Big Data and Information Analytics

https://www.aimspress.com/journal/BDIA

BDIA, 9: 231–264.
DOI:10.3934/bdia.2025011
Received: 11 November 2025
Revised: 14 November 2025
Accepted: 15 November 2025
Published: 25 November 2025

Research article

Under-bagging nearest neighbors for long-tailed regression

Hanyuan Hang[1], **Hongwei Wen**[2] **and Zhouchen Lin**[3,*]

[1] AI Energy Lab, EcoFlow Technology Inc., Shenzhen 518052, China

[2] School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

[3] State Key Lab. of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing 100871, China

* **Correspondence:** Email: zlin@pku.edu.cn.

Abstract: Long-tailed regression, also known as imbalanced regression, poses significant challenges in real-world prediction tasks where the target labels follow highly skewed or heavy-tailed distributions. In such scenarios, conventional regressors tend to bias toward high-density regions and perform poorly on tail samples, which often carry critical information in scientific and industrial applications. To address this issue, we propose an ensemble-based under-sampling algorithm named under-bagging nearest neighbors for long-tailed regression (UNNLR). The method employs a data-driven histogram density estimation (HDE) to adaptively estimate the label densities and determine sampling probabilities, thereby generating approximately uniform label distributions in subsampled datasets. To mitigate the information loss caused by under-sampling, a bagging mechanism is introduced, allowing multiple k-nearest neighbor (k-NN) regressors to be trained on bootstrap sub-samples in parallel. The theoretical analysis establishes, for the first time, minimax-optimal convergence rates for sampling-based regression under label imbalance. Experiments on both synthetic and real-world datasets demonstrate that UNNLR consistently outperforms existing sampling-based methods such as SMOGN, IRRCE, and ReBagg in terms of balanced mean squared error (BMSE), while achieving superior computational efficiency. The proposed framework bridges theoretical guarantees with scalable regression learning under long-tailed label distributions.

Keywords: long-tailed regression; imbalanced regression; ensemble learning; under-sampling; bagging; histogram density estimation; convergence analysis; k-nearest neighbors

1. Introduction

Regression tasks under highly skewed or heavy-tailed label distributions—commonly referred to as *long-tailed regression* or *imbalanced regression*—are pervasive in modern data-driven applications

[1, 2]. In such settings, most samples concentrate in a few high-density regions, while the tail regions, containing rare but often critical observations, are sparsely populated. When a regressor is trained on such data, it tends to overfit the dominant regions, resulting in biased predictions and poor performance on the tail labels. However, accurate prediction of low-density labels is vital in numerous domains. For example, in financial markets, forecasting extreme returns is essential despite their rarity, as these rare events can induce disproportionate economic impacts [3].

To tackle this challenge, two mainstream paradigms have been developed: cost-sensitive learning and re-sampling strategies. Cost-sensitive approaches assign higher loss weights to samples in low-density regions, forcing the regressor to better capture tail behaviors within an optimization framework [2, 4–6]. Although theoretically principled, these methods often lead to complex loss functions that are computationally expensive to optimize, especially for large-scale regression tasks. Re-sampling strategies, on the other hand, aim to rebalance the labels' distribution in the training data by adjusting the sample frequencies, offering a simpler and more flexible solution in practice.

Existing re-sampling algorithms can be broadly divided into over-sampling and under-sampling approaches. Over-sampling methods rebalance the dataset by generating synthetic instances for low-density labels. Notable examples include SMOGN [7], DistSMOGN [8], and their variants [9, 11], which are inspired by the SMOTE algorithm [10] and use a k-nearest neighbor (k-NN) criterion to interpolate synthetic labels. While effective in some scenarios, these methods significantly increase the training costs and may introduce noise or overfitting due to the artificial data generation process, especially under severe label imbalance.

To improve computational efficiency, under-sampling approaches have been proposed, which subsample high-density regions until the labels' distribution becomes approximately uniform. Early work such as [12] divided labels into coarse high- and low-density categories using a threshold and randomly subsampled labels from the dense regions. Subsequent studies, including [13], incorporated the bagging technique to bootstrap multiple subsampled datasets, while [11] further refined the idea by partitioning high-density regions into a fixed number of intervals. However, these approaches rely on predefined partitions or limited density granularity, making their estimated sampling probabilities too coarse to accurately reflect the true label imbalance. As a result, the subsampled datasets often remain internally imbalanced, leading to inconsistent training performance and unstable regression predictions.

Motivated by these limitations, we propose a theoretically grounded and computationally efficient ensemble learning framework named under-bagging nearest neighbors for long-tailed regression (UNNLR). Our approach introduces a histogram density estimation (HDE) strategy to adaptively partition the label space and estimate the sampling probabilities with fine granularity. Unlike fixed-partition strategies, HDE dynamically adjusts to the underlying label density, enabling more accurate estimation across regions with drastically different sample frequencies. An under-sampling k-NN regressor is then constructed with acceptance probabilities that are inversely proportional to the estimated density. However, when the estimated density becomes extremely small in the tail regions, the corresponding sampling probability may also diminish, potentially leading to information loss. To address this, we integrate a bagging mechanism that aggregates multiple under-sampled k-NN regressors trained on bootstrap subsets, thereby recovering under-represented information and enhancing parallel computational efficiency.

The main contributions of this paper are summarized as follows.

(i) Data-driven under-sampling. We propose, for the first time, a histogram-based density estimation strategy for sampling-based long-tailed regression, enabling accurate sample weighting and adaptive partitioning of the label space.
(ii) Theoretical guarantees. We establish new learning-theoretic results for sampling-based regression, proving that our under-bagging k-NN regressor achieves minimax-optimal convergence rates with respect to the balanced least squares loss.
(iii) Empirical superiority. Through comprehensive experiments on both synthetic and real-world datasets, we validate the effectiveness of UNNLR, demonstrating consistent improvements over existing sampling-based baselines in both predictive performance and computational efficiency.

2. Preliminaries

2.1. Notations

We first introduce the basic notation and conventions used throughout this paper. For any vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $1 \leq p < \infty$, the L_p-norm is defined as

$$\|x\|_p := \left(|x_1|^p + \cdots + |x_d|^p\right)^{1/p},$$

and the L_∞-norm as

$$\|x\|_\infty := \max_{1 \leq i \leq d} |x_i|.$$

For any $x \in \mathbb{R}^d$ and radius $r > 0$, we

$$B_r(x) := \{x' \in \mathbb{R}^d : \|x' - x\|_2 \leq r\}$$

to denote the closed Euclidean ball centered at x with a radius r.

For two sequences (a_n) and (b_n), we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there is a positive constant c independent of n such that $a_n \leq c\, b_n$ for all $n \in \mathbb{N}$. Similarly, $a_n \gtrsim b_n$ denotes that a constant $c \in (0, 1)$ exists such that $a_n \geq c^{-1} b_n$.

For any measurable set $A \subset \mathbb{R}^d$, $\#(A)$ denotes its cardinality and $\mathbf{1}_A$ (or equivalently $\mathbf{1}\{A\}$) denotes the indicator function of A.

2.2. Long-tailed regression problem

In regression analysis, the goal is to learn a function that predicts the value of an unobserved output variable Y based on an observed input variable X. We observe a training dataset

$$D_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

consisting of n independent and identically distributed (i.i.d.) samples drawn from an unknown joint distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$.

Let $f : \mathcal{X} \to \mathcal{Y}$ be a measurable regression function, and let

$$L(x, y, f(x)) : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$$

be a loss function that quantifies the penalty incurred when predicting $f(x)$ given the true label y. The expected risk of f under L and P is defined as

$$\mathcal{R}_{L,P}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, f(x))\, dP(x, y).$$

The following optimal regression function, also called the Bayes function, minimizes this risk:

$$f^*_{L,P} := \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_{L,P}(f).$$

In the case of the classical *least squares* (LS) loss

$$L_{\mathrm{LS}}(x, y, f(x)) := (y - f(x))^2,$$

the Bayes regressor takes the form

$$f^*_{L_{\mathrm{LS}},P}(x) = \mathbb{E}[Y \mid X = x],$$

which corresponds to the conditional expectation of Y given $X = x$.

When the marginal distribution of Y, denoted $p_Y(y)$, is highly non-uniform (as is common in long-tailed regression), the least squares loss inherently biases the learned function toward high-density label regions. Consequently, the model tends to perform poorly on tail regions where $p_Y(y)$ is small, leading to imbalanced prediction behavior.

2.3. Balanced loss for long-tailed regression

To mitigate the bias induced by non-uniform label densities, a balanced least squares (BLS) loss [4, 15] is introduced. This loss re-weights the standard least squares loss according to the inverse of the label density as follows:

$$L_{\mathrm{bal}}(x, y, f(x)) := \frac{(y - f(x))^2}{p_Y(y)}. \tag{2.1}$$

Intuitively, this formulation amplifies the contribution of rare labels during training, forcing the regressor to pay equal attention to both the head and tail regions of the label space.

The corresponding balanced risk becomes

$$\begin{aligned}\mathcal{R}_{L_{\mathrm{bal}},P}(f) &= \int_{\mathcal{X}\times\mathcal{Y}} L_{\mathrm{bal}}(x, y, f(x))\, dP(x, y) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} (y - f(x))^2\, p_{X|Y}(x|y)\, dx\, dy,\end{aligned}$$

which effectively neutralizes the influence of the marginal label distribution $p_Y(y)$ and evaluates the model's performance uniformly across labels.

In practice, the label density $p_Y(y)$ is unknown and must be estimated from the data. Let $\widehat{p_Y}(y)$ denote a nonparametric estimate (e.g., a histogram or kernel density estimator) obtained from D_n. The balanced mean squared error (BMSE) is then defined as

$$\mathrm{BMSE}(D_n, \widehat{f}) := \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \widehat{f}(X_i))^2}{\widehat{p_Y}(Y_i)}. \tag{2.2}$$

The BMSE thus serves as a performance metric that evaluates regression models in a label-density-agnostic manner, and will be used throughout this work as a key evaluation criterion for long-tailed regression algorithms.

3. Methodology

In this section, we introduce the proposed under-sampling algorithm for long-tailed regression, named UNNLR. Section 3.1 presents a data-driven under-sampling strategy that constructs a balanced dataset with nearly uniform label density. Subsequently, Section 3.2 describes how the bagging technique is incorporated to mitigate information loss and improve the performance of the under-sampling k-NN regressor.

3.1. Under-sampling k-NN for long-tailed regression

The core component of any under-sampling algorithm for long-tailed regression lies in the sampling mechanism—how to generate a balanced dataset that mitigates the labels' skewness. In this subsection, we first analyze why datasets generated by existing under-sampling methods still exhibit nonuniform label density. We then introduce a data-driven under-sampling strategy that uses accurate density estimation to construct a uniformly distributed dataset.

Problems in existing under-sampling strategies. To obtain a balanced dataset via under-sampling, the number of high-density labels must be reduced to match that of the low-density labels. However, most existing approaches rely on *partition-based* methods, where the number of label partitions is fixed and pre-defined. As shown in Figure 1(a), we demonstrate an example with a bandwidth $h = 1/6$, where the density threshold equals one and the high-density region is marked in red. By sub-sampling only from these high-density areas (Figure 1(b)), the resulting dataset remains long-tailed: Not only does the relative proportion between dense and sparse regions persist, but the intra-partition distributional patterns also remain unchanged. This phenomenon reveals that partition-based methods fail to produce truly uniform label distributions.

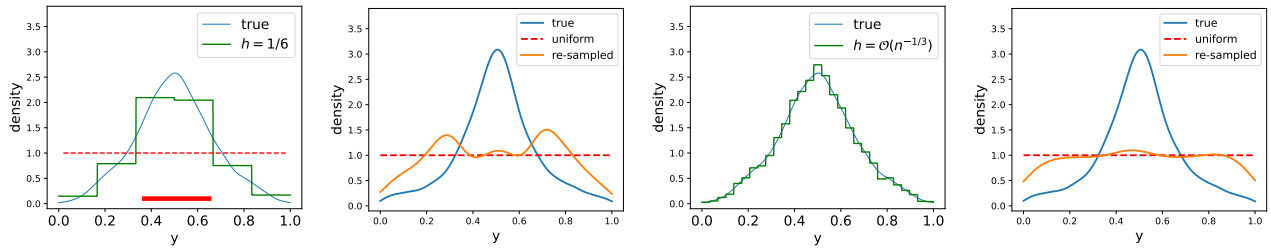


Figure 1. Label density estimation and under-sampling results for the synthetic `tan` dataset. Panels (a)–(b) show the traditional fixed partition method, while (c)–(d) demonstrate our data-driven HDE approach, which produces an approximately uniform label density.

Label density estimation. Accurate estimation of the labels' density is crucial for assigning appropriate sampling probabilities to each instance. We adopt a data-driven *HDE* approach, as suggested in [14], which can be regarded as a special case of kernel density estimation suitable for one-dimensional variables. We assume, without loss of generality, that the label space is normalized to $\mathcal{Y} := [0, 1]$ (e.g., using min-max scaling). Given the bandwidth parameter $h \in (0, 1]$, we divide $\mathcal{Y}$

into $N := \lfloor h^{-1} \rfloor$ equally spaced bins. For a dataset $D_n = \{(X_i, Y_i)\}_{i=1}^n$, the histogram density estimator is defined as

$$\widehat{p}_Y(y) = \frac{1}{nh} \sum_{i=1}^{N} n_i \mathbf{1}\{y \in [(i-1)h, ih)\}, \tag{3.1}$$

where $n_i := \sum_{j=1}^n \mathbf{1}\{Y_j \in [(i-1)h, ih)\}$ denotes the number of samples within the i-th bin. In practice, the bandwidth h is set as $h \asymp n^{-1/(2\alpha+1)}$ according to the theoretical results in Section 4, where n is the dataset size and α is the Hölder continuity parameter of the underlying density function. As shown in Figure 1(c), this data-driven estimator provides an accurate fit to the true label density and effectively captures its non-uniformity.

Data-driven sub-sampling strategy. On the basis of the estimated label density $\widehat{p}_Y(y)$, we define a probabilistic sub-sampling mechanism that eliminates the within-bin distributional patterns. Specifically, for each sample $(X_i, Y_i) \in D_n$, we compute the acceptance probability:

$$a(X_i, Y_i) := \frac{\widehat{p}_Y^{-1}(Y_i)}{\max_{j=1,\ldots,n} \widehat{p}_Y^{-1}(Y_j)} = \rho \cdot \widehat{p}_Y^{-1}(Y_i), \qquad i = 1, \ldots, n, \tag{3.2}$$

where $\rho := \min_{j=1,\ldots,n} \widehat{p}_Y(Y_j)$ serves as a normalization factor ensuring $a(X_i, Y_i) \in (0, 1]$. This definition assigns higher acceptance probabilities to low-density samples and lower probabilities to those from high-density regions.

The sub-sampling procedure follows two simple steps.

(1) For each (X_i, Y_i), draw $Z(X_i, Y_i) \sim \text{Bernoulli}(a(X_i, Y_i))$.
(2) If $Z(X_i, Y_i) = 1$, accept (X_i, Y_i) into the sub-sampled dataset D_n^u; otherwise, reject it.

The resulting dataset is $D_n^u = \{(X_1^u, Y_1^u), \ldots, (X_{s_u}^u, Y_{s_u}^u)\}$, where $s_u = \#(D_n^u)$ denotes its size. As illustrated in Figure 1(d), the label density of D_n^u is nearly uniform, indicating successful balancing.

Under-sampling k-NN regressor. Once the balanced dataset D_n^u is obtained, we train a k-NN regressor. Let $(X_{(i)}^u(x), Y_{(i)}^u(x))$ denote the i-th nearest neighbor of a query point x, satisfying $\|X_{(1)}^u(x) - x\| \le \cdots \le \|X_{(s_u)}^u(x) - x\|$. The under-sampling k-NN regressor is then defined as:

$$\widehat{f}^{k,u}(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}^{k,u}(x). \tag{3.3}$$

Although this regressor benefits from a uniform label density, excessive sub-sampling may remove valuable information, especially in low-density boundary regions.

3.2. Under-bagging k-NN for long-tailed regression

While the under-sampling k-NN regressor achieves label balance, it may suffer from information loss when the minimal estimated density ρ becomes extremely small (e.g., $\rho = 1/(nh)$). This occurs when certain partitions contain only a single instance, leading to tiny acceptance probabilities in Eq (3.2). Consequently, many informative samples are excluded, degrading the regression's

accuracy—an effect that is especially common in long-tailed regression, where the labels are continuous and the tail densities are inherently sparse.

To mitigate information loss, we incorporate the bagging technique, leveraging unused samples in $D_n \setminus D_n^u$. Specifically, let $1 \le s \le \rho n$ denote the expected number of bootstrap samples. We define the acceptance probability with a bootstrap sampling ratio as follows:

$$a(X_j, Y_j) = \frac{s}{n} \cdot \widehat{p}_Y(Y_j), \quad j = 1, \ldots, n. \tag{3.4}$$

At each bagging iteration $b = 1, \ldots, B$, we follow the sampling strategy described in Section 3.1 to obtain a bootstrap balanced subset as shown below:

$$D_b^u := \{(X_1^{b,u}, Y_1^{b,u}), \ldots, (X_{s_b}^{b,u}, Y_{s_b}^{b,u})\}.$$

We then train a base learner $\widehat{f}^{b,u}$ on D_b^u using the k-NN rule:

$$\widehat{f}^{b,u}(x) := \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}^{b,u}(x). \tag{3.5}$$

The final ensemble regressor, termed the under-bagging k-NN regressor, aggregates all B base models as:

$$\widehat{f}^{B,u}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{f}^{b,u}(x). \tag{3.6}$$

This bagging mechanism exploits complementary information from multiple balanced subsets, effectively reducing variance and alleviating information loss.

Algorithm 1: UNNLR

Input: Training dataset $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$;
Bagging iterations B; expected under-sampling size s; nearest neighbor parameter k.

Construct the data-driven HDE based on $\{Y_i\}_{i=1}^n$ as $\widehat{p}_Y(\cdot)$;
for $b = 1 \to B$ **do**
  Define the acceptance probability $a_s(x, y)$ by Eq (3.4) with size parameter s;
  Randomly sample D_b^u from D_n according to $a_s(x, y)$;
  Compute the b-th k-NN regressor $\widehat{f}^{b,u}$ via Eq (3.5);
Aggregate the regressors to obtain $\widehat{f}^{B,u}$ using Eq (3.6).
Output: Final under-bagging nearest neighbors regressor $\widehat{f}^{B,u}$.

The proposed UNNLR framework integrates data-driven density estimation with ensemble learning to achieve both balance and robustness. The HDE adaptively captures the labels' distribution characteristics, while the bagging ensemble compensates for data sparsity and stabilizes the prediction. Together, these two components yield a theoretically sound and computationally efficient solution for long-tailed regression.

4. Theoretical results

In this section, we establish theoretical guarantees for the proposed UNNLR algorithm. Our analysis builds upon the classical smoothness assumption on the regression function, a standard foundation in nonparametric regression theory [16, 17]. Specifically, we assume that the underlying regression function belongs to a Hölder continuous function space.

Definition 4.1. Let $\alpha \in (0, 1]$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be *α-Hölder continuous* if a constant $c_L \in (0, \infty)$ exists such that

$$|f(x) - f(x')| \le c_L \|x - x'\|^\alpha, \quad \forall x, x' \in \mathcal{X}.$$

The collection of all such functions is denoted by $C^\alpha(\mathcal{X})$.

4.1. Convergence rates for the under-sampling k-NN regressor

We begin by analyzing the convergence rate of the under-sampling k-NN regressor under BLS loss.

Assumption 4.2. Let $p(x, y)$ denote the joint probability density function with respect to the probability measure $\mathrm{P}_{\mathcal{X}\times\mathcal{Y}}$. We assume that the constants $\alpha, \beta \in (0, 1]$ and $c_\alpha, c_\beta > 0$ exist such that

$$|p(x, y) - p(x', y)| \le c_\alpha \|x - x'\|^\alpha, \qquad |p(x, y) - p(x, y')| \le c_\beta |y - y'|^\beta.$$

Moreover, there are positive constants $\underline{c}$ and $\overline{c}$ satisfying $0 < \underline{c} \le p(x, y) \le \overline{c} < \infty$. We further assume that P_X has bounded compact support.

Theorem 4.3. *Let $\widehat{f}^{k,u}$ denote the under-sampling k-NN regressor defined in* (3.3)*, where the acceptance probability follows Eq* (3.2)*. Assume that* P *satisfies Assumption 4.2, and define* $\gamma := \min\{\alpha, \beta\}$*. We then have, for sufficiently large n, by choosing*

$$h = n^{-\frac{1}{2\beta+1}}, \qquad k = s_u^{2\alpha/(2\alpha+d)} (\log s_u)^{d/(2\alpha+d)}, \tag{4.1}$$

where $s_u = \#(D_n^u)$ denotes the size of the under-sampled dataset, we obtain

$$\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\widehat{f}^{k,u}) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} \lesssim (n/\log n)^{-2\gamma/(2\gamma+d)} \tag{4.2}$$

with a probability of at least $1 - 4/n^2$ under $\mathrm{P}_Z \otimes \mathrm{P}^n$.

Compared with the standard k-NN regressor, where k typically scales as $n^{2\alpha/(2\alpha+d)}$ (up to logarithmic factors), Theorem 4.3 shows that, with under-sampling, k effectively scales as $(\rho n)^{2\alpha/(2\alpha+d)}$. Hence, for highly long-tailed data with a small ρ, the optimal neighborhood size k is significantly smaller, leading to improved computational efficiency.

The next theorem establishes the minimax lower bound for the long-tailed regression problem under BLS loss.

Theorem 4.4. *Let $\mathcal{F}$ denote the set of all measurable functions $f_n : (\mathbb{R}^d \times \mathbb{R})^n \times \mathbb{R}^d \to \mathbb{R}$, and let $\mathcal{P}$ denote the family of probability distributions satisfying Assumption 4.2. In this case,*

$$\inf_{f_n \in \mathcal{F}} \sup_{\mathrm{P} \in \mathcal{P}} \left(\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} \right) \gtrsim n^{-2\alpha/(2\alpha+d)}.$$

Theorem 4.4 provides, for the first time, a lower bound for long-tailed regression under BLS loss. This rate coincides with the classical minimax bound for standard regression under the mean squared error [16, 17]. Together with Theorem 4.3, this result confirms that the convergence rate achieved by the under-sampling k-NN regressor is indeed optimal.

Furthermore, observe that the convergence rate in (4.2) depends on $\gamma = \min\{\alpha, \beta\}$. Since the variables in higher-dimensional spaces are typically less smooth, α is often smaller than β, justifying the rate exponent. When $\alpha > \beta$, Theorem 4.3 indicates that the accuracy of estimation of the labels' density becomes the limiting factor for convergence. Hence, the proposed data-driven HDE strategy in Section 3.1 plays a critical role in achieving optimal rates for long-tailed regression.

4.2. Convergence rates for the under-bagging k-NN regressor

We next establish the convergence behavior of the proposed under-bagging k-NN regressor under the same BLS framework.

Theorem 4.5. *Let $\widehat{f}^{B,u}(x)$ denote the under-bagging k-NN regressor defined in* (3.6)*. Assume that* P *satisfies Assumption 4.2, and define $\gamma := \min\{\alpha, \beta\}$. Then, for a sufficiently large n, by selecting*

$$h \asymp n^{-\frac{1}{2\beta+1}}, \tag{4.3}$$

$$s \gtrsim \begin{cases} (\rho n)^{d/(2\alpha+d)} (\log(\rho n))^{2\alpha/(2\alpha+d)}, & \text{if } d > 2\alpha, \\ (\rho n \log(\rho n))^{1/2}, & \text{if } d \le 2\alpha, \end{cases} \tag{4.4}$$

$$k = s(\log(\rho n)/\rho n)^{d/(2\alpha+d)}, \tag{4.5}$$

$$B = k\rho n/s = (\rho n)^{2\alpha/(2\alpha+d)} (\log(\rho n))^{d/(2\alpha+d)}, \tag{4.6}$$

the following bound holds:

$$\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\widehat{f}^{B,u}) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} \lesssim (n/\log n)^{-2\gamma/(2\gamma+d)} \tag{4.7}$$

with a probability of at least $1 - 5/n^2$ under $\mathrm{P}_Z^B \otimes \mathrm{P}^n$.

Theorem 4.5, together with the lower bound in Theorem 4.4, implies that the convergence rate (4.7) achieved by the under-bagging k-NN regressor is minimax-optimal (up to logarithmic factors) under the BLS loss. Moreover, from Eqs (4.5) and (4.6), we observe that both k and B scale proportionally with the sub-sample size s. Therefore, only a small number of bootstrap subsets are required to achieve convergence. In particular, k is reduced to $O(\log n)$, yielding substantial computational advantages.

4.3. Complexity analysis

We now analyze the computational complexity of the proposed algorithm. As in standard nearest-neighbor methods, a k-d tree structure [18] is adopted for an efficient neighbor search. We show that under-bagging substantially reduces the complexity of both the construction and query time while maintaining comparable space usage.

Tree construction. According to [18], building a k-d tree for n samples requires a time $O(nd \log n)$ and a space $O(nd)$. By Theorem 4.5, choosing

$$B = O((\rho n)^{2\alpha/(2\alpha+d)} (\log(\rho n))^{d/(2\alpha+d)}), \quad s = (\rho n)^{d/(2\alpha+d)} (\log(\rho n))^{2\alpha/(2\alpha+d)},$$

the construction cost per base learner reduces to

$$O((\rho n)^{d/(2\alpha+d)}\, d \log(\rho n)),$$

assuming parallel computation across bagging rounds. The overall space complexity becomes $O(Bsd) = O(k\rho nd) = O(\rho n \log(\rho n)d)$, comparable with the standard $O(nd)$ requirement.

Neighbor search. For query operations, the standard k-NN requires $O(k \log n)$ time [18]. Since $k = O(n^{2\alpha/(2\alpha+d)})$ for the classical k-NN [20, 21], its query complexity is $O(n^{2\alpha/(2\alpha+d)} \log n)$. In contrast, under Theorem 4.5, each base learner of UNNLR searches only $O(\log(\rho n))$ neighbors among $s = (\rho n)^{d/(2\alpha+d)}(\log(\rho n))^{2\alpha/(2\alpha+d)}$ samples. Thus, the search time reduces to $O(\log^2(\rho n))$, a substantial improvement.

In summary, the under-bagging mechanism markedly enhances computational efficiency, especially when combined with parallel computation. Although higher-dimensional spaces typically suffer from the curse of dimensionality—requiring exponentially more samples—the proposed approach mitigates this effect by lowering the dependence on n and enabling efficient distributed implementation.

5. Experiments

In this section, we present both synthetic and real-data experiments to validate the theoretical properties of the proposed UNNLR algorithm and to demonstrate its empirical superiority over representative state-of-the-art long-tailed regression approaches.

5.1. Synthetic data experiments

We first conduct synthetic experiments to analyze the influence of the hyperparameters in UNNLR, including the bandwidth h of HDE, the number of bagging iterations B, the under-sampling size s, and the nearest-neighbor parameter k.

Synthetic dataset. To construct an long-tailed regression scenario, we employ the tangent function, whose derivative $1/\cos^2(x)$ induces non-uniform variation over its domain. Because $\tan(x)$ changes more slowly near $x = 0$ and diverges near $\pm\frac{\pi}{2}$, its value distribution naturally exhibits long-tailed imbalance. We generate the synthetic dataset by sampling $X \in [-\frac{2\pi}{5}, \frac{2\pi}{5}]$ uniformly and define the regression function as

$$f(x) = \tan(x) + \sigma,$$

where $\sigma \sim \mathcal{N}(0, 0.5)$ represents additive Gaussian noise. An example of the resulting data distribution is illustrated in Figure 2.

Experimental setup. Each dataset is split into training and test subsets with a ratio of 7 : 3. A five-fold cross-validation is performed on the training subset to determine the optimal hyperparameters, and a final evaluation is carried out on the held-out test subset using the selected parameters.

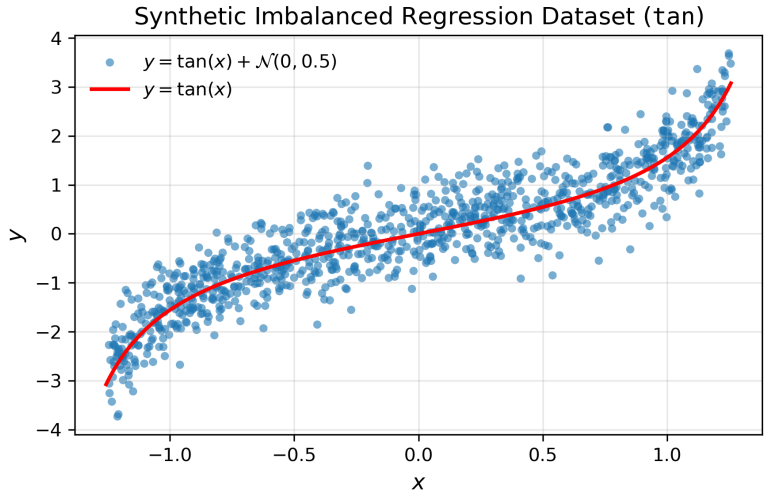
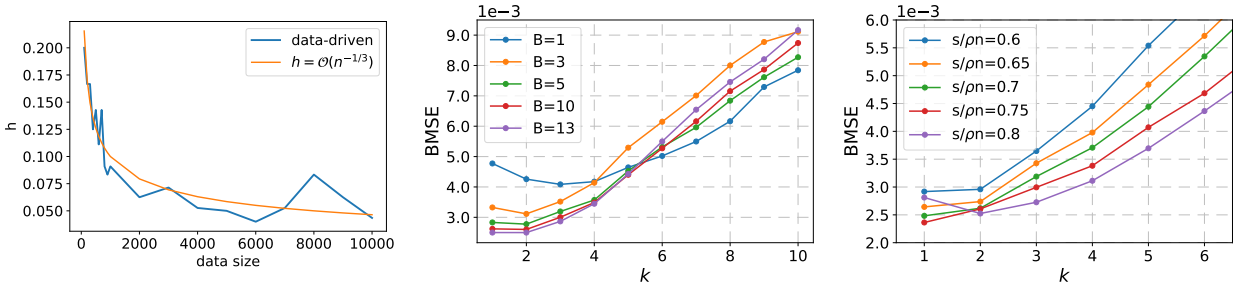


Figure 2. Synthetic data distribution for the `tan` function. Samples are drawn uniformly from $x \in [-\frac{2\pi}{5}, \frac{2\pi}{5}]$ and labeled by $y = \tan(x) + \mathcal{N}(0, 0.5)$. The dense region near $x = 0$ and the sparse tails near the boundaries demonstrate the long-tailed property of the label distribution.

Verification of the data-driven under-sampling strategy. To verify the theoretical relationship between the dataset size n and the optimal bandwidth h, we generate multiple `tan` datasets with n ranging from 100 to 10,000. As shown in Figure 3(a), the empirically optimal h closely follows the theoretical scaling law $h = O(n^{-1/3})$, supporting our analysis in Section 4.



(a) Relationship between n and h

(b) Relationship between B and k

(c) Relationship between s and k

Figure 3. Parameter sensitivity analysis of UNNLR on the synthetic `tan` dataset. Panel (a) verifies the theoretical scaling of h, (b) shows the stabilizing effect of bagging on k, and (c) analyzes the influence of the sampling ratio $s/(\rho n)$.

Effect of bagging on parameter selection. Bagging not only alleviates information loss but also stabilizes parameter tuning. Fixing $n = 10{,}000$ and using the optimal h, we evaluate the method's performance over different combinations of B and k. Figure 3(b) shows that as B increases, (i) the minimum BMSE loss decreases from approximately 0.004 to 0.001, and (ii) the optimal k decreases from 3 to 1. This demonstrates that bagging both enhances stability and simplifies the model's configuration.

Influence of the sampling ratio. Figure 3(c) investigates the impact of different sampling ratios $s/(\rho n)$. Both excessively small and excessively large s values degrade performance: A small s causes severe information loss (requiring a large B to compensate), while a large s reduces sampling diversity and thus weakens the ensemble effect. Moderate ratios yield an optimal balance between bias and variance.

5.2. Real data experiments

We next evaluate UNNLR on 10 real-world datasets from the UCI Machine Learning Repository [22] and the OpenML platform [23], namely `Abalone` [24], `Accel` [25], `Airfoild` [22], `AvailPwr` [23], `Bank8FM` [26], `ConcrStr` [27], `CpuSm` [28], `DAiler` [29], `FuelCons` [23], and `MaxTorq` [23]. Basic dataset characteristics and the minimum estimated density ρ under HDE are summarized in Table 1.

Table 1. Summary of real datasets used in experiments. N denotes the number of samples, *Dim* the feature dimension, and ρ the minimum estimated density obtained from data-driven HDE.

Dataset	Abalone	Accel	Airfoild	AvailPwr	Bank8FM	ConcrStr	CpuSm	DAiler	FuelCons	MaxTorq
Abbreviation	`ABA`	`ACC`	`AIR`	`AVA`	`BAN`	`CON`	`CPU`	`DEL`	`FUE`	`MAX`
N	4177	1732	1503	1802	4499	1030	8192	7129	1764	1802
Dim	8	15	5	16	9	8	13	5	38	33
ρ	3.83×10^{-3}	1.39×10^{-2}	6.59×10^{-2}	1.33×10^{-2}	3.56×10^{-3}	1.84×10^{-1}	7.32×10^{-3}	2.67×10^{-3}	1.36×10^{-2}	6.66×10^{-3}

Comparison methods. We compare UNNLR with three representative sampling-based long-tailed regression algorithms (IRRCE [9], SMOGN [7], and ReBagg [13]) as well as with the standard k-NN regressor as a nonsampling baseline.

Parameter settings. The hyperparameter grids for all methods are configured as follows.

- For `SMOGN`, `IRRCE`, and `ReBagg`, we follow the authors' recommended settings and search over $k_1 \in$ [2:1:10], noise level $\sigma \in$ [0.01:0.01:0.05], and threshold $\in$ [0.1:0.1:0.5].
- For `UNNLR`, we search $B \in \{1, 2, 5, 10, 20\}$, subsampling ratio $s/n \in$ [0.1:0.1:0.9], the number of bins $h^{-1} \in$ [1:2:25], and the nearest-neighbor parameter $k \in$ [1:2:30].
- For the standard k-`NN`, k is tuned within [1:1:50].

Evaluation metrics. We adopt two regression metrics: The BMSE to assess performance under label imbalance, and the standard mean square error (MSE) for conventional accuracy evaluation. To assess efficiency, we also report the runtime (in seconds) for each algorithm using the optimal parameter configuration.

Results and discussion. Table 2 reports the BMSE, MSE, and runtime for all 10 datasets. UNNLR consistently achieves the lowest BMSE in most cases, confirming its effectiveness in addressing imbalance by equalizing the label density and stabilizing the variance through bagging. Interestingly,

while the existing sampling-based methods (IRRCE, SMOGN, and ReBagg) often perform better than plain k-NN on BMSE, they underperform on MSE, suggesting that they partially alleviate—but do not fully resolve—imbalance effects. In contrast, UNNLR attains the best overall trade-off between balance-aware accuracy and computational efficiency. Notably, its runtime remains among the lowest, demonstrating both scalability and practicality.

Table 2. Performance comparison of UNNLR and baseline methods on 10 real-world regression datasets. The best result for each dataset and criterion is highlighted in **bold**.

Dataset	Criteria	UNNLR	UsNNIR	SMOGN	IRRCE	ReBagg	k-NN
	BMSE	2.020×10^{-1}	2.056×10^{-1}	3.007×10^{-1}	2.903×10^{-1}	3.020×10^{-1}	3.102×10^{-1}
ABA	MSE	1.999×10^{-2}	2.517×10^{-2}	7.137×10^{-3}	7.255×10^{-3}	7.059×10^{-3}	8.085×10^{-3}
	Time(s)	3.950×10^{-3}	3.831×10^{-3}	1.570×10^{2}	2.440×10^{0}	6.311×10^{1}	7.550×10^{-2}
	BMSE	2.873×10^{-3}	2.911×10^{-3}	3.068×10^{-3}	2.906×10^{-3}	4.123×10^{-3}	3.399×10^{-3}
ACC	MSE	2.088×10^{-3}	2.156×10^{-3}	2.733×10^{-3}	2.148×10^{-3}	2.210×10^{-3}	2.053×10^{-3}
	Time(s)	2.420×10^{-3}	1.386×10^{-2}	1.100×10^{1}	1.060×10^{1}	5.266×10^{1}	4.030×10^{-3}
	BMSE	5.029×10^{-3}	5.255×10^{-3}	6.371×10^{-3}	6.775×10^{-3}	5.230×10^{-3}	5.255×10^{-3}
AIR	MSE	5.189×10^{-3}	5.622×10^{-3}	7.902×10^{-3}	5.353×10^{-3}	5.759×10^{-3}	5.622×10^{-3}
	Time(s)	2.180×10^{-3}	6.587×10^{-2}	4.320×10^{0}	1.030×10^{1}	3.125×10^{1}	3.600×10^{-3}
	BMSE	6.957×10^{-3}	7.454×10^{-3}	7.813×10^{-3}	7.442×10^{-3}	6.175×10^{-3}	7.458×10^{-3}
AVA	MSE	8.443×10^{-4}	8.689×10^{-4}	1.062×10^{-3}	8.606×10^{-4}	9.385×10^{-4}	8.704×10^{-4}
	Time(s)	2.330×10^{-3}	1.332×10^{-2}	2.130×10^{1}	1.210×10^{1}	5.740×10^{1}	4.030×10^{-3}
	BMSE	2.277×10^{-2}	2.948×10^{-2}	2.771×10^{-2}	3.051×10^{-2}	9.531×10^{-3}	2.895×10^{-2}
BAN	MSE	3.639×10^{-3}	8.362×10^{-3}	6.259×10^{-3}	4.813×10^{-3}	7.188×10^{-3}	4.753×10^{-3}
	Time(s)	2.390×10^{-3}	3.556×10^{-3}	3.900×10^{1}	1.190×10^{1}	1.416×10^{2}	3.400×10^{-3}
	BMSE	1.185×10^{-2}	1.484×10^{-2}	1.517×10^{-2}	1.539×10^{-2}	1.212×10^{-2}	1.235×10^{-2}
CON	MSE	1.297×10^{-2}	1.452×10^{-2}	1.423×10^{-2}	1.094×10^{-2}	1.211×10^{-2}	1.128×10^{-2}
	Time(s)	2.000×10^{-3}	1.845×10^{-1}	7.370×10^{0}	1.290×10^{1}	2.786×10^{1}	3.580×10^{-3}
	BMSE	2.473×10^{-3}	3.110×10^{-3}	2.503×10^{-3}	4.566×10^{-3}	3.527×10^{-3}	3.367×10^{-3}
CPU	MSE	1.736×10^{-3}	2.172×10^{-3}	1.317×10^{-3}	1.630×10^{-3}	1.560×10^{-3}	1.563×10^{-3}
	Time(s)	2.140×10^{-3}	7.324×10^{-3}	1.710×10^{2}	1.620×10^{1}	5.374×10^{1}	3.650×10^{-3}
	BMSE	3.005×10^{-2}	4.260×10^{-2}	4.280×10^{-2}	1.379×10^{-1}	3.274×10^{-2}	9.550×10^{-2}
DEL	MSE	5.620×10^{-3}	7.015×10^{-3}	3.096×10^{-3}	1.608×10^{-3}	3.149×10^{-3}	1.910×10^{-3}
	Time(s)	2.060×10^{-3}	2.665×10^{-3}	9.650×10^{1}	1.580×10^{0}	3.070×10^{2}	3.730×10^{-3}
	BMSE	2.812×10^{-3}	3.123×10^{-3}	3.365×10^{-3}	3.283×10^{-3}	3.839×10^{-3}	3.123×10^{-3}
FUE	MSE	1.835×10^{-3}	1.954×10^{-3}	2.519×10^{-3}	2.071×10^{-3}	3.108×10^{-3}	1.954×10^{-3}
	Time(s)	2.090×10^{-3}	1.361×10^{-2}	3.100×10^{1}	1.450×10^{1}	8.396×10^{2}	8.530×10^{-3}
	BMSE	3.159×10^{-2}	3.161×10^{-2}	3.338×10^{-2}	3.161×10^{-2}	3.492×10^{-2}	4.081×10^{-2}
MAX	MSE	1.346×10^{-3}	1.344×10^{-3}	2.621×10^{-3}	1.347×10^{-3}	2.067×10^{-3}	1.469×10^{-3}
	Time(s)	2.410×10^{-3}	6.659×10^{-3}	2.820×10^{1}	1.410×10^{1}	1.169×10^{2}	1.180×10^{-2}

6. Conclusions

In this paper, we propose an ensemble-based sampling-based long-tailed regression algorithm called UNNLR based on a data-driven under-sampling strategy to obtain a consistent density estimation for label density. The bagging technique is used for dealing with the information loss caused by the small sampling probability derived from the accurate density estimation. To our best knowledge, we are the first to provide a theoretically guaranteed sampling-based long-tailed

regression algorithm, and we even provide the optimal convergence rates of our UNNLR under mild assumptions. From the practical perspective, we conduct parameter analysis as guidance for parameter tuning in applications and empirically verify the superiority compared with other sampling-based long-tailed regression algorithms.

Use of AI tools declaration

The authors declare that artificial intelligence (AI) tools were used in the preparation of this article. Specifically, the AI language model ChatGPT (OpenAI, San Francisco, USA) was employed to assist in English language polishing, grammar improvement, and stylistic refinement of the manuscript. All technical content, equations, experimental designs, theoretical results, and interpretations were conceived, written, and verified by the authors. The AI-generated text was carefully reviewed and edited to ensure its accuracy, originality, and consistency with the authors' intent. The relevant sections where AI assistance was applied include the abstract, introduction, methodology, and discussion sections.

Acknowledgments

Zhouchen Lin was supported by the NSF China (No. 62276004) and the State Key Laboratory of General Artificial Intelligence.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Yang Y, Zha K, Chen Y, Wang H, Katabi D, (2021) Delving into deep imbalanced regression, In: *International Conference on Machine Learning*, 11842–11851.
2. Ren J, Zhang M, Yu C, Liu Z, (2022) Balanced MSE for imbalanced visual regression, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7926–7935. https://doi.org/10.1109/CVPR52688.2022.00777
3. Branco P, Torgo L, (2019) A study on the impact of data characteristics in imbalanced regression tasks, In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 193–202. https://doi.org/10.1109/DSAA.2019.00034
4. Steininger M, Kobs K, Davidson P, Krause A, Hotho A, (2021) Density-based weighting for imbalanced regression, *Mach Learn* 110: 2187–2211. https://doi.org/10.1007/s10994-021-06023-5
5. Ding Y, Jia M, Zhuang J, Ding P, (2022) Deep imbalanced regression using cost-sensitive learning and deep feature transfer for bearing remaining useful life estimation. *Appl Soft Comput* 127: 109271. https://doi.org/10.1016/j.asoc.2022.109271

6. Islam A, Belhaouari SB, Rehman AU, Bensmail H, (2022) KNNOR: An oversampling technique for imbalanced datasets. *Appl Soft Comput* 115: 108288. https://doi.org/10.1016/j.asoc.2021.108288
7. Branco P, Torgo L, Ribeiro RP, (2017) SMOGN: A pre-processing approach for imbalanced regression, In: *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 36–50.
8. Song XY, Dao N, Branco P, (2022) DistSMOGN: Distributed SMOGN for imbalanced regression problems, In: *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 38–52.
9. Orhobor OI, Grinberg NF, Soldatova LN, King RD, (2022) Imbalanced regression using regressor-classifier ensembles. *Mach Learn* 112: 1365–1387. https://doi.org/10.1007/s10994-022-06199-4
10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP, (2002) SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 16: 321–357. https://doi.org/10.1613/jair.953
11. Branco P, Torgo L, Ribeiro RP, (2019) Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing* 343: 76–99. https://doi.org/10.1016/j.neucom.2018.11.100
12. Torgo L, Branco P, Ribeiro RP, Pfahringer B, (2015) Resampling strategies for regression. *Exp Syst* 32: 465–476. https://doi.org/10.1111/exsy.12081
13. Branco P, Torgo L, Ribeiro RP, (2018) Rebagg: Resampled bagging for imbalanced regression, In: *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 67–81.
14. Hang H, Steinwart I, Feng Y, Suykens JAK, (2018) Kernel density estimation for dynamical systems. *J Mach Learn Res* 19: 1–49.
15. Sadouk L, Gadi T, Essoufi EH, (2021) A novel cost-sensitive algorithm and new evaluation strategies for regression in imbalanced domains. *Exp Syst* 38: e12680. https://doi.org/10.1111/exsy.12680
16. Tsybakov AB, (2004) *Introduction to Nonparametric Estimation*, Springer New York. https://doi.org/10.1007/b13794
17. Bartlett PL, Long PM, Lugosi G, Tsigler A, (2020) Benign overfitting in linear regression. *Proc Natl Acad Sci* 117: 30063–30070. https://doi.org/10.1073/pnas.1907378117
18. Friedman JH, Bentley JL, Finkel RA, (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans Math Software* 3: 209–226. https://doi.org/10.1145/355744.355745
19. Hang H, Cai Y, Yang H, Lin Z, (2022) Under-bagging nearest neighbors for imbalanced classification. *J Mach Learn Res* 23: 1–63.
20. Chaudhuri K, Dasgupta S, (2014) Rates of convergence for nearest neighbor classification. *Adv Neural Inf Process Syst* 27: 3437–3445.
21. Zhao P, Lai L, (2021) Minimax rate optimal adaptive nearest neighbor classification and regression. *IEEE Trans Inf Theory* 67: 3155–3182. https://doi.org/10.1109/TIT.2021.3062078
22. Dua D, Graff C, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2017. Available from: http://archive.ics.uci.edu/ml.

23. Vanschoren J, van Rijn JN, Bischl B, Torgo L, (2014) OpenML: Networked science in machine learning. *ACM SIGKDD Explor Newsl* 15: 49–60. https://doi.org/10.1145/2641190.2641198

24. Nash WJ, Sellers TL, Talbot SR, Cawthorn AJ, Ford WB, (1994) The population biology of abalone (Haliotis species) in Tasmania. *Sea Fish Div Tech Rep* 48: 411.

25. Fanaee-T H, Gama J, (2014) Event labeling combining ensemble detectors and background knowledge. *Prog Artif Intell* 2: 113–127. https://doi.org/10.1007/s13748-013-0040-3

26. Zhao Y, Sun J, (2009) Recursive reduced least squares support vector regression. *Pattern Recognit* 42: 837–842. https://doi.org/10.1016/j.patcog.2008.09.028

27. Yeh IC, (1998) Modeling of strength of high-performance concrete using artificial neural networks. *Cem Concr Res* 28: 1797–1808. https://doi.org/10.1016/S0008-8846(98)00165-3

28. Alhamdoosh M, Wang D, (2014) Fast decorrelated neural network ensembles with random weights. *Inf Sci* 264: 104–117. https://doi.org/10.1016/j.ins.2013.12.016

29. Torgo L, Ribeiro R, (2003) Predicting outliers, In: *Knowledge Discovery in Databases: PKDD 2003*, 447–458. https://doi.org/10.1007/978-3-540-39804-2_40

30. Bernstein SN, (1946) *The Theory of Probabilities*, Moscow: Gastehizdat Publishing House.

31. Hoeffding W, (1963) Probability inequalities for sums of bounded random variables. *J Am Stat Assoc* 58: 13–30. https://doi.org/10.1080/01621459.1963.10500830

32. Steinwart I, Christmann A , (2008) *Support Vector Machines*, Springer Science & Business Media, 1–312. https://doi.org/10.1007/978-0-387-77242-4_1

33. Torgo L, Ribeiro R, (2007) Utility-based regression, In: *European Conference on Principles of Data Mining and Knowledge Discovery*, 597–604. https://doi.org/10.1007/978-3-540-74976-9_63

34. Torgo L, Ribeiro RP, Pfahringer B, Branco P, (2013) Smote for regression, In: *Portuguese Conference on Artificial Intelligence*, 378–389. https://doi.org/10.1007/978-3-642-40669-0_33

Appendices

A. Error analysis

A.1. Error analysis for the BLS loss

To bound the excess risk with respect to the BLS loss, we begin by constructing an auxiliary probability distribution that allows the excess balanced risk to be reformulated as a standard mean square error (MSE). This transformation enables the use of classical approximation theory and Bernstein-type concentration inequalities in the analysis.

Let $\mathrm{P}(X, Y)$ denote the joint probability distribution of the samples, and let $p(x, y)$ be its corresponding density function. We define the *balanced probability distribution* $\mathrm{P}^b(X, Y)$ according to the density function

$$p^b(x, y) = \frac{p(x, y)}{p_Y(y)}, \tag{A.1}$$

where $p_Y(y)$ denotes the marginal density of Y.

The following results demonstrate that the excess BLS risk under P is equivalent to the standard least square risk under P^b. This equivalence plays a fundamental role in deriving the convergence rates of the proposed UNNLR algorithm with respect to the BLS loss.

Theorem A.1. *Let* P *be the underlying probability measure and* P^b *be the balanced measure defined by* (A.1). *Let* L_{LS} *and* L_{bal} *denote the standard and BLS losses, respectively, as in* (2.1). *Use* $f^*_{L_{\mathrm{bal}},\mathrm{P}}(x)$ *and* $f^*_{L_{\mathrm{LS}},\mathrm{P}^b}(x)$ *denote their respective Bayes regression functions. Then*

$$f^*_{L_{\mathrm{LS}},\mathrm{P}^b}(x) = f^*_{L_{\mathrm{bal}},\mathrm{P}}(x) = \frac{\int_{\mathcal{Y}} y\, p(x,y)/p_Y(y)\, dy}{\int_{\mathcal{Y}} p(x,y)/p_Y(y)\, dy}. \tag{A.2}$$

Theorem A.2. *Under the same notation as above, for any measurable regression function* $f : \mathcal{X} \to \mathcal{Y}$, *the following equality holds:*

$$\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} = \mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b}.$$

Therefore, standard regression techniques and theoretical results for the MSE loss can be directly extended to the BLS setting by considering the transformed measure P^b. This theoretical foundation supports the convergence analysis of both the under-sampling k-NN regressor (UkNNR) and the UNNLR.

A.2. Error analysis for the UkNNR

In this subsection, we analyze the estimation error of the UkNNR by decomposing the term $\|\widehat{f}^{k,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty$ into interpretable components. Specifically, we separately bound the *sample error*, *approximation error*, and *under-sampling error*. These three terms constitute the foundation for the convergence results and minimax lower bounds presented in Theorems 4.3 and 4.4 in Section 4.1. Detailed proofs for all results in Sections A.2.1–A.2.3 are provided in Section B.3.

A key difficulty lies in the fact that the under-sampling procedure alters the distribution of the training data. In particular, the under-sampled subset D^u_n no longer follows P. To handle this issue, we first derive the explicit form of the probability distribution of the accepted samples, as shown in the following lemma.

Lemma A.3. *Let* $p(x,y)$ *be the density function corresponding to* P, *and let* $\widehat{p}_Y(y)$ *be the histogram density estimator defined in* (3.1). *Let* P^u *denote the probability distribution of the accepted samples obtained by the under-sampling strategy described in Section 3. Then the density function of* P^u *can be expressed as*

$$p^u(x,y) = \frac{p(x,y)/\widehat{p}_Y(y)}{\int_{\mathcal{Y}} p(y)/\widehat{p}_Y(y)\, dy}.$$

Furthermore, the Bayes regression function with respect to the least square loss under P^u *is given by*

$$f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x) = \frac{\int_{\mathcal{Y}} y\, p(x,y)/\widehat{p}_Y(y)\, dy}{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\, dy}. \tag{A.3}$$

From Lemma A.3, we have the following inequality:

$$\|\widehat{f}^{k,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty \leq \|\widehat{f}^{k,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty + \|f^*_{L_{\mathrm{LS}},\mathrm{P}^u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty. \tag{A.4}$$

The first term on the right-hand side corresponds to the estimation error of applying the k-NN regressor to the under-sampled subset D_n^u, while the second term—termed the *under-sampling error*—arises from the distributional discrepancy introduced by the under-sampling strategy.

To analyze the first term, we define $\overline{f}^{k,u} : \mathcal{X} \to \mathbb{R}^d$ as the conditional expectation of $\widehat{f}^{k,u}$ given D_n^u

$$\overline{f}^{k,u}(x) = \mathbb{E}\Big[\widehat{f}^{k,u}(x) \mid D_n^u\Big] = \frac{1}{k}\sum_{i=1}^{k} Y^u_{(i)}(x). \tag{A.5}$$

Here, $\overline{f}^{k,u}$ represents the expected prediction of $\widehat{f}^{k,u}$ conditioned on the under-sampled dataset. This leads to the following error decomposition:

$$\|\widehat{f}^{k,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty \leq \|f^*_{L_{\mathrm{LS}},\mathrm{P}^u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty + \|\widehat{f}^{k,u} - \overline{f}^{k,u}\|_\infty + \|\overline{f}^{k,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty. \tag{A.6}$$

Each term in (A.6) corresponds to a distinct component of the total estimation error:

- The first term represents the *under-sampling bias*, quantifying the deviation introduced by the reweighted sampling distribution.
- The second term, the *sample error*, captures the stochastic variability associated with finite-sample estimation on D_n^u.
- The third term, the *approximation error*, measures how closely the k-NN estimator approximates the Bayes regressor under P^u.

This three-term decomposition provides the analytical basis for the theoretical convergence results of the UkNNR and extends naturally to the UNNLR framework, where multiple under-sampled regressors are aggregated to further reduce variance and stabilize learning under long-tailed label distributions.

A.2.1. Bounding the sample error term

We now establish the predicted inequality for the under-sampling posterior function $\widehat{f}^{k,u}$ under the L_∞-norm. This inequality serves as a key component in the convergence analysis of the proposed estimator.

Proposition A.4. *Let $\widehat{f}^{k,u}$ and $\overline{f}^{k,u}$ be defined by* (3.3) *and* (A.5)*, respectively. For all sufficiently large n, with a probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/n^2$, we then have*

$$\|\widehat{f}^{k,u} - \overline{f}^{k,u}\|_\infty \lesssim \sqrt{\log s_u / k}. \tag{A.7}$$

This result provides a probabilistic upper bound on the deviation of the under-sampling k-NN estimator from its conditional expectation, which quantifies the stochastic fluctuation introduced by random sampling.

A.2.2. Bounding the approximation error term

We next consider the deterministic approximation error, which measures the L_∞-distance between $\overline{f}^{k,u}$ and the Bayes function $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}$. The following proposition shows that this error can be made arbitrarily small by selecting k appropriately.

Proposition A.5. *Let $\overline{f}^{k,u}$ be defined by* (A.5)*. Assume that Assumption 4.2 holds. For all sufficiently large n, with a probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/n^2$, the following then holds*

$$\|\overline{f}^{k,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty \lesssim (k/s_u)^{\alpha/d}.$$

This inequality captures the deterministic bias induced by local averaging in the k-NN regression, which depends on the smoothness exponent α and the intrinsic data dimension d.

A.2.3. Bounding the under-sampling error term

The following result characterizes the discrepancy between the two Bayes functions, $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}$ and $f^*_{L_{\mathrm{LS}},\mathrm{P}^b}$, caused by the histogram-based under-sampling approximation of the labels' distribution. The distance between them decays polynomially with respect to the number of training samples, which is crucial for establishing the overall convergence rates of both the under-sampling and the under-bagging regressors.

Proposition A.6. *Let $f^*_{L_{\mathrm{LS}},\mathrm{P}^b}(x)$ and $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)$ be defined by* (A.2) *and* (A.3)*, respectively. Assume that Assumption 4.2 holds. For all sufficiently large values of n, with a probability P^n of at least $1 - 1/n^2$, the following then holds*

$$\|f^*_{L_{\mathrm{LS}},\mathrm{P}^b} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty \leq (n/\log n)^{-\beta/(2\beta+d)}.$$

This bound quantifies how the histogram density estimation of the label distribution influences the model bias, and serves as a foundation for analyzing the asymptotic behavior of UNNLR.

A.3. Error analysis for UNNLR

We now turn to the complete convergence analysis of the proposed UNNLR. We first show that the UNNLR estimator can be equivalently expressed as a weighted k-NN model, which facilitates its theoretical treatment within the standard nonparametric regression framework.

Let $X_{(i)}(x)$ denote the i-th nearest neighbor of x in D_n under the Euclidean distance, and let $Y_{(i)}(x)$ be its corresponding label. For each bootstrap iteration $1 \leq b \leq B$, the b-th under-sampled estimator can be written as

$$\widehat{f}^{b,u}(x) = \sum_{i=1}^{n} V_i^{b,u}(x) Y_{(i)}(x),$$

where $V_i^{b,u}(x) = 1/k$ if $\sum_{j=1}^{i} Z^b(X_{(j)}(x), Y_{(j)}(x)) \leq k$, but 0 otherwise. Here, $Z^b(x,y)$, $1 \leq b \leq B$, are i.i.d. Bernoulli random variables with the parameter $a(x,y)$.

Accordingly, the UNNLR estimator in (3.6) can be expressed as

$$\widehat{f}^{B,u}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}^{b,u}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} V_i^{b,u}(x) Y_{(i)}(x). \tag{A.8}$$

To analyze the term $\|\widehat{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty$, we introduce the idealized bagged nearest-neighbor estimator by averaging over infinitely many bootstrap replicates. Define

$$\widetilde{f}^{B,u}(x) = \mathbb{E}^B_{\mathrm{P}_Z}\Big[\widehat{f}^{B,u}(x) \mid \{(X_i, Y_i)\}_{i=1}^n\Big] := \sum_{i=1}^n \overline{V}_i^u(x) Y_{(i)}(x), \tag{A.9}$$

where the averaged weight is

$$\overline{V}_i^u(x) = \mathbb{E}_{\mathrm{P}_Z}\Big[V_i^{b,u}(x) \mid \{(X_i, Y_i)\}_{i=1}^n\Big]. \tag{A.10}$$

By the law of large numbers, for any $x \in \mathcal{X}$, we have $\widehat{f}^{B,u}(x) \to \widetilde{f}^{B,u}(x)$ almost surely as $B \to \infty$. The corresponding population version of $\widetilde{f}^{B,u}$ is defined by

$$\overline{f}^{B,u}(x) = \mathbb{E}\Big[\widetilde{f}^{B,u}(x) \mid X_1, \ldots, X_n\Big] = \sum_{i=1}^n \overline{V}_i^u(x) f^u(X_{(i)}(x)), \tag{A.11}$$

where the conditional expectation is taken with respect to $(\mathrm{P}^u)^n_{Y|X}$.

On the basis of the above definitions, we can decompose the overall estimation error of UNNLR as

$$\begin{aligned}\|\widehat{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty \le \|\widehat{f}^{B,u} - \widetilde{f}^{B,u}\|_\infty + \|\overline{f}^{B,u} - \widetilde{f}^{B,u}\|_\infty \\ + \|\overline{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty + \|f^*_{L_{\mathrm{LS}},\mathrm{P}^u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty.\end{aligned} \tag{A.12}$$

Each term on the right-hand side of (A.12) has a distinct interpretation.

- The first term, $\|\widehat{f}^{B,u} - \widetilde{f}^{B,u}\|_\infty$, is the *bagging error*, reflecting the finite number of bootstrap replications.
- The second term represents the *bagged sample error*, arising from the finite-sample stochastic variability in the ensemble.
- The third term quantifies the *bagged approximation error*, corresponding to the deterministic bias in the ensemble of base regressors.
- The final term, $\|f^*_{L_{\mathrm{LS}},\mathrm{P}^u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty$, is the *under-sampling error*, inherited from the single UkNNR model.

This comprehensive decomposition highlights how bagging effectively reduces both the variance and bias components of the UkNNR, yielding the improved stability and generalization performance observed in UNNLR under long-tailed regression settings.

A.3.1. Bounding the bagging error term

We first provide a probabilistic upper bound for the bagging error term, which quantifies the deviation between the finite bagging estimator $\widehat{f}^{B,u}$ and its infinite bagging counterpart $\widetilde{f}^{B,u}$. This term naturally depends on the number of bagging rounds B.

Proposition A.7. *Let $\widehat{f}^{B,u}$ and $\widetilde{f}^{B,u}$ be defined as in* (A.8) *and* (A.9)*, respectively. Suppose that*

$$9B \ge 2(2d+3)\log n.$$

For all sufficiently large values of n, with a probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ of at least $1 - 1/n^2$, we then have

$$\|\widehat{f}^{B,u} - \widetilde{f}^{B,u}\|_\infty \lesssim \sqrt{\log n / B}.$$

This result demonstrates that the bagging error decreases at a rate of $O((\log n/B)^{1/2})$, implying that a moderate number of bootstrap rounds is sufficient to ensure statistical stability. Consequently, UNNLR achieves both computational efficiency and robust variance reduction without requiring excessively large ensemble sizes.

A.3.2. Bounding the bagged approximation error term

We now establish an upper bound on the deterministic bias term between the expected bagged estimator $\overline{f}^{B,u}$ and the Bayes function $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}$. The resulting inequality consists of two components: The first term arises from the local averaging bias associated with the ratio k/s and the regression function's smoothness, while the second term captures the influence of the under-sampling mechanism.

Proposition A.8. *Let* $\overline{f}^{B,u}$ *be defined by* (A.11) *with* $k \geq \lceil 48(2d+9)\log n \rceil$, *and let* $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}$ *be given by* (A.3). *Assume that* P *satisfies Assumption 4.2. Moreover, suppose that*

$$s\exp\left(-\frac{(sn_{(1)}/n-k)^2}{2n_{(N)}}\right) \leq nh\underline{c}/2.$$

For all sufficiently large values of n, with a probability P^n *not less than* $1-1/n^2$, *the following then holds:*

$$\|\overline{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty \lesssim (k/s)^{\alpha/d} + \exp(-(s-k)^2/(2n)).$$

The first term, $(k/s)^{\alpha/d}$, represents the approximation bias inherited from local smoothing in the k-NN regression. The exponential term $\exp[-(s-k)^2/(2n)]$ quantifies the residual effect of stochastic variation in the subsampling process. Together, these results imply that, with appropriate parameter selection, the bias of the bagged ensemble is asymptotically negligible.

A.3.3. Bounding the bagged sample error term

We next derive the oracle inequality for the under-bagged regression estimator $\widetilde{f}^{B,u}$ in terms of the L_∞-norm. This inequality provides the statistical foundation for the convergence analysis of the UNNLR algorithm.

Proposition A.9. *Let* $\widetilde{f}^{B,u}(x)$ *and* $\overline{f}^{B,u}(x)$ *be defined by* (A.9) *and* (A.11), *respectively. For all sufficiently large values of n, with a probability* $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ *of at least* $1-1/n^2$, *the following then holds*

$$\|\widetilde{f}^{B,u} - \overline{f}^{B,u}\|_\infty \lesssim \sqrt{\frac{s\log n}{nk}}.$$

This result bounds the random fluctuation of the empirical bagged estimator around its conditional mean, showing that the deviation diminishes with the sample size n and the number of nearest neighbors k. In combination with the previous bounds, it ensures that the overall estimation error of UNNLR converges optimally in both rate and probability.

B. Proofs

Before presenting the detailed proofs, we first recall two classical concentration inequalities that will be repeatedly used throughout this section. Lemma B.1 is Hoeffding's inequality [31], and Lemma B.2 is Bernstein's inequality [30]. Both results are standard in statistical learning theory and can be found in textbooks such as [32].

Lemma B.1 (Hoeffding's inequality)**.** *Let $a < b$ be two real numbers, $n \geq 1$, and let $\xi_1, \ldots, \xi_n$ be independent random variables satisfying $\xi_i \in [a, b]$ for all $1 \leq i \leq n$. For all $\tau > 0$, we then have*

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i) \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq e^{-\tau}.$$

Lemma B.2 (Bernstein's inequality)**.** *Let $B > 0$ and $\sigma > 0$ be real numbers, and let $n \geq 1$. Suppose that $\xi_1, \ldots, \xi_n$ are independent random variables such that $\mathbb{E}_{\mathrm{P}}\xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_{\mathrm{P}}\xi_i^2 \leq \sigma^2$ for all $i = 1, \ldots, n$. For all $\tau > 0$, we then have*

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq e^{-\tau}.$$

B.1. Proofs related to HDE

Although we only outline the key steps of the HDE procedure, this simplified version already captures the essential intuition and is of considerable practical relevance.

Let $\Delta_h := (A_j)_{1 \leq j \leq h^{-1}}$ be a cubic partition of bandwidth $h > 0$, where each cell is given by $A_j := [(j-1)h, jh)$. The histogram density estimator in (3.1) can then be written as

$$\widehat{p}_Y(y) := \frac{1}{h}\sum_{j=1}^{h^{-1}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{Y_i \in A_j\}\right)\mathbf{1}_{A_j}(y). \tag{B.1}$$

To analyze the estimation error $\|\widehat{p}_Y - p_Y\|_\infty$, we introduce its population version as follows:

$$\widetilde{p}_Y(y) := \frac{1}{h}\sum_{j=1}^{h^{-1}}\mathrm{P}_Y(Y \in A_j)\,\mathbf{1}_{A_j}(y). \tag{B.2}$$

Thus, the total estimation error can be decomposed as

$$\|\widehat{p}_Y - p_Y\|_\infty \leq \|\widehat{p}_Y - \widetilde{p}_Y\|_\infty + \|\widetilde{p}_Y - p_Y\|_\infty,$$

where the first term corresponds to the *sample error*, and the second term represents the *approximation error*.

The following proposition establishes the β-Hölder smoothness of the label density p_Y under Assumption 4.2.

Proposition B.3. *Let Assumption 4.2 hold. Then $p_Y(y)$ is β-Hölder continuous and $\|p_Y\|_\infty < \infty$.*

Proof of Proposition B.3. Since $p_Y(y) = \int p(x,y)\,dx$, for any $y, y' \in [0,1]$, we have

$$\begin{aligned}|p_Y(y) - p_Y(y')| &= \left|\int_{\mathcal{X}} p(x,y)\,dx - \int_{\mathcal{X}} p(x,y')\,dx\right| \\ &\le \int_{\mathcal{X}} |p(x,y) - p(x,y')|\,dx \le c_\beta \mu(\mathcal{X})|y - y'|^\beta.\end{aligned}$$

Moreover

$$|p_Y(y)| = \left|\int_{\mathcal{X}} p(x,y)\,dx\right| \le \overline{c}\,\mu(\mathcal{X}) < \infty.$$

This completes the proof. □

Proposition B.4. *Let Δ_h be a cubic partition with a bandwidth $h > 0$. If p_Y is β-Hölder continuous, then for all $h > 0$*

$$\|\widetilde{p}_Y - p_Y\|_\infty \lesssim h^\beta.$$

Proof of Proposition B.4. By the β-Hölder continuity of p_Y, for any $y, y' \in A_j$, we have

$$|p_Y(y) - p_Y(y')| \lesssim |y - y'|^\beta \lesssim h^\beta.$$

For any $y \in \mathbb{R}^d$, we then have

$$\begin{aligned}|p_Y(y) - \widetilde{p}_Y(y)| &= \left|\frac{1}{h}\int_{A(y)} p_Y(y)\,d\mu(y') - \frac{1}{h}\int_{A(y)} p_Y(y')\,d\mu(y')\right| \\ &\le \frac{1}{h}\int_{A(y)} |p_Y(y) - p_Y(y')|\,d\mu(y') \lesssim \frac{1}{h}\cdot h^{1+\beta} \lesssim h^\beta,\end{aligned}$$

where $A(y)$ denotes the cell containing y. Hence, the result follows. □

We next provide an oracle inequality for the histogram estimator under the L_∞-norm.

Proposition B.5. *Let Δ_h be a cubic partition with a bandwidth $h \in (0,1]$. Suppose that the density function p_Y satisfies $\|p_Y\|_\infty < \infty$. For all $\tau > 0$ and $n \ge 1$, with a probability of at least $1 - e^{-\tau}$, the following then holds:*

$$\|\widehat{p}_Y - \widetilde{p}_Y\|_\infty \lesssim \sqrt{\frac{\|p_Y\|_\infty(\tau + \log(\frac{1}{h}))}{nh}} + \frac{\|p_Y\|_\infty(\tau + \log(\frac{1}{h}))}{nh}.$$

Proof of Proposition B.5. The proof follows from applying Bernstein's inequality to the empirical process over the partition cells $\{A_j\}$. Let $A \subset \mathbb{R}$ be measurable, and define

$$\xi_i := \mathbf{1}_A \pi_i - \mathbb{E}_{\mathrm{P}} \mathbf{1}_A,$$

where π_i is the projection onto the i-th observation. Clearly, $\mathbb{E}_{\mathrm{P}^n}\xi_i = 0$, $\|\xi_i\|_\infty \le 1$, and $\mathbb{E}_{\mathrm{P}^n}\xi_i^2 \le \mathrm{P}(A)$. Applying Bernstein's inequality yields, with a probability of at most $2e^{-\tau}$

$$|\mathbb{E}_{\mathrm{D}}\mathbf{1}_A - \mathbb{E}_{\mathrm{P}}\mathbf{1}_A| \ge \sqrt{\frac{2\mathrm{P}(A)\tau}{n}} + \frac{2\tau}{3n}. \tag{B.3}$$

Following a union bound over $J := \{j : A_j \cap [0,1] \ne \emptyset\}$ with $|J| \le (2/h)^d$, and using standard capacity arguments, we obtain the desired bound. The remaining steps follow from bounding $\mathrm{P}_Y(A_j)/\mu(A_j) \le \|p_Y\|_\infty$ and $\log(2|J|) \le 2\log(2/h)$. Substituting these estimates into (B.3) completes the proof. □

Theorem B.6. *Let $\widehat{p_Y}$ be the histogram estimator as in* (3.1) *with a bandwidth $h > 0$, and suppose that p_Y is β-Hölder continuous with $\|p_Y\|_\infty < \infty$. By choosing*

$$h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+d}},$$

with a probability of at least $1 - 2/n^2$, the following holds:

$$\|\widehat{p_Y} - p_Y\|_\infty \lesssim \left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+d}}. \tag{B.4}$$

Proof of Theorem B.6. Combining Propositions B.5 and B.4, we obtain, with a probability of at least $1 - 2e^{-\tau}$,

$$\|\widehat{p_Y} - p_Y\|_\infty \le \|\widehat{p_Y} - \widetilde{p_Y}\|_\infty + \|\widetilde{p_Y} - p_Y\|_\infty \lesssim \sqrt{\frac{\|p_Y\|_\infty(\tau + \log(\frac{1}{h}))}{nh}} + h^\beta.$$

Setting $\tau = 2\log n$ and $h_n = (\log n/n)^{1/(2\beta+d)}$ yields the stated convergence rate. □

B.2. Proofs related to Section A.1

Proof of Theorem A.1. We first consider the Bayes function associated with the loss L_{bal} under the probability measure P. By the law of total expectation, we have

$$\begin{aligned}\mathcal{R}_{L_{\text{bal}},\mathrm{P}}(f) &= \mathbb{E}_{\mathrm{P}_X}\left[\mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{(f(x)-y)^2}{p_Y(y)}\,\middle|\, X = x\right]\right] \\ &= \mathbb{E}_{\mathrm{P}_X}\left[f(x)^2\mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{1}{p_Y(y)}\middle| X = x\right] - 2f(x)\mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{y}{p_Y(y)}\middle| X = x\right] + \mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{y^2}{p_Y(y)}\middle| X = x\right]\right].\end{aligned}$$

Minimizing the inner expression with respect to $f(x)$ gives

$$f^*_{L_{\text{bal}},\mathrm{P}}(x) = \frac{\mathbb{E}_{\mathrm{P}_{Y|X}}[y/p_Y(y)\,|\,X = x]}{\mathbb{E}_{\mathrm{P}_{Y|X}}[1/p_Y(y)\,|\,X = x]} = \frac{\int_{\mathcal{Y}} y\, p(y|X = x)/p_Y(y)\,dy}{\int_{\mathcal{Y}} p(y|X = x)/p_Y(y)\,dy}. \tag{B.5}$$

Since $p(y|X = x) = p(x, y)/p_X(x)$, the expression above can be rewritten as

$$f^*_{L_{\text{bal}},\mathrm{P}}(x) = \frac{\int_{\mathcal{Y}} y\, p(x, y)/p_Y(y)\,dy}{\int_{\mathcal{Y}} p(x, y)/p_Y(y)\,dy}.$$

Next, consider the Bayes function $f^*_{L_{\text{LS}},\mathrm{P}^b}$ corresponding to the least squares loss under P^b. We have $f^*_{L_{\text{LS}},\mathrm{P}^b}(x) = \mathbb{E}_{\mathrm{P}^b_{Y|X}}[Y|X = x]$. From (A.1), the conditional density under P^b satisfies

$$p^b(y|X = x) = \frac{p^b(x, y)}{p^b_X(x)} = \frac{p(x, y)/p_Y(y)}{\int_{\mathcal{Y}} p(x, y)/p_Y(y)\,dy}.$$

Hence,

$$f^*_{L_{\text{LS}},\mathrm{P}^b}(x) = \int_{\mathcal{Y}} y\, p^b(y|X = x)\,dy = \frac{\int_{\mathcal{Y}} y\, p(x, y)/p_Y(y)\,dy}{\int_{\mathcal{Y}} p(x, y)/p_Y(y)\,dy}.$$

Comparing with (B.5) confirms the desired identity. □

Proof of Theorem A.2. We now establish the connection between the balanced risk and the least squares risk under the reweighted measure P^b.

Starting with the balanced risk, by the law of total expectation, we have

$$\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} = \mathbb{E}_{\mathrm{P}_X}\left[\mathbb{E}_{\mathrm{P}_{Y|X}}\left(\frac{(f(x)-y)^2 - (f^*_{L_{\mathrm{bal}},\mathrm{P}}(x)-y)^2}{p_Y(y)} \,\middle|\, X = x\right)\right]. \tag{B.6}$$

A straightforward expansion gives

$$\begin{aligned}\mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{(f(x)-y)^2 - (f^*_{L_{\mathrm{bal}},\mathrm{P}}(x)-y)^2}{p_Y(y)} \,\middle|\, X = x\right] &= (f(x) - f^*_{L_{\mathrm{bal}},\mathrm{P}}(x))\,\mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{f(x) + f^*_{L_{\mathrm{bal}},\mathrm{P}}(x) - 2y}{p_Y(y)} \,\middle|\, X = x\right] \\ &= (f(x) - f^*_{L_{\mathrm{bal}},\mathrm{P}}(x))^2 \mathbb{E}_{\mathrm{P}_{Y|X}}\left[\frac{1}{p_Y(y)} \,\middle|\, X = x\right], \end{aligned} \tag{B.7}$$

where we have used Theorem A.1 to substitute $f^*_{L_{\mathrm{bal}},\mathrm{P}}(x)\,\mathbb{E}_{\mathrm{P}_{Y|X}}[1/p_Y(y)|X = x] = \mathbb{E}_{\mathrm{P}_{Y|X}}[y/p_Y(y)|X = x]$. Consequently,

$$\begin{aligned}\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} &= \int_{\mathcal{X}} p_X(x)(f(x) - f^*_{L_{\mathrm{bal}},\mathrm{P}}(x))^2 \int_{\mathcal{Y}} \frac{p(x,y)}{p_Y(y)p_X(x)}\,dy\,dx \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} (f(x) - f^*_{L_{\mathrm{bal}},\mathrm{P}}(x))^2 \frac{p(x,y)}{p_Y(y)}\,dy\,dx. \end{aligned} \tag{B.8}$$

Next, consider the least squares risk under the balanced distribution P^b. By definition, we have

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b} = \int_{\mathcal{X}} (f(x) - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}(x))^2 p^b_X(x)\,dx. \tag{B.9}$$

The marginal density of P^b satisfies

$$p^b_X(x) = \int_{\mathcal{Y}} p^b(x,y)\,dy = \int_{\mathcal{Y}} \frac{p(x,y)}{p_Y(y)}\,dy.$$

Substituting into (B.9) yields

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b} = \int_{\mathcal{X}}\int_{\mathcal{Y}} (f(x) - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}(x))^2 \frac{p(x,y)}{p_Y(y)}\,dy\,dx.$$

Combining this with (B.8) and the identity (A.2) completes the proof. □

B.3. Proofs related to Section A.2

Proof of Lemma A.3. Let $\mathrm{P}_{X,Y,Z} = \mathrm{P}_{X,Y} \times \mathrm{P}_{Z|(X,Y)}$ denote the joint probability measure. The probability that a sample (X, Y) is accepted under the under-sampling strategy is

$$\begin{aligned}\mathrm{P}_{X,Y,Z}(Z(X,Y) = 1) &= \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{P}(Z(X,Y) = 1|(X,Y) = (x,y))\,d\mathrm{P}(x,y) \\ &= \int_{\mathcal{Y}} p_Y(y) \int_{\mathcal{X}} a(x,y)p(x|Y = y)\,dx\,dy.\end{aligned}$$

For any measurable sets $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$

$$\mathrm{P}_{X,Y,Z}(X \in A, Y \in B, Z(X,Y) = 1) = \int_B p_Y(y) \int_A a(x,y) p(x|Y=y)\, dx\, dy.$$

Hence, the distribution of accepted samples satisfies

$$\mathrm{P}^u(X \in A, Y \in B) = \frac{\int_B \int_A a(x,y) p_Y(y) p(x|Y=y)\, dx\, dy}{\int_{\mathcal{Y}} \int_{\mathcal{X}} a(x,y) p_Y(y) p(x|Y=y)\, dx\, dy}.$$

Using $p_Y(y)p(x|Y=y) = p(x,y)$ and (3.4), we obtain

$$\mathrm{P}^u(X \in A, Y \in B) = \frac{\int_B \int_A p(x,y)/\widehat{p}_Y(y)\, dx\, dy}{\int_{\mathcal{Y}} \int_{\mathcal{X}} p(x,y)/\widehat{p}_Y(y)\, dx\, dy}. \tag{B.10}$$

Thus, the density function under P^u is

$$p^u(x,y) = \frac{p(x,y)/\widehat{p}_Y(y)}{\int_{\mathcal{Y}} p(y)/\widehat{p}_Y(y)\, dy}.$$

For the Bayes function under P^u, note that the marginal density is

$$p^u_X(x) = \frac{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\, dy}{\int_{\mathcal{Y}} p(y)/\widehat{p}_Y(y)\, dy},$$

so that

$$p^u(y|X=x) = \frac{p^u(x,y)}{p^u_X(x)} = \frac{p(x,y)/\widehat{p}_Y(y)}{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\, dy}.$$

Therefore,

$$f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x) = \mathbb{E}_{\mathrm{P}^u_{Y|X}}[Y|X=x] = \int_{\mathcal{Y}} y\, p^u(y|X=x)\, dy = \frac{\int_{\mathcal{Y}} y\, p(x,y)/\widehat{p}_Y(y)\, dy}{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\, dy}.$$

This completes the proof. □

B.3.1. Proofs related to Section A.2.1

The following lemma shows that if the marginal density functions of the probability measure P are α-Hölder continuous, then $f^*_{L_{\mathrm{LS}},P^u}$ also retains this property. This result will be frequently invoked in the sequel and plays a crucial role in proving Propositions A.4 and A.5.

Lemma B.7. *Let $f_{L_{\mathrm{LS}},\mathrm{P}^u}$ be defined as in* (A.3)*. Assume that* P *satisfies Assumption 4.2 and that* P_X *is uniformly distributed over* $[0,1]^d$*. Then, with a probability of at least* $1 - 1/n^2$ *under* P^n*, for any* $x, x' \in \mathcal{X}$*, we have*

$$|f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x') - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)| \lesssim \|x - x'\|^\alpha. \tag{B.11}$$

Proof of Lemma B.7. By the definition of $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)$, we have

$$\begin{aligned}|f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x')| &\le \left|\frac{\int_{\mathcal{Y}} yp(x,y)/\widehat{p}_Y(y)\,dy}{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\,dy} - \frac{\int_{\mathcal{Y}} yp(x',y)/\widehat{p}_Y(y)\,dy}{\int_{\mathcal{Y}} p(x',y)/\widehat{p}_Y(y)\,dy}\right| \\ &\le \frac{\int_{\mathcal{Y}} yp(x',y)/\widehat{p}_Y(y)\,dy \,|\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\,dy - \int_{\mathcal{Y}} p(x',y)/\widehat{p}_Y(y)\,dy|}{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\,dy \cdot \int_{\mathcal{Y}} p(x',y)/\widehat{p}_Y(y)\,dy} \\ &\quad + \frac{|\int_{\mathcal{Y}} yp(x,y)/\widehat{p}_Y(y)\,dy - \int_{\mathcal{Y}} yp(x',y)/\widehat{p}_Y(y)\,dy|}{\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\,dy}.\end{aligned}$$

By Theorem B.6, for a sufficiently large n, with a probability of at least $1-2/n^2$, we have $\underline{c}\mu(\mathcal{X})/2 \le \widehat{p}_Y(y) \le 2\mu(\mathcal{X})\overline{c}$. Consequently,

$$\begin{aligned}\int_{\mathcal{Y}} yp(x',y)/\widehat{p}_Y(y)\,dy &\le 2/\underline{c}\int_{\mathcal{Y}} p(x',y)\,dy \le 2\overline{c}/(\underline{c}\mu(\mathcal{X})), \\ \int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\,dy &\ge \underline{c}/(2\overline{c}\mu(\mathcal{X})).\end{aligned}$$

Thus,

$$\begin{aligned}|f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x') - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)| &\lesssim \left|\int_{\mathcal{Y}} p(x,y)/\widehat{p}_Y(y)\,dy - \int_{\mathcal{Y}} p(x',y)/\widehat{p}_Y(y)\,dy\right| \\ &\lesssim \int_{\mathcal{Y}} |p(x,y) - p(x',y)|\,dy \lesssim \|x - x'\|^\alpha,\end{aligned}$$

where the last step follows from Assumption 4.2. This completes the proof. □

Lemma B.8. *Let Assumption 4.2 hold. Let n_i denote the number of samples in interval $[(i-1)h, ih)$ in Section 3.1. Sort these counts in ascending order:*

$$n_{(1)} \le n_{(2)} \le \cdots \le n_{(N)}. \tag{B.12}$$

For a sufficiently large n and all $1 \le i \le n$, we then have

$$nh\underline{c}\mu(\mathcal{X})/2 \le n_{(i)} \le 2nh\overline{c}\mu(\mathcal{X})$$

with a probability of at least $1 - 1/n^2$.

Proof of Lemma B.7. For any $1 \le i \le h^{-1}$, define

$$\zeta_i := \mathbf{1}\{(i-1)h \le Y_i < ih\} - \mathrm{P}_Y((i-1)h \le Y_i < ih).$$

Then the values of ζ_i's are independent with $\mathbb{E}_{\mathrm{P}}[\zeta_i] = 0$ and $\mathbb{E}_{\mathrm{P}}\zeta_i^2 \le 1/4$. Applying Bernstein's inequality (Lemma B.2), for $\tau > 0$, we have

$$\mathrm{P}^n\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{(i-1)h \le Y_i < ih\} \ge \mathrm{P}_Y((i-1)h \le Y_i < ih) + \sqrt{\tfrac{\tau}{2n}} + \tfrac{2\tau}{3n}\right) \le e^{-\tau}.$$

Setting $\tau = 3\log n$ gives

$$\mathrm{P}^n(n_{(i)}/n \geq \mathrm{P}_Y((i-1)h \leq Y_i < ih) + 2\sqrt{\log n/n}) \leq 1/n^3.$$

A similar argument for the complementary event yields

$$\mathrm{P}^n(\mathrm{P}_Y((i-1)h \leq Y_i < ih) \geq n_{(i)}/n + 2\sqrt{\log n/n}) \leq 1/n^3.$$

Since $\underline{c}h\mu(\mathcal{X}) \leq \mathrm{P}_Y((i-1)h \leq Y_i < ih) \leq \overline{c}h\mu(\mathcal{X})$, a union bound yields

$$\mathrm{P}^n\Big(nh\underline{c}\mu(\mathcal{X})/2 \leq n_{(i)} \leq 2nh\overline{c}\mu(\mathcal{X}),\ \forall 1 \leq i \leq h^{-1}\Big) \geq 1 - 2/n^2.$$

This completes the proof. □

Lemma B.9. *Let $Z_1, \ldots, Z_n$ be independent zero-mean real random variables with $|Z_i| \leq C$. Let $(v_1, \ldots, v_n)$ be a weight vector with $v_{\max} = \max_i |v_i| > 0$. Then for all $\varepsilon > 0$,*

$$\mathrm{P}\left(\sum_{i=1}^n v_i Z_i \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{2C^2 v_{\max}\sum_{i=1}^n |v_i|}\right).$$

Proof of Lemma B.9. For $1 \leq i \leq n$, note that $|v_i Z_i| \leq C|v_i|$. By Hoeffding's inequality (Lemma B.1),

$$\mathrm{P}\left(\sum_{i=1}^n v_i Z_i \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (2Cv_i)^2}\right) = 2\exp\left(-\frac{\varepsilon^2}{2C^2 v_{\max}\sum_{i=1}^n |v_i|}\right).$$

□

B.3.2. Proofs related to Section A.3.2

To bound the bagged approximation error $\|\overline{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty$, we first examine the theoretical properties of the weight function $\overline{V}^u_i(x)$. By the definition of $\overline{f}^{B,u}$ and $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}$, we have

$$\begin{aligned}
|\overline{f}^{B,u}(x) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)| &= \left|\sum_{i=1}^n \overline{V}^u_i(x) f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{(i)}(x)) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)\right| \\
&\leq \left|\sum_{i=1}^n \overline{V}^u_i(x)\big(f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{(i)}(x)) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)\big)\right| + \left|\sum_{i=1}^n \overline{V}^u_i(x) - 1\right|.
\end{aligned}$$

Let $\xi_i := Z^b(X_{(i)}(x), Y_{(i)}(x))$ for $i \geq 1$. According to (A.10), the weight $\overline{V}^u_i(x)$ can be rewritten as

$$\overline{V}^u_i(x) = \frac{1}{k}\mathrm{P}_Z\Big(\sum_{l=1}^i \xi_l \leq k,\ \xi_i = 1\Big|\{(X_i, Y_i)\}_{i=1}^n\Big) = \frac{1}{k}\sum_{j=1}^k p^u_{i,j}(x), \tag{B.13}$$

where

$$p^u_{i,j}(x) = \mathrm{P}_Z\Big(\sum_{l=1}^i \xi_l = j,\ \xi_i = 1\Big|\{(X_i, Y_i)\}_{i=1}^n\Big). \tag{B.14}$$

Hence, $p^u_{i,j}(x)$ gives the probability that, within the independent Bernoulli sequence $\{\xi_i\}$ with success probabilities $a(X_{(i)}(x), Y_{(i)}(x))$, the j-th success occurs exactly at the i-th trial.

Generalized Pascal distribution. To characterize this probability, we adopt the *Generalized Pascal (GP)* distribution [19]. Suppose that $\{\xi_i\}_{i\geq 1}$ are independent Bernoulli trials with success probabilities $\mathrm{P}(\xi_i = 1) = p_i$ and failure probabilities $\mathrm{P}(\xi_i = 0) = 1 - p_i$. The random variable X, representing the number of trials until j successes are observed, follows a GP distribution with the parameters (j, p), denoted $X \sim \mathrm{GP}(j, p)$. For $i \geq j$, define

$$\Omega(j, i) := \{\omega = \{\omega_1, \ldots, \omega_j\} : 1 \leq \omega_1 < \cdots < \omega_{j-1} < \omega_j = i\}.$$

The probability mass function is then given by

$$\mathrm{P}_{\mathrm{GP}}(X = i) = f_{\mathrm{GP}}(i; j, p) = \sum_{\omega\in\Omega(j,i)} p_\ell \prod_{r=1}^{i-1} p_r^{\mathbf{1}\{r\in\omega\}}(1 - p_r)^{\mathbf{1}\{r\notin\omega\}}, \qquad i \geq j. \tag{B.15}$$

According to (B.14), we have

$$p_{i,j}^u(x) = f_{\mathrm{GP}}(i; j, p(x)), \tag{B.16}$$

where $p(x) = (p_1(x), \ldots, p_n(x), \ldots)$ with $p_i(x) = a(X_{(i)}(x), Y_{(i)}(x))$. In Section 3.2, the acceptance probability (3.4) involves n_i, the number of samples in $[(i-1)h, ih)$, sorted as

$$n_{(1)} \leq n_{(2)} \leq \cdots \leq n_{(N)}, \quad N = h^{-1}.$$

Hence, the GP parameter vector $p(x)$ belongs to

$$\begin{aligned} S_\nu := \Big\{ p = \{p_i\}_{i=1}^\infty : (p_1, \ldots, p_n) = (\nu_{\sigma_1}, \ldots, \nu_{\sigma_n}),\ p_i = \nu_n \text{ for } i > n, \\ \text{where } (\sigma_1, \ldots, \sigma_n) \text{ is a permutation of } (1, \ldots, n) \Big\}, \end{aligned} \tag{B.17}$$

with $\nu = (\nu_1, \nu_2, \ldots)$ being defined by

$$\nu_i = \begin{cases} \dfrac{s \cdot n_{(1)}}{n \cdot n_{(N)}}, & 1 \leq i \leq n_{(N)}, \\ \dfrac{s \cdot n_{(1)}}{n \cdot n_{(j)}}, & \sum_{\ell=j+1}^N n_{(\ell)} < i \leq \sum_{\ell=j}^N n_{(\ell)},\ 1 \leq j \leq N-1, \\ \dfrac{s}{n}, & i > n. \end{cases} \tag{B.18}$$

Therefore, our analysis on $p_{i,j}^u(x)$ reduces to studying the GP distribution $f_{\mathrm{GP}}(i; j, p)$ restricted to S_ν.

Auxiliary results. We recall two key lemmas on the GP distribution from [19].

Lemma B.10 (Tail bound)**.** *Let $p \in S_\nu$ and suppose that $\sum_{i=1}^\ell p_i \geq j$. Then*

$$\sum_{i=\ell+1}^\infty f_{\mathrm{GP}}(i; j, p) \leq \exp\left(-\frac{1}{2\ell}\Big(\sum_{i=1}^\ell p_i - j\Big)^2\right).$$

Theorem B.11 (Expected truncation bound)**.** *Let $p \in S_\nu$. For any $j \leq u \leq n$ satisfying $\sum_{\ell=1}^u \nu_\ell > j$, the following holds:*

$$\sum_{i=j}^n i f_{\mathrm{GP}}(i; j, p) \leq \frac{j}{\nu_1}\Big(\frac{\nu_u}{\nu_1}\Big)^j + n\exp\left(-\frac{1}{2u}\Big(\sum_{\ell=1}^u \nu_\ell - j\Big)^2\right).$$

Bounding the bagged approximation error. The next two lemmas are essential for the proof of Proposition A.8. Lemma B.12 provides a uniform upper bound on the weighted sum of nearest-neighbor distances, while Lemma B.13 bounds the total weight $\sum_i \overline{V}_i^u(x)$.

Lemma B.12. *Let $\overline{V}_i^u(x)$ be defined as in* (A.10) *and let $R_{(i)}(x) := \|X_{(i)}(x) - x\|$. If $s\exp(-(sn_{(1)}/n - k)^2/(2n_{(N)})) \le nhc\underline{\mu}(\mathcal{X})/2$, then for a sufficiently large n, with a probability at least $1 - 3/n^2$ under P^n, we have*

$$\sum_{i=1}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) \lesssim (k/s)^{\alpha/d}, \quad \forall x \in \mathcal{X}.$$

Proof of Lemma B.12. Let $a_n = \lceil 48(2d+9)\log n \rceil$. Then for $n > n_2$ and all $x \in \mathcal{X}$, we have

$$\sup_{i \ge a_n} R_{(i)}(x) \le (2i/n)^{1/d}$$

with a probability of at least $1 - 1/n^3$. Then

$$\begin{aligned}\sum_{i=1}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) &= \sum_{i=1}^{a_n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) \\ &\le R_{(a_n)}^{\alpha}(x) + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) \lesssim (k/s)^{\alpha/d} + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x). \end{aligned} \tag{B.19}$$

For the second term in (B.19), we have

$$\sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) = k^{-1} \sum_{j=1}^{k} \sum_{i=1}^{n} p_{i,j}^u(x) (2i/n)^{\alpha/d} \lesssim k^{-1}(1/n)^{\alpha/d} \sum_{j=1}^{k} \sum_{i=1}^{n} i^{\alpha/d} p_{i,j}^u(x). \tag{B.20}$$

By Jensen's inequality, we have

$$\sum_{i=1}^{n} i^{\alpha/d} p_{i,j}^u(x) \le \Big(\sum_{i=1}^{n} i p_{i,j}^u(x) \Big)^{\alpha/d}.$$

Since $p_{i,j}^u(x) = f_{\mathrm{GP}}(i; j, p(x))$ with $p(x) \in S_\nu$, Theorem B.11 (with $u = n_{(N)}$) gives

$$\sum_{i=1}^{n} i p_{i,j}^u(x) \le n j n_{(N)}/(sn_{(1)}) + n\exp(-(sn_{(1)}/n - j)^2/(2n_{(N)})).$$

By Lemma B.8, for a sufficiently large n, $n_{(N)} \ge nhc\underline{\mu}(\mathcal{X})/2$ with a probability of at least $1 - 2/n^2$. Combining these estimates yields

$$\sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) \le k^{-1} \sum_{j=1}^{k} (2jn_{(N)}/(sn_{(1)}))^{\alpha/d}.$$

Since $g(t) = t^{\alpha/d}$ is increasing on $[0, 1]$, we have

$$k^{-1} \sum_{j=1}^{k} (j/k)^{\alpha/d} \le 2(1 + \alpha/d),$$

hence

$$\sum_{i=1}^{n} \overline{V}_i^u(x) R_{(i)}^{\alpha}(x) \lesssim (k/s)^{\alpha/d}.$$

This completes the proof. □

Lemma B.13. *Let $\overline{V}_i^u(x)$ be defined as in* (A.10) *and suppose that $k \le s$. Then for all $x \in \mathcal{X}$, we have*

$$1 - \sum_{i=1}^{n} \overline{V}_i^u(x) \le \exp(-(sNn_{(1)}/n - k)^2/(2n)),$$

where $N = h^{-1}$ and $n_{(1)}$ are defined in (B.12).

Proof of Lemma B.13. From (A.10), we have

$$\sum_{i=1}^{n} \overline{V}_i^u(x) = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{n} p_{i,j}^u(x).$$

Using (B.16), we have $\sum_{i=1}^n p_{i,j}^u(x) = \sum_{i=1}^n f_{\mathrm{GP}}(i; j, p(x))$ with $p(x) \in S_\nu$. Since $\sum_{i=1}^n p_i(x) = \frac{sNn_{(1)}}{n}$, Lemma B.10 gives

$$\sum_{i=1}^{n} p_{i,j}^u(x) \ge 1 - \exp(-(sNn_{(1)}/n - j)^2/(2n)) \ge 1 - \exp(-(sNn_{(1)}/n - k)^2/(2n)).$$

Therefore,

$$\sum_{i=1}^{n} \overline{V}_i^u(x) \ge 1 - \exp(-(sNn_{(1)}/n - k)^2/(2n)),$$

which proves the lemma. □

Proof of Proposition A.8. By the definition of $\overline{f}^{B,u}$ and $f^*_{L_{\mathrm{LS}},\mathrm{P}^u}$,

$$|\overline{f}^{B,u}(x) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)| \le \left|\sum_{i=1}^{n} \overline{V}_i^u(x)\big(f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{(i)}(x)) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(x)\big)\right| + \left|\sum_{i=1}^{n} \overline{V}_i^u(x) - 1\right|.$$

By Lemma B.7, for a sufficient large n, with a probability of at least $1 - 1/n^2$, we have

$$\|\overline{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty \lesssim \sup_{x\in\mathcal{X}}\left(\sum_{i=1}^{n} \overline{V}_i^u(x)\|X_{(i)}(x) - x\|^{\alpha}\right) + \sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n} \overline{V}_i^u(x) - 1\right|.$$

Applying Lemmas B.12 and B.13, we obtain

$$\|\overline{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty \lesssim (k/s)^{\alpha/d} + \exp(-(sNn_{(1)}/n - k)^2/(2n)), \tag{B.21}$$

which completes the proof. □

B.3.3. Proofs related to Section A.3.3

To establish Proposition A.9, we begin with a key lemma that provides a uniform bound on the bagged weights $\overline{V}_i^u(x)$ defined in (A.10).

Lemma B.14. *Let $\overline{V}_i^u(x)$ be defined as in* (A.10)*. Then for any $x \in \mathbb{R}^d$, we have*

$$\max_{1\le i\le n} \overline{V}_i^u(x) \le \frac{s}{kn}.$$

Proof of Lemma B.14. From (B.13) and (B.14), we have

$$\begin{aligned}\overline{V}_i^u(x) &= \frac{1}{k}\sum_{j=1}^{k} p_{i,j}^u(x) = \frac{1}{k}\sum_{j=1}^{k} \mathrm{P}_Z\left(\sum_{\ell=1}^{i} Z^b(X_{(\ell)}(x), Y_{(\ell)}(x)) = j,\, Z^b(X_{(i)}(x), Y_{(i)}(x)) = 1 \;\middle|\; \{(X_i, Y_i)\}_{i=1}^n\right)\\ &\le k^{-1}\mathrm{P}_Z\left(Z^b(X_{(i)}(x), Y_{(i)}(x)) = 1 \;\middle|\; \{(X_i, Y_i)\}_{i=1}^n\right) = k^{-1}a(X_{(i)}(x), Y_{(i)}(x)) \le \frac{s}{kn}.\end{aligned}$$

This completes the proof. □

Proof of Proposition A.9. By the definitions of $\widetilde{f}^{B,u}$ and $\overline{f}^{B,u}$, we can write

$$|\widetilde{f}^{B,u}(x) - \overline{f}^{B,u}(x)| = \left|\sum_{i=1}^{n} \overline{V}_i^u(x)(Y_{(i)}(x) - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{(i)}(x)))\right|.$$

For any fixed $x \in \mathcal{X}$, Lemmas B.9 and B.14 imply

$$(\mathrm{P}^u)^n_{Y|X}\left(|\widetilde{f}^{B,u}(x) - \overline{f}^{B,u}(x)| \ge \varepsilon \;\middle|\; D_n\right) \le 2\exp\left(-\frac{\varepsilon^2 kn}{2s}\right).$$

Setting $\varepsilon := \sqrt{2(2d+3)s\log n/(kn)}$, we have

$$(\mathrm{P}^u)^n_{Y|X}\left(|\widetilde{f}^{B,u}(x) - \overline{f}^{B,u}(x)| \ge \varepsilon \;\middle|\; D_n\right) \le 2n^{-(2d+3)}. \tag{B.22}$$

The bound above holds pointwise in x. To extend it uniformly over $\mathcal{X}$, let $\mathcal{S} = \{(\sigma_1, \ldots, \sigma_n) :$ all permutations of $(1, \ldots, n)$ induced by varying $x \in \mathbb{R}^d\}$. In this case,

$$\begin{aligned}&(\mathrm{P}^u)^n_{Y|X}\left(\sup_{x\in\mathbb{R}^d}(|\widetilde{f}^{B,u}(x) - \overline{f}^{B,u}(x)| - \varepsilon) > 0 \;\middle|\; D_n\right)\\ &\le (\mathrm{P}^u)^n_{Y|X}\left(\bigcup_{(\sigma_1,\ldots,\sigma_n)\in\mathcal{S}} \left|\sum_{i=1}^{n} \overline{V}^u_{i,\sigma}(Y_{\sigma_i} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{\sigma_i}))\right| > \varepsilon \;\middle|\; D_n\right)\\ &\le \sum_{(\sigma_1,\ldots,\sigma_n)\in\mathcal{S}} (\mathrm{P}^u)^n_{Y|X}\left(\left|\sum_{i=1}^{n} \overline{V}^u_{i,\sigma}(Y_{\sigma_i} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{\sigma_i}))\right| > \varepsilon \;\middle|\; D_n\right),\end{aligned}$$

where $\overline{V}^u_{i,\sigma}(x) = k^{-1}\mathrm{P}_Z(\sum_{j=1}^{i} Z^b(X_{\sigma_j}(x), Y_{\sigma_j}(x)) \le k \mid \{X_i, Y_i\}_{i=1}^n)$. For each $(\sigma_1, \ldots, \sigma_n) \in \mathcal{S}$, inequality (B.22) gives

$$(\mathrm{P}^u)^n_{Y|X}\left(\left|\sum_{i=1}^{n} \overline{V}^u_{i,\sigma}(Y_{\sigma_i} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}(X_{\sigma_i}))\right| > \varepsilon \;\middle|\; D_n\right) \le 2n^{-(2d+3)}.$$

Thus we have

$$(\mathrm{P}^u)^n_{Y|X}\left(\sup_{x\in\mathbb{R}^d}(|\widetilde{f}^{B,u}(x)-\overline{f}^{B,u}(x)|-\varepsilon)>0 \;\middle|\; D_n\right)\le 2(25/d)^d/n^3,\quad n\ge 2d.$$

By the law of total probability, we have

$$\mathrm{P}_Z^B\otimes\mathrm{P}^n\left(\|\widetilde{f}^{B,u}-\overline{f}^{B,u}\|_\infty\le\sqrt{\frac{2(2d+3)s\log n}{kn}}\right)\ge 1-\frac{2(25/d)^d}{n^3}.$$

Therefore, for all sufficiently large values of n, we have

$$\mathrm{P}_Z^B\otimes\mathrm{P}^n\left(\|\widetilde{f}^{B,u}-\overline{f}^{B,u}\|_\infty\le\sqrt{\frac{2(2d+3)s\log n}{kn}}\right)\ge 1-\frac{1}{n^2}.$$

This completes the proof of Proposition A.9. □

B.4. Proofs related to Section 4

B.4.1. Proofs related to Section 4.1

Proof of Theorem 4.3. Propositions A.4 and A.5 imply that for a sufficiently large n, we have

$$\|\overline{f}^{k,u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty\lesssim(k/s_u)^{\alpha/d},\quad \|\widehat{f}^{k,u}-\overline{f}^{k,u}\|_\infty\lesssim\sqrt{\frac{\log s_u}{k}},$$

with a probability of at least $1-2/n^2$ under $\mathrm{P}_Z\otimes\mathrm{P}^n$. Hence,

$$\begin{aligned}\|\widehat{f}^{k,u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty&\le\|\widehat{f}^{k,u}-\overline{f}^{k,u}\|_\infty+\|\overline{f}^{k,u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty\\&\lesssim(k/s_u)^{\alpha/d}+\sqrt{\frac{\log s_u}{k}}+(n/\log n)^{-\beta/(2\beta+d)}\lesssim(\log s_u/s_u)^{\alpha/(2\alpha+d)}.\end{aligned}$$

By Lemma B.8, $s_u\ge n\underline{c}\mu(\mathcal{X})/2$ with a probability of at least $1-1/n^2$. Since $g(x)=\log(x)/x$ is decreasing on $[e,\infty)$, we have

$$\|\widehat{f}^{k,u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty\lesssim(\log(n\underline{c}/2)/(n\underline{c}/2))^{\alpha/(2\alpha+d)}\lesssim(n/\log n)^{-\alpha/(2\alpha+d)}.$$

Combining this with Proposition A.6, we have for a sufficient large n,

$$\begin{aligned}\|\widehat{f}^{k,u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty&\le\|\widehat{f}^{k,u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty+\|f^*_{L_{\mathrm{LS}},\mathrm{P}^u}-f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty\\&\lesssim(n/\log n)^{-\alpha/(2\alpha+d)}+(n/\log n)^{-\beta/(2\beta+d)}\lesssim(n/\log n)^{-\gamma/(2\gamma+d)},\end{aligned}$$

with a probability of at least $1-4/n^2$, where $\gamma=\alpha\wedge\beta$. Taking the expectation with respect to P_X^b, it follows that

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(\widehat{f}^{k,u})-\mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b}\lesssim(\log n/n)^{2\alpha/(2\alpha+d)}.$$

This, together with Theorem A.2, establishes the claim. □

Proof of Theorem 4.4. By Theorem A.2,

$$\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} = \mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b}.$$

By applying Theorem 2.6.1 in [16], $C > 0$ exists such that

$$\inf_{f_n} \sup_{\mathrm{P}^b} (\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b}) \geq Cn^{-2\alpha/(2\alpha+d)},$$

where P^b denotes the class of probability measures defined by (A.1). Combining this with the equivalence above gives

$$\inf_{f_n} \sup_{\mathrm{P}} (\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}}) \geq Cn^{-2\alpha/(2\alpha+d)}.$$

This completes the proof. □

B.4.2. Proofs related to Section 4.2

Proof of Theorem 4.5. Choosing s, B, and k according to (4.4), (4.6), and (4.5), respectively, Propositions A.7, A.8, and A.9 imply that, for a sufficiently large n, we have

$$\|\widehat{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^u}\|_\infty \lesssim \sqrt{\frac{\log n}{B}} + (k/s)^{\alpha/d} + \exp(-(s-k)^2/(2n)) + \sqrt{\frac{s\log n}{kn}} \lesssim (n/\log n)^{-\alpha/(2\alpha+d)},$$

with a probability of at least $1 - 3/n^2$ under $\mathrm{P}_Z^B \otimes \mathrm{P}^n$. Combining this with Theorem A.2, we have

$$\|\widehat{f}^{B,u} - f^*_{L_{\mathrm{LS}},\mathrm{P}^b}\|_\infty \lesssim (n/\log n)^{-\gamma/(2\gamma+d)}, \qquad \gamma = \alpha \wedge \beta.$$

Taking the expectation over P_X^b, we finally obtain

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}^b}(\widehat{f}^{B,u}) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}^b} \lesssim (n/\log n)^{-2\gamma/(2\gamma+d)}.$$

This completes the proof. □