On Theoretical Limits of Learning with Label Differential Privacy

Anonymous Author(s)
Affiliation
Address
email

Abstract

Label differential privacy (DP) is designed for learning problems with private labels and public features. Although various methods have been proposed for learning under label DP, the theoretical limits remain unknown. The main challenge is to take infimum over all possible learners with arbitrary model complexity. In this paper, we investigate the fundamental limits of learning with label DP under both central and local models. To overcome the challenge above, we derive new lower bounds on testing errors that are adaptive to the model complexity. Our analyses indicate that ϵ-local label DP only enlarges the sample complexity with respect to ϵ, without affecting the convergence rate over the sample size N, except the case with heavy-tailed label. Under the central model, the performance loss due to the privacy mechanism is further weakened, such that the additional sample complexity becomes negligible. Overall, our analysis validates the promise of learning under the label DP from a theoretical perspective and shows that the learning performance can be significantly improved by weakening the DP definition to only labels.

1 Introduction

Many modern machine learning tasks require sensitive training samples that need to be protected from leakage [1]. As a standard approach for privacy protection, differential privacy (DP) [2] has been extensively studied [3–9]. However, the learning performances under original DP definition are usually far from satisfactory [10–13]. Therefore, researchers attempt to design weakened DP requirements, under which the performances can be significantly improved, while still securing sensitive information. Under such background, label DP has emerged in recent years [14], which regards features as public, while only labels are sensitive and need to be protected. Such setting is realistic in many applications, such as computational advertising [15], recommendation systems [16] and medical diagnosis [17]. These tasks usually use some basic demographic information as features, which can be far less sensitive.

Despite various approaches for learning with label DP [14, 18–21], the fundamental limits are still unknown. An interesting question is: By weakening the DP definitions to only labels, how much accuracy improvement is possible? From an information-theoretic perspective [22], the underlying limits of statistical problems are characterized by the minimax lower bound, which takes the supremum over all possible distributions from a general class, and infimum over all learners. Deriving minimax lower bounds for learning under the label DP is challenging in two aspects. Firstly, under label DP, each sample has both public (i.e. the feature) and private (i.e. the label) components. Directly applying the methods for original DP [23–27] treats all components as private, and thus does not yield tight results. Secondly, the classical packing method [47] is only suitable for fixed model structures with fixed dimensionality. However, to establish lower bounds, one needs to take infimum over all possible learners with arbitrary model complexity.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

	Classification	Regression Bounded label noise	Regression Unbounded label noise
Local	$\tilde{O}((N(\epsilon^2 \wedge 1))^{-\frac{\beta(\gamma+1)}{2\beta+d}})$	$\tilde{O}\left((N(\epsilon^2 \wedge 1)^{-\frac{2\beta}{d+2\beta}}\right)$	$O\left((N\epsilon^2)^{-\frac{2\beta(p-1)}{2p\beta+d(p-1)}} \vee N^{-\frac{2\beta}{2\beta+d}}\right)$
Central	$\tilde{O}\left(N^{-\frac{\beta(\gamma+1)}{2\beta+d}} + (\epsilon N)^{-\frac{\beta(\gamma+1)}{\beta+d}}\right)$	$O\left(N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta}{d+\beta}}\right)$	$O\left(N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta(p-1)}{p\beta+d(p-1)}}\right)$
Local full	$O((N(\epsilon^2 \wedge 1))^{-\frac{\beta(\gamma+1)}{2\beta+2d}})$	$O((N(\epsilon^2 \wedge 1))^{-\frac{\beta}{\beta+d}})$	$O((N(\epsilon^2 \wedge 1))^{-\frac{\beta(p-1)}{p\beta+d(p-1)}})$
Non-priv.	$O(N^{-\frac{\beta(\gamma+1)}{2\beta+d}})$	$O(N^{-\frac{2\beta}{2\beta+d}})$	$O(N^{-\frac{2\beta}{2\beta+d}})$

Table 1: Minimax rate of convergence under label differential privacy. d is the dimension of features.

In this paper, we investigate the theoretical limits of classification and regression problems under label DP. Our analysis involves both central and local models. For each problem, we derive the information-theoretic minimax lower bound of the risk function over a wide class of distributions satisfying the β-Hölder smoothness and the γ-Tsybakov margin assumption [28] (see Assumption 1 for details). The general idea is to convert the problem to multiple hypothesis testing. To overcome the challenges above, we provide a bound of Kullback-Leibler divergence over joint distributions of private and public random variables, which is tighter than the bound between fully private variables. Moreover, under the central model, instead of using the packing method, we develop a new lower bound on the minimum testing error for each pair of hypotheses based on the group privacy property [4], which is suitable for arbitrary model complexity. After deriving minimax lower bounds, we also propose algorithms with matching upper bounds to validate the tightness of our results.

The results are shown in Table 1, in which the third row refers to the bounds under the original local DP definition, while the fourth row lists the non-private baselines. To the best of our knowledge, minimax rates under central DP have not been established, and are thus not listed here. The main findings are summarized as follows.

- Under ϵ-local label DP, for classification and regression with bounded label noise, the sample complexity is larger by a factor of $O(1/\epsilon^2)$. However, the convergence rate remains unaffected, which is in clear contrast with the original DP, under which the convergence rate is slower.
- Under ϵ-local label DP constraint, for regression with heavy-tailed label noise, the convergence rate of risk over N becomes slower, indicating that heavy-tailed labels increase the difficulty of privacy protection.
- Under ϵ-central label DP constraint, the performance loss caused by the privacy mechanism becomes further weakened. The risk only increases by a term that decays faster than the non-private rate, indicating that the additional sample complexity caused by the privacy mechanism becomes negligible with large N.

In general, our analysis provides a theoretical perspective of understanding label DP. The result shows that by weakening the DP definition to protecting labels only, the learning performances can be significantly improved.

2 Related Work

Label DP. Under the local model, labels are randomized before training. The simplest method is randomized response [30]. An important improvement is proposed in [14], called RRWithPrior, which incorporates prior distribution. [19] proposes ALIBI, which further improves randomized response by generating soft labels through Bayesian inference. There are also several methods for regression under label DP [18, 31]. Under central label DP, [20] proposes a clustering approach. [19] proposes private aggregation of teacher ensembles (PATE), which is then further improved in [21].

Minimax analysis for public data. Minimax theory provides a rigorous framework for the best possible performance of an algorithm given some assumptions. Classical methods include Le Cam [32], Fano [33] and Assouad [34]. Using these methods, minimax lower bounds have been widely established for both classification and regression problems [28, 29, 35–41]. If the feature vector has bounded support, then the minimax rate of classification and regression are $O(N^{-\frac{\beta(\gamma+1)}{2\beta+d}})$ and $O(N^{-\frac{2\beta}{2\beta+d}})$, respectively.

Minimax analysis for private data. Under the local model, [42] finds the relation between label DP and stochastic query. [23] and [24] develop the variants of Le Cam, Fano, and Assouad's method under local DP. Lower bounds are then established for various statistical problems, such as mean estimation [43–46], classification [26] and regression [27]. Under central model, for pure DP, the standard approach is the packing method [47], which is then used in hypothesis testing [48], mean estimation [49,50], and learning of distributions [51–53]. There are also several works on approximate DP, such as [54, 55].

This work studies the theoretical limits of label DP, under which each sample is a mixture of public feature and private labels, thus existing methods can not be directly applied here. Under the central model, the minimax analysis becomes more challenging, since the packing method is only suitable for fixed model structures (i.e. the dimensionality of model output is fixed), while we need to find the minimum possible error over all possible learners with arbitrary output dimensions. As a result, the lower bounds of general classification and regression problems have not been established even under the original DP definition. To overcome such challenge, we develop a new approach to bound the error of hypothesis testing (see Lemma 1 in Appendix D).

3 Preliminaries

In this section, we show some necessary definitions, background information, and notations.

3.1 Label DP

To begin with, we review the definition of DP. Suppose the dataset consists of N samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, in which $\mathbf{x}_i \in \mathcal{X}$ is the feature vector, while $y_i \in \mathcal{Y} \subset \mathbb{R}^d$ is the label.

Definition 1. *(Differential Privacy (DP) [2]) Let $\epsilon \geq 0$. A randomized function $\mathcal{A} : (\mathcal{X}, \mathcal{Y})^N \rightarrow \Theta$ is ϵ-DP if for any two adjacent datasets $D, D' \in (\mathcal{X}, \mathcal{Y})^N$ and any $S \subseteq \Theta$,*

$$P(\mathcal{A}(D) \in S) \leq e^{\epsilon} P(\mathcal{A}(D') \in S), \tag{1}$$

in which D and D' are adjacent if they differ only on a single sample, including both the feature vector and the label.

In machine learning tasks, the output of $\mathcal{A}$ is the model parameters, while the input is the training dataset. Definition 1 requires that both features and labels are privatized. Consider that in some applications, the features may be much less sensitive, the notion of label DP is defined as follows.

Definition 2. *(Central label DP) A randomized function $\mathcal{A}$ is ϵ-label DP if for any two datasets D and D' that differ on the label of only one training sample and any $S \subseteq \Theta$,* (1) *holds.*

Compared with Definition 1, Definition 2 only requires the output to be insensitive to the replacement of a label. Therefore label DP is a weaker requirement. Correspondingly, the local label DP is defined as follows.

Definition 3. *(Local label DP) A randomized function $M : (\mathcal{X}, \mathcal{Y}) \rightarrow \mathcal{Z}$ is ϵ-local label DP if*

$$\sup_{y, y' \in \mathcal{Y}} \sup_{S \subseteq \mathcal{Z}} \ln \frac{P(M(\mathbf{x}, y) \in S)}{P(M(\mathbf{x}, y') \in S)} \leq \epsilon. \tag{2}$$

Definition 3 requires that each label is privatized locally before running any machine learning algorithms. It is straightforward to show that local label DP ensures central label DP. To be more precise, we have the following proposition.

Proposition 1. *Let $\mathbf{z}_i = M(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, N$. If $\mathcal{A}$ is a function of $(\mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, N$, then $\mathcal{A}$ is ϵ-label DP.*

3.2 Risk of Classification and Regression

In supervised learning problems, given N samples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$ drawn from a common distribution, the task is to learn a function $g : \mathcal{X} \rightarrow \mathcal{Y}$. For a loss function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, the goal is to minimize the *risk function*, which is defined as the expectation of loss function between the predicted value and the ground truth:

$$R = \mathbb{E}[l(\hat{Y}, Y)]. \tag{3}$$

The minimum risk among all function g is called Bayes risk, i.e. $R^* = \min_g \mathbb{E}[l(g(\mathbf{X}, Y))]$. In practice, the sample distribution is unknown, and we need to learn g from samples. Therefore, the risk of any practical classifiers is larger than Bayes risk. The gap $R - R^*$ is called excess risk, and we hope that $R - R^*$ to be as small as possible. Now we discuss classification and regression problems separately.

1) Classification. For classification problems, the size of $\mathcal{Y}$ is finite. For convenience, we denote $\mathcal{Y} = [K]$, in which $[K] := \{1, \ldots, K\}$. In this paper, we use $0-1$ loss, i.e. $l(\hat{Y}, Y) = \mathbf{1}(\hat{Y} \neq Y)$, then $R = \text{P}(\hat{Y} \neq Y)$. Define K functions $\eta_1, \ldots, \eta_K$ as the conditional class probabilities:

$$\eta_k(\mathbf{x}) = \text{P}(Y = k | \mathbf{X} = \mathbf{x}), k = 1, \ldots, K. \tag{4}$$

Under this setting, the Bayes optimal classifier and the corresponding Bayes risk is

$$c^*(\mathbf{x}) = \underset{j \in [K]}{\arg\max} \eta_j(\mathbf{x}), \tag{5}$$

$$R^*_{cls} = \text{P}(c^*(\mathbf{X}) \neq Y). \tag{6}$$

2) Regression. Now we consider the case with $\mathcal{Y}$ having infinite size. We use ℓ_2 loss in this paper, i.e. $l(\hat{Y}, Y) = (\hat{Y} - Y)^2$. Then the Bayes risk is

$$R^*_{reg} = \mathbb{E}[(Y - \eta(\mathbf{X}))^2]. \tag{7}$$

Then the following proposition gives a bound of the excess risk for classification and regression problems.

Proposition 2. *For any classifier $c : \mathcal{X} \to [K]$, the excess risk of classification is bounded by*

$$R_{cls} - R^*_{cls} = \int (\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c(\mathbf{x})}(\mathbf{x})]) f(\mathbf{x}) d\mathbf{x}. \tag{8}$$

For any regression estimate $\hat{\eta} : \mathcal{X} \to \mathcal{Y}$, the excess risk of regression is bounded by

$$R_{reg} - R^*_{reg} = \mathbb{E}[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2]. \tag{9}$$

The proof of Proposition 2 is shown in Appendix A. Finally, we state some basic assumptions that will be used throughout this paper.

Assumption 1. *There exists some constants L, β, C_T, γ, c, D and $\theta \in (0, 1]$ such that*

(a) For all $j \in [K]$ and any $\mathbf{x}$, $\mathbf{x}'$, $|\eta_j(\mathbf{x}) - \eta_j(\mathbf{x}')| \leq L \left\|\mathbf{x} - \mathbf{x}'\right\|^\beta$;

(b) For any $t > 0$, $P(0 < \eta^*(\mathbf{X}) - \eta_s(\mathbf{X}) < t) \leq C_T t^\gamma$, in which $\eta_s(\mathbf{x})$ is the second largest one among $\{\eta_1(\mathbf{x}), \ldots, \eta_K(\mathbf{x})\}$;

(c) The feature vector $\mathbf{X}$ has a probability density function (pdf) f which is bounded from below, i.e. $f(\mathbf{x}) \geq c$;

(d) For all $r < D$, $V_r(\mathbf{x}) \geq \theta v_d r^d$, in which $V_r(\mathbf{x})$ is the volume (Lebesgue measure) of $B(\mathbf{x}, r) \cap \mathcal{X}$, v_d is the volume of a unit ball.

Assumption 1 (a) requires that all η_j are Hölder continuous. This condition is common in literatures about nonparametric statistics [28]. (b) is generalized from the Tsybakov noise assumption for binary classification, which is commonly used in many existing works in the field of both nonparametric classification [29, 37, 40, 41] and differential privacy [26, 27]. If $K = 2$, then η^* and η_s refer to the larger and smaller class conditional probability, respectively. An intuitive understanding of (b) is that in the majority of the support, the maximum value among $\{\eta_1(\mathbf{x}), \ldots, \eta_K(\mathbf{x})\}$ should have some gap to the second largest one. With sufficiently large sample size and model complexity, assumption (b) ensures that for test samples within the majority of the support $\mathcal{X}$, the algorithm is highly likely to correctly identify the class with the maximum conditional probability. Therefore, in (b), we only care about $\eta^*(\mathbf{x})$ and $\eta_s(\mathbf{x})$, while other classes with small conditional probabilities can be ignored. (c) is usually called "strong density assumption" in existing works [39, 40], which is quite strong. It is possible to relax this assumption so that the theoretical analysis becomes suitable for general cases. However, we do not focus on such generalization in this paper. Assumption (d) prevents the corner of the support $\mathcal{X}$ from being too sharp. In the remainder of this section, denote $\mathcal{F}_{cls}$ as the set of all pairs (f, η) satisfying Assumption 1.

4 Classification

In this section, we derive the upper and lower bounds of learning under central and local label DP, respectively.

4.1 Local Label DP

1) Lower bound. The following theorem shows the minimax lower bound, which characterizes the theoretical limit.

Theorem 1. *Denote $\mathcal{M}_\epsilon$ as the set of all privacy mechanisms satisfying ϵ-local label DP (Definition 3). Then*

$$\inf_{\hat{Y}} \inf_{M\in\mathcal{M}_\epsilon} \sup_{(f,\eta)\in\mathcal{F}_{cls}} (R_{cls} - R_{cls}^*) \gtrsim \left[N\left(\epsilon^2 \wedge 1\right)\right]^{-\frac{\beta(\gamma+1)}{2\beta+d}}. \tag{10}$$

Proof. (Outline) It suffices to derive (10) with $K = 2$. We convert the problem into multiple binary hypothesis testing problems. In particular, we divide the support into G bins. For some of them, we construct two opposite hypotheses such that they are statistically not distinguishable. Our proof uses some techniques in local DP [24] and some classical minimax theory [28]. The detailed proof is shown in Appendix B. □

In Theorem 1, (10) takes supremum over all joint distributions of $(\mathbf{X}, Y)$, and infimum over all classifiers and privacy mechanisms satisfying ϵ-local label DP.

2) Upper bound. We then show that the bound (10) is achievable. Let the privacy mechanism $M(\mathbf{x}, y)$ outputs a K dimensional vector, with each component being either 0 or 1, such that

$$\mathrm{P}(M(\mathbf{x}, y)(j) = 1) = \begin{cases} \frac{e^{\frac{\epsilon}{2}}}{e^{\frac{\epsilon}{2}}+1} & \text{if} \quad y = j \\ \frac{1}{e^{\frac{\epsilon}{2}}+1} & \text{if} \quad y \neq j, \end{cases} \tag{11}$$

and $\mathrm{P}(M(\mathbf{x}, y)(j) = 0) = 1 - \mathrm{P}(M(\mathbf{x}, y)(j) = 1)$, in which $M(\mathbf{x}, y)(j)$ is the j-th component of $M(\mathbf{x}, y)$. For N random training samples $(\mathbf{X}_i, Y_i)$, let $\mathbf{Z}_i = M(\mathbf{X}_i, Y_i)$, and correspondingly, $Z_i(j)$ is the j-th component of $\mathbf{Z}_i$.

Divide the support $\mathcal{X}$ into G bins, named $B_1, \ldots, B_G$, such that the length of each bin is h. $B_1, \ldots, B_G$ are disjoint, and these bins form a covering of $\mathcal{X}$, i.e. $\mathcal{X} \subset \cup_{l=1}^G B_l$. Then calculate

$$S_{lj} = \sum_{i:\mathbf{X}_i\in B_l} Z_i(j), l = 1, \ldots, G, j = 1, \ldots, K, \tag{12}$$

The classification within the l-th bin is

$$c_l = \arg\max_j S_{lj}, \tag{13}$$

such that the the prediction given $\mathbf{x}$ is $c(\mathbf{x}) = c_l$ for all $\mathbf{x} \in B_l$. The next theorem shows the privacy guarantee, as well as the bound of the excess risk.

Theorem 2. *The privacy mechanism M is ϵ-local label DP. Moreover, under Assumption 1, with $h \sim \left(N(\epsilon^2 \wedge 1)/\ln K\right)^{-\frac{1}{2\beta+d}}$, the excess risk of the classifier described above can be upper bounded as follows:*

$$R_{cls} - R_{cls}^* \lesssim \left(\frac{N(\epsilon^2 \wedge 1)}{\ln K}\right)^{-\frac{\beta(\gamma+1)}{2\beta+d}}. \tag{14}$$

Proof. (Outline) For privacy guarantee, we need to show that (11) is ϵ-local label DP:

$$\begin{aligned} \frac{\mathrm{P}(M(\mathbf{x}, y) = \mathbf{z})}{\mathrm{P}(M(\mathbf{x}, y') = \mathbf{z})} &= \Pi_{j=1}^K \frac{\mathrm{P}(M(\mathbf{x}, y)(j) = \mathbf{z}(j))}{\mathrm{P}(M(\mathbf{x}, y')(j) = \mathbf{z}(j))} \\ &= \frac{\mathrm{P}(M(\mathbf{x}, y)(y) = \mathbf{z}(y))}{\mathrm{P}(M(\mathbf{x}, y')(y) = \mathbf{z}(y))} \frac{\mathrm{P}(M(\mathbf{x}, y)(y') = \mathbf{z}(y'))}{\mathrm{P}(M(\mathbf{x}, y')(y') = \mathbf{z}(y'))} \\ &\leq e^{\frac{\epsilon}{2}} e^{\frac{\epsilon}{2}} = e^\epsilon. \end{aligned} \tag{15}$$

According to Definition 3, M is ϵ-local label DP. For the performance guarantee (14), according to Proposition 2, we need to bound $\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c(\mathbf{x})}(\mathbf{x})]$ for each $\mathbf{x}$. If $\eta^*(\mathbf{x}) - \eta_s(\mathbf{x})$ is large, then with high probability, $c(\mathbf{x}) = c^*(\mathbf{x})$, and then $\eta^*(\mathbf{x}) = \eta_{c(\mathbf{x})}(\mathbf{x})$. Thus we mainly consider the case with small $\eta^*(\mathbf{x}) - \eta_s(\mathbf{x})$. The details of proof are shown in Appendix C. □

The lower bound (10) and the upper bound (14) match up to a logarithm factor, indicating that the results are tight. Now we comment on the results.

Remark 1. ***1) Comparison with non-private bound.*** *The classical minimax lower bound for non-private classification problem is* $N^{-\frac{\beta(\gamma+1)}{2\beta+d}}$. *Therefore, the lower bound* (10) *reaches the non-private bound with* $\epsilon \gtrsim 1$. *With small* ϵ, N *training samples with privatized labels roughly equals* $N\epsilon^2$ *non-privatized samples in terms of performance.*

2) Comparison with local DP that protects both features and labels. *In this case, the optimal excess risk is* $(N\epsilon^2)^{-\beta(\gamma+1)/(2\beta+2d)} \vee N^{-\beta(\gamma+1)/(2\beta+d)}$, *which is worse than the right hand side of* (10). *Such result indicates that compared with classical DP, label DP incurs significantly weaker performance loss.*

3) Comparison with other baseline methods. *If we use the randomized response method instead of the privacy mechanism* (11), *then the performance decreases sharply with the number of classes* K. *Several methods have been proposed to improve the randomized response method, such as RRWithPrior [14] and ALIBI [19]. However, these methods are not guaranteed in theory.*

4.2 Central Label DP

1) Lower bound. The following theorem shows the minimax lower bound under the central label DP.

Theorem 3. *Denote* $\mathcal{A}_\epsilon$ *as the set of all learning algorithms satisfying* ϵ*-label DP (Definition 2). Then*

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon} \sup_{(f,\eta)\in\mathcal{F}_{cls}} (R_{cls} - R_{cls}^*) \gtrsim N^{-\frac{\beta(\gamma+1)}{2\beta+d}} + (\epsilon N)^{-\frac{\beta(\gamma+1)}{\beta+d}}. \tag{16}$$

Proof. (Outline) Lower bounds under central DP are usually constructed by packing method [47], which works for fixed output dimensions. However, to achieve a desirable bias and variance tradeoff, the model complexity needs to increase with N. In our proof, we still divide the support into G bins and construct two hypotheses for each bin, but we develop a new tool (see Lemma 1) to give a lower bound of the minimum error of hypothesis testing. We then use the group privacy property [4] to get the overall lower bound. The details can be found in Appendix D. □

2) Upper bound. Now we show that (16) is achievable. Similar to the local label DP problem, now divide the support into G bins, such that the length of each bin is h. Now the classification within the l-th bin follows a exponential mechanism [56]:

$$\mathrm{P}(c_l = j | \mathbf{X}_{1:N}, Y_{1:N}) = \frac{e^{\epsilon n_{lj}/2}}{\sum_{k=1}^K e^{\epsilon n_{lk}/2}}, \tag{17}$$

in which $n_{lj} = \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l, Y_i = j)$. Then let $c(\mathbf{x}) = c_l$ for $\mathbf{x} \in B_l$. The excess risk is bounded in the next theorem.

Theorem 4. *The privacy mechanism* (17) *is* ϵ*-label DP. Moreover, under Assumption 1, if* h *scales as* $h \sim (\ln K/\epsilon N)^{\frac{1}{\beta+d}} + (\ln K/N)^{\frac{1}{2\beta+d}}$, *then the excess risk can be bounded as follows:*

$$R - R^* \lesssim \left(\frac{N}{\ln K}\right)^{-\frac{\beta(\gamma+1)}{2\beta+d}} + \left(\frac{\epsilon N}{\ln K}\right)^{-\frac{\beta(\gamma+1)}{\beta+d}}. \tag{18}$$

Proof. (Outline) The privacy guarantee of the exponential mechanism has been analyzed in [4]. Following these existing analyses, it can be shown that (17) is ϵ-label DP. It remains to show (18). Note that if $\eta^*(\mathbf{x}) - \eta_s(\mathbf{x})$ is large, then the difference between the largest and the second largest one from $\{n_{lj} | j = 1, \ldots, K\}$ will also be large. From (17), the following inequality holds with high probability: $c_l = \arg\max_j n_{lj} = \arg\max_j \eta_j(\mathbf{x}) = c^*(\mathbf{x})$, which means that the classifier makes optimal prediction. Hence we mainly consider the case with small $\eta^*(\mathbf{x}) - \eta_s(\mathbf{x})$. The details of the proof can be found in Appendix E. $\square$

The upper and lower bounds match up to logarithmic factors. In (18), the first term is just the non-private convergence rate, while the second term $(\epsilon N)^{-\frac{\beta(\gamma+1)}{\beta+d}}$ can be regarded as the additional risk caused by the privacy mechanism. It decays faster with N compared with the first term, thus the additional performance loss caused by the privacy mechanism becomes negligible as N increases. This result is crucially different from the local model, under which the privacy mechanism always induces higher sample complexity by a factor of $O(1/(\epsilon^2 \wedge 1))$.

5 Regression with Bounded Noise

Now we analyze the theoretical limits of regression problems under local and central label DP. Throughout this section, we assume that the label is restricted within a bounded interval.

Assumption 2. *Given any* $\mathbf{x} \in \mathcal{X}$, $P(|Y| < T|\mathbf{X} = \mathbf{x}) = 1$.

Assumption 1 remains the same here. In the remainder of this section, denote $\mathcal{F}_{reg1}$ as the set of (f, η) that satisfies Assumption 1 and 2.

5.1 Local Label DP

1) Lower bound. Theorem 5 shows the minimax lower bound.

Theorem 5. *Denote $\mathcal{M}_\epsilon$ as the set of all privacy mechanisms satisfying ϵ-label DP. Then*

$$\inf_{\hat{\eta}} \inf_{M \in \mathcal{M}_\epsilon} \sup_{(f,\eta) \in \mathcal{F}_{reg1}} (R_{reg} - R_{reg}^*) \gtrsim (N(\epsilon^2 \wedge 1))^{-\frac{2\beta}{d+2\beta}}. \tag{19}$$

The proof of Theorem 5 is similar to that of Theorem 1, except for some details in hypotheses construction and the final bound of excess risk. The details are shown in Appendix F.

2) Upper bound. The privacy mechanism is $Z = Y + W$, in which $W \sim \text{Lap}(2T/\epsilon)$. Then the privacy mechanism satisfies ϵ-label DP. In this case, the real regression function $\eta(\mathbf{x})$ can be estimated using the nearest neighbor approach. Let

$$\hat{\eta}(\mathbf{x}) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(\mathbf{x})} Z_i, \tag{20}$$

in which $\mathcal{N}_k(\mathbf{x})$ is the set of k nearest neighbors of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

Theorem 6. *The method described above is ϵ-local label DP. Moreover, with $k \sim N^{\frac{2\beta}{d+2\beta}} (\epsilon \wedge 1)^{-\frac{2d}{d+2\beta}}$, then under Assumption 1 and 2,*

$$R_{reg} - R_{reg}^* \lesssim (N(\epsilon^2 \wedge 1))^{-\frac{2\beta}{d+2\beta}}. \tag{21}$$

Proof. (Outline) Since $|Y| < T$, $W \sim \text{Lap}(2T/\epsilon)$, it is obvious that $Z = Y + W$ is ϵ-local label DP. For the performance (21), the bias can be bounded by the k nearest neighbor distances based on Assumption 1(a). The variance of $\hat{\eta}(\mathbf{x})$ scales inversely with k. An appropriate k can be selected to achieve a good tradeoff between bias and variance. The details are shown in Appendix G. $\square$

From standard minimax analysis on regression problems, the non-private convergence rate is $N^{-2\beta/(d+2\beta)}$. From Theorem 5 and 6, the privatization process makes sample complexity larger by a $O(1/\epsilon^2)$ factor.

5.2 Central Label DP

1) Lower bound. The following theorem shows the minimax lower bound.

Theorem 7. *Let $\mathcal{A}_\epsilon$ be the set of all algorithms satisfying ϵ-central DP. Then*

$$\inf_{\mathcal{A} \in \mathcal{A}_\epsilon} \sup_{(f,\eta) \in \mathcal{F}_{reg1}} (R_{reg} - R_{reg}^*) \gtrsim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta}{d+\beta}}. \tag{22}$$

2) Upper bound. For each bin B_l, let $n_l = \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l)$ be the number of samples in B_l. If $n_l > 0$, then

$$\hat{\eta}_l = \frac{1}{n_l} \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l) Y_i + W_l, \tag{23}$$

in which $W_l \sim \text{Lap}(2/(n_l \epsilon))$. If $n_l = 0$, i.e. no sample falls in B_l, then just let $\hat{\eta}_l = 0$. For all $\mathbf{x} \in B_l$, let $\hat{\eta}(\mathbf{x}) = \hat{\eta}_l$. The excess risk can be bounded with the following theorem.

Theorem 8. (23) *is ϵ-label DP. Moreover, under Assumption 1 and 2, if h scales as $h \sim N^{-\frac{1}{2\beta+d}} + (\epsilon N)^{-\frac{1}{d+\beta}}$, then the excess risk is bounded by*

$$R - R^* \lesssim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta}{d+\beta}}. \tag{24}$$

The upper and lower bounds match, indicating that the results are tight. Again, the second term in (24) converges faster than the first one with respect to N, the performance loss caused by privacy constraints becomes negligible as N increases.

6 Regression with Heavy-tailed Noise

In this section, we consider the case such that the noise has tails. We make the following assumption.

Assumption 3. *For all $\mathbf{x} \in \mathcal{X}$, $\mathbb{E}[|Y|^p | \mathbf{X} = \mathbf{x}] \leq M_p$ for some $p \geq 2$.*

Instead of requiring $|Y| < T$ for some T, now we only assume that the p-th order moment is bounded. For non-private cases, given fixed noise variance, the tail does not affect the mean squared error of regression. As a result, as long as $p \geq 2$, the convergence rate of regression risk is the same as the case with bounded noise. However, the label DP requires the output to be insensitive to the worst case replacement of labels, which can be harder if the noise has tails. To achieve ϵ-DP, the clipping radius decreases with ϵ, thus the noise strength needs to grow faster than $O(1/\epsilon)$. As a result, the convergence rate becomes slower than the non-private case. In the remainder of this section, denote $\mathcal{F}_{reg2}$ as the set of (f, η) that satisfies Assumption 1 and 3.

6.1 Local Label DP

1) Lower bound. In earlier sections about classification and regression with bounded noise, the impact of privacy mechanisms is only a polynomial factor on ϵ, while the convergence rate of excess risk with respect to N is not changed. However, this rule no longer holds when the noise has heavy tails.

Theorem 9. *Denote $\mathcal{M}_\epsilon$ as the set of all privacy mechanisms satisfying ϵ-label DP. Then for small ϵ,*

$$\inf_{\hat{\eta}} \inf_{M \in \mathcal{M}_\epsilon} \sup_{(f,\eta) \in \mathcal{F}} (R_{reg} - R_{reg}^*) \gtrsim (N(e^\epsilon - 1)^2)^{-\frac{2\beta(p-1)}{2p\beta + d(p-1)}} + N^{-\frac{2\beta}{2\beta+d}}. \tag{25}$$

2) Upper bound. Since now the noise has unbounded distribution, without preprocessing, the sensitivity is unbounded, thus simply adding noise to Y can no longer protect the privacy. Therefore, a solution is to clip Y into $[-T, T]$, and add noise proportional to T/ϵ to achieve ϵ-local label DP. Such truncation will inevitably introduce some bias. To achieve a tradeoff between clipping bias and sensitivity, the value of T needs to be tuned carefully. Based on such intuition, the method is precisely stated as follows. Let $Z_i = Y_{Ti} + W_i$, in which Y_{Ti} is the truncation of Y_i, i.e. $Y_{Ti} = (Y_i \wedge T) \vee (-T)$, and $W \sim \text{Lap}(2T/\epsilon)$. The result is shown in the next theorem.

Theorem 10. *The method above is ϵ-local label DP. Moreover, with $k \sim (N\epsilon^2)^{\frac{2p\beta}{2p\beta + d(p-1)}} \vee N^{\frac{2\beta}{2\beta+d}}$, and $T \sim (k\epsilon^2)^{\frac{1}{2p}}$, the risk is bounded by*

$$R_{reg} - R_{reg}^* \lesssim (N\epsilon^2)^{-\frac{2\beta(p-1)}{2p\beta + d(p-1)}} + N^{-\frac{2\beta}{2\beta+d}}. \tag{26}$$

Proof. (Outline) It can be shown that the clipping bias scales as $T^{2(1-p)}$. To meet the ϵ-label DP, an additional error that scales as T/ϵ is needed. By averaging over k nearest neighbors, the variance caused by noise W scales with $T^2/(k\epsilon^2)$. From standard analysis on nearest neighbor methods [29], the non-private mean squared error scales as $1/k + (k/N)^{2\beta/d}$. Put all these terms together, Theorem 10 can be proved. Details can be found in Appendix K. ☐

With the limit of $p \to \infty$, the problem reduces to the case with bounded noise, and the growth rate of k and the convergence rate of risk are the same as those in Theorem 6. For finite p, $2\beta(p-1)/(2p\beta + d(p-1)) < 2\beta/(2\beta+d)$, thus the convergence rate becomes slower due to the privacy mechanism.

6.2 Central Label DP

1) Lower bound. The minimax lower bound is shown in Theorem 11.

Theorem 11. *The minimax lower bound is*

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon} \sup_{(f,\eta)\in\mathcal{F}_{reg2}} (R_{reg} - R_{reg}^*) \gtrsim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta(p-1)}{p\beta+d(p-1)}} \tag{27}$$

2) Upper bound. Now we derive the upper bound. To restrict the sensitivity, instead of estimating with (23) directly, now we calculate an average of clipped label values:

$$\hat{\eta}_l = \frac{1}{n_l}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l)\,\mathrm{Clip}(Y_i, T) + W_l, \tag{28}$$

in which $W_l \sim \mathrm{Lap}(2T/(n_l\epsilon))$. Then for all $\mathbf{x} \in B_l$, let $\hat{\eta}(\mathbf{x}) = \hat{\eta}_l$. The following theorem bounds the excess risk.

Theorem 12. (28) *is ϵ-label DP. Moreover, under Assumption 1 and 3, if h and T scales as $h \sim N^{-\frac{1}{2\beta+d}} + (\epsilon N)^{-\frac{1}{p\beta+d(p-1)}}$, and $T \sim (\epsilon N h^d)^{1/p}$, then the excess risk can be bounded by*

$$R_{reg} - R_{reg}^* \lesssim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta(p-1)}{p\beta+d(p-1)}}. \tag{29}$$

The proof of Theorem 11 and 12 follow that of Theorem 7 and 8. The details are shown in Appendix L and M respectively. With $p = 2$, the right hand side of (29) becomes $(\epsilon \wedge 1)^{-\frac{2\beta}{2\beta+d}}$, indicating that the privacy constraint blows up the sample complexity by a constant factor. With larger p, the second term in (29) becomes negligible compared with the first one.

The theoretical analyses in this section are summarized as follows. In general, with fixed noise variance, if the label noise is heavy-tailed, while the non-private convergence rates remain unaffected, the additional risk caused by privacy mechanisms becomes significantly higher, indicating the difficulty of privacy protection for heavy-tailed distributions.

7 Conclusion

In this paper, we have derived the minimax lower bounds of learning under label DP for both central and local models. Furthermore, we propose methods whose upper bounds match these lower bounds. The results indicate the theoretical limits of learning under the label DP. From these results, it is discovered that under local label DP constraints, the sample complexity blows up by a factor of at least $O(1/\epsilon^2)$. Under central label DP requirements, the additional error caused by privacy mechanisms is significantly smaller. Finally, it is shown that for regression problem with heavy-tailed label distribution, the additional risk induced by privacy requirement becomes inevitably higher.

Limitations: The limitations of our work include the following aspects. Some assumptions can be weakened. For example, current analysis assumes that feature distributions have bounded supports, which may be extended to the unbounded case. One can let the bin splitting and nearest neighbor method be adaptive in the tails of features, such as [41]. Moreover, the bounds derived in this paper require that samples increase exponentially with dimensionality. However, in practice, the performance of learning under the label DP can be quite well even in high dimensions. The discrepancy can be explained by the fact that the minimax lower bound considers the worst-case distribution over a wide range of distributions. However, in most realistic cases, the distributions satisfy significantly better properties. A better modeling is to assume that these samples lie on a low dimensional manifold [57, 58]. In this case, it is possible to achieve a much better convergence rate. Finally, it is not sure whether approximate DP (i.e. (ϵ, δ)-DP) can improve the convergence rates.

References

[1] Rao, B., J. Zhang, D. Wu, et al. Privacy inference attack and defense in centralized and federated learning: A comprehensive survey. *IEEE Transactions on Artificial Intelligence*, 2024.

[2] Dwork, C., F. McSherry, K. Nissim, et al. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[3] Abadi, M., A. Chu, I. Goodfellow, et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. 2016.

[4] Dwork, C., A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[5] Bassily, R., A. Smith, A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[6] Bassily, R., V. Feldman, K. Talwar, et al. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Wang, D., H. Xiao, S. Devadas, et al. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

[8] Asi, H., V. Feldman, T. Koren, et al. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021.

[9] Das, R., S. Kale, Z. Xu, et al. Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*, pages 7066–7101. PMLR, 2023.

[10] Tramer, F., D. Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*. 2021.

[11] Bu, Z., J. Mao, S. Xu. Scalable and efficient training of large convolutional neural networks with differential privacy. *Advances in Neural Information Processing Systems*, 35:38305–38318, 2022.

[12] De, S., L. Berrada, J. Hayes, et al. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[13] Wei, J., E. Bao, X. Xiao, et al. Dpis: An enhanced mechanism for differentially private sgd with importance sampling. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2885–2899. 2022.

[14] Ghazi, B., N. Golowich, R. Kumar, et al. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34:27131–27145, 2021.

[15] McMahan, H. B., G. Holt, D. Sculley, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230. 2013.

[16] McSherry, F., I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636. 2009.

[17] Bussone, A., B. Kasadha, S. Stumpf, et al. Trust, identity, privacy, and security considerations for designing a peer data sharing platform between people living with hiv. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.

[18] Ghazi, B., P. Kamath, R. Kumar, et al. Regression with label differential privacy. In *The Eleventh International Conference on Learning Representations*. 2022.

[19] Malek Esmaeili, M., I. Mironov, K. Prasad, et al. Antipodes of label differential privacy: Pate and alibi. *Advances in Neural Information Processing Systems*, 34:6934–6945, 2021.

[20] Esfandiari, H., V. Mirrokni, U. Syed, et al. Label differential privacy via clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 7055–7075. PMLR, 2022.

[21] Tang, X., M. Nasr, S. Mahloujifar, et al. Machine learning with differentially private labels: Mechanisms and frameworks. *Proceedings on Privacy Enhancing Technologies*, 2022.

[22] Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

[23] Duchi, J. C., M. I. Jordan, M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[24] —. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[25] Gopi, S., G. Kamath, J. Kulkarni, et al. Locally private hypothesis selection. In *Conference on Learning Theory*, pages 1785–1816. PMLR, 2020.

[26] Berrett, T., C. Butucea. Classification under local differential privacy. *arXiv preprint arXiv:1912.04629*, 2019.

[27] Berrett, T. B., L. Györfi, H. Walk. Strongly universally consistent nonparametric regression and classification with privatised data. *Electronic Journal of Statistics*, 15:2430–2453, 2021.

[28] Tsybakov, A. B. *Introduction to Nonparametric Estimation*. 2009.

[29] Audibert, J.-Y., A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 2007.

[30] Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[31] Badanidiyuru Varadaraja, A., B. Ghazi, P. Kamath, et al. Optimal unbiased randomizers for regression with label differential privacy. *Advances in Neural Information Processing Systems*, 36, 2023.

[32] LeCam, L. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[33] Verdú, S., et al. Generalizing the fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, 1994.

[34] Assouad, P. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

[35] Yang, Y. Minimax nonparametric classification. i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

[36] —. Minimax nonparametric classification. ii. model selection for adaptation. *IEEE Transactions on Information Theory*, 45(7):2285–2292, 1999.

[37] Chaudhuri, K., S. Dasgupta. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014.

[38] Yang, Y., S. T. Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, pages 652–674, 2015.

[39] Döring, M., L. Györfi, H. Walk. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18(227):1–16, 2018.

[40] Gadat, S., T. Klein, C. Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Annals of Statistics*, 2016.

[41] Zhao, P., L. Lai. Minimax rate optimal adaptive nearest neighbor classification and regression. *IEEE Transactions on Information Theory*, 67(5):3155–3182, 2021.

[42] Kasiviswanathan, S. P., H. K. Lee, K. Nissim, et al. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[43] Li, M., T. B. Berrett, Y. Yu. On robustness and local differential privacy. *The Annals of Statistics*, 51(2):717–737, 2023.

[44] Feldman, V., T. Koren, K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449. 2020.

[45] Duchi, J., R. Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pages 1161–1191. PMLR, 2019.

[46] Huang, Z., Y. Liang, K. Yi. Instance-optimal mean estimation under differential privacy. *Advances in Neural Information Processing Systems*, 34:25993–26004, 2021.

[47] Hardt, M., K. Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714. 2010.

[48] Bun, M., G. Kamath, T. Steinke, et al. Private hypothesis selection. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Narayanan, S. Better and simpler lower bounds for differentially private statistical estimation. *arXiv preprint arXiv:2310.06289*, 2023.

[50] Kamath, G., V. Singhal, J. Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pages 2204–2235. PMLR, 2020.

[51] Kamath, G., J. Li, V. Singhal, et al. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

[52] Alabi, D., P. K. Kothari, P. Tankala, et al. Privately estimating a gaussian: Efficient, robust, and optimal. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 483–496. 2023.

[53] Arbas, J., H. Ashtiani, C. Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *International Conference on Machine Learning*, pages 1018–1040. 2023.

[54] Bun, M., J. Ullman, S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10. 2014.

[55] Kamath, G., A. Mouzakis, V. Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in neural information processing systems*, 35:24405–24418, 2022.

[56] McSherry, F., K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[57] Kpotufe, S. k-nn regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.

[58] Carter, K. M., R. Raich, A. O. Hero III. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2):650–663, 2009.

A Proof of Proposition 2

From (5) and (6), the Bayes risk is

$$R^*_{cls} = \mathrm{P}(Y \neq c^*(\mathbf{X})) = \int \mathrm{P}(Y \neq c^*(\mathbf{x})|\mathbf{X}=\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int (1-\eta^*(\mathbf{x})) f(\mathbf{x}) d\mathbf{x}. \tag{30}$$

The risk of classifier c is

$$R_{cls} = \mathrm{P}(Y \neq c(\mathbf{X})) = \mathbb{E}\left[\int \left(1-\eta_{c(\mathbf{x})}(\mathbf{x})\right) f(\mathbf{x}) d\mathbf{x}\right]. \tag{31}$$

From (31) and (6),

$$R_{cls} - R^*_{cls} = \int (\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c(\mathbf{x})}(\mathbf{x})]) f(\mathbf{x}) d\mathbf{x}. \tag{32}$$

The proof is complete.

B Proof of Theorem 1

In this section, we prove the minimax lower bound of multi-class classification. The problem with K classes with $K > 2$ is inherently harder than that with $K = 2$. Therefore, we just need to prove the lower bound for binary classification, in which $\mathcal{Y} = \{1, 2\}$. Let

$$\eta(\mathbf{x}) = \eta_2(\mathbf{x}) - \eta_1(\mathbf{x}). \tag{33}$$

Since $\eta_1(\mathbf{x}) + \eta_2(\mathbf{x}) = 1$ always holds, we have

$$\eta_1(\mathbf{x}) = \frac{1-\eta(\mathbf{x})}{2}, \eta_2(\mathbf{x}) = \frac{1+\eta(\mathbf{x})}{2}. \tag{34}$$

Therefore, $\eta(\mathbf{x})$ captures the conditional distribution of Y given $\mathbf{x}$.

Find G disjoint cubes $B_1, \ldots, B_G \subset \mathcal{X}$, such that the length of each cube is h. Denote $\mathbf{c}_1, \ldots, \mathbf{c}_G$ as the centers of these cubes. Let $\phi(\mathbf{u})$ be some function supported at $[-1/2, 1/2]^d$, such that

$$0 \leq \phi(\mathbf{u}) \leq 1. \tag{35}$$

Let $f(\mathbf{x}) = c$ over $\mathbf{x} \in \mathcal{X}$. For $\mathbf{v} \in \mathcal{V} := \{-1, 1\}^m$, let

$$\eta_{\mathbf{v}}(\mathbf{x}) = \sum_{k=1}^m v_k \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) h^\beta. \tag{36}$$

It can be proved that if for some constant C_M,

$$m \leq C_M h^{\gamma\beta - d}, \tag{37}$$

then for any $\eta = \eta_{\mathbf{v}}$, η_1 and η_2 satisfies Assumption 1(b). Denote

$$\hat{v}_k = \underset{s \in \{-1,1\}}{\arg\max} \int_{B_k} \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) \mathbf{1}(\mathrm{sign}(\hat{\eta}(\mathbf{x})) = s) f(\mathbf{x}) d\mathbf{x}. \tag{38}$$

Then the excess risk is bounded by

$$\begin{aligned} R - R^* &= \int |\eta_{\mathbf{v}}(\mathbf{x})| \mathrm{P}(\mathrm{sign}(\hat{\eta}(\mathbf{x})) \neq \mathrm{sign}(\eta_{\mathbf{v}}(\mathbf{x}))) f(\mathbf{x}) d\mathbf{x} \\ &\geq \sum_{k=1}^m \int_{B_k} |\eta_{\mathbf{v}}(\mathbf{x})| \mathrm{P}(\mathrm{sign}(\hat{\eta}(\mathbf{x})) \neq \mathrm{sign}(\eta_{\mathbf{v}}(\mathbf{x}))) f(\mathbf{x}) d\mathbf{x} \\ &= \sum_{k=1}^m h^\beta \int_{B_k} \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) \mathrm{P}(\mathrm{sign}(\hat{\eta}(\mathbf{x}))) f(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{39}$$

If $\hat{v}_k \neq v_k$, then from (38),

$$\int_{B_k} \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) \mathbf{1}(\mathrm{sign}(\hat{\eta}(\mathbf{x}))) f(\mathbf{x}) d\mathbf{x} \geq \int_{B_k} \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) \mathbf{1}(\mathrm{sign}(\hat{\eta}(\mathbf{x})) = v_k) f(\mathbf{x}) d\mathbf{x}. \tag{40}$$

Therefore

$$\int_{B_k} \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) \mathbf{1}(\text{sign}(\hat{\eta}(\mathbf{x})) \neq v_k) f(\mathbf{x}) d\mathbf{x} \geq \frac{1}{2}\int_{B_k} \phi\left(\frac{\mathbf{x}-\mathbf{c}_k}{h}\right) f(\mathbf{x}) d\mathbf{x} \geq \frac{1}{2} c h^d \left\|\phi\right\|_1. \quad (41)$$

Hence

$$\begin{aligned} R - R^* &\geq \frac{1}{2} c h^{\beta+d} \left\|\phi\right\|_1 \sum_{k=1}^m \mathrm{P}(\hat{v}_k \neq v_k) \\ &= \frac{1}{2} c h^{\beta+d} \left\|\phi\right\|_1 \mathbb{E}[\rho_H(\hat{\mathbf{v}}, \mathbf{v})], \end{aligned} \quad (42)$$

in which ρ_H denotes the Hamming distance. Then

$$\inf_{\hat{Y}} \inf_{M \in \mathcal{M}_\epsilon} \sup_{(f,\eta)\in\mathcal{P}} (R - R^*) \geq \frac{1}{2} h^{\beta+d} \left\|\phi\right\|_1 \inf_{\hat{\mathbf{v}}} \inf_{M\in\mathcal{M}_\epsilon} \max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}[\rho_H(\hat{\mathbf{v}}, \mathbf{v})]. \quad (43)$$

Define

$$\delta = \sup_{M\in\mathcal{M}_\epsilon} \max_{\mathbf{v},\mathbf{v}':\rho_H(\mathbf{v},\mathbf{v}')=1} D_{KL}(P_{(X,Z)_{1:N}|\mathbf{v}} || P_{(X,Z)_{1:N}|\mathbf{v}'}), \quad (44)$$

in which $P_{(X,Z)_{1:N}|\mathbf{v}}$ denotes the distribution of $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_N, Z_N)$ with $\eta = \eta_\mathbf{v}$. D_{KL} denotes the Kullback-Leibler divergence. Then from [28], Theorem 2.12(iv),

$$\inf_{\hat{\mathbf{v}}} \inf_{M} \max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}[\rho_H(\hat{\mathbf{v}}, \mathbf{v})] \geq \frac{m}{2}\left(\frac{1}{2}e^{-\delta}, 1 - \sqrt{\frac{\delta}{2}}\right). \quad (45)$$

It remains to bound δ. Without loss of generality, suppose $v_1 \neq v_1'$, and $v_i = v_i'$ for $i \neq 1$. Then

$$\begin{aligned} D_{KL}(P_{(X,Z)_{1:N}|\mathbf{v}} || P_{(X,Z)_{1:N}|\mathbf{v}'}) &\overset{(a)}{=} N D_{KL}(P_{X,Z|\mathbf{v}} || P_{X,Z|\mathbf{v}'}) \\ &\overset{(b)}{=} N \int_{B_1} f(\mathbf{x}) D_{KL}(P_{Z|\mathbf{X}=\mathbf{x},\mathbf{v}} || P_{Z|\mathbf{X}=\mathbf{x},\mathbf{v}'}) d\mathbf{x} \\ &\overset{(c)}{\leq} N \int_{B_1} f(\mathbf{x}) (e^\epsilon - 1)^2 \mathbb{TV}^2(P_{Z|\mathbf{X}=\mathbf{x},\mathbf{v}}, P_{Z|\mathbf{X}=\mathbf{x},\mathbf{v}'}) d\mathbf{x} \\ &= N \int_{B_1} f(\mathbf{x}) (e^\epsilon - 1)^2 \eta_\mathbf{v}^2(\mathbf{x}) d\mathbf{x} \\ &= N(e^\epsilon - 1)^2 \int_{B_1} f(\mathbf{x}) \phi^2\left(\frac{\mathbf{x}-\mathbf{c}_1}{h}\right) h^{2\beta} d\mathbf{x} \\ &\overset{(d)}{=} N(e^\epsilon - 1)^2 h^{2\beta+d} \left\|\phi\right\|_2^2. \end{aligned} \quad (46)$$

In (a), $P_{X,Z|\mathbf{v}}$ denotes the distribution of a single sample with privatized label (X, Z), with $\eta = \eta_\mathbf{v}$. In (b), $P_{Z|\mathbf{X}=\mathbf{x},\mathbf{v}}$ denotes the conditional distribution of Z given $\mathbf{X} = \mathbf{x}$, with $\eta = \eta_\mathbf{v}$. (c) uses [24], Theorem 1. In (d), $\left\|\phi\right\|_2^2 = \int \phi^2(\mathbf{u}) d\mathbf{u}$, which is a constant. Moreover,

$$\begin{aligned} D_{KL}(P_{X,Z|\mathbf{v}} || P_{X,Z|\mathbf{v}'}) &\overset{(a)}{\leq} D_{KL}(P_{X,Y|\mathbf{v}} || P_{X,Y|\mathbf{v}'}) \\ &= \int_{B_1} f(\mathbf{x}) \left[\mathrm{P}(Y=1|\mathbf{v}) \ln \frac{\mathrm{P}(Y=1|\mathbf{v})}{\mathrm{P}(Y=1|\mathbf{v}')} + \mathrm{P}(Y=-1|\mathbf{v}) \ln \frac{\mathrm{P}(Y=-1|\mathbf{v})}{\mathrm{P}(Y=-1|\mathbf{v}')}\right] d\mathbf{x} \\ &= \int_{B_1} f(\mathbf{x}) \left[\frac{1+\eta_\mathbf{v}(\mathbf{x})}{2} \ln \frac{1+\eta_\mathbf{v}(\mathbf{x})}{1-\eta_\mathbf{v}(\mathbf{x})} + \frac{1-\eta_\mathbf{v}(\mathbf{x})}{2} \ln \frac{1-\eta_\mathbf{v}(\mathbf{x})}{1+\eta_\mathbf{v}(\mathbf{x})}\right] d\mathbf{x} \\ &\overset{(b)}{\leq} 3 \int_{B_1} f(\mathbf{x}) \eta_\mathbf{v}^2(\mathbf{x}) d\mathbf{x} \\ &\leq 3 h^{2\beta+d} \left\|\phi\right\|_2^2. \end{aligned} \quad (47)$$

For (a), note that Z is generated from Y. From data processing inequality, (a) holds. For (b), without loss of generality, suppose that $v_1 = 1$, thus $\eta_\mathbf{v}(\mathbf{x}) \geq 0$ in B_1. Then $\ln(1 + \eta_\mathbf{v}(\mathbf{x})) \leq \eta_\mathbf{v}(\mathbf{x})$. From (35) and (36), $|\eta_\mathbf{v}(\mathbf{x})| \leq 1/2$. Therefore, $-\ln(1 - \eta_\mathbf{v}(\mathbf{x})) \leq 2\eta_\mathbf{v}(\mathbf{x})$. Therefore (b) holds.

From (46) and (47),

$$\delta \leq N \left[(e^{\epsilon} - 1)^2 \wedge 3 \right] h^{2\beta+d} \left\| \phi \right\|_2^2. \tag{48}$$

Let

$$h \sim \left(N \left(\epsilon^2 \wedge 1 \right) \right)^{-\frac{1}{2\beta+d}}. \tag{49}$$

Then $\delta \lesssim 1$. From (45), with $m \sim h^{\gamma\beta-d}$,

$$\inf_{\hat{\mathbf{v}}} \inf_{M \in \mathcal{M}_\epsilon} \max_{\mathbf{v} \in \mathcal{V}} \mathbb{E}[\rho_H(\hat{\mathbf{v}}, \mathbf{v})] \gtrsim h^{\gamma\beta-d}. \tag{50}$$

Hence

$$\inf_{\hat{Y}} \inf_{M \in \mathcal{M}_\epsilon} \sup_{(f,\eta) \in \mathcal{P}} (R - R^*) \gtrsim h^{\beta+d} h^{\gamma\beta-d} \sim h^{\beta(\gamma+1)} \sim \left[N \left(\epsilon^2 \wedge 1 \right) \right]^{-\frac{\beta(\gamma+1)}{2\beta+d}}. \tag{51}$$

The proof is complete.

C Proof of Theorem 2

Denote

$$n_l = \sum_{i=1}^{N} \mathbf{1}(\mathbf{X}_i \in B_l), \tag{52}$$

and for $\mathbf{Z} = M(\mathbf{X}, Y)$, let

$$\begin{aligned} \tilde{\eta}_j(\mathbf{x}) &:= \mathbb{E}[\mathbf{Z}(j) | \mathbf{X} = \mathbf{x}] \\ &= \frac{e^{\frac{\epsilon}{2}}}{e^{\frac{\epsilon}{2}} + 1} \eta_j(\mathbf{x}) + \frac{1}{e^{\frac{\epsilon}{2}} + 1} (1 - \eta_j(\mathbf{x})) \end{aligned} \tag{53}$$

as the number of training samples whose feature vectors fall in B_l, and

$$v_{lj} := \frac{1}{n_l} \sum_{i : \mathbf{X}_i \in B_l} \tilde{\eta}_j(\mathbf{X}_i). \tag{54}$$

Recall (12) that defines S_{lj}. From Hoeffding's inequality,

$$\mathrm{P}\left(|S_{lj} - n_l v_{lj}| > t | \mathbf{X}_{1:N} \right) \leq 2 \exp\left[-\frac{2t^2}{n_l} \right], \tag{55}$$

in which $\mathbf{X}_{1:N}$ denotes $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

Define

$$v_l^* := \max_j v_{lj}, \tag{56}$$

and

$$c_l^* := \arg\max_j v_{lj}. \tag{57}$$

Now we bound $\mathrm{P}(v_l^* - v_{lc_l} > t)$, in which c_l is defined in (13). c_l can be viewed as the prediction at the l-th bin. We would like to show that the even if the prediction is wrong, the value (i.e. conditional probability) of the predicted class is close to the ground truth. $v_l^* - v_{lc_l} > t$ only if $\exists j$, $v_l^* - v_{lj} > t$, and $S_{lj} > S_{lc_l^*}$. Therefore either $S_{lj} - n_l v_{lj} > t/2$ or $S_{lc_l^*} - n_l v_l^* > t/2$ holds. Hence

$$\mathrm{P}\left(v_l^* - v_{lc_l} \geq t \right) \leq \mathrm{P}\left(\exists j, |S_{lj} - n_l v_{lj}| \geq \frac{1}{2} n_l t \right) \leq 2K \exp\left(-\frac{1}{2} n_l t^2 \right). \tag{58}$$

Define

$$t_0 = \sqrt{\frac{2 \ln(2K)}{n_l}}. \tag{59}$$

Then

$$\begin{aligned}
v_l^* - \mathbb{E}[v_{lc_l}|\mathbf{X}_{1:N}] &= \int_0^1 \mathrm{P}(v_l^* - v_{lc_l} > t)dt \\
&\leq t_0 + \int_{t_0}^{\infty} 2K \exp\left(-\frac{1}{2}n_l t^2\right) dt \\
&\overset{(a)}{\leq} t_0 + 2\sqrt{\frac{2\pi}{n_l}} K \exp\left(-\frac{1}{2}n_l t_0^2\right) \\
&= \sqrt{\frac{2\ln(2K)}{n_l}} + \sqrt{\frac{2\pi}{n_l}} \\
&\leq 3\sqrt{\frac{\ln(2K)}{n_l}}. \qquad (60)
\end{aligned}$$

In (a), we use the inequality

$$\int_t^{\infty} e^{-\frac{u^2}{2\sigma^2}} du \leq \sqrt{2\pi}\sigma e^{-\frac{t^2}{2\sigma^2}}. \qquad (61)$$

Now we bound the excess risk.

$$\begin{aligned}
R - R^* &= \int \left(\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c(\mathbf{x})}(\mathbf{x})]\right) f(\mathbf{x}) d\mathbf{x} \\
&= \sum_{l=1}^{G} \int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c(\mathbf{x})}(\mathbf{x})]\right) f(\mathbf{x}) d\mathbf{x}. \qquad (62)
\end{aligned}$$

We need to bound $\int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c(\mathbf{x})}(\mathbf{x})]\right) f(\mathbf{x}) d\mathbf{x}$ for each l. From Assumption 1(a), for any $\mathbf{x}, \mathbf{x}' \in B_l$, the distance is bounded by $\|\mathbf{x} - \mathbf{x}'\| \leq \sqrt{d}L$. Thus

$$|\eta_j(\mathbf{x}) - \eta_j(\mathbf{x}')| \leq L_d h^\beta, \qquad (63)$$

in which L_d is defined as $L_d := L\sqrt{d}$. From (63) and (53),

$$|\tilde{\eta}_j(\mathbf{x}) - \tilde{\eta}_j(\mathbf{x}')| \leq \frac{e^{\frac{\epsilon}{2}} - 1}{e^{\frac{\epsilon}{2}} + 1} L_d h^\beta. \qquad (64)$$

Define

$$\tilde{\eta}^*(\mathbf{x}) = \max_j \tilde{\eta}_j(\mathbf{x}), \qquad (65)$$

then

$$\begin{aligned}
\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c_l}(\mathbf{x})|\mathbf{X}_{1:N}] &\leq \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \left(\tilde{\eta}^*(\mathbf{x}) - \mathbb{E}[\tilde{\eta}_{c_l}(\mathbf{x})|\mathbf{X}_{1:N}]\right) \\
&\leq \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \left(v_l^* - \mathbb{E}[v_{lc_l}|\mathbf{X}_{1:N}]\right) + 2L_d h^\beta \\
&\leq 3\frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \sqrt{\frac{2\ln(2K)}{n_l}} + 2L_d h^\beta. \qquad (66)
\end{aligned}$$

Take integration over cube B_l, we get

$$\begin{aligned}
&\int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c_l}(\mathbf{x})]\right) f(\mathbf{x}) d\mathbf{x} \\
\leq\ & \mathrm{P}\left(n_l < \frac{1}{2}Np(B_l)\right) \int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c_l}(\mathbf{x})|n_l < \frac{1}{N}p(B_l)]\right) f(\mathbf{x}) d\mathbf{x} \\
&+ \int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c_l}(\mathbf{x})|n_l \geq \frac{1}{N}p(B_l)]\right) f(\mathbf{x}) d\mathbf{x} \\
\leq\ & p(B_l) e^{-\frac{1}{2}(1-\ln 2)Np(B_l)} + \left[3\frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \sqrt{\frac{2\ln(2K)}{Np(B_l)}} + 2L^d h^\beta\right] p(B_l), \qquad (67)
\end{aligned}$$

in which $p(B_l) = \mathrm{P}(\mathbf{X} \in B_l)$ is the probability mass of B_l. Moreover, define

$$\Delta_l = \inf_{\mathbf{x} \in B_l} \left(\eta^*(\mathbf{x}) - \eta_s(\mathbf{x})\right), \tag{68}$$

and

$$\tilde{\Delta}_l = \inf_{\mathbf{x} \in B_l} \left(\tilde{\eta}^*(\mathbf{x}) - \tilde{\eta}_s(\mathbf{x})\right) = \frac{e^{\frac{\epsilon}{2}} - 1}{e^{\frac{\epsilon}{2}} + 1} \Delta_l, \tag{69}$$

in which the $\tilde{\eta}_s$ is the second largest value of $\tilde{\eta}_j$ among $j = 1, \ldots, K$, which follows the definition of η_s.

If $\Delta_l > 0$, then $c^*(\mathbf{x})$ is the same over B_l. Then either $v_l^* - v_{lc_l} = 0$ or $v_l^* - v_{lc_l} \geq \Delta_l$ holds. Hence

$$\begin{aligned}
& \tilde{\eta}^*(\mathbf{x}) - \mathbb{E}[\tilde{\eta}_{c_l}(\mathbf{x}) | \mathbf{X}_{1:N}] \\
= \; & \int_0^1 \mathrm{P}\left(\tilde{\eta}^*(\mathbf{x}) - \tilde{\eta}_{c_l}(\mathbf{x}) > t | \mathbf{X}_{1:N}\right) dt \\
\leq \; & \int_0^1 \mathrm{P}\left(v_l^* - v_{lc_l} > t - 2L_d h^\beta \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} | \mathbf{X}_{1:N}\right) dt \\
\leq \; & \int_0^{\tilde{\Delta}_l + 2L_d h^\beta} \mathrm{P}(v_l^* - v_{lc_l} \geq \Delta_l) dt + \int_{\tilde{\Delta}_l + 2L_d h^\beta}^\infty 2K \exp\left[-\frac{1}{2} n_l (t - 2L_d h^\beta)^2\right] dt \\
\leq \; & 2K \exp\left(-\frac{1}{2} n_l \tilde{\Delta}_l^2\right) (\tilde{\Delta}_l + 2L_d h^\beta \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1}) + 2K \sqrt{\frac{2\pi}{n_l}} \exp\left(-\frac{1}{2} n_l \tilde{\Delta}_l^2\right) \\
= \; & \left[2K \left(\tilde{\Delta}_l + 2L_d h^\beta \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1}\right) + 2K \sqrt{\frac{2\pi}{n_l}}\right] \exp\left(-\frac{1}{2} n_l \tilde{\Delta}_l^2\right).
\end{aligned} \tag{70}$$

Take expectation over $\mathbf{X}_{1:N}$, we get

$$\begin{aligned}
& \int_{B_l} (\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c_l}(\mathbf{x})]) f(\mathbf{x}) d\mathbf{x} \leq p(B_l) e^{-\frac{1}{2}(1 - \ln 2) N p(B_l)} \\
& + 2K p(B_l) \left(\Delta_l + 2L_d h^\beta + \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \sqrt{\frac{2\pi}{N p(B_l)}}\right) \exp\left[-\frac{1}{2} N p(B_l) \Delta_l^2 \left(\frac{e^{\frac{\epsilon}{2}} - 1}{e^{\frac{\epsilon}{2}} + 1}\right)^2\right].
\end{aligned} \tag{71}$$

Define

$$a_l = \left[3 \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \sqrt{\frac{2 \ln(2K)}{c N h^d}} + 2L_d h^\beta\right] p(B_l), \tag{72}$$

and

$$b_l = 2K p(B_l) \left(\Delta_l + 2L_d h^\beta + \frac{e^{\frac{\epsilon}{2}} + 1}{e^{\frac{\epsilon}{2}} - 1} \sqrt{\frac{2\pi}{c N h^d}}\right) \exp\left[-\frac{1}{2} c N h^d \Delta_l^2 \left(\frac{e^{\frac{\epsilon}{2}} - 1}{e^{\frac{\epsilon}{2}} + 1}\right)^2\right]. \tag{73}$$

From Assumption 1(c), $p(B_l) \geq c N h^d$. Therefore, from (67) and (71)

$$\begin{aligned}
R - R^* \quad & \leq \quad \sum_{l=1}^G \left[p(B_l) e^{-\frac{1}{2}(1 - \ln 2) N p(B_l)} + \min\{a_l, b_l\}\right] \\
& \leq \quad e^{-\frac{1}{2}(1 - \ln 2) c N h^d} + \sum_{l=1}^G \min\{a_l, b_l\}.
\end{aligned} \tag{74}$$

It remains to bound $\sum_{l=1}^G \min\{a_l, b_l\}$. Note that for all $\mathbf{x} \in B_l$, $\eta^*(\mathbf{x}) - \eta_s(\mathbf{x}) \leq \Delta_l + 2L_d h^\beta$. Thus

$$\sum_{l: \Delta_l \leq u} p(B_l) \leq \mathrm{P}\left(\eta^*(\mathbf{X}) - \eta_s(\mathbf{X}) \leq u + 2L_d h^\beta\right) \leq M(u + 2L_d h^\beta)^\gamma. \tag{75}$$

Let

$$\Delta_0 = \frac{e^{\frac{\epsilon}{2}}+1}{e^{\frac{\epsilon}{2}}-1}\sqrt{\frac{2\ln(2K)}{cNh^d}}, \tag{76}$$

and

$$\begin{aligned} I_0 &= \{l|\Delta_l \le \Delta_0\}, & (77)\\ I_k &= \{l|2^{k-1}\Delta_0 < \Delta_l \le 2^k\Delta_0\}, k=1,2,\dots & (78) \end{aligned}$$

Then

$$\begin{aligned} \min_{l\in I_0}\{a_l,b_l\} &\le \sum_{l\in I_0} a_l \\ &\le \left(\sum_{l:\Delta_l\le\Delta_0} p(B_l)\right)\left[3\frac{e^{\frac{\epsilon}{2}}+1}{e^{\frac{\epsilon}{2}}-1}\sqrt{\frac{2\ln(2K)}{cNh^d}}+2L_dh^\beta\right] \\ &\le M(\Delta_0+2L_dh^\beta)^\gamma\left[3\frac{e^{\frac{\epsilon}{2}}+1}{e^{\frac{\epsilon}{2}}-1}\sqrt{\frac{2\ln(2K)}{cNh^d}}+2L_dh^\beta\right] \\ &\lesssim \left(\frac{1}{\epsilon^2\wedge 1}\frac{\ln K}{Nh^d}\right)^{\frac{\gamma+1}{2}} + h^{\beta(\gamma+1)}. \end{aligned} \tag{79}$$

For I_k with $k\ge 1$,

$$\begin{aligned} \min_{l\in I_k}\{a_l,b_l\} &\le \sum_{l\in I_k} b_l \\ &\le \left(\sum_{l:\Delta_l\le 2^k\Delta_0} p(B_l)\right)\cdot 2K\left(2^k\Delta_0+2L_dh^\beta+\Delta_0\right)\exp\left[-\frac{1}{2}\left(\frac{e^{\frac{\epsilon}{2}}-1}{e^{\frac{\epsilon}{2}}+1}\right)^2 cNh^d 2^{2k-2}\Delta_0^2\right] \\ &\le M(2^k\Delta_0+2L_dh^\beta)^\gamma\left((2^k+1)\Delta_0+2L_dh^\beta\right)(2K)^{-2^{2k-2}+1} \\ &\le M(\Delta_0+2L_dh^\beta)^{\gamma+1}2^{k\gamma+k-2^{2k-2}+2}. \end{aligned} \tag{80}$$

It is obvious that there exists a finite constant $C'<\infty$ that depends on γ, such that

$$\sum_{k=1}^\infty 2^{k\gamma+k-2^{2k-2}+2} \le C'. \tag{81}$$

Therefore

$$\sum_{k=1}^\infty\sum_{l\in I_k}\min\{a_l,b_l\} \lesssim \left(\frac{1}{\epsilon^2\wedge 1}\frac{\ln K}{Nh^d}\right)^{\frac{\gamma+1}{2}} + h^{\beta(\gamma+1)}. \tag{82}$$

Combine (74), (79) and (82),

$$R-R^* \lesssim \left(\frac{1}{\epsilon^2\wedge 1}\frac{\ln K}{Nh^d}\right)^{\frac{\gamma+1}{2}} + h^{\beta(\gamma+1)}. \tag{83}$$

To minimize the overall excess risk, let

$$h \sim \left(\frac{N(\epsilon^2\wedge 1)}{\ln K}\right)^{-\frac{1}{2\beta+d}}, \tag{84}$$

then

$$R-R^* \lesssim \left(\frac{N(\epsilon^2\wedge 1)}{\ln K}\right)^{-\frac{\beta(\gamma+1)}{2\beta+d}}. \tag{85}$$

Compare to the simple random response method, the bin splitting avoids the polynomial decrease over K.

D Proof of Theorem 3

We still divide the support as the local label DP setting, except that the value of h is different, which will be specified later in this section. Note that (42) still holds here. Let $\mathbf{V}$ takes values from $\{-1,1\}^m$ randomly with equal probability, and V_k is the k-th element. Then $\eta_{\mathbf{V}}(\mathbf{x})$ is a random function. The corresponding random output of hypothesis testing is denoted as $\hat{V}_k$, which is calculated by (38). Then

$$\begin{aligned}\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{cls}}(R-R^*) &\geq \frac{1}{2}ch^{\beta+d}\|\phi\|_1\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\max_{\mathbf{v}\in\mathcal{V}}\sum_{k=1}^m \mathrm{P}(\hat{v}_k\neq v_k)\\ &\geq \frac{1}{2}h^{\beta+d}\|\phi\|_1\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sum_{k=1}^m \mathrm{P}(\hat{V}_k\neq V_k)\\ &= \frac{1}{2}h^{\beta+d}\|\phi\|_1\sum_{k=1}^m\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\mathrm{P}(\hat{V}_k\neq V_k), \qquad (86)\end{aligned}$$

in which the last step holds since $\hat{V}_k$ for different k are calculated independently.

It remains to give a lower bound of $\mathrm{P}(\hat{V}_k\neq V_k)$. Denote n_k as the number of samples falling in B_k, $\bar{Y}_k$ as the average label values in B_k:

$$\begin{aligned} n_k &:= \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i\in B_k), \qquad (87)\\ \bar{Y}_k &:= \frac{1}{n_k}\sum_{i=1}^N Y_i\mathbf{1}(X_i\in B_k). \qquad (88)\end{aligned}$$

Moreover, define

$$\begin{aligned} a_k &:= \frac{1}{n_k}\sum_{i=1}^N |\eta(\mathbf{X}_i)|\mathbf{1}(\mathbf{X}_i\in B_k)\\ &= \frac{h^\beta}{n_k}\sum_{i=1}^N \phi\left(\frac{\mathbf{X}_i-\mathbf{c}_k}{h}\right)\mathbf{1}(\mathbf{X}_i\in B_k), \qquad (89)\end{aligned}$$

in which the last step comes from (36). Then

$$\mathbb{E}[\bar{Y}_k|\mathbf{X}_{1:N},V_k]=V_ka_k, \qquad (90)$$

in which $\mathbf{X}_{1:N}$ means $\mathbf{X}_1,\ldots,\mathbf{X}_N$. We then show the following lemma.

Lemma 1. *If $0\leq t\leq \ln 2/(\epsilon n_k)$, and n_kt is an integer, then*

$$P(\hat{V}_k=1|\mathbf{X}_{1:N},\bar{Y}_k=-t)+P(\hat{V}_k=-1|\mathbf{X}_{1:N},\bar{Y}_k=t)\geq\frac{2}{3}. \qquad (91)$$

Proof. Construct D' by changing the label values of $l=n_kt$ items from these n_k samples falling in B_k, from -1 to 1. Then the average label values in B_k is denoted as $\bar{Y}_k'$ after such replacement. $\hat{V}_k$ also becomes $\hat{V}_k'$. Then from the ϵ-label DP requirement,

$$\begin{aligned}\mathrm{P}(\hat{V}_k=1|\mathbf{X}_{1:N},\bar{Y}_k=-t) &\overset{(a)}{\geq} e^{-l\epsilon}\mathrm{P}\left(\hat{V}_k'=1|\mathbf{X}_{1:N},\bar{Y}_k'=-t+\frac{2l}{n_k}\right)\\ &\overset{(b)}{\geq} e^{-l\epsilon}\mathrm{P}\left(\hat{V}_k=1|\mathbf{X}_{1:N},\bar{Y}_k=-t+\frac{2l}{n_k}\right)\\ &\geq e^{-n_kt\epsilon}\left[1-\mathrm{P}\left(\hat{V}_k=-1|\mathbf{X}_{1:N},\bar{Y}_k=-t+\frac{2l}{n_k}\right)\right]\\ &\geq \frac{1}{2}\left[1-\mathrm{P}\left(\hat{V}_k=-1|\mathbf{X}_{1:N},\bar{Y}_k=t\right)\right]. \qquad (92)\end{aligned}$$

in which (a) uses the group privacy property. The Hamming distance between D and D' is l, thus the ratio of probability between D and D' is within $[e^{-l\epsilon}, e^{l\epsilon}]$. (b) holds because the algorithm does not change after changing D to D'. Similarly,

$$\mathrm{P}(\hat{V}_k = -1|\mathbf{X}_{1:N}, \bar{Y}_k = t) \geq \frac{1}{2}\left[1 - \mathrm{P}\left(\hat{V}_k = 1|\mathbf{X}_{1:N}, \bar{Y}_k = -t\right)\right]. \tag{93}$$

Then (91) can be shown by adding up (92) and (93). □

Now we use Lemma 1 to bound the excess risk. With sufficiently large n_k, $\hat{Y}_k$ will be close to Gaussian distribution with mean a_k. To be more rigorous, by Berry-Esseen theorem [**?**], for some absolute constant C_E,

$$\mathrm{P}\left(\bar{Y}_k \leq a_k|\mathbf{X}_{1:N}, V_k = 1\right) \geq \frac{1}{2} - \frac{C_E}{\sqrt{n_k}}. \tag{94}$$

Similarly,

$$\mathrm{P}\left(\bar{Y}_k \geq -a_k|\mathbf{X}_{1:N}, V_k = -1\right) \geq \frac{1}{2} - \frac{C_E}{\sqrt{n_k}}. \tag{95}$$

We first analyze cubes with

$$n_k > 16C_E^2, a_k < \frac{\ln 2}{\epsilon n_k}. \tag{96}$$

Under condition (96), the right hand side of (94) and (95) are at least $1/4$. Therefore

$$\begin{aligned}
\mathrm{P}(\hat{V}_k \neq V_k|\mathbf{X}_{1:N}) &= \frac{1}{2}\mathrm{P}(\hat{V}_k = 1|\mathbf{X}_{1:N}, V_k = -1) + \frac{1}{2}\mathrm{P}(\hat{V}_k = -1|\mathbf{X}_{1:N}, V_k = 1)\\
&\geq \frac{1}{8}\mathrm{P}\left(\hat{V}_k = 1|\mathbf{X}_{1:N}, \bar{Y}_k \geq -\frac{\ln 2}{\epsilon n_k}\right) + \frac{1}{8}\mathrm{P}\left(\hat{V}_k = -1|\mathbf{X}_{1:N}, \bar{Y}_k \leq \frac{\ln 2}{\epsilon n_k}\right)\\
&\geq \frac{1}{12}.
\end{aligned}\tag{97}$$

From (86),

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{cls}}(R - R^*) \geq \frac{1}{2}h^{\beta+d}\left\|\phi\right\|_1\sum_{k=1}^m \frac{1}{12}\mathrm{P}\left(a_k < \frac{\ln 2}{\epsilon n_k}, n_k > 16C_E^2\right) \tag{98}$$

From (35), (89) and (87), $a_k \leq h^\beta$. Therefore

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{cls}}(R - R^*) \geq \frac{1}{24}h^{\beta+d}\left\|\phi\right\|_1\sum_{k=1}^m \mathrm{P}\left(16C_E^2 < n_k < \frac{\ln 2}{\epsilon h^\beta}\right). \tag{99}$$

Recall that each cube has probability mass ch^d. Select h such that

$$2Nch^d = \frac{\ln 2}{\epsilon h^\beta}. \tag{100}$$

From Chernoff inequality, $16C_E^2 < n_k < \ln 2/(\epsilon h^\beta)$ holds with high probability. (100) yields

$$h \sim (\epsilon N)^{-\frac{1}{d+\beta}}. \tag{101}$$

Recall the bound of m in (37). Let $m \sim h^{\gamma\beta - d}$, then (99) becomes

$$\begin{aligned}
\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{cls}}(R - R^*) &\gtrsim h^{\beta(\gamma+1)}\\
&\gtrsim (\epsilon N)^{-\frac{\beta(\gamma+1)}{d+\beta}}.
\end{aligned}\tag{102}$$

Moreover, the standard lower bound for classification [28] is

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{cls}}(R - R^*) \gtrsim N^{-\frac{\beta(\gamma+1)}{2\beta+d}}. \tag{103}$$

Therefore

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{cls}}(R - R^*) \gtrsim N^{-\frac{\beta(\gamma+1)}{2\beta+d}} + (\epsilon N)^{-\frac{\beta(\gamma+1)}{d+\beta}}. \tag{104}$$

E Proof of Theorem 4

Denote

$$n_l^* = \max_j n_{lj}, \tag{105}$$

$$n_l := \sum_{j=1}^K n_{lj} = \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l). \tag{106}$$

For all j such that $n_l^* - n_{lj} > t$,

$$\begin{aligned} \mathrm{P}(c_l = j | \mathbf{X}_{1:N}, Y_{1:N}) &= \frac{e^{\epsilon n_{lj}/2}}{\sum_{k=1}^K e^{\epsilon n_{lk}/2}} \\ &\leq \frac{e^{\epsilon n_l^*/2}}{\sum_{k=1}^K e^{\epsilon n_{lk}/2}} e^{-\frac{1}{2}\epsilon t} \\ &\leq e^{-\frac{1}{2}\epsilon t}. \end{aligned} \tag{107}$$

Therefore

$$\mathrm{P}(n_l^* - n_{lc_l} > t) = \sum_{j: n_l^* - n_{lj} > t} \mathrm{P}(c_l = j | \mathbf{X}_{1:N}, Y_{1:N}) \leq K e^{-\frac{1}{2}\epsilon t}. \tag{108}$$

Hence

$$\begin{aligned} \mathbb{E}[n_l^* - n_{lc_l}] &= \int_0^\infty \mathrm{P}(n_l^* - n_{lj} > t) dt \\ &\leq \int_0^{2\ln K/\epsilon} 1 dt + \int_{2\ln K/\epsilon}^\infty K e^{-\frac{1}{2}\epsilon t} dt \\ &= \frac{2}{\epsilon}(\ln K + 1). \end{aligned} \tag{109}$$

Define

$$v_{lj} = \frac{1}{n_l} \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l) \eta_j(\mathbf{X}_i), \tag{110}$$

then

$$\mathbb{E}[n_{lj} | \mathbf{X}_{1:N}] = n_l v_{lj}. \tag{111}$$

From Hoeffding's inequality,

$$\mathrm{P}(|n_{lj} - n_l v_{lj}| > t) \leq 2 e^{-\frac{1}{2n_l} t^2}. \tag{112}$$

Thus

$$\begin{aligned} \mathbb{E}\left[\max_j |n_{lj} - n_l v_{lj}|\right] &= \int_0^\infty \mathrm{P}\left(\cup_{j=1}^K \{|n_{lj} - n_l v_{lj}| > t\}\right) dt \\ &\leq \int_0^\infty \min\left(1, 2K e^{-\frac{1}{2n_l} t^2}\right) dt \\ &= \sqrt{2 n_l \ln(2K)} + \int_{\sqrt{2 n_l \ln(2K)}}^\infty 2K e^{-\frac{1}{2n_l} t^2} dt \\ &< 2\sqrt{2 n_l \ln(2K)}, \end{aligned} \tag{113}$$

in which the last step uses the inequalit $\int_t^\infty e^{-u^2/(2\sigma^2)} du \leq \sqrt{2\pi}\sigma e^{-t^2/(2\sigma^2)}$. Then

$$\begin{aligned} \mathbb{E}[v_l^* - v_{lc_l} | \mathbf{X}_{1:N}] &= \frac{1}{n_l} \mathbb{E}[n_l v_l^* - n_l v_{lc_l}] \\ &= \frac{1}{n_l} \mathbb{E}\left[n_l^* - n_{lc_l} + n_l v_l^* - n_l^* + n_{lc_l} - n_l v_{lc_l}\right] \\ &\leq \frac{1}{n_l} \mathbb{E}[n_l^* - n_{lc_l}] + \frac{2}{n_l} \mathbb{E}\left[\max_j |n_{lj} - n_l v_{lj}|\right] \\ &\leq \frac{2}{\epsilon n_l}(\ln K + 1) + 4\sqrt{\frac{2\ln(2K)}{n_l}}. \end{aligned} \tag{114}$$

By Hölder continuity assumption (Assumption 1(a)), for $\mathbf{x} \in B_l$,

$$|v_{lj} - \eta_j(\mathbf{x})| \leq \frac{1}{n_l} \sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l) |\eta_j(\mathbf{X}_i) - \eta_j(\mathbf{x})| \leq L_d h^\beta, \tag{115}$$

in which $L_d = L\sqrt{d}$, L is the constant in Assumption 1(a). Thus

$$\mathbb{E}[\eta^*(\mathbf{x}) - \eta_{c_l}(\mathbf{x}) | \mathbf{X}_{1:N}] \leq \frac{2}{\epsilon n_l}(\ln K + 1) + 4\sqrt{\frac{2\ln(2K)}{n_l}} + 2L_d h^\beta. \tag{116}$$

Now take integration over B_l.

$$\begin{aligned}
&\int_{B_l} (\eta^*(\mathbf{x}) - \mathbb{E}[\eta_{c_l}(\mathbf{x})]) f(\mathbf{x}) d\mathbf{x} \\
\leq\;& \mathrm{P}\left(n_l < \frac{1}{2} N p(B_l)\right) \int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}\left[\eta_{c_l}(\mathbf{x}) | n_l < \frac{1}{2} N p(B_l)\right]\right) f(\mathbf{x}) d\mathbf{x} \\
&+ \int_{B_l} \left(\eta^*(\mathbf{x}) - \mathbb{E}\left[\eta_{c_l}(\mathbf{x}) | n_l \geq \frac{1}{2} N p(B_l)\right]\right) f(\mathbf{x}) d\mathbf{x} \\
\leq\;& p(B_l) \exp\left[-\frac{1}{2}(1 - \ln 2) N p(B_l)\right] + \left[\frac{2(\ln K + 1)}{\epsilon N p(B_l)} + 4\sqrt{\frac{2\ln(2K)}{N p(B_l)}} + 2L_d h^\beta\right] p(B_l),
\end{aligned} \tag{117}$$

in which $p(B_l) = \mathrm{P}(\mathbf{X} \in B_l) = \int_{B_l} f(\mathbf{x}) d\mathbf{x}$. (117) is the central label DP counterpart of (67). The remainder of the proof follows arguments of the local label DP. We omit detailed steps. The result is

$$R - R^* \lesssim \left(\frac{\ln K}{\epsilon N h^d} + \sqrt{\frac{\ln K}{N h^d}} + h^\beta\right)^{\gamma+1}. \tag{118}$$

Let

$$h \sim \left(\frac{\ln K}{\epsilon N}\right)^{\frac{1}{\beta+d}} + \left(\frac{\ln K}{N}\right)^{\frac{1}{2\beta+d}}, \tag{119}$$

then

$$R - R^* \lesssim \left(\frac{\ln K}{\epsilon N}\right)^{\frac{\beta(\gamma+1)}{\beta+d}} + \left(\frac{\ln K}{N}\right)^{\frac{\beta(\gamma+1)}{2\beta+d}}. \tag{120}$$

The proof is complete.

F Proof of Theorem 5

Find G cubes in the support and the length of each cube is h. Let $\phi(\mathbf{u})$ be the same as the classification case shown in appendix B. For $\mathbf{v} \in \mathcal{V} := \{-1, 1\}^G$, let

$$\eta_{\mathbf{v}}(\mathbf{x}) = \sum_{k=1}^K v_k \phi\left(\frac{\mathbf{x} - \mathbf{c}_k}{h}\right) h^\beta. \tag{121}$$

Let $\mathrm{P}(Y = 1|\mathbf{x}) = (1 + \eta_{\mathbf{v}}(\mathbf{x}))/2$, $\mathrm{P}(Y = -1|\mathbf{x}) = (1 - \eta_{\mathbf{v}}(\mathbf{x}))$, then $\eta(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \eta_{\mathbf{v}}(\mathbf{x})$.

The overall volume of the support is bounded. Thus, we have

$$G \leq C_G h^{-d} \tag{122}$$

for some constant C_G.

Denote

$$\hat{v}_k = \mathrm{sign}\left(\int_{B_k} \hat{\eta}(\mathbf{x}) \phi\left(\frac{\mathbf{x} - \mathbf{c}_k}{h}\right) f(\mathbf{x}) d\mathbf{x}\right), \tag{123}$$

then the excess risk is bounded by

$$\begin{aligned} R &= \mathbb{E}\left[(\hat{\eta}(\mathbf{X}) - \eta_{\mathbf{v}}(\mathbf{X}))^2\right] \\ &= \sum_{k=1}^{K} \int_{B_k} \mathbb{E}\left[(\hat{\eta}(\mathbf{x}) - \eta_{\mathbf{v}}(\mathbf{x}))^2\right] f(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{124}$$

If $\hat{v}_k \neq v_k$, from (123),

$$\int_{B_k} \left(\hat{\eta}(\mathbf{x}) - v_k \phi\left(\frac{\mathbf{x} - \mathbf{c}_k}{h}\right) h^\beta\right)^2 f(\mathbf{x}) d\mathbf{x} \geq \int_{B_k} \left(\hat{\eta}(\mathbf{x}) + v_k \phi\left(\frac{\mathbf{x} - \mathbf{c}_k}{h}\right) h^\beta\right)^2 f(\mathbf{x}) d\mathbf{x}. \tag{125}$$

Therefore, if $\hat{v}_k \neq v_k$, then

$$\int_{B_k} (\hat{\eta}(\mathbf{x}) - \eta_{\mathbf{v}}(\mathbf{x}))^2 d\mathbf{x} \geq \frac{1}{2} \int_{B_k} \phi^2 \left(\frac{\mathbf{x} - \mathbf{c}_k}{h}\right) h^{2\beta} f(\mathbf{x}) d\mathbf{x} = \frac{1}{2} c h^{2\beta+d} \|\phi\|_2^2. \tag{126}$$

Therefore

$$\begin{aligned} R - R^* &\geq \mathbb{E}\left[\frac{1}{2} c h^{2\beta+d} \|\phi\|_2^2 \mathbf{1}(\hat{v}_k \neq v_k)\right] \\ &= \frac{1}{2} c h^{2\beta+d} \|\phi\|_2^2 \mathbb{E}[\rho_H(\hat{\mathbf{v}}, \mathbf{v})]. \end{aligned} \tag{127}$$

Similar to the classification problem analyzed in Appendix B, let

$$h \sim \left(N(\epsilon \wedge 1)^2\right)^{-\frac{1}{2\beta+d}}, \tag{128}$$

then $\delta \lesssim 1$, and

$$\inf_{\hat{\mathbf{v}}} \sup_{M \in \mathcal{M}_\epsilon} \max_{\mathbf{v} \in \mathcal{V}} \mathbb{E}[\rho_H(\hat{\mathbf{v}}, \mathbf{v})] \gtrsim G \sim h^{-d}. \tag{129}$$

Thus

$$\inf_{\hat{\eta}} \inf_{M \in \mathcal{M}_\epsilon} \sup_{P_{X,Y} \in \mathcal{F}_{reg1}} R \gtrsim h^{2\eta+d} h^{-d} \sim h^{2\beta} \sim (N(\epsilon \wedge 1)^2)^{-\frac{2\beta}{2\beta+d}}. \tag{130}$$

G Proof of Theorem 6

According to Assumption 2, $|Y| < T$ with probability 1, thus $\text{Var}[Y|\mathbf{x}] \leq T^2$ for any $\mathbf{x}$. A Laplacian distribution with parameter λ has variance $2\lambda^2$, thus

$$\text{Var}[W] = 2\lambda^2 = 2\left(\frac{2T}{\epsilon}\right)^2 = \frac{8T^2}{\epsilon^2}. \tag{131}$$

Hence

$$\text{Var}[Z] = \text{Var}[Y] + \text{Var}[W] \leq T^2 \left(1 + \frac{8}{\epsilon^2}\right). \tag{132}$$

Now we analyze the bias first.

$$\mathbb{E}[\hat{\eta}(\mathbf{x})] = \mathbb{E}\left[\frac{1}{k} \sum_{i \in \mathcal{N}_k(\mathbf{x})} Z_i\right] = \mathbb{E}\left[\frac{1}{k} \sum_{i \in \mathcal{N}_k(\mathbf{x})} \eta(\mathbf{X}_i)\right]. \tag{133}$$

Thus

$$\begin{aligned}
|\mathbb{E}[\hat{\eta}(\mathbf{x})]-\eta(\mathbf{x})| &\leq \mathbb{E}\left[\frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}|\eta(\mathbf{X}_i)-\eta(\mathbf{x})|\right] \\
&\leq \mathbb{E}\left[\frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\min\left\{L\left\|\mathbf{X}_i-\mathbf{x}\right\|^{\beta},2T\right\}\right] \\
&\leq \mathbb{E}\left[\frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\min\left\{L\rho^{\beta}(\mathbf{x}),2T\right\}\right] \\
&\leq 2T\mathrm{P}(\rho(\mathbf{x})>r_0)+Lr_0^{\beta} \\
&\leq 2Te^{-(1-\ln 2)k}+L\left(\frac{2k}{Ncv_d\theta}\right)^{\frac{\beta}{d}} \\
&\leq C_1\left(\frac{k}{N}\right)^{\frac{\beta}{d}}, \qquad (134)
\end{aligned}$$

for some constant C_1.

It remains to bound the variance.

$$\mathrm{Var}[\hat{\eta}(\mathbf{x})]=\mathbb{E}\left[\mathrm{Var}\left[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N\right]\right]+\mathrm{Var}[\mathbb{E}[\hat{\eta}(\mathbf{x})]|\mathbf{X}_1,\ldots,\mathbf{X}_N]. \qquad (135)$$

For the first term in (135),

$$\begin{aligned}
\mathrm{Var}[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N] &= \mathrm{Var}\left[\frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}Z_i|\mathbf{X}_1,\ldots,\mathbf{X}_N\right] \\
&= \frac{1}{k^2}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\mathrm{Var}[Z_i|\mathbf{X}_1,\ldots,\mathbf{X}_N] \\
&\leq \frac{1}{k}T^2\left(1+\frac{8}{\epsilon^2}\right). \qquad (136)
\end{aligned}$$

For the second term in (135),

$$\begin{aligned}
\mathrm{Var}[\mathbb{E}[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N]] &= \mathrm{Var}\left[\frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\eta(\mathbf{X}_i)\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\eta(\mathbf{X}_i)-\eta(\mathbf{x})\right)^2\right] \\
&= \frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\mathbb{E}\left[(\eta(\mathbf{X}_i)-\eta(\mathbf{x}))^2\right] \\
&\leq \frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\mathbb{E}\left[\min\left\{L^2\left\|\mathbf{X}_i-\mathbf{x}\right\|^{2\beta},4T^2\right\}\right] \\
&\leq 4T^2e^{-(1-\ln 2)k}+L^2r_0^{2\beta} \\
&\leq C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}. \qquad (137)
\end{aligned}$$

Therefore (135) becomes

$$\mathrm{Var}[\hat{\eta}(\mathbf{x})]\leq\frac{1}{k}T^2\left(1+\frac{8}{\epsilon^2}\right)+C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}. \qquad (138)$$

Combine the analysis of bias and variance,

$$\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2] \leq \frac{1}{k}T^2\left(1+\frac{8}{\epsilon^2}\right) + 2C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}. \tag{139}$$

Therefore the overall risk is bounded by

$$R = \mathbb{E}[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2] \lesssim \frac{1}{k}T^2\left(1+\frac{8}{\epsilon^2}\right) + 2C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}. \tag{140}$$

The optimal growth rate of k over N is

$$k \sim N^{\frac{2\beta}{d+2\beta}}(\epsilon \wedge 1)^{-\frac{2d}{d+2\beta}}. \tag{141}$$

Then the convergence rate of the overall risk becomes

$$R \lesssim (N(\epsilon\wedge 1)^2)^{-\frac{2\beta}{d+2\beta}}. \tag{142}$$

H Proof of Theorem 7

From (127),

$$\begin{aligned} R - R^* &\geq \frac{1}{2}ch^{2\beta+d}\left\|\phi\right\|_2^2 \mathbb{E}[\rho_H(\hat{\mathbf{V}}, \mathbf{V})] \\ &= \frac{1}{2}ch^{2\beta+d}\left\|\phi\right\|_2^2 \sum_{k=1}^G \mathrm{P}(\hat{V}_k \neq V_k). \end{aligned} \tag{143}$$

Follow the analysis of lower bounds of classification in Appendix D, let h scales as (101), then $\mathrm{P}(\hat{V}_k \neq V_k) \gtrsim 1$. Moreover, $G \sim h^{-d}$. Hence

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{reg1}}(R-R^*) \gtrsim h^{2\beta} \sim (\epsilon N)^{-\frac{2\beta}{d+\beta}}. \tag{144}$$

Moreover, note that the non-private lower bound of regression is

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{reg1}}(R-R^*) \gtrsim N^{-\frac{2\beta}{2\beta+d}}. \tag{145}$$

Combine (144) and (145),

$$\inf_{\mathcal{A}\in\mathcal{A}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}_{reg1}}(R-R^*) \gtrsim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta}{d+\beta}}. \tag{146}$$

I Proof of Theorem 8

1) Analysis of bias. Note that

$$\mathbb{E}[\hat{\eta}_l|\mathbf{X}_{1:N}] = \mathbb{E}[Y|\mathbf{X}\in B_l] = \frac{1}{p(B_l)}\int \eta(\mathbf{u})f(\mathbf{u})d\mathbf{u}. \tag{147}$$

Therefore, for all $\mathbf{x}\in B_l$,

$$\begin{aligned} |\mathbb{E}[\hat{\eta}_l|\mathbf{X}_{1:N}] - \eta(\mathbf{x})| &\leq \frac{1}{p(B_l)}\int |\eta(\mathbf{u}) - \eta(\mathbf{x})| f(\mathbf{u})d\mathbf{u} \\ &\leq L_d h^\beta. \end{aligned} \tag{148}$$

Therefore for all $\mathbf{x}\in B_l$,

$$|\mathbb{E}[\hat{\eta}_l] - \eta(\mathbf{x})| \leq L_d h^\beta. \tag{149}$$

2) Analysis of variance. If $n_l > 0$,

$$\mathrm{Var}\left[\frac{1}{n_l}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i\in B_l)Y_i|\mathbf{X}_{1:N}\right] = \frac{1}{n_l}\mathrm{Var}[Y|\mathbf{X}\in B_l] \leq \frac{1}{n_l}. \tag{150}$$

Therefore

$$\begin{aligned}\operatorname{Var}\left[\frac{1}{n_l}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i\in B_l)Y_i\right] &\leq \mathrm{P}\left(n_l<\frac{1}{2}Np(B_l)\right)+\mathrm{P}\left(n_l\geq\frac{1}{2}Np(B_l)\right)\frac{2}{Np(B_l)}\\ &\leq \exp\left[-\frac{1}{2}(1-\ln 2)Np(B_l)\right]+\frac{2}{Nch^d}. \end{aligned} \tag{151}$$

Similarly,

$$\begin{aligned}\operatorname{Var}[W_l] &\leq \mathrm{P}\left(n_l<\frac{1}{2}Np(B_l)\right)\frac{1}{\epsilon^2}+\mathrm{P}\left(n_l\geq\frac{1}{2}Np(B_l)\right)\frac{8}{\left(\frac{1}{2}Np(B_l)\right)^2\epsilon^2}\\ &\lesssim \frac{1}{N^2h^{2d}\epsilon^2}. \end{aligned} \tag{152}$$

The mean squared error can then be bounded by the bounds of bias and variance.

$$\mathbb{E}\left[(\hat{\eta}(\mathbf{x})-\eta(\mathbf{x}))^2\right]\lesssim h^{2\beta}+\frac{1}{Nh^d}+\frac{1}{N^2h^{2d}\epsilon^2}. \tag{153}$$

Let

$$h\sim N^{-\frac{1}{2\beta+d}}+(\epsilon N)^{-\frac{1}{d+\beta}}. \tag{154}$$

Then

$$R-R^*\lesssim N^{-\frac{2\beta}{2\beta+d}}+(\epsilon N)^{-\frac{2\beta}{d+\beta}}. \tag{155}$$

J Proof of Theorem 9

Now we prove the minimax lower bound of nonparametric regression under label DP constraint. We focus on the case in which ϵ is small.

Similar to the steps of the proof of Theorem 5 in Appendix F, we find B cubes in the support. The definition of $\eta_{\mathbf{v}}$, $\hat{v}_k$ are also the same as (121) and (123). Compared with the case with bounded noise, now Y can take values in $\mathbb{R}$.

For given $\mathbf{x}$, let

$$Y=\begin{cases} T & \text{with probability} \quad \frac{1}{2}\left(\frac{M_p}{T^p}+\frac{\eta_{\mathbf{v}}(\mathbf{x})}{T}\right)\\ 0 & \text{with probability} \quad 1-\frac{M_p}{T^p}\\ -T & \text{with probability} \quad \frac{1}{2}\left(\frac{M_p}{T^p}-\frac{\eta_{\mathbf{v}}(\mathbf{x})}{T}\right). \end{cases} \tag{156}$$

In Appendix F about the case with bounded noise, T is a fixed constant. However, here T is not fixed and will change over N. It is straightforward to show that the distribution of Y in (156) satisfies Assumption 3:

$$\mathbb{E}[|Y|^p|\mathbf{x}]=M_p. \tag{157}$$

Moreover, by taking expectation over Y, it can be shown that $\eta_{\mathbf{v}}$ is still the regression function:

$$\mathbb{E}[Y|\mathbf{x}]=\eta_{\mathbf{v}}(\mathbf{x}). \tag{158}$$

Let

$$T=\left(\frac{1}{2}M_p h^{-\beta}\right)^{\frac{1}{p-1}}. \tag{159}$$

Here we still define

$$\delta=\sup_{M\in\mathcal{M}_\epsilon}\max_{\mathbf{v},\mathbf{v}':\rho_H(\mathbf{v},\mathbf{v}')=1} D(P_{(X,Z)_{1:N}|\mathbf{v}}||P_{(X,Z)_{1:N}|\mathbf{v}'}). \tag{160}$$

Without loss of generality, suppose that $v_1 = v_1'$ for $i \neq 1$. Then

$$\begin{aligned}
D(P_{(X,Z)_{1:N}|\mathbf{v}}||P_{(X,Z)_{1:N}|\mathbf{v}'}) &= ND(P_{X,Z|\mathbf{v}}||P_{X,Z|\mathbf{v}'}) \\
&= N\int_{B_1} f(\mathbf{x}) D(P_{Z|\mathbf{X},\mathbf{v}}||P_{Z|\mathbf{X},\mathbf{v}'}) d\mathbf{x} \\
&\leq N\int_{B_1} f(\mathbf{x})(e^\epsilon - 1)^2 \mathbb{TV}^2\left(P_{Z|X,\mathbf{v}}, P_{Z|X,\mathbf{v}'}\right) d\mathbf{x} \\
&= N\int_{B_1} f(\mathbf{x})(e^\epsilon-1)^2 \eta_{\mathbf{v}}^2(\mathbf{x})\frac{1}{T^2} d\mathbf{x} \\
&= N(e^\epsilon-1)^2\frac{h^{2\beta}}{T^2}\int_{B_1} f(\mathbf{x})\phi^2\left(\frac{\mathbf{x}-\mathbf{c}_1}{h}\right) d\mathbf{x} \\
&= N(e^\epsilon-1)^2 h^{2\beta+d}\left\|\phi\right\|_2^2 T^{-2} \\
&= N(e^\epsilon-1)^2\left\|\phi\right\|_2^2\left(\frac{1}{2}M_p\right)^{-\frac{2}{p-1}} h^{2\beta+d+\frac{2\beta}{p-1}}. \qquad (161)
\end{aligned}$$

Let

$$h \sim (N(e^\epsilon-1)^2)^{-\frac{p-1}{2p\beta+d(p-1)}}, \qquad (162)$$

then $\delta \lesssim 1$. Hence

$$\inf_{\hat{\eta}}\inf_{M\in\mathcal{M}_\epsilon}\sup_{(f,\eta)\in\mathcal{F}} R \gtrsim h^{2\beta} \sim (N(e^\epsilon-1)^2)^{-\frac{2\beta(p-1)}{2p\beta+d(p-1)}}. \qquad (163)$$

K Proof of Theorem 10

Define

$$\eta_T(\mathbf{x}) := \mathbb{E}[Y_T|\mathbf{x}]. \qquad (164)$$

Then

$$\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}) = \eta_T(\mathbf{x}) - \eta(\mathbf{x}) + \mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta_T(\mathbf{x}) + \hat{\eta}(\mathbf{x}) - \mathbb{E}[\hat{\eta}(\mathbf{x})]. \qquad (165)$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[(\hat{\eta}(\mathbf{x})-\eta(\mathbf{x}))^2\right] &\leq 3(\eta_T(\mathbf{x})-\eta(\mathbf{x}))^2 + 3(\mathbb{E}[\hat{\eta}(\mathbf{x})]-\eta_T(\mathbf{x}))^2 + 3\operatorname{Var}[\hat{\eta}(\mathbf{x})] \\
&:= 3(I_1+I_2+I_3). \qquad (166)
\end{aligned}$$

Now we bound I_1, I_2 and I_3 separately.

Bound of I_1. We show the following lemma (which will also be used later).

Lemma 2.

$$|\eta_T(\mathbf{x}) - \eta(\mathbf{x})| \leq \frac{M_p}{p-1} T^{1-p}. \qquad (167)$$

Proof. Firstly, we decompose $\eta_T(\mathbf{x})$ and $\eta(\mathbf{x})$:

$$\eta_T(\mathbf{x}) = \mathbb{E}[Y_T|\mathbf{x}] = \mathbb{E}[Y\mathbf{1}(-T\leq Y\leq T)|\mathbf{x}] + T\text{P}(Y>T|\mathbf{x}) - TP(Y<T|\mathbf{x}), \qquad (168)$$

$$\eta(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \mathbb{E}[Y\mathbf{1}(-T\leq Y\leq T)|\mathbf{x}] + \mathbb{E}[Y\mathbf{1}(Y>T)|\mathbf{x}] - \mathbb{E}[Y\mathbf{1}(Y<T)|\mathbf{x}]. \qquad (169)$$

The first term is the same between (168) and (169). Therefore we only need to compare the second and the third term.

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}(Y>T)|\mathbf{x}] &= \int_0^T \text{P}(Y>T|\mathbf{x})dt + \int_T^\infty \text{P}(Y>T|\mathbf{x})dt \\
&\leq T\text{P}(Y>T|\mathbf{x}) + \int_T^\infty M_p t^{-p} dt \\
&= T\text{P}(Y>T|\mathbf{x}) + \frac{M_p}{p-1}T^{1-p}. \qquad (170)
\end{aligned}$$

Therefore

$$\mathbb{E}[Y\mathbf{1}(Y>T)|\mathbf{x}] - T\mathrm{P}(Y>T|\mathbf{x}) \leq \frac{M_p}{p-1}T^{1-p}. \tag{171}$$

Similarly,

$$TP(Y<T|\mathbf{x}) - \mathbb{E}[Y\mathbf{1}(Y<T)|\mathbf{x}] \leq \frac{M_p}{p-1}T^{1-p}. \tag{172}$$

A Combination of these two inequalities yields the (167). □

With Lemma 2,

$$I_1 \leq \frac{M_p^2}{(p-1)^2}T^{2(1-p)}. \tag{173}$$

Bound of I_2. Follow the steps in (134),

$$I_2 \leq C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}. \tag{174}$$

Bound of I_3. We decompose $\mathrm{Var}[\hat{\eta}(\mathbf{x})]$ as following:

$$\mathrm{Var}[\hat{\eta}(\mathbf{x})] = \mathbb{E}[\mathrm{Var}[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N]] + \mathrm{Var}[\mathbb{E}[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N]]. \tag{175}$$

For the first term in (175), from Assumption 3, $\mathbb{E}[|Y|^p|\mathbf{x}] \leq M_p$. Since $p \geq 2$, we have $\mathbb{E}[Y^2|\mathbf{x}] = M_p^{\frac{2}{p}}$. Therefore

$$\mathrm{Var}[Z_i|\mathbf{X}_1,\ldots,\mathbf{X}_N] = \mathrm{Var}[Y_T] + \mathrm{Var}[W] \leq M_p^{\frac{2}{p}} + \frac{8T^2}{\epsilon^2}. \tag{176}$$

Recall (20), we have

$$\begin{aligned}\mathrm{Var}[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N] &= \frac{1}{k^2}\sum_{i\in\mathcal{N}_k(\mathbf{x})}\mathrm{Var}[Z_i|\mathbf{X}_1,\ldots,\mathbf{X}_N] \\ &\leq \frac{1}{k}\left(M_p^{\frac{2}{p}} + \frac{8T^2}{\epsilon^2}\right).\end{aligned} \tag{177}$$

For the second term in (175), (137) still holds, thus

$$\mathrm{Var}[\mathbb{E}[\hat{\eta}(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_N]] \leq C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}, \tag{178}$$

and

$$I_3 \leq \frac{1}{k}\left(M_p^{\frac{2}{p}} + \frac{8T^2}{\epsilon^2}\right) + C_1^2\left(\frac{k}{N}\right)^{\frac{2\beta}{d}}. \tag{179}$$

Plug (173), (174) and (179) into (166), and take expectations, we get

$$\begin{aligned}R &= \mathbb{E}[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2] \\ &\lesssim T^{2(1-p)} + \frac{1}{k} + \frac{T^2}{k\epsilon^2} + \left(\frac{k}{N}\right)^{\frac{2\beta}{d}}.\end{aligned} \tag{180}$$

Let

$$T \sim (k\epsilon^2)^{\frac{1}{2p}}, k \sim (N\epsilon^2)^{\frac{2p\beta}{d(p-1)+2p\beta}} \vee N^{\frac{2\beta}{2\beta+d}}, \tag{181}$$

then

$$R \lesssim (N\epsilon^2)^{-\frac{2\beta(p-1)}{d(p-1)+2p\beta}} \vee N^{-\frac{2\beta}{2\beta+d}}. \tag{182}$$

L Proof of Theorem 11

Let Y be distributed as (156). Recall Lemma 1 for the problem of classification and regression with bounded noise.

Now we show the corresponding lemma for regression with unbounded noise.

Lemma 3. *If $0 \leq t \leq T \ln 2/(\epsilon n_k)$, and $n_k t/T$ is an integer, then*

$$P(\hat{V}_k = 1|\mathbf{X}_{1:N}, \bar{Y}_k = -t) + P(\hat{V}_k = -1|\mathbf{X}_{1:N}, \bar{Y}_k = t) \geq \frac{2}{3}. \tag{183}$$

Here we briefly explain the condition $n_k t$ is an integer. Recall the definition of $\bar{Y}_k$ in (88). Now since Y take values in $\{-T, 0, T\}$, $n_k \bar{Y}_k/T$ must be an integer. Therefore, in Lemma 3, we only need to consider the case such that $n_k t/T$ is an integer.

Proof. The proof follows the proof of Lemma 1 closely. We provide the proof here for completeness. Construct D' by changing the label values of $l = n_k t/T$ items from these n_k samples falling in B_k, from $-T$ to T. Then the average label values in B_k is denoted as $\bar{Y}_k'$ after such replacement. $\hat{V}_k$ also becomes $\hat{V}_k'$. Then from the ϵ-label DP requirement,

$$\begin{aligned}
\mathrm{P}(\hat{V}_k = 1|\mathbf{X}_{1:N}, \bar{Y}_k = -t) &\overset{(a)}{\geq} e^{-l\epsilon}\mathrm{P}\left(\hat{V}_k' = 1|\mathbf{X}_{1:N}, \bar{Y}_k' = -t + \frac{2l}{n_k}\right) \\
&\overset{(b)}{\geq} e^{-l\epsilon}\mathrm{P}\left(\hat{V}_k = 1|\mathbf{X}_{1:N}, \bar{Y}_k = -t + \frac{2l}{n_k}\right) \\
&\geq e^{-n_k t\epsilon}\left[1 - \mathrm{P}\left(\hat{V}_k = -1|\mathbf{X}_{1:N}, \bar{Y}_k = -t + \frac{2l}{n_k}\right)\right] \\
&\geq \frac{1}{2}\left[1 - \mathrm{P}\left(\hat{V}_k = -1|\mathbf{X}_{1:N}, \bar{Y}_k = t\right)\right].
\end{aligned} \tag{184}$$

in which (a) uses the group privacy property. The Hamming distance between D and D' is l, thus the ratio of probability between D and D' is within $[e^{-l\epsilon}, e^{l\epsilon}]$. (b) holds because the algorithm does not change after changing D to D'. Similarly,

$$\mathrm{P}(\hat{V}_k = -1|\mathbf{X}_{1:N}, \bar{Y}_k = t) \geq \frac{1}{2}\left[1 - \mathrm{P}\left(\hat{V}_k = 1|\mathbf{X}_{1:N}, \bar{Y}_k = -t\right)\right]. \tag{185}$$

Then (183) can be shown by adding up (184) and (185). $\square$

We then follow the proof of Theorem 3 in Appendix D. (101) becomes

$$h \sim \left(\frac{\epsilon N}{T}\right)^{-\frac{1}{d+\beta}}. \tag{186}$$

In (156), note that $\mathrm{P}(Y = T) \geq 0$ and $\mathrm{P}(Y = -T) \geq 0$. Therefore $M_p/T^p \geq \eta_{\mathbf{v}}(\mathbf{x})/T$. This requires $h^\beta T^{p-1} \leq M_p$. Let $T \sim h^{-\frac{\beta}{p-1}}$, then

$$h \sim (\epsilon N)^{-\frac{1}{d+\beta}} h^{\frac{\beta}{(d+\beta)(p-1)}}, \tag{187}$$

i.e.

$$h \sim (\epsilon N)^{-\frac{p-1}{p\beta + d(p-1)}}. \tag{188}$$

Combine with standard minimax rate, the lower bound of regression with unbounded noise is

$$\inf_{\mathcal{A} \in \mathcal{A}_\epsilon} \sup_{(f,\eta) \in \mathcal{F}_{reg2}} (R - R^*) \gtrsim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta(p-1)}{p\beta + d(p-1)}}. \tag{189}$$

M Proof of Theorem 12

1) Analysis of bias. Note that Lemma 2 still holds here. Moreover, recall (149). Therefore

$$|\mathbb{E}[\hat{\eta}_l] - \eta(\mathbf{x})| \leq |\mathbb{E}[\hat{\eta}_l - \eta_T(\mathbf{x})]| + |\eta_T(\mathbf{x}) - \eta(\mathbf{x})| \leq L_d h^\beta + \frac{M_p}{p-1} T^{1-p}. \tag{190}$$

2) Analysis of variance. Similar to (151), it can be shown that

$$\mathrm{Var}\left[\frac{1}{n_l}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in B_l) Y_i\right] \lesssim \frac{1}{Nh^d}. \tag{191}$$

Moreover, the noise variance can be bounded by

$$\mathrm{Var}[W_l] \lesssim \frac{T^2}{N^2 h^{2d} \epsilon^2}. \tag{192}$$

The mean squared error is then bounded by

$$\mathbb{E}\left[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2\right] \lesssim h^{2\beta} + T^{2(1-p)} + \frac{T^2}{N^2 h^{2d}\epsilon^2} + \frac{1}{Nh^d}. \tag{193}$$

Let $T \sim (\epsilon N h^d)^{1/p}$, then

$$R - R^* = \mathbb{E}\left[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2\right] \lesssim h^{2\beta} + \frac{1}{Nh^d} + (\epsilon N h^d)^{-2(1-1/p)}. \tag{194}$$

To minimize (194), let

$$h \sim N^{-\frac{1}{2\beta+d}} + (\epsilon N)^{-\frac{p-1}{p\beta + d(p-1)}}, \tag{195}$$

then

$$R - R^* \lesssim N^{-\frac{2\beta}{2\beta+d}} + (\epsilon N)^{-\frac{2\beta(p-1)}{p\beta+d(p-1)}}. \tag{196}$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution (i.e. proposing a new Huber loss minimization approach which is more suitable to realistic cases, and providing theoretical analysis) has been made clear in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It is explained at the end of conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs are shown in the appendix, and intuition is provided in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This is a theoretical paper without experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This is a theoretical paper without experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This is a theoretical paper without experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: No experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper does not violate code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

   Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

   Answer: [NA]

   Justification: This paper is foundational and theoretical research and not tied to particular applications.

   Guidelines:

   - The answer NA means that there is no societal impact of the work performed.
   - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
   - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper has no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.