

$$\|\partial_j (g_{\varrho,i} - s_i^*)\|_{L^\infty(Q_\varrho)} \leq (2\varrho)^{-1} c_1 H_0 (2\varrho)^m (UV)^{-2(m-1)/d},$$

which is (D.20).

Step 3: Account for the vector architecture.

The unit-cube construction uses a common architecture for all target coordinates, with only coefficients changing. Put the d scalar realizations in parallel using block-diagonal hidden matrices and one d-dimensional affine output. This uses total width at most $d\bar{W}_{m,d}(U)$ and the same depth and activation phases. The affine input transformation is absorbed into the first affine layer. No factor of d is omitted from the actual width budget. □

Lemma D.5 (Global bounded extension inside the mixed class)**.** *For the network in Lemma D.4, suppose*

$$b_0 \le M - H_0, \qquad b_1 \le B_J/d - H_0. \tag{D.21}$$

Then there is $f_\varrho \in \mathcal{F}_n$ whenever

$$W_n \ge \max\{2d, d\bar{W}_{m,d}(U)\}, \qquad L_n \ge \bar{L}(V) + 1, \qquad c_{\mathrm{term}} \ge c_{\mathrm{app}}, \tag{D.22}$$

and $f_\varrho = g_\varrho$ on Q_ϱ.

Proof. We prepend coordinatewise input saturation to the local network. Pointwise output control and Lipschitz control on the closed cube preserve the global envelopes after composition.

Step 1: Specify an exact network extension.

Let $\Pi_\varrho : \mathbb{R}^d \to Q_\varrho$ be the coordinate saturation map

$$(\Pi_\varrho x)_j = -\varrho + \mathrm{ReLU}(x_j + \varrho) - \mathrm{ReLU}(x_j - \varrho). \tag{D.23}$$

It is the identity on Q_ϱ, is coordinatewise nonexpansive, and is realized by one ReLU hidden layer of width $2d$ and an affine output. Set $f_\varrho = g_\varrho \circ \Pi_\varrho$. Compose the saturation affine output with the first affine map of g_ϱ. This adds one ReLU hidden layer, preserves the terminal ReQU block, and gives exactly the capacity accounting in (D.22).

Step 2: Verify the global output envelope.

By (D.19) and Assumption 2, $|g_{\varrho,i}(x)| \le H_0 + b_0$ almost everywhere on Q_ϱ. The network and the target score have continuous representatives, so this inequality holds on its closed cube. For every $x \in \mathbb{R}^d$, $\Pi_\varrho x \in Q_\varrho$. Therefore

$$\|f_\varrho(x)\|_\infty \le H_0 + b_0 \le M.$$

Step 3: Control derivatives also on saturated faces.

On the cube, (D.20) implies $|\partial_j g_{\varrho,i}| \le H_0 + b_1$ almost everywhere. For two points of the cube differing only in coordinate j, the corresponding values of $g_{\varrho,i}$ differ by at most $(H_0 + b_1)|x_j - y_j|$. To cover lines on exceptional slices or boundary faces, prove this first on almost every interior line by absolute continuity on lines, approximate an arbitrary line by these lines, and use continuity of $g_{\varrho,i}$. The bound then holds for every such pair on the closed cube.

Varying only coordinate j before applying Π_ϱ varies only that coordinate afterward, by no larger an amount. Consequently each $f_{\varrho,i}$ is globally Lipschitz in each coordinate with constant $H_0 + b_1$. Its weak partial derivatives satisfy $|\partial_j f_{\varrho,i}| \le H_0 + b_1$ almost everywhere on all of $\mathbb{R}^d$, including regions mapped onto faces of the cube. Summing squares over d^2 entries gives

$$\left\|J_{f_\varrho}\right\|_{F,\infty} \le d(H_0 + b_1) \le B_J.$$

This argument does not incorrectly apply an almost-everywhere derivative bound through a map that sends positive-measure sets onto a null boundary. The representative convention of Lemma C.1 supplies the same envelope pointwise. Both constraints in (10) are verified, so $f_\varrho \in \mathcal{F}_n$. □

D.3 The resulting weighted approximation error

Lemma D.6 (Full-space comparison risk)**.** *Under the conditions of Lemma D.5,*

$$\|f_\varrho - s^*\|^2_{L^2(p)} \le db_0^2 + d(M + H_0)^2 \mathbb{P}(\|X\|_\infty > \varrho), \tag{D.24}$$

$$\mathfrak{A}_n \le \frac{\Lambda d}{2}\left[b_0^2 + (M + H_0)^2 \mathbb{P}(\|X\|_\infty > \varrho)\right]. \tag{D.25}$$

Proof. We split the score error inside and outside the approximation cube. The uniform local error controls the first part, and the global envelopes control the second.

Step 1: Split the population integral, not the observations.

Write $\|f_\varrho - s^*\|^2_{L^2(p)}$ as the sum of its integrals over Q_ϱ and Q_ϱ^c. On Q_ϱ, the squared Euclidean error is at most db_0^2 by (D.19), and $\int_{Q_\varrho} p \le 1$. The first integral is at most db_0^2.

Step 2: Bound the part outside the approximation cube.

On Q_ϱ^c, $\|f_\varrho - s^*\|_\infty \le M + H_0$ by the global envelopes. Thus its squared Euclidean norm is at most $d(M+H_0)^2$. Multiplying by the probability of Q_ϱ^c gives the second term in (D.24). The observations have not been truncated or discarded.

Step 3: Convert to excess Bregman risk.

Lemma D.5 supplies $f_\varrho \in \mathcal{F}_n$, hence $\mathfrak{A}_n \le \mathcal{E}_\phi(f_\varrho)$. By Proposition B.1,

$$\mathcal{E}_\phi(f_\varrho) \le \frac{\Lambda}{2} \|f_\varrho - s^*\|^2_{L^2(p)} .$$

Substituting (D.24) gives (D.25). □

D.4 Conversion from auxiliary indices to actual width and depth

For actual total width W and depth L, define

$$\Xi(W,L) = \frac{WL}{\log(eW)\log(eL)}, \qquad W, L \ge 2. \tag{D.26}$$

The output and entrywise first-derivative error bounds, with constants specified in (D.33), are

$$e_0(\varrho, W, L) = C_{\mathrm{val}} H_0 (2\varrho)^m \Xi(W,L)^{-2m/d}. \tag{D.27}$$

$$e_1(\varrho, W, L) = C_{\mathrm{der}} H_0 (2\varrho)^{m-1} \Xi(W,L)^{-2(m-1)/d}. \tag{D.28}$$

They ensure the candidate envelopes when

$$e_0(\varrho, W, L) \le M - H_0, \qquad e_1(\varrho, W, L) \le B_J/d - H_0. \tag{D.29}$$

Lemma D.7 (Full-space approximant with actual capacities)**.** *There are $C_{\mathrm{val}}, C_{\mathrm{der}}, W_0, L_0$ depending only on m, d such that, for $W \ge W_0$, $L \ge L_0$, $\varrho \ge 1$, and $c_{\mathrm{term}} \ge c_{\mathrm{app}}$, the inequalities* (D.29) *imply the existence of $f_\varrho \in \mathcal{F}_n$ with* (D.24)–(D.25)*, replacing b_0, b_1 by $e_0(\varrho, W, L), e_1(\varrho, W, L)$. Here W, L are the actual total width and depth in* (10)*.*

Proof. We choose auxiliary construction indices inside the prescribed width and depth. Rescaling and input saturation then produce a globally admissible comparator with separate value and derivative errors.

Step 1: Upper-bound the realized capacities.

For $U \ge \max\{2, d\}$ and $V \ge 2$, the formulas (D.1) imply

$$d\bar{W}_{m,d}(U) \le A_W U \log(eU), \qquad \bar{L}(V) \le A_L V \log(eV),$$

where, for example, the deliberately generous choices

$$A_W = \frac{10d\, 2^{d+6} m^{d+1}}{\log 2}, \qquad A_L = \frac{100}{\log 2} \tag{D.30}$$

suffice. The inequalities follow from $U + d \le 2U$, $V + 1 \le 2V$, and the definitions of the ceiling and logarithm. They involve only the fixed unit-cube architecture, not ϱ or H_0.

Step 2: Select auxiliary indices that fit the actual network.

Set

$$U = \left\lfloor \frac{W}{4A_W \log(eW)} \right\rfloor, \qquad V = \left\lfloor \frac{L-1}{4A_L \log(eL)} \right\rfloor. \tag{D.31}$$

A sufficient fixed choice is

$$W_0 = \max\{2d, \lceil (16A_W U_0)^2 \rceil\}, \qquad L_0 = \max\{2, \lceil (32A_L V_0)^2 \rceil\}.$$

Indeed, $\log(ex) \le 2\sqrt{x}$ for $x \ge 1$, and $L - 1 \ge L/2$ for $L \ge 2$. Hence the two floor arguments in (D.31) are at least $\sqrt{W}/(8A_W) \ge 2U_0$ and $\sqrt{L}/(16A_L) \ge 2V_0$, respectively. Thus the indices meet the fixed local-construction thresholds and $W \ge 2d$.

Since $U \le W$ and $V \le L$, the capacity upper bounds give

$$d\bar{W}_{m,d}(U) \le W/4, \qquad \bar{L}(V) + 1 \le (L-1)/4 + 1 \le L.$$

The additional width-$2d$ layer is already covered by $W \ge W_0$. Moreover, $\lfloor z \rfloor \ge z/2$ for $z \ge 2$ yields

$$U \ge \frac{W}{8A_W \log(eW)}, \qquad V \ge \frac{L}{16A_L \log(eL)}.$$

Consequently

$$UV \ge \frac{\Xi(W,L)}{A_T}, \qquad A_T = 128 A_W A_L. \tag{D.32}$$

Step 3: Track the approximation constants.

Insert (D.32) into (D.19) and (D.20). The choices

$$C_{\mathrm{val}} = c_0 A_T^{2m/d}, \qquad C_{\mathrm{der}} = c_1 A_T^{2(m-1)/d} \tag{D.33}$$

give $b_0 \le e_0$ and $b_1 \le e_1$. Conditions (D.29) imply (D.21). Lemmas D.5 and D.6 now prove the result. The factor d for parallel outputs and the extra saturation layer have been paid for before defining Ξ, rather than hidden in a change of the candidate class. □

Proposition D.1 (Full-space comparison network)**.** *Suppose Assumptions 1–3 hold, and use the unit-cube realization* (D.1)–(D.13)*. Let $W = W_n \ge W_0$, $L = L_n \ge L_0$, $c_{\mathrm{term}} \ge c_{\mathrm{app}}$ and $\varrho \ge 1$. If* (D.29) *holds, there exists a comparison network $f_\varrho \in \mathcal{F}_n$ satisfying*

$$\|f_\varrho - s^*\|_{\infty, Q_\varrho} \le e_0(\varrho, W, L), \qquad \max_{i,j} \|\partial_j (f_{\varrho,i} - s_i^*)\|_{L^\infty(Q_\varrho)} \le e_1(\varrho, W, L), \tag{D.34}$$

and

$$\mathfrak{A}_n \le \frac{\Lambda d}{2} \left[e_0(\varrho, W, L)^2 + (M + H_0)^2 \mathbb{P}(\|X\|_\infty > \varrho) \right]. \tag{D.35}$$

Proof. Lemma D.7 constructs the same comparison network with both estimates in (D.34) and the two global envelopes. Applying Lemma D.6 gives (D.35). □

D.5 Tail control supplied by the fourth assumption

Lemma D.8 (An exponential tail consequence of the local MGF bound)**.** *Assumption 4 provides fixed $a_{\mathrm{mgf}} > 0$ and $K_{\mathrm{mgf}} < \infty$ such that*

$$\mathbb{E} e^{a_{\mathrm{mgf}} X_j} \le K_{\mathrm{mgf}}, \qquad \mathbb{E} e^{-a_{\mathrm{mgf}} X_j} \le K_{\mathrm{mgf}}, \qquad j = 1, \dots, d. \tag{D.36}$$

For all $\varrho \ge 0$,

$$\mathbb{P}(\|X\|_\infty > \varrho) \le 2dK_{\mathrm{mgf}} e^{-a_{\mathrm{mgf}} \varrho}. \tag{D.37}$$

No coordinate independence is needed.

Proof. Choose a closed Euclidean ball of radius $a_{\mathrm{mgf}} > 0$ contained in the neighborhood in Assumption 4, and let K_{mgf} be the uniform MGF bound there. Each of $a_{\mathrm{mgf}} e_j$ and $-a_{\mathrm{mgf}} e_j$ belongs to the ball, proving (D.36). Exponential Markov inequalities give

$$\mathbb{P}(X_j > \varrho) \le e^{-a_{\mathrm{mgf}} \varrho} \mathbb{E} e^{a_{\mathrm{mgf}} X_j} \le K_{\mathrm{mgf}} e^{-a_{\mathrm{mgf}} \varrho}, \qquad \mathbb{P}(X_j < -\varrho) \le K_{\mathrm{mgf}} e^{-a_{\mathrm{mgf}} \varrho}.$$

The event $\{\|X\|_\infty > \varrho\}$ is contained in the union of these $2d$ coordinate events. The union bound proves (D.37). □

E PROOFS OF THE MAIN THEOREMS AND COROLLARIES

E.1 EMPIRICAL RISK DECOMPOSITION

We use the standard empirical-minimization comparison (Anthony & Bartlett, 1999). The two empirical-process deviations are controlled uniformly, while the population comparator need not attain the approximation infimum.

Lemma E.1 (Four-term excess-risk decomposition)**.** *For any estimator satisfying* (11)*,*

$$\mathcal{E}_\phi(\widehat{s}) \le 2Z(\mathcal{D}) + \mathfrak{A}_n. \tag{E.1}$$

Taking expectations gives

$$\mathbb{E}_{\mathcal{D}}\mathcal{E}_\phi(\widehat{s}) \le 2\mathfrak{S}_n + \mathfrak{A}_n. \tag{E.2}$$

Proof. We compare the empirical minimizer with an arbitrary admissible function. After adding and subtracting empirical losses, exact minimization eliminates one term and uniform deviation controls two others.

Step 1: Compare with an arbitrary admissible function.

Fix a deterministic $f \in \mathcal{F}_n$. Add and subtract its empirical risk and the empirical risk of $\widehat{s}$:

$$\begin{aligned}\mathcal{E}_\phi(\widehat{s}) = \{\mathcal{L}_\phi(\widehat{s}) - \widehat{\mathcal{L}}_\phi(\widehat{s})\} + \{\widehat{\mathcal{L}}_\phi(\widehat{s}) - \widehat{\mathcal{L}}_\phi(f)\} \\ + \{\widehat{\mathcal{L}}_\phi(f) - \mathcal{L}_\phi(f)\} + \{\mathcal{L}_\phi(f) - \mathcal{L}_\phi(s^*)\}.\end{aligned} \tag{E.3}$$

Because $\widehat{s}, f \in \mathcal{F}_n$, the first and third terms are each at most $Z(\mathcal{D})$. Equation (11) makes the second nonpositive, and the fourth is $\mathcal{E}_\phi(f)$. Thus

$$\mathcal{E}_\phi(\widehat{s}) \le 2Z(\mathcal{D}) + \mathcal{E}_\phi(f).$$

Step 2: Take the approximation infimum without assuming attainment.

The preceding inequality holds for every deterministic $f \in \mathcal{F}_n$. Taking its infimum gives (E.1). Equivalently, choose a sequence whose population excess risks decrease to $\mathfrak{A}_n$ and pass to the limit. No population minimizer in $\mathcal{F}_n$ is required. Taking the training-sample expectation and using (A.2) proves the second assertion. □

E.2 AN APPROXIMATION-AGNOSTIC ORACLE INEQUALITY

Theorem E.1 (Approximation-agnostic oracle inequality)**.** *Under Assumptions 1–3, every $\mathcal{F}_n$-valued estimator satisfying* (11) *obeys*

$$\mathbb{E}_{\mathcal{D}}\mathcal{E}_\phi(\widehat{s}) \le C_{\mathrm{sto}}\sqrt{\frac{V_n}{n}} + \mathfrak{A}_n, \tag{E.4}$$

$$\mathbb{E}_{\mathcal{D}}\,\|\widehat{s} - s^*\|^2_{L^2(p)} \le \frac{2}{\lambda}\left(C_{\mathrm{sto}}\sqrt{\frac{V_n}{n}} + \mathfrak{A}_n\right). \tag{E.5}$$

Remark E.1 (High-probability oracle inequality). Under the same conditions, for $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\mathcal{E}_\phi(\widehat{s}) \le C_{\mathrm{sto}}\sqrt{\frac{V_n}{n}} + 2B_\ell\sqrt{\frac{2\log(1/\delta)}{n}} + \mathfrak{A}_n. \tag{E.6}$$

Here B_ℓ is defined in (C.7). Multiplying the right side by $2/\lambda$ gives the corresponding score bound.

Proof of Theorem E.1. Insert Proposition C.1 into the expected inequality of Lemma E.1: $2\mathfrak{S}_n \le C_{\mathrm{sto}}\sqrt{V_n/n}$. This gives (E.4). The lower inequality in Proposition B.1, applied for each realization of $\widehat{s}$, gives (E.5) after expectation.

For the probability bound, use (C.19) in (E.1). On its event of probability at least $1 - \delta$,

$$\mathcal{E}_\phi(\widehat{s}) \le 2\mathfrak{S}_n + 2B_\ell\sqrt{2\log(1/\delta)/n} + \mathfrak{A}_n.$$

Apply Proposition C.1 once more. This is (E.6); the same pointwise quadratic conversion supplies the score version. No use of Assumption 4 occurred. □

E.3 Proof of Theorem 1

Proof. The proof first verifies the envelope conditions for the constructed comparator. Its approximation error is inserted into the oracle inequality, after which the architectural pseudo-dimension bound gives the stated width–depth form.

Step 1: Recover the construction thresholds and envelope margins.

Let $C_{\mathrm{val}}, C_{\mathrm{der}}, W_0, L_0$ be the constants in Lemma D.7, and take

$$c_{\mathrm{cap}} = \max\Bigg\{1, \frac{W_0}{\log(eW_0)}, \frac{L_0}{\log(eL_0)}, 2^{d/4}\left(\frac{C_{\mathrm{val}}H_0}{M-H_0}\right)^{d/(4m)}, 2^{d/4}\left(\frac{C_{\mathrm{der}}H_0}{B_J/d-H_0}\right)^{d/[4(m-1)]}\Bigg\}. \tag{E.7}$$

Both envelope margins are positive. The map $x \mapsto x/\log(ex)$ is increasing for $x \geq 1$, since its derivative is $\log x/[\log(ex)]^2$. Thus (13) and $\varrho \geq 1$ imply $W_n \geq W_0$ and $L_n \geq L_0$. With $W = W_n$ and $L = L_n$, multiplying the two lower bounds in (13) gives

$$\Xi(W, L) \geq c_{\mathrm{cap}}^2 \varrho^{d/2}.$$

The fourth term in the maximum in (E.7) therefore yields

$$\Xi(W, L)^{-2m/d} \leq (2\varrho)^{-m}\frac{M-H_0}{C_{\mathrm{val}}H_0},$$

and substitution into (D.27) gives $e_0 \leq M - H_0$. The fifth term similarly gives $e_1 \leq B_J/d - H_0$ through (D.28). Hence (D.29) holds. Together with the recovered width and depth thresholds and the standing terminal-layer convention, Lemma D.7 supplies $f_\varrho \in \mathcal{F}_n$ with both required global envelopes.

Step 2: Bound the deterministic excess risk.

Lemma D.6, with b_0 replaced by e_0, yields

$$\begin{aligned}\mathfrak{A}_n &\leq \frac{\Lambda d}{2}\left[e_0(\varrho, W, L)^2 + (M+H_0)^2\mathbb{P}(\|X\|_\infty > \varrho)\right] \\ &= \frac{\Lambda d}{2}C_{\mathrm{val}}^2 H_0^2 (2\varrho)^{2m}\Xi(W, L)^{-4m/d} + \frac{\Lambda d}{2}(M+H_0)^2\mathbb{P}(\|X\|_\infty > \varrho).\end{aligned}$$

Only the output error is squared here. The derivative error has been used to establish $f_\varrho \in \mathcal{F}_n$.

Step 3: Combine approximation and stochastic errors.

Substituting the preceding estimate into (E.4) gives the intermediate excess-risk bound

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\mathcal{E}_\phi(\widehat{s}) \leq C_{\mathrm{sto}}\sqrt{\frac{V_n}{n}} &+ \frac{\Lambda d}{2}C_{\mathrm{val}}^2 H_0^2 (2\varrho)^{2m}\Xi(W, L)^{-4m/d} \\ &+ \frac{\Lambda d}{2}(M+H_0)^2\mathbb{P}(\|X\|_\infty > \varrho).\end{aligned} \tag{E.8}$$

Multiplying by $2/\lambda$ and using (B.6) yields

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\|\widehat{s} - s^*\|_{L^2(p)}^2 \leq \frac{2C_{\mathrm{sto}}}{\lambda}\sqrt{\frac{V_n}{n}} &+ \frac{\Lambda d}{\lambda}C_{\mathrm{val}}^2 H_0^2 (2\varrho)^{2m}\Xi(W, L)^{-4m/d} \\ &+ \frac{\Lambda d}{\lambda}(M+H_0)^2\mathbb{P}(\|X\|_\infty > \varrho).\end{aligned} \tag{E.9}$$

Step 4: Substitute the architectural capacity and collect fixed constants.

Lemma C.2 permits a common bound $V_n \le C_{\dim} W^2 L^2 \log(eWL)$. Thus the first term in (E.9) is at most

$$\frac{2C_{\mathrm{sto}}\sqrt{C_{\dim}}}{\lambda} WL\sqrt{\frac{\log(eWL)}{n}}.$$

Choose

$$C_{\mathrm{main}} = \max\left\{\frac{2C_{\mathrm{sto}}\sqrt{C_{\dim}}}{\lambda},\ \frac{\Lambda d}{\lambda}C_{\mathrm{val}}^2 H_0^2 2^{2m},\ \frac{\Lambda d}{\lambda}(M+H_0)^2\right\}. \tag{E.10}$$

Expanding Ξ by (D.26) and writing $(2\varrho)^{2m} = 2^{2m}\varrho^{2m}$ gives (14), with C_{main} as a permissible hidden multiplier. The dependencies of c_{cap} and C_{main} follow from (E.7), (E.10), Lemmas C.2 and D.7, and Proposition C.1. □

Remark E.2 (High-probability form of the main bound). For $0 < \delta < 1$, (E.8) holds with probability at least $1-\delta$ after removing $\mathbb{E}_{\mathcal{D}}$ on the left and adding $2B_\ell\sqrt{2\log(1/\delta)/n}$ to the right side. For the score bound (E.9), the corresponding added term is $(4B_\ell/\lambda)\sqrt{2\log(1/\delta)/n}$. Both statements follow by the same comparator substitution into (E.6).

E.4 Proofs of Corollaries 1 and 2

Proof. We choose a logarithmic approximation radius to control the tail. The polynomial network schedule makes both admissibility errors vanish and balances the stochastic and output-approximation powers of the sample size.

Set $\zeta = 2m/(d+4m)$, choose $a_{\mathrm{mgf}}, K_{\mathrm{mgf}}$ as in Lemma D.8, and use the network schedule (15). For the proof introduce

$$\kappa = \frac{d}{4d+16m}, \qquad W_n = L_n = \lceil n^\kappa \rceil, \qquad \varrho_n = \max\left\{1, \frac{2}{a_{\mathrm{mgf}}}\log(en)\right\}. \tag{E.11}$$

The radius and the exponent κ are auxiliary to the proof. All constants below are independent of n and have the dependence specified in the corollary.

Step 1: Relate the actual capacities to their effective product.

Since $W_n = L_n = \lceil n^\kappa \rceil$, for $n \ge 3$,

$$W_n = L_n \asymp n^\kappa, \qquad \log(eW_n) = \log(eL_n) \lesssim \log(en).$$

The hidden constants here depend only on d, m. It follows that

$$\Xi(W_n, L_n) \gtrsim \frac{n^{2\kappa}}{[\log(en)]^2}. \tag{E.12}$$

Also $1 \le \varrho_n \lesssim \log(en)$ and, by the definition of ϱ_n, $e^{-a_{\mathrm{mgf}}\varrho_n} \le (en)^{-2}$.

Step 2: Verify both admissibility margins for finite networks.

Equations (D.27), (D.28), and (E.12) give

$$e_0(\varrho_n, W_n, L_n) \lesssim n^{-4m\kappa/d}[\log(en)]^{\,m+4m/d}, \tag{E.13}$$

$$e_1(\varrho_n, W_n, L_n) \lesssim n^{-4(m-1)\kappa/d}[\log(en)]^{\,m-1+4(m-1)/d}. \tag{E.14}$$

Because $\kappa > 0$ and $m \ge 2$, both powers of n are negative. For any $c > 0$ and finite b, $n^{-c}(\log(en))^b \to 0$; this follows by setting $t = \log n$ and comparing t^b with e^{ct}. Therefore (E.13) and (E.14) are eventually smaller than the fixed positive quantities $M - H_0$ and $B_J/d - H_0$.

The balanced schedule also satisfies the capacity condition in Theorem 1. Indeed,

$$\frac{1}{\varrho_n^{d/4}}\min\left\{\frac{W_n}{\log(eW_n)}, \frac{L_n}{\log(eL_n)}\right\} \gtrsim \frac{n^\kappa}{[\log(en)]^{1+d/4}} \longrightarrow \infty.$$

Choose n_0 so that (13) and both envelope inequalities hold for every $n \ge n_0$. The terminal-block coefficient was fixed before Theorem 1; its sufficient value is recorded in Lemma D.2. These choices verify all conditions needed to apply the theorem.

Step 3: Bound the stochastic term and compute its power.

From (C.4),

$$V_n \lesssim n^{4\kappa}\log(en), \qquad \sqrt{V_n/n} \lesssim n^{2\kappa-1/2}[\log(en)]^{1/2}.$$

The chosen exponent obeys

$$\frac{1}{2} - 2\kappa = \frac{1}{2} - \frac{d}{2(d+4m)} = \frac{2m}{d+4m} = \zeta, \qquad \frac{8m\kappa}{d} = \frac{2m}{d+4m} = \zeta. \tag{E.15}$$

Thus the stochastic term is $\lesssim n^{-\zeta}[\log(en)]^{1/2}$.

Step 4: Bound local approximation and tails.

Squaring (E.13) and using the second equality in (E.15) gives

$$e_0(\varrho_n, W_n, L_n)^2 \lesssim n^{-\zeta}[\log(en)]^{2m+8m/d}.$$

Lemma D.8 yields

$$\mathbb{P}(\|X\|_\infty > \varrho_n) \le 2dK_{\mathrm{mgf}}(en)^{-2}.$$

Since $0 < \zeta < 1/2$ and $2m + 8m/d > 1/2$, the tail and stochastic terms are bounded by a constant times the displayed local-approximation rate for $n \ge 3$.

Step 5: Insert the three terms into the theorem.

Applying (E.9) proves the squared score bound (16). The intermediate excess-risk inequality (E.8) also gives

$$\mathbb{E}_{\mathcal{D}}\mathcal{E}_\phi(\widehat{s}) \lesssim n^{-\frac{2m}{d+4m}}[\log(en)]^{2m+8m/d}. \tag{E.16}$$

By the population identity (B.5), this is (17). The excess-risk calculation uses (E.8) directly, so its multiplier does not contain the strong-convexity conversion factor $2/\lambda$. Uniformity over a family of densities requires a common neighborhood and bound in Assumption 4; these supply the fixed $a_{\mathrm{mgf}}, K_{\mathrm{mgf}}$ used in this proof. □

Remark E.3 (An alternative explicit logarithmic prefactor). The auxiliary-parameter form also allows $U_n = V_n^{\mathrm{aux}} = \lceil n^\kappa \rceil$, where V_n^{aux} is an approximation index, not the capacity V_n. Choose actual width $\max\{2d, d\bar{W}_{m,d}(U_n)\}$ and depth $\bar{L}(V_n^{\mathrm{aux}}) + 1$. Both are $O(n^\kappa \log(en))$. With the same ϱ_n, the estimator has the bound

$$\mathbb{E}_{\mathcal{D}}\|\widehat{s} - s^*\|_{L^2(p)}^2 \lesssim n^{-\zeta}[\log(en)]^{2m}. \tag{E.17}$$

Indeed, (D.19) gives squared approximation $\lesssim n^{-8m\kappa/d}(\log(en))^{2m}$; the actual-capacity bound gives $\sqrt{V_n/n} \lesssim n^{2\kappa-1/2}(\log(en))^{5/2}$. Since $m \ge 2$, $2m \ge 4 > 5/2$, and the same margin and tail arguments apply. This optional schedule pays an extra logarithm in network size. Corollary 1 retains the original convention of equal polynomial-order actual width and depth.

E.5 The conservative upper-bound calculation

Remark E.4 (Conservative rate corresponding to the original proof). Bounding the output error by the uniform $W^{1,\infty}$ error yields a squared approximation term of order $H_0^2(2\varrho)^{2m}\Xi(W,L)^{-4(m-1)/d}$. Under the same assumptions and terminal-block condition, take $W_n = L_n = \lceil n^{d/[4d+16(m-1)]} \rceil$ and keep ϱ_n from (E.11). For sufficiently large n,

$$\mathbb{E}_{\mathcal{D}}\|\widehat{s} - s^*\|_{L^2(p)}^2 \lesssim n^{-\frac{2(m-1)}{d+4(m-1)}}[\log(en)]^{2m+8(m-1)/d}. \tag{E.18}$$

This secondary result recovers the original polynomial exponent. The refined exponent in Corollary 1 comes from retaining the output bound (D.27), while using (D.28) for admissibility. Both analyses use the same estimator family.

Proof. We reuse the same admissible comparator with a coarser Sobolev error bound, then balance that approximation error against the unchanged stochastic term.

Step 1: Recover the original uniform-Sobolev upper bound.

Lemma D.3 implies, with $c_1 \geq c_0$ and $UV \geq 1$,

$$\|q - g\|_{W^{1,\infty}([0,1]^d)} \leq c_1 G (UV)^{-2(m-1)/d}.$$

For the output we deliberately use this larger error rather than (D.17). Rescaling with $G = H_0(2\varrho)^m$ and using (D.32) gives

$$e_{W^1} = C_{\mathrm{der}} H_0 (2\varrho)^m \Xi(W, L)^{-2(m-1)/d}$$

as a common upper bound for the local coordinatewise output error and, since $\varrho \geq 1$, for the first-derivative error as well. The same saturation extension applies when this bound is below the two fixed admissibility margins. Lemma D.6 then yields an approximation term bounded by a constant times $e_{W^1}^2$ plus the unchanged tail probability.

Step 2: Balance at the original polynomial exponent.

Put $r = m - 1 > 0$, $\kappa_0 = d/(4d + 16r)$, and $\zeta_0 = 2r/(d + 4r)$. With $W_n = L_n = \lceil n^{\kappa_0} \rceil$, the same actual-capacity calculation gives

$$\sqrt{V_n/n} \lesssim n^{-\zeta_0} [\log(en)]^{1/2},$$

where $1/2 - 2\kappa_0 = \zeta_0$. With $\varrho_n \asymp \log n$,

$$e_{W^1}^2 \lesssim n^{-8r\kappa_0/d} [\log(en)]^{2m+8r/d} = n^{-\zeta_0} [\log(en)]^{2m+8r/d}.$$

The common error tends to zero, ensuring both margins for sufficiently large n, and the tail remains $O(n^{-2})$. Insert these three bounds into Theorem E.1. This proves (E.18). □

E.6 Comparison of the two upper-bound exponents

The simultaneous approximation in Lemma D.3 uses the same regularity order m as Assumption 2. Derivative approximation uses the exponent $m - 1$, which suffices for admissibility. The excess-risk approximation uses the exponent m, because (B.5) depends only on output values. Thus the refinement does not assume an extra derivative of s^* or a growing output/Jacobian envelope.

For $d \geq 1$ and $m \geq 2$, the difference between the two leading polynomial exponents is

$$\frac{2m}{d + 4m} - \frac{2(m-1)}{d + 4(m-1)} = \frac{2d}{(d + 4m)(d + 4m - 4)} > 0.$$

This compares the two upper-bound derivations. It is not a lower-bound argument and does not establish that a general Bregman estimator is statistically superior to a quadratic one. The curvature constants, and hence finite-sample constants, still depend on the chosen generator.

F Quadratic score bounds under finite Fisher information

This appendix proves Theorem 2 and Corollary 3, and verifies Remark 5. The population identity, comparison bound, and pointwise consistency require only Assumption 1 and (18). The explicit approximation bound also requires the local Sobolev regularity in Assumption 5. The rate uses the Gaussian-mixture structure in Corollary 3; neither Assumption 2 nor Assumption 4 is imposed on this larger class. Throughout this appendix, $A = A^\top \succ 0$ is fixed, $\phi_A(u) = u^\top A u/2$, and

$$\lambda_A = \lambda_{\min}(A), \qquad \Lambda_A = \lambda_{\max}(A).$$

We use the same selected weak Jacobians as in Appendix C.1, with the guard threshold $B_{J,n}$. All envelope sequences are deterministic. Constants hidden by $\lesssim$ in the oracle calculation depend only on d, A, c_{term}, not on $M_n, B_{J,n}, W_n, L_n, n$, or the density. Additional dependencies in distribution-specific bounds are stated separately.

F.1 Population identity with finite Fisher information

Lemma F.1 (Quadratic identity under a score second moment)**.** *Under Assumption 1 and* (18)*, every bounded $f \in W^{1,\infty}(\mathbb{R}^d; \mathbb{R}^d)$ satisfies*

$$\mathcal{L}_{\phi_A}(f) = \frac{1}{2}\mathbb{E}\big[(f(X) - s^*(X))^\top A (f(X) - s^*(X))\big] - \frac{1}{2}\mathbb{E}\big[s^*(X)^\top A s^*(X)\big]. \tag{F.1}$$

Every term in this identity is finite. No differentiability or integrability of J_{s^*} is required.

Proof. We adapt the quadratic score-matching identity of Hyvärinen (2005, Theorem 1). Finite score energy supplies integrability for the full-space cutoff argument, and completing the square retains the additive constant.

Step 1: Verify the integration identity.

The relation $\nabla p = ps^*$ and Cauchy–Schwarz give

$$\int_{\mathbb{R}^d} \|\nabla p(x)\|_2 \, dx = \mathbb{E}\,\|s^*(X)\|_2 \le I(p)^{1/2}. \tag{F.2}$$

For each fixed bounded candidate, $|\operatorname{Tr}(AJ_f)| \le \Lambda_A\sqrt{d}B_{J,n}$ and $|(Af)^\top s^*| \le \Lambda_A\sqrt{d}M_n\,\|s^*\|_2$. Both are integrable under p. Repeat the weak product-rule and cutoff calculation in Lemma B.3 with $v_f = Af$. Its boundary remainder is at most $\Lambda_A\sqrt{d}M_n\,\|\nabla\chi\|_\infty/T$, so dominated convergence gives

$$\mathbb{E}\operatorname{Tr}(AJ_f(X)) = -\mathbb{E}\big[s^*(X)^\top Af(X)\big]. \tag{F.3}$$

The limit is taken for each fixed n, so no uniform bound on a growing envelope is required.

Step 2: Complete the square.

Insert (F.3) into $\ell_{\phi_A}(f;x) = f(x)^\top Af(x)/2 + \operatorname{Tr}(AJ_f(x))$ and expand $(f-s^*)^\top A(f-s^*)$. This proves (F.1); its two quadratic expectations are finite by boundedness of f and (18). No loss or derivative of the true score is evaluated. □

F.2 Uniform stochastic control with growing envelopes

Lemma F.2 (Quadratic loss fluctuation)**.** *For the class $\mathcal{F}_n$, write*

$$Z_n = \sup_{f\in\mathcal{F}_n} \big|\widehat{\mathcal{L}}_{\phi_A}(f) - \mathcal{L}_{\phi_A}(f)\big|.$$

Then

$$\mathbb{E}_{\mathcal{D}} Z_n \lesssim (M_n^2 + B_{J,n})W_nL_n\sqrt{\frac{\log(eW_nL_n)}{n}}. \tag{F.4}$$

This bound holds for any sampling distribution, with a constant depending only on d, A, c_{term}.

Proof. We retain the architectural pseudo-dimension bound and normalize the output and Jacobian covers by their growing envelopes. The same chaining argument then controls the quadratic loss class.

Step 1: Retain the architecture bound.

The polynomial sign-pattern argument in Lemma C.2 gives a common pseudo-dimension bound for each output coordinate and selected Jacobian entry:

$$V_n \lesssim W_n^2L_n^2\log(eW_nL_n).$$

The values of the envelope thresholds do not affect the polynomial degrees, parameter counts, or number of comparisons. Dividing an output by M_n, or a Jacobian entry by $B_{J,n}$, preserves its pseudo-dimension. The normalized scalar classes therefore have envelope one and pseudo-dimension at most V_n.

Step 2: Normalize the covers before forming the loss class.

Set

$$b_n = \Lambda_A\left(\frac{d}{2}M_n^2 + \sqrt{d}B_{J,n}\right). \tag{F.5}$$

The bound $|\ell_{\phi_A}(f;x)| \le b_n$ follows from (5) and the Frobenius Cauchy–Schwarz inequality. For $f, g \in \mathcal{F}_n$,

$$|\ell_{\phi_A}(f;x) - \ell_{\phi_A}(g;x)| \le \Lambda_A\sqrt{d}M_n\,\|f(x)-g(x)\|_2 + \Lambda_A\sqrt{d}\,\|J_f(x)-J_g(x)\|_F. \tag{F.6}$$

Indeed, the first term follows from $f^\top Af - g^\top Ag = (f-g)^\top A(f+g)$, and the second follows from linearity of the trace.

For an arbitrary probability measure Q, cover each normalized scalar class in $L^2(Q)$ with radius $\eta \le 1$. As in Lemma C.5, select an actual network from each occupied combination of cover indices. Two candidates assigned to the same combination then satisfy

$$\big\|\|f-g\|_2\big\|_{L^2(Q)} \le 2\sqrt{d}M_n\eta, \qquad \big\|\|J_f - J_g\|_F\big\|_{L^2(Q)} \le 2dB_{J,n}\eta.$$

Minkowski's inequality and (F.6) give a loss distance at most

$$2\Lambda_A\left(dM_n^2 + d\sqrt{d}B_{J,n}\right)\eta \le 4d\,b_n\eta.$$

For $0 < \varepsilon \le b_n$, choose $\eta = \varepsilon/(4db_n)$. Lemma C.4, applied to the $d + d^2$ normalized classes and then to their product cover, yields

$$\log N\big(\varepsilon, \mathcal{G}_n^{\pm}, L^2(Q)\big) \le [4(d+d^2)+2]V_n \log\left(\frac{64ed\,b_n}{\varepsilon}\right), \tag{F.7}$$

where $\mathcal{G}_n^{\pm}$ is the quadratic loss class together with its negative and zero. The added factor for these two operations is absorbed by the final $+2$ in the coefficient. In particular, the logarithm contains no envelope ratio that could hide an extra dependence on n.

Step 3: Chain and symmetrize.

Apply the conditional chaining proof of Lemma C.6 to (F.7). Its small-scale integral satisfies

$$\begin{aligned}\int_0^{b_n}\sqrt{\log(64ed\,b_n/u)}\,du &= b_n\int_0^1\sqrt{\log(64ed)+\log(1/v)}\,dv\\ &\le b_n\left[\sqrt{\log(64ed)}+\frac{\sqrt{\pi}}{2}\right].\end{aligned}$$

The inequality uses $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ and $\int_0^1\sqrt{\log(1/v)}\,dv = \Gamma(3/2)$. The conditional Rademacher average is consequently at most a constant depending on d times $b_n\sqrt{V_n/n}$. Symmetrization, as in the proof of Proposition C.1, multiplies this by two. Inserting (F.5) and the capacity bound from Step 1 proves (F.4). □

F.3 A COMPARISON BOUND AND CONSISTENCY UNDER FINITE SCORE ENERGY

Lemma F.3 (Comparison inequality and pointwise consistency)**.** *Suppose Assumption 1 and* (18) *hold. For $n \ge 2$, $W_n \ge d$, and $L_n \ge 2$, every deterministic $f \in \mathcal{F}_n$ and every measurable exact empirical minimizer satisfy*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\,\|\widehat{s}-s^*\|_{L^2(p)}^2 \lesssim (M_n^2+B_{J,n})W_nL_n\sqrt{\frac{\log(eW_nL_n)}{n}}\\ +\|f-s^*\|_{L^2(p)}^2.\end{aligned} \tag{F.8}$$

The hidden constant depends only on d, A, c_{term}. If all four sequences $M_n, B_{J,n}, W_n, L_n$ diverge and the sampling term tends to zero, the mean-squared error tends to zero for each fixed density. Local Sobolev regularity is not needed for this qualitative statement.

Proof. The quadratic identity compares the estimator with any deterministic candidate. For consistency, smooth compactly supported fields approximate a finite-energy score, and fixed mixed-network approximants eventually belong to the expanding candidate class.

Step 1: Compare the empirical estimator with an arbitrary candidate.

For any deterministic $f \in \mathcal{F}_n$, exact empirical minimization gives

$$\mathcal{L}_{\phi_A}(\widehat{s}) \le \mathcal{L}_{\phi_A}(f) + 2Z_n.$$

This follows by adding and subtracting the two empirical risks and bounding each empirical-minus-population difference by Z_n. Lemma F.1 cancels the common finite constant and yields

$$\begin{aligned}&\mathbb{E}\big[(\widehat{s}(X)-s^*(X))^\top A(\widehat{s}(X)-s^*(X))\mid\mathcal{D}\big]\\ &\quad\le \mathbb{E}\big[(f(X)-s^*(X))^\top A(f(X)-s^*(X))\big]+4Z_n.\end{aligned}$$

The eigenvalue inequalities $\lambda_A I_d \preceq A \preceq \Lambda_A I_d$, followed by expectation over $\mathcal{D}$, give

$$\mathbb{E}_{\mathcal{D}} \|\widehat{s} - s^*\|^2_{L^2(p)} \leq \frac{\Lambda_A}{\lambda_A} \|f - s^*\|^2_{L^2(p)} + \frac{4}{\lambda_A} \mathbb{E}_{\mathcal{D}} Z_n. \tag{F.9}$$

The candidate f is arbitrary and need not minimize a population criterion. Substitution of (F.4) proves (F.8).

Step 2: Approximate a finite-energy score by a smooth compactly supported field.

Fix one density satisfying the lemma and a tolerance $\delta > 0$. The function $s^* = \nabla p/p$ is continuous by Assumption 1. Multiplying it by smooth cutoffs converging pointwise to one gives compactly supported continuous fields converging to s^* in $L^2(p)$, because the squared error is bounded by the integrable function $\|s^*\|_2^2$. A compactly supported continuous field is uniformly continuous on $\mathbb{R}^d$. Convolution with a smooth compactly supported approximate identity therefore converges uniformly and hence in $L^2(p)$, since p integrates to one. Select a resulting $g \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$ with

$$\|g - s^*\|_{L^2(p)} < \delta.$$

Step 3: Place a bounded mixed approximation in the expanding class.

Choose ϱ so that $\operatorname{supp} g$ lies strictly inside $[-\varrho, \varrho]^d$. Applying Shen et al. (2023, Theorem 5, with RePU order two) to each smooth coordinate on this compact cube, and running the resulting networks in parallel, gives a finite ReQU network q_δ with $\|q_\delta - g\|_\infty \leq \delta/\sqrt{d}$. Only existence at fixed accuracy is needed here; no width–depth rate is transferred from this result. Prepend the ReLU input saturation

$$f_\delta(x) = q_\delta(\Pi_\varrho x), \qquad (\Pi_\varrho x)_j = -\varrho + \mathrm{ReLU}(x_j + \varrho) - \mathrm{ReLU}(x_j - \varrho).$$

This preserves the ReLU-then-ReQU order. Since g vanishes near the cube boundary, $g(\Pi_\varrho x) = g(x)$ on all of $\mathbb{R}^d$. Consequently

$$\|f_\delta - g\|_\infty \leq \delta/\sqrt{d}, \qquad \|f_\delta - g\|_{L^2(p)} \leq \delta.$$

The finite ReQU network and its first derivatives are bounded on the cube. Coordinatewise saturation is Lipschitz, so f_δ has finite global output and weak-Jacobian envelopes.

The network f_δ has a fixed finite width, depth, and number of terminal ReQU layers. Because $M_n, B_{J,n}, W_n, L_n \to \infty$ and $c_{\mathrm{term}} > 0$, each fixed requirement is eventually permitted by $\mathcal{F}_n$; in particular its terminal depth is eventually at most $c_{\mathrm{term}} \log(eL_n)$. For all sufficiently large n,

$$\|f_\delta - s^*\|^2_{L^2(p)} \leq 4\delta^2.$$

Use $f = f_\delta$ in (F.8), let $n \to \infty$, and then let $\delta \downarrow 0$. This proves consistency for the fixed density. This argument is pointwise in p and supplies no density-uniform approximation rate from (18) alone. □

For example, the deterministic choice

$$M_n = B_{J,n} = \log(en), \qquad W_n = \max\{d, \lceil n^{1/8} \rceil\}, \qquad L_n = \max\{2, \lceil n^{1/8} \rceil\} \tag{F.10}$$

makes the stochastic term $O(n^{-1/4}[\log(en)]^{5/2})$. All four sequences diverge, so Lemma F.3 gives consistency. This order describes only the stochastic contribution; the approximation term need not have a prescribed power rate for a general finite-Fisher-information density.

F.4 Explicit local approximation and proof of Theorem 2

The finite-energy comparison does not specify how fast a given score can be approximated. Local Sobolev control supplies that information through the construction in Appendix D. Only the score on the approximation cube is used to build the network; outside it, the error is controlled by the actual score energy.

Lemma F.4 (A local comparison network for an unbounded score)**.** *Let $m \geq 2$, $\varrho \geq 1$, and $s^* \in W^{m,\infty}_{\mathrm{loc}}(\mathbb{R}^d;\mathbb{R}^d)$. Use the local bound H_ϱ defined in Section 3.2. Under the architecture and envelope conditions of Theorem 2, there is a deterministic $f_\varrho \in \mathcal{F}_n$ such that*

$$\|f_\varrho - s^*\|_{\infty,[-\varrho,\varrho]^d} \leq C_{\mathrm{val}} H_\varrho (2\varrho)^m \Xi(W_n, L_n)^{-2m/d}, \tag{F.11}$$

where Ξ is defined in (D.26) *and C_{val} is the same fixed constant as in* (D.27)*. If* (18) *holds, then*

$$\begin{aligned}\|f_\varrho - s^*\|^2_{L^2(p)} \leq\ & dC^2_{\mathrm{val}} H^2_\varrho (2\varrho)^{2m} \Xi(W_n, L_n)^{-4m/d} \\ & + 2dM_n^2 \mathbb{P}(\|X\|_\infty > \varrho) + 2\mathbb{E}\Big[\|s^*(X)\|_2^2 \mathbf{1}\{\|X\|_\infty > \varrho\}\Big].\end{aligned} \tag{F.12}$$

All implementation constants depend only on d, m.

Proof. The local construction in Lemmas D.3 and D.7 is homogeneous in the target Sobolev bound. We apply it with H_ϱ in place of H_0, check the envelopes after saturation, and split the weighted error at the cube boundary.

Step 1: Rescale the local target and fit it within the architecture.

Set $Q_\varrho = [-\varrho, \varrho]^d$. For $h(z) = s^*(2\varrho z - \varrho \mathbf{1})$, the affine weak chain rule gives $\|h\|_{W^{m,\infty}([0,1]^d)} \leq H_\varrho (2\varrho)^m$. The proof of Lemma D.4 uses only this local bound. Its vector construction and the conversion in Lemma D.7 give a network g_ϱ with the reserved saturation layer included in the width–depth budget, and errors

$$\|g_\varrho - s^*\|_{\infty,Q_\varrho} \leq C_{\mathrm{val}} H_\varrho (2\varrho)^m \Xi(W_n, L_n)^{-2m/d},$$

$$\max_{i,j} \|\partial_j (g_{\varrho,i} - s^*_i)\|_{\infty,Q_\varrho} \leq C_{\mathrm{der}} H_\varrho (2\varrho)^{m-1} \Xi(W_n, L_n)^{-2(m-1)/d}.$$

The same constants $C_{\mathrm{val}}, C_{\mathrm{der}}$ and W_0, L_0 apply; no bound outside Q_ϱ is used. A sufficient capacity threshold is

$$c_{\mathrm{size}} = 2^{d/2} \max\left\{1, C_{\mathrm{val}}^{d/(2m)}, C_{\mathrm{der}}^{d/(2(m-1))}\right\}. \tag{F.13}$$

Substituting (19) makes both errors at most H_ϱ.

Step 2: Extend the network and verify membership.

Let $f_\varrho = g_\varrho \circ \Pi_\varrho$, with Π_ϱ from (D.23). On the cube, $\|s^*\|_\infty \leq H_\varrho$ and every first partial derivative has magnitude at most H_ϱ. Thus $\|f_\varrho\|_\infty \leq 2H_\varrho \leq M_n$. The coordinatewise Lipschitz argument on closed faces in Lemma D.5 gives $\|J_{f_\varrho}\|_{F,\infty} \leq 2dH_\varrho \leq B_{J,n}$, including regions saturated onto the cube boundary. The selected representative uses the same guard convention. The saturation adds one ReLU layer before the original mixed network, so the activation order and the already reserved width–depth budget are preserved. Hence $f_\varrho \in \mathcal{F}_n$ and (F.11) holds.

Step 3: Integrate the local and tail errors.

On Q_ϱ, the squared Euclidean error is at most d times the squared bound in (F.11), and $p(Q_\varrho) \leq 1$. On the complement, use

$$\|f_\varrho(x) - s^*(x)\|_2^2 \leq 2\|f_\varrho(x)\|_2^2 + 2\|s^*(x)\|_2^2 \leq 2dM_n^2 + 2\|s^*(x)\|_2^2.$$

Adding the two integrals gives (F.12). Finite Fisher information makes the score-tail integral finite without a global score envelope. □

Proof of Theorem 2. Use the deterministic network $f = f_\varrho$ supplied by Lemma F.4 in the comparison inequality (F.8). Substituting (F.12) and the definition of Ξ gives (20). The constants from local approximation depend on d, m; the quadratic risk conversion depends on A. The fixed terminal-layer coefficient was chosen in terms of d, m, so it introduces no additional varying parameter. All dependence on H_ϱ, M_n, $B_{J,n}$, and ϱ is displayed. The proof uses no assertion that finite score energy implies local Sobolev regularity. □

F.5 GAUSSIAN-MIXTURE REGULARITY AND PROOF OF COROLLARY 3

Lemma F.5 (Local smoothness and score tails of a finite Gaussian mixture)**.** *Let $p = \sum_{k=1}^{K_{\mathrm{mix}}} \pi_k p_k$, where p_k is the $N(\mu_k, \Sigma_k)$ density, $\pi_k > 0$, $\sum_k \pi_k = 1$, and $\Sigma_k \succ 0$. For each fixed integer $m \geq 2$, p has finite Fisher information and*

$$H_\varrho \lesssim (1+\varrho)^{m+1}, \qquad \varrho \geq 1. \tag{F.14}$$

There are fixed $a_{\mathrm{G}} > 0$ and $\varrho_0 \geq 1$ such that, for $\varrho \geq \varrho_0$,

$$\mathbb{P}(\|X\|_\infty > \varrho) + \mathbb{E}\Big[\|s^*(X)\|_2^2 \, \mathbf{1}\{\|X\|_\infty > \varrho\}\Big] \lesssim e^{-a_{\mathrm{G}}\varrho^2}. \tag{F.15}$$

The constants in (F.14) *depend on d, m and the fixed mixture parameters. The tail constants depend only on the fixed mixture parameters and d.*

Proof. The score is a posterior-weighted average of affine component scores. Ratios of Gaussian derivatives give polynomial bounds for its higher derivatives, and Gaussian coordinate tails control the weighted tail energy.

Step 1: Bound the score and its second moment.

Write $s_k(x) = -\Sigma_k^{-1}(x - \mu_k)$ and $w_k(x) = \pi_k p_k(x)/p(x)$. Differentiating the mixture density yields $s^*(x) = \sum_k w_k(x) s_k(x)$. Thus $\|s^*(x)\|_2 \lesssim 1 + \|x\|_\infty$. Conditional Jensen's inequality gives

$$I(p) \leq \sum_k \pi_k \mathbb{E}_{p_k} \|s_k(X)\|_2^2 = \sum_k \pi_k \operatorname{Tr}(\Sigma_k^{-1}) < \infty.$$

Step 2: Bound the local Sobolev norm.

For each multi-index α, $D^\alpha p_k = p_k P_{k,\alpha}$, where $P_{k,\alpha}$ is a polynomial of degree at most $|\alpha|$. This follows from $P_{k,0} = 1$ and the recursion

$$P_{k,\alpha+e_i} = \partial_i P_{k,\alpha} + s_{k,i} P_{k,\alpha},$$

since $s_{k,i}$ is affine. Set $r_\alpha = (D^\alpha p)/p$. Then

$$r_\alpha(x) = \sum_k w_k(x) P_{k,\alpha}(x), \qquad |r_\alpha(x)| \lesssim (1 + \|x\|_\infty)^{|\alpha|}, \qquad \partial_i r_\alpha = r_{\alpha+e_i} - r_\alpha r_{e_i}.$$

Starting with $\partial_i \log p = r_{e_i}$, the last identity shows inductively that a derivative of $\log p$ of total order $q \geq 1$ is a finite linear combination of products $\prod_\ell r_{\alpha_\ell}$ with $\sum_\ell |\alpha_\ell| = q$. Each product is bounded by a constant times $(1 + \|x\|_\infty)^q$. Therefore

$$|D^\alpha s_j^*(x)| \lesssim (1 + \|x\|_\infty)^{|\alpha|+1}, \qquad |\alpha| \leq m.$$

Taking maxima on $[-\varrho, \varrho]^d$ proves (F.14). This argument allows different covariance matrices; no common-covariance assumption is imposed.

Step 3: Control the score-weighted tail.

Let $\bar{\mu} = \max_k \|\mu_k\|_\infty$ and $\bar{\sigma}^2 = \max_k \|\Sigma_k\|_{\mathrm{op}}$. A Gaussian Chernoff bound, the union bound over coordinates, and averaging over mixture components give, for $\varrho \geq 2\bar{\mu}$,

$$\mathbb{P}(\|X\|_\infty > \varrho) \leq 2d \exp\left(-\frac{\varrho^2}{8\bar{\sigma}^2}\right).$$

A fixed finite Gaussian mixture has a finite fourth moment. Using the linear score bound from Step 1 and Cauchy–Schwarz,

$$\mathbb{E}[\|s^*(X)\|_2^2 \, \mathbf{1}\{\|X\|_\infty > \varrho\}] \lesssim \{\mathbb{E}(1 + \|X\|_\infty^2)^2\}^{1/2} \mathbb{P}(\|X\|_\infty > \varrho)^{1/2} \lesssim e^{-\varrho^2/(16\bar{\sigma}^2)}.$$

Hence (F.15) holds with $a_{\mathrm{G}} = (16\bar{\sigma}^2)^{-1}$ and $\varrho_0 = \max\{1, 2\bar{\mu}\}$. □

Proof of Corollary 3. The local norm grows polynomially in the approximation radius, whereas the weighted tail decays as a Gaussian tail. A square-root-logarithmic radius therefore leaves only logarithmic factors in the approximation and sampling terms.

Step 1: Choose the analysis radius and check admissibility.

Write $\kappa = d/(4d+16m)$ and $\ell_n = \log(en)$ within this proof. Choose

$$\varrho_n = \max\left\{\varrho_0, \sqrt{2\ell_n/a_{\mathrm{G}}}\right\}. \tag{F.16}$$

Lemma F.5 gives $H_{\varrho_n} \lesssim \ell_n^{(m+1)/2}$. Sufficiently large fixed multipliers in (21) ensure $M_n \geq 2H_{\varrho_n}$ and $B_{J,n} \geq 2dH_{\varrho_n}$. Also,

$$W_n L_n \asymp n^{2\kappa}, \qquad \Xi(W_n, L_n) \gtrsim \frac{n^{2\kappa}}{\ell_n^2}, \qquad \varrho_n^{d/2} \lesssim \ell_n^{d/4}.$$

A positive power of n dominates every fixed logarithmic power. Thus the capacity inequality in (19), as well as $W_n \geq W_0$ and $L_n \geq L_0$, holds for all sufficiently large n.

Step 2: Substitute the three contributions.

The sampling term in (20) satisfies

$$(M_n^2 + B_{J,n}) W_n L_n \sqrt{\frac{\log(eW_nL_n)}{n}} \lesssim n^{2\kappa - 1/2} \ell_n^{m+3/2}.$$

Using (F.14), its approximation term satisfies

$$H_{\varrho_n}^2 \varrho_n^{2m} \Xi(W_n, L_n)^{-4m/d} \lesssim n^{-8m\kappa/d} \ell_n^{2m+1+8m/d}.$$

Finally, (F.15) and (F.16) give

$$\mathbb{E}[(M_n^2 + \|s^*(X)\|_2^2)\mathbf{1}\{\|X\|_\infty > \varrho_n\}] \lesssim (1+M_n^2) e^{-a_{\mathrm{G}}\varrho_n^2} \lesssim n^{-2}\ell_n^{m+1}.$$

Step 3: Balance the polynomial orders.

The chosen κ satisfies

$$2\kappa - \tfrac{1}{2} = -\frac{8m\kappa}{d} = -\frac{2m}{d+4m}.$$

Since $m \geq 2$, $2m + 1 + 8m/d \geq m + 3/2$, and the tail term is of smaller order. Substitution into Theorem 2 proves (22). The constants may depend on m and the fixed mixture parameters; the assertion fixes m before taking $n \to \infty$ and makes no minimax claim. □

F.6 Gaussian, Gaussian-mixture, and Student examples

Proposition F.1 (A Gaussian bound with fixed architecture)**.** *Let $p = N(\mu, \Sigma)$ with $\Sigma \succ 0$, and let $\kappa_{\mathrm{G}} \geq \left\|\Sigma^{-1}\right\|_{\mathrm{op}}$ be a fixed positive upper bound. Fix $A = A^\top \succ 0$, $B_J \geq \max\{1, \left\|\Sigma^{-1}\right\|_F\}$, $W \geq 4d$, and $L \geq 2$ with $c_{\mathrm{term}} \log(eL) \geq 1$. Use $\mathcal{F}_n$ with $W_n = W$, $L_n = L$, $B_{J,n} = B_J$, and*

$$M_n = \max\{1, 2\sqrt{\kappa_{\mathrm{G}} \log(en)}\}. \tag{F.17}$$

Then every measurable empirical minimizer in this class for ϕ_A satisfies

$$\mathbb{E}_{\mathcal{D}} \|\widehat{s} - s^*\|_{L^2(p)}^2 \lesssim \frac{\log(en)}{\sqrt{n}}. \tag{F.18}$$

The hidden constant depends only on $d, A, \kappa_{\mathrm{G}}, B_J, W, L, c_{\mathrm{term}}$. No boundedness assumption is imposed on s^*.

Proof. The Gaussian score is affine. We realize its coordinatewise clipping by a small mixed network, control the clipping error using Gaussian tails, and apply the finite-moment oracle bound.

Step 1: Calculate the score and its energy.

Differentiating the Gaussian log density gives

$$s^*(x) = -\Sigma^{-1}(x-\mu), \qquad J_{s^*} = -\Sigma^{-1}, \qquad I(p) = \mathrm{Tr}(\Sigma^{-1}). \tag{F.19}$$

The last equality uses $\mathbb{E}[(X-\mu)(X-\mu)^\top] = \Sigma$. These are unbounded affine scores with finite score second moment, as in the Gaussian example of Hyvärinen (2005).

Step 2: Realize a clipped affine comparator in the mixed class.

For $M > 0$, let $\mathrm{clip}_M(t) = \max\{-M, \min\{M, t\}\}$ and put $f_M = \mathrm{clip}_M \circ s^*$ coordinatewise. The identity

$$\mathrm{clip}_M(t) = -M + \mathrm{ReLU}(t + M) - \mathrm{ReLU}(t - M) \tag{F.20}$$

realizes all clipped coordinates with $2d$ neurons in one ReLU hidden layer. To follow it by a terminal ReQU layer without changing its values, use

$$t = \frac{1}{4}\big[\mathrm{ReQU}(t + 1) + \mathrm{ReQU}(-t - 1) - \mathrm{ReQU}(t - 1) - \mathrm{ReQU}(1 - t)\big]. \tag{F.21}$$

The identity follows from $\mathrm{ReQU}(z)+\mathrm{ReQU}(-z) = z^2$. It uses $4d$ neurons for all output coordinates and has an affine output. The resulting network has one ReLU hidden layer followed by one ReQU hidden layer, with width at most $4d$. Its weak Jacobian equals $-D_M(x)\Sigma^{-1}$, where $D_M(x)$ is diagonal with entries in $\{0, 1\}$ almost everywhere. Hence

$$\|f_M\|_\infty \le M, \qquad \|J_{f_M}\|_{F,\infty} \le \left\|\Sigma^{-1}\right\|_F.$$

The conditions on W, L, c_{term} and B_J therefore place f_{M_n} in $\mathcal{F}_n$.

Step 3: Control the clipping error and apply the oracle inequality.

Each coordinate $Z_j = s_j^*(X)$ is centered Gaussian with variance $v_j = (\Sigma^{-1})_{jj} \le \kappa_{\mathrm{G}}$. From its moment generating function, Chernoff's inequality gives $\mathbb{P}(|Z_j| > M) \le 2\exp(-M^2/(2v_j))$, and $\mathbb{E}Z_j^4 = 3v_j^2$. Thus Cauchy–Schwarz yields

$$\begin{aligned}\mathbb{E}(|Z_j| - M)_+^2 &\le \mathbb{E}\big[Z_j^2 \mathbf{1}\{|Z_j| > M\}\big] \\ &\le (\mathbb{E}Z_j^4)^{1/2}\mathbb{P}(|Z_j| > M)^{1/2} \le \sqrt{6}\, v_j \exp\left(-\frac{M^2}{4v_j}\right).\end{aligned}$$

After summing over coordinates,

$$\|f_M - s^*\|_{L^2(p)}^2 \le \sqrt{6}\ \mathrm{Tr}(\Sigma^{-1}) \exp\left(-\frac{M^2}{4\kappa_{\mathrm{G}}}\right). \tag{F.22}$$

With (F.17), this is at most $\sqrt{6}\ \mathrm{Tr}(\Sigma^{-1})/(en)$. For fixed W, L, B_J, the stochastic term of (F.8) is $O(\log(en)/\sqrt{n})$. Substitution proves (F.18). This rate follows from a uniform-deviation bound and is not asserted to be the parametric minimax rate. □

Finite Gaussian mixtures. Let $p = \sum_{k=1}^{K_{\mathrm{mix}}} \pi_k p_k$, where $\pi_k > 0$, $\sum_k \pi_k = 1$, and p_k is the $N(\mu_k, \Sigma_k)$ density with $\Sigma_k \succ 0$. Then p is positive and smooth. If Z_{mix} denotes the component label, direct differentiation gives

$$s^*(x) = \sum_k \frac{\pi_k p_k(x)}{p(x)} s_k(x) = \mathbb{E}[s_{Z_{\mathrm{mix}}}(X) \mid X = x], \qquad s_k(x) = -\Sigma_k^{-1}(x - \mu_k).$$

The convexity of squared Euclidean norm and conditional Jensen's inequality imply

$$I(p) \le \mathbb{E}\,\|s_{Z_{\mathrm{mix}}}(X)\|_2^2 = \sum_{k=1}^{K_{\mathrm{mix}}} \pi_k\, \mathrm{Tr}(\Sigma_k^{-1}) < \infty. \tag{F.23}$$

Thus Lemma F.3 gives pointwise consistency for any such fixed mixture. Lemma F.5 additionally verifies the quantitative conditions needed for Corollary 3. The clipped affine comparator in Proposition F.1 is specific to a single Gaussian.

Multivariate Student distributions. Let $\nu_{\mathrm{t}} > 0$, $\Sigma \succ 0$, and

$$p(x) \propto \left(1 + \frac{(x - \mu)^\top \Sigma^{-1}(x - \mu)}{\nu_{\mathrm{t}}}\right)^{-(\nu_{\mathrm{t}}+d)/2}.$$

Here Σ is the scale matrix. The density is positive and smooth, and its log-density derivatives are bounded on every compact cube. Thus the local Sobolev condition in Assumption 5 holds for each fixed m. Direct differentiation gives

$$s^*(x) = -\frac{(\nu_t + d)\Sigma^{-1}(x-\mu)}{\nu_t + (x-\mu)^\top \Sigma^{-1}(x-\mu)}. \tag{F.24}$$

To compute its energy, set $Y = \Sigma^{-1/2}(X-\mu)$ and $r = \|Y\|_2$. Rotational symmetry implies that the direction of Y is uniform on the unit sphere, independently of r, with second moment matrix I_d/d. The radial density is proportional to $r^{d-1}(1+r^2/\nu_t)^{-(\nu_t+d)/2}$. Under $T = r^2/(\nu_t + r^2)$, change of variables gives the beta density with parameters $d/2$ and $\nu_t/2$. Its beta-integral moment is

$$\mathbb{E}[T(1-T)] = \frac{d\nu_t}{(d+\nu_t)(d+\nu_t+2)}.$$

Since $r^2/(\nu_t + r^2)^2 = T(1-T)/\nu_t$, substituting (F.24) and averaging first over directions yields

$$I(p) = \frac{\nu_t + d}{\nu_t + d + 2}\,\mathrm{Tr}(\Sigma^{-1}) < \infty. \tag{F.25}$$

In contrast, the radial integral for $\mathbb{E}r^2$ has an integrand of order $r^{1-\nu_t}$ at infinity, and diverges for $0 < \nu_t \le 2$. These cases distinguish the score second moment in (18) from an input second moment. No moment generating function assumption is needed in Theorem 2.

G Additional Experimental Details

G.1 Synthetic score recovery on grid mixtures

Data generation. We evaluate score recovery on a synthetic grid-mixture distribution following the kernel score-estimation setup used in prior work. For each dimension d, we construct $K_{\mathrm{mix}} = d$ Gaussian components. The component centers $\{\mu_k\}_{k=1}^{K_{\mathrm{mix}}} \subset \mathbb{R}^d$ are generated as random binary grid vertices, with each coordinate taking values in $\{0, 1\}$. The data density is

$$p(x) = \frac{1}{d} \sum_{k=1}^{d} \varphi_{I_d}(x - \mu_k), \qquad x \in \mathbb{R}^d,$$

where φ_{I_d} denotes the density of a d-dimensional standard Gaussian distribution. Equivalently, a sample is generated by first drawing a component index Z_{mix} uniformly from $\{1, \dots, d\}$, and then sampling

$$X = \mu_{Z_{\mathrm{mix}}} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d).$$

The population score is available in closed form:

$$s^*(x) = \nabla \log p(x) = \sum_{k=1}^{d} \omega_k(x) \big[-(x - \mu_k)\big],$$

where

$$\omega_k(x) = \frac{\varphi_{I_d}(x - \mu_k)}{\sum_{\ell=1}^{d} \varphi_{I_d}(x - \mu_\ell)}$$

is the posterior responsibility of the k-th component. In implementation, the true score is computed by differentiating the mixture log-density.

Sample sizes and dimensions. We consider dimensions

$$d \in \{8, 32, 64\}$$

and training sample sizes

$$n \in \{500, 1000, 2000\}.$$

For each pair (d, n), all methods are evaluated on an independent test set of size

$$n_{\mathrm{test}} = 1024.$$

Each configuration is repeated over

$$N_{\mathrm{rep}} = 10$$

independent trials. Figure 1 reports the mean normalized score-recovery error over these repeated trials.

Mixed-network architecture and training. The mixed-network score estimator is a vector-valued network with input dimension d and output dimension d. The hidden width is 128. The architecture is

$$d \longrightarrow 128 \overset{\mathrm{ReLU}}{\longrightarrow} 128 \overset{\mathrm{ReLU}}{\longrightarrow} 128 \overset{\mathrm{SquaredReLU}}{\longrightarrow} d,$$

where

$$\mathrm{SquaredReLU}(u) = (\max\{u, 0\})^2.$$

Thus, the network uses ReLU layers in the main body and a terminal squared-ReLU activation before the output layer, matching the mixed architecture in Definition 1.

The model is trained with the quadratic Bregman score matching loss,

$$\widehat{\mathcal{L}}_{\mathrm{SM}}(s) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2} \|s(X_i)\|_2^2 + \operatorname{div} s(X_i) \right],$$

where the divergence term is computed by automatic differentiation. We use AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 256, and a maximum of 200 epochs. Gradients are clipped with maximum norm 1.0. The learning rate is reduced by a factor 0.5 every 80 epochs. Early stopping is based on validation score MSE, with patience 25 and minimum improvement 10^{-4}.

Baselines. We compare the mixed estimator with representative kernel and Stein-type score estimators: Stein, SSGE, KEF-CG, NKEF$_2$, and ν-Method. Stein and SSGE use diagonal IMQ kernels, while KEF-CG, NKEF$_2$, and ν-Method use curl-free inverse multiquadric (IMQ) kernels. For Stein, KEF-CG, NKEF$_2$, and ν-Method, the regularization parameter is searched over

$$\lambda_{\mathrm{reg}} \in [10^{-6}, 10^{-2}]$$

using six logarithmically spaced values. For SSGE, the spectral keep rate is searched over six equally spaced values in $[0.1, 0.95]$. For all kernel methods, the kernel bandwidth parameter is searched over six equally spaced values in $[5, 40]$. For NKEF$_2$, we use a Nyström subsample rate of 0.2. For each method and trial, the original experiment record describes the selected value as the best "validation/test-grid score error" over this grid. That record does not unambiguously distinguish the split used for baseline selection. The reported values are retained, and an untouched-test interpretation requires confirmation from the original logs. The neural early-stopping criterion above uses oracle score information on the validation split.

Evaluation metric. Throughout the paper, NormL2 denotes the test-set statistic in (23). For this benchmark,

$$\mathrm{NormL2} = \frac{1}{d\, n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left\| \widehat{s}(X_i^{\mathrm{test}}) - s^*(X_i^{\mathrm{test}}) \right\|_2^2, \qquad n_{\mathrm{test}} = 1024.$$

This statistic estimates the per-coordinate population error $d^{-1}\|\widehat{s} - s^*\|^2_{L^2(p)}$; the population quantity is not separately denoted by NormL2. Figure 1 reports its mean over the $N_{\mathrm{rep}} = 10$ independent trials. Lower values indicate more accurate score recovery.

Additional numerical results. Table 6 reports the mean and standard deviation of $\mathrm{NormL2}$ over $N_{\mathrm{rep}} = 10$ independent trials for the grid-mixture benchmark. These values correspond to the curves shown in Figure 1.

Table 6: Grid-mixture score-recovery results. Each entry reports $\mathrm{NormL2}$ mean $\pm$ standard deviation over $N_{\mathrm{rep}} = 10$ independent trials. Lower is better.

d	Method	$n = 500$	$n = 1000$	$n = 2000$
8	Ours	$\mathbf{0.0167 \pm 0.0014}$	$\mathbf{0.0132 \pm 0.0007}$	$\mathbf{0.0123 \pm 0.0015}$
8	Stein	0.0507 ± 0.0077	0.0417 ± 0.0056	0.0384 ± 0.0026
8	SSGE	0.1432 ± 0.0138	0.1017 ± 0.0083	0.0929 ± 0.0042
8	KEF-CG	0.0378 ± 0.0047	0.0310 ± 0.0040	0.0226 ± 0.0018
8	NKEF$_2$	0.0382 ± 0.0049	0.0311 ± 0.0040	0.0226 ± 0.0021
8	ν-Method	0.0369 ± 0.0018	0.0319 ± 0.0015	0.0246 ± 0.0014
32	Ours	$\mathbf{0.0163 \pm 0.0008}$	$\mathbf{0.0124 \pm 0.0004}$	$\mathbf{0.0105 \pm 0.0006}$
32	Stein	0.1847 ± 0.0033	0.1468 ± 0.0019	0.1315 ± 0.0011
32	SSGE	0.2066 ± 0.0010	0.1529 ± 0.0026	0.1167 ± 0.0007
32	KEF-CG	0.1155 ± 0.0016	0.1064 ± 0.0006	0.0957 ± 0.0005
32	NKEF$_2$	0.1165 ± 0.0011	0.1073 ± 0.0007	0.0972 ± 0.0005
32	ν-Method	0.1107 ± 0.0023	0.1043 ± 0.0012	0.0965 ± 0.0010
64	Ours	$\mathbf{0.0137 \pm 0.0001}$	$\mathbf{0.0128 \pm 0.0002}$	$\mathbf{0.0116 \pm 0.0004}$
64	Stein	0.3044 ± 0.0013	0.2471 ± 0.0030	0.2374 ± 0.0022
64	SSGE	0.2610 ± 0.0028	0.2350 ± 0.0032	0.1940 ± 0.0018
64	KEF-CG	0.1786 ± 0.0024	0.1708 ± 0.0020	0.1627 ± 0.0015
64	NKEF$_2$	0.1793 ± 0.0022	0.1716 ± 0.0021	0.1638 ± 0.0015
64	ν-Method	0.1789 ± 0.0024	0.1702 ± 0.0020	0.1642 ± 0.0012

G.2 Matched activation comparison

This experiment examines the effect of the activation sequence on score recovery while keeping the quadratic loss and parameter count fixed. We compare pure ReLU, pure ReQU, and mixed ReLU–ReQU networks on a correlated Gaussian with an analytic score and Jacobian. Table 4 in Section 4.3 reports the results for NormL2, cosine similarity, and Jacobian-RelF. The setup and metric definitions are given below.

Target and preprocessing. The target before standardization is $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu_j = -1 + 2(j-1)/7$ and $\Sigma_{ij} = 0.4^{|i-j|}$. The score is $s^*(x) = -\Sigma^{-1}(x-\mu)$ and its input Jacobian is $-\Sigma^{-1}$. The training-set mean and coordinate standard deviations are used to standardize all three splits. If $z = D^{-1}(x-b)$ with a diagonal scaling matrix D, the analytic score and Jacobian in the standardized coordinates are

$$s_z^*(z) = -D\Sigma^{-1}(Dz + b - \mu), \qquad J_{s_z^*} = -D\Sigma^{-1}D. \tag{G.1}$$

Equation (G.1) expresses the analytic target in the coordinates used for training; it does not introduce a new experiment.

Architecture and selection. Every model has affine dimensions $8 \to 128 \to 128 \to 128 \to 8$. The three activation sequences are ReLU–ReLU–ReLU, ReQU–ReQU–ReQU, and ReLU–ReLU–ReQU. Each model therefore has $35{,}208$ parameters, with no activation-specific learned parameters. The five paired runs share data, affine-layer initialization, optimizer choice, preprocessing, and batch order across architectures. The selected learning rates are shown in Table 4. Learning rates are tuned separately by architecture, and epoch selection uses validation score MSE. The reported test set is distinct from the validation set.

Metrics. The three metrics are evaluated in the training-standardized coordinates in (G.1). NormL2 is defined in (23). Cosine similarity measures directional agreement between the estimated and analytic score vectors, with larger values indicating closer alignment. It complements NormL2 by separating direction from magnitude. We report

$$\text{Cosine} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{\langle \widehat{s}(z_i), s_z^*(z_i) \rangle}{\|\widehat{s}(z_i)\|_2 \|s_z^*(z_i)\|_2}.$$

We evaluate the relative Jacobian error on 256 independent test observations:

$$\text{Jacobian-RelF} = \frac{\left(\frac{1}{256} \sum_{i=1}^{256} \|J_{\widehat{s}}(z_i) - J_{s_z^*}\|_F^2 \right)^{1/2}}{\|J_{s_z^*}\|_F}. \tag{G.2}$$

The full Jacobian is formed only for this evaluation. All three metrics in Table 4 are means and sample standard deviations across five runs; no standard errors or significance tests are substituted for those deviations.

Results. Table 4 reports cosine similarities of 0.8419 ± 0.0133 for pure ReLU, 0.9676 ± 0.0037 for pure ReQU, and 0.9800 ± 0.0039 for the mixed network. The mixed network has the highest mean directional agreement and the lowest mean NormL2 and Jacobian-RelF in this comparison.

G.3 Generator geometry experiment

Data and network. The bounded product-density diagnostic is given in (24). Its analytic score and nonzero Jacobian entries are

$$s_j^*(x) = -2\pi m_j \kappa_j \sin(2\pi m_j x_j), \qquad \partial_j s_j^*(x) = -4\pi^2 m_j^2 \kappa_j \cos(2\pi m_j x_j). \tag{G.3}$$

The cross-coordinate partial derivatives vanish. The diagnostic uses coordinatewise rejection sampling, with 500 training, $1{,}000$ validation, and $5{,}000$ test observations for each of five paired runs. Its mixed network has affine dimensions $8 \to 128 \to 128 \to 128 \to 8$ and the activation sequence ReLU–ReLU–ReQU.

Dual parameterization. The network represents the generator-dual field of the score. For the mixed-network output h_θ, the boundary-compatible implementation uses

$$v_{\theta,j}(x) = 4x_j(1-x_j)h_{\theta,j}(x), \qquad s_\theta(x) = (\nabla\phi)^{-1}(v_\theta(x)). \tag{G.4}$$

For the quadratic baseline the inverse is the identity. For the anisotropic family it is $s_\theta = A_a^{-1} v_\theta$; for hypentropy it is $s_{\theta,j} = \beta \sinh(v_{\theta,j})$, as follows from (8). The gate makes the matching normal component of v_θ zero on each face. It is retained as part of this particular experiment and is not a condition in the full-space class (10).

Writing ϕ^* for the convex conjugate, the loss used by the dual formulation can be expressed as

$$\ell_\phi(s_\theta; x) = \phi^*(v_\theta(x)) + \operatorname{div} v_\theta(x). \tag{G.5}$$

Indeed, $v_\theta = \nabla\phi(s_\theta)$ gives $\phi^*(v_\theta) = \langle v_\theta, s_\theta\rangle - \phi(s_\theta)$, and the chain rule identifies $\operatorname{div} v_\theta$ with the trace term in (5). This explains the implemented objective without identifying the dual network class with the direct-score class. The inverse hypentropy map is nonlinear, and the primal classes vary with the generator. The empirical trade-offs therefore describe the generator-dependent implementations as a whole.

Selection and repeated runs. Model selection uses the observable validation score-matching loss. Tables 2–3 aggregate five paired runs with the dual-coordinate parameterization in (G.4). The $a = 32$ point is a stress test; the directional trend is assessed over the tested stable range $a \le 16$.

Coordinate and magnitude errors. For a coordinate set I, define

$$\mathrm{MSE}_I = \frac{1}{|I| n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \sum_{j \in I} (s_{\theta,j}(x_i) - s_j^*(x_i))^2. \tag{G.6}$$

We use $I = \{1, 2\}$ and $I = \{3, \dots, 8\}$. For magnitude stratification, all observation–coordinate pairs are grouped according to $|s_j^*(x_i)|$, using its 50th and 90th percentiles. Within each bin the squared coordinate errors are averaged over the included pairs. These are true-score bins. They should not be confused with a separate check of the predicted-score curvature weights $(s_{\theta,j}(x_i)^2 + 1)^{-1/2}$, which decreases across predicted-magnitude bins in the reported diagnostic. The latter diagnostic is not used to define the errors in Table 3.

Precision and uncertainty. Table 2 uses four-decimal rounding. In Table 3, magnitude-bin errors use three decimals and global NormL2 uses four. All uncertainty summaries are sample standard deviations across five runs, rather than standard errors. The comparisons are descriptive and do not assert statistical significance.

G.4 Derivative computation

For the quadratic objective, the implementation computes the divergence by applying reverse-mode automatic differentiation to one output coordinate at a time and retaining its matching input derivative. Training thus uses d input-gradient evaluations per minibatch without storing the complete batch of $d \times d$ Jacobians. The parameter update retains the graph needed to differentiate the divergence with respect to network weights. For a general generator, the same coordinatewise procedure can be applied to $\nabla\phi(s_\theta)$; (G.5) already exposes that vector field in the dual experiment. This describes the computational procedure. It is not a measured runtime comparison, and no stochastic trace estimator is used in the reported theory.

G.5 Network-capacity sensitivity at fixed sample size

We examine how the mixed ReLU–ReQU estimator responds to a joint increase in width and depth at fixed $n = 2000$ and $d = 8$. Table 7 records four configurations, from two hidden layers of width 64 to five hidden layers of width 512. Depth counts hidden layers, and the parameter counts include affine weights and biases. NormL2 follows the test-set convention in (23); the Jacobian-RelF column retains the relative Jacobian error reported in the capacity experiment.

Table 7: Network-capacity sensitivity at fixed $n = 2000$ and $d = 8$. Lower errors are better.

Configuration	Depth	Width	Parameters	NormL2	Jacobian-RelF
Small	2	64	5,256	0.0854 ± 0.0034	0.5033 ± 0.0083
Medium	3	128	35,208	0.0905 ± 0.0097	0.4876 ± 0.0275
Large	4	256	201,736	$\mathbf{0.0822 \pm 0.0047}$	$\mathbf{0.4347 \pm 0.0135}$
XLarge	5	512	1,059,336	0.1091 ± 0.0077	0.4397 ± 0.0176

The mean score error is not monotone in network size: the Large configuration has the lowest NormL2, whereas XLarge has a higher error than all three smaller configurations. Mean Jacobian-RelF decreases from Small to Large and increases slightly at XLarge. Thus the largest network need not provide the most accurate estimate at this fixed sample size. The sweep varies width and depth together and measures their combined finite-sample effect; it does not isolate either factor or verify the asymptotic exponents in Section 3.

G.6 WAE EXPERIMENTAL DETAILS

Implementation. The WAE experiments are implemented in PyTorch, following the MNIST WAE setting of Zhou et al. (2020). The nonparametric score estimators are adapted from prior score-estimation implementations, while the mixed-network score estimator is integrated into the same WAE training framework. All methods use the same data pipeline, WAE architecture, and evaluation protocol whenever applicable.

Dataset and preprocessing. We use the real-valued MNIST dataset stored in `mnist.pkl.gz`, following the preprocessing convention of the original WAE experiment rather than switching to the torchvision MNIST loader. Images are represented as 28×28 grayscale images and flattened into 784-dimensional vectors for the fully connected WAE architecture.

Compared methods. We compare the proposed mixed-network score estimator with the nonparametric score estimators used in the main experiments, including Stein, SSGE, KEF-CG, NKEF_2, and the ν-Method. In the WAE setting, each score estimator is applied in the latent space to approximate the score of the aggregated posterior $q_\eta(z)$, which is then used to estimate the entropy-gradient component of the KL regularization term.

Training protocol. For each method and latent dimension $d_z \in \{8, 32, 64, 128\}$, we train a WAE using the same encoder–decoder architecture and training protocol. Each configuration is repeated over 10 independent runs, and we report the mean and standard deviation of AIS-based negative log-likelihood in Table 5. The implemented WAE methods include `stein`, `ssge`, `nkef2`, `kef_cg`, `nu`, and `DSRN`; the legacy code identifier `DSRN` corresponds to "Ours" in the reported results.

Mixed-network score estimator. For the WAE using the mixed score estimator, we use a buffer-based latent score estimator. During training, latent samples are periodically collected from the current encoder and standardized. The mixed score network is then refitted using score matching and frozen during subsequent WAE updates. The default stabilizing choices include a latent buffer, periodic score refitting, gradient clipping, score-norm clipping, a floor on the latent standard deviation, entropy scaling, and entropy warmup. This design separates latent score fitting from WAE parameter updates. The present experiments do not isolate the contribution of each stabilizing mechanism.

AIS evaluation. Generative performance is evaluated using Annealed Importance Sampling (AIS) to estimate the test negative log-likelihood. For paper-scale evaluation, AIS uses multiple chains and a large number of intermediate temperatures, with prior-to-joint AIS, automatic Hamiltonian Monte Carlo (HMC) step-size calibration, float64 AIS computation, and multiple leapfrog steps. The same AIS protocol is used across methods and latent dimensions. Lower AIS-based negative log-likelihood indicates better generative performance.

Generated samples. Figure 2 shows generated MNIST samples from WAE models trained with different latent-space score estimators and latent dimensions. Rows correspond to score estimators, and columns correspond to $d_z \in \{8, 32, 64, 128\}$. The figure is intended as qualitative evidence complementing the AIS-based quantitative evaluation in Table 5.

Reporting scope. The original records specify the comparison pipeline but do not provide every encoder/decoder layer size, buffer/refit setting, or numerical AIS control. These quantities cannot be reconstructed uniquely from the reported NLL table. Hardware, wall-clock time, and an isolated stabilization ablation are not reported here.

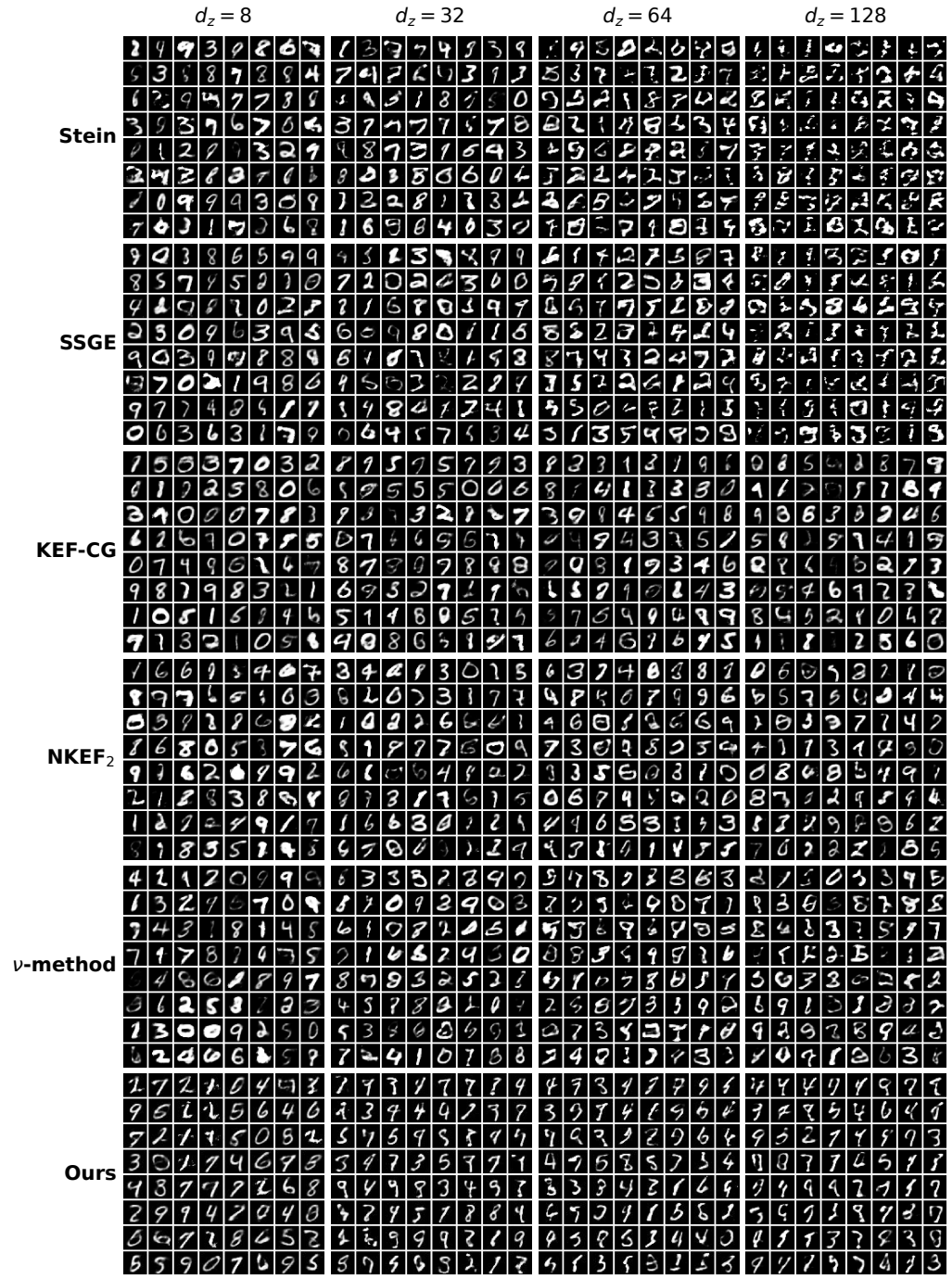

Figure 2: Generated MNIST samples from WAE models using different latent-space score estimators and latent dimensions. Columns correspond to $d_z \in \{8, 32, 64, 128\}$, and rows correspond to Stein, SSGE, KEF-CG, NKEF$_2$, ν-Method, and Ours (mixed ReLU–ReQU).