Nonparametric Indirect Active Learning

Shashank Singh SHASHANKSSINGH44@GMAIL.COM
Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Abstract

Typical models of active learning assume a learner can directly manipulate or query a covariate X to study its relationship with a response Y. However, if X is a feature of a complex system, it may be possible only to indirectly influence X by manipulating a control variable Z, a scenario we refer to as Indirect Active Learning. Under a nonparametric fixed-budget model of Indirect Active Learning, we study minimax convergence rates for estimating a local relationship between X and Y, with different rates depending on the complexities and noise levels of the relationships between Z and X and between X and Y. We also derive minimax rates for passive learning under comparable assumptions, finding in many cases that, while there is an asymptotic benefit to active learning, this benefit is fully realized by a simple two-stage learner that runs two passive experiments in sequence.

1. Introduction

Traditional models of active learning and experimental design assume a learner can directly manipulate or query a covariate X in order to probe its influence on a response Y. In many cases, however, a learner's ability to influence X may be limited. For example, the covariate X might be the result of a complex process that the learner can influence and measure but not completely control. Genetic modifications intended to target specific genes often have limited precision, causing unintended random or systematic changes in other off-target genes and making it challenging to measure the effect of the intended modification (Cellini et al., 2004; Hendel et al., 2015; Wang and Wang, 2019). Similarly, it is often unclear whether manipulations conducted in psychological experiments reliably induce a desired psychological state in a participant, necessitating the use of manipulation checks (Festinger, 1953; Lench et al., 2014). In the most general case, studied here, the learner's influence over X might be governed by an unknown blackbox function and subject to noise.

This paper analyzes nonparametric regression under an *Indirect Active Learning* model, illustrated in Figure 1, in which, rather than specifying the covariate X, the learner specifies a control variable Z and then observes both a resulting covariate value $X = g(Z) + \sigma_X \varepsilon_X$ and a resulting response $Y = f(X) + \sigma_Y \varepsilon_Y$. Here, f and g are unknown functions, ε_X is unobserved centered additive *covariate noise* of level $\sigma_X \geq 0$, and ε_Y is unobserved centered additive *response noise* of level $\sigma_Y \geq 0$. The learner repeats this process iteratively, using past observations of (X, Y, Z) to select a new value for the control variable Z in each iteration, until a prespecified learning budget has been met. The learner's goal is to estimate

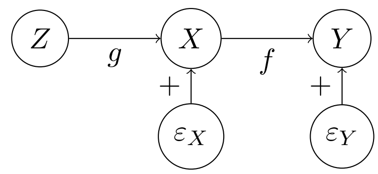

Figure 1: Graphical model of a single iteration of Indirect Active Learning.

the expected response $f(x_0)$ corresponding to a prespecified target covariate value x_0.[1] To do this efficiently, the learner must query "informative" values of Z, i.e., those such that $g(Z) + \sigma_X \varepsilon_X$ is likely to be near to x_0. Thus, in contrast to traditional active learning, which involves only exploitation (selecting the most informative values of X), Indirect Active Learning involves an exploration-exploitation trade-off, as the learner must both search for the most informative choices of Z and query these choices repeatedly to mitigate noise.

The present paper begins to characterize the statistical properties of optimal learning under this Indirect Active Learning model. Specifically, in Section 5, we study minimax convergence rates under a nonparametric model of Indirect Active Learning with a fixed budget, obtaining different rates depending on the complexities of the functions f and g and the noise levels σ_X and σ_Y. Our upper bounds utilize an algorithm (Algorithm 1 in Section 5) consisting of two passive learning stages, a "pure exploration" stage followed by a "pure exploitation" stage. We derive high-probability upper bounds on the error of this algorithm as well as minimax lower bounds (on the error of any algorithm) that match our upper bounds in most, though not all, cases. Because passive experiments are typically much easier to carry out (and parallelize) than sequential active experiments, the optimality of this simple two-stage algorithm in many settings has important practical implications for optimal experimental design.

To understand the benefits of active learning under indirect control, Section 4 derives minimax rates for passive learning under comparable assumptions, which differ, along with the resulting minimax rates, from those of typical nonparametric regression. In the passive case, the standard k-nearest neighbor (k-NN) regressor turns out to be minimax rate-optimal. However, if either the covariate or response noise level σ_X or σ_Y is sufficiently large, the two-stage learner described above asymptotically outperforms the k-NN regressor, demonstrating an advantage of (a very limited form of) active learning over passive learning.

2. Problem Setup

We now describe the passive and active learning problems we study. Due to space limitations, some notation and regularity conditions (specifically, "Hölder ball", "dimension", and "sub-Gaussian") used here are formally defined in Appendix A.

Passive Setting In the passive setting, for some known $x_0 \in \mathcal{X} \subseteq \mathbb{R}^d$, the learner must estimate $f(x_0)$ from pre-collected IID data $\{(X_i, Y_i, Z_i)\}_{i=1}^n \in (\mathcal{X} \times \mathbb{R} \times \mathcal{Z})^n$, assuming:
A1) The data are drawn from the model $X = g(Z) + \varepsilon_X$, $Y = f(X) + \varepsilon_Y$.
A2) The distribution P_Z of Z has dimension d_Z around some $z_0 \in \mathcal{Z}$ with $g(z_0) = x_0$.
A3) The distribution of ε_X has dimension d_X around 0.

1. We note that our results extend straightforwardly to the problem of estimating a contrast, or the difference $f(x_1) - f(x_0)$ in the conditional expectation of Y at two different covariate values x_0 and x_1.

A4) g and f lie in Hölder balls $\mathcal{C}^{s_g}(\mathcal{Z};\mathcal{X};L_g)$ and $\mathcal{C}^{s_f}(\mathcal{X};\mathbb{R};L_f)$, respectively.
A5) ε_X and ε_Y are sub-Gaussian.

We note that our bounds will depend on the intrinsic dimensions d_Z of Z and d_X of ε_X (defined in Appendix A), but not on the extrinsic dimension d of X.

Active Setting In the active setting, at each time step i, the learner must select a new sample Z_i based on the past data $\{(X_i, Y_i, Z_i)\}_{j=1}^{i-1}$. The learner may also utilize the "prior" distribution P_Z of Z from the passive case.[2] After selecting Z_i, the learner observes the covariate response pair (X_i, Y_i), and the process repeats. After n iterations of this process, the learner must estimate $f(x_0)$ (for some $x_0 \in \mathcal{X}$ known in advance), using the dataset $\{(X_i, Y_i, Z_i)\}_{i=1}^{n}$. The learner may also assume **A1)-A5)** from the passive setting.

3. Related Work

While we are not aware of prior theoretical analyses of active learning under the above Indirect Active Learning model, some closely related problems have been studied previously:

Active Learning with Feature Noise Our generative model (Figure 1) generalizes the Berkson errors-in-variables regression model (Berkson, 1950), the special case when $\mathcal{X} = \mathcal{Z}$ and g is known to be the identify function. The Berkson model, which has been studied extensively in passive supervised learning (Buonaccorsi, 2010; Carroll et al., 2006), was used by Ramdas et al. (2014) to study the effects of feature noise in active binary nonparametric classification. To the best of our knowledge, this is the only prior theoretical analysis of active learning in the setting where the learner cannot directly select the covariate.

Active Learning on Manifolds If the dimension of Z is smaller than that of X and g is sufficiently smooth, $g(Z)$ lies on a low-dimensional manifold in $\mathcal{X}$. This is a natural model under which to study active learning on an unknown manifold (i.e., under unknown structural constraints on the covariate $\mathcal{X}$). To the best of our knowledge, no statistical theory has been developed for this problem, although practical methods have been proposed (Zhou and Sun, 2014; Li et al., 2020; Huang et al., 2020; Sreenivasaiah et al., 2021).

Instrumental Variables Methods Our model of Indirect Active Learning (Figure 1) assumes that the relationship between X and Y is unconfounded given Z, i.e., that ε_Y is independent of X and that Y is conditionally independent of ε_X given X. In general, there may exist confounders that affect both X and Y, leading to the causal model assumed by instrumental variable methods, which attempt to use an unconfounded *instrumental variable* (here, Z) to deconfound the relationship between X and Y, typically by projecting X onto its variation caused by Z (Baiocchi et al., 2014). In the passive case, Singh et al. (2019) provided minimax rates for nonparametric instrumental variable regression, but we are not aware of work characterizing the potential benefits of actively manipulating the instrumental variable Z. Indeed, one motivation for the present work is to provide a baseline for such investigations by characterizing minimax-optimal learning in the simpler unconfounded case.

2. Although not a prior on z_0 in the typical Bayesian sense, P_Z plays a similar role in that the estimators' performance will likely depend on how concentrated P_Z is around z_0.

4. Bounds for Passive Learning

To provide an appropriate baseline for performance in Indirect Active Learning, we first consider the problem of *passively* estimating $f(x_0)$ from an IID dataset $\{(X_i, Y_i, Z_i)\}_{i=1}^n\}$. Upon first glance, since observations of Z provide no information about the relationship between X and Y, one might suppose that this passive problem reduces to standard passive regression, in which we have only IID observations of (X, Y). However, while observations of Z are indeed uninformative here, the assumption $X = g(Z) + \sigma_X \varepsilon_X$ imposes constraints on the distribution of X that differ from the standard passive regression setting, resulting in significantly different convergence rates. Obtaining minimax rates for the passive case will therefore facilitate interpretation of our main results for the active case, especially to understand the benefit of active learning over passive learning.

In the passive case, we will use a k-nearest neighbor (k-NN) regression estimate of $f(x_0)$:

Definition 1 (Nearest-Neighbor Order and k-Nearest Neighbor Regressor) *Given a point x_0 lying in a metric space $(\mathcal{X}, \rho)$ and a finite dataset $X_1, ..., X_n \in \mathcal{X}$, the* nearest-neighbor order $\pi(x_0; \{X_1, ..., X_n\})$ *is a permutation of $[n]$ such that*

$$\rho\left(x_0, X_{\pi_1(x_0;\{X_1,...,X_n\})}\right) \leq \rho\left(x_0, X_{\pi_2(x_0;\{X_1,...,X_n\})}\right) \leq \cdots \leq \rho\left(x_0, X_{\pi_n(x_0;\{X_1,...,X_n\})}\right),$$

with ties broken arbitrarily. The k-nearest neighbor regressor $\widehat{f}_{x_0,k}$ of f at x_0 is defined by

$$\widehat{f}_{x_0,k} := \frac{1}{k}\sum_{i=1}^{k} Y_{\pi_i(x_0;\{X_1,...,X_n\})}, \tag{1}$$

Crucially, in contrast to most other approaches to nonparametric regression, k-NN regression automatically adapts to unknown low-dimensional structure in the distribution of $X_1, ..., X_n$ (Kpotufe, 2011), which is important in our setting because the model $X = g(Z) + \sigma_X \varepsilon_X$ can constrain the intrinsic dimension of X. In contrast, we expect many methods, such as kernel methods with uniform bandwidth, to be sub-optimal in our setting. The following result, proven in Appendix B, bounds the error of the k-NN estimate (1):

Theorem 2 (Passive Upper Bound) *Under Assumptions A1)-A5), there exists a constant $C > 0$ such that, for k chosen according to Eq. (5) in Appendix B, for any $\delta \in (0, 1)$, with probability $\geq 1 - \delta$, $\left|\widehat{f}_{x_0,k} - f(x_0)\right| \leq C\Phi_P \log 1/\delta$, where*

$$\Phi_P = \max\left\{\left(\frac{1}{n}\right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n}\right)^{\frac{s_f}{d_X + d_Z/s_g}}, \left(\frac{\sigma_Y^2}{n}\right)^{\frac{s_f}{2s_f + d_Z/s_g}}, \left(\frac{\sigma_X^{d_X}\sigma_Y^2}{n}\right)^{\frac{s_f}{2s_f + d_X + d_Z/s_g}}\right\}. \tag{2}$$

Before discussing this result further, we present a matching lower bound, proven in Appendix C, showing that the rate Φ_P in Theorem 2 is the optimal in the passive case:

Theorem 3 (Passive Lower Bound) *There exist constants $c_1, c_2, n_0 > 0$, depending only on L_g, s_g, L_f, s_f, d_X, d_Z, such that, for all $n \geq n_0$, for any passive estimator $\widehat{f}_{x_0}$,*

$$\sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq c_1\Phi_P\right] \geq c_2,$$

where the supremum is taken over all g, f, P_Z, ε_X, ε_Y satisfying Assumptions A1)-A5).

Comparing Theorems 2 and 3, we see that the k-NN regressor in Eq. (1) achieves the minimax optimal rate Φ_P, given in Eq. (2), for the passive case. Each of the terms in Φ_P can dominate, depending on the relative magnitudes of σ_X, σ_Y, and n. If the covariate noise level σ_X is sufficiently small, one can show that the distribution of X concentrates near a manifold of dimension d_Z/s_g within $\mathcal{X}$. Hence, depending on the response noise level σ_Y, we obtain the rate $\asymp n^{-\frac{s_f s_g}{d_Z}}$ or $\asymp n^{-\frac{s_f}{2s_f + d_Z/s_g}}$, corresponding to the rate of nonparametric regression over a covariate space of dimension d_Z/s_g, in the absence of response noise (Kohler, 2014) or the presence of response noise (Tsybakov, 2008), respectively. Increasing the level σ_X of covariate noise leads to slower rates, respectively $\asymp n^{-\frac{s_f}{d_X + d_Z/s_g}}$ or $\asymp n^{-\frac{s_f}{2s_f + d_X + d_Z/s_g}}$, equivalent to increasing the dimension of the covariate space by d_X.

5. Bounds for Active Learning

We now turn to the active case. Our upper bounds use a "two-stage" algorithm, detailed in Algorithm 1, which sequentially performs two passive experiments, a "pure exploration" experiment followed by a "pure exploitation" experiment. In the first stage, the algorithm randomly samples the control variable Z according to the given "prior" distribution P_Z (as in the passive algorithm). Each sample value of Z is repeated several times, and the resulting values of X are averaged to determine which observed value of Z is most likely to return a value of X near x_0. The algorithm then repeatedly samples this value of Z (hopefully obtaining a large number of samples with X near x_0), and finally applies the k-NN regressor from Eq. (1) over the resulting dataset. Because passive experiments are

Algorithm 1: Two-stage active learning algorithm.

Input: Budget n, integers $\ell, k \in [1, n/2]$, distribution P_Z on $\mathcal{Z}$, target point $x_0 \in \mathcal{X}$
Output: Estimate $\widehat{f}_{x_0}$ of x_0

$m \leftarrow \lfloor n/2 \rfloor$
for $i \in [\lfloor m/\ell \rfloor]$ **do** `// "pure exploration" stage`
  Draw $Z_i \sim P_Z$
  for $l \in [\ell]$ **do**
    Observe $X_{i,l} = g(Z_i) + \sigma_X \varepsilon_{X,i,l}$ and $Y_{i,l} = f(X_{i,l}) + \sigma_Y \varepsilon_{Y,i,l}$
  Compute $\overline{X}_i = \sum_{l=1}^{\ell} X_{i,l}$
$i^* \leftarrow \operatorname{argmin}_{i \in [\lfloor m/\ell \rfloor]} \left\| \overline{X}_i - x_0 \right\|$
for $i \in [n - m]$ **do** `// "pure exploitation" stage`
  Sample $X_{m+i} = g(Z_{i^*}) + \sigma_X \varepsilon_{X,m+i}$ and $Y_{m+i} = f(X_{m+i}) + \sigma_Y \varepsilon_{Y,m+i}$
return $\widehat{f}_{x_0} = \frac{1}{k} \sum_{j=1}^{k} Y_{\pi_j(x_0; \{X_1, \ldots, X_n\})}$

often much easier to carry out (and parallelize) than sequential active experiments, this two-stage design may be practical in many settings where a fully iterative procedure would be expensive or time-consuming to implement. Nevertheless, as discussed below, Algorithm 1 achieves the optimal convergence rate in most, although not all, experimental settings, and typically converges much faster than the passive minimax rate. In particular, we have the following upper bound error, proven in Appendix D:

Theorem 4 (Active Upper Bound) *Let $\delta \in (0,1)$, and let $\widehat{f}_{x_0}$ denote the estimator proposed in Algorithm 1, with k and ℓ chosen according to Eqs.* (30) *and* (31) *in Appendix D. Then, under Assumptions A1)-A5), for some constant $C > 0$ depending only on L_g, s_g, L_f, s_f, d_X, and d_Z, with probability $\geq 1-\delta$, $\left|\widehat{f}_{x_0} - f(x_0)\right| \leq C\Phi_A^* \log 1/\delta$, where*

$$\Phi_A^* := \max\left\{\left(\frac{1}{n}\right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n}\right)^{\frac{s_f}{d_X}}, \frac{\sigma_Y}{\sqrt{n}}, \left(\frac{\sigma_Y^2 \sigma_X^{d_X}}{n}\right)^{\frac{s_f}{2s_f+d_X}}, \left(\frac{\sigma_X^2 \log n}{n}\right)^{\frac{s_f s_g}{2s_g+d_Z}}\right\}. \tag{3}$$

Meanwhile, we have the following minimax lower bound, proven in Appendix E, on the error of any active learner:

Theorem 5 (Lower Bounds for the Active Case) *There exist constants $C, c > 0$, depending only on L_g, s_g, L_f, s_f, d_X, and d_Z, such that, for any active estimator $\widehat{f}_{x_0}$*

$$\sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_{\{(Z_i,X_i,Y_i)\}_{i=1}^n}\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq C\Phi_{A,*}\right] \geq c,$$

where the supremum is over all g, f, P_Z, ε_X, and ε_Y satisfying Assumptions A1)-A5), and

$$\Phi_{A,*} = \max\left\{\left(\frac{1}{n}\right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n}\right)^{s_f/d_X}, \frac{\sigma_Y}{\sqrt{n}}, \left(\frac{\sigma_X^{d_X}\sigma_Y^2}{n}\right)^{\frac{s_f}{2s_f+d_X}}\right\}. \tag{4}$$

Comparison of Active Upper and Lower Bounds Comparing the upper bound Φ_A^* (Eq. (3)) and lower bound $\Phi_{A,*}$ (Eq. (4)), $\Phi_A^* = \max\left\{\Phi_{A,*}^*, \left(\frac{\sigma_X^2 \log n}{n}\right)^{\frac{s_f s_g}{2s_g+d_Z}}\right\}$; i.e., the first four terms in Φ_A^* match the corresponding terms in $\Phi_{A,*}$, while the fifth term in Φ_A^* appears to be sub-optimal. Hence, Algorithm 1 is optimal whenever σ_Y is sufficiently large or both d_X and σ_X are sufficiently small; specifically, for σ_Y small, the sub-optimal term $\left(\frac{\sigma_X^2 \log n}{n}\right)^{\frac{s_f s_g}{2s_g+d_Z}}$ dominates Φ_A^* only when $n^{-\frac{s_g}{d_Z}} \lesssim_{\log} \sigma_X \lesssim_{\log} n^{\frac{2s_g+d_Z}{d_X d_Z} - \frac{s_g}{d_Z}}$.

Comparison of Active and Passive Bounds The first terms of the passive minimax rate Φ_P (Eq. (2)) and our upper bound rate Φ_A^* (Eq. (3)) match. Meanwhile, Φ_A^* improves on the second, third, and fourth terms of Φ_P by removing the d_Z/s_g terms from the exponents. Meanwhile, the last term of Φ_A^* dominates Φ_P only if σ_Y is sufficiently small and both $d_X \leq 2$ and $\sigma_X \gtrsim_{\log} n^{-d_Z/s_g}$. In particular, whenever σ_Y is sufficiently large or both $d_X > 2$ and $\sigma_X \gtrsim_{\log} n^{-d_Z/s_g}$, Algorithm 1 strictly outperforms the best passive algorithm.

6. Conclusions & Future Work

This paper studied the nonparametric minimax theory of Indirect Active Learning, as well as a corresponding passive baseline. We showed that active learning asymptotically outperforms passive learning on many cases, and that much of this improvement is realized by a two-stage algorithm, which uses only a very limited form of active learning and may be much easier to implement than a fully active algorithm. Besides closing the gap between our active upper and lower bounds, interesting directions for future work include exploring active methods for learning *global parametric* relationships between X and Y (e.g., $f(x) = x^\intercal\beta$) and extending our results to the instrumental variable setting, as discussed in Section 3.

Acknowledgments

The author would like to thank Jia-Jie Zhu, Krikamol Maundet, and Bernhard Schölkopf for helpful discussions. This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B).

References

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

Joseph Berkson. Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180, 1950.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

John P Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, 2010.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Francesco Cellini, Andrew Chesson, Ian Colquhoun, Anne Constable, HV Davies, Kjeld H Engel, Angharad MR Gatehouse, Sirpa Kärenlampi, Esther J Kok, J-J Leguay, et al. Unintended effects and their detection in genetically modified crops. *Food and Chemical Toxicology*, 42(7):1089–1125, 2004.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014.

Leon Festinger. Laboratory experiments. *Research methods in the behavioral sciences*, pages 136–172, 1953.

Ayal Hendel, Eli J Fine, Gang Bao, and Matthew H Porteus. Quantifying on-and off-target genome editing. *Trends in biotechnology*, 33(2):132–140, 2015.

Lei Huang, Yuqing Ma, and Xianglong Liu. A general non-parametric active learning framework for classification on multiple manifolds. *Pattern Recognition Letters*, 130: 250–258, 2020.

Michael Kohler. Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis*, 132:197–208, 2014.

Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.

Heather C Lench, Aaron B Taylor, and Shane W Bench. An alternative approach to analysis of mental states in experimental social cognition research. *Behavior Research Methods*, 46(1):215–228, 2014.

Hang Li, Enrique Del Castillo, and George Runger. On active learning methods for manifold data. *Test*, 29(1):1–33, 2020.

Aaditya Ramdas, Barnabas Poczos, Aarti Singh, and Larry Wasserman. An analysis of active learning with uniform feature noise. In *Artificial Intelligence and Statistics*, pages 805–813. PMLR, 2014.

Philippe Rigollet. 18. s997: High dimensional statistics. *Lecture Notes), Cambridge, MA, USA: MIT Open-CourseWare*, 2015.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. *Advances in neural information processing systems*, 29, 2016.

Deepthi Sreenivasaiah, Johannes Otterbach, and Thomas Wollmann. Meal: Manifold embedding-based active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1029–1037, 2021.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Diane Catherine Wang and Xiangdong Wang. Off-target genome editing: A new discipline of gene science and a new class of medicine, 2019.

Jin Zhou and Shiliang Sun. Active learning of gaussian processes with manifold-preserving graph reduction. *Neural Computing and Applications*, 25(7):1615–1625, 2014.

Appendix A. Notation

This appendix formally defines some notation and regularity conditions used throughout the main paper.

Given sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ of positive real numbers, we write $a_n \lesssim b_n$ if $\limsup_{n \to \infty} a_n/b_n < \infty$, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. When ignoring polylogarithmic factors, we write $\lesssim_{\log}$ (to mean that, for some $p > 0$, $a_n \lesssim b_n (\log b_n)^p$), etc. For a positive integer n, $[n] = \{1, ..., n\}$ denotes the set of positive integers at most n. For $x \in \mathbb{R}$, $\lfloor x \rfloor = \max\{z \in \mathbb{Z} : z \leq x\}$ is the greatest integer at most x. For $x \in \mathbb{R}^d$, $\|x\| = \max_j |x_j|$ and $\|x\|_2 = \sqrt{\sum_j x_j^2}$ are the ℓ_∞- and ℓ_2-norms of x.

We now define three regularity conditions needed for our learning problem. The first is *Hölder smoothness*, used to measure the complexities of the functions f and g:

Definition 6 (Hölder Seminorm, Space, and Ball) *For any* $s \in [0, 1]$ *and metric spaces* $(\mathcal{Z}, \rho_{\mathcal{Z}})$ *and* $(\mathcal{X}, \rho_{\mathcal{X}})$*, the* Hölder seminorm $\|\cdot\|_{\mathcal{C}^s(\mathcal{Z};\mathcal{X})}$ *is defined for* $f : \mathcal{Z} \to \mathcal{X}$ *by*

$$\|f\|_{\mathcal{C}^s(\mathcal{Z};\mathcal{X})} := \sup_{x \neq y \in \mathcal{X}} \frac{\rho_{\mathcal{Z}}\left(f(x), f(y)\right)}{\rho_{\mathcal{X}}^s(x, y)}.$$

The Hölder space $\mathcal{C}^s(\mathcal{Z};\mathcal{X})$ *is*

$$\mathcal{C}^s(\mathcal{Z};\mathcal{X}) := \{f : \mathcal{Z} \to \mathcal{X} \text{ such that } \|f\|_{\mathcal{C}^s(\mathcal{Z};\mathcal{X})} < \infty\}.$$

For $L \geq 0$*, the* Hölder ball $\mathcal{C}^s(\mathcal{Z};\mathcal{X};L)$ *is*

$$\mathcal{C}^s(\mathcal{Z};\mathcal{X};L) := \{f : \mathcal{Z} \to \mathcal{X} \mid \|f\|_{\mathcal{C}^s(\mathcal{Z};\mathcal{X})} \leq L\}.$$

When the spaces $\mathcal{Z}$ *and* $\mathcal{X}$ *are clear from context, we omit these and simply write* $\mathcal{C}^s$*.*

We next provide a definition of *local dimension* that will be used to measure intrinsic complexities of the random variables Z and ε_X:

Definition 7 (Local Dimension) *A random variable* X *taking values in a metric space* $(\mathcal{X}, \rho)$ *has* local dimension d *around a point* $x \in \mathcal{X}$ *if* $\liminf_{r \downarrow 0} r^{-d} \mathbb{P}\left[\rho(X, x) \leq r\right] > 0$*.*

Finally, we state a *sub-Gaussian tail condition*, required for noise variables ε_X and ε_Y:

Definition 8 (Sub-Gaussian Random Variable) *For any integer* $d > 0$*, an* $\mathbb{R}^d$*-valued random variable* ε *is said to be sub-Gaussian if, for all* $t \in \mathbb{R}^d$*,* $\mathbb{E}\left[e^{\langle \varepsilon, t\rangle}\right] \leq e^{\|t\|_2^2/2}$*.*

Sub-Gaussianity is usually defined with a parameter $\sigma \geq 0$ indicating the scale of ε. In this work, to be consistent with other assumptions (namely, the local dimension assumption on ε_X), we separate the scale parameter σ, explicitly writing $\sigma\varepsilon$ as needed.

Appendix B. Upper Bound for the Passive Case (Theorem 2)

In this Appendix, we prove our main upper bound for the passive case. We begin by reiterating the result:

Theorem 2 (Main Passive Upper Bound) *Suppose that* Z *has dimension* d_Z *around some* $z_0 \in \mathcal{Z}$ *with* $g(z_0) = x_0$*. Suppose* ε_X *is sub-Gaussian and has dimension* d_X *around* 0*. Suppose that the response noise* ε_Y *is sub-Gaussian. Suppose that* $g \in \mathcal{C}^{s_g}(\mathcal{Z}, \mathcal{X})$ *and* $f \in \mathcal{C}^{s_f}(\mathcal{X}, \mathcal{Y})$*. Then, there exists a constant* $C > 0$ *such that, for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta/2 - 2e^{-k/4}$*,*

$$\left|\widehat{f}_{x_0,k} - f(x_0)\right| \leq C \left(\sigma_Y \sqrt{\frac{\log 1/\delta}{k}} + \max\left\{ \left(\left(\frac{k}{n}\right)^{\frac{s_g}{d_Z}} \sqrt{\log \frac{n}{\delta}} \right)^{s_f}, \left(\frac{k \sigma_X^{d_X}}{n} \right)^{\frac{s_f}{d_X + d_Z/s_g}} \right\} \right).$$

In particular, setting

$$k = \max\left\{ 4 \log \frac{4}{\delta}, \min\left\{ \sigma_Y^{\frac{2(d_X + d_Z/s_g)}{2s_f + d_X + d_Z/s_g}} \left(\frac{n}{\sigma_X^{d_X}} \right)^{\frac{2s_f}{2s_f + d_X + d_Z/s_g}}, \sigma_Y^{\frac{2d_Z}{2s_f s_g + d_Z}} n^{\frac{2s_f s_g}{2s_f s_g + d_Z}} \right\} \right\}, \quad (5)$$

gives, with probability $\geq 1-\delta$, $\left|\widehat{f}_{x_0,k} - f(x_0)\right| \lesssim \Phi_P \log \frac{1}{\delta}$, *where*

$$\Phi_P = \max\left\{ \left(\frac{1}{n}\right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n}\right)^{\frac{s_f}{d_X + d_Z/s_g}}, \left(\frac{\sigma_Y^2}{n}\right)^{\frac{s_f}{2s_f + d_Z/s_g}}, \left(\frac{\sigma_X^{d_X}\sigma_Y^2}{n}\right)^{\frac{s_f}{2s_f + d_X + d_Z/s_g}} \right\}.$$

Before proving this theorem, we state and prove a few useful lemmas. Our first lemma shows that, under the Local Dimension assumption (Definition 7), the distribution of k-NN distances is tightly concentrated. Such results are commonly used in the analysis of k-NN methods (Kpotufe, 2011; Chaudhuri and Dasgupta, 2014; Biau and Devroye, 2015; Singh and Póczos, 2016):

Lemma 9 (Convergence of k-NN Distance) *Let X be a random variable taking values in a metric space $(\mathcal{X}, \rho)$, and let $X_1, ..., X_n$ be IID observations of X. Suppose X has local dimension d around a point $x \in \mathcal{X}$. Then there exist constants $C, \kappa > 0$ such that, whenever $k/n \leq \kappa$, with probability at least $1 - e^{-k/4}$, the distance $\rho(x, X_{\pi_k(x;\{X_1,...,X_n\})})$ from x to its k-nearest neighbor in $\{X_1, ..., X_n\}$ is at most*

$$\rho(x, X_{\pi_k(x;\{X_1,...,X_n\})}) \leq C\left(\frac{k}{n}\right)^{1/d}.$$

Proof By definition of local dimension, there exist $c, r^* > 0$ such that, for all $r \in (0, r^*]$, $\mathbb{P}[\rho(X,x) \leq r] \geq cr^d$. Suppose $\frac{k}{n} \leq \frac{c(r^*)^d}{2}$, so that, for $r := \left(\frac{2k}{cn}\right)^{1/d}$, $r \leq r^*$. Then,

$$\mathbb{P}[\rho(X,x) \leq r] \geq \frac{2k}{n}.$$

Since, for each $i \in [n]$, $1\{\rho(X_i,x) \leq r\} \sim \text{Bernoulli}(\mathbb{P}[\rho(X,x) \leq r])$, by a multiplicative Chernoff bound, with probability at least $1 - e^{-k/4}$,

$$\sum_{i=1}^n 1\{\rho(X_i,x) \leq r\} \geq k,$$

which implies $\rho(x, X_{\pi_k(x;\{X_1,...,X_n\})}) \leq r$. ∎

Our next lemma shows that local dimension is (sub-)additive:

Lemma 10 *Consider two independent $\mathbb{R}^d$-valued random variables X_1 and X_2. Under any norm $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$, suppose X_1 has local dimension d_1 around x_1 and X_2 has local dimension d_2 around x_2. Then, $X_1 + X_2$ has local dimension $d_1 + d_2$ around $x_1 + x_2$.*

Proof [Proof of Lemma 10] By the triangle inequality and independence of X and ε,

$$\begin{aligned}\Pr[\|X_1 + X_2 - (x_1 + x_2)\| \leq r] &\geq \Pr[\|X_1\| \leq r/2 \text{ AND } \|X_2\| \leq r/2] \\ &= \Pr[\|X_1\| \leq r/2]\Pr[\|X_2\| \leq r/2].\end{aligned}$$

The result follows from the definition of local dimension. ■

Finally, the following lemma shows how local dimension of a random variable changes when applying a Hölder-smooth function:

Lemma 11 (Dimension of Pushforward) *Let $(\mathcal{X}, \rho_\mathcal{X})$ and $(\mathcal{Y}, \rho_\mathcal{Y})$ be metric spaces. Let X be a random variable taking values in $\mathcal{X}$, and let $f \in \mathcal{C}^s(\mathcal{X}, \mathcal{Y})$ be a Hölder continuous map from $\mathcal{X}$ to $\mathcal{Y}$. If X has local dimension d around a point $x \in \mathcal{X}$ in $(\mathcal{X}, \rho_\mathcal{X})$, then $f(X)$ has local dimension d/s around $f(x)$ in $(\mathcal{Y}, \rho_\mathcal{Y})$.*

Proof If $\|f\|_{\mathcal{C}^s(\mathcal{X},\mathcal{Y})} = 0$, then f is constant and the result follows trivially. Otherwise, by definition of Hölder continuity,

$$\begin{aligned}
\liminf_{r \downarrow 0} \frac{\mathbb{P}\left[\rho_\mathcal{Y}(f(X), f(x)) \leq r\right]}{r^{d/s}} &= \|f\|_{\mathcal{C}^s(\mathcal{X},\mathcal{Y})}^{-d} \liminf_{r \downarrow 0} \frac{\mathbb{P}\left[\rho_\mathcal{Y}(f(X), f(x)) \leq L r^s\right]}{r^d} \\
&\geq \|f\|_{\mathcal{C}^s(\mathcal{X},\mathcal{Y})}^{-d} \liminf_{r \downarrow 0} \frac{\mathbb{P}\left[\rho_\mathcal{X}(X, x) \leq r\right]}{r^d} > 0,
\end{aligned}$$

since X has dimension d around x. ■

We are now ready to prove Theorem 2. The key idea of the proof is that, combining Lemmas 10 and 11, X has dimension $d_X + d_Z/s_g$ around x_0. Together with Lemma 9, this implies that the distance between x and its k-nearest neighbor among $\{X_1, ..., X_n\}$ is of order $\left(\frac{k}{n}\right)^{\frac{1}{d_X + d_Z/s_g}}$, determining the bias due to smoothing implicit in the k-NN regressor. The full proof is as follows:

Proof [Proof of Theorem 2] For any $X_1, ..., X_n \in \mathcal{X}$, let

$$\widetilde{f}_{x_0, X_1, \ldots, X_n, k} = \frac{1}{k} \sum_{i=1}^{k} f(X_{\pi_i(x; \{X_1, \ldots, X_n\})})$$

denote the conditional expectation of $\widehat{f}_{x_0,k}$ given $X_1, ..., X_n$. By the triangle inequality,

$$\left|\widehat{f}_{x_0,k} - f(x_0)\right| \leq \left|\widehat{f}_{x_0,k} - \widetilde{f}_{x_0,X_1,\ldots,X_n,k}\right| + \left|\widetilde{f}_{x_0,X_1,\ldots,X_n,k} - f(x_0)\right|, \tag{6}$$

where the first term captures stochastic error due to the response noise ε_Y and the second term captures smoothing bias.

For the first term in Eq. (6), since the response noise $\{\varepsilon_{Y,i}\}_{i=1}^n$ is assumed to be additive,

$$\left|\widehat{f}_{x_0,k} - \widetilde{f}_{x_0,X_1,\ldots,X_n,k}\right| = \left|\frac{1}{k} \sum_{i=1}^{k} \sigma_Y \varepsilon_{Y, \pi_i(x_0; \{X_1, \ldots, X_n\})}\right|.$$

Hence, noting that the response noise $\{\varepsilon_{Y,i}\}_{i=1}^n$ is independent of the covariates $\{X_i\}_{i=1}^n$, by a standard concentration inequality for sub-Gaussian random variables (e.g., Lemma 22 in Appendix F), with probability at least $1 - \delta/2$,

$$\left|\widehat{f}_{x_0,k} - \widetilde{f}_{x_0,X_1,\ldots,X_n,k}\right| \leq \sigma_Y \sqrt{\frac{2}{k} \log \frac{4}{\delta}}. \tag{7}$$

For the second term in Eq. (6), by the triangle inequality and Hölder smoothness of f, for some constant $C_1 > 0$,

$$\begin{aligned}
\left|\widetilde{f}_{x_0,X_1,...,X_n,k} - f(x_0)\right| &= \left|\frac{1}{k}\sum_{i=1}^{k} f(X_{\pi_i(x_0;\{X_1,...,X_n\})}) - f(x_0)\right| \\
&\leq \frac{1}{k}\sum_{i=1}^{k} \left|f(X_{\pi_i(x_0;\{X_1,...,X_n\})}) - f(x_0)\right| \\
&\leq C_1 \frac{1}{k}\sum_{i=1}^{k} \left\|x_0 - X_{\pi_i(x_0;\{X_1,...,X_n\})}\right\|^{s_f} \\
&\leq C_1 \left\|x_0 - X_{\pi_k(x_0;\{X_1,...,X_n\})}\right\|^{s_f}. \qquad (8)
\end{aligned}$$

Hence, the remainder of this proof is devoted to bounding the distance $\left\|x_0 - X_{\pi_k(x_0;\{X_1,...,X_n\})}\right\|$ from x_0 to its k-nearest neighbor in $X_1, ..., X_n$.

Combining Lemmas 9, 10, and 11 gives, for some constant $C_2 > 0$, with probability at least $1 - e^{-k/4}$,

$$\|x_0 - X_{\pi_k(x_0;\{X_1,...,X_n\})}\| \leq C_2 \left(\frac{k\sigma_X^{d_X}}{n}\right)^{\frac{1}{d_X + d_Z/s_g}}. \qquad (9)$$

Now suppose that $(k/n)^{s_g/d_Z} \geq \sigma_X$. By the triangle inequality,

$$\|x_0 - X_{\pi_k(x_0;\{X_1,...,X_n\})}\| \leq \|x_0 - X_{\pi_k(x_0;\{g(Z_1),...,g(Z_n)\})}\| + \max_{i\in[n]} \|\varepsilon_i\|. \qquad (10)$$

Combining Lemmas 9 and 11, there exists a constant $C_4 > 0$ such that

$$\|x_0 - X_{\pi_k(x_0;\{g(Z_1),...,g(Z_n)\})}\| \leq C_4 \left(\frac{k}{n}\right)^{\frac{s_g}{d_Z}}. \qquad (11)$$

Meanwhile, by a standard maximal inequality for sub-Gaussian random variables (Lemma 21 in Appendix F), for some constant $C_3 > 0$, with probability at least $1 - \delta/2$,

$$\max_{i\in[n]} \|\varepsilon_i\| \leq \sigma_X \sqrt{2\log\frac{4n}{\delta}} \leq C_3 \left(\frac{k}{n}\right)^{s_g/d_Z} \sqrt{\log\frac{n}{\delta}}. \qquad (12)$$

Plugging Inequalities (11) and (12) into Inequality (10) gives, with probability at least $1 - e^{-k/4}$,

$$\|x_0 - X_{\pi_k(x_0;\{X_1,...,X_n\})}\| \leq (C_3 + C_4) \left(\frac{k}{n}\right)^{\frac{s_g}{d_Z}} \sqrt{\log\frac{n}{\delta}}. \qquad (13)$$

One can check that, whenever $(k/n)^{s_g/d_Z} < \sigma_X$, the bound (9) dominates the bound (13). Hence, can take the maximum of these two bounds and omit the condition $(k/n)^{s_g/d_Z} \geq \sigma_X$; i.e., for some constant $C_5 > 0$, we always have

$$\|x_0 - X_{\pi_k(x_0;\{X_1,...,X_n\})}\| \leq C_5 \max\left\{\left(\frac{k\sigma_X^{d_X}}{n}\right)^{\frac{1}{d_X + d_Z/s_g}}, \left(\frac{k}{n}\right)^{\frac{s_g}{d_Z}} \sqrt{\log\frac{n}{\delta}}\right\}.$$

plugging this into Inequality (8) gives the desired result. ■

Appendix C. Lower Bounds for the Passive Case (Theorem 3)

In this section, we prove our lower bounds on minimax risk in the passive case. Throughout this section, we assume $\mathcal{Z} = [0,1]^{d_Z}$ and $\mathcal{X} = \mathbb{R}^{d_X}$.

Theorem 3 (Main Passive Lower Bound) *In the passive case, there exist constants $c_1, c_2, n_0 > 0$, depending only on L_g, s_g, L_f, s_f, d_X, and d_Z, such that, for all $n \geq n_0$,*

$$\sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq c_1 \Phi_P \right] \geq c_2, \tag{14}$$

where

$$\Phi_P = \max \left\{ \left(\frac{1}{n} \right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n} \right)^{\frac{s_f}{d_X + d_Z/s_g}}, \left(\frac{\sigma_Y^2}{n} \right)^{\frac{s_f s_g}{2 s_f s_g + d_Z}}, \left(\frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{s_f}{2 s_f + d_X + d_Z/s_g}} \right\}.$$

The remainder of this section is devoted to proving Theorem 3. Since Inequality (14) is a maximum of four lower bounds, we can prove each of these lower bounds separately. Proposition 15 will provide the two terms depending on σ_Y. Proposition 16 will provide the $n^{-\frac{s_f s_g}{d_Z}}$ term. Finally, Proposition 17 will provide the term depending on σ_X but not on σ_Y.

We begin by construct a family $\Theta_{g,M}$ of functions $g : \mathcal{Z} \to \mathcal{X}$ that will be used throughout the proofs:

Lemma 12 *For any $M \in \mathbb{N}$, $s_g, L_g > 0$, there exists a family $\Theta_{g,M} \subseteq \mathcal{C}^{s_g}(\mathcal{Z}; \mathcal{X}; L_g)$ of s_g-Hölder continuous functions from $\mathcal{Z}$ to $\mathcal{X}$ such that:*

1. *$|\Theta_{g,M}| = M^{d_Z}$.*
2. *for each $g \in \Theta_{g,M}$, there exists $z_g \in \mathcal{Z}$ with $g(z_g) = x_0$.*
3. *for some $g_0 \in \mathcal{X}$ and some constant $C_{d_Z,s_g} > 0$ depending only on d_Z and s_g,*
   - *(a) $\|g_0 - x_0\|_\infty = C_{d_Z,s_g} L_g M^{-s_g}$.*
   - *(b) for $g \neq g' \in \Theta_{g,M}$, the functions $x \mapsto 1\{g(x) \neq g_0\}$, $x \mapsto 1\{g'(x) \neq g_0\}$ have disjoint supports in $\mathcal{Z}$.*

Intuitively, the family $\Theta_{g,M}$ is constructed such that, for any $z \in \mathcal{Z}$, there is no more than one $g \in \Theta_{g,M}$ such that $\|g(z) - x_0\|_\infty \lesssim M^{-s_g}$; hence, for the typical $g \in \Theta_{g,M}$, $\lesssim 1/|\Theta_{g,M}|$ of the samples $X_1, ..., X_n$ will lie within $\asymp M^{-s_g}$ of x_0. The formal construction of $\Theta_{g,M}$ is as follows:

Proof For any positive integer d, let $K_d : \mathbb{R}^d \to \mathbb{R}$ denote the standard d-dimensional "bump function" defined by

$$K_d(x) = \exp\left(1 - \frac{1}{1 - \|x\|_2^2}\right) 1\{\|x\|_2 \leq 1\} \quad \text{for all } x \in \mathbb{R}^d. \tag{15}$$

Note in particular that K_d satisfies the following:

1. $K_d \in \mathcal{C}^\infty(\mathbb{R}^d)$; i.e., K_d is infinitely differentiable.
2. For all $x \notin (-1,1)^d$, $K_d(x) = 0$.
3. $\|K_d\|_\infty = K(0) = 1$.

For any $m \in [M]^{d_Z}$, define $g_m : \mathcal{Z} \to \mathcal{X}$ by

$$g_m(z) = x_0 + \frac{L_g}{M^{s_g} \|K_{d_Z}\|_{\mathcal{C}^{s_g}}} \left(K_{d_Z}(0) - K_{d_Z}\left((M+1)z - m \right) \right) v, \quad \text{for all } z \in \mathcal{Z}, \tag{16}$$

where v is an arbitrary unit vector in $\mathbb{R}^{d_X}$. Then, define $\Theta_{g,M} := \{g_m : m \in [M]^{d_Z}\}$. Clearly, $|\Theta_{g,M}| = M^{d_Z}$. By Eq. (16), $\Theta_{g,M} \subseteq \mathcal{C}^{s_g}(\mathcal{Z}; \mathcal{X}; L_g)$. Moreover, for each $m \in [M]^d$, $g_m(m/(M+1)) = x_0$, while, whenever $\|(M+1)z - m\|_\infty \geq 1$, $g_m(z) = g_0 := x_0 + \frac{L_g}{\|K_{d_Z}\|_{\mathcal{C}^{s_g}}} M^{-s_g} v$. ■

Next, we present a general information-theoretic lower bound in the "stochastic case", i.e., in the presence of response noise ($\sigma_Y > 0$). Together with the construction of the family $\Theta_{g,M}$ in Lemma 12, this lemma will be the basis for our lower bounds depending on σ_Y, in both the active and passive cases:

Lemma 13 (General Lower Bound with Response Noise) *Suppose $Y_i | X_i \sim \mathcal{N}(f(X_i), \sigma_Y^2)$, and suppose that the distribution of (Z_i, X_i) depends only on the preceding observations $\{(Z_j, X_j, Y_j)\}_{j=1}^{i-1}$. Let $h > 0$, and let P_g be any probability distribution over $\mathcal{C}^{s_g}(\mathcal{Z}; \mathcal{X}; L_g)$. Then, there exist $f_0, f_1 \in \mathcal{C}^{s_f}(\mathcal{X}; \mathcal{Y}; L_f)$ such that, if $P_0 = P_g \times \delta_{f_0}$ is the product distribution of P_g and a unit mass on f_0, and P_1 is the product distribution of P_g and a unit mass on f_1, for some constant $C_{d_X, s_f} > 0$ depending only on d_X and s_f,*

$$\inf_{\widehat{f}_{x_0}} \sup_{g, f, P_Z, \varepsilon_X, \varepsilon_Y} \Pr_D \left[|\widehat{f}_{x_0} - f(x_0)| \geq C_{d_X, s_f} L_f h^{s_f} \right] \geq \frac{1}{2} \left(1 - C_{d_X, s_f} \frac{L_f h^{s_f}}{\sigma_Y} \sqrt{ \sum_{i=1}^n \mathop{\mathbb{E}}_{D, g \sim P_0} \left[1\{ \|X_i - x_0\|_\infty < h \} \right] } \right). \tag{17}$$

Here, the infimum is taken over all estimators $\widehat{f}_{x_0}$ (i.e., all functions $\widehat{f}_{x_0} : (\mathcal{Z} \times \mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$, the supremum is taken over all $g \in \mathcal{C}^{s_g}(\mathcal{Z}; \mathcal{X}; L_g)$, $f \in \mathcal{C}^{s_f}(\mathcal{X}; \mathcal{Y}; L_f)$, sub-Gaussian covariate noise random variables ε_X with dimension d_X around 0, and sub-Gaussian response noise random variables ε_Y, and we used the shorthand $D := \{(Z_i, X_i, Y_i)\}_{i=1}^n$ to denote the entire dataset.

Intuitively, Lemma 13 provides a lower bound in terms of the expected number of covariate observations X_i that are "informative" about $f(x_0)$, i.e., those that lie sufficiently near x_0. The family $\Theta_{g,M}$ was constructed in Lemma 12 such that Hence, it remains only to bound the expected number of such X_i, which will be small for g lying in the family $\Theta_{g,M}$. Note that inequality (17) applies in both active and passive settings. However, in the passive setting, the distribution of each Z_i will be independent of g, whereas, in the active setting, Z_i may depend on g through the preceding observations (specifically, $(Z_1, X_1), ..., (Z_{i-1}, X_{i-1})$).

The proof of Lemma 13 is based on the method of two fuzzy hypotheses, specifically the following lemma:

Lemma 14 (Theorem 2.15 of Tsybakov (2008), Total Variation Case) *Let P_0 and P_1 be probability distributions over Θ, and let $F : \Theta \to \mathbb{R}$. Suppose*

$$\sup_{\theta \in \operatorname{supp}(P_0)} F(\theta) \leq 0, \quad \text{and} \quad s := \inf_{\theta \in \operatorname{supp}(P_1)} F(\theta) > 0. \tag{18}$$

Then,

$$\inf_{\widehat{F}} \sup_{\theta \in \Theta} \Pr_{\theta} \left[|\widehat{F} - F(\theta)| \geq s/2 \right] \geq \frac{1 - D_{TV}(P_0, P_1)}{2}.$$

The method of two fuzzy hypotheses is one of several standard methods for deriving minimax lower bounds on statistical estimation error (see Chapter 2 of Tsybakov (2008) for further discussion of this and other methods). The method of two fuzzy hypotheses, in particular, is typically used in semiparametric problems where one wishes to estimate a parametric quantity (here, $\widehat{f}(x_0)$) that is subject to variation in a nonparametric "nuisance" quantity (here, g). The key idea of the method is to lower bound the supremum over the nuisance quantity (here, $\sup_g$) by in terms of an appropriate information-theoretic measure (here D_{TV}) computed over an appropriately chosen distribution of the nuisance variable (here, P_g). By choosing this distribution carefully (as we will do later in this section, using the family $\Theta_{g,M}$ constructed in Lemma 12), the information-theoretic measure will give a tight lower bound on the probability of error.

We now present the proof of Lemma 13.

Proof Let $h > 0$. Define $f_0, f_1 : \mathcal{X} \to \mathcal{Y}$ by

$$f_0(x) = 0 \quad \text{and} \quad f_1(x) = \frac{L_f h^{s_f}}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}} K_{d_X}\left(\frac{x - x_0}{h}\right) \quad \text{for all } x \in \mathcal{X}, \tag{19}$$

where K_{d_X} is as defined in Eq. (15). Clearly,

$$\sup_{(g,f) \in \operatorname{supp}(P_0)} f(x_0) = 0,$$

while

$$s := \inf_{(g,f) \in \operatorname{supp}(P_1)} f(x_0) = \frac{L_f h^{s_f}}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}} K_{d_X}(0) = \frac{L_f h^{s_f}}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}}, \tag{20}$$

satisfying the condition (18). It remains only to bound $D_{\text{TV}}(P_0, P_1)$. For each $i \in [n]$, let $P_{0,i}$ and $P_{1,i}$ denote the conditional distributions of

$$(Z_i, X_i, Y_i) \quad \text{given} \quad (Z_1, X_1, Y_1), ..., (Z_{i-1}, X_{i-1}, Y_{i-1})$$

under P_0 and P_1, respectively, and let $P_{Y,0,i}$ and $P_{Y,1,i}$ denote the conditional distributions of

$$Y_i \quad \text{given} \quad (Z_1, X_1, Y_1), ..., (Z_{i-1}, X_{i-1}, Y_{i-1}), (Z_i, X_i)$$

under P_0 and P_1, respectively. By a standard formula for the KL divergence between two Gaussian random variables,

$$D_{\text{KL}}\left(P_{Y,0,i} \| P_{Y,1,i}\right) = \frac{f_1^2(X_i)}{2\sigma_Y^2}.$$

Moreover, by construction of P_0 and P_1, the conditional distributions of

$$(Z_i, X_i) \quad \text{given} \quad (Z_1, X_1, Y_1), ..., (Z_{i-1}, X_{i-1}, Y_{i-1})$$

are identical under P_0 and P_1. Therefore, by the chain rule for KL divergence,

$$\begin{aligned}
D_{\text{KL}}(P_0, P_1) &= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[D_{\text{KL}}(P_{0,i} \| P_{1,i}) \right] \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[D_{\text{KL}}(P_{Y,0,i} \| P_{Y,1,i}) \right] \\
&= \frac{1}{2\sigma_Y^2} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[f_1^2(X_i) \right] \\
&\leq \frac{\|f_1\|_\infty^2}{2\sigma_Y^2} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[1\{f_1(X_i) \neq 0\} \right] \\
&= \frac{L_f^2 h^{2s_f}}{2\sigma_Y^2 \|K_{d_X}\|_{\mathcal{C}^{s_f}}^2} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[1\{f_1(X_i) \neq 0\} \right] \\
&\leq \frac{L_f^2 h^{2s_f}}{2\sigma_Y^2 \|K_{d_X}\|_{\mathcal{C}^{s_f}}^2} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[1\{\|X_i - x_0\|_\infty < h\} \right],
\end{aligned}$$

where the last inequality follows from the fact that f_1 is non-zero only on the open ℓ_∞ ball of radius h centered at x_0 (see Eq. (19)). Thus, by Pinsker's Inequality,

$$D_{\text{TV}}(P_0, P_1) \leq \sqrt{2 D_{\text{KL}}(P_0, P_1)} \leq \frac{L_f h^{s_f}}{\sigma_Y \|K_{d_X}\|_{\mathcal{C}^{s_f}}} \sqrt{\sum_{i=1}^{n} \mathop{\mathbb{E}}_{D,g\sim P_0} \left[1\{\|X_i - x_0\|_\infty < h\} \right]}$$

The result now follows by plugging this bound on $D_{\text{TV}}(P_0, P_1)$ into Lemma 14. ■

We now utilize Lemma 13 to prove the following minimax lower bound for the passive case under response noise:

Proposition 15 (Minimax Lower Bound, Passive Case, σ_Y large) *In the passive case, there exists a constant $C > 0$ depending only on d_Z, d_X, s_g, L_g, s_f, L_f such that*

$$\sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq C \max \left\{ \left(\frac{\sigma_Y^2}{n} \right)^{\frac{s_f s_g}{2 s_f s_g + d_Z}}, \left(\frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{s_f}{2 s_f + d_X + d_Z/s_g}} \right\} \right] \geq \frac{1}{4}.$$

Proof Consider the case in which $\sigma_X \varepsilon_{X,i}$ is uniformly distributed over the cube $[-\frac{\sigma_X}{\sqrt{d_X}}, \frac{\sigma_X}{\sqrt{d_X}}]^{d_X}$. Let

$$h := \begin{cases} \left(\frac{\|K_{d_X}\|_{\mathcal{C}^{s_f}}^2}{2^{2+d_Z/s_g}} \frac{L_g^{d_Z/s_g}}{L_f^2} \frac{\sigma_Y^2}{n} \right)^{\frac{1}{2s_f + d_Z/s_g}} : & \text{if } \sigma_X \leq d_X^{-1/2} \left(\frac{\|K_{d_X}\|_{\mathcal{C}^{s_f}}^2}{2^{2+d_Z/s_g}} \frac{L_g^{d_Z/s_g}}{L_f^2} \frac{\sigma_Y^2}{n} \right)^{\frac{1}{2s_f + d_Z/s_g}} \\ \left(\frac{\|K_{d_X}\|_{\mathcal{C}^{s_f}}^2}{2^{2+d_Z/s_g} d_X^{d_X/2}} \frac{L_g^{d_Z/s_g}}{L_f^2} \frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{1}{2s_f + d_X + d_Z/s_g}} : & \text{else} \end{cases} \tag{21}$$

and

$$M := \left(\frac{L_g}{\|K_{d_Z}\|_{\mathcal{C}^{s_g}} \left(h + \frac{\sigma_X}{\sqrt{d_X}} \right)} \right)^{1/s_g}. \tag{22}$$

Suppose P_g is the uniform distribution over $\Theta_{g,M}$, given in Lemma 12.

By the Law of Total Probability,

$$\begin{aligned}
\mathop{\mathbb{E}}_{D,g\sim P_0} [1\{\|X_i - x_0\|_\infty < h\}] &= \Pr_{D,g\sim P_0} [\|g(Z_i) + \sigma_X \varepsilon_{X,i} - x_0\|_\infty < h] \\
&\leq \Pr_{D,g\sim P_0} [\|g(Z_i) + \sigma_X \varepsilon_{X,i} - x_0\|_\infty < h | g(Z_i) \neq g_0] \Pr_{D,g\sim P_0} [g(Z_i) \neq g_0] \\
&+ \Pr_{D,g\sim P_0} [\|g(Z_i) + \sigma_X \varepsilon_{X,i} - x_0\|_\infty < h | g(Z_i) = g_0] \qquad (23)
\end{aligned}$$

Since in the passive case, g is uniformly distributed over $\Theta_{g,M}$ independently of Z_i and the elements of $\Theta_{g,M}$ have disjoint supports,

$$\Pr_{D,g\sim P_0} [g(Z_i) \neq g_0] \leq \frac{1}{|\Theta_{g,M}|} = M^{-d_Z} = \left(\frac{\|K_{d_Z}\|_{\mathcal{C}^{s_g}}}{L_g} \left(h + \frac{\sigma_X}{\sqrt{d_X}} \right) \right)^{d_Z/s_g}.$$

Since $\varepsilon_{X,i}$ has a uniform distribution with mean 0, the function $x \mapsto \Pr[\|\sigma_X \varepsilon_{X,i} + x\| \leq h]$ is maximized at $x = 0$, and so, similar to Inequality (39),

$$\begin{aligned}
\Pr_{D,g\sim P_0} [\|g(Z_i) + \sigma_X \varepsilon_{X,i} - x_0\|_\infty < h | g(Z_i) \neq g_0] &\leq \Pr_{D,g\sim P_0} [\|\sigma_X \varepsilon_{X,i}\|_\infty < h | g(Z_i) \neq g_0] \\
&= \min \left\{ 1, \left(\frac{h\sqrt{d_X}}{\sigma_X} \right)^{d_X} \right\}.
\end{aligned}$$

Finally, since $\sigma_X \varepsilon_{X,i}$ is uniformly distributed over the cube $[-\frac{\sigma_X}{\sqrt{d_X}}, \frac{\sigma_X}{\sqrt{d_X}}]^{d_X}$, by our choice (Eq. (22)) of M, if $g(Z_i) = g_0$, then there is no overlap between the supports of X_i and f_1; formally,

$$\begin{aligned}
\Pr_{D,g\sim P_0} [\|g(Z_i) + \sigma_X \varepsilon_{X,i} - x_0\|_\infty < h | g(Z_i) = g_0] &= \left(\max \left\{ 0, x_0 + h - g_0 + \frac{\sigma_X}{\sqrt{d_X}} \right\} \right)^{d_X} \\
&= \left(\max \left\{ 0, h + \frac{\sigma_X}{\sqrt{d_X}} - \frac{L_g}{M^{s_g} \|K_{d_Z}\|_{\mathcal{C}^{s_g}}} \right\} \right)^{d_X} = 0.
\end{aligned}$$

Plugging these into Inequality (23) gives

$$\mathop{\mathbb{E}}_{D,g\sim P_0} [1\{\|X_i - x_0\|_\infty < h\}] \leq \min \left\{ 1, \left(\frac{h\sqrt{d_X}}{\sigma_X} \right)^{d_X} \right\} \left(\frac{\|K_{d_Z}\|_{\mathcal{C}^{s_g}}}{L_g} \left(h + \frac{\sigma_X}{\sqrt{d_X}} \right) \right)^{d_Z/s_g}, \tag{24}$$

and plugging this into Lemma (13) gives

$$\begin{aligned}
&\inf_{\widehat{f}_{x_0}} \sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D \left[|\widehat{f}_{x_0} - f(x_0)| \geq C h^{s_f} \right] \\
&\qquad \geq \frac{1}{2} \left(1 - C \frac{h^{s_f}}{\sigma_Y} \sqrt{n \min \left\{ 1, \left(\frac{h\sqrt{d_X}}{\sigma_X} \right)^{d_X} \right\} \left(h + \frac{\sigma_X}{\sqrt{d_X}} \right)^{d_Z/s_g}} \right),
\end{aligned}$$

where $C > 0$ is a constant depending only on d_X, L_f, s_f, d_Z, L_g, and s_g. Plugging in the appropriate value of h, according to Eq. (21), depending on whether σ_X is greater or less than $h/\sqrt{d_X}$, gives

$$\inf_{\widehat{F}} \sup_{\theta \in \Theta} \Pr_{\theta} \left[|\widehat{F} - F(\theta)| \geq s/2 \right] \geq \frac{1}{4},$$

where

$$s \asymp \begin{cases} \left(\frac{\sigma_Y^2}{n} \right)^{\frac{s_f s_g}{2 s_f s_g + d_Z}} : & \text{if } \sigma_X \sqrt{d_X} \leq \left(\frac{1}{2^{2+d_Z/s_g}} \frac{L_g^{d_Z/s_g}}{L_f^2} \frac{\|K_{d_X}\|_{\mathcal{C}^{s_f}}^2}{\|K_{d_Z}\|_{\mathcal{C}^{s_g}}^{s_f/s_g}} \frac{\sigma_Y^2}{n} \right)^{\frac{1}{2s_f + d_Z/s_g}} \\ \left(\frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{s_f}{2s_f + d_X + d_Z/s_g}} : & \text{else} \end{cases}.$$

■

Proposition 15 provided minimax lower bounds for the passive case under response noise of scale $\sigma_Y > 0$, based on information-theoretic bounds showing that the information obtainable from a single pair (X_i, Y_i) decays as σ_Y^{-2}; hence, a different approach is needed to obtain tight bounds when σ_Y is very small or 0. Our next results provide tight lower bounds for the case when σ_Y is small or 0 using a combinatorial, rather than information-theoretic, approach. In particular, we show that, given a sufficiently large family $\Theta_{g,M}$ of functions g with $x \mapsto 1\{g(x) \neq g_0\}$ having disjoint supports, the probability of observing any X_i with $g(X_i) \neq g_0$ is small.

Proposition 16 (Minimax Lower Bound, Active and Passive Case, σ_X small, σ_Y small) *In both active and passive cases,*

$$\inf_{\widehat{f}_{x_0}} \sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq C \left(\frac{1}{n} \right)^{\frac{s_f s_g}{d_Z}} \right] \geq \frac{1}{4}, \tag{25}$$

where C is a constant (given in Eq. (26)) depending only on L_g, s_g, L_f, s_f, d_X, and d_Z.

We note that, in contrast to other the other lower bounds in this section, Proposition 16 applies to to both the active and passive settings, as its proof makes no assumptions on the distribution of the control variables $Z_1, ..., Z_n$.

Proof Let $\Theta_{g,M}$ be as given in Lemma 12, with $M := \lceil (2n)^{1/d_Z} \rceil$. Additionally, let f_0 and f_1 be as in Eq. (19). For each $g \in \Theta_{g,M}$, let E_g denote the event $g(Z_1) = g(Z_2) = \cdots = g(Z_n) = g_0$. Since, for any $g_1 \neq g_2 \in \Theta_{g,M}$, the functions $z \mapsto 1\{g_1(z) \neq g_0\}$ and $z \mapsto 1\{g_2(z) \neq g_0\}$ have disjoint support, for any $Z_1, ..., Z_n$,

$$\sum_{g \in \Theta_{g,M}} 1 - 1_{E_g} \leq n.$$

Since, by our choice of M, $|\Theta_{g,M}| = M^{d_Z} \geq 2n$, for any (potentially randomized) estimator $\widehat{f}_{x_0}$,

$$\begin{aligned}
|\Theta_{g,M}| \max_{g \in \Theta_{g,M}} \Pr[E_g] &\geq \sum_{g \in \Theta_{g,M}} \Pr[E_g] \\
&= \mathbb{E}\left[\sum_{g \in \Theta_{g,M}} 1_{E_g}\right] \\
&= \mathbb{E}\left[\sum_{g \in \Theta_{g,M}} 1 - (1 - 1_{E_g})\right] \geq |\Theta_{g,M}| - n \geq n,
\end{aligned}$$

and so there exists $g_{\widehat{f}_{x_0}} \in \Theta_{g,M}$ such that

$$\Pr\left[E_{g_{\widehat{f}_{x_0}}} \middle| g = g_{\widehat{f}_{x_0}}\right] \geq \frac{n}{|\Theta_{g,M}|} \geq \frac{n}{((2n)^{1/d_Z} + 1)^{d_Z}} \geq 2^{-d_Z - 1}.$$

For $h := \|x_0 - g_0\|_\infty = \frac{L_g}{M^{s_g} \|K_{d_Z}\|_{\mathcal{C}^{s_g}}}$, under the event E_g, the distribution of the data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is independent of whether $f = f_0$ or $f = f_1$. In particular, either

$$\Pr\left[\widehat{f}_{x_0} \geq \frac{f_0(x_0) + f_1(x_0)}{2} \middle| f = f_0\right] \geq \frac{1}{2} \quad \text{or} \quad \Pr\left[\widehat{f}_{x_0} \leq \frac{f_0(x_0) + f_1(x_0)}{2} \middle| f = f_1\right] \geq \frac{1}{2},$$

and so

$$\max_{g \in \Theta_{g,M}, f \in \{f_0, f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq \frac{|f_1(x_0) - f_0(x)|}{2}\right] \geq \frac{1}{2} \Pr[E_{g_{\widehat{f}_{x_0}}}] \geq 2^{-d_Z - 2}.$$

By construction of f_0 and f_1,

$$|f_1(x_0) - f_0(x_0)| = \frac{L_f h^{s_f}}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}},$$

and so plugging in our choices of M and h gives

$$\max_{g \in \Theta_{g,M}, f \in \{f_0, f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq \frac{L_f L_g^{s_f}}{2\|K_{d_X}\|_{\mathcal{C}^{s_f}} \|K_{d_Z}\|_{\mathcal{C}^{s_g}}^{s_f} \left((2n)^{s_f s_g / d_Z} + 1\right)}\right] \geq 2^{-d_Z - 2}. \tag{26}$$

■

Proposition 17 (Minimax Lower Bound, Passive Case, σ_X large, σ_Y small) *In the passive case, for some constant $C > 0$, whenever $n \geq C \sigma_X^{-d_Z/s_g}$,*

$$\inf_{\widehat{f}_{x_0}} \sup_{g, f, P_Z, \varepsilon_X, \varepsilon_Y} \Pr_D\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq C \left(\frac{\sigma_X^{d_X}}{n}\right)^{\frac{s_f}{d_X + d_Z / s_g}}\right] \geq \frac{1}{2e} > 0, \tag{27}$$

where $C > 0$ (given below in Eq. (29)) depends only on L_g, s_g, L_f, s_f, d_X, and d_Z.

We note that the condition $n \gtrsim \sigma_X^{-d_Z/s_g}$ is precisely when the rate $\left(\sigma_X^{d_X}/n\right)^{\frac{s_f}{d_X+d_Z/s_g}}$ in Proposition 17 dominates the rate $n^{\frac{-s_f s_g}{d_Z}}$ obtained earlier in Proposition 16; hence, this condition can be omitted when we take the maximum of these lower bounds in Theorem 3.

Proof Suppose Z is uniformly distributed over $\mathcal{Z}$ and $\sigma_X \varepsilon_{X,i}$ is uniformly distributed on the set $\{0\} \times [-\frac{\sigma_X}{\sqrt{d_X}}, \frac{\sigma_X}{\sqrt{d_X}}]^{d_X} \subseteq \mathbb{R}^{d_X+1}$ (i.e., the first coordinate of ε_{X_i} is always 0 and the remaining d_X coordinates are uniformly random). Define $g : \mathcal{Z} \to \mathcal{X}$ by $g(z) = L_g \|z\|_\infty^{s_g} e_1$, where $e_1 \in \mathbb{R}^{d_X+1}$ denotes the first canonical basis vector. It is straightforward to verify that (a) Z has dimension d_Z around 0, (b) $\varepsilon_{X,i}$ is sub-Gaussian, (c) $\varepsilon_{X,i}$ has dimension d_X around 0, and (d) $g \in \mathcal{C}^{s_g}(\mathcal{Z}; \mathcal{X})$. Let f_0 and f_1 be as in Eq. (19) with $x_0 = 0$.

Let $r^* := \min\{L_g, \frac{\sigma_X}{\sqrt{d_X}}\}$. Then, for each $i \in [n]$ and any $r \in [0, r^*]$, since Z_i and $\varepsilon_{X,i}$ are independent,

$$\begin{aligned}
\Pr\left[\|X_i\|_\infty \leq r\right] &= \Pr\left[\max\{\|g(Z_i)\|, \|\varepsilon_i\|_\infty\} \leq r\right] \\
&= \Pr\left[\|Z_i\|_\infty \leq (r/L_g)^{1/s_g}\right] \Pr\left[\|\varepsilon_i\|_\infty \leq r\right] \\
&= \left(\frac{r}{L_g}\right)^{d_Z/s_g} \left(\frac{r\sqrt{d_X}}{\sigma_X}\right)^{d_X}.
\end{aligned}$$

Since, in the passive case, $X_1, ..., X_n$ are IID,

$$\begin{aligned}
\Pr\left[\min_{i\in[n]} \|X_i\|_\infty \geq r\right] &= \left(\Pr\left[\|X_i\|_\infty \geq r\right]\right)^n \\
&= \left(1 - \Pr\left[\|X_i\|_\infty \leq r\right]\right)^n \\
&= \left(1 - \left(\frac{r}{L_g}\right)^{d_Z/s_g} \left(\frac{r\sqrt{d_X}}{\sigma_X}\right)^{d_X}\right)^n.
\end{aligned}$$

For n sufficiently large that $r \leq r^*$ (i.e., $n \geq C\sigma_X^{-d_Z/s_g}$, as assumed), plugging in

$$r = \left(\left(\frac{\sigma_X}{\sqrt{d_X}}\right)^{d_X} \frac{L_g^{d_Z/s_g}}{n}\right)^{\frac{1}{d_X+d_Z/s_g}} \tag{28}$$

gives

$$\Pr\left[\min_{i\in[n]} \|X_i\|_\infty \geq \left(\left(\frac{\sigma_X}{\sqrt{d_X}}\right)^{d_X} \frac{L_g^{d_Z/s_g}}{n}\right)^{\frac{1}{d_X+d_Z/s_g}}\right] = \left(1 - \frac{1}{n}\right)^n \geq 1/e > 0,$$

since $\left(1 - \frac{1}{n}\right)^n$ approaches $1/e$ from above as $n \to \infty$.

Now let f_0 and f_1 be as in Eq. (19), with $h = r$ above. Given the event $\min_{i\in[n]} \|X_i\|_\infty \geq h$, the distribution of the data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is independent of whether $f = f_0$ or $f = f_1$. Hence, as argued in the proof of the previous Proposition (16),

$$\max_{f\in\{f_0,f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq \frac{|f_1(x_0) - f_0(x)|}{2}\right] \geq \frac{1}{2} \Pr\left[\min_{i\in[n]} \|X_i\|_\infty \leq h\right] \geq \frac{1}{2e}.$$

Since, by construction of f_0 and f_1,

$$|f_1(x_0) - f_0(x_0)| = \frac{L_f h^{s_f}}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}},$$

it follows that

$$\max_{f \in \{f_0, f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq \frac{L_f h^{s_f}}{2\|K_{d_X}\|_{\mathcal{C}^{s_f}}}\right] \geq \frac{1}{2e}.$$

Plugging in our choice of $h = r$ from Eq (28) above gives

$$\max_{f \in \{f_0, f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq C\left(\frac{\sigma_X}{n}\right)^{\frac{s_f}{d_X + d_Z/s_g}}\right] \geq \frac{1}{2e},$$

where

$$C = \frac{L_f\left(d_X^{-d_X/2} L_g^{d_Z/s_g}\right)^{\frac{s_f}{d_X + d_Z/s_g}}}{2\|K_{d_X}\|_{\mathcal{C}^{s_f}}}. \tag{29}$$

■

Appendix D. Upper Bound for the Active Case (Theorem 4)

Theorem 4 (Main Active Upper Bound) *Let $\delta \in (0, 1)$ Let $\widehat{f}_{x_0}$ denote the estimator proposed in Algorithm 1, with $k, \ell \geq 4 \log \frac{6}{\delta}$. Then, for some constant $C > 0$ depending only on L_g, s_g, L_f, s_f, d_X, and d_Z, with probability at least $1 - \delta$,*

$$\left|\widehat{f}_{x_0,k} - f(x_0)\right| \leq C_4\left(\sigma_Y\sqrt{\frac{1}{k}\log\frac{1}{\delta}} + \sigma_X^{s_f}\left(\frac{k}{n}\right)^{\frac{s_f}{d_X}} + \left(\frac{\ell}{n}\right)^{\frac{s_f s_g}{d_Z}} + \left(\sigma_X\sqrt{\frac{1}{\ell}\log\frac{n}{\delta}}\right)^{s_f}\right),$$

In particular, setting

$$k \asymp \max\left\{\log\frac{1}{\delta}, \min\left\{n, \left(\frac{\sigma_Y}{\sigma_X^{s_f}}\right)^{\frac{2d_X}{2s_f + d_X}} n^{\frac{2s_f}{2s_f + d_X}}\right\}\right\} \tag{30}$$

and

$$\ell \asymp \max\left\{\log\frac{1}{\delta}, \min\left\{n, \sigma_X^{\frac{2d_Z}{2s_g + d_Z}} n^{\frac{2s_g}{2s_g + d_Z}} (\log n)^{\frac{d_Z}{2s_g + d_Z}}\right\}\right\}, \tag{31}$$

we have, with probability at least $1 - \delta$, $\left|\widehat{f}_{x_0} - f(x_0)\right| \leq C\Phi_A^* \log\frac{1}{\delta}$, where

$$\Phi_A^* := \max\left\{\left(\frac{1}{n}\right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n}\right)^{\frac{s_f}{d_X}}, \frac{\sigma_Y}{\sqrt{n}}, \left(\frac{\sigma_Y^2 \sigma_X^{d_X}}{n}\right)^{\frac{s_f}{2s_f + d_X}}, \left(\frac{\sigma_X^2 \log n}{n}\right)^{\frac{s_f s_g}{2s_g + d_Z}}\right\}.$$

Proof As in the proof of the Passive Upper Bound (Theorem 2), for any $X_1, ..., X_n \in \mathcal{X}$, let

$$\widetilde{f}_{x_0,X_1,...,X_n,k} = \frac{1}{k} \sum_{i=1}^{k} f(X_{\pi_i(x_0;\{X_1,...,X_n\})})$$

denote the conditional expectation of $\widehat{f}_{x_0,k}$ given $X_1, ..., X_n$. By the triangle inequality,

$$\left| \widehat{f}_{x_0,k} - f(x_0) \right| \leq \left| \widehat{f}_{x_0,k} - \widetilde{f}_{x_0,X_1,...,X_n,k} \right| + \left| \widetilde{f}_{x_0,X_1,...,X_n,k} - f(x_0) \right|, \tag{32}$$

where the first term captures stochastic error due to the response noise ε_Y and the second term captures smoothing bias.

For the first term in Eq. (32), as argued in the passive case (see Eq. (33)), since the response noise ε_Y is sub-Gaussian, with probability at least $1 - \delta/2$,

$$\left| \widehat{f}_{x_0,k} - \widetilde{f}_{x_0,X_1,...,X_n,k} \right| \leq \sigma_Y \sqrt{\frac{2}{k} \log \frac{4}{\delta}}. \tag{33}$$

For the second term in Eq. (32), by the triangle inequality and Hölder continuity of f, for some constant $C_1 > 0$,

$$\begin{aligned} \left| \widetilde{f}_{x_0,X_1,...,X_n,k} - f(x_0) \right| &\leq \frac{1}{k} \sum_{i=1}^{k} \left| f(X_{\pi_i(x_0;\{X_1,...,X_n\})}) - f(x_0) \right| \\ &\leq \frac{C_1}{k} \sum_{i=1}^{k} \left\| X_{\pi_i(x_0;\{X_1,...,X_n\})} - x_0 \right\|^{s_f} \\ &\leq C_1 \left\| X_{\pi_k(x_0;\{X_1,...,X_n\})} - x_0 \right\|^{s_f}. \end{aligned} \tag{34}$$

By Lemmas 9 and 11, for some constant $C_2 > 0$, with probability at least $1 - \delta/6$, if $\ell \geq 4 \log \frac{6}{\delta}$,

$$\|g(Z_{\pi_\ell(x_0,\{g(Z_1),...,g(Z_i)\})}) - x_0\| \leq C_2 \left(\frac{\ell}{m} \right)^{s_g/d_Z}, \tag{35}$$

where $m = \lfloor n/2 \rfloor$ as in Algorithm 1. Meanwhile, by a standard maximal inequality for sub-Gaussian random variables (Lemma 21 in Appendix F), with probability at least $1 - \delta/6$,

$$\max_{i \in [\lfloor m/\ell \rfloor]} \left\| g(Z_i) - \overline{X}_i \right\| \leq \sigma_X \sqrt{\frac{2}{\ell} \log \frac{12m}{\delta}}. \tag{36}$$

By Inequality 36 and Inequality 35, with probability at least $1 - 2\delta/6$,

$$\begin{aligned} \left\| g(Z_{\pi_\ell(x_0,\{\overline{X}_1,...,\overline{X}_i\})}) - x_0 \right\| &\leq \left\| \overline{X}_{\pi_\ell(x_0,\{\overline{X}_1,...,\overline{X}_{\lfloor m/\ell \rfloor}\})} - x_0 \right\| + \sigma_X \sqrt{\frac{2}{\ell} \log \frac{8m}{\delta}} \\ &\leq \left\| \overline{X}_{\pi_\ell(x_0,\{\overline{X}_1,...,\overline{X}_{\lfloor m/\ell \rfloor}\})} - x_0 \right\| + \sigma_X \sqrt{\frac{2}{\ell} \log \frac{8m}{\delta}} \\ &\leq C_2 \left(\frac{\ell}{m} \right)^{s_g/d_Z} + \sigma_X \sqrt{\frac{2}{\ell} \log \frac{8m}{\delta}}. \end{aligned}$$

Since $Z_{m+1} = \cdots = Z_n = Z_{\pi_\ell(x_0, \{\overline{X}_1, \ldots, \overline{X}_{\lfloor m/\ell \rfloor}\})}$, by Lemma 9, for some constant $C_3 > 0$, with probability at least $1 - \delta/6$, if $k \geq 4 \log \frac{6}{\delta}$,

$$\left\| X_{\pi_k(x; \{X_{m+1}, \ldots, X_n\})} - g(Z_{\pi_\ell(x_0, \{\overline{X}_1, \ldots, \overline{X}_{\lfloor m/\ell \rfloor}\})}) \right\| \leq C_3 \sigma_X \left(\frac{k}{n - m} \right)^{1/d_X},$$

and it follows from the triangle inequality and the preceding inequality, with probability at least $1 - \delta/2$,

$$\begin{aligned} &\left\| X_{\pi_k(x_0; \{X_1, \ldots, X_n\})} - x_0 \right\| \\ &\leq \left\| X_{\pi_k(g(Z_{\pi_\ell(x_0, \{\overline{X}_1, \ldots, \overline{X}_{\lfloor m/\ell \rfloor}\})}); \{X_{m+1}, \ldots, X_n\})} - g(Z_{\pi_\ell(x_0, \{\overline{X}_1, \ldots, \overline{X}_{\lfloor m/\ell \rfloor}\})}) \right\| \\ &+ \| g(Z_{\pi_\ell(x_0, \{\overline{X}_1, \ldots, \overline{X}_{\lfloor m/\ell \rfloor}\})}) - x_0 \| \\ &\leq C_3 \sigma_X \left(\frac{k}{n - m} \right)^{1/d_X} + C_2 \left(\frac{\ell}{m} \right)^{s_g/d_Z} + \sigma_X \sqrt{\frac{2}{\ell} \log \frac{12m}{\delta}}. \end{aligned}$$

Combining this with inequalities (34), (33), and (32) gives, with probability at least $1 - \delta$, as long as $k, \ell \geq 4 \log \frac{6}{\delta}$,

$$\left| \widehat{f}_{x_0,k} - f(x_0) \right| \leq C_4 \left(\sigma_Y \sqrt{\frac{1}{k} \log \frac{1}{\delta}} + \left(\sigma_X^{d_X} \frac{k}{n} \right)^{\frac{s_f}{d_X}} + \left(\frac{\ell}{n} \right)^{\frac{s_f s_g}{d_Z}} + \left(\sigma_X \sqrt{\frac{1}{\ell} \log \frac{n}{\delta}} \right)^{s_f} \right),$$

for some constant $C_4 > 0$, since $m = \lfloor n/2 \rfloor$. ■

Appendix E. Lower Bounds for the Active Case (Theorem 5)

Theorem 5 (Main Active Lower Bound) *In the active case, there exist constants $C, c > 0$, depending only on L_g, s_g, L_f, s_f, d_X, and d_Z, such that, for any estimator $\widehat{f}_{x_0}$*

$$\inf_{\widehat{f}_{x_0}} \sup_{g, f, P_Z, \varepsilon_X, \varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq C \Phi_{A,*} \right] \geq c,$$

where

$$\Phi_{A,*} = \max \left\{ \left(\frac{1}{n} \right)^{\frac{s_f s_g}{d_Z}}, \left(\frac{\sigma_X^{d_X}}{n} \right)^{s_f/d_X}, \frac{\sigma_Y}{\sqrt{n}}, \left(\frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{s_f}{2s_f + d_X}} \right\}.$$

Proposition 18 (Minimax Lower Bound, Active Case, σ_X large, σ_Y large) *In the active case,*

$$\inf_{\widehat{f}_{x_0}} \sup_{g, f, P_Z, \varepsilon_X, \varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq C \left(\frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{s_f}{2s_f + d_X}} \right] \geq \frac{1}{4}, \tag{37}$$

where $C > 0$ (given in Eq. (40)) depends only on L_f, s_f, and d_X.

Proof As in the proofs of Propositions 15 and 17, consider the case in which $\sigma_X \varepsilon_{X,i}$ is uniformly distributed on the cube $[-\frac{\sigma_X}{\sqrt{d_X}}, \frac{\sigma_X}{\sqrt{d_X}}]^{d_X}$. Let

$$h := \left(\frac{\|K_{d_X}\|_{\mathcal{C}^{s_f}}^2}{4 d_X^{d_X/2} L_f^2} \frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{1}{2s_f + d_X}}. \tag{38}$$

Since $\varepsilon_{X,i}$ has a uniform conditional distribution independent of Z_i and g, the function

$$x \mapsto \mathop{\mathbb{E}}_{D,g \sim P_0} \left[1\{\|x + \sigma_X \varepsilon_{X,i}\|_\infty < h\} \right]$$

is maximized over $x \in \mathbb{R}^{d_X}$ when $x = 0$. Note that, since this bound depends *only* on the distribution of $\varepsilon_{X,i}$, and not on the distribution of Z_i or g, this applies even in the active case.

$$\begin{aligned} \mathop{\mathbb{E}}_{D,g \sim P_0} \left[1\{\|X_i - x_0\|_\infty < h\} \right] &= \mathop{\mathbb{E}}_{D,g \sim P_0} \left[1\{\|g(Z_i) + \sigma_X \varepsilon_{X,i} - x_0\|_\infty < h\} \right] \\ &\leq \mathop{\mathbb{E}}_{D,g \sim P_0} \left[1\{\|\sigma_X \varepsilon_{X,i}\|_\infty < h\} \right] = \min \left\{ 1, \left(\frac{h\sqrt{d_X}}{\sigma_X} \right)^{d_X} \right\}. \end{aligned} \tag{39}$$

where ϕ_{d_X} denotes the d_X-dimensional standard normal density. Plugging this into Lemma 13 gives

$$\sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D \left[|\widehat{f}_{x_0} - f(x_0)| \geq C_{d_X,s_f} L_f h^{s_f} \right] \geq \frac{1}{2} \left(1 - \frac{L_f h^{s_f}}{\sigma_Y \|K_{d_X}\|_{\mathcal{C}^{s_f}}} \sqrt{n \left(\frac{h\sqrt{d_X}}{\sigma_X} \right)^{d_X}} \right).$$

Plugging in our choice (Eq. (38)) of h gives

$$\inf_{\widehat{F}} \sup_{\theta \in \Theta} \Pr_\theta \left[|\widehat{F} - F(\theta)| \geq s \right] \geq \frac{1}{4},$$

for

$$s = \frac{1}{2} \left(\frac{L_f}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}} \right)^{\frac{d_X}{2s_f + d_X}} \left(4\sqrt{d_X^{d_X}} \right)^{\frac{-s_f}{2s_f + d_X}} \left(\frac{\sigma_X^{d_X} \sigma_Y^2}{n} \right)^{\frac{s_f}{2s_f + d_X}}. \tag{40}$$

■

Next, we prove a tight lower bound for the case that the covariate noise level σ_X is large but the response noise level σ_Y is small. This proof is quite similar to that of Proposition 17 for the passive case, except that, (a) rather that considering Z uniformly distributed on $\mathcal{Z}$, we make no assumptions on Z and utilize only randomness coming from ε_X, and (b) we must reason somewhat more carefully about the interdependence between samples.

Proposition 19 (Minimax Lower Bound, Active Case, σ_X large, σ_Y small) *In the active case,*

$$\inf_{\widehat{f}_{x_0}} \sup_{g,f,P_Z,\varepsilon_X,\varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq C \left(\frac{\sigma_X^{d_X}}{n} \right)^{s_f/d_X} \right] \geq \frac{1}{2e} \approx 0.18, \tag{41}$$

where $C > 0$ (given below in Eq. (43)) depends only on L_g, s_g, L_f, s_f, d_X, and d_Z.

Proof Suppose $\sigma_X \varepsilon_{X,i}$ is uniformly distributed on $[-\frac{\sigma_X}{\sqrt{d_X}}, \frac{\sigma_X}{\sqrt{d_X}}]^{d_X} \subseteq \mathbb{R}^{d_X+1}$. It is straightforward to verify that $\varepsilon_{X,i}$ is sub-Gaussian with and that $\varepsilon_{X,i}$ has dimension d_X around 0. Let $r^* := \frac{\sigma_X}{\sqrt{d_X}}$. For any $r \in [0, r^*]$, since each ε_i is independent of the preceding data $\{(X_j, Y_j, Z_j)\}_{j=1}^n$, and since the function $x \mapsto \Pr[\|x + \varepsilon_i\|_\infty \leq r]$ is maximized at $x = 0$,

$$\begin{aligned}
\Pr\left[\min_{i\in[n]} \|X_i\|_\infty \geq r\right] &= \prod_{i=1}^n \Pr\left[\|X_i\|_\infty \geq r \,\middle|\, \min_{j\in[i-1]} \|X_j\|_\infty \geq r\right] \\
&= \prod_{i=1}^n \left(1 - \Pr\left[\|X_i\|_\infty \leq r \,\middle|\, \min_{j\in[i-1]} \|X_j\|_\infty \geq r\right]\right) \\
&= \prod_{i=1}^n \left(1 - \Pr\left[\|g(Z_i) + \varepsilon_i\|_\infty \leq r \,\middle|\, \min_{j\in[i-1]} \|X_j\|_\infty \geq r\right]\right) \\
&\geq \prod_{i=1}^n \left(1 - \Pr\left[\|\varepsilon_i\|_\infty \leq r \,\middle|\, \min_{j\in[i-1]} \|X_j\|_\infty \geq r\right]\right) \\
&= \prod_{i=1}^n \left(1 - \Pr\left[\|\varepsilon_i\|_\infty \leq r\right]\right) \\
&= \left(1 - \left(\frac{r\sqrt{d_X}}{\sigma_X}\right)^{d_X}\right)^n.
\end{aligned}$$

For n sufficiently large that $r \leq r^*$, plugging in

$$r = \frac{\sigma_X}{\sqrt{d_X}} n^{-1/d_X}. \tag{42}$$

gives

$$\Pr\left[\min_{i\in[n]} \|X_i\|_\infty \geq \frac{\sigma_X}{\sqrt{d_X}} n^{-1/d_X}\right] = \left(1 - \frac{1}{n}\right)^n \geq 1/e > 0,$$

since $\left(1 - \frac{1}{n}\right)^n$ approaches $1/e$ from above as $n \to \infty$.

Now let f_0 and f_1 be as in Eq. (19), with $h = r$ above. Given the event $\min_{i\in[n]} \|X_i\|_\infty \geq h$, the distribution of the data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is independent of whether $f = f_0$ or $f = f_1$. Hence, as argued in the proof of Proposition (16),

$$\max_{f\in\{f_0,f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq \frac{|f_1(x_0) - f_0(x)|}{2}\right] \geq \frac{1}{2} \Pr\left[\min_{i\in[n]} \|X_i\|_\infty \leq h\right] \geq \frac{1}{2e}.$$

Since, by construction of f_0 and f_1,

$$|f_1(x_0) - f_0(x_0)| = \frac{L_f h^{s_f}}{\|K_{d_X}\|_{\mathcal{C}^{s_f}}},$$

it follows that

$$\max_{f\in\{f_0,f_1\}} \Pr\left[\left|\widehat{f}_{x_0} - f(x_0)\right| \geq \frac{L_f h^{s_f}}{2\|K_{d_X}\|_{\mathcal{C}^{s_f}}}\right] \geq \frac{1}{2e}.$$

Plugging in our choice of $h = r$ from Eq (42) above gives

$$\max_{f \in \{f_0, f_1\}} \Pr \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq C \left(\frac{\sigma_X^{d_X}}{n} \right)^{s_f/d_X} \right] \geq \frac{1}{2e},$$

where

$$C = \frac{L_f d_X^{-s_f/2}}{2\|K_{d_X}\|_{\mathcal{C}^{s_f}}}. \tag{43}$$

■

Finally, we prove a lower bound for the Active case that is tight when σ_X is small but σ_Y is large:

Proposition 20 (Minimax Lower Bound, Active Case, σ_X small, σ_Y large) *In the active case,*

$$\inf_{\widehat{f}_{x_0}} \sup_{g, f, P_Z, \varepsilon_X, \varepsilon_Y} \Pr_D \left[\left| \widehat{f}_{x_0} - f(x_0) \right| \geq \frac{\sigma_Y}{4\sqrt{n}} \right] \geq \frac{1}{4}. \tag{44}$$

Proof Plugging the trivial bound

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{D, g \sim P_0} \left[1\{\|X_i - x_0\|_\infty < h\} \right] \leq n \quad \text{and} \quad h = \left(\frac{\sigma_Y \|K_{d_X}\|_{\mathcal{C}^{s_f}}}{2L_f\sqrt{n}} \right)^{1/s_f}$$

into Lemma 13 gives

$$\inf_{\widehat{F}} \sup_{\theta \in \Theta} \Pr_\theta \left[|\widehat{F} - F(\theta)| \geq \frac{\sigma_Y}{4\sqrt{n}} \right] \geq \frac{1}{4}.$$

■

Appendix F. Supplementary Lemmas

For the reader's convenience, here we state two standard concentration inequalities for sub-Gaussian random variables, used in our upper bounds.

Lemma 21 (Theorem 1.14 of Rigollet (2015)) *Suppose that $X_1, ..., X_n$ are IID observations of a sub-Gaussian random variable. Then, for any $\delta \in (0, 1)$,*

$$\Pr_{X_1, ..., X_n} \left[\max_{i \in [n]} |X_i| \leq \sqrt{2 \log \left(\frac{2n}{\delta} \right)} \right] \leq \delta.$$

Lemma 22 (Proposition 2.1 of Wainwright (2019)) *Suppose that $X_1, ..., X_n$ are IID observations of a sub-Gaussian random variable. Then, for any $\delta \in (0, 1)$,*

$$\Pr_{X_1, ..., X_n} \left[\frac{1}{n} \sum_{i=1}^{n} X_i \leq \sqrt{\frac{2}{n} \log \frac{1}{\delta}} \right] \leq \delta.$$