SHARP BOUNDS FOR CONDITIONAL-VARIANCE ESTIMATION IN DIFFUSION REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the conditional variance $v_0(x) = \mathrm{Var}(Y \mid X = x)$ from M conditional diffusion samples involves model-law mismatch and Monte Carlo error. Small Kullback–Leibler divergence alone cannot control variance, even for smooth Gaussian mixtures. To limit the influence of extreme model samples, we analyze a clipped pairwise estimator with exact variance correction for early-stopped Ornstein–Uhlenbeck diffusion. Under a target pairwise q-moment bound and a target-to-model conditional KL bound, we establish matching upper and lower bounds for estimators using only model samples. For scalar covariates, we also prove a new-covariate risk bound for common-radius selection using independent validation data. We evaluate the estimator through counterexample, controlled, and learned-model experiments. The experiments distinguish clipping gains from validation selection loss and show diminishing improvements as the Monte Carlo sample size increases.

1 INTRODUCTION

Estimating conditional variance with conditional diffusion regression is a multistage problem involving score learning, numerical sampling, and finitely many model samples. The target $v_0(x) = \mathrm{Var}(Y \mid X = x)$ describes how predictive dispersion varies with the covariate. Classical heteroskedastic neural networks estimate the conditional mean and variance within a prescribed response family (Nix & Weigend, 1994; Seitzer et al., 2022). Heteroscedastic Gaussian processes and neural variance estimators address related estimation and optimization problems (Le et al., 2005; Skafte et al., 2019; Zhang et al., 2023; Wong-Toi et al., 2024). Generalized additive distributional models and natural-gradient boosting model covariate-dependent parameters of flexible response families, while distributional random forests estimate conditional laws through data-adaptive weights (Rigby & Stasinopoulos, 2005; Duan et al., 2020; Cevid et al., 2022). In conditional diffusion regression, the conditional variance is not directly parameterized but is recovered from the learned conditional law.

Conditional diffusion regression learns flexible predictive laws, with generated responses used to summarize predictive uncertainty. CARD learns conditional response distributions using a denoising diffusion model, while subsequent work develops tree-based score estimators, diffusion boosting, and alternative noise parameterizations for probabilistic regression (Han et al., 2022; Beltran Velez et al., 2024; Han & Zhou, 2024; Zhao et al., 2025; Kneissl et al., 2025). Diffusion ensembles have also been evaluated for zero-shot uncertainty assessment (Shu & Farimani, 2024). Existing theory provides CARD-specific Wasserstein bounds, statistical rates for general conditional diffusion models, and asymptotic inference for conditional means based on model samples (Li, 2025; Fu et al., 2024; Tang et al., 2025; Jiao et al., 2025). General diffusion-sampling and distribution-estimation guarantees provide complementary law-level control (Lee et al., 2023; Benton et al., 2024; Oko et al., 2023). Separate work studies solver-dependent conditional-variance propagation for a fixed learned score (Song & Jeon, 2026). We study how accurately $v_0(x)$ can be recovered from a learned conditional law using finitely many model samples.

Target-to-model conditional KL divergence alone cannot control the unbounded variance functional without additional tail control, even for smooth Gaussian mixtures. We study estimation of the target variance from model samples under a pairwise q-moment bound on the target alone. We analyze the pairwise clipping estimator of Ke et al. (2019) under target-to-model KL mismatch

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.

and apply exact OU variance inversion at the early-stopped endpoint. Under common pointwise moment and KL budgets, the RMSE for $2 < q \leq 4$ is bounded by $C_q e^{\delta} B^{2/q}(\kappa + M^{-1})^{(q-2)/q}$. The minimax lower bound for estimators using only model samples matches this dependence on B, κ, and M, up to constants depending only on q. No moment bound on the model law is required. The analysis combines a Hellinger–KL comparison, the pairwise variance identity, and clipped U-statistic arguments (Sason & Verdú, 2016; Hoeffding, 1948; Minsker & Wei, 2019; Joly & Lugosi, 2016; Bubeck et al., 2013). Under a high-probability score-risk budget, diffusion-sampling bounds (Chen et al., 2023; Benton et al., 2024) yield a target-to-model conditional KL bound averaged over covariates. Combining this bound with an average pairwise q-moment bound on the target gives an integrated L^1 guarantee for variance estimation.

The radius prescribed by the error bound depends on moment and KL bounds that are generally unknown. We therefore select a common radius R over a continuous range using independent validation data and a pairwise difference criterion that avoids conditional-mean estimation (Wang et al., 2008). For scalar covariates, we establish a new-covariate risk bound and a vanishing expected selection-loss bound along a fixed clipping path, conditional on the fitted model and realized model samples.

Controlled experiments examine the dependence on κ and M. With the target and model laws held fixed, the benefit of additional validation data depends on whether clipping can correct the variance distortion. In a learned heteroscedastic-mixture experiment, common-radius clipping reduces RMSE relative to the unclipped estimator. Repeated validation with fitted models and model samples fixed identifies when selection loss outweighs the available clipping gain and isolates validation-response noise as a major source of that loss. Appendix D gives the protocols and additional results.

2 Conditional Diffusion and the Variance Estimator

Let $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ be i.i.d., where Y is a scalar response and X takes values in a standard Borel space. Let U_n be fitting randomness independent of $\mathcal{D}_n$, and define the fitting sigma-field $\mathscr{G}_n = \sigma(\mathcal{D}_n, U_n)$. Set $Y_0 = Y$ and let $x \mapsto P_0^x$ be a regular conditional law of Y_0 given $X = x$. Unless stated otherwise, statements at a fixed covariate are understood to hold for P_X-almost every x.

2.1 Conditional OU Model

In conditional diffusion regression, the forward diffusion acts only on the scalar response, while the covariate is held fixed. Let $(W_t)_{t\geq 0}$ be a standard Brownian motion independent of (X, Y_0). We use the variance-preserving Ornstein–Uhlenbeck process, whose solution at each fixed t admits the representation

$$\mathrm{d}Y_t = -\tfrac{1}{2} Y_t \,\mathrm{d}t + \mathrm{d}W_t, \qquad Y_0 = Y, \qquad Y_t = \alpha_t Y_0 + \sigma_t \varepsilon, \qquad \alpha_t = e^{-t/2}, \quad \sigma_t^2 = 1 - e^{-t}, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is independent of (X, Y_0). The process has invariant law $\mathcal{N}(0,1)$. Write $P_t^x = \mathrm{Law}(Y_t \mid X = x)$ and let $p_t(\cdot \mid x)$ denote its density. Since $\alpha_t \to 0$ and $\sigma_t^2 \to 1$, P_t^x converges weakly to $\mathcal{N}(0,1)$ as $t \to \infty$. For every $t > 0$, Gaussian smoothing makes $p_t(\cdot \mid x)$ strictly positive and smooth. Its conditional score is $s^\star(t, y, x) = \partial_y \log p_t(y \mid x)$.

For $0 < \delta < T$, let τ denote reverse time, so that the corresponding forward time is $T - \tau$. The exact reverse process uses the conditional score $s^\star$ and is initialized from the true terminal law P_T^x. It satisfies

$$\mathrm{d}Z_\tau = \{\tfrac{1}{2} Z_\tau + s^\star(T - \tau, Z_\tau, x)\}\mathrm{d}\tau + \mathrm{d}B_\tau, \qquad Z_0 \sim P_T^x, \qquad 0 \leq \tau \leq T - \delta, \tag{2}$$

where $(B_\tau)_{\tau \geq 0}$ is a standard Brownian motion for the reverse dynamics. Its marginal at reverse time τ is $P_{T-\tau}^x$ (Anderson, 1982; Haussmann & Pardoux, 1986; Song et al., 2021). In particular, $Z_{T-\delta} \sim P_\delta^x$.

Define $v_t(x) = \mathrm{Var}(Y_t \mid X = x)$. The OU transition in (1) gives the exact variance identities

$$v_t(x) = e^{-t} v_0(x) + 1 - e^{-t}, \qquad v_0(x) = e^t v_t(x) - (e^t - 1). \tag{3}$$

Thus stopping at $\delta > 0$ avoids the smallest-noise regime, while the exact inversion removes OU-smoothing bias from the variance functional at the population level. Any error in estimating the endpoint variance $v_\delta(x)$ is multiplied by e^δ after inversion. This correction does not invert the full conditional law. The exact reverse endpoint remains P_δ^x rather than P_0^x.

2.2 Score Learning and ExpInt Sampling

Simulating the reverse process in (2) requires the conditional score $s^\star(t, y, x)$, which is unknown because $p_t(\cdot \mid x)$ is unknown. We estimate it by conditional denoising score matching (Hyvärinen, 2005; Vincent, 2011).

Let $\delta = t_0 < \cdots < t_N = T$, $h_k = t_k - t_{k-1}$, and $\omega_k = h_k/(T - \delta)$. For a score s, define

$$\overline{\mathcal{R}}_{N,T,\delta}(s) = \sum_{k=1}^{N} \omega_k \mathbb{E}_X \mathbb{E}_{P^X_{t_k}} \left| s(t_k, Y, X) - s^\star(t_k, Y, X) \right|^2 . \tag{4}$$

The OU representation in (1) yields the grid-based population DSM objective

$$\overline{\mathcal{L}}_{\mathrm{DSM}}(s) = \mathbb{E}_{X,Y_0} \sum_{k=1}^{N} \omega_k \mathbb{E}_\varepsilon \left| s(t_k, \alpha_{t_k} Y_0 + \sigma_{t_k} \varepsilon, X) + \frac{\varepsilon}{\sigma_{t_k}} \right|^2 . \tag{5}$$

Because $s^\star(t, Y_t, X) = \mathbb{E}[-\varepsilon/\sigma_t \mid Y_t, X]$, the population excess DSM risk equals $\overline{\mathcal{R}}_{N,T,\delta}(s)$. The fitted score is denoted by $\widehat{s}_n$. Appendix A.1 gives the projection identity and a sufficient training-risk condition.

The exponential integrator (ExpInt) traverses the grid from T to δ, integrates the linear OU drift exactly, and freezes the score at the beginning of each interval (Zhang & Chen, 2023; De Bortoli, 2022; Chen et al., 2023; Benton et al., 2024). For $k = N, \ldots, 1$, its update is

$$\widehat{Y}_{t_{k-1}} = e^{h_k/2} \widehat{Y}_{t_k} + 2(e^{h_k/2} - 1) \widehat{s}_n(t_k, \widehat{Y}_{t_k}, x) + \sqrt{e^{h_k} - 1}\, \xi_k, \qquad \xi_k \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1). \tag{6}$$

Starting from $\widehat{Y}_T \sim \pi_\infty = \mathcal{N}(0, 1)$, with all Monte Carlo randomness independent of $\mathscr{G}_n$, this recursion defines the conditional endpoint law $Q^x_{\delta,n,N,T}$.

2.3 Target and the Clipped Pairwise Estimator

Following Ke et al. (2019), we estimate variance from clipped pairwise differences, an approach developed for covariance estimation under heavy tails. Pairwise differences avoid separate mean estimation, while clipping limits the contribution of extreme model samples. We combine this construction with the exact OU variance inversion in (3).

For $R > 0$ and $M \geq 2$, draw $\widehat{Y}_1, \ldots, \widehat{Y}_M$ independently from $Q^x_{\delta,n,N,T}$ conditionally on $(x, \mathscr{G}_n)$ and set

$$\psi_R(y, y') = \tfrac{1}{2} \min\{(y - y')^2, R^2\}, \tag{7}$$

$$\widehat{C}_{R,M}(x) = \binom{M}{2}^{-1} \sum_{i<j} \psi_R(\widehat{Y}_i, \widehat{Y}_j), \tag{8}$$

$$\widehat{v}_{0,R,M}(x) = [e^\delta \widehat{C}_{R,M}(x) - (e^\delta - 1)]_+, \tag{9}$$

where R is the clipping radius, M is the Monte Carlo sample size at each covariate, and $[a]_+ = \max\{a, 0\}$. At $R = \infty$, $\widehat{C}_{\infty,M}(x)$ is the sample variance.

3 From Score Risk to Conditional KL

Score risk, time discretization, and terminal initialization yield an average target-to-model conditional KL bound. This bound alone, however, does not control the unbounded variance functional.

3.1 Assumptions and the ExpInt Grid

Let

$$m_2(x) = \mathbb{E}[Y_0^2 \mid X = x], \qquad m_2 = \mathbb{E}_X m_2(X) = \mathbb{E}[Y_0^2].$$

Assumption 1 (Second moment)**.** $m_2 < \infty$.

Assumption 2 (Score-risk budget). *For each $0 < \zeta < 1$, let $r_{n,N,T,\delta}(\zeta) < \infty$ be a deterministic score-risk budget. The fitted score satisfies*

$$\mathbb{P}\{\overline{\mathcal{R}}_{N,T,\delta}(\widehat{s}_n) \le r_{n,N,T,\delta}(\zeta)\} \ge 1 - \zeta,$$

where the probability is over $\mathcal{D}_n$ and U_n.

Assumption 3 (Reverse-process admissibility). *Conditional on $\mathscr{G}_n$, for P_X-almost every x, the exact reverse process and continuous ExpInt interpolations have nonexplosive weak solutions that are unique in law. For comparisons with the same initial law, the drift difference is progressively measurable and has locally finite quadratic energy under both path laws. The stopped uniqueness and concatenation conditions for the localized Girsanov comparison are stated in Appendix A.2.*

Lemma 1 (Admissible ExpInt grid). *Let $0 < \delta \le 1$, $T \ge 2$, and $L_{\delta,T} = T + \log(1/\delta)$. For every even integer $N \ge 2L_{\delta,T}$, there exists an ExpInt grid satisfying*

$$\rho_N = \max_k \frac{h_k}{\min\{1, t_{k-1}\}} \le \frac{4L_{\delta,T}}{N}. \tag{10}$$

Proof in Appendix A.3, following Benton et al. (2024, Corollary 1 and Appendix D).

3.2 Average Conditional KL Bound

Conditional on $\mathscr{G}_n$, define

$$\overline{K}_{n,N,T,\delta} = \mathbb{E}_X \, \mathrm{KL}\big(P_\delta^X \| Q_{\delta,n,N,T}^X\big), \tag{11}$$

where $\mathbb{E}_X$ averages over an independent $X \sim P_X$.

Proposition 1 (Average conditional KL bound). *Under Assumptions 1 and 3, let $0 < \delta \le 1$ and $T \ge 2$, and use an ExpInt grid satisfying* (10)*. There exists a universal constant $C_{\mathrm{KL}} > 0$ such that, almost surely over $(\mathcal{D}_n, U_n)$,*

$$\overline{K}_{n,N,T,\delta} \le C_{\mathrm{KL}} \left[(T - \delta)\overline{\mathcal{R}}_{N,T,\delta}(\widehat{s}_n) + \frac{L_{\delta,T}^2}{N} + (1 + m_2)e^{-T} \right]. \tag{12}$$

Proof in Appendix A.4.

Terms in the KL bound. The three terms in (12) correspond to score risk, time discretization, and terminal initialization. For fixed N and δ, increasing T decreases the initialization term but increases the discretization term. The score-risk contribution may also vary with T. Appendix A.3 derives the reverse-clock score-process identity and adapts the sampling arguments of Chen et al. (2023); Benton et al. (2024).

3.3 Why Conditional KL Does Not Control Variance

The following counterexample keeps the target fixed and varies only the model law.

Proposition 2 (KL does not control OU-corrected variance). *For every $\delta > 0$, let $P_\delta = \mathcal{N}(0, 1)$. There exists a family of model endpoint laws $\{Q_{\varepsilon,\delta}\}_{0<\varepsilon<1}$ with smooth positive densities, each obtained from a smooth centered time-zero density, such that, as $\varepsilon \downarrow 0$,*

$$\mathrm{KL}(P_\delta \| Q_{\varepsilon,\delta}) \to 0 \quad \textit{but} \quad |\mathcal{T}_\delta(Q_{\varepsilon,\delta}) - \mathcal{T}_\delta(P_\delta)| \to \infty,$$

where $\mathcal{T}_\delta(\nu) = e^\delta \operatorname{Var}_\nu(Z) - (e^\delta - 1)$.

Proof in Appendix A.5.

Vanishing model-law mass at increasingly distant locations causes the variance error to diverge.

4 Conditional-Variance Estimation and Radius Selection

Pairwise clipping limits the influence of extreme model samples, while a target pairwise q-moment bound controls clipping bias. We first establish sharp estimation bounds from model samples alone, then bound the additional error from validation-based selection of a common radius.

4.1 Sharp Estimation Bounds from Model Samples

Condition on $\mathscr{G}_n$ and fix a covariate x. Write $P = P_\delta^x$, $Q = Q_{\delta,n,N,T}^x$, and $Z, Z' \overset{\text{iid}}{\sim} P$. The following bound requires a pairwise moment under P and no moment condition on Q.

Theorem 1 (RMSE upper bound from model samples)**.** *Let $2 < q \le 4$, $\mathrm{KL}(P\|Q) \le \kappa$, and $\mathbb{E}_{P^{\otimes 2}}|Z - Z'|^q \le B$. For $M \ge 2$ and $R > 0$,*

$$\begin{aligned}&\left(\mathbb{E}_{Q^{\otimes M}}|\widehat{v}_{0,R,M}(x) - v_0(x)|^2\right)^{1/2}\\ &\quad\le C_q e^{\delta}\left[BR^{2-q} + \sqrt{B(\kappa + M^{-1})}\,R^{2-q/2} + R^2(\kappa + M^{-1})\right].\end{aligned} \tag{13}$$

If $B > 0$, the bound-balancing radius $R_{\mathrm{bd}} = (B/(\kappa + M^{-1}))^{1/q}$ gives

$$\left(\mathbb{E}_{Q^{\otimes M}}|\widehat{v}_{0,R_{\mathrm{bd}},M}(x) - v_0(x)|^2\right)^{1/2} \le C_q e^{\delta} B^{2/q}(\kappa + M^{-1})^{(q-2)/q}, \tag{14}$$

where the finite constant $C_q > 0$ depends only on q. If the same B, κ hold uniformly in x, both bounds also hold after averaging squared error over an independent $X \sim P_X$.

Proof in Appendix B.1.

The radius R_{bd} balances the bound terms and need not minimize risk for a particular (P, Q). To assess whether the rate can be improved using only model samples, let P_0 be a time-zero law and P_δ its OU endpoint law. For $Z, Z' \overset{\text{iid}}{\sim} P_\delta$ and $\xi, \xi' \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, define

$$\mathfrak{P}_{\delta,q}(B) = \{P_0 \mid \mathbb{E}|Z - Z'|^q \le B\}, \qquad b_{\delta,q}^{(0)} = \mathbb{E}|\sqrt{1 - e^{-\delta}}(\xi - \xi')|^q, \tag{15}$$

where $b_{\delta,q}^{(0)}$ is the pairwise q-moment of the OU noise and enters the moment condition of Theorem 2.

Theorem 2 (Minimax RMSE lower bound from model samples)**.** *Fix $2 < q \le 4$, $\delta > 0$, $M \ge 2$, $0 \le \kappa \le 1/4$, and $B \ge 2^q b_{\delta,q}^{(0)}$. Define*

$$\mathfrak{C}_{\delta,q}(B,\kappa) = \{(P_0, Q) \mid P_0 \in \mathfrak{P}_{\delta,q}(B), \mathrm{KL}(P_\delta\|Q) \le \kappa\}.$$

There is a constant $c_q > 0$, depending only on q, such that

$$\inf_{\widetilde{v}} \sup_{(P_0,Q)\in\mathfrak{C}_{\delta,q}(B,\kappa)} \left\{\mathbb{E}_{Q^{\otimes M}}|\widetilde{v}(\widehat{Y}_1, \dots, \widehat{Y}_M) - \mathrm{Var}_{P_0}(Y_0)|^2\right\}^{1/2} \ge c_q e^{\delta} B^{2/q}(\kappa + M^{-1})^{(q-2)/q}. \tag{16}$$

The infimum is over measurable estimators observing only the M model samples, without labeled training or validation data.

Proof in Appendix B.3.

Over the stated class, the matching bounds separate finite-M error from a worst-case mismatch floor that persists as $M \to \infty$ for fixed $\kappa > 0$.

Under the score-risk budget in Assumption 2 and an average target pairwise q-moment bound, Corollary 1 converts the average conditional KL bound into an integrated L^1 guarantee. The integrated RMSE extension of Theorem 1 requires common pointwise moment and KL budgets.

4.2 Common-Radius Selection

To select a common radius without estimating the target moment and conditional KL budgets needed for R_{bd}, we use independent observations from the target law. Let $\mathcal{V}_B$ contain $n_{\mathrm{val}} = 2m$ i.i.d. observations from the law of (X, Y_0), independent of $\mathscr{G}_n$ and Monte Carlo randomness.

Fix ordered anchors $x_1 < \cdots < x_K$ in a compact interval $I \subset \mathbb{R}$. At each anchor, reuse $M \ge 2$ model samples $(\widehat{Y}_{k,a})_{a=1}^M$, conditionally i.i.d. from $Q_{\delta,n,N,T}^{x_k}$ given $\mathscr{G}_n$ and independent of $\mathcal{V}_B$. Parameterize the common radius $R \ge 0$ by $\lambda = R^2$. For nonnegative linear-interpolation weights w_k summing to one, with constant boundary extension, define

$$f_\lambda(x) = \sum_{k=1}^K w_k(x)\,\Pi_{[\ell, V_{\max}]}\widehat{v}_{0,\sqrt{\lambda},M}(x_k), \qquad \lambda \ge 0, \tag{17}$$

where $\Pi_{[\ell,V_{\max}]}$ denotes projection onto $[\ell, V_{\max}]$, $0 \le \ell < V_{\max} < \infty$, and $\widehat{v}_{0,0,M} = 0$ by continuity.

Pairwise response differences provide a validation criterion without a separate conditional-mean fit (Wang et al., 2008). For $m \ge 1$, sort the $2m$ covariates in $\mathcal{V}_B$ and form disjoint adjacent pairs $(X_i^-, Y_i^-), (X_i^+, Y_i^+)$. Set $z_i = (X_i^- + X_i^+)/2$ and $U_i = (Y_i^- - Y_i^+)^2/2$. Choose a measurable $\widehat{\lambda}$ satisfying

$$\begin{aligned} \widehat{\mathcal{L}}_{\mathrm{val}}(\lambda) &= m^{-1} \sum_i \{ f_\lambda(z_i)^2 - 2U_i f_\lambda(z_i) \}, \\ \widehat{\mathcal{L}}_{\mathrm{val}}(\widehat{\lambda}) &\le \inf_{\lambda \ge 0} \widehat{\mathcal{L}}_{\mathrm{val}}(\lambda) + \eta_{\mathrm{opt}}, \end{aligned} \tag{18}$$

for deterministic $\eta_{\mathrm{opt}} \ge 0$. The selected predictor is $f_{\widehat{\lambda}}$, with common radius $\widehat{R} = \sqrt{\widehat{\lambda}}$.

Theorem 3 (New-covariate validation guarantee)**.** *Let $\mu_0(x) = \mathbb{E}[Y_0 \mid X = x]$. Suppose $X \in I$ almost surely and fix versions of μ_0, v_0 with bounded variation on I and $0 \le v_0 \le V_{\max}$. For $2 < q \le 4$ and $p = q/2$, assume $\mathbb{E}[|U_i - \bar{U}_i|^p \mid \mathbf{X}^{\mathrm{val}}] \le S_p < \infty$ for all i, where S_p is deterministic, $\mathbf{X}^{\mathrm{val}}$ collects all selection covariates, and $\bar{U}_i = \mathbb{E}[U_i \mid \mathbf{X}^{\mathrm{val}}]$. Set $\mathcal{E}^\star_{\mathrm{cont}} = \inf_{\lambda \ge 0} \|f_\lambda - v_0\|_{L^2(P_X)}$. Conditionally on $\mathscr{G}_n$ and the realized model samples,*

$$\begin{aligned} &\left\{ \mathbb{E}_{\mathrm{val}} \| f_{\widehat{\lambda}} - v_0 \|^2_{L^2(P_X)} \right\}^{1/2} \\ &\quad \le \sqrt{2}\, \mathcal{E}^\star_{\mathrm{cont}} + C_1 \sqrt{\frac{\log(2Km+4)}{m}} + C_2 m^{-(q-2)/4} + \sqrt{3\eta_{\mathrm{opt}}/2}, \end{aligned} \tag{19}$$

where $\mathbb{E}_{\mathrm{val}}$ averages over $\mathcal{V}_B$, and C_1, C_2 depend only on $K, V_{\max}, S_p, q$ and the total variations of μ_0, v_0.

Proof in Appendix C.2.

The comparator $\mathcal{E}^\star_{\mathrm{cont}}$ is the minimum error along the realized clipping path, which additional validation data cannot reduce. For fixed K and $q = 4$, the validation and optimization terms are $O(\sqrt{\log m/m} + \sqrt{\eta_{\mathrm{opt}}})$. Under the anchor conditions of Corollary 3, the comparator's root mean square over model samples is bounded by the rate in Theorem 1 plus interpolation error.

Proposition 5 further bounds expected excess MSE over the best member of the fixed path. It tends to zero for fixed K as $m \to \infty$ and $\eta_{\mathrm{opt}} \to 0$.

5 Numerical Study

We examine the KL–variance gap, the (M, κ) dependence in Theorems 1 and 2, and the approximation and selection errors in Theorem 3. Controlled laws isolate these effects, while learned-model experiments assess clipping gains and the loss from validation-based radius selection. Appendix D.1 gives the shared evaluation and training settings, and Appendix D.4 defines the evaluation metrics.

5.1 Conditional KL and Variance Distortion

Rare, large responses can substantially change variance while keeping target-to-model KL small. We take $X \sim \mathrm{Unif}[-1, 1]$ and $P_0^x = \mathcal{N}(0, \varsigma_0(x)^2)$, where $\varsigma_0(x) = 1 + 0.5 \sin(3\pi x)$. To test clipping under this tail perturbation, we train noise-prediction networks on the mixture from Proposition 2,

$$Q_{0,w}^{\mathrm{mix},x} = (1-w) P_0^x + \frac{w}{2} \sum_{u \in \{-1,1\}} \mathcal{N}(u\varsigma_0(x)/w, \varsigma_0(x)^2).$$

For these specified laws, $\mathrm{KL}(P_0^x \| Q_{0,w}^{\mathrm{mix},x}) \le -\log(1-w)$, whereas their variance difference is $\varsigma_0(x)^2/w$. The bound also holds after OU smoothing of the specified mixture, but does not certify the fitted sampler's KL. The model is trained to approximate this mixture, while the estimation target remains $v_0(x)$. Independent observations from P_0^x select the common radius. For each $w \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$, we fit five models. Clipped and unclipped estimates share the same model samples.

Figure 1 shows that common-radius clipping reduces variance RMSE at small estimated conditional KL. The fitted model captures little rare-tail mass at the smallest weight, so this sweep does not reproduce the counterexample's variance divergence (Appendix D.3).

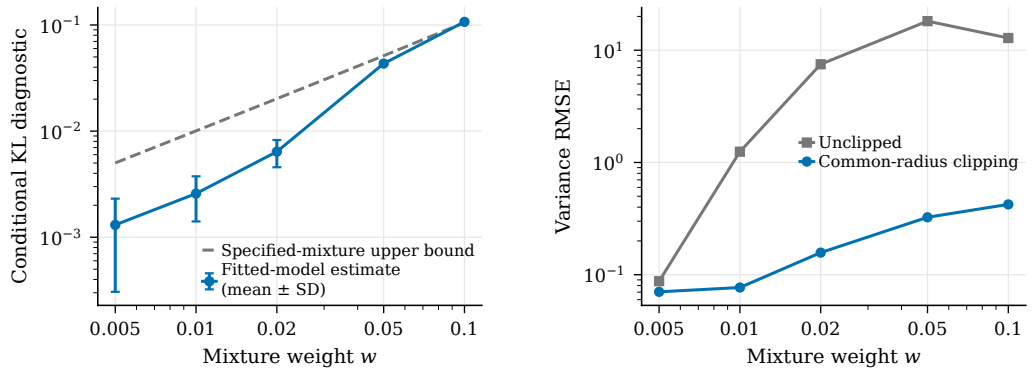


Figure 1: Conditional KL and variance estimation error. Five training repetitions per mixture weight w. Left: estimated target-to-model conditional endpoint KL (mean ± 1 SD across models) and the specified-mixture KL upper bound. Right: variance RMSE for unclipped and common-radius estimates across mixture weights (log scale).

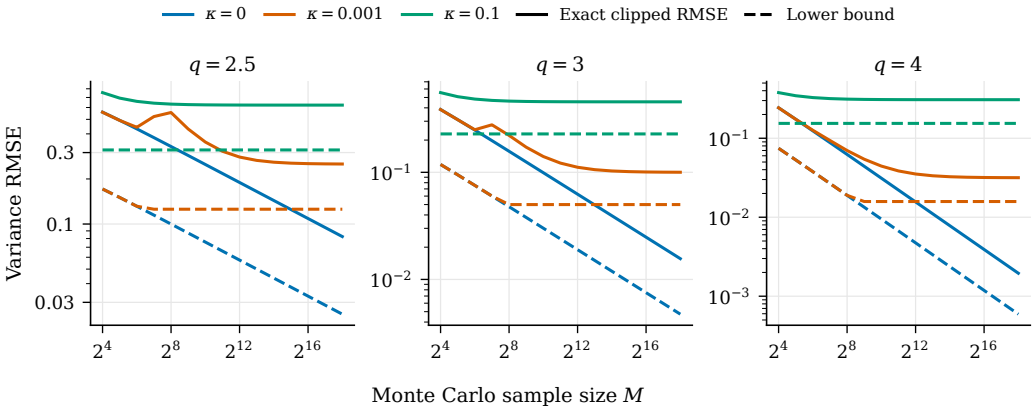


Figure 2: Variance estimation rates and lower bounds. Exact variance RMSE with $R = R_{\mathrm{bd}}$ and $B = 2^q$ (solid), and two-point lower bounds (dashed), for $q = 2.5, 3, 4$. We use $\delta = 0$ and vary the target and model laws with M and κ.

5.2 Estimation Error and Validation Size

To examine how the estimation error depends on M and κ, we calculate RMSE directly for specified rare-event distributions with $q \in \{2.5, 3, 4\}$ (Figure 2). In these examples, RMSE decreases at rate $M^{-(q-2)/q}$ when the laws agree. With mismatch, the improvement levels off at an error of order $\kappa^{(q-2)/q}$. We calculate RMSE exactly at R_{bd} with $B = 2^q$ and compare it with two-point lower bounds. The laws vary with (M, κ) to examine worst-case dependence on sample size and mismatch. Appendix D.2 gives all 270 configurations.

Radius selection and remaining variance error. The preceding calculation uses R_{bd} from known budgets. We next examine radius selection without these budgets, using a single anchor and holding the target and model laws fixed as validation size varies. Both laws are supported on the same three values, but the model assigns more or less probability to the two extreme values. Its variance is therefore larger or smaller than the target variance, with $\mathrm{KL}(P\|Q) = 0.01$ in both cases. For each $q \in \{2.5, 3, 4\}$, we compare R_{bd} computed from the known budgets, the fixed radius $R = 4$, and a common radius selected by validation.

Clipping reduces the contribution of large response differences. In these examples, it offsets the upward bias when the model variance is too large, but adds downward bias when it is too small.

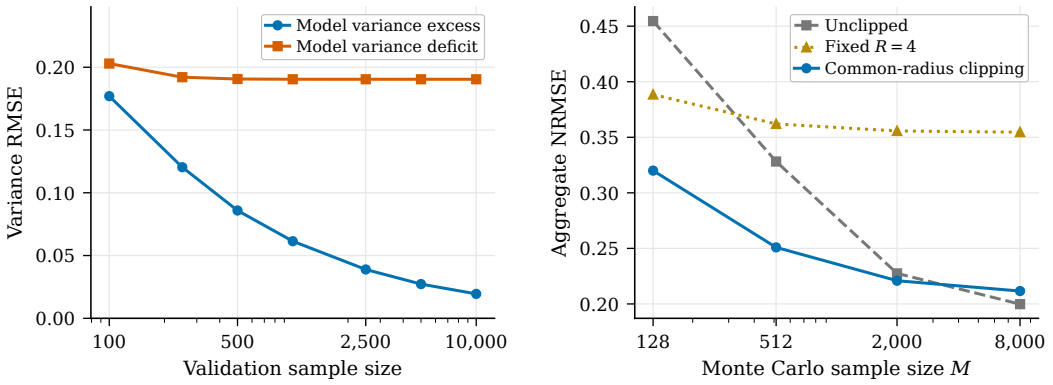


Figure 3: Effects of validation and Monte Carlo sample sizes. Left: variance RMSE of common-radius selection versus validation sample size for fixed target–model laws at $q = 3$, $M = 2,000$, and $\kappa = 0.01$ (20,000 repetitions). Right: variance NRMSE versus Monte Carlo sample size across twelve regression conditions at $(n_{\text{fit}}, n_{\text{val}}) = (10{,}000, 2{,}500)$.

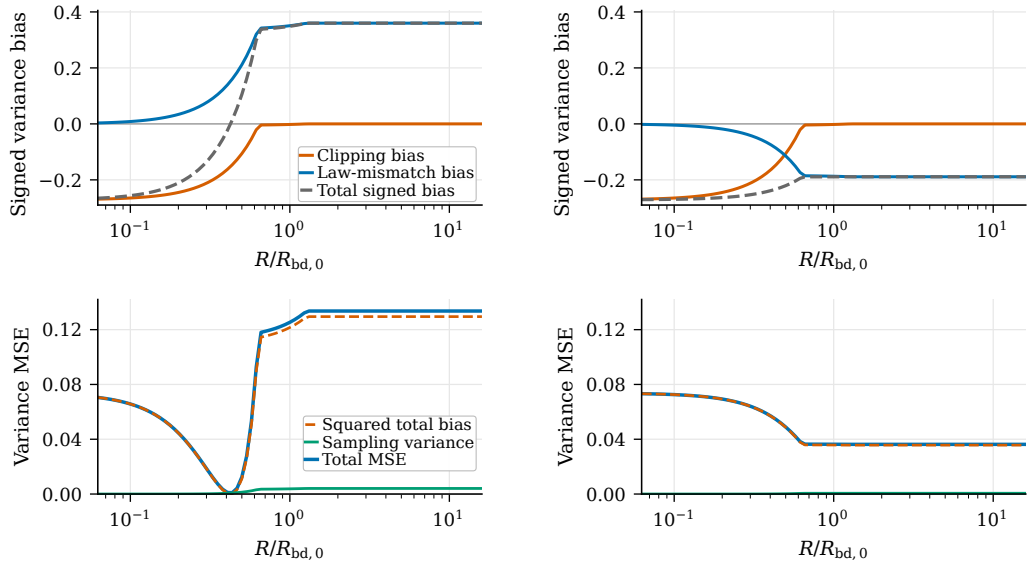


Figure 4: Bias and MSE decompositions across clipping radii. Exact signed-bias (top) and MSE (bottom) decompositions for three-point laws with excess (left) and deficient (right) model variance. We use $q = 3$, $\kappa = 0.01$, $M = 2,000$, and $\delta = 0$, without output projection. The reference radius $R_{\text{bd},0} \approx 5.79$ balances the upper bound with $B = 2.04$.

Figure 4 separates clipping and law-mismatch biases and adds Monte Carlo variance to squared total bias to obtain MSE. More validation data substantially reduce RMSE when the model variance is larger than the target variance, but provide limited gains when it is smaller (Figure 3, left). Thus, more validation data can reduce selection loss, but not the error remaining at the best common radius, represented by $\mathcal{E}^{\star}_{\text{cont}}$ in Theorem 3. Appendix D.2 specifies the laws and exact bias and MSE calculations.

5.3 Conditional-Variance Estimation in Regression

Across four regression functions and three noise laws, we compare unclipped estimation, $R = 4$, and common-radius selection for diffusion regression against MC Dropout (Gal & Ghahramani, 2016) and DeepHetero (Nix & Weigend, 1994). Each condition uses five independent fits and three Monte Carlo repetitions per fit. Between the two budgets in Table 1, both fitting and validation sample sizes

Table 1: Synthetic regression performance. Left/right entries use $(n_{\mathrm{fit}}, n_{\mathrm{val}}) = (10{,}000, 2{,}500)$ / $(40{,}000, 8{,}192)$, with $M = 2{,}000$. Clipping changes only variance estimates. Quantile interval coverage error (QICE) is evaluated against Y_0. Metric definitions appear in Appendix D.4.

Estimator	Mean RMSE	Variance RMSE	Variance NRMSE	Sample RMSE	QICE (%)
Unclipped	0.1402/0.0914	5.5981/4.7401	0.2275/0.1924	3.0603/3.0590	0.3286/0.3366
$R = 4$	0.1402/0.0914	16.6993/16.6015	0.3557/0.3501	3.0603/3.0590	0.3286/0.3366
Common radius	0.1402/0.0914	7.2086/5.0738	0.2209/0.1573	3.0603/3.0590	0.3286/0.3366
MC Dropout	0.1863/0.1486	5.4833/4.3348	0.2076/0.1283	3.0622/3.0609	2.6060/2.6290
DeepHetero	0.1618/0.0994	5.8724/5.3069	0.2127/0.1479	3.0623/3.0582	2.4247/2.4325

increase. Baselines train on the combined samples, whereas diffusion reserves validation data for radius selection.

With fitted models fixed and the radius reselected at each M, variance NRMSE decreases most at small M (Figure 3, right). At both budgets, diffusion achieves lower aggregate mean RMSE and substantially lower QICE than both baselines (Table 1). These strengths in mean prediction and aggregate quantile coverage coexist with higher common-radius variance RMSE than MC Dropout's, motivating our analysis of variance recovery from diffusion samples. Relative to unclipped estimation, common-radius clipping lowers aggregate variance NRMSE after normalization by each condition's target variance scale, but raises raw variance RMSE. Cubic, whose target variance reaches 100, accounts for over 99.7% of aggregate common-radius variance MSE. Theorem 1 provides a fixed-radius error bound, not a guarantee of improvement over unclipped estimation.

To examine the approximation and selection errors in Theorem 3, we repeat validation in four conditions at both budgets, holding fitted models and model samples fixed. The best common radius reduces unclipped MSE by G, while choosing the radius by validation adds an excess MSE of S. Selected clipping therefore improves estimation exactly when $S < G$, which holds on average for Curve with t_3 noise but not for Cubic (Appendix D.4). For contaminated Cubic, the best common radius reduces average unclipped MSE by less than 0.6% at both budgets. For diagnosis only, replacing U_i by its conditional expectation $\bar{U}_i$ sharply reduces average selection loss in all eight condition–budget combinations. Validation-response noise also dominates selection-loss variability (Appendix D.4).

6 Conclusion

Conditional-variance estimation from diffusion samples is limited by target-to-model mismatch and by Monte Carlo error, and KL control of the generated law is not enough. Under target pairwise q-moment and target-to-model conditional KL bounds, we derive an error bound for pairwise clipping. At R_{bd}, pairwise clipping attains the minimax rate over the stated class without a moment condition on the model law. The average conditional KL analysis links this bound to score risk, discretization, and initialization through an integrated L^1 guarantee. Independent validation selects a common radius without estimating the moment or KL budgets. For a fixed path, we bound both new-covariate risk and expected selection loss. The experiments illustrate the predicted rates and highlight the additional challenge of recovering conditional variance from models with strong mean prediction and aggregate quantile coverage.

The validation guarantee is limited to scalar covariates, fixed anchors, and bounded-variation conditional mean and variance. If the clipping path poorly approximates the target variance, more validation data cannot remove the resulting error. The minimax result concerns estimators using only model samples; optimality for procedures that also use target observations remains open. Future work could also extend validation guarantees to higher-dimensional covariates and validation-dependent anchors.

AI USE STATEMENT

The authors developed the initial conditional-law error analysis and its associated assumptions through their own research. We used OpenAI's ChatGPT and Codex for literature search, including identifying and organizing references relevant to Theorem 3 and Appendix C and comparing the assumptions and conclusions of related results. These tools also helped us explore proof ideas and check the convergence rates and error decompositions in these arguments. For the remaining theoretical results, we used the tools to review proofs previously developed by the authors for possible errors, gaps, and missing assumptions. The authors reviewed the suggestions and determined the final mathematical statements and proofs.

We also used these tools to refine sentence structure and wording in the abstract and main text, check the consistency of terminology and mathematical notation, and translate the abstract and technical explanations between the authors' native language and English.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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

Nicolas Beltran Velez, Alessandro A Grande, Achille Nazaret, Alp Kucukelbir, and David Blei. Treeffuser: probabilistic prediction via conditional diffusions with gradient-boosted trees. *Advances in Neural Information Processing Systems*, 37:118296–118325, 2024.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, volume 2024, pp. 36916–36936, 2024.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 9780199535255.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL http://jmlr.org/papers/v23/21-0585.html.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. NGBoost: Natural gradient boosting for probabilistic prediction. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2690–2700. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/duan20a.html.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Xizewen Han and Mingyuan Zhou. Diffusion boosted trees. *arXiv preprint arXiv:2406.01813*, 2024.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pp. 1188–1205, 1986.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005. URL https://www.jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf.

Yuling Jiao, Lican Kang, Jin Liu, Heng Peng, and Heng Zuo. Model free prediction with uncertainty assessment. *IEEE Transactions on Information Theory*, 2025.

Emilien Joly and Gábor Lugosi. Robust estimation of u-statistics. *Stochastic Processes and their Applications*, 126(12):3760–3773, 2016.

Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

Carlo Kneissl, Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Improved probabilistic regression using diffusion models. In *Northern Lights Deep Learning Conference Abstracts 2026*, 2025.

Quoc V. Le, Alexander J. Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005. URL https://www.machinelearning.org/proceedings/icml2005/papers/033_Heteroskedastic_LeSmolaCanu.pdf.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/lee23a.html.

Mengze Li. Non-asymptotic convergence bound of conditional diffusion models. *arXiv preprint arXiv:2508.10944*, 2025.

Stanislav Minsker and Xiaohan Wei. Robust modifications of u-statistics and applications to covariance estimation problems. *Bernoulli*, 26(1):694–727, 2019.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/oko23a.html.

Iosif Pinelis. Multidimensional probability inequalities via spherical symmetry, 2022. URL https://arxiv.org/abs/2210.04391.

Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3):507–554, 2005.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2203.09168.

Dule Shu and Amir Barati Farimani. Zero-shot uncertainty quantification using diffusion probabilistic models. *arXiv preprint arXiv:2408.04718*, 2024.

Bruno O. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388, 1972.

Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Jaemin Song and Jaegi Jeon. Understanding solver-induced variance distortion in conditional diffusion regression. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026.

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*, 2021.

Rong Tang, Lizhen Lin, and Yun Yang. Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. URL https://www.iro.umontreal.ca/~vincentp/Publications/smdae_techreport.pdf.

Lie Wang, Lawrence D. Brown, T. Tony Cai, and Michael Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664, 2008.

Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5): 2392–2405, 2009.

Eliot Wong-Toi, Alex Boyd, Vincent Fortuin, and Stephan Mandt. Understanding pathologies of deep heteroskedastic regression. In *Uncertainty in Artificial Intelligence*, pp. 3765–3790. PMLR, 2024.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023.

Yikai Zhang, Jiahe Lin, Fengpei Li, Yeshaya Adler, Kashif Rasul, Anderson Schneider, and Yuriy Nevmyvaka. Risk bounds on aleatoric uncertainty recovery. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6015–6036. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/zhang23h.html.

Changyuan Zhao, Hongyang Du, Guangyuan Liu, and Dusit Niyato. Supervised score-based modeling by gradient boosting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22768–22776, 2025.

A Conditional KL Bounds in Reverse Time

We prove Proposition 1 by bounding score risk, time discretization, and terminal initialization at a fixed covariate, then averaging over P_X. The discretization argument adapts Benton et al. (2024, Lemmas 2–3 and Appendix E) to the reverse clock $\tau = T - t$ and the unit-diffusion normalization of (2).

Condition on $\mathscr{G}_n$ and take $0 < \delta \le 1$ and $T \ge 2$ unless a different range is stated. The fixed-covariate arguments apply for P_X-almost every x with $m_2(x) < \infty$ for which Assumption 3 holds. Write $S = T - \delta$ and $\tau_k = T - t_k$, so that $0 = \tau_N < \cdots < \tau_0 = S$. Let P_x denote the path law of the exact reverse process in (2), initialized from P_T^x.

A.1 Conditioning and Score-Risk Transfer

Measurability. Assume that $\widehat{s}_n(t, y, x)$ is jointly measurable in (t, y, x, ω) and $\mathscr{G}_n$-measurable in ω. The finite ExpInt recursion defines a probability kernel $x \mapsto Q^x_{\delta,n,N,T}$ measurable in x and the fitting randomness. Together with the target kernels in Section 2, this makes the conditional moments and divergences measurable.

Conditional denoising identity. The conditional denoising identity in Section 2.2 and squared-loss projection (Vincent, 2011) give

$$\overline{\mathcal{L}}_{\mathrm{DSM}}(s) - \overline{\mathcal{L}}_{\mathrm{DSM}}(s^\star) = \overline{\mathcal{R}}_{N,T,\delta}(s). \tag{20}$$

The cutoff $\delta > 0$ ensures $\mathbb{E}(\varepsilon/\sigma_{t_k})^2 < \infty$ at every grid point.

A sufficient training-risk condition. Let $\mathcal{F}$ be a measurable score class with finite population losses, and assume that $\widehat{s}_n \in \mathcal{F}$ almost surely. Let $\widehat{\mathcal{L}}_{\mathrm{DSM}}$ be the empirical counterpart of $\overline{\mathcal{L}}_{\mathrm{DSM}}$ in (5). Let $a_n, \epsilon_{\mathrm{opt,n}} \ge 0$ be deterministic tolerances. On an event where

$$\sup_{s\in\mathcal{F}} |\widehat{\mathcal{L}}_{\mathrm{DSM}}(s) - \overline{\mathcal{L}}_{\mathrm{DSM}}(s)| \le a_n, \qquad \widehat{\mathcal{L}}_{\mathrm{DSM}}(\widehat{s}_n) \le \inf_{s\in\mathcal{F}} \widehat{\mathcal{L}}_{\mathrm{DSM}}(s) + \epsilon_{\mathrm{opt,n}},$$

the empirical-risk decomposition and (20) give

$$\overline{\mathcal{R}}_{N,T,\delta}(\widehat{s}_n) \le \inf_{s\in\mathcal{F}} \overline{\mathcal{R}}_{N,T,\delta}(s) + 2a_n + \epsilon_{\mathrm{opt,n}}. \tag{21}$$

A deterministic bound on the right-hand side that holds with probability at least $1 - \zeta$ supplies the budget in Assumption 2. A concrete rate for a_n requires class-specific complexity and tail bounds.

A.2 Path-Law Entropy and Terminal Initialization

On $C([0, S], \mathbb{R})$ with its raw canonical filtration $(\mathscr{F}_\tau)_{0\le\tau\le S}$, let Q^ν_x be the continuous ExpInt law initialized from $\nu \in \{P_T^x, \pi_\infty\}$. The exact and ExpInt drifts on $[\tau_k, \tau_{k-1})$ are, respectively,

$$b_\tau(Z) = \tfrac{1}{2}Z_\tau + s^\star(T - \tau, Z_\tau, x), \qquad \widehat{b}_\tau(Z) = \tfrac{1}{2}Z_\tau + \widehat{s}_n(t_k, Z_{\tau_k}, x).$$

The exact marginal is $P^x_{T-\tau}$ by diffusion time reversal (Anderson, 1982; Haussmann & Pardoux, 1986). Integrating the second SDE on each reverse interval gives (6), so the endpoint of $\mathsf{Q}^{\pi_\infty}_x$ has law $Q^x_{\delta,n,N,T}$.

The stopping conditions in Assumption 3 require the stopped martingale problems for these drifts to be unique in law and stable under concatenation. The three path laws just specified are nonexplosive. The drift difference $\Delta_\tau = b_\tau - \widehat{b}_\tau$ is progressively measurable with locally finite quadratic energy under both laws with initial distribution P_T^x. More precisely, for every $\rho > 0$, with $\sigma_\rho = \inf\{\tau \ge 0 \mid |Z_\tau| \ge \rho\} \wedge S$, assume $\int_0^{\sigma_\rho} |\Delta_\tau|^2 \mathrm{d}\tau < \infty$ almost surely under P_x and $\mathsf{Q}^{P_T^x}_x$.

Lemma 2 (Target-to-sampler path entropy)**.** *Under these conditions, if*

$$\mathbb{E}_{\mathsf{P}_x} \int_0^S |\Delta_\tau|^2 \mathrm{d}\tau < \infty,$$

then

$$\mathrm{KL}(\mathsf{P}_x \| \mathsf{Q}^{\pi_\infty}_x) \le \mathrm{KL}(P_T^x \| \pi_\infty) + \frac{1}{2}\mathbb{E}_{\mathsf{P}_x} \int_0^S |\Delta_\tau|^2 \mathrm{d}\tau. \tag{22}$$

Proof of Lemma 2. Write $\mathsf{Q}_x^0 = \mathsf{Q}_x^{P_T^x}$ and, for $\ell \geq 1$, set

$$\eta_\ell = S \wedge \inf\left\{\tau \geq 0 \;\middle|\; \int_0^\tau |\Delta_u|^2 \mathrm{d}u \geq \ell \text{ or } |Z_\tau| \geq \ell\right\}.$$

Under Q_x^0, the stochastic integral in

$$L_\ell = \exp\left\{\int_0^{\eta_\ell} \Delta_u \mathrm{d}B_u^{\mathsf{Q}_x^0} - \frac{1}{2}\int_0^{\eta_\ell} |\Delta_u|^2 \mathrm{d}u\right\}$$

has quadratic variation at most ℓ, so its stochastic exponential is a martingale with conditional mean one given Z_0. Girsanov's theorem and stopped uniqueness identify the tilted stopped law with the exact stopped law. The localized entropy comparison of Benton et al. (2024, Proposition 3 and Appendix F) gives

$$\mathrm{KL}\big(\mathrm{Law}_{\mathsf{P}_x}(Z_{\cdot\wedge\eta_\ell})\big\|\,\mathrm{Law}_{\mathsf{Q}_x^0}(Z_{\cdot\wedge\eta_\ell})\big) \leq \frac{1}{2}\mathbb{E}_{\mathsf{P}_x}\int_0^{\eta_\ell} |\Delta_u|^2 \mathrm{d}u.$$

Nonexplosion and local finite energy give $\eta_\ell \uparrow S$ under both laws. Path continuity yields uniform convergence of the stopped paths and hence weak convergence of their laws. Lower semicontinuity of relative entropy and monotone convergence give

$$\mathrm{KL}(\mathsf{P}_x\|\mathsf{Q}_x^{P_T^x}) \leq \frac{1}{2}\mathbb{E}_{\mathsf{P}_x}\int_0^S |\Delta_u|^2 \mathrm{d}u.$$

Let P_x^z be the conditional exact path law given $Z_0 = z$. The continuous ExpInt interpolation defines a common conditional path kernel $K_x(z,\cdot)$ for both sampler initializations. Disintegration over Z_0 gives

$$\mathrm{KL}(\mathsf{P}_x\|\mathsf{Q}_x^0) = \int \mathrm{KL}(\mathsf{P}_x^z\|K_x(z,\cdot))P_T^x(\mathrm{d}z),$$

$$\mathrm{KL}(\mathsf{P}_x\|\mathsf{Q}_x^{\pi_\infty}) = \mathrm{KL}(P_T^x\|\pi_\infty) + \int \mathrm{KL}(\mathsf{P}_x^z\|K_x(z,\cdot))P_T^x(\mathrm{d}z).$$

The initial entropy is finite by Lemma 3. Combining these identities proves (22). □

Lemma 3 (Terminal initialization)**.** *For $T \geq 1$,*

$$\mathrm{KL}(P_T^x\|\pi_\infty) \leq \tfrac{1}{2}\{e^{-T}m_2(x) - e^{-T} - \log(1-e^{-T})\} \leq C\{1+m_2(x)\}e^{-T}.$$

Proof of Lemma 3. Apply convexity of KL to $P_T^x = \int \mathcal{N}(e^{-T/2}y, 1-e^{-T})P_0^x(\mathrm{d}y)$, then use the Gaussian KL formula and $-a - \log(1-a) \leq Ca^2$ for $0 \leq a \leq e^{-1}$. This is the terminal-mixture argument of Chen et al. (2023, Lemma C.4) and Benton et al. (2024, Proposition 5). □

A.3 Fixed-Covariate ExpInt Discretization

We construct the grid of Lemma 1, then use the reverse-clock score process to prove Lemma 4. This bound controls the discretization term in Proposition 1.

Proof of Lemma 1. Use the geometric–linear construction of Benton et al. (2024, Corollary 1 and Appendix D). For $0 < \delta < 1$, write $N_{\mathrm{geo}} = N/2$ and set

$$\begin{aligned} t_k &= \delta^{1-k/N_{\mathrm{geo}}}, && 0 \leq k \leq N_{\mathrm{geo}}, \\ t_{N_{\mathrm{geo}}+j} &= 1 + j(T-1)/N_{\mathrm{geo}}, && 1 \leq j \leq N_{\mathrm{geo}}. \end{aligned}$$

These nodes define the reverse grid $\tau_k = T - t_k$. For $a = 2\log(1/\delta)/N \leq 1$,

$$\frac{h_k}{\min\{1, t_{k-1}\}} = \begin{cases} e^a - 1, & k \leq N_{\mathrm{geo}}, \\ 2(T-1)/N, & k > N_{\mathrm{geo}}. \end{cases}$$

Since $e^a - 1 \leq (e-1)a$, (10) follows. When $\delta = 1$, use $t_k = 1 + k(T-1)/N$ for all $0 \leq k \leq N$. Then $\rho_N = (T-1)/N$, with all N steps positive. □

Lemma 4 (Fixed-covariate discretization)**.** *For $0 < \delta \le 1$, $T \ge 2$, and a grid with $h_k \le \rho_N \min\{1, t_{k-1}\}$,*

$$D_x = \sum_{k=1}^{N} \int_{\tau_k}^{\tau_{k-1}} \mathbb{E}_{\mathsf{P}_x} \big| s^\star(T-\tau, Z_\tau, x) - s^\star(t_k, Z_{\tau_k}, x) \big|^2 \mathrm{d}\tau \le C\{\rho_N^2 N + \rho_N T\}. \tag{23}$$

Proof of Lemma 4. Write

$$A_\tau = s^\star(T-\tau, Z_\tau, x), \quad H_\tau = \partial_y s^\star(T-\tau, Z_\tau, x),$$
$$I_\tau = \mathbb{E}_{\mathsf{P}_x} A_\tau^2, \qquad\qquad J_\tau = \mathbb{E}_{\mathsf{P}_x} H_\tau^2.$$

The score equation obtained by differentiating the forward Fokker–Planck equation is

$$\partial_t s^\star = \tfrac{1}{2} s^\star + \tfrac{1}{2} y\, \partial_y s^\star + s^\star \partial_y s^\star + \tfrac{1}{2} \partial_{yy} s^\star.$$

Apply Itô's formula along the reverse dynamics $\mathrm{d}Z_\tau = (Z_\tau/2 + A_\tau)\mathrm{d}\tau + \mathrm{d}B_\tau$. The derivative in $T - \tau$ has a minus sign, and the spatial derivative terms cancel.

$$\mathrm{d}A_\tau = -\tfrac{1}{2} A_\tau \mathrm{d}\tau + H_\tau \mathrm{d}B_\tau, \qquad \frac{\mathrm{d}I_\tau}{\mathrm{d}\tau} = -I_\tau + J_\tau. \tag{24}$$

Gaussian smoothing justifies these identities without a fourth moment of Y_0. Differentiating the kernel in $Y_t = \alpha_t Y_0 + \sigma_t \varepsilon$ gives

$$s^\star(t, Y_t, x) = -\frac{\mathbb{E}[\varepsilon \mid Y_t, x]}{\sigma_t}, \qquad \partial_y s^\star(t, Y_t, x) = \frac{\mathrm{Var}(\varepsilon \mid Y_t, x) - 1}{\sigma_t^2}.$$

Since $\mathrm{Var}(\varepsilon \mid Y_t, x) \le \mathbb{E}[\varepsilon^2 \mid Y_t, x]$, conditional Jensen's inequality gives, for $r \in \{2, 4\}$,

$$\mathbb{E}_{\mathsf{P}_x} |A_\tau|^r \le \sigma_{T-\tau}^{-r} \mathbb{E}|\varepsilon|^r,$$
$$\mathbb{E}_{\mathsf{P}_x} |H_\tau|^r \le 2^{r-1} \sigma_{T-\tau}^{-2r} \{\mathbb{E}|\varepsilon|^{2r} + 1\}.$$

In particular,

$$I_\tau \le \frac{1}{1 - e^{-(T-\tau)}}, \qquad J_\tau \le \frac{8}{(1 - e^{-(T-\tau)})^2}. \tag{25}$$

Since $\sigma_{T-\tau}^2 \ge 1 - e^{-\delta} > 0$, the fourth-moment bounds and Cauchy–Schwarz imply

$$\mathbb{E}_{\mathsf{P}_x} \int_0^S A_\tau^2 H_\tau^2 \, \mathrm{d}\tau < \infty.$$

Thus the martingale terms for A_τ and A_τ^2 are square-integrable and have mean zero, justifying (24).

Solving the first equation in (24) from τ_k to τ and applying Itô isometry gives

$$\mathbb{E}|A_\tau - A_{\tau_k}|^2 = (1 - e^{-(\tau - \tau_k)/2})^2 I_{\tau_k} + \int_{\tau_k}^{\tau} e^{-(\tau - u)} J_u \mathrm{d}u.$$

The cross term vanishes by conditioning on $\mathscr{F}_{\tau_k}$. Using $(1 - e^{-r/2})^2 \le r$ and $w_\tau = \min\{1, T - \tau\}$, integration over each interval and summation give

$$D_x \le \frac{1}{2} \sum_{k=1}^{N} h_k^2 I_{\tau_k} + \rho_N \int_0^S w_\tau J_\tau \mathrm{d}\tau. \tag{26}$$

Indeed, the stochastic integral is weighted by at most h_k, and $h_k \le \rho_N w_u$ throughout its reverse interval. The energy and step-size bounds give

$$I_{\tau_k} \le \frac{C}{\min\{1, t_k\}} \le \frac{C}{\min\{1, t_{k-1}\}},$$
$$\sum_k h_k^2 I_{\tau_k} \le C \rho_N \sum_k h_k \le C \rho_N T.$$

For the second term, integrate (24) against w_τ. This weight is continuous and has derivative $-\mathbf{1}_{\{\tau>T-1\}}$ almost everywhere. Consequently,

$$\int_0^S w_\tau J_\tau \mathrm{d}\tau = \delta I_S - I_0 + \int_{T-1}^S I_\tau \mathrm{d}\tau + \int_0^S w_\tau I_\tau \mathrm{d}\tau \\ \leq C\{1 + \log(1/\delta) + T\} \leq C\{\log(1/\delta) + T\}.$$

At $\delta = 1$ the middle integral is zero. Finally,

$$\log(1/\delta) = \int_0^S \frac{\mathbf{1}_{\{T-\tau\leq 1\}}}{T-\tau}\mathrm{d}\tau \leq \sum_{k=1}^N \frac{h_k}{\min\{1, t_{k-1}\}} \leq \rho_N N.$$

Substitution into (26) gives

$$D_x \leq C\rho_N T + C\rho_N\{\log(1/\delta) + T\} \leq C\{\rho_N^2 N + \rho_N T\},$$

which proves (23). □

A.4 Proof of the Average Conditional KL Bound

Proof of Proposition 1. If $\overline{\mathcal{R}}_{N,T,\delta}(\widehat{s}_n) = \infty$, the bound is immediate. Otherwise, define

$$E_x = \sum_{k=1}^N h_k \, \mathbb{E}_{Y\sim P^x_{t_k}} |\widehat{s}_n(t_k, Y, x) - s^\star(t_k, Y, x)|^2.$$

Adding and subtracting $s^\star(t_k, Z_{\tau_k}, x)$ inside Δ_τ, and using the exact reverse marginals, gives

$$\mathbb{E}_{\mathsf{P}_x} \int_0^S |\Delta_\tau|^2 \mathrm{d}\tau \leq 2D_x + 2E_x.$$

By the definition of the grid weights,

$$\int E_x \, P_X(\mathrm{d}x) = (T-\delta)\overline{\mathcal{R}}_{N,T,\delta}(\widehat{s}_n) < \infty.$$

Thus E_x and, by Lemma 4, the drift-error energy are finite for P_X-almost every x. Applying Lemmas 2 to 4 and the data-processing inequality at the endpoint gives

$$\mathrm{KL}(P^x_\delta \| Q^x_{\delta,n,N,T}) \leq E_x + C\{\rho_N^2 N + \rho_N T + (1 + m_2(x))e^{-T}\}.$$

Integrating this bound over P_X yields

$$\overline{K}_{n,N,T,\delta} \leq C\left[(T-\delta)\overline{\mathcal{R}}_{N,T,\delta}(\widehat{s}_n) + \rho_N^2 N + \rho_N T + (1+m_2)e^{-T}\right]. \tag{27}$$

For any grid satisfying (10), $\rho_N \leq 4L_{\delta,T}/N$ and $T \leq L_{\delta,T}$. Therefore,

$$\rho_N^2 N + \rho_N T \leq C\frac{L^2_{\delta,T}}{N},$$

which proves (12). □

A.5 KL Alone Does Not Control Unclipped Variance

Proof of Proposition 2. Fix $\delta > 0$. For $0 < \varepsilon < 1$, take $P_0 = \mathcal{N}(0,1)$ and

$$Q_{\varepsilon,0} = (1-\varepsilon)P_0 + \frac{\varepsilon}{2}\{\mathcal{N}(a,1) + \mathcal{N}(-a,1)\}, \qquad a = \varepsilon^{-1}.$$

Each $Q_{\varepsilon,0}$ is centered and has a smooth positive density. After forward OU evolution, $P_\delta = \mathcal{N}(0,1)$ and $Q_{\varepsilon,\delta} \geq (1-\varepsilon)P_\delta$, so

$$\mathrm{KL}(P_\delta \| Q_{\varepsilon,\delta}) \leq -\log(1-\varepsilon).$$

The endpoint variance difference is

$$\mathrm{Var}_{Q_{\varepsilon,\delta}}(Z) - \mathrm{Var}_{P_\delta}(Z) = e^{-\delta}\varepsilon a^2.$$

Consequently,

$$|\mathcal{T}_\delta(Q_{\varepsilon,\delta}) - \mathcal{T}_\delta(P_\delta)| = \varepsilon a^2 = \varepsilon^{-1}.$$

As $\varepsilon \downarrow 0$, the KL divergence tends to zero, whereas the OU-corrected variance difference diverges. □

B Proofs for Sharp Conditional-Variance Estimation

We prove the RMSE bounds in Section 4.1 and derive the integrated L^1 consequence of the average KL bound. Throughout, we condition on $\mathscr{G}_n$ and use the OU variance identity (3) and clipping kernel (7). The clipped-estimator bounds require pairwise moments only under the target law.

B.1 Moment-Sensitive Upper Bound

A one-sided Hellinger comparison controls the change in the clipped-kernel expectation, and a U-statistic variance bound controls Monte Carlo error. For densities p and r of P and Q with respect to a common dominating measure μ, let $H^2(P,Q) = \int(\sqrt{p}-\sqrt{r})^2\,\mathrm{d}\mu$. With this normalization, $H^2(P,Q) \le \mathrm{KL}(P\|Q)$ and $H^2(P^{\otimes 2},Q^{\otimes 2}) \le 2H^2(P,Q)$ (Sason & Verdú, 2016). For an integrable function f, write $Pf = \int f\,\mathrm{d}P$ and $Qf = \int f\,\mathrm{d}Q$. For $0 \le f \le b$,

$$|Qf - Pf| \le 2H(P,Q)\sqrt{Pf^2} + bH^2(P,Q), \qquad \sqrt{Qf^2} \le \sqrt{Pf^2} + bH(P,Q). \tag{28}$$

The first inequality follows from

$$r - p = 2\sqrt{p}(\sqrt{r}-\sqrt{p}) + (\sqrt{r}-\sqrt{p})^2$$

and Cauchy–Schwarz. The second follows from Minkowski.

For $M \ge 2$ and $Z_1,\dots,Z_M \overset{\text{iid}}{\sim} Q$, define

$$U_M(f) = \binom{M}{2}^{-1} \sum_{1\le i<j\le M} f(Z_i,Z_j).$$

For the model sample at a fixed covariate, $U_M(\psi_R) = \widehat{C}_{R,M}(x)$. For a square-integrable symmetric kernel f, Hoeffding's order-two decomposition (Hoeffding, 1948) gives

$$\mathrm{Var}_{Q^{\otimes M}}(U_M(f)) = \frac{4\zeta_1}{M} + \frac{2\zeta_2}{M(M-1)} \le \frac{2}{M}Q^{\otimes 2}f^2, \tag{29}$$

where ζ_1 and ζ_2 are the variances of the first and canonical second Hoeffding projections, so $2\zeta_1+\zeta_2 = \mathrm{Var}_{Q^{\otimes 2}}(f)$.

Proof of Theorem 1. Fix x and put $P = P_\delta^x$, $Q = Q_{\delta,n,N,T}^x$, $b_x = P^{\otimes 2}|Z-Z'|^q$, $k_x = \mathrm{KL}(P\|Q)$, and $\theta_{P,R} = P^{\otimes 2}\psi_R$. Since $2 < q \le 4$,

$$0 \le \mathrm{Var}_P(Z) - \theta_{P,R} \le \tfrac{1}{2}b_xR^{2-q}, \qquad P^{\otimes 2}\psi_R^2 \le \tfrac{1}{4}b_xR^{4-q}. \tag{30}$$

Apply (28) to the product laws with $b = R^2/2$, and then apply (29). The bias and standard deviation satisfy

$$|Q^{\otimes 2}\psi_R - \mathrm{Var}_P(Z)| \le C\{b_xR^{2-q} + \sqrt{b_xk_x}R^{2-q/2} + R^2k_x\}, \tag{31}$$

$$\sqrt{\mathrm{Var}_{Q^{\otimes M}}(U_M(\psi_R))} \le C\{\sqrt{b_x/M}R^{2-q/2} + R^2\sqrt{k_x/M}\}. \tag{32}$$

The RMSE is at most the absolute bias plus the standard deviation. Using $2\sqrt{k_x/M} \le k_x + M^{-1}$ gives

$$\begin{aligned}&\left\{\mathbb{E}_{Q^{\otimes M}}|U_M(\psi_R) - \mathrm{Var}_P(Z)|^2\right\}^{1/2}\\ &\quad\le C_q\left[b_xR^{2-q} + \sqrt{b_x(k_x+M^{-1})}R^{2-q/2} + R^2(k_x+M^{-1})\right].\end{aligned} \tag{33}$$

The OU correction multiplies the endpoint estimation error by e^δ. Since $v_0(x) \ge 0$, the projection onto $[0,\infty)$ in $\widehat{v}_{0,R,M}(x)$ cannot increase its absolute error. Substituting $b_x \le B$ and $k_x \le \kappa$ therefore proves (13). If these bounds hold uniformly in x, squaring and integrating over an independent $X \sim P_X$ gives the integrated RMSE bound. For $B > 0$, choosing $R = R_{\mathrm{bd}} = (B/(\kappa+M^{-1}))^{1/q}$ makes all three terms equal to $B^{2/q}(\kappa+M^{-1})^{(q-2)/q}$, proving (14). □

B.2 Integrated Variance Bounds under Average Budgets

The average KL bound in Proposition 1 yields an integrated L^1 guarantee under an average target pairwise-moment bound. Define

$$B_{\delta,q} = \mathbb{E}_X \mathbb{E}_{(P_\delta^X)^{\otimes 2}} |Z - Z'|^q.$$

A finite time-zero pairwise moment supplies this budget through the OU flow.

Lemma 5 (OU pairwise-moment propagation)**.** *Fix $2 < q \le 4$. Conditionally on X, let Y_0, Y_0' be independent with law P_0^X, and suppose*

$$B_{0,q} = \mathbb{E}_X \mathbb{E}[|Y_0 - Y_0'|^q \mid X] < \infty.$$

Then

$$B_{\delta,q} \le 2^{q-1} \left[e^{-q\delta/2} B_{0,q} + (1 - e^{-\delta})^{q/2} \mathbb{E}|\xi - \xi'|^q \right], \qquad \xi, \xi' \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{34}$$

Proof of Lemma 5. Take ξ, ξ' independent of (X, Y_0, Y_0'). Two independent forward OU evolutions give

$$Z - Z' = e^{-\delta/2}(Y_0 - Y_0') + \sqrt{1 - e^{-\delta}}(\xi - \xi').$$

Apply $|a + b|^q \le 2^{q-1}(|a|^q + |b|^q)$ and take expectations. □

Proposition 3 (Integrated L^1 variance transfer)**.** *Fix $2 < q \le 4$. Under Assumption 1, if $B_{\delta,q} < \infty$, then conditionally on $\mathscr{G}_n$, for $M \ge 2$ and $R > 0$,*

$$\mathbb{E}[|\widehat{v}_{0,R,M}(X) - v_0(X)| \mid \mathscr{G}_n] \le C_q e^{\delta} \left[B_{\delta,q} R^{2-q} + R^2 \{ \overline{K}_{n,N,T,\delta} + M^{-1} \} \right]. \tag{35}$$

Proof of Proposition 3. At each x, the expected absolute error is bounded by the RMSE. Apply (33), multiply by e^{δ}, and use the contraction of the projection onto $[0, \infty)$. After averaging over X, Cauchy–Schwarz bounds the mixed term by

$$\sqrt{B_{\delta,q}(\overline{K}_{n,N,T,\delta} + M^{-1})} R^{2-q/2}.$$

This is the geometric mean of $B_{\delta,q} R^{2-q}$ and $R^2(\overline{K}_{n,N,T,\delta} + M^{-1})$. Bounding it by half their sum proves (35). □

Corollary 1 (End-to-end integrated L^1 bound)**.** *Under Assumptions 1 to 3 and the ExpInt-grid conditions of Proposition 1, fix $2 < q \le 4$ and $M \ge 2$. Suppose $B_{\delta,q} \le \bar{B}_{\delta,q}$ for a deterministic bound $0 < \bar{B}_{\delta,q} < \infty$. For $0 < \zeta < 1$, with $L_{\delta,T} = T + \log(1/\delta)$, set*

$$\mathcal{K}(\zeta) = C_{\mathrm{KL}} \left[(T - \delta) r_{n,N,T,\delta}(\zeta) + \frac{L_{\delta,T}^2}{N} + (1 + m_2) e^{-T} \right]. \tag{36}$$

For $R_{\mathrm{e2e}} = (\bar{B}_{\delta,q} / (\mathcal{K}(\zeta) + M^{-1}))^{1/q}$, with probability at least $1 - \zeta$ over $(\mathcal{D}_n, U_n)$,

$$\mathbb{E}[|\widehat{v}_{0,R_{\mathrm{e2e}},M}(X) - v_0(X)| \mid \mathscr{G}_n] \le C_q e^{\delta} \bar{B}_{\delta,q}^{2/q} \{ \mathcal{K}(\zeta) + M^{-1} \}^{(q-2)/q}, \tag{37}$$

where the conditional expectation averages over an independent $X \sim P_X$ and the M model samples at that covariate.

Proof of Corollary 1. On the score-risk event $\mathcal{A}_\zeta$ in Assumption 2, Proposition 1 gives $\overline{K}_{n,N,T,\delta} \le \mathcal{K}(\zeta)$. This event has probability at least $1 - \zeta$. Put $u = \mathcal{K}(\zeta) + M^{-1}$. The radius $R_{\mathrm{e2e}} = (\bar{B}_{\delta,q}/u)^{1/q}$ satisfies

$$\bar{B}_{\delta,q} R_{\mathrm{e2e}}^{2-q} = R_{\mathrm{e2e}}^2 u = \bar{B}_{\delta,q}^{2/q} u^{(q-2)/q}.$$

Apply (35) with $B_{\delta,q} \le \bar{B}_{\delta,q}$ and $\overline{K}_{n,N,T,\delta} \le \mathcal{K}(\zeta)$. The displayed identity then proves (37) on $\mathcal{A}_\zeta$. □

Unclipped model-law variance under a moment envelope. A pairwise-moment bound on the model law also controls the unprojected OU-corrected variance $\mathcal{T}_\delta$ from Proposition 2. This additional bound is not required for the clipped estimator.

Corollary 2 (Unclipped model-law variance transfer)**.** *Conditionally on $\mathscr{G}_n$, fix $2 < q \le 4$. Suppose $B_{\delta,q} < \infty$, $\overline{K}_{n,N,T,\delta} < \infty$, and, for a deterministic $\bar{C}_Q < \infty$,*

$$C_{Q,q} = \mathbb{E}_X \mathbb{E}_{(Q^X_{\delta,n,N,T})^{\otimes 2}} |Z - Z'|^q \le \bar{C}_Q.$$

Then

$$\mathbb{E}_X \left| \mathcal{T}_\delta(Q^X_{\delta,n,N,T}) - v_0(X) \right| \le C_q e^\delta (B_{\delta,q} + \bar{C}_Q)^{2/q} \overline{K}^{(q-2)/q}_{n,N,T,\delta}. \tag{38}$$

Proof. For a fixed x, put $P = P^x_\delta$, $Q = Q^x_{\delta,n,N,T}$, $b_x = P^{\otimes 2}|Z - Z'|^q$, $c_x = Q^{\otimes 2}|Z - Z'|^q$, and $k_x = \mathrm{KL}(P\|Q)$. The model clipping remainder satisfies

$$0 \le \mathrm{Var}_Q(Z) - Q^{\otimes 2}\psi_R \le \tfrac{1}{2} c_x R^{2-q}.$$

Add this to (31), average over X, and apply Cauchy–Schwarz to obtain

$$\begin{aligned} &\mathbb{E}_X \left| \mathrm{Var}_{Q^X_{\delta,n,N,T}}(Z) - \mathrm{Var}_{P^X_\delta}(Z) \right| \\ &\qquad \le C_q \left[\mathsf{M}_q R^{2-q} + \sqrt{\mathsf{M}_q \overline{K}_{n,N,T,\delta}}\, R^{2-q/2} + R^2 \overline{K}_{n,N,T,\delta} \right], \end{aligned} \tag{39}$$

where $\mathsf{M}_q = B_{\delta,q} + \bar{C}_Q$. When both M_q and $\overline{K}_{n,N,T,\delta}$ are positive, choose

$$R = \left(\frac{\mathsf{M}_q}{\overline{K}_{n,N,T,\delta}} \right)^{1/q}.$$

All three terms in (39) are then bounded by a constant times $\mathsf{M}_q^{2/q} \overline{K}^{(q-2)/q}_{n,N,T,\delta}$. If the average KL is zero, $P^x_\delta = Q^x_{\delta,n,N,T}$ for P_X-almost every x. If $\mathsf{M}_q = 0$, both pairwise variance functionals vanish almost everywhere. Finally, exact OU inversion multiplies the endpoint variance difference by e^δ, which proves (38). □

B.3 Minimax Lower Bound from Model Samples

Lower-bound mechanism. The finite-M term comes from a rare component with $Q = P_\delta$, whereas the mismatch term comes from two targets with different variances but the same model-sample distribution. We use the two-point reduction (Tsybakov, 2009, Sections 2.2–2.3).

Proof of Theorem 2. Let δ_a denote the point mass at a. For $0 < \varepsilon \le 1/4$, define

$$P_{0,0} = \delta_0, \qquad P_{1,0} = (1-\varepsilon)\delta_0 + \frac{\varepsilon}{2}(\delta_a + \delta_{-a}), \qquad a = e^{\delta/2}\{B/(2^{2q}\varepsilon)\}^{1/q}.$$

Let $P_{i,\delta}$ be the forward OU evolution of $P_{i,0}$ and put $v_i = \mathrm{Var}_{P_{i,0}}(Y_0)$. Then $v_0 = 0$, $v_1 = \varepsilon a^2$, and

$$\Delta = v_1 - v_0 = \varepsilon a^2 = \tfrac{1}{16} e^\delta B^{2/q} \varepsilon^{(q-2)/q}. \tag{40}$$

The target pairwise q-moment bound holds for both laws. For independent $Y_0, Y_0' \sim P_{1,0}$,

$$\mathbb{E}|Y_0 - Y_0'|^q \le 2^q \varepsilon a^q.$$

Hence (34) gives

$$\mathbb{E}_{P^{\otimes 2}_{1,\delta}} |Z - Z'|^q \le B/2 + 2^{q-1} b^{(0)}_{\delta,q} \le B,$$

where the last inequality uses $B \ge 2^q b^{(0)}_{\delta,q}$. For $P_{0,\delta}$, the pairwise moment equals $b^{(0)}_{\delta,q} \le B$. Both endpoint densities are smooth and strictly positive.

For the finite-M lower bound, take $Q_i = P_{i,\delta}$ and $\varepsilon = (2M)^{-1}$. Both $Q_0^{\otimes M}$ and $Q_1^{\otimes M}$ dominate the measure

$$\mathcal{A} = (1-\varepsilon)^M P^{\otimes M}_{0,\delta}, \qquad \mathcal{A}(\mathbb{R}^M) = (1-\varepsilon)^M \ge 1 - M\varepsilon = \tfrac{1}{2}.$$

For any estimator $\widetilde{v}$, let $\mathbb{E}_i$ denote expectation under $Q_i^{\otimes M}$. Since $t^2 + (t-\Delta)^2 \geq \Delta^2/2$,

$$\max_{i=0,1} \mathbb{E}_i|\widetilde{v} - v_i|^2 \geq \frac{1}{2}\int \left\{|\widetilde{v} - v_0|^2 + |\widetilde{v} - v_1|^2\right\} \mathrm{d}\mathcal{A} \geq \frac{\Delta^2}{4}\mathcal{A}(\mathbb{R}^M) \geq \frac{\Delta^2}{8}.$$

Together with (40), this yields an RMSE lower bound of $c_q e^{\delta} B^{2/q} M^{-(q-2)/q}$.

For the mismatch lower bound, let $0 < \kappa \leq 1/4$, set $\varepsilon = 1 - e^{-\kappa}$, and use $Q = P_{1,\delta}$ for both targets. Then

$$\mathrm{KL}(P_{0,\delta}\|Q) \leq -\log(1-\varepsilon) = \kappa, \qquad \mathrm{KL}(P_{1,\delta}\|Q) = 0.$$

The common observation law is $Q^{\otimes M}$, so

$$\begin{aligned}\max_{i=0,1} \mathbb{E}_{Q^{\otimes M}}|\widetilde{v} - v_i|^2 &\geq \frac{1}{2}\mathbb{E}_{Q^{\otimes M}}\left[|\widetilde{v} - v_0|^2 + |\widetilde{v} - v_1|^2\right] \\ &\geq \frac{\Delta^2}{4}.\end{aligned}$$

Thus the RMSE is at least $\Delta/2$. Since $1 - e^{-\kappa} \geq e^{-1/4}\kappa$, this yields an RMSE lower bound of $c_q e^{\delta} B^{2/q} \kappa^{(q-2)/q}$.

For $\kappa > 0$, all pairs used in these two constructions belong to $\mathfrak{C}_{\delta,q}(B,\kappa)$. Taking the smaller of the two constants, the minimax RMSE is at least

$$c_q e^{\delta} B^{2/q} \max\{\kappa^{(q-2)/q}, M^{-(q-2)/q}\} \geq c_q 2^{-(q-2)/q} e^{\delta} B^{2/q} (\kappa + M^{-1})^{(q-2)/q}.$$

For $\kappa = 0$, the finite-M construction gives the same conclusion. Absorbing $2^{-(q-2)/q}$ into c_q proves (16). □

C Validation-Based Clipping-Radius Selection

C.1 Pairwise Validation Criterion

We use the common-radius path in (17) and the selector in (18). The criterion omits the λ-independent term U_i^2, whose expectation need not be finite when $q < 4$.

Until Corollary 3, condition on $\mathscr{G}_n$ and the realized model samples. The path is then fixed and independent of $\mathcal{V}_B$. Set $H = V_{\max}$ and write

$$\begin{aligned}C_k(\lambda) &= \frac{1}{M(M-1)} \sum_{a<b} \min\{(\widehat{Y}_{k,a} - \widehat{Y}_{k,b})^2, \lambda\}, \\ u_k(\lambda) &= \Pi_{[\ell,H]}\{e^{\delta} C_k(\lambda) - (e^{\delta} - 1)\}.\end{aligned}$$

Since $\ell \geq 0$, this equals the projected estimator in (17).

Resolve covariate ties by original index in the adjacent pairing of Section 4.2. With $a_i = X_i^-$ and $b_i = X_i^+$, conditional independence of the responses gives

$$\bar{U}_i = \mathbb{E}[U_i \mid \mathbf{X}^{\mathrm{val}}] = \tfrac{1}{2}\{v_0(a_i) + v_0(b_i)\} + \tfrac{1}{2}\{\mu_0(a_i) - \mu_0(b_i)\}^2. \tag{41}$$

A sufficient condition for the pair-label moment bound in Theorem 3, with $p = q/2$, is

$$\sup_{x \in I} \mathbb{E}[|Y_0 - \mu_0(x)|^q \mid X = x] \leq K_q.$$

With $A_\mu = \mathrm{TV}_I(\mu_0)$, one may take $S_p = 3^{q-1}(2K_q + A_\mu^q)$. Indeed, for $D_i = Y_i^- - Y_i^+$,

$$\mathbb{E}[|U_i - \bar{U}_i|^p \mid \mathbf{X}^{\mathrm{val}}] \leq 2^p \mathbb{E}[U_i^p \mid \mathbf{X}^{\mathrm{val}}] = \mathbb{E}[|D_i|^q \mid \mathbf{X}^{\mathrm{val}}].$$

Decompose D_i into its two centered responses and its mean difference, and apply $|a+b+c|^q \leq 3^{q-1}(|a|^q + |b|^q + |c|^q)$.

Proposition 4 (Continuous radius search)**.** *Let the model samples at each anchor be finite, and put*

$$\lambda_{\max} = \max_k \left(\max_{1\le a\le M} \widehat{Y}_{k,a} - \min_{1\le a\le M} \widehat{Y}_{k,a} \right)^2.$$

The predictions for $\lambda \ge \lambda_{\max}$ equal those at $\lambda_{\max}$. For each x, the map $\lambda \mapsto f_\lambda(x)$ is continuous, nondecreasing, piecewise linear, and $e^\delta/2$-Lipschitz. For an interval $[a,b] \subset [0,\lambda_{\max}]$, set $l_i = f_a(z_i)$, $h_i = f_b(z_i)$, and $r_i = \Pi_{[l_i,h_i]} U_i$. Then

$$\begin{aligned} B_{\mathrm{ord}}(a,b) &= \frac{1}{m}\sum_i (r_i^2 - 2U_i r_i), \\ L_{a,b} &= \frac{e^\delta}{m}\sum_i \max\{|l_i - U_i|, |h_i - U_i|\}, \\ B_{\mathrm{lip}}(a,b) &= \{\widehat{\mathcal{L}}_{\mathrm{val}}(a) + \widehat{\mathcal{L}}_{\mathrm{val}}(b) - L_{a,b}(b-a)\}/2 \end{aligned} \tag{42}$$

satisfy

$$\max\{B_{\mathrm{ord}}(a,b), B_{\mathrm{lip}}(a,b)\} \le \inf_{\lambda\in[a,b]} \widehat{\mathcal{L}}_{\mathrm{val}}(\lambda).$$

For a partition with interval lower bounds B_j, let V_{inc} be the smallest evaluated criterion value. The gap $V_{\mathrm{inc}} - \min\{V_{\mathrm{inc}}, \min_j B_j\}$ bounds the optimization error.

Proof. All pairwise differences saturate at $\lambda_{\max}$. The slope of C_k lies in $[0, 1/2]$ almost everywhere. Projection and convex interpolation preserve monotonicity, continuity, and piecewise linearity, giving the stated Lipschitz bound.

For $\lambda \in [a,b]$, each prediction belongs to $[l_i, h_i]$. The minimum of $r^2 - 2U_i r$ on this interval occurs at r_i, giving B_{ord}. Almost everywhere,

$$|\widehat{\mathcal{L}}'_{\mathrm{val}}(\lambda)| \le \frac{e^\delta}{m}\sum_i |f_\lambda(z_i) - U_i| \le L_{a,b}.$$

The endpoint Lipschitz inequalities give B_{lip}. Every point lies in an interval of the partition, so the smallest interval lower bound is a global lower bound. An interval whose lower bound is at least V_{inc} can be discarded because it cannot improve the incumbent. □

For $\eta_{\mathrm{opt}} > 0$, set $B(a,b) = \max\{B_{\mathrm{ord}}(a,b), B_{\mathrm{lip}}(a,b)\}$. Starting from $[0,\lambda_{\max}]$, evaluate interval endpoints and bisect an active interval with smallest lower bound. Deterministic initialization and tie rules give a measurable selector.

If $\lambda_{\max} = 0$, the path is constant. Otherwise, define

$$L_{\mathrm{glob}} = \frac{e^\delta}{m}\sum_i \max\{|\ell - U_i|, |H - U_i|\}.$$

This constant is finite and positive, and $L_{a,b} \le L_{\mathrm{glob}}$ for every interval. Since $B(a,b) \ge B_{\mathrm{lip}}(a,b)$,

$$\min\{\widehat{\mathcal{L}}_{\mathrm{val}}(a), \widehat{\mathcal{L}}_{\mathrm{val}}(b)\} - B(a,b) \le \frac{L_{\mathrm{glob}}(b-a)}{2}.$$

When the active interval with smallest lower bound has width at most $2\eta_{\mathrm{opt}}/L_{\mathrm{glob}}$, an endpoint gives an optimization gap at most η_{opt}. Only finitely many dyadic intervals are wider than this threshold, so refinement terminates in exact arithmetic.

For $\eta_{\mathrm{opt}} = 0$, continuity ensures an exact minimizer. After writing $\lambda = t\lambda_{\max}$ with $t \in [0,1]$, the objective is jointly measurable in the data and continuous in t. For its smallest minimizer t_* and every $a \in [0,1]$, $\{t_* \le a\}$ is the event that the minima on $[0,a]$ and $[0,1]$ coincide. Both minima are measurable by continuity and countable rational approximation, so $\lambda_{\max} t_*$ is measurable.

The refinement follows the Lipschitz global-search principle of Shubert (1972). The statistical guarantee requires (18), not merely a local minimum. A floating-point gap is a numerical diagnostic rather than an interval-arithmetic certificate.

C.2 Proof of the New-Covariate Guarantee

Set $p = q/2 \in (1, 2]$ and

$$A_\mu = \mathrm{TV}_I(\mu_0), \qquad A_v = \mathrm{TV}_I(v_0),$$
$$A_c = \sup_{\lambda \geq 0} \mathrm{TV}_I(f_\lambda) \leq (K-1)H, \qquad H_{\mathrm{val}} = H + A_\mu^2/2.$$

Define

$$D_m = \frac{A_c + A_v}{\sqrt{2m}}, \qquad B_m = \frac{A_v + A_\mu^2}{2\sqrt{m}},$$
$$G_m = 2H\sqrt{\frac{\log\{2(Km+2)\}+1}{2m}} + \frac{2H}{m}, \qquad A_m = D_m + B_m + G_m, \tag{43}$$
$$\Pi_m = c_p 3^p S_p H_{\mathrm{val}}^{2-p} K^{p/2} m^{-(p-1)}, \qquad c_p = \begin{cases} 2, & 1 < p < 2, \\ 1, & p = 2. \end{cases}$$

The following bound implies Theorem 3.

$$\left\{\mathbb{E}_{\mathrm{val}} \|f_{\widehat{\lambda}} - v_0\|_{L^2(P_X)}^2\right\}^{1/2} \leq \sqrt{2(\mathcal{E}^\star_{\mathrm{cont}} + A_m)^2 + \Pi_m + 3\eta_{\mathrm{opt}}/2} + A_m. \tag{44}$$

The constants in (19) are uniform over realizations of the model samples for fixed K, H, S_p, q, A_v, A_μ.

Lemma 6 (Projected pair-label moment)**.** *Condition on all validation covariates. Let $\mathbf{W}_{ik} = w_k(z_i)$, let $\mathbf{P}_W$ be the orthogonal projection onto the column space of $\mathbf{W}$, and put $d = \mathrm{rank}(\mathbf{W})$. For $\varepsilon_i^{\mathrm{val}} = U_i - \bar{U}_i$ and $\boldsymbol{\varepsilon} = (\varepsilon_i^{\mathrm{val}})_{i=1}^m$,*

$$\mathbb{E}[\|\mathbf{P}_W \boldsymbol{\varepsilon}\|_2^p \mid \mathbf{X}^{\mathrm{val}}] \leq c_p S_p \sum_i (\mathbf{P}_W)_{ii}^{p/2} \leq c_p S_p d^{p/2} m^{1-p/2}.$$

Proof. Let e_i denote the ith Euclidean basis vector. Conditionally on $\mathbf{X}^{\mathrm{val}}$, the vectors $\varepsilon_i^{\mathrm{val}} \mathbf{P}_W e_i$ are independent and have mean zero. The Hilbert-space von Bahr–Esseen inequality (Pinelis, 2022, Theorem 3.1) bounds the pth moment of their sum by

$$c_p \sum_i \mathbb{E}[|\varepsilon_i^{\mathrm{val}}|^p \mid \mathbf{X}^{\mathrm{val}}] \|\mathbf{P}_W e_i\|_2^p.$$

For $p < 2$ its constant is at most two. At $p = 2$, addition of variances gives constant one. Since $\|\mathbf{P}_W e_i\|_2^2 = (\mathbf{P}_W)_{ii}$ and $\sum_i (\mathbf{P}_W)_{ii} = d$, concavity of $t^{p/2}$ proves the second inequality. □

Write

$$\|g\|_z^2 = m^{-1} \sum_i g(z_i)^2,$$
$$\|g\|_{P_{2m}}^2 = (2m)^{-1} \sum_{i=1}^{2m} g(X_i^{\mathrm{val}})^2,$$
$$\|g\|_{P_X}^2 = \mathbb{E}_X g(X)^2.$$

Lemma 7 (Norm transfer for the ordered path)**.** *Let P_{2m} be the empirical measure of the original independent validation covariates, and define*

$$\Gamma_X = \sup_{\lambda \geq 0} \left| \|f_\lambda - v_0\|_{P_{2m}} - \|f_\lambda - v_0\|_{P_X} \right|.$$

Then $\{\mathbb{E}\Gamma_X^2\}^{1/2} \leq G_m$.

Proof. The anchor path $(u_1(\lambda), \ldots, u_K(\lambda))$ is coordinatewise nondecreasing and belongs to $[0, H]^K$. Partition it by increments of H/m in $\sum_k u_k(\lambda)$, including both endpoints. Continuity gives a net of at most $N_{\mathrm{net}} = Km + 2$ path members. Every member differs from a net member by at most H/m

in each anchor coordinate and therefore in sup norm after interpolation. This net depends only on the frozen path.

For a fixed net function f, put $e = f - v_0$ and $r = \|e\|_{P_X}$. Then $|e| \le H$ and $\mathrm{Var}(e(X)^2) \le H^2 r^2$. Bernstein's inequality (Boucheron et al., 2013, Corollary 2.11) gives, with probability at least $1 - 2e^{-t}$,

$$\left|\|e\|_{P_{2m}}^2 - r^2\right| \le Hr\sqrt{\frac{2t}{2m}} + \frac{2H^2 t}{3(2m)}.$$

Writing $s = \|e\|_{P_{2m}}$ and $a = H\sqrt{t/(2m)}$, the inequality $|s^2 - r^2| \le \sqrt{2}\, ar + 2a^2/3$ implies $|s - r| \le 2a$. A union bound with $t = \log(2N_{\mathrm{net}}) + u$ and tail integration give the following L^2 bound for the maximum over the net

$$2H\sqrt{\frac{\log(2N_{\mathrm{net}}) + 1}{2m}}.$$

Each empirical- and population-norm approximation costs at most H/m. Minkowski's inequality proves the claim. The finite net is used only in the proof. □

Proof of Theorem 3. First condition on the complete validation covariate vector. The pair labels are independent, with conditional means (41). Let $\mathcal{L}_\lambda^{\bar{U}} = m^{-1}\sum_i (f_\lambda(z_i) - \bar{U}_i)^2$ and choose its smallest minimizer λ_* on $[0, \lambda_{\max}]$. For every λ, the vector

$$\big(f_\lambda(z_i) - f_{\lambda_*}(z_i)\big)_{i=1}^m$$

belongs to the column space of $\mathbf{W}$. Cauchy–Schwarz and $\|f_\lambda - f_{\lambda_*}\|_z^2 \le 2(\mathcal{L}_\lambda^{\bar{U}} + \mathcal{L}_{\lambda_*}^{\bar{U}})$ give, simultaneously for every λ,

$$\frac{\left|2m^{-1}\sum_i (f_\lambda(z_i) - f_{\lambda_*}(z_i))\varepsilon_i^{\mathrm{val}}\right|}{\sqrt{\mathcal{L}_\lambda^{\bar{U}} + \mathcal{L}_{\lambda_*}^{\bar{U}}}} \le C_X = \frac{2\sqrt{2}}{\sqrt{m}}\|\mathbf{P}_W \boldsymbol{\varepsilon}\|_2.$$

A zero denominator gives zero numerator. Approximate empirical minimization implies

$$\mathcal{L}_{\widehat{\lambda}}^{\bar{U}} - \mathcal{L}_{\lambda_*}^{\bar{U}} \le C_X\sqrt{\mathcal{L}_{\widehat{\lambda}}^{\bar{U}} + \mathcal{L}_{\lambda_*}^{\bar{U}}} + \eta_{\mathrm{opt}}.$$

Using $c\sqrt{t} \le t/3 + 3c^2/4$ yields $\mathcal{L}_{\widehat{\lambda}}^{\bar{U}} \le 2\mathcal{L}_{\lambda_*}^{\bar{U}} + 9C_X^2/8 + 3\eta_{\mathrm{opt}}/2$. Since $0 \le f_\lambda(z_i), \bar{U}_i \le H_{\mathrm{val}}$,

$$(\mathcal{L}_{\widehat{\lambda}}^{\bar{U}} - 2\mathcal{L}_{\lambda_*}^{\bar{U}} - 3\eta_{\mathrm{opt}}/2)_+ \le \min\{H_{\mathrm{val}}^2, 9C_X^2/8\} \le H_{\mathrm{val}}^{2-p}(9/8)^{p/2}C_X^p.$$

Lemma 6 and $d \le K$ therefore give

$$\mathbb{E}[\mathcal{L}_{\widehat{\lambda}}^{\bar{U}} \mid \mathbf{X}^{\mathrm{val}}] \le 2\mathcal{L}_{\lambda_*}^{\bar{U}} + \Pi_m + 3\eta_{\mathrm{opt}}/2, \tag{45}$$

where we used $(9/8)^{p/2}(2\sqrt{2})^p = 3^p$.

For a function g of bounded variation, comparing its midpoint evaluation with its two endpoint evaluations gives

$$|\|g\|_z - \|g\|_{P_{2m}}| \le \left\{\frac{1}{2m}\sum_i [(g(a_i) - g(z_i))^2 + (g(b_i) - g(z_i))^2]\right\}^{1/2} \le \frac{\mathrm{TV}_I(g)}{\sqrt{2m}}. \tag{46}$$

Since $b_i \le a_{i+1}$, the variations over the pair intervals sum to at most $\mathrm{TV}_I(g)$, including when covariates tie. For $s_i = \mathrm{TV}_{[a_i,b_i]}(v_0)$ and $t_i = \mathrm{TV}_{[a_i,b_i]}(\mu_0)$, $|\bar{U}_i - v_0(z_i)| \le (s_i + t_i^2)/2$. Since $\sum_i s_i \le A_v$ and $\sum_i t_i \le A_\mu$,

$$\|\bar{U} - v_0\|_z \le \frac{A_v + A_\mu^2}{2\sqrt{m}} = B_m. \tag{47}$$

Applying (46) to $f_\lambda - v_0$ gives the uniform midpoint transfer cost D_m. Lemma 7 supplies the remaining transfer from P_{2m} to P_X.

Because the path is constant beyond $\lambda_{\max}$, continuity and compactness give a smallest population minimizer

$$\lambda_{\text{pop}} \in \operatorname*{arg\,min}_{\lambda \in [0,\lambda_{\max}]} \|f_\lambda - v_0\|_{P_X}.$$

Comparing λ_* with λ_{pop} gives, pathwise,

$$\|f_{\widehat{\lambda}} - v_0\|_{P_X} \le \sqrt{\mathcal{L}^{\bar{U}}_{\widehat{\lambda}}} + D_m + B_m + \Gamma_X, \qquad \sqrt{\mathcal{L}^{\bar{U}}_{\lambda_*}} \le \mathcal{E}^\star_{\text{cont}} + D_m + B_m + \Gamma_X.$$

Take L^2 norms over validation observations, apply (45), and use Minkowski's inequality. This proves (44). Finally, $A_m \le \{(A_c + A_v)/\sqrt{2} + (A_v + A_\mu^2)/2 + 4H\}\sqrt{\log(2Km+4)/m}$, which gives (19). □

C.3 Expected Selection Loss

Proposition 5 (Expected selection loss)**.** *Assume the conditions of Theorem 3. Condition on the fitted model and the realized model samples. Write*

$$L(\lambda) = \|f_\lambda - v_0\|^2_{L^2(P_X)}, \qquad L_* = \inf_{\lambda \ge 0} L(\lambda) = (\mathcal{E}^\star_{\text{cont}})^2, \qquad S_{\text{val}} = L(\widehat{\lambda}) - L_*.$$

With $p = q/2$ and the constants in (43)*,*

$$0 \le \mathbb{E}_{\text{val}} S_{\text{val}} \le 4H(D_m + G_m) + \frac{H(A_v + A_\mu^2)}{m} + 2H(c_p S_p)^{1/p}\left(\frac{K}{m}\right)^{1-1/p} + \eta_{\text{opt}}. \tag{48}$$

For fixed K, H, S_p, A_v, A_μ, q, the right-hand side is $O(\sqrt{\log m/m} + m^{-(q-2)/q} + \eta_{\text{opt}})$. Thus expected selection loss vanishes as $m \to \infty$ and $\eta_{\text{opt}} \to 0$.

Proof. Let λ_* minimize L on $[0, \lambda_{\max}]$. Existence follows as in the proof of Theorem 3. Condition first on all validation covariates and set

$$\varepsilon_i^{\text{val}} = U_i - \bar{U}_i, \qquad \beta_i = \bar{U}_i - v_0(z_i), \qquad a_k = \frac{1}{m}\sum_i w_k(z_i)\varepsilon_i^{\text{val}}.$$

Approximate empirical minimization gives

$$\|f_{\widehat{\lambda}} - v_0\|_z^2 - \|f_{\lambda_*} - v_0\|_z^2 \le \frac{2}{m}\sum_i (\varepsilon_i^{\text{val}} + \beta_i)\{f_{\widehat{\lambda}}(z_i) - f_{\lambda_*}(z_i)\} + \eta_{\text{opt}}.$$

Since $f_\lambda = \sum_k w_k u_k(\lambda)$ and $u_k(\lambda) \in [0, H]$, the noise term is at most $2H\sum_k |a_k|$. Conditional independence and the scalar von Bahr–Esseen inequality (Pinelis, 2022, Theorem 3.1) give

$$\mathbb{E}[|a_k|^p \mid \mathbf{X}^{\text{val}}] \le \frac{c_p S_p}{m^p}\sum_i w_k(z_i)^p.$$

Jensen's and Hölder's inequalities, $0 \le w_k \le 1$, and $\sum_k w_k = 1$ imply

$$\begin{aligned}\mathbb{E}\left[\sum_k |a_k| \mid \mathbf{X}^{\text{val}}\right] &\le \frac{(c_p S_p)^{1/p}}{m}\sum_k \left\{\sum_i w_k(z_i)^p\right\}^{1/p} \\ &\le (c_p S_p)^{1/p} K^{1-1/p} m^{-(1-1/p)}.\end{aligned}$$

The adjacent-pair calculation in (47) gives $\sum_i |\beta_i| \le (A_v + A_\mu^2)/2$. Since $|f_{\widehat{\lambda}} - f_{\lambda_*}| \le H$, the matching term is at most $H(A_v + A_\mu^2)/m$.

By (46) and Lemma 7,

$$\sup_{\lambda \ge 0} |\|f_\lambda - v_0\|_z - \|f_\lambda - v_0\|_{P_X}| \le D_m + \Gamma_X, \qquad \mathbb{E}\Gamma_X \le G_m.$$

Both norms are at most H. Transferring the two squared norms to population risk therefore costs at most $4H(D_m + \Gamma_X)$. Taking expectations proves (48). Nonnegativity follows from the definition of L_*. □

The bound controls excess MSE with leading constant one. It complements Theorem 3, whose RMSE bound can be faster when $\mathcal{E}^\star_{\text{cont}} = 0$.

C.4 Combining Estimation and Selection Errors

Corollary 3 (Validation-selected variance estimation)**.** *Condition on $\mathscr{G}_n$ and assume the conditions of Theorem 3, with the sampling construction above. Suppose $B_* > 0$ and $\kappa \geq 0$ satisfy, at every anchor,*

$$\mathbb{E}_{(P_\delta^{x_k})^{\otimes 2}}|Z - Z'|^q \leq B_*, \qquad \mathrm{KL}(P_\delta^{x_k} \| Q_{\delta,n,N,T}^{x_k}) \leq \kappa,$$

and $v_0(x_k) \in [\ell, V_{\max}]$. Define

$$(\mathcal{I}v_0)(x) = \sum_k w_k(x) v_0(x_k).$$

Then

$$\begin{aligned} &\left\{\mathbb{E}_{\mathrm{MC,val}} \|f_{\widehat{\lambda}} - v_0\|_{L^2(P_X)}^2\right\}^{1/2} \\ &\quad \leq \sqrt{2}\left[C_q e^\delta B_*^{2/q}(\kappa + M^{-1})^{(q-2)/q} + \|\mathcal{I}v_0 - v_0\|_{L^2(P_X)}\right] \\ &\qquad + C_1\sqrt{\frac{\log(2Km+4)}{m}} + C_2 m^{-(q-2)/4} + \sqrt{3\eta_{\mathrm{opt}}/2}, \end{aligned} \tag{49}$$

where $\mathbb{E}_{\mathrm{MC,val}}$ averages over Monte Carlo sampling and $\mathcal{V}_B$, conditionally on $\mathscr{G}_n$.

Proof of Corollary 3. Set

$$u = \kappa + M^{-1}, \qquad R_{\mathrm{bd}} = (B_*/u)^{1/q}, \qquad \lambda_{\mathrm{bd}} = R_{\mathrm{bd}}^2.$$

At every anchor, apply Theorem 1 with common radius R_{bd} and moment budget B_*. Since $v_0(x_k) \in [\ell, V_{\max}]$, projection cannot increase the anchor error, so

$$\left\{\mathbb{E}_{\mathrm{MC}}|u_k(\lambda_{\mathrm{bd}}) - v_0(x_k)|^2\right\}^{1/2} \leq C_q e^\delta B_*^{2/q} u^{(q-2)/q}.$$

The path contains $f_{\lambda_{\mathrm{bd}}}$, including its saturated value when $\lambda_{\mathrm{bd}} > \lambda_{\max}$. Hence

$$\mathcal{E}_{\mathrm{cont}}^\star \leq \|f_{\lambda_{\mathrm{bd}}} - v_0\|_{L^2(P_X)}.$$

Convexity of squared error and nonnegative interpolation weights with $\sum_k w_k(x) = 1$ give

$$\mathbb{E}_{\mathrm{MC}}\|f_{\lambda_{\mathrm{bd}}} - \mathcal{I}v_0\|_{L^2(P_X)}^2 \leq C_q^2 e^{2\delta} B_*^{4/q} u^{2(q-2)/q}.$$

Taking square roots and applying Minkowski's inequality yields

$$\left\{\mathbb{E}_{\mathrm{MC}}(\mathcal{E}_{\mathrm{cont}}^\star)^2\right\}^{1/2} \leq C_q e^\delta B_*^{2/q} u^{(q-2)/q} + \|\mathcal{I}v_0 - v_0\|_{L^2(P_X)}.$$

Apply Theorem 3 conditionally on each realization of the model samples. Its remainder is uniform over these realizations. Taking the L^2 norm over Monte Carlo sampling and using Minkowski's inequality proves (49). □

Under the same assumptions, let ε_m be the right-hand side of (48) with A_c replaced by $(K-1)H$. Then

$$\begin{aligned} &\mathbb{E}_{\mathrm{MC,val}}\|f_{\widehat{\lambda}} - v_0\|_{L^2(P_X)}^2 \\ &\quad \leq \left[C_q e^\delta B_*^{2/q}(\kappa + M^{-1})^{(q-2)/q} + \|\mathcal{I}v_0 - v_0\|_{L^2(P_X)}\right]^2 + \varepsilon_m. \end{aligned}$$

To see this, apply Proposition 5 conditionally on the model samples and average over Monte Carlo randomness. The bound on $\mathbb{E}_{\mathrm{MC}}(\mathcal{E}_{\mathrm{cont}}^\star)^2$ in the preceding proof gives the result. This preserves a leading constant of one on the squared estimation and interpolation bound.

D Experimental Details and Additional Results

D.1 Experimental Setup

Data. For synthetic regression, we generate $Y_0 = \mu_0(X) + \sigma_0(X)\xi$ using the functions in Table 2, following Song & Jeon (2026), where X is uniform on the listed domain and independent of ξ. We use $\mathcal{N}(0,1)$, $t_3/\sqrt{3}$, and $C\varepsilon/\sqrt{2.98}$, where $\varepsilon \sim \mathcal{N}(0,1)$, $\mathbb{P}(C=1) = 0.98$, $\mathbb{P}(C=10) = 0.02$, and $C \perp \varepsilon$. All three noise laws have unit variance, so $v_0(x) = \sigma_0(x)^2$. For Student's t_3 noise, the moment assumptions hold for $2 < q < 3$. Fitting, validation, and test samples follow the same law within each synthetic regression condition.

Table 2: Synthetic datasets.

Dataset	X domain	$\mu_0(x)$	Standard deviation $\sigma_0(x)$
Sine	$[0, 1]$	$2\sin(4\pi x)$	$1.25 + \sin(6\pi x)$
Cubic	$[-3, 3]$	$0.3x^3$	$\begin{cases} 1, & x < -1, \\ 3, & -1 \le x < 1, \\ 10, & x \ge 1 \end{cases}$
$x \sin x$	$[-5, 5]$	$x \sin x$	$0.3(1 + \lvert x \rvert)$
Curve	$[0, 3]$	$\log(1 + x^2) + 0.5\sin(3x)$	$0.2x + 0.2$

Table 3: Sample sizes and repetitions. We use $M = 2,000$ unless varied. Repetitions denote independent fits / Monte Carlo repetitions per fit.

Study	n_{fit}	Monitor	n_{val}	Anchors	Test X	Repetitions
Synthetic (basic)	10,000	2,500	2,500	128	2,500	5/3
Synthetic (expanded)	40,000	2,500	8,192	128	2,500	5/3
Mixture	40,000	4,000	2,500	161	4,096	5/1

Table 4: Model settings.

Component	Diffusion	MC Dropout / DeepHetero
Hidden layers	Four layers of width 128	Three layers of width 128
Activation / dropout	ReLU / 0	ReLU / 0.1 (MC Dropout), 0 (DeepHetero)
Conditioning	Input covariate, response, and time embedding	Input covariate
Training objective	Noise-prediction loss	Gaussian negative log likelihood
Optimizer / rate	AdamW, weight decay 10^{-5}; 0.002 with cosine decay	Adam; 0.001
Batch / duration	1,024 samples; 16,000 updates	256 samples; 500 epochs
Denoising draws	Eight per sample	————
Gradient clipping	Norm threshold 10	Norm threshold 1
Reported weights	Final-update EMA, decay 0.999	Final-epoch weights

Samples and preprocessing. Table 3 gives sample sizes and repetitions. Each Monte Carlo repetition uses M model samples per anchor, a fixed covariate where variance is estimated before interpolation. Diffusion variance estimators share these model samples. We standardize covariates using the fit-sample mean and standard deviation. Working response units are $(Y_0 - c_{\text{tr}})/a_{\text{tr}}$, where c_{tr} is the fit-sample median and $a_{\text{tr}} = \max\{\text{IQR}(Y_0)/1.3489795003921634, 10^{-6}\}$. Monitoring samples are used only for diagnostics.

Models. Table 4 gives the training settings. The noise-prediction network predicts $-\varepsilon$, with $s_\theta(t, z, x) = n_\theta(t, z, x)/\sqrt{1 - e^{-t}}$. Its time embedding uses eight sinusoidal frequencies of normalized $\log\sqrt{1 - e^{-t}}$. MC Dropout (Gal & Ghahramani, 2016) and DeepHetero (Nix & Weigend, 1994) predict Gaussian means and variances. Both use $n_{\text{fit}} + n_{\text{val}}$ responses and fit-sample preprocessing. MC Dropout estimates $\mathbb{E}_{\text{mask}}\sigma^2 + \text{Var}_{\text{mask}}(\mu)$ from 2,000 mask passes. DeepHetero uses its variance head.

Table 5: Radius selection.

Setting	Values
Moment indices	$q \in \{2.5, 3, 4\}$
KL budget	$\kappa = 0.01$
Model samples	$M \in \{64, 128, 256, 512, 1000, 2000, 4000, 8000, 16000\}$
Validation samples	$n_{\mathrm{val}} \in \{100, 250, 500, 1000, 2500, 5000, 10000\}$
Fixed quantities	$n_{\mathrm{val}} = 2,500$ for the M sweep and $M = 2,000$ for the validation sweep
Repetitions	20,000 per law

Time grid and score loss. We use $T = 5$, $\delta = 0.001$, 868 training intervals, and 217 ExpInt sampling intervals. Sampling starts from $\mathcal{N}(0, 1)$ in working units. Of N grid intervals, $\mathrm{round}(0.15N)$ are geometric across 1.5 decades above δ, and the rest are uniform in $t^{1/8.3587}$ up to T. On the training grid, set $Z_{\mathrm{time}} = \sum_k \omega_k/(1 - e^{-t_k})$ with $\omega_k = h_k/(T - \delta)$. We sample k with probability $p_k = \omega_k/\{(1 - e^{-t_k})Z_{\mathrm{time}}\}$ and multiply the negative-epsilon squared loss by Z_{time}. Its expectation equals the grid-based score DSM objective in (5).

Radius selection. Common-radius selection forms disjoint adjacent-covariate pairs from $2m = n_{\mathrm{val}}$ validation samples and minimizes (18) over $\lambda = R^2$. It reuses the same model samples and uses a float64 objective-gap tolerance of 10^{-4}. Predictions use linear interpolation with constant boundary extension.

D.2 Controlled Experiments

Estimation bounds. Figure 2 uses specified laws without model fitting or validation. Let $G_0 = \delta_0$ and

$$G_\varepsilon = (1 - \varepsilon)\delta_0 + \frac{\varepsilon}{2}(\delta_{\varepsilon^{-1/q}} + \delta_{-\varepsilon^{-1/q}}), \qquad 0 < \varepsilon \le 1,$$

where δ_a denotes a point mass at a. The pairwise qth moment of G_ε is at most 2^q. Finite-M controls use $P = Q = G_{M^{-1}}$, isolating Monte Carlo error. Mismatch controls use the same $Q = G_{1-e^{-\kappa}}$ with targets δ_0 and Q. Model samples cannot distinguish these targets, although their variances differ. We compute clipped RMSE at $R_{\mathrm{bd}} = \{2^q/(\kappa + M^{-1})\}^{1/q}$ and the maximum of the two separate two-point lower bounds, without OU evolution or output projection. We vary $q \in \{2.5, 3, 4\}$, $\kappa \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, and $M \in \{2^4, \ldots, 2^{18}\}$.

Validation size. To separate radius-selection error from changes in the laws, we keep each target–model pair fixed as M and n_{val} vary. For each $q \in \{2.5, 3, 4\}$, the target has mass $1 - p$ at zero and $p/2$ at each of $\pm a$, with $p = 0.02$ and $a = p^{-1/q}$. The model has the same support and tail mass p_Q satisfying

$$p \log(p/p_Q) + (1 - p) \log\{(1 - p)/(1 - p_Q)\} = 0.01.$$

The roots $p_Q \approx 0.0465159211$ and 0.0060767391 give excess and deficient model variance, respectively. The target budget is $B = \{2p(1 - p) + 2^{q-1}p^2\}a^q$. All moments are finite, and q specifies the budget used in R_{bd}. We compare R_{bd}, $R = 4$, and common-radius selection with one anchor, covariates independent of responses, and $\delta = 0$. With one anchor, the selected variance estimate is the projection of $m^{-1} \sum_i U_i$ onto $[0, \text{model sample variance}]$, where U_i is half the squared response difference within a validation pair. Radius rules share model-sample counts, accumulated across sample sizes within each repetition. Validation uses independent disjoint pairs. Fixed-radius risks are analytic. Selected-radius RMSE and its Monte Carlo integration intervals use the repetitions in Table 5.

Exact bias and MSE calculations. To explain when clipping reduces error, Figure 4 uses the preceding two law pairs at $q = 3$, $M = 2,000$, $\kappa = 0.01$, and $\delta = 0$, without projection. For $A \in \{P, Q\}$, write $C_R(A) = \mathbb{E}_{A^{\otimes 2}}\psi_R(Y, Y')$. The total bias b_R is the sum of clipping bias $b_{\mathrm{clip}}(R)$

and law-mismatch bias $b_{\mathrm{law}}(R)$,

$$b_R = \underbrace{C_R(P) - \mathrm{Var}_P(Y)}_{b_{\mathrm{clip}}(R)} + \underbrace{C_R(Q) - C_R(P)}_{b_{\mathrm{law}}(R)},$$

$$\mathbb{E}_{Q^{\otimes M}}[(\widehat{C}_{R,M} - \mathrm{Var}_P(Y))^2] = b_R^2 + \mathrm{Var}_{Q^{\otimes M}}(\widehat{C}_{R,M}).$$

For independent $Z_1, Z_2 \sim Q$,

$$\mathrm{Var}_{Q^{\otimes M}}(\widehat{C}_{R,M}) = \frac{2\,\mathrm{Var}\{\psi_R(Z_1, Z_2)\} + 4(M-2)\,\mathrm{Var}\{\mathbb{E}[\psi_R(Z_1, Z_2) \mid Z_1]\}}{M(M-1)}.$$

We evaluate these expectations by finite sums, retaining the dependence between pairs that share a response. The reference $R_{\mathrm{bd},0} = \{B/(\kappa + 1/2000)\}^{1/3} \approx 5.79$ uses $B = 2.04$ and balances the bound, not the exact risk.

D.3 Learned Mixture Experiments

The model is trained on $Q_{0,w}^{\mathrm{mix},x}$ from Section 5.1, while validation follows P_0^x and the target remains $v_0(x)$. We apply raw-response radii to $\widehat{Y}_{\delta,j}^{\mathrm{raw}} = a_{\mathrm{tr}}\widehat{Y}_{\delta,j} + e^{-\delta/2}c_{\mathrm{tr}}$ and use the corrected variance $e^{\delta}\widehat{C}_{R,M}^{\mathrm{raw}}(x) - a_{\mathrm{tr}}^2(e^{\delta} - 1)$. Clipped predictions are projected onto $[0, 100]$ before interpolation, with the upper cap inactive. The unclipped comparator has no upper cap.

The fitted sampler captures little rare-tail mass at the smallest mixture weight. At $w = 0.005$, the frequency of $|\widehat{Y}_{\delta}^{\mathrm{raw}}| > 5e^{-\delta/2}\varsigma_0(x)$ is 1.37×10^{-5}, versus 0.005 for the specified mixture. Thus the learned sweep does not reproduce the counterexample's variance divergence. We estimate $\mathbb{E}_X\,\mathrm{KL}(P_{\delta}^X \| Q_{\delta,n,N,T}^X)$ using the five-neighbor estimator of Wang et al. (2009) and trapezoidal integration over covariates. We check sensitivity using three and ten neighbors and sample sizes of 500, 1,000, and 2,000. Equal-law errors are of order 10^{-3}, comparable to the smallest estimates. Negative estimates are retained. The analytic KL curve applies to the OU-smoothed specified mixture, not the fitted sampler.

A direct pairwise smoother has lower RMSE than common-radius selection in all five paired repetitions at $w = 0.05$, with RMSE 0.1967 versus 0.3250. It uses the same 2,500 target-validation samples and no model samples, applying Gaussian kernel smoothing to U_i at pair midpoints. Five-fold cross-validation over disjoint pairs selects its bandwidth from $\{0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3\}$ in raw covariate units.

D.4 Synthetic Regression

Tables 6 and 7 report condition-level results. Anchors are midpoints of equal covariate intervals, and validation data are fixed across Monte Carlo repetitions within each fitted model. After OU correction, variance is projected onto $[10^{-4}, 100]$ in working units, multiplied by a_{tr}^2, and interpolated. Baselines use the same output interval and interpolation.

Metrics. Mean and variance RMSE compare predictions with $\mu_0(X)$ and $v_0(X)$. Sample RMSE compares the predicted mean with an independent test response Y_0. Diffusion means are $c_{\mathrm{tr}} + a_{\mathrm{tr}}e^{\delta/2}M^{-1}\sum_a \widehat{Y}_{k,a}$. We average squared errors within fits, across fits, and equally across conditions before taking the square root. Variance NRMSE first divides each condition–fit MSE by the empirical mean of $v_0(X)^2$ on its test covariates.

QICE (Han et al., 2022) uses ten model-quantile bins and 64 independent target responses Y_0 at each of 128 anchors, with 2,000 predictive samples. Diffusion samples are $c_{\mathrm{tr}} + a_{\mathrm{tr}}e^{\delta/2}\widehat{Y}_{k,a}$ at $\delta = 0.001$ and retain residual OU noise. Baselines use their time-zero predictive laws. We report $100\,\mathrm{QICE}$, where $\mathrm{QICE} = \frac{1}{10}\sum_{b=1}^{10} |\widehat{\pi}_b - 1/10|$ and $\widehat{\pi}_b$ is the target-response proportion in bin b across anchors. QICE is averaged within fits, across fits, and across conditions. It measures coverage aggregated across anchors and does not control conditional-variance error.

Model comparison. Under t_3 and contaminated noise, diffusion has lower QICE than both baselines in every condition, but its variance RMSE is not uniformly lower. At the basic budget, clipping worsens all three Cubic conditions.

Table 6: Regression performance by condition. Entries use the basic / expanded budgets. G denotes Gaussian noise and C contaminated noise. Diffusion variance RMSE uses common-radius selection.

	Variance RMSE			QICE (%)		
Case	Diffusion	MC Dropout	DeepHetero	Diffusion	MC Dropout	DeepHetero
Sine, G	0.2246/0.1245	0.2904/0.2093	0.2913/0.1977	0.2702/0.3215	0.4997/0.5337	0.6521/0.3688
Sine, t_3	0.5202/0.4357	0.5318/0.3182	0.5830/0.3730	0.4182/0.4296	3.0761/3.2576	2.8295/2.9943
Sine, C	0.9636/0.6701	0.7626/0.5655	0.8582/0.5824	0.4120/0.4077	4.1835/4.1662	4.0341/3.8880
Cubic, G	6.0466/4.0645	5.5242/5.0876	8.0313/5.0809	0.3022/0.3114	0.4331/0.3651	0.4595/0.3785
Cubic, t_3	12.1965/10.1338	9.5083/6.6306	10.0706/11.7930	0.3023/0.2983	3.2733/3.1794	3.0435/3.0080
Cubic, C	20.8978/13.7434	15.4412/12.4525	15.6968/13.1304	0.2911/0.3241	4.0885/4.3128	4.0322/3.8503
$x \sin x$, G	0.1281/0.0600	0.1741/0.1283	0.1183/0.0688	0.2720/0.2690	0.3944/0.4366	0.3626/0.2674
$x \sin x$, t_3	0.2874/0.2595	0.2377/0.1555	0.2930/0.2406	0.3607/0.3306	3.1042/3.1020	2.7181/2.8038
$x \sin x$, C	0.4262/0.3022	0.6277/0.2497	0.4809/0.2808	0.4773/0.4452	4.5362/4.3470	4.1023/4.1538
Curve, G	0.0243/0.0153	0.0214/0.0145	0.0211/0.0110	0.2429/0.2653	0.4607/0.3646	0.3659/0.2617
Curve, t_3	0.0559/0.0420	0.0488/0.0318	0.0594/0.0327	0.2821/0.3145	3.0669/3.1162	2.9850/3.0296
Curve, C	0.0968/0.0624	0.0876/0.0431	0.1043/0.0528	0.3122/0.3220	4.1551/4.3669	3.5121/4.1852

	Mean RMSE			Sample RMSE		
Case	Diffusion	MC Dropout	DeepHetero	Diffusion	MC Dropout	DeepHetero
Sine, G	0.0804/0.0400	0.1226/0.0968	0.1148/0.0796	1.4403/1.4383	1.4427/1.4404	1.4422/1.4396
Sine, t_3	0.0649/0.0394	0.1275/0.0997	0.1105/0.0616	1.5033/1.5030	1.5080/1.5059	1.5076/1.5045
Sine, C	0.0571/0.0398	0.1217/0.0938	0.1401/0.0956	1.4397/1.4388	1.4441/1.4419	1.4446/1.4423
Cubic, G	0.2955/0.1744	0.3722/0.2312	0.2987/0.1567	6.0028/5.9985	6.0057/5.9960	6.0063/5.9950
Cubic, t_3	0.2544/0.1799	0.3539/0.2347	0.2840/0.1769	5.6192/5.6187	5.6223/5.6186	5.6217/5.6154
Cubic, C	0.2446/0.1693	0.2400/0.3007	0.2344/0.1613	5.8196/5.8181	5.8203/5.8275	5.8208/5.8166
$x \sin x$, G	0.0499/0.0336	0.1224/0.0996	0.1069/0.0697	1.1407/1.1403	1.1435/1.1438	1.1455/1.1417
$x \sin x$, t_3	0.0547/0.0359	0.1295/0.0998	0.1086/0.0758	1.1058/1.1050	1.1122/1.1088	1.1089/1.1074
$x \sin x$, C	0.0585/0.0330	0.1139/0.1220	0.1303/0.0670	1.0885/1.0872	1.0922/1.0949	1.0951/1.0899
Curve, G	0.0215/0.0136	0.0377/0.0271	0.0374/0.0211	0.5335/0.5329	0.5341/0.5333	0.5337/0.5333
Curve, t_3	0.0211/0.0146	0.0411/0.0344	0.0333/0.0239	0.4810/0.4803	0.4823/0.4815	0.4812/0.4804
Curve, C	0.0205/0.0139	0.0335/0.0328	0.0414/0.0356	0.5231/0.5227	0.5240/0.5235	0.5236/0.5234

Table 7: Variance NRMSE at the basic / expanded budgets.

Function	Noise	Unclipped	$R = 4$	Common	MC Dropout	DeepHetero
Sine	Gaussian	0.0725/0.0470	0.0703/0.0456	0.0824/0.0456	0.1062/0.0765	0.1068/0.0724
Sine	t_3	0.2024/0.1939	0.2335/0.2197	0.1871/0.1571	0.1913/0.1150	0.2103/0.1345
Sine	Contam.	0.3028/0.2570	0.4648/0.4582	0.3517/0.2438	0.2775/0.2052	0.3123/0.2119
Cubic	Gaussian	0.0777/0.0724	0.3571/0.3498	0.1033/0.0697	0.0944/0.0871	0.1377/0.0874
Cubic	t_3	0.1765/0.1398	0.5973/0.5984	0.2111/0.1755	0.1640/0.1151	0.1745/0.2054
Cubic	Contam.	0.2731/0.2352	0.7179/0.7125	0.3632/0.2370	0.2663/0.2150	0.2707/0.2270
$x \sin x$	Gaussian	0.0651/0.0378	0.0650/0.0378	0.0802/0.0378	0.1101/0.0810	0.0746/0.0434
$x \sin x$	t_3	0.2330/0.2625	0.1452/0.1307	0.1809/0.1634	0.1487/0.0977	0.1838/0.1516
$x \sin x$	Contam.	0.3865/0.2248	0.2522/0.2730	0.2685/0.1907	0.3931/0.1579	0.3030/0.1776
Curve	Gaussian	0.0601/0.0389	0.0599/0.0387	0.0739/0.0467	0.0649/0.0441	0.0640/0.0334
Curve	t_3	0.2806/0.3040	0.1736/0.1518	0.1697/0.1272	0.1478/0.0959	0.1792/0.0987
Curve	Contam.	0.2848/0.2086	0.3464/0.3231	0.2952/0.1900	0.2657/0.1312	0.3160/0.1604

Clipping gain and selection loss. We compare the improvement available at the best common radius with the loss from validation selection (Figure 5). For a fitted model and Monte Carlo repetition b, let $f_{b,R}$ be the final variance estimator. We use the same projection and interpolation at every radius, including $R = \infty$, and define

$$L_b(R) = \mathbb{E}_X[(f_{b,R}(X) - v_0(X))^2], \quad L_b^\star = \inf_{R \geq 0} L_b(R),$$
$$G_b = L_b(\infty) - L_b^\star, \qquad S_b = L_b(\widehat{R}_b) - L_b^\star.$$

Table 8: Clipping gain and selection loss under repeated validation. Population MSE is averaged within fits (three Monte Carlo repetitions), then across five independent fits. Entries for Selected, S, and S^{cp} are also averaged over validation repetitions. Parentheses give SE across fits for G, S, and S^{cp}. Curve entries are in units of 10^{-3}.

Condition	Budget	Unclipped	Best common	Selected	G (SE)	S (SE)	S^{cp} (SE)
Cubic Gaussian	Basic	20.250	18.656	26.796	1.59 (0.95)	8.14 (0.35)	0.00727 (0.0021)
Cubic Gaussian	Expanded	17.475	15.893	18.687	1.58 (0.77)	2.79 (0.21)	0.00122 (0.00015)
Cubic t_3	Basic	105.295	92.949	143.486	12.3 (7.9)	50.5 (5.1)	0.0325 (0.014)
Cubic t_3	Expanded	65.868	63.846	85.084	2.02 (0.57)	21.2 (1.8)	0.00952 (0.004)
Cubic Contam.	Basic	253.938	252.445	331.493	1.49 (1.2)	79 (6.7)	0.014 (0.0086)
Cubic Contam.	Expanded	189.944	189.406	212.920	0.538 (0.48)	23.5 (2.1)	0.0074 (0.0027)
Curve t_3	Basic	8.254	2.206	3.913	6.05 (2.7)	1.71 (0.22)	0.000991 (0.00042)
Curve t_3	Expanded	9.771	1.126	2.178	8.64 (4.6)	1.05 (0.19)	0.000255 (0.000044)

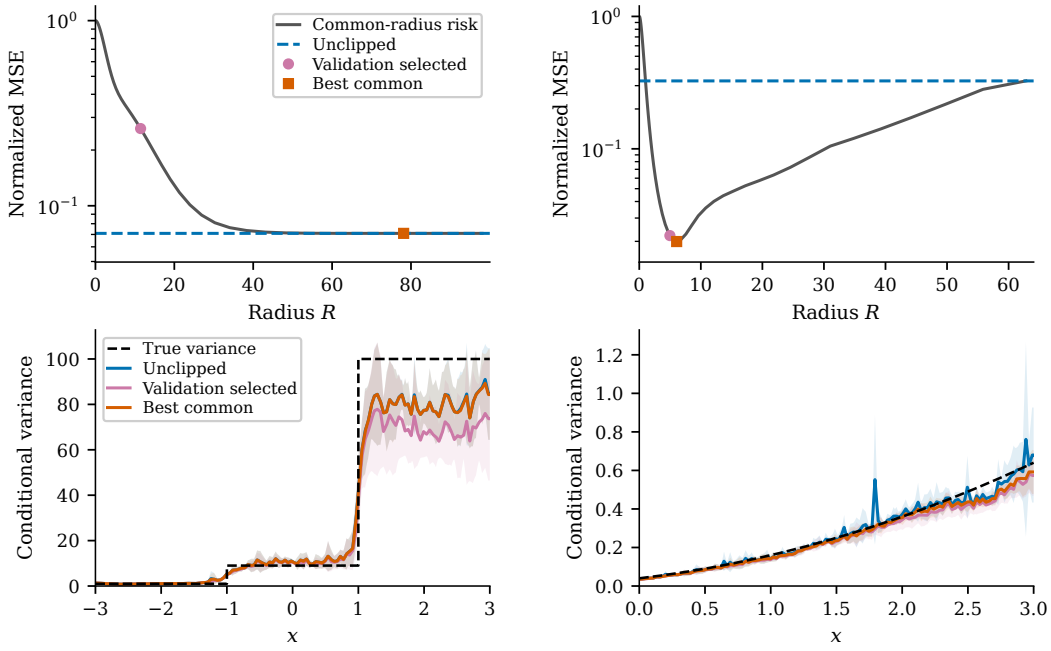


Figure 5: Clipping gains and selection loss. Cubic with contaminated noise (left) and Curve with t_3 noise (right), at the basic budget with the original validation sample of each fit. Top: risk for the first fit and Monte Carlo repetition, normalized by $\mathbb{E}_X[v_0(X)^2]$. Bottom: variance predictions averaged over three Monte Carlo repetitions within each of five fits, with ± 1 SD across fits. Best common uses the known target variance for diagnosis only.

Thus G_b is the available gain and S_b the selection loss. Since $L_b(\widehat{R}_b) - L_b(\infty) = S_b - G_b$, selected clipping improves on unclipped estimation exactly when $S_b < G_b$. The best common radius uses v_0 for diagnosis and differs from R_{bd} and the validation-selected $\widehat{R}_b$. We integrate squared error over the uniform covariate law and minimize over the full radius range, including the selected and unclipped estimates. The double-precision optimality gap, divided by $\mathbb{E}_X[v_0(X)^2]$, is at most 10^{-6}. Minimization precedes averaging. Table 8 reports this population MSE, whereas Table 1 uses 2,500 test covariates. With fitted models and model samples fixed, we draw 100 independent validation covariate sets of size n_{val} and 100 conditionally independent response realizations per set. Each response realization is shared across the three model-sample sets within a fit. We also compute S^{cp} by replacing U_i with $\bar{U}_i = \mathbb{E}[U_i \mid \mathbf{X}^{\mathrm{val}}]$, using the known data-generating law only for diagnosis.

Table 9: Sampler sensitivity. Variance NRMSE across sampling grids and early-stop times.

Sampling-grid sweep				Early-stop sweep			
Grid	Unclipped	$R = 4$	Common	δ	Unclipped	$R = 4$	Common
54	0.2500	0.2618	0.2294	0.001	0.1852	0.3895	0.2094
108	0.2081	0.2901	0.2321	0.003	0.1852	0.3893	0.2094
217	0.2144	0.3030	0.2349	0.01	0.1853	0.3884	0.2095
434	0.2069	0.3093	0.2384	0.03	0.1857	0.3862	0.2097
868	0.2109	0.3123	0.2379				

Validation response noise. To separate covariate and response variability, let A be selection loss averaged over the three Monte Carlo repetitions and define

$$V_X = \mathrm{Var}_{\mathbf{X}^{\mathrm{val}}}(\mathbb{E}[A \mid \mathbf{X}^{\mathrm{val}}]), \qquad V_Y = \mathbb{E}_{\mathbf{X}^{\mathrm{val}}}[\mathrm{Var}(A \mid \mathbf{X}^{\mathrm{val}})].$$

We estimate V_Y by the mean within-set variance and V_X by the variance of set means minus $\widehat{V}_Y/100$. We average each component across five fits before forming $100\widehat{V}_X/\widehat{V}_Y$. We jointly bootstrap set means and within-set variances within each fit, using 10,000 resamples. The approximate 95% normal intervals are pointwise and conditional on the saved fits and model samples. Negative corrected estimates are retained. Repeated-validation selection uses a double-precision criterion gap of at most $10^{-6}a_{\mathrm{tr}}^4$.

Removing response noise sharply reduces mean selection loss in all eight condition–budget combinations. The estimated ratio $100\widehat{V}_X/\widehat{V}_Y$ ranges from -0.054% to 0.189%, with all upper conditional Monte Carlo interval endpoints below 0.35%. Mean $S - G$ is positive for Cubic and negative for Curve with t_3 noise, so radius selection can erase the available clipping gain. For contaminated Cubic, substantial error remains even at the best common radius.

Sampler sensitivity. For the basic-budget models at $(T, \delta) = (5, 0.001)$, we reselect the radius at $M \in \{128, 512, 2000, 8000\}$ across all twelve conditions (Figure 3, right). The grid sweep uses the three Sine conditions. The early-stop sweep adds Cubic with Gaussian and t_3 noise, sharing trajectories on a grid containing all endpoints and recomputing correction and radius selection. Both use $M = 2,000$. Finer grids do not consistently reduce error, and changing δ has little effect after correction (Table 9).