Undersmoothing Black-Box Models for Functional Estimation

Yue Yu
University of Michigan

Debarghya Mukherjee
Boston University

Moulinath Banerjee
University of Michigan

Ya'acov Ritov
University of Michigan

Abstract

We study functional estimation using black-box models through a model-agnostic undersmoothing framework. The proposed procedure `Rep` operates by augmenting the original dataset through replicating a proportion of samples multiple times, and subsequently applying the black-box algorithm to the augmented dataset. This construction automatically induces undersmoothing and reduces the functional estimation error. We provide several empirical demonstrations (including neural network based learners) showing that compared to the plug-in estimator, the proposed algorithm `Rep` improves the estimation accuracy of functional estimation *without* requiring explicit expressions for the associated influence functions. Furthermore, we develop a theoretical analysis in two representative settings, the Nadaraya–Watson estimator and the random feature model, establishing that replication provides explicit prescriptions for the replication proportion and number of copies, and yields optimal convergence rates for functional estimation. In the classical nonparametric regression setting, we extend `Rep` with a Lepski-style method that adapts to unknown structural features of the regression function.

1 INTRODUCTION

In this work, we study the functional estimation problem by undersmoothing a generic black-box regression model. Given training data $\mathcal{D}_{\mathsf{train}} = \{\boldsymbol{x}_i, y_i\}_{1\le i\le n} \subseteq \Omega \times \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^d$ is compact and $y = f^\star(\boldsymbol{x}) + \varepsilon$, let $\widehat{f}$ denote the fitted predictor and let $\tau(f^\star)$ be the target functional of the true regression function $f^\star$. Estimation and inference of a smooth functional at a function is naturally tied to semi-parametric inference (Bickel et al., 1993), which naturally arises in machine learning (Kandasamy et al., 2014), information theory (Paninski and Yajima, 2008; Wu and Yang, 2016; Kozachenko, 1987), and in the modern causal inference literature (Cui et al., 2024; Kennedy, 2016; Zhang et al., 2024).

A common and convenient approach for estimating target functional $\tau(f^\star)$ is to estimate $f^\star$ through the black-box model $\widehat{f} \leftarrow \mathtt{Alg}(\mathcal{D}_{\mathsf{train}})$ and then estimate $\tau(f^\star)$ via the plug-in estimator $\tau(\widehat{f})$. However, it is well known that when the estimator $\widehat{f}$ is minimax-optimal with respect to standard function norms (e.g., $\|\cdot\|_{L^2(\mathbb{P})}$ or $\|\cdot\|_\infty$), the plug-in estimator $\tau(\widehat{f})$ for a smooth functional $\tau(f^\star)$ is generally suboptimal and suffers from a slow rate of convergence for a broad class of statistical functionals (Goldstein and Messer, 1992; Newey et al., 1998; Chernozhukov et al., 2017). To mitigate this issue, a common strategy is to apply bias correction based on the influence function (Bickel et al., 1993; Bickel and Ritov, 1988; Chernozhukov et al., 2017; Kennedy, 2023), the main ideas of which are reviewed in Appendix C. This approach, however, requires knowledge of the exact influence function associated with the functional τ, which may be unavailable or difficult to derive in practice.

To address this limitation, we pursue undersmoothing as an alternative. Prior work has demonstrated that undersmoothing yields valid inference and mitigates bias (Fisher and Fisher, 2023; Laan et al., 2021; Paninski and Yajima, 2008; Giné and Nickl, 2008; Newey et al., 1998; Hall, 1992; McGrath and Mukherjee, 2022). Our approach diverges from these methods by *removing the reliance on explicit hyperparameter tuning*, such as bandwidth selection in kernel density estimation (Paninski and Yajima, 2008); and yields a model-agnostic procedure: it modifies the

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

training sample $\mathcal{D}_{\mathsf{train}}$ solely through the construction of an augmented dataset $\mathcal{D}_{\mathsf{aug}}$, formalized in Section 2.

We adopt a deliberately broad notion of a black-box model to preserve generality. This perspective highlights two aspects: (i) the practitioner remains agnostic to the underlying model, and (ii) explicitly regulating the complexity of `Alg` is often infeasible in practice. Such a model-agnostic framework is not only of theoretical interest but also of practical importance. This approach is particularly appropriate when the model is too complex for practitioner to fully understand its underlying complexity. For example, debates concerning model complexity continue to persist even in the context of fully connected neural networks, (Golowich et al., 2018; Sellke, 2024; Bartlett et al., 2017; Dwivedi et al., 2023) and the references therein. Although these problems are significant in their own right, they lie beyond the scope of the present work.

Main contributions: We summarize our main contributions as follows: **(1)** We propose a model-agnostic framework for undersmoothing black-box estimators in functional estimation via Algorithm 1 (`Rep`), which modifies only the input data through replication. Algorithm 1 draws a subset of the training data and replicates each selected point a fixed number of times. This controlled replication increases the weights of selected samples in a balanced way, inducing the desired undersmoothing effect for a black-box model. **(2)** We provide theory explaining why `Rep` improves convergence rates for functional estimation in two concrete settings: the Nadaraya-Watson estimator and random-feature models. The analysis yields practical guidance for tuning the replication proportion ρ_n and the number of copies R_n. **(3)** In the spirit of the nonparametric regression problem, we augment `Rep` with a Lepski-style adaptation to remove the need for oracle structural information $\bar{\beta}$ (e.g., smoothness of the function). **(4)** The theory is validated by extensive synthetic experiments, which corroborate the efficacy of the proposed procedures.

Organization In section 2.1, we revisit undersmoothing as a means for functional estimation and specify the class of functional under discussion. Section 2.2 describes the replication procedure for both known and unknown structural indices $\bar{\beta}$ (*e.g.*, Hölder smoothness of the true function $f^\star$ or the eigenvalue decay rate of $f^\star$ with certain decomposition with respect to certain basis). Section 3 presents two case studies: Nadaraya-Watson estimator (Section 3.1) and random feature models (Section 3.2) to illustrate how replication induces undersmoothing in these two models and improves functional estimation. Section 4 reports simulations, including settings where `Alg` is a deep neural network, that corroborate the theory and suggest new findings and directions for future work.

2 UNDERSMOOTHING VIA REPLICATION PROCEDURE

2.1 Undersmoothing and Plug-in Estimator

We first review the undersmoothing approach when estimating a smooth functional $\tau(f^\star)$. Principled undersmoothing for modern nonparametric models, such as feed-forward neural networks, remains largely unexplored, especially in a model-agnostic setting where the practitioner would like to avoid heavily manual tuning of hyperparameters (for example, kernel bandwidth, learning rate, or network architecture). In the present work, we consider a smooth functional τ.

Assumption 2.1 (SF)**.** Let $\Omega = [0,1]^d$, and let $\beta, \zeta \in \mathbb{N}$ with $0 \le \zeta \le \beta$. Let $\tau : \mathcal{W}_{\beta,2} \to \mathbb{R}$ be a functional, where $\mathcal{W}_{\beta,2} := \mathcal{W}^{\beta,2}(\Omega)$ is the Sobolev space. Suppose that for any $f^* \in \mathcal{W}_{\beta,2}$ and any perturbation $h \in \mathcal{C}^\beta(\Omega)$ with $\|h\|_{\zeta,\infty}$ sufficiently small, τ admits the expansion

$$\tau(f^* + h) = \tau(f^*) + T_{f^*}(h) + \mathcal{O}\big(\|h\|_{\zeta,2}^2\big), \qquad \text{(SF)}$$

where T_{f^*} denotes the Fréchet derivative of τ at f^* and $\|\cdot\|_{\zeta,\infty}$ is the Hölder norm on $\mathcal{C}^\zeta(\Omega)$.

For clarity of exposition, we assume $\beta, \zeta \in \mathbb{N}$. The case $\beta \notin \mathbb{N}$ can be handled by the same arguments within our analysis routine. All relevant definitions are gathered in Appendix B.1 for ease of reference.

Compared with Goldstein and Messer (1992), we adopt a closely related but more natural setup. They require the first order term $T_{f^\star}(h)$ to depend only on h and not on its derivatives, which rules out many standard functionals. For instance, for the integrated squared derivative $\tau_1(f) = \int (f')^2\, dx$, one has $T_{f^\star}(h) = 2\int h' f^\star\, dx$, which involves h'. Assumption (SF) in our framework is a generic condition tied *only* to the smoothness of the functional estimand. We therefore work with the present formulation for its broader applicability.

For any consistent estimator $\widehat{f} \in \mathcal{W}^{\beta,2}$, the Fréchet expansion implies

$$\tau(\widehat{f}) - \tau(f^\star) = T_{f^\star}\big(\widehat{f} - f^\star\big) + \mathcal{O}_P\big(\,\|\widehat{f} - f^\star\|_{\zeta,2}^2\big).$$

The linear term $T_{f^\star}(\widehat{f}-f^\star)$ dominates, while the quadratic remainder $\|\widehat{f}-f^\star\|^2_{\zeta,2}$ is controlled by the squared Sobolev norm of order ζ. Crucially, ζ reflects the structure of τ rather than the smoothness of $f^\star$. For example, for the canonical functionals $\tau_0(f^\star)=\int (f^\star)^2 \mathrm{d}x$ and $\tau_1(f^\star)=\int (f^\star)'^2 \mathrm{d}x$, we have $\zeta_0=0$ and $\zeta_1=1$, even if $f^\star$ is infinitely smooth (i.e., $\beta=\infty$).

Undersmoothing We first sketch the heuristics behind undersmoothing approach through the classical Nadaraya–Watson kernel regression estimator. Suppose $\widehat{f}_h$ is a d-dimensional kernel estimator with bandwidth $h>0$ and that the regression function $f^\star \in \mathcal{W}^{\beta,2}$ is β-times differentiable. To illustrate the main ideas, we focus on the case $\zeta=0$ in this section. Let $\mathsf{Bias}(\widehat{f}_h)=\mathbb{E}(\widehat{f}_h)-f^\star$. Under the usual kernel conditions (see Section 3.1 or Tsybakov (2008, Section 1.5) for precise regularity conditions), we have $\|\widehat{f}_h-f^\star\|^2_{\mathcal{W}^{0,2}}=\mathcal{O}_P\big(h^{2\beta}+(nh^d)^{-1}\big)$ and $T_{f^\star}(\widehat{f}_h-f^\star)=T_{f^\star}\big(\widehat{f}_h-\mathbb{E}_{\boldsymbol{\varepsilon}}(\widehat{f}_h)\big)+T_{f^\star}\big(\mathbb{E}_{\boldsymbol{\varepsilon}}(\widehat{f}_h)-f^\star\big) \lesssim \frac{1}{n}\sum_{i=1}^n \varepsilon_i\, T_{f^\star}\Big(\frac{1}{h}K\Big(\frac{\cdot-X_i}{h}\Big)\Big)+\big|\mathsf{Bias}(\widehat{f}_h)\big|$, where the first term in the last expression is of order $\mathcal{O}_P(n^{-1/2})$ and independent of the choice of bandwidth h while the bias term $\big|\mathsf{Bias}(\widehat{f}_h)\big|$ is of the canonical order $\mathcal{O}_P(h^\beta)$ under mild regularity conditions, which will be stated precisely in Section 3. That is, the bandwidth affects the first-order term of a smooth functional solely through the bias component, *not* through the variance. From the expansion in (SF), the estimation error $\tau(\widehat{f}_h)-\tau(f^\star)=\mathcal{O}_P\big(h^\beta+n^{-1/2}\big)+\mathcal{O}_P\big(h^{2\beta}+(nh^d)^{-1}\big)$. Balancing the leading bias proxy h^β with the leading variance proxy $(nh^d)^{-1}$ yields that the optimal functional estimation bandwidth is less than the optimal function estimation bandwidth: $h_{\tau\text{-opt}} \asymp n^{-\frac{1}{\beta+d}} \le h_{f^\star\text{-opt}} \asymp n^{-\frac{1}{2\beta+d}}$, achieving the full parametric rate nevertheless requires additional, albeit mild, regularity conditions on β, d. We can plug-in $h_{\tau\text{-opt}}$ and obtain the rate of convergence for functional estimator $\tau(\widehat{f}_h)$, namely, $\mathcal{O}_P\big(n^{-(\frac{1}{2}\wedge\frac{\beta}{\beta+d})}\big)$.

2.2 An Overview of the Replication Procedure

Let `Alg` be a black-box model that automatically selects hyperparameters optimally with respect to mean-squared error (MSE). The protypical machinery of black-box model `Alg` will automate the choice of hyperparameters via procedures like various cross-validation techniques and then refit the model using the original dataset $\mathcal{D}_{\mathsf{train}}$.

Assume $f^\star \in \mathcal{F}_{\bar\beta}$, where the index $\bar\beta$ encodes structural information such as smoothness or eigenvalue decay rate. Let $\tau(f^\star)$ be the functional of interest, and let `Alg` denote a black-box learner whose internal tuning (e.g., bandwidth, learning rate) is inaccessible.

The replication scheme (Algorithm 1) modifies only the data input: **(1)** uniformly select[1] a proportion $\rho_n=\rho_n(\bar\beta,\zeta)$ of the training set $\mathcal{D}_{\mathsf{train}}$[2], where ζ is a functional-specific constant determined by the closed form of τ; denote the selected subset by $\mathcal{D}_{\mathsf{sel}}$. **(2)** duplicate $\mathcal{D}_{\mathsf{sel}}$ R_n-1 times and merge it with the remaining training examples to form the augmented dataset $\mathcal{D}_{\mathsf{aug}}$, where $\mathsf{R}_n\in\mathbb{N}^+, \mathsf{R}_n>2$. Precisley, we define the multiset $\texttt{Duplicate}(\mathcal{D}_{\mathsf{sel}},\mathsf{R}_n)=\{\underbrace{\mathcal{D}_{\mathsf{sel}},\cdots,\mathcal{D}_{\mathsf{sel}}}_{\mathsf{R}_n-1 \text{ times}}\}$. **(3)** fit `Alg` on $\mathcal{D}_{\mathsf{aug}}$ and plug the resulting predictor $\widehat{f}$ into τ to estimate $\tau(f^\star)$.

The rule governing the replication proportion ρ_n and the number of copies R_n plays a central role in our procedure and will be made explicit in Section 3. We analyze separately the cases in which the structural index $\bar\beta$ is known and in which it is unknown. We emphasize that, although `Rep` still requires tuning of hyperparameters, it substantially reduces the effective number of hyperparameters within the black-box model to the pair (ρ_n,R_n).

1. **Known $\bar\beta$.** When the structural index $\bar\beta$ is known, we set the proportion of training examples and the number of copies to $\rho_n \asymp n^{-\varrho(\bar\beta,\zeta,d,c)}$ and $\mathsf{R}_n \asymp n^c$, respectively, where the log-proportion function ϱ describes the replication level. Define $\widehat{f}_\rho=\texttt{Alg}\big(\mathcal{D}_{\mathsf{aug}}\big)$ and $\widehat{\tau}=\tau(\widehat{f}_\rho)$. The procedure is stated in Algorithm 1 `Rep`.

2. **Unknown $\bar\beta$.** In practice, the structural index $\bar\beta$ is unknown. We address this by an adaptive replication procedure, Algorithm 2, built on top of the basic replication procedure.

In certain examples, the form of the log-proportion function ϱ is known. For example, Section 3.1 derives the form explicitly under a two-fold cross-validation scheme embedded in the black-box learner. Hence, when the structural index $\bar\beta$ is known, one can tune the sampling proportion $\varrho(\bar\beta,\zeta,d,c)$ accordingly,

[1] "Uniform selection" means no observation is underweighted or overweighted. In practice, one may sample without replacement uniformly at random, or use a symmetric deterministic rule, for example take the first $\lfloor\rho_n n\rfloor$ observations after a random permutation, or choose approximately equispaced indices with spacing $\lfloor 1/\rho_n\rfloor$.

[2] The proportion ρ_n depends on n and, in general, satisfies $n^{-1}\lesssim\rho_n\lesssim 1$.

Algorithm 1: Replication procedure `Rep` (with a black-box model)

Input:
`Option A:` $(\zeta, \bar{\beta}, c)$: ζ: highest derivative order in (SF); $\bar{\beta}$: structural index of $f^\star$; c: the growth of number of copies.
`Option B:` (ρ_n, R_n)

$\rho_n \leftarrow n^{-\varrho(\bar{\beta},\zeta,d,c)}$
$\mathsf{R}_n \leftarrow n^c$ `// Only for option A`
Select uniformly a subset $\mathcal{D}_{\mathsf{sel}} \subseteq \mathcal{D}_{\mathsf{train}}$ with $|\mathcal{D}_{\mathsf{sel}}| = \lfloor \rho_n n \rfloor$
$\mathcal{D}_{\mathsf{aug}} \leftarrow \mathtt{Duplicate}(\mathcal{D}_{\mathsf{sel}}, \mathsf{R}_n) \cup \mathcal{D}_{\mathsf{train}}$
$\widehat{f} \leftarrow \mathtt{Alg}(\mathcal{D}_{\mathsf{aug}})$
Output: $\tau(\widehat{f})$

yielding a function estimator $\widehat{f}_\rho$ and a plug-in estimator $\tau(\widehat{f}_\rho)$ that attains an accelerated rate of convergence. Because the replication is fully determined by the choice of ρ_n and R_n, we write $\mathtt{Rep}(\rho_n, \mathsf{R}_n)$ when the two hyperparameters are specified in Option B. Inspired by the Lepski method (see Lepski and Spokoiny (1997)), we propose a data-driven adaptive estimator to address the situation when the structural information $\bar{\beta}$ is unknown. The number of grid points N is set to be greater than $\log n$ and the threshold function $\mathbf{r}(n, \rho_j)$ depends on the black-box model and on the target functional τ, which essential measures the bias and variance tradeoff. See Appendix D for detailed explanations.

Algorithm 2: Adaptive replication procedure

Input: $\bar{\beta}_{\max}, \bar{\beta}_{\min}$: Upper and lower bound on the structural index $\bar{\beta}$, and ζ, c: same as Algorithm 1
Discretize $[\bar{\beta}_{\min}, \bar{\beta}_{\max}]$ with $\mathcal{B}_N = \{\bar{\beta}_{\min} = \bar{\beta}_1 < \cdots < \bar{\beta}_{N-1} < \bar{\beta}_N = \bar{\beta}_{\max}\}$.
for $i = 1$ *to* N **do**
  $\rho_i \leftarrow n^{-\varrho(\bar{\beta}_i,\zeta,d,c)}$
  $\tau(\widehat{f}_{\rho^{(i)}}) \leftarrow \mathtt{Rep}(\zeta, \bar{\beta}_i, c)$
Set

$$\tilde{\rho} \leftarrow \max_{1 \le j \le N} \left\{ \forall i < j, \left| \tau(\widehat{f}_{\rho_i}) - \tau(\widehat{f}_{\rho_j}) \right| \le \mathbf{r}(n, \rho_j) \right\}$$

Output: $\tau(\widehat{f}_{\tilde{\rho}})$

Our approach differs from the conventional Lepski's method in two respects. **(1)** Operating externally to black-box models, replication gets rid of structural hyperparameter grids and tunes only the replication proportion ρ. **(2)** Because the estimand is $\tau(f^\star)$, the threshold $\mathbf{r}(n, \rho_j)$ exhibits the elbow phenomenon of functional estimation (Bickel and Ritov, 1988; Tsybakov, 2008; Kennedy et al., 2023), rather than the rate merely depending on the true function $f^\star$.

3 MAIN RESULTS

3.1 Kernel Nonparametric Regression

We begin our analysis by considering a black-box algorithm `Alg`, which uses a Nadaraya-Watson estimator with two-fold cross-validation, which dates back to Stone (1974); Geisser (1975) (See a modern survey (Arlot and Celisse, 2010)). Recall that the Nadaraya–Watson estimator constructed from a generic dataset $\mathcal{D}$ is

$$\widehat{f}_h(\boldsymbol{x}; \mathcal{D}) = \sum_{(\boldsymbol{x}_j, y_j) \in \mathcal{D}} \mathsf{W}_h(\boldsymbol{x} - \boldsymbol{x}_j; \mathcal{D}) y_j, \qquad \text{(NW)}$$

with weights $\mathsf{W}_h(\boldsymbol{x} - \boldsymbol{x}_j; \mathcal{D}) = \frac{K((\boldsymbol{x}-\boldsymbol{x}_j)/h)}{\sum_{k \in \mathcal{D}} K((\boldsymbol{x}-\boldsymbol{x}_k)/h)} = \frac{K_h(\boldsymbol{x}-\boldsymbol{x}_j)}{\sum_{k \in \mathcal{D}} K_h(\boldsymbol{x}-\boldsymbol{x}_k)}$, where $K_h(\boldsymbol{u}) = K(\boldsymbol{u}/h)$. This example acts as a running example to illustrate our proposed approach to functional estimation via replication (Algorithm 1) and serves as an initial step toward a comprehensive theoretical understanding of how replication undersmooths (nonparametric) black-box models, thereby enhancing the performance of the corresponding downstream plug-in functional estimators.

The rationale is to deliberately undersmooth the function estimator. This intentional undersmoothing enables the plug-in estimator to automatically achieve the optimal convergence rate for estimating $\varphi(f^\star)$ unlike the bias-correction method[3], which requires a functional-specific influence function. As shown in Figure 1, when the estimation curve is undersmoothed (i.e., the bandwidth is substantially smaller than the optimal bandwidth for function estimation), the plug-in estimator attains a lower estimation error and an improved rate of convergence.

As mentioned earlier, black-box models typically include automatic hyperparameter tuning. Let $\mathcal{D}_{\mathsf{aug}}$ be split into two disjoint subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ of equal size m so that $|\mathcal{D}_{\mathsf{aug}}| = 2m = \rho_n n \mathsf{R}_n + (1 - \rho_n)n$. For a bandwidth $h > 0$, define the empirical two-fold

[3]We revisit the idea of bias-correction in Appendix C.

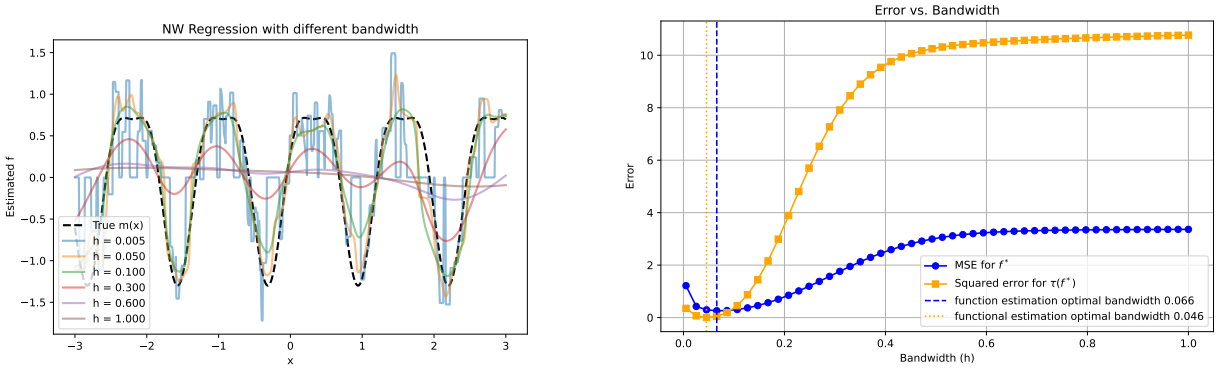


Figure 1: **(Left)**: NW kernel regression estimates for a periodic function at varying bandwidths $5 \times 10^{-3} \leq h \leq 1$. The black dashed curve shows the true function $f^\star(x)$. **(Right)**: Mean squared error curves for estimating $f^\star$ (blue) and its integrated squared functional $\tau(f^\star) = \int_{x\in[-3,3]} f^{\star 2}(x)\,\mathrm{d}x$ as functions of bandwidth, with dashed lines marking the optimal bandwidth for each.

cross-validation risk

$$\mathcal{L}_{\mathsf{CV}}(h, \rho_n, \mathsf{R}_n) = \frac{1}{2m}\Big[\sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_1} \big(y_i - \widehat{f}_h(\boldsymbol{x}_i; \mathcal{D}_2)\big)^2 + \sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_2} \big(y_i - \widehat{f}_h(\boldsymbol{x}_i; \mathcal{D}_1)\big)^2\Big],$$

and the selected bandwidth

$$\widehat{h}_n = \arg\min_{h\in(0,1]} \mathcal{L}_{\mathsf{CV}}(h, \rho_n, \mathsf{R}_n), \tag{3.1}$$

where we suppress the dependence of $\widehat{h}_n$ on (ρ_n, R_n) for brevity.

To facilitate the analysis, we impose the following standard assumptions from the nonparametric literature.

Assumption 3.1 (subG)**.** The noise ε_i is i.i.d. uniformly mean-zero subgaussian wih K_ε and σ^2 denoting its ψ_2 norm and variance respectively.

Assumption 3.2 (kernel)**.** We impose three standard conditions on the kernel $K : \mathbb{R}^d \to \mathbb{R}$:(i) $\int_{\mathcal{X}} K(x)\,dx = 1$. (ii) $\mathsf{supp}(K) = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. (iii) There exists a constant L such that $\forall x, y \in \mathbb{R}^d, |K(x) - K(y)| \leq L\|x - y\|$.

Assumption 3.3 (density)**.** Let p_X denote the density of X with respect to Lebesgue measure and assume it is bounded: $0 < p_{\min} \leq \inf_{x\in\Omega} p_X \leq \sup_{x\in\Omega} p_X(x) \leq p_{\max} < \infty$.

Assumption (subG) is standard in nonparametric analysis and restricts the tail behavior of the error distribution. The kernel conditions in Assumption (kernel) mainly simplify the proofs and are satisfied, for example, by the Epanechnikov–type kernel $K(s) = c_{\alpha,d}(1 - \|s\|_2^2)_+^\alpha$ with $\alpha \geq 1$ and $c_{\alpha,d} = \big(\int_{\mathcal{X}}(1-\|s\|_2^2)_+^\alpha\, ds\big)^{-1}$ when $\Omega = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq 1\}$. Then we present our main theorem.

Theorem 1. *With $\mathcal{L}_{\mathsf{CV}}$ defined as in equation (3.1), under Assumptions (subG), (kernel), and (density), the following statements hold:*

$$\Big|\mathcal{L}_{\mathsf{CV}}(h, \rho_n, \mathsf{R}_n) - \Big[\mathcal{L}^\dagger(h, \rho_n, \mathsf{R}_n) + \sigma^2\Big]\Big| \leq \Delta_{n,\rho,R}, \tag{3.2}$$

where $\mathcal{L}^\dagger(h, \rho_n, \mathsf{R}_n) = \rho n \mathsf{R}\Big(\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R}+nh^d(1-\rho+\rho R))^2}\Big) + (1-\rho)n\Big[\sigma^2 + h^{2\beta} + \frac{1}{\rho n h^d}\Big]$, with probability $1 - o(1)$ and $\Delta_{n,\rho,R} \ll \mathcal{L}^\dagger(h, \rho_n, \mathsf{R}_n)$. Therefore, the optimal bandwidth to minimize $\mathcal{L}_{\mathsf{CV}}$ is $\frac{1}{C}h^\dagger \leq (\widehat{h})^d \leq Ch^\dagger$, for some constant $C > 0$, where $h^\dagger = n^{-1}\mathsf{R}_n^{-1/2}\rho_n^{-3/2}$. If $-3\varrho + c = -2\frac{\beta+\zeta}{\beta+\zeta+d}$, then, with the same probability as above, $\frac{1}{C}n^{-\frac{1}{\beta+\zeta+d}} \leq \widehat{h} \leq Cn^{-\frac{1}{\beta+\zeta+d}}$, for a constant $C > 0$.

We remark a few heuristic observations on the choice of the proportion of replication $|\mathcal{D}_{\mathsf{sel}}|/|\mathcal{D}_{\mathsf{train}}| = \rho_n$. The empirical cross-validation risk (3.1) exhibits a clear tradeoff in ρ_n when $\mathsf{R}_n \in \mathbb{N}^+$ is fixed. When $\rho_n \approx 1$, the selected bandwidth h^* is driven toward 0, so that on average each kernel window contains only one observation, because many samples appear in both the training and validation splits. In this regime the NW regression nearly interpolates most observations in $\mathcal{D}_{\mathsf{train}}$. Conversely, when ρ_n is very small, for example when only a single observation is replicated,

the overlap is negligible and no undersmoothing occurs. Therefore, for a fixed $\mathsf{R}_n = n^c$ (equivalently, c), there is an optimal order of the replication proportion ρ_n in terms of functional. Therefore, one can set $\rho_n = n^{-\varrho}$ appropriately to control the selected bandwidth $\widehat{h}_n$ is of the desired order. The optimal choice of ϱ depends on $(\bar{\beta}, \zeta, d, c)$, which in turn motivates the selection of ρ_n in Algorithm 1.

Theorem 2. *Under Assumptions* (subG), (kernel), *and* (density), *if the smoothness index β is known and greater than $3\zeta + d$, the plug-in estimator $\tau(\widehat{f}_h)$ attains the $\sqrt{n}$ convergenece rate and semiparametric efficiency for estimating the functional $\tau(f^\star)$:*

$$\sqrt{n}\Big(\tau(\widehat{f}_h) - \tau(f^\star)\Big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2 \mathbb{E}_{\boldsymbol{X}\sim P_X}(\boldsymbol{t}^2(\boldsymbol{X}))\big), \tag{3.3}$$

where $\boldsymbol{t} \in L^2(P_X)$ refers to the (unique) Riesz representer of $T_{f^\star}$ in $L^2(P_X)$. Further, if the smoothness index β is unknown, if $\widehat{f}$ is the output of Alg 2, *we obtain that the plug-in estimator satisfies*

$$\sup_{\beta\in[\beta_{\min},\beta_{\max}]} \sup_{f^\star\in\mathcal{W}^{\beta,2}} \mathbb{E}\big(\tau(\widehat{f}_h) - \tau(f^\star)\big)^2 \tag{3.4}$$
$$\lesssim n^{-1} \vee \Big(n^{-\frac{2(\beta-\zeta)}{\beta+\zeta+d}}\Big) \operatorname{poly}(\log n).$$

We now provide interpretations of Theorem 2: adaptive replication procedure chooses ρ_n in a data-driven manner. If $\beta > 3\zeta + d$, the adaptive functional estimator attains parametric rate up to a logarithmic factor.

3.2 Random Features Model

In this subsection, we analyze the replication procedure for undersmoothing in the setting where the black-box estimator is a random features model (RFM), *i.e.*, when the nuisance regression function $f^\star$ is estimated via RFM. The random feature models (RFM) (Rahimi and Recht, 2007) are defined by model class $\mathcal{F}_{\mathsf{RF}} = \{\widehat{f}(x;a) = \frac{1}{\sqrt{p}}\sum_{j=1}^p a_j\varphi(x,\omega_j) : \boldsymbol{a} = (a_j)_{j=1}^p \in \mathbb{R}^p\}$, and approximate kernel methods by replacing the kernel with a random d-dimensional sub-Gaussian feature. Whereas standard kernel estimators scale poorly due to the inversion of an $n \times n$ Gram matrix, RFM approximate the kernel matrix $K(x,y) \approx \sum_{j=1}^p \varphi_j(x)\varphi_j(y) \approx \sum_{j=1}^p \varphi(x,\omega_j)\varphi(y,\omega_j)$, with a moderate number of features $p \ll n$, where φ denotes kernel function and $\omega_1, \ldots, \omega_p \overset{\text{i.i.d.}}{\sim} \pi$. This reduces nonlinear learning to a linear problem in the transformed space and yields a tractable linear problem with p predictors, providing substantial computational savings while retaining accuracy comparable to the full kernel estimator.

Beyond its computational advantages, RFM provides a convenient framework for analyzing optimization dynamics and generalization properties of neural networks in high-dimensional regimes (Mei and Montanari, 2020; Ghorbani et al., 2021; Celentano et al., 2021; Mei et al., 2021; Hu and Lu, 2022; Defilippis et al., 2024). RFM can be interpreted as one-hidden-layer neural networks with frozen first-layer weights, thereby isolating the linear readout while preserving the nonlinear representational capacity of the architecture. Moreover, RFM naturally connect to the Neural Tangent Kernel (NTK) perspective (Jacot et al., 2018): in the lazy-training regime, neural networks behave like kernel methods under the NTK, which in turn can be efficiently approximated by RFMs. Thus, it provides a computationally scalable surrogate that retains the essential theoretical structure of kernel and NTK-based learning.

For the RFM model class $\mathcal{F}_{\mathsf{RF}}$, the parameter $\hat{\boldsymbol{a}}$ is given by the minimizer of the ℓ_2-regularized loss $\widehat{\boldsymbol{a}}_\lambda(\boldsymbol{Z}, y) = \arg\min_{\boldsymbol{a}\in\mathbb{R}^p}\{\sum_{(x,y)\in\mathcal{D}_{\mathsf{train}}}(y_i - \widehat{f}(x_i,\boldsymbol{a}))^2 + \lambda\|\boldsymbol{a}\|_2^2\} = (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}$, where the normalized feature matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ with $\boldsymbol{Z}_{ij} = p^{-1/2}\varphi(x_i,\omega_j)$. Therefore, the estimator of $f^\star$ in RFM takes a closed form of

$$\widehat{f}_\lambda(x) = p^{-1/2}\boldsymbol{\varphi}(x)(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}, \quad \text{(RFM)}$$

where $\boldsymbol{\varphi}(x) = [\varphi(x,\omega_1), \cdots, \varphi(x,\omega_p)]^\mathsf{T} \in \mathbb{R}^p$.

We consider a square-integrable kernel $\varphi \in L_2(\mathcal{X}\times\mathcal{W})$, and define the Fredholm integral operator $\mathbb{T} : L_2(\mathcal{X}) \to \mathcal{V} \subseteq L_2(\mathcal{W})$ by $\mathbb{T}h(w) := \int_\mathcal{X} \varphi(x;w)h(x)\mathrm{d}P_X, \forall h \in L_2(\mathcal{X})$, where we set $\mathcal{V} = \mathrm{Im}(\mathbb{T})$. This operator is compact, and thus admits a spectral decomposition: $\mathbb{T} = \sum_{k=1}^\infty \xi_k\psi_k\varphi_k^*$, where $(\xi_k)_{k\ge1} \subset \mathbb{R}$ are the eigenvalues, and $(\psi_k)_{k\ge1}$, $(\varphi_k)_{k\ge1}$ are orthonormal bases of $L_2(\mathcal{X})$ and $\mathcal{V}$, respectively: $\langle\psi_k,\psi_{k'}\rangle_{L_2(\mathcal{X})} = \delta_{kk'}, \langle\varphi_k,\varphi_{k'}\rangle_{L_2(\mathcal{W})} = \delta_{kk'}$. Without loss of generality, we arrange the eigenvalues in non-increasing order of their absolute values, i.e., $|\xi_1| \ge |\xi_2| \ge \cdots$. For simplicity of exposition, we assume that all eigenvalues are strictly positive, so that $\ker(\mathbb{T}) = \{0\}$. Define $\boldsymbol{\Sigma} = \mathrm{diag}(\xi_1^2, \xi_2^2, \ldots) \in \mathbb{R}^{\infty\times\infty}$ as the diagonal matrix of squared eigenvalues. Moreover, since $f^\star \in L_2(P_X)$, it admits an expansion in the basis $(\psi_k)_{k\ge1}$: $f^\star = \sum_{k\ge1}\beta_k\psi_k$.

We next state the source and capacity (also known as power-law decay) condition in the kernel literature (Rahimi and Recht, 2007, 2008; Rudi and Rosasco, 2021).

Assumption 3.4 (SC). Let $f^\star$ satisfy

$$\mathsf{Tr}(\mathbf{\Sigma}^{1/\alpha}) < \infty, \qquad \|\mathbf{\Sigma}^{-r}\boldsymbol{\beta}^*\|_2 < \infty, \quad \text{(SC)}$$

where $\alpha > 1$ and $r \geq 1/2$.

Specifically, when $r = 1/2$, the true data generation function $f^\star$ belongs to the reproducing kernel Hilbert space of $\mathbb{E}_\pi(\varphi(x,\omega)\varphi(x,\omega))$. Note that the optimal minimax rate for function estimation is $\mathcal{O}(n^{-\frac{2\alpha r}{2\alpha r+1}})$ under source and capacity conditions (Caponnetto and De Vito, 2007; Rahimi and Recht, 2007; Defilippis et al., 2024).

Theorem 3. *Let* $\Delta_{n,\rho,R} = \sqrt{\frac{\log n}{\rho n R + (1-\rho)n}}$ *and* $d_\lambda = \mathsf{Tr}(\mathbf{\Sigma}(\mathbf{\Sigma}+\lambda\mathbf{I})^{-1})$. *With* $\mathcal{L}_{\mathsf{CV}}$ *defined as in equation* (3.1) *and* $n \gg p > p^* \asymp n^{\frac{\alpha-1+2r}{2\alpha r+1}}$ *features, under Assumption 3.4, we have:*

$$\Big|\mathcal{L}_{\mathsf{CV}}(\lambda,\rho_n,R_n) - \Big[C\big((\lambda/m)^{2r} + \frac{d_\lambda}{m}\big) + C'\big(\frac{m}{m+Rd_\lambda}\big)^2 + \sigma^2\Big]\Big| \lesssim \Delta_{n,\rho,R}(\lambda) \cdot \Big(\sigma^2 + (\lambda/m)^{2r} + d_\lambda/m\Big), \tag{3.5}$$

with probability at least $1 - \widetilde{C}(p\mathsf{exp}(-cm/R) + n^{-\delta})$, *where* $\widetilde{C}, \delta > 0$.

Let $\lambda_{f^\star-\mathsf{opt}}$ and $\lambda_{\tau-\mathsf{opt}}$ be the optimal orders of regularization coefficient for the function and functional estimation respectively. Similar to Theorem 1, the selected $\widehat{\lambda} = \arg\min_\lambda \mathcal{L}_{\mathsf{CV}}(\lambda,\rho_n,R_n)$ satisfies $|\widehat{\lambda} - \lambda_{\tau-\mathsf{opt}}| \ll |\lambda_{f^\star-\mathsf{opt}} - \lambda_{\tau-\mathsf{opt}}|$. The details are available in Appendix G.

4 NUMERICAL EXPERIMENTS

We conduct simulation studies to evaluate whether our plug-in estimator, as part of the proposed framework, indeed achieves improved convergence and efficiency in functional estimation. More details are stated in Appendix H.

4.1 Validation of Theoretical Guarantees

To gauge practical utility, in Table 1, we compare estimation errors for several canonical functionals and observe systematic improvement in estimation from the replication procedure.

Alg	n	τ_1	τ_2
NW	100	$\mathbf{2.36 \times 10^{-2}}$	$\mathbf{2.46 \times 10^{-2}}$
		7.91×10^{-2}	4.66×10^{-2}
	1000	$\mathbf{5.43 \times 10^{-3}}$	$\mathbf{6.26 \times 10^{-3}}$
		4.63×10^{-2}	2.73×10^{-2}
	10000	$\mathbf{1.70 \times 10^{-3}}$	$\mathbf{1.88 \times 10^{-3}}$
		1.43×10^{-2}	1.80×10^{-2}
RFM	100	$\mathbf{2.20 \times 10^{-2}}$	$\mathbf{1.22 \times 10^{-2}}$
		6.37×10^{-2}	5.63×10^{-2}
	1000	$\mathbf{5.86 \times 10^{-3}}$	$\mathbf{4.54 \times 10^{-3}}$
		2.53×10^{-2}	1.79×10^{-2}
	10000	$\mathbf{1.45 \times 10^{-3}}$	$\mathbf{1.33 \times 10^{-3}}$
		8.07×10^{-3}	7.87×10^{-3}
NN	100	$\mathbf{1.65 \times 10^{-2}}$	$\mathbf{1.11 \times 10^{-2}}$
		4.91×10^{-2}	4.79×10^{-2}
	1000	$\mathbf{4.63 \times 10^{-3}}$	$\mathbf{4.58 \times 10^{-3}}$
		1.33×10^{-2}	1.22×10^{-2}
	10000	$\mathbf{1.60 \times 10^{-3}}$	$\mathbf{1.39 \times 10^{-3}}$
		8.24×10^{-3}	7.52×10^{-3}
EL	100	$\mathbf{1.62 \times 10^{-2}}$	$\mathbf{1.35 \times 10^{-2}}$
	1000	$\mathbf{5.13 \times 10^{-3}}$	$\mathbf{4.29 \times 10^{-3}}$
	10000	$\mathbf{1.62 \times 10^{-3}}$	$\mathbf{1.36 \times 10^{-3}}$

Table 1: Estimation error (*i.e.*, $|\widehat{\tau} - \tau|$) for the integrated squared functionals $\tau_1(f^\star) = \int_{\mathcal{X}}(f^\star)^2 \mathrm{d}x$ and $\tau_2(f^\star) = \int_0^1 f^\star \log(f^\star)\mathrm{d}x$ under $\sigma = 0.1$. Note that the estimation error of the undersmoothed estimator from our `Rep` procedure is systematically better than that of the direct plug-in estimator.

Boldface numbers correspond to estimation errors from our proposed `Rep` method, whereas standard-font numbers represent direct plug-in estimators. The **last three lines** correspond to the efficiency lower bound for functionals with different choices of data generation functions $f^\star$. We set $f_1^* = \sin(20x) + x^2$. By straight calculation and the condition that $P_X = \mathsf{Unif}[0,1]$, the (semiparametric) efficiency lower bounds for the two functional τ_1, τ_2 are given as $\mathsf{Eff}(\tau_1) = \frac{4\sigma^2}{n}\int_0^1 (f^\star)^2 \mathrm{d}x$ and $\mathsf{Eff}(\tau_2) = \frac{\sigma^2}{n}\int_0^1 (1 + \log f^\star)^2 \mathrm{d}x$.

4.2 `Rep` in Black-Box model

In our subsequent experiments, we apply `Rep` to a range of deep learning models. There are subtle yet important distinctions between training deep neural networks and working with models that admit closed-form solutions. First, even the simplest multilayer perceptron requires an explicit optimization procedure and a substantially more intricate hyperparameter tuning process. This tuning involves not only architectural decisions, such as the num-

ber of layers and choice of activation functions, but also optimization parameters, including the choice of optimizer, step size, and initialization strategy. In contrast, estimators such as NW and RFM do *not* require similar optimization steps in their implementation. Second, while classical nonparametric regression methods derive predictions directly from the training data, deep neural networks encode the data into a set of learned parameters, such that the training set no longer appears explicitly in the final predictor. Therefore, we employ two-fold cross validation along the training dynamics. Specifically, given fixed ρ_n and R_n, for a maximum training epoch T, we set our estimator as $\widehat{f} = f_{\widehat{\boldsymbol{\theta}}}$, where $\widehat{\boldsymbol{\theta}} = \arg\min_{1\le t\le T} \min\{\mathcal{L}^{(1)}_{\mathsf{CV}}(f_{\boldsymbol{\theta}^1_t}), \mathcal{L}^{(2)}_{\mathsf{CV}}(f_{\boldsymbol{\theta}^2_t})\}$, and $\{\boldsymbol{\theta}^k_t\}_{\substack{1\le t\le T\\ k\in\{1,2\}}}$ denotes the sequence of weights generated by the optimization dynamics (*e.g.*, stochastic gradient descent).

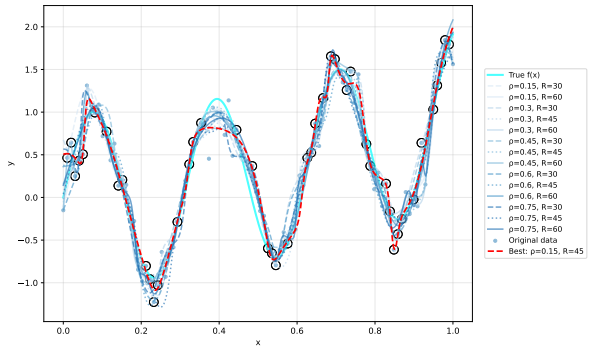
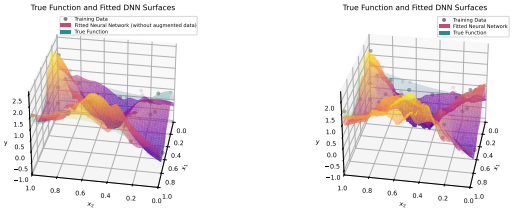


Figure 2: Curve plots for DNN across different (ρ_n, R_n) combinations in `Rep`.

Choosing ρ and R: We now address the choice of ρ and R for generic black box models. Since there is no analogue of the theoretical choice ϱ available for the cases in Section 3, we adopt a simple empirical rule: search over grids $\rho_n \in \{0.15, 0.3, \cdots, 0.75\}$ and $\mathsf{R}_n \in \{30, 45, 60\}$, for example in Figure 2. For each pair candidate (ρ_n, R_n), we construct the replicated dataset via $\texttt{Rep}(\rho_n, \mathsf{R}_n)$, fit the undersmoothed black box model, and select $(\rho_\star, \mathsf{R}_\star)$ by minimizing prediction errors on the replicated observations, in line with the observed interpolation phenomenon. The final estimator is the fit obtained with $\texttt{Rep}(\rho_\star, R_\star)$. In Figure 2, we plot curves for DNN with different replication hyperparameters to illustrate the choice of ρ and n across grid, where fitted DNN associated with the optimal choice is the red curve and the black circle dots represent $\mathcal{D}_{\mathsf{sel}}$.

In Figure 3, we compare the fitted surfaces of DNN with and without `Rep` and illustrate the occurrence of



(a) without replication (b) with replication

Figure 3: Comparison of the true function (green surface) and fitted deep neural networks.

the undersmoothing phenomenon in DNN. We simulate a two-variable nonparametric regression problem by generating $n = 100$ samples with inputs uniformly drawn from $[0, 1]^2$ and responses computed as $Y = \sin(6X_1)\cos(6X_2) + X_1^2 + X_2^2 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$. The deep neural network features an architecture comprising an input layer with 2 nodes, three hidden layers each with 1024 neurons activated by the hyperbolic tangent function, and a linear output layer producing a scalar response. We train the network using `Adam` (Kingma and Ba, 2017) with a learning rate of 10^{-3}, a mini-batch size of 32, and over 3000 epochs, using the cross-validation scheme described above. Relative to (a), the fitted DNN in (b) displays greater curvature and reduced smoothing. Additional details on the simulation setup, including experiments in higher dimensional settings and alternative specifications for the true function, are provided in Appendix H.

5 CONCLUDING DISCUSSION

We devote this section to describing some natural directions for future research.

GCV: In this work, we focus on its theoretical justification via two-fold cross validation for the proposed replication method and its adaptive variant. An interesting direction is to investigate the universality and compatibility of the algorithm when alternative forms of cross validation, such as generalized cross validation, are employed in black-box models. Through simulations, we demonstrate the effectiveness of the proposed method when `Rep` is incorporated into the training dynamics of deep neural networks, highlighting its potential universality.

Nonlinear estimator: For the theoretical analysis, we exploit the linearity of $\widehat{f}$ in the response vector Y to decompose $\widehat{f} - \mathbb{E}_{\boldsymbol{\varepsilon}}[\widehat{f}]$. This covers a broad family of linear estimators, from classical linear smoothers

to kernel ridge regression and random features. Empirically, our replication scheme also works well with *nonlinear* estimators, including deep neural networks. A general theory for such black-box learners remains open.

Computational burden: Replication raises the computational load, since the black-box learner `Alg` processes extra copies of the data. A natural extension is to design a leaner variant that *actively* chooses which points to replicate, framing the task as data selection or active learning; see, for example, Hanneke et al. (2025); Dasgupta et al. (2019). Another interesting problem is that, given a pre-trained nonparametric estimator, for example, an empirical risk minimizer $\widehat{f}_{\mathsf{ERM}}$, how to effectively undersmooth the esstimator.

Overparametrization: In Section 3.2, we established guarantees under the source–capacity conditions (Assumption 3.4) when the number of features is moderate ($p \ll n$) in the random feature models. Given the connections between the random features model and DNN, and the empirical evidence from our DNN simulations, we expect that `Rep` likewise induces undersmoothing for the RFM in the overparameterized regime ($p \gtrsim n$).

References

Melissa Adrian, Jake A. Soloff, and Rebecca Willett. Stabilizing black-box model selection with the inflated argmax, January 2025.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none), January 2010. ISSN 1935-7516. doi: 10.1214/09-SS054. arXiv:0907.4728 [math].

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sivaraman Balakrishnan, Edward H. Kennedy, and Larry Wasserman. The Fundamental Limits of Structure-Agnostic Functional Estimation, June 2025. arXiv:2305.04116 [math].

Peter L. Bartlett, Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks, October 2017. arXiv:1703.02930 [cs].

P. J. Bickel and Y. Ritov. Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 50 (3):381–393, 1988. ISSN 0581-572X. Publisher: Springer.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration Inequalities. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 208–240. Springer, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_9.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368, 2007.

Michael Celentano, Theodor Misiakiewicz, and Andrea Montanari. Minimum complexity interpolation in random features models. *arXiv preprint arXiv:2103.15996*, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Causal Parameters, December 2017. arXiv:1608.00060 [econ, stat].

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming, January 2020. arXiv:1812.07956 [math].

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, and Xiaojin Zhu. Teaching a black-box learner. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1547–1555. PMLR, 09–15 Jun 2019.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression, November 2024. arXiv:2405.15699 [stat].

Raaz Dwivedi, Chandan Singh, Bin Yu, and Martin Wainwright. Revisiting minimum description

length complexity in overparameterized models. 2023.

Weinan E, Chao Ma, and Lei Wu. The Barron Space and the Flow-induced Function Spaces for Neural Network Models, March 2021.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021. ISSN 0012-9682. doi: 10.3982/ECTA16901. arXiv:1809.09953 [econ].

Aaron Fisher and Virginia Fisher. Three-way Cross-Fitting and Pseudo-Outcome Regression for Estimation of Conditional Effects and other Linear Functionals, June 2023. arXiv:2306.07230 [stat].

Seymour Geisser. The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975. ISSN 0162-1459. doi: 10.2307/2285815. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Annals of Statistics*, 2021.

Evarist Giné and Richard Nickl. A Simple Adaptive Estimator of the Integrated Square of a Density. *Bernoulli*, 14(1):47–61, 2008. ISSN 1350-7265. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

Larry Goldstein and Karen Messer. Optimal Plug-in Estimators for Nonparametric Functional Estimation. *The Annals of Statistics*, 20(3):1306–1328, 1992. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-Independent Sample Complexity of Neural Networks. In *Proceedings of the 31st Conference On Learning Theory*, pages 297–299. PMLR, July 2018. ISSN: 2640-3498.

Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. A Rainbow in Deep Network Black Boxes, October 2024. arXiv:2305.18512 [cs].

Peter Hall. Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *The Annals of Statistics*, 20 (2):675–694, June 1992. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176348651. Publisher: Institute of Mathematical Statistics.

Steve Hanneke, Shay Moran, Alexander Shlimovich, and Amir Yehudayoff. Data Selection for ERMs, April 2025. arXiv:2504.14572 [cs].

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jikai Jin and Vasilis Syrgkanis. Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation, May 2025. arXiv:2402.14264 [stat].

Jikai Jin, Lester Mackey, and Vasilis Syrgkanis. It's Hard to Be Normal: The Impact of Noise on Structure-agnostic Estimation, July 2025. arXiv:2507.02275 [stat].

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James M Robins. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations. *arXiv preprint arXiv:1411.4342*, 2014.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review, January 2023. arXiv:2203.06469 [stat].

Edward H. Kennedy, Sivaraman Balakrishnan, James M. Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation, December 2023. arXiv:2203.00837 [math].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

Leonenko Kozachenko. Sample estimate of the entropy of a random vector. *Probl. Pered. Inform.*, 23:9, 1987.

Mark J. van der Laan, David Benkeser, and Weixin Cai. Efficient Estimation of Pathwise Differentiable Target Parameters with the Undersmoothed Highly Adaptive Lasso, July 2021. arXiv:1908.05607 [math].

Erin LeDell and Sebastien Poirier. H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, December 1997. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1030741083. Publisher: Institute of Mathematical Statistics.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan A. K. Suykens. Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, October 2022. ISSN 1939-3539.

Sean McGrath and Rajarshi Mukherjee. Nuisance function tuning and sample splitting for optimal doubly robust estimation. *arXiv preprint arXiv:2212.14857*, 2022.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve, December 2020. arXiv:1908.05355 [math].

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pages 3351–3418. PMLR, 2021.

Whitney K. Newey, Fushing Hsieh, and James Robins. Undersmoothing and bias corrected functional estimation. Working Paper, Cambridge, Mass. : Massachusetts Institute of Technology, 1998. Accepted: 2011-06-10T05:17:48Z.

Liam Paninski and Masanao Yajima. Undersmoothed Kernel Entropy Estimators. *IEEE Transactions on Information Theory*, 54(9):4384–4388, September 2008. ISSN 1557-9654. doi: 10.1109/TIT.2008.928251. Conference Name: IEEE Transactions on Information Theory.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Alessandro Rudi and Lorenzo Rosasco. Generalization Properties of Learning with Random Features, April 2021. arXiv:1602.04474 [stat].

Shubhra Kanti Karmaker Santu, Md Mahadi Hassan, Micah J. Smith, Lei Xu, ChengXiang Zhai, and Kalyan Veeramachaneni. AutoML to Date and Beyond: Challenges and Opportunities, May 2021. arXiv:2010.10777 [cs].

Mark Sellke. On Size-Independent Sample Complexity of ReLU Networks, February 2024. arXiv:2306.01992 [cs].

Jake A Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging Provides Assumption-free Stability.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2): 111–147, 1974. ISSN 0035-9246. Publisher: [Royal Statistical Society, Oxford University Press].

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, October 2008. ISBN 978-0-387-79052-7. Google-Books-ID: mwB8rUBsbqoC.

Martin J. Wainwright. Wild refitting for black box prediction, July 2025. arXiv:2506.21460 [stat].

Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. FLAML: A Fast and Lightweight AutoML Library, May 2021. arXiv:1911.04706 [cs].

E. Weinan, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, 63(7):1235–1258, July 2020. ISSN 1674-7283. doi: 10.1007/s11425-019-1628-5. Publisher: Science China Press.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Archer Gong Zhang, Nancy Reid, and Qiang Sun. A semiparametric approach to causal inference. *arXiv preprint arXiv:2411.00950*, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We define the statistical estimand and undersmoothing mechanism in Section 1 and 2. We list Assumptions 2.1, 3.1, 3.2, and 3.3 for Theorem 1 and Theorem 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We derive the rate of convergence for functional estimation for both cases in Section 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See question 1 (a) in checklist.
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes] For all the assumptions used, we provide justifications after the assumptions are stated.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See the provided URL and additional experiment setup in the supplemental material.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not applicable]
   (b) The license information of the assets, if applicable. [Not applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not applicable]
   (d) Information about consent from data providers/curators. [Not applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not applicable]

Supplementary Materials to "Undersmoothing Black-Box Models for Functional Estimation"

The organization of the supplement is as follows. Section A provides an extended review of related work on undersmoothing techniques, inference for black-box models, and random feature representations. Section B establishes several preliminary results concerning the replication procedure. Section C reviews alternative approaches to efficient functional estimation based on influence functions. Section D specifies the theoretical recipe for hyperparameter tuning in the Nadaraya–Watson estimator. Sections E and G contain the detailed proofs of Theorems 1 and 3, respectively. Finally, Section H presents additional implementation details and supplementary simulation results.

A Related work

Undersmoothing. As noted earlier, undersmoothing is a standard device for functional estimation (Fisher and Fisher, 2023; Laan et al., 2021; Newey et al., 1998; Hall, 1992), with applications to the integrated squared density (Giné and Nickl, 2008), differential entropy (Paninski and Yajima, 2008), and various causal estimands. Our contribution departs from prior work in two respects. First, the proposed `Rep` (Algorithm 1) operates *externally* to the black-box models. Second, the framework accommodates a wide class of function estimators and, under mild regularity conditions, enables efficient inference for a broad family of smooth functionals τ.

Inference for black-box models. The widespread adoption of black-box learners, driven by strong empirical performance (Santu et al., 2021; LeDell and Poirier, 2020; Dasgupta et al., 2019; Wang et al., 2021; Guth et al., 2024), has sharpened the need for rigorous statistical inference for such procedures (Adrian et al., 2025; Soloff et al.). Recent contributions include Wainwright (2025), who develop a method for risk estimation of general black-box predictors representable as penalized M-estimators, and related work that investigates fundamental limits in a model-agnostic minimax framework (Jin and Syrgkanis, 2025; Chernozhukov et al., 2017; Balakrishnan et al., 2025; Jin et al., 2025). These papers posit a convergence rate for a nonparametric estimator $\widehat{f}$, typically expressed through the mean-squared risk $\|\widehat{f} - f\|^2_{L^2(P_X)}$, and then employ debiasing or double machine learning techniques (Bickel et al., 1993; Kennedy, 2023). We depart from this paradigm. By deliberately undersmoothing, we impose no explicit rate requirement on $\widehat{f}$ itself and show that, provided the black-box learner can be refitted repeatedly and tuned via a data-adaptive rule such as cross-validation, it is still possible to attain the parametric $n^{-1/2}$ rate for estimating the target functional $\tau(f)$.

Random feature model (RFM). RFMs serve as finite-dimensional approximations of kernel methods or as models for two-layer neural networks trained in a lazy regime (Chizat et al., 2020). Originally proposed for computational efficiency, they have seen widespread use and gained prominence as proxies for overparametrized neural networks, see (Weinan et al., 2020; E et al., 2021; Arora et al., 2019), and survey (Liu et al., 2022). Indeed, for particular choices of feature maps such as $\sigma(x^\mathsf{T}\omega)$ for some activation function σ, it can be seen as a two-layer neural network with fixed first-layer weights. A line of theoretical work investigates the generalization error of RFM (Rudi and Rosasco, 2021; Rahimi and Recht, 2007; Defilippis et al., 2024), which aligns with our analysis of `Rep` in the RFM setting.

B Auxillary Results and Definitions

Throughout the analysis we use the following shorthands:

(A) $\mathcal{M} = \mathcal{D}_1 \cap \mathcal{D}_2$ denotes the overlapped set. By construction, we have that $\mathcal{M} \subseteq \mathcal{D}_{\mathsf{sel}}$, and under Bernoulli sample splitting, the event $\mathcal{M} \neq \mathcal{D}_{\mathsf{sel}}$ occurs with probability at most $\mathsf{exp}(-R_n)$.

(B) Define a multiset (i.e., whose order does not matter, but repetitions are allowed) $\mathcal{N}_1 = \big\{(X_i, Y_i) \in \mathcal{D}_1 \setminus \mathcal{D}_2 : i \in [n]\big\}$, the observations that appear only in $\mathcal{D}_1$. Let $\mathcal{N}_2$ denote the analogous set for $\mathcal{D}_2$.

(C) For every overlapped index $i \in \mathcal{M}$, we take multisets $\mathcal{J}_{i,1} = \big\{(X_i^{(k)}, Y_i^{(k)}) \in \mathcal{D}_1 : k \in \{1, \ldots, R_n\}\big\}$, the replicated copies of (X_i, Y_i) in $\mathcal{D}_1$, and define multisets $\mathcal{J}_{i,2}$ similarly for $\mathcal{D}_2$.

We can write the sample splits as $\mathcal{D}_1 = \mathcal{N}_1 \cup \Big(\bigcup_{i\in\mathcal{M}} \mathcal{J}_{i,1}\Big)$ and $\mathcal{D}_2 = \mathcal{N}_2 \cup \Big(\bigcup_{i\in\mathcal{M}} \mathcal{J}_{i,2}\Big)$. We prove the following non-asymptotic results for the proof of Proposition **??**.

Lemma 4. *Let $\mathcal{M} = \mathcal{D}_1 \cap \mathcal{D}_2$ and $\mathcal{N}_1 \subseteq \mathcal{D}_1, \mathcal{N}_2 \subseteq \mathcal{D}_2$ denote the overlapping index set and unique sample set in Section 3. Then for any positive $t > 0$,*

$$\mathbb{P}\Big(\big||\mathcal{M}| - \rho n(1 - 2^{-R})\big| > t\Big) \leq 2\mathsf{exp}\Big(-\frac{2t^2}{\rho n}\Big). \tag{B.1}$$

Proof. We first note that $|\mathcal{M}| = \sum_{i=1}^n \mathbb{1}\big\{i\text{-th sample} \in \mathcal{D}_1 \cap \mathcal{D}_2\big\}$. We proceed as follows:

$$\begin{aligned}\mathbb{E}|\mathcal{M}| &= \sum_{i=1}^n \mathbb{P}\Big(i\text{-th sample} \in \mathcal{D}_1 \cap \mathcal{D}_2, i\text{-th sample} \in \mathcal{D}_{\mathsf{sel}}\Big) = \rho \sum_{i=1}^n \mathbb{P}\Big(i\text{-th sample} \in \mathcal{D}_1 \cap \mathcal{D}_2 \mid i\text{-th sample} \in \mathcal{D}_{\mathsf{sel}}\Big) \\ &= \rho n\Big(1 - 2 \cdot 2^{-(R+1)}\Big) = \rho n(1 - 2^{-R}).\end{aligned}$$

For the concentration inequality, we denote Z_i as the centered indicator functions: $Z_i = \mathbb{1}\big\{i\text{-th sample} \in \mathcal{D}_1 \cap \mathcal{D}_2\big\} - \mathbb{P}\big(i\text{-th sample} \in \mathcal{D}_1 \cap \mathcal{D}_2\big)$ such that $|\mathcal{M}| - \mathbb{E}|\mathcal{M}| = \sum_{i=1}^n Z_i$. We condition on the configuration of the subset $\mathcal{D}_{\mathsf{sel}} \subseteq \mathcal{D}_{\mathsf{train}}$ which is chosen to be duplicated. Therefore, we obtain that

$$\begin{aligned}\mathbb{P}\Big(|\mathcal{M}| - \mathbb{E}|\mathcal{M}| > t\Big) &= \mathbb{E}\Big(\mathbb{P}\Big(|\mathcal{M}| - \mathbb{E}|\mathcal{M}| > t \mid \mathcal{D}_{\mathsf{sel}}\Big)\Big) \\ &= \mathbb{E}\Big(\mathbb{P}\Big(\sum_{i\in\mathcal{D}_{\mathsf{sel}}} Z_i - \rho n(1 - 2^{-R}) > t \mid \mathcal{D}_{\mathsf{sel}}\Big)\Big) \\ &\leq 2\mathsf{exp}\Big(-\frac{2t^2}{\rho n}\Big),\end{aligned}$$

where the last step is justified by bounded differences inequality (Boucheron et al., 2004, Theorem 6.2). □

Lemma 5. *For $i \in \mathcal{M}$, the following inequalties hold*

$$\mathbb{P}\Big(\Big||\mathcal{J}_{i,1}| - R_n/2\Big| > t\Big) \vee \mathbb{P}\Big(\Big||\mathcal{J}_{i,2}| - R_n/2\Big| > t\Big) \leq 2\mathsf{exp}\Big(-\frac{2t^2}{R_n}\Big). \tag{B.2}$$

Proof. Recall that we define $\mathcal{J}_{i,1} = \{(X_i^{(k)}, Y_i^{(k)}) \in \mathcal{D}_1 : k \in R_n\}$. We have R_n copies of X_i in total. By the independent and identically distributed nature of random assignment, for any $i \in \mathcal{M}$, we have

$$\mathbb{P}\Big(\Big||\mathcal{J}_{i,1}| - R_n/2\Big| > t\Big) = \mathbb{P}\Big(\Big|\sum_{i=1}^{R_n} \mathbb{1}\big\{X_i^{(k)} \in \mathcal{D}_1\big\} - R_n/2\Big| > t\Big) \leq 2\mathsf{exp}\Big(-\frac{2t^2}{R_n}\Big).$$

Similarly, we have the same bound for the set $\mathcal{J}_{i,2}$ for all $i \in \mathcal{M}$. □

B.1 Function spaces

Definition 6 (Sobolev space $\mathcal{W}^{\beta,2}(\Omega)$)**.** *Let $\Omega = [0,1]^d$. Throughout the paper we work with integer smoothness indices $\beta, \zeta \in \mathbb{N}$.*

For an integer $k \geq 0$, define

$$\mathcal{W}^{k,2}(\Omega) := \Big\{ f \in L^2(\Omega) : D^\gamma f \in L^2(\Omega) \text{ for all multi–indices } \gamma \in \mathbb{N}^d \text{ with } |\gamma| \leq k \Big\},$$

equipped with the norm

$$\|f\|^2_{\mathcal{W}^{k,2}(\Omega)} := \sum_{|\gamma| \leq k} \|D^\gamma f\|^2_{L^2(\Omega)}.$$

When $\beta \in \mathbb{N}$ we write $\mathcal{W}_{\beta,2} := \mathcal{W}^{\beta,2}(\Omega), \|f\|_{\beta,2} := \|f\|_{\mathcal{W}^{\beta,2}(\Omega)}$.

Definition 7 (Hölder space $\mathcal{C}^\beta(\Omega)$)**.** *Let $\Omega = [0,1]^d$. For $\beta > 0$ set $k = \lfloor \beta \rfloor \in \mathbb{N}$ and $\alpha = \beta - k \in (0,1]$. For $g : \Omega \to \mathbb{R}$ define the Hölder seminorm*

$$[g]_{\mathcal{C}^{0,\alpha}(\Omega)} := \sup_{x \neq y} \frac{|g(x) - g(y)|}{|x - y|^\alpha}.$$

Then

$$\mathcal{C}^\beta(\Omega) := \Big\{ f \in \mathcal{C}^k(\overline{\Omega}) : \sum_{|\gamma| = k} [D^\gamma f]_{\mathcal{C}^{0,\alpha}(\Omega)} < \infty \Big\},$$

with norm

$$\|f\|_{\mathcal{C}^\beta(\Omega)} := \sum_{|\gamma| \leq k} \|D^\gamma f\|_{L^\infty(\Omega)} + \sum_{|\gamma| = k} [D^\gamma f]_{\mathcal{C}^{0,\alpha}(\Omega)}.$$

When $\beta \in \mathbb{N}$ this coincides with the usual $\mathcal{C}^k$ norm.

For integer ζ we will use the shorthand $\mathcal{C}^\zeta := \mathcal{C}^\zeta(\Omega), \|f\|_{\zeta,\infty} := \|f\|_{\mathcal{C}^\zeta(\Omega)}$.

C Influence Function-based Estimators

Bias correction is well understood in the literature (Bickel et al., 1993; Chernozhukov et al., 2017; Farrell et al., 2021). To remove this leading bias, another approach is to let $\psi(x; f)$ be the pathwise influence function of τ and define the bias-corrected estimator: $\widehat{\tau}_{\mathsf{corr}} = \tau(\widehat{f}) + \frac{1}{n}\sum_{i=1}^n \psi\big(X_i; \widehat{f}\big)$. Choose the mean-squared-error-optimal bandwidth $h \asymp n^{-1/(2\beta+d)}$, so that under regularity,

$$\|\widehat{f} - f\|^2_{\mathcal{W}^{\beta,2}} = \mathcal{O}_P\big(n^{-(2\beta-2\zeta)/(2\beta+d)}\big), \tag{C.1}$$

If $\beta \geq 2\zeta + \frac{d}{2}$, one can show that

$$\widehat{\tau}_{\mathsf{corr}} - \tau(f) = \frac{1}{n}\sum_{i=1}^n \psi\big(X_i; f\big) + o_P\big(n^{-1/2}\big),$$

so that $\widehat{\tau}_{\mathrm{corr}}$ is asymptotically normal at parametric rate $n^{-1/2}$.

Under a strengthened smoothness requirement such as $\beta \geq 2\zeta + d$, both undersmoothing and bias correction achieve the parametric rate $n^{-1/2}$; even when this condition is not met, each still improves the convergence rate relative to a naive plug-in. A practical advantage of undersmoothing is that it avoids explicit construction of an influence function, offering a less ad hoc route to efficiency.

D Recipe for Replication Parameter Tuning

In this section, we explain how to choose ρ and R based on the nonasymptotic result established in theorem 1.

Theorem 8 (Restatement of Theorem 1)**.** *With $\mathcal{L}_{\mathsf{CV}}$ defined as in equation (3.1), under Assumptions (subG), (kernel), and (density), the following statements hold:*

$$\left|\mathcal{L}_{\mathsf{CV}}(h,\rho_n,\mathsf{R}_n)-\left[\mathcal{L}^\dagger(h,\rho_n,\mathsf{R}_n)+\sigma^2\right]\right|\le\Delta_{n,\rho,R}, \tag{D.1}$$

where $\mathcal{L}^\dagger(h,\rho_n,\mathsf{R}_n)=\rho n\mathsf{R}\Big(\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R}+nh^d(1-\rho+\rho R))^2}\Big)+(1-\rho)n\Big[\sigma^2+h^{2\beta}+\frac{1}{\rho nh^d}\Big]$, with probability $1-o(1)$ and $\Delta_{n,\rho,R}\ll\mathcal{L}^\dagger(h,\rho_n,\mathsf{R}_n)$. Therefore, the optimal bandwidth to minimize $\mathcal{L}_{\mathsf{CV}}$ is $\frac{1}{C}h^\dagger\le(\widehat{h})^d\le Ch^\dagger$, for some constant $C>0$, where $h^\dagger=n^{-1}\mathsf{R}_n^{-1/2}\rho_n^{-3/2}$. If $-3\varrho+c=-2\frac{\beta+\zeta}{\beta+\zeta+d}$, then, with the same probability as above, $\frac{1}{C}n^{-\frac{1}{\beta+\zeta+d}}\le\widehat{h}\le Cn^{-\frac{1}{\beta+\zeta+d}}$, for a constant $C>0$.

We begin by characterizing the empirical cross-validation risk $\mathcal{L}_{\mathsf{CV}}$. By Theorem 1, we obtain that, if $\rho_n\mathsf{R}_n\to\infty$, asymptotically, we have

$$2m\mathcal{L}_{\mathsf{CV}}(h,\rho_n,R_n)\asymp\rho^2\mathsf{R}n^2h^d+\frac{1}{\rho h^d},$$

The minimizer of the right hand side is of the order

$$(h^\star)^d\asymp n^{-1+\frac{3}{2}\varrho-\frac{1}{2}c} \tag{D.2}$$

with the same high probability stated in Theorem 1. From (D.2), the two controllable quantities ρ_n and R_n affect the selected bandwidth. We therefore use them as levers to influence the chosen bandwidth and, more generally in black-box settings, the selected model. We then expand how to pick $\varrho(\bar{\beta},\zeta,d)$, provided known $\bar{\beta}$ (i.e. simply the smoothness β in this kernel nonparametric regression case). Recall that we want to set ρ_n and R_n such that $h^*=h_{\tau\text{-opt}}\asymp n^{-\frac{1}{\beta+\zeta+d}}$.

Parametrization. Let $\mathsf{R}_n=n^c$ and $\rho_n=n^{-\varrho}$ with $c>0$ and $\varrho\in[0,1]$. To enforce $h^*_{\tau-\mathsf{opt}}\asymp n^{-\frac{1}{\beta+\zeta+d}}$, it suffices to set

$$-3\varrho+c=-2\mathsf{A},\quad \mathsf{A}=\frac{\beta+\zeta}{\beta+\zeta+d}. \tag{D.3}$$

Equation (D.3) refers to the red segment in Figure 4.

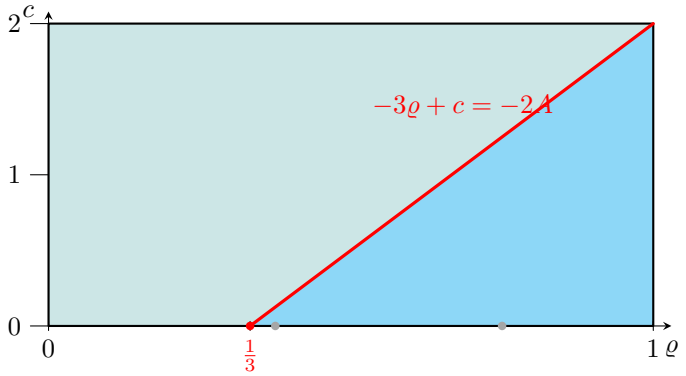


Figure 4: The (ϱ,c) plane. The red segment marks optimal choices of (ϱ,c) under the constraint $\varrho\in[0,1]$, ensuring that $\mathcal{D}_{\mathsf{sel}}$ is asymptotically nonnegligible. The teal region indicates dominance of the variance term (from the squared second–order remainder), whereas the cyan region indicates dominance of the bias term (from the first–order component $T_{f^*}(\widehat{f}-f^*)$). If $\beta\ge2\zeta+d$, the red segment lies between the two gray guide lines, further clarifying admissible specifications of (ϱ,c). Selecting (ϱ,c) on the red segment yields the rate $\mathbf{r}(n,\beta)=n^{-\left(\frac{1}{2}\wedge\frac{\beta+\zeta}{\beta+\zeta+d}\right)}$ for the plug–in estimator of the target functional τ as proved in Theorem 1.

E Proof of Theorem 1

Proof sketch. From Appendix B, if $\mathsf{R}_n \to \infty$, the following are true

$$|\mathcal{M}| = \rho n + \mathcal{O}_P(\sqrt{\rho n}), \quad |\mathcal{J}_{i,1}|, |\mathcal{J}_{i,2}| = \frac{1}{2}\mathsf{R}_n + \mathcal{O}_P(\sqrt{\mathsf{R}_n}), \quad |\mathcal{N}_1| \asymp |\mathcal{N}_2| = \frac{(1-\rho)n}{2} + \mathcal{O}_P(n^{1/2}). \tag{E.1}$$

The NW estimator built from $\mathcal{D}_2$ is

$$\widehat{f}_h(\cdot; \mathcal{D}_2) = \frac{\sum_{(\boldsymbol{x},y)\in\mathcal{D}_2} yK_h(\cdot - \boldsymbol{x})}{\sum_{(\boldsymbol{x},y)\in\mathcal{D}_2} K_h(\cdot - \boldsymbol{x})}. \tag{E.2}$$

We proceed by expanding (3.1) as follows:

$$\begin{aligned} 2m \cdot \mathcal{L}_{\mathsf{CV}}(h,\rho,R) &= \sum_{i\in\mathcal{M}}\sum_{k\in\mathcal{J}_{i,1}} (y_i^{(k)} - \widehat{f}_h(\boldsymbol{x}_i^{(k)}; \mathcal{D}_2))^2 + \sum_{i\in\mathcal{M}}\sum_{k\in\mathcal{J}_{i,2}} (y_i^{(k)} - \widehat{f}_h(\boldsymbol{x}_i^{(k)}; \mathcal{D}_1))^2 \\ &\quad + \Big(\sum_{i\in\mathcal{N}_1} (y_i - \widehat{f}_h(\boldsymbol{x}_i; \mathcal{D}_2))^2 + \sum_{i\in\mathcal{N}_2} (y_i - \widehat{f}_h(\boldsymbol{x}_i; \mathcal{D}_1))^2\Big) \\ &=: T_{11} + T_{12} + T_{21} + T_{22}. \end{aligned}$$

Throughout the proof we omit the subscript n and write (R, ρ) for (R_n, ρ_n). We condition on the randomness introduced by $(\mathcal{M}, \{\mathcal{J}_{i,1}, \mathcal{J}_{i,2}\}_{i\in\mathcal{M}}, \mathcal{N}_1, \mathcal{N}_2)$. It suffices to show the order of the terms T_{11} and T_{21}.

Overlapped term T_{11}. By construction, we obtain that

$$T_{11} = \sum_{i\in\mathcal{M}}\sum_{k\in\mathcal{J}_{i,1}} (y_i^{(k)} - \widehat{f}_h(\boldsymbol{x}_i^{(k)}; \mathcal{D}_2))^2 = \sum_{i\in\mathcal{M}} |\mathcal{J}_{i,1}|\mathsf{e}_i^2, \tag{E.3}$$

where the error can be written as

$$\mathsf{e}_i = y_i - \widehat{f}_h(\boldsymbol{x}_i; \mathcal{D}_2) = \frac{\sum_{j\in\mathcal{D}_2\setminus\mathcal{J}_{i,2}} K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)(y_i - y_j)}{|\mathcal{J}_{i,2}| + \sum_{j\in\mathcal{D}_2\setminus\mathcal{J}_{i,2}} K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)} \tag{E.4}$$

$$= \frac{\mathsf{A}_i\varepsilon_i - \mathsf{U}_i + \mathsf{B}_i}{|\mathcal{J}_{i,2}| + \mathsf{A}_i}, \tag{E.5}$$

where

$$\mathsf{A}_i = \sum_{j\in\mathcal{D}_2\setminus\mathcal{J}_{i,2}} K_h(\boldsymbol{x}_i - \boldsymbol{x}_j), \tag{E.6}$$

$$\mathsf{B}_i = \sum_{j\in\mathcal{D}_2\setminus\mathcal{J}_{i,2}} K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)(f^\star(\boldsymbol{x}_i) - f^\star(\boldsymbol{x}_j)) \asymp \mathsf{A}_i^2 h^\beta, \tag{E.7}$$

$$\mathsf{U}_i = \sum_{j\in\mathcal{D}_2\setminus\mathcal{J}_{i,2}} K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)\varepsilon_j. \tag{E.8}$$

Further by conditioning on $\boldsymbol{x}_i$, we obtain that

$$\mathbb{E}(\mathsf{e}_i^2 \mid \boldsymbol{x}_i) = \frac{\mathsf{A}_i^2\sigma^2 + \mathsf{Var}(\mathsf{U}_i \mid \boldsymbol{x}_i) + \mathsf{B}_i^2}{(|\mathcal{J}_{i,2}| + \mathsf{A}_i)^2}, \tag{E.9}$$

yielding that

$$\begin{aligned} &\mathbb{E}(\mathsf{e}_i^2 \mid \boldsymbol{x}_i) \\ &\asymp \sigma^2\Big(\frac{nh^d(1-\rho+\rho\mathsf{R})}{\mathsf{R} + nh^d(1-\rho+\rho\mathsf{R})}\Big)^2 + \mathcal{O}_P\Big(\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R} + nh^d(1-\rho+\rho R))^2}\Big) + \mathcal{O}_P\Big(h^{2\beta}\Big(\frac{nh^d(1-\rho+\rho\mathsf{R})}{\mathsf{R} + nh^d(1-\rho+\rho\mathsf{R})}\Big)^2\Big), \end{aligned}$$

where the omitted constant only depends on the marginal density of $\boldsymbol{x}$ if one reads the noise level $\sigma \asymp 1$.

It's straightforward to check that

$$h \to 0, \quad \rho n h^d \to 0, \quad \rho \mathsf{R} \to \infty,$$

implies that

$$\frac{T_{11}}{2m} = \mathsf{C}\sigma^2 \frac{\rho \mathsf{R}}{2(1-\rho+\rho\mathsf{R})} \Big(\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R}+nh^d(1-\rho+\rho R))^2} \Big) \Big(1+o_P(1)\Big). \tag{E.10}$$

Non-overlapped term T_{21}. We decompose

$$\mathbb{E}\Big((y_i - \widehat{f}_h(\boldsymbol{x}_i;\mathcal{D}_2))^2 \mid \boldsymbol{x}_i\Big) = \sigma^2 + \mathsf{Bias}(\boldsymbol{x}_i)^2 + \mathsf{Var}(\widehat{f}_h(\boldsymbol{x}_i;\mathcal{D}_2) \mid \boldsymbol{x}_i) \tag{E.11}$$

$$\asymp \sigma^2 + h^{2\beta} + \frac{\sigma^2}{nh^d} \frac{2(1-\rho)+\rho(\mathsf{R}^2+\mathsf{R})}{(1-\rho+\rho\mathsf{R})^2}(1+o_P(1)). \tag{E.12}$$

In the regime when $\rho\mathsf{R} \to \infty$, we have

$$\frac{\sigma^2}{nh^d} \frac{2(1-\rho)+\rho(\mathsf{R}^2+\mathsf{R})}{(1-\rho+\rho\mathsf{R})^2}(1+o_P(1)) \asymp \frac{1}{\rho n h^d}.$$

Combine T_{11} and T_{21}. In the final step, we balance the dominating terms

$$2m\mathcal{L}_{\mathsf{CV}} \asymp \underbrace{\rho n \mathsf{R}\Big(\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R}+nh^d(1-\rho+\rho R))^2} \Big)\Big(1+o_P(1)\Big)}_{\text{overlapped}} + \underbrace{(1-\rho)n\Big[\sigma^2 + h^{2\beta} + \frac{1}{\rho n h^d}\Big]}_{\text{non-overlapped}} \tag{E.13}$$

$$\asymp \rho^2 \mathsf{R} n^2 h^d + \frac{1}{\rho h^d}, \tag{E.14}$$

leading that

$$\widehat{h} = \arg\min_{h>0} \mathcal{L}_{\mathsf{CV}}(h) \asymp n^{\frac{1}{d}(-1+\frac{3}{2}\varrho-\frac{1}{2}c)},$$

if we follow the parametrization in Appendix D. The optimal choice for (ρ, R) pairs are given by

$$3\varrho - c = 2\frac{\beta+\zeta}{\beta+\zeta+d}. \tag{E.15}$$

□

We first provide a classical non-asymptotic result for function estimation, which is used to bound the terms T_{21} and T_{22} in the proof of Theorem 1 as stated in the preceding proof sketch.

E.1 Non-overlapped Term

Proposition 9. *Under Assumptions* (subG) *and* (kernel)*, following the notation convention in Section 3.1, with probability* $1-\delta$ *for* $\delta \in (0, 1/2)$*,*

$$\sup_{x\in\mathcal{X}} \Big|\widehat{f}_h(x;\mathcal{D}_2) - f^\star(x)\Big| \le \mathsf{C}_0 h^\beta + \mathsf{C}_1 \sqrt{\frac{\mathsf{R}\log(\mathsf{C}_1 h^{-d}\delta^{-1})}{mh^d}},$$

for some universal constants $\mathsf{C}_0, \mathsf{C}_1$*.*

Remark 10. *Recall that* $m = \frac{1}{2}|\mathcal{D}_{\mathsf{aug}}| = \frac{1}{2}\big((1-\rho)n + \rho n\mathsf{R}\big)$*. Therefore, we obtain that* $m \asymp \rho n\mathsf{R}$ *if* $\rho\mathsf{R} \to \infty$*.*

Proof of Proposition 9. We first decompose the sup norm as

$$\begin{aligned}
|\widehat{f}_h(x;\mathcal{D}_2) - f^\star(x)| &\le \frac{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)|f^\star(X_i)-f^\star(x)|}{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)} + \frac{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)\varepsilon_i}{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)} \\
&= \frac{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)|f^\star(X_i)-f^\star(x)|\mathbb{1}\big(\|X_i-x\|\le h\big)\big)}{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)} + \frac{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)\varepsilon_i}{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)} \\
&\lesssim h^\beta + \frac{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)\varepsilon_i}{\sum_{i\in\mathcal{D}_2} K_h(x-X_i)} \\
&= h^\beta + \sum_{i\in\mathcal{D}_2} \mathsf{W}_h(x-X_i;\mathcal{D}_2)\varepsilon_i.
\end{aligned}$$

For the variance term, we first note that

$$\mathsf{Var}\Big(\sum_{i\in\mathcal{D}_2} \mathsf{W}_h(\boldsymbol{x}-\boldsymbol{x}_i;\mathcal{D}_2)\varepsilon_i \mid \{\varepsilon_i\}_{1\le i\le n}\Big) \tag{E.16}$$

$$= \sigma^2 \sum_{i\in\mathcal{D}_2} Q_i^2 \mathsf{W}_h^2(\boldsymbol{x}-\boldsymbol{x}_i;\mathcal{D}_2) \quad \text{where } Q_i := \begin{cases} |\mathcal{J}_{i,2}| & \text{if } i\in\mathcal{M} \\ 1 & \text{if } i\in\mathcal{N}_2 \end{cases} \tag{E.17}$$

$$\asymp \sigma^2 \frac{2(1-\rho)+\rho\mathsf{R}^2}{nh^d(1-\rho+\rho\mathsf{R})^2}. \tag{E.18}$$

Therefore, we obtain that

$$\mathbb{P}\Big(\sum_{i\in\mathcal{D}_2} \mathsf{W}_h(x-X_i;\mathcal{D}_2)\varepsilon_i \ge K_\varepsilon\sqrt{\frac{2(1-\rho)+\rho\mathsf{R}^2}{nh^d(1-\rho+\rho\mathsf{R})^2}}\Big) \mid \{\boldsymbol{x}_i\}_{1\le i\le n}\Big) \le \delta,$$

implying that with high probability $1-\delta$,

$$|\widehat{f}_h(\boldsymbol{x};\mathcal{D}_2) - f^\star(\boldsymbol{x})| \lesssim h^\beta + \frac{1}{\sqrt{\rho nh^d}},$$

for fixed $\boldsymbol{x}\in\Omega$.

To have a uniform lower bound for $\sum_{i\in\mathcal{D}_2} K_h(\boldsymbol{x}-\boldsymbol{x}_i)$, we take one step discretization onto

$$\Omega_{\mathsf{grid}} = \Big\{\boldsymbol{x} : (x_i)_{1\le i\le d} \in \{a_1,\cdots,a_N\} \text{ where } a_1<a_2<\cdots<a_N : a_t - a_{t-1} = \Delta = \frac{p_{\max}h}{L_k\sqrt{d}}, \forall 2\le t\le N\}. \tag{E.19}$$

For $\boldsymbol{y}\in\mathcal{X}_{\mathsf{grid}}$, the nearest grid point with respect to $\boldsymbol{x}$, we have

$$\sum_{i\in\mathcal{D}_2}\big(K(\boldsymbol{x}-\boldsymbol{x}_i)-K(\boldsymbol{y}-\boldsymbol{x}_i)\big) \le L_K mh^{-(d+1)}\|\boldsymbol{x}-\boldsymbol{y}\| \le L_k mh^{-(d+1)}\sqrt{d\Delta^2} = mh^{-d}p_{\max}$$

we have, for any $\eta>0$,

$$\begin{aligned}
&\mathbb{P}\Big(\exists \boldsymbol{x}\in\Omega : \mathsf{C}_1\sum_{i\in\mathcal{D}_2} K_h(x-X_i) > \sqrt{\frac{m\mathsf{R}}{h^d}}+\eta\Big) \\
&\le \mathbb{P}\Big(\exists y\in\Omega_{\mathsf{grid}} : \sum_{i\in\mathcal{D}_2} K_h(y-X_i) > \mathsf{C}_1\sum_{i\in\mathcal{D}_2} K_h(x-X_i) > \sqrt{\frac{m\mathsf{R}}{h^d}}+\delta\Big) \\
&\le \mathbb{P}\Big(\exists y\in\Omega_{\mathsf{grid}}, \sum_{i\in\mathcal{D}_2} K_h(y-X_i) - \mathbb{E}\Big(\sum_{i\in\mathcal{D}_2} K_h(y-X_i)\Big) \ge \delta\Big),
\end{aligned}$$

where the last step is due to the fact that

$$\mathbb{E}\Big(\sum_{i\in\mathcal{D}_2} K_h(y-X_i)\Big) = m\int_{u\in\mathcal{X}} K\big(\frac{y-u}{h}\big)p(u)\mathrm{d}u \le mh^{-d}p_{\max}$$

Applying Hoeffding inequality, we have

$$\mathbb{P}\Big(\sup_{\boldsymbol{x}\in\Omega}\big|\sum_{i\in\mathcal{D}_2} K_h(\boldsymbol{x}-\boldsymbol{x}_i)\big| \ge \mathsf{C}\sqrt{\frac{m\mathsf{R}}{h^d}\log\big(\frac{1}{h^d}\delta\big)}\Big) < \delta,$$

which proves the statment. □

E.2 Overlapped Term

Recall that

$$T_{11} = \sum_{i\in\mathcal{M}}\sum_{k\in\mathcal{J}_{i,1}} (y_i^{(k)} - \widehat{f}_h(\boldsymbol{x}_i^{(k)};\mathcal{D}_2))^2 = \sum_{i\in\mathcal{M}} |\mathcal{J}_{i,1}|\mathsf{e}_i^2, \tag{E.20}$$

where e_i^2 is given by equation (E.4).

Proposition 11. *Suppose that Assumption (density),(subG), (kernel), and $f^\star \in \mathcal{W}_{\beta,2}$ hold. We assume the event $\mathcal{E}_n = \{\max_{i\in\mathcal{M}}\big||\mathcal{J}|_{i,1} - \frac{1}{2}\mathsf{R}\big| \vee \max_{i\in\mathcal{M}}\big||\mathcal{J}|_{i,2} - \frac{1}{2}\mathsf{R}\big| = o(\mathsf{R})\}$ holds true with probability $1-o(1)$. There exists fixed constants $0<\mathsf{c}<\mathsf{C}<\infty$ such that, with probability $1-o(1)$,*

$$\begin{aligned}&\mathsf{c}\sigma^2\frac{\rho\mathsf{R}}{2(1-\rho+\rho\mathsf{R})}\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R}))^2}\\ &\le \frac{T_{11}}{2m} \le \mathsf{C}\sigma^2\frac{\rho\mathsf{R}}{2(1-\rho+\rho\mathsf{R})}\frac{nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R}))^2}.\end{aligned} \tag{E.21}$$

Remark 12. *Appendix B justifies the probability of $\mathcal{E}_n$ converges to 1 as $n\to\infty$.*

Proof. By equation (E.4), we've shown that We continue from

$$T_{11} = \sum_{i\in M}\frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\big(\mathsf{A}_i\varepsilon_i - \mathsf{U}_i\big)^2 + 2\sum_{i\in M}\frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\mathsf{B}_i\big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i\big) + \sum_{i\in M}\frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\mathsf{B}_i^2. \tag{E.22}$$

On $\mathcal{E}_n$, uniformly over $i\in M$, we have $|\mathcal{J}_{i,1}| = \frac{\mathsf{R}}{2}(1+o(1))$ and $|\mathcal{J}_{i,2}| = \frac{\mathsf{R}}{2}(1+o(1))$. Also, by the same Bernstein argument used in Proposition 9, applied to $K_h(\boldsymbol{x}_i-\cdot)$ and $K_h(\boldsymbol{x}_i-\cdot)^2$, with probability $1-o(1)$, $\mathsf{A}_i \asymp nh^d(1-\rho+\rho\mathsf{R})$ and $\mathsf{c}\sigma^2 nh^d\big(2(1-\rho)+\rho\mathsf{R}^2\big) \le \mathsf{Var}(U_i \mid x_1,\dots,x_n,\text{split}) \le C\,\sigma^2 nh^d\big(2(1-\rho)+\rho\mathsf{R}^2\big)$ uniformly over $i\in\mathcal{M}$, where the contribution of every overlapped family to $\mathsf{Var}(U_i \mid \boldsymbol{x}_1,\dots,\boldsymbol{x}_n,\text{split})$ is weighted by $|\mathcal{J}_{i,2}|^2$, exactly as in the definition of U_i. Consequently,

$$\mathsf{c}\frac{\mathsf{R}}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2} \le \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2} \le \mathsf{C}\frac{\mathsf{R}}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}$$

uniformly over $i\in M$, with probability $1-o(1)$.

We first treat the last sum. Since $K_h(\boldsymbol{x}_i-\boldsymbol{x}_j)=0$ unless $\|x_i-x_j\|\le h$, the smoothness of $f^\star$ gives

$$|f^\star(\boldsymbol{x}_i) - f^\star(\boldsymbol{x}_j)| \le \mathsf{C}h^\beta$$

whenever $K_h(\boldsymbol{x}_i-\boldsymbol{x}_j)\neq 0$. Hence

$$|\mathsf{B}_i| \le \mathsf{C}h^\beta \sum_{j\in\mathcal{D}_2\setminus\mathcal{J}_{i,2}} K_h(\boldsymbol{x}_i-\boldsymbol{x}_j) = \mathsf{C}\mathsf{A}_i h^\beta \le \mathsf{C}nh^d(1-\rho+\rho\mathsf{R})h^\beta$$

uniformly over $i \in \mathcal{M}$, and therefore

$$\sum_{i\in M} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\mathsf{B}_i^2 \le \mathsf{C}|\mathcal{M}|\frac{\mathsf{R}\, n^2h^{2d}(1-\rho+\rho\mathsf{R})^2h^{2\beta}}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}. \tag{E.23}$$

Since $\rho nh^d \to 0$, $\rho\mathsf{R}\to\infty$, and $h\to 0$, we have $\frac{n^2h^{2d}(1-\rho+\rho\mathsf{R})^2h^{2\beta}}{nh^d(2(1-\rho)+\rho\mathsf{R}^2)} \to 0$. Next, by Cauchy-Schwarz inequality, we obtain that

$$\left|2\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\mathsf{B}_i\big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i\big)\right| \le 2\left(\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i\big)^2\right)^{1/2}\left(\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+A_i)^2}\mathsf{B}_i^2\right)^{1/2}.$$

Hence, once the first sum is shown to be of order $|\mathcal{M}|\frac{\mathsf{R}\,nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}$, the middle sum in is of smaller order as well.

It remains to treat the first sum. Conditional on $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ and the sample split, it is a quadratic form in $(\varepsilon_1,\dots,\varepsilon_n)$. Its conditional expectation is

$$\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+A_i)^2}\,\mathbb{E}\Big(\big(A_i\varepsilon_i-U_i\big)^2 \mid x_1,\dots,x_n,\text{split}\Big) = \sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\Big(\sigma^2\mathsf{A}_i^2+\mathsf{Var}(\mathsf{U}_i\mid \boldsymbol{x}_1,\dots,\boldsymbol{x}_n)\Big),$$

because ε_i is independent of U_i. By the bounds above,

$$\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\mathsf{Var}(\mathsf{U}_i\mid \boldsymbol{x}_1,\dots,\boldsymbol{x}_n) \asymp |\mathcal{M}|\frac{\mathsf{R}\,nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}$$

with probability $1-o(1)$. On the other hand, we have

$$\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\sigma^2\mathsf{A}_i^2 \le \mathsf{C}|\mathcal{M}|\frac{\mathsf{R}n^2h^{2d}(1-\rho+\rho\mathsf{R})^2}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}$$
$$\frac{n^2h^{2d}(1-\rho+\rho\mathsf{R})^2}{nh^d(2(1-\rho)+\rho\mathsf{R}^2)} \to 0$$

under $\rho nh^d\to 0$ and $\rho\mathsf{R}\to\infty$. Therefore the conditional expectation of the first sum is bounded from above and below by fixed multiples of $\sigma^2|\mathcal{M}|\frac{\mathsf{R}\,nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}$ with probability $1-o(1)$.

We now apply Hanson-Wright conditionally on $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ and the sample split. The Frobenius norm squared of the corresponding coefficient matrix is bounded by a fixed multiple of

$$\frac{\mathsf{R}^2}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^4}\sum_{i\in\mathcal{M}}\sum_{\ell\in\mathcal{M}}\Big|\mathsf{Cov}\Big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i,\mathsf{A}_\ell\varepsilon_\ell-U_\ell\mid x_1,\dots,x_n,\text{split}\Big)\Big|^2.$$

Because K_h is supported on the unit ball, the covariance in the last display vanishes unless $\|x_i-x_\ell\|\le 2h$. By the same kernel-counting argument as in Proposition 9, $\#\big\{(i,\ell)\in\mathcal{M}\times\mathcal{M} : \|\boldsymbol{x}_i-\boldsymbol{x}_\ell\|\le 2h\big\} = \mathcal{O}_P\big(|\mathcal{M}|+|\mathcal{M}|^2h^d\big)=\mathcal{O}_P(|\mathcal{M}|)$, since $|\mathcal{M}|\asymp\rho n$ and $\rho nh^d\to 0$. For every such pair,

$$\Big|\mathsf{Cov}\Big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i,\, A_\ell\varepsilon_\ell-U_\ell\mid \boldsymbol{x}_1,\dots,\boldsymbol{x}_n\Big)\Big| \le C\sigma^2nh^d\big(2(1-\rho)+\rho\mathsf{R}^2\big),$$

again by the same variance calculation used above. Hence the Frobenius norm squared is $O_P\big(|\mathcal{M}|\big[\frac{\mathsf{R}\,nh^d(2(1-\rho)+\rho\mathsf{R}^2)}{\big(\mathsf{R}+nh^d(1-\rho+\rho\mathsf{R})\big)^2}\big]^2\big)$. The operator norm is bounded by the Frobenius norm. Therefore Hanson-Wright yields

$$\mathbb{P}\Bigg(\Bigg|\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i\big)^2 - \mathbb{E}\Big(\sum_{i\in M} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i\big)^2 \;\Big|\; \boldsymbol{x}_1,\dots,\boldsymbol{x}_n\Big)\Bigg|$$
$$> \frac{1}{2}\mathbb{E}\Big(\sum_{i\in\mathcal{M}} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}|+\mathsf{A}_i)^2}\big(\mathsf{A}_i\varepsilon_i-\mathsf{U}_i\big)^2 \;\Big|\; \boldsymbol{x}_1,\dots,\boldsymbol{x}_n\Big)\;\Bigg|\; \boldsymbol{x}_1,\dots,\boldsymbol{x}_n\Bigg) \le 2e^{-c|\mathcal{M}|}$$

for some fixed $c > 0$. Since $|\mathcal{M}| \to \infty$, this probability is $o(1)$. Thus, with probability $1 - o(1)$,

$$\sum_{i \in M} \frac{|\mathcal{J}_{i,1}|}{(|\mathcal{J}_{i,2}| + \mathsf{A}_i)^2} \big(\mathsf{A}_i \varepsilon_i - \mathsf{U}_i\big)^2 \asymp \sigma^2 |\mathcal{M}| \frac{\mathsf{R}\, nh^d(2(1-\rho) + \rho\mathsf{R}^2)}{\big(\mathsf{R} + nh^d(1 - \rho + \rho\mathsf{R})\big)^2}.$$

Combining this with the bounds for the middle and last sums gives

$$T_{11} \asymp \sigma^2 |\mathcal{M}| \frac{R\, nh^d(2(1-\rho) + \rho\mathsf{R}^2)}{\big(\mathsf{R} + nh^d(1 - \rho + \rho\mathsf{R})\big)^2}$$

with probability $1 - o(1)$. Finally, since $|\mathcal{M}| = \rho n(1 + o(1))$ and $2m = n(1 - \rho + \rho\mathsf{R})$,

$$\frac{T_{11}}{2m} \asymp \sigma^2 \frac{\rho\mathsf{R}}{2(1 - \rho + \rho\mathsf{R})} \frac{nh^d(2(1-\rho) + \rho\mathsf{R}^2)}{\big(\mathsf{R} + nh^d(1 - \rho + \rho\mathsf{R})\big)^2},$$

which concludes the proof. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1.

(I) By Bernstein inequality for $\{\varepsilon_i^2\}$, for any $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$,

$$\sum_{i \in \mathcal{N}_1} \big(Y_i - \widehat{f}_h(X_i; \mathcal{D}_2)\big)^2 = |\mathcal{N}_1| \Big(\sigma^2 + O\Big(h^{2\beta} + \frac{1}{\rho R n h^d}\Big)\Big) + O\Big(\sqrt{|\mathcal{N}_1| \log(1/\delta)} + \log(1/\delta)\Big),$$

and the same bound holds for $\sum_{i \in \mathcal{N}_2} \big(Y_i - \widehat{f}_h(X_i; \mathcal{D}_1)\big)^2$. Dividing by $2m$ gives

$$\frac{T_{21} + T_{22}}{2m} = \sigma^2 + C' \Big(h^{2\beta} + \frac{1}{\rho R n h^d}\Big) \;\pm\; o\Big(h^{2\beta} + \frac{1}{\rho R n h^d}\Big), \tag{E.24}$$

with probability at least $1 - Ce^{-c \log(1/\delta)}$.

(II) For $i \in \mathcal{M}$ and $k \in \mathcal{J}_{i,1}$, using $K(0) = 1$ and the NW weights,

$$Y_i^{(k)} - \widehat{f}_h(X_i^{(k)}; \mathcal{D}_2) = \frac{\sum_{j \in \mathcal{D}_2 \setminus \mathcal{J}_{i,2}} (Y_i^{(k)} + Y_j) K_h(X_i^{(k)} - X_j)}{|\mathcal{J}_{i,2}| + \sum_{j \in \mathcal{D}_2 \setminus \mathcal{J}_{i,2}} K_h(X_i^{(k)} - X_j)}.$$

With the shorthands,

$$\kappa_i^{(1)} = \frac{1}{h^d} \mathbb{E}\big[(Y_i + Y) K_h(X_i - X) \,\big|\, X_i, Y_i\big], \quad \kappa_i^{(0)} = \frac{1}{h^d} \mathbb{E}\big[K_h(X_i - X) \,\big|\, X_i\big],$$

a direct algebraic manipulation yields (E.11):

$$\begin{aligned} T_{11} &= \sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{J}_{i,1}} \left(\frac{\frac{1}{h^d} \sum_{j \in \mathcal{D}_2 \setminus \mathcal{J}_{i,2}} (Y_i^{(k)} + Y_j) K_h(X_i^{(k)} - X_j)}{\frac{1}{h^d} |\mathcal{J}_{i,2}| + \frac{1}{h^d} \sum_{j \in \mathcal{D}_2 \setminus \mathcal{J}_{i,2}} K_h(X_i^{(k)} - X_j)} \right)^2 \\ &= \sum_{i \in \mathcal{M}} \big(R - |\mathcal{J}_{i,2}|\big) \Big(\frac{h^d(m - |\mathcal{J}_{i,2}|)\kappa_i^{(1)}}{|\mathcal{J}_{i,2}| + h^d(m - |\mathcal{J}_{i,2}|)\kappa_i^{(0)}} \Big)^2 \;+\; R_1, \end{aligned} \tag{E.25}$$

where the remainder R_1 is the difference between the two lines.

Control of the main term in (E.25). By Hoeffding inequality, $|\mathcal{J}_{i,2}| = R/2 + c_i\Delta_n(\delta)$ with $|c_i| \le 1$ and $\Delta_n(\delta) = \sqrt{R\{\log(\rho n) - \log\delta\}}$ for all $i \in \mathcal{M}$ with probability $1-\delta$. Using $|\mathcal{M}| = \rho n(1+o_{\mathbb{P}}(1))$, we get

$$\frac{T_{11}}{2m} = C\,\rho R\Big(\frac{\rho R n h^d}{\rho R n h^d + R}\Big)^2 \;\pm\; o\Big(\rho R\Big(\frac{\rho R n h^d}{\rho R n h^d + R}\Big)^2\Big) + \frac{R_1}{2m}. \tag{E.26}$$

The same analysis applies to T_{12}, producing an identical leading term and a remainder R_1'.

Remainders R_1 and R_1'. Decomposing $R_1 = R_{11} + R_{12}$ yields $|R_{11}| + |R_{12}| \le C\,C_n(\rho)\,n^{2/3}$ with probability at least $1 - 4e^{-c\kappa_0^2 n^{1/6}/(B^2\vee 1)} - 2e^{-n^{1/6}\log n/\rho} - 2e^{-n^{1/3}/(2\rho)}$. The same bound holds for the symmetric remainder R_1'. Since $C_n(\rho)\,n^{2/3} = o\Big(\rho R\big(\frac{\rho R n h^d}{\rho R n h^d + R}\big)^2\Big)$ on the bandwidth scales of interest (in particular for $h\rho \gtrsim n^{-1}$), the remainder contribution is negligible in (E.26).

(III) Add the two overlapped parts and use (E.24) for the non-overlapped pieces: with probability at least $1 - Ce^{-\xi(n,\rho,R)}$,

$$\mathcal{L}_{\mathsf{CV}}(h,\rho,R) = C\,\rho R\Big(\frac{\rho R n h^d}{\rho R n h^d + R}\Big)^2 \;+\; C'\Big(h^{2\beta} + \frac{1}{\rho R n h^d}\Big) \;+\; \sigma^2 \;\pm\; \Delta_{n,\rho,R}, \tag{E.27}$$

where $\Delta_{n,\rho,R} \ll \rho R(\frac{\rho R n h^d}{\rho R n h^d + R})^2 + h^{2\beta} + (\rho R n h^d)^{-1}$ collects the bracket fluctuations, the $T_{21}+T_{22}$ fluctuations, and the remainders R_1, R_1' just bounded. This is (3.2).

(IV) Let $S(h) := \rho R n h^d$. For $S \ll R$ the first term in (E.27) behaves like $C\,\rho R\,(S/R)^2 = C\,\rho^3 R\,n^2 h^{2d}$, while the variance term is $C'/(S) = C'/(\rho R n h^d)$. Balancing these two dominates the optimization and yields

$$h^{3d} \asymp \frac{1}{\rho^4 R^2 n^3} \qquad\Longrightarrow\qquad (h^\dagger)^d \asymp n^{-1}R^{-2/3}\rho^{-4/3}.$$

A standard constant-factor comparison shows that: for some universal $C > 0$,

$$\frac{1}{C}\,(h^\dagger)^d \;\le\; (\widehat{h})^d \;\le\; C\,(h^\dagger)^d \quad \text{with probability at least } 1 - Ce^{-\xi(n,\rho,R)}.$$

Equivalently, $\widehat{h}$ lies within a constant factor of $h^\dagger$. Finally, parametrize $R = n^c$ and $\rho = n^{-\varrho}$. Imposing the condition $-2\varrho + c = -\frac{3}{2}\frac{\beta+\zeta}{\beta+\zeta+d}$ gives

$$(h^\dagger)^d \asymp n^{-\frac{\beta+\zeta}{\beta+\zeta+d}} \qquad\Longrightarrow\qquad \widehat{h} \asymp n^{-\frac{1}{\beta+\zeta+d}},$$

with high probability, which completes the proof. □

F Proof of Theorem 2

We first work on an intermediate step, fully inspired by Lepskii's method but in functional estimation task. We define a rate function

$$\mathbf{r}(n,\beta) = \Big(n^{-\frac{\beta-\zeta}{\beta+\zeta+1}} \vee n^{-1/2}\Big)\log n.$$

Algorithm 3: Adaptive functional estimation procedure

Input: $\beta_{\max}, \beta_{\min}$: Upper and lower bound on β

Discretize $[\beta_{\min}, \beta_{\max}]$ with $\mathcal{B}_N = \{\beta_{\min} = \beta_1 < \cdots < \beta_{N-1} < \beta_N = \beta_{\max}\}$, where $\Delta = \beta_{i+1} - \beta_i := \frac{1}{\log n}$ for all $i \in [N-1]$.

Choose $\tilde{\beta} = \max_{\beta_j \in \mathcal{B}_N} \Big\{\forall i < j, \big|\tau(\widehat{f}_{\beta_i}) - \tau(\widehat{f}_{\beta_j})\big| \le \mathbf{r}(n,\beta_i)\Big\}$.

Output: $\tau(\widehat{f}_{\tilde{\beta}})$

In the preceding algorithm, we manipulate the bandwidth according to the smoothness index grid. Mirroring the control of proportion ρ_n and number of copies, we have the following adaptive procedure without the demand to control the bandwith explicitly. We first cite a classical non-asymptotic bound for NW estimator, see Tsybakov (2008). As stated in Theorem 1, we set oracle bandwidth $h_n = n^{-\frac{1}{\beta+\zeta+d}}$.

Proposition 13. *Under the same Assumptions, we obtain that there exists $C > 0$ such that for any $s, s' > 0$, the following inequality holds*

$$\mathbb{P}\left(\left|\tau(\widehat{f}_\beta) - \tau(f^\star)\right|^2 > C\Big(\frac{s}{n^2 h^{2+4\zeta}} + h^{2\beta-2\zeta} + n^{-1}(s')^2\Big)\right) \le 2\exp(-s \wedge s'). \tag{F.1}$$

Lemma 14. *Let $\Delta = \beta_i - \beta_{i-1} := \frac{1}{\log n}$, then $r(n, \beta_{i-1})/r(n, \beta_i) = o(n^\varepsilon)$ for any $\varepsilon > 0$.*

Proof. Let β^* be the solution to $\alpha(\beta) := \frac{\beta-\zeta}{\beta+\zeta+1} = \frac{1}{2}$. It's obvious that it suffices to consider the case where (i) $\beta_{j-1} < \beta_j < \beta^*$ and (ii) $\beta_{j-1} \le \beta^* \le \beta_j$. Consider

$$\delta_j = \log\Big(\frac{r(n, \beta_{j-1})}{r(n, \beta) j)}\Big) = \begin{cases} \Big(\alpha(\beta_j) - \alpha(\beta_{j-1})\Big)\log n & \text{case (i)} \\ \frac{1}{2} - \alpha(\beta_{j-1}) & \text{case (ii)} \end{cases} \asymp 1, \tag{F.2}$$

implying that the ratio $r(n, \beta_{i-1})/r(n, \beta_i)$ is less than any power of n. □

Lemma 15 (Uniform expansion on the oracle scale)**.** *Let c_0 and C_0 be two aboslute positve constants. Uniformly over $h \in [\mathsf{c}_0 h_n, \mathsf{C}_0 h_n]$,*

$$\tau(\widehat{f}_h) - \tau(f^\star) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{t}(\boldsymbol{x}_i)\varepsilon_i + \mathcal{O}_P(h^{\beta-\zeta}) + \mathcal{O}_P\big(h^{2\beta-2\zeta} + (nh^{d+2\zeta})^{-1}\big) + o_P(n^{-1/2}). \tag{F.3}$$

Further, we have

$$\sup_{f^\star \in \mathcal{W}^{\beta,2}} \mathbb{E}\big(\tau(\widehat{f}_h) - \tau(f^\star)\big)^2 \lesssim n^{-1} \vee h^{2\beta-2\zeta} \vee (nh^{d+2\zeta})^{-1}. \tag{F.4}$$

Proof. Assumption (SF) yields the following expansion

$$\tau(\widehat{f}_h) - \tau(f^\star) = T_{f^\star}(\widehat{f}_h - f^\star) + \mathcal{O}_P\big(\|\widehat{f}_h - f^\star\|_{\zeta,2}^2\big). \tag{F.5}$$

Standard kernel bounds yield

$$\|\mathbb{E}_\varepsilon(\widehat{f}_h) - f^\star\|_{\zeta,2} \lesssim h^{\beta-\zeta}, \quad \|\widehat{f}_h - \mathbb{E}_\varepsilon \widehat{f}_h\|_{\zeta,2} = \mathcal{O}_P\big((nh^{d+2\zeta})^{-1/2}\big). \tag{F.6}$$

Thus the quadratic remainder is of the order $\mathcal{O}_P\big(h^{2\beta-2\zeta} + (nh^{d+2\zeta})^{-1}\big)$ and the bias term

$$T_{f^\star}(\mathbb{E}_\varepsilon \widehat{f}_h - f^\star) = O_P(h^{\beta-\zeta}). \tag{F.7}$$

By defining $\overline{P}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n K_h(\boldsymbol{x} - \boldsymbol{x}_i), \overline{R}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n K_h(\boldsymbol{x} - \boldsymbol{x}_i)\varepsilon_i$, we have $\widehat{f}_h - \mathbb{E}_\varepsilon \widehat{f}_h = \frac{\overline{R}_h}{\overline{P}_h}$. Let $\delta_h(\boldsymbol{x}) := \frac{p_X(\boldsymbol{x})}{\overline{P}_h(\boldsymbol{x})} - 1$ and $\|\delta_h\|_\infty = o_P(1)$.

Then, we decompose the linear term into two terms

$$T_{f^\star}(\widehat{f}_h - \mathbb{E}_\varepsilon \widehat{f}_h) = \int t(\boldsymbol{x})\overline{R}_h(\boldsymbol{x})\,dx + \int t(\boldsymbol{x})\delta_h(\boldsymbol{x})\overline{R}_h(\boldsymbol{x})\,dx \tag{F.8}$$

$$=: \mathsf{L}_{1,h} + \mathsf{L}_{2,h}. \tag{F.9}$$

For $\mathsf{L}_{1,h}$, we have

$$\mathsf{L}_{1,h} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{t}_h(\boldsymbol{x}_i)\varepsilon_i, \quad \boldsymbol{t}_h(\boldsymbol{u}) := \int \boldsymbol{t}(\boldsymbol{x})K_h(\boldsymbol{x}-\boldsymbol{u})\mathrm{d}\boldsymbol{x}. \tag{F.10}$$

Since $\boldsymbol{t}_h \to \boldsymbol{t}$ in $L^2(P_X)$, as $h \to 0$, we obtain that $\mathsf{L}_{1,h} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i)\varepsilon_i + o_P(n^{-1/2})$.

For $\mathsf{L}_{2,h}$, define $s_h := K_h * (t\delta_h)$. Then, we write the second term $\mathsf{L}_{2,h} = \frac{1}{n}\sum_{i=1}^{n} s_h(\boldsymbol{x}_i)\varepsilon_i$, and

$$\mathsf{Var}(\sqrt{n}L_{2,h} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \lesssim \|s_h\|^2_{L^2(P_X)} = o_P(1), \tag{F.11}$$

so the second term is asymptotically negligible $\mathsf{L}_{2,h} = o_P(n^{-1/2})$. Combining all terms yields the result. □

Proof of Theorem 2.

CLT part. We want to show that, if $\beta > 3\zeta + d$, then

$$\sqrt{n}\big(\tau_{\widehat{h}} - \tau(f^\star)\big) \Rightarrow \mathcal{N}\big(0, \sigma^2\mathbb{E}[\boldsymbol{t}^2(\boldsymbol{X})]\big). \tag{F.12}$$

From Lemma 15, the following asymptotically linear expansion holds

$$\tau_{\widehat{h}} - \tau(f^\star) = \frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{x}_i)\varepsilon_i + r_n, \tag{F.13}$$

with $r_n = o_P(n^{-1/2})$ iff $\beta > 3\zeta + d$. Thus, $\sqrt{n}(\tau_{\widehat{h}} - \tau(f^\star)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} t(X_i)\varepsilon_i + o_P(1)$ leads, by Lindeberg-Feller central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} t(\boldsymbol{x}_i)\varepsilon_i \xrightarrow{d} \mathcal{N}(0, \sigma^2\mathbb{E}[\boldsymbol{t}^2(\boldsymbol{X})]). \tag{F.14}$$

Adaptivity part. Let $\beta_1 < \cdots < \beta_N$ with spacing $\Delta = (\log n)^{-1}$. Define

$$\mathbf{r}(n,\beta) = M\Big(n^{-1/2} \vee n^{-(\beta-\zeta)/(\beta+\zeta+d)}\Big)\log n. \tag{F.15}$$

The Lepski index is defined as $\widehat{j} = \max\Big\{j : \ |\widehat{\tau}_i - \widehat{\tau}_j| \le 4\mathbf{r}(n,\beta), \ \forall i < j\Big\}$. We want to show that, for $\beta \in [\beta_{\min}, \beta_{\max}]$,

$$\sup_{f^\star \in W^{\beta,2}} \mathbb{E}\big(\widehat{\tau}_{\widehat{j}} - \tau(f^\star)\big)^2 \lesssim n^{-1} \vee n^{-2(\beta-\zeta)/(\beta+\zeta+d)} \cdot \mathrm{poly}(\log n). \tag{F.16}$$

Let k satisfy $\beta_k \le \beta < \beta_{k+1}$. By defining $\mathcal{G}_k = \bigcap_{j\le k}\{|\widehat{\tau}_j - \tau(f^\star)| \le r_n(\beta_j)\}$, we have $\mathbb{P}(\mathcal{G}_k^c) \lesssim (\log n)n^{-A}$. On $\mathcal{G}_k$, by Proposition 13, one shows $\widehat{j} \ge k$ and $|\widehat{\tau}_{\widehat{j}} - \tau(f^\star)| \le 5\mathbf{r}(n,\beta_k)$. Therefore, we obtain $\mathbb{E}(\widehat{\tau}_{\widehat{j}} - \tau(f^\star))^2 \lesssim \mathbf{r}(n,\beta_k)^2$, which yields the result. □

G Proof of Theorem 3

Throughout the analysis, we denote L_φ as the Lipschitz constant of the nonlinear function φ. For example, if $\varphi = \mathsf{ReLU}$, then we have $L_\varphi = 1$.

Lemma 16 (Conditional sub-Gaussian linear forms)**.** *Fix the feature draw* $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p)$*. For any* $\boldsymbol{u} \in \mathbb{R}^p$*,*

$$\Big\|\big\langle \boldsymbol{u}, \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}) - \boldsymbol{\mu}_{\boldsymbol{\Omega}}\big\rangle\Big\|_{\psi_2} \le CK_X L_\varphi \frac{\|\boldsymbol{W}_{\boldsymbol{\Omega}}\boldsymbol{u}\|_2}{\sqrt{p}}, \tag{G.1}$$

where $\boldsymbol{W}_{\boldsymbol{\Omega}} = [\boldsymbol{\omega}_1 \ \ldots \ \boldsymbol{\omega}_p] \in \mathbb{R}^{d\times p}$ *and* $\boldsymbol{\mu}_{\boldsymbol{\Omega}} = \mathbb{E}(\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}))$.

Proof. For $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$,

$$\begin{aligned}\left|\langle \boldsymbol{u},\, \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{x}) - \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{x}')\rangle\right| &= \frac{1}{\sqrt{p}}\left|\sum_{j=1}^{p} u_j\big(\varphi(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\omega}_j) - \varphi(\boldsymbol{x}'^{\mathsf{T}}\boldsymbol{\omega}_j)\big)\right| \\ &\le \frac{L_\varphi}{\sqrt{p}}\left|\left\langle \boldsymbol{x}-\boldsymbol{x}',\, \sum_{j=1}^{p} u_j\boldsymbol{\omega}_j\right\rangle\right| = \frac{L_\varphi}{\sqrt{p}}\|\boldsymbol{W}_{\boldsymbol{\Omega}}\boldsymbol{u}\|_2\|\boldsymbol{x}-\boldsymbol{x}'\|_2.\end{aligned}$$

Thus $\boldsymbol{x} \mapsto \langle \boldsymbol{u}, \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{x})\rangle$ is L-Lipschitz with $L = L_\varphi\|\boldsymbol{W}_{\boldsymbol{\Omega}}\boldsymbol{u}\|_2/\sqrt{p}$. Since $\boldsymbol{X}$ is sub-Gaussian with parameter K_X, concentration for Lipschitz functionals yields

$$\Big\|\langle \boldsymbol{u},\, \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}) - \boldsymbol{\mu}_{\boldsymbol{\Omega}}\rangle\Big\|_{\psi_2} \le CK_X L_\varphi \frac{\|\boldsymbol{W}_{\boldsymbol{\Omega}}\boldsymbol{u}\|_2}{\sqrt{p}}.$$

□

Corollary 17 (Conditional second-moment bounds)**.** *Let* $\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} = \mathbb{E}\big(\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X})\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X})^{\mathsf{T}}\big)$ *and* $C_{\boldsymbol{\Omega}} = \mathrm{Cov}(\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}))$. *On any event where* $\frac{1}{p}\sum_{j=1}^{p}\|\boldsymbol{\omega}_j\|_2^2 \le M_\omega$,

$$\mathrm{Tr}\,\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} \le 2\varphi(0)^2 + 2L_\varphi^2\|\boldsymbol{\Sigma}_X\|M_\omega, \qquad \|C_{\boldsymbol{\Omega}}\| \le CL_\varphi^2K_X^2M_\omega. \tag{G.2}$$

Proof. Using $|\varphi(t)| \le L_\varphi|t| + |\varphi(0)|$,

$$\mathrm{Tr}\,\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} = (\|\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X})\|_2^2) = \frac{1}{p}\sum_{j=1}^{p}(\varphi(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\omega}_j)^2) \le \frac{2L_\varphi^2}{p}\sum_{j=1}^{p}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\omega}_j)^2 + 2\varphi(0)^2 \le 2L_\varphi^2\|\boldsymbol{\Sigma}_X\|M_\omega + 2\varphi(0)^2.$$

For $\|C_{\boldsymbol{\Omega}}\| = \sup_{\|\boldsymbol{u}\|=1}\mathrm{Var}(\langle \boldsymbol{u}, \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X})\rangle)$, apply Lemma 16 with $\|\boldsymbol{u}\| = 1$ and $\|\boldsymbol{W}_{\boldsymbol{\Omega}}\boldsymbol{u}\|_2^2 \le \sum_j\|\boldsymbol{\omega}_j\|_2^2$. □

Then apply standard Bernstein inequality to $\|\boldsymbol{\omega}_j\|_2^2$ and a union bound to $\max_j\|\boldsymbol{\omega}_j\|_2$ using sub-Gaussian tails.

Lemma 18. *There exists an event* $\mathcal{G}_p$ *with* $\mathbb{P}_{\boldsymbol{\Omega}}(\mathcal{G}_p) \ge 1 - e^{-cp}$ *such that*

$$\frac{1}{p}\sum_{j=1}^{p}\|\boldsymbol{\omega}_j\|_2^2 \le 2\,(\|\boldsymbol{\omega}\|_2^2), \qquad \max_{1\le j\le p}\|\boldsymbol{\omega}_j\|_2 \le CK_\omega\big(\sqrt{d} + \sqrt{2\log p}\big). \tag{G.3}$$

On $\mathcal{G}_p$, *Corollary 17 gives uniform constants* C_Σ, C'_Σ *with* $\|\Sigma_\Omega\| \le C_\Sigma$ *and* $\mathsf{Tr}\Sigma_\Omega \le C'_\Sigma$.

Lemma 19. *Let*

$$\mathcal{H}_{n,\boldsymbol{\Omega}} := \left\{\max_{1\le i\le n}\|\boldsymbol{X}_i\|_2^2 \le CK_X^2(d + \log n)\right\}. \tag{G.4}$$

Then $\mathbb{P}(\mathcal{H}_{n,\boldsymbol{\Omega}} \mid \boldsymbol{\Omega}) \ge 1 - 2n^{-4}$ *and on* $\mathcal{H}_{n,\boldsymbol{\Omega}}$,

$$\max_{1\le i\le n}\|\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}_i)\|_2^2 \le B_{n,\boldsymbol{\Omega}}^2 := 2\varphi(0)^2 + 2L_\varphi^2\left(\frac{1}{p}\sum_{j=1}^{p}\|\boldsymbol{\omega}_j\|_2^2\right)\max_{i\le n}\|\boldsymbol{X}_i\|_2^2. \tag{G.5}$$

Proof. From $|\varphi(t)| \le L_\varphi|t| + |\varphi(0)|$ and Cauchy–Schwarz inequality,

$$\|\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}_i)\|_2^2 = \frac{1}{p}\sum_{j=1}^{p}\varphi(\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\omega}_j)^2 \le \frac{2L_\varphi^2}{p}\sum_{j=1}^{p}(\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\omega}_j)^2 + 2\varphi(0)^2 \le 2L_\varphi^2\left(\frac{1}{p}\sum_{j=1}^{p}\|\boldsymbol{\omega}_j\|_2^2\right)\|\boldsymbol{X}_i\|_2^2 + 2\varphi(0)^2.$$

The probability bound for $\mathcal{H}_{n,\boldsymbol{\Omega}}$ is standard for sub-Gaussian maxima. □

Proposition 20. *Assume $m \geq CR\log(pn)$. For any $\varepsilon \in (0,1)$,*

$$(1-\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} \preceq \boldsymbol{S}_{\boldsymbol{\Omega}} \preceq (1+\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}$$

with $\mathbb{P}_{\boldsymbol{\Omega}}$-probability at least $1 - 2p\,e^{-c\varepsilon^2 m/R} - 2n^{-4}$.

Proof. Let

$$\boldsymbol{S}_{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}_{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \boldsymbol{\Delta}_i, \qquad \boldsymbol{\Delta}_i := w_i\big(\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}_i)\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}_i)^{\mathsf{T}} - \boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\big),$$

for fixed $\boldsymbol{\Omega}$, the $\boldsymbol{X}_i$ are independent, mean-zero, self-adjoint. On $\mathcal{H}_{n,\boldsymbol{\Omega}}$,

$$\|\boldsymbol{\Delta}_i\| \leq w_i\big(\|\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}_i)\|_2^2 + \|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|\big) \leq \frac{R}{m}\big(B_{n,\boldsymbol{\Omega}}^2 + \|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|\big) =: L_{\boldsymbol{\Omega}}.$$

For the variance proxy,

$$\boldsymbol{V}_{\boldsymbol{\Omega}} := \sum_{i=1}^{n}(\boldsymbol{X}_i^2) = \left(\sum_{i=1}^{n} w_i^2\right)((\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}} - \boldsymbol{\Sigma}_{\boldsymbol{\Omega}})^2).$$

By sub-Gaussian fourth-moment comparison,

$$\|\ (\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}} - \boldsymbol{\Sigma}_{\boldsymbol{\Omega}})^2\ \| \leq C\big(\operatorname{Tr}\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\| + \|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|^2\big) \leq C''\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|^2,$$

and since $\sum_i w_i^2 \leq R/m$,

$$\|\boldsymbol{V}_{\boldsymbol{\Omega}}\| \leq C''\frac{R}{m}\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|^2 =: v_{\boldsymbol{\Omega}}.$$

Matrix Bernstein inequality with $t = \varepsilon\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|$ gives

$$\mathbb{P}_{\boldsymbol{\Omega}}\big(\|\boldsymbol{S}_{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\| \geq \varepsilon\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|\big) \leq 2p\exp\left(-\frac{\varepsilon^2\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|^2/2}{v_{\boldsymbol{\Omega}} + (\varepsilon\|\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\|/3)L_{\boldsymbol{\Omega}}}\right) \leq 2p\,e^{-c\varepsilon^2 m/R} + 2n^{-4}.$$

□

Lemma 21. *Let $\boldsymbol{A}_{\boldsymbol{\Omega}} = \boldsymbol{Z}_{-i}^{\mathsf{T}}\boldsymbol{Z}_{-i}$ remove all copies of family i from the training fold and $\boldsymbol{S}_{i,\boldsymbol{\Omega}} = \boldsymbol{A}_{\boldsymbol{\Omega}}/(m - r_i)$. On $\mathcal{H}_{n,\boldsymbol{\Omega}}$ and the event in Proposition 20,*

$$(1-\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} - \gamma_{n,\boldsymbol{\Omega}} I \preceq \boldsymbol{S}_{i,\boldsymbol{\Omega}} \preceq (1+\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} + \gamma_{n,\boldsymbol{\Omega}} I,$$

where $\gamma_{n,\boldsymbol{\Omega}} = \frac{2r_i}{m}B_{n,\boldsymbol{\Omega}}^2$.

Proof. We write

$$\boldsymbol{S}_{i,\boldsymbol{\Omega}} - \boldsymbol{S}_{\boldsymbol{\Omega}} = \frac{r_i}{m(m-r_i)}\sum_{j\neq i} c_j \boldsymbol{z}_j\boldsymbol{z}_j^{\mathsf{T}} - \frac{r_i}{m}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}.$$

On $\mathcal{H}_{n,\boldsymbol{\Omega}}$, $\|\boldsymbol{S}_{i,\boldsymbol{\Omega}} - \boldsymbol{S}_{\boldsymbol{\Omega}}\| \leq \frac{2r_i}{m}B_{n,\boldsymbol{\Omega}}^2$. Combine with $(1\pm\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} \preceq \boldsymbol{S}_{\boldsymbol{\Omega}}$ from Proposition 20. □

Lemma 22 (Non-centered Hanson–Wright)**.** *Let $\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \boldsymbol{\mu}_{\boldsymbol{\Omega}} + \boldsymbol{w}$ with $\|\langle \boldsymbol{u}, \boldsymbol{w}\rangle\|_{\psi_2} \leq K\sqrt{\boldsymbol{u}^{\mathsf{T}}C_{\boldsymbol{\Omega}}\boldsymbol{u}}$ and $\boldsymbol{B} \succeq 0$. Then for any $t > 0$,*

$$\mathbb{P}_{\boldsymbol{\Omega}}\Big(|\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X})^{\mathsf{T}}\boldsymbol{B}\boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}) - \operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B})| \geq t\Big)$$
$$\leq 4\exp\left[-c\min\left(\frac{t^2}{K^4(\|C_{\boldsymbol{\Omega}}^{1/2}\boldsymbol{B}C_{\boldsymbol{\Omega}}^{1/2}\|_F^2 + \|C_{\boldsymbol{\Omega}}\|\|\boldsymbol{B}\|\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B}))}, \frac{t}{K^2(\|C_{\boldsymbol{\Omega}}^{1/2}\boldsymbol{B}C_{\boldsymbol{\Omega}}^{1/2}\| + \sqrt{\|C_{\boldsymbol{\Omega}}\|\|\boldsymbol{B}\|\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B})})}\right)\right].$$

Proof. Decompose $\boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z} - (\boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z}) = ((\boldsymbol{w}^\top \boldsymbol{B}\boldsymbol{w}) - (\boldsymbol{w}^\top \boldsymbol{B}\boldsymbol{w})) + 2\boldsymbol{\mu}_{\boldsymbol{\Omega}}^\top \boldsymbol{B}\boldsymbol{w}$ and apply centered HW plus a sub-Gaussian tail. For $\boldsymbol{B} \succeq 0$, $\boldsymbol{\mu}_{\boldsymbol{\Omega}}^\top \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{\Omega}} \le \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B})$. □

Let $u = t/\delta$. Consider $g(u) = \frac{u^{2r}}{(au+1)^2}$. For $r \in [0,1]$, $\sup_{u\ge 0} g(u) = C_{a,r} < \infty$ attained at $u^\star = \frac{r}{a(1-r)}$. Substitute back $t = \delta u$. We obtain the following result:

Lemma 23. *Fix $r \in [0,1]$ and $a \in (0,1]$. For all $t \ge 0$ and $\delta > 0$,*

$$\left(\frac{\delta}{at+\delta}\right)^2 t^{2r} \le C_{a,r}\,\delta^{2r}.$$

Lemma 24. *Let $\boldsymbol{A}_{\boldsymbol{\Omega}} = \boldsymbol{Z}_{-i}^\top \boldsymbol{Z}_{-i}$, $\boldsymbol{B}_{\boldsymbol{\Omega}} = \boldsymbol{A}_{\boldsymbol{\Omega}} + \lambda I$, $\boldsymbol{z}_i = \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{X}_i)$, and*

$$u_i(\lambda) := \boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i, \qquad \tilde{f}_{i,\lambda}(x) := \boldsymbol{z}_{\boldsymbol{\Omega}}(x)^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{Z}_{-i}^\top \boldsymbol{y}_{-i}.$$

If r_i copies of family i are in the training fold used to predict at $\boldsymbol{X}_i$, then

$$\big(Y_i - \hat{f}_\lambda(\boldsymbol{X}_i)\big)^2 = \frac{\big(Y_i - \tilde{f}_{i,\lambda}(\boldsymbol{X}_i)\big)^2}{\big(1 + r_i u_i(\lambda)\big)^2}.$$

Proof. In the full fold $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{A}_{\boldsymbol{\Omega}} + r_i\boldsymbol{z}_i\boldsymbol{z}_i^\top$ and $\boldsymbol{Z}^\top\boldsymbol{y} = \boldsymbol{Z}_{-i}^\top\boldsymbol{y}_{-i} + r_i\boldsymbol{z}_i Y_i$. With $\boldsymbol{G} := \boldsymbol{Z}^\top\boldsymbol{Z} + \lambda I = \boldsymbol{B}_{\boldsymbol{\Omega}} + r_i\boldsymbol{z}_i\boldsymbol{z}_i^\top$, Sherman–Morrison gives

$$\boldsymbol{G}^{-1} = \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1} - \frac{r_i \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}}{1 + r_i\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i} = \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1} - \frac{r_i \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}}{1 + r_i u_i}.$$

Then

$$\begin{aligned}\hat{\boldsymbol{a}}_\lambda &= \boldsymbol{G}^{-1}\boldsymbol{Z}^\top\boldsymbol{y} \\ &= \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{Z}_{-i}^\top\boldsymbol{y}_{-i} + r_i\boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i Y_i - \frac{r_i \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{Z}_{-i}^\top\boldsymbol{y}_{-i}}{1 + r_i u_i} - \frac{r_i^2 \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i Y_i}{1 + r_i u_i}.\end{aligned}$$

Evaluating at $\boldsymbol{X}_i$ and using $\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{Z}_{-i}^\top\boldsymbol{y}_{-i} = \tilde{f}_{i,\lambda}(\boldsymbol{X}_i)$ and $\boldsymbol{z}_i^\top \boldsymbol{B}_{\boldsymbol{\Omega}}^{-1}\boldsymbol{z}_i = u_i$,

$$\hat{f}_\lambda(\boldsymbol{X}_i) = \frac{1}{1 + r_i u_i}\tilde{f}_{i,\lambda}(\boldsymbol{X}_i) + \frac{r_i u_i}{1 + r_i u_i} Y_i,$$

so we obtain that

$$Y_i - \hat{f}_\lambda(\boldsymbol{X}_i) = \frac{Y_i - \tilde{f}_{i,\lambda}(\boldsymbol{X}_i)}{1 + r_i u_i}.$$

□

Proposition 25. *Let $\boldsymbol{S}_{i,\boldsymbol{\Omega}} = \boldsymbol{A}_{\boldsymbol{\Omega}}/(m - r_i)$ and $\tau_i = \lambda/(m - r_i)$. Then*

$$u_i(\lambda) = \frac{1}{m - r_i}\,\boldsymbol{z}_i^\top (\boldsymbol{S}_{i,\boldsymbol{\Omega}} + \tau_i I)^{-1}\boldsymbol{z}_i,$$

and there exist constants $0 < c < C < \infty$ such that with $\mathbb{P}_{\boldsymbol{\Omega}}$-probability at least $1 - \tilde{C}n^{-3}$ (uniformly over replicated families i),

$$c\,\frac{d_\lambda(\boldsymbol{\Omega})}{m} \le u_i(\lambda) \le C\,\frac{d_\lambda(\boldsymbol{\Omega})}{m},$$

where $d_\lambda(\boldsymbol{\Omega}) = \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} + \frac{\lambda}{m} I)^{-1})$.

Proof. For $\boldsymbol{B} = (\boldsymbol{S}_{i,\boldsymbol{\Omega}} + \tau_i I)^{-1}$, $\|\boldsymbol{B}\| \le 1/\tau_i$. Lemma 21 implies

$$\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B}\right) \in \left[\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\big((1+\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} + \tau_i + \gamma_{n,\boldsymbol{\Omega}} I\big)^{-1}\right), \operatorname{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\big((1-\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} + \tau_i - \gamma_{n,\boldsymbol{\Omega}} I\big)^{-1}\right)\right].$$

Since $\tau_i = \lambda/(m - r_i) = \Theta(\lambda/m)$ and $\gamma_{n,\boldsymbol{\Omega}} = \mathcal{O}((R/m)B^2_{n,\boldsymbol{\Omega}})$, both endpoints are $\Theta(d_\lambda(\boldsymbol{\Omega}))$. Next apply Lemma 22 with

$$\|C_{\boldsymbol{\Omega}}^{1/2}\boldsymbol{B}C_{\boldsymbol{\Omega}}^{1/2}\|_F^2 = \operatorname{Tr}(C_{\boldsymbol{\Omega}}\boldsymbol{B}C_{\boldsymbol{\Omega}}\boldsymbol{B}) \le \|C_{\boldsymbol{\Omega}}\|\operatorname{Tr}(C_{\boldsymbol{\Omega}}\boldsymbol{B}^2) \le \frac{\|C_{\boldsymbol{\Omega}}\|}{\tau_i}\operatorname{Tr}(C_{\boldsymbol{\Omega}}\boldsymbol{B}) \le \frac{1}{\tau_i}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B}),$$

which yields, for $\eta = \sqrt{3\log n/\tau_i} \wedge \frac{1}{2}$,

$$\mathbb{P}_{\boldsymbol{\Omega}}\left(\left|\boldsymbol{z}_i^\top\boldsymbol{B}\boldsymbol{z}_i - \operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B})\right| \ge \eta\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B})\right) \le 4n^{-3}.$$

Divide by $m - r_i = \Theta(m)$ and use the trace comparison to get

$$u_i(\lambda) = \frac{1}{m-r_i}\boldsymbol{z}_i^\top\boldsymbol{B}\boldsymbol{z}_i = \frac{\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B})}{m-r_i} \pm C\sqrt{\frac{d_\lambda(\boldsymbol{\Omega})\log n}{m^2\tau_i}} = \frac{d_\lambda(\boldsymbol{\Omega})}{m} \pm C\sqrt{\frac{d_\lambda(\boldsymbol{\Omega})\log n}{m\lambda}}.$$

For $m \gtrsim \log n$ and a finite λ grid, absorb the deviation into constants and take a union bound over families. □

Proposition 26. *Let $\boldsymbol{S}_{\boldsymbol{\Omega}} = (1/m)\boldsymbol{Z}^\top\boldsymbol{Z}$, $\delta = \lambda/m$, and*

$$\mathsf{Risk}_m(\lambda \mid \boldsymbol{\Omega}) := ((Y - \hat{f}_\lambda(\boldsymbol{X}))^2 \mid \boldsymbol{Z}, \boldsymbol{\Omega}).$$

Under the source–capacity assumption (SC),

$$\mathsf{Risk}_m(\lambda \mid \boldsymbol{\Omega}) \le \boldsymbol{\Sigma}^2 + C\,\delta^{2r} + C\,\frac{d_\lambda(\boldsymbol{\Omega})}{m}.$$

Proof. Write $\boldsymbol{y} = f^\star(\boldsymbol{X}_{1:m}) + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon}\mathcal{N} \sim (0, \boldsymbol{\sigma}^2\mathbf{I}_)$ and $\boldsymbol{B} = (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda I)^{-1} = \frac{1}{m}(\boldsymbol{S}_{\boldsymbol{\Omega}} + \delta I)^{-1}$. Then we decompose

$$\hat{f}_\lambda(\boldsymbol{x}) = \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{x})^\top\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{y} = \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{x})^\top\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{Z}\beta^\star + \boldsymbol{z}_{\boldsymbol{\Omega}}(\boldsymbol{x})^\top\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{\varepsilon} =: \hat{f}_{\lambda,0}(\boldsymbol{x}) + \hat{\eta}_\lambda(\boldsymbol{x}).$$

(I) Since $\beta^\star - \hat{\boldsymbol{a}}_{\lambda,0} = \lambda\boldsymbol{B}\beta^\star$,

$$(\,(f^\star(\boldsymbol{X}) - \hat{f}_{\lambda,0}(\boldsymbol{X}))^2 \mid \boldsymbol{Z}, \boldsymbol{\Omega}) = \lambda^2\beta^{\star\top}\boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B}\beta^\star = \delta^2\beta^{\star\top}(\boldsymbol{S}_{\boldsymbol{\Omega}} + \delta I)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}(\boldsymbol{S}_{\boldsymbol{\Omega}} + \delta I)^{-1}\beta^\star.$$

By Proposition 20, $(\boldsymbol{S}_{\boldsymbol{\Omega}} + \delta I)^{-1} \preceq ((1-\varepsilon)\boldsymbol{\Sigma}_{\boldsymbol{\Omega}} + \delta I)^{-1}$. Diagonalize $\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}$: $\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{v}_k = \xi_k^2\boldsymbol{v}_k$ and $\beta^\star = \boldsymbol{\Sigma}_{\boldsymbol{\Omega}}^r\boldsymbol{v}$. Then

$$\mathsf{Bias}^2 \le \sum_k \left(\frac{\delta}{(1-\varepsilon)\xi_k^2 + \delta}\right)^2 \xi_k^{4r} v_k^2 \le C\delta^{2r}\sum_k v_k^2$$

by Lemma 23.

(II) Since $\boldsymbol{B}^{-1} = \boldsymbol{Z}^\top\boldsymbol{Z} + \lambda I \succeq \boldsymbol{Z}^\top\boldsymbol{Z}$,

$$\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{B} \preceq \boldsymbol{B}.$$

Hence

$$(\hat{\eta}_\lambda(\boldsymbol{X})^2 \mid \boldsymbol{Z}, \boldsymbol{\Omega}) = \boldsymbol{\Sigma}^2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{B}) \le \boldsymbol{\Sigma}^2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\boldsymbol{B}) = \frac{\boldsymbol{\Sigma}^2}{m}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\Omega}}(\boldsymbol{S}_{\boldsymbol{\Omega}} + \delta I)^{-1}\right) \asymp \frac{\boldsymbol{\Sigma}^2}{m}d_\lambda(\boldsymbol{\Omega}).$$

Adding $\boldsymbol{\Sigma}^2$ gives the claim. □

Lemma 27. *Fix the folds. For each replicated family i, let $\tilde{f}^{(1)}_{i,\lambda}$ and $\tilde{f}^{(2)}_{i,\lambda}$ be trained on $\mathcal{D}_1 \setminus \{\text{copies of } i\}$ and $\mathcal{D}_2 \setminus \{\text{copies of } i\}$, respectively. Then, conditional on the folds and $\mathbf{\Omega}$,*

$$S^{(1)}_{i,\lambda} := \big(Y_i - \tilde{f}^{(1)}_{i,\lambda}(\boldsymbol{X}_i)\big)^2, \qquad S^{(2)}_{i,\lambda} := \big(Y_i - \tilde{f}^{(2)}_{i,\lambda}(\boldsymbol{X}_i)\big)^2$$

are independent across families i, and

$$\big(S^{(\ell)}_{i,\lambda} \mid \textit{folds}, \mathbf{\Omega}\big) \;\le\; \mathsf{Risk}_m(\lambda \mid \mathbf{\Omega}), \qquad \ell \in \{1,2\}.$$

Proof. Removing family i ensures $\tilde{f}^{(\ell)}_{i,\lambda}$ is a functional of the training data *excluding* $(\boldsymbol{X}_i, Y_i)$, hence $\tilde{f}^{(\ell)}_{i,\lambda} \perp (\boldsymbol{X}_i, Y_i)$ conditional on the folds and $\mathbf{\Omega}$. Across families, $(\boldsymbol{X}_i, Y_i)$ are i.i.d., so $\{S^{(\ell)}_{i,\lambda}\}_i$ are independent. The mean bound follows from Proposition 26:

$$\big(S^{(\ell)}_{i,\lambda} \mid \text{folds}, \mathbf{\Omega}\big) = \big((Y_i - \tilde{f}^{(\ell)}_{i,\lambda}(\boldsymbol{X}_i))^2 \mid \text{folds}, \mathbf{\Omega}\big) \le \mathsf{Risk}_m(\lambda \mid \mathbf{\Omega}).$$

□

Lemma 28 (Weighted empirical Bernstein)**.** *Let $U_1, \dots, U_N$ be independent centered sub-exponential random variables with $\|U_j\|_{\psi_1} \le V$ and deterministic weights a_j. Then, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{j=1}^N a_j U_j\right| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{V^2\sum_{j=1}^N a_j^2}, \frac{t}{V\max_{1\le j\le N}|a_j|}\right\}\right).$$

Proposition 29 (Replicated CV block, conditional)**.** *Let $\mathcal{V}_{\text{rep}}$ be the set of validated copies of replicated families (size $\rho n R$). Define*

$$a^{(1)}_i = \frac{R - r_i}{2m}\frac{1}{\big(1 + r_i u^{(2)}_i(\lambda)\big)^2}, \qquad a^{(2)}_i = \frac{r_i}{2m}\frac{1}{\big(1 + (R - r_i) u^{(1)}_i(\lambda)\big)^2}.$$

With $\mathbb{P}_{\mathbf{\Omega}}$-probability at least $1 - \tilde{C}n^{-3}$,

$$\left|\frac{1}{2m}\sum_{j\in\mathcal{V}_{\text{rep}}}\big(Y_j - \hat{f}_\lambda(\boldsymbol{X}_j)\big)^2 - \left(\frac{m}{m + R\,d_\lambda(\mathbf{\Omega})}\right)^2 \mathsf{Risk}_m(\lambda \mid \mathbf{\Omega})\right| \le C\sqrt{\frac{\log n}{\rho n R}}\left(\sigma^2 + \left(\frac{\lambda}{m}\right)^{2r} + \frac{d_\lambda(\mathbf{\Omega})}{m}\right).$$

Proof. By Lemma 24 and grouping by family,

$$\frac{1}{2m}\sum_{j\in\mathcal{V}_{\text{rep}}}\big(Y_j - \hat{f}_\lambda(\boldsymbol{X}_j)\big)^2 = \sum_{i\in\mathcal{M}}\big[a^{(1)}_i S^{(2)}_{i,\lambda} + a^{(2)}_i S^{(1)}_{i,\lambda}\big].$$

On the event of Proposition 25, $u^{(\ell)}_i(\lambda) \asymp d_\lambda(\mathbf{\Omega})/m$ and $r_i \asymp R$, hence

$$a^{(1)}_i + a^{(2)}_i = \Theta\left(\frac{1}{2m}\left(\frac{m}{m + R\,d_\lambda(\mathbf{\Omega})}\right)^2\right)$$

uniformly in i. Summing over $|\mathcal{M}| = \rho n$ families and computing the squared-sum,

$$\sum_{i\in\mathcal{M}}(a^{(1)}_i + a^{(2)}_i) = \Theta\left(\left(\frac{m}{m + R\,d_\lambda(\mathbf{\Omega})}\right)^2\right),$$
$$\sum_{i\in\mathcal{M}}(a^{(1)}_i + a^{(2)}_i)^2 = \Theta\left(\frac{\rho n}{m^2}\left(\frac{m}{m + R\,d_\lambda(\mathbf{\Omega})}\right)^4\right).$$

By Lemma 27 and Proposition 26, the centered variables $S_{i,\lambda}^{(\ell)} - \big(S_{i,\lambda}^{(\ell)} \mid \text{folds}, \mathbf{\Omega}\big)$ are independent with ψ_1-norm bounded by

$$V(\lambda) := \sigma^2 + \left(\frac{\lambda}{m}\right)^{2r} + \frac{d_\lambda(\mathbf{\Omega})}{m}.$$

Apply Lemma 28 to the weighted sum to obtain the deviation of order

$$V(\lambda)\sqrt{\log n}\,\sqrt{\sum_{i\in\mathcal{M}}(a_i^{(1)}+a_i^{(2)})^2} \asymp V(\lambda)\sqrt{\frac{\log n}{\rho n R}},$$

which gives the stated bound. □

Proposition 30. *Let $\mathcal{V}_{\text{nrep}}$ be the validated non-replicated points (size $(1-\rho)n$). With $\mathbb{P}_{\mathbf{\Omega}}$-probability at least $1-\tilde{C}n^{-3}$,*

$$\left|\frac{1}{2m}\sum_{j\in\mathcal{V}_{\text{nrep}}}\big(Y_j - \hat{f}_\lambda(\boldsymbol{X}_j)\big)^2 - \frac{(1-\rho)n}{2m}\,\mathsf{Risk}_m(\lambda \mid \mathbf{\Omega})\right| \le C\sqrt{\frac{\log n}{m}}\left(\sigma^2 + \left(\frac{\lambda}{m}\right)^{2r} + \frac{d_\lambda(\mathbf{\Omega})}{m}\right).$$

Proof. Conditional on $\mathbf{\Omega}$ and the folds, the summands are i.i.d. with mean at most $\mathsf{Risk}_m(\lambda \mid \mathbf{\Omega})$ and sub-exponential norm $V(\lambda)$ from Proposition 26. Apply Lemma 28 with $N = (1-\rho)n$ terms of weight $1/(2m)$. □

Proposition 31. *Define $\kappa_{\mathbf{\Omega}}(\lambda) := \left(\frac{m}{m+R\,d_\lambda(\mathbf{\Omega})}\right)^2 + \frac{(1-\rho)n}{2m}$. With $\mathbb{P}_{\mathbf{\Omega}}$-probability at least $1-\tilde{C}n^{-3}$,*

$$\big|\mathcal{L}_{\text{CV}}(\lambda,\rho,R) - \kappa_{\mathbf{\Omega}}(\lambda)\,\mathsf{Risk}_m(\lambda \mid \mathbf{\Omega})\big| \le C\left(\sqrt{\frac{\log n}{m}} + \sqrt{\frac{\log n}{\rho n R}}\right)\left(\sigma^2 + \left(\frac{\lambda}{m}\right)^{2r} + \frac{d_\lambda(\mathbf{\Omega})}{m}\right).$$

Proof. Sum the deviations in Propositions 29 and 30 and use that $\kappa_{\mathbf{\Omega}}(\lambda) \in (0,1]$. □

Now we recall Theorem 3 as follow.

Theorem 32. *Assume $m \ge CR\log(pn)$ with $p^* \ll p \ll n$, for every $\lambda > 0$, with probability at least $1 - e^{-cp} - 2p\,e^{-cm/R} - \tilde{C}n^{-3}$, we have*

$$\begin{aligned}&\left|\mathcal{L}_{\text{CV}}(\lambda,\rho,R) - \left[\sigma^2 + C\left(\left(\frac{\lambda}{m}\right)^{2r} + \frac{d_\lambda}{m}\right) + C'\left(\frac{m}{m+R\,d_\lambda}\right)^2\right]\right| \\ &\le \left(\sqrt{\frac{\log n}{m}} + \sqrt{\frac{\log n}{\rho n R}}\right)\left(\sigma^2 + \left(\frac{\lambda}{m}\right)^{2r} + \frac{d_\lambda}{m}\right),\end{aligned}$$

where $d_\lambda = (\mathrm{Tr}(\Sigma_{\mathbf{\Omega}}(\Sigma_{\mathbf{\Omega}} + \frac{\lambda}{m}I)^{-1}))$ and the constants are uniform on the event $\mathcal{G}_p$.

Proof. On $\mathcal{G}_p$ (Lemma 18) all spectral constants are uniform in $\mathbf{\Omega}$. Condition on $\mathbf{\Omega} \in \mathcal{G}_p$ and apply Proposition 31 together with Proposition 26 to upper bound $\mathsf{Risk}_m(\lambda \mid \mathbf{\Omega})$. Remove conditioning via the tower property and multiply probabilities. Replace $d_\lambda(\mathbf{\Omega})$ by d_λ up to uniform constants and absorb $\kappa_{\mathbf{\Omega}}(\lambda) \in (0,1]$ into C, C'. □

H Supplemental Simulations

H.1 Implementation Details in Section 4

This section provides additional details and results complementing Section 4.

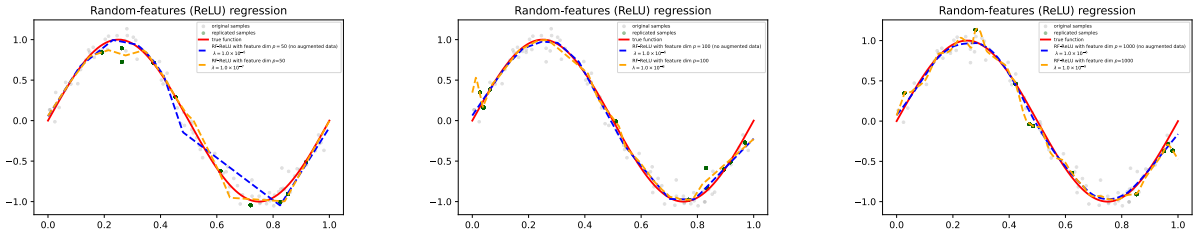


Figure 5: See Figure 7.

Across all panels, the estimator trained on augmented data is visibly more curvilinear, especially near regions of high curvature, reflecting deliberate undersmoothing and reduced bias. The effect persists over a wide range of p/n, indicating that the replication scheme robustly induces the desired bias–variance rebalancing even as the random-feature dimension grows.

NW estimator We begin by conducting simulations using the precise setup described in Theorem 1 and Theorem 2. We generate datasets with sample sizes $n \in \{100, 1000, 10000\}$, and consider noise levels $\sigma^2 \in \{0.1, 0.5, 2\}$. For each dataset, we draw i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \sim \text{Unif}([0,1])$, and the responses are generated as $Y_i = X_i^2 \sin(1/X_i^2) + \varepsilon_i$, where $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. We first perform a sanity check for Theorem 1 to verify that the bandwidth selected by two-fold cross-validation indeed decreases at a smaller order.

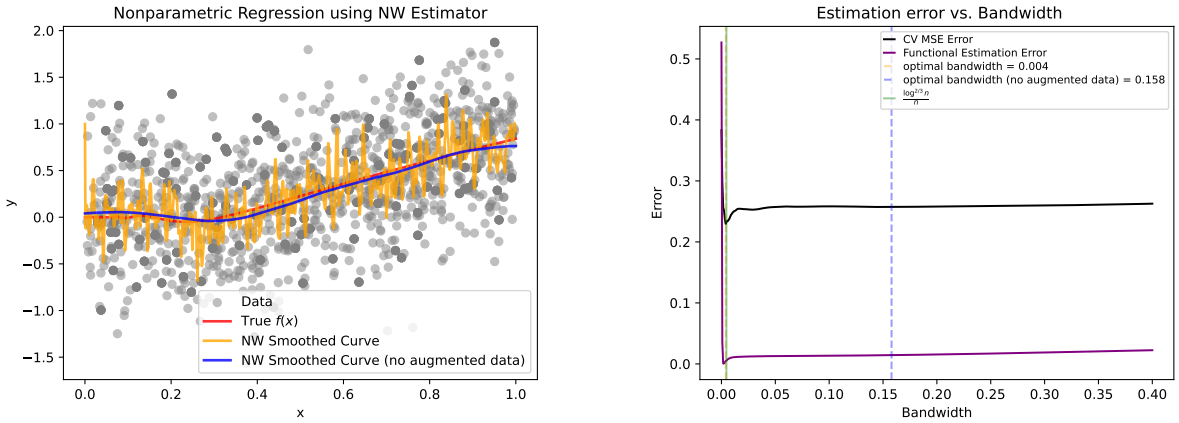


Figure 6: **(Left)**: Illustration of Nadaraya–Watson (NW) regression for nonparametric estimation. The gray dots represent the observed data, the red line indicates the true function $f(x)$, and the orange and blue curves show NW-smoothed estimates (with and without augmented data). **(Right)**: The plot shows the estimation error versus the bandwidth parameter. The vertical dashed lines mark the optimal bandwidth that minimizes the CV risk with and without the replicated data.

From the left subplot of Figure 6, it is evident that the fitted model is substantially more undersmoothed under cross-validation. This observation further corroborates our theoretical explanation that the replication

procedure leads to a smaller-order bandwidth, thereby reducing the first-order bias while maintaining balance with the variance term arising from the second-order residual, ultimately achieving a faster convergence rate for targeted smooth functionals.

Random Features Model In Section 3.2, we focus on the case where the number of random features stays in the regime $p = \mathcal{O}(n)$.

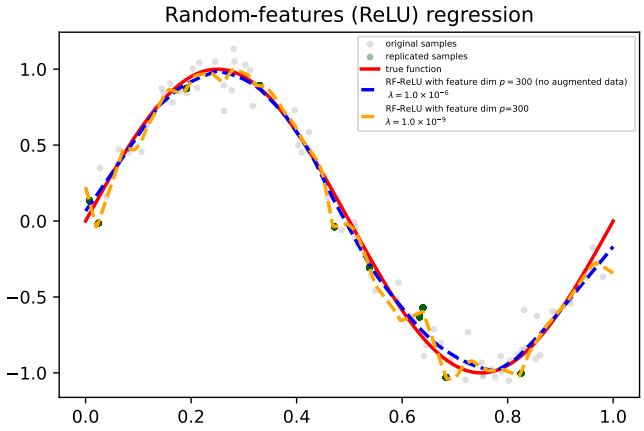


Figure 7: Representative comparison for a random-feature model with $\omega_i \in \mathbb{R}^{300}$ and ReLU activation. Green dots mark the replicated samples. The blue curve fits the original data $\mathcal{D}_{\mathsf{train}}$, while the yellow curve fits the augmented data $\mathcal{D}_{\mathsf{aug}}$. Parameters: n= 100, $\rho_n = 0.2$, $R_n = 20$.

Here we set the feature distribution as $\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{I}_d)$ or $\mathsf{Unif}(\mathbf{S}^{d-1})$. Compared to the baseline estimator trained on the original sample (blue), whose curves remain smooth, the estimator trained on augmented data (yellow) appears more fluctuant and nearly interpolates the replicated samples.

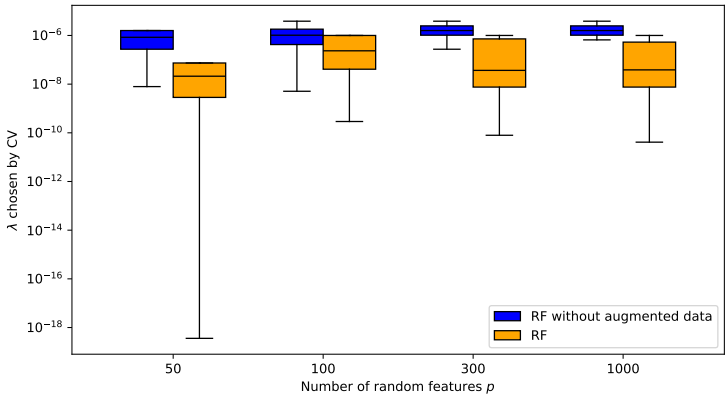


Figure 8: Boxplot of the cross-validated regularization parameter λ after applying the replication procedure. The λ axis is shown on a logarithmic scale.

Figure 8 reports the cross-validated regularization constants selected after replication. For each feature dimension p, we generate 100 independent data sets and plot the resulting λ values on a logarithmic scale. Across all settings, the replication scheme drives λ downward, indicating systematic undersmoothing. The shrinkage is more pronounced once the random-feature dimension exceeds a moderate threshold. This is precisely the regime in which the black-box model has adequate capacity, supporting the assumptions that

underlie our theory.

H.2 Kernel Ridge Regression

Analogous to Section H.1, we conduct synthetic-data experiments with kernel ridge regression (KRR) and observe the same undersmoothing effect induced by replication.

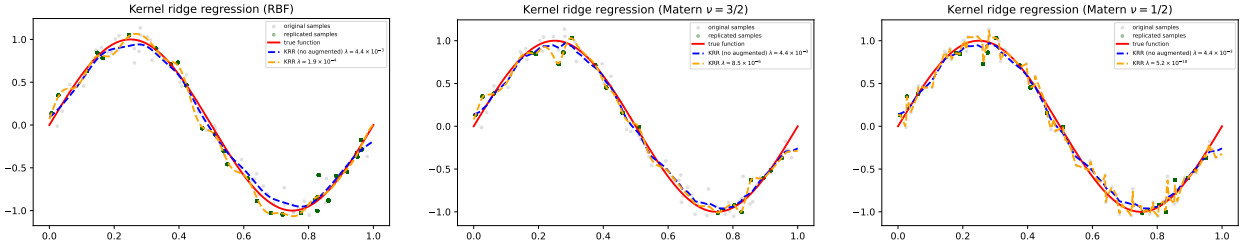


Figure 9: Instantiation of `Alg` as KRR with three kernels (left to right): $k_{\mathsf{RBF}}(x,x') = \exp(-\gamma\|x-x'\|_2^2)$, $k_{\mathsf{Matern},3/2}(x,x') = \big(1+\sqrt{3}\|x-x'\|_2/\ell\big)\exp\big(-\sqrt{3}\|x-x'\|_2/\ell\big)$, and $k_{\mathsf{Matern},1/2}(x,x') = \exp(-\|x-x'\|_2/\ell)$. Experimental settings match those of Figure 7.

H.3 Additional numerical results on estimation error

`Alg`	n	f_2		f_3	
		τ_1	τ_2	τ_1	τ_2
NW	100	$\mathbf{1.59\times10^{-2}}$	$\mathbf{7.82\times10^{-3}}$	$\mathbf{8.23\times10^{-3}}$	$\mathbf{1.29\times10^{-3}}$
		6.28×10^{-2}	4.02×10^{-2}	2.83×10^{-2}	2.52×10^{-2}
	1000	$\mathbf{5.34\times10^{-3}}$	$\mathbf{2.55\times10^{-3}}$	$\mathbf{2.35\times10^{-3}}$	$\mathbf{4.54\times10^{-3}}$
		3.56×10^{-2}	2.42×10^{-2}	2.73×10^{-2}	2.36×10^{-2}
	10000	$\mathbf{1.24\times10^{-3}}$	$\mathbf{6.96\times10^{-4}}$	$\mathbf{7.93\times10^{-4}}$	$\mathbf{5.95\times10^{-4}}$
		1.12×10^{-2}	1.09×10^{-2}	1.38×10^{-2}	1.06×10^{-2}
RF	100	$\mathbf{2.12\times10^{-2}}$	$\mathbf{6.80\times10^{-3}}$	$\mathbf{9.54\times10^{-3}}$	$\mathbf{1.40\times10^{-3}}$
		9.12×10^{-2}	4.80×10^{-2}	8.54×10^{-2}	2.26×10^{-2}
	1000	$\mathbf{4.77\times10^{-3}}$	$\mathbf{2.32\times10^{-3}}$	$\mathbf{1.95\times10^{-3}}$	$\mathbf{3.78\times10^{-3}}$
		2.55×10^{-2}	1.61×10^{-2}	1.91×10^{-2}	2.48×10^{-2}
	10000	$\mathbf{1.38\times10^{-3}}$	$\mathbf{6.37\times10^{-4}}$	$\mathbf{9.16\times10^{-4}}$	$\mathbf{6.33\times10^{-4}}$
		8.87×10^{-3}	5.70×10^{-3}	8.91×10^{-3}	8.22×10^{-3}
NN	100	$\mathbf{2.01\times10^{-2}}$	$\mathbf{7.92\times10^{-3}}$	$\mathbf{8.09\times10^{-3}}$	$\mathbf{1.51\times10^{-3}}$
		7.39×10^{-2}	4.89×10^{-2}	7.23×10^{-2}	2.08×10^{-2}
	1000	$\mathbf{5.56\times10^{-3}}$	$\mathbf{2.34\times10^{-3}}$	$\mathbf{2.25\times10^{-3}}$	$\mathbf{4.07\times10^{-3}}$
		1.85×10^{-2}	1.37×10^{-2}	1.48×10^{-2}	1.88×10^{-2}
	10000	$\mathbf{1.18\times10^{-3}}$	$\mathbf{6.41\times10^{-4}}$	$\mathbf{8.29\times10^{-4}}$	$\mathbf{5.46\times10^{-4}}$
		7.80×10^{-3}	4.48×10^{-3}	7.45×10^{-3}	9.47×10^{-3}
EL	100	$\mathbf{1.22\times10^{-2}}$	$\mathbf{6.52\times10^{-3}}$	$\mathbf{7.21\times10^{-3}}$	$\mathbf{1.19\times10^{-3}}$
	1000	$\mathbf{3.85\times10^{-3}}$	$\mathbf{2.06\times10^{-3}}$	$\mathbf{2.28\times10^{-3}}$	$\mathbf{3.77\times10^{-3}}$
	10000	$\mathbf{1.22\times10^{-3}}$	$\mathbf{6.52\times10^{-4}}$	$\mathbf{7.21\times10^{-4}}$	$\mathbf{1.19\times10^{-4}}$

Table 2: Estimation error $|\widehat{\tau}-\tau|$ under $\sigma = 0.1$ for $f_2(x) = 2-(e^{-x}+x^x)$ and $f_3(x) = 100x^3(1-x)^3(0.5\sin(10\pi x)+0.25\sin(50\pi x))$. See Table 1.

In both Table 1 and Table 2, each entry summarizes results over 100 Monte Carlo replicates, computed on a high-performance cluster using CPU nodes (4 cores per job).