

Random Unbiased Perturbations in Nonparametric Regression

Anna Lyubarskaja and Dominik Rothenhäusler

27 November 2025

Abstract

We study nonparametric regression with covariates X and outcome Y under random unbiased perturbations (RUPs) of the conditional distribution $Y|X$, where the marginal distribution of covariates, P^X , remains fixed but the conditional law, $P^{Y|X}$, varies randomly across datasets. Unlike adversarial distribution shift frameworks that yield conservative worst-case guarantees, RUPs induce dataset-level variance inflation rather than systematic bias. We provide examples of RUPs and show that this distributional uncertainty reduces the effective sample size to $n_{\text{eff}} = n/(1+n\tau)$, where $\tau \in [0, 1]$ quantifies the perturbation strength. For local polynomial estimators, we derive an extended bias-variance decomposition that includes a distributional variance term with the same bandwidth scaling as classical sampling variance. This leads to a modified bandwidth selection principle: when distributional uncertainty dominates sampling uncertainty ($\tau \gg 1/n$), optimal bandwidths scale as $\tau^{1/(2\beta+1)}$ rather than the usual $n^{-1/(2\beta+1)}$, where β indicates the smoothness of the function class considered. We also establish matching minimax lower bounds showing that there exists an RUP for which this effective sample size n_{eff} is fundamental. Our results demonstrate that random dataset-level perturbations create a distinct mode of uncertainty that affects both practical tuning and fundamental statistical limits.

1 Introduction

Data are rarely sampled i.i.d. from the true distribution of interest. Instead, we often observe data from a slightly shifted version of our target distribution. For instance, the true cost of living for an average household is typically approximated by consumer price index (CPI), which is constructed by choosing a basket of representative goods and services. Data can then be collected to estimate this CPI, but the discrepancy between the true average cost of living and the CPI depends on the choice of goods and services included in the CPI basket, and is not reduced by collecting additional data. Similarly, clinical outcomes observed at a single hospital differ slightly from regional population-level outcomes as they depend on the hospital's particularities, such as specific staff or protocols. As a final example, flight delays on a given day are influenced by that day's weather patterns and will thus deviate from a long-run average distribution of delays. In expectation over many such days, hospitals or consumer baskets, one recovers a true target distribution P_0 , but any single dataset is generated under just one realization P_ξ , a shifted version of the truth.

In the examples above, the distribution of covariates X can be regarded as fixed (consumer spending patterns, patient populations, scheduled flights), while the conditional law of outcomes $Y|X$ is perturbed by measurement procedure or environment. We refer to such shifts as *random unbiased perturbations* (RUPs): random, mean-zero deviations from a target law that preserve the covariate distribution. We define this notion formally in Definition 1 (Section 2).

Our RUP framework thus involves two distinct layers of randomness. First, Ξ represents a distribution over possible perturbations, where a random realization $\xi \sim \Xi$ parameterizes the distributional shift. Second, conditional on the chosen ξ , the data $\{(X_i, Y_i)\}_{i=1}^n$ are sampled i.i.d. with $X_i \sim P^X$ as usual and $Y_i | X_i \sim P_\xi(\cdot | X_i)$. This perspective introduces a new source of uncertainty alongside the sampling error, as now the discrepancy between the observed empirical law, \hat{P}_ξ , and the target, P_0 , naturally decomposes as

$$\hat{P}_\xi - P_0 = (\hat{P}_\xi - P_\xi) + (P_\xi - P_0). \quad (1)$$

The first term is the a familiar sampling error due to observing only n samples from a given distribution, while the second term reflects a distributional shift from the perturbation $P_\xi \neq P_0$.

Much of the distribution shift literature relies on distributionally robust optimization, and treats the conditional $Y|X$ shift in the second term in (1) adversarially. P_ξ is assumed to lie within a divergence ball

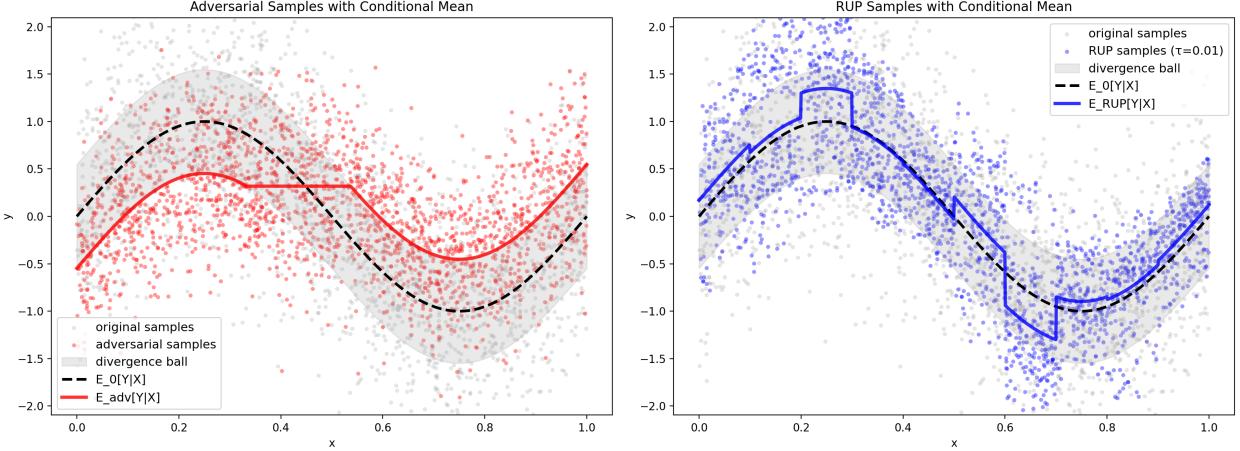


Figure 1: Examples of points sampled adversarially (left, red) and from an RUP (right, blue). The two distribution shifts have the same KL divergence κ from the baseline distribution P_0 (sampled in gray). However, the adversarial conditional mean (left, red) provides a much poorer approximation of the target P_0 conditional mean (black, dashed), whereas the RUP conditional mean (right, blue) remains much closer to the true relationship. The shaded gray divergence ball illustrates the range of conditional means that distributions at KL divergence κ from P_0 can take.

around P_0 , and statistical methods are conservatively designed to be valid for all distributions within this ball[6, 5, 14]. In contrast, we study perturbations that are random and mean-zero. Figure 1 illustrates why the DRO approach is too conservative for the study of RUPs. We simulate two different $Y|X$ shifts of a baseline distribution P_0 , each with the same KL divergence κ . On the left we visualize an adversarially shifted distribution, while the right side pictures a representative sample of an RUP. A DRO approach is optimized for adversarial shifts, which may distort the conditional mean systematically producing biased approximations of $\mathbb{E}_0[Y|X]$. RUPs, on the other hand, introduce only random, mean-zero fluctuations in statistics such as $\mathbb{E}[Y|X]$ that average out and stay near the target function.

To build intuition for the behavior of RUPs, recall the flight delay example. Let X represent the route information and Y the delay. The marginal distribution of routes is fixed by the flight schedule, but a perturbation ξ (i.e. a single day's weather) changes the conditional distribution of delays. These shifts are not uniform across all X : a snowstorm may delay flights in and out of Chicago, while flights elsewhere remain largely unaffected. Thus, RUPs may induce correlated changes in $Y|X$ across subsets of X , reflecting the geography of the covariates. We summarize the overall impact of an RUP by a perturbation strength parameter τ , which we later decompose into a variance scale δ^2 and an X -dependency parameter $\bar{\rho}$ (Definition 1).

An extended bias–variance decomposition. Under the RUP setting, the distributional shift term in (1), $P_\xi - P_0$, appears as an additional variance term in the error. Consider for instance the pointwise risk. Let $f^0(x) = \mathbb{E}_0[Y|X=x]$ denote the baseline regression function under P_0 , and let $f^\xi(x) = \mathbb{E}_\xi[Y|X=x]$ denote the regression function under a perturbed law P_ξ , with estimator \hat{f}_n^ξ constructed from n samples drawn i.i.d. from P_ξ . We let the subscript Ξ denote "with respect to the distribution over perturbations $\xi' \sim \Xi$ ", and the subscript ξ denote "conditional on the realization ξ ".

Applying the law of total variance to $\hat{f}_n^\xi(x_0)$ yields

$$\mathbb{E}_\Xi \mathbb{E}_\xi [(\hat{f}_n^\xi(x_0) - f^0(x_0))^2] = \underbrace{\left(\mathbb{E}_\Xi \mathbb{E}_\xi [\hat{f}_n^\xi(x_0)] - f^0(x_0) \right)^2}_{\text{bias squared}} + \underbrace{\mathbb{E}_\Xi [\text{Var}_\xi(\hat{f}_n^\xi(x_0))] + \text{Var}_\Xi (\mathbb{E}_\xi[\hat{f}_n^\xi(x_0)])}_{\text{sampling variance}} + \underbrace{\text{Var}_\Xi (\mathbb{E}_\xi[\hat{f}_n^\xi(x_0)])}_{\text{distributional variance}}. \quad (2)$$

The first two terms coincide with the usual bias variance decomposition under i.i.d. sampling from P_0 . The third term is new: it captures dataset-level variability introduced by the perturbation.

Our contributions. We study nonparametric regression under RUPs of the conditional law $Y|X$. A central consequence of RUPs is that they reduce the effective sample size available for estimation

$$n_{\text{eff}} = \frac{n}{1+n\tau}, \quad (3)$$

where τ is a measure of strength of the RUP (as outlined in Definition 1).

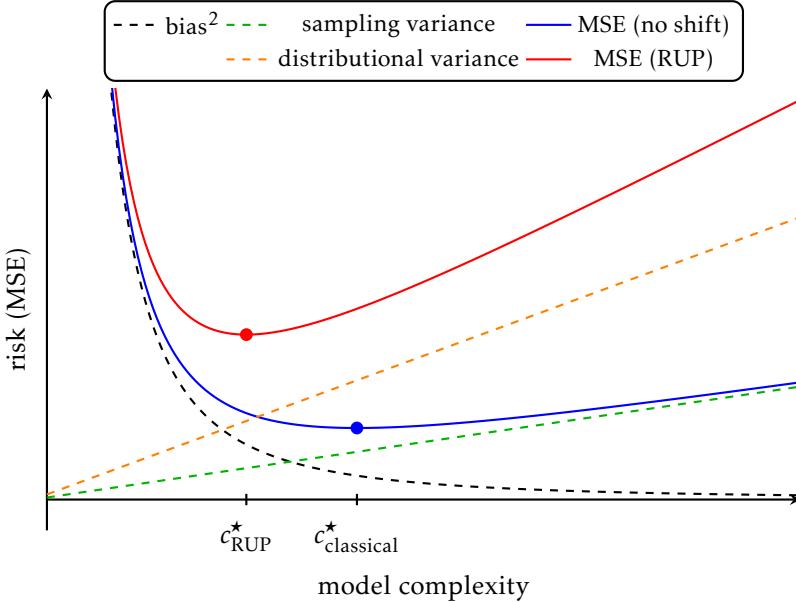


Figure 2: Schematic bias–variance tradeoff under random perturbations. Bias decreases with model complexity, while sampling variance and distributional variance both increase. The additional distributional variance raises the total RUP-adjusted curve and shifts the optimal complexity leftward (toward more regularized models).

1. **Variance inflation and optimal tuning.** For local polynomial estimators (LPEs), perturbations introduce an additional variance term with the same complexity dependence as classical sampling variance. Ignoring this term leads to overfitting and underestimation of risk. Accounting for it yields a modified tuning principle: bandwidths should be chosen by replacing n with n_{eff} . An illustration of this adjusted bandwidth tuning is presented in Figure 2. This interpretation highlights that under RUPs, distributional uncertainty acts like variance rather than bias.
2. **Fundamental limits.** Using minimax lower bounds, we show that n_{eff} also captures the fundamental information limit. We provide an example of an RUP set up where the amount of information scales with n_{eff} rather than with n , showing that distributional uncertainty fundamentally caps the rate of convergence of any estimator. In particular, we show that in the regime where the sampling and distributional variance are of the same order, i.e. the perturbation strength scales as $\tau_n \propto \frac{1}{n}$, n_{eff} captures the correct rescaling between the two for the optimal convergence rate.

Together, these results demonstrate that random unbiased perturbations of $Y | X$ create a new mode of uncertainty: neither adversarial nor per-sample, but dataset-level, and strong enough to shape both practical estimator tuning and minimax rates.

Related work. One line of work models distribution shift via distributionally robust optimization (DRO) [2, 3, 6], treating the deviation between a source P' and a target P_0 adversarially, guaranteeing performance in the worst case over all distributions within a divergence ball. This literature assumes a fixed shift and provides robust but pessimistic guarantees. If this method were applied to the RUP setting, the random, mean-zero shifts would be treated as deterministic and biased, which would lead to overly conservative bounds (see Figure 1).

A second line of work assumes benign but smooth source-target shifts [5, 13, 14]. This involves a fixed shift with a smooth or bounded density ratio, or a smoothness assumption in the classification boundary. Thus, this literature does not allow for the highly irregular density ratios that arise naturally under RUPs, as it controls the shift via regularity rather than randomness.

Closest to our work are recent models that introduce random shifts of the full joint law. These works analyze how random shift models may be used to predict $Y|X$ shift from X shift [12], how to construct confidence intervals that allow for distributional shifts [10], and how to exploit multiple shifted distributions [11]. We build on the models introduced in these works by studying the setting that only allows for $Y|X$ shift and consider the question of optimal convergence rates for nonparametric regression. Unlike classical measurement-error and surrogate-outcome literatures [4], which introduce noise at the level of individual

observations, our distributional randomness is on the dataset-level as it stems from a single realization ξ per dataset.

Finally, our minimax lower bounds connect to foundational work on nonparametric rates and information-theoretic lower bounds [9, 16, 17]. By isolating random unbiased perturbations of $Y|X$, we identify a new mode of uncertainty—neither adversarial nor per-sample—that inflates estimator variance by decreasing the effective sample size from n to n_{eff} .

The rest of the paper is organized as follows. In Section 2, we define RUPs and present the partition model (Section 2.2.1) and the correlated noise model (Section 2.2.2) as two examples for how such RUPs may be generated. In Section 3, we provide upper bounds for the convergence rate of pointwise risk when sampling from an RUP. To do this we derive (2) in the context of local polynomial estimators. In Section 4, we derive corresponding lower bounds for the correlated noise model. Finally, in Section 5 we provide numerical results from simulated RUPs, which exhibit the expected convergence rates.

2 Random Unbiased Perturbations (RUPs)

We now formalize the notion of *random unbiased perturbations* introduced above. Recall that our motivating examples involved datasets drawn from a single randomly shifted conditional law $P_\xi(Y|X)$ with a fixed covariate distribution P^X . In this section, we give a precise definition of such perturbations, specify how their strength and correlation structure are parameterized, and provide two concrete generative examples. These constructions form the foundation for the upper and lower bound results that follow.

2.1 RUP Definition

Let P_0 be a baseline distribution over (X, Y) , where $X \sim P_0^X$, and, setting $f(X) = \mathbb{E}_0[Y|X]$, we write $Y = f(X) + \varepsilon$ for some mean zero random ε . In this paper, we focus on the setting with the following simplifying assumption (although RUPs could be generalized beyond this).

Assumption 1. Suppose, under P_0 , that $X \sim \text{Unif}[0, 1]$, and that $Y = f(X) + \varepsilon$, where $X \perp \varepsilon$, $\mathbb{E}_0[\varepsilon] = 0$, and $\text{Var}_0(\varepsilon) = \sigma^2$.

Under RUPs we shift the distribution of $\varepsilon|X$, with shifts that may vary across different values of X . We let $\rho : \mathcal{X} \times \mathcal{X} \rightarrow [-1, 1]$ be a correlation function such that $\rho(x_1, x_2)$ captures how correlated the shifts are at $X = x_1$ and $X = x_2$. Moreover, we define the average correlation parameter

$$\bar{\rho} := \mathbb{E}_{X_1, X_2} [\rho(X_1, X_2)], \quad X_1, X_2 \stackrel{\text{i.i.d.}}{\sim} P_0^X.$$

Definition 1 (Random unbiased perturbation of P_0). Let P_0 be a baseline law on (X, Y) and let Ξ be a law on perturbation indices ξ . For each ξ , let P_ξ be a probability measure on (X, Y) (on the same sample space as P_0). We parameterize a shift in $Y|X$ as a shift in $\varepsilon := Y - f(X)$ given X . Define

$$\Delta_\xi(x) := \mathbb{E}_\xi[\varepsilon | X = x] - \mathbb{E}_0[\varepsilon | X = x].$$

We say that the pair $\mathcal{R} = (\Xi, (P_\xi)_\xi)$ is a random unbiased perturbation (RUP) of P_0 with variance scale δ^2 and X -correlation kernel ρ if the following hold for P_0^X -a.e. x, x' :

1. **Fixed marginal.** $P_\xi^X = P_0^X$ a.s.

2. **Centering.** $\mathbb{E}_\Xi[\Delta_\xi(x)] = 0$

3. **Variance scale.**

$$\text{Var}_\Xi(\Delta_\xi(x)) = \delta^2 \sigma^2.$$

4. **X -dependency.**

$$\mathbb{C}\text{ov}_\Xi(\Delta_\xi(x), \Delta_\xi(x')) = \rho(x, x') \delta^2 \sigma^2.$$

We say P_ξ is an RUP with strength $\tau = \delta^2 \bar{\rho}$.

Remark 2.1 (On the correlation kernel). The function $\rho(x_1, x_2)$ acts as a correlation kernel across covariate values. To be consistent with a covariance structure, ρ must be symmetric with $\rho(x, x) = 1$, and for every finite set $\{x_i\}_{i=1}^n$, $[\rho(x_i, x_j)]$ is positive semidefinite.

Finally, we note that first sampling $\xi \sim \Xi$, then n samples from P_ξ , is equivalent to sampling from the mixture distribution

$$\bar{P}^n(D) := \mathbb{E}_\Xi[P_\xi^{\otimes n}(D)]. \quad (4)$$

We now describe an example of how to generate shifts satisfying the RUP definition above.

2.2 Examples of RUP-generating processes

2.2.1 The Partition model

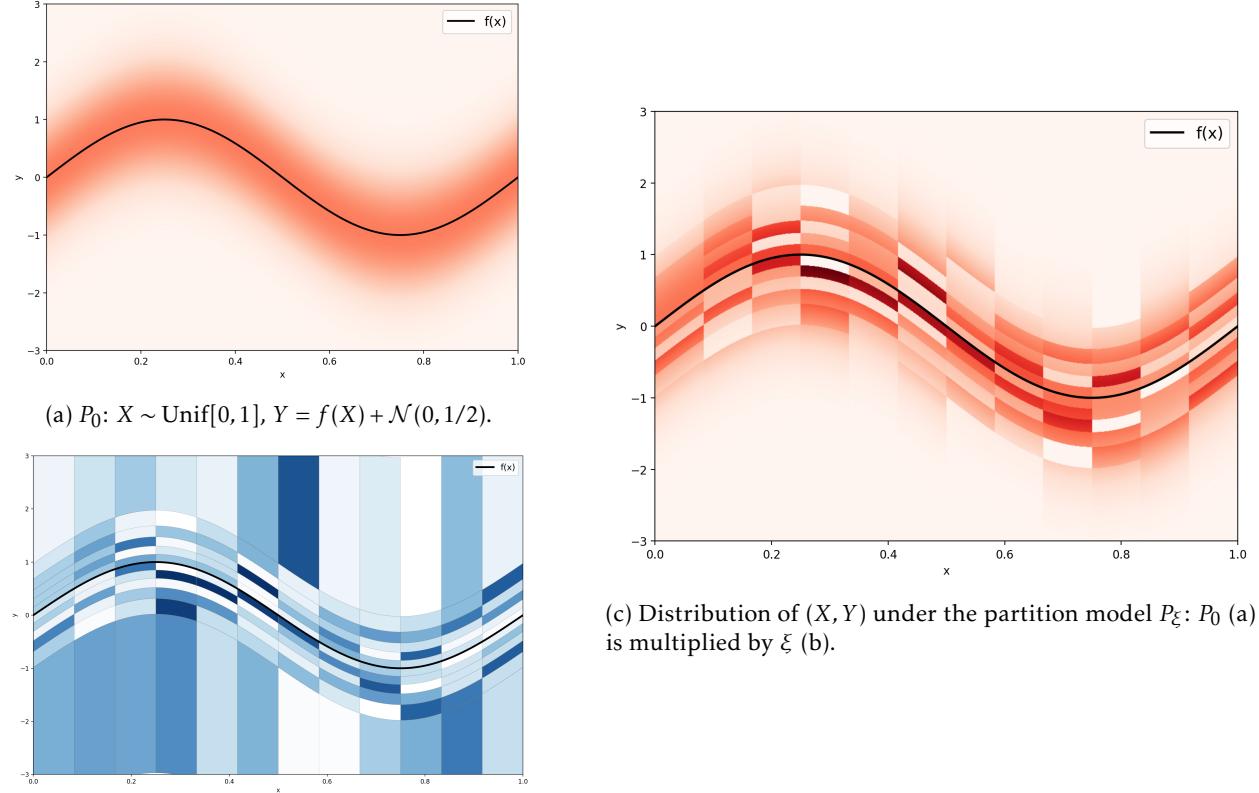


Figure 3: The partition model. The original density P_0 (shown in (a)), is split into X and ε quantiles which are each assigned i.i.d. (normalized) random weights (shown in (b)). The resulting RUP (shown in (c)) is the product of P_0 and these weights.

We provide an example of an RUP by defining a distribution Ξ over weight arrays $\xi = (\xi_{ij})$, and for each ξ a perturbed law P_ξ on (X, Y) . We construct P_ξ by specifying its density in the (X, ε) coordinates and then view it as a law on (X, Y) via $Y = f(X) + \varepsilon$. We will not distinguish notationally between the law on (X, ε) and its pushforward to (X, Y) .

This model, to which we will refer to as the *partition model*, is based on a construction first introduced by Jeong and Rothenhäusler [10]. We partition the joint distribution of (X, ε) under P_0 into quantiles and generate a random perturbation by randomly upweighting or downweighting the density in each quantile while keeping the marginal distribution over X constant. Formally, we choose integers B_X, B_ε , and set $\{r_i\}_{i=0}^{B_X}$ to be P_0^X -quantiles and $\{q_j\}_{j=0}^{B_\varepsilon}$ to be P_0^ε -quantiles. Then we form bins

$$I_i = [r_{i-1}, r_i), \quad J_j = [q_{j-1}, q_j), \quad i \in [B_X], j \in [B_\varepsilon].$$

Thus under P_0 , $P_0(X \in I_i) = 1/B_X$ and, by $\varepsilon \perp X$, $P_0(\varepsilon \in J_j | X) = 1/B_\varepsilon$. Next, draw i.i.d. positive weights $\{\xi_{ij}\}_{i \leq B_X, j \leq B_\varepsilon}$ (e.g. $\xi_{ij} \sim \text{Exp}(1)$). For $x \in I_i, \varepsilon \in J_j$ we define the perturbed noise law by

$$P_\xi(x, \varepsilon) = \frac{\xi_{ij}}{\frac{1}{B_\varepsilon} \sum_{j'=1}^{B_\varepsilon} \xi_{ij'}} P_0(x, \varepsilon) \quad (5)$$

In other words, as visualized in Section 2.2.1, we draw i.i.d. weights for each of $B_X B_\varepsilon$ quantile bins, then normalize them so that the marginal density of X is unchanged.

Remark 2.2. Equation (5) is equivalent to the multiplicative tilt form $p_\xi(y \mid X = x) = \frac{\xi_{ij}}{\sum_{j'} \xi_{ij'}} p_0(y \mid X = x)$ whenever $y - f(x) \in J_j$ and $x \in I_i$.

Proposition 2.1. Let Ξ be the law of iid positive weights $\{\xi_{ij}\}_{i \in [B_X], j \in [B_\varepsilon]}$, and let $\mathbb{E}[\xi^{-3}] < \infty$ (e.g. $\xi \sim \text{Exp}(1)$). The partition model produces a random unbiased perturbation of P_0 , $(\Xi, (P_\xi)_\xi)$ with:

$$\text{Variance scale } \delta^2 = \frac{1}{B_\varepsilon} \left(\frac{\text{Var}_\Xi(\xi_{ij})}{\mathbb{E}_\Xi[\xi_{ij}]^2} + o(1) \right), \quad X\text{-correlation } \rho(x_1, x_2) = \mathbf{1}_{I_{x_1} = I_{x_2}}, \text{ and } \bar{\rho} = \frac{1}{B_X}.$$

where I_x represents the P_0^X -quantile containing x .

See Section A.1 for proof. A dataset D of n samples from a distribution perturbed following the partition model follows the mixture model $\tilde{P}^n(D)$ as in (4).

2.2.2 Correlated noise model

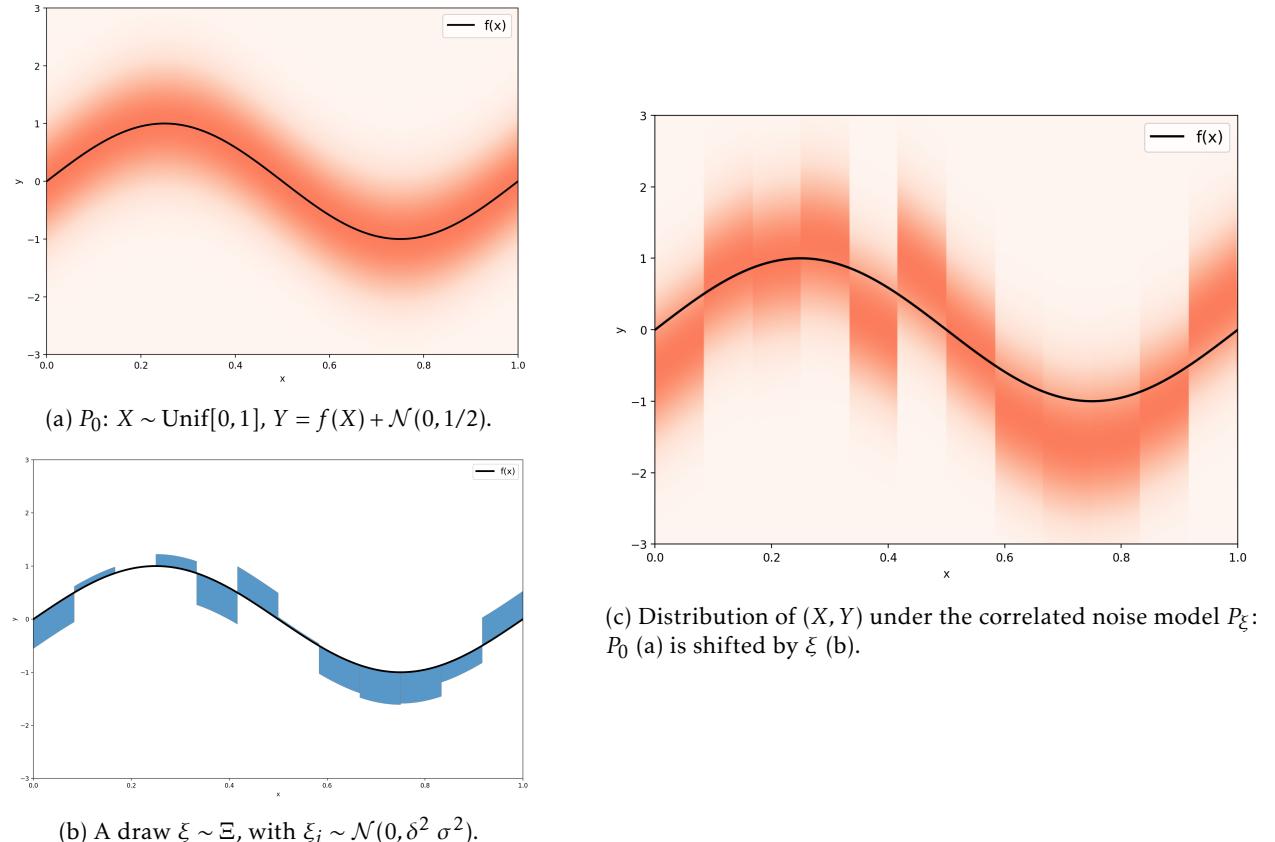


Figure 4: The correlated noise model. The original density P_0 (shown in (a)), is split into X -quantiles which are each assigned i.i.d. random shifts (shown in (b)). The resulting RUP (shown in (c)) is the original P_0 and shifted by (b).

We now introduce an alternative RUP instance, which we will refer to as *the correlated noise model*. Again, we partition the distribution of X into B_X quantiles. This time, instead of multiplicative weights, we have additive noise terms in each quantile.

Recall that $\sigma^2 = \text{Var}_0(\varepsilon)$. The perturbation is parameterized by $\xi = \{\xi_b\}_{b=1}^{B_X}$, where $\xi_b \sim^{i.i.d.} \mathcal{N}(0, \sigma^2 \delta^2)$. For observation $X = x$, let $b(x)$ be the B_X -bucket and set

$$P_\xi(x, \varepsilon) = P_0^X(x) P_0^\varepsilon(\varepsilon - \xi_{b(x)})$$

Thus, within every bucket b , under P_ξ , Y is shifted to be centered at $f(X) + \xi_b$.

Proposition 2.2. *The correlated noise model produces a random unbiased perturbation of $Y|X$ with:*

$$\text{Variance scale } \delta^2, \quad X\text{-correlation } \rho(x_1, x_2) = \mathbf{1}_{I_{x_1} = I_{x_2}}, \text{ and } \bar{\rho} = \frac{1}{B_X},$$

where I_x represents the P_0^X -quantile containing x .

See Section A.2 for proof.

3 Upper Bounds under Random Unbiased Perturbations

In this section, we recall basic properties and convergence rates of local polynomial estimators (LPEs) in the absence of distribution shift. We then analyze how LPEs behave in the RUP setting by deriving the additional distributional variance term. Finally, we discuss the consequences of this additional term for convergence rates and bandwidth tuning. We assume throughout this section that $P_0^X = \text{Unif}[0, 1]$.

3.1 Local polynomial estimator (LPE) Preliminaries

Local polynomial (LP) regression offers a flexible and widely used nonparametric method for estimating regression functions. We work under the standard local polynomial design assumptions (LP1)–(LP3) of [16], which ensure that the local Gram matrix is uniformly non-degenerate and the LP weights are well-behaved. Under an i.i.d. design, these conditions hold automatically (up to events of vanishing probability).

For a given bandwidth h , an LP estimator of order ℓ uses kernel weighting to locally fit polynomials of degree up to ℓ , adapting to the structure of the data around each target point. Such estimators admit a linear representation [See 16, Chapter 1.6 for details] and can be written as

$$\hat{f}_n(x) = \sum_{k=1}^n W_k(x) Y_k,$$

where the weights $W_k(x)$ depend on the target point x , covariates X_1, \dots, X_n , a kernel function K and the bandwidth h . In the standard no-shift setting, where samples are drawn i.i.d. from the target distribution, local polynomial estimators achieve minimax-optimal rates for pointwise risk. For a fixed bandwidth h , their bias and variance satisfy the following bounds.

Proposition 3.1 (based on Tsybakov [16], Proposition 1.13). *Let $f \in \Sigma(\beta, L)$ on $[0, 1]$ and $\hat{f}_n(x)$ be the LP(ℓ) estimator of f . Under standard assumptions (see Lemma B.1), the pointwise risk for any $x_0 \in [0, 1]$ admits the following bias variance decomposition*

$$\mathbb{E}_0[(\hat{f}_n(x_0) - f(x_0))^2] = \underbrace{\left(\mathbb{E}_0[\hat{f}_n(x_0)] - f(x_0) \right)^2}_{b^2(x_0; \hat{f}_n)} + \underbrace{\text{Var}_0(\hat{f}_n(x_0))}_{v_{\text{sampling}}^2(x_0; \hat{f}_n)},$$

where

$$b^2(x_0; \hat{f}_n) \leq C_1 h^{2\beta}, \quad v_{\text{sampling}}^2(x_0; \hat{f}_n) \leq \frac{C_2 \sigma^2}{nh} + O\left(\frac{h^{2\beta}}{nh}\right),$$

where $C_1 = \frac{C_* L}{\ell!}$ and $C_2 = \sigma^2 C_*^2$, with C_* depending on the choice of kernel and $\inf_x P_0^X(x)$. The last term in the sampling variance comes from the random design setup.

Thus, in the no-shift setting, the pointwise risk is minimized at $h \asymp n^{-1/(2\beta+1)}$, giving the optimal convergence rate

$$\mathbb{E}_0[(\hat{f}_n(x_0) - f(x_0))^2] \leq C n^{-2\beta/(2\beta+1)}$$

for some constant depending on ℓ, L, σ^2 and the choice of kernel.

3.2 LPEs under RUPs

We present the following analog to Proposition 3.1, for when sampling from a RUP.

Theorem 3.1. *For any $x_0 \in [0, 1]$, let $\hat{f}_n^0(x_0), \hat{f}_n'(x_0)$ be LP(ℓ) estimators of bandwidth h using $(X_{1:n}, Y_{1:n})$ from $P_0^{\otimes n}$ and an RUP mixture \tilde{P}^n of strength $\tau = \delta^2 \bar{\rho}$ respectively. Moreover, assume that $\rho(X_1, X_2)$ has finite correlation length $\lambda < h$, i.e. $|x_1 - x_2| > \lambda \implies \rho(x_1, x_2) = 0$, and that the choice of x_0 is independent of ρ . Then*

$$\mathbb{E}_{\Xi} \mathbb{E}_{\xi}[(\hat{f}_n(x_0) - f(x_0))^2] \leq b^2(x_0; \hat{f}_n^0) + v_{\text{sampling}}^2(x_0; \hat{f}_n^0) + v_{\text{dist}}^2(x_0; \hat{f}_n'), \quad (6)$$

where

$$b^2(x_0; \hat{f}_n^0) \leq C_1 h^{2\beta}, \quad v_{\text{sampling}}^2(x_0; \hat{f}_n^0) \leq \frac{C_2 \sigma^2}{nh} + O\left(\frac{h^{2\beta}}{nh}\right), \quad \text{and } v_{\text{dist}}^2(x_0; \hat{f}'_n) \leq \frac{C_2 \sigma^2 \tau}{h} + O\left(\frac{h^{2\beta}}{nh}\right),$$

where C_1, C_2 are as in Proposition 3.1.

Note that the distributional variance term v_{dist}^2 has the same bandwidth dependence as the classical v_{sampling}^2 and the new error decomposition is equivalent to replacing the sample size n in Proposition 3.1 with $n_{\text{eff}} = \left(\frac{1}{n} + \tau\right)^{-1} = \frac{n}{1+n\tau}$

The perturbation strength, $\tau = \delta^2 \bar{\rho}$ quantifies both the variance scale δ^2 of the random perturbations and their spatial correlation $\bar{\rho}$ across covariates. This dependence on τ contrasts with previous distribution generalization results, which bound the risk in terms of deterministic, pointwise measures of shift. In domain adaptation and DRO frameworks, [1, 6], the discrepancy between source and target distributions is fixed and measured globally through f -divergences, which, in the case of RUPs will not account for the random cancellations of the model. Consider for instance the KL divergence between P_0 and an RUP P_ξ under the partition model.

$$KL(P_0||P_\xi) = \int p_0(x, \varepsilon) \log \frac{p_0(x, \varepsilon)}{p_\xi(x, \varepsilon)} d(x, \varepsilon) = \sum_{b=1}^{B_X B_\varepsilon} \log \frac{1}{\xi_b} \int_b p_0(x, \varepsilon) d(x, \varepsilon) = \frac{1}{B_X B_\varepsilon} \sum_{b=1}^{B_X B_\varepsilon} (-\log \xi_b) \approx \mathbb{E}[-\log \xi]$$

Thus, as the number of bins $B_X B_\varepsilon$ increases, this KL divergence converges to the negative log-moment of the weight distribution Ξ . However, the perturbation strength τ tends to zero as $B_X B_\varepsilon \rightarrow \infty$, and indeed as shown by the theorem above, the error tends to that of the no-shift setting.

Similarly, recent transfer-based formulations [14, 15] characterize distribution shift through smooth or bounded transfer functions $T(x) = p_T(x)/p_S(x)$, which capture systematic bias between domains but not stochastic variation of $T(x)$. Our results show that even under irregular changes of the density ratio, transfer is possible.

3.2.1 Bandwidth selection under RUPs

The classical approach to bandwidth selection relies on the i.i.d. assumption, where methods such as cross-validation or plug-in estimators balance bias against sampling variance. Under RUPs, however, the additional distributional variance term requires a modified approach.

Oracle bandwidth. When τ is known, the optimal bandwidth minimizes the total MSE:

$$h^\star = \arg \min_h \left\{ C_1 h^{2\beta} + \frac{C_2}{nh} + \frac{C_2 \tau}{h} \right\},$$

where $C_1, C_2 > 0$ are constants.

Corollary 3.1. *Balancing the bias term against the combined variance yields*

$$h^\star \asymp \left(\frac{1}{n} + \tau \right)^{1/(2\beta+1)} = n_{\text{eff}}^{-1/(2\beta+1)}.$$

Thus, inserting this bandwidth into the results from Theorem 3.1, we have that the LPE pointwise estimator converges at a rate of $n_{\text{eff}}^{-\frac{2\beta}{2\beta+1}}$.

In the regime where $\tau \gg 1/n$, this simplifies to $\tau^{\frac{2\beta}{2\beta+1}}$, independent of sample size.

Practical estimation. In practice, τ is unknown and must be estimated from data. A rigorous treatment of optimal estimation procedures for τ is beyond the scope of this work; here we outline two practical approaches that can be employed depending on the available information.

1. *Domain-structured holdouts.* When the target of inference involves generalization across a known domain structure (e.g., different days, geographical regions, or institutions), practitioners often construct validation sets that respect this structure—holding out entire days, regions, or institutions rather than randomly

sampling observations. Our RUP framework provides a model for this common practice: such domain-structured holdout sets naturally capture both sampling and distributional variance, whereas standard random splits only reflect sampling variability. Consequently, bandwidth selection via cross-validation on domain-structured holdouts will appropriately account for the effective sample size n_{eff} rather than the nominal sample size n . This perspective aligns with empirical findings in the domain generalization literature [8], where validation procedures that respect the structure of distribution shifts consistently yield better-tuned models.

2. Estimating τ from summary statistics. When summary data from the target distribution are available, one can estimate the distributional uncertainty directly. Suppose we have access to population-level statistics (e.g., means of the outcome) for the target distribution across multiple realizations. By comparing the training data to these summary statistics, we can estimate the strength of the distributional shift τ . Specifically, let $\{\theta_j\}_{j=1}^J$ be summary statistics from J independent realizations of the target distribution. The variance of θ_j across realizations provides an estimate of the distributional uncertainty, depending on how the statistic aggregates over X . Summaries that preserve global symmetry over X (e.g. averages over values of X) vary with the global shift strength τ , while pointwise-in- x summaries vary with δ^2 . Fixed, nonvanishing local windows (e.g., an LP equivalent kernel at x_0 with bandwidth h) target the corresponding local shift factor: δ^2 times the window's X -correlation—which is exactly the variance inflation entering the LP risk. Thus such summary statistics may be used to calibrate the bandwidth h^* .

4 Minimax Lower Bounds under Random Unbiased Perturbations

We now show that random unbiased perturbations (RUPs) slow the attainable rate of convergence by effectively reducing the sample size. While the optimal learning rate depends on the exact perturbation mechanism, we show that in the case of the correlated noise model, in the regime where the sampling and distributional uncertainties are of the same order, or the sampling uncertainty dominates, the rate derived in Corollary 3.1 is the optimal rate of convergence for pointwise risk. We do so under the following additional assumption.

Assumption 2. *The noise ε is normally distributed under P_0 , i.e. $P_0^\varepsilon \sim \mathcal{N}(0, \sigma^2)$.*

Recall that under the correlated noise model, a perturbation is parameterized by $\xi = \{\xi_b\}_{b=1}^{B_X}$, where $\xi_b \sim \mathcal{N}(0, \delta^2 \sigma^2)$, and observations follow for $X_i = x$, $Y_i = f(x) + \varepsilon$, where $P_\xi^\varepsilon(\varepsilon) = P_0^\varepsilon(\varepsilon - \xi_{b(x)})$. We study the regime where the number of buckets B_X grows with n such that $\lim_{n \rightarrow \infty} \frac{n}{B_X(n)} < \infty$.

Remark 4.1. *Depending on the smoothness class of functions we assume, the restriction of the regime we study can be weakened. However, due to the discreteness of the RUP examples provided, fixing B_X and sending $n \rightarrow \infty$ leads to almost identifiable P_0 for sufficiently smooth functions f . Thus we recover the behavior of the no-shift setting. If, on the other hand, we fix n and send $B_X \rightarrow \infty$, the correlation $\rho \rightarrow 0$, and this effectively increases the variance of ε , sending the effective sample size back to n .*

Theorem 4.1. *Suppose that $\beta > 0$ and $L > 0$. Let $(\Xi, (P_\xi)_\xi)$ be an RUP following the correlated noise model (defined in Section 2.2.2) of strength $\tau = \tau_n$. For any fixed point x_0 , let $\hat{f}_n(x_0)$ be an estimator of $f(x_0)$ constructed using $(X_{1:n}, Y_{1:n}) \sim \bar{P}^n$. Let $n_{\text{eff}} = \frac{n}{1+n\tau_n}$. Then, as long as $\lim_{n \rightarrow \infty} n\tau_n < \infty$, the minimax lower bound satisfies*

$$\liminf_{n \rightarrow \infty} \inf_{\hat{f}_n} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_\Xi \mathbb{E}_{\xi, f}^n \left[n_{\text{eff}}^{\frac{2\beta}{2\beta+1}} (\hat{f}_n(x_0) - f(x_0))^2 \right] \geq c_1 \quad (7)$$

for some constant c_1 depending on σ^2 , β and L .

We provide a full proof in Section C. The idea follows standard minimax arguments; the key new step is the scaling of the KL divergence under RUPs. In traditional minimax arguments for pointwise risk around x_0 , one usually constructs two hypothesis functions (see for instance [16]), $f_0(x) \equiv 0$ and $f_1(x) = Lh^\beta K\left(\frac{x-x_0}{h}\right)$. Here K is some kernel function and h is the bandwidth. Setting P_0, P_1 to be the probability distributions given data is generated via f_0, f_1 respectively, one shows that $KL(P_0^{\otimes n} \| P_1^{\otimes n}) = C_* nh^{2\beta+1}$, so letting the bandwidth scale as $h \propto n^{-\frac{1}{2\beta+1}}$, results in a constant bound on the KL divergence as n grows.

Under RUPs of two close parametric distributions P_0, P_1 , the usual factorization breaks because the mixture distribution \bar{P}_i^n no longer factorizes. The next lemma shows how the same expansion holds with n replaced by n_{eff} .

Lemma 4.1 (KL scaling under the correlated noise model). *Under the correlated noise model, let $f_0 \equiv 0$, and $f_1(x) = Lh^\beta K\left(\frac{x-x_0}{h}\right)$ for some fixed x_0 and fixed kernel function K . Let \bar{P}_0^n, \bar{P}_1^n be the mixture probability distributions for the noise-correlated RUP data generated by f_0, f_1 respectively. Then, there exist a constant C , depending on $L, \beta, K_{\max} = \|K\|_\infty$, and σ^2 such that, in the regime where $\lim_{n \rightarrow \infty} \frac{n}{B_X(n)} < \infty$,*

$$\lim_{n \rightarrow \infty} \frac{1}{n_{\text{eff}}} KL(\bar{P}_0^n \parallel \bar{P}_1^n) = Ch^{2\beta+1}$$

In other words, setting the bandwidth to be $h \propto n_{\text{eff}}^{-\frac{1}{2\beta+1}}$ leads to a KL divergence that is bounded by a constant in the limit $n \rightarrow \infty$.

The lemma is proved in Section C.1, but we outline the main argument here. Under the correlated noise RUP, datapoints $(X_i, Y_i)_{i=1}^n$ are no longer iid, however, the X_i s are iid, and the $Y|X$ follows a multivariate gaussian distribution, $\mathcal{N}(f(X), \Sigma)$, where

$$\Sigma_{ij} = \begin{cases} (1 + \delta^2)\sigma^2 & \text{if } i = j \\ \delta^2\sigma^2 & \text{if } i \neq j, \text{ but } b(X_i) = b(X_j) \\ 0 & \text{if } b(X_i) \neq b(X_j) \end{cases}$$

Computing Σ^{-1} (see Section C.1) shows that the within-bucket average direction is downweighted by a factor $(1 + n\tau)^{-1}$, so the Fisher information grows like an *effective* sample size $n_{\text{eff}} < n$ rather than n itself. This is exactly the n_{eff} appearing in Lemma 4.1, and explains why the optimal bandwidth scales as $h \propto n_{\text{eff}}^{-1/(2\beta+1)}$ in the correlated noise model.

Remark 4.2 (Extension to multi-hypothesis minimax arguments.). *The same LAN-based reasoning extends directly to minimax lower bound proofs that rely on more than two hypotheses, such as Fano, Le Cam, or Assouad constructions for mean integrated squared error (MISE) risks. In those arguments, the pairwise KL divergences for RUPs under the correlated noise model $KL(\bar{P}_i^{(n)} \parallel \bar{P}_j^{(n)})$ appearing in the packing or testing steps scale as $n_{\text{eff}} KL(P_i^0 \parallel P_j^0)$ by Lemma 4.1.*

5 Numerical results

We illustrate the theoretical results above through simple simulations using an RUP model. For each dataset, covariates $X_i \sim \text{Unif}[0, 1]$, outcomes are generated as $Y_i = f(X_i) + \varepsilon_i$ with $f(x) = \sin(2\pi x)$, and ε_i drawn according to the perturbed law $p_\xi(\varepsilon | X)$ following the correlated noise model. We compute the local-polynomial estimator of order 1 with Epanechnikov kernel over a range of bandwidths h and report the empirical mean integrated squared error (MISE) across 100 perturbation realizations.

Figure 5 illustrates the MISE as a function of different bandwidths for three distributions: a baseline P_0 , and a sequence of perturbed distributions, $\{P_\xi\}$. As expected, we see that the bandwidth achieving minimal MISE increases, (i.e. optimal model complexity decreases) as we consider stronger perturbations. In particular, simply tuning the bandwidth size to the shift can lead to significant improvements in the case of larger random shifts.

In Figure 6 we analyze how the optimal bandwidth behaves as n grows. Under no shift, the best h is scaled with n and we see it decrease throughout the range of n 's considered. However, for perturbed distributions, the optimal bandwidth depends on both n and τ , and once $1/n < \tau$, increasing n no longer leads to smaller optimal bandwidths.

6 Discussion

We have introduced a framework for nonparametric regression under random unbiased perturbations (RUPs) of the conditional distribution $Y|X$. This setting captures a common data collection scenario where datasets are drawn from randomly shifted versions of a target distribution, with the marginal distribution of covariates held fixed but the conditional law perturbed by uncontrollable factors such as measurement protocols, environmental conditions, or sampling procedures.

Our analysis reveals that RUPs introduce a distinct mode of uncertainty that differs fundamentally from both classical sampling variability and adversarial distribution shifts. The key insight is that distributional uncertainty manifests as variance inflation rather than bias. This variance inflation reduces the effective

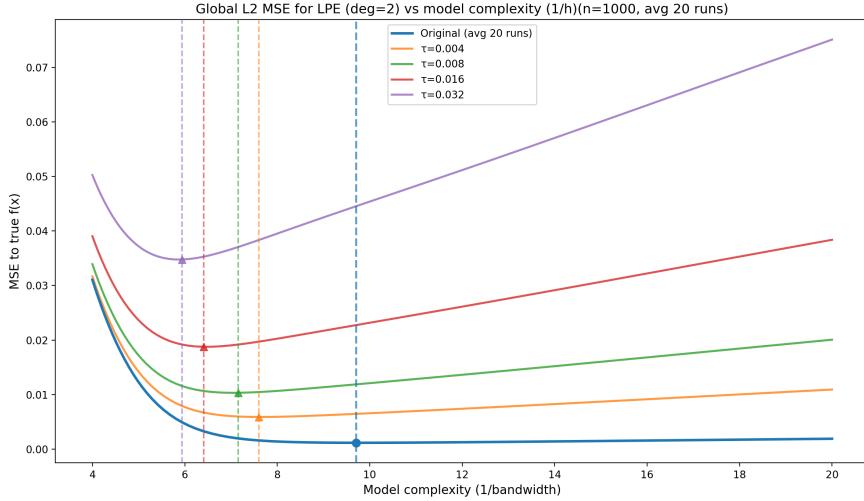


Figure 5: Simulation illustrating the predicted behavior from Figure 2. Under i.i.d. sampling assumptions, a model would choose a smaller bandwidth and hence a higher complexity model. As random distribution shift is introduced, the optimal model complexity decreases to account for the additional variance.

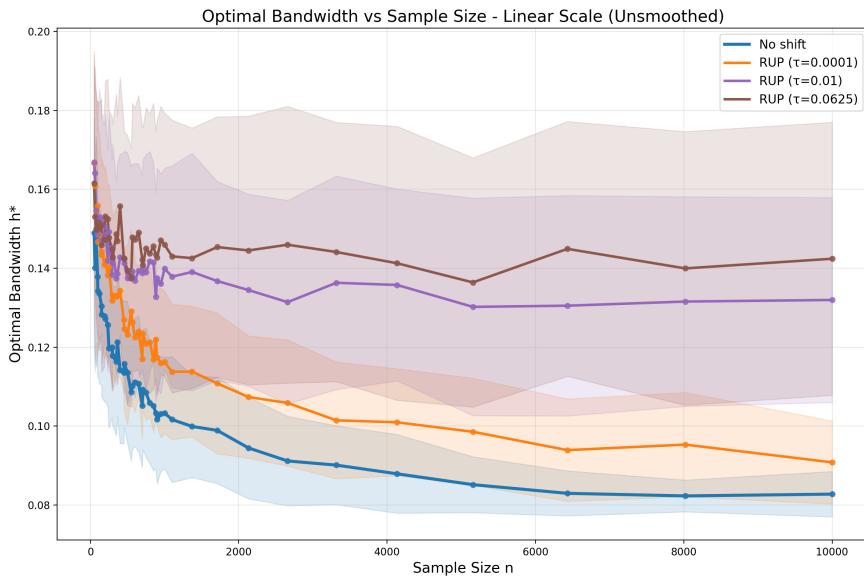


Figure 6: Empirical optimal bandwidth h^* as a function of sample size n under different perturbation strengths. When $\tau = 0$, h^* decreases as $n^{-1/(2\beta+1)}$ (blue); for $\tau > 0$, the curve flattens once $1/n < \tau$, consistent with the theoretical transition to the $\tau^{1/(2\beta+1)}$ scaling.

sample size to $n_{\text{eff}} = n/(1 + n\tau)$, where τ quantifies the perturbation strength as the product of variance scale δ^2 and average correlation $\bar{\rho}$.

For local polynomial estimators, we have shown that this effective sample size affects both practical bandwidth selection and fundamental convergence rates. When distributional uncertainty dominates sampling uncertainty ($\tau \gg 1/n$), optimal bandwidths must be chosen to balance bias against distributional variance rather than sampling variance, leading to the scaling $h \propto \tau^{1/(2\beta+1)}$ and convergence rate $\tau^{2\beta/(2\beta+1)}$. Our minimax lower bounds demonstrate that these rates are optimal, at least in the setting $\lim_{n \rightarrow \infty} n\tau_n < \infty$, by studying the correlated noise model and establishing that distributional uncertainty fundamentally limits the rate of convergence.

The RUP framework offers several advantages over existing approaches to distribution shift. Unlike DRO or transfer-learning approaches, which assume fixed or adversarial shifts, our framework models dataset-level randomness that averages to the target law, yielding sharper and more realistic risk characterizations. Unlike transfer learning and domain adaptation settings, which treat the source-to-target shift as fixed, we model dataset-level randomness that averages to the target distribution. This perspective is more appropriate for applications where perturbations represent idiosyncratic variation rather than systematic domain differences.

Several directions for future work emerge from our analysis. First, while we have focused on local polynomial estimators, the effective sample size principle should extend to other nonparametric methods such as kernel ridge regression, smoothing splines, and nearest-neighbor estimators. From a methodology perspective, it would also be interesting to see whether the sample size principle can be applied to random forest, gradient boosting, or deep learning techniques. Second, our bandwidth selection discussion has emphasized the theoretical optimum; developing robust procedures that estimate τ based on principles discussed in Section 3.2.1 remains an important part of the puzzle. From the lower bound side, we have focused on perturbation models that are discrete in that they depend on quantile binning of the distribution and therefore only allow lower bounds in the regime where $\lim_{n \rightarrow \infty} n\tau_n < \infty$. However, there may be continuous RUPs based on random processes that could extend these bounds to all regimes.

Moreover, extending the framework to structured perturbations with known correlation patterns could broaden applicability. More generally, one could decompose real-world distribution shifts into systematic and random components, thus capturing both persistent biases **across domains** and transient, mean-zero fluctuations **within domains**. Finally, similar variance inflation phenomena may occur in other statistical problems, such as density estimation, classification, or high-dimensional inference. Investigating these would deepen our understanding of how dataset-level randomness affects statistical inference more broadly.

In summary, random unbiased perturbations provide a tractable model that captures plausible patterns of distributional uncertainty and leads to concrete modifications of classical nonparametric methods. By recognizing that such uncertainty inflates variance rather than bias, practitioners can make more informed choices about model complexity and achieve better performance than methods tuned under idealized i.i.d. assumptions or overly conservative distributional robustness constraints.

7 Acknowledgments

Rothenhäusler gratefully acknowledges support as a David Huntington Faculty Scholar, Chamber Fellow, and from the Dieter Schwarz Foundation.

References

- [1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, May 2010.
- [2] J. Blanchet, J. Li, S. Lin, and X. Zhang. Distributionally Robust Optimization and Robust Statistics. *Statistical Science*, 40(3):351–377, Aug. 2025. Publisher: Institute of Mathematical Statistics.
- [3] J. Blanchet, K. Murthy, and F. Zhang. Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes. *Mathematics of Operations Research*, 47(2):1500–1529, May 2022. Publisher: INFORMS.
- [4] P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.
- [5] T. T. Cai and H. Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), Feb. 2021. Publisher: Institute of Mathematical Statistics.
- [6] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3), June 2021.
- [7] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability* 66. CRC Press, Mar. 1996. Google-Books-ID: BM1ckQKCXP8C.
- [8] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.
- [9] C. Huber. Lower Bounds for Function Estimation. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 245–258. Springer, New York, NY, 1997.
- [10] Y. Jeong and D. Rothenhäusler. Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty. *In press at the Journal of Machine Learning Research*, 2025+.
- [11] Y. Jeong and D. Rothenhäusler. Out-of-distribution generalization under random, dense distributional shifts, Apr. 2024. arXiv:2404.18370.
- [12] Y. Jin, N. Egami, and D. Rothenhäusler. Beyond Reweighting: On the Predictive Role of Covariate Shift in Effect Generalization. *Proceedings of the National Academy of Sciences of the United States of America*, 2025.
- [13] C. Ma, R. Pathak, and M. J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.
- [14] H. W. J. Reeve, T. I. Cannings, and R. J. Samworth. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, Dec. 2021. Publisher: Institute of Mathematical Statistics.
- [15] R. Sahoo, L. Lei, and S. Wager. Learning from a Biased Sample, Sept. 2025. arXiv:2209.01754.
- [16] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009.
- [17] B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. Springer, New York, NY, 1997.

A Validity of RUP models

A.1 Validity of the partition model

Here we prove Proposition 2.1, that the partition model does indeed define a family of RUPs satisfying Definition 1.

Proof of Proposition 2.1. Recall that for $x \in I_i$, $\varepsilon \in J_j$ we define the perturbed noise law by

$$P_\xi(x, \varepsilon) = \frac{\xi_{ij}}{\bar{\xi}_i} P_0(x, \varepsilon),$$

where $\bar{\xi}_i = \frac{1}{B_\varepsilon} \sum_{j'=1}^{B_\varepsilon} \xi_{ij'}$.

1. Fixed marginal. We write out for any x , wlog assume $x \in I_i$,

$$\begin{aligned} P_\xi(x) &= \int P_\xi(x, \varepsilon) d\varepsilon = \sum_{j=1}^{B_\varepsilon} \int_{J_j} P_\xi(x, \varepsilon) d\varepsilon \\ &= \sum_{j=1}^{B_\varepsilon} \int_{J_j} \frac{\xi_{ij}}{\bar{\xi}_i} P_0(x, \varepsilon) d\varepsilon \end{aligned}$$

By Assumption 1,

$$\begin{aligned} &= \sum_{j=1}^{B_\varepsilon} \frac{\xi_{ij}}{\bar{\xi}_i} \int_{J_j} P_0(x) P_0(\varepsilon) d\varepsilon \\ &= \sum_{j=1}^{B_\varepsilon} \frac{\xi_{ij}}{\bar{\xi}_i} \frac{1}{B_\varepsilon} P_0(x) = P_0(x) \end{aligned}$$

2. Centering. Recall $\Delta_\xi(x) := \mathbb{E}_\xi[\varepsilon | X = x] - \mathbb{E}_0[\varepsilon | X = x]$. Then, under the partition model,

$$\begin{aligned} \Delta_\xi(x) &= \int \varepsilon \left(P_\xi(\varepsilon | X = x) - P_0(\varepsilon | X = x) \right) d\varepsilon \\ &= \sum_{j=1}^{B_\varepsilon} \int_{J_j} \varepsilon \left(\frac{\xi_{ij}}{\bar{\xi}_i} - 1 \right) P_0(\varepsilon | X = x) d\varepsilon \\ &= \frac{1}{B_\varepsilon} \sum_{j=1}^{B_\varepsilon} \left(\frac{\xi_{ij}}{\bar{\xi}_i} - 1 \right) \mathbb{E}_0[\varepsilon | \varepsilon \in J_j] \end{aligned}$$

Now, taking expectations over Ξ , we have $\mathbb{E}_\Xi \left[\frac{\xi_{ij}}{\bar{\xi}_i} - 1 \right] = 0$ for all $i \in [B_X], j \in [B_\varepsilon]$.

3. Variance scale. Let $m_j = \mathbb{E}_0[\varepsilon | \varepsilon \in J_j]$. From above, we have

$$\mathbb{V}\text{ar}_\Xi(\Delta_\xi(x)) = \frac{1}{B_\varepsilon^2} \sum_{j=1}^{B_\varepsilon} \mathbb{V}\text{ar}_\Xi \left(\frac{\xi_{ij}}{\bar{\xi}_i} \right) m_j^2 + \frac{1}{B_\varepsilon^2} \sum_{\substack{1 \leq j, k \leq B_\varepsilon \\ j \neq k}} \mathbb{C}\text{ov}_\Xi \left(\frac{\xi_{ij}}{\bar{\xi}_i}, \frac{\xi_{ik}}{\bar{\xi}_i} \right) m_j m_k$$

Now by Taylor expansion since $\mathbb{E}[\bar{\xi}^{-3}] \leq \mathbb{E}[\xi^{-3}] < \infty$, we can write $\mathbb{V}\text{ar}_\Xi \left(\frac{\xi_{ij}}{\bar{\xi}_i} \right) = \frac{\mathbb{V}\text{ar}(\xi)}{\mathbb{E}[\xi]^2} (1 - O(1/B_\varepsilon))$.

Moreover, for $j \neq k$, $\mathbb{C}\text{ov}_\Xi \left(\frac{\xi_{ij}}{\bar{\xi}_i}, \frac{\xi_{ik}}{\bar{\xi}_i} \right) = -O(1/B_\varepsilon) \frac{\mathbb{V}\text{ar}(\xi)}{\mathbb{E}[\xi]^2}$. Thus

$$\mathbb{V}\text{ar}_\Xi(\Delta_\xi(x)) = \frac{\mathbb{V}\text{ar}(\xi)}{\mathbb{E}[\xi]^2} \frac{1}{B_\varepsilon} \left(\sum_{j=1}^{B_\varepsilon} m_j^2 (1 - O(1/B_\varepsilon)) - O(1/B_\varepsilon) \left(\left(\sum_{1 \leq j \leq B_\varepsilon} m_j \right)^2 - \sum_{j=1}^{B_\varepsilon} m_j^2 \right) \right)$$

Note that $\sum_{1 \leq j \leq B_\varepsilon} m_j = 0$

$$= \frac{\text{Var}(\xi)}{\mathbb{E}[\xi]^2} \frac{1}{B_\varepsilon^2} \left(\sum_{j=1}^{B_\varepsilon} m_j^2 (1 - O(1/B_\varepsilon)) \right)$$

Finally, we will show that $\frac{1}{B_\varepsilon} \sum_{j=1}^{B_\varepsilon} m_j^2 = \text{Var}_0(\varepsilon) + o(1)$, which will give us the desired result.

$$\frac{1}{B_\varepsilon} \sum_{j=1}^{B_\varepsilon} m_j^2 = \mathbb{E}_0[\mathbb{E}_0[\varepsilon|J]^2] = \text{Var}_0(\mathbb{E}_0[\varepsilon|J]) + \underbrace{\mathbb{E}_0[\mathbb{E}_0[\varepsilon|J]]^2}_{0} = \text{Var}_0(\varepsilon) - \underbrace{\mathbb{E}_0[\text{Var}_0(\varepsilon|J)]}_{R_{B_\varepsilon}},$$

where since $\mathbb{E}[\varepsilon|J] \rightarrow \varepsilon$ in L^2 , we also have that $R_{B_\varepsilon} = \mathbb{E}_0[(\varepsilon - \mathbb{E}_0[\varepsilon|J])^2] \rightarrow 0$ as $B_\varepsilon \rightarrow \infty$.

4. X-dependency. We write out the covariance. Let $x_1 \in I_{i_1}$ and $x_2 \in I_{i_2}$

$$\begin{aligned} \text{Cov}_\Xi(\Delta_\xi(x_1), \Delta_\xi(x_2)) &= \text{Cov}_\Xi\left(\frac{1}{B_\varepsilon} \sum_{j=1}^{B_\varepsilon} \left(\frac{\xi_{i_1 j}}{\bar{\xi}_{i_1}} - 1\right) m_j, \frac{1}{B_\varepsilon} \sum_{j=1}^{B_\varepsilon} \left(\frac{\xi_{i_2 j}}{\bar{\xi}_{i_2}} - 1\right) m_j\right) \\ &= \begin{cases} \text{Var}_\Xi(\Delta_\xi(x_1)) & \text{if } i_1 = i_2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

By previous parts, we are done. □

A.2 Validity of the correlated noise model

Here we prove Proposition 2.2, that the correlated noise model does indeed define a family of RUPs satisfying Definition 1.

Proof of Proposition 2.2. Recall that under the correlated noise model, we observe

$$P_\xi(x, \varepsilon) = P_0^X(x) P_0^\varepsilon(\varepsilon - \xi_{b(x)})$$

with $\xi_{b(x)} \sim \mathcal{N}(0, \delta^2 \sigma^2)$, where $\sigma^2 = \text{Var}_0(\varepsilon)$.

1. Fixed marginal. This holds by definition. For $x \in I_i$,

$$P_\xi(x) = \int P_\xi(x, \varepsilon) d\varepsilon = \int P_0^X(x) P_0^\varepsilon(\varepsilon - \xi_{b(x)}) d\varepsilon = P_0^X(x)$$

2. Centering. Recall $\Delta_\xi(x) := \mathbb{E}_\xi[\varepsilon | X = x] - \mathbb{E}_0[\varepsilon | X = x]$. Then,

$$\Delta_\xi(x) = \int \varepsilon' dP_0^\varepsilon(\varepsilon' - \xi_{b(x)}) - \int \varepsilon dP_0^\varepsilon(\varepsilon) = \int (\varepsilon + \xi_{b(x)} - \varepsilon) dP_0^\varepsilon(\varepsilon) = \xi_{b(x)}$$

Now, taking expectations over Ξ , we have for all x , $\mathbb{E}_\Xi[\xi_{b(x)}] = 0$.

3. Variance scale. From above, we have

$$\text{Var}_\Xi(\Delta_\xi(x)) = \text{Var}_\Xi(\xi_{b(x)}) = \delta^2 \sigma^2$$

4. X-dependency. We write out the covariance. Let $x_1 \in I_{i_1}$ and $x_2 \in I_{i_2}$

$$\text{Cov}_\Xi(\Delta_\xi(x_1), \Delta_\xi(x_2)) = \text{Cov}_\Xi(\xi_{b_1}, \xi_{b_2}) = \mathbf{1}_{I_{i_1} = I_{i_2}} \delta^2 \sigma^2$$

□

B Proof of Upper Bounds (Theorem 3.1)

We outline the following properties of the weights, which we will use in our analysis. These follow under mild regularity conditions, including a minimal eigenvalue condition on the local design matrix as in [16, Section 1.6.1].

Lemma B.1 (Tsybakov [16], Lemma 1.3). *Assume the kernel function K is bounded by some K_{\max} and has compact support in $[-1, 1]$. Moreover, assume that the local design matrix is uniformly non-degenerate (via some λ_0) for large enough n , and the points X_i are dense in $[0, 1]$. Then, almost surely,*

$$(i) \sum_{k=1}^n W_k(x) = 1, \text{ with } \sum_{k=1}^n |W_k(x)| \leq C_*$$

$$(ii) \sup_{k,x} |W_k(x)| \leq \frac{C_*}{nh}$$

$$(iii) W_k(x) = 0 \text{ if } |X_k - x| > h$$

where C_* is a constant depending only on λ_0 and K_{\max} .

The assumptions required for the lemma above hold almost surely for the setting with iid distributed X_i , where the probability density of X over $[0, 1]$ is bounded away from 0. The minimum value of this density will determine λ_0 . Moreover, in this setting we will satisfy the assumption that the points X_i are dense in $[0, 1]$ almost surely.

Remark B.1 (Regarding Proposition 3.1). *While this proposition is based on Proposition 1.13 from [16], the version in the book only considers the fixed design regime. To extend the result to random design regime, we must account for the variance from X which is the same expression as we consider in (8) and is the source of the $O\left(\frac{h^{2\beta}}{nh}\right)$ term.*

Proof of Theorem 3.1. By the same decomposition as in (2), we decompose the risk into bias, sampling variance and distributional variance by the law of total variance:

$$\begin{aligned} R_n(x_0; \hat{f}) &= \mathbb{E}_{\Xi} \mathbb{E}_{\xi} \left[(\hat{f}_n^{\xi}(x_0) - f^0(x_0))^2 \right] \\ &= \underbrace{\left(\mathbb{E}_{\Xi} \mathbb{E}_{\xi} [\hat{f}_n(x_0)] - f(x_0) \right)^2}_{\text{bias squared}} + \underbrace{\mathbb{E}_{\Xi} \text{Var}_{\xi} (\hat{f}_n^{\xi}(x_0))}_{\text{sampling variance}} + \underbrace{\text{Var}_{\Xi} \mathbb{E}_{\xi} [\hat{f}_n^{\xi}(x_0)]}_{\text{distributional variance}} \end{aligned}$$

The first two terms follow standard LPE analysis under random design setting. Recall that $\hat{f}_n(x_0) = \sum_{k=1}^n W_k(x_0) Y_k$. Then, for the **squared bias**, we use the centering property and standard LPE analysis (see [7, 16]).

$$\begin{aligned} \left(\mathbb{E}_{\Xi} \mathbb{E}_{\xi} [\hat{f}_n(x_0)] - f(x_0) \right)^2 &= \left(\mathbb{E}_{\Xi} \mathbb{E}_{\xi} \left[\sum_{k=1}^n Y_k W_k(x_0) \right] - f(x_0) \right)^2 \\ &= \left(\mathbb{E}_0 \left[\sum_{k=1}^n W_k(x_0) f(X_k) \right] - f(x_0) \right)^2 = \frac{C_*^2 L^2}{(\ell!)^2} h^{2\beta} \end{aligned}$$

For the **sampling variance**, we have

$$\mathbb{E}_{\Xi} \text{Var}_{\xi} (\hat{f}_n^{\xi}(x_0)) = \underbrace{\mathbb{E}_{\Xi} \mathbb{E}_X [\text{Var}_{\xi} (\hat{f}_n^{\xi}(x_0) | X)]}_{I^s} + \underbrace{\mathbb{E}_{\Xi} \text{Var}_X (\mathbb{E}_{\xi} [\hat{f}_n^{\xi}(x_0) | X])}_{II^s}$$

Note that given ξ, X, ε_k are independent under \mathbb{P}_ξ .

$$\begin{aligned}
I^s &= \mathbb{E}_\Xi \mathbb{E}_X \left[\text{Var}_\xi \left(\sum_{k=1}^n W_k(x_0) Y_k | X \right) \right] = \mathbb{E}_\Xi \mathbb{E}_X \left[\text{Var}_\xi \left(\sum_{k=1}^n W_k(x_0) \varepsilon_k | X \right) \right] \\
&= \mathbb{E}_\Xi \mathbb{E}_X \left[\sum_{k=1}^n W_k(x_0)^2 \text{Var}_\xi (\varepsilon_k | X) \right] = \mathbb{E}_X \left[\sum_{k=1}^n W_k(x_0)^2 \mathbb{E}_\Xi [\mathbb{E}_\xi [\varepsilon_k^2 | X] - \mathbb{E}_\xi [\varepsilon_k | X]^2] \right] \\
&= \mathbb{E}_X \left[\sum_{k=1}^n W_k(x_0)^2 (\mathbb{E}_0 [\varepsilon_k^2 | X] - \text{Var}_\Xi (\mathbb{E}_\xi [\varepsilon_k | X])) \right] \\
&\leq \sum_{k=1}^n \mathbb{E}_X [W_k(x_0)^2] \sigma^2 \leq nh \left(\frac{C_*}{nh} \right)^2 \sigma^2 = \frac{\sigma^2 C_*^2}{nh},
\end{aligned}$$

where we use Lemma B.1 and C_* depends only on λ_0 and K_{\max} . For the second, smaller order term (which comes from the choice to include a random design setup), we have

$$II^s = \mathbb{E}_\Xi \text{Var}_X (\mathbb{E}_\xi [\hat{f}_n^\xi(x_0) | X]) = \mathbb{E}_\Xi \text{Var}_X \left(\sum_{k=1}^n W_k(x_0) f(X_k) \right) = O \left(\frac{h^{2\beta}}{nh} \right) \quad (8)$$

Finally, for the **distributional variance**,

$$\text{Var}_\Xi \mathbb{E}_\xi [\hat{f}_n^\xi(x_0)] = \underbrace{\mathbb{E}_X \text{Var}_\Xi (\mathbb{E}_\xi [\hat{f}_n^\xi(x_0) | X])}_{I^D} + \underbrace{\text{Var}_X (\mathbb{E}_\Xi \mathbb{E}_\xi [\hat{f}_n^\xi(x_0) | X])}_{II^D}$$

For the first, main term, we have:

$$\begin{aligned}
I^D &= \mathbb{E}_X \text{Var}_\Xi (\mathbb{E}_\xi [\hat{f}_n^\xi(x_0) | X]) \\
&= \mathbb{E}_X \text{Var}_\Xi \left(\sum_{k=1}^n W_k(x_0) \mathbb{E}_\xi [\varepsilon_k | X_k] \right) \\
&= \mathbb{E}_X \left[\sum_{k,\ell} W_k(x_0) W_\ell(x_0) \text{Cov}_\Xi (\mathbb{E}_\xi [\varepsilon_k | X_k], \mathbb{E}_\xi [\varepsilon_\ell | X_\ell]) \right] \\
&= \delta^2 \sigma^2 \mathbb{E}_X \left[\sum_{k,\ell} W_k(x_0) W_\ell(x_0) \rho(X_k, X_\ell) \right] \\
&= \delta^2 \sigma^2 \mathbb{E}_X \left[\sum_{k:|X_k-x_0|<h} W_k(x_0) \sum_{\ell:|X_k-X_\ell|<\lambda} W_\ell(x_0) \rho(X_k, X_\ell) \right]
\end{aligned}$$

Since we sum over $W_k(x_0)$ where $|X_k - x_0| < h$, we are summing about nh terms, and for each of these, we sum over ℓ such that $|X_k - X_\ell| < \lambda$ which is at most about $n\lambda$ terms. Moreover, for all k , $|W_k(x)| \leq \frac{C_*}{nh}$. Hence, (using the fact that the choice of x_0 is independent from the correlation kernel ρ),

$$\begin{aligned}
&\leq \delta^2 \sigma^2 nh \cdot n\lambda \left(\frac{C_*}{nh} \right)^2 \mathbb{E}_X [\rho(X_k, X_\ell) | |X_k - X_\ell| < \lambda] \\
&\leq \frac{\delta^2 \sigma^2 C_*^2 \bar{\rho}}{h} = \frac{C_*^2 \sigma^2 \tau}{h},
\end{aligned}$$

where the last line follows from

$$\begin{aligned}
\bar{\rho} &= \mathbb{P}(|X_1 - X_2| \geq \lambda) \underbrace{\mathbb{E}_X [\rho(X_1, X_2) | |X_1 - X_2| \geq \lambda]}_0 + \mathbb{P}(|X_1 - X_2| < \lambda) \mathbb{E}_X [\rho(X_1, X_2) | |X_1 - X_2| < \lambda] \\
&\geq \lambda \mathbb{E}_X [\rho(X_1, X_2) | |X_1 - X_2| < \lambda]
\end{aligned}$$

Again, for the second, smaller order term,

$$II^D = \text{Var}_X (\mathbb{E}_\Xi \mathbb{E}_\xi [\hat{f}_n^\xi(x_0) | X]) = \text{Var}_X \left(\sum_{k=1}^n W_k(x_0) (f(X_k) + \mathbb{E}_\Xi \mathbb{E}_\xi [\varepsilon | X]) \right) = \text{Var}_X \left(\sum_{k=1}^n W_k(x_0) f(X_k) \right) = O \left(\frac{h^{2\beta}}{nh} \right)$$

□

C Proof of Lower Bounds (Theorem 4.1)

By Markov's inequality, we write out

$$\begin{aligned} \inf_{\hat{f}_n} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_{\Xi} \mathbb{E}_{\xi, f, n} [|\hat{f}_n(x_0) - f(x_0)|] &\geq \inf_{\hat{f}_n} \sup_{f \in \Sigma(\beta, L)} s \mathbb{E}_{\Xi} \mathbb{P}_{\xi, f}^n (|\hat{f}_n(x_0) - f(x_0)| \geq s) \\ &\geq s \inf_{\hat{f}_n} \max_{f \in \{f_0, f_1\}} \bar{\mathbb{P}}_{\xi, f}^n (|\hat{f}_n(x_0) - f(x_0)| \geq s) \end{aligned}$$

For any two $f_0, f_1 \in \Sigma(\beta, L)$. In particular, if we set $f_0(x) \equiv 0$, and $f_1(x) = Lh^\beta K\left(\frac{x-x_0}{h}\right)$, where we choose K to be a bounded kernel function satisfying $K \in \Sigma(1/2, L) \cap C^\infty$, and vanishing outside of $[-1/2, 1/2]$. Let $\psi(D) = \arg \min_{i \in \{0, 1\}} |\hat{f}_n(D, x_0) - f_i(x_0)|$ be the minimal distance test, i.e. returns the index of the function in $\{f_0, f_1\}$ closest to the estimator. Note that in this case, $|f_0(x_0) - f_1(x_0)| = Lh^\beta K_{\max}$, so setting $s = \frac{Lh^\beta}{2} K_{\max}$ guarantees that the event $\{\psi(D) \neq j\} \subseteq \{|\hat{f}_n(x_0) - f(x_0)| > s\}$ because there exists a k such that $|\hat{f}_n(x_0) - f_k(x_0)| \leq |\hat{f}_n(x_0) - f_j(x_0)|$, thus

$$2s \leq |f_k(x_0) - f_j(x_0)| \leq |\hat{f}_n(x_0) - f_k(x_0)| + |\hat{f}_n(x_0) - f_j(x_0)| \leq 2|\hat{f}_n(x_0) - f_j(x_0)|$$

Thus we can write

$$\inf_{\hat{f}_n} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_{\Xi} \mathbb{E}_{\xi, f, n} [|\hat{f}_n(x_0) - f(x_0)|] \geq \frac{Lh^\beta}{2} \underbrace{\inf_{\hat{f}_n} \max_{f_j \in \{f_0, f_1\}} \bar{\mathbb{P}}_{f_j}^n (\psi(D) \neq j)}_{p_{e,1}}$$

Now, by Tsybakov Thm 2.2 (iii), if $KL(\bar{\mathbb{P}}_0^n, \bar{\mathbb{P}}_1^n) \leq \beta$, then $p_{e,1} \geq \max\left(\frac{1}{4} \exp(-\beta), \frac{1-\sqrt{\beta/2}}{2}\right)$. Now, by Lemma 4.1 we know that $KL(\bar{\mathbb{P}}_0^n, \bar{\mathbb{P}}_1^n) \leq n_{\text{eff}} KL(\mathbb{P}_0, \mathbb{P}_1)$. In particular, for Hölder- $\Sigma(\beta, L)$ kernel-bump alternatives of bandwidth $h \rightarrow 0$,

$$KL\left(\bar{P}_{\theta_1}^{(n)} \parallel \bar{P}_{\theta_0}^{(n)}\right) = n_{\text{eff}} c L^2 h^{2\beta+1} + o(1),$$

for a constant $c > 0$ depending on the kernel and noise law. Thus, choosing $h = n_{\text{eff}}^{-\frac{1}{2\beta+1}} \left(\frac{\beta}{K_{\max}^2 L^2 c}\right)^{\frac{1}{2\beta+1}}$, guarantees $KL(\bar{\mathbb{P}}_0^n, \bar{\mathbb{P}}_1^n) \leq \beta$ and thus

$$\inf_{\hat{f}_n} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_{\Xi} \mathbb{E}_{\xi, f, n} [(\hat{f}_n(x_0) - f(x_0))^2] \geq n_{\text{eff}}^{-\frac{2\beta}{2\beta+1}} c_\beta$$

where c_β is a constant that depends on L, β and σ^2 , as we can choose the kernel to be bounded by 1. Taking limits as $n \rightarrow \infty$ completes the proof.

C.1 Proof of Lemma 4.1

Proof of Lemma 4.1. Recall the correlated noise model. We split the domain of X into B_X buckets and parameterize the shift by $\xi = \{\xi_b\}_{b=1}^{B_X}$, with $\xi_b \sim^{i.i.d.} \mathcal{N}(0, \delta^2 \sigma^2)$.

$$P_\xi(x, \varepsilon) = P_0^X(x) P_0^\varepsilon(\varepsilon - \xi_{b(x)}),$$

where $b(x)$ is the bucket for x . Let $m_b = |\{X \in \{X_1, \dots, X_n\} | X \in b\}|$ be the number of points in bucket b and let $m = \frac{n}{B_X}$ denote the average. By Assumption 2, when data is drawn from the mixture model $\bar{P}^n = \int_{\Xi} P^{\xi, n} d\xi$, $(X_i, Y_i)_{i=1}^n \sim \bar{P}^n$ actually follows a multivariate gaussian distribution, where

$$Y_{1:n} \sim \mathcal{N}(f(X_{1:n}), \Sigma), \quad \Sigma = \sigma^2 I_{B_X} \otimes \Sigma_{\text{blk}} \quad (\in \mathbb{R}^{n \times n}), \quad \Sigma_{\text{blk}, b} = I_{m_b} + \delta^2 \mathbf{1}_{m_b} \mathbf{1}_{m_b}^\top \quad (\in \mathbb{R}^{m_b \times m_b})$$

Thus, the within-bucket $m_b \times m_b$ matrix has diagonal entries $(1 + \delta^2)\sigma^2$ and off-diagonal entries $\delta^2\sigma^2$. By Sherman-Morrison we have that the inverse matrices are

$$\Sigma_{\text{blk}, b}^{-1} = \frac{1}{\sigma^2} \left(I_{m_b} - \frac{\delta^2}{1 + m_b \delta^2} \mathbf{1}_{m_b} \mathbf{1}_{m_b}^\top \right)$$

Since the X distribution is fixed, $KL(\bar{P}_0^n \| \bar{P}_1^n) = \mathbb{E}_X [KL(\bar{P}_0^n(Y_{1:n}|X_{1:n}) \| \bar{P}_1^n(Y_{1:n}|X_{1:n}))]$. Under \bar{P}_j , $Y_{1:n}|X_{1:n} \sim \mathcal{N}(f_j(X_{1:n}), \Sigma)$, so

$$\begin{aligned}
KL(\bar{P}_0^n(Y_{1:n}|X_{1:n})\|\bar{P}_1^n(Y_{1:n}|X)_{1:n}) &= \frac{1}{2}(f_1(X_{1:n}) - f_0(X_{1:n}))^T \Sigma^{-1} (f_1(X_{1:n}) - f_0(X_{1:n})) \\
(f_0 \equiv 0) &= \frac{1}{2\sigma^2} \sum_{b=1}^{B_X} (f_1(X_{1:n})_b^T \Sigma_{blk,b}^{-1} f_1(X_{1:n})_b) \\
&= \frac{1}{2\sigma^2} \sum_{b=1}^{B_X} \left(\sum_{i \in b} f_1(X_i)^2 - \frac{\delta^2}{1+m_b\delta^2} \left(\sum_{i \in b} f_1(X_i) \right)^2 \right)
\end{aligned}$$

Now set $\hat{f}_b = \frac{1}{m_b} \sum_{i \in b} f_1(X_i)$, and for $X_i = x$, set $d_i = f_1(X_i) - \hat{f}_b(x)$. Note that for all b , $\sum_{i \in b} d_i = 0$.

$$\begin{aligned}
&= \frac{1}{2\sigma^2} \sum_{b=1}^{B_X} \left(\sum_{i \in b} (\hat{f}_b + d_i)^2 - \frac{\delta^2}{1+m_b\delta^2} (m_b \hat{f}_b)^2 \right) \\
&= \frac{1}{2\sigma^2} \sum_{b=1}^{B_X} \left(m_b \hat{f}_b^2 + \sum_{i \in b} d_i^2 - \frac{\delta^2}{1+m_b\delta^2} m_b^2 \hat{f}_b^2 \right)
\end{aligned}$$

Now taking the expectation over the distribution of $X_{1:n}$,

$$\begin{aligned}
KL(\bar{P}_0^n\|\bar{P}_1^n) &= \mathbb{E}_X \left[\frac{1}{2\sigma^2} \sum_{b=1}^{B_X} \left(m_b \hat{f}_b^2 + \sum_{i \in b} d_i^2 - \frac{\delta^2}{1+m_b\delta^2} m_b^2 \hat{f}_b^2 \right) \right] \\
(\text{Jensen's}) &\leq \frac{1}{2\sigma^2} m \left(1 - \frac{m\delta^2}{1+m\delta^2} \right) \mathbb{E}_X \left[\sum_{b=1}^{B_X} \hat{f}_b^2 \right] + \mathbb{E}_X \left[\sum_{i=1}^n d_i^2 \right] \\
&= \frac{1}{2\sigma^2} m \left(\frac{1}{1+m\delta^2} \right) \mathbb{E}_X \left[\sum_{b=1}^{B_X} \hat{f}_b^2 \right] + \mathbb{E}_X \left[\sum_{i=1}^n d_i^2 \right] \\
&= \frac{1}{2\sigma^2} \left(\frac{n}{1+\frac{n}{B_X}\delta^2} \right) \frac{1}{B_X} \mathbb{E}_X \left[\sum_{b=1}^{B_X} \hat{f}_b^2 \right] + \mathbb{E}_X \left[\sum_{i=1}^n d_i^2 \right]
\end{aligned}$$

Now, setting $f_b = \mathbb{E}[f(X)|X \in b]$, we have

$$\mathbb{E}[\hat{f}_b^2] \leq \mathbb{E}[\hat{f}_b^2|m_b > 0] = f_b^2 + \text{Var}(f(X)|X \in b) \mathbb{E}\left[\frac{1}{m_b}|m_b > 0\right] \leq f_b^2 + \text{Var}(f(X)|X \in b)$$

Since $f_1(x) = Lh^\beta K\left(\frac{x-x_0}{h}\right)$, it is $L_k h^{\beta-1}$ Lipschitz, where $L_k = L\|\nabla K\|_\infty$. Combining this with the fact that for X supported on $[a, b]$, $\text{Var}(X) \leq \frac{1}{4}(b-a)^2$, and plugging it in above,

$$\mathbb{E}[\hat{f}_b^2] \leq \left(\frac{1}{4} \left(\frac{L_k h^{\beta-1}}{B_X} \right)^2 + (Lh^\beta \|K\|_\infty)^2 \right) \mathbf{1}\{b \cap [x_0 - h/2, x_0 + h/2] \neq \emptyset\}$$

Where the indicator reflects that K vanishes outside of $[-1/2, 1/2]$. Since an h fraction of the buckets have nonzero f values

$$\frac{1}{B_X} \sum_{b=1}^{B_X} \mathbb{E}[\hat{f}_b^2] \leq h \times h^{2\beta} \left(\frac{L_k^2}{4B_X^2 h^2} + L^2 K_{\max}^2 \right) = h^{2\beta+1} CL^2 K_{\max}^2 + O\left(\frac{h^{2\beta-1}}{B_X^2}\right)$$

Again, since f is $L_k h^{\beta-1}$ -Lipschitz, we know $d_i \leq \frac{L_k h^{\beta-1}}{B_X} \mathbf{1}\{b \cap [x_0 - h/2, x_0 + h/2] \neq \emptyset\}$

$$\sum_{i=1}^n d_i^2 = \sum_{b=1}^{B_X} \sum_{i \in b} d_i^2 \leq \sum_{b=1}^{B_X} \left(\mathbf{1}\{b \cap [x_0 - h/2, x_0 + h/2] \neq \emptyset\} \sum_{i \in b} \left(\frac{L_k h^{\beta-1}}{B_X} \right)^2 \right) \leq L_k^2 h^{2\beta-1} \frac{n}{B_X^2}$$

Putting everything together we have

$$\begin{aligned}
KL(\bar{P}_0\|\bar{P}_1) &\leq \frac{1}{2\sigma^2} \left(\frac{n}{1+n\frac{\delta^2}{B_X}} \right) \left(CL^2 K_{\max}^2 h^{2\beta+1} + O(h^{2\beta-1}/(B_X^2)) \right) + L_k^2 h^{2\beta-1} \frac{n}{B_X^2} \\
&= C_{\sigma, L, K} n_{\text{eff}} h^{2\beta+1} + O\left(\frac{h^{2\beta-1} n}{B_X^2}\right)
\end{aligned}$$

We set $h = n_{\text{eff}}^{-1/(2\beta+1)}$ to get that the first term is constant and the second term behaves as

$$n_{\text{eff}}^{-\frac{2\beta-1}{2\beta+1}} n B_X^{-2} \approx n^{1-\frac{2\beta-1}{2\beta+1}} B_X^{-2} (1 + n B_X)^{\frac{2\beta-1}{2\beta+1}} \approx \max(n^{1-\frac{2\beta-1}{2\beta+1}} B_X^{-2}, n B_X^{-2+\frac{2\beta-1}{2\beta+1}}) = \max(n^{\frac{2}{2\beta+1}} B_X^{-2}, n B_X^{-\frac{2\beta+3}{2\beta+1}})$$

Since $2 > \frac{2}{2\beta+1}$, and $\frac{2\beta+3}{2\beta+1} > 1$ for any $\beta > 0$, in the regime where $n/B_X \rightarrow c$, these terms are $o(1)$.

□