Supplement to "Minimax And Adaptive Transfer Learning for Nonparametric Classification under Distributed Differential Privacy Constraints"

Arnab Auddy

University of Pennsylvania, Philadelphia, Pennsylvania, United States.

T. Tony Cai

University of Pennsylvania, Philadelphia, Pennsylvania, United States.

Abhinav Chakraborty

University of Pennsylvania, Philadelphia, Pennsylvania, United States.

Summary. This supplementary material contains the proofs to all results in the main paper Auddy et al. [2024]. We restate the results in their respective sections and provide the corresponding proofs.

1. Introduction

The rest of the paper is organized as follows. In Section 2, we provide the background and formulate our problem in detail. Section 3 presents the minimax rate of excess risk for our problem across various specific cases, as well as the most general case. Section 4 introduces our kernel based classifier, derives its excess risk bounds, and states the minimax lower bound. Section 5 describes the data-driven adaptive procedure for bandwidth and weight selection. The proofs of all results are given in Section 6.

2. Problem Formulation

In this section, we outline the general framework for federated transfer learning under privacy constraints. Our dataset is distributed across $m+1$ servers, indexed by the set $\{0, 1, \ldots, m\}$. The dataset is categorized as target and source. On server 0 (also called the target server), we have n_0 i.i.d. samples from the distribution P_0, while on server j (the source servers) for $j \in 1, \ldots, m$, we have n_j i.i.d. samples from distribution P_j. All of the probability measures $\{P_j\}_{j=0}^m$ are defined on the measurable space $(\mathcal{Z}, \mathscr{Z})$.

Let $Z^{(0)} = \{Z_i^{(0)}\}_{i=1}^{n_0}$ denote the n_0 realizations from P_0 on the target server. Let us denote by $Z^{(j)} = \{Z_i^{(j)}\}_{i=1}^{n_j}$ the n_j realizations from P_j on the jth server for $j = 1, \ldots, m$. These servers serve as the source data, and our goal is to learn the model for our target distribution P_0.

For each source server i.e $j = 1, \ldots, m$, we send a (randomized) transcript $\widetilde{T}^{(j)}$ based on $Z^{(j)}$ to the target server 0, where the law of the transcript is given by a distribution conditional on $Z^{(j)}$, $\mathbb{P}(\cdot|Z^{(j)})$, on a measurable space $(\mathcal{T}, \mathscr{T})$. For $j = 1, \ldots, m$ the transcript $\widetilde{T}^{(j)}$ has to satisfy a $(\varepsilon_j, \delta_j)$-differential privacy constraint.

DEFINITION 2.1. *The transcript $\widetilde{T}^{(j)}$ is $(\varepsilon_j, \delta_j)$-differentially private if for all $A \in \mathcal{A}$ and z, z' differing in one individual datum, it holds that*

$$\mathbb{P}\left(\widetilde{T}^{(j)} \in A \,|\, Z^{(j)} = z\right) \leq e^{\varepsilon_j}\mathbb{P}\left(\widetilde{T}^{(j)} \in A \,|\, Z^{(j)} = z'\right) + \delta_j.$$

The target server can look at the private transcripts $\{\widetilde{T}^{(j)}\}_{j=1}^m$ and the target data $Z^{(0)}$ while constructing the final private transcript $\widetilde{T}$. Hence $\widetilde{T}$ satisfies $(\varepsilon_0, \delta_0)$-interactive differential privacy constraint, which is defined as follows:

DEFINITION 2.2. *The transcript $\widetilde{T}$ is $(\varepsilon_0, \delta_0)$-differentially private if for all $A \in \mathcal{A}$ and z, z' differing in one individual datum and for all $t_j \in \mathcal{T}$ for $j = 1, \ldots, m$, it holds that*

$$\begin{aligned}&\mathbb{P}\left(\widetilde{T} \in A \,|\, Z^{(0)} = z, \widetilde{T}^{(j)} = t_j \text{ for } 1 \leq j \leq m\right)\\ &\leq e^{\varepsilon_0}\mathbb{P}\left(\widetilde{T} \in A \,|\, Z^{(0)} = z', \widetilde{T}^{(j)} = t_j \text{ for } 1 \leq j \leq m\right) + \delta_0.\end{aligned}$$

This privacy constraint can be understood as follows: if we condition on the outcome of all other servers then the distribution of the final private transcript $\widetilde{T}$ does not change much if one of the datum on the target server is changed.

In transfer learning, our focus is on scenarios where multiple parties, such as hospitals, possess heterogeneous data with differing underlying distributions. Employing distributed protocols in such contexts ensures differential privacy while yielding outputs from each participating party. Within this framework, transcripts generated by each source server rely solely on its local data, with no exchange of information occurring between source servers. Communication is solely between the source and target servers. Each of the source server transmits its transcripts to the target server. The target server utilizing all the transcripts $(\widetilde{T}^{(1)}, \ldots, \widetilde{T}^{(m)})$ from the other servers and target data $Z^{(0)}$, computes the final private transcript $\widetilde{T}$. This scenario often arises when multiple trials involving a population similar to that of the target server are conducted, yet individual locations, such as hospitals, opt against consolidating their original data due to privacy apprehensions.

In the context of transfer learning for nonparametric classification our data looks like a couple $Z_i^{(j)} := (X_i^{(j)}, Y_i^{(j)})$, for $i = 1, \ldots, n_j$; $j = 1, \ldots, m$ for the source servers, and $Z_i^{(0)} := (X_i^{(0)}, Y_i^{(0)})$, for $i = 1, \ldots, n_0$ for the target server. We assume that $Z_i^{(j)}$ takes values in $\mathcal{Z} := [0,1]^d \times \{0,1\}$. We regard $X \in [0,1]^d$ as a vector of features corresponding to an object and $Y \in \{0,1\}$ as a label indicating that the object belongs to one of two classes. Our goal is to propose distributed DP protocols $\widetilde{T}^{(j)}$ for each server and construct classifier $\widehat{f} : [0,1]^d \to \{0,1\}$ based on the final private transcript $\{\widetilde{T}\}$. Unlike the traditional federated learning framework, there's no central server; alternatively, we can consider the target server as acting in a central capacity.

We denote the vector of privacy budgets as $(\boldsymbol{\varepsilon}, \boldsymbol{\delta}) = \{(\varepsilon_j, \delta_j)\}_{j=0}^m$ and the class of distributed DP classifiers $\widehat{f}$ by $\mathcal{M}_{\boldsymbol{\varepsilon},\boldsymbol{\delta}}$.

Next we denote

$$\begin{aligned}&\eta_j(X^{(j)}) := \mathbb{P}(Y^{(j)} = 1 | X^{(j)}) \text{ for the source servers } j = 1, \ldots, m; \text{ and}\\ &\eta_0(X^{(0)}) := \mathbb{P}(Y^{(0)} = 1 | X^{(0)}) \text{ for the target server,}\end{aligned}$$

as the (source and target) regression functions of Y on X. We denote the marginal distribution of X for the jth server, $j = 0, \ldots, m$ as P_j^X. Define the classification error of a classifier f under the target distribution P_0 as

$$R_0(f) := P_0(Y \neq f(X))$$

The Bayes decision rule is a minimizer of the of the risk $R_0(f)$ which has the form $f_0^*(X) = \mathbb{1}\{\eta_0(X) \geq 1/2\}$. The goal of transfer learning is to transfer the knowledge gained from the source data together with the information in the target data to construct a classifier which minimizes the excess risk on the target data

$$\mathcal{E}_0(\widehat{f}) = \mathbb{E}[R_0(\widehat{f})] - R_0(f_0^*)$$

Under the posterior drift model we quantify the similarity between the regression functions $\{\eta_j\}_{j=1}^m$ and η_0 as follows:

Definition 2.3 (**Relative Signal Exponent (RSE)**). *The class $\Gamma(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}})$ with relative signal exponent $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in \mathbb{R}_+^m$ and constants $C_{\boldsymbol{\gamma}} = (C_1, \ldots, C_m) \in \mathbb{R}_+^m$, is the set of distribution tuples $(P_0, P_1, \ldots, P_m)$ that satisfy for $1 \leq j \leq m$*

(a) $\operatorname{sign}\left(\eta_j(x) - \frac{1}{2}\right) = \operatorname{sign}\left(\eta_0(x) - \frac{1}{2}\right)$ *for all $1 \leq j \leq m$ and all $x \in [0,1]^d$.*

(b) $\left|\eta_j(x) - \frac{1}{2}\right| \geq C_j \left|\eta_0(x) - \frac{1}{2}\right|^{\gamma_j}$ *for some $\gamma_j > 0$, for all $1 \leq j \leq m$ and all $x \in [0,1]^d$.*

In addition to the RSE assumption we also need to assume smoothness of η_0 and characterize it behavior near 1/2.

Definition 2.4 (**Hölder Smoothness**). *The regression function η_0 belongs to the Hölder class of functions denoted by $\Sigma(\beta, L)$ $(0 < \beta \leq 1)$ which is defined as the set of functions satisfying:*

$$|\eta_0(x) - \eta_0(x')| \leq L\|x - x'\|^\beta \quad \text{for } x, x' \in [0,1]^d.$$

Definition 2.5 (**Margin Assumption (MA)**). *The margin class $\mathcal{M}(\alpha, C_\alpha)$ with $\alpha \geq 0$ and $C_\alpha > 0$ is defined as the set of distributions P_0 such that*

$$P_0^X(0 \leq |\eta_0(X) - 1/2| \leq t) \leq C_\alpha t^\alpha \text{ for all } t > 0.$$

Another definition is about marginal density of X, P_j^X for $j = 1, \ldots, m$.

Definition 2.6 (**Common Support and Strong Density Assumption (SD)**). *We assume that P_j^X for $j = 0, \ldots, m$ have the identical support on a compact (c_μ, r_μ) regular set $A \subset [0,1]^d$ and has a density g_j w.r.t. the Lebesgue measure bounded away from zero and infinity on A:*

$$g_{\min} \leq g_j(x) \leq g_{\max} \text{ for } x \in A \text{ and } g_j(x) = 0 \text{ otherwise,}$$

where $c_0, r_0 > 0$ and $0 < g_{\min} < g_{\max} < \infty$ are fixed constants. We denote the set of marginal distributions $(P_0^X, \ldots, P_m^X)$ which satisfy the above constraints as $\mathcal{S}(\mu, c_\mu, r_\mu)$ where $\mu = (g_{\min}, g_{\max})$.

REMARK 2.1. In this paper we focus our attention to the case when the marginal densities have regular support and are bounded from below and above on their support. Moreover we assume that $\alpha\beta \leq d$ throughout the paper. This is because in the other regime ($\alpha\beta > d$), there is no distribution P_0^X such that the regression function η_0 crosses $1/2$ in the interior of the support of P_0^X (Audibert and Tsybakov [2007]) and hence this case only contains the trivial cases for classification.

We put all the definitions together to define the class of distributions we consider in the posterior drift model as

$$\begin{aligned}&\Pi(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}, \beta, L, \alpha, C_\alpha, \mu, c_\mu, r_\mu)\\ :=\ &\{(P_0, P_1, \ldots, P_m) : (P_0, P_1, \ldots, P_m) \in \Gamma(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}), \eta_0 \in \Sigma(\beta, L),\\ &\qquad\qquad \mathbb{P}_0^X \in \mathcal{M}(\alpha, C_\alpha), (P_0^X, P_1^X, \ldots, P_m^X) \in \mathcal{S}(\mu, c_\mu, r_\mu)\}\end{aligned}$$

For the rest of the paper we will use the shorthand $\Pi(\alpha, \beta, \boldsymbol{\gamma}, \mu)$ or Π if there is no confusion.

3. Main Results

In this section, we present the key findings of our paper, where we establish the minimax rate of convergence for transfer learning under differential privacy constraints, specifically addressing the nonparametric classification problem.We divide our results into two subsections: Section 3.1 covers the homogeneous case, while Section 3.2 addresses the general heterogeneous case.

3.1. Minimax Rates under Source Homogeneity

To derive meaningful and interpretable insights from our minimax rate, we first examine the scenario where the source servers are exchangeable in terms of the distributed classification problem under transfer learning and privacy constraints. This homogeneous scenario is characterized by equal sample sizes ($n_j = n$), privacy parameters ($\varepsilon_j = \varepsilon$, $\delta_j = \delta$) and transfer exponents ($\gamma_j = \gamma$) for all $j = 1, \ldots, m$.

THEOREM 3.1. *Suppose $n_j = n, \varepsilon_j = \varepsilon, \delta_j = \delta$ and $\gamma_j = \gamma$ for all $j = 1, \ldots, m$ and assume that $\delta = o((nm)^{-1})$. Then the minimax rate for the excess risk satisfies*

$$\begin{aligned}\inf_{\widehat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{(P_0, \ldots, P_m) \in \Pi} \mathcal{E}_0(\widehat{f}) \asymp \Bigg[L_N \Bigg\{ &\left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2 \varepsilon_0^2)^{\frac{1}{2\beta+2d}} \right)\\ &+ \left((mn)^{\frac{1}{2\beta\gamma+d}} \wedge (mn^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}} \right) \Bigg\}^{-\beta(1+\alpha)} \wedge 1 \Bigg]\end{aligned}$$

for a sequence L_N of order at most $(\log((\delta \wedge \delta_0)^{-1}))^{\frac{\beta(1+\alpha)}{2\beta(\gamma \wedge 1)+d}}$.

In order to further our understanding about interplay between the transfer exponent γ and privacy parameters we restrict our attention to the case where the target server

has same privacy budget, i.e., $\varepsilon_0 = \varepsilon$, $\delta_0 = \delta$ and the number of target samples is between n and mn. Other sample size regimes can be described similarly. As is clear from Theorem 3.1, the minimax rate of decay for the excess risk is given in this case by:

$$\mathcal{E}_0(\widehat{f}) \asymp \left[L_N \left\{ \left(n_0^{-\frac{\beta(1+\alpha)}{2\beta+d}} \vee (n_0^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta+2d}} \right) \bigwedge \left((mn)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} \vee (mn^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+2d}} \right) \right\} \wedge 1 \right]. \tag{1}$$

We will refer to the four terms on the right hand side in (1) as the non-private target rate (NP_t), the private target rate (P_t), the non-private source rate (NP_s), and the private source rate (P_s) respectively. Depending on the value of the common privacy parameter ε and the transfer exponent γ, the overall rate will be determined by one of these four rates, as demonstrated by the following table. Corollary 3.2 formally states the results of Table 1 along with the endpoints of the various transfer and privacy regimes.

Table 1. Minimax rate of excess risk at different transfer and privacy regimes. See Corollary 3.2.

<table>
<tr><th>Transfer \ Privacy</th><th>$\varepsilon \in (0, \varepsilon^{(1)}]$</th><th>$\varepsilon \in (\varepsilon^{(1)}, \varepsilon^{(2)}]$</th><th>$\varepsilon \in (\varepsilon^{(2)}, \varepsilon^{(3)}]$</th><th>$\varepsilon \in (\varepsilon^{(3)}, 1]$</th></tr>
<tr><td>$\gamma \in (0, 1]$</td><td rowspan="3">1</td><td>$\begin{cases} \mathrm{P}_t & \text{if } \varepsilon \le \varepsilon^{(11)} \\ \mathrm{P}_s & \text{if } \varepsilon > \varepsilon^{(11)} \end{cases}$</td><td colspan="2">$\mathrm{NP}_s$</td></tr>
<tr><td>$\gamma \in (1, \gamma^{(*)}]$</td><td rowspan="2">$\begin{cases} \mathrm{P}_s & \text{if } \varepsilon \le \varepsilon^{(11)} \\ \mathrm{P}_t & \text{if } \varepsilon > \varepsilon^{(11)} \end{cases}$</td><td>$\begin{cases} \mathrm{NP}_t & \text{if } \varepsilon \le \varepsilon^{(21)} \\ \mathrm{P}_s & \text{if } \varepsilon > \varepsilon^{(21)} \end{cases}$</td><td>$\mathrm{NP}_s$</td></tr>
<tr><td>$\gamma \in (\gamma^{(*)}, \infty)$</td><td colspan="2">$\mathrm{NP}_t$</td></tr>
</table>

COROLLARY 3.2. *Suppose $n_j = n, \gamma_j = \gamma \, \forall 1 \le j \le m$, $n \le n_0 \le mn$, and equal privacy budget $\varepsilon_j = \varepsilon$, $\delta_j = \delta \, \forall 0 \le j \le m$. Further assume that $\delta = o((mn)^{-1})$. Then the minimax rate for the excess risk are as given in Table 1 with the various regimes characterized by the following endpoints:*

(a) $\gamma^{(*)} = \frac{1}{2\beta}\left[\frac{(2\beta+d)\log mn}{\log n_0} - d\right]$.

(b) $\varepsilon^{(1)} = (\sqrt{mn})^{-1} \wedge n_0^{-1}$; $\varepsilon^{(2)} = n_0^{-\frac{\beta}{2\beta+d}}$; $\varepsilon^{(3)} = \left(m^{d/2} n^{-\beta\gamma}\right)^{\frac{1}{2\beta\gamma+d}}$.

(c) $\varepsilon^{(11)} = \begin{cases} \left[(\sqrt{mn})^{\beta+d} n_0^{-(\beta\gamma+d)}\right]^{\frac{1}{\beta(\gamma-1)}} & \text{if } \gamma \ne 1, \\ \varepsilon^{(1)} & \text{if } \gamma = 1,\ n_0 \le \sqrt{mn^2}, \\ \varepsilon^{(2)} & \text{if } \gamma = 1,\ n_0 > \sqrt{mn^2}. \end{cases}$

(d) $\varepsilon^{(21)} = (\sqrt{mn})^{-1} n_0^{\frac{\beta\gamma+d}{\beta+d}}$.

In the rest of this subsection, we describe another specialized setting where we allow one of the source servers to be public. To demonstrate the effect of publicly available data, we take $m = 2$ sources, with one private and one public source server. The minimax rate for excess risk is then given by the following corollary:

COROLLARY 3.3. *Suppose that* $\gamma_1 = \gamma_2 = \gamma$, $\varepsilon_1 = \infty$, $\varepsilon_0 = \varepsilon_2 = \varepsilon$ *and* $n_2 > n_0^{\frac{2\beta\gamma+d}{2\beta+d}}$, $\delta_0 \vee \delta_2 = o(n_2^{-1})$. *Then the minimax rate for the excess risk satisfies the following.*

(a) If $n_1 > n_2$, $\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \asymp n_1^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}}$.

(b) If $n_0^{\frac{2\beta\gamma+d}{2\beta+d}} \le n_1 \le n_2$, *then*

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \asymp \begin{cases} L_N n_1^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} & \text{if } \varepsilon \le n_2^{-1} n_1^{\frac{\beta\gamma+d}{2\beta\gamma+d}} \\ L_N (n_2^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+2d}} & \text{if } n_2^{-1} n_1^{\frac{\beta\gamma+d}{2\beta\gamma+d}} < \varepsilon \le L_N n_2^{-\frac{\beta\gamma}{2\beta\gamma+d}} \\ L_N n_2^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} & \text{if } n_2^{-\frac{\beta\gamma+d}{2\beta\gamma+d}} < \varepsilon \le 1. \end{cases}$$

(c) If $n_1 \le n_0^{\frac{2\beta\gamma+d}{2\beta+d}} \le n_2$, *then*

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f})$$
$$\asymp \begin{cases} L_N n_1^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} & \text{if } \varepsilon \le \widetilde{n}^{-\beta} \\ \left[L_N \left\{ \left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2\varepsilon^2)^{\frac{1}{2\beta+2d}} \right) + \left(n_2^{\frac{1}{2\beta\gamma+d}} \wedge (n_2^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}} \right) \right\}^{-\beta(1+\alpha)} \wedge 1 \right] & \text{otherwise,} \end{cases}$$

where $\widetilde{n} = n_0^{\frac{1}{2\beta+d}} \wedge n_2^{\frac{\gamma}{2\beta\gamma+d}}$. *Here* L_N *is a sequence of order at most* $(\log((\delta_0\wedge\delta_2)^{-1}))^{\frac{\beta(1+\alpha)}{2\beta(\gamma\wedge 1)+d}}$.

3.2. *Minimax Rates in General Setting*

We now turn our attention to the general case where the sample sizes n_j, transfer exponents γ_j, privacy parameters $(\varepsilon_j, \delta_j)$ are all allowed to vary for $0 \le j \le m$. Our main result, captured in Theorem 3.4, quantifies the rate. The homogeneous case described earlier can be thought of as a special case of this vastly more general setting.

THEOREM 3.4. *Let* $r \in \mathbb{R}_+$ *be the solution to the following equation:*

$$(n_0 \wedge n_0^2\varepsilon_0^2 r^d) r^{2\beta+d} + \sum_{j=1}^{m} (n_j \wedge n_j^2\varepsilon_j^2 r^d) r^{2\beta\gamma_j+d} = 1 \tag{2}$$

The minimax rate for excess risk is given by

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \asymp \left(L_N r^{\beta(1+\alpha)} \wedge 1 \right). \tag{3}$$

whenever $\sum_j n_j\delta_j \to 0$, *for a sequence* L_N *of order at most* $(-\log(\delta_{\min}))^{\frac{\beta(1+\alpha)}{2\beta\gamma_{\min}+d}}$.

4. Minimax Optimal Classification Procedure

This section has two goals, divided respectively into the corresponding subsections. In the first subsection we derive a nonparametric classifier for the target population, that suitably utilizes information from the sources while satisfying privacy requirements for each server. In the second subsection we prove a minimax lower bound showing that our classifier is minimax rate optimal in the distributed private transfer learning context.

4.1. Classifier

We now describe a classifier for transfer learning with distributed privacy. Our method has three main steps. First, we use a kernel estimator to estimate $(\eta_j(x) - \frac{1}{2})g(x)$ for $j = 0, 1, \ldots, m$. Second, then use a convex combination of these estimators, where the weights are designed to borrow strength from the source servers under the transfer learning setup. The third step is to add a Gaussian noise to the weighted kernel estimator to satisfy privacy requirements. Our classifier is given by the sign of the noise perturbed weighted estimator. See Section 4 of Auddy et al. [2024] for details.

PROPOSITION 4.1. *For any $h \in [0, 1]$ the transcripts $\{T_h^{(j)}(x_0) + \widetilde{\xi}_h^{(j)}(x_0) : 0 \leq j \leq m\}$ described above satisfies $(\varepsilon_j, \delta_j)$ differential privacy distributed across servers $j \in \{0, 1 \ldots, m\}$.*

The optimal bandwidth choice h_{opt} is given by the solution to (2). To account for the additional δ factor for approximate privacy, we now define $h_{opt,\delta}$ which is the solution to:

$$(n_0 \wedge n_0^2\varepsilon_0^2 r^d)r^{2\beta+d} + \sum_{j=1}^{m}(n_j \wedge n_j^2\varepsilon_j^2 r^d)r^{2\beta\gamma_j+d} = \log\left(\frac{2}{\delta_{\min}}\right) \tag{4}$$

Finally our classifier is given by

$$\widehat{f}(x_0) := \mathbb{1}(\widetilde{T}_{h_{opt,\delta}}(x_0) \geq 0) \tag{5}$$

where $h_{opt,\delta}$ is the solution to (4). The following theorem provides an upper bound for the excess risk of this classifier.

THEOREM 4.2. *Let r be the solution to (2). Let $\widehat{f}$ be the classifier defined in (5) based on the weighted kernel estimator in Section 4 of Auddy et al. [2024]. Then,*

$$\sup_{(P_0,\ldots,P_m)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \leq C_* r^{\beta(1+\alpha)} \left(\log\left(\frac{1}{\delta_{\min}}\right)\right)^{\frac{\beta(1+\alpha)}{2\beta\gamma_{\min}+d}}$$

where C_* is a constant depending on $L, d, \alpha, \beta, \gamma_j$, while $\gamma_{\min} = \min\{1, \gamma_1, \ldots, \gamma_m\}$ and $\delta_{\min} = \min\{\delta_0, \ldots, \delta_m\}$.

4.2. Minimax Lower Bounds

The above theorem bounds the error rate of our kernel based classifier. Alongside the upper bound above, in this subsection we derive the minimax lower bound on the excess

risk, to establish that our kernel based classifier is minimax optimal upto logarithmic terms.

We introduce a general data processing inequality which extends the findings presented in Cai et al. [2023a]. This new result provides a bound on the total variation (TV) distance between the push forward measures of the transcripts $\mathbb{P}_{\sigma}^{\widetilde{T}}$ and $\mathbb{P}_{\sigma'}^{\widetilde{T}}$, utilizing the TV distance between their underlying distributions. Such an inequality may be of independent interest beyond the current setting due to its broader applicability.

LEMMA 1. *For any subset $S \subseteq \{0, \dots, m\}$, the TV distance is bounded as follows:*

$$\mathrm{TV}\left(\mathbb{P}_{\sigma}^{T^{(0)}}, \mathbb{P}_{\sigma'}^{T^{(0)}}\right) \leq \sqrt{2}\sqrt{\sum_{j\in S} \bar{\varepsilon}_j \left(e^{\bar{\varepsilon}_j} - 1\right) + \sum_{j\in S^c} n_j \mathrm{KL}(P_{j,\sigma}, P_{j,\sigma'})} + 4\sum_{j\in S} e^{\bar{\varepsilon}_j} n_j \delta_j \rho_j, \quad (6)$$

where $\bar{\varepsilon}_j = 6n_j\varepsilon_j\rho_j$ and $\rho_j = \mathrm{TV}(P_{j,\sigma}, P_{j,\sigma'})$.

The following theorem then establishes the fundamental cost of privacy for the nonparametric classification problem in the distributed privacy setting.

THEOREM 4.3. *Suppose δ_j's are such that $\sum_j n_j\delta_j = o(1)$, then there exists a $c > 0$ not depending on n_j for $j = 0, \dots, m$ such that*

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\dots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \geq cr^{\beta(1+\alpha)}$$

where r is the solution to (2).

5. Data-driven Adaptive Classifier

In practice we do not know the smoothness and transfer exponent parameters of the unknown regression function. Choosing the correct bandwidth h thus becomes nontrivial. We will use an estimator based on the Lepski method to choose h from a grid of possible values. To choose the best candidate bandwidth, we define a grid of possible choices for h as:

$$\mathcal{H} = \{2^{-j} : j = 0, 1, \dots, (\log n_*)/d\}, \text{ where } n_* = \sum_{j=0}^{m} n_j \wedge n_j^2\varepsilon_j^2.$$

Let $\Delta^m = \{w : w_i \in [0,1], \sum_{i=0}^m w_i = 1\}$ denote the m-dimensional simplex. For a weight vector $w = (w_0, w_1, \dots, w_m) \in \Delta^m$ we define

$$\widetilde{T}(x_0, h, w) := \sum_{j=0}^{m} w_j T_h^{(j)}(x_0) + \sum_{j=0}^{m} w_j \frac{\sqrt{2c_K \log(2|\mathcal{H}|/\delta_j)|\mathcal{H}|}}{n_j\varepsilon_j h^d} \xi^{(j)}(x_0) \text{ for } h, w \in [0,1] \quad (7)$$

where $T_h^{(j)}(\cdot)$ is as defined in Section 4 of Auddy et al. [2024] and $\xi^{(j)}(\cdot)$ are mean zero Gaussian processes with covariance kernel $K(\cdot/h)$.

5.1. Adaptation under Source Homogeneity

In this subsection we consider the sources to have transfer homogeneity, i.e., every source has identical transfer exponent $\gamma_j = \gamma$ for $j = 1, \ldots, m$. It is then intuitive to weigh the estimators $T_h^{(j)}$ and the noise $\xi^{(j)}$ with the weights proportional to $n_j \wedge n_j^2 \varepsilon_j^2 h^d$ for the sources. Then the adaptive choice of h in the transfer homogeneous setting is given by

$$h_0 = \begin{cases} \min\{h \in \mathcal{H} : \widehat{\rho}_0(h) > 4.5 \log(2n_*|\mathcal{H}|)\} & \text{if } \max_{h \in \mathcal{H}} \widehat{\rho}_0(h) > 4.5 \log(2n_*|\mathcal{H}|) \\ \text{argmax}_h \widehat{\rho}_0(h) & \text{otherwise.} \end{cases}$$

Define

$$w_0^* = \text{argmax}_{w \in \mathcal{W}(h_0)} \frac{(\widetilde{T}(x_0, h_0, w))^2}{v_0(h_0, w)}.$$

The adaptive classifier is now defined as

$$\widehat{f}_0(x) := \mathbb{1}(\widetilde{T}(x, h_0, w_0^*) > 0). \tag{8}$$

The following theorem states the excess risk of the adaptive classifier in terms of the regression function parameter α, β, the transfer exponent γ and the privacy constraints.

THEOREM 5.1. *Let r be the solution to (2) with $\gamma_j = \gamma$ for $j = 1, \ldots, m$. Let $\widehat{f}_0$ be the data adaptive classifier defined in (8). Then,*

$$\sup_{(P_0, \ldots, P_m) \in \Pi(\alpha, \beta, \gamma, \mu)} \mathcal{E}_0(\widehat{f}_0) \leq C'_* r^{\beta(1+\alpha)} \left[(\log(n_*|\mathcal{H}|) \log(2|\mathcal{H}|/\delta_{\min}))^{\frac{\beta(1+\alpha)}{2\beta(1\wedge\gamma)+d}} \vee |\mathcal{H}|^{\frac{2\beta(1+\alpha)}{d}} \right]$$

where C'_* is a constant depending on $m, L, d, \alpha, \beta, \gamma$, while $\delta_{\min} = \min\{\delta_0, \ldots, \delta_m\}$.

5.2. General Adaptation for Multiple Sources

We now shift to the general setting where we no longer constrain $\gamma_1, \ldots, \gamma_m$ to be all equal. Note that for optimal estimation (as in Theorem 4.2), one requires knowledge of potentially m many different parameters $\gamma_1, \ldots, \gamma_m$. The adaptation procedure therefore requires optimizing over all possible weights in $w \in \Delta^m$. When m increases with n, the adaptation to this growing number of parameters necessarily worsens the rate of decay for the excess risk. We will not delve further into issue and focus instead on the case where m is finite and does not increase with n.

Let us define the signal-to-noise ratio index $\widehat{\rho}(h)$:

$$\widehat{\rho}(h) = \max_{w \in \Delta^m} \frac{(\widetilde{T}(x_0, h, w))^2}{v(h, w)},$$

In this general setting, the adaptive choice of h is given by

$$h_* = \begin{cases} \min\{h \in \mathcal{H} : \widehat{r}(h) > C_* \log(n_*|\mathcal{H}|(m+1))\} & \text{if } \max_{h \in \mathcal{H}} \widehat{\rho}(h) > C_* \log(n_*|\mathcal{H}|(m+1)) \\ \text{argmax}_h \ \widehat{\rho}(h) & \text{otherwise.} \end{cases}$$

where $C_* = 2.25(m+1)$. Defining

$$w_* = \text{argmax}_{w \in \Delta^m} \frac{(\widetilde{T}(x_0, h_*, w))^2}{v(h_*, w)}.$$

we obtain the adaptive classifier

$$\widehat{f}_a(x) := \mathbb{1}(\widetilde{T}(x, h_*, w_*) > 0). \tag{9}$$

The following theorem verifies the efficacy of the general adaptive procedure.

THEOREM 5.2. *Let r be the solution to (2). Let $\widehat{f}_a$ be the data adaptive classifier defined in (9). Then,*

$$\sup_{(P_0,\dots,P_m)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \le C'_* r^{\beta(1+\alpha)} \left[(\log(n_*|\mathcal{H}|) \log(2|\mathcal{H}|/\delta_{\min}))^{\frac{\beta(1+\alpha)}{2\beta\gamma_{\min}+d}} \vee |\mathcal{H}|^{\frac{2\beta(1+\alpha)}{d}} \right]$$

where C'_* is a constant depending on $m, L, d, \alpha, \beta, \gamma_j$, while $\gamma_{\min} = \min\{1, \gamma_1, \dots, \gamma_m\}$ and $\delta_{\min} = \min\{\delta_0, \dots, \delta_m\}$.

An important special case of the above theorem is the server homogeneous case, where sample sizes and the privacy parameters are the same for every server, i.e., $n_j = n$, $\varepsilon_j = \varepsilon$ and $\delta_j = \delta$ for $j = 0, 1, \dots, m$. The following corollary bounds the excess risk of the adaptive estimator for this special case.

COROLLARY 5.3. *Let r be the solution to (2) with $n_j = n, \varepsilon_j = \varepsilon, \delta_j = \delta$ and $\gamma_j = \gamma$ for all $j = 1, \dots, m$. Let $\widehat{f}_0$ be the data adaptive classifier defined in (8). Then,*

$$\sup_{(P_0,\dots,P_m)\in\Pi(\alpha,\beta,\gamma,\mu)} \mathcal{E}_0(\widehat{f}_0) \le C'_* \Bigg[L_N^{(ada)} \Bigg\{ \left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2\varepsilon_0^2)^{\frac{1}{2\beta+2d}} \right) + \left((mn)^{\frac{1}{2\beta\gamma+d}} \wedge (mn^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}} \right) \Bigg\}^{-\beta(1+\alpha)} \wedge 1 \Bigg]$$

where $L_N^{(ada)}$ is given by $\left[(\log(n_*|\mathcal{H}|) \log(2|\mathcal{H}|/\delta))^{\frac{\beta(1+\alpha)}{2\beta(1\wedge\gamma)+d}} \vee |\mathcal{H}|^{\frac{2\beta(1+\alpha)}{d}} \right]$, and C'_* is a constant depending on $m, L, d, \alpha, \beta, \gamma$.

6. Proofs

We divide this section into three subsections where the proofs for the main results, the upper bound, the lower bound, and the adaptive estimator are respectively given.

6.1. Proofs of Results in Section 3

We first prove the theorems and corollaries in Section 3.

PROOF (PROOF OF THEOREM 3.1). The proof follows from Theorem 3.4 by precisely deriving the value of r in the homogeneous setting. In particular, when $n_j = n$, $\gamma_j = \gamma$, $\varepsilon_j = \varepsilon$, and $\delta_j = \delta$ for $j = 1, \ldots, m$, the solution r to the equation (2) satisfies

$$\frac{1}{2} \leq (n_0 r^{2\beta+d} \wedge n_0^2 \varepsilon_0^2 r^{2\beta+2d}) \vee (mnr^{2\beta\gamma+d} \wedge mn^2\varepsilon^2 r^{2\beta\gamma+2d}) \leq 1.$$

In conjunction with the result from Theorem 3.4, the different parts of the min-max inequality above finishes the proof of Theorem 3.1.

PROOF (PROOF OF COROLLARY 3.2). To simplify notation we define

$$a_1 := n_0^{\frac{1}{2\beta+d}},\ a_2 := (n_0^2\varepsilon^2)^{\frac{1}{2\beta+2d}},\ a_3 := (mn)^{\frac{1}{2\beta\gamma+d}},\ \text{ and }\ a_4 := (mn^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}}. \tag{10}$$

By Theorem 3.1 we have that the minimax rate of excess risk in this case is given by

$$\inf_{\widehat{f}\in\mathcal{M}(\varepsilon,\delta)} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \asymp [L_N a_5 \wedge 1] \text{ where } a_5 := \{(a_1 \wedge a_2) \vee (a_3 \wedge a_4)\}^{-\beta(1+\alpha)}. \tag{11}$$

Case 1: ($\gamma \in (\mathbf{1}, \gamma^{(*)}]$) We first consider the regime where $1 \leq \gamma \leq \gamma^{(*)}$ for $\gamma^{(*)} := \frac{1}{2\beta}\left[\frac{(2\beta+d)\log mn}{\log n_0} - d\right]$, since this shows the most number of phase transitions. The proofs for other regimes follow similarly. Note first that:

$$\begin{aligned}\gamma \leq \frac{1}{2\beta}\left[\frac{(2\beta+d)\log mn}{\log n_0} - d\right] &\iff (2\beta\gamma + d)\log(n_0) \leq (2\beta+d)\log mn \\ &\iff a_1 \leq a_3. \end{aligned} \tag{12}$$

The rest of the proof proceeds in sub-cases based on the range of ε.
Case a): ($0 < \varepsilon \leq \varepsilon^{(1)}$) Since $\varepsilon < \varepsilon^{(1)} := (\sqrt{mn})^{-1} \wedge n_0^{-1}$, we have $mn^2\varepsilon^2 \vee n_0^2\varepsilon^2 \leq 1$ and hence $a_2 \vee a_4 \leq 1$, and thus $a_5 \geq 1$. Thus $\mathcal{E}_0(\widehat{f}) = 1$ in this case by equation (11).
Case b): ($\varepsilon^{(1)} < \varepsilon \leq \varepsilon^{(2)}$) Since $\varepsilon < \varepsilon^{(2)} := n_0^{-\frac{\beta}{2\beta+d}}$, we have

$$a_2 \geq a_1 \text{ and } a_4 < a_3.$$

Now a direct comparison yields that since $\gamma > 1$

$$\varepsilon \leq \varepsilon^{(11)} := \left[(\sqrt{mn})^{\beta+d} n_0^{-\beta\gamma-d}\right]^{\frac{1}{\beta(\gamma-1)}} \iff a_4 > a_2.$$

This means

$$\{(a_1 \wedge a_2) \vee (a_3 \wedge a_4)\} = a_4$$

when $\varepsilon \leq \varepsilon^{(11)}$. Similarly $\{(a_1 \wedge a_2) \vee (a_3 \wedge a_4)\} = a_2$ when $\varepsilon > \varepsilon^{(11)}$. This finishes the proof for case b).
Case c): ($\varepsilon^{(2)} < \varepsilon \leq \varepsilon^{(3)}$) Comparing the expression we arrive at $a_1 < a_2$ and $a_3 \geq a_4$ in this case. It is left to compare a_1 and a_4. It can be checked that

$$\varepsilon \leq \varepsilon^{(21)} \iff a_1 > a_4.$$

Thus $\{(a_1 \wedge a_2) \vee (a_3 \wedge a_4)\} = a_1$ when $\varepsilon \leq \varepsilon^{(21)}$, and $\{(a_1 \wedge a_2) \vee (a_3 \wedge a_4)\} = a_4$ when $\varepsilon > \varepsilon^{(21)}$. This finishes the proof for case b).

Case d): ($\varepsilon > \varepsilon^{(3)}$) In this case, we have $a_3 < a_4$. Since $(a_1 \wedge a_2) \leq a_1 \leq a_3$ by (12), we find that

$$\{(a_1 \wedge a_2) \vee (a_3 \wedge a_4)\} = a_3.$$

This finishes the proof for case d). The proofs for other ranges of γ follow analogously and are hence skipped.

Proof (Proof of Corollary 3.3). The proof follows from Theorem 3.4 by setting $m = 2$ and $\varepsilon_1 = \infty$.

Proof (Proof of Theorem 3.4). The proof follows by combining the conclusions of Theorem 4.2 and Theorem 4.3.

6.2. *Proofs of Results in Section 4*

Proof (Proof of Proposition 4.1). We first show the RKHS norm bounds, i.e., for $T_h^{(j)}$ and $T_h^{(j)'}$ as defined in Section 4 of Auddy et al. [2024], we have

$$\begin{aligned}
&\|T_h^{(j)} - T_h^{(j)'}\|_{\mathcal{K}} \\
&= \sqrt{\frac{1}{n_{P_j}^2 h^{2d}} \left[\frac{1}{2}K(0) - 2\left(Y_1^{(j)} - \frac{1}{2}\right)\left(Y_1^{(j)'} - \frac{1}{2}\right) K\left(\frac{X_1^{(j)} - X_1^{(j)'}}{h}\right)\right]} \\
&\leq \frac{\sqrt{c_K}}{n_j h^d}
\end{aligned}$$

for $j = 0, 1, \ldots, m$. The proof then follows using Corollary 3.5 of Hall et al. [2013].

Proof (Proof of Theorem 4.2). To get started, we state the following concentration inequality on $T_h^{(j)}$, which holds for any fixed $h \in [0, 1]$.

Lemma 2. *For any $t \in [0, c_K 2^{d-1})$ we have*

$$\mathbb{P}\left(|T_h(x_0) - \mathbb{E}(T_h(x_0))| \geq t\right) \leq \exp\left[-\frac{t^2 h^d}{C_{up}}\left\{\sum_{j=0}^{m} u_j^2 \left(\frac{1}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2 h^d}\right) + \max_{0 \leq j \leq m} \frac{u_j}{n_j}\right\}^{-1}\right].$$

where $C_{up} = (c_K^2 \vee 1)[2^{d-2} g_{\max} + \log(2/\delta_{\min})/4]$.

It follows from the definition of $\{T_h^{(j)} : 0 \leq j \leq m\}$ that for $x_0 \in \mathbb{R}^d$ we have

$$\begin{aligned}
\mathbb{E}T_h(x_0) &= \frac{1}{h^d} \int \sum_{j=0}^{m} \left[g_j(x) u_j \left(\eta_j(x) - \frac{1}{2}\right)\right] K\left(\frac{x - x_0}{h}\right) dx \\
&= \frac{1}{h^d} \int_{x : \|x - x_0\|_\infty \leq h} \sum_{j=0}^{m} \left[g_j(x) u_j \left(\eta_j(x) - \frac{1}{2}\right)\right] K\left(\frac{x - x_0}{h}\right) dx.
\end{aligned}$$

where the second inequality follows due to the support of K. Note that:

$$|\mathbb{E}T_h(x_0)| = \sum_{j=0}^{m} \frac{u_j}{h^d} \int_{x:\|x-x_0\|_\infty \le h} \left[g_j(x) \left| \eta_j(x) - \frac{1}{2} \right| \right] K\left(\frac{x-x_0}{h}\right) dx$$
$$\le \frac{c_K}{2h^d} \int_{x:\|u\|_\infty \le h} g_j(x)dx \le \; c_K 2^{d-1}. \tag{13}$$

Suppose $\eta_Q(x_0) \ge \frac{1}{2} + L(2h\sqrt{d})^\beta$. This implies for any $x \in \mathcal{X}_h(x_0) := \{x : \|x - x_0\|_\infty \le h\}$, we have by the Hölder condition that

$$\eta_Q(x_0) \ge \frac{1}{2} + L(2h\sqrt{d})^\beta \implies \eta_Q(x) \ge \eta_Q(x_0) - \sup_{x \in \mathcal{X}_h(x_0)} L\|x - x_0\|^\beta \ge \frac{1}{2} \text{ for all } x \in \mathcal{X}_h(x_0).$$

By part 1 of the RSE assumption, and similarly arguing for x_0 such that $\eta_Q(x_0) \le \frac{1}{2} - L(2h\sqrt{d})^\beta$, we have

$$\eta_0(x_0) \ge \frac{1}{2} + L(2h\sqrt{d})^\beta \implies \eta_j(x) \ge \frac{1}{2} \text{ for all } x \in \mathcal{X}_h(x_0)$$
$$\eta_0(x_0) \le \frac{1}{2} - L(2h\sqrt{d})^\beta \implies \eta_j(x) \le \frac{1}{2} \text{ for all } x \in \mathcal{X}_h(x_0).$$

On the other hand, it can be checked that $B_{x_0}(h/\sqrt{2}) \subset \mathcal{X}_h(x_0)$. Consequently for $x_0 \in \mathbb{R}^d$, $\eta_0(x_0) \ge \frac{1}{2} + L(2h\sqrt{d})^\beta$ implies,

$$\begin{aligned}
&\mathbb{E}T_h(x_0) \\
&= \frac{1}{h^d} \int_{\mathcal{X}_h(x_0)} \sum_{j=0}^{m} \left[g_j(x) u_j \left(\eta_j(x) - \frac{1}{2} \right) \right] K\left(\frac{x-x_0}{h}\right) dx \\
&\ge \frac{1}{h^d} \int_{\mathcal{X}_h(x_0)} \left[g_0(x) u_0 \left(\eta_0(x) - \frac{1}{2} \right) + \sum_{j=1}^{m} g_j(x) u_j \left(\eta_0(x) - \frac{1}{2} \right)^{\gamma_j} \right] K\left(\frac{x-x_0}{h}\right) dx \\
&\ge \frac{1}{h^d} \left[\sum_{j=0}^{m} u_j C_j \left(\eta_0(x_0) - \frac{1}{2} - L(2h\sqrt{d})^\beta \right)^{\gamma_j} \int_{B_{x_0}(h/2) \cap A} g_j(x) K\left(\frac{x-x_0}{h}\right) dx \right] \\
&\ge \frac{b_K}{h^d} \left[\sum_{j=0}^{m} u_j C_j \left(\eta_0(x_0) - \frac{1}{2} - L(2h\sqrt{d})^\beta \right)^{\gamma_j} \int_{B_{x_0}(h/2) \cap A} g_j(x) dx \right] \\
&\ge \frac{b_K c_0}{2^d} \left[\sum_{j=0}^{m} u_j C_j \left(\eta_0(x_0) - \frac{1}{2} - L(2h\sqrt{d})^\beta \right)^{\gamma_j} \right].
\end{aligned} \tag{14}$$

where we use the notation $C_0 = \gamma_0 = 1$. Here the three inequalities follow from (14), the Hölder condition on η_0, and the c_0 regularity of P_j^X for $j = 0, \ldots, m$ respectively. The rest of our proof follows the general principle laid out in Audibert and Tsybakov [2007]. We define

$$\mathcal{X}_k := \{x \in [0,1]^d : 2^k L(2h\sqrt{d})^\beta < |\eta_0(x) - 1/2| \le 2^{k+1} L(2h\sqrt{d})^\beta\} \text{ for } k = 0, 1, 2, \ldots$$

Note that if $x \in \mathcal{X}_k$,

$$\left|\eta_0(x) - \frac{1}{2} - L(2h\sqrt{d})^\beta\right| \geq 2^k L(2h\sqrt{d})^\beta - L(2h\sqrt{d})^\beta \geq 2^k L(2h\sqrt{d})^\beta/2$$

for $k \geq 1$. Since by (13), we have $|\mathbb{E}T_h(x_0)| < c_K 2^{d-1}$ for all x_0, we have using Lemma 2 that

$$\begin{aligned}
&\mathcal{E}(\widehat{f}) \\
&= \int g(x_0)\left\{\mathbb{P}(T_h(x_0) < 0) - \mathbb{1}\left(\eta_0(x_0) < \frac{1}{2}\right)\right\}(2\eta(x_0) - 1)dx_0 \\
&\leq 4L(2h\sqrt{d})^\beta \mathbb{P}(X \in \mathcal{X}_0) + 8L(2h\sqrt{d})^\beta \mathbb{P}(X \in \mathcal{X}_1) \\
&\quad + \sum_{k=2}^{\infty} \int_{\mathcal{X}_k} g(x_0)(2\eta_0(x_0) - 1) \exp\left\{-\frac{(\mathbb{E}T_h(x_0))^2 h^d}{C_{up}}\left(\sum_{j=0}^{m} \frac{2u_j^2}{n_j \wedge n_j^2\varepsilon_j^2 h^d} + \max_{0\leq j\leq m} \frac{u_j}{n_j}\right)^{-1}\right\} dx_0 \\
&\leq 4L(2h\sqrt{d})^\beta \mathbb{P}(X \in \mathcal{X}_0) + 8L(2h\sqrt{d})^\beta \mathbb{P}(X \in \mathcal{X}_1) \\
&\quad + \sum_{k=2}^{\infty} \int_{\mathcal{X}_k} g(x_0)(2\eta_0(x_0) - 1)dx_0 \times \\
&\quad \times \exp\left\{-\frac{b_K^2 c_0^2 h^d}{2^{2d+5} C_{up}}\left[\sum_{j=0}^{m} u_j C_j \left(2^k L(2\sqrt{d})^\beta\right)^{\gamma_j} h^{\beta\gamma_j}\right]^2 \left(\sum_{j=0}^{m} \frac{u_j^2}{n_j \wedge n_j^2\varepsilon_j^2 h^d} + \max_{0\leq j\leq m} \frac{u_j}{n_j}\right)^{-1}\right\}.
\end{aligned} \tag{15}$$

In the rest of the proof we will use the choice of $h = h_{opt,\delta}$ where $h_{opt,\delta}$ is the solution

to (4). By definitions of u_j we have for $h = h_{opt,\delta}$ that

$$
\begin{aligned}
& h_{opt,\delta}^d \left(\sum_{j=0}^m u_j C_j \left(2^k L (2\sqrt{d})^\beta \right)^{\gamma_j} h_{opt,\delta}^{\beta\gamma_j} \right)^2 \left\{ \sum_{j=0}^m \frac{u_j^2}{n_j \wedge n_j^2 \varepsilon_j^2 h_{opt,\delta}^d} + \max_{0 \le j \le m} \frac{u_j}{n_j} \right\}^{-1} \\
= \; & h_{opt,\delta}^d \left(\sum_{j=0}^m C_j (n_j \wedge n_j^2 \varepsilon_j^2 h_{opt,\delta}^d) h_{opt,\delta}^{\beta\gamma_j} \left(2^k L (2\sqrt{d})^\beta \right)^{\gamma_j} h_{opt,\delta}^{\beta\gamma_j} \right)^2 \\
& \times \left[\left\{ \sum_{j=0}^m (n_j \wedge n_j^2 \varepsilon_j^2 h^d) h_{opt,\delta}^{2\beta\gamma_j} \right\} \left(1 + \max_{0 \le j \le m} \frac{n_j \wedge n_j^2 \varepsilon_j^2 h^d}{n_j} \right) \right]^{-1} \\
\ge \; & \min\{1, C_1, \ldots, C_m\}^2 (2^{2k+3} L^2 d)^{\gamma_{\min}} \left(\sum_{j=0}^m (n_j \wedge n_j^2 \varepsilon_j^2 h_{opt,\delta}^d) h_{opt,\delta}^{\beta\gamma_j} h_{opt,\delta}^{\beta\gamma_j} \right)^2 h_{opt,\delta}^d \\
& \times \left\{ \sum_{j=0}^m (n_j \wedge n_j^2 \varepsilon_j^2 h_{opt,\delta}^d) h_{opt,\delta}^{2\beta\gamma_j} \right\}^{-1} \\
\ge \; & \min\{1, C_1, \ldots, C_m\}^2 (2^{2k+3} L^2 d)^{\gamma_{\min}} \left(\sum_{j=0}^m (n_j \wedge n_j^2 \varepsilon_j^2 h_{opt,\delta}^d) h_{opt,\delta}^{2\beta\gamma_j + d} \right) \\
\ge \; & \min\{1, C_1, \ldots, C_m\}^2 (2^{2k+3} L^2 d)^{\gamma_{\min}} \log \left(\frac{2}{\delta_{\min}} \right)
\end{aligned}
$$

where the last two inequalities follow since $h_{opt,\delta}$ solves (4). Plugging this lower bound into (15) we have for $h = h_{opt,\delta}$ that

$$
\begin{aligned}
\mathcal{E}(\widehat{f}) & \le 8L(2h_{opt,\delta}\sqrt{d})^\beta \mathbb{P}(|\eta_Q(X) - 1/2| \le 4L(2h_{opt,\delta}\sqrt{d})^\beta) \\
& \quad + \sum_{k=2}^\infty (2^{k+2} L (2h_{opt,\delta}\sqrt{d})^\beta) \mathbb{P}(|\eta_Q(X) - 1/2| \le 2^{k+1} L (2h_{opt,\delta}\sqrt{d})^\beta) \times \\
& \quad \times \exp\left\{ - \frac{b_K^2 c_0^2 \log(2/\delta_{\min}) \min\{1, C_1, \ldots, C_m\}^2}{2^{2d+5} C_{up}} (2^{2k+3} L^2 d)^{\gamma_{\min}} \right\} \\
& \le CL(2h_{opt,\delta}\sqrt{d})^\beta (2L(2h_{opt,\delta}\sqrt{d})^\beta)^\alpha \\
& \le CL^{1+\alpha} \sqrt{d}^{\beta(1+\alpha)} h_{opt,\delta}^{\beta(1+\alpha)}.
\end{aligned}
$$

for a numerical constant $C > 0$. The second inequality uses the margin assumption. To finish the proof we relate h_{opt} and $h_{opt,\delta}$ as follows.

Let us define a function $\lambda_0 : [0,1] \to \mathbb{R}_+$ as $\lambda_0(r) = \sum_{j=0}^m (n_j \wedge n_j^2 \varepsilon_j^2 r^d) r^{2\beta\gamma_j + d}$. Then,

for $\kappa_0 = (\log(2/\delta_{\min}))^{1/(2\beta\gamma_{\min}+d)}$ we have

$$\begin{aligned}\lambda_0(\kappa_0 h_{opt}) &= \sum_{j=0}^{m}(n_j \wedge n_j^2\varepsilon_j^2(h_{opt}\kappa)^d)(h_{opt}\kappa_0)^{2\beta\gamma_j+d} \\ &\geq \sum_{j=0}^{m}(n_j \wedge n_j^2\varepsilon_j^2 h_{opt}^d)(h_{opt}\kappa_0)^{2\beta\gamma_j+d} \\ &\geq \kappa_0^{2\beta\gamma_{\min}+d}\sum_{j=0}^{m}(n_j \wedge n_j^2\varepsilon_j^2 h_{opt}^d)h_{opt}^{2\beta\gamma_j+d} = \kappa_0^{2\beta\gamma_{\min}+d} = \log\left(\frac{2}{\delta_{\min}}\right),\end{aligned}$$

where the first equality follows since h_{opt} is the solution to (2). Note that the function $\lambda_0(\cdot)$ is increasing. By the definition of $h_{opt,\delta}$ from (4) this implies the relation $h_{opt,\delta} \leq \kappa_0 h_{opt} \leq h_{opt}(\log(2/\delta_{\min}))^{1/(2\beta\gamma_{\min}+d)}$. This finishes the proof.

Proof (Proof of Theorem 4.3). The main tool for accomplishing the lower bound is the Assouad's Lemma. We construct $m+1$ family of distributions $P_{j,\sigma}$ for $j = 1, \ldots, m$ and Q_σ, $\sigma \in \{-1, 1\}^M$ and applying Assouad's lemma on the family $\mathbb{P}_\sigma^{\widetilde{T}} = P_\sigma^{\widetilde{T}^{(1)}} \times \ldots \times P_\sigma^{\widetilde{T}^{(m)}} \times Q_\sigma^{\widetilde{T}^{(0)}}$, where $P_\sigma^{\widetilde{T}^{(j)}}$ denotes the distribution of the $T^{(j)}$ ($T^{(j)}$ is the DP transcript obtained from n_{P_j} samples coming from $P_{j,\sigma}$) and $Q_\sigma^{T^{(0)}}$ is defined similarly.

We borrow the construction of $P_{j,\sigma}$ and Q_σ from Cai and Wei [2021]. To apply Assouad's lemma we look at $\sigma, \sigma' \in \{-1, 1\}^M$ that differ at only one element i.e. $\sigma_k \neq \sigma_k'$ for some k and $\sigma_i = \sigma_i'$ for all $i \neq k$. We bound the total variation distance between P_σ and $P_{\sigma'}$ and also between Q_σ and $Q_{\sigma'}$.

$$\begin{aligned}TV(Q_\sigma, Q_{\sigma'}) &= \frac{1}{2}\int \mu(x) 2\,|\eta_{0,\sigma}(x) - \eta_{0,\sigma'}(x)|\,dx \\ &\leq \frac{1}{2}\int_{B(x_k,r)} \frac{w}{\lambda(B(x_k,r))}.2\left(\frac{1}{2}(1 + C_\beta r^\beta) - \frac{1}{2}(1 - C_\beta r^\beta)\right)dx \\ &= C_\beta w r^\beta\end{aligned}$$

Similarly we can show $TV(P_{j,\sigma}, P_{j,\sigma'}) \leq C_{\gamma_j} C_\beta^{\gamma_j} w r^{\beta\gamma}$ Next we would bound the KL divergence between $Q_\sigma, Q_{\sigma'}$.

$$\begin{aligned}KL(Q_\sigma, Q_{\sigma'}) &= \frac{1}{2}\int \mu(x)\left[\eta_{Q,\sigma}(x)\log\left(\frac{\eta_{Q,\sigma}(x)}{\eta_{Q,\sigma'}(x)}\right) + (1 - \eta_{Q,\sigma}(x))\log\left(\frac{1 - \eta_{Q,\sigma}(x)}{1 - \eta_{Q,\sigma'}(x)}\right)\right]dx \\ &\leq \frac{1}{2}\int_{B(x_k,r)} \frac{w}{\lambda(B(x_k,r))} C_\beta r^\beta \log\left(\frac{1 + C_\beta r^\beta}{1 - C_\beta r^\beta}\right)dx \\ &\leq c C_\beta^2 w r^{2\beta}\end{aligned}$$

The last line follows if r is chosen such that $C_\beta r^\beta \leq 1$. Similarly, we can show that $KL(P_{j,\sigma}, P_{j,\sigma'}) \leq c C_\gamma^2 C_\beta^{2\gamma} w r^{2\beta\gamma}$. For the sake of brevity we denote $P_{0,\sigma} := Q_\sigma$.

We now use Lemma 1. A similar lemma for the federated non-interactive setting can be found in Cai et al. [2023b].

Let us define $\mathcal{S} := \{j : \varepsilon_j \leq (r^d n_j)^{-1/2}\}$. Choose $w = r^d$ and $M = \left(\frac{1}{r}\right)^{d-\alpha\beta}$, first we show that for $j \in S$, $\bar{\varepsilon}_j \leq c_{\beta,\gamma_j} := C_{\gamma_j} C_\beta^{\gamma_j}$.

$$\bar{\varepsilon}_j \leq n_j \varepsilon_j \rho_j \leq n_j \varepsilon_j C_{\gamma_j} C_\beta^{\gamma_j} r^{\beta\gamma_j}$$

Now we upper bound $n_j \varepsilon_j r^{\beta\gamma_j}$.

$$\begin{aligned} n_j \varepsilon_j r^{\beta\gamma_j + d} &= \frac{n_j \varepsilon_j r^{\beta\gamma_j + d}}{\left(\sum_{j=0}^m (n_j \wedge n_j^2 \varepsilon_j^2 r^d) r^{2\beta\gamma_j + d}\right)^{1/2}} \\ &\leq \frac{n_j \varepsilon_j r^{\beta\gamma_j + d}}{\left((n_j \wedge n_j^2 \varepsilon_j^2 r^d) r^{2\beta\gamma_j + d}\right)^{1/2}} \\ &= \frac{n_j \varepsilon_j r^{\beta\gamma_j + d}}{\left(n_j^2 \varepsilon_j^2 r^d r^{2\beta\gamma_j + d}\right)^{1/2}} = 1, \end{aligned}$$

where we have used the fact that $j \in \mathcal{S}$. Hence we have that $\bar{\varepsilon}_j \leq c_{\beta,\gamma_j}$ which implies that $\bar{\varepsilon}_j (e^{\bar{\varepsilon}_j} - 1) \leq c'_{\beta,\gamma_j} \bar{\varepsilon}_j^2$. Next let us upper bound

$$\begin{aligned} \sum_{j \in \mathcal{S}} \bar{\varepsilon}_j \left(e^{\bar{\varepsilon}_j} - 1\right) + \sum_{j \in S^c} n_j \mathrm{KL}(P_{j,\sigma}, P_{j,\sigma'}) &\leq c'_{\beta,\gamma_j} \sum_{j \in S} \bar{\varepsilon}_j^2 + \sum_{j \in S^c} n_j c C_\gamma^2 C_\beta^{2\gamma} r^{2\beta\gamma_j + d} \\ &\leq c''_{\beta,\gamma} \left(\sum_{j \in S} n_j^2 \varepsilon_j^2 r^{2\beta\gamma_j} + \sum_{j \in S^c} n_j r^{2\beta\gamma_j + d} \right) \\ &= c''_{\beta,\gamma} \left(\sum_{j \in S} n_j^2 \varepsilon_j^2 r^{2\beta\gamma_j} \wedge n_j r^{2\beta\gamma_j + d} \right) = c''_{\beta,\gamma} \end{aligned}$$

where we have used our choice of r as in (2). We can choose the constants appropriately such that $c''_{\beta,\gamma}$ is arbitrarily small. Next lets look at the last term in (6) and observe that

$$\sum_{j \in S} e^{\bar{\varepsilon}_j} n_j \delta_j \rho_j \leq e^{c_{\beta,\gamma_j}} \sum_j n_j \delta_j = o(1)$$

where we have used the fact that $\bar{\varepsilon}_j \leq c_{\beta,\gamma_j}$ and $\rho_j \leq 1$. Hence we have that $\mathrm{TV}\left(\mathbb{P}_\sigma^T, \mathbb{P}_{\sigma'}^T\right) \leq 1/2$ for appropriate choice of the constants. Next in order to apply Assouad's we need to

lower bound the risk for neighboring distributions.

$$\begin{aligned}
R_{Q_\sigma}(\widehat{f}) + R_{Q_{\sigma'}}(\widehat{f}) =& 2\mathbb{E}_{X\sim Q_{X,\sigma}}\left(\left|\eta_Q(X) - \frac{1}{2}\right| \mathbb{1}_{\left\{\widehat{f}(X)=f^*_{Q_\sigma}(X)\right\}}\right) \\
&+ 2\mathbb{E}_{X\sim Q_{X,\sigma'}}\left(\left|\eta_Q(X) - \frac{1}{2}\right| \mathbb{1}_{\left\{\widehat{f}(X)=f^*_{Q^*_{\sigma'}}(X)\right\}}\right) \\
=& 2\sum_{i=1}^{M}\int_{B(x_i,r)} \mu(x)\cdot\frac{1}{2}C_\beta r^\beta \cdot \left(\mathbb{1}_{\left\{\widehat{f}(x)=f^*_{Q_\sigma}(x)\right\}} + \mathbb{1}_{\left\{\widehat{f}(x)=f^*_{Q_{\sigma'}}(x)\right\}}\right) dx \\
\geq& \int_{B(x_k,r)} \mu(x)\cdot C_\beta r^\beta \cdot \left(\mathbb{1}_{\left\{\widehat{f}(x)=f^*_{Q_\sigma}(x)\right\}} + \mathbb{1}_{\left\{\widehat{f}(x)=f^*_{Q^*_{\sigma'}}(x)\right\}}\right) dx \\
=& C_\beta w r^\beta.
\end{aligned}$$

Hence by Assouad's Lemma we obtain that for all distributed DP estimators $\widehat{f}$,

$$\inf_{\widehat{f}}\sup_{P_1,\ldots,Q} \mathcal{E}_Q(\widehat{f}) \geq \frac{M}{2}C_\beta w r^\beta\left(1-\frac{1}{2}\right) = cr^{\beta(1+\alpha)}$$

6.3. *Proofs of Results in Section 5*

Proof (Proof of Theorem 5.1). We break the proof into several steps following Cai and Wei [2021]. The first step is to derive a concentration bound for $\widetilde{T}(x,h,w)$ around its expectation, uniformly over $x\in[0,1]^d$, $h\in\mathcal{H}$, and $w\in\mathcal{W}(h)$.

Lemma 3.

$$\mathbb{P}\left(\sup_{x\in[0,1]^d, h\in\mathcal{H}, w\in\mathcal{W}(h)}\left|\frac{\widetilde{T}(x,h,w)-\mathbb{E}\widetilde{T}(x,h,w)}{\sqrt{v_0(h,w)}}\right| \geq \frac{3}{\sqrt{2}}\sqrt{\log(2n_*|\mathcal{H}|)}\right) \leq \frac{2}{n_*}.$$

To allow the probability guarantees to hold uniformly over the adaptation procedure, we define a new optimal bandwidth similar to (2) and (4). Let $h_{ada,0}$ be the solution to

$$(n_0\wedge(n_0^2(\varepsilon_0/|\mathcal{H}|)^2r^d))r^{2\beta+d} + \sum_{j=1}^{m}(n_j\wedge(n_j^2(\varepsilon_j/|\mathcal{H}|)^2r^d))r^{2\beta\gamma+d} = 2\log(2n_*|\mathcal{H}|)\log\left(\frac{2|\mathcal{H}|}{\delta_{\min}}\right). \tag{16}$$

Next, let $E_{A,0}$ be the high probability event described in Lemma 7, we know that $\mathbb{P}(E_{A,0})\geq 1-2n_*^{-1}$. Next we define

$$\psi_0 := 2C_{\psi_0}h_{ada,0}^\beta \tag{17}$$

for $C_{\psi_0} := 2L(2\sqrt{d})^\beta \vee [2^{d+3}/(c_0b_K)]^{1/\gamma_{\min}}$. We also define

$$G_{\psi_0} := \{x : |\eta_Q(x) - 1/2| \geq \psi_0\}.$$

Lemma 4. *If $h\leq h_{ada,0}$ and $x\in G_{\psi_0}$, we have that*

i) when $f^*(x) = 1$,
$\mathbb{E}T_h^{(0)}(x) \geq c_0 b_K(\psi_0/2)(1/2)^d$, *and* $\mathbb{E}T_h^{(j)}(x) \geq c_0 b_K(\psi_0/2)^\gamma(1/2)^d$, *for* $j = 1, \ldots, m$

ii) when $f^*(x) = 0$,
$\mathbb{E}T_h^{(0)}(x) \leq -c_0 b_K(\psi_0/2)(1/2)^d$, *and* $\mathbb{E}T_h^{(j)}(x) \leq -c_0 b_K(\psi_0/2)^\gamma(1/2)^d$, *for* $j = 1, \ldots, m$.

LEMMA 5. *Under* $E_{A,0}$, *if* $x \in G_{\psi_0}$ *and* $h_0 \leq h_{ada,0}$, *then the output of the algorithm is correct i.e* $\widehat{f}_0(x) = f^*(x)$.

LEMMA 6. *For all* $x \in G_{\psi_0}$, $\widehat{f}_0(x) = f^*(x)$ *with probability at least* $1 - \psi_0^{1+\alpha}$.

Now to complete the proof of Theorem 5.1, let us denote by $E_{ada,0}$ the event that $\widehat{f}_0(x) = f^*(x)$ for all $x \in G_{\psi_0}$. We then have

$$\begin{aligned}
\mathbb{E}\mathcal{E}_Q(\widehat{f}_0) &= \mathbb{E}_X\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}(\widehat{f}_0(X) \neq f^*(X))\right] \\
&\leq \mathbb{E}_X\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}(\widehat{f}_0(X) \neq f^*(X))\mathbb{1}(E_{ada,0})\right] + \mathbb{P}(E_{ada,0}^c) \\
&\leq \mathbb{E}_X\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}(X \notin G_{\psi_0})\mathbb{1}(E_{ada,0})\right] + \mathbb{P}(E_{ada,0}^c) \\
&\leq \psi_0 \mathbb{P}\left(\left|\eta(X) - \frac{1}{2}\right| < \psi_0\right) + \mathbb{P}(E_{ada,0}^c) \\
&\leq (C_0 + 1)\psi_0^{1+\alpha}
\end{aligned}$$

To complete the proof we now relate $h_{ada,0}$ and h_{opt}. Let $\lambda_1 : [0,1] \to \mathbb{R}_+$ be a function defined as

$$\lambda_1(r) = (n_0 \wedge (n_0^2(\varepsilon_0/|\mathcal{H}|)^2 r^d))r^{2\beta+d} + \sum_{j=1}^m (n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2 r^d))r^{2\beta\gamma+d}.$$

Let us then define

$$\kappa_1 = (2\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min}))^{1/(2\beta(1\wedge\gamma)+d)} \vee |\mathcal{H}|^{2/d}.$$

Then

$$\begin{aligned}
&\lambda_1(\kappa_1 h_{opt}) \\
&= (n_0 \wedge (n_0^2(\varepsilon_0/|\mathcal{H}|)^2(\kappa_1 h_{opt})^d))(\kappa_1 h_{opt})^{2\beta+d} + \sum_{j=1}^m (n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2(\kappa_1 h_{opt})^d))(\kappa_1 h_{opt})^{2\beta\gamma+d} \\
&\geq \kappa_1^{2\beta(1\wedge\gamma)+d} \sum_{j=0}^m (n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2|\mathcal{H}|^2 h_{opt}^d))h_{opt}^{2\beta\gamma_j+d} \\
&= \kappa_1^{2\beta(1\wedge\gamma)+d} \geq 2\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})
\end{aligned}$$

where the equality holds since h_{opt} solves (2). Since $\lambda_1(\cdot)$ is increasing, we have $h_{ada,0} \leq \kappa_1 h_{opt}$, and thus

$$\psi \leq 2C_\psi h_{opt}^\beta ((2\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min}))^{\frac{\beta}{2\beta(1\wedge\gamma)+d}} \vee |\mathcal{H}|^{\frac{2\beta}{d}})$$

This finishes the proof.

Proof (Proof of Theorem 5.2). The proof is identical in structure to the proof of Theorem 5.1.

Lemma 7.

$$\mathbb{P}\left(\sup_{x\in[0,1]^d, h\in\mathcal{H}, w\in\Delta^m}\left|\frac{\widetilde{T}(x,h,w)-\mathbb{E}\widetilde{T}(x,h,w)}{\sqrt{v(h,w)}}\right| \geq \frac{3}{2}\sqrt{(m+1)\log(n_*|\mathcal{H}|(m+1))}\right) \leq \frac{2}{n_*}.$$

Similar to (16), let h_{ada} be the solution to

$$\sum_{j=0}^{m}(n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2 r^d))r^{2\beta\gamma_j+d} = (m+1)^2\log(n_*|\mathcal{H}|(m+1))\log\left(\frac{2|\mathcal{H}|}{\delta_{\min}}\right). \tag{18}$$

Let E_A be the high probability event described in Lemma 7, we know that $\mathbb{P}(E_A) \geq 1-2n_*^{-1}$. Next we define

$$\psi := 2C_\psi h_{ada}^\beta \tag{19}$$

for $C_\psi := 2L(2\sqrt{d})^\beta \vee [2^{d+2}C_{unif}/(c_0 b_K)]^{1/\gamma_{\min}}$. As before we will utilize the set

$$G_\psi := \{x : |\eta_Q(x) - 1/2| \geq \psi\}.$$

Lemma 8. *If $h \leq h_{ada}$ and $x \in G_\psi$, denoting $\gamma_0 = 1$, we have that*

i) when $f^*(x) = 1$, $\mathbb{E}T_h^{(j)}(x) \geq c_0 b_K(\psi/2)^{\gamma_j}(1/2)^d$, for $j = 0, 1, \ldots, m$

ii) when $f^*(x) = 0$, $\mathbb{E}T_h^{(j)}(x) \leq -c_0 b_K(\psi/2)^{\gamma_j}(1/2)^d$ for $j = 0, 1, \ldots, m$.

Lemma 9. *Under E_A, if $x \in G_\psi$ and $h_* \leq h_{ada}$, then the output of the transfer homogeneous adaptation algorithm is correct i.e $\widehat{f}_0(x) = f^*(x)$.*

Lemma 10. *For all $x \in G_\psi$, $\widehat{f}_a(x) = f^*(x)$ with probability at least $1-\psi^{1+\alpha}$.*

Now to complete the proof of Theorem 5.2, let us denote by E_{ada} the event that $\widehat{f}_a(x) = f^*(x)$ for all $x \in G_\psi$. We then have

$$\begin{aligned}
\mathbb{E}\mathcal{E}_Q(\widehat{f}_a) &= \mathbb{E}_X\left[\left|\eta(X)-\frac{1}{2}\right| \mathbb{1}(\widehat{f}_a(X) \neq f^*(X))\right] \\
&\leq \mathbb{E}_X\left[\left|\eta(X)-\frac{1}{2}\right| \mathbb{1}(\widehat{f}_a(X) \neq f^*(X))\mathbb{1}(E_{ada})\right] + \mathbb{P}(E_{ada}^c) \\
&\leq \mathbb{E}_X\left[\left|\eta(X)-\frac{1}{2}\right| \mathbb{1}(X \notin G_\psi)\mathbb{1}(E_{ada})\right] + \mathbb{P}(E_{ada}^c) \\
&\leq \psi\mathbb{P}\left(\left|\eta(X)-\frac{1}{2}\right| < \psi\right) + \mathbb{P}(E_{ada}^c) \\
&\leq (C_0+1)\psi^{1+\alpha}
\end{aligned}$$

To complete the proof we now relate h_{ada} and h_{opt}. Let $\lambda_2 : [0,1] \to \mathbb{R}_+$ be a function defined as $\lambda_2(r) = \sum_{j=0}^m (n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2 r^d))r^{2\beta\gamma_j+d}$. Define

$$\kappa_2 = ((m+1)^2\log(n_*|\mathcal{H}|(m+1))\log(2|\mathcal{H}|/\delta_{\min}))^{1/(2\beta\gamma_{\min}+d)} \vee |\mathcal{H}|^{2/d}.$$

Then

$$\begin{aligned}\lambda_2(\kappa_2 h_{opt}) &= \sum_{j=0}^{m}(n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2(\kappa_1 h_{opt})^d))(\kappa_1 h_{opt})^{2\beta\gamma_j+d} \\ &\geq \kappa_2^{2\beta\gamma_{\min}+d}\sum_{j=0}^{m}(n_j \wedge (n_j^2(\varepsilon_j/|\mathcal{H}|)^2|\mathcal{H}|^2 h_{opt}^d))h_{opt}^{2\beta\gamma_j+d} \\ &= \kappa_2^{2\beta\gamma_{\min}+d} \geq (m+1)^2\log(n_*|\mathcal{H}|(m+1))\log\left(2|\mathcal{H}|/\delta_{\min}\right)\end{aligned}$$

where the equality holds since h_{opt} solves (2). Since $\lambda_2(\cdot)$ is increasing, we have $h_{ada} \leq \kappa_2 h_{opt}$, and thus

$$\psi \leq 2C_\psi h_{opt}^\beta(((m+1)^2\log(n_*|\mathcal{H}|(m+1))\log\left(2|\mathcal{H}|/\delta_{\min}\right))^{\frac{\beta}{2\beta\gamma_{\min}+d}} \vee |\mathcal{H}|^{\frac{2\beta}{d}})$$

This finishes the proof.

Proof (of Corollary 5.3). The proof follows from Theorem 5.1 by plugging in the analytical form of the solution r to (2), which can be found from the proof of Theorem 3.1.

7. Proofs of Lemmas

Proof (Proof of Lemma 1). Let us denote the conditional distribution $\widetilde{T} \mid \widetilde{T}^{(1)} = t_1, \widetilde{T}^{(2)} = t_2, \ldots, \widetilde{T}^{(m)} = t_m$ by $\mathbb{P}_\sigma^{\widetilde{T}(\boldsymbol{t})}$ where $\boldsymbol{t} = (t_1, \ldots, t_m)$ and the data is generated from $P_{0,\sigma}$. The notation $\widetilde{T}(\boldsymbol{t})$ denotes the random variable $\widetilde{T}$ when we fix the values of $(T^{(1)}, \ldots, T^{(m)}) = \boldsymbol{t}$ making the dependence of $\widetilde{T}$ on $(T^{(1)}, \ldots, T^{(m)})$ explicit.

Now we can obtain as a corollary of Karwa and Vadhan [2017] (by conditioning on $\{\widetilde{T}^{(j)}\}_{j=1}^m$ throughout) that

$$D_\infty^{\delta_0'}(\mathbb{P}_\sigma^{\widetilde{T}(\boldsymbol{t})}, \mathbb{P}_{\sigma'}^{\widetilde{T}(\boldsymbol{t})}) \leq e^{\varepsilon_0'} \quad \text{and} \quad D_\infty^{\delta_0'}(\mathbb{P}_{\sigma'}^{\widetilde{T}(\boldsymbol{t})}, \mathbb{P}_\sigma^{\widetilde{T}(\boldsymbol{t})}) \leq e^{\varepsilon_0'} \quad \forall\, \boldsymbol{t} \in \mathcal{T}^m$$

where $\varepsilon_0' = 6n_0\varepsilon_0 TV(P_{0,\sigma}, P_{0,\sigma'})$ and $\delta_0' = e^{\varepsilon_0'} n_0\delta_0 TV(P_{0,\sigma}, P_{0,\sigma'})$. We denote for $j = 1, \ldots, m$, $\widetilde{T}_\sigma^{(j)}$ the random variable $\widetilde{T}^{(j)}$ when the underlying data is generated from $P_{j,\sigma}$. Let us denote the marginal distributions of $\widetilde{T}_\sigma^{(j)}$ by $\mathbb{P}_\sigma^{\widetilde{T}^{(j)}}$. Similarly we can obtain that

$$D_\infty^{\delta_j'}(\mathbb{P}_\sigma^{\widetilde{T}^{(j)}}, \mathbb{P}_{\sigma'}^{\widetilde{T}^{(j)}}) \leq e^{\varepsilon_j'} \quad \text{and} \quad D_\infty^{\delta_j'}(\mathbb{P}_{\sigma'}^{\widetilde{T}^{(j)}}, \mathbb{P}_\sigma^{\widetilde{T}^{(j)}}) \leq e^{\varepsilon_j'} \quad \text{for } j = 1, \ldots, m$$

where $\varepsilon_j' = 6n_j\varepsilon_j TV(P_{j,\sigma}, P_{j,\sigma'})$ and $\delta_j' = e^{\varepsilon_j'} n_j\delta_j TV(P_{j,\sigma}, P_{j,\sigma'})$. We will use the following lemma.

Lemma 11. *Let Y and Z be such that $\mathbb{P}^Y \ll \mathbb{P}^Z$, $D_\infty^\delta(Y\|Z) \leq \varepsilon$ and $D_\infty^\delta(Z\|Y) \leq \varepsilon$. Then, there exists random variables Y', Z' such that*

$$D_{\mathrm{TV}}(Y, Y') \leq \delta, D_{\mathrm{TV}}(Z, Z') \leq \delta \text{ and } D_\infty(Y'\|Z') \leq \varepsilon,\ D_\infty(Z'\|Y') \leq \varepsilon. \tag{20}$$

Using Lemma 11 there exists random variables $\check{T}_\sigma(\boldsymbol{t})$ and $\check{T}_{\sigma'}(\boldsymbol{t})$ with corresponding measures denoted by $\mathbb{P}_\sigma^{\check{T}(\boldsymbol{t})}$ and $\mathbb{P}_{\sigma'}^{\check{T}(\boldsymbol{t})}$ such that

$$D_\infty(\mathbb{P}_\sigma^{\check{T}(\boldsymbol{t})}, \mathbb{P}_{\sigma'}^{\check{T}(\boldsymbol{t})}) \le \varepsilon_0' \quad \text{and} \quad D_\infty(\mathbb{P}_\sigma^{\check{T}(\boldsymbol{t})}, \mathbb{P}_{\sigma'}^{\check{T}(\boldsymbol{t})}) \le \varepsilon_0' \quad \forall\, \boldsymbol{t} \in \mathcal{T}^m$$

with $D_{\mathrm{TV}}(\mathbb{P}_\kappa^{\widetilde{T}(\boldsymbol{t})}, \mathbb{P}_\kappa^{\check{T}(\boldsymbol{t})}) \le 2\delta_0'$ for all $\boldsymbol{t}$ and $\kappa \in \{\sigma, \sigma'\}$. Similarly, for $j = 1, \dots, m$ we have that there exists random variables $\check{T}_\sigma^{(j)}$ and $\check{T}_{\sigma'}^{(j)}$ with corresponding measures denoted by $\mathbb{P}_\sigma^{\check{T}(j)}$ and $\mathbb{P}_{\sigma'}^{\check{T}(j)}$ such that

$$D_\infty(\mathbb{P}_\sigma^{\check{T}^{(j)}}, \mathbb{P}_{\sigma'}^{\check{T}^{(j)}}) \le \varepsilon_j' \quad \text{and} \quad D_\infty(\mathbb{P}_\sigma^{\check{T}^{(j)}}, \mathbb{P}_{\sigma'}^{\check{T}^{(j)}}) \le \varepsilon_j'$$

with $D_{\mathrm{TV}}(\mathbb{P}_\kappa^{\widetilde{T}^{(j)}}, \mathbb{P}_\kappa^{\check{T}^{(j)}}) \le 2\delta_j'$ for all $\kappa \in \{\sigma, \sigma'\}$.

Now fix a set $\mathcal{S} \subseteq [m]$. Next let us define for $j = 1 \dots, m$ $\bar{T}_\sigma^{(j)}$ as $\check{T}_\sigma^{(j)}$ if $j \in S$ and $\widetilde{T}_\sigma^{(j)}$ otherwise. Similarly define $\bar{T}_\sigma(\boldsymbol{t})$ as $\check{T}_\sigma(\boldsymbol{t})$ if $0 \in \mathcal{S}$ and $\widetilde{T}_\sigma(\boldsymbol{t})$ otherwise. Next we define $\widetilde{T}_\sigma'$, $\check{T}_\sigma'$ as the marginal distribution of $\widetilde{T}_\sigma(\boldsymbol{t})$, $\check{T}_\sigma(\boldsymbol{t})$ respectively where $\boldsymbol{t} \overset{d}{=} \bar{T}^{(1:m)} = (\bar{T}_\sigma^{(1)}, \dots, \bar{T}_\sigma^{(m)})$. Also define $\check{T}_\sigma$ as the marginal distribution of $\check{T}_\sigma(\boldsymbol{t})$ where $\boldsymbol{t} \overset{d}{=} T^{(1:m)} = (\widetilde{T}_\sigma^{(1)}, \dots, T_\sigma^{(m)})$. Finally we define $\bar{T}_\sigma$ as $\check{T}_\sigma'$ if $0 \in \mathcal{S}$ and $\widetilde{T}_\sigma'$ otherwise. By the triangle inequality,

$$D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_{\sigma'}^{\widetilde{T}}) \le D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\bar{T}}, \mathbb{P}_{\sigma'}^{\bar{T}}) + D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_\sigma^{\bar{T}}) + D_{\mathrm{TV}}(\mathbb{P}_{\sigma'}^{\bar{T}}, \mathbb{P}_{\sigma'}^{\widetilde{T}}).$$

Next we bound the second term, consider the case that $0 \notin S$

$$\begin{aligned} D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_\sigma^{\bar{T}}) &= D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_\sigma^{\widetilde{T}'}) \\ &\le D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}^{(1:m)}}, \mathbb{P}_\sigma^{\bar{T}^{(1:m)}}) \\ &\le \sum_{j \in S} D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\check{T}^{(j)}}, \mathbb{P}_\sigma^{\widetilde{T}^{(j)}}) \le 2 \sum_{j \in S} \delta_j'. \end{aligned}$$

where the second inequality follows from the data processing inequality. Now we consider the case when $0 \in \mathcal{S}$, and in that case

$$\begin{aligned} D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_\sigma^{\bar{T}}) &= D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_\sigma^{\check{T}'}) \\ &\le D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}}, \mathbb{P}_\sigma^{\check{T}}) + D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\check{T}}, \mathbb{P}_\sigma^{\check{T}'}) \\ &\le \mathbb{E}_{\boldsymbol{t} \sim T^{(1:m)}} D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\widetilde{T}(\boldsymbol{t})}, \mathbb{P}_\sigma^{\check{T}(\boldsymbol{t})}) + D_{\mathrm{TV}}(\mathbb{P}_\sigma^{T^{(1:m)}}, \mathbb{P}_\sigma^{\bar{T}^{(1:m)}}) \\ &\le \sum_{j \in S} D_{\mathrm{TV}}(\mathbb{P}_\sigma^{\check{T}^{(j)}}, \mathbb{P}_\sigma^{\widetilde{T}^{(j)}}) \le 2 \sum_{j \in S} \delta_j'. \end{aligned}$$

where the second inequality follows from triangle inequality. The third inequality is a consequence of the convexity of the TV distance and data processing inequality. Similarly we have that $D_{\mathrm{TV}}(\mathbb{P}_{\sigma'}^{\bar{T}}, \mathbb{P}_{\sigma'}^{\widetilde{T}}) \le 2 \sum_{j \in S} \delta_j'$.

We now state another lemma.

LEMMA 12. *If* $D_\infty(Y, Z) \le \varepsilon$ *then* $D_{\mathrm{KL}}(Y, Z) \le \varepsilon(e^\varepsilon - 1)$.

Now we can use Lemma 12 to conclude that $D_{KL}(\mathbb{P}_{\sigma}^{\check{T}(\boldsymbol{t})}, \mathbb{P}_{\sigma'}^{\check{T}(\boldsymbol{t})}) \leq \varepsilon_0'(e^{\varepsilon_0'}-1)$. Similarly, we can show that $D_{KL}(\mathbb{P}_{\sigma}^{\check{T}(j)}, \mathbb{P}_{\sigma'}^{\check{T}(j)}) \leq \varepsilon_j'(e^{\varepsilon_j'}-1)$. We now tend to the term $D_{\mathrm{TV}}(\mathbb{P}_{\sigma}^{\bar{T}}, \mathbb{P}_{\sigma'}^{\bar{T}})$. By Pinsker's inequality, the independence of the transcripts given the data generating process and the chain rule for the KL-divergence (we only need to condition for the first server as that is the only place interaction is happening) using Theorem 5.3.1. from Gray [2011] we have that

$$
\begin{aligned}
&D_{\mathrm{TV}}(\mathbb{P}_{\sigma}^{\bar{T}}, \mathbb{P}_{\sigma'}^{\bar{T}}) \\
&\leq \sqrt{2D_{\mathrm{KL}}(\mathbb{P}_{\sigma}^{\bar{T}}, \mathbb{P}_{\sigma'}^{\bar{T}})} \\
&\leq \sqrt{2D_{\mathrm{KL}}(\mathbb{P}_{\sigma}^{\bar{T},\bar{T}^{(1:m)}}, \mathbb{P}_{\sigma'}^{\bar{T},\bar{T}^{(1:m)}})} \\
&= \sqrt{2\int D_{\mathrm{KL}}(\mathbb{P}_{\sigma}^{\bar{T}(\boldsymbol{t})}, \mathbb{P}_{\sigma'}^{\bar{T}(\boldsymbol{t})}) d\mathbb{P}_{\sigma}^{\widetilde{T}^{(1)}} \times \ldots \times d\mathbb{P}_{\sigma}^{\widetilde{T}^{(m)}}(\boldsymbol{t}) + 2\sum_{j=1}^{m} D_{\mathrm{KL}}(\mathbb{P}_{\sigma}^{\bar{T}^{(j)}}, \mathbb{P}_{\sigma'}^{\bar{T}^{(j)}})}
\end{aligned}
$$

If $0 \in \mathcal{S}$ then the first term is bounded by $\varepsilon_0'(e^{\varepsilon_0'}-1)$ because each term is uniformly bounded by this very term. And if $0 \notin \mathcal{S}$ we use data processing inequality to conclude that $D_{\mathrm{KL}}(\mathbb{P}_{\sigma}^{\bar{T}(\boldsymbol{t})}, \mathbb{P}_{\sigma'}^{\bar{T}(\boldsymbol{t})}) \leq D_{KL}(\mathbb{P}_{\sigma}, \mathbb{P}_{\sigma'}) = n_0 D_{KL}(P_{0,\sigma}, P_{0,\sigma'})$. We can similarly show that if $j \in S$ $D_{\mathrm{KL}}(\mathbb{P}_{\sigma}^{\bar{T}^{(j)}}, \mathbb{P}_{\sigma'}^{\bar{T}^{(j)}})$ is upper bounded by $\varepsilon_j'(e^{\varepsilon_j'}-1)$ and by $n_j D_{KL}(P_{j,\sigma}, P_{j,\sigma'})$ if $j \notin \mathcal{S}$. Combining everything we have the lemma that we desired to prove.

Proof (Proof of Lemma 2). For any $1 \leq i \leq n_j$, $j = 0, \ldots, m$ we define

$$
T_i^{(j)} = \frac{1}{h^d}\left(Y_i^{(j)} - \tfrac{1}{2}\right) K\left(\frac{X_i^{(j)} - x_0}{h}\right)
$$

Note that since $|Y_i^{(j)} - 1/2| = 1/2$ for all i, j, the variance of $T_i^{(j)}$ is bounded as

$$
\begin{aligned}
\sigma^2 = \max_{i,j} \mathrm{Var}(T_i^{(j)}) &\leq \frac{1}{4h^{2d}}\left(h^d \int_{[-1,1]^d} (K(t))^2 g(x_0 + th) dt\right) \\
&\leq \frac{c_K^2}{2h^d}\left(\int_{[-1,1]^d} g(x_0 + th) dt\right) \\
&\leq \frac{c_K^2}{2h^d}\left(2^d g_{\max}\right).
\end{aligned}
$$

Similarly we have the almost sure upper bound

$$
|T_i^{(j)}| \leq \frac{1}{h^d}\left|Y_i^{(j)} - \frac{1}{2}\right| \max_{u \in \mathbb{R}^d} K(u) \leq \frac{c_K}{2h^d}.
$$

Therefore we can use Bernstein inequality to write:

$$
\begin{aligned}
\mathbb{P}\left(\left|\sum_{j=0}^{m} \frac{u_j}{n_j} \sum_{i=1}^{n_j} (T_i^{(j)} - \mathbb{E}T_i^{(j)})\right| > t\right) &\leq \exp\left\{-t^2 h^d \left(c_K^2 (2^{d-1} g_{\max}) \sum_{j=0}^{m} \frac{u_j^2}{n_j} + \frac{c_K t}{2} \max_j \frac{u_j}{3n_j}\right)^{-1}\right\} \\
&\leq \exp\left\{-\frac{t^2 h^d}{C_0}\left(\sum_{j=0}^{m} \frac{u_j^2}{n_j} + \max_j \frac{u_j}{3n_j}\right)^{-1}\right\} \qquad (21)
\end{aligned}
$$

for any $0 < t < c_K(2^{d-1}g_{\max})$, where $C_0 = c_k 2^{d-1}$. By Gaussian concentration inequalities, the following bound holds true.

$$\mathbb{P}\left(\left|\sum_{j=0}^{m} \frac{u_j\sqrt{2c_K\log(2/\delta_j)}}{n_j\varepsilon_j h^d}\xi^{(j)}(x_0)\right| \geq t\right) \leq \exp\left\{-\frac{t^2}{c_K^2\log(2/\delta_{\min})}\left(\sum_{j=0}^{m}\frac{1}{n_j^2\varepsilon_j^2h^{2d}}\right)^{-1}\right\}. \tag{22}$$

Note that

$$\begin{aligned}\widetilde{T}(x_0) &:= \sum_{j=0}^{m}\left\{\frac{u_j}{n_jh^d}\sum_{i=1}^{n_j}\left(Y_i^{(j)}-\frac{1}{2}\right)K\left(\frac{X_i^{(j)}-x_0}{h}\right) + \frac{u_j\sqrt{2c_K\log(2/\delta_j)}}{n_j\varepsilon_j h^d}\xi^{(j)}(x_0)\right\}\\ &= \sum_{j=0}^{m}u_j\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}T_i^{(j)} + \frac{\sqrt{2c_K\log(2/\delta_j)}}{n_j\varepsilon_j h^d}\xi^{(j)}(x_0)\right\}\end{aligned}$$

where $\xi^{(j)}(x_0) \overset{iid}{\sim} N(0, K(0))$ for $j = 1, \ldots, m$. Then, combining (21) and (22) we get

$$\mathbb{P}\left(|T(x_0) - \mathbb{E}T(x_0)| \geq t\right) \leq \exp\left(-\frac{t^2h^d}{C_{up}}\left\{\sum_{j=0}^{m}u_j^2\left(\frac{1}{n_j}+\frac{1}{n_j^2\varepsilon_j^2h^d}\right) + \max_{0\leq j\leq m}\frac{u_j}{n_j}\right\}^{-1}\right).$$

for any $t \in (0, 1)$, where $C_{up} = (c_K^2 \vee 1)[2^{d-3}g_{\max} + \log(2/\delta_{\min})/4]$. This finishes the proof.

Proof (Proof of Lemma 3). We will first prove that a uniform concentration bound over x for T_h for each fixed $h \in \mathcal{H}$, whence a union bound over h will prove the result. For $(a, y, z, w) \in [0, 1]^d \times \{0, 1\} \times [0, 1] \times \Delta^m$ let us define the class of functions

$$\mathcal{F}_h(a, y, z, w) := \left\{z\left(y-\frac{1}{2}\right)K\left(\frac{a-x}{h}\right) - \sum_{j=0}^{m}w_j\mathbb{E}_{P_j}\left(Y-\frac{1}{2}\right)K\left(\frac{X-x}{h}\right) : x \in [0,1]^d\right\}$$

Since $K(\cdot)$ is L_K-Lipschitz, and $|y - \frac{1}{2}| = \frac{1}{2}$ for $y \in \{0, 1\}$, we can follow the proof of Lemma 14 in Kim et al. [2019], to obtain that the covering number $\mathcal{N}(\mathcal{F}_h, L_2, \eta)$ satisfies

$$\mathcal{N}(\mathcal{F}_h, L_2, \eta) \leq \left(\frac{((L_K/c_K)h^{-1}+1)c_K}{\eta}\right)^d \text{ for all } \eta \in (0, c_K) \text{ and } j \in \{0, 1\ldots, m\}. \tag{23}$$

Since $|Y_i - 1/2| = 1/2$ for all i, we have

$$a := \sup_x \frac{1}{h^d}\left|(Y-1/2)K\left(\frac{X-x}{h}\right)\right| \leq \frac{c_K}{2h^d} \tag{24}$$

$$\begin{aligned}\sigma^2 &:= \sup_{\mathbf{x}\in[0,1]^d}\mathbb{E}\left\{\frac{1}{h^{2d}}(Y-1/2)^2K^2\left(\frac{X-x}{h}\right)\right\}\\ &= \frac{1}{4h^{2d}}\int K^2\left(\frac{z-x}{h}\right)g(z)dz\\ &= \frac{1}{4h^d}\int K^2(t)g(x+th)dt \leq \frac{c_Kg_{\max}}{4h^d}\int K(t)dt = \frac{c_Kg_{\max}}{4h^d}.\end{aligned} \tag{25}$$

Let us define the vector

$$w^\dagger(h) = \left(\frac{n_1 \wedge n_1^2\varepsilon_1^2 h^d}{\sum_{j=1}^m n_j \wedge n_j^2\varepsilon_j^2 h^d}, \frac{n_2 \wedge n_2^2\varepsilon_2^2 h^d}{\sum_{j=1}^m n_j \wedge n_j^2\varepsilon_j^2 h^d}, \ldots, \frac{n_m \wedge n_m^2\varepsilon_m^2 h^d}{\sum_{j=1}^m n_j \wedge n_j^2\varepsilon_j^2 h^d} \right)^\top \in \Delta^{m-1}$$

so that we can write $\mathcal{W}(h) = \{(w_0, (1-w_0)w^\dagger(h)) : w_0 \in [0,1]\}$. We define

$$T_h^{(s)}(x) = \sum_{j=1}^m (w^\dagger(h))_j [T_h^{(j)}(x) - \mathbb{E}T_h^{(j)}(x)]$$

Note that for $j \in \{0, 1, \ldots, m\}$ we have

$$\sup_{x\in[0,1]^d} |T_h^{(s)}(x)| = \frac{1}{h^d} \sup_{f\in\mathcal{F}_h} \left| \sum_{j=1}^m \sum_{i=1}^{n_j} f\left(X_i^{(j)}, Y_i^{(j)}, \frac{(w^\dagger(h))_j}{n_j}, w^\dagger(h) \right) \right|$$

leading to the upper bound

$$\mathbb{E} \sup_{x\in[0,1]^d} |T_h^{(s)}(x)| \le C \sqrt{(d+1)c_K g_{\max} \log(8(L_K + c_K)/\sqrt{c_K g_{\max}}) \sum_{j=1}^m \frac{(w^\dagger(h))^2}{n_j h^d}} \tag{26}$$

where we use the bounds from (23) and (25) along with Proposition 2.1 from Giné and Guillou [2001]. We next define:

$$Z_j = \sup_{x\in[0,1]^d} [T_h^{(j)}(x) - \mathbb{E}T_h^{(j)}(x)]$$

By Talagrand's concentration inequality, in particular the form given in Theorems 1.1 and 1.2 of Klein and Rio [2005] we obtain:

$$\mathbb{P}\left(|Z_j - \mathbb{E}Z_j| \ge t_j\right) \le \exp\left[-\frac{1}{2} \min\left\{ \frac{n_j t_j^2}{\sigma^2 + 2\mathbb{E}aZ_j}, \frac{n_j t_j}{3a} \right\} \right] \quad \text{for } j = 0, \ldots, m.$$

where a and σ^2 are as defined in (24) and (25) respectively. Since a fixed linear combination of sub-exponential random variables is sub-exponential, it follows that $T_h^{(s)}(x)$ satisfies

$$\begin{aligned} &\mathbb{P}\left(\left| \sup_x T_h^{(s)}(x) - \mathbb{E}\sup_x T_h^{(s)}(x) \right| \ge t \right) \\ &\le \exp\left[-\frac{1}{2} \min\left\{ t^2 \left(\sum_{j=1}^m \frac{(w^\dagger(h))_j^2}{n_j} (\sigma^2 + 2a\mathbb{E}Z_j) \right)^{-1}, \frac{t}{3a \min\{(w^\dagger(h))_j/n_j\}} \right\} \right]. \end{aligned} \tag{27}$$

We next bound the Gaussian processes $\xi^{(j)}(\cdot)$. By Dudley's theorem [see, e.g., De la Pena and Giné, 2012] we have for a numerical constant $C > 0$ that

$$\mathbb{E} \sup_{x\in[0,1]^d} \xi^{(j)}(x) \le K(0) + C \int_0^1 \sqrt{\log\left\{ \left(1 + \frac{2}{\sigma}\right)^d \right\}} d\sigma \le C\sqrt{d} \quad \text{for } j = 0, 1, \ldots, m. \tag{28}$$

Moreover, $\sup_x \mathbb{E}(\xi^{(j)}(x))^2 = K(0) \le c_K$ for $j = 0, \dots, m$. Notice that writing the noise variance as $\sigma^2_{j,pvt} := \dfrac{2c_K \log(2|\mathcal{H}|/\delta_j)|\mathcal{H}|^2}{h^{2d}}$ we obtain that

$$\xi_h^{(s)}(\cdot) = \sum_{j=1}^{m} (w^\dagger(h))_j \frac{\sigma_{j,pvt}}{n_j \varepsilon_j} \xi^{(j)}(\cdot)$$

is a mean zero Gaussian process with covariance kernel

$$\mathrm{Cov}(\xi_h^{(s)}(t_1), \xi_h^{(s)}(t_2)) = K\left(\frac{t_1 - t_2}{h}\right) \sum_{j=1}^{m} (w^\dagger(h))_j^2 \frac{\sigma^2_{j,pvt}}{n_j^2 \varepsilon_j^2}.$$

By Gaussian concentration inequalities we therefore have that

$$\mathbb{P}\left(\left|\sup_x \xi_h^{(s)}(x) - \mathbb{E}\sup_x \xi_h^{(s)}(x)\right| \ge t\right) \le \exp\left[-\frac{t^2}{2}\left\{\sum_{j=1}^{m} \frac{(w^\dagger(h))_j^2 \sigma^2_{j,pvt}}{n_j^2 \varepsilon_j^2}\right\}^{-1}\right]. \tag{29}$$

Combining (27) and (29) we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\sup_x (T_h^{(s)}(x) + \xi_h^{(s)}(x)) - \mathbb{E}\sup_x (T_h^{(s)}(x) + \xi_h^{(s)}(x))\right| \ge t\right) \\
&\le \exp\left[-\frac{1}{2}\min\left\{t^2\left(\sum_{j=1}^{m} (w^\dagger(h))_j^2 \left(\frac{\sigma^2 + 2a\mathbb{E}Z_j}{n_j} + \frac{\sigma^2_{j,pvt}}{n_j^2\varepsilon_j^2}\right)\right)^{-1}, \frac{t}{3a\min\{(w^\dagger(h))_j/n_j\}}\right\}\right]
\end{aligned}$$

Now using (26) and (28) it follows after some calculation that

$$\begin{aligned}
&\sup_x |T_h^{(s)}(x) + \xi_h^{(s)}(x) - \mathbb{E}T_h^{(s)}(x)| \\
&\le C\sum_{j=1}^{m} (w^\dagger(h))j \left(\sqrt{\frac{(d+1)c_K g_{\max} \log(8(L_K + c_K)/\sqrt{c_K g_{\max}})}{n_j h^d}} + \frac{\sigma_{j,pvt}\sqrt{d}}{n_j \varepsilon_j}\right) \\
&\quad + \frac{3}{2}\sqrt{\log\left(\frac{2}{\nu}\right)\left(\sum_{j=1}^{m} (w^\dagger(h))_j^2 \left(\frac{\sigma^2}{n_j} + \frac{\sigma^2_{j,pvt}}{n_j^2\varepsilon_j^2}\right)\right)} \\
&\le \sqrt{\log\left(\frac{2}{\nu}\right)\left(\sum_{j=1}^{m} (w^\dagger(h))_j^2 \left(\frac{3\sigma^2}{n_j} + \frac{9\sigma^2_{j,pvt}}{4n_j^2\varepsilon_j^2}\right)\right)} = \frac{3}{2}\sqrt{v_0(h, (0, w^\dagger(h)))\log\left(\frac{2}{\nu}\right)}
\end{aligned} \tag{30}$$

with probability at least $1 - \nu/2$, provided $\nu \le 3n_*^{-1}$ for a sufficiently large constant $C > 0$. By an identical calculation one can show that

$$\begin{aligned}
&\sup_x \left|T_h^{(0)}(x) - \mathbb{E}T_h^{(0)}(x) + \frac{\sqrt{2c_K \log(2|\mathcal{H}|/\delta_j)|\mathcal{H}|}}{n_0 \varepsilon_0 h^d} \xi^{(0)}(x)\right| \\
&\le \frac{3}{2}\sqrt{v_0(h, (1, 0, 0, \dots, 0))\log\left(\frac{2}{\nu}\right)}
\end{aligned} \tag{31}$$

with probability at least $1-\nu/2$ provided $\nu \leq n_0^{-1}$ for sufficiently large n_0. Note that $\widetilde{T}(x,h,w) - \mathbb{E}\widetilde{T}(x,h,w)$ is a convex combination of the two quantities on the left hand side of equations (30) and (31), with weights $(1-w_0)$ and w_0 respectively. Moreover, note that

$$w_0\sqrt{v_0(h,(0,w^\dagger(h)))} + (1-w_0)\sqrt{v_0(h,(1,0,0,\ldots,0))} \leq \sqrt{2}\sqrt{v_0(h,w)} \text{ for all } w \in \mathcal{W}(h).$$

Therefore, choosing $\nu = \dfrac{1}{|\mathcal{H}|n_*}$, we now combine equations (30) and (31) by a union bound to obtain that, for all $w \in \mathcal{W}(h)$,

$$\sup_{x\in[0,1]^d} \left|\widetilde{T}(x,h,w) - \mathbb{E}\widetilde{T}(x,h,w)\right| \geq \frac{3\sqrt{2}}{2}\sqrt{\log(2|\mathcal{H}|n_*)}\sqrt{v_0(h,w)}$$

with probability at most $\dfrac{2}{|\mathcal{H}|n_*}$. Finally taking a union bound over all possible $h \in \mathcal{H}$ finishes the proof.

Proof (Proof of Lemma 4). Assume $f^*(x) = 1$, we know that $|\eta_Q(x) - 1/2| \geq \psi_0$ because $x \in G_{\psi_0}$. By (17), since $h \leq h_{ada,0}$ we obtain $(L(2h\sqrt{d})^\beta) \leq \psi_0/2$. Moreover, since $f^*(x) = 1$ we have that $\eta_Q(x) \geq \frac{1}{2} + L(2h\sqrt{d})^\beta$, so that we can use arguments identical to (14) to get

$$\mathbb{E}T_h^{(j)}(x) \geq \frac{b_K c_0}{2^d}\left(\eta_Q(x_0) - \frac{1}{2} - L(2h\sqrt{d})^\beta\right)^{\gamma_j}.$$

Since $(L(2h\sqrt{d})^\beta) \leq \psi_0/2$, we have that

$$\mathbb{E}T_h^{(j)}(x) \geq c_0 b_K (\psi_0/2)^{\gamma_j}(1/2)^d,$$

for $j = 0, 1, \ldots, m$. The case when $f^*(x) = 0$ follows similarly.

Proof (Proof of Lemma 5). Since $h_0 \leq h_{ada,0}$ this implies that $h_0 < \infty$ so the algorithm stops at h_0. By the stopping rule we know that:

$$\widehat{\rho}_0(h_0) \geq C_{unif}^2 \log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})$$

where

$$\sqrt{\widehat{\rho}_0(h_0)} = \max_{w\in\mathcal{W}(h_0)} \frac{|\widetilde{T}(x,h_0,w)|}{\sqrt{v_0(h_0,w)}}.$$

Let w_* be one of the values for which the RHS takes its maximum. Then we would have

$$|\widetilde{T}(x,h_0,w_*)| > \frac{3}{\sqrt{2}}\sqrt{\log(2n_*|\mathcal{H}|)}\sqrt{v_0(h_0,w_*)}.$$

By definition of $E_{A,0}$ we have that under $E_{A,0}$:

$$\left|\widetilde{T}(x, h_0, w_*) - \mathbb{E}\widetilde{T}(x, h_0, w_*)\right| \leq \frac{3}{\sqrt{2}}\sqrt{\log(2n_*|\mathcal{H}|)}\sqrt{v_0(h_0, w_*)}.$$

Combining the two inequalities we have

$$\left|\widetilde{T}(x, h_0, w_*) - \mathbb{E}\widetilde{T}(x, h_0, w_*)\right| < |\widetilde{T}(x, h_0, w_*)|,$$

which implies that

$$\operatorname{sign}\left(\widetilde{T}(x, h_0, w_*)\right) = \operatorname{sign}\left(\mathbb{E}\widetilde{T}(x, h_0, w_*)\right) \neq 0.$$

Note that since $h_0 \leq h_{ada,0}$, given $x \in G_{\psi,0}$, by Lemma 8 we have when $f^*(x) = 1$,

$$\mathbb{E}\widetilde{T}(x, h_0, w_*) \geq c_0 b_K(\psi_0/2)(1/2)^d + \sum_{j=1}^{m} c_0 b_K(\psi_0/2)^\gamma(1/2)^d > 0,$$

and when $f^*(x) = 0$

$$\mathbb{E}\widetilde{T}(x, h_0, w_*) \leq -c_0 b_K(\psi_0/2)(1/2)^d - \sum_{j=1}^{m} c_0 b_K(\psi_0/2)^\gamma(1/2)^d < 0.$$

So

$$\operatorname{sign}\left(\widetilde{T}(x, h_0, w_*)\right) = \operatorname{sign}\left(\mathbb{E}\widetilde{T}(x, h_0, w_*)\right) = \begin{cases} 1 & \text{if } f^*(x) = 1 \\ -1 & \text{if } f^*(x) = 0 \end{cases}$$

Hence $\widehat{f}_0(x) = f^*(x)$.

Proof (Proof of Lemma 6). We will show that under the event $E_{A,0}$ the algorithm stops at $h_{ada,0}$ if it does not stop earlier.

For all $x \in G_{\psi_0} \cap \{x : f^*(x) = 1\}$, we apply Lemma 4 to get that:

$$\begin{aligned} \mathbb{E}T^{(0)}_{h_{opt,0}}(x) &\geq c_0 b_K(\psi_0/2)(1/2)^d \geq \frac{c_0 b_K C_{\psi_0}}{2^d} \times h^{\beta}_{ada,0} \\ \mathbb{E}T^{(j)}_{h_{opt,0}}(x) &\geq c_0 b_K(\psi_0/2)^{\gamma_j}(1/2)^d \geq \frac{c_0 b_K C^{\gamma}_{\psi_0}}{2^d} \times h^{\beta\gamma}_{ada,0} \quad \text{for } j = 1, \ldots, m. \end{aligned} \tag{32}$$

The rest of the proof can be separated into two cases.

Case 1: $(n_0 \wedge n_0^2\varepsilon_0^2 h^d_{ada,0})h^{2\beta+d}_{ada,0} \geq \sum_{j=1}^m (n_j \wedge n_j^2\varepsilon_j^2 h^d_{ada,0})h^{2\beta\gamma+d}_{ada,0}$.

Case 2: $(n_0 \wedge n_0^2\varepsilon_0^2 h^d_{ada,0})h^{2\beta+d}_{ada,0} < \sum_{j=1}^m (n_j \wedge n_j^2\varepsilon_j^2 h^d_{ada,0})h^{2\beta\gamma+d}_{ada,0}$.

Since the steps are analogous, we write the rest of the proof only for Case 1. This implies

$$\begin{aligned} h_{ada,0}^{\beta} &\geq \frac{1}{\sqrt{2}} \frac{1}{h_{ada,0}^{d/2}} \sqrt{2(h_{ada,0})^{2\beta+d}} \\ &\geq \sqrt{\frac{1}{2(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d}} \sqrt{2(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d)(h_{ada,0})^{2\beta+d}} \\ &\geq \sqrt{\frac{1}{2(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d}} \times \sqrt{\sum_{k=0}^{m} (n_k \wedge n_k^2(\varepsilon_k/|\mathcal{H}|)^2 h_{ada}^d) h_{ada,0}^{2\beta\gamma_k+d}} \\ &= \sqrt{\frac{\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d}} \end{aligned} \tag{33}$$

where in the second last line we write $\gamma_0 = 1$ and $\gamma_k = \gamma$ for $k = 1, \ldots, m$; while the last line follows by (18). Thus for any $x \in G_{\psi,0} \cap \{x : f^*(x) = 1\}$ we have by (32) that

$$\mathbb{E}T_{h_{ada,0}}^{(0)}(x) \geq 6 \sqrt{\frac{\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d}}$$

where the last step uses (33). Along with the uniform concentration inequality from Lemma 3 used with $w = (1, 0, \ldots, 0)$ and $h = h_{ada,0}$, we have

$$T_{h_{ada,0}}^{(0)}(x) \geq \mathbb{E}T_{h_{ada,0}}^{(0)}(x) - |T_{h_{ada,0}}^{(0)}(x) - \mathbb{E}T_{h_{ada,0}}^{(0)}(x)| > 3 \sqrt{\frac{\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d}}.$$

whenever $x \in G_{\psi,0} \cap \{x : f^*(x) = 1\}$. In the other case, i.e., when $x \in G_{\psi,0} \cap \{x : f^*(x) = 0\}$ we similarly have

$$T_{h_{ada,0}}^{(0)}(x) \leq \mathbb{E}T_{h_{ada,0}}^{(0)}(x) + |T_{h_{ada,0}}^{(0)}(x) - \mathbb{E}T_{h_{ada,0}}^{(0)}(x)| < -3 \sqrt{\frac{\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d}}.$$

Combining the two cases above we obtain

$$(T_{h_{ada,0}}^{(0)}(x))^2 > 4.5 \left(\frac{\log(2n_*|\mathcal{H}|)\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d) h_{ada,0}^d} \right).$$

By definition of $\widehat{\rho}(\cdot)$, we have for the choice of weights $w_{(0)} = (1, 0 \ldots, 0)$ that

$$\begin{aligned} \widehat{\rho}_0(h_{ada,0}) &= \sup_{w \in \mathcal{W}(h_{ada,0})} \frac{(\widetilde{T}(x, h_{ada,0}, w))^2}{v_0(h_{ada,0}, w)} \geq \frac{(\widetilde{T}(x, h_{ada}, w_{(0)}))^2}{v_0(h_{ada,0}, w_{(0)})} = \frac{(T_{h_{ada,0}}^{(0)}(x))^2}{v_0(h_{ada}, w_{(0)})} \\ &> 4.5 \log(2n_*|\mathcal{H}|). \end{aligned}$$

Thus under event $E_{A,0}$, for all $x \in G_{\psi,0}$, we obtain by the definition of h_0 that since $\widehat{r}(\cdot)$ is monotonically increasing, we must have $h_0(x) \leq h_{ada,0}(x)$. By Lemma 5, it follows that

$$\widehat{f}_0(x) = f^*(x) \text{ for all } x \in G_{\psi}.$$

Next note that under case 1 above,

$$\psi_0^{1+\alpha} = c_{\psi_0} h_{ada,0}^{\beta(1+\alpha)} \geq c'_{\psi,0} \left(\frac{\log(n_* |\mathcal{H}|)}{n_0 \wedge n_0^2 (\varepsilon_0/|\mathcal{H}|)^2 h_{ada,0}^d} \right)^{1/2} \geq 2n_*^{-1}$$

by definition of n_*. Since by Lemma 7, the event $E_{A,0}$ holds with probability at least $1 - 2n_*^{-1} \geq 1 - \psi_0^{1+\alpha}$, this finishes the proof.

Proof (Proof of Lemma 7). The proof is very similar to the proof of Lemma 3 and we highlight the important differences. Similar to (26) we can write for $0 \leq j \leq m$ that

$$\begin{aligned} &\sup_x |T_h^{(j)}(x) - \mathbb{E}T_h^{(j)}(x)| \\ &\leq C\sqrt{\frac{(d+1)c_K g_{\max} \log(8L_K\sqrt{c_K}/\sqrt{g_{\max}})}{n_j h^d}} + \frac{3}{2}\sqrt{\frac{c_K g_{\max} \log((\nu/(m+1))^{-1})}{4n_j h^d}} \\ &\leq \sqrt{\frac{3c_K g_{\max} \log((m+1)/\nu)}{4n_j h^d}} \end{aligned}$$

with probability at least $1 - \nu/(m+1)$ for large enough n_j, provided $\nu \leq n_j^{-1}$. Adding the above bounds we have by the triangle inequality that

$$\mathbb{P}\left(\sup_{x\in[0,1]^d, w\in\Delta^m} \left| \sum_{j=0}^m w_j T_h^{(j)}(x) - \mathbb{E}[\sum_{j=0}^m w_j T_h^{(j)}(x)] \right| \geq \sum_{j=0}^m w_j \sqrt{\frac{3c_K g_{\max} \log((m+1)/\nu)}{4n_j h^d}} \right) \leq \nu, \tag{34}$$

assuming $\nu \leq \min_j n_j^{-1}$. We next bound the Gaussian processes $\xi^{(j)}(\cdot)$. By Dudley's theorem [see, e.g., De la Pena and Giné, 2012] we have for a numerical constant $C > 0$ that

$$\mathbb{E} \sup_{x\in[0,1]^d} \xi^{(j)}(x) \leq K(0) + C\int_0^1 \sqrt{\log\left\{\left(1 + \frac{2}{\sigma}\right)^d\right\}} d\sigma \leq C\sqrt{d}.$$

Notice moreover that $\sup_x \mathbb{E}(\xi^{(j)}(x))^2 = K(0) \leq c_K$. Thus we have by Gaussian concentration inequalities that

$$\mathbb{P}\left(\sup_{x\in[0,1]^d} |\xi^{(j)}(x)| \geq \frac{3}{2}\sqrt{c_K \log((m+1)/\nu)} \right) \leq \frac{\nu}{m+1}$$

which implies for $\nu \leq n_j^{-1}$ and sufficiently large n_j that

$$\begin{aligned} &\sup_{x\in[0,1]^d, w\in\Delta^m} \left| \sum_{j=0}^m w_j \frac{\sqrt{2c_K \log(2|\mathcal{H}|/\delta_j)}|\mathcal{H}|}{n_j \varepsilon_j h^d} \xi^{(j)}(x_0) \right| \\ &\leq \sum_{j=0}^m w_j \frac{\sqrt{4.5c_K^2 \log(2|\mathcal{H}|/\delta_j)\log((m+1)/\nu)}}{n_j(\varepsilon_j/|\mathcal{H}|)h^d} \end{aligned} \tag{35}$$

with probability at most ν. Choosing $\nu = \frac{1}{|\mathcal{H}|n_*}$ we then have

$$\sup_{x\in[0,1]^d, w\in\Delta^m} \left|\widetilde{T}(x,h,w) - \mathbb{E}\widetilde{T}(x,h,w)\right| \le \frac{3}{2}\sqrt{(m+1)c_K \log(|\mathcal{H}|n_*(m+1))}\sqrt{v(h,w)}$$

with probability at most $\frac{2}{|\mathcal{H}|n_*}$, where the last line follows by the definition of $v(h,w)$ in comparison to the right hand sides of (34) and (35). Then taking a union bound over all possible $h \in \mathcal{H}$ finishes the proof.

Proof (Proof of Lemma 8). The proof is identical to the proof of Lemma 4 and hence omitted.

Proof (Proof of Lemma 9). The proof is identical to the proof of Lemma 5 and hence omitted.

Proof (Proof of Lemma 10). Similar to the proof of Lemma 6, we will show that under the event E_A the algorithm stops at h_{ada} if it does not stop earlier.

For all $x \in G_\psi \cap \{x : f^*(x) = 1\}$, we apply Lemma 8 to get that for any $j = 0, 1, \ldots, m$,

$$\mathbb{E}T^{(j)}_{h_{opt}}(x) \ge c_0 b_K (\psi/2)^{\gamma_j} (1/2)^d \ge \frac{c_0 b_K C_\psi^{\gamma_j}}{2^d} \times h_{ada}^{\beta\gamma_j}. \tag{36}$$

The rest of the proof can be separated into $(m+1)$ cases.

Case j: $(n_j \wedge n_j^2\varepsilon_j^2 h_{ada}^d)h_{ada}^{2\beta\gamma_j+d} = \max\left\{(n_k \wedge n_k^2\varepsilon_k^2 h_{ada}^d)h_{ada}^{2\beta\gamma_k+d} : 0 \le k \le m\right\}$ for $j \in \{0, 1, \ldots, m\}$.

Since the steps are analogous, we write the rest of the proof only for Case 0. This implies

$$\begin{aligned}
h_{ada}^\beta &\ge \frac{1}{\sqrt{m+1}} \frac{1}{h_{ada}^{d/2}} \sqrt{(m+1)(h_{ada})^{2\beta+d}} \\
&\ge \sqrt{\frac{1}{(m+1)(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^d}} \sqrt{(m+1)(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)(h_{ada})^{2\beta+d}} \\
&\ge \sqrt{\frac{1}{(m+1)(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^d}} \times \sqrt{\sum_{k=0}^m (n_k \wedge n_k^2(\varepsilon_k/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^{2\beta\gamma_k+d}} \\
&= \sqrt{\frac{(m+1)\log(n_*|\mathcal{H}|(m+1))\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^d}}
\end{aligned} \tag{37}$$

where the last line follows by (18). Thus for any $x \in G_\psi \cap \{x : f^*(x) = 1\}$ we have by (36) that

$$\mathbb{E}T^{(0)}_{h_{ada}}(x) \ge 4C_{unif}\sqrt{\frac{(m+1)\log(n_*|\mathcal{H}|(m+1))\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{opt}^d)h_{opt}^d}}$$

where the last step uses (37). Along with the uniform concentration inequality from Lemma 7 used with $w = 0$ and $h = h_{ada}$, we have

$$T^{(0)}_{h_{ada}}(x) \geq \mathbb{E}T^{(0)}_{h_{ada}}(x) - |T^{(0)}_{h_{ada}}(x) - \mathbb{E}T^{(0)}_{h_{ada}}(x)| > 3\sqrt{\frac{(m+1)\log(n_*|\mathcal{H}|(m+1))\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^d}}.$$

whenever $x \in G_\psi \cap \{x : f^*(x) = 1\}$. In the other case, i.e., when $x \in G_\psi \cap \{x : f^*(x) = 0\}$ we similarly have

$$\begin{aligned} T^{(0)}_{h_{ada}}(x) &\leq \mathbb{E}T^{(0)}_{h_{ada}}(x) + |T^{(0)}_{h_{ada}}(x) - \mathbb{E}T^{(0)}_{h_{ada}}(x)| \\ &< -3\sqrt{\frac{(m+1)\log(n_*|\mathcal{H}|(m+1))\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^d}}. \end{aligned}$$

Combining the two cases above we obtain

$$(T^{(0)}_{h_{ada}}(x))^2 > 9\left(\frac{\log(n_*|\mathcal{H}|(m+1))\log(2|\mathcal{H}|/\delta_{\min})}{(n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d)h_{ada}^d}\right).$$

By definition of $\widehat{\rho}(\cdot)$, we have for the choice of weights $w_0 = (1, 0 \ldots, 0)$ that

$$\widetilde{T}(x, h_{ada}, w_0) = T^{(0)}_{h_{ada}}(x) \quad \text{and} \quad v(h_{ada}, w_0) = \frac{1}{n_0 h_{ada}^d} + \frac{2\log(2|\mathcal{H}|/\delta_{\min})}{n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^{2d}}.$$

and hence

$$\begin{aligned} \widehat{\rho}(h_{ada}) &= \sup_{w \in \Delta^m} \frac{(\widetilde{T}(x, h_{ada}, w))^2}{v(h_{ada}, w)} \geq \frac{(\widetilde{T}(x, h_{ada}, w_0))^2}{v(h_{ada}, w_0)} = \frac{(T^{(0)}_{h_{ada}}(x))^2}{v(h_{ada}, w_0)} \\ &> 2.25(m+1)\log(n_*|\mathcal{H}|(m+1)). \end{aligned}$$

Thus under event E_A, for all $x \in G_\psi$, we obtain by the definition of h_* that since $\widehat{\rho}(\cdot)$ is monotonically increasing, we must have $h_*(x) \leq h_{ada}(x)$. By Lemma 9, it follows that

$$\widehat{f}_a(x) = f^*(x) \text{ for all } x \in G_\psi.$$

Next note that under case 0,

$$\psi^{1+\alpha} = c_\psi h_{ada}^{\beta(1+\alpha)} \geq c'_\psi \left(\frac{\log(n_*|\mathcal{H}|)}{n_0 \wedge n_0^2(\varepsilon_0/|\mathcal{H}|)^2 h_{ada}^d}\right)^{1/2} \geq 2n_*^{-1}$$

by definition of n_*. Since by Lemma 7, the event E_A holds with probability at least $1 - 2n_*^{-1} \geq 1 - \psi^{1+\alpha}$, this finishes the proof.

Proof (Proof of Lemma 11). This is earlier stated and proved as Lemma C4 in Cai et al. [2023b].

Proof (Proof of Lemma 12). This is earlier stated and proved as Lemma C5 in Cai et al. [2023b].

References

Arnab Auddy, Abhinav Chakraborty, and T. Tony Cai. Minimax and adaptive nonparametric classification for transfer learning under distributed differential privacy constraints. *Technical Report*, 2024.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007. doi: 10.1214/009053606000001217. URL https://doi.org/10.1214/009053606000001217.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100 – 128, 2021. doi: 10.1214/20-AOS1949. URL https://doi.org/10.1214/20-AOS1949.

T. Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *Technical Report*, 2023a.

T Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Supplement to "Optimal federated learning for nonparametric regression with heterogenous distributed differential privacy constraints", 2023b.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Evarist Giné and Armelle Guillou. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. In *Annales de l'IHP Probabilités et statistiques*, volume 37, pages 503–522, 2001.

Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727, 2013.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning*, pages 3398–3407. PMLR, 2019.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060 – 1077, 2005. doi: 10.1214/009117905000000044. URL https://doi.org/10.1214/009117905000000044.