WEIRDO: WEAK RESIDUAL REGULARIZED DOOB'S H-TRANSFORM DIFFUSION ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of estimating the guidance that steers the distribution learned by a diffusion generative model toward a tilted target $q_0 \propto w\, p_0$ at inference time. Relying on the stochastic optimal control approach, we observe that the exact drift correction is the gradient of the logarithm of Doob's h-function, and we study the problem of estimating it from a sample. In the present paper, we assume that the score of the pretrained model is available, that the tilting weight is bounded and positive, and that the reference distribution has a bounded support, no smoothness of the weight is required. Introducing a penalized least-squares risk in which the penalty is the residual of the space-time harmonicity equation satisfied by the h-function, measured in a dual Sobolev norm, we derive high-probability bounds on the squared error of the resulting guidance estimate. Since the penalty vanishes at the target, the estimator is free of regularization bias, and in favourable scenarios its rate of convergence is faster than the minimax rate of estimating first-order derivatives of a smooth regression function. Assuming that w is bounded and positive with $\mathbb{E}_{p_0}[w^{-\mathrm{s}}] < \infty$ for some $\mathrm{s} \in (0, \infty]$, and that the reference data are compactly supported, we prove that the guidance is estimable in squared L^2 at rate $\varepsilon_n^{\mathrm{s}/(\mathrm{s}+4)}$, where $\varepsilon_n = n^{-2(\beta-1)/(2(\beta-1)+d)}$. We also transfer the obtained bounds to the total variation distance between the marginals of the estimated and the exactly guided samplers, and illustrate the performance of the suggested approach with numerical experiments.

1 INTRODUCTION

Score-based diffusion models (Song et al., 2021b) have become a standard tool of generative modelling. Such a model transforms data into noise by a forward stochastic differential equation and produces new samples by running the corresponding time reversal. Let p_0 denote the data distribution on $\mathbb{R}^d$ and let the forward dynamics be the variance-preserving Ornstein–Uhlenbeck process

$$dX_\tau = -X_\tau\, d\tau + \sqrt{2}\, dB_\tau, \qquad 0 \le \tau \le T, \qquad X_0 \sim p_0, \tag{1}$$

where B is a standard d-dimensional Brownian motion and $T > 0$ is a fixed time horizon. We denote the law of X_τ by p_τ, which for $\tau > 0$ has a smooth positive density denoted in the same way, and note that p_τ approaches the standard Gaussian density as τ grows. It is well known (Anderson, 1982; Haussmann & Pardoux, 1986) that the time reversal $\overleftarrow{X}_t \overset{d}{=} X_{T-t}$ of the process (1) is again a diffusion, governed by the SDE

$$d\overleftarrow{X}_t = b(t, \overleftarrow{X}_t)\, dt + \sqrt{2}\, dB_t, \qquad b(t,x) = x + 2\, s(t,x), \qquad s(t,x) = \nabla \log p_{T-t}(x), \tag{2}$$

started at $\overleftarrow{X}_0 \sim p_T$ and terminating at $\overleftarrow{X}_T \sim p_0$. The function s is referred to as the score. It is not available in closed form and, in practice, is learned by a neural network. Throughout the paper, a pretrained diffusion model is understood as a supplier of s, and we treat s as known.

Once such a model is available, one usually wants to adapt it to a particular task or reward, and retraining it is expensive. Let $w : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ be a weight expressing the new objective: a reward exponential $e^{r/\alpha}$, a likelihood in an inverse problem, or a class probability. The aim is then to sample not from p_0 but from the tilted distribution

$$q_0 \;\propto\; w\, p_0. \tag{3}$$

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.

Several routes are available. One can fine-tune the network (Uehara et al., 2024), resample from a heuristic proposal (Wu et al., 2023), or add a drift to (2) (Dhariwal & Nichol, 2021; Ho & Salimans, 2021). We are interested in the third kind, known as inference-time alignment, because it leaves the pretrained weights untouched: the pretrained network is never retrained, only an auxiliary estimator of the correction is fitted per reward, and there is no need to draw many samples in order to keep a few. Unlike the usual heuristics, however, we work with the drift that is exact.

That drift is determined by a single function. From the optimal control point of view, turning p_0 into q_0 reduces to finding

$$h^*(t,x) \;:=\; \mathbb{E}\left[w(X_0) \mid X_{T-t} = x\right], \tag{4}$$

known as Doob's h-function (Doob, 1984). Adding $2\nabla_x \log h^*$ to the drift of (2) produces

$$d\overleftarrow{Z}_t = \big(b(t, \overleftarrow{Z}_t) + 2\nabla_x \log h^*(t, \overleftarrow{Z}_t)\big)dt + \sqrt{2}\, dB_t, \tag{5}$$

whose terminal law is exactly q_0, when it is started from the tilted law $h^*(0,\cdot)p_T/Z$, $Z := \int w dp_0$. What the sampler uses is the guidance $g^* := \nabla \log h^*$, and since p_0 is unknown, h^* has to be estimated from data. This is done on a window of reverse times $I = [t_0, t_1] \subset (0, T)$ kept away from both endpoints, since as $t \uparrow T$ the noise vanishes and recovery is ill-posed, while as $t \downarrow 0$ there is nothing to recover.

Our focus lies on statistical aspects of guidance estimation, and here the following difficulty arises. By (4), the function h^* is a conditional expectation, so a natural strategy is to estimate it by nonparametric regression. The sampler (5), however, does not consume h^* but $\nabla \log h^*$, and a small value error does not by itself imply a small gradient error. Estimating the derivative is also intrinsically harder than estimating the function: for a β-smooth regression function observed in noise, the optimal squared L^2 rate is $n^{-2\beta/(2\beta+d)}$ for the function and $n^{-2(\beta-1)/(2\beta+d)}$ for its first derivative (Stone, 1982), one derivative's worth slower. In order to do better, one has to exploit a property of h^* that an arbitrary conditional expectation does not possess.

The property we use is that h^* is space-time harmonic for the generator of (2), that is, it satisfies the PDE

$$\mathcal{M}h^* \;:=\; \partial_t h^* + b(t,x)\cdot\nabla h^* + \Delta h^* \;=\; 0. \tag{6}$$

Every term in (6) is either a derivative of the unknown function or, through b, the score of the pretrained model. The residual $\mathcal{M}h$ can therefore be evaluated on the sample and added to the regression objective as a penalty. Since it vanishes at h^*, this penalty introduces no bias at the population level, whatever weight it is given.

However, the Laplacian in (6) makes residual second order in the unknown, and second derivatives are exactly what the rate we are aiming for cannot control. We therefore pass to a weak form: we test (6) against a function ψ and integrate by parts, which moves one derivative from h onto ψ. This step is available to us for a specific reason. Integrating by parts against the marginal p_{T-t} produces its logarithmic derivative as a weight, and that is again the score already appearing in (2). What results is a penalty that is first order in both arguments and computable from the pretrained model. From it we obtain an estimator of h^* whose guidance converges to g^* in the squared L^2 norm at a rate we make explicit, and the same bound transfers to the marginals of (5) in total variation.

Throughout we take $\beta = 2$ as the leading case, treating the approximation of h^* as that of a $W^{\beta,\infty}$ function. Although h^* is smooth for any bounded w, as we show below, $\beta = 2$ is the smallest integer for which the minimax rate $n^{-2(\beta-1)/(2\beta+d)}$ of Stone (1982) is informative, and hence the smallest for which a comparison against it is meaningful. This should be read not as a assumption that h^* belong to that smoothness class, but as the primary example in which the difference between our estimate and the benchmark is clearest.

Contribution.

- The main contribution of this paper is a non-asymptotic high-probability upper bound on the squared L^2-error of the guidance estimate (Theorem 5), which holds on all but an arbitrarily small end-portion of the estimation window. In favourable scenarios, when the weight w is not too close to zero, the obtained rate is faster than the generic minimax rate of first-derivative estimation for a smooth regression function.

- The proof relies on an exact energy identity for the reverse diffusion, which allows us to control the gradient by the value error and the weak residual (Theorem 2). We would like to emphasize that this result requires no inf-sup condition on the trial and test classes, a hypothesis that a weak formulation would ordinarily need, it is replaced by a containment between the two classes.
- We transfer the obtained bound to the total variation distance between the marginals of the estimated and the exactly guided samplers on the estimation window (Corollary 6), so that an error in the estimated guidance is controlled as an error in the law that is sampled.

Related work

Inference-time alignment. The oldest way of steering a diffusion sampler is to add a heuristic drift: classifier guidance (Dhariwal & Nichol, 2021) uses the gradient of a noise-conditioned classifier, and classifier-free guidance (Ho & Salimans, 2021) extrapolates between conditional and unconditional scores. Both are effective but not exact: classifier guidance reproduces the conditional h-transform only with an exact noise-conditioned classifier at unit scale and the matching initial law, and other scales change the terminal law by an unquantified amount. The alternatives split by what they modify. Fine-tuning changes the weights, by reinforcement learning or stochastic control, and is reviewed by Uehara et al. (2024). Sequential Monte Carlo leaves the model alone and corrects a heuristic proposal by importance weights, as in the twisted sampler of Wu et al. (2023), paying for exactness with a particle system and with weight degeneracy in high dimension. The line closest to ours keeps the sampler exact and learns the h-transform itself. Denker et al. (2024) fit a generalised h-transform on top of a frozen model, without a rate, and Zhu et al. (2026) estimate it at sampling time by Monte Carlo, with a total-variation guarantee governed by the number of look-ahead draws rather than by a sample size. Chang et al. (2026) derive convergence rates for a variational estimator of h^*, but their smoothness penalty does not vanish at the target, so the rate reflects a bias–variance trade-off in the penalty weight. In Guo et al. (2026), who estimate h and its gradient through martingale identities, and Kawata et al. (2025), who sample dual-averaging iterates through the h-transform, the drift error enters the bound as a parameter rather than being estimated at a rate. The question here is accordingly not whether the drift correction admits a statistical theory, but what it costs: which rate is attainable when the penalty incurs no bias at all, and what has to be assumed to get it.

Statistical theory for diffusion models. The nonparametric theory of diffusion models has concentrated on the score. Oko et al. (2023) prove minimax-optimal distribution estimation under empirical score matching, and a parallel line takes the score as given and asks what sampling error follows (Chen et al., 2023b; Li et al., 2024). Because rates in the ambient dimension degrade quickly, much recent effort targets the intrinsic dimension instead: Chen et al. (2023a) under an exact manifold hypothesis, Yakovlev & Puchkin (2025) under a substantially relaxed one — a nonparametric Gaussian mixture in which samples may leave the manifold, with rates set by the intrinsic dimension and valid even when the ambient dimension grows polynomially in n — and Li & Yan (2024) for the sampler, showing DDPM adapts to unknown low-dimensional structure. The guidance is a log-gradient in the same geometry as the score, but is estimated from a regression whose response is the terminal weight, and the error that matters is in the gradient.

Empirical process theory. Localization by local Rademacher complexities (Bartlett et al., 2005; Koltchinskii, 2006), in the form set out by Wainwright (2019), yields multiplicative rather than additive comparisons between empirical and population quantities: on a star hull satisfying a Bernstein condition $Pf^2 \lesssim Pf$, the deviation is governed by a critical radius read off a uniform entropy bound. That theory is built for bounded classes. For suprema with unbounded envelopes, Adamczak (2008) supplies a Talagrand-type tail inequality that applies after truncation and peeling. We need both, the drift coefficient $x + s$ being only sub-Gaussian under ρ_t.

2 Setup

Notation and conventions. We keep the processes of Section 1 and fix $T > 0$. The forward SDE (1) has Gaussian transition kernel $X_\tau \mid X_0 = x_0 \sim \mathcal{N}(\mu_\tau x_0, \sigma_\tau^2 I_d)$ with $\mu_\tau = e^{-\tau}$ and $\sigma_\tau^2 = 1 - e^{-2\tau}$, so $\mu_\tau^2 + \sigma_\tau^2 = 1$, and we write $p_\tau := \mathrm{Law}(X_\tau)$. The reversal (2) is started at $\overleftarrow{X}_0 \sim p_T$, its marginals

$\rho_t := \mathrm{Law}(\overleftarrow{X}_t) = p_{T-t}$ are the laws from which the estimation sample is drawn. Write

$$(\mathcal{L}_t f)(x) := b(t,x)\cdot\nabla f(x) + \Delta f(x), \qquad \mathcal{M} := \partial_t + \mathcal{L}_t \tag{7}$$

for the generator of (2) and the associated space-time operator, the Laplacian rather than $\frac{1}{2}\Delta$ appears because the diffusion coefficient is $\sqrt{2}$. The score s is supplied by the pretrained model, and we treat it as known throughout. As is standard in the statistical analysis of guidance, we treat s as exact, which separates the approximation error of the pretrained model from the error of estimating h^*. The image experiments of Section 7 deliberately drop this idealization and use the model's learned score, to test the estimator where it would be used in practice.

Target and guidance. Given a known weight $w \geq 0$, the target is $q_0 = wp_0/Z$ with $Z = \int w\, p_0$. Doob's h-transform adds the drift $2\nabla \log h^*$ to (2), giving the controlled process

$$d\overleftarrow{Z}_t = \big(\overleftarrow{Z}_t + 2s(t, \overleftarrow{Z}_t) + 2\nabla \log h^*(t, \overleftarrow{Z}_t)\big)dt + \sqrt{2}\, dB_t, \tag{8}$$

whose terminal law is q_0, when started from $\rho_0^* := h^*(0,\cdot)p_T/Z$ by Lemma I.1. Here $h^*(t,x) := \mathbb{E}[w(\overleftarrow{X}_T) \mid \overleftarrow{X}_t = x] = \mathbb{E}[w(X_0) \mid X_{T-t} = x]$ is Doob's h-function and $g^* := \nabla \log h^*$ the guidance. By Bayes' rule for the Gaussian likelihood, $h^*(t,x) = \mathbb{E}_{X_0\sim\pi_{t,x}}[w(X_0)]$ for the posterior $\pi_{t,x}(dx_0) \propto \phi_{\sigma^2_{T-t}}(x - \mu_{T-t}x_0)p_0(dx_0)$.

Assumptions. Exactly two hypotheses are used anywhere below.

Assumption 1 (Weight)**.** $w : \mathbb{R}^d \to \mathbb{R}$ is measurable with $0 < w \leq \bar{B} < \infty$ on $\mathrm{supp}(p_0)$, and there is $\mathrm{s} \in (0,\infty]$ with $\Xi_{\mathrm{s}} := \mathbb{E}_{p_0}[w^{-\mathrm{s}}] < \infty$, where $\mathrm{s} = \infty$ is read as $w \geq B > 0$ on $\mathrm{supp}(p_0)$.

Assumption 2 (Compact reference support)**.** p_0 is supported in the cube $\{x_0 : |x_0|_\infty \leq 1\}$.

Reward and likelihood tilts with bounded reward are of this form, Ξ_{s} quantifies how close w comes to zero, degenerating at $\mathrm{s} = \infty$ to a uniform lower bound. Assumption 2 is the usual compact-support hypothesis, and its only role is to make the posterior compactly supported, hence to yield the regularity bounds of Appendix B. No smoothness of w is assumed anywhere: the smoothness index β below is a free parameter of the construction, and the requisite derivative envelopes for h^* are derived from Assumptions 1–2 (Lemma F.8, Step 1). Assumption 1 lets w vanish, which the corresponding hypotheses of prior work do not: Chang et al. (2026) and Kawata et al. (2025) bound w and the density ratio, respectively, from both sides, and Zhu et al. (2026) bounds h from below. Guo et al. (2026) work under conditions of a different kind, strong log-concavity of the marginals and of h with the score and the guidance error controlled in sup norm, which are not directly comparable to ours. The order s is free here and enters the rate, so how far w may vanish is traded explicitly against how fast the guidance converges, a trade we measure in Appendix J.10.

Window and function spaces. As $t \uparrow T$ the noise level $\sigma_{T-t} \to 0$ and estimation is ill-posed, as $t \downarrow 0$ there is nothing to recover. We fix an early–late window

$$I := [t_0, t_1] \subset (0,T), \quad \ell := t_1 - t_0 \leq 1, \quad \sigma_* := \sigma_{T-t_1} > 0, \tag{9}$$

and a probability weight ν on I, taken to be normalized Lebesgue measure. For each t, $L^2(\rho_t)$ and $H^1(\rho_t) = \{f : f, \nabla f \in L^2(\rho_t)\}$ carry the reverse marginal as reference measure, and $H^{-1}(\rho_t)$ is the dual of $H^1(\rho_t)$ under the pairing

$$\|F\|_{H^{-1}(\rho_t)} := \sup_{0\neq\psi\in H^1(\rho_t)} \frac{\langle F,\psi\rangle_{\rho_t}}{\|\psi\|_{H^1(\rho_t)}} \quad \leq \quad \|F\|_{\rho_t}. \tag{10}$$

On the window we use the space-time spaces $\mathcal{V} := L^2_\nu(I; H^1(\rho_t))$ and its dual $\mathcal{V}^*$, with $\|F\|^2_{\mathcal{V}^*} = \int_I \|F(t,\cdot)\|^2_{H^{-1}(\rho_t)}\nu(dt)$. Constants written σ^{-c} are of the form $C_\star\sigma_*^{-c}$, where c is absolute and $C_\star$ depends only on $(d, \beta, \bar{B}, \ell, C_t, \kappa_\Theta)$ and not on n, N, ζ, δ or σ_*, the relations $\lesssim$ and $\asymp$ hide constants of the same form.

3 Harmonicity and its weak form

Write $\mathcal{R}_t[h] := (\mathcal{M}h)(t,\cdot)$ for the residual and

$$\mathcal{D}(h) := \int_I \|h(t,\cdot) - h^*(t,\cdot)\|^2_{\rho_t}\nu(dt), \qquad \mathcal{P}^w(h) := \|\mathcal{R}[h]\|^2_{\mathcal{V}^*} \tag{11}$$

for the value functional and the weak penalty. Here $\mathcal{D}$ is exactly the population excess least-squares risk, since at fixed t the minimizer of $\mathbb{E}[(h(t, X_{T-t}) - w(X_0))^2]$ is the conditional mean $h^*(t,\cdot)$.

Proposition 1 (Harmonicity and its weak form)**.** *Under Assumptions 1–2, h^* is space-time harmonic for the reverse generator, $\mathcal{M}h^* = 0$ on $(0,T)\times\mathbb{R}^d$ with $h^*(T,\cdot) = w$. Moreover, for $t \in I$, $h \in C^{1,2}$ and $\psi \in H^1(\rho_t)$ of polynomial growth, writing $\varphi := h - h^*$,*

$$\langle \mathcal{R}_t[h], \psi\rangle_{\rho_t} \;=\; a_t(h,\psi) \;:=\; \int_{\mathbb{R}^d} \Big[\partial_t h\,\psi + \big((x+s)\cdot\nabla h\big)\psi - \nabla h\cdot\nabla\psi\Big]\rho_t\,dx, \tag{12}$$

and $|a_t(h,\psi)| \le \big(\|\partial_t\varphi\|_{\rho_t} + \|\nabla\varphi\|_{\rho_t} + M_4\||\nabla\varphi|\|_{L^4(\rho_t)}\big)\|\psi\|_{H^1(\rho_t)}$ with $M_4 \le C\sqrt{d}\sigma_*^{-2}$.

There are three main consequences. The form a_t is computable: it needs first derivatives of h, first derivatives of ψ, and the pretrained score, because the weight generated by integrating Δh by parts against ρ_t is $\nabla\log\rho_t = s$. It is gap-free: $a_t(h^*,\psi) = 0$ for all ψ, so h^* minimizes $\mathcal{D} + \mu\mathcal{P}^w$ for every $\mu \ge 0$ and no bias is traded against variance in the penalty weight. And — the point that drives the rate — the continuity bound is first order in φ, so the penalty inherits a first-order approximation floor: for $n \ge n_0$ and suitable budgets there is an $h_N \in \mathcal{H}_N$ with

$$\|h_N - h^*\|_{L^4} + \sup_{t\in I}\|h_N - h^*\|_{\rho_t} + \||\nabla(h_N - h^*)|\|_{L^4} + \|\partial_t(h_N - h^*)\|_{L^2} \le e_N, \tag{13}$$

$$e_N := 8Q_n\varepsilon_N, \quad \varepsilon_N := \mathrm{A}_\beta N^{-(\beta-1)}, \tag{14}$$

where A_β is the constant, depending only on $\beta, \bar{B}, \sigma_*$ and, through R, polylogarithmically on n, and $Q_n \le \sigma^{-c}(\log n)^4$ is given in (F.14). Such an h_N exists in a class with $\mathbb{V}_N \lesssim N^d\,\mathrm{polylog}(n)$, and gives $\inf_{h\in\mathcal{H}_N}\mathcal{P}^w(h) \lesssim Cd\sigma_*^{-4}e_N^2$ (Lemmas F.1 and F.8). Table 1 places this against the alternatives.

Table 1: Why the residual is measured weakly. The floor is $\inf_{h\in\mathcal{H}_N}\mathcal{P}(h)$ at resolution N, only the last is both gap-free and informative at $\beta = 2$.

Penalty $\mathcal{P}(h)$	order in h	value at h^*	approximation floor
Sobolev, $\int_I \|\nabla h\|^2_{\rho_t}\nu(dt)$	1	$\|\nabla h^*\|^2 \neq 0$	— (nonzero gap for $\lambda > 0$)
strong residual, $\|\mathcal{R}[h]\|^2_{L^2(\nu\otimes\rho)}$	2	0	$N^{-2(\beta-2)}$, non-informative at $\beta = 2$
weak residual, $\|\mathcal{R}[h]\|^2_{\mathcal{V}^*}$	1	0	$N^{-2(\beta-1)}$

4 From residual to gradient

Two things remain: that the weak penalty still controls the gradient, and that the part of it an estimator can see suffices. Both rest on an exact identity. Put $\omega(t) := \nu([t,t_1]) = (t_1 - t)/\ell$, $A(h,\psi) := \int_I a_t(h,\psi(t,\cdot))\,\nu(dt)$, and $E_\omega := \int_I \omega\|\nabla\varphi\|^2_{\rho_t}\nu(dt)$. Itô's formula applied to φ^2 along the reverse process, averaged over a random endpoint $r \sim \nu$, gives (Theorem D.2)

$$2\ell E_\omega \;+\; \|\varphi(t_0,\cdot)\|^2_{\rho_{t_0}} \;=\; \mathcal{D}(h) \;-\; 2\ell\,A(h,\omega\varphi). \tag{15}$$

The endpoint averaging is what produces ω, and is forced: the identity on a fixed $[t_0,r]$ leaves a terminal value at a single instant, which the statistical analysis cannot control. Its cost is that ω vanishes at t_1, so guarantees are stated on

$$I_\delta := [t_0,\, t_1 - \delta\ell], \qquad \delta \in (0,1), \qquad \omega \ge \delta \;\text{ on } I_\delta, \tag{16}$$

covering a $(1-\delta)$ fraction of the window at a cost of δ^{-1} in the constant.

Bounding the cross term in (15) by $\|\mathcal{R}[h]\|_{\mathcal{V}^*}\|\omega\varphi\|_{\mathcal{V}}$ and absorbing already yields coercivity (Corollary D.3). But no estimator sees $\|\mathcal{R}[h]\|_{\mathcal{V}^*}$. What a finite test class Ψ delivers is the projected residual

$$\mathcal{N}_{\Psi,\tau}(h) := \sup_{0\neq\psi\in\Psi}\frac{A(h,\psi)}{\|\psi\|_{\mathcal{V},\tau}} \;\le\; \|\mathcal{R}[h]\|_{\mathcal{V}^*}, \qquad \|\psi\|^2_{\mathcal{V},\tau} := \|\psi\|^2_{\mathcal{V}} + \tau^2\|\psi\|^2_\bullet, \tag{17}$$

with $\|\cdot\|_\bullet$ a positively homogeneous scale on the cone Ψ and τ a slack. Recovering the left-hand side of (17) from the right is a discrete inf-sup condition: the requirement that some $\kappa > 0$ satisfy

$$\|\mathcal{R}[h]\|_{\mathcal{V}^*} \;\le\; \kappa^{-1}\,\mathcal{N}_{\Psi_N}(h) \qquad \text{for all } h \in \mathcal{H}_N, \tag{18}$$

that is, that the finite test class detect a fixed fraction of any residual the trial class can produce. It is the standard price of a weak formulation, and κ would enter the final rate multiplicatively. We do not assume (18). Instead we impose two structural properties on the classes, neither of which requires linearity or convexity:

(S1) **Symmetric cone:** $\psi \in \Psi \Rightarrow -\psi \in \Psi$ and $r\psi \in \Psi$ for $r \geq 0$. This makes the inner maximization closed-form up to a scalar problem (Lemma G.2) and costs nothing: leave the output layer unnormalized.

(S2) **Absorption:** $\omega \cdot (\mathcal{H} - \mathcal{H}) \subseteq \Psi$. This is what replaces (18). It costs a test network of twice the width and one more layer than the trial network, since a difference of two networks is a network (Lemma E.2).

Finally, write Λ_* for the uniform L^4 envelope of the coefficients generated by the integration by parts below,

$$\Lambda_* := \sup_{t\in I} \left(\| |x+s| \|_{L^4(\rho_t)} + \|d + \nabla \cdot s\|_{L^4(\rho_t)} + \| |x+s||s| \|_{L^4(\rho_t)} + \|\partial_t \log \rho_t\|_{L^4(\rho_t)} \right), \quad (19)$$

which under Assumptions 1-2 satisfies $\Lambda_* \lesssim d^2 \sigma_*^{-4}$.

Theorem 2 (Coercivity without an inf-sup constant)**.** *Let Assumptions 1–2 hold, $\mathcal{H}_N, \Psi_N$ satisfy (S1)–(S2), every $h \in \mathcal{H}_N$ be $C^{1,2}$ on $I \times \mathbb{R}^d$ with $|h| \leq 3\bar{B}$ and bounded $\partial_t h, \nabla h, \nabla^2 h$, and $\|\omega(h - h')\|_\bullet \leq C_0$ for all $h, h' \in \mathcal{H}_N$. If $h_N \in \mathcal{H}_N$ satisfies the inequality in* (13) *with some $e_N > 0$, then for every $h \in \mathcal{H}_N$, $\tau \geq 0$, $\delta \in (0,1)$ with $\varphi := h - h^*$,*

$$\int_{I_\delta} \|\nabla\varphi\|^2_{\rho_t} \nu(dt) \leq \frac{C}{\delta} \left(\frac{1+\ell}{\ell} \mathcal{D}(h) + \mathcal{N}_{\Psi,\tau}(h)^2 + C_0^2 \tau^2 + (1 + \Lambda_*^2 + \ell^{-2}) e_N^2 \right), \quad (20)$$

where C absolute and Λ_* as in (19)*. For the classes of Definition 4 the regularity conditions hold and $C_0 = 1$.*

5 Estimator and main result

Let $\varrho(u) = (u \vee 0)^3$ be the rectified cubic activation function, which we will use for our proof sketch of main result below.

Definition 3. (Network class) Put $Q_R := \{x : |x|_\infty \leq R\}$ and $Z := I \times Q_R$. For $W, L, P_0 \in \mathbb{N}$ and $\Lambda, \mathcal{A} \geq 1$, let $\mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A})$ be the set of $f_\vartheta(z) = A_L x^{(L)}(z) + b_L$ with $x^{(0)}(z) := z$ and $x^{(\ell+1)}(z) := \varrho(A_\ell x^{(\ell)}(z) + b_\ell)$, subject to: hidden widths at most W, parameters $\vartheta = (A_0, b_0, \ldots, A_L, b_L)$ bounded in modulus by Λ, at most P_0 of them nonzero and $\sup_{z\in Z} \|x^{(\ell)}(z)\|_\infty \leq \mathcal{A}$ for every hidden layer. We will write $\mathcal{NN}_{\varrho,\Theta}$ for the same class with Θ as the activation of the last hidden layer.

Definition 4. (Trial and test classes) Fix a resolution N, a time degree K, a radius R and budgets $\Lambda, P_0, \mathcal{A}$, and set $W_N \asymp KN^d$ with $W_N \geq d+1$. Let $\Theta \in C^\infty(\mathbb{R}, [-2\bar{B}, 3\bar{B}])$ be a fixed 1-Lipschitz retraction with $\Theta = \mathrm{id}$ on $[-\bar{B}, 2\bar{B}]$ and bounded second derivative, and let χ_R be a fixed smooth coordinatewise clip, the identity on $[-(R-1), R-1]$, of modulus at most R, derivative at most 1 and bounded second derivative. Write $f^\chi(t,x) := f(t, \chi_R(x))$, then

$$\mathcal{H}_N := \{(\Theta \circ g)^\chi : g \in \mathcal{NN}_\varrho(W_N, L, \Lambda, P_0, \mathcal{A}), \sup_{I\times\mathbb{R}^d} |\nabla h| \leq G, \sup_{I\times\mathbb{R}^d} |\partial_t h| \leq G_t\}, \quad (21)$$

$$\Psi_N := \{c\omega\phi^\chi : c \in \mathbb{R}, \phi \in \mathcal{NN}_{\varrho,\Theta}(2W_N, L+1, \Lambda, 2P_0+2, \mathcal{A} \vee 3\bar{B}), \quad (22)$$

$$\sup_{I\times\mathbb{R}^d} (|\phi^\chi| + |\nabla\phi^\chi| + |\partial_t\phi^\chi|) \leq \Gamma)\}, \quad (23)$$

the constraint in (21) being on $h = (\Theta \circ g)^\chi$ itself, with $G := 4\sqrt{d}\bar{B}\sigma_*^{-2}$, $G_t := C_t d\bar{B}(R+1)\sigma_*^{-4}$, $\Gamma := 6\bar{B} + 2G + 2G_t + 6\bar{B}/\ell$ and C_t absolute.

With $\{(X_0^i, \varepsilon^i)\}_{i=1}^n$ i.i.d., $X_{T-t}^i := \mu_{T-t} X_0^i + \sigma_{T-t}\varepsilon^i \sim \rho_t$, and $\hat{a}_t, \hat{\mathcal{D}}, \|\cdot\|_{\hat{\mathcal{V}}}$ the empirical counterparts,

$$\hat{h} \in \operatorname*{argmin}_{h\in\mathcal{H}_N} \hat{\mathcal{D}}(h) + \mu \sup_{\psi\in\Psi_N} \left[\int_I \hat{a}_t(h,\psi)\nu(dt) - \frac{\alpha}{2}\|\psi\|^2_{\hat{\mathcal{V}},\delta_n} \right], \ \hat{g} := \Pi_{G^*}\left(\frac{\nabla\hat{h}}{\hat{h} \vee \mathrm{b}} \right). \quad (24)$$

For $\psi \in \Psi_N$ let $\|\psi\|_\bullet := \inf\{|c| : \psi = c\omega\phi^\chi \text{ as in (23)}\}$ be its cone scale, the scale entering (24). Writing $\mathbb{V}_N$ for the metric entropy index, in the supremum norm and normalized by $\log(e/\varepsilon)$, of

the first order objects $(h, \partial_t h, \nabla h)$, $h \in \mathcal{H}_N$, and $(\psi, \nabla\psi)$, $\psi \in \Psi_N$, $\|\psi\|_\bullet \le 1$ (see (E.14)), it is admissible that

$$\mathbb{V}_N \lesssim N^d \,\mathrm{polylog},(n) \tag{25}$$

once R, K and the budgets are as specified below. The test class is normalized by the cone scale.

The spatial radius, time degree and the approximation constant are

$$R := 2 + \sqrt{(8\beta + 16)\log(2dn)}, \quad \eta := \min\{\frac{\ell}{2}, \frac{\pi\sigma_*^4}{12d(5R+3)}\}, \quad K := \left\lceil \frac{\ell}{\eta}(\beta+2)\log n \right\rceil, \tag{26}$$

$$\mathrm{A}_\beta := C_\beta \bar{B}\sigma_*^{-2\beta}(2R)^{\beta-1}, \quad C_\beta := \mathfrak{B}_{\beta+1}\beta!\,(2\sec(\pi/6))^{\beta+1}, \tag{27}$$

with $\mathfrak{B}_{\beta+1}$ the Bell number. Here η is the width of a complex neighborhood of I on which h^* is analytic in t, $C_\beta\bar{B}\sigma_*^{-2\beta}$ bounds its x-derivatives of order at most β there, and $(2R)^{\beta-1}$ is the cost of rescaling Q_R to the unit cube. Thus $R \asymp \sqrt{\log n}$, $K \le \sigma^{-c}(\log n)^{3/2}$ and $\mathrm{A}_\beta = \mathrm{polylog}(n)$ at fixed β.

Theorem 5 (Guidance rate)**.** *Grant Assumptions 1–2 and fix $\zeta, \delta \in (0,1)$ and $\kappa \ge 1$. There are constants $C^\#, c^\# \ge 1$ and $L_0 \in \mathbb{N}$ depending only on (d, β), a constant $C_0 = C_\star\sigma_*^{-c}$ and n_0 not depending on n, such that the following holds for every $n \ge n_0$. Let R, K and A_β be given by* (26)–(27)*, and choose*

$$\begin{aligned} &N := \left\lceil (n\mathrm{A}_\beta^2)^{\frac{1}{2(\beta-1)+d}} \right\rceil,\ \varepsilon_n := \mathrm{A}_\beta^{\frac{2d}{2(\beta-1)+d}} n^{-\frac{2(\beta-1)}{2(\beta-1)+d}},\ \mu, \alpha \in [\kappa^{-1}, \kappa], \\ &\mathrm{b} := (\varepsilon_n/\Xi_{\mathrm{s}})^{\frac{1}{\mathrm{s}+4}} \wedge \bar{B},\ \delta_n^2 := C_0\frac{(\mathbb{V}_N + \log(1/\zeta))\log^2(n/\zeta)}{n}. \end{aligned} \tag{28}$$

Let $\mathcal{H}_N, \Psi_N$ be the classes of Definition 4 with these N, R, K and the budgets

$$\begin{aligned} &W_N := \lceil C^\# K(N+1)^d \rceil,\ L := L_0 + \lceil \log_2(K+1) \rceil + 1, \\ &P_0 := \lceil C^\# K((N+1)^d + L) \rceil,\ \Lambda := \mathcal{A} := n^{c^\#}, \end{aligned} \tag{29}$$

let $\mathbb{V}_N$ be given by (25)*, and let $\hat{h}, \hat{g}$ be the estimator* (24)*. Then, with probability at least $1-\zeta$*

$$\|\hat{g} - g^*\|^2_{L^2(\nu\otimes\rho; I_\delta)} \le \delta^{-1}\sigma^{-c}\log^8(en/\zeta)\Xi_{\mathrm{s}}^{\frac{4}{\mathrm{s}+4}}\varepsilon_n^{\frac{\mathrm{s}}{\mathrm{s}+4}}. \tag{30}$$

In the boundary case $\mathrm{s} = \infty$, where $w \ge B > 0$ on $\mathrm{supp}(p_0)$, the right hand side is $\delta^{-1}B^{-4}\sigma^{-c}\log^8(en/\zeta)\varepsilon_n$, of order $n^{-2/(d+2)}$ at $\beta = 2$ up to logarithmic factors.

Proof sketch. Write $\Pi_\Psi(h) := \mathcal{N}_{\Psi_N,\delta_n}(h)^2$ and let $h_N \in \mathcal{H}_N$ be the approximant of Section 3, satisfying (13). Since $Q_n \le \sigma^{-c}(\log n)^4$ and $\varepsilon_N^2 \le \varepsilon_n$ for the N of (28), $e_N^2 \le \sigma^{-c}(\log n)^8\varepsilon_n$. The floor gives $\mathcal{D}(h_N) \le e_N^2$ directly, and (17) together with the penalty bound of Section 3 gives $\Pi_\Psi(h_N) \le \mathcal{P}^w(h_N) \le Cd\sigma_*^{-4}e_N^2$. The argument runs on the event of the deviation bound (Theorem H.7), which holds with probability $1-\zeta$ and is proved, not assumed.

Step 1 (the inner maximum is the projected residual). Since Ψ_N is a symmetric cone (S1), the supremum in (24) factors into a radial and a directional part, the radial part being a scalar quadratic (Lemma G.2). The deviation bound transfers between $\hat{a}$ and a and between the empirical and population $\mathcal{V}$-norms multiplicatively, pinning the inner supremum $\hat{F}(h)$ between $\Pi_\Psi(h)/(8\alpha) - \delta_n^2/(2\alpha)$ and $(4\Pi_\Psi(h) + 8\delta_n^2)/\alpha$, no additive term appears outside the ridge, which is what the ridge secures.

Step 2 (basic inequality). $\hat{h}$ minimizes and h_N is feasible, so $\hat{\mathcal{D}}(\hat{h}) + \mu\hat{F}(\hat{h}) \le \hat{\mathcal{D}}(h_N) + \mu\hat{F}(h_N)$. Subtracting $\hat{\mathcal{D}}(h^*)$ and applying Step 1 gives

$$\mathcal{D}(\hat{h}) + \Pi_\Psi(\hat{h}) \ \lesssim\ \mathcal{D}(h_N) + \Pi_\Psi(h_N) + \delta_n^2 \ \lesssim\ d\sigma_*^{-4}e_N^2 + \delta_n^2. \tag{31}$$

Calibration enters only here, and only through $\mu/\alpha \in [\kappa^{-2}, \kappa^2]$. No choice of (μ, α) trades bias against variance, because gap-freeness left no bias to trade.

Step 3 (coercivity). Every $h \in \mathcal{H}_N$ is $C^{1,2}$ with bounded derivatives up to second order and $\|\omega(h - h')\|_\bullet \le 1$ (Proposition E.4), so Theorem 2 applies to $\hat{h}$ at $\tau = \delta_n$ with $C_0 = 1$. With (31) this gives $\int_{I_\delta} \|\nabla\hat{\varphi}\|^2_{\rho_t}\nu(dt) + \mathcal{D}(\hat{h}) \lesssim \delta^{-1}\sigma^{-c}(e_N^2 + \delta_n^2)$.

Step 4 (balance). With the budgets (29), $\mathbb{V}_N \le \sigma^{-c} N^d (\log n)^{7/2}$, so $\delta_n^2 \le \sigma^{-c}(N^d/n)\log^{13/2}(en/\zeta)$. Balancing $\varepsilon_N^2 = \mathrm{A}_\beta^2 N^{-2(\beta-1)}$ against N^d/n gives $N^{2(\beta-1)+d} \asymp n\mathrm{A}_\beta^2$, the choice in (28), with common value of order ε_n, hence $e_N^2 + \delta_n^2 \le \sigma^{-c}\log^8(en/\zeta)\,\varepsilon_n$. Here the weak form is decisive: with the strong penalty the floor would be $N^{-2(\beta-2)}$ and the balance empty at $\beta = 2$.

Step 5 (clipping). Corollary D.4 converts gradient and value bounds into $\delta^{-1}\sigma^{-c}\big(\log^8(en/\zeta)\,\mathrm{b}^{-4}\varepsilon_n + \Xi_{\mathrm{s}}\mathrm{b}^{\mathrm{s}}\big)$, the two terms being equal to $\Xi_{\mathrm{s}}^{4/(\mathrm{s}+4)}\varepsilon_n^{\mathrm{s}/(\mathrm{s}+4)}$, up to the logarithmic factor, at $\mathrm{b} = (\varepsilon_n/\Xi_{\mathrm{s}})^{1/(\mathrm{s}+4)}$. If this value exceeds $\bar{B}$, then $\Xi_{\mathrm{s}}\mathrm{b}^{\mathrm{s}} \ge \Xi_{\mathrm{s}}\bar{B}^{\mathrm{s}} \ge 1$ because $w \le \bar{B}$, and the trivial bound $|\hat{g} - g^*| \le 2G^*$ already gives (30). At $\mathrm{s} = \infty$, $\mathrm{b} = B \le h^*$ and the second term vanishes. □

6 From guidance error to sampling error

Corollary 6 (Local stability of the controlled marginals)**.** *Let $J := [t_0, \bar{t}] \subseteq I_\delta$ and, conditionally on the sample, let $\mathbb{Q}^*$ and $\hat{\mathbb{Q}}$ be the laws on $C(J;\mathbb{R}^d)$ of the solutions of*

$$d\overleftarrow{Z}_t = (b + 2g^*)(t, \overleftarrow{Z}_t)dt + \sqrt{2}dB_t, \quad d\overleftarrow{Y}_t = (b + 2\hat{g})(t, \overleftarrow{Y}_t)dt + \sqrt{2}dB_t, \tag{32}$$

with b as in (2), both started at time t_0 from $\rho_{t_0}^* := h^*(t_0,\cdot)\rho_{t_0}/Z$, and let ρ_t^* and $\hat{\rho}_t$ be their marginals at time $t \in J$. Then the two laws are equivalent, $\mathrm{KL}(\mathbb{Q}^*\|\hat{\mathbb{Q}}) = \int_J \|\hat{g} - g^*\|_{\rho_t^*}^2 dt \le \ell\Upsilon\|\hat{g} - g^*\|_{L^2(\nu\otimes\rho;J)}^2$ with $\Upsilon := \sup_{t\in J}\|d\rho_t^*/d\rho_t\|_\infty \le \bar{B}\,\Xi_{\mathrm{s}}^{1/\mathrm{s}}$, and consequently, on the event of Theorem 5 and up to σ^{-c} and $\mathrm{polylog}(n)$ factors,

$$\mathrm{TV}(\rho_{\bar{t}}^*, \hat{\rho}_{\bar{t}})^2 \;\lesssim\; \delta^{-1}\sigma^{-c}\log^8(en/\zeta)\ell\bar{B}\,\Xi_{\mathrm{s}}^{\frac{1}{\mathrm{s}}+\frac{4}{\mathrm{s}+4}}\,\varepsilon_n^{\frac{\mathrm{s}}{\mathrm{s}+4}}, \tag{33}$$

which in the boundary case $\mathrm{s} = \infty$ reads $\delta^{-1}\ell\bar{B}B^{-5}\sigma^{-c}\log^8(en/\zeta)\varepsilon_n$.

The projection Π_{G^*} is what keeps the Novikov constant independent of n, and the reweighting bound $\Upsilon \le \bar{B}/Z$ is what lets a guarantee stated under the uncontrolled marginals ρ_t — the ones we can sample — be read under the controlled ones.

7 Numerical Experiments

We test the rate of Theorem 5 where every quantity it bounds is closed-form, and apply the estimator (24) to pretrained image diffusion models, where none is (details in Appendix J). The image experiments ask one question: does the estimator the theorem analyses work on a real pretrained model, and at what cost to the base distribution? We measure that cost and compare with the two estimators closest to it in kind.

The rate. p_0 is atomic on 512 points in a ball of radius 0.8 and $0.5 \le w \le 3$, so Assumptions 1–2 hold with $\mathrm{s} = \infty$, and the score is closed-form. In tensor cubic B-splines ($\beta = 4$, $\mathbb{V}_N \asymp N^d$), with (N, μ) chosen by oracle, the measured exponent over three decades of n is 0.842 ± 0.026 (Figure 1a), against 0.857 predicted and 0.667 for recovering a first derivative from noisy values (Stone, 1982). At $n = 10^6$, the strong residual, unpenalized least squares and the Sobolev penalty give 1.4, 5.9 and 8.5 times the guidance error of the weak residual. At $\beta = 3$, the strong residual's approximation floor decays as $N^{-1.0}$, against $N^{-1.8}$ for the weak one (Appendix J.4).

Alignment of a pretrained model. On a model whose score is only a network, we tilt the CIFAR-10 DDPM of Ho et al. (2020) (100 DDIM steps, Song et al., 2021a) towards animals, with a bounded, positive weight w built from a classifier (Appendix J.1). The reference sample is drawn from the model, so the score entering the penalty is the pretrained network itself, and the trial and test classes are convolutional instances of Definition 4, with absorption holding by construction (Appendix J.2). We sample along $c\,\nabla\log\hat{h}$: $c = 1$ is the case covered by Theorem 5, while $c > 1$, common in practice (Dhariwal & Nichol, 2021), targets a sharper law and is reported separately as a heuristic. Samples are scored by a classifier head that w never saw.

Exact guidance ($c = 1$). The theorem covers a non-degenerate regime: at $c = 1$ the estimated guidance applied only on the estimation window already achieves 11.7% of the exact tilt's reward

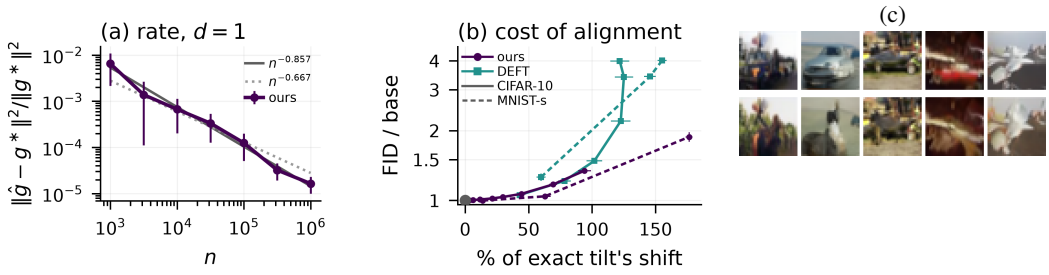


Figure 1: (a) Guidance error at $d = 1$, ± 1 s.d. over 8 seeds, against slopes 0.857 and 0.667. (b) Held-out reward shift, as a share of the exact tilt's, against the FID it costs over that dataset's base (24.5, 7.4), so grey is both base models and lower is cheaper. Colour is the method, line style the dataset, $c \in [0.25, 20]$, bars ± 1 s.d. over 3 seeds, DOIT is off the panel at 4.8 and 21 times its base, absolute values in Figure 3. (c) The five largest rises in held-out probability among 64 samples, base above and ours at $c = 20$ below on the same initial noise, all 64 in Figure 4.

shift, at an FID to the base distribution of 24.7 against 24.5. This measures practical steering rather than the guarantee itself.

Scaled guidance ($c > 1$), a heuristic. Scaling the same direction traces the frontier of Figure 1b: the share of animals rises from 0.653 to 0.831 at $c = 20$, against 0.854 for 32-fold importance resampling at $32\times$ the cost. Under the same weight, sample, sweep and cap we compare DEFT (Denker et al., 2024) and DOIT (Zhu et al., 2026), from their authors' public code (Appendix J.5). Of the three only DEFT is given the gradient of the log weight, ours never differentiating the reward model. Read at a fixed departure the frontiers coincide: 62.5% of the exact tilt's shift against 62.7% at FID 28, 77.0% against 79.2% at FID 30, DOIT reaches 70.6% only at FID 116.6. A sample that changes class keeps its counterpart's layout and palette (Figure 1c). The gain transfers to a head that w never saw: at 94% shift the ratio of the gain on the classifier defining w to the gain on the held-out one is 1.19 for us and 1.20 for DEFT, while for DOIT it is negative, the two heads moving apart.

MNIST. On a public MNIST DDPM with the CIFAR-10 architecture, tilted towards odd digits (Appendix J.6), the ordering is ours: at $c = 20$ the share of odd digits rises from 0.523 to 0.819 at FID 13.8, which is 176% of the exact tilt's shift, where DEFT at the same FID reaches 84% and needs FID 29.5 for 155% (Figure 1b). DOIT reaches 296% at FID 153, its cap binding on every step (Appendix J.8).

8 Discussion and Limitations

Constants, and the choice of β. Theorem 5 is a family indexed by β, and β is not free: A_β enters (28) as $\mathrm{A}_\beta^{2d/(2(\beta-1)+d)}$ rather than as a prefactor, and grows factorially, so raising β improves the exponent and worsens the constant, and at moderate n and large d the smallest admissible value wins the trade. We therefore read $\beta = 2$, giving $n^{-2/(d+2)}$, as the operative case — and it is exactly the case in which the strong residual is non-informative (Table 1, see Appendix J.9).

The score enters as a known coefficient. This is what makes the penalty computable from a frozen model and nothing else: the weak form needs exactly the quantity the pretrained sampler already supplies, and no second network is fitted to obtain it. The theory reads that coefficient as exact, and the closed-form instance supplies it exactly. On images it is the base network's own output, and the estimator remains effective there, which is evidence that the construction tolerates the score error a pretrained model carries. Quantifying how the guarantees degrade with that error is a natural next step.

Towards intrinsic dimension. The ambient dimension enters the exponent only through the approximation budget $\mathbb{V}_N \asymp N^d$. Since natural data concentrate near low-dimensional sets (Pope et al., 2021) and score estimation adapts to the intrinsic dimension (Chen et al., 2023a; Yakovlev & Puchkin, 2025), we expect the same theorem with d replaced by an intrinsic k, as the bilinear form is integrated against the d-dimensional ρ_t, the gain has to come from the approximation side.

AI USE STATEMENT

In this work, we used generative AI tools, large language models including Claude Opus 5, for two kinds of task. For the text: writing and editing, work with LaTeX, spelling and punctuation, and improving the exposition of the theoretical arguments and shortening proofs. For the experiments: implementing the estimator and the evaluation pipeline, re-implementing the two baselines against their authors' public code, and producing the figures and the analysis scripts. We have reviewed all AI-assisted work. The text was read, revised and proofread by the authors throughout, and was substantially altered in the course of writing. The experimental software was verified by reproducing what it reports: every number in the paper is rebuilt from the stored run records by the released code, with no step that depends on re-running the models, and this rebuild is checked. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical and involves no human subjects, no data collection, and no new datasets, its experiments use publicly available pretrained models and the CIFAR-10 and MNIST datasets. It analyses the statistical cost of steering a pretrained generative model toward a reweighted target.

REPRODUCIBILITY STATEMENT

All results are stated with their hypotheses in the main text and proved in full in the appendices, with complete proofs of every intermediate lemma. Assumptions 1 and 2 are the only hypotheses used, and are stated in Section 2. The constants that determine the rate are given explicitly, the remaining prefactors are of the form $C_\star \sigma_*^{-c}$ fixed in Section 2, displayed in their dependence on σ_*, while the dependence of $C_\star$ on $(d, \beta, \bar{B}, \ell)$ is not made explicit. Appendix B collects the regularity facts derived from those assumptions, Appendices C–D the structural results, Appendices E–F the architecture and approximation constructions, and Appendices G–I the statistical analysis. External results are restated in the exact form used (Theorems H.3 and H.4). The accompanying code runs both experiments from a single entry point. In the synthetic one the ground truth, the score included, is closed-form and every number is an average over 8 seeds, the image ones use public checkpoints, with the configuration, the empirical objective and the evaluation protocol in Appendices J.1–J.3 and J.7, and every number computed on 3 seeds of 2000 samples with paired initial noise. Both baselines are implemented against their authors' public code at the commits named in Appendix J.5, under the weight, reference sample, guidance sweep and step cap shared by every method, where a method's own requirements conflict with that protocol, as the sampler does for Zhu et al. (2026), the departure is stated and measured separately.

REFERENCES

Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

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

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Jinyuan Chang, Chenguang Duan, Yuling Jiao, Yi Xu, and Jerry Zhijian Yang. Inference-time alignment for diffusion models via variationally stable Doob's matching. *arXiv preprint arXiv:2601.06514*, 2026.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the*

40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023b.

Carl de Boor. *A Practical Guide to Splines*. Springer, 1978.

Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Vincent Dutordoir, Riccardo Barbano, Emile Mathieu, Urszula Julia Komorowska, and Pietro Liò. DEFT: Efficient fine-tuning of diffusion models by learning the generalised h-transform. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19636–19682, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Joseph L. Doob. *Classical Potential Theory and Its Probabilistic Counterpart*. Springer, 1984.

Daniel Voigt Godoy. `dvgodoy/ddpm-cifar10-32-mnist`: DDPM checkpoint fine-tuned on MNIST. `https://huggingface.co/dvgodoy/ddpm-cifar10-32-mnist`, 2023.

Google. `google/ddpm-cifar10-32`: Unconditional DDPM checkpoint for CIFAR-10. `https://huggingface.co/google/ddpm-cifar10-32`, 2022.

Ingo Gühring and Mones Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021.

Zhengyi Guo, Wenpin Tang, and Renyuan Xu. Conditional diffusion guidance under hard constraint: A stochastic analysis approach. *arXiv preprint arXiv:2602.05533*, 2026.

Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

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

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1991.

Ryotaro Kawata, Kazusato Oko, Atsushi Nitanda, and Taiji Suzuki. Direct distributional optimization for provable alignment of diffusion models. In *International Conference on Learning Representations*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer, Berlin, 1991.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126297–126331, 2024.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *International Conference on Learning Representations*, 2024.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582, 2023.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition, 2003.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, 2013.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024.

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31372–31403, 2023.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5824–5891, 2025.

Qijie Zhu, Zeqi Ye, Han Liu, Zhaoran Wang, and Minshuo Chen. Training-free adaptation of diffusion models via Doob's h-transform. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2026.

A NOTATION RECAP

For the reader's convenience we recall the objects fixed in Section 2. The forward process is the variance-preserving Ornstein–Uhlenbeck SDE

$$dX_\tau = -X_\tau\, d\tau + \sqrt{2}\, dB_\tau, \quad \tau \in (0,T), \quad X_0 \sim p_0, \tag{A.1}$$

with Gaussian transition kernel

$$X_\tau \mid X_0 = x_0 \sim \mathcal{N}(\mu_\tau x_0, \sigma_\tau^2 I_d), \quad \mu_\tau = e^{-\tau}, \quad \sigma_\tau^2 = 1 - e^{-2\tau}, \tag{A.2}$$

so that $\mu_\tau^2 + \sigma_\tau^2 = 1$, and $p_\tau := \mathrm{Law}(X_\tau)$. The reverse process (2) has marginals $\rho_t = p_{T-t}$ and generator $(\mathcal{L}_t f)(x) = b(t,x)\cdot\nabla f(x) + \Delta f(x)$, with $\mathcal{M} = \partial_t + \mathcal{L}_t$. The Laplacian, rather than $\frac{1}{2}\Delta$, appears because the diffusion coefficient is $\sqrt{2}$: Itô's rule reads $df(t,X_t) = (\partial_t f + b\cdot\nabla f + \Delta f)\, dt + \sqrt{2}\,\nabla f \cdot dB_t$, the second-order term being $\frac{1}{2}\mathrm{tr}(2I\nabla^2 f) = \Delta f$.

Definition A.1 (Doob's h-function)**.** For $t \in [0,T]$ and $x \in \mathbb{R}^d$,

$$h^*(t,x) := \mathbb{E}[w(\overleftarrow{X}_T) \mid \overleftarrow{X}_t = x] = \mathbb{E}[w(X_0) \mid X_{T-t} = x], \tag{A.3}$$

the two expressions coinciding because $(\overleftarrow{X}_t, \overleftarrow{X}_T) \overset{d}{=} (X_{T-t}, X_0)$ under time reversal. The guidance is $g^* := \nabla \log h^*$.

By Bayes' rule for (A.2), the conditional law of X_0 given $X_{T-t} = x$ is

$$\pi_{t,x}(dx_0) \propto \phi_{\sigma^2_{T-t}}(x - \mu_{T-t}x_0)\, p_0(dx_0), \tag{A.4}$$

and $h^*(t,x) = \mathbb{E}_{X_0\sim\pi_{t,x}}[w(X_0)]$, so $0 < w \le \bar{B}$ implies $0 < h^* \le \bar{B}$.

We write $\|\varphi\|^2_{L^2(\nu\otimes\rho)} := \int_I \|\varphi\|^2_{\rho_t}\nu(dt)$, similarly for L^4, and $\|\cdot\|_{\mathcal{C}^\ell}$ for the supremum norm of derivatives up to order ℓ (the reservation of $\mathcal{H}$ for hypothesis classes and of H^1 for Sobolev spaces makes this notation necessary). Throughout, $\beta \ge 2$ is an integer and the approximation analysis of Appendix F measures h^* in $W^{\beta,\infty}$ uniformly on the window. This is not an assumption: by Proposition B.2, $h^* \in C^\infty((0,T)\times\mathbb{R}^d)$, and Step 1 of Lemma F.8 produces from Assumptions 1–2 alone the explicit envelope $\sup|\partial_x^\gamma h^*| \le C_\beta \bar{B}\sigma_*^{-2|\gamma|}$ at every order $|\gamma| \le \beta$. Thus β is a free parameter of the construction, traded against a constant growing factorially in β.

Finally, we write

$$I^\circ := [t_0, \tfrac{t_0+t_1}{2}] \tag{A.5}$$

for the first half of the window (9). It is where the weight ω of Theorem D.1 satisfies $\omega \ge 1/2$, which is why the guarantees below are stated there. Remark I.3 explains how they transfer to a prescribed sampling window.

B REGULARITY OF THE DOOB h-FUNCTION

Lemma B.1. *(Score identity and drift envelope) Under Assumptions 1–2, for every $t \in (0,T)$ and $x \in \mathbb{R}^d$,*

$$s(t,x) = -\frac{x - \mu_{T-t}\mathbb{E}[X_0|X_{T-t}=x]}{\sigma^2_{T-t}}, \quad \textit{hence } x + s(t,x) = \left(1 - \frac{1}{\sigma^2_{T-t}}\right)x + \frac{\mu_{T-t}}{\sigma^2_{T-t}}\mathbb{E}[X_0|x]. \tag{B.1}$$

Consequently there is an absolute constant C such that, for all $t \in I$,

$$\|\|x + s(t,\cdot)\|\|_{\psi_2,\rho_t} \le \frac{C\sqrt{d}}{\sigma^2_{T-t}}, \quad M_4 := \sup_{t\in I}\|\|x + s(t,\cdot)\|\|_{L^4(\rho_t)} \le \frac{C\sqrt{d}}{\sigma_*^2} < \infty. \tag{B.2}$$

Proof. Differentiating $p_\tau(x) = \int \phi_{\sigma_\tau^2}(x - \mu_\tau x_0)p_0(dx_0)$ under the integral (legitimate by compact support and Gaussian tails) gives $\nabla \log p_\tau(x) = \int(-\frac{x-\mu_\tau x_0}{\sigma_\tau^2})\pi_{\tau,x}(dx_0)$, which is (B.1) at $\tau = T - t$, the second display follows from $b - s = x + s$ by rearrangement.

By Assumption 2, $|X_0| \le \sqrt{d}$ almost surely, hence $|\mathbb{E}[X_0|x]| \le \sqrt{d}$ for every x. Under ρ_t we may write $X_{T-t} = \mu_{T-t}X_0 + \sigma_{T-t}\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_d)$ independent of X_0, so by the triangle inequality for $\|\cdot\|_{\psi_2}$ and $\mu_{T-t}, \sigma_{T-t} \le 1$, $\||X_{T-t}|\|_{\psi_2} \le \mu_{T-t}\sqrt{d} + \sigma_{T-t}\||\varepsilon|\|_{\psi_2} \le C\sqrt{d}$. Applying the triangle inequality to the second identity in (B.1)

$$\||x+s|\|_{\psi_2,\rho_t} \le \left|1 - \frac{1}{\sigma_{T-t}^2}\right| \||X_{T-t}|\|_{\psi_2} + \frac{\mu_{T-t}}{\sigma_{T-t}^2}\sqrt{d} \le \frac{C\sqrt{d}}{\sigma_{T-t}^2}, \tag{B.3}$$

using $\sigma_{T-t} \le 1$. Finally $\|Z\|_{L^4} \le C\|Z\|_{\psi_2}$ for any Z, and $\sigma_{T-t} \ge \sigma_*$ on I. □

Lemma B.1 does more than make M_4 finite: it supplies a sub-Gaussian envelope for the unbounded coefficient of the weak form, which is what the deviation bound of Section H needs.

The Gaussian likelihood smooths the bounded terminal datum w. The following makes the smoothness and its σ-scaling quantitative. The first-derivative identity is Tweedie's formula and it exhibits the guidance as a posterior covariance.

Proposition B.2. *(Smoothness and derivative bounds) Under Assumptions 1–2, $h^* \in C^\infty((0,T)\times\mathbb{R}^d)$, and for all $t \in (0,T)$, $x \in \mathbb{R}^d$ and coordinates k, l*

$$\partial_{x_k} h^*(t,x) = \frac{\mu_{T-t}}{\sigma_{T-t}^2}\operatorname{Cov}(w(X_0), X_{0,k}|X_{T-t} = x), \tag{B.4}$$

where the covariance is under the posterior (A.4)*. Consequently*

$$0 < h^* \le \bar{B}, \quad |\partial_{x_k} h^*| \le \frac{2\bar{B}}{\sigma_{T-t}^2}, \quad |\partial^2_{x_k x_l} h^*| \le \frac{C_2\bar{B}}{\sigma_{T-t}^4}, \tag{B.5}$$

for an absolute constant C_2. Moreover the log-density of the reverse marginal satisfies

$$-\frac{1}{\sigma_{T-t}^2} I_d \preceq \nabla^2 \log \rho_t(x) = -\frac{1}{\sigma_{T-t}^2} I_d + \frac{\mu_{T-t}^2}{\sigma_{T-t}^4}\operatorname{Cov}(X_0|x) \preceq \left(\frac{d\mu_{T-t}^2}{\sigma_{T-t}^4} - \frac{1}{\sigma_{T-t}^2}\right) I_d. \tag{B.6}$$

Proof. Smoothness: $h^*(t,x) = (\int \phi_{\sigma_{T-t}^2}(x - \mu_{T-t}x_0)w(x_0)p_0(x_0)dx_0)/(\int \phi_{\sigma_{T-t}^2}(x - \mu_{T-t}x_0)p_0(x_0)dx_0)$ is a ratio of Gaussian convolutions with a strictly positive denominator on $(0,T)$. Both are C^∞ in x by differentiation under the integral, justified by the compact support and Gaussian tails, and the denominator is positive, so $h^* \in C^\infty$. Smoothness in t is likewise clear since μ_{T-t}, σ_{T-t} are smooth on $(0,T)$.

Tweedie identity: From (A.4), $\partial_{x_k}\log \pi_{t,x}(x_0) = \partial_{x_k}\left[-\frac{|x-\mu_{T-t}x_0|^2}{2\sigma_{T-t}^2} - \log Z_t(x)\right]$ with $Z_t(x) = \int \phi_{\sigma_{T-t}^2}(x - \mu_{T-t}x_0)p_0(x_0)dx_0$. The first term differentiates to $-(x_k - \mu_{T-t}x_{0,k})/\sigma_{T-t}^2$, and $\partial_{x_k}\log Z_t(x) = \int\left(-\frac{x_k - \mu_{T-t}x_{0,k}}{\sigma_{T-t}^2}\right)\pi_{t,x}(x_0)dx_0 = -\frac{x_k}{\sigma_{T-t}^2} + \frac{\mu_{T-t}}{\sigma_{T-t}^2}\mathbb{E}[X_{0,k}|x]$. Subtracting, $\partial_{x_k}\log \pi_{t,x}(x_0) = \frac{\mu_{T-t}}{\sigma_{T-t}^2}(x_{0,k} - \mathbb{E}[X_{0,k}|x])$. Hence, since $\partial_{x_k}\pi_{t,x} = \pi_{t,x}\partial_{x_k}\log \pi_{t,x}$,

$$\begin{aligned}\partial_{x_k} h^* = \int w(x_0)\partial_{x_k}\pi_{t,x}(x_0)dx_0 &= \frac{\mu_{T-t}}{\sigma_{T-t}^2}\int w(x_0)(x_{0,k} - \mathbb{E}[X_{0,k}|x])\pi_{t,x}(x_0)dx_0 \\ &= \frac{\mu_{T-t}}{\sigma_{T-t}^2}\operatorname{Cov}(w(X_0), X_{0,k}|x),\end{aligned}$$

which is (B.4) coordinatewise.

Bounds: With $\bar{w} := \mathbb{E}[w|x]$, Cauchy-Schwarz gives $|\operatorname{Cov}(w, X_{0,k}|x)| \le \|w - \bar{w}\|_\infty\sqrt{\operatorname{Var}(X_{0,k}|x)} \le \bar{B}$, using $\|w - \bar{w}\|_\infty \le \bar{B}$ and $\operatorname{Var}(X_{0,k}|x) \le 1$. With $\mu_{T-t} \le 1$ this yields $|\partial_{x_k} h^*| \le \bar{B}/\sigma_{T-t}^2$, which we record in the weaker form (B.5) the form in which it is used below. Differentiating (B.4) once more and using that every conditional expectation obeys the same Tweedie rule expresses $\partial^2_{x_k,x_l}h^*$ as $\mu_{T-t}^2\sigma_{T-t}^{-4}$ times a fixed combination of posterior central third moments of $(w(X_0), X_{0,k}, X_{0,l})$, each bounded by an absolute multiple of $\bar{B}$ under Assumption 2.

Hessian: By (B.1), $\nabla\log\rho_t(x) = \sigma_{T-t}^{-2}(\mu_{T-t}\mathbb{E}[X_0|x] - x)$, and $\nabla_x\mathbb{E}[X_0|x] = \mu_{T-t}\sigma_{T-t}^{-2}\operatorname{Cov}(X_0|x)$ by the Tweedie rule applied coordinatewise. Combining gives the identity in (B.6). The bounds follow from $0 \preceq \operatorname{Cov}(X_0|x) \preceq \mathbb{E}[|X_0|^2|x]I_d \preceq dI_d$. □

Equation (B.4) exhibits the guidance as a rescaled posterior covariance between the weight and the clean signal. The σ^{-2}, σ^{-4} scalings are the quantitative form of the endpoint singularity that (9) excises.

Lemma B.3. *(The guidance is bounded, with no lower bound on w) Under Assumptions 1–2, write $\pi^w_{t,x}(dx_0) \propto w(x_0)\pi_{t,x}(dx_0)$ for the w-tilted posterior. Then for all $t \in (0,T)$ and $x \in \mathbb{R}^d$*

$$g^*(t,x) = \frac{\mu_{T-t}}{\sigma^2_{T-t}}\left(\mathbb{E}_{\pi^w_{t,x}}[X_0] - \mathbb{E}_{\pi_{t,x}}[X_0]\right), \quad \textit{hence } |g^*(t,x)| \le \frac{2\sqrt{d}}{\sigma^2_{T-t}}, \tag{B.7}$$

and in particular $|g^*(t,x)| \le G^* := 2\sqrt{d}/\sigma^2_*$ for $t \in (0,t_1]$.

Proof. Since $w > 0$ on $\mathrm{supp}(p_0)$ and $\pi_{t,x}$ is supported there, $h^*(t,x) = \mathbb{E}_{\pi_{t,x}}[w] > 0$ and $\pi^w_{t,x}$ is a well-defined probability measure. Dividing (B.4) by h^* and expanding the covariance

$$\frac{\mathrm{Cov}(w(X_0), X_{0,k}|x)}{\mathbb{E}[w(X_0)|x]} = \frac{\mathbb{E}[w(X_0)X_{0,k}|x]}{\mathbb{E}[w(X_0)|x]} - \mathbb{E}[X_{0,k}|x] = \mathbb{E}_{\pi^w_{t,x}}[X_{0,k}] - \mathbb{E}_{\pi_{t,x}}[X_{0,k}], \tag{B.8}$$

which is (B.7). Both expectations are averages of X_0 over probability measures supported in $\mathrm{supp}(p_0) \subseteq \{|x_0|_\infty \le 1\}$, hence lie in the convex hull of that cube and differ by at most $2\sqrt{d}$ in norm. Since $\mu_{T-t} \le 1$ this gives the first bound, and $\sigma_{T-t} \ge \sigma_*$ for $t \le t_1$ gives the second. □

Lemma B.4. *(Negative moments transfer to h^*) Under Assumptions 1–2, for every $t \in (0,T)$ and every $\mathrm{b} > 0$,*

$$\mathbb{E}_{\rho_t}[(h^*)^{-\mathrm{s}}] \le \Xi_{\mathrm{s}}, \quad \textit{hence } \rho_t(h^* < \mathrm{b}) \le \Xi_{\mathrm{s}}\mathrm{b}^{\mathrm{s}}. \tag{B.9}$$

Proof. The map $u \mapsto u^{-\mathrm{s}}$ is convex on $(0,\infty)$, so conditional Jensen applied to $h^*(t,x) = \mathbb{E}[w(X_0)|X_{T-t} = x]$ gives $h^*(t,x)^{-\mathrm{s}} \le \mathbb{E}[w(X_0)^{-\mathrm{s}}|X_{T-t} = x]$ pointwise. Taking expectations over $X_{T-t} \sim \rho_t$ and using the tower property gives $\mathbb{E}_{\rho_t}[(h^*)^{-\mathrm{s}}] \le \mathbb{E}_{p_0}[w^{-\mathrm{s}}] = \Xi_{\mathrm{s}}$, uniformly in t. Markov's inequality applied to $(h^*)^{-\mathrm{s}}$ at level $\mathrm{b}^{-\mathrm{s}}$ gives the second claim. □

C HARMONICITY, THE WEAK FORM, AND GAP-FREENESSS

Proposition C.1. *(Martingale property) Under Assumption 1, the process $L_t := h^*(t, \overleftarrow{X}_t)$, $t \in [t_0, t_1]$ is a martingale for the law $\mathbb{P}$ of the reverse SDE* (2)*, with respect to its natural filtration $\mathcal{F}_t = \sigma(\overleftarrow{X}_s : 0 \le s \le t)$.*

Proof. By the Markov property of (2) and the tower property of conditional expectation

$$\mathbb{E}[w(\overleftarrow{X}_T)|\mathcal{F}_t] = \mathbb{E}[w(\overleftarrow{X}_T)|\overleftarrow{X}_t] = h^*(t, \overleftarrow{X}_t) = L_t, \tag{C.1}$$

using Definition A.1. Thus (L_t) is a Doob martingale – a conditional expectation of the single terminal random variable $w(\overleftarrow{X}_T)$ along the filtration. Integrability is following from $|w| \le \bar{B}$. □

Theorem C.2. *(Harmonicity of h^*) Suppose Assumptions 1–2 hold. Then h^* is space-time harmonic for the reverse generator:*

$$\mathcal{M}h^* = \partial_t h^* + b(t,x)\cdot\nabla h^* + \Delta h^* = 0 \quad \textit{on } (0,T)\times\mathbb{R}^d, \quad h^*(T,\cdot) = w. \tag{C.2}$$

Moreover h^* admits the martingale representation $dh^*(t, \overleftarrow{X}_t) = \sqrt{2}\nabla h^*(t, \overleftarrow{X}_t)\cdot dB_t$.

Proof. $h^* \in C^{1,2}(I\times\mathbb{R}^d)$, by Proposition B.2. Itô's formula for (2) gives

$$dL_t = (\partial_t h^* + \mathcal{L}_t h^*)(t, \overleftarrow{X}_t)dt + \sqrt{2}\nabla h^*(t, \overleftarrow{X}_t)\cdot dB_t = (\mathcal{M}h^*)(t, \overleftarrow{X}_t)dt + \sqrt{2}\nabla h^*(t, \overleftarrow{X}_t)\cdot dB_t. \tag{C.3}$$

By Proposition C.1, (L_t) is a martingale, so its drift part must vanish almost surely: $(\mathcal{M}h^*)(x, \overleftarrow{X}_t) = 0$ for a.e. t, $\mathbb{P}$-a.s. The law of $\overleftarrow{X}_t$ is $\rho_t = p_{T-t}$, a Gaussian mixture with full support om $\mathbb{R}^d$. Since $\mathcal{M}h^*$ is continuous and vanishes ρ_t-a.e. for a dense set of t, it vanishes identically on $(0,T)\times\mathbb{R}^d$. The terminal condition is $h^*(T,x) = \mathbb{E}[w(\overleftarrow{X}_T)|\overleftarrow{X}_T = x] = w(x)$. Substituting $\mathcal{M}h^* = 0$ into (C.2) leaves the representation. □

Definition C.3. (Strong residual and penalized risk) For $h \in C^{1,2}$ the residual is $\mathcal{R}_t[h] := (\mathcal{M}h)(t,\cdot) = \partial_t h + b \cdot \nabla h + \Delta h$. By Theorem C.2, $\mathcal{R}_t[h^*] \equiv 0$ and, writing $\varphi := h - h^*$, $\mathcal{M}\varphi = \mathcal{R}_t[h]$. The population value-fidelity and strong PDE functionals are

$$\mathcal{D}(h) := \int_I \|h(t,\cdot) - h^*(t,\cdot)\|^2_{\rho_t} \nu(dt), \quad \mathcal{P}^s(h) := \int_I \|\mathcal{R}_t[h]\|^2_{\rho_t} \nu(dt). \tag{C.4}$$

The value functional (C.4) is exactly the population excess least-squares risk: at fixed t the minimizer over measurable $h(t,\cdot)$ of $\mathbb{E}[(h(t, X_{T-t}) - w(X_0))^2]$ is the conditional mean $h^*(t,\cdot)$, and the excess risk equals $\|h(t,\cdot) - h^*(t,\cdot)\|^2_{\rho_t}$, so $\mathcal{D}$ is what regression naturally drives down.

Theorem C.4. *(Gap-freeness) Under Assumptions 1–2, for every penalty weight $\mu \geq 0$ the function h^* is the unique minimizer, over all measurable h with $\varphi \in L^2(\nu \otimes \rho)$ and $\mathcal{R}[h] \in L^2(\nu \otimes \rho)$, of the penalized population risk $J_\mu := \mathcal{D} + \mu\mathcal{P}^s$. In particular the penalized target does not move with μ : there is no regularization gap. The same holds with $\mathcal{P}^s$ replaced by any nonnegative functional vanishing at h^*.*

Proof. By Theorem C.2, $\mathcal{R}_t[h^*] = 0$, so $\mathcal{P}^s(h^*) = 0$, and $\mathcal{D}(h^*) = 0$ by definition. Hence, $J_\mu(h^*) = 0$. Since $\mathcal{D} \geq 0$ and $\mathcal{P}^s \geq 0$, we have $J_\mu(h) \geq 0 = J_\mu(h^*)$ for all h and all $\mu \geq 0$, so h^* minimizes. If $J_\mu(h) = 0$ then in particular $\mathcal{D}(h) = 0$, i.e. $\|h - h^*\|_{L^2(\nu\otimes\rho)} = 0$, whence $h = h^*$ ($\nu \otimes \rho$-a.e.). Uniqueness follows. □

Contrast this with a Sobolev regularized objective $J^\lambda(h) := \mathcal{D}(h) + \lambda \int_I \|\nabla h\|^2_{\rho_t} \nu(dt)$, $\lambda > 0$: its penalty does not vanish at h^*. If $\int_I \|\nabla h^*\|^2_{\rho_t} \nu(dt) > 0$, then h^* does not minimize J^λ, since $J^\lambda((1-\epsilon)h^*) - J^\lambda(h^*) = \epsilon^2 \int_I \|h^*\|^2_{\rho_t} \nu(dt) - \lambda(2\epsilon - \epsilon^2) \int_I \|\nabla h^*\|^2_{\rho_t} \nu(dt) < 0$ for all sufficiently small $\epsilon > 0$. The resulting shrinkage bias must be traded against variance in λ. Gap-freeness removes that trade-off.

The bottleneck. Gap-freeness alone does not improve the rate. Since $\mathcal{M}\varphi$ contains $\Delta\varphi$, the best second-derivative control a resolution-N approximant of a $W^{\beta,\infty}$ function supplies is $\|\nabla^2(h_N - h^*)\|_{L^2} \lesssim N^{-(\beta-2)}$, so the corresponding upper bound is $\inf_{h\in\mathcal{H}} \mathcal{P}^s(h)$ is $N^{-2(\beta-2)}$, which is $O(1)$ vacuous at a $\beta = 2$. Removing this floor is the sole purpose of the weak form.

The idea is to test the residual against a function ψ and move one derivative off φ onto ψ by integration by parts. The key algebraic point is that the weight generated by integrating against ρ_t is exactly the base score, because $\nabla \log \rho_t = \nabla \log p_{T-t} = s$.

Lemma C.5. *(Weak form of residual) Let $t \in I$, $h \in C^{1,2}$ and $\psi \in H^1(\rho_t)$, with $h, \nabla h, \psi$ of polynomial growth. Then, with $\nabla \log \rho_t = s(t,\cdot)$ and $b - s = x + s$,*

$$\langle \mathcal{R}_t[h], \psi \rangle_{\rho_t} = \underbrace{\int_{\mathbb{R}^d} [\partial_t h \psi \; + \; ((x+s)\cdot\nabla h)\psi \; - \; \nabla h \cdot \nabla\psi] \, \rho_t dx}_{:= a_t(h,\psi)} . \tag{C.5}$$

This bilinear form a_t is first order in both h and ψ (no Δh and no $\Delta\psi$), and is computable from the covariates and the known score s. Moreover $a_t(h^*, \psi) = 0$ for every ψ, that is, h^* is the weak solution of (C.2).

Proof. Expand $\langle \mathcal{R}_t[h], \psi \rangle_{\rho_t} = \int (\partial_t h + b\cdot\nabla h + \Delta h)\psi\rho_t$ and integrate the Laplacian term by parts. With no boundary contribution at infinity (polynomial growth against the Gaussian-tailed ρ_t)

$$\int \Delta h \psi \rho_t = -\int \nabla h \cdot \nabla(\psi\rho_t) = -\int \nabla h \cdot \nabla\psi \rho_t - \int (\nabla h \cdot \nabla \rho_t)\psi. \tag{C.6}$$

Since $\nabla\rho_t = (\nabla \log \rho_t)\rho_t = s\rho_t$, the last integrand is $(s\cdot\nabla h)\psi\rho_t$. Collecting the three terms

$$\langle \mathcal{R}_t[h], \psi \rangle_{\rho_t} = \int \partial_t h \psi \rho_t + \int ((b-s)\cdot\nabla h)\psi\rho_t - \int \nabla h \cdot \nabla\psi\rho_t, \tag{C.7}$$

and $b - s = (x + 2s) - s = x + s$, which is (C.5). Finally, taking $h = h^*$ and reversing the integration by parts $a_t(h^*, \psi) = \langle \mathcal{M}h^*, \psi \rangle_{\rho_t} = 0$ by Theorem C.2. □

That the integration-by-parts weight is exactly s is what makes the weak residual free: no second derivatives, and the only coefficient it needs is the output of the pre-trained model.

Lemma C.6. *(Continuity) Under the Assumptions 1–2, with $M_4 \le C\sqrt{d}\sigma_*^{-2}$ from Lemma B.1, for all $h \in C^{1,2}$ of polynomial growth, $\psi \in H^1(\rho_t)$ and $\varphi := h - h^*$*

$$|a_t(h,\psi)| \le \big(\|\partial_t\varphi\|_{\rho_t} + \|\nabla\varphi\|_{\rho_t} + M_4\||\nabla\varphi|\|_{L^4(\rho_t)}\big)\,\|\psi\|_{H^1(\rho_t)}. \tag{C.8}$$

Consequently $\|\mathcal{R}_t[h]\|_{H^{-1}(\rho_t)} \le \|\partial_t\varphi\|_{\rho_t} + \|\nabla\varphi\|_{\rho_t} + M_4\||\nabla\varphi|\|_{L^4(\rho_t)}$.

Proof. Because a_t is linear in its first argument and $a_t(h^*,\psi) = 0$ by Lemma C.5, $a_t(h,\psi) = a_t(\varphi,\psi)$. Bound the three terms of $a_t(\varphi,\psi)$ in (C.5):

$$\left|\int \partial_t\varphi\psi\rho_t\right| \le \|\partial_t\varphi\|_{\rho_t}\|\psi\|_{\rho_t}, \tag{C.9}$$

$$\left|\int \nabla\varphi\cdot\nabla\psi\rho_t\right| \le \|\nabla\varphi\|_{\rho_t}\|\nabla\psi\|_{\rho_t}, \tag{C.10}$$

$$\left|\int ((x+s)\cdot\nabla\varphi)\psi\rho_t\right| \le \|(x+s)\cdot\nabla\varphi\|_{\rho_t}\|\psi\|_{\rho_t} \le M_4\|\nabla\varphi\|_{L^4(\rho_t)}\|\psi\|_{\rho_t}, \tag{C.11}$$

where the last line uses Cauchy-Schwarz twice: first $\int((x+s)\cdot\nabla\varphi)\psi\rho_t \le \|(x+s)\cdot\nabla\varphi\|_{\rho_t}\|\psi\|_{\rho_t}$, then $\|(x+s)\cdot\nabla\varphi\|_{\rho_t} \le \||x+s|\|_{L^4(\rho_t)}\||\nabla\varphi|\|_{L^4(\rho_t)} \le M_4\|\nabla\varphi\|_{L^4(\rho_t)}$. Summing and bounding $\|\psi\|_{\rho_t}$, $\|\nabla\psi\|_{\rho_t} \le \|\psi\|_{H^1(\rho_t)}$ gives (C.8). The dual-norm consequence is (10). □

Definition C.7. (Weak residual norm and penalty) The weak (dual) residual norm at time t is $\|\mathcal{R}_t[h]\|_{H^{-1}(\rho_t)}$ as in (10), with $\langle\mathcal{R}_t[h],\psi\rangle_{\rho_t} = a_t(h,\psi)$ by Lemma C.5. The associated population penalty is

$$\mathcal{P}^w(h) := \int_I \|\mathcal{R}_t[h]\|^2_{H^{-1}(\rho_t)}\nu(dt) = \|\mathcal{R}[h]\|^2_{\mathcal{V}^*}. \tag{C.12}$$

By Lemma C.5, $\mathcal{P}^w(h^*) = 0$, so the gap-freeness argument of Theorem C.4 applies verbatim with $\mathcal{P}^w$ in place of $\mathcal{P}^s$. $\mathcal{P}^w$ is measured in the weaker H^{-1} norm and the two remaining tasks are to shown this weaker penalty still controls the gradient and has better approximation floor.

D THE ENERGY IDENTITY AND COERCIVITY

The bridge from residual to gradient is an exact identity obtained by applying Itô's formula to φ^2 along the reverse process.

Theorem D.1. *(Energy identity) Let $\varphi \in C^{1,2}(I\times\mathbb{R}^d)$ be bounded, with $\nabla\varphi \in L^2(dt\otimes\rho)$ and $\mathcal{M}\varphi \in L^1(dt\otimes\rho)$. Then, for every $r \in I$,*

$$2\int_{t_0}^r \|\nabla\varphi(t,\cdot)\|^2_{\rho_t}dt = \|\varphi(r,\cdot)\|^2_{\rho_r} - \|\varphi(t_0,\cdot)\|^2_{\rho_{t_0}} - 2\int_{t_0}^r \langle\varphi(t,\cdot),\mathcal{M}\varphi(t,\cdot)\rangle_{\rho_t}dt. \tag{D.1}$$

Proof. For $\varphi \in C^{1,2}$ a direct computation gives the pointwise identity

$$\mathcal{M}(\varphi^2) = 2\varphi\,\mathcal{M}\varphi + 2|\nabla\varphi|^2. \tag{D.2}$$

Indeed $\partial_t(\varphi^2) = 2\varphi\partial_t\varphi$, $b\cdot\nabla(\varphi^2) = 2\varphi\, b\cdot\nabla\varphi$, and $\Delta(\varphi^2) = 2\varphi\Delta\varphi + 2|\nabla\varphi|^2$. Summing and grouping the factor 2φ against $\mathcal{M}\varphi$ yields (D.2), in which the term $|\nabla\varphi|^2$ is the carré du champ $\Gamma(\varphi,\varphi)$ of $\mathcal{M}$ (Bakry et al., 2014, §1.4.2). Itô's formula for (2) applied to $t\mapsto\varphi^2(t,\overleftarrow{X}_t)$ then reads

$$d[\varphi^2(t,\overleftarrow{X}_t)] = (2\varphi\mathcal{M}\varphi + 2|\nabla\varphi|^2)(t,\overleftarrow{X}_t)dt + 2\sqrt{2}(\varphi\nabla\varphi)(t,\overleftarrow{X}_t)\cdot dB_t. \tag{D.3}$$

For $R > 0$ let $\tau_R := \inf\{t \ge t_0 : |\overleftarrow{X}_t| \ge R\} \wedge r$. On the stochastic interval $[t_0,\tau_R]$ the state stays in the ball of radius R, on which the continuous integrand $\varphi\nabla\varphi$ is bounded, hence $\mathbb{E}\int_{t_0}^{\tau_R}|\varphi\nabla\varphi|^2(t,\overleftarrow{X}_t)\,dt < \infty$ and the stopped stochastic integral $\int_{t_0}^{\cdot\wedge\tau_R}\varphi\nabla\varphi\cdot dB$ is a continuous square-integrable martingale started at 0, so it has mean zero (Karatzas & Shreve, 1991, §3.2),

see also (Øksendal, 2003, Ch. 3). Taking expectations in the stopped version of (D.3), the martingale term drops:

$$\mathbb{E}[\varphi^2(\tau_R, \overleftarrow{X}_{\tau_R})] - \mathbb{E}[\varphi^2(t_0, \overleftarrow{X}_{t_0})] = \mathbb{E}\int_{t_0}^{\tau_R} (2\varphi\mathcal{M}\varphi + 2|\nabla\varphi|^2)(t, \overleftarrow{X}_t)dt. \tag{D.4}$$

Now let $R \to \infty$. Since $|\overleftarrow{X}_t| < \infty$ a.s. on $[t_0, r]$, $\tau_R \uparrow r$ a.s. The left side converges by bounded convergence since φ is bounded. On the right, the $2|\nabla\varphi|^2 \geq 0$ contribution converges by monotone convergence and the $2\varphi\mathcal{M}\varphi$ contribution by dominated convergence, because $|\varphi| \leq \|\varphi\|_\infty$ and $\mathcal{M}\varphi \in L^1(dt \otimes \rho)$ make it absolutely integrable. Using $\overleftarrow{X}_t \sim \rho_t$ to write $\mathbb{E}[G(t, \overleftarrow{X}_t)] = \int G(t,x)\rho_t(dx)$ for the surviving deterministic integrands,

$$\|\varphi(r)\|^2_{\rho_r} - \|\varphi(t_0)\|^2_{\rho_{t_0}} = 2\int_{t_0}^{r} \left(\langle\varphi, \mathcal{M}\varphi\rangle_{\rho_t} + \|\nabla\varphi\|^2_{\rho_t}\right) dt. \tag{D.5}$$

Rearranging gives (D.1). □

We now show that the weak residual controls the gradient. The mechanism is to recognize the cross term in (D.1) as the bilinear form a_t evaluated at the test function $\psi = \varphi$ and to bound it by the H^{-1} norm of the residual times the H^1 norm of φ, absorbing the resulting gradient factor.

Throughout this section and the next, ν is normalized Lebesgue measure on I, $\ell := t_1 - t_0$, $\ell \leq 1$ and

$$\omega(t) := \nu([t, t_1]) = \frac{t_1 - t}{t_1 - t_0} \in [0,1], \quad \omega(t_0) = 1,\ \omega(t_1) = 0,\ \omega \geq \frac{1}{2} \text{ on } I \circ . \tag{D.6}$$

The weight ω is not a modeling choice. The energy identity of Theorem D.1 holds on $[t_0, r]$ for each fixed r and leaves a terminal term $\|\varphi(r,\cdot)\|^2_{\rho_r}$, a value error at a single instant, which the statistical analysis does not control – it controls only the ν-average $\mathcal{D}(h)$. Averaging the identity over a random endpoint $r \sim \nu$ turns that term into exactly $\mathcal{D}(h)$, and exchanging the order of integration on the left produces $\int_I f(t)\nu([t,t_1])dt$. Thus ω is the survival function of ν, and its affine form is simply the uniformity of ν. Its three relevant properties each do work below: $\omega(t_1) = 0$ removes a boundary term in the integration by parts of Theorem G.4, $\omega^2 \leq \omega$ allows the in-class term of Theorem G.4 to be absorbed into the same weighted energy that appears on the left, and $\omega \geq \frac{1}{2}$ on I° converts the weighted bound back into an unweighted one –which is the reason the guarantee is stated on the half-window rather that on I.

For a space-time test function ψ we write $A(h,\psi) := \int_I a_t(h, \psi(t,\cdot))\nu(dt)$, so that $\|\mathcal{R}[h]\|_{\mathcal{V}^*} = \sup_{\psi \neq 0} A(h,\psi)/\|\psi\|_{\mathcal{V}}$.

Theorem D.2. *(Weighted energy identity) Under Assumptions 1–2, let $h \in C^{1,2}(I \times \mathbb{R}^d)$ with $|h| \leq 3\bar{B}$ and $\nabla\varphi \in L^2(dt \otimes \rho)$ and*

$$\sup_{I \times \mathbb{R}^d} (|\partial_t h| + |\nabla^2 h|) < \infty, \tag{D.7}$$

and write $\varphi := h - h^*$. Then, with $E_\omega := \int_I \omega(t)\|\nabla\varphi(t,\cdot)\|^2_{\rho_t}\nu(dt)$,

$$2\ell E_\omega + \|\varphi(t_0,\cdot)\|^2_{\rho_{t_0}} = \mathcal{D}(h) - 2\ell A(h, \omega\varphi). \tag{D.8}$$

Proof. The hypotheses of Theorem D.1 hold. Boundedness is immediate $|\varphi| \leq 4\bar{B}$. Let's split

$$\mathcal{M}\varphi = (\partial_t h - \partial_t h^*) + b \cdot (\nabla h - \nabla h^*) + (\Delta h - \Delta h^*) \tag{D.9}$$

and bound the six terms. By (D.7), $\partial_t h$ and Δh are bounded. By (B.5), $|\nabla h^*| \leq 2\sqrt{d}\bar{B}\sigma_*^{-2}$ and $|\Delta h^*| \leq dC_2\bar{B}\sigma_*^{-4}$, whence by harmonicity

$$|\partial_t h^*| = |b \cdot \nabla h^* + \Delta h^*| \leq 2b\sqrt{d}\bar{B}\sigma_*^{-2} + dC_2\bar{B}\sigma_*^{-4}. \tag{D.10}$$

Finally, $|b| \leq |x + s| + |s|$ lies in $L_2(\rho_t)$ uniformly on I by Lemma B.1 for the first summand and by $|s| \leq \sigma_{T-t}^{-2}(|x| + \sqrt{d})$ from (B.1) for the second, so Cauchy-Schwarz in $dt \otimes \rho$ gives

$$\int_I \int |b||\nabla h - \nabla h^*|\rho_t dt \leq \sqrt{\ell}\sup_{t \in I}\||b|\|_{L^2(\rho_t)} \cdot \||\nabla\varphi|\|_{L^2(dt \otimes \rho_t)} < \infty \tag{D.11}$$

by the hypothesis $\nabla\varphi \in L^2(dt \otimes \rho_t)$. Every remaining term of (D.8) is bounded by affine function of $|b|$, hence integrable against Gaussian-tailed ρ_t. Thus $\mathcal{M}\varphi \in L^1(dt \otimes \rho_t)$. Its cross term is $\langle \varphi, \mathcal{M}\varphi \rangle_{\rho_t} = a_t(h, \varphi)$ by Lemma C.5. Integrating (D.1) over the endpoint r against ν and using $\int_I (\int_{t_0}^r f(t)dt)\nu(dr) = \int_I f(t)\omega(t)dt = \ell \int_I f\omega\nu(dt)$ for $f \in L^1(I)$, together with $\int_I \|\varphi(r,\cdot)\|^2_{\rho_r}\nu(dr) = \mathcal{D}(h)$, gives (D.8). □

Identity (D.8) is exact. Two consequences follow, according to how the cross term $A(h, \omega\varphi)$ is bounded.

Corollary D.3. *(Coercivity, dual-norm form) Under the hypotheses of Theorem D.2, for every $\theta \in (0, 2\ell)$*

$$E_\omega \le \frac{1+\theta}{2\ell-\theta}\mathcal{D}(h) + \frac{\ell^2/\theta}{2\ell-\theta}\mathcal{P}^w(h), \tag{D.12}$$

and in particular, taking $\theta = \ell$ and using $\omega \ge \frac{1}{2}$ on I°

$$\int_{I^\circ} \|\nabla\varphi\|^2_{\rho_t}\nu(dt) \le \frac{2(1+\ell)}{\ell}\mathcal{D}(h) + 2\mathcal{P}^w(h). \tag{D.13}$$

Proof. Abbreviate $E := E_\omega,\ D := \mathcal{D}(h)$ and $P := \mathcal{P}^w(h) = \|\mathcal{R}[h]\|^2_{\mathcal{V}^*}$.

Step 1: Both terms on the left of (D.8) are nonnegative, so dropping $\|\varphi(t_0,\cdot)\|^2_{\rho_{t_0}}$ gives

$$2\ell E \le D + 2\ell|A(h, \omega\varphi)|. \tag{D.14}$$

Note that D enters here without a factor ℓ, because it arises from the terminal value $\|\varphi(r,\cdot)\|^2_{\rho_r}$ of the energy identity, which is averaged over the endpoint r and not integrated in t. Every contribution coming from the cross term, by contrast, carries ℓ.

Step 2: The function $\omega\varphi$ belongs to $\mathcal{V}$, so by the definition of $\|\cdot\|_{\mathcal{V}^*}$

$$|A(h, \omega\varphi)| \le \|\mathcal{R}[h]\|_{\mathcal{V}^*}\|\omega\varphi\|_{\mathcal{V}} = \sqrt{P}\|\omega\varphi\|_{\mathcal{V}}. \tag{D.15}$$

Step 3: Since ω depends only on t, we have $\nabla(\omega\varphi) = \omega\nabla\varphi$, and since $0 \le \omega \le 1$ we have both $\omega^2 \le 1$ and $\omega^2 \le \omega$. Using the first on the value part and the second on the gradient part

$$\|\omega\varphi\|^2_{\mathcal{V}} = \int_I \omega^2(\|\varphi\|^2_{\rho_t} + \|\nabla\varphi\|^2_{\rho_t})\nu(dt) \le \int_I \|\varphi\|^2_{\rho_t}\nu(dt) + \int_I w\|\nabla\varphi\|^2_{\rho_t}\nu(dt) = D + E. \tag{D.16}$$

The second inequality is where $\omega^2 \le \omega$ is used: it produces the same weighted energy E that stands on the left of (D.14), which is what makes absorption possible.

Step 4: Combining last steps and applying $2ab \le \theta a^2 + \theta^{-1}b^2$ with $a = \sqrt{D+E},\ b = \ell\sqrt{P}$ and any $\theta > 0$

$$2\ell E \le D + 2\ell\sqrt{P}\sqrt{D+E} \le D + \theta(D+E) + \frac{\ell^2}{\theta}P. \tag{D.17}$$

Collecting the E terms, which requires $\theta < 2\ell$

$$(2\ell - \theta)E \le (1+\theta)D + \frac{\ell^2}{\theta}P, \tag{D.18}$$

which is (D.12). Setting $\theta = \ell$ gives $\ell E \le (1+\ell)D + \ell P$, i.e. $E \le \frac{1+\ell}{\ell}D + P$.

Step 5: For $\delta \in (0,1)$ put $I_\delta := [t_0, t_1 - \delta\ell]$, on which $\omega \ge \delta$, so that

$$\int_{I_\delta} \|\nabla\varphi\|^2_{\rho_t}\nu(dt) \le \frac{1}{\delta}\int_{I_\delta} \omega\|\nabla\varphi\|^2_{\rho_t}\nu(dt) \le \frac{1}{\delta}E. \tag{D.19}$$

The case $\delta = 1/2$, for which $I_\delta = I^\circ$ and $\delta^{-1} = 2$, is (D.13). Remark I.3 explains how (D.19) places the guarantee on a prescribed sampling window. □

The estimator, however, does not control $\mathcal{P}^w(h)$ but only its projection onto the test class. The second consequence, proved in Section G, is that with the test class chosen appropriately the projected quantity suffices, and no inf-sup constant is incurred.

Corollary D.4. *(Guidance control) Under the hypotheses of Theorem D.2, let* $\mathrm{b} \in (0, \bar{B}]$ *be a clipping level and define the clipped, projected guidance*

$$\hat{g} := \Pi_{G^*}\left(\frac{\nabla h}{h \vee \mathrm{b}}\right), \quad \Pi_{G^*}(v) := v \min\{1, G^*/|v|\}, \tag{D.20}$$

with $G^* = 2\sqrt{d}\sigma_*^{-2}$ *from Lemma B.3. Then*

$$\|\hat{g} - g^*\|^2_{L^2(\nu\otimes\rho;I^\circ)} \le \frac{2}{\mathrm{b}^2}\int_{I^\circ}\|\nabla\varphi\|^2_{\rho_t}\nu(dt) + \frac{8d\bar{B}^2}{\mathrm{b}^4}\sigma_*^{-4}\mathcal{D}(h) + 4(G^*)^2\Xi_{\mathrm{s}}\mathrm{b}^{\mathrm{s}}. \tag{D.21}$$

Proof. By Lemma B.3, $|g^*| \le G^*$, so $\Pi_{G^*}(g^*) = g^*$ and Π_{G^*} is the Euclidean projection onto a convex set, hence 1-Lipschitz. Therefore $|\hat{g} - g^*| \le |\nabla h/(h \vee \mathrm{b}) - g^*|$ pointwise, and it suffices to bound the right-hand side. We split according to the size of the true h^*.

On $\{h^* \ge \mathrm{b}\}$. Put $u := h \vee \mathrm{b} \ge \mathrm{b}$ and $v := h^* \ge \mathrm{b}$, so that

$$\frac{\nabla h}{u} - \frac{\nabla h^*}{v} = \frac{\nabla\varphi}{u} + \nabla h^*\frac{v-u}{uv}. \tag{D.22}$$

Here $|v-u| \le |\varphi|$: if $h \ge \mathrm{b}$ then $u = h$ and $v - u = -\varphi$. If $h < \mathrm{b}$ then $u = \mathrm{b}$ and $0 \le v - u = h^* - \mathrm{b} \le h^* - h = -\varphi$. With $|\nabla h^*| \le 2\sqrt{d}\bar{B}\sigma_*^{-2}$ from (B.7) and $u, v \ge \mathrm{b}$

$$|\nabla h/(h \vee \mathrm{b}) - g^*| \le \mathrm{b}^{-1}|\nabla\varphi| + 2\sqrt{d}\bar{B}\mathrm{b}^{-2}\sigma_*^{-2}|\varphi|, \tag{D.23}$$

and squaring with $(a_1 + a_2)^2 \le 2a_1^2 + 2a_2^2$ and integrating over I° gives the first two of (D.21).

On $\{h^* < \mathrm{b}\}$. Both $\hat{g}$ and g^* lie in the ball of radius G^*, the former by construction and the latter by Lemma B.3, so $|\hat{g} - g^*| \le 2G^*$ regardless of how small h^* is. The contribution is at most $4(G^*)^2\int_{I^\circ}\rho_t(h^* < \mathrm{b})\nu(dt)$, which Lemma B.4 bounds by $4(G^*)^2\Xi_s\mathrm{b}^s$. □

E THE HYPOTHESIS CLASSES

Neither property requires $\mathcal{H}$ or Ψ to be linear or convex. We record the second as a lemma.

Definition E.1. (Network classes) Fix the evaluation domain and the spatial cube

$$Q_R := \{x \in \mathbb{R}^d : |x|_\infty \le R\}, \quad Z := I \times Q_R, \quad R_* := \max(t_1, R). \tag{E.1}$$

The sup-norm ball is the right choice, since it is exactly the range of the spatial clip χ_R in Definition E.3, so a constraint verified on Z is automatically global for clipped functions. Note that $|x| \le \sqrt{d}R$ on Q_R, and $\|z\|_\infty \le R_*$ for $z = (t, x) \in Z$. Let ϱ be an activation function. For $W, L \in \mathbb{N}$ and $\Lambda, \mathcal{A} \ge 1$ and $P_0 \in \mathbb{N}$, let $\mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A})$ be the set of functions

$$f_\vartheta(z) = A_L x^{(L)}(z) + b_L,\ x^{(0)}(z) := z,\ x^{(\ell+1)}(z) := \varrho(A_\ell x^{(\ell)}(z) + b_\ell) \quad (0 \le \ell \le L-1), \tag{E.2}$$

with parameter vector $\vartheta := (A_0, b_0, \ldots, A_L, b_L) \in \mathbb{R}^P$, subject to:

(N1) all hidden widths at most W, where $W \ge d + 1$, so that $P \le (L + 1)(W + 1)^2$,

(N2) all entries of ϑ bounded in modulus by Λ, and at most P_0 of them nonzero,

(N3) all hidden activations bounded on the evaluation domain: $\sup_{z\in Z}\|x^{(\ell)}(z)\|_\infty \le \mathcal{A}$ for $1 \le \ell \le L$.

Condition (N3) is a constraint on the realized network, not on its parameters. It is checkable and satisfied by construction we invoke and is what makes the entropy bound in Lemma E.8 polylogarithmic in the parameters. Write $\mathcal{A}_{\max} := \mathcal{A} \vee R_*$ and

$$\mathrm{m} := 3\Lambda W(\mathcal{A}_{\max} + 1) \ge 3, \tag{E.3}$$

a single quantity in terms of which all the constants below are expressed.

Lemma E.2. *(Differences of networks are networks) Let $\Theta : \mathbb{R} \to \mathbb{R}$ be bounded, write $\|\Theta\|_\infty := \sup_{u\in\mathbb{R}} |\Theta(u)|$, and put $\Theta \circ \mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A}) := \{\Theta \circ g : g \in \mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A})\}$. Then, for $\Lambda \geq 1$,*

$$\Theta \circ \mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A}) \;-\; \Theta \circ \mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A}) \subseteq \mathcal{NN}_{\varrho,\Theta}\big(2W,\, L+1,\, \Lambda,\, 2P_0+2,\, \mathcal{A} \vee \|\Theta\|_\infty\big), \quad \text{(E.4)}$$

where $\mathcal{NN}_{\varrho,\Theta}$ denotes the same architecture with Θ used as the activation of the last hidden layer.

Proof. Let $g_1, g_2 \in \mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A})$ have parameters $(A_\ell^{(i)}, b_\ell^{(i)})_{0\leq\ell\leq L}$. Build a new network whose hidden layers are the two originals stacked side by side: duplicate the input by $\tilde{A}_0 := (A_0^{(1)}; A_0^{(2)})$, and for $1 \leq \ell \leq L-1$ take block-diagonal weight matrices $\tilde{A}_\ell := \mathrm{diag}(A_\ell^{(1)}, A_\ell^{(2)})$ and stacked biases $\tilde{b}_\ell := (b_\ell^{(1)}, b_\ell^{(2)})$. Because ϱ acts coordinatewise, no information crosses between the two blocks, so after L hidden layers the two halves of the activation vector are the $x^{(L)}$ of g_1 and of g_2.

Append one more layer, of width 2, with weights $\mathrm{diag}(A_L^{(1)}, A_L^{(2)})$ and biases $(b_L^{(1)}, b_L^{(2)})$ and with Θ as its activation: its pre-activations are $(g_1(z), g_2(z))$ and its activations are $(\Theta(g_1(z)), \Theta(g_2(z)))$. Read out with the output weights $(1, -1)$ and zero bias. The result is $\Theta(g_1(z)) - \Theta(g_2(z))$, which is the general element of the left-hand side of (E.4).

It remains to check (N1)–(N3) for the new network. (N1): the hidden widths of layers $1, \ldots, L$ are at most $2W$, and the appended layer has width $2 \leq 2W$, the depth is $L+1$. (N2): every parameter used is a parameter of g_1 or of g_2, a structural zero, or one of the two output weights ± 1, so all entries are bounded in modulus by $\Lambda \vee 1 = \Lambda$, and the number of nonzero entries is at most $2P_0$ (those of g_1 and g_2, all of which are reused, the readout weights $A_L^{(i)}$ and biases $b_L^{(i)}$ having been moved into the appended layer) plus the two output weights, i.e. at most $2P_0 + 2$. (N3): the activations of layers $1, \ldots, L$ are those of g_1 and g_2, bounded by $\mathcal{A}$ on Z. Those of the appended layer lie in the range of Θ, hence are bounded by $\|\Theta\|_\infty$ everywhere. Since (N3) is a single bound over all hidden layers, it holds with budget $\mathcal{A} \vee \|\Theta\|_\infty$. □

Definition E.3. (Trial and test classes) Fix a resolution N, a time degree K, a truncation radius R and a weight budget $\Lambda = \Lambda(N)$, a sparsity budget $P_0 = P_0(N)$ and an activation budget $\mathcal{A} = \mathcal{A}(N)$. Let ϱ be the rectified cubic $\varrho(u) = (u \vee 0)^3$, let $\Theta \in C^\infty(\mathbb{R}, [-2\bar{B}, 3\bar{B}])$ be a fixed 1-Lipschitz retraction with $\Theta = \mathrm{id}$ on $[-\bar{B}, 2\bar{B}]$ and bounded second derivative, $\kappa_\Theta := \|\Theta''\|_\infty < \infty$, and set $W_N \asymp KN^d$ with $W_N \geq d+1$. Fix also a spatial clip $\chi_R : \mathbb{R}^d \to \mathbb{R}^d$, acting coordinatewise by a fixed C^∞ function that is the identity on $[-(R-1), R-1]$, has modulus at most R, derivative at most 1 in modulus and bounded second derivative. For a function f on $I \times \mathbb{R}^d$ write

$$f^\chi(t, x) := f(t, \chi_R(x)). \quad \text{(E.5)}$$

Like the gate ω, the clip is a fixed map with no free parameters. Define

$$\mathcal{H}_N := \{(\Theta \circ g)^\chi : g \in \mathcal{NN}_\varrho(W_N, L, \Lambda, P_0, \mathcal{A}),\ \sup_{I\times\mathbb{R}^d} |\nabla h| \leq G,\ \sup_{I\times\mathbb{R}^d} |\partial_t h| \leq G_t\}, \quad \text{(E.6)}$$

$$\Psi_N := \{c\,\omega\,\phi^\chi : c \in \mathbb{R},\ \phi \in \mathcal{NN}_{\varrho,\Theta}(2W_N, L+1, \Lambda, 2P_0+2, \mathcal{A} \vee 3\bar{B}), \sup_{I\times\mathbb{R}^d} (|\phi^\chi| + |\nabla\phi^\chi| + |\partial_t\phi^\chi|) \leq \Gamma\}, \quad \text{(E.7)}$$

the constraint in (E.6) being on $h = (\Theta \circ g)^\chi$ itself. The three constants are

$$G := \frac{4\sqrt{d}\bar{B}}{\sigma_*^2},\ G_t = \frac{C_t d\bar{B}(R+1)}{\sigma_*^4},\ \Gamma := 6\bar{B} + 2G + 2G_t + \frac{6\bar{B}}{\ell}, \quad \text{(E.8)}$$

with C_t absolute.

Proposition E.4. *(The classes have properties (S1)–(S2)) With Definition E.3:*

(i) Ψ_N is a symmetric cone, i.e. (S1) holds,

(ii) $\omega \cdot (\mathcal{H}_N - \mathcal{H}_N) \subseteq \Psi_N$, i.e. (S2) holds,

(iii) $\mathcal{H}_N \cup \Psi_N \subset C^{1,2}(I \times \mathbb{R}^d)$, *every element has space-time derivatives up to second order bounded on all* $I \times \mathbb{R}^d$, *and in particular the qualitative condition* (D.7) *holds at every* $h \in \mathcal{H}_N$, *so the hypotheses of Lemmas C.5–C.6 and Theorems D.1 and D.2 are met there,*

(iv) h^* *satisfies both sup constraints in* (E.6) *on* Z, *and* $\Theta \circ h^* = h^*$,

(v) every $\varphi_1 \in \mathcal{H}_N - \mathcal{H}_N$ *satisfies* $\sup_{I \times \mathbb{R}^d} (|\omega\varphi_1| + |\nabla(\omega\varphi_1)| + |\partial_t(\omega\varphi_1)|) \le \Gamma$, *so the inclusion in (ii) lands inside the constrained cone* (E.7).

Proof. (i) The scalar c in (E.7) ranges over all of $\mathbb{R}$, so Ψ_N is closed under multiplication by any real number, in particular under negation and under nonnegative scaling.

(ii) The clip commutes with differences, $(\Theta \circ g_1)^\chi - (\Theta \circ g_2)^\chi = ((\Theta \circ g_1) - (\Theta \circ g_2))^\chi$, so it suffices to difference the un-clipped networks. By Lemma E.2, $(\Theta \circ g_1) - (\Theta \circ g_2) \in \mathcal{NN}_{\varrho,\Theta}(2W_N, L + 1, \Lambda, 2P_0 + 2, \mathcal{A} \vee 3\bar{B})$ – the constraints in (E.6) only shrink the set being differed, so they do not affect the inclusion. Multiplying by ω and taking $c = 1$ in (E.7) gives the claim. Note that ω is a fixed function of t, not a parameter: the test architecture is "network, then multiply the output by $\omega(t)$," a gate with no free parameters.

(iii) $\varrho(u) = (u \vee 0)^3$ has $\varrho'(u) = 3(u \vee 0)^2$ and $\varrho''(u) = 6(u \vee 0)$, both continuous, so $\varrho \in C^2$ and any finite composition of affine maps and coordinatewise ϱ is C^2 in (t, x). Θ and χ_R are C^∞ and ω is affine, so composing, clipping and gating preserves C^2. In particular the classes are $C^{1,2}$, which is what Theorem D.1 requires and what a C^1 activation such as the rectified quadratic would not give. For the boundedness claim, every $h = (\Theta \circ g)^\chi$ factors through the clip, so each $\sup_{I \times \mathbb{R}^d} |\partial^\alpha h|$ with $|\alpha| \le 2$ is a supremum of derivatives of $\Theta \circ g$ over the compact set Z, multiplied by fixed powers of $\|D\chi_R\|_\infty \le 1$ and $\|D^2\chi_R\|_\infty < \infty$, the derivatives of $\Theta \circ g$ up to second order are continuous, because $\varrho \in C^2$ and $\Theta \in C^\infty$, hence bounded on the compact set Z.

(iv) All three constraints are verified on Z, which is what a clipped function needs. By (B.5), $|\partial_{x_k} h^*| \le 2\bar{B}\sigma_{T-t}^{-2}$, so $|\nabla h^*| \le 2\sqrt{d}\bar{B}\sigma_*^{-2} = G/2$. For the time derivative, harmonicity (C.2) gives $\partial_t h^* = -b \cdot \nabla h^* - \Delta h^*$, and on Z we have $|x| \le \sqrt{d}R$ by (E.1), so $|b| = |x + 2s| \le |x| + 2\sigma_{T-t}^{-2}(|x| + \sqrt{d}) \le 3\sqrt{d}\sigma_*^{-2}(R + 1)$ by (B.1) and $\sigma_* \le 1$, while $|\Delta h^*| \le dC_2\bar{B}\sigma_*^{-4}$ by (B.5). Hence $|\partial_t h^*| \le 6d\bar{B}\sigma_*^{-4}(R + 1) + dC_2\bar{B}\sigma_*^{-4} \le Cd\bar{B}\sigma_*^{-4}(R + 1) = G_t/2$ for C_t large enough. Finally, $0 < h^* \le \bar{B}$ lies in the range on which Θ is the identity.

(v) For $h, h' \in \mathcal{H}_N$ and $\varphi_1 = h - h'$, we have, on all of $I \times \mathbb{R}^d$ now that (E.6) is global, $|\varphi_1| \le 6\bar{B}$ (both lie in $[-2\bar{B}, 3\bar{B}]$), $|\nabla\varphi_1| \le 2G$ and $|\partial_t\varphi_1| \le 2G_t$. Since $0 \le \omega \le 1$ and $|\dot{\omega}| = \ell^{-1}$, the function $\omega\varphi_1$ obeys $|\omega\varphi_1| \le 6\bar{B}$, $|\nabla(\omega\varphi)| = \omega|\nabla\varphi_1| \le 2G$ and $|\partial_t(\omega\varphi_1)| \le 2G_t + 6\bar{B}/\ell$, and the sum of the three is at most Γ by (E.8). □

The statistical analysis measures a test function on two scales: its size in $\mathcal{V}$, and the size of the scalar that generates it. For $\psi \in \Psi_N$ put

$$\|\psi\|_\bullet := \inf\{|c| : \psi = c\,\omega\,\phi^\chi \text{ as in (E.7)}\}, \tag{E.9}$$

the cone scale of ψ. It is symmetric and positively homogeneous – the only two properties used below – and it dominates the supremum norm globally, $\sup_{I \times \mathbb{R}^d} (|\psi| + |\nabla\psi|) \le \Gamma\|\psi\|_\bullet$, because every representation $\psi = c\,\omega\,\phi^\chi$ gives $|\psi| + |\nabla\psi| = |c|\,\omega\,(|\phi^\chi| + |\nabla\phi^\chi|) \le \Gamma|c|$ by (E.7) and $0 \le \omega \le 1$, so that the quantity

$$\Gamma_{\mathcal{V}} := \sup\{\|\psi\|_{\mathcal{V}} : \psi \in \Psi_N, \|\psi\|_\bullet \le 1\} \tag{E.10}$$

satisfies $\Gamma_{\mathcal{V}} \le \Gamma$, ρ_t being a probability measure. By parts (ii) and (v) of Proposition E.4, every displacement obeys

$$\|\omega\varphi_1\|_\bullet \le 1, \quad \text{for all } \varphi_1 \in \mathcal{H}_N - \mathcal{H}_N, \tag{E.11}$$

since $c = 1$ is an admissible representation. The two scales are genuinely different: under the mild conditions of Remark E.6 the ratio $\sup(|\psi| + |\nabla\psi|)/\|\psi\|_{\mathcal{V}}$, and a fortiori $\|\psi\|_\bullet/\|\psi\|_{\mathcal{V}}$, is unbounded on Ψ_N, a generator being able to be small in $L^2(\nu \otimes \rho)$ without being small in the supremum norm. This is why the deviation bounds of Section H are two-norm bounds, why the estimator carries a ridge at the scale $\|\cdot\|_\bullet$, and why the complexity of the test class is measured in (E.14) on the unit ball of $\|\cdot\|_\bullet$ and not on that of $\|\cdot\|_{\mathcal{V}}$.

Remark E.5. (Why the rectified cubic) Three requirements pull in different directions and $\varrho(u) = (u \vee 0)^3$ meets all of them. The energy identity needs $C^{1,2}$ trial functions, which excludes ReLU (not C^1) and the rectified quadratic (only C^1). Exact emulation of products is needed for the space-time construction of Lemma F.8, and it is available here, since

$$(u+1)^3 - (u-1)^3 = 6u^2 + 2, \quad u^3 = \varrho(u) - \varrho(-u), \tag{E.12}$$

a two-node layer computes $u \mapsto u^2$ exactly, hence $uv = \frac{1}{4}[(u+v)^2 - (u-v)^2]$ exactly, hence any polynomial exactly. Smooth activations such as tanh satisfy the first requirement but emulate products only approximately.

The quantity that the statistical analysis consumes is the metric entropy of first-order jets. For a function F on $I \times \mathbb{R}^d$ with values in a Euclidean space write $\|F\|_\infty := \sup_{I\times\mathbb{R}^d} |F|$, with $|\cdot|$ the Euclidean norm, and put

$$\begin{aligned} &\Psi_N^\bullet := \{\psi \in \Psi_N : \|\psi\|_\bullet \le 1\}, \\ &J\mathcal{H}_N := \{(h, \partial_t h, \nabla h) : h \in \mathcal{H}_N\}, \qquad J\Psi_N^\bullet := \{(\psi, \nabla\psi) : \psi \in \Psi_N^\bullet\}, \end{aligned} \tag{E.13}$$

$$\mathbb{V}_N := 1 \vee \sup_{0<\varepsilon\le 1} \frac{\max\{\log N(\varepsilon, J\mathcal{H}_N, \|\cdot\|_\infty),\, \log N(\varepsilon, J\Psi_N^\bullet, \|\cdot\|_\infty)\}}{\log(e/\varepsilon)}, \tag{E.14}$$

the normalization by $\log(e/\varepsilon)$ being the one under which a P-dimensional parametric class has $\mathbb{V}_N \asymp P$. Three features of (E.14) are used below. First, jets are covered jointly, because the summands of Section H depend on $(\partial_t h, \nabla h)$ and on $(\psi, \nabla\psi)$ simultaneously, $\partial_t h$ enters through the bilinear form and $\nabla\psi$ because the adversary is normed in H^1. Second, no lower cutoff is imposed on ε, so that every scale at which an entropy integral is evaluated in Section H is covered. Third, the free scalar c of (E.7) is removed by the cone scale: on $\Psi_N^\bullet$ one has $\sup(|\psi| + |\nabla\psi|) \le \Gamma$. The unit ball of $\|\cdot\|_{\mathcal{V}}$ would define the same projected residual, since every ray of the cone meets both balls and $\psi \mapsto A(h,\psi)/\|\psi\|_{\mathcal{V},\tau}$ is invariant under positive scaling, but it is not a totally bounded index set: by Remark E.6 its gradients are in general unbounded in $\|\cdot\|_\infty$, so its covering numbers are infinite at every radius.

Remark E.6. (The unit $\mathcal{V}$-ball of Ψ_N is not totally bounded) Suppose that χ_R is nondecreasing and that $R \ge 2$, $W_N \ge 3$, $2P_0 + 2 \ge 40L + 10$ and $\mathcal{A} \ge (R+1)^3$. Then

$$\sup\{\|\nabla\psi\|_\infty : \psi \in \Psi_N,\ \|\psi\|_{\mathcal{V}} \le 1\} = \infty. \tag{E.15}$$

Consequently $\{\nabla\psi : \psi \in \Psi_N,\ \|\psi\|_{\mathcal{V}} \le 1\}$ has infinite covering numbers in $\|\cdot\|_\infty$ at every radius, and, since $\|\nabla\psi\|_\infty \le \Gamma\|\psi\|_\bullet$, the ratio $\|\psi\|_\bullet/\|\psi\|_{\mathcal{V}}$ is unbounded on $\Psi_N \setminus \{0\}$.

Proof. Let B be the centred cardinal cubic B-spline, supported on $[-2,2]$, which satisfies $0 \le B \le 2/3$, $|B'| \le 2/3$ and $B'(1) = -1/2$, and put $\kappa_B := \|B\|^2_{L^2(\mathbb{R})} + \|B'\|^2_{L^2(\mathbb{R})} < \infty$. Its truncated-power representation gives, for $\varepsilon > 0$ and $y \in \mathbb{R}$,

$$g_\varepsilon(y) := \varepsilon^3 B(y/\varepsilon) = \sum_{k=-2}^{2} a_k\, \varrho(y - k\varepsilon), \qquad (a_{-2}, \dots, a_2) := \tfrac{1}{6}(1, -4, 6, -4, 1).$$

Fix $0 < \varepsilon \le \min\{1/4, (3\bar{B})^{1/3}, \Gamma^{1/2}\}$. Then $0 \le g_\varepsilon \le 2\varepsilon^3/3 \le \min\{1, 2\bar{B}\}$, $\operatorname{supp} g_\varepsilon \subseteq [-2\varepsilon, 2\varepsilon]$ and $\|g_\varepsilon\|_\infty + \|g_\varepsilon'\|_\infty \le \frac{2}{3}(\varepsilon^3 + \varepsilon^2) \le \varepsilon^2 \le \Gamma$.

Realization. We exhibit $\phi_\varepsilon \in \mathcal{NN}_{\varrho,\Theta}(2W_N, L+1, \Lambda, 2P_0+2, \mathcal{A} \vee 3\bar{B})$ with $\phi_\varepsilon(t,x) = g_\varepsilon(x_1)$ on $I \times \mathbb{R}^d$. Hidden layer 1 consists of the five units $\varrho(x_1 - k\varepsilon)$, $|k| \le 2$, so that the affine combination $\sum_k a_k \varrho(x_1 - k\varepsilon)$ of its outputs is $u := g_\varepsilon(x_1)$. If $L \ge 2$, each of the hidden layers $2, \dots, L$ consists of the six units $\varrho(u+1), \varrho(-u-1), \varrho(u-1), \varrho(1-u), \varrho(u), \varrho(-u)$, whose pre-activations are affine combinations of the outputs of the previous layer, and u is recovered exactly from them by

$$6u = \varrho(u+1) - \varrho(-u-1) + \varrho(u-1) - \varrho(1-u) - 2\varrho(u) + 2\varrho(-u),$$

which follows from $\varrho(v) - \varrho(-v) = v^3$ and $(u+1)^3 + (u-1)^3 - 2u^3 = 6u$. Hidden layer $L+1$ is the single unit $\Theta(u)$, read out with weight 1 and bias 0. Since $u \in [0, 2\bar{B}]$ and $\Theta = \mathrm{id}$ there, $\phi_\varepsilon(t,x) = g_\varepsilon(x_1)$ everywhere. All entries lie in $\{0, \pm 1/6, \pm 1/3, \pm 2/3, \pm 1, \pm\varepsilon, \pm 2\varepsilon\}$ and are therefore bounded by $1 \le \Lambda$. Hidden widths are at most $6 \le 2W_N$. The number of nonzero entries is 15 for $L = 1$ and $9 + 34 + 40(L-2) + 6 + 1$ for $L \ge 2$, in both cases at most $40L + 10 \le 2P_0 + 2$,

and on Z, where $|x_1| \le R$, the units of layer 1 take values in $[0, (R+1)^3]$, those of layers $2, \dots, L$ in $[0, 8]$ because $0 \le u \le 1$, and the last one in $[0, 2\bar{B}]$, all within $\mathcal{A} \vee 3\bar{B}$.

Clip and constraint. If $|y| \le R-1$ then $\chi_R(y) = y$. If $y > R-1$ then $y > 2\varepsilon$ and, χ_R being nondecreasing, $\chi_R(y) \ge R-1 > 2\varepsilon$, symmetrically for $y < -(R-1)$. In all cases $g_\varepsilon(\chi_R(y)) = g_\varepsilon(y)$, so $\phi^\chi_\varepsilon(t,x) = g_\varepsilon(x_1)$ and $|\phi^\chi_\varepsilon| + |\nabla\phi^\chi_\varepsilon| + |\partial_t\phi^\chi_\varepsilon| \le \Gamma$. Thus $\psi_\varepsilon(t,x) := \omega(t)\, g_\varepsilon(x_1)$ belongs to Ψ_N, with $c = 1$.

Two norms. The x_1-marginal of ρ_t, the law of $\mu_{T-t}X_0 + \sigma_{T-t}\xi$ with $\xi \sim \mathcal{N}(0, I_d)$ independent of X_0, has a positive density bounded by $(2\pi)^{-1/2}\sigma_{T-t}^{-1} \le (2\pi)^{-1/2}\sigma_*^{-1}$ for $t \in I$. Since $0 \le \omega \le 1$ and $\varepsilon \le 1$,

$$0 < \|\psi_\varepsilon\|_{\mathcal{V}}^2 = \int_I \omega^2 \left(\|g_\varepsilon(x_1)\|_{\rho_t}^2 + \|g_\varepsilon'(x_1)\|_{\rho_t}^2 \right) \nu(dt) \le \frac{\varepsilon^7 \|B\|_{L^2(\mathbb{R})}^2 + \varepsilon^5 \|B'\|_{L^2(\mathbb{R})}^2}{\sqrt{2\pi}\,\sigma_*} \le \frac{\kappa_B\, \varepsilon^5}{\sqrt{2\pi}\,\sigma_*},$$

with positivity because $g_\varepsilon \not\equiv 0$ and $\omega > 0$ on $[t_0, t_1)$. On the other hand $\|\nabla\psi_\varepsilon\|_\infty \ge \omega(t_0)\,|g_\varepsilon'(\varepsilon)| = \varepsilon^2 |B'(1)| = \varepsilon^2/2$. Hence $\bar{\psi}_\varepsilon := \psi_\varepsilon / \|\psi_\varepsilon\|_{\mathcal{V}} \in \Psi_N$ has $\|\bar{\psi}_\varepsilon\|_{\mathcal{V}} = 1$ and

$$\|\nabla\bar{\psi}_\varepsilon\|_\infty \ge \tfrac{1}{2}(2\pi)^{1/4}\sigma_*^{1/2}\kappa_B^{-1/2}\,\varepsilon^{-1/2} \longrightarrow \infty \quad (\varepsilon \downarrow 0),$$

which is (E.15). The first consequence holds because a set covered by finitely many $\|\cdot\|_\infty$-balls of finite radius is bounded, the second because $\|\bar{\psi}_\varepsilon\|_\bullet \ge \Gamma^{-1}\|\nabla\bar{\psi}_\varepsilon\|_\infty$ by (E.9). □

We first record three elementary facts about covering numbers. Throughout, for a subset S of a metric space (X, d), $N(\varepsilon, S, \mathrm{d})$ denotes the minimal cardinality of an ε-net of S, that is, of a set $\mathcal{C} \subseteq X$ such that every point of S lies at distance at most ε from $\mathcal{C}$, in particular $N(\varepsilon, S, \mathrm{d}) \le N(\varepsilon, S', \mathrm{d})$ whenever $S \subseteq S'$.

Lemma E.7. *(Three covering facts)*

(E1) (Cube) For $Q := [-\Lambda, \Lambda]^m$ and $0 < \delta \le \Lambda$, one has $N(\delta, Q, \|\cdot\|_\infty) \le \lceil \Lambda/\delta \rceil^m \le (2\Lambda/\delta)^m$,

(E2) (Internal nets) If $S \subseteq X$ and $N(\delta/2, S, \mathrm{d}) < \infty$, then S contains a δ-net of S consisting of itself points and of cardinality at most $N(\delta/2, S, \mathrm{d})$,

(E3) (Sparse supports) For $1 \le m \le P$, the number of subsets of $\{1, \dots, P\}$ of cardinality at most m is at most $(eP/m)^m$.

Proof. (E1) Partition $[-\Lambda, \Lambda]$ into $\lceil \Lambda/\delta \rceil$ intervals of length at most 2δ and take their midpoints. The product grid has the stated cardinality and every point of Q lies within δ of a grid point in each coordinate, hence within δ in $\|\cdot\|_\infty$. The final inequality uses $\lceil u \rceil \le 2u$ for $u \ge 1$.

(E2) Call $\mathcal{S} \subseteq S$ δ-separated if distinct points of $\mathcal{S}$ are at distance larger than δ. Let $\mathcal{C}$ be a $(\delta/2)$-net of S of minimal size and assign to each point of a δ-separated $\mathcal{S}$ a point of $\mathcal{C}$ within $\delta/2$ of it. The assignment is injective, since two points with the same image are within δ of each other. Hence every δ-separated subset of S has at most $|\mathcal{C}| = N(\delta/2, S, \mathrm{d})$ points, and one of maximal cardinality is maximal under inclusion. By maximality no point of S is at distance larger than δ from all of $\mathcal{S}$, which is exactly the δ-net property.

(E3) $\sum_{k \le m} \binom{P}{k} \le (eP/m)^m$. Indeed, for $\lambda := m/P \le 1$, $\sum_{k \le m} \binom{P}{k}\lambda^k \le (1+\lambda)^P \le e^{\lambda P} = e^m$, while every term on the left with $k \le m$ satisfies $\binom{P}{k}\lambda^k \ge \binom{P}{k}\lambda^m$. Dividing by λ^m gives the claim. □

Content of the next lemma is that the map from the parameters to functions is Lipschitz with a constant that is exponential in depth but polynomial in everything else.

Lemma E.8. *(Complexity) Let $\mathcal{F}$ be either $\mathcal{NN}_\varrho(W, L, \Lambda, P_0, \mathcal{A})$ with $\varrho(u) = (u \vee 0)^3$, or $\mathcal{NN}_{\varrho,\Theta}(W, L, \Lambda, P_0, \mathcal{A})$ with Θ as in Definition E.3. Let* m *be as in* (E.3) *in the first case and* $\mathrm{m} := 3\Lambda W(\mathcal{A}_{\max} + 1) \vee \kappa_\Theta$ *in the second, put* $\mathrm{L}_* := 2L\mathrm{m}^{3L+1}$ *and* $\mathrm{L}'_* := 12L^2\mathrm{m}^{9L+3}$, *and write* $\nabla_z := (\partial_t, \nabla_x)$ *for the space-time gradient and* $|\cdot|_1, |\cdot|$ *for the ℓ^1 and the Euclidean norm. Then:*

(i) (Parameter Lipschitz bound) For any two parameter vectors $\vartheta, \tilde{\vartheta} \in \mathbb{R}^P$ of elements of $\mathcal{F}$, realized in the common layout of the proof,

$$\sup_{z\in Z} |f_\vartheta(z) - f_{\tilde{\vartheta}}(z)| \le \mathrm{L}_* \|\vartheta - \tilde{\vartheta}\|_\infty, \qquad \sup_{z\in Z} |\nabla_z f_\vartheta(z) - \nabla_z f_{\tilde{\vartheta}}(z)|_1 \le \mathrm{L}'_* \|\vartheta - \tilde{\vartheta}\|_\infty. \tag{E.16}$$

(ii) (Entropy of jets) Suppose $1 \le P_0 \le P$. For every nonempty $\mathcal{F}' \subseteq \mathcal{F}$ and $0 < \varepsilon \le 1$, the jet class $J\mathcal{F}' := \{(f, \nabla_z f) : f \in \mathcal{F}'\}$, normed by $\|F\|_{\infty,Z} := \sup_{z\in Z} |F(z)|$, has an ε-net consisting of jets of elements of $\mathcal{F}'$ and of cardinality at most $(eP/P_0)^{P_0}(8\Lambda \mathrm{L}'_*/\varepsilon)^{P_0}$, in particular

$$\log N(\varepsilon, J\mathcal{F}', \|\cdot\|_{\infty,Z}) \le P_0 \left[\log \frac{eP}{P_0} + \log(8\Lambda \mathrm{L}'_*) + \log \frac{1}{\varepsilon}\right]. \tag{E.17}$$

(iii) (Effective dimension) Let $\mathcal{H}_N, \Psi_N$ be as in Definition E.3, let $P_{\mathcal{H}}$ be the number of entries of a parameter vector of $\mathcal{NN}_\varrho(W_N, L, \Lambda, P_0, \mathcal{A})$ in that layout, and assume $1 \le P_0 \le P_{\mathcal{H}}$, a larger sparsity budget being no constraint. Put

$$\bar{P}_0 := 2P_0 + 2, \quad \bar{P} := (L+2)(2W_N+1)^2, \quad \bar{\mathrm{m}} := 6\Lambda W_N\big((\mathcal{A} \vee 3\bar{B} \vee R_*) + 1\big) \vee \kappa_\Theta. \tag{E.18}$$

Then $\mathbb{V}_N$ of (E.14) *satisfies*

$$\mathbb{V}_N \le \log(1 + 4\Gamma) + \bar{P}_0 \left[\log \frac{e\bar{P}}{\bar{P}_0} + \log\big(384\Lambda(L+1)^2\big) + (9L + 12)\log \bar{\mathrm{m}}\right]. \tag{E.19}$$

In particular, if $P_0 \lesssim KN^d$, $W_N \lesssim KN^d$, $2 \le N \le n$, $L \lesssim \log N$ and $\Lambda, \mathcal{A} \le \mathrm{poly}(n)$, then

$$\mathbb{V}_N \lesssim KN^d \log N \log n \lesssim N^d \mathrm{polylog}(n). \tag{E.20}$$

Proof. (i) Layout. Padding a hidden layer with a unit whose incoming weights, bias and outgoing weights vanish changes neither f_ϑ nor (N2)–(N3), since $\varrho(0) = \Theta(0) = 0$. We therefore realize every element of $\mathcal{F}$ with all hidden widths equal to W, so that $A_0 \in \mathbb{R}^{W\times(d+1)}$, $A_\ell \in \mathbb{R}^{W\times W}$ for $1 \le \ell \le L-1$ and $A_L \in \mathbb{R}^{1\times W}$, and a parameter vector is an element of $\mathbb{R}^P$ with

$$P = W(d+2) + (L-1)W(W+1) + W + 1. \tag{E.21}$$

Write σ_ℓ for the activation producing $x^{(\ell+1)}$, so that $\sigma_\ell = \varrho$ except for $\sigma_{L-1} = \Theta$ in the second case. The only properties of the activations used below are, for $u, u' \in [-\mathrm{m}, \mathrm{m}]$,

$$|\sigma'_\ell(u)| \le 3\mathrm{m}^2, \qquad |\sigma'_\ell(u) - \sigma'_\ell(u')| \le 6\mathrm{m}|u - u'|. \tag{E.22}$$

For ϱ they follow from $\varrho'(u) = 3(u \vee 0)^2$ and $\varrho''(u) = 6(u \vee 0)$, and for Θ from $|\Theta'| \le 1 \le 3\mathrm{m}^2$ and $|\Theta''| \le \kappa_\Theta \le \mathrm{m}$.

Write $x^{(\ell)}, \tilde{x}^{(\ell)}$ for the hidden activations (E.2) of $f_\vartheta, f_{\tilde{\vartheta}}$, set $v^{(\ell)} := A_\ell x^{(\ell)} + b_\ell$ and $\tilde{v}^{(\ell)} := \tilde{A}_\ell \tilde{x}^{(\ell)} + \tilde{b}_\ell$ and put

$$e_\ell := \sup_{z\in Z} \|x^{(\ell)}(z) - \tilde{x}^{(\ell)}(z)\|_\infty, \quad \delta := \|\vartheta - \tilde{\vartheta}\|_\infty. \tag{E.23}$$

Since $x^{(0)} = \tilde{x}^{(0)} = z$, we have $e_0 = 0$.

First, the pre-activations are bounded. By (N1)-(N3), each row of A_ℓ has at most W entries of modulus at most Λ (for $\ell = 0$ because $d + 1 \le W$), and $\|x^{(\ell)}\|_\infty \le \mathcal{A}_{\max}$ for every ℓ (including $\ell = 0$, where $\|z\|_\infty \le R_*$), so

$$\|v^{(\ell)}\|_\infty \le \Lambda(W\mathcal{A}_{\max} + 1) \le \mathrm{m}, \tag{E.24}$$

and identically for $\tilde{v}^{(\ell)}$, since $f_{\tilde{\vartheta}}$ also lies in $\mathcal{F}$. This is the step that (N3) exists to supply.

Second, one layer of the recursion. Splitting the difference into a term from the activations and a term from the parameters

$$\|v^{(\ell)}-\tilde v^{(\ell)}\|_\infty \le \|A_\ell(x^{(\ell)}-\tilde x^{(\ell)})\|_\infty+\|(A_\ell-\tilde A_\ell)\tilde x^{(\ell)}\|_\infty+\|b_\ell-\tilde b_\ell\|_\infty \le \Lambda W e_\ell+\delta(W\mathcal{A}_{\max}+1). \tag{E.25}$$

On the interval $[-\mathrm{m},\mathrm{m}]$ containing both $v^{(\ell)}$ and $\tilde v^{(\ell)}$ by (E.24), the mean value theorem and (E.22) give $|\sigma_\ell(u)-\sigma_\ell(u')|\le 3\mathrm{m}^2|u-u'|$. Hence

$$e_{\ell+1}\le 3\mathrm{m}^2\left[\Lambda W e_\ell+\delta(W\mathcal{A}_{\max}+1)\right]=\kappa e_\ell+c_0\delta,\quad \kappa:=3\mathrm{m}^2\Lambda W,\quad c_0:=3\mathrm{m}^2(W\mathcal{A}_{\max}+1). \tag{E.26}$$

Note $\kappa\le \mathrm{m}^3$ and $c_0\le \mathrm{m}^3$, both by the definition of m, and $\kappa\ge 1$.

Third, unroll. From $e_0=0$ and $e_{\ell+1}\le \kappa e_\ell+c_0\delta$ one gets, for $1\le \ell\le L$, $e_\ell\le c_0\delta\sum_{j=0}^{\ell-1}\kappa^j\le c_0\delta\ell\kappa^{\ell-1}\le \delta L\mathrm{m}^{3L}$. Finally the output layer contributes

$$|f_\vartheta(z)-f_{\tilde\vartheta}(z)|\le \Lambda W e_L+\delta(W\mathcal{A}_{\max}+1)\le \delta(\mathrm{m}\cdot L\mathrm{m}^{3L}+\mathrm{m})\le 2L\mathrm{m}^{3L+1}\delta, \tag{E.27}$$

which is the first bound in (E.16).

By the chain rule applied to (E.2), $\nabla_z f_\vartheta = A_L D_{L-1}A_{L-1}\cdots D_0A_0$ with $D_\ell:=\mathrm{diag}(\sigma_\ell'(v^{(\ell)}))$. Each factor is bounded in the $\ell_\infty\to\ell_\infty$ operator norm, the maximal absolute row sum: $\|A_\ell\|\le \Lambda W\le \mathrm{m}$ and $\|D_\ell\|\le 3\mathrm{m}^2\le \mathrm{m}^3$ by (E.24) and (E.22). Each factor is also stable: $\|A_\ell-\tilde A_\ell\|\le W\delta\le \mathrm{m}\delta$, while (E.22) gives $\|D_\ell-\tilde D_\ell\|\le 6\mathrm{m}\|v^{(\ell)}-\tilde v^{(\ell)}\|_\infty\le 6\mathrm{m}(\Lambda W e_\ell+\delta\mathrm{m})\le 12L\mathrm{m}^{3L+2}\delta\le 4L\mathrm{m}^{3L+3}\delta$, the last step by $\mathrm{m}\ge 3$. Writing the product difference as a telescopic sum

$$\prod_{i=1}^{2L+1}M_i-\prod_{i=1}^{2L+1}\tilde M_i=\sum_{i=1}^{2L+1}\left(\prod_{j<i}M_j\right)(M_i-\tilde M_i)\left(\prod_{j>i}\tilde M_j\right), \tag{E.28}$$

and bounding every retained factor by m^3 and every difference by $4L\mathrm{m}^{3L+3}\delta$, we obtain $\|\nabla_z f_\vartheta-\nabla_z f_{\tilde\vartheta}\|_\infty\le (2L+1)\mathrm{m}^{6L}\cdot 4L\mathrm{m}^{3L+3}\delta\le 12L^2\mathrm{m}^{9L+3}\delta$ using $2L+1\le 3L$, at every $z\in Z$. The $\ell_\infty\to\ell_\infty$ operator norm of a row vector is its ℓ^1 norm, so this is the second bound in (E.16).

(ii) Let $\mathcal{P}'$ be the set of parameter vectors, in the layout above, of elements of $\mathcal{F}'$. For $S\subseteq\{1,\dots,P\}$ with $|S|=P_0$ let $\mathcal{P}'_S$ consist of those $\vartheta\in\mathcal{P}'$ whose entries outside S vanish. By (N2) and $P_0\le P$ every element of $\mathcal{P}'$ lies in some $\mathcal{P}'_S$, and $\mathcal{P}'_S$ is contained in $Q_S:=\{\vartheta:\vartheta_i=0 \text{ for } i\notin S,\ \|\vartheta\|_\infty\le\Lambda\}$, which is isometric to $[-\Lambda,\Lambda]^{P_0}$. Put $\delta:=\varepsilon/(2\mathrm{L}'_*)$, so that $\delta/2\le 1\le\Lambda$. By (E1), $N(\delta/2,\mathcal{P}'_S,\|\cdot\|_\infty)\le N(\delta/2,Q_S,\|\cdot\|_\infty)\le(4\Lambda/\delta)^{P_0}$, so by (E2) each nonempty $\mathcal{P}'_S$ contains a δ-net $\mathcal{S}_S\subseteq\mathcal{P}'_S$ with $|\mathcal{S}_S|\le(4\Lambda/\delta)^{P_0}=(8\Lambda\mathrm{L}'_*/\varepsilon)^{P_0}$. For $\vartheta\in\mathcal{P}'_S$ and $\tilde\vartheta\in\mathcal{S}_S$ with $\|\vartheta-\tilde\vartheta\|_\infty\le\delta$, part (i), $|\cdot|\le|\cdot|_1$ and $\mathrm{L}_*\le\mathrm{L}'_*$ give, for every $z\in Z$,

$$\big|(f_\vartheta,\nabla_z f_\vartheta)(z)-(f_{\tilde\vartheta},\nabla_z f_{\tilde\vartheta})(z)\big|\le|f_\vartheta(z)-f_{\tilde\vartheta}(z)|+|\nabla_z f_\vartheta(z)-\nabla_z f_{\tilde\vartheta}(z)|_1\le(\mathrm{L}_*+\mathrm{L}'_*)\delta\le\varepsilon.$$

Hence the jets of $f_{\tilde\vartheta}$, $\tilde\vartheta\in\bigcup_S\mathcal{S}_S$, form an ε-net of $J\mathcal{F}'$ consisting of jets of elements of $\mathcal{F}'$. By (E3) there are at most $(eP/P_0)^{P_0}$ sets S, which gives the cardinality bound and, taking logarithms, (E.17).

(iii) Put $\mathfrak{F}:=\mathcal{NN}_{\varrho,\Theta}(2W_N,L+1,\Lambda,\bar P_0,\mathcal{A}\vee 3\bar B)$, the architecture of Ψ_N. Its constant m in part (i) is $\bar{\mathrm{m}}$ of (E.18) and its depth is $L+1$, so its constant L'_* is $\bar{\mathrm{L}}':=12(L+1)^2\bar{\mathrm{m}}^{9L+12}$. By (E.21) and $d+1\le W_N$, its number of parameter entries $P_{\mathfrak{F}}$ satisfies

$$P_{\mathfrak{F}}-(2P_{\mathcal{H}}+2)=2(L+1)W_N^2+2W_N-3>0,\qquad P_{\mathfrak{F}}\le(2W_N+1)\big(2(L+1)W_N+1\big)\le\bar P.$$

Hence $1\le\bar P_0\le P_{\mathfrak{F}}$, part (ii) applies to every nonempty subset of $\mathfrak{F}$, and since $P\mapsto\log(eP/\bar P_0)$ is increasing, $\bar P$ may replace $P_{\mathfrak{F}}$ in (E.17). Fix $0<\varepsilon\le 1$ and write $\mathfrak{e}(\varepsilon):=\bar P_0[\log(e\bar P/\bar P_0)+\log(8\Lambda\bar{\mathrm{L}}')+\log(1/\varepsilon)]$.

Trial class. Every $h\in\mathcal{H}_N$ equals f^χ for some $f\in\mathfrak{F}$. Indeed, let $h=(\Theta\circ g)^\chi$ with $g=A_Lx^{(L)}+b_L\in\mathcal{NN}_\varrho(W_N,L,\Lambda,P_0,\mathcal{A})$. Turning the readout of g into a hidden layer of width one with activation Θ, and reading this layer out with weight 1 and bias 0, gives a network that computes $\Theta\circ g$ on all of $I\times\mathbb{R}^d$, has depth $L+1$ and hidden widths at most W_N, entries bounded by $\Lambda\ge1$, at most $P_0+1\le\bar P_0$ nonzero entries, and hidden activations bounded on Z by $\mathcal{A}\vee 3\bar B$, as $|\Theta|\le 3\bar B$. Let $\mathcal{F}_{\mathcal{H}}:=\{f\in\mathfrak{F}:f^\chi\in\mathcal{H}_N\}$. For $f\in\mathfrak{F}$ and $(t,x)\in I\times\mathbb{R}^d$ the chain rule gives

$$(f^\chi,\partial_t f^\chi,\nabla f^\chi)(t,x)=\big(f,\partial_t f,D\chi_R(x)^\top\nabla_x f\big)(t,\chi_R(x)), \tag{E.29}$$

where $D\chi_R(x)$ is diagonal with entries in $[-1,1]$ and $(t,\chi_R(x)) \in Z$. Hence, for $f,\tilde f \in \mathfrak{F}$, $\|J(f^\chi) - J(\tilde f^\chi)\|_\infty \le \|(f,\nabla_z f) - (\tilde f,\nabla_z \tilde f)\|_{\infty,Z}$, and part (ii) applied to $\mathcal{F}_{\mathcal{H}}$ yields an ε-net of $J\mathcal{H}_N$ consisting of jets of elements of $\mathcal{H}_N$, so that $\log N(\varepsilon, J\mathcal{H}_N, \|\cdot\|_\infty) \le \mathfrak{e}(\varepsilon)$.

Generators. Let $\Phi_N := \{\phi^\chi : \phi \in \mathfrak{F},\ \sup_{I\times\mathbb{R}^d}(|\phi^\chi| + |\nabla\phi^\chi| + |\partial_t\phi^\chi|) \le \Gamma\}$, so that $\Psi_N = \{c\,\omega\,u : c\in\mathbb{R},\ u\in\Phi_N\}$ by (E.7), and $J\Phi_N := \{(u,\nabla u) : u\in\Phi_N\}$. Deleting the time component in (E.29) does not increase distances, so the same argument gives $\log N(\varepsilon, J\Phi_N, \|\cdot\|_\infty) \le \mathfrak{e}(\varepsilon)$.

Unit ball of the cone scale. Let $\psi \in \Psi_N^\bullet$. By (E.9) there is a representation $\psi = c\,\omega\,u$ with $u\in\Phi_N$ and $|c|\le 2$, and since ω depends on t only, $(\psi,\nabla\psi) = c\,\omega\,(u,\nabla u)$. Put $\eta := \varepsilon/(2\Gamma)$ and let $\mathcal{C}$ consist of the points $\max\{-2,\min\{2,-2+(2k-1)\eta\}\}$, $1\le k\le\lceil 2/\eta\rceil$. Every point of $[-2,2]$ lies within η of $\mathcal{C}$, because the intervals $[-2+2(k-1)\eta, -2+2k\eta]$ cover $[-2,2]$ and projecting onto $[-2,2]$ does not increase distances to its points, and $|\mathcal{C}| \le 1 + 2/\eta = 1 + 4\Gamma/\varepsilon$. Let $\mathcal{U}$ be an $(\varepsilon/4)$-net of $J\Phi_N$. If $c'\in\mathcal{C}$ and $U'\in\mathcal{U}$ satisfy $|c-c'|\le\eta$ and $\|(u,\nabla u) - U'\|_\infty \le \varepsilon/4$, then, using $0\le\omega\le 1$, $|(u,\nabla u)| \le |u| + |\nabla u| \le \Gamma$ and $|c'|\le 2$,

$$|c\,\omega\,(u,\nabla u) - c'\omega\,U'| \le |c-c'|\,|(u,\nabla u)| + |c'|\,|(u,\nabla u) - U'| \le \eta\Gamma + \varepsilon/2 = \varepsilon$$

everywhere on $I\times\mathbb{R}^d$. Hence $N(\varepsilon, J\Psi_N^\bullet, \|\cdot\|_\infty) \le (1+4\Gamma/\varepsilon)\,N(\varepsilon/4, J\Phi_N, \|\cdot\|_\infty)$, and with $\log(1+4\Gamma/\varepsilon) \le \log(1+4\Gamma) + \log(1/\varepsilon)$ for $\varepsilon\le 1$,

$$\log N(\varepsilon, J\Psi_N^\bullet, \|\cdot\|_\infty) \le \log(1+4\Gamma) + \bar P_0\left[\log\frac{e\bar P}{\bar P_0} + \log(32\Lambda\bar{\mathrm{L}}')\right] + (\bar P_0+1)\log\frac{1}{\varepsilon}.$$

Conclusion. With $u := \log(1/\varepsilon) \ge 0$, both logarithms are at most $a + (\bar P_0+1)u$, where $a := \log(1+4\Gamma) + \bar P_0[\log(e\bar P/\bar P_0) + \log(32\Lambda\bar{\mathrm{L}}')]$. The quotient $(a+(\bar P_0+1)u)/(1+u)$ is a convex combination of a and $\bar P_0+1$, and $a \ge \bar P_0(1+\log 32) \ge \bar P_0 + 1 \ge 1$, since $\bar P_0 \le \bar P$ and $\Lambda, \bar{\mathrm{L}}' \ge 1$. Therefore $\mathbb{V}_N \le a$, and substituting $\log\bar{\mathrm{L}}' = \log(12(L+1)^2) + (9L+12)\log\bar{\mathrm{m}}$ gives (E.19).

For (E.20), insert the stated magnitudes. Then $\bar P_0 \lesssim KN^d$, $\log(e\bar P/\bar P_0) \le \log(e\bar P) \lesssim \log(LKN^d) \lesssim \log n$ and $\log(384\Lambda(L+1)^2) \lesssim \log n$. Next, $\log\bar{\mathrm{m}} \lesssim \log n$, because $\Lambda, W_N, \mathcal{A}$ are $\mathrm{poly}(n)$ while $\bar B, \kappa_\Theta$ are fixed and $R_* = \mathrm{polylog}(n)$, and likewise $\log(1+4\Gamma) \lesssim \log n$, since Γ depends on n only through R. As $9L+12 \le 21L$ and $L \lesssim \log N$ with $N\ge 2$, the bracket in (E.19) is $\lesssim \log N\log n$, and multiplying by $\bar P_0$ gives $\mathbb{V}_N \lesssim KN^d\log N\log n$. With $K \le \sigma^{-c}(\log n)^{3/2}$ from (F.13) this is $N^d\,\mathrm{polylog}(n)$. □

Remark E.9. (The sparsity budget is attained) Equation (E.20) rests on the sparsity budget $P_0 \lesssim KN^d$ of (N2): a dense layer of width $W_N \asymp KN^d$ carries $\bar P \asymp K^2N^{2d}$ parameters, and (E.19) would then give $\mathbb{V}_N \asymp N^{2d}\mathrm{polylog}(n)$ and the exponent $2(\beta-1)/(2(\beta-1)+2d)$ in Theorem H.8. Two facts from (Gühring & Raslan, 2021) help us. Proposition 4.8 there produces $O(\varepsilon^{-d/(\beta-k)})$ nonzero weights, which at the accuracy $\varepsilon \asymp N^{-(\beta-k)}$ used in Lemma F.7 is exactly $O(N^d)$ per Chebyshev mode, hence $O(N^d\mathrm{polylog}(n))$ in total. Also Corollary 3.8 shows that any architecture with encodable weights achieving accuracy ε in $W^{k,\infty}$ on the unit ball of $W^{\beta,\infty}$ needs at least $C\varepsilon^{-d/(\beta-k)}/\log_2(1/\varepsilon)$ weights.

F THE APPROXIMATION FLOOR

Every constant in this section and the next depends on β. Write

$$\mathrm{A}_\beta := C_\beta\bar B\sigma_*^{-2\beta}(2R)^{\beta-1}, \quad C_\beta := \mathfrak{B}_{\beta+1}\beta!\,(2\sec(\pi/6))^{\beta+1}, \tag{F.1}$$

with $\mathfrak{B}_{\beta+1}$ the Bell number. The factor $C_\beta\sigma_*^{-2\beta}$ is the derivative envelope of Step 1 of Lemma F.8. The factor $(2R)^{\beta-1}$ is the price of rescaling Q_R onto the unit cube in Step 3. Since $R \asymp \sqrt{\log n}$ by (F.13), the whole of A_β is $\mathrm{polylog}(n)$ at fixed β, but the bound grows factorially in β, see the remark on constants in Section 8. We write

$$\varepsilon_N := \mathrm{A}_\beta N^{-(\beta-1)} \tag{F.2}$$

for the approximation scale, and from here on A_β replaces the σ^{-c} used to this point wherever a β-dependent constant occurs.

Lemma F.1. *(Weak residual approximation) Suppose Assumptions 1–2 hold and that $h_N \in \mathcal{H}_N$ and $e_N > 0$ satisfy, with $\varphi_N := h_N - h^*$,*

$$\|\varphi_N\|_{L^4(\nu\otimes\rho)} + \sup_{t\in I}\|\varphi_N(t,\cdot)\|_{\rho_t} + \||\nabla\varphi_N|\|_{L^4(\nu\otimes\rho)} + \|\partial_t\varphi_N\|_{L^2(\nu\otimes\rho)} \le e_N. \tag{F.3}$$

Then $\inf_{h\in\mathcal{H}_N}\mathcal{P}^w(h) \le \mathcal{P}^w(h_N) \le 4(1+M_4)^2 e_N^2 \le Cd\sigma_*^{-4}e_N^2$.

Proof. Lemma C.6 bounds the dual norm of the residual by first-order quantities only: $\|\mathcal{R}_t[h_N]\|_{H^{-1}(\rho_t)} \le \|\partial_t\varphi_N\|_{\rho_t} + \||\nabla\varphi_N|\|_{\rho_t} + M_4\||\nabla\varphi_N|\|_{L^4(\rho_t)}$. Since ρ_t is a probability measure, $\|\cdot\|_{\rho_t} \le \|\cdot\|_{L^4(\rho_t)}$, so the right-hand side is at most $\|\partial_t\varphi_N\|_{\rho_t} + (1+M_4)\||\nabla\varphi_N|\|_{L^4(\rho_t)}$. Squaring, using $(a+b)^2 \le 2a^2+2b^2$ and integrating against ν,

$$\mathcal{P}^w(h_N) \le 2\|\partial_t\varphi_N\|^2_{L^2(\nu\otimes\rho)} + 2(1+M_4)^2\int_I \||\nabla\varphi_N|\|^2_{L^4(\rho_t)}\nu(dt).$$

By Jensen's inequality for the concave map $\sqrt{\cdot}$, $\int_I(\int|\nabla\varphi_N|^4\rho_t)^{1/2}\nu(dt) \le \||\nabla\varphi_N|\|^2_{L^4(\nu\otimes\rho)}$. Hence $\mathcal{P}^w(h_N) \le 2e_N^2 + 2(1+M_4)^2e_N^2 \le 4(1+M_4)^2e_N^2$, and $M_4 \le C\sqrt{d}\sigma_*^{-2}$ by Lemma B.1. □

It remains to exhibit an $h_N \in \mathcal{H}_N$ satisfying (F.3) with e_N of order $\mathrm{A}_\beta N^{-(\beta-1)}$ up to polylogarithmic factors. The construction separates the two variables – Chebyshev truncation in t and spline quasi-interpolation in x.

We will use fact that a function analytic in a neighbourhood of an interval has geometrically decaying Chebyshev coefficients. For $\mathrm{r} > 1$ the Bernstein ellipse $E_\mathrm{r} \subset \mathbb{C}$ is the image of the circle $\{|w| = \mathrm{r}\}$ under the map $w \mapsto \frac{1}{2}(w + w^{-1})$. It is the open ellipse with foci ±1, semi-axes

$$a_\mathrm{r} = \frac{1}{2}(\mathrm{r} + \mathrm{r}^{-1}), \quad b_\mathrm{r} = \frac{1}{2}(\mathrm{r} - \mathrm{r}^{-1}), \tag{F.4}$$

so that E_r shrinks to $[-1,1]$ as $\mathrm{r} \downarrow 1$. Write T_j for the Chebyshev polynomials on $[-1,1]$ and, for $g : [-1,1] \to \mathbb{R}$ integrable against the Chebyshev weight, $a_j(g) := \frac{2}{\pi}\int_{-1}^1 g(\theta)T_j(\theta)(1-\theta^2)^{-1/2}d\theta$ for $j \ge 1$ (and half that for $j = 0$).

Lemma F.2. *(Chebyshev truncation on a strip) Let $J := [t_0, t_1]$ with $\ell = t_1 - t_0$, let $\iota(\theta) := t_0 + \frac{\ell}{2}(\theta+1)$ map $[-1,1]$ onto J, and for $\eta > 0$ let $S_\eta := \{z \in \mathbb{C} : \mathrm{dist}(z, J) < \eta\}$ be the η-neighbourhood of J. Suppose g is holomorphic on S_η with $|g| \le M$ there. Then:*

(i) $\iota^{-1}(S_\eta)$ contains the Bernstein ellipse E_r for every $\mathrm{r} \le 1 + 2\eta/\ell$,

(ii) the Chebyshev coefficients of $g\circ\iota$ obey $|a_j(g\circ\iota)| \le 2M\mathrm{r}^{-j}$,

(iii) for every $K \ge 1$, writing $\Pi_K g := \sum_{j\le K} a_j(g\circ\iota)T_j\circ\iota^{-1}$,

$$\sup_{t\in J}|g(t) - \Pi_K g(t)| \le \frac{2M\mathrm{r}^{-K}}{\mathrm{r}-1}, \quad \sup_{t\in J}|g'(t) - (\Pi_K g)'(t)| \le \frac{2}{\ell}\cdot\frac{10M(K+1)^2\,\mathrm{r}^{2-K}}{(\mathrm{r}-1)^3}. \tag{F.5}$$

Proof. (i) A point of E_r is $a_\mathrm{r}\cos\vartheta + ib_\mathrm{r}\sin\vartheta$. Its distance to $[-1,1]$ is $b_\mathrm{r}|\sin\vartheta| \le b_\mathrm{r}$ where $|a_\mathrm{r}\cos\vartheta| \le 1$. Elsewhere, with $c := |\cos\vartheta| > 1/a_\mathrm{r}$, the squared distance is $F(c) = (a_\mathrm{r}c-1)^2 + b_\mathrm{r}^2(1-c^2)$ and $F'(c) = 2c(a_\mathrm{r}^2 - b_\mathrm{r}^2) - 2a_\mathrm{r} = 2(c - a_\mathrm{r}) < 0$ by (F.4), so $F(c) \le F(1/a_\mathrm{r}) = b_\mathrm{r}^4/a_\mathrm{r}^2 \le b_\mathrm{r}^2$. Hence $\mathrm{dist}(E_\mathrm{r}, [-1,1]) = b_\mathrm{r}$ exactly. Under ι lengths scale by $\ell/2$, and writing $\mathrm{r} = 1+\varepsilon$ with $\varepsilon \le 1$ one has $b_\mathrm{r} = \frac{\varepsilon(2+\varepsilon)}{2(1+\varepsilon)} \le \varepsilon$, so the distance is at most $\ell\varepsilon/2 \le \eta$ when $\varepsilon \le 2\eta/\ell$.

(ii) and the first half of (iii) are the Bernstein estimates: for $g\circ\iota$ holomorphic and bounded by M on E_r, $|a_j| \le 2M\mathrm{r}^{-j}$ and the truncation error is at most $2M\mathrm{r}^{-K}/(\mathrm{r}-1)$ (Trefethen, 2013) (Theorems 8.1–8.2). The second follows from the first by summing the geometric tail, $\sum_{j>K}|a_j|\|T_j\|_\infty \le 2M\sum_{j>K}\mathrm{r}^{-j}$.

For the derivative bound, Markov's inequality gives $\|T_j'\|_{\infty,[-1,1]} = j^2$, so $\sum_{j>K} |a_j| \|T_j'\|_\infty \le 2M \sum_{j>K} j^2 \mathrm{r}^{-j}$. For $x \in (0,1)$ and $m \ge 1$

$$\sum_{j \ge m} j^2 x^j = x^m \sum_{i \ge 0} (i+m)^2 x^i \le x^m \left[\frac{m^2}{1-x} + \frac{2mx}{(1-x)^2} + \frac{x(1+x)}{(1-x)^3} \right] \le \frac{5m^2 x^m}{(1-x)^3}, \tag{F.6}$$

using $m^2 + 2m + 2 \le 5m^2$, with $x = \mathrm{r}^{-1}$, so $1 - x = (\mathrm{r}-1)/\mathrm{r}$, and $m = K+1$ this is $5(K+1)^2 \mathrm{r}^2 \mathrm{r}^{-K} (\mathrm{r}-1)^{-3}$. The chain rule contributes $|(\iota^{-1})'| = 2/\ell$. □

Lemma F.3. *(Exact arithmetic with the rectified cubic) Let $\varrho(u) = (u \vee 0)^3$ and let $M \ge 1$. Then as identities of functions:*

(i) (Cube) $u^3 = \varrho(u) - \varrho(-u)$ for all $u \in \mathbb{R}$,

(ii) (Square) $u^2 = \frac{1}{6M} \left[\varrho(u+M) + \varrho(M-u) - 2M^3 \right]$ for all $|u| \le M$,

(iii) (Identity) $u = \frac{1}{6M^2} \left[\varrho(u+M) - \varrho(M-u) - 2\varrho(u) + 2\varrho(-u) \right]$ for all $|u| \le M$,

(iv) (Product) $uv = \frac{1}{4} \left[(u+v)^2 - (u-v)^2 \right]$ for all $|u|, |v| \le M/2$, hence $(u,v) \mapsto uv$ is realized exactly by one hidden layer of four ϱ-units,

(v) (Cubic B-splines) For knots $\tau_0 < \ldots < \tau_4$ the cardinal cubic B-spline $B(u) := (\tau_4 - \tau_0) \sum_{i=0}^4 (\prod_{j \ne i} (\tau_i - \tau_j))^{-1} \varrho(\tau_i - u)$ is realized exactly by one hidden layer of five ϱ-units, with inner weights ± 1 and outer weights of modulus $O(\mathfrak{h}^{-3})$ for a uniform mesh of width $\mathfrak{h}$.

Proof. (i) For $u \ge 0$ the right side is $u^3 - 0$, for $u < 0$ its is $0 - (-u)^3 = u^3$.

(ii) For $|u| \le M$ both $u+M$ and $M-u$ are nonnegative, so $\varrho(u+M) + \varrho(M-u) = (u+M)^3 + (M-u)^3 = (u+M)^3 - (u-M)^3 = 6u^2 M + 2M^3$, using $(a+b)^3 - (a-b)^3 = 6a^2 b + 2b^3$, with $a = u, b = M$.

(iii) Similarly $\varrho(u+M) - \varrho(M-u) = (u+M)^3 + (u-M)^3 = 2u^3 + 6uM^2$, using $(a+b)^3 + (a-b)^3 = 2a^3 + 6ab^2$. Subtracting $2u^3$ in the form (i) leaves $6uM^2$.

(iv) The polarization identity, with (ii) applied to $u \pm v$, both of modulus at most M.

(v) The divided-difference representation of the B-spline of degree 3 in the knots $\tau_0, \ldots, \tau_4$ is exactly the stated combination of truncated cubics $(\tau_i - u)_+^3 = \varrho(\tau_i - u)$ (de Boor, 1978) (Chapter IX). On a uniform mesh of width $\mathfrak{h}$, $\prod_{j \ne i} |\tau_i - \tau_j| \asymp \mathfrak{h}^4$ and $\tau_4 - \tau_0 = 4\mathfrak{h}$, giving outer weights of modulus $O(\mathfrak{h}^{-3})$. □

Definition F.4. (Exact partition of unity of radius a_*) Let $j \in \mathbb{N}_0$ and $\mathrm{a}_* \in \mathbb{N}$. For $N \in \mathbb{N}$ put $M_N := \{-\mathrm{a}_*, \ldots, N + \mathrm{a}_*\}^d$ A family $\Psi^{(j,N)} := \{\varphi_m : m \in M_N\}$ of functions $\varphi_m : \mathbb{R}^d \to \mathbb{R}$ is an exact j-PU of radius a_* for the activation ϱ if there is $C = C(d,j)$ such that for every N and every $k \in \{0, \ldots, j\}$:

(i) $\|\varphi_m\|_{W^{k,\infty}(\mathbb{R}^d)} \le C N^k$ for every $m \in M_N$,

(ii) $\varphi_m \equiv 0$ on $\{x : \|x - m/N\|_\infty \ge \mathrm{a}_*/N\}$,

(iii) $\sum_{m \in M_N} \varphi_m = 1$ on $(0,1)^d$,

(iv) for each m there is a network Φ_m with d-dimensional input and output, two layers and at most C nonzero weights, such that $\prod_{l=1}^d [R_\varrho(\Phi_m)]_l = \varphi_m$, $\|R_\varrho(\Phi_m)\|_{W^{k,\infty}((0,1)^d)} \le C N^k$ and $\|\Phi_m\|_{\max} \le CN$.

Lemma F.5. *(The cubic B-spline family is an exact 3-PU of radius 2 for ϱ_3) Let B be the cardinal cubic B-spline of Lemma F.3 with unit knot spacing, i.e. $\tau_i = i - 2$ for $i = 0, \ldots, 4$, and set*

$$\varphi_m(x) := \prod_{l=1}^d B(N x_l - m_l), \quad m \in M_N = \{-2, \ldots, N+2\}^d. \tag{F.7}$$

Then $\Psi^{(3,N)} := \{\varphi_m : m \in M_N\}$ is an exact 3-PU of radius $\mathrm{a}_* = 2$ for ϱ_3, with $C = 3^d \vee 15d$.

Proof. Lemma F.3 writes $B(u) = 4\sum_{i=0}^{4}(\prod_{j\neq i}(\tau_i - \tau_j))^{-1}\varrho_3(\tau_i - u)$, whose outer coefficients are $(\frac{1}{6}, -\frac{2}{3}, 1-\frac{2}{3}, \frac{1}{6})$. The standard properties of the cardinal cubic B-spline (de Boor, 1978, Ch. IX) are $B \geq 0$, $\mathrm{supp}B = [-2, 2]$, $\sum_{m\in\mathbb{Z}} B(\cdot - m) \equiv 1$, $B \in C^2(\mathbb{R}) \cap W^{3,\infty}(\mathbb{R})$, with

$$\|B\|_\infty = \|B'\|_\infty = \frac{2}{3}, \quad \|B''\|_\infty = 2, \quad \|B'''\|_\infty = 3. \tag{F.8}$$

(i) For $|\alpha| = k \leq 3$, $\partial^\alpha \varphi_m(x) = N^k \prod_l B^{(\alpha_l)}(Nx_l - m_l)$, so $\|\varphi_m\|_{W^{k,\infty}} \leq 3^d N^k$ by (F.8).

(ii) $B(Nx_l - m_l) = 0$ once $|x_l - m_l/N| \geq 2/N$, so φ_m vanishes off $\{\|x - m/N\|_\infty < 2/N\}$. This is where the radius $\mathrm{a}_* = 2$ enters. The radius is a convenience: $b(u) := B(3u - 1) + B(3u) + B(3u + 1)$ vanishes for $|u| \geq 1$, is realized by fifteen ϱ_3-units, and satisfies $\sum_{m\in\mathbb{Z}} b(u - m) = \sum_{k\in\mathbb{Z}} B(3u - k) = 1$, so radius 1 is also attainable, radius 2 is used because it keeps the construction and its constants simplest.

(iii) Fix $x \in (0, 1)^d$ and l. Then $B(Nx_l - m_l) \neq 0$ forces $m_l \in (Nx_l - 2, Nx_l + 2) \subset (-2, N + 2)$, so every index contributing at x already lies on M_N and

$$\sum_{m\in M_N} \varphi_m(x) = \prod_{l=1}^{d} \sum_{m_l\in\mathbb{Z}} B(Nx_l - m_l) = 1. \tag{F.9}$$

The overhang is needed, since over $\{0, \ldots, N\}^d$ the sum equals $(5/6)^d$ at the corners of the cube and differs from 1 on a boundary layer of width $2/N$.

(iv) $u \mapsto B(Nu - m_l)$ is one hidden layer of five ϱ_3 units with inner weight $-N$, biases $\tau_i + m_l$ of modulus at most $N + 4$, and outer weights of modulus at most 1. Parallelizing the d coordinates (Lemma C.2 of (Gühring & Raslan, 2021)) gives Φ_m with two layers, at most $15d$ nonzero weights, $\prod_l [R_{\varrho_3}(\Phi_m)]_l = \varphi_m$, $\|R_{\varrho_3}(\Phi_m)\|_{W^{k,\infty}} \leq 3N^k$ and $\|\Phi_m\|_{\max} \leq N + 4 \leq 3N$ for $N \geq 2$.

□

Lemma F.6. *(Localization under a widened partition of unity) Let $\Psi^{(j,N)}$ be an exact j-PU of radius a_*, $k \in \{0, \ldots, j\}$, $n \in \mathbb{N}_{\geq k+1}$ and $1 \leq p \leq \infty$. Then the conclusion of Lemma D.1 of (Gühring & Raslan, 2021) holds: there are polynomials $p_{f,m}$ of degree $\leq n - 1$, the averaged Taylor polynomials of f on the balls $\hat{\Omega}_{m,N} := B_{\mathrm{a}_*/N,\|\cdot\|_\infty}(m/N)$, independent of k, such that $f_N := \sum_{m\in M_N} \varphi_m p_{f,m}$ satisfies*

$$\|f - f_N\|_{W^{k,p}((0,1)^d)} \leq C\|f\|_{W^{n,p}((0,1)^d)} N^{-(n-k)}, \quad C = C(d, n, p, k, \mathrm{a}_*). \tag{F.10}$$

Proof. First, in Step 1 the Bramble-Hilbert polynomial is taken from their Lemma B.4 on $\hat{\Omega}_{m,N}$ rather than on $\Omega_{m,N} = B_{1/N,\|\cdot\|_\infty}(m/N)$. That lemma holds on a ball of any radius c/N with a constant $C(n, d, c)$, and returns $(\mathrm{a}_*/N)^{n-k}$, i.e. the same rate with the constant multiplied by a_*^{n-k}. The polynomial is still the averaged Taylor polynomial, hence still independent of k, which is what makes the simultaneity in Lemma F.7 possible.

Second, their Steps 2 and 3 split the index set at $\|\tilde{m} - m\|_\infty > \mathrm{a}_* + 1$ instead of > 1. By (ii) of Definition F.4, $\mathrm{supp}\,\varphi_m \cap \Omega_{\tilde{m},N} = \emptyset$ as soon as $\|\tilde{m} - m\|_\infty > \mathrm{a}_* + 1$, so the far contribution vanishes identically, as it does for an exact PU in their Step 2. The near contribution now runs over at most $(2\mathrm{a}_* + 3)^d$ indices instead of 3^d and is estimated as in their Step 1, on $\hat{\Omega}_{m,N}$. For $\mathrm{a}_* = 2$ the count is 7^d.

Third, in their Step 4a the term $\|\tilde{f}(\mathbf{1}_{(0,1)^d} - \sum_m \varphi_m)$ vanishes identically by (iii), and in their Step 4b Hölder inequality is applied with $(2\mathrm{a}_* + 3)^d$ in place of 3^d while the overlap decomposition preceding their Eq. (D.9) uses $(2\mathrm{a}_*)^d$ disjoint classes in place of 2^d. Both are d-dependent constants. □

Lemma F.7. *(Simultaneous approximation by ϱ_3-networks) Let $\beta \geq 2$ be an integer and let f lie in the unit ball of $W^{\beta,\infty}((0, 1)^d)$. There are constants $C, \tilde{N}$ and a depth L_0 depending only on*

(d, β) such that for every $N \geq \tilde{N}$ there is a ϱ_3-network Φ_N with at most L_0 layers and at most $C(N+1)^d$ nonzero weights, all of modulus at most CN^C and encodable with $\lceil C\beta \log_2 N \rceil$ bits, whose realization satisfies simultaneously

$$\|f - R(\Phi_N)\|_{W^{0,\infty}} \leq CN^{-\beta}, \quad \|f - R(\Phi_N)\|_{W^{1,\infty}} \leq CN^{-(\beta-1)}. \tag{F.11}$$

Proof. We assemble this from (Gühring & Raslan, 2021) with the partition of unity supplied by Lemma F.5 rather than by Remark 4.6 of that paper.

Localization. By Lemma F.5 the family (F.7) is an exact 3-PU of radius 2 for ϱ_3, so Lemma F.6 applies with $j = 3, \mathrm{a}_* = 2, n = \beta, p = \infty$. It produces one localized approximant $f_N = \sum_{m \in M_N} \varphi_m p_{f,m}$, whose polynomials are averaged Taylor polynomials and therefore do not depend on k, with

$$\|f - f_N\|_{W^{k,\infty}((0,1)^d)} \leq CN^{-(\beta-k)} \quad \text{simultaneously for } k \in \{0, 1\}. \tag{F.12}$$

Lemma F.6 requires $k \leq j = 3$ and $n = \beta \geq k+1$, so (F.12) holds for $k \in \{0, \ldots, \min(3, \beta-1)\}$, which contains $\{0, 1\}$ because $\beta \geq 2$. Only $k \in \{0, 1\}$ is used, so neither this restriction nor the fact that $\varrho_3 \notin C^3$ makes the chain rule of Corollary B.6 of (Gühring & Raslan, 2021) unavailable at order three ever binds.

Realization. Lemma D.5 of (Gühring & Raslan, 2021) is now applied unchanged. It is insensitive to the widening, since its proof invokes only conditions (i) and (iv) of the partition of unity – the bump network of Definition F.4 and the derivative bound $\|R_\varrho(\Phi_m)\|_{W^{k,\infty}} \leq CN^k$ of (i) – and never the support condition (ii) or the summation condition (iii), which are the two that Lemma F.6 had to absorb. The only trace of the enlarged index set is that its count $C(N+1)^d$ becomes $C|M_N| = C(N+5)^d \leq C'(N+1)^d$. Its side hypothesis that ϱ be three times continuously differentiable near some x_0 with $\varrho''(x_0) \neq 0$ – holds at any $x_0 > 0$, where ϱ_3 agrees with $u \mapsto u^3$ and $\varrho_3''(x_0) = 6x_0 \neq 0$. The lemma realizes f_N to accuracy ε in $W^{k,\infty}$ with at most C layers and $C(N+1)^d$ nonzero weights independently of ε, only the weight magnitude growing, as ε^{-2} times a power of N.

Simultaneity. Apply it once, at the same N, with $k = 1$ and $\varepsilon := N^{-\beta}$. Since $\|\cdot\|_{W^{0,\infty}} \leq \|\cdot\|_{W^{1,\infty}}$, the resulting network approximates f_N to $N^{-\beta}$ in both norms at once, and the triangle inequality against (F.12) gives (F.11). It is the ε-independence of the weight count that makes this free. One may spend accuracy $N^{-\beta}$ on the realization while the localization error is only $N^{-(\beta-1)}$, and pay for it in weight magnitude rather than in weight count.

Encodability. The weights are $\mathrm{poly}(N)$. Round each of them to the grid of spacing $N^{-C\beta}$. The resulting perturbation is controlled by the parameter-Lipschitz bound of Lemma E.8: $\|R(\Phi_\vartheta) - R(\Phi_{\tilde{\vartheta}})\|_{C^1} \leq \mathrm{L}'_* \|\vartheta - \tilde{\vartheta}\|_\infty$ with $\mathrm{L}'_* = 12L^2\mathrm{m}^{9L+3} = \mathrm{poly}(N)$ at the constant depth L_0. Taking C large enough makes the perturbation at most $N^{-\beta}$, absorbed into (F.11), and each surviving weight is encoded by $\lceil C\beta \log_2 N \rceil$ bits. □

Lemma F.8. *(Space-time approximation) Grant Assumptions 1–2, let $\beta \geq 2$ be an integer and $n \geq 2$, and put*

$$R := 2 + \sqrt{(8\beta + 16)\log(2dn)}, \quad \eta := \min\Big\{\frac{\ell}{2}, \frac{\pi\sigma_*^4}{12d(5R+3)}\Big\}, \quad K := \Big\lceil \frac{\ell}{\eta}(\beta+2)\log n \Big\rceil, \tag{F.13}$$

so that $R \asymp \sqrt{\log n}$ and $K \lesssim (\log n)^{3/2}$, and

$$Q_n := 16 C_{\mathrm{F}} \sqrt{d}\, R K^2 (1 + \eta^{-1}) \leq \sigma^{-c} (\log n)^4, \tag{F.14}$$

where $C_{\mathrm{F}} \geq 1, \tilde{N}$ and L_0 are the constants of Lemma F.7. There are constants $C^\sharp, c^\sharp \geq 1$ depending only on (d, β), and n_0 depending only on $(d, \beta, \bar{B}, \ell, \sigma_*, C_t, \kappa_\Theta)$, with the following property. Suppose $n \geq n_0$,

$$\tilde{N} \leq N \leq n, \qquad Q_n \mathrm{A}_\beta N^{-(\beta-1)} \leq \min\{\bar{B},\, G/2,\, G_t/2\}, \tag{F.15}$$

and that the budgets of Definition E.3 satisfy

$$\begin{aligned} &W_N \geq C^\sharp K (N+1)^d, \quad L = L_0 + \lceil \log_2(K+1) \rceil + 1, \quad P_0 \geq C^\sharp K\big((N+1)^d + L\big), \\ &\Lambda \geq n^{c^\sharp}, \quad \mathcal{A} \geq n^{c^\sharp}. \end{aligned} \tag{F.16}$$

Then $\mathcal{H}_N$ contains an h_N such that $\varphi_N := h_N - h^*$ satisfies

$$\begin{aligned} \|\varphi_N\|_{L^4(\nu\otimes\rho)} + \sup_{t\in I}\|\varphi_N(t,\cdot)\|_{\rho_t} &\le 4Q_n \mathrm{A}_\beta N^{-\beta}, \\ \||\nabla\varphi_N|\|_{L^4(\nu\otimes\rho)} + \|\partial_t\varphi_N\|_{L^2(\nu\otimes\rho)} &\le 4Q_n \mathrm{A}_\beta N^{-(\beta-1)}, \end{aligned} \tag{F.17}$$

so that (F.3) *holds with $e_N := 8Q_n\mathrm{A}_\beta N^{-(\beta-1)}$.*

Proof. The residual is a space-time object, so h must depend on t as well as x. A class of resolution N in $d+1$ variables would have complexity N^{d+1}. The proof shows that the time direction costs only a polylogarithmic factor, because h^* is analytic in t on the window while being merely controlled in $W^{\beta,\infty}$ in x. Norms of vectors are Euclidean, and for a function f on Q_R we write $\|f\|_{W^{\beta,\infty}(Q_R)} := \max_{|\gamma|\le\beta}\sup_{Q_R}|\partial^\gamma f|$. The bound in (F.14) follows from $R \le C_\star\sqrt{\log n}$, $\eta^{-1} \le \sigma^{-c}R$ and $K \le \sigma^{-c}R\log n$.

Step 1: holomorphic extension in t. Let $S_\eta := \{z \in \mathbb{C} : \mathrm{dist}(z, I) < \eta\}$ and write $z = s + i\tau$. Since $\sigma_*^2 = 1 - e^{-2(T-t_1)} \le 2(T-t_1)$ and $\eta \le \sigma_*^2/8$, every $z \in S_\eta$ has $s < t_1 + \eta < T$. The functions $\mu_{T-z} = e^{-(T-z)}$ and $\sigma^2_{T-z} = 1 - e^{-2(T-z)}$ are entire, with $|\mu_{T-z}| = e^{-(T-s)} \le 1$ and $\partial_z\mu_{T-z} = \mu_{T-z}$, $\partial_z\sigma^2_{T-z} = -2\mu^2_{T-z}$. Moreover $s \mapsto \sigma^2_{T-s}$ is nonincreasing with derivative in $[-2, 0]$, so $\sigma^2_{T-s} \ge \sigma_*^2 - 2\eta$ for $s < t_1 + \eta$, and $|\sigma^2_{T-z} - \sigma^2_{T-s}| = e^{-2(T-s)}|e^{2i\tau} - 1| \le 2\eta$, hence $|\sigma^2_{T-z}| \ge \sigma_*^2 - 4\eta \ge \sigma_*^2/2$ on S_η. By Bayes' rule, cancelling the factor $e^{-|x|^2/2\sigma^2_{T-t}}$ from numerator and denominator,

$$h^*(t,x) = \frac{\int w(x_0)e^{u_t(x,x_0)}p_0(dx_0)}{\int e^{u_t(x,x_0)}p_0(dx_0)}, \quad u_z(x,x_0) := \frac{\mu_{T-z}\,x\cdot x_0}{\sigma^2_{T-z}} - \frac{\mu^2_{T-z}|x_0|^2}{2\sigma^2_{T-z}}, \tag{F.18}$$

and for $x \in Q_R$ both integrals extend holomorphically to $z \in S_\eta$, their integrands being holomorphic in z and locally uniformly bounded because p_0 is supported in the unit cube. For $z \in S_\eta$, $|\partial_z(\mu/\sigma^2)| = |\mu/\sigma^2 + 2\mu^3/\sigma^4| \le 2\sigma_*^{-2} + 8\sigma_*^{-4} \le 10\sigma_*^{-4}$ and $|\partial_z(\mu^2/2\sigma^2)| = |\mu^2/\sigma^2 + \mu^4/\sigma^4| \le 6\sigma_*^{-4}$, and $|x\cdot x_0| \le |x|_1|x_0|_\infty \le dR$, $|x_0|^2 \le d$. Since u_s is real for real s and the vertical segment from s to z stays in S_η,

$$|\mathrm{Im}\,u_z(x,x_0)| \le |\tau|\sup_{S_\eta}|\partial_z u| \le \eta\,\sigma_*^{-4}d(10R+6), \quad \text{hence} \quad \mathrm{osc}_{x_0}\mathrm{Im}\,u_z \le \frac{\pi}{3}$$

by the choice of η. Then all numbers $e^{i\,\mathrm{Im}\,u_z(x,x_0)}$ lie in a sector of half-angle $\pi/6$, so $|\int f e^{u_z}p_0(dx_0)| \le \sec(\pi/6)\sup_{\mathrm{supp}\,p_0}|f|\,|\int e^{u_z}p_0(dx_0)|$ for bounded measurable f, and in particular the denominator does not vanish. Writing $\langle f\rangle := \int f e^{u_z}p_0 / \int e^{u_z}p_0$, we get $|\langle f\rangle| \le \sec(\pi/6)\sup|f|$ and $|h^*| \le \sec(\pi/6)\bar{B}$ on $S_\eta \times Q_R$. For the x-derivatives, u_z is affine in x with $\partial_{x_k}u_z = \lambda_k(x_0) := \mu_{T-z}x_{0,k}/\sigma^2_{T-z}$, $|\lambda_k| \le 2\sigma_*^{-2}$, so $\partial_{x_k}\langle f\rangle = \langle f\lambda_k\rangle - \langle f\rangle\langle\lambda_k\rangle$ for f independent of x. Iterating, $\partial_x^\gamma h^*$ with $|\gamma| = m$ is the joint cumulant of $(w, \lambda_{k_1}, \dots, \lambda_{k_m})$ under $\langle\cdot\rangle$, that is $\sum_\pi(-1)^{|\pi|-1}(|\pi|-1)!\prod_{B\in\pi}\langle\prod_{i\in B}\cdot\rangle$ over the $\mathfrak{B}_{m+1}$ partitions π of the $m+1$ symbols. Each product is bounded by $\sec(\pi/6)^{m+1}\bar{B}(2\sigma_*^{-2})^m$, so, with C_β as in (F.1),

$$\sup_{S_\eta\times Q_R}|\partial_x^\gamma h^*| \le \mathfrak{B}_{m+1}m!\,(2\sec(\pi/6))^{m+1}\bar{B}\sigma_*^{-2m} \le M_\beta := C_\beta\bar{B}\sigma_*^{-2\beta}, \quad |\gamma| = m \le \beta, \tag{F.19}$$

and $\partial_x^\gamma h^*(\cdot, x)$ is holomorphic on S_η for every $x \in Q_R$. Note $M_\beta(2R)^{\beta-1} = \mathrm{A}_\beta$.

Step 2: truncation in t. Put $\mathrm{r} := 1 + 2\eta/\ell \in (1, 2]$ and let Π_K be the Chebyshev truncation of Lemma F.2, acting in t only. Since $\log(1+y) \ge y/2$ on $[0,1]$, $\mathrm{r}^{-K} \le e^{-K\eta/\ell} \le n^{-(\beta+2)}$, and $\mathrm{r}/(\mathrm{r}-1) \le \ell/\eta$. Lemma F.2 applied to $\partial_x^\gamma h^*(\cdot, x)$, $|\gamma| \le 1$, with $M = M_\beta$, and the fact that Π_K commutes with ∂_x^γ, give uniformly in $(t,x) \in I \times Q_R$

$$\begin{aligned} |h^* - \Pi_K h^*| &\le M_\beta\ell\eta^{-1}n^{-(\beta+2)}, \qquad |\nabla(h^* - \Pi_K h^*)| \le \sqrt{d}M_\beta\ell\eta^{-1}n^{-(\beta+2)}, \\ |\partial_t(h^* - \Pi_K h^*)| &\le 10M_\beta(K+1)^2\ell^2\eta^{-3}n^{-(\beta+2)}. \end{aligned}$$

Let n_0 be so large that $\sqrt{d}\eta^{-1} \le n^2$ and $10(K+1)^2\eta^{-3} \le n^3$ for $n \ge n_0$, this is possible since the left-hand sides are $\sigma^{-c}\mathrm{polylog}(n)$. Using $M_\beta \le \mathrm{A}_\beta$, $\ell \le 1$ and $N \le n$, the three errors are then at most $\mathrm{A}_\beta N^{-\beta}$, $\mathrm{A}_\beta N^{-\beta}$ and $\mathrm{A}_\beta N^{-(\beta-1)}$ respectively.

Step 3: approximation of the coefficients. Write $\Pi_K h^*(t,x) = \sum_{j=0}^{K} c_j^*(x) T_j(\tilde{t})$, $\tilde{t} := 2(t - t_0)/\ell - 1$. Since $\partial_x^\gamma c_j^*$ is the j-th Chebyshev coefficient of $\partial_x^\gamma h^*(\cdot, x)$, Lemma F.2(ii) and (F.19) give $\|c_j^*\|_{W^{\beta,\infty}(Q_R)} \le 2M_\beta \mathrm{r}^{-j}$. Put $S_j := 2M_\beta (2R)^\beta \mathrm{r}^{-j}$ and $f_j(y) := c_j^*(2Ry - R\mathbf{1})/S_j$, $y \in [0,1]^d$, since $2R \ge 1$, f_j lies in the unit ball of $W^{\beta,\infty}((0,1)^d)$. Let Φ_j be the network of Lemma F.7 for f_j and put $\hat{c}_j(x) := S_j R(\Phi_j)((x + R\mathbf{1})/2R)$. By (F.11), with $|\nabla| \le \sqrt{d} \max_k |\partial_k|$, on Q_R

$$|\hat{c}_j - c_j^*| \le C_{\mathrm{F}} S_j N^{-\beta} = 4 C_{\mathrm{F}} R \mathrm{A}_\beta \mathrm{r}^{-j} N^{-\beta},$$
$$|\nabla(\hat{c}_j - c_j^*)| \le \frac{\sqrt{d} C_{\mathrm{F}} S_j}{2R} N^{-(\beta-1)} = 2\sqrt{d} C_{\mathrm{F}} \mathrm{A}_\beta \mathrm{r}^{-j} N^{-(\beta-1)}.$$

Set $H_N(t,x) := \sum_{j=0}^{K} \hat{c}_j(x) T_j(\tilde{t})$. Using $|T_j| \le 1$, $|T_j'| \le j^2 \le K^2$ on $[-1,1]$, $\partial_t T_j(\tilde{t}) = \frac{2}{\ell} T_j'(\tilde{t})$ and $\sum_{j \le K} \mathrm{r}^{-j} \le \ell/\eta$, on $I \times Q_R$

$$|H_N - \Pi_K h^*| \le 4 C_{\mathrm{F}} R \ell \eta^{-1} \mathrm{A}_\beta N^{-\beta}, \qquad |\nabla(H_N - \Pi_K h^*)| \le 2\sqrt{d} C_{\mathrm{F}} \ell \eta^{-1} \mathrm{A}_\beta N^{-(\beta-1)},$$
$$|\partial_t (H_N - \Pi_K h^*)| \le 8 C_{\mathrm{F}} R K^2 \eta^{-1} \mathrm{A}_\beta N^{-\beta}.$$

Step 4: bounds on $I \times Q_R$, constraints, and tails. By Steps 2–3 and the definition of Q_n, on $I \times Q_R$

$$|H_N - h^*| \le Q_n \mathrm{A}_\beta N^{-\beta}, \qquad |\nabla(H_N - h^*)| + |\partial_t (H_N - h^*)| \le Q_n \mathrm{A}_\beta N^{-(\beta-1)}, \tag{F.20}$$

and by (F.15) each left-hand side is at most $\bar{B}$, resp. $\min\{G/2, G_t/2\}$. Let $h_N := (\Theta \circ H_N)^\chi$. Step 5 shows that H_N is a network of the required architecture. Constraints: for $(t,x) \in I \times \mathbb{R}^d$ put $y := \chi_R(x) \in Q_R$. By the chain rule, $|\Theta'| \le 1$ and $|D\chi_R| \le 1$, $|\nabla h_N(t,x)| \le |\nabla H_N(t,y)| \le |\nabla h^*(t,y)| + G/2 \le G$ and $|\partial_t h_N(t,x)| \le |\partial_t H_N(t,y)| \le |\partial_t h^*(t,y)| + G_t/2 \le G_t$, using $|\nabla h^*| \le G/2$ and $|\partial_t h^*| \le G_t/2$ on Z (Proposition E.4(iv)). Hence $h_N \in \mathcal{H}_N$. Inside: on $I \times Q_{R-1}$ the clip is the identity with $D\chi_R = I$, and H_N takes values in $(-\bar{B}, 2\bar{B}]$ since $0 < h^* \le \bar{B}$, as $\Theta = \mathrm{id}$ on $[-\bar{B}, 2\bar{B}]$, also $\Theta' = 1$ there, so $\varphi_N = H_N - h^*$ together with its first derivatives on $I \times Q_{R-1}$, and (F.20) applies. Tails: on $\{|x|_\infty > R - 1\}$ we have $|\varphi_N| \le 3\bar{B}$ because $h_N \in [-2\bar{B}, 3\bar{B}]$ and $h^* \in (0, \bar{B}]$, $|\nabla \varphi_N| \le 3G/2$, and $|\partial_t \varphi_N| \le G_t + |\partial_t h^*|$. Under ρ_t, $x = \mu_{T-t} x_0 + \sigma_{T-t} \xi$ with $|x_0|_\infty \le 1$ and $\xi \sim \mathcal{N}(0, I_d)$, so a union bound over coordinates and the choice of R give $p_R := \rho_t(|x|_\infty > R - 1) \le 2d e^{-(R-2)^2/2} \le n^{-(4\beta+8)}$. By harmonicity, $|\partial_t h^*| \le |x + 2s| \, |\nabla h^*| + |\Delta h^*| \le 3\sigma_*^{-2}(|x| + \sqrt{d}) \cdot 2\sqrt{d}\bar{B}\sigma_*^{-2} + dC_2 \bar{B} \sigma_*^{-4}$ on I by Lemma B.3 and (B.5), and $\||x|\|_{L^4(\rho_t)} \le \sqrt{d} + (d^2 + 2d)^{1/4} \le 3\sqrt{d}$, so $C_h := \sup_{t \in I} \|\partial_t h^*\|_{L^4(\rho_t)} \le (24 + C_2) d \bar{B} \sigma_*^{-4}$. Enlarge n_0 so that $3G + 2G_t + 2C_h \le 2\mathrm{A}_\beta n^3$ for $n \ge n_0$ (the left-hand side is $\sigma^{-c}\sqrt{\log n}$). Splitting each norm into $I \times Q_{R-1}$ and its complement, and using Hölder on the complement,

$$\|\varphi_N(t,\cdot)\|_{L^4(\rho_t)} + \|\varphi_N(t,\cdot)\|_{\rho_t} \le 2Q_n \mathrm{A}_\beta N^{-\beta} + 6\bar{B} p_R^{1/4} \le 4 Q_n \mathrm{A}_\beta N^{-\beta},$$
$$\||\nabla \varphi_N(t,\cdot)|\|_{L^4(\rho_t)} + \|\partial_t \varphi_N(t,\cdot)\|_{\rho_t} \le 2Q_n \mathrm{A}_\beta N^{-(\beta-1)} + \big(\tfrac{3}{2}G + G_t + C_h\big) p_R^{1/4}$$
$$\le 4 Q_n \mathrm{A}_\beta N^{-(\beta-1)},$$

because $p_R^{1/4} \le n^{-(\beta+2)}$, $N \le n$ and $n \ge 2$ give $6\bar{B} p_R^{1/4} \le 6\bar{B} n^{-2} N^{-\beta} \le 2\mathrm{A}_\beta N^{-\beta} \le 2Q_n \mathrm{A}_\beta N^{-\beta}$ (as $\bar{B} \le \mathrm{A}_\beta$ and $Q_n \ge 1$), and $(\frac{3}{2}G + G_t + C_h) p_R^{1/4} \le (\frac{3}{2}G + G_t + C_h) n^{-3} N^{-(\beta-1)} \le \mathrm{A}_\beta N^{-(\beta-1)}$. For the ∂_t term on the complement we used $\|\partial_t \varphi_N \mathbf{1}\|_{\rho_t} \le \|G_t + |\partial_t h^*|\|_{L^4(\rho_t)} p_R^{1/4}$. Taking the supremum over $t \in I$, and noting that $L^4(\nu \otimes \rho)$ and $L^2(\nu \otimes \rho)$ norms are bounded by the corresponding suprema over t, gives (F.17).

Step 5: the network. All constants $C^\sharp, c^\sharp$ below depend only on (d, β) and may be enlarged from line to line, n_0 is enlarged finitely many times. By Lemma F.7, each Φ_j has at most L_0 hidden layers, at most $C_{\mathrm{F}}(N+1)^d$ nonzero weights (hence width at most $C_{\mathrm{F}}(N+1)^d$) and weights of modulus at most $C_{\mathrm{F}} N^{C_{\mathrm{F}}}$. For a network with at most L_0 hidden layers, widths at most $W' \ge d + 1$, weights of modulus at most $\Lambda' \ge 1$ and inputs in $[-1,1]^{d+1}$, induction over the layers gives hidden activations bounded by $(2W'\Lambda')^{3^{L_0+1}}$, since each layer maps $a := \max(1, \|x^{(\ell)}\|_\infty)$ to at most $(2W'\Lambda' a)^3$. On Z the inputs $y = (x + R\mathbf{1})/2R \in [0,1]^d$ and $\tilde{t} \in [-1,1]$ are affine in z, so they are absorbed in the first-layer parameters.

(a) Coefficient blocks. Realize $\hat{c}_j$ as Φ_j composed with the input map, with output weights multiplied by S_j. Since $S_j \le 4M_\beta(2R)^\beta = \sigma^{-c}\mathrm{polylog}(n)$ and $\sup_{Q_R}|\hat{c}_j| \le 2M_\beta + C_{\mathrm{F}}S_j$, for $n \ge n_0$ we have $S_j \le n$ and $M_c := 1 + \max_j \sup_{Q_R}|\hat{c}_j| \le n$.

(b) Chebyshev block. With $T_0 = 1, T_1 = \tilde{t}, T_{2m} = 2T_mT_m - 1, T_{2m+1} = 2T_mT_{m+1} - T_1$, hidden layer $k \le k_* := \lceil \log_2(K+1) \rceil$ of this block outputs, as affine functions of its units, all $T_j(\tilde{t})$ with $j \le \min\{2^k, K\}$: each new T_j is a product of two values available at layer $k-1$, realized by four units via Lemma F.3(iv) with $M = 2$, and each value already available, including T_1, is carried by four units via Lemma F.3(iii) with $M = 2$. This is exact because $|T_j| \le 1$ on $[-1, 1]$. Each layer has at most $8(K+1)$ units, each with at most 17 nonzero incoming parameters of modulus at most 4, and activations at most 64 on Z.

(c) Alignment and products. Let $L_1 := L - 1 = L_0 + \lceil \log_2(K+1) \rceil$, which exceeds the depth of every block. Pad every block to L_1 hidden layers by carrying its outputs through Lemma F.3(iii) with $M = M_c$ (resp. $M = 2$). Hidden layer L computes the $K+1$ products $\hat{c}_jT_j$ by Lemma F.3(iv) with $M := 2M_c + 2$, and the readout sums them, which gives H_N exactly. The resulting network has depth L, width at most $(K+1)(C_{\mathrm{F}}(N+1)^d + 20) \le C^\sharp K(N+1)^d$, at most $6(K+1)C_{\mathrm{F}}(N+1)^d + 320(K+1)L \le C^\sharp K((N+1)^d + L)$ nonzero parameters (biases created by the input map, and the first carrying layer after each coefficient block, whose units read the whole last layer of Φ_j, account for the factor 6), parameters of modulus at most $n^{c^\sharp}$, and hidden activations on Z at most $n^{c^\sharp}$ by the induction above, for $n \ge n_0$ (using $N \le n$, $M_c \le n$, $2/\ell + 2t_1/\ell + 2 \le n$). Under (F.16) and after padding widths with inactive units, $H_N \in \mathcal{NN}_\varrho(W_N, L, \Lambda, P_0, \mathcal{A})$. □

G Removing the inf-sup assumption

Definition G.1. (Projected residual) Let $\Psi \subset \mathcal{V}$ be a test class carrying a symmetric, positively homogeneous scale $\|\cdot\|_\bullet$ – for Ψ_N that of (E.9) – and let $\tau \ge 0$. Set

$$\|\psi\|_{\mathcal{V},\tau}^2 := \|\psi\|_{\mathcal{V}}^2 + \tau^2\|\psi\|_\bullet^2, \quad \mathcal{N}_{\Psi,\tau}(h) := \sup_{0\neq\psi\in\Psi} \frac{A(h,\psi)}{\|\psi\|_{\mathcal{V},\tau}} \le \mathcal{N}_\Psi(h) \le \|\mathcal{R}[h]\|_{\mathcal{V}^*}, \tag{G.1}$$

where $\mathcal{N}_\Psi := \mathcal{N}_{\Psi,0}$, the largest amount of the residual that the test class can detect at slack τ. The second inequality is immediate from $\Psi \subseteq \mathcal{V}$, and the first from $\|\cdot\|_{\mathcal{V},\tau} \ge \|\cdot\|_{\mathcal{V}}$. Together they make the approximation input of Lemma F.1 available at every τ without any further hypothesis. The slack is $\tau = 0$ at the population level and $\tau = \delta_n$ in the estimator (H.1), where it is what makes the empirical and population adversary norms comparable multiplicatively.

Lemma G.2. *(Fenchel identity for a cone) Let Ψ satisfy (S1) and let $\alpha > 0$. Then*

$$\sup_{\psi\in\Psi}\Big[A(h,\psi) - \frac{\alpha}{2}\|\psi\|_{\mathcal{V}}^2\Big] = \frac{\mathcal{N}_\Psi(h)^2}{2\alpha}. \tag{G.2}$$

The same identity holds with $\|\cdot\|_{\mathcal{V}}$ replaced by any other symmetric positively homogeneous functional – the proof uses nothing else – in particular by $\|\cdot\|_{\mathcal{V},\tau}$, by the empirical norm $\|\cdot\|_{\hat{\mathcal{V}}}$ and by the ridged empirical norm $\|\cdot\|_{\hat{\mathcal{V}},\tau}$ of (H.1)*.*

Proof. Every nonzero $\psi \in \Psi$ factors as $\psi = r\hat{\psi}$ with $r := \|\psi\|_{\mathcal{V}} > 0$ and $\|\hat{\psi}\|_{\mathcal{V}} = 1$. By (S1) both $\hat{\psi}$ and $r'\hat{\psi}$ lie in Ψ for every $r' \ge 0$. Hence the supremum may be computed in two stages, first over the radius and then over the direction:

$$\sup_{\psi\in\Psi}\Big[A(h,\psi) - \frac{\alpha}{2}\|\psi\|^2\Big] = \sup_{\hat{\psi}}\sup_{r\ge0}\Big[rA(h,\hat{\psi}) - \frac{\alpha}{2}r^2\Big] = \sup_{\hat{\psi}}\frac{(A(h,\hat{\psi})_+^2)}{2\alpha}, \tag{G.3}$$

the inner supremum being attained at $r = (A(h,\hat{\psi}))_+/\alpha$, which is admissible precisely because Ψ is a cone. Since Ψ is symmetric, $\sup_{\hat{\psi}}(A(h,\hat{\psi}))_+ = \sup_{\hat{\psi}}|A(h,\hat{\psi})| = \mathcal{N}_\Psi(h)$. □

Lemma G.3. *(Coefficient envelopes) Under Assumptions 1–2,*

$$\Lambda_* := \sup_{t\in I}\big(\||x+s|\|_{L^4(\rho_t)} + \|d + \nabla\cdot s\|_{L^4(\rho_t)} + \||x+s||s|\|_{L^4(\rho_t)} + \|\partial_t\log\rho_t\|_{L^4(\rho_t)}\big) \lesssim d^2\sigma_*^{-4}. \tag{G.4}$$

Proof. The first term is Lemma B.1, together with $\|Z\|_{L^4} \le C\|Z\|_{\psi_2}$. For the second, (B.6) gives the deterministic bound $|\nabla \cdot s| = |\mathrm{tr}\nabla^2 \log \rho_t| \le d\sigma_{T-t}^{-2}(1 + d\mu_{T-t}^2\sigma_{T-t}^{-2})$. For the third (B.1) gives $|s| \le \sigma_{T-t}^{-2}(|x| + \sqrt{d})$, whence $\||s|\|_{\psi_2,\rho_t} \lesssim \sqrt{d}\sigma_{T-t}^{-2}$ exactly as in Lemma B.1, and Cauchy-Schwarz in L^8 bounds the product. For the fourth, Fokker-Planck equation for $dX_\tau = -X_\tau d\tau + \sqrt{2}dB_\tau$ reads $\partial_\tau p_\tau = \nabla \cdot (xp_\tau) + \Delta p_\tau$ and dividing by p_τ and using $\rho_t = p_{T-t}$ (so that $\partial_t = -\partial_\tau$) gives $\partial_t \log \rho_t = -(d + x \cdot s + \nabla \cdot s + |s|^2)$, each summand of which was just bounded in $L^4(\rho_t)$. All bounds are largest at $t = t_1$, where $\sigma_{T-t} = \sigma_*$. □

Theorem G.4. *(Coercivity without an inf-sup constant) Let Assumptions 1–2 hold and let $\mathcal{H}_N, \Psi_N$ satisfy (S1)–(S2) and the following two conditions: every $h \in \mathcal{H}_N$ is $C^{1,2}$ on $I \times \mathbb{R}^d$ with $|h| \le 3\bar{B}$ and $\sup_{I\times\mathbb{R}^d}(|\partial_t h| + |\nabla h| + |\nabla^2 h|) < \infty$, and, for some $C_0 \ge 0$ and the scale $\|\cdot\|_\bullet$ of Definition G.1,*

$$\|\omega(h - h')\|_\bullet \le C_0 \quad \text{for all } h, h' \in \mathcal{H}_N. \tag{G.5}$$

Let $h_N \in \mathcal{H}_N$ satisfy (F.3) *with some $e_N > 0$, and let $\tau \ge 0$ and $\delta \in (0,1)$. Then for every $h \in \mathcal{H}_N$, writing $\varphi := h - h^*$ and $I_\delta := [t_0, t_1 - \delta\ell]$,*

$$\int_{I_\delta} \|\nabla\varphi\|_{\rho_t}^2 \nu(dt) \le \frac{C}{\delta}\left(\frac{1+\ell}{\ell}\mathcal{D}(h) + \mathcal{N}_{\Psi_N,\tau}(h)^2 + C_0^2\tau^2 + (1 + \Lambda_*^2 + \ell^{-2})e_N^2\right), \tag{G.6}$$

with C absolute and $\Lambda_* \lesssim d^2\sigma_*^{-4}$ the coefficient envelope of Lemma G.3. For the classes of Definition E.3 all hypotheses other than (F.3) *hold, with $C_0 = 1$, by Proposition E.4(iii) and* (E.11)*, for $\tau = 0$ the scale condition is not used.*

Proof. By the first hypothesis and Lemma B.3, h satisfies the hypotheses of Theorem D.2, whose exact identity reads

$$2\ell E_\omega + \|\varphi(t_0,\cdot)\|_{\rho_{t_0}}^2 = \mathcal{D}(h) - 2\ell A(h, \omega\varphi), \quad E_\omega = \int_I \omega\|\nabla\varphi\|_{\rho_t}^2\nu(dt). \tag{G.7}$$

Both left-hand terms are favourable, so the whole problem is to bound the cross term $A(h,\omega\varphi)$. The naive bound $|A(h,\omega\varphi)| \le \|\mathcal{R}[h]\|_{\mathcal{V}^*}\|\omega\varphi\|_{\mathcal{V}}$ is what produces the inf-sup constant, since the estimator controls $\mathcal{N}_{\Psi_N}(h)$ and not $\|\mathcal{R}[h]\|_{\mathcal{V}^*}$. We split the test function into a part that lies in Ψ_N and a part that is small.

Set $\varphi_1 := h - h_N$ and $\varphi_2 := h_N - h^*$, so that $\varphi = \varphi_1 + \varphi_2$, and abbreviate $E := E_\omega$, $D := \mathcal{D}(h)$, $\mathcal{N} := \mathcal{N}_{\Psi_N,\tau}(h)$. Since $A(h^*,\cdot) = 0$ and A is linear in its first argument, $A(h,\psi) = A(\varphi,\psi)$ for every admissible ψ. Throughout we use $0 \le \omega \le 1$, hence $\omega^2 \le \omega$, and the consequences $\int_I \|\varphi_2\|_{\rho_t}^2\nu \le e_N^2$, $\int_I \|\nabla\varphi_2\|_{\rho_t}^2\nu \le \||\nabla\varphi_2|\|_{L^4(\nu\otimes\rho)}^2 \le e_N^2$ of (F.3).

(a) The in-class part. Both h and h_N lie in $\mathcal{H}_N$, so $\omega\varphi_1 \in \Psi_N$ by (S2) and $\|\omega\varphi_1\|_\bullet \le C_0$ by (G.5). By the definition (G.1) of $\mathcal{N}$, $|A(h,\omega\varphi_1)| \le \mathcal{N}\|\omega\varphi_1\|_{\mathcal{V},\tau}$. Since ω depends only on t, $\nabla(\omega\varphi_1) = \omega\nabla\varphi_1$, and using $\omega^2 \le \omega$ and $(a+b)^2 \le 2a^2 + 2b^2$ inside each integral,

$$\|\omega\varphi_1\|_{\mathcal{V},\tau}^2 \le \int_I \omega(\|\varphi_1\|_{\rho_t}^2 + \|\nabla\varphi_1\|_{\rho_t}^2)\nu(dt) + C_0^2\tau^2 \le 2D + 2E + 4e_N^2 + C_0^2\tau^2.$$

The E produced here is the weighted energy on the left of (G.7). Young's inequality $2\ell\mathcal{N}\kappa \le \frac{\ell}{4}\kappa^2 + 4\ell\mathcal{N}^2$ with $\kappa := \|\omega\varphi_1\|_{\mathcal{V},\tau}$ gives

$$2\ell|A(h,\omega\varphi_1)| \le \frac{\ell}{2}D + \frac{\ell}{2}E + \ell e_N^2 + \frac{\ell}{4}C_0^2\tau^2 + 4\ell\mathcal{N}^2.$$

(b) The out-of-class part, gradient term. By (C.5) the cross term $A(\varphi,\omega\varphi_2)$ has three pieces. The last is $-\int_I \omega \int \nabla\varphi \cdot \nabla\varphi_2\rho_t\,dx\,\nu(dt)$, which by Cauchy–Schwarz against the weight ω and $ab \le \frac{1}{8}a^2 + 2b^2$ is at most $\frac{1}{8}E + 2\int_I \omega\|\nabla\varphi_2\|_{\rho_t}^2\nu \le \frac{1}{8}E + 2e_N^2$.

(c) The out-of-class part, drift term. The piece $\int_I \omega \int((x+s)\cdot\nabla\varphi)\varphi_2\rho_t\,dx\,\nu(dt)$ cannot be treated by Cauchy–Schwarz directly: that would produce $\||\nabla\varphi|\|_{L^4}$, which the estimator does not control. For $V := x + s$ and g bounded, C^1 and with bounded gradient,

$$\int (V\cdot\nabla\varphi)g\rho_t = -\int \varphi\,\nabla\cdot(gV\rho_t) = -\int \varphi[(\nabla\cdot V)g + V\cdot\nabla g + (V\cdot s)g]\rho_t,$$

where the boundary term at infinity vanishes because φ and g are bounded, $|V|$ has at most linear growth and ρ_t has Gaussian tails, and $\nabla\rho_t = s\rho_t$. Take $g := \omega\varphi_2$ and note $\nabla\cdot V = d + \nabla\cdot s$. Cauchy–Schwarz in x, Hölder with the envelopes of Lemma G.3, and Cauchy–Schwarz in t give

$$\left|\int_I \omega \int (V\cdot\nabla\varphi)\varphi_2\rho_t\,dx\,\nu(dt)\right| \le \Lambda_*\sqrt{D}\left(2\|\varphi_2\|_{L^4(\nu\otimes\rho)} + \||\nabla\varphi_2|\|_{L^4(\nu\otimes\rho)}\right)$$
$$\le 3\Lambda_*\sqrt{D}e_N \le \frac{1}{8}D + 18\Lambda_*^2 e_N^2.$$

(d) The out-of-class part, time term. With $F(t) := \int \varphi\,\omega\varphi_2\,\rho_t\,dx$, the function F is absolutely continuous on I with $F' = \int \partial_t(\varphi\omega\varphi_2\rho_t)\,dx$, by dominated convergence: φ, φ_2 are bounded, $|\partial_t\varphi|, |\partial_t\varphi_2|$ grow at most linearly in x by the first hypothesis and harmonicity, and $\partial_t\rho_t = \rho_t\,\partial_t\log\rho_t$ with $\partial_t\log\rho_t$ of polynomial growth. Hence, with $\nu(dt) = dt/\ell$,

$$\int_I \omega\int \partial_t\varphi\,\varphi_2\rho_t\,dx\,\nu(dt) = \frac{F(t_1)-F(t_0)}{\ell} - \int_I\int \varphi\,\partial_t(\omega\varphi_2\rho_t)\,dx\,\nu(dt).$$

The boundary term at t_1 vanishes because $\omega(t_1) = 0$. At t_0, $\ell^{-1}|F(t_0)| \le \frac{1}{4\ell}\|\varphi(t_0)\|^2_{\rho_{t_0}} + \frac{1}{\ell}\|\varphi_2(t_0)\|^2_{\rho_{t_0}} \le \frac{1}{4\ell}\|\varphi(t_0)\|^2_{\rho_{t_0}} + \frac{1}{\ell}e_N^2$. For the bulk term, $\partial_t(\omega\varphi_2\rho_t) = \rho_t[\dot\omega\varphi_2 + \omega\partial_t\varphi_2 + \omega\varphi_2\partial_t\log\rho_t]$ with $|\dot\omega| = \ell^{-1}$, so by Cauchy–Schwarz, Hölder and Lemma G.3 it is at most $\sqrt{D}(\ell^{-1}\|\varphi_2\|_{L^2(\nu\otimes\rho)} + \|\partial_t\varphi_2\|_{L^2(\nu\otimes\rho)} + \Lambda_*\|\varphi_2\|_{L^4(\nu\otimes\rho)}) \le (1+\ell^{-1}+\Lambda_*)\sqrt{D}e_N \le \frac{1}{8}D + 2(1+\ell^{-1}+\Lambda_*)^2e_N^2$.

(e) Assembly. Multiplying the bounds of (b), (c), (d) by 2ℓ and inserting them with (a) into (G.7): the terms in E total $\frac{\ell}{2}E + \frac{\ell}{4}E = \frac{3\ell}{4}E$, absorbed by $2\ell E$, the boundary term of (d) contributes $\frac{1}{2}\|\varphi(t_0)\|^2_{\rho_{t_0}}$, absorbed by the favourable term, the terms in D total $(1+\ell)D$, the projected residual contributes $4\ell\mathcal{N}^2$ and the slack $\frac{\ell}{4}C_0^2\tau^2$. The terms in e_N^2 are $\ell e_N^2 + 4\ell e_N^2 + 36\ell\Lambda_*^2 e_N^2 + 2e_N^2 + 4\ell(1+\ell^{-1}+\Lambda_*)^2 e_N^2 \le C\ell(1+\Lambda_*^2+\ell^{-2})e_N^2$, using $2 \le \ell(1+\ell^{-2})$ and $(1+\ell^{-1}+\Lambda_*)^2 \le 3(1+\ell^{-2}+\Lambda_*^2)$. Hence

$$\frac{5\ell}{4}E \le (1+\ell)D + 4\ell\mathcal{N}^2 + \frac{\ell}{4}C_0^2\tau^2 + C\ell(1+\Lambda_*^2+\ell^{-2})e_N^2,$$

and dividing by ℓ gives $E \le C(\frac{1+\ell}{\ell}D + \mathcal{N}^2 + C_0^2\tau^2 + (1+\Lambda_*^2+\ell^{-2})e_N^2)$. Finally $\omega(t) = (t_1-t)/\ell \ge \delta$ on I_δ, so $\int_{I_\delta}\|\nabla\varphi\|^2_{\rho_t}\nu(dt) \le \delta^{-1}\int_{I_\delta}\omega\|\nabla\varphi\|^2_{\rho_t}\nu(dt) \le \delta^{-1}E$, which is (G.6). The bound $\Lambda_* \lesssim d^2\sigma_*^{-4}$ is Lemma G.3. □

H Statistical guarantee

Let $\{(X_0^i, \varepsilon^i)\}_{i=1}^n$ be i.i.d. with $X_0^i \sim p_0$, $\varepsilon^i \sim \mathcal{N}(0, I_d)$ independent, and $X_{T-t}^i := \mu_{T-t}X_0^i + \sigma_{T-t}\varepsilon^i \sim \rho_t$. With $\hat a_t, \hat{\mathcal{D}}$ the empirical versions of a_t, $\mathcal{D}$ and $\|\cdot\|_{\hat{\mathcal{V}}}$ the empirical $\mathcal{V}$-norm, the estimator is

$$\hat h \in \operatorname*{argmin}_{h\in\mathcal{H}_N} \hat{\mathcal{D}}(h) + \mu \sup_{\psi\in\Psi_N}\left[\int_I \hat a_t(h,\psi)\nu(dt) - \frac{\alpha}{2}\|\psi\|^2_{\hat{\mathcal{V}},\delta_n}\right], \quad \|\psi\|^2_{\hat{\mathcal{V}},\tau} := \|\psi\|^2_{\hat{\mathcal{V}}} + \tau^2\|\psi\|^2_\bullet, \tag{H.1}$$

with $\hat g = \Pi_{G^*}(\nabla\hat h/(\hat h \vee \mathrm{b}))$ as in (D.20) and δ_n as in (H.15).

The second term in the adversary norm is a ridge at the statistical scale. $\|\psi\|_\bullet/\|\psi\|_{\mathcal{V}}$ is unbounded on the cone (Remark E.6) and no bound of the form $\delta_n\|\psi\|_{\mathcal{V}} + \delta_n^2$ holds uniformly on Ψ_N. The naive estimator – ridged in $\|\cdot\|_{\hat{\mathcal{V}}}$ alone – has an inner supremum that the deviation bounds do not control in directions where the empirical $\mathcal{V}$-norm nearly degenerates. As we will see at Theorem H.7, adding $\delta_n^2\|\cdot\|_\bullet^2$ removes the difficulty. The cost is one extra term of order δ_n^2 in Theorem G.4, taken there at $\tau = \delta_n$.

$$G_{h,\psi}(z) := \int_I \left[\partial_t h\,\psi + ((x+s)\cdot\nabla h)\psi - \nabla h\cdot\nabla\psi\right](t, X_{T-t}(z))\nu(dt), \tag{H.2}$$

$$D_h(z) := \int_I \left[(h - w(x_0))^2 - (h^* - w(x_0))^2\right](t, X_{T-t}(z))\nu(dt), \tag{H.3}$$

$$Q_\psi(z) := \int_I \left[\psi^2 + |\nabla\psi|^2\right](t, X_{T-t}(z))\nu(dt), \tag{H.4}$$

for $h \in \mathcal{H}_N$ and $\psi \in \Psi_N$.

All three quantities are ordinary empirical processes over a single i.i.d. sample and the integration in t having been absorbed into the summand. Indeed, put $Z_i := (X_0^i, \varepsilon^i) \in \mathcal{Z} := \mathrm{supp}(p_0) \times \mathbb{R}^d$, which are i.i.d., and write $X_{T-t}(z) := \mu_{T-t}x_0 + \sigma_{T-t}e$ for $z = (z_0, e)$.

Writing P_n for the empirical and P for population measure of Z

$$\int_I \hat{a}_t(h,\psi)\nu(dt) = P_n G_{h,\psi},\ \hat{\mathcal{D}}(h) - \hat{\mathcal{D}}(h^*) = P_n D_h,\ \|\psi\|^2_{\hat{\mathcal{V}}} = P_n Q_\psi, \tag{H.5}$$

with population counterparts $PG_{h,\psi} = A(h,\psi),\ PD_h = \mathcal{D}(h)$ and $PQ_\psi = \|\psi\|^2_{\mathcal{V}}$. Only $\mathcal{G} := \{G_{h,\psi}\}$ has an unbounded envelope, and only through the factor $x + s$.

We will also use that $\mathcal{G}$ is linear in ψ and $\mathcal{Q}$ quadratic, while Ψ_N is a cone whose scale is measured by $\|\cdot\|_\bullet$. Exact homogeneity therefore removes the radial variable from the empirical process problem: it suffices to work at generator scale $\|\psi\|_\bullet \le 1$ and multiply back. No restriction of the adversary to a norm ball is imposed: with the ridge of (H.1) the bounds below hold on all of Ψ_N.

Lemma H.1. *(Envelopes and variances) Let $\mathcal{T} > 0$ and, for $z = (x_0, e)$, write $\bar{V}(z) := \sigma_*^{-2}(2\sqrt{d} + |e|)$. Under Assumptions 1–2 and Definition E.3:*

(i) (Uniform drift envelope) $\sup_{t\in I}|(x+s)(t, X_{T-t}(z))| \le \bar{V}(z)$ for every z, and $\|\bar{V}\|_{\psi_2} \le C\sqrt{d}\sigma_*^{-2}$.

(ii) (Envelope) $|G_{h,\psi}| \le \Gamma\|\psi\|_\bullet(\max(G_t, G) + G\bar{V})$ with $\|\psi\|_\bullet$ the cone scale (E.9), and for the envelope $F := \Gamma(\max(G_t, G) + G\bar{V})$ of $\{G_{h,\psi} : \|\psi\|_\bullet \le 1\}$,

$$\|F\|_{\psi_2} \le C\sigma^{-c}\log n, \quad \text{hence } \|\max_{i\le n} F(Z_i)\|_{\psi_1} \le C\log n\|F\|_{\psi_2} \le C\sigma^{-c}\log^2 n. \tag{H.6}$$

(iii) (Bounded surrogates) Let $G^{\mathcal{T}}_{h,\psi}$ be defined as in (H.2) with $|x+s|$ replaced by $(x+s)\mathbf{1}\{|x+s| \le \mathcal{T}\}$. Then $\|G^{\mathcal{T}}_{h,\psi}\|_\infty \le \Gamma(G_t + G + G\mathcal{T})\|\psi\|_\bullet$ and

$$P(G^{\mathcal{T}}_{h,\psi})^2 \le 3(G_t^2 + G^2\mathcal{T}^2 + G^2)\|\psi\|^2_{\mathcal{V}}, \tag{H.7}$$

moreover $G_{h,\psi} = G^{\mathcal{T}}_{h,\psi}$ for every (h,ψ) simultaneously on the event $\{\max_{i\le n}\bar{V}(Z_i) \le \mathcal{T}\}$, whose complement has probability at most $2ne^{-c\mathcal{T}^2\sigma_*^4/d}$, and, if $\mathcal{T} \ge C\sqrt{d}\sigma_*^{-2}$, $P|G_{h,\psi} - G^{\mathcal{T}}_{h,\psi}| \le \Gamma G\|\psi\|_\bullet P[\bar{V}\mathbf{1}\{\bar{V} > \mathcal{T}\}] \le C\mathcal{T}\Gamma G\|\psi\|_\bullet e^{-c\mathcal{T}^2\sigma_*^4/d}$.

(iv) (Bernstein condition for $\mathcal{D}$) $\|D_h\|_\infty \le 12\bar{B}^2$ and $PD_h^2 \le 16\bar{B}^2 PD_h$,

(v) (Bernstein condition for $\mathcal{Q}$) $0 \le Q_\psi \le \Gamma^2\|\psi\|^2_\bullet$ and $PQ_\psi^2 \le \Gamma^2\|\psi\|^2_\bullet PQ_\psi$.

Proof. (i) By Tweedie's formula, $x + s(t,x) = (1 - \sigma^{-2}_{T-t})x + \mu_{T-t}\sigma^{-2}_{T-t}\mathbb{E}[X_0|x]$, so $|x + s| \le \sigma^{-2}_{T-t}(|x| + \sqrt{d})$ using $|1 - \sigma^{-2}| \le \sigma^{-2}$ for $\sigma \le 1, \mu \le 1$ and $|X_0| \le \sqrt{d}$. At $x = X_{T-t}(z) = \mu_{T-t}x_0 + \sigma_{T-t}e$ we have $|x| \le \sqrt{d} + |e|$, so the supremum over $t \in I$ is at most $\bar{V}(z)$. Since $\||e|\|_{\psi_2} \le C\sqrt{d}$, also $\|\bar{V}\|_{\psi_2} \le C\sqrt{d}\sigma_*^{-2}$.

(ii) The three terms of (H.2) are bounded pointwise by $G_t|\psi|$, $|x + s|G|\psi|$ and $G|\nabla\psi|$, using (E.6), whose suprema are over $I \times \mathbb{R}^d$ and so are valid at the sample points $(t, X_{T-t}(z))$ – this is what the clip of Definition E.3 is for. Grouping the first and third through $G_t|\psi| + G|\nabla\psi| \le \max(G_t, G)(|\psi| + |\nabla\psi|) \le \max(G_t, G)\Gamma\|\psi\|_\bullet$, bounding the middle by $\bar{V}(z)G\Gamma\|\psi\|_\bullet$ through (i), and integrating the probability measure ν gives the envelope. For (H.6), F is affine in $\bar{V}$, so $\|F\|_{\psi_2} \le C\Gamma(\max(G_t, G) + G\sqrt{d}\sigma_*^{-2})$. Since $\Gamma \asymp 2G_t$ and $G_t \asymp \sigma^{-c}\sqrt{\log n}$ through $R \asymp \sqrt{\log n}$, this is $C\sigma^{-c}\log n$, and the standard $\|\max_{i\le n}\xi_i\|_{\psi_1} \le C\log n \max_i\|\xi_i\|_{\psi_1}$ with $\|\cdot\|_{\psi_1} \le C\|\cdot\|_{\psi_2}$ gives $C\sigma^{-c}\log^2 n$. This term is tight against δ_n^2 rather than of lower order.

(iii) The sup bound is immediate. For (H.7), Cauchy-Schwarz in ν gives $(G^{\mathcal{T}}_{h,\psi})^2 \le \int_I (\dots)^2 \nu(dt)$, and $(\partial_t h\psi + (\cdot)\psi - \nabla h\cdot\nabla\psi)^2 \le 3(G_t^2 + G^2\mathcal{T}^2)\psi^2 + 3G^2|\nabla\psi|^2$ pointwise. Taking P and using $P\int_I \psi^2\nu = \|\psi\|^2_{L^2(\nu\otimes\rho)}$ and likewise for $\nabla\psi$ gives (H.7), since both are at most $\|\psi\|^2_{\mathbb{V}}$. On the event $\{\max_i \bar V(Z_i) \le \mathcal{T}\}$ part (i) gives $|x+s|(t, X_{T-t}(Z_i)) \le \mathcal{T}$ for every t and every i, so the truncation is inactive on the sample. Its complement has probability at most $2ne^{-c\mathcal{T}^2\sigma_*^4/d}$ by a union bound over $i \le n$ alone. Using the averaged $\int_I |x+s|\nu$ here would not help, since that quantity being below $\mathcal{T}$ does not force $|x+s| \le \mathcal{T}$ at every t, and the event would involve an uncountable index set. The population remainder uses $P[\bar V \mathbf{1}\{\bar V > \mathcal{T}\}] \le C\mathcal{T}\,\|\bar V\|_{\psi_2} e^{-c\mathcal{T}^2/\|\bar V\|^2_{\psi_2}}$ for $\mathcal{T} \ge 2\|\bar V\|_{\psi_2}$.

(iv) Pointwise in t, $(h-w)^2 - (h^*-w)^2 = \varphi(h+h^*-2w)$ with $\varphi := h - h^*$. Since $h \in [-2\bar B, 3\bar B]$, $h^* \in (0,\bar B]$ and $w \in (0,\bar B]$, the extreme values are $|\varphi| \le 3\bar B$ – the constraint $h^* > 0$ cutting one side – and $|h+h^*-2w| \le 4\bar B$, so $\|D_h\|_\infty \le 12\bar B^2$. By Cauchy-Schwarz in ν, $D_h^2 \le \int_I \varphi^2\nu \cdot \int_I (h+h^*-2w)^2\nu \le 16\bar B^2 \int_I \varphi^2\nu$, and taking P gives $PD_h^2 \le 16\bar B^2 \mathcal{D}(h) = 16\bar B^2 PD_h$.

(v) $Q_\psi \ge 0$, and $\psi^2 + |\nabla\psi|^2 \le (|\psi| + |\nabla\psi|)^2 \le \Gamma^2\|\psi\|^2_\bullet$ by (E.7), with no factor 2. For a nonnegative bounded function, $PQ^2 \le \|Q\|_\infty PQ$. □

Lemma H.2. *(Entropy of the index classes) Let $\mathcal{T} > 0$ and let $\mathcal{G}_1^{\mathcal{T}} := \{G^{\mathcal{T}}_{h,\psi} : h \in \mathcal{H}_N, \psi \in \Psi_N^\bullet\}$, $\mathcal{D}_\bullet := \{D_h : h \in \mathcal{H}_N\}$ and $\mathcal{Q}_1 := \{Q_\psi : \psi \in \Psi_N^\bullet\}$, where $\Psi_N^\bullet = \{\psi \in \Psi_N : \|\psi\|_\bullet \le 1\}$ as in* (E.13)*. Put*

$$\mathfrak{b} := \Gamma(\Gamma + G\mathcal{T}) + 12\bar B^2, \qquad \mathfrak{l} := (1+\Gamma)^2(1+\mathcal{T})(1+G_t+G) + 12\bar B^2. \tag{H.8}$$

Then every member of $\mathcal{G}_1^{\mathcal{T}} \cup \mathcal{D}_\bullet \cup \mathcal{Q}_1$ is a Borel function on $\mathcal{Z}$ bounded in modulus by $\mathfrak{b}$, one has $1 \vee \mathfrak{b} \le \mathfrak{l}$, and for every probability measure Q on $\mathcal{Z}$ and every $0 < \varepsilon \le \mathfrak{l}$

$$\log N(\varepsilon, \mathcal{G}_1^{\mathcal{T}} \cup \mathcal{D}_\bullet \cup \mathcal{Q}_1,\, L^2(Q)) \le 6\,\mathbb{V}_N \log\frac{e\mathfrak{l}}{\varepsilon}, \tag{H.9}$$

with $\mathbb{V}_N$ as in (E.14)*, bounded in Lemma E.8(iii). For $\varepsilon \ge \mathfrak{b}$ the covering number equals* 1*. Each of the three classes, and each of their subclasses, is separable in the supremum norm on $\mathcal{Z}$.*

Proof. If $\mathbb{V}_N = \infty$ there is nothing to prove, so let $\mathbb{V}_N < \infty$.

Envelopes. For $h \in \mathcal{H}_N$ and $\psi \in \Psi_N^\bullet$ we have, globally on $I \times \mathbb{R}^d$, $h \in [-2\bar B, 3\bar B]$, $|\partial_t h| \le G_t$ and $|\nabla h| \le G$ by (E.6), and $|\psi| + |\nabla\psi| \le \Gamma\|\psi\|_\bullet \le \Gamma$. Write $v^{\mathcal{T}} := (x+s)\mathbf{1}\{|x+s| \le \mathcal{T}\}$, so that $|v^{\mathcal{T}}| \le \mathcal{T}$. The integrands of $G^{\mathcal{T}}_{h,\psi}, D_h, Q_\psi$ are bounded Borel functions of (t,z), so by Fubini the three functions are bounded Borel functions on $\mathcal{Z}$, and their $L^2(Q)$-distances are at most their supremum distances. Pointwise, $|\partial_t h\,\psi + (v^{\mathcal{T}}\cdot\nabla h)\psi - \nabla h\cdot\nabla\psi| \le (\max(G_t, G) + G\mathcal{T})(|\psi| + |\nabla\psi|) \le \Gamma(\Gamma + G\mathcal{T})$, using $\Gamma \ge 2G + 2G_t$, while $|D_h| \le 12\bar B^2$ and $0 \le Q_\psi \le \Gamma^2$ by Lemma H.1. Each bound is at most $\mathfrak{b}$. Moreover $(1+\Gamma)^2(1+\mathcal{T})(1+G_t+G) \ge (1+\Gamma)^2(1+G\mathcal{T}) \ge \Gamma^2 + \Gamma G\mathcal{T}$, so $\mathfrak{b} \le \mathfrak{l}$, and clearly $\mathfrak{l} \ge 1$. The single function 0 covers the union at every radius $\varepsilon \ge \mathfrak{b}$.

Lipschitz bounds in the jets. Let also $h' \in \mathcal{H}_N$ and $\psi' \in \Psi_N^\bullet$, and put, pointwise on $I \times \mathbb{R}^d$, $\Delta_h := |(h, \partial_t h, \nabla h) - (h', \partial_t h', \nabla h')|$ and $\Delta_\psi := |(\psi, \nabla\psi) - (\psi', \nabla\psi')|$. Adding and subtracting, and using the bounds above for h and for ψ',

$$\begin{aligned}
|\partial_t h\,\psi - \partial_t h'\,\psi'| &\le |\psi'|\,|\partial_t h - \partial_t h'| + |\partial_t h|\,|\psi - \psi'| \le \Gamma\Delta_h + G_t\Delta_\psi,\\
|(v^{\mathcal{T}}\cdot\nabla h)\psi - (v^{\mathcal{T}}\cdot\nabla h')\psi'| &\le \mathcal{T}|\psi'|\,|\nabla h - \nabla h'| + \mathcal{T}|\nabla h|\,|\psi - \psi'| \le \Gamma\mathcal{T}\Delta_h + G\mathcal{T}\Delta_\psi,\\
|\nabla h\cdot\nabla\psi - \nabla h'\cdot\nabla\psi'| &\le |\nabla\psi'|\,|\nabla h - \nabla h'| + |\nabla h|\,|\nabla\psi - \nabla\psi'| \le \Gamma\Delta_h + G\Delta_\psi.
\end{aligned}$$

Also $|(h-w)^2 - (h'-w)^2| = |h-h'|\,|(h-w) + (h'-w)| \le 6\bar B\Delta_h$, since $w \in (0,\bar B]$ gives $|h-w|, |h'-w| \le 3\bar B$, and $|\psi^2 + |\nabla\psi|^2 - \psi'^2 - |\nabla\psi'|^2| \le 2\Gamma(|\psi-\psi'| + |\nabla\psi - \nabla\psi'|) \le 2\sqrt{2}\,\Gamma\Delta_\psi$. Integrating against the probability measure ν at $x = X_{T-t}(z)$ and taking the supremum over $z \in \mathcal{Z}$,

$$\begin{aligned}
&\|G^{\mathcal{T}}_{h,\psi} - G^{\mathcal{T}}_{h',\psi'}\|_\infty \le \mathrm{L}_{\mathcal{G}}\max(\|\Delta_h\|_\infty, \|\Delta_\psi\|_\infty),\\
&\|D_h - D_{h'}\|_\infty \le 6\bar B\|\Delta_h\|_\infty, \qquad \|Q_\psi - Q_{\psi'}\|_\infty \le 3\Gamma\|\Delta_\psi\|_\infty,
\end{aligned} \tag{H.10}$$

with $\mathrm{L}_{\mathcal{G}} := \Gamma(\mathcal{T}+2) + G_t + G(\mathcal{T}+1)$. All three constants are at most $\mathfrak{l}$: indeed $6\bar B \le \Gamma \le 3\Gamma \le (1+\Gamma)^2$, and $\mathrm{L}_{\mathcal{G}} \le (1+\mathcal{T})(2\Gamma + G_t + G) \le (1+\mathcal{T})(1+2\Gamma)(1+G_t+G) \le (1+\mathcal{T})(1+\Gamma)^2(1+G_t+G)$.

Separability. By (E.29) and Lemma E.8(i), the map from admissible parameters to $J(f^{\chi})$ is continuous into the supremum norm, for $\mathcal{H}_N$ and for the internal nets of Ψ_N, hence $J\mathcal{H}_N$ is separable, and $J\Psi_N^{\bullet}$, contained in the continuous image of $[-2,2]\times\{\text{parameters}\}$ under $(c,\vartheta)\mapsto c\,\omega J(\phi_{\vartheta}^{\chi})$, is separable. By (H.10) the three classes are Lipschitz images of $J\mathcal{H}_N\times J\Psi_N^{\bullet}$, $J\mathcal{H}_N$ and $J\Psi_N^{\bullet}$, hence separable, and subsets of separable metric spaces are separable.

Nets. Fix $0<\varepsilon\le \mathfrak{l}$ and put $\eta:=\varepsilon/(2\mathfrak{l})\le 1/2$. Since $\eta/2\le 1$, (E.14) gives $\log N(\eta/2, J\mathcal{H}_N, \|\cdot\|_\infty)\le \mathbb{V}_N\log(4e\mathfrak{l}/\varepsilon)$, and the same for $J\Psi_N^{\bullet}$. By (E2) there are $\mathcal{S}_{\mathcal{H}}\subseteq\mathcal{H}_N$ and $\mathcal{S}_{\Psi}\subseteq\Psi_N^{\bullet}$, each of cardinality at most $\exp(\mathbb{V}_N\log(4e\mathfrak{l}/\varepsilon))$, whose jets form η-nets of $J\mathcal{H}_N$ and $J\Psi_N^{\bullet}$ respectively. Given $h\in\mathcal{H}_N$ and $\psi\in\Psi_N^{\bullet}$, choose $h'\in\mathcal{S}_{\mathcal{H}}$ and $\psi'\in\mathcal{S}_{\Psi}$ with $\|\Delta_h\|_\infty,\|\Delta_\psi\|_\infty\le\eta$. By (H.10), $G^{\mathcal{T}}_{h',\psi'}, D_{h'}, Q_{\psi'}$ lie within $\mathfrak{l}\eta=\varepsilon/2$ of $G^{\mathcal{T}}_{h,\psi}, D_h, Q_\psi$ in the supremum norm on $\mathcal{Z}$, hence in $L^2(Q)$. The union class therefore has an ε-net with at most $|\mathcal{S}_{\mathcal{H}}|\,|\mathcal{S}_{\Psi}|+|\mathcal{S}_{\mathcal{H}}|+|\mathcal{S}_{\Psi}|\le 3|\mathcal{S}_{\mathcal{H}}|\,|\mathcal{S}_{\Psi}|$ elements, and

$$\log N(\varepsilon,\mathcal{G}_1^{\mathcal{T}}\cup\mathcal{D}_{\bullet}\cup\mathcal{Q}_1,\,L^2(Q))\le\log 3+2\,\mathbb{V}_N\log\frac{4e\mathfrak{l}}{\varepsilon}.$$

Finally, $\log(e\mathfrak{l}/\varepsilon)\ge 1$ and $\mathbb{V}_N\ge 1$ give $\log 3\le(\log 3)\,\mathbb{V}_N\log(e\mathfrak{l}/\varepsilon)$ and $\log(4e\mathfrak{l}/\varepsilon)\le(1+\log 4)\log(e\mathfrak{l}/\varepsilon)$, and $\log 3+2(1+\log 4)<6$ yields (H.9). □

We restate the external results in the form used below. Throughout, $\varpi_1,\dots,\varpi_n$ are i.i.d. Rademacher signs independent of the sample, and for a class $\mathcal{G}$ of functions on $\mathcal{Z}$ we write $\mathfrak{R}_n(\mathcal{G}):=\sup_{f\in\mathcal{G}}|n^{-1}\sum_{i=1}^n\varpi_i f(Z_i)|$. A class is called separable if it is a separable subset of the bounded functions on $\mathcal{Z}$ with the supremum norm. Suprema of empirical and Rademacher processes over a separable class coincide with suprema over a countable dense subclass, so they are measurable and the results below, stated for countable classes, apply.

Theorem H.3. *(Localization, Theorem 3.3 of (Bartlett et al., 2005) with $K=2$) Let $\mathcal{F}$ be a separable class of functions on $\mathcal{Z}$, star-shaped around 0, with $\|f\|_\infty\le b$ and $Pf^2\le\beta_0 Pf$ for all $f\in\mathcal{F}$. Let $\mathfrak{s}$ be a sub-root function with fixed point r^* such that $\mathfrak{s}(r)\ge\beta_0\mathbb{E}\mathfrak{R}_n\{f\in\mathcal{F}:Pf^2\le r\}$ for all $r\ge r^*$. Then for every $u>0$, with probability at least $1-2e^{-u}$, simultaneously for all $f\in\mathcal{F}$*

$$\frac12 Pf-c_1\left(\frac{r^*}{\beta_0}+\frac{(b+\beta_0)u}{n}\right)\le P_nf\le 2Pf+c_1\left(\frac{r^*}{\beta_0}+\frac{(b+\beta_0)u}{n}\right),\tag{H.11}$$

with c_1 absolute. (The Rademacher average of (Bartlett et al., 2005) carries no absolute value and is dominated by $\mathfrak{R}_n$.)

Theorem H.4. *(Unbounded suprema, Theorem 4 of (Adamczak, 2008) with $\eta=\delta=\alpha=1$) Let $\mathcal{F}$ be a separable class of functions on $\mathcal{Z}$ with a measurable envelope $F\ge\sup_{f\in\mathcal{F}}|f|$ such that $\|\max_{i\le n}F(Z_i)\|_{\psi_1}<\infty$, and let $\varsigma_{\mathcal{F}}^2:=\sup_{f\in\mathcal{F}}\mathrm{Var}(f)$. Then for every $u\ge1$, with probability at least $1-e^{-u}$*

$$\sup_{f\in\mathcal{F}}|P_nf-Pf|\le 2\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf-Pf|+c_2\varsigma_{\mathcal{F}}\sqrt{\frac{u}{n}}+c_2\frac{u\,\|\max_{i\le n}F(Z_i)\|_{\psi_1}}{n},\tag{H.12}$$

with c_2 absolute.

Proof. Theorem 4 of (Adamczak, 2008) applies to the centred class $\{f-Pf\}$, whose summed variance is $n\varsigma_{\mathcal{F}}^2$ and whose envelope satisfies $\|\max_i\sup_f|f(Z_i)-Pf|\|_{\psi_1}\le(1+1/\log 2)\|\max_i F(Z_i)\|_{\psi_1}=:\mathfrak{M}$, since $\sup_f|Pf|\le PF\le\|F(Z_1)\|_{\psi_1}\le\|\max_i F(Z_i)\|_{\psi_1}$ and a constant $a\ge0$ has ψ_1-norm $a/\log 2$. With $Y:=n\sup_f|P_nf-Pf|$ it gives $\mathbb{P}(Y\ge2\mathbb{E}Y+t)\le\exp(-t^2/(4n\varsigma_{\mathcal{F}}^2))+3\exp(-t/(C\mathfrak{M}))$ with C absolute. Choosing $t:=2\sqrt{n\varsigma_{\mathcal{F}}^2u'}+C\mathfrak{M}u'$ with $u':=u+\log 4\le(1+\log 4)u$ makes the right-hand side at most $4e^{-u'}=e^{-u}$, dividing by n gives the claim. □

Lemma H.5. *(Closure) Let $\mathcal{F}$ be a separable class with $\sup_{f\in\mathcal{F}}\|f\|_\infty\le b$, and suppose that for some $V\ge1$ and $\mathfrak{l}\ge b$*

$$\log N(\varepsilon,\mathcal{F},L^2(Q))\le V\log(e\mathfrak{l}/\varepsilon)\quad\text{for every probability measure } Q \text{ on } \mathcal{Z} \text{ and every } 0<\varepsilon\le\mathfrak{l}.\tag{H.13}$$

Then every subclass of $\mathcal{F}$ is separable and satisfies (H.13)*, and the star hull* $\operatorname{star}\mathcal{F}:=\{\lambda f:\lambda\in[0,1],f\in\mathcal{F}\}$ *is separable, star-shaped around 0, bounded by b, and satisfies* (H.13) *with $4V$ in place of V.*

Proof. Subsets of separable metric spaces are separable, and covering numbers, with centres anywhere, do not increase under inclusion. The star hull is the image of $[0,1] \times \mathcal{F}$ under the continuous map $(\lambda, f) \mapsto \lambda f$, hence separable. Fix Q and $0 < \varepsilon \le \mathfrak{l}$, an $(\varepsilon/2)$-net $\{g_k\}$ of $\mathcal{F}$ in $L^2(Q)$, and the points $\lambda_m := \min\{1, m\varepsilon/b\}$, $0 \le m \le \lceil b/\varepsilon \rceil$, so that every $\lambda \in [0,1]$ is within $\varepsilon/(2b)$ of some λ_m. Then $\|\lambda f - \lambda_m g_k\|_{L^2(Q)} \le |\lambda - \lambda_m|\, b + \lambda_m \|f - g_k\|_{L^2(Q)} \le \varepsilon$, so $N(\varepsilon, \operatorname{star} \mathcal{F}, L^2(Q)) \le (2 + b/\varepsilon) N(\varepsilon/2, \mathcal{F}, L^2(Q))$. Since $2 + b/\varepsilon \le 3\mathfrak{l}/\varepsilon$ and $\log(e\mathfrak{l}/\varepsilon) \ge 1$, we get $\log(3\mathfrak{l}/\varepsilon) \le 2\log(e\mathfrak{l}/\varepsilon)$ and $V \log(2e\mathfrak{l}/\varepsilon) \le V(1 + \log 2)\log(e\mathfrak{l}/\varepsilon)$, and $2 + 1.7V \le 4V$. □

Lemma H.6. *(Critical radius from entropy) Let $\mathcal{F}$ be a separable class with $0 \in \mathcal{F}$, $\sup_{f \in \mathcal{F}} \|f\|_\infty \le b$ for some $b > 0$, satisfying* (H.13) *for some $V \ge 1$ and $\mathfrak{l} \ge b$, and put $\mathrm{L} := \log(2e\mathfrak{l}n/b)$. Then for every $r > 0$*

$$\mathbb{E}\mathfrak{R}_n\{f \in \mathcal{F} : Pf^2 \le r\} \le C_1 \left(\sqrt{\frac{rV\mathrm{L}}{n}} + \frac{bV\mathrm{L}}{n} \right), \tag{H.14}$$

with C_1 absolute. Consequently, if $\beta_0 \ge b$, then $\mathfrak{s}(r) := \beta_0 C_1(\sqrt{rV\mathrm{L}/n} + bV\mathrm{L}/n)$ is a sub-root function dominating $\beta_0 \mathbb{E}\mathfrak{R}_n\{f \in \mathcal{F} : Pf^2 \le r\}$ for all $r > 0$, and its fixed point satisfies $r^* \le 4C_1^2\beta_0^2 V\mathrm{L}/n$.

Proof. Fix $r > 0$, let $\mathcal{F}_r := \{f \in \mathcal{F} : Pf^2 \le r\}$, which contains 0, and write $\mathfrak{r} := \mathbb{E}\mathfrak{R}_n(\mathcal{F}_r) \le b$ and $\hat{\sigma}^2 := \sup_{f \in \mathcal{F}_r} P_n f^2 \le b^2$. Conditionally on the sample, $X_f := n^{-1/2} \sum_i \varpi_i f(Z_i)$ is a centred sub-Gaussian process with respect to $\mathrm{d}(f,g) := \|f - g\|_{L^2(P_n)}$, and $|X_f - X_g| \le \sqrt{n}\, \mathrm{d}(f,g)$. The diameter of $\mathcal{F}_r$ is at most $2\hat{\sigma}$. Since $X_0 = 0$, Theorem 5.22 of (Wainwright, 2019) with $\delta := 4b/n$ (or, when δ exceeds the diameter D_r of $\mathcal{F}_r$, the trivial bound $\sup(X_f - X_g) \le \sqrt{n} D_r < \sqrt{n}\delta$) gives

$$\mathbb{E}_\varpi \mathfrak{R}_n(\mathcal{F}_r) \le \frac{1}{\sqrt{n}} \mathbb{E}_\varpi \sup_{f,g \in \mathcal{F}_r} (X_f - X_g) \le \frac{8b}{n} + \frac{32}{\sqrt{n}} \int_{b/n}^{2\hat{\sigma}} \sqrt{\log N_{\mathrm{int}}(u, \mathcal{F}_r, L^2(P_n))}\, du,$$

with the integral read as 0 if $2\hat{\sigma} \le b/n$, where N_{int} counts nets with centres in $\mathcal{F}_r$. For $u \le 2\hat{\sigma} \le 2\mathfrak{l}$, choosing a point of $\mathcal{F}_r$ in each ball of an external $(u/2)$-net shows $\log N_{\mathrm{int}}(u, \mathcal{F}_r, L^2(P_n)) \le \log N(u/2, \mathcal{F}, L^2(P_n)) \le V \log(2e\mathfrak{l}/u) \le V\mathrm{L}$ for $u \ge b/n$. Hence $\mathbb{E}_\varpi \mathfrak{R}_n(\mathcal{F}_r) \le 8b/n + 64\hat{\sigma}\sqrt{V\mathrm{L}/n}$, and taking expectations, with Jensen, $\mathfrak{r} \le 8b/n + 64\sqrt{V\mathrm{L}/n}\,(\mathbb{E}\hat{\sigma}^2)^{1/2}$. Next, $\hat{\sigma}^2 \le r + \sup_{f \in \mathcal{F}_r} |P_n f^2 - Pf^2|$. By symmetrization (Lemma 2.3.1 of (van der Vaart & Wellner, 1996)) and the contraction principle (Theorem 4.12 of (Ledoux & Talagrand, 1991)) applied to $\phi(v) := \min\{v^2, b^2\}/(2b)$, which is a contraction with $\phi(0) = 0$ and agrees with $v^2/(2b)$ on $[-b, b]$,

$$\mathbb{E} \sup_{f \in \mathcal{F}_r} |P_n f^2 - Pf^2| \le 2\mathbb{E}\mathfrak{R}_n\{f^2 : f \in \mathcal{F}_r\} = 4b\, \mathbb{E}\mathfrak{R}_n\{\phi \circ f : f \in \mathcal{F}_r\} \le 8b\,\mathfrak{r}.$$

Therefore $\mathfrak{r} \le 8b/n + 64\sqrt{V\mathrm{L}/n}\,(\sqrt{r} + \sqrt{8b\mathfrak{r}})$. By $xy \le x^2/2 + y^2/2$ with $x := \sqrt{\mathfrak{r}}$, $64\sqrt{8bV\mathrm{L}/n}\sqrt{\mathfrak{r}} \le \mathfrak{r}/2 + 16384\, bV\mathrm{L}/n$, hence $\mathfrak{r} \le 16b/n + 128\sqrt{rV\mathrm{L}/n} + 32768\, bV\mathrm{L}/n$, which is (H.14) with $C_1 := 32784$, because $V\mathrm{L} \ge 1$.

The function $\mathfrak{s}$ is nonnegative, nondecreasing, and $\mathfrak{s}(r)/\sqrt{r}$ is nonincreasing, so it is sub-root and has a unique positive fixed point. At it, $r^* = X + Y$ with $X := \beta_0 C_1 \sqrt{r^* V\mathrm{L}/n}$ and $Y := \beta_0 C_1 bV\mathrm{L}/n$, so $r^* \le 2\max(X, Y)$. If $X \ge Y$, dividing $r^* \le 2X$ by $\sqrt{r^*}$ and squaring gives $r^* \le 4\beta_0^2 C_1^2 V\mathrm{L}/n$. If $Y > X$, then $r^* \le 2Y \le 2C_1\beta_0^2 V\mathrm{L}/n$ by $b \le \beta_0$. Both are at most $4C_1^2\beta_0^2 V\mathrm{L}/n$. □

Theorem H.7. *(Deviation bound) Grant Assumptions 1–2, let $\mathcal{H}_N, \Psi_N$ be as in Definition E.3 with R as in* (F.13)*, and fix $\zeta \in (0,1)$. Put*

$$\delta_n^2 := C_0 \sigma^{-c} \frac{(\mathbb{V}_N + \log(1/\zeta)) \log^2(n/\zeta)}{n}, \tag{H.15}$$

where $C_0\sigma^{-c}$ is a sufficiently large constant of the form fixed in Appendix A. Then with probability at least $1 - \zeta$, simultaneously for all $h \in \mathcal{H}_N$ and all $\psi \in \Psi_N$:

(i) $|\int_I [\hat{a}_t - a_t]\,(h, \psi)\nu(dt)| \le \delta_n \|\psi\|_{\mathcal{V}} + \delta_n^2 \|\psi\|_\bullet \le \sqrt{2}\delta_n \|\psi\|_{\mathcal{V}, \delta_n}$,

(ii) $\frac{1}{2}\mathcal{D}(h) - \delta_n^2 \le \hat{\mathcal{D}}(h) - \hat{\mathcal{D}}(h^*) \le 2\mathcal{D}(h) + \delta_n^2$,

(iii) $\frac{1}{2}\|\psi\|_{\mathcal{V}}^2 - \delta_n^2\|\psi\|_\bullet^2 \le \|\psi\|_{\hat{\mathcal{V}}}^2 \le 2\|\psi\|_{\mathcal{V}}^2 + \delta_n^2\|\psi\|_\bullet^2$, *and consequently*

$$\frac{1}{4}\|\psi\|_{\mathcal{V},\delta_n}^2 \le \|\psi\|_{\hat{\mathcal{V}},\delta_n}^2 \le 2\|\psi\|_{\mathcal{V},\delta_n}^2, \tag{H.16}$$

a purely multiplicative comparison.

For $\zeta \ge n^{-C'}$ *one has* $\log(n/\zeta) \le (C'+1)\log n$, *so that* $\delta_n^2 \le C''\sigma^{-c}(\mathbb{V}_N + \log(1/\zeta))\log^2 n/n$.

Proof. Constants. Write $\eta_n := \delta_n/3$ and fix an absolute $C_T \ge 1$ with $cC_T^2 \ge 2$ and $C_T\sqrt{\log 8} \ge C$, where c, C are the absolute constants of Lemma H.1(iii). Put $\mathcal{T} := C_T\sqrt{d}\sigma_*^{-2}\sqrt{\log(8n/\zeta)}$ and $\Sigma^2 := 3(G_t^2 + G^2\mathcal{T}^2 + G^2)$, and let $\mathfrak{b}, \mathfrak{l}$ be as in (H.8) at this $\mathcal{T}$. From (E.8), (F.13), $n \ge 2$ and $\log(8n/\zeta) \le 4\log(n/\zeta)$, with σ^{-c} as in Appendix A:

$$\begin{aligned} &\Gamma^2 + G_t^2 \le \sigma^{-c}\log n, \quad \Sigma^2 + \mathfrak{b} \le \sigma^{-c}\log(n/\zeta), \quad \mathfrak{l} \le \sigma^{-c}\log^2(n/\zeta), \\ &\log(2e\mathfrak{l}n/b) \le \sigma^{-c}\log(n/\zeta) \quad \text{for every } b \ge 12\bar{B}^2. \end{aligned} \tag{H.17}$$

Choosing $C_0 \ge 9/\log^2 2$ gives $\eta_n^2 \ge 1/n$, hence $J := \max\{0, \lceil \log_2(\Gamma/\eta_n)\rceil\}$ obeys $J \le \log_2(\Gamma\sqrt{n}) + 1$ and $\log(J+1) \le \sigma^{-c}\log n$. Every inequality "$\le \delta_n^2$" or "$\le \eta_n^2$" below holds once C_0 is large enough, because its left-hand side is at most $\sigma^{-c}(\mathbb{V}_N + \log(1/\zeta))\log^2(n/\zeta)/n$ by (H.17) and $\mathbb{V}_N \ge 1$. By Lemma H.2, the three index classes are separable, bounded by $\mathfrak{b} \le \mathfrak{l}$, and satisfy (H.13) with $V = 6\mathbb{V}_N$. Put $u := \log(8/\zeta)$.

Step 1 (value class). $\mathcal{F}_1 := \operatorname{star}\mathcal{D}_\bullet$ is separable, star-shaped, bounded by $b_1 := 12\bar{B}^2$, and satisfies (H.13) with $V = 24\mathbb{V}_N$ by Lemma H.5. Since $PD_h = \mathcal{D}(h) \ge 0$ and $PD_h^2 \le 16\bar{B}^2 PD_h$ (Lemma H.1(iv)), $P(\lambda D_h)^2 \le \lambda\beta_1 P(\lambda D_h) \le \beta_1 P(\lambda D_h)$ for $\lambda \in [0,1]$ with $\beta_1 := 16\bar{B}^2 \ge b_1$. Lemma H.6 gives $r_1^* \le 96C_1^2\beta_1^2\mathbb{V}_N\mathrm{L}_1/n$ with $\mathrm{L}_1 := \log(2e\mathfrak{l}n/b_1)$, and Theorem H.3 gives, on an event E_1 of probability at least $1 - 2e^{-u} = 1 - \zeta/4$, for all $h \in \mathcal{H}_N$,

$$\frac{1}{2}\mathcal{D}(h) - \Delta_1 \le \hat{\mathcal{D}}(h) - \hat{\mathcal{D}}(h^*) \le 2\mathcal{D}(h) + \Delta_1, \qquad \Delta_1 := c_1\Big(96C_1^2\beta_1\frac{\mathbb{V}_N\mathrm{L}_1}{n} + \frac{28\bar{B}^2u}{n}\Big) \le \delta_n^2,$$

using $PD_h = \mathcal{D}(h)$ and $P_nD_h = \hat{\mathcal{D}}(h) - \hat{\mathcal{D}}(h^*)$. This is (ii).

Step 2 (norm class, generator scale). $\mathcal{F}_2 := \operatorname{star}\mathcal{Q}_1$ is separable, star-shaped, bounded by $b_2 := \Gamma^2$, satisfies (H.13) with $V = 24\mathbb{V}_N$, and $Pf^2 \le \Gamma^2 Pf$ on it by Lemma H.1(v). With $\beta_2 := \Gamma^2$ and $\mathrm{L}_2 := \log(2e\mathfrak{l}n/\Gamma^2)$, Lemma H.6 and Theorem H.3 give an event E_2 of probability at least $1 - \zeta/4$ on which, for all $\psi \in \Psi_N^\bullet$,

$$\frac{1}{2}\|\psi\|_{\mathcal{V}}^2 - \eta_n^2 \le \|\psi\|_{\hat{\mathcal{V}}}^2 \le 2\|\psi\|_{\mathcal{V}}^2 + \eta_n^2,$$

because $PQ_\psi = \|\psi\|_{\mathcal{V}}^2$, $P_nQ_\psi = \|\psi\|_{\hat{\mathcal{V}}}^2$ and $c_1(96C_1^2\Gamma^2\mathbb{V}_N\mathrm{L}_2 + 2\Gamma^2u)/n \le \eta_n^2$ by (H.17).

Step 3 (bilinear class, generator scale). (a) Truncation. By Lemma H.1(iii) and $cC_T^2 \ge 2$, the event $E_{\mathcal{T}} := \{\max_{i\le n}\bar{V}(Z_i) \le \mathcal{T}\}$ has probability at least $1 - 2n(\zeta/8n)^{cC_T^2} \ge 1 - \zeta/4$, and on it $P_nG_{h,\psi} = P_nG_{h,\psi}^{\mathcal{T}}$ for all (h,ψ). Since $\mathcal{T} \ge C\sqrt{d}\sigma_*^{-2}$, the same lemma gives, for $\psi \in \Psi_N^\bullet$, $|PG_{h,\psi} - PG_{h,\psi}^{\mathcal{T}}| \le C\mathcal{T}\Gamma G(\zeta/8n)^2 \le \sigma^{-c}\log(8n/\zeta)(\zeta/8n)^2 \le \sigma^{-c}/n \le \eta_n^2$, using $y^{-2}\log y \le y^{-1}$ for $y \ge 1$.

(b) Shells. For $\psi \in \Psi_N^\bullet$ we have $\|\psi\|_{\mathcal{V}} \le \Gamma_{\mathcal{V}} \le \Gamma$. For $0 \le j \le J$ put $r_j := 2^{-j}\Gamma$ and $\mathcal{G}_j := \{G_{h,\psi}^{\mathcal{T}} : h \in \mathcal{H}_N,\ \psi \in \Psi_N^\bullet,\ \|\psi\|_{\mathcal{V}} \le r_j\}$. Each $\mathcal{G}_j$ is a separable subclass of $\mathcal{G}_1^{\mathcal{T}}$ containing 0, bounded by $\mathfrak{b}$, satisfying (H.13) with $V = 6\mathbb{V}_N$, and with $\operatorname{Var}(f) \le Pf^2 \le \varsigma_j^2 := \Sigma^2 r_j^2$ by (H.7). Symmetrization and Lemma H.6 with $\mathrm{L}_3 := \log(2e\mathfrak{l}n/\mathfrak{b})$ give

$$\mathbb{E}\sup_{f\in\mathcal{G}_j}|P_nf - Pf| \le 2\mathbb{E}\mathfrak{R}_n\{f \in \mathcal{G}_1^{\mathcal{T}} : Pf^2 \le \varsigma_j^2\} \le 2C_1\Big(\varsigma_j\sqrt{\frac{6\mathbb{V}_N\mathrm{L}_3}{n}} + \frac{6\mathfrak{b}\mathbb{V}_N\mathrm{L}_3}{n}\Big).$$

Theorem H.4 with the constant envelope $\mathfrak{b}$ (so $\|\max_i \mathfrak{b}\|_{\psi_1} = \mathfrak{b}/\log 2$) and $u_3 := \log(4(J+1)/\zeta) \geq 1$ gives, with probability at least $1-\zeta/(4(J+1))$,

$$\sup_{f\in\mathcal{G}_j} |P_n f - Pf| \leq \varsigma_j \frac{4C_1\sqrt{6\mathbb{V}_N \mathrm{L}_3} + c_2\sqrt{u_3}}{\sqrt{n}} + \frac{24C_1\mathfrak{b}\mathbb{V}_N \mathrm{L}_3 + 2c_2\mathfrak{b}u_3}{n} \leq r_j\eta_n + \eta_n^2,$$

where the last step uses $\varsigma_j = \Sigma r_j$, $u_3 \leq \sigma^{-c}\log(n/\zeta)$ and (H.17), so that $\Sigma^2(\mathbb{V}_N \mathrm{L}_3 + u_3) \leq \sigma^{-c}(\mathbb{V}_N + \log(1/\zeta))\log^2(n/\zeta)$ and $\mathfrak{b}(\mathbb{V}_N \mathrm{L}_3 + u_3) \leq \sigma^{-c}(\mathbb{V}_N + \log(1/\zeta))\log^2(n/\zeta)$. Let E_3 be the intersection of these $J+1$ events, $\mathbb{P}(E_3) \geq 1-\zeta/4$.

(c) Combination. On $E_{\mathcal{T}} \cap E_3$, let $h \in \mathcal{H}_N$, $\psi \in \Psi_N^\bullet$ and $j_* := \max\{j \leq J : \|\psi\|_{\mathcal{V}} \leq r_j\}$. If $j_* < J$ then $\|\psi\|_{\mathcal{V}} > r_{j_*+1} = r_{j_*}/2$, if $j_* = J$ then $r_J \leq \eta_n$. Hence $r_{j_*} \leq 2\|\psi\|_{\mathcal{V}} + \eta_n$ and, by (a) and (b),

$$|P_n G_{h,\psi} - PG_{h,\psi}| \leq r_{j_*}\eta_n + \eta_n^2 + \eta_n^2 \leq 2\eta_n\|\psi\|_{\mathcal{V}} + 3\eta_n^2, \quad \psi \in \Psi_N^\bullet. \tag{H.18}$$

Step 4 (from generator scale to the cone). Work on $E_1 \cap E_2 \cap E_{\mathcal{T}} \cap E_3$, which has probability at least $1-\zeta$. Fix h and $\psi \in \Psi_N$ and let $\psi = c\,\omega\phi^\chi$ be any representation as in (E.7), then $\upsilon := \omega\phi^\chi \in \Psi_N^\bullet$. Since $\psi \mapsto G_{h,\psi}$ is linear, (H.18) applied to υ gives

$$|P_n G_{h,\psi} - PG_{h,\psi}| = |c|\,|P_n G_{h,\upsilon} - PG_{h,\upsilon}| \leq 2\eta_n\|\psi\|_{\mathcal{V}} + 3\eta_n^2|c|.$$

The left-hand side does not depend on the representation, so taking the infimum over representations replaces $|c|$ by $\|\psi\|_\bullet$. As $2\eta_n \leq \delta_n$ and $3\eta_n^2 \leq \delta_n^2$, and by (H.5), this is the first inequality of (i), the second is $a+b \leq \sqrt{2}(a^2+b^2)^{1/2}$. Likewise $Q_\psi = c^2 Q_\upsilon$, and Step 2 applied to υ gives $\frac{1}{2}\|\psi\|_{\mathcal{V}}^2 - \eta_n^2 c^2 \leq \|\psi\|_{\hat{\mathcal{V}}}^2 \leq 2\|\psi\|_{\mathcal{V}}^2 + \eta_n^2 c^2$ for every representation, whence (iii) after taking the infimum and using $\eta_n \leq \delta_n$. Adding $\delta_n^2\|\psi\|_\bullet^2$ to (iii) yields $\|\psi\|_{\hat{\mathcal{V}},\delta_n}^2 \geq \frac{1}{2}\|\psi\|_{\mathcal{V}}^2$ and $\|\psi\|_{\hat{\mathcal{V}},\delta_n}^2 \leq 2\|\psi\|_{\mathcal{V},\delta_n}^2$, since also $\|\psi\|_{\hat{\mathcal{V}},\delta_n}^2 \geq \delta_n^2\|\psi\|_\bullet^2$, averaging the two lower bounds gives (H.16). Statement (ii) is Step 1. □

Theorem H.8. *(Guidance rate) Grant Assumptions 1–2, fix $\zeta \in (0,1)$ and $\kappa \geq 1$, let R, K, Q_n be as in* (F.13)–(F.14)*, and choose*

$$N := \left\lceil \left(n\mathrm{A}_\beta^2\right)^{\frac{1}{2(\beta-1)+d}} \right\rceil, \quad \mu, \alpha \in [\kappa^{-1}, \kappa], \quad \mathrm{b} := (\varepsilon_n/\Xi_{\mathrm{s}})^{\frac{1}{\mathrm{s}+4}} \wedge \bar{B} \ \ (\mathrm{b} := B \text{ if } \mathrm{s} = \infty), \tag{H.19}$$

$$\varepsilon_n := \mathrm{A}_\beta^{\frac{2d}{2(\beta-1)+d}} n^{-\frac{2(\beta-1)}{2(\beta-1)+d}}. \tag{H.20}$$

Let $\mathcal{H}_N, \Psi_N$ be as in Definition E.3 with budgets equal to the right-hand sides of (F.16)*, and let $\hat{h}, \hat{g}$ be given by* (H.1) *with δ_n as in* (H.15)*. There is n_0, depending only on $(d, \beta, \bar{B}, \ell, \sigma_*, C_t, \kappa_\Theta)$, such that for $n \geq n_0$, on the event of Theorem H.7,*

$$\|\hat{g} - g^*\|_{L^2(\nu\otimes\rho; I^\circ)}^2 \leq \sigma^{-c}\log^8(en/\zeta)\,\Xi_{\mathrm{s}}^{\frac{4}{\mathrm{s}+4}}\varepsilon_n^{\frac{\mathrm{s}}{\mathrm{s}+4}}, \tag{H.21}$$

where σ^{-c} may also depend on κ, and where at $\mathrm{s} = \infty$ the right-hand side is read as $\sigma^{-c}\log^8(en/\zeta)B^{-4}\varepsilon_n$. At fixed β the exponent of n in (H.20) *is $2(\beta-1)/(2(\beta-1)+d)$, so* (H.21) *is $n^{-2/(d+2)}$ up to logarithmic factors at $\beta = 2$ and $\mathrm{s} = \infty$. The bound contains no regularization-gap term, for any fixed κ.*

Proof. Write $D(h) := \mathcal{D}(h)$, $\Pi_\Psi(h) := \mathcal{N}_{\Psi_N,\delta_n}(h)^2$ and $\mathfrak{D} := 2(\beta-1)+d$, and let $\hat{F}(h)$ denote the supremum in (H.1).

Step 0 (regime and approximant). Put $x_n := (n\mathrm{A}_\beta^2)^{1/\mathfrak{D}} \geq 1$, so $x_n \leq N \leq 2x_n$. Then $\mathrm{A}_\beta^2 N^{-2(\beta-1)} \leq \mathrm{A}_\beta^2 x_n^{-2(\beta-1)} = \varepsilon_n$ and $N^d/n \leq 2^d x_n^d/n = 2^d\varepsilon_n$. Since A_β and Q_n grow only polylogarithmically in n while x_n grows polynomially, enlarging n_0 ensures (F.15) and the conditions of Lemma F.8 for $n \geq n_0$. That lemma gives $h_N \in \mathcal{H}_N$ satisfying (F.3) with $e_N := 8Q_n\mathrm{A}_\beta N^{-(\beta-1)}$, so $e_N^2 \leq 64Q_n^2\varepsilon_n \leq \sigma^{-c}\log^8(n)\,\varepsilon_n$ by (F.14). By Lemma F.1 and (G.1), $\Pi_\Psi(h_N) \leq \mathcal{P}^w(h_N) \leq 4(1+M_4)^2 e_N^2$ and $D(h_N) \leq \sup_t \|h_N - h^*\|_{\rho_t}^2 \leq e_N^2$. With the budgets fixed, Lemma E.8(iii) (whose side conditions $W_N \geq d+1$ and $P_0 \leq P_{\mathcal{H}}$ hold once $C^\sharp \geq d+1$, because $L, W_N \geq 2$ give $(L-1)W_N^2 \geq C^\sharp K(N+1)^d L \geq P_0$) gives $\mathbb{V}_N \leq \sigma^{-c}N^d(\log n)^{7/2}$,

because $\bar{P}_0 \le \sigma^{-c} K N^d$ (as $L \le (N+1)^d$ for $n \ge n_0$), $K \le \sigma^{-c}(\log n)^{3/2}$, and the bracket in (E.19) and $\log(1+4\Gamma)$ are at most $\sigma^{-c}\log^2 n$. Hence, using $\mathbb{V}_N \ge 1$ and $1+\log(1/\zeta) \le \log(en/\zeta)$,

$$\delta_n^2 \le \sigma^{-c}\frac{N^d}{n}(\log n)^{7/2}\big(1+\log(1/\zeta)\big)\log^2(n/\zeta) \le \sigma^{-c}\log^8(en/\zeta)\,\varepsilon_n,$$
$$\bar{\varepsilon}_n := e_N^2 + \delta_n^2 \le \sigma^{-c}\log^8(en/\zeta)\,\varepsilon_n.$$

Throughout the remaining steps we work on the event of Theorem H.7, on which the adversary norms $\|\cdot\|_{\hat{\mathcal{V}},\delta_n}$ and $\|\cdot\|_{\mathcal{V},\delta_n}$ are equivalent within the factors $\frac{1}{4}$ and 2 by (H.16).

Step 1 (the inner maximum). For every $\psi \in \Psi_N$, the definition (G.1) and Theorem H.7(i) give

$$\int_I \hat{a}_t(h,\psi)\nu(dt) \le A(h,\psi) + \sqrt{2}\delta_n\|\psi\|_{\mathcal{V},\delta_n} \le \Big(\sqrt{\Pi_\Psi(h)} + \sqrt{2}\delta_n\Big)\,\|\psi\|_{\mathcal{V},\delta_n}, \tag{H.22}$$

and $\|\psi\|_{\mathcal{V},\delta_n} \le 2\|\psi\|_{\hat{\mathcal{V}},\delta_n}$ by (H.16). Writing $r := \|\psi\|_{\hat{\mathcal{V}},\delta_n}$, the bracket in (H.1) is at most $2(\sqrt{\Pi_\Psi}+\sqrt{2}\delta_n)r - \frac{\alpha}{2}r^2$, whose maximum over $r \ge 0$ is $2(\sqrt{\Pi_\Psi}+\sqrt{2}\delta_n)^2/\alpha$, so

$$\hat{F}(h) \le \frac{4\Pi_\Psi(h) + 8\delta_n^2}{\alpha}. \tag{H.23}$$

In the other direction, fix $\epsilon > 0$ and pick $\hat{\psi} \in \Psi_N$ with $\|\hat{\psi}\|_{\mathcal{V},\delta_n} = 1$ and $A(h,\hat{\psi}) \ge \sqrt{\Pi_\Psi(h)} - \epsilon$, possible because Ψ_N is a symmetric cone. Evaluating the bracket at $\psi = r\hat{\psi}$, using Theorem H.7(i) in the other direction and $\|r\hat{\psi}\|^2_{\hat{\mathcal{V}},\delta_n} \le 2r^2$, and letting $\epsilon \downarrow 0$,

$$\hat{F}(h) \ge \sup_{r\ge 0}\Big[r(\sqrt{\Pi_\Psi(h)} - \sqrt{2}\delta_n) - \alpha r^2\Big] = \frac{(\sqrt{\Pi_\Psi(h)} - \sqrt{2}\delta_n)_+^2}{4\alpha} \ge \frac{\Pi_\Psi(h)}{8\alpha} - \frac{\delta_n^2}{2\alpha}, \tag{H.24}$$

the last step by $(x-y)_+^2 \ge \frac{1}{2}x^2 - y^2$ for $x, y \ge 0$.

Step 2 (main inequality). Since $\hat{h}$ minimizes and $h_N \in \mathcal{H}_N$, Theorem H.7(ii) gives

$$\begin{aligned}\tfrac{1}{2}D(\hat{h}) - \delta_n^2 + \mu\hat{F}(\hat{h}) &\le \hat{\mathcal{D}}(\hat{h}) - \hat{\mathcal{D}}(h^*) + \mu\hat{F}(\hat{h}) \le \hat{\mathcal{D}}(h_N) - \hat{\mathcal{D}}(h^*) + \mu\hat{F}(h_N)\\ &\le 2D(h_N) + \delta_n^2 + \mu\hat{F}(h_N).\end{aligned}$$

Inserting (H.24) on the left and (H.23) on the right, with $q := \mu/\alpha \in [\kappa^{-2}, \kappa^2]$, gives $\min\{\frac{1}{2}, \frac{q}{8}\}(D(\hat{h}) + \Pi_\Psi(\hat{h})) \le \max\{2, 4q\}(D(h_N) + \Pi_\Psi(h_N)) + (2 + 8.5q)\delta_n^2$, hence

$$D(\hat{h}) + \Pi_\Psi(\hat{h}) \le C_\kappa\big(D(h_N) + \Pi_\Psi(h_N) + \delta_n^2\big) \le \sigma^{-c}\bar{\varepsilon}_n, \tag{H.25}$$

with C_κ depending only on κ, using Step 0 and $(1+M_4)^2 \le Cd\sigma_*^{-4}$. Only the ratio μ/α enters, and no choice of it trades bias against variance.

Step 3 (coercivity). By Theorem G.4 for the classes of Definition E.3 ($C_0 = 1$), applied to $\hat{h}$ with $\tau = \delta_n$ and $\delta = 1/2$, together with (H.25), $\Lambda_* \lesssim d^2\sigma_*^{-4}$ and $\ell \le 1$,

$$\int_{I^\circ} \|\nabla\hat{\varphi}\|_{\rho_t}^2\nu(dt) + D(\hat{h}) \le \sigma^{-c}\bar{\varepsilon}_n \le \sigma^{-c}\log^8(en/\zeta)\,\varepsilon_n, \qquad \hat{\varphi} := \hat{h} - h^*.$$

Step 4 (clipping). Corollary D.4, applicable since $\hat{h} \in \mathcal{H}_N$ and $\mathrm{b} \in (0, \bar{B}]$, with $\mathrm{b}^{-2} \le \bar{B}^2\mathrm{b}^{-4}$, gives

$$\|\hat{g} - g^*\|^2_{L^2(\nu\otimes\rho;I^\circ)} \le \sigma^{-c}\log^8(en/\zeta)\,\mathrm{b}^{-4}\varepsilon_n + 4(G^*)^2\,\mathcal{E}_\mathrm{b},$$

where, as in its proof, $\mathcal{E}_\mathrm{b} := \int_{I^\circ}\rho_t(h^* < \mathrm{b})\nu(dt) \le \Xi_\mathrm{s}\mathrm{b}^\mathrm{s}$ for $\mathrm{s} < \infty$. Let $\mathrm{s} < \infty$, $\mathrm{b}_0 := (\varepsilon_n/\Xi_\mathrm{s})^{1/(\mathrm{s}+4)}$ and $\mathcal{V}_n := \Xi_\mathrm{s}^{4/(\mathrm{s}+4)}\varepsilon_n^{\mathrm{s}/(\mathrm{s}+4)} = \mathrm{b}_0^{-4}\varepsilon_n = \Xi_\mathrm{s}\mathrm{b}_0^\mathrm{s}$. If $\mathrm{b}_0 \le \bar{B}$, then $\mathrm{b} = \mathrm{b}_0$ and both terms are at most $\sigma^{-c}\log^8(en/\zeta)\mathcal{V}_n$, as $G^* \le \sigma^{-c}$. If $\mathrm{b}_0 > \bar{B}$, then, because $w \le \bar{B}$ on $\mathrm{supp}(p_0)$ gives $\Xi_\mathrm{s}\bar{B}^\mathrm{s} = \mathbb{E}_{p_0}(\bar{B}/w)^\mathrm{s} \ge 1$, we have $\mathcal{V}_n = \Xi_\mathrm{s}\mathrm{b}_0^\mathrm{s} > \Xi_\mathrm{s}\bar{B}^\mathrm{s} \ge 1$; since $|\hat{g}| \le G^*$ by construction and $|g^*| \le G^*$ on I by Lemma B.3, $\|\hat{g} - g^*\|^2_{L^2(\nu\otimes\rho;I^\circ)} \le 4(G^*)^2 \le 4(G^*)^2\mathcal{V}_n$. In both cases (H.21) follows. If $\mathrm{s} = \infty$, then $w \ge B$ on $\mathrm{supp}(p_0)$ gives $h^* \ge B$, so $\mathcal{E}_B = 0$ and the first term with $\mathrm{b} = B \le \bar{B}$ gives the stated bound. □

Remark H.9. (No variance amplification) Heuristically, the gap is a factor N^2 of variance: a resolution-N value estimator has variance N^d/n, differentiating it amplifies this to N^{d+2}/n, and balancing against the gradient bias $N^{-2(\beta-1)}$ then forces $N^{2\beta+d} \asymp n$. Our estimator also differentiates its fit, $\hat{g} = \Pi_{G^*}(\nabla\hat{h}/(\hat{h} \vee \mathrm{b}))$, but $\hat{h}$ is not a pure value estimate: by (12) each test function supplies a linear moment condition on the first derivatives of h^*, so the gradient of the fit is controlled at complexity $\mathbb{V}_N \asymp N^d$ with no amplification, and the balance is $N^{2(\beta-1)+d} \asymp n$, as in Step 4.

I From guidance error to sampling error

Lemma I.1. *(The controlled marginals are a bounded reweighting) Under Assumptions 1–2, define*

$$\rho_t^*(dx) := \frac{h^*(t,x)}{Z}\rho_t(dx), \quad Z = \mathbb{E}_{p_0}[w] \in (0, \bar{B}]. \tag{I.1}$$

Then each ρ_t^* is a probability measure, $\rho_T^* = q_0$, and $(\rho_t^*)_{t\in[0,T]}$ is the marginal flow of the exactly guided SDE (8) *started from ρ_0^*. Moreover*

$$\Upsilon := \sup_{t\in[0,T]} \left\| \frac{d\rho_t^*}{d\rho_t} \right\|_\infty \le \frac{\bar{B}}{Z} \le \bar{B}\,\Xi_{\mathrm{s}}^{1/\mathrm{s}}, \tag{I.2}$$

the last expression read as $\bar{B}/B$ when $\mathrm{s} = \infty$.

Proof. Total mass: $\int h^*(t,\cdot)\rho_t = \mathbb{E}[h^*(t, \overleftarrow{X}_t)] = \mathbb{E}[w(X_0)] = Z$, by Proposition C.1 and the tower property, so ρ_t^* is a probability measure. At $t = T$ it is $wp_0/Z = q_0$.

For the flow statement, $L_t := h^*(t, \overleftarrow{X}_t)$ is a positive $\mathbb{P}$-martingale with $dL_t = \sqrt{2}\nabla h^* \cdot dB_t$ by Theorem C.2, so $dL_t/L_t = \sqrt{2}\nabla \log h^* \cdot dB_t$ and $L_t/L_0 = \mathcal{E}(\int \sqrt{2}\nabla \log h^* \cdot dB)_t$. Put $d\mathbb{Q}/d\mathbb{P}|_{\mathcal{F}_t} := L_t/Z$. Girsanov's theorem makes $\tilde{B}_t := B_t - \int_0^t \sqrt{2}\nabla \log h^* ds$ as a $\mathbb{Q}$-Brownian motion, so under $\mathbb{Q}$

$$d\overleftarrow{X}_t = (b + \sqrt{2}\cdot\sqrt{2}\nabla \log h^*)dt + \sqrt{2}d\tilde{B}_t = (b + 2\nabla \log h^*)dt + \sqrt{2}d\tilde{B}_t, \tag{I.3}$$

which is (8), and the $\mathbb{Q}$-marginals are $\mathbb{E}_{\mathbb{P}}[f(\overleftarrow{X}_t)L_t]/Z = \int f h^* \rho_t/Z$, that is ρ_t^*. Note that the tilt acts on the initial condition as well: the exactly guided process must be started from $\rho_0^* = h^*(0,\cdot)p_T/Z$ rather than from p_T. Initialization error is not addressed here.

For (I.2), $h^* \le \bar{B}$ by Assumption 1, while Jensen applied to the convex map $u \mapsto u^{-\mathrm{s}}$ gives $\mathbb{E}_{p_0}[w^{-\mathrm{s}}] \ge (\mathbb{E}_{p_0}[w])^{-\mathrm{s}}$, i.e. $Z \ge \Xi_{\mathrm{s}}^{-1/\mathrm{s}}$. At $\mathrm{s} = \infty$ one has $Z \ge B$ directly. □

Corollary I.2. *(Sampling guarantee in the controlled metric) Let $J := [t_0, \bar{t}] \subseteq I^\circ$, and let $\mathbb{Q}^*$ and $\hat{\mathbb{Q}}$ be the laws on $C(J;\mathbb{R}^d)$ of*

$$d\overleftarrow{Z}_t = (b + 2g^*)(t, \overleftarrow{Z}_t)dt + \sqrt{2}dB_t, \quad d\overleftarrow{Y}_t = (b + 2\hat{g})(t, \overleftarrow{Y}_t)dt + \sqrt{2}dB_t, \tag{I.4}$$

both started at time t_0 from $\rho_{t_0}^*$. Then $\hat{\mathbb{Q}}$ and $\mathbb{Q}^*$ are equivalent,

$$\mathrm{KL}(\mathbb{Q}^*||\hat{\mathbb{Q}}) = \int_J \|\hat{g} - g^*\|_{\rho_t^*}^2 dt \le \ell\,\Upsilon\,\|\hat{g} - g^*\|_{L^2(\nu\otimes\rho;J)}^2, \tag{I.5}$$

and, writing $\hat{\rho}_{\bar{t}} := \mathrm{Law}(\overleftarrow{Y}_{\bar{t}})$ for the time-$\bar{t}$ marginal of the estimated sampler,

$$\mathrm{TV}(\rho_{\bar{t}}^*, \hat{\rho}_{\bar{t}})^2 \le \frac{1}{2}\ell\,\Upsilon\,\|\hat{g} - g^*\|_{L^2(\nu\otimes\rho;J)}^2. \tag{I.6}$$

In particular, under the hypotheses of Theorem H.8, for $n \ge n_0$ and on the event of Theorem H.7,

$$\mathrm{TV}(\rho_{\bar{t}}^*, \hat{\rho}_{\bar{t}})^2 \le \sigma^{-c}\log^8(en/\zeta)\,\ell\bar{B}\,\Xi_{\mathrm{s}}^{\frac{1}{\mathrm{s}}+\frac{4}{\mathrm{s}+4}}\varepsilon_n^{\frac{\mathrm{s}}{\mathrm{s}+4}}, \tag{I.7}$$

which at $\mathrm{s} = \infty$ reads $\sigma^{-c}\log^8(en/\zeta)\,\ell\bar{B}B^{-5}\varepsilon_n$, of order $n^{-2/(d+2)}$ up to logarithmic factors at $\beta = 2$.

Proof. The drift difference is $2(\hat{g} - g^*)$, of modulus at most $4G^*$ since $|\hat{g}| \le G^*$ by construction (D.20) and $|g^*| \le G^*$ by Lemma B.3. With $\theta := \sqrt{2}(\hat{g} - g^*)$, Novikov's condition holds trivially

$$\mathbb{E}_{\mathbb{Q}^*} \exp\left(\frac{1}{2}\int_J |\theta|^2 dt\right) \le \exp(4(G^*)^2\ell) < \infty, \tag{I.8}$$

so the two path laws are equivalent and

$$\mathrm{KL}(\mathbb{Q}^*||\hat{\mathbb{Q}}) = \mathbb{E}_{\mathbb{Q}^*}\left[\frac{1}{2}\int_J |\theta|^2 dt\right] = \int_J \mathbb{E}_{\rho_t^*}|\hat{g} - g^*|^2 dt, \tag{I.9}$$

which is the identity in (I.5). This is where the projection Π_{G^*} earns its keep beyond Corollary D.4. The unprojected ratio obeys only $|\nabla\hat{h}/(\hat{h} \vee \mathrm{b})| \le G/\mathrm{b}$, and $\mathrm{b} \asymp (\varepsilon_n/\Xi_\mathrm{s})^{1/(\mathrm{s}+4)} \downarrow 0$ by (H.19), so that bound degrades with n and the Novikov constant with it, whereas $G^* = 2\sqrt{d}\sigma_*^{-2}$ is independent of n.

The inequality in (I.5) is Lemma I.1: $\int_J \|\cdot\|^2_{\rho_t^*} dt \le \Upsilon \int_J \|\cdot\|^2_{\rho_t} dt = \ell\,\Upsilon \int_J \|\cdot\|^2_{\rho_t} \nu(dt)$, using $dt = \ell\,\nu(dt)$. Then (I.6) follows from Pinsker's inequality and the data-processing inequality for the coordinate map $\pi_{\bar{t}} : C(J;\mathbb{R}^d) \to \mathbb{R}^d$, and (I.7) from Theorem H.8 together with $\|\cdot\|_{L^2(\nu\otimes\rho;J)} \le \|\cdot\|_{L^2(\nu\otimes\rho;I^\circ)}$ and (I.2). □

Remark I.3. (The window I°) Corollary D.3 and Theorem G.4 deliver the gradient bound on the half-window I° and not on I. The restriction is intrinsic to the endpoint-averaging device and not an artifact of taking ν uniform: for any probability measure ν on I without an atom at t_1 the weight $\omega(t) = \nu([t, t_1])$ is a survival function vanishing at t_1. The argument leaves $\|\varphi(t_1,\cdot)\|^2_{\rho_{t_1}}$ uncontrolled near the right endpoint, i.e. a value error at a single instant, since the statistical analysis controls only the ν-average $\mathcal{D}(h)$, and no reweighting of the spatial marginal manufactures an instantaneous quantity out of a time-average.

At the same time, given a target sampling window $J = [t_0, \bar{t}]$ and any $\delta > 0$, we can run the whole construction on the enlarged estimation window $\tilde{I} := [t_0, \bar{t} + \delta]$, for which $J = \tilde{I}_{\delta/\tilde{\ell}}$ in the notation of (D.19), with $\tilde{\ell} := \bar{t} + \delta - t_0$, provided $\tilde{\ell} \le 1$ and $\bar{t} + \delta < T$. Every statement from Section 4 onwards holds verbatim with ℓ replaced by $\tilde{\ell}$, with the constant 2 of (D.13) replaced by $\tilde{\ell}/\delta$, and with σ_* replaced by $\sigma_{T-\bar{t}-\delta}$. Taking $\delta = \bar{t} - t_0$ doubles the window and keeps the constant at 2. So, the estimator must be run closer to the singular end than the sampler is, and the σ^{-c} prefactors of Theorem H.8 and Corollary I.2 are then evaluated at the deeper cutoff. Nothing else in the argument changes, and in particular Corollary I.2 applies with $J = \tilde{I}_{\delta/\tilde{\ell}}$ in place of I°.

Remark I.4 (From the half-window to I_δ). Several statements above are proved on $I^\circ = [t_0, (t_0 + t_1)/2]$, where $\omega \ge 1/2$. Fix $\delta \in (0,1)$ and set $I_\delta := [t_0, t_1 - \delta\ell]$, so that $\omega(t) = (t_1 - t)/\ell \ge \delta$ on I_δ and $I_{1/2} = I^\circ$. The weight enters only when E_ω is converted into an unweighted integral, and there $\omega \ge \delta$ gives

$$\int_{I_\delta} \|\nabla\varphi\|^2_{\rho_t} \nu(dt) \le \frac{1}{\delta} E_\omega, \tag{I.10}$$

so the left-hand side is at most δ^{-1} times the right-hand side of (D.12) at $\theta = \ell$, equivalently $(2\delta)^{-1}$ times the right-hand side of (D.13), while Theorem G.4 is already stated on I_δ. Corollary D.4 holds verbatim with I° replaced by I_δ, its proof being unchanged. Hence Theorem H.8 and Corollary I.2 hold on I_δ, resp. for $J \subseteq I_\delta$, with an extra factor δ^{-1} : apply Theorem G.4 with this δ in Step 3 of the proof of Theorem H.8, and use $|\hat{g} - g^*| \le 2G^*$ on $I_\delta \subseteq I$ in the capped case of Step 4. What cannot be reached is t_1 itself, where $\omega(t_1) = 0$.

J Further details of the numerical experiments

This section records how the experiments of Section 7 are implemented, and gives the full sweeps behind the numbers quoted there.

J.1 General implementation details

Base models. Both are public unconditional DDPM checkpoints with ϵ-prediction and 1000 training steps, frozen throughout and sampled in 100 DDIM steps at $\eta = 0$, that is deterministically (Table

2) (Google, 2022; Godoy, 2023). Throughout this section η is the DDIM noise level, unrelated to the width of the analytic strip carrying the same name in the approximation lemmas. The target is the superclass of the six animal classes on CIFAR-10 (Krizhevsky, 2009) and of the odd digits on MNIST (LeCun et al., 1998). MNIST model was fine-tuned with random horizontal flips and draws some digits mirrored, so the MNIST classifier is applied to an image and to its mirror and keeps the more confident prediction.

Table 2: Base models. "Share" is the fraction of base samples the held-out head assigns to the target superclass, "SIR-32" the same after 32-fold importance resampling with weights w, and FID is to the model's own reference sample (Appendix J.3).

	checkpoint	parameters	share	SIR-32	FID
CIFAR-10	`google/ddpm-cifar10-32`	35.7M	0.653	0.854	24.5
MNIST	`dvgodoy/ddpm-cifar10-32-mnist`	35.7M	0.523	0.630	7.4

Weight and heads. $r(x_0) = \log \sum_{k \in C} p_\varphi(k \mid x_0)$ and $w = \epsilon + (1-\epsilon)\,\mathrm{sigmoid}((r - r_{\mathrm{ref}})/\alpha)$, with r_{ref} the median of r on the reference sample and α half its interquartile range, so $\epsilon \le w \le 1$ and Assumption 1 holds with $\mathrm{s} = \infty$. Two heads of one architecture are trained on disjoint halves of the labelled set: the tilt head defines w and is used by every estimator, the held-out head scores every reported number and is used by none.

Reference sample. The n reference points are samples of the base model itself, so ρ_t are its own marginals and s in (12) is its network, they are noised at 12 levels in the window $\sigma_{T-t} \in [0.35, 0.95]$. The guidance is zero above the window, and below it $\hat{h}$ is evaluated at the window's lower edge.

Fitted classes. Those of Definition E.3. The trial class is a convolutional network of three blocks, width 32 doubling with depth, with the rectified cubic $\varrho(u) = (u \vee 0)^3$ under a smooth saturation at the activation budget $\mathcal{A} = 3$, the retraction Θ on the output, the spatial clip χ_R on the input, and the weight budget $\Lambda = 5$ imposed by projection after every step, a convolution is a sparse linear map, so such a network is a member at these budgets in $d = 3072$, where a dense map is not. The test class is the cone over differences of two trial networks, which by Lemma E.2 is the network of twice the width and one more layer with Θ as the activation of the appended layer, so (S2) holds as an identity, the ratio measured before each fit is 1.000000. Four things depart from the letter of the definition: the sup-norm constraints are read on the batch rather than over $I \times \mathbb{R}^d$ (measured, $\sup|\nabla h| = 0.24$ against $G = 1.8 \cdot 10^3$ and $\sup|\partial_t h| = 2.9$ against $G_t = 8.2 \cdot 10^5$, so they never bind), the sparsity budget is whatever the convolutions realise, a group normalisation and a smooth time embedding are kept for trainability, both C^∞ so the class stays $C^{1,2}$, neither of the form $\varrho(Ax + b)$, and the penalty is a plug-in square of a batch mean, with a positive bias of order $1/\mathrm{batch}$.

Sampling and cost. The guided sampler replaces $\hat{\epsilon}$ by $\hat{\epsilon} - \sigma\, c\, \nabla \log \hat{h}$, capped at 10% of $\|\hat{\epsilon}\|$ for every guided method. Each configuration draws 3 seeds of 2000 samples with the initial noise of sample i shared across methods. The CIFAR-10 sweep took about three hours on one RTX 4090, the MNIST sweep about two. The code runs from a single entry point with one configuration file per base model, in PyTorch on top of `diffusers`, with FID from `torchmetrics`.

J.2 The empirical objective

With $\omega(u) = (u - u_{\mathrm{lo}})/(u_{\mathrm{hi}} - u_{\mathrm{lo}})$, $(t_i, X_i)_{i \le m}$ the noised reference points, $w_i = w(X_0^i)$ and $s = -\hat{\epsilon}/\sigma$, the estimator minimises over $\mathcal{H}$

$$\frac{1}{m}\sum_i \big(h(t_i,X_i) - w_i\big)^2 + \frac{\mu}{2}\sup_{0 \neq \psi \in \Psi} \frac{\hat{a}(h,\psi)^2}{\|\psi\|_{\hat{\mathcal{V}}}^2}, \quad \hat{a}(h,\psi) = \frac{1}{m}\sum_i \Big[\big(\partial_t h + (x+s)\cdot\nabla h\big)\psi - \nabla h \cdot \nabla \psi\Big](t_i, X_i),$$

the second term being $\mathcal{N}_\Psi(h)^2/2$: Ψ carries a free real scalar and the quotient is invariant under it, so the radial part is solved in closed form and only the direction is optimised. Reverse time runs backwards in u, so $\partial_t h = -\partial_u h$ in the coordinate the reference points are indexed by.

Trial and test networks are trained at two time scales by Adam, two ascent steps per descent step, with a warmup on the learning rate and on μ and a projection of every weight onto Λ after each step, the scale of the test network is held near one, which does not move the maximiser. Each batch draws eight of the twelve levels. The first tenth of the steps runs at $\mu = 0$, after which μ is set so that the

penalty equals the value term at that point rather than tuned: on CIFAR-10 this gives $\mu = 0.0599$, and the penalty then falls by more than three orders of magnitude.

The inner supremum over a network class is nonconvex, so the penalty the fit reports is a lower bound on $\mathcal{N}_\Psi(\hat{h})$. It is therefore also measured against cones the fit never saw, $\omega \cdot \operatorname{span} \phi$ over $Q = 1024$ random cosine features of pooled pixels, where the supremum is a closed-form quadratic and the plug-in bias is removed by cross-fitting over two halves of $24{,}576$ points: $\mathcal{N}^2 = 2.1 \cdot 10^{-3}$ at 8×8 pooling, $1.8 \cdot 10^{-3}$ against a second cone with independent frequencies, $1.1 \cdot 10^{-3}$ at 16×16. A value-only least-squares fit gives $5.7 \cdot 10^{-3}$ on the same cones and a linear instance of the estimator, whose inner supremum is exact, gives $5.1 \cdot 10^{-4}$.

J.3 Metrics used in the experiments

1. *Percentage of the exact tilt's shift*: 100 times the paired gain in held-out reward over the base model, divided by $\mathbb{E}_{q_0} r - \mathbb{E}_{p_0} r$ computed by self-normalized importance sampling on the 6000 base samples, values above 100% are possible for $c > 1$. Standard errors are over the 6000 paired samples, the reported standard deviations over the three seeds, and are the larger of the two.
2. *Share*: the fraction of samples the held-out head assigns to the target superclass.
3. *FID* (Heusel et al., 2017): Fréchet distance between 2048-dimensional Inception (Szegedy et al., 2016) features of the 2000 samples of a seed and of 2000 points of the model's own reference sample, averaged over the 3 seeds, so it measures departure from the base distribution.
4. *Transfer ratio*: the paired gain on the tilt head divided by that on the held-out head, a check that a reported gain is not specific to the head defining w.

J.4 The synthetic instance

The instance. p_0 is atomic on $M = 512$ points drawn uniformly from the ball of radius 0.8, so Assumption 2 holds verbatim, and the tilt obeys $0.5 \le w \le 3$, so Assumption 1 holds with $\bar{B} = 3$ at the boundary case $\mathrm{s} = \infty$. On such a p_0, h^* and its derivatives, the score s and $\nabla \cdot s$ are softmax averages over the atoms, closed-form at any d, the score is never learned. With $T = 1$ the window is $I = [0.40, 0.85]$ and errors are read on I_δ at $\delta = 1/2$, relative to $\|g^*\|^2_{L^2(\nu \otimes \rho; I_\delta)}$, by Monte Carlo on an 8000-point cloud. A replicate resamples atoms, masses, tilt, knots, sample and evaluation cloud, bars are ± 1 standard deviation over 8 replicates.

Trial class and sweeps. Tensor B-splines times a Chebyshev basis in time, cubic because (15) needs $C^{1,2}$, hence $\beta = 4$, here (24) is a single linear solve, the inner supremum being a closed-form quadratic in the coefficients. Seven sample sizes over three decades, resolution chosen by oracle on a grid of nine values and the penalty weight over eleven decades, slopes are fitted on the seed mean and bootstrapped over seeds. A fitted exponent is the rate only if the oracle's resolution stays interior to its grid, which happens at $d = 1$ but not at $d = 2, 3$, where the oracle sits at or near the coarsest cell, the exponent is therefore read at $d = 1$. The comparison with least squares does not depend on the exponent being readable, and is reported at all three dimensions in Table 4: the weak residual is ahead in every one of the 20 cells swept, by a margin that grows with d.

Penalty ablation. Four estimators are fitted on one sample and read on one evaluation cloud, each at its own oracle (N, μ) over the same grids: least squares on the values ($\mu = 0$), the Sobolev penalty $\int_I \|\nabla h\|^2_{\rho_t} \nu(dt)$, the strong residual $\|\mathcal{R}[h]\|^2_{L^2(\nu \otimes \rho)}$ and the weak residual of (24). The strong residual carries the reverse drift, $\partial_t + (x + 2s) \cdot \nabla + \Delta$, the form with $(x + s)$ and $-\Delta$ being its integration by parts against ρ_t and valid only under the integral, the weak penalty is a quadratic form of a sample mean, whose plug-in estimate carries a bias of order P/n, so it is read across two independent halves, as the image fits do.

J.5 Evaluation on CIFAR-10

Estimators compared. Both baselines are implemented against their authors' public code rather than from the papers alone: DEFT from `github.com/alexdenker/DEFT` (commit `2495d46`), DOIT from `github.com/liamyzq/Doob_training_free_adaptation` (commit `9cbc541`), where code and paper differ we follow the code. DEFT (Denker et al., 2024)

Table 3: Penalty ablation at $d = 1$ in the cubic class ($\beta = 4$), 8 seeds, each penalty at its own oracle (N, μ): relative guidance error at the smallest and the largest sample size, against the exponent the theory predicts for it. A rate is predicted only for the weak residual and, from the classical derivative-recovery bound, for least squares, the Sobolev penalty carries a regularization bias and the strong residual an approximation floor (Table 1), and neither is analysed here. "Value LS" is the $\mu = 0$ baseline of Figure 2.

penalty	error at $n = 10^3$	error at $n = 10^6$	predicted exponent
value LS ($\mu = 0$)	6.46×10^{-2}	9.71×10^{-5}	0.667
Sobolev	4.41×10^{-2}	1.40×10^{-4}	—
strong residual	4.46×10^{-3}	2.34×10^{-5}	—
weak residual (ours)	6.51×10^{-3}	1.64×10^{-5}	**0.857**

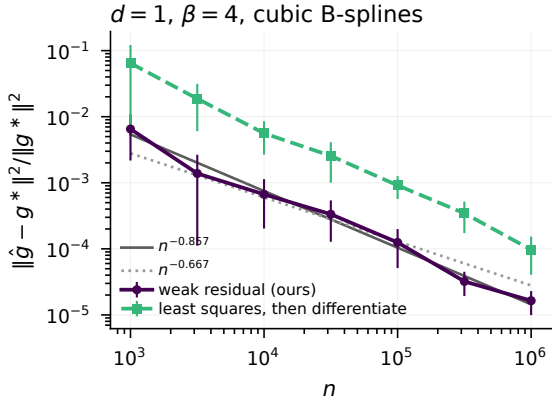


Figure 2: Guidance error against n at $d = 1$ in the tensor cubic B-spline class ($\beta = 4$), for the weak-residual estimator and for least squares on the values followed by differentiation, against the predicted slopes 0.857 and 0.667, intercepts fitted only. Bars are ± 1 standard deviation over 8 seeds. The weak residual's fitted exponent, 0.842 ± 0.026, matches its prediction, least squares runs above its own, 0.909 ± 0.038 against 0.667, so over this range it is measured far from its asymptote and its steeper slope is a climb out of a worse level rather than a faster rate.

trains a head for the h-transform on the same reference sample and weight, in its authors' parametrisation, their equation (12): the head reads the noisy state, the Tweedie estimate $\hat{x}_0$ and the gradient $\nabla_{\hat{x}_0} \log w(\hat{x}_0)$, and returns a network output plus a learned time-dependent multiple of that gradient. It shares the scale c and the cap with every other method, and the deterministic sampler is its authors' own setting rather than a liberty we take: their objective is denoising score matching against the forward process, with no sampler entering it, and the configuration shipped with their code sets the DDIM η to 0.

DOIT (Zhu et al., 2026) is training-free, estimating h and ∇h at each step by Monte Carlo over M one-step look-ahead samples. Two of its choices are not ours to make. Its terminal condition is the exponential tilt $h(\cdot, 0) \propto \exp(r/\tau)$ rather than the bounded w of Assumption 1, and has to be: the estimator resolves only the contrast between the M weights, and forcing the shared w on it leaves them too close together to steer, the field then having cosine 0.01 with the reward gradient where the tilt gives 0.16–0.20. And what it differentiates is the sampler's own transition density, so it is defined only for a stochastic sampler, its reference implementation drawing the look-ahead cloud from the kernel it samples with. Our headline protocol is deterministic, where that density is a point mass, so DOIT is reported at $\eta = 0.4$ as well, the value its authors use, with ours resampled there, the reported DEFT head was not, so that protocol carries ours and DOIT only. Under $\eta = 0$ the look-ahead spread has to be supplied by hand, a departure from the method which we report as such. We sweep DOIT over both its parameters, the temperature τ and the strength γ, the latter playing the role of c under the same cap, and below the shared sweep as well, its authors' own range being $\gamma \leq 0.5$. That range is the only one in which the method is competitive at all, and only on one of the

Table 4: The two estimators at equal budget, 8 seeds, each at its own oracle (N, μ) on the same grids. "Differentiate" is least squares on the values, then differentiated. Entries are relative squared guidance error, the largest n is 10^6, and $3.2 \cdot 10^5$ at $d = 3$.

	weak residual (ours)		differentiate		ratio	
d	$n = 10^3$	largest n	$n = 10^3$	largest n	smallest	largest
1	6.5×10^{-3}	1.6×10^{-5}	6.5×10^{-2}	9.7×10^{-5}	5.9×	13.5×
2	2.3×10^{-2}	6.9×10^{-5}	3.5×10^{-1}	4.2×10^{-4}	5.4×	15.3×
3	8.3×10^{-2}	3.7×10^{-4}	2.5	3.6×10^{-3}	9.7×	30.2×

two datasets: on MNIST $\gamma = 0.01$ buys 88.5% of the exact tilt's shift at FID 14.1 against a base of 7.3, its one point on a usable frontier, where we reach 176% at FID 13.8, on CIFAR-10 the same γ moves the held-out reward the wrong way, -43.8%, and the best small-γ point is 116% at FID 222.

Table 5: CIFAR-10, animals, along the guidance scale (γ for DOIT): held-out reward shift as % of the exact tilt's, and FID to the model's own reference sample. Base FID 24.5 ± 0.2, share 0.653. Bold marks, at each scale, the largest shift and the smallest FID, a row holds c fixed, which leaves the three methods at different FIDs, so the operative reading is the frontier of Figure 3 and not the row. DOIT is entered at the three γ of the shared sweep, where its cap binds on every guided step so that the three agree to 0.2 points, and at the small γ its authors' own setting calls for. Daggered rows are that setting — $\eta = 0.4$ and no cap, read against the base of the same sampler — and are not comparable with the rest of the row they sit in, the wider grid over γ, τ and the sampler is in Appendix J.8. Here and in Table 6, $\pm$ is one standard deviation over the three seeds, which dominates the standard error over the 6000 paired samples (at most 2.3 for the shift).

	ours		DEFT		DOIT	
c	%	FID	%	FID	%	FID
$0.01^\dagger$	—	—	—	—	-43.8 ± 5.1	43.1 ± 0.7
$0.05^\dagger$	—	—	—	—	38.7 ± 2.0	87.0 ± 1.2
$0.1^\dagger$	—	—	—	—	116.0 ± 2.5	221.6 ± 1.0
0.25	3.2 ± 0.3	24.5 ± 0.2	—	—	—	—
0.5	6.1 ± 0.4	24.6 ± 0.2	—	—	—	—
1	11.7 ± 0.8	24.7 ± 0.2	43.0 ± 3.0	25.8 ± 0.4	70.5 ± 8.0	116.7 ± 0.9
2	21.4 ± 1.6	25.0 ± 0.3	78.2 ± 5.2	29.7 ± 0.9	—	—
3	29.7 ± 2.3	25.3 ± 0.3	101.8 ± 6.1	36.2 ± 1.0	—	—
5	44.3 ± 3.3	26.1 ± 0.4	122.6 ± 7.7	53.8 ± 0.4	70.7 ± 7.8	116.7 ± 0.8
10	69.3 ± 3.1	28.7 ± 0.5	125.1 ± 7.3	83.4 ± 0.5	—	—
20	94.0 ± 5.1	32.9 ± 0.7	121.8 ± 7.4	97.7 ± 0.8	70.6 ± 7.8	116.6 ± 0.9

Transfer ratios fall towards one along the sweep, $1.44 \to 1.19$ for ours and $1.37 \to 1.15$ for DEFT. The sweeps are not aligned, so we read them at a matched shift as the frontiers are read, there the two are indistinguishable, 1.23 against 1.26 at 78% of the exact tilt's shift and 1.19 against 1.20 at 94%, and we claim no difference, having no standard error for the ratio. For DOIT the ratio is -0.46 at every γ, and its sign is the reading that matters: the held-out head gains 0.63 while the head defining w, from which DOIT's own terminal condition is built, loses 0.29. At FID 117 the samples are far enough outside the model's distribution that the two classifiers disagree about which way the reward moved, so the 70.6% in the table is not an alignment gain. The pairs of Figure 1c are the five largest rises in the held-out probability among the 64 samples of seed 0 at $c = 20$, a rule rather than a hand-picked set, the same 64 are shown uncurated in Figure 4. Over all 6000 pairs that probability rises from below 0.3 to above 0.7 in 13.4% of them and falls the other way in 0.07%.

J.6 Evaluation on MNIST

The shift exceeds that of the exact tilt from $c = 5$ on, as a scale $c > 1$ allows. Read at a fixed departure from the base distribution rather than at a fixed c, the ordering is the one of CIFAR-10: at the FID DEFT reaches at $c = 1$ our shift is about 92% against its 59.8%, and at our $c = 20$ it is 176.4% against about 84% for DEFT interpolated to the same FID. Transfer ratios grow with the scale here, $1.37 \to 1.54$ for ours and $1.56 \to 1.66$ for DEFT, so ours is the smaller of the two at

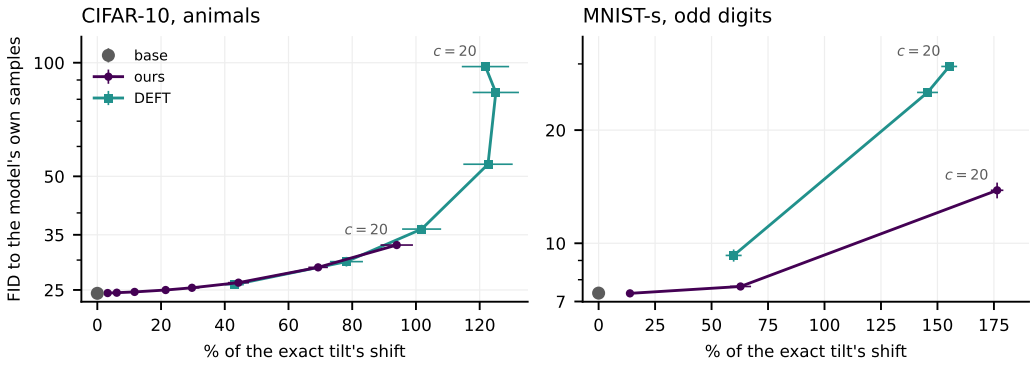


Figure 3: What the alignment costs. Horizontal is the held-out reward shift as a percentage of the exact tilt's, vertical the FID to the model's own reference sample, so the grey point is the base model and a lower curve is cheaper, bars are ± 1 s.d. over three seeds on both axes. Left: CIFAR-10, $c \in [0.25, 20]$. Right: MNIST, $c \in \{1, 5, 20\}$. DOIT is off both panels, at FID 117 and 153 against bases of 24.5 and 7.4, drawing it would compress the region the figure exists to show, and its numbers are in Tables 5 and 6.

Table 6: MNIST, odd digits, tilted under the recipe of Appendix J.1: held-out reward shift as % of the exact tilt's and FID to the model's own reference sample. Base share 0.523, FID 7.4 ± 0.1, SIR-32 share 0.630. Bold marks, at each scale, the largest shift and the smallest FID, and as in Table 5 the row, which holds c fixed, is not the operative comparison. Daggered rows are DOIT at the small γ and in the setting its authors use, $\eta = 0.4$ and no cap, read against the base of that sampler, $\gamma = 0.01$ is the one point at which DOIT reaches a usable frontier at all. $\pm$ is one standard deviation over the three seeds.

	ours		DEFT		DOIT	
c	%	FID	%	FID	%	FID
$0.01^\dagger$	—	—	—	—	88.5 ± 3.7	14.1 ± 0.8
$0.05^\dagger$	—	—	—	—	280.0 ± 7.9	99.4 ± 2.2
$0.1^\dagger$	—	—	—	—	283.4 ± 7.9	108.0 ± 1.4
1	13.9 ± 1.0	7.4 ± 0.1	59.8 ± 3.5	9.3 ± 0.3	296.4 ± 4.1	153.3 ± 3.6
5	62.8 ± 4.7	7.7 ± 0.1	145.6 ± 4.6	25.2 ± 0.2	296.4 ± 4.0	153.1 ± 3.5
20	176.4 ± 2.8	13.8 ± 0.7	155.2 ± 3.5	29.5 ± 0.7	296.3 ± 4.0	152.7 ± 3.5

every scale. DOIT reports 296% at FID 153, twenty times the base, here, unlike on CIFAR-10, both heads move the same way, ratio 1.50, but they are reading a sample the cap could not hold together.

J.7 Final hyperparameter values

Shared by both base models unless stated otherwise.

- *Weight and reference sample*: $n = 20{,}000$ samples from the base model, 12 noise levels per sample in $\sigma_{T-t} \in [0.35, 0.95]$, $\epsilon = 0.05$ in w. Classifier heads: a five-layer convolutional network, the MNIST heads work on a canonical 28×28 grey image, with label smoothing 0.1 and noise augmentation.
- *Our estimator*: trial network of three blocks, width 32 doubling with depth, activation budget $\mathcal{A} = 3$, weight budget $\Lambda = 5$, test network the difference of two such networks. Minimax: 6000 descent steps at batch 128, two ascent steps per descent step, Adam at 10^{-4} and $4 \cdot 10^{-4}$ with $\beta = (0, 0.9)$, warmup over the first tenth and a cosine decay after it, an exponential moving average of the trial weights at 0.999, gradient norm clipped at 5, μ calibrated so that the penalty equals the value term at the end of the warmup.
- *DEFT*: the head of their equation (12) — a U-Net trunk on the concatenation of the noisy state, the Tweedie estimate and the log-weight gradient, 48 base channels, 3 downsampling stages on CIFAR-10 and 2 on MNIST, its last layer initialised at zero, plus a two-layer time embedding

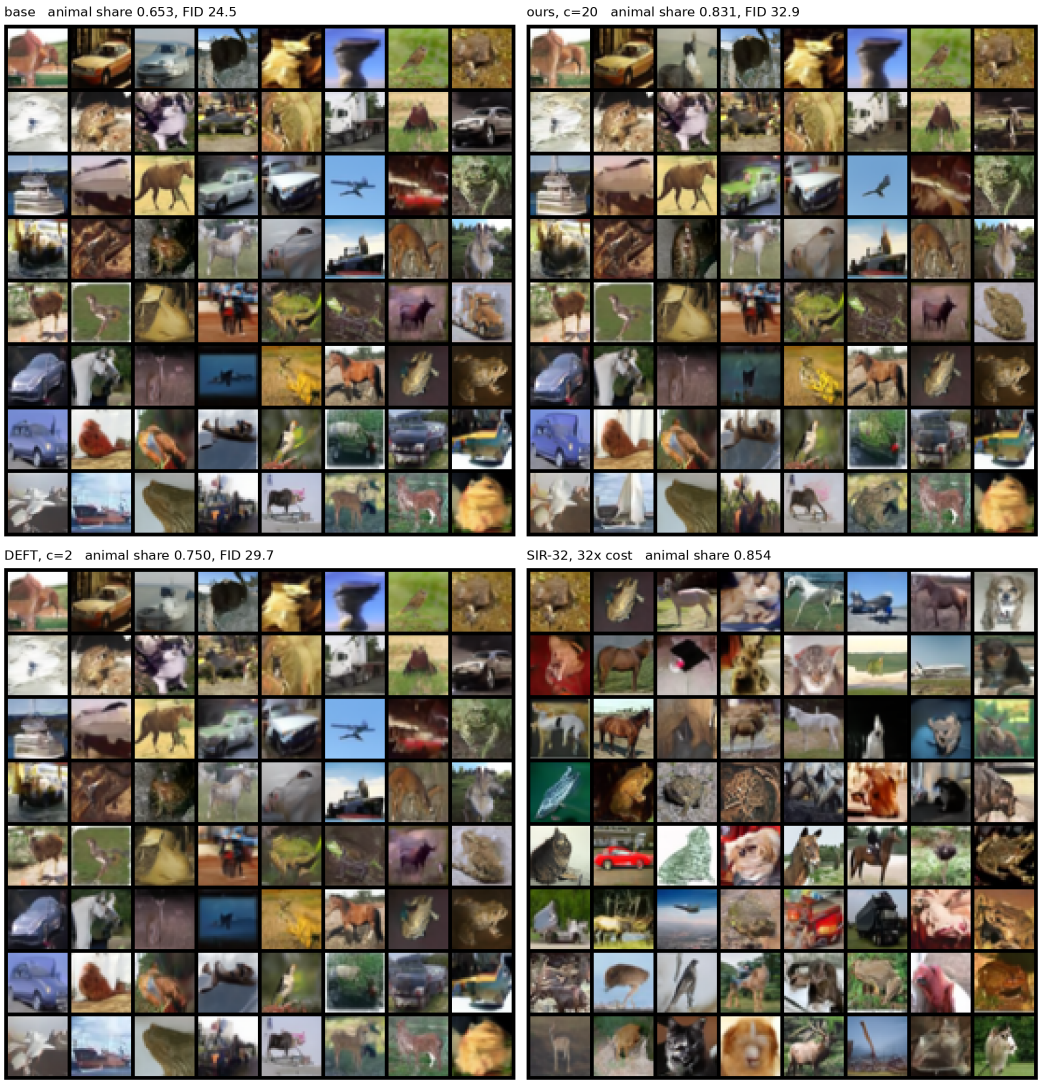


Figure 4: CIFAR-10, animals, the first 64 samples of seed 0, uncurated. Base and ours at $c = 20$ above, DEFT and SIR-32 below, each panel labelled with the share of animals it carries and with its FID. DEFT is drawn at $c = 2$, the scale of its sweep whose FID is closest to the one our own panel is drawn at, because a matched departure from the base distribution is the reading under which the two frontiers are compared, at a matched c it would instead sit at FID 98. The guided panels share the initial noise with the base panel, so a sample can be compared with its counterpart in the same position, SIR-32 does not, by construction.

returning the multiple of the gradient, initialised at one, Adam (Kingma & Ba, 2015), learning rate $2 \cdot 10^{-4}$, batch 128, 8000 steps, gradient norm clipped at 5.

- *DOIT*: $M = 32$ antithetic look-ahead samples at $\eta = 0.4$, correction over steps $1 \dots \lfloor 0.67L \rfloor$ of the $L = 100$, $h(\cdot, 0) \propto \exp(r/\tau)$ with τ swept, these are the reference implementation's values.

- *Sampling*: 100 DDIM steps at $\eta = 0$, and at $\eta = 0.4$ where a second regime is reported, correction capped at 10% of $\|\hat{\epsilon}\|$ for every guided method, $c \in \{0.25, 0.5, 1, 2, 3, 5, 10, 20\}$ on CIFAR-10 and $\{1, 5, 20\}$ on MNIST, 3 seeds of 2000 samples.

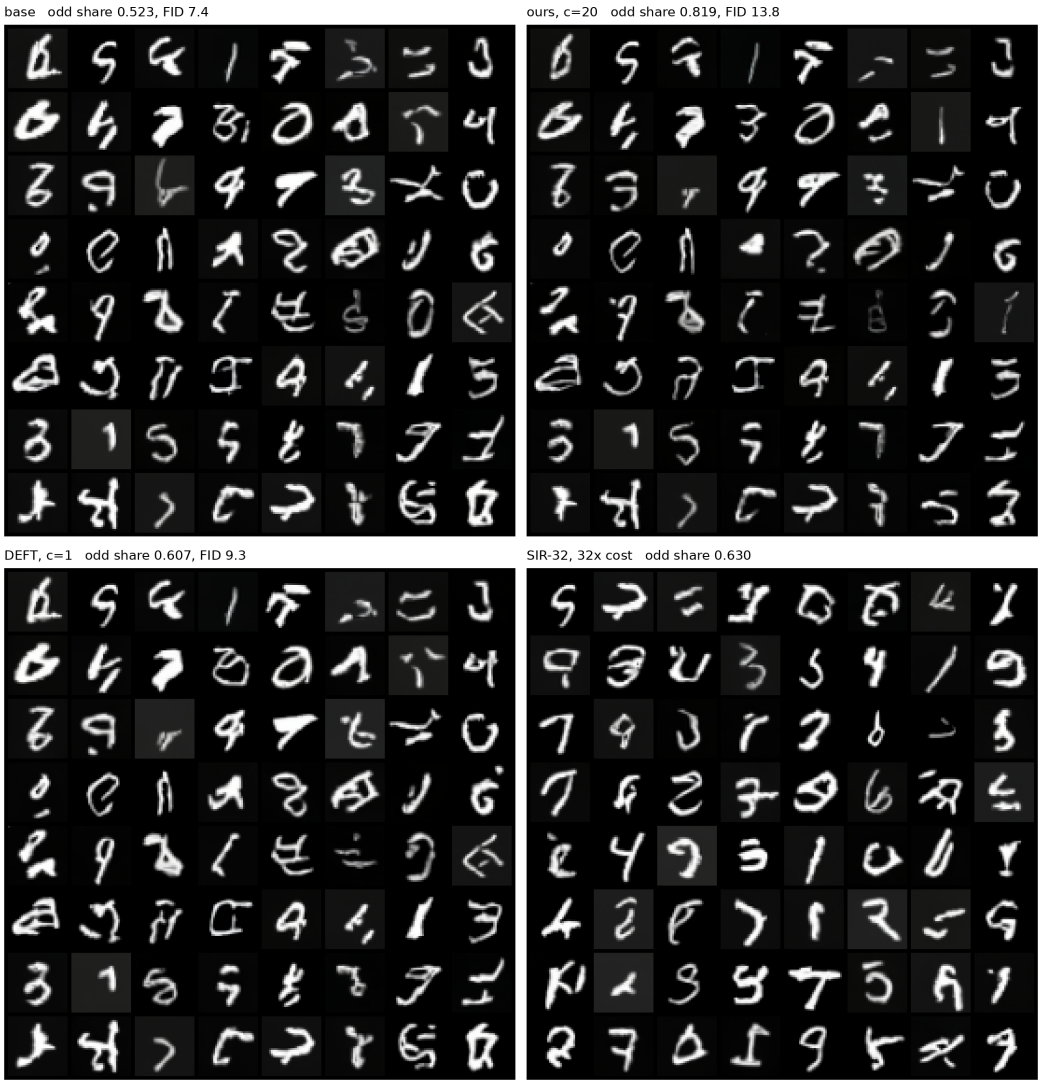


Figure 5: MNIST, odd digits, the first 64 samples of seed 0, uncurated. Base and ours at $c = 20$ above, DEFT at $c = 1$ and SIR-32 below, each panel labelled with the share of odd digits and with its FID, as in Figure 4, DEFT is drawn at the scale of its sweep whose FID is closest to ours. The guided panels share the initial noise with the base panel. The base model draws some digits mirrored.

J.8 What the guidance cap does

Every guided method runs under one shared constraint, a correction capped at 10% of $\|\hat{\epsilon}\|$ per sample and step. A constraint that binds unevenly would shape the comparison, so each run records how hard it works.

The cap does different work for each. Ours asks for less than a third of it on average and removing it costs nine points of share and ten of FID, DEFT asks for more than the cap itself at $c = 20$ and is clipped on four fifths of its steps, for DOIT the cap is the only thing between the run and divergence, and at $\gamma = 20$ the uncapped run collapses to FID 719 with no animals left. What matters for the paper is that the comparison does not rest on it: the two frontiers meet between FID 28 and 30, which is DEFT's $c \in \{1, 2\}$ and our $c \in \{10, 20\}$, and there the cap binds on none of DEFT's steps and at most a twentieth of ours.

Table 7: The cap at $c = 20$ on CIFAR-10 ($\gamma = 1$ for DOIT, and $\gamma = 0.1$, the small-γ setting of its authors), where it binds at all: the mean uncapped correction against the cap of 0.1, the share of guided steps it binds on, and what removing it does. At $c = 5$ and below it binds on at most 1% of the steps of either trained method. The uncapped point was not sampled for the reported DEFT head, nor was that head run under the stochastic sampler, so those cells are empty.

sampler	method	$\overline{\lvert\delta\rvert}$	binds	capped → uncapped
$\eta = 0$	ours	0.031	5.3%	$0.831/32.9 \to 0.739/43.4$
	DEFT	0.158	77.9%	$0.941/97.7 \to$ —
	DOIT	2.4	100%	$0.689/116.7 \to 0.995/463.2$
$\eta = 0.4$	ours	0.031	5.0%	$0.854/34.6 \to 0.805/46.2$
	DOIT, $\gamma = 1$	2.3	100%	$0.464/70.2 \to 0.996/437.1$
	DOIT, $\gamma = 0.1$	0.23	95.2%	$0.462/70.0 \to 0.889/221.6$

J.9 Approximation floors at the operative smoothness

Table 1 claims the strong residual carries a first-order approximation floor, non-informative at $\beta = 2$. A floor belongs to the class, not the estimator, so it is measured directly, at the member of the class approximating h^* in value, gradient and time derivative at once. Orders 3, 2, 1 give $\beta = 4$, 3, 2, order 1, being only C^0, is not an admissible trial class, but its floor is defined.

Table 8: Approximation floors against resolution, $d = 1$, 4 replicates, slopes over $N \in [2, 24]$, predictions $-2(\beta - 1)$ and $-2(\beta - 2)$.

β (order)	weak floor	strong floor	strong / weak at $N = 24$
4 (3)	$N^{-1.44}$	$N^{-1.89}$	7.4
3 (2)	$N^{-1.81}$	$N^{-1.00}$	311
2 (1)	$N^{-2.46}$	$\mathbf{N^{+0.29}}$	8500

At $\beta = 2$ the strong floor stops decaying while the weak one keeps falling, the gap widening from 7 through 311 to 8500. The fitted slopes are not the predicted ones, the grid binding before the smoothness does at $\beta = 4$ and 3, and h^* being smoother than the $W^{2,\infty}$ the prediction assumes at $\beta = 2$, the table supports the ordering and the flattening, not the exponents.

J.10 The finite-s regime

The rate experiment runs at the boundary case $\mathrm{s} = \infty$, where the clipping never binds. What Assumption 1 buys beyond it is the factor $\Xi_{\mathrm{s}}^{4/(\mathrm{s}+4)} \varepsilon_n^{\mathrm{s}/(\mathrm{s}+4)}$, so what is worth testing is whether the error degrades where that envelope says it should. A fraction 0.8 of the atoms of Appendix J.4 is therefore given a weight w_{low} descending the ladder $1, 0.3, 0.1, 0.03, 3 \cdot 10^{-3}$, the rest keeping 3, with class, sample sizes, oracle grids and four replicates unchanged. The window is not: on $I = [0.40, 0.85]$ the noise is comparable to the radius of the data, the smoothing averages the low-weight region away, and h^* never falls below b, however small w_{low} is. The ladder is read on $I = [0.96, 0.995]$ instead, with $r = 2$ and $N \le 32$ for the smaller σ, b is the level the theorem prescribes at the minimising s, exact here because p_0 is atomic.

Table 9: The finite-s ladder, $d = 1$, 4 replicates, each cell at its own oracle (N, μ), slopes over the seven sample sizes, "clip" the mass on which $h^* < \mathrm{b}$ at $n = 10^6$.

$\min w$	$n = 10^3$	$n = 10^6$	measured	envelope	clip
0.5 (base)	3.0×10^{-3}	1.0×10^{-5}	0.787	0.779	0.00
1	1.2×10^{-2}	2.2×10^{-5}	0.919	0.779	0.00
0.3	2.6×10^{-2}	2.0×10^{-5}	1.014	0.779	0.00
0.1	4.3×10^{-2}	2.2×10^{-5}	1.076	0.337	0.00
0.03	1.5×10^{-1}	2.9×10^{-3}	0.547	0.004	0.40
3×10^{-3}	2.6×10^{-1}	$\mathbf{2.1 \times 10^{-1}}$	**0.028**	0.002	0.44

The clipping mass separates the two regimes. Down to $\min w = 0.1$ it does not bind at $n = 10^6$ and the error there is the reference rung's 2×10^{-5}: a small w on a set the smoothing averages away

costs a constant, not a rate. At 0.03 it binds at every sample size, the guaranteed exponent collapses and the error is two orders of magnitude worse, at $3 \cdot 10^{-3}$ there is no convergence at all over three decades of n. The envelope being an upper bound, the measured slope need only be at least as steep, which it is at every rung, what the table shows is that the bound goes flat exactly where the estimator stops converging.