arXiv:2103.07095v1 [math.ST] 12 Mar 2021

Minimax Optimal Conditional Density Estimation under Total Variation Smoothness

Michael Li Matey Neykov Sivaraman Balakrishnan

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

{mli4, mneykov, sbalakri}@andrew.cmu.edu

Abstract

This paper studies the minimax rate of nonparametric conditional density estimation under a weighted absolute value loss function in a multivariate setting. We first demonstrate that conditional density estimation is impossible if one only requires that $p_{X|Z}$ is smooth in x for all values of z. This motivates us to consider a sub-class of absolutely continuous distributions, restricting the conditional density $p_{X|Z}(x|z)$ to not only be Hölder smooth in x, but also be total variation smooth in z. We propose a corresponding kernel-based estimator and prove that it achieves the minimax rate. We give some simple examples of densities satisfying our assumptions which imply that our results are not vacuous. Finally, we propose an estimator which achieves the minimax optimal rate adaptively, i.e., without the need to know the smoothness parameter values in advance. Crucially, both of our estimators (the adaptive and non-adaptive ones) impose no assumptions on the marginal density p_Z, and are not obtained as a ratio between two kernel smoothing estimators which may sound like a go to approach in this problem.

1 Introduction

A significant yet challenging problem in statistical inference is how to learn from complex, multidimensional data. While the nonparametric regression problem of estimating conditional mean $\mathbb{E}(x|z)$ from an *i.i.d.* sample of (X, Z) is well studied, the alternative problem of estimating the full conditional density $p_{X|Z}(x|z)$ remains largely unexplored. There is little literature studying minimax optimal conditional density estimation, and particularly not when both X and Z are in a multivariate setting. However, the advantages of estimating $p_{X|Z}(x|z)$ over just the conditional mean are numerous. Fundamentally, the conditional mean is a summary of the conditional density. It follows that conditional density yields more information about the data and can be more useful for subsequent analysis. This is especially important when there exists multi-modality, asymmetry, or heteroscedastic noise in $p_{X|Z}(x|z)$, in which case the conditional mean $\mathbb{E}(x|z)$ would be insufficient to explain the data and to do inference. Furthermore, the problem of nonparametric quantile regression [1] can be solved via conditional density estimation. Finally, when forecasting and making predictions in fields such as economics, conditional density has been proven to be a key component of interest [2]. However, although the advantages of conditional density estimation are clear, it is a harder problem than conditional mean estimation, which in turn raises the need to impose stronger assumptions.

To the best of our knowledge in this paper we give the first matching minimax upper and lower bounds for conditional density estimation in a multivariate setting. Concretely, the problem we consider is the following. Suppose here and throughout the paper that we have random variables

$X \in [0,1]^{d_X}$, $Z \in [0,1]^{d_Z}$, and n independent and identically distributed (i.i.d.) observations $\mathcal{D}_n = \{(X_1, Z_1), \ldots, (X_n, Z_n)\}$, coming from a joint distribution $p_{X,Z}$ that is absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$, where $d = d_X + d_Z$. Our goal is to estimate the conditional density $p_{X|Z}(x|z)$ with the estimate $\widehat{p}_{X|Z}(x|z)$ where $x = (x_1, \ldots, x_{d_X}), z = (z_1, \ldots, z_{d_Z})$. We will focus on the following loss function

$$\mathbb{E} \int \int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz, \tag{1.1}$$

where p_Z denotes the marginal density of Z, the expectation is taken with respect to n i.i.d. samples from $p_{X,Z}$, and dx and dz are shorthands for $\prod_{i \in [d_X]} dx_i$, $\prod_{i \in [d_Z]} dz_i$ respectively. The loss function (1.1) is largely inspired by the works [3, 4], where the authors consider the squared version of this loss. The L_1-based loss function that we use has several benefits, and has been argued for in past works (see, for instance, [5, 6] in the context of density estimation, and [7, 8] in the context of density testing). The L_1-based distance metric induced by this loss function is invariant to monotonic transformations, and in contrast to the L_2-based distance, closeness in this distance has a clear probabilistic interpretation. Equivalently, the loss function we define may be interpreted as the L_1 distance between the joint distribution $\widehat{p}_{X|Z}(x|z)p_Z(z)$ and the joint distribution $p_{X,Z}(x,z)$. Furthermore, we can decompose the loss function into two parts. First we have $\int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| dx$ which is equal to the L_1 distance between the estimated density and the target density. Next we weigh this distance by p_Z to stress on the regions where Z is more common, and downweight regions where Z is less common. An important point that is worth making is that in this work we do not impose any assumptions on p_Z, which is enabled by the fact that the true density p_Z is present in the loss function. Hence our estimators can handle situations where p_Z may be non-differentiable and not even continuous. This is in stark contrast with an approach that one may be willing to take, i.e., to assume that p_Z is Hölder smooth and estimate the conditional distribution as a ratio between kernel smoothed estimators of the joint $p_{X,Z}$ and the marginal p_Z.

We note that minimax rates with respect to this loss function have not been previously studied in the literature, and in fact any analysis of the minimax rates of conditional density estimation is scarce. The closest minimax analysis is given by Efromovich [9], where the author studied minimax rates of conditional density estimation under an unweighted squared loss function. Unlike in the present work, [9] only focused on the one dimensional setting, i.e., when $d_X = d_Z = 1$. Additionally, there exist more significant differences in the assumptions made and the problem settings, which will be elaborated on later.

1.1 Relevant Literature

In this section we review some of the relevant literature. In a classical work, Rosenblatt [10] proposed a kernel based estimate of $p_{X,Z}$ and p_Z and combined them using the formula $p_{X|Z} = \frac{p_{X,Z}}{p_Z}$. Assuming that $p_{X|Z}$, p_Z and the conditional mean of $X|Z$ have continuous second derivatives, Hyndman et al. [11] analyzed the mean integrated squared error of a ratio between two kernel smoother estimators in the $d_X = 1$ dimensional case. Bashtannyk and Hyndman [3] looked into optimal bandwidth selection in the aforementioned kernel smoother estimate. Fan et al. [12] used locally polynomial regression to develop nonparametric estimate of the conditional density function in nonlinear dynamical systems. In a follow-up work, Fan and Yim [13] used cross-validation to select the bandwidth of the double-kernel estimator developed by Fan et al. [12]. Hall et al. [14]

used cross-validation to automatically reduce the number of relevant covariates when estimating the conditional density, but they did not study the minimax rates of estimation. In a related work, Hall et al. [15] proposed a different method for estimating the density using dimension reduction. Hall et al. [16] studied methods for conditional distribution estimation based on parametric and nonparametric techniques, including a logistic model and a Nadaraya-Watson estimator. A different method using dimension reduction was proposed by Efromovich [17] where the author used an orthogonal series based approach. Chagny [18] used an expansion of a "warped" conditional density onto a space spanned by orthonormal bases. Recently, Ćevid et al. [19] studied conditional density estimation using an adapted Random Forest algorithm. In conclusion, although there has been some work on conditional density estimation, the minimax optimal rate is an open question. In this paper we address this question for the loss function (1.1) under certain smoothness assumptions on the conditional distributions $p_{X|Z=z}$.

1.2 Summary of Results

We begin by showing that conditional density estimation is impossible if one does not impose sufficient assumptions on the class of distributions. In particular, assuming that $p_{X|Z}$ is smooth in x for all z is not enough and further assumptions are needed. We formally prove this fact by arguing that for any sample size $n \in \mathbb{N}$, there exists a finite class of distributions whose conditional densities $p_{X|Z}$ are Hölder smooth (see Definition 2.1) in x for all z, for which the worst case loss is bounded from below by a constant. This result motivates the assumptions that we impose next.

We formalize a class of distributions $\mathcal{P}_{\beta,\gamma}$ consisting of conditional densities that are Hölder smooth with smoothness β in x, and γ-total variation (γ-TV) smooth in z (see Definition 3.1). We show the following result:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}_{\beta,\gamma}} \mathbb{E}_p \int \int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \asymp n^{\frac{-1}{\beta^{-1} d_X + \gamma^{-1} d_Z + 2}},$$

where $\asymp$ means equality up to constant factors, and $\mathbb{E}_p$ is the expectation over n i.i.d. samples, each of which comes from the distribution p. This minimax rate is achieved by a kernel-based estimator, which is defined in (3.1). Furthermore, observe that there is a curse of dimensionality, where the dimensions d_X, d_Z may have different effects on the rate depending on the corresponding smoothing parameters β and γ.

Finally, we devise an adaptive estimator to achieve the minimax optimal rate without the need to know the values of the smoothness parameters β and γ in advance. Our estimator is based on the work of Yatracos [20], but requires delicate care and crucial modifications, since we do not possess knowledge of, and are not willing to make any assumptions on the marginal p_Z.

1.3 Notation

The following notations will be used throughout the paper. We use $p_{X,Z} = p_{X|Z} \cdot p_Z$ to denote any joint distribution (and density function) of the pair of random variables (X, Z). We also use $p_{X|Z}(x|z)$, $p_{X|Z=z}$ to denote the conditional density function and the conditional distribution of $X|Z = z$ respectively, and p_Z to denote the marginal distribution (and density function) of Z. For an integer n we use the shorthand $[n] = \{1, \ldots, n\}$.

We also use multi-index notations. Suppose we have vectors $x = (x_1, \ldots, x_{d_X})$, $\alpha = (\alpha_1, \ldots, \alpha_{d_X})$ such that $x \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}_+^{d_X}$, where $\mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\}$, then we have

$$\|\alpha\|_1 = \sum_{i=1}^{d_X} |\alpha_i|, \quad \alpha! = \prod_{i=1}^{d_X} \alpha_i!, \quad x^\alpha = \prod_{i=1}^{d_X} x_i^{\alpha_i}.$$

Furthermore

$$D^\alpha f = \frac{\partial^{\|\alpha\|_1} f}{\partial x_1^{\alpha_1} \ldots \partial x_{d_X}^{\alpha_{d_X}}}.$$

We let $\lfloor \beta \rfloor$ denote the greatest integer strictly less than the real number β. We also use $\lesssim, \gtrsim$ to mean inequalities up to universal constants, and we write $f(n) \asymp g(n)$ if both $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold.

2 Impossibility

In this section, we will show that it is, in general, impossible to estimate the conditional density at a reasonable rate unless some assumptions on the class of distributions are imposed. Importantly, we show that even if one is willing to assume that $p_{X|Z}$ is smooth in x for all z, it is still insufficient, and more assumptions are needed. Intuitively, when the conditioning variable Z has a continuous density we observe no replicates (multiple samples with identical Z values) and it is necessary to impose that the conditional densities $p_{X|Z}$ are smooth in z in order to reliably estimate $p_{X|Z}$. Our impossibility result, Theorem 2.2, formalizes this intuition.

As detailed in the introduction, for an estimate $\widehat{p}_{X|Z}(x|z)$, based on a dataset $\mathcal{D}_n = \{(X_1, Z_1), \ldots, (X_n, Z_n)\}$ with n observations and a density $p_{X|Z}(x|z)$, we will use the loss function (1.1). In order to formally state our result, we first define Hölder smoothness.

Definition 2.1 (Hölder smoothness)**.** *We say that the collection of conditional densities $p_{X|Z}(x|z)$ for $Z \in [0,1]^{d_Z}$ is Hölder smooth with some constant W_1 and smoothness β, where β, W_1 are positive numbers, if it is $\ell = \lfloor \beta \rfloor$ times differentiable, and for all $x, x' \in [0,1]^{d_X}$, $z \in [0,1]^{d_Z}$ satisfies*

$$\sup_\alpha |D^\alpha p_{X|Z}(x|z) - D^\alpha p_{X|Z}(x'|z)| \leq W_1 \|x - x'\|_1^{\beta - \ell}, \quad \text{for all } \alpha \text{ such that } \|\alpha\|_1 = \ell, \alpha \in \mathbb{N}_0^{d_X},$$

where $\alpha = (\alpha_1, ..., \alpha_{d_X})$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

We then have the following result:

Theorem 2.2 (Impossibility of Conditional Density Estimation)**.** *Let the sample size n be any fixed integer. Then for any constants β, W_1 and $\varepsilon > 0$, there exists a finite class of distributions $\mathcal{C}(\beta, W_1)$ (whose cardinality depends on n and ε) on $[0,1]^{d_X + d_Z}$ satisfying the following three properties:*

i. the marginal Z densities are absolutely continuous with respect to the Lebesgue measure on $[0,1]^{d_Z}$, with density equal to p_Z (which can be specified by the user),

ii. the conditional distributions $p_{X|Z}$ are Hölder smooth with constant W_1 and smoothness β, and

iii. the following inequality holds

$$\inf_{\widehat{p}} \max_{p \in \mathcal{C}(\beta, W_1)} \mathbb{E}_p \int \int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \geq \kappa - \epsilon,$$

where κ is some positive constant that depends on β and W_1. More specifically, if $\beta \leq 1$ and $W_1 < \frac{2}{d_X^\beta}$, we have $\kappa = \frac{W_1^2 d_X^{(2\beta-1)}}{144}$, and otherwise $\kappa = \frac{1}{36 d_X}$.

Remarks:

1. It is important to note that this result holds for any arbitrary marginal Z density p_Z (i.e. it can be chosen to be arbitrarily smooth and known to the statistician). Our result shows that consistent conditional density estimation is impossible when the only assumptions made are that the marginal density p_Z, and the conditional densities $p_{X|Z}$ are smooth (no matter how smooth they are).

2. A straightforward extension of our proof shows that one can further relax the condition on the marginal p_Z. If the absolutely continuous part of Z has probability mass at least θ for some $\theta > 0$ then an identical argument will show that the minimax error can be made arbitrarily close to $\kappa\theta$.

Here we provide a sketch of the proof, while the full proof is deferred to Section 7.1.

Proof Sketch. We define a "null" distribution by taking $p_{X|Z}$ to be uniform on $[0,1]^{d_X}$ for all values of Z. We then construct a family of "alternate" distributions which are perturbations of the null distribution constructed in the following way. We first construct a pair of smooth distributions p_1, p_2 such that they yield that uniform distribution when mixed with equal weights, but which are individually sufficiently far from uniform. We divide the support of Z into many small intervals, and in each interval, we randomly (with equal probability) set $p_{X|Z}$ to be either p_1 or p_2. This constructs a large family of possible alternate distributions. We then argue that with high probability each sample point Z_i falls into different intervals, and show that this in turn makes it impossible to distinguish whether the samples came from the null distribution or the uniform mixture over the possible alternate distributions. □

Theorem 2.2 illustrates that if we only assume that $p_{X|Z}$ is smooth in x for all z (e.g. Hölder smooth), we can construct a finite collection of distributions such that any estimate will produce an expected error of at least κ in the worst case sense. Importantly, the proof makes use of the fact that we do not see replications of the densities $p_{X|Z}$ for different Z values. We can remedy this by assuming that the distributions $p_{X|Z=z}$ vary smoothly with z. In the following section, we do so by imposing a Total Variation smoothness assumption on z, and show that under such conditions we can obtain reasonable (and minimax-optimal) bounds (i.e., bounds decreasing with the sample size) on the loss function. Intuitively, this happens since with additional smoothness assumptions, one can group observations whose Z_i values are close, while this strategy is unavailable in the general setting.

3 Upper Bound

In this section we propose an estimate $\widehat{p}_{X|Z}(x|z)$ under certain smoothness assumptions on the class of distributions. Formally, suppose we have a joint distribution of two variables (X, Z): $p_{X,Z}(x, z)$ where $x = (x_1, \ldots, x_{d_X}), z = (z_1, \ldots, z_{d_Z})$ and $X \in [0,1]^{d_X}$, $Z \in [0,1]^{d_Z}$. We assume that the conditional density $p_{X|Z}(x|z)$ satisfies Hölder smoothness in x (see Definition 2.1) and the following γ-total variation (γ-TV) smoothness in z. We denote the class of densities which satisfy our smoothness conditions by $\mathcal{P}_{\beta,\gamma}$.

Definition 3.1 (γ-TV smoothness)**.** *We say that the distribution is γ-total variation (γ-TV) smooth if the following inequality holds for some $0 < \gamma \leq 1$, and for all $x \in [0,1]^{d_X}$ and $z, z' \in [0,1]^{d_Z}$:*

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \leq W_2 \|z - z'\|_1^{\gamma},$$

for some sufficiently large constant W_2.

In the above, the L_1 distance between probability densities (equal to 2 times the TV distance) is defined as:

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 = 2\,\mathrm{TV}(p_{X|Z=z}, p_{X|Z=z'}) = \int |p_{X|Z}(x|z) - p_{X|Z}(x|z')| dx.$$

In other words, TV smoothness requires that the distributions $p_{X|Z=z}$ vary smoothly with z in the L_1 sense. This assumption is inspired by [21], where the authors used a similar assumption to establish the minimax rate for conditional independence testing. Furthermore, $\gamma \leq 1$ is required due to the following lemma:

Lemma 3.2. *Suppose $\gamma > 1$, and the inequality $\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \leq W_2 \|z - z'\|_1^{\gamma}$ from Definition 3.1 holds. Then it must be that $p_{X|Z=z} \equiv p_{X|Z=z'}$ for all $z, z' \in [0,1]^{d_Z}$.*

Proof. Fix any two points z, z' in $[0,1]^{d_Z}$. Take $\alpha_j = \frac{j}{k+1}$, $j = 0, 1, \ldots, k+1$. Let $z_j = \alpha_j z + (1 - \alpha_j) z'$. Then by γ-TV smoothness with $\gamma > 1$ we have

$$\mathrm{TV}(p_{X|Z=z}, p_{X|Z=z'}) \leq \sum_{i=0}^{k} \mathrm{TV}(p_{X|Z=z_i}, p_{X|Z=z_{i+1}}) \leq W_2 \left(\frac{\|z - z'\|_1}{k+1} \right)^{\gamma} (k+1).$$

Taking $k \to \infty$ lets us conclude that $p_{X|Z=z} = p_{X|Z=z'}$ as desired. □

Finally, the estimator we propose under the Hölder smoothness assumption makes use of kernels. Below we define a class of kernels that can be used in the estimator to achieve the minimax optimal rate.

Definition 3.3 (Appropriate Kernels)**.** *We say that a kernel $K : \mathbb{R}^{d_X} \to \mathbb{R}$ is appropriate if*

$$\int K(u) du = 1, \quad \int u^{\alpha} K(u) du = 0, \text{ for all } \boldsymbol{\alpha} \text{ such that } \|\alpha\|_1 \leq \ell, \alpha \in \mathbb{N}_0^{d_X},$$

where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. In addition the kernel should satisfy

$$\int K^2(u) du < \infty$$

and

$$\int |K(u)| \cdot |u^\alpha| < \infty, \text{ for all } \alpha \text{ such that } \|\alpha\|_1 \leq \beta, \alpha \in \mathbb{R}_+^{d_X},$$

where $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Importantly, appropriate kernels do exist, and one method of constructing them is detailed in Lemma 3.4 below.

Lemma 3.4 (Appropriate Kernels' Construction)**.** *We can construct appropriate kernels* $K : \mathbb{R}^{d_X} \to \mathbb{R}$ *using a product kernel* $K(u) = \prod K_i(u_i)$*, where each* K_i *is a kernel of order* ℓ *as defined in [22, Proposition 1.3].*

We prove this lemma in Appendix B. We now formally define the estimator. Recall that we are interested in estimating the conditional density $p_{X|Z}(x|z)$ with the estimate $\widehat{p}_{X|Z}(x|z)$ under the loss function (1.1). We propose a histogram-type estimator that uses kernel smoothing. Namely, bin $[0,1]$ into intervals $A_1, \ldots, A_m$ of equal length (m^{-1}), and consider the hyper-rectangles created from the Cartesian product of such intervals over $[0,1]^{d_Z}$. We define the following shorthand notations: let $\bar{j} = (j_1, ..., j_{d_Z}) \in [m]^{d_Z}$ denote the bin indices for some d_Z dimensional hyper-rectangle. Then $A_{\bar{j}} = \prod_{k=1}^{d_Z} A_{j_k}$ denotes that hyper-rectangle itself, where $\prod$ stands for the Cartesian product between sets. Finally, define the estimate

$$\widehat{p}_{X|Z}(x|z) = \sum_{\bar{j} \in [m]^{d_Z}} \mathbb{1}\left(z \in A_{\bar{j}}\right) \frac{\sum_{i \in [n]} \mathbb{1}(Z_i \in A_{\bar{j}}) K(\frac{X_i - x}{h})}{h^{d_X} \sum_{i \in [n]} \mathbb{1}(Z_i \in A_{\bar{j}})}, \tag{3.1}$$

where $K : \mathbb{R}^{d_X} \to \mathbb{R}$ is an appropriate multi-dimensional kernel as in Definition 3.3 and $\frac{0}{0}$ is understood as 0. Note that we only apply binning to Z here, and apply kernel smoothing to X. Our estimator (3.1) need not be a proper density since it need not be positive, but we provide a simple modification below. This modified estimator has the same properties as $\widehat{p}_{X|Z}$ but is in fact a proper density. This will play an important role when we devise our adaptive estimator in Section 6.

Theorem 3.5 (Hölder Upper Bound)**.** *Suppose that* $p_{X,Z} \in \mathcal{P}_{\beta,\gamma}$*. Using the estimate* (3.1) *with an appropriate kernel as per Definition 3.3, and selecting the parameters* $h \asymp n^{\frac{-1}{d_X + d_Z \beta \gamma^{-1} + 2\beta}}$ *and* $m \asymp n^{\frac{\beta}{d_X + d_Z \beta \gamma^{-1} + 2\beta}}$ *(for some appropriately selected constants) we have that*

$$\mathbb{E} \int \int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \lesssim n^{\frac{-1}{\beta^{-1} d_X + \gamma^{-1} d_Z + 2}} =: r_n(\beta, \gamma, d_X, d_Y). \tag{3.2}$$

Theorem 3.5 provides an upper bound for the estimator (3.1), and in Section 4 we will derive a matching lower bound. Combining the two proves that (3.2) is in fact the minimax optimal rate, and therefore no estimator can do better than $\widehat{p}_{X|Z}$ up to constant factors. We defer the proof of Theorem 3.5 to Section 7.2. Roughly, the proof consists of a "bias" and "variance" decomposition (in quotation marks due to the fact that the two terms in the decomposition are not exactly bias and variance since our loss function is not the squared loss) and carefully controlling both ensures that the rate exhibited in (3.2) holds.

Remarks:

1. Equation (3.2) shows that the estimator we propose exhibits the curse of dimensionality. In particular, the presence of β and γ smoothness parameters makes sense intuitively. Recall that the class of distributions $\mathcal{P}_{\beta,\gamma}$ assumes conditional densities that are β-Hölder smooth in x, and γ-TV smooth in z, which matches the effects observed here. We note that when holding both dimensions d_X, d_Z constant, our estimator performs better for higher values of β or γ smoothness (i.e. smoother densities). Indeed, as β increases, the effect of d_X on the estimator's effectiveness diminishes, and when $\beta \to \infty$, the dimension of X has no effect on the minimax rate at all. In addition, since our proof does not require the restriction $\gamma \leq 1$, we note that when we take $\gamma \to \infty$, the minimax rate simply reduces to that of classical unconditional density estimation of x in a multivariate setting with Hölder smoothness assumptions. In fact, by Lemma 3.2 any value of $\gamma > 1$ can be thought of as $\gamma \to \infty$.

2. Once again, we would like to stress the fact that we make no assumptions on the marginal density of Z. This is in stark contrast to an approach one may be compelled to take, by first estimating the joint density $p_{X,Z}(x,z)$, then the marginal density of z by integrating x out, and dividing the two to arrive at an estimate of $p_{X|Z}$. This implies consistently estimating $p_{X,Z}(x,z)$ and $p_Z(z)$, which likely requires assumptions on both of these densities, whereas that is not needed in our case. In fact, Theorem 5.1 provides examples of one such class of Hölder smooth densities where our approach is minimax optimal regardless of what p_Z is, whereas the aforementioned approach will likely fail.

3. Finally, as discussed earlier, there does not exist work closely comparable to ours. The most related paper is [9], where the author studied local minimax rates for conditional density estimation in a bivariate case (i.e. $d_X = d_Z = 1$). Furthermore, [9] imposed vastly different assumptions (e.g. the conditional densities are assumed to belong to a perturbed Sobolev class), and utilized different loss functions. The difference in the two problem settings renders the comparison of the resulting minimax rates nonproductive.

We now provide a modified estimator of (3.1) to ensure that it is a proper density. Consider the following estimator:

- if $\widehat{p}_{X|Z=z} \not\equiv 0$ (which implies $\int \widehat{p}_{X|Z}(x|z)dx = 1$), define $\bar{p}_{X|Z}(x|z) = C^{-1}(\widehat{p}_{X|Z}(x|z))_+$ where $C = \int (\widehat{p}_{X|Z}(x|z))_+ dx$, and for a function $f(x)$ we denote $(f(x))_+ = f(x)\mathbb{1}(f(x) \geq 0)$;
- else if $\widehat{p}_{X|Z=z} \equiv 0$, define $\bar{p}_{X|Z=z} \equiv 1$.

Lemma 3.6. *$\bar{p}_{X|Z}(x|z)$ satisfies (3.2), and is a proper density.*

The details of Lemma 3.6 are given in Appendix B. Finally, recall our result in equation (3.2), which shows an upper bound for the loss function (1.1) by a quantity which is of the same order as $r_n := r_n(\beta, \gamma, d_X, d_Y)$. Now consider our modified estimator $\bar{p}_{X|Z}(x|z)$. We have the following result which is a simple consequence of Markov's inequality:

Lemma 3.7. *For any $\epsilon > 0$ there exists a set A_ϵ satisfying $\mathbb{P}(Z \in A_\epsilon) \geq 1 - \epsilon$, such that for all $z \in A_\epsilon$ we have*

$$\mathrm{TV}(\bar{p}_{X|Z}(x|z), p_{X|Z}(x|z)) \lesssim \frac{1}{\epsilon} r_n.$$

Proof. Since $\bar{p}_{X|Z}$ is a density this allows us to write

$$\int |\bar{p}_{X|Z}(x|z) - p_{X|Z}(x|z)|dx = 2\,\mathrm{TV}(\bar{p}_{X|Z}(x|z), p_{X|Z}(x|z)).$$

Let $f(z) := \mathrm{TV}(\bar{p}_{X|Z}(x|z), p_{X|Z}(x|z))$, and note that by Theorem 3.5 and Lemma 3.6 we know that

$$\mathbb{E}f(z) \lesssim r_n.$$

By the Markov's inequality we have

$$\mathbb{P}\left(f(z) > \frac{1}{\epsilon}\mathbb{E}[f(z)]\right) \leq \epsilon.$$

This implies that there exists a subset of the support of Z – call it A_ϵ – which has probability of at least $1-\epsilon$ to occur, and for all $z \in A_\epsilon$, $f(z) \leq \frac{1}{\epsilon}\mathbb{E}[f(z)] \lesssim \frac{1}{\epsilon}r_n$. This completes the proof. □

Lemma 3.7 illustrates that we can estimate well an overwhelming majority (in terms of the Z distribution) of conditional densities in terms of TV distance. Next we move on to establish a lower bound.

4 Minimax Lower Bound

In this section we produce a minimax lower bound for the estimation problem with the loss function (1.1). We recall the definition of the class $\mathcal{P}_{\beta,\gamma}$ in Section 3.

Theorem 4.1 (Hölder Lower Bound)**.** *For any $p_Z(z) \geq c$ where c is some constant, we have that*

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}_{\beta,\gamma}} \mathbb{E}_p \int\int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)|p_Z(z)dxdz \gtrsim n^{\frac{-1}{\beta^{-1}d_X + \gamma^{-1}d_Z + 2}}.$$

Remarks:

1. Theorem 3.5 and this theorem together show that the proposed estimate (3.1) achieves the minimax rate $n^{\frac{-1}{\beta^{-1}d_X + \gamma^{-1}d_Z + 2}}$ under the loss function (1.1).
2. It is worth comparing and contrasting the results of this Theorem with our earlier impossibility result in Theorem 2.2. In rough terms, they convey the same basic intuition that when γ is very small (or 0) conditional density estimation is difficult (or impossible) in a minimax sense. This result is more quantitative, capturing in a more precise sense the dependence on γ. On the other hand, the result of Theorem 2.2 is more flexible, allowing essentially any marginal density p_Z (not requiring it to be lower bounded by a constant), as long as the marginal density has some non-trivial absolutely continuous component (as discussed in the remarks following Theorem 2.2).

Here we provide a sketch of the proof for the minimax optimal lower bound. The full proof is deferred to Section 7.3.

Proof Sketch. We first define a class of conditional density functions and show that their joint distributions indeed belong to $\mathcal{P}_{\beta,\gamma}$. Then, we apply Fano's inequality to derive the minimax lower bound.

The conditional density functions are defined as

$$p^{\Delta}_{X|Z}(x|z) = 1 + \sum_{\bar{i}\in[m]^{d_X}} \sum_{\bar{j}\in[m]^{d_Z}} \Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{i_k}(x_k) \prod_{k\in[d_Z]} g_{j_k}(z_k),$$

where recall the shorthands $\bar{i} \in [r]^{d_X}, \bar{j} \in [m]^{d_Z}$ (r, m are integers chosen later), and $\Delta_{\bar{i},\bar{j}} \in \{\pm 1\}$. The intuition for such a construction is to add multiple small perturbations to the uniform conditional density function by using infinitely differentiable bump functions h_{i_k}, g_{j_k}. We proceed to verify that the constructed conditional density functions are indeed density functions (i.e. always positive and integrates to 1), and follow both the γ-TV smoothness condition as in Definition 3.1 and the Hölder smoothness condition as in Definition 2.1.

In order to apply Fano's inequality [23], we first show that there exists a subset of the conditional density functions defined above, such that the distance between any pair as measured by the loss function is sufficiently large (more specifically, is lower bounded by some ϵ). This is done by using Varshamov-Gilbert's construction [22, Lemma 2.9]. We then find an upper bound on the Kullback-Leibler (KL) divergence between any pair of our conditional density functions. Finally, applying Markov's inequality to the Fano's inequality allows us to express the minimax lower bound in terms of the distance lower bound and KL divergence upper bound we just derived. Making some optimal selection of parameter values completes the proof and produces the desired matching minimax optimal lower bound. □

Importantly, note that in the process of proving Theorem 4.1, we constructed a class of conditional density functions and showed that it satisfies all our assumptions. In order to better understand the class $\mathcal{P}_{\beta,\gamma}$, we develop further examples of densities belonging to this class in the next section.

5 Examples

In this section we provide examples of distributions which belong to $\mathcal{P}_{\beta,\gamma}$. Recall that this class of distributions requires conditional densities to be Hölder smooth (see Definition 2.1) and γ-TV smooth (see Definition 3.1). We already saw examples of such distributions in the proof of Theorem 4.1. Below we give two additional classes of examples for different values of the smoothness β.

Theorem 5.1 (Examples for $\beta > 1$)**.** *Suppose $g(x,z) : [0,1]^d \mapsto \mathbb{R}$ is such that $g(x,z) \geq a > 0$ for some constant a, and is Hölder smooth with smoothness $\beta > 1$ in both x and z, i.e.,*

$$\sup_{\alpha} |D^{\alpha} g(x,z) - D^{\alpha} g(x',z')| \leq C(\|x - x'\|_1 + \|z - z'\|_1)^{\beta-\ell},$$

for all α such that $\|\alpha\|_1 = \ell, \alpha \in \mathbb{N}_0^{d_X}$, where $\ell = \lfloor \beta \rfloor$. Then if $p_{X|Z}(x|z) = \frac{g(x,z)}{\int g(x,z)dx}$, we have $p_{X|Z} \in \mathcal{P}_{\beta,1} \subseteq \mathcal{P}_{\beta,\gamma}$, for any $\gamma \leq 1$.

Theorem 5.2 (Examples for $\beta \leq 1$)**.** *Suppose $g(x,z) : [0,1]^d \mapsto [-M, M]$ is a bounded function such that*

$$|g(x,z) - g(x',z')| \leq C(\|x - x'\|_1^{\beta} + \|z - z'\|_1^{\gamma}).$$

Then if $p_{X|Z}(x|z) = \frac{\exp(g(x,z))}{\int \exp(g(x,z))dx}$, *we have* $p_{X,Z} \in \mathcal{P}_{\beta,\gamma}$.

The proofs of Theorem 5.1 and 5.2 are given in Appendix C.

6 Hyperparameter Tuning and Selection

We have shown that our proposed estimate (3.1) achieves the minimax-optimal rate $n^{\frac{-1}{\beta^{-1}d_X+\gamma^{-1}d_Z+2}}$ under the loss function (1.1). However, in doing so we manually picked the values of the hyperparameters h and m, which depend on the smoothness parameters β and γ of the true distribution. Here we introduce an adaptive method of selecting the hyperparameters without needing to assume the knowledge of β and γ.

Towards the goal of hyperparameter tuning we first design and analyze a selection procedure for conditional density estimation which satisfies a type of oracle inequality with respect to the loss (1.1). Given a collection of candidate conditional density estimates we devise a procedure which selects one which has nearly minimal loss. Our procedure is inspired by a minimum distance estimate described in the work of Yatracos [20], and further developed in the works [5, 6]. However, in contrast to these works our selection procedure only has access to conditional density estimates (as opposed to joint density estimates), and we aim to design a selection procedure tailored to the loss (1.1) (as opposed to the usual L_1 loss on the joint densities). Furthermore, our goal is to avoid smoothness assumptions which would be required to estimate the marginal of Z, and this necessitates careful modifications of the minimum distance procedure.

We describe our oracle inequality in Theorem 6.1 and use this result to develop an adaptive conditional density estimate which achieves the same minimax-optimal rates as the estimate in (3.1) without knowledge of the smoothness parameters in Section 6.2.

6.1 A Modified Selection Procedure for Conditional Density Estimates

To begin with we consider the following setup. We are given access to a collection of conditional density estimates $\widehat{f}_1, \ldots, \widehat{f}_N$, which are either fixed, or estimated on a separate sample. Our goal is to select an estimate of (nearly) minimal loss.

We associate each estimate with an oracle joint distribution $\widetilde{f}_j(x,z) = \widehat{f}_j(x|z)p_Z(z)$ where p is the unknown true density of the samples. Associated with each pair (i,j) of density estimates we define the so-called Yatracos set:

$$A_{ij} = \{(x,z) : \widetilde{f}_i(x,z) > \widetilde{f}_j(x,z)\}.$$

We note that we can compute A_{ij} even without access to the unknown density p. Denote the collection of such sets $\mathcal{A}$. For a set A we let A^z denote the subset with $Z = z$.

Given n samples $\{(X_1, Z_1), \ldots, (X_n, Z_n)\}$ from p, we use the following minimum distance estimator:

$$\psi = \underset{\widehat{f}_j : j \in [N]}{\operatorname{argmin}} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \int_{A^{Z_i}} \widehat{f}(x|Z_i)dx - \mathbb{P}_n(A) \right|.$$

In rough terms our selection procedure compares, for each candidate $\widehat{f}_j$, an estimate of the mass of the Yatracos sets under $\widetilde{f}_j$ to an estimate of the population mass of these sets, selecting the candidate for which the largest discrepancy is smallest.

We show the following result:

Theorem 6.1. *With probability at least* $1-\delta$,

$$\int_z \|\psi(x|z) - p(x|z)\|_1 p_Z(z)dz \leq 3 \min_j \int_z \|\widehat{f}_j - p(x|z)\|_1 p_Z(z)dz + 14\sqrt{\frac{\log(N/\delta)}{n}}.$$

Remarks:

1. Our method and analysis are inspired by those of Yatracos [20]. The crucial insight of Yatracos is that when our goal is to select one of a collection of candidates, the supremum over the relatively small collection of Yatracos sets is adequate as a (statistically and computationally) tractable proxy for the supremum over all measurable sets (the TV/L_1 distance).

2. The guarantee of the theorem is extraordinary in that we are able to obtain an oracle inequality with an excess error which scales as $1/\sqrt{n}$, despite the fact that accurately estimating the loss (1.1) of even a single estimate would require many more samples. Furthermore, the guarantee degrades only logarithmically in the number of estimates N we are aiming to select from and in the failure probability δ. This, for instance, will be important in the next section when we use the method to select from a large collection of candidate density estimates constructed using different values of the tuning parameter.

3. Although not the main focus of our paper, the computational costs of constructing the Yatracos sets and computing the minimum distance estimate are discussed extensively in [5, 6]. At the expense of a slightly worse guarantee one might use a tournament-based selection rule which has a computational cost which scales linearly (as opposed to quadratically) in the number of estimates N.

6.2 Adaptive Conditional Density Estimation

With our previous result in place we now describe an adaptive conditional density estimate which achieves the rate $n^{\frac{-1}{\beta^{-1}d_X+\gamma^{-1}d_Z+2}}$, without knowledge of the smoothness parameters β and γ. We assume throughout that β and γ are upper bounded by some (unknown, but fixed) universal constants. We split our sample in two halves, using one half to construct a collection of candidate density estimates, and the second half to select one of these candidates following the procedure in Section 6.1. In practice, one might choose instead to use cross-fitting, where we repeat this procedure swapping the roles of the two samples, and average the two resulting estimates. It is straightforward to show that our guarantees continue to hold for the cross-fit variant as well.

We postulate two intervals where h and m are assumed to lie in respectively. Recall our choice of optimal hyperparameter values while proving Theorem 3.5: $h \asymp n^{\frac{-1}{d_X+d_Z\beta\gamma^{-1}+2\beta}}$ and $m \asymp n^{\frac{\beta}{d_X+d_Z\beta\gamma^{-1}+2\beta}}$. Based on this result we consider values of the tuning parameters in two sets $\mathcal{I}_1 = \{n^{-1/d_X}, 2\times n^{-1/d_X}, \ldots, 2^{\lceil \log_2 n^{1/d_X}\rceil} n^{-1/d_X}\}$, and $\mathcal{I}_2 = \{1, 2, 4, \ldots, 2^{\lceil (\log_2 n)/2\rceil}\}$. We consider all pairs of tuning parameters $(h, m) \in \mathcal{I}_1 \times \mathcal{I}_2$, noting that there are at most $N := \mathcal{O}(\log^2 n)$ such choices. For each possible hyperparameter combination, we compute our conditional density estimate $\bar{p}_j(x|z)$, for $j \in [N]$ (we use the truncated and renormalized estimate analyzed in Lemma 3.6). At least one of these estimates achieves the minimax rate of $n^{\frac{-1}{\beta^{-1}d_X+\gamma^{-1}d_Z+2}}$ and it thus only remains to select a sufficiently good candidate.

We apply the Yatracos procedure from Section 6.1 to select a candidate $\widehat{\psi}$ using the second half of the sample, and we obtain the following result.

Theorem 6.2 (Adaptive Conditional Density Estimation)**.** *Suppose that $p_{X,Z} \in \mathcal{P}_{\beta,\gamma}$. The tuning-parameter free procedure described above yields an estimate $\widehat{\psi}$ such that,*

$$\mathbb{E} \int \int |\widehat{\psi}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \lesssim n^{\frac{-1}{\beta^{-1} d_X + \gamma^{-1} d_Z + 2}}.$$

Remarks:

1. The proof of this result is fairly straightforward given the result of the previous section, and only requires us to convert the high-probability bound from Theorem 6.1 so that we may use our previously derived upper bound in Theorem 3.5. We present the details in Section 7.5.

2. We note that, in contrast to works on pointwise adaptation over smoothness classes [24] where typically a logarithmic price is unavoidable, we design an adaptive estimator which achieves the (oracle) minimax-rate (of Theorem 4.1).

7 Proofs

In this section we present the proofs of the main results of our paper, deferring remaining technical aspects to the Appendix.

7.1 Proof of Theorem 2.2

We start by constructing a joint distribution $p_{X,Z}(x,z)$ such that $p_{X|Z}$ is uniform on $[0,1]^{d_X}$ for all values of Z (i.e. X is independent of Z), and the marginal of Z is equal to p_Z which is specified by the user. We will now construct multiple distributions out of $p_{X,Z}(x,z)$. Take disjoint Borel measurable sets $C_1, \ldots, C_m \subset [0,1]^{d_Z}$ such that $\int_{C_i} p_Z(z) dz = m^{-1}$ and $\cup_{i \in [m]} C_i = [0,1]^{d_Z}$.

For a given β and W_1, we construct two distributions p_1, p_2 which are Hölder smooth with constant W_1 and smoothness β, and have sufficiently large total variation (TV) distance from the uniform distribution. In addition, their mixture distribution with equal weights produces $p_{X|Z}$, i.e.

$$\frac{1}{2} p_1 + \frac{1}{2} p_2 = p_{X|Z} = U([0,1]^{d_X}).$$

For a constant $0 < c < 1$ which we will set appropriately in the sequel, we define p_1, p_2 as linear functions of the x_i as follows:

$$p_1(x) = 2(1-c) \sum_{i \in [d_x]} \frac{x_i}{d_X} + c, \quad p_2(x) = 2 - p_1(x).$$

The following result develops some properties of the distributions p_1 and p_2.

Lemma 7.1. *1. The densities p_1, p_2 are positive, integrate to 1, and satisfy the property that $\frac{1}{2} p_1 + \frac{1}{2} p_2 = U([0,1]^{d_X})$.*

2. When $\beta \leq 1$ the densities are Hölder with $W_1 = 2(1-c)/d_X^{\beta}$ and if $\beta > 1$ they are Hölder for any value $W_1 \geq 0$.

3. Furthermore, $\mathrm{TV}(U([0,1]^{d_X}), p_1) = \mathrm{TV}(U([0,1]^{d_X}), p_2) \geq \frac{(1-c)^2}{18 d_X} =: c_{TV}$.

Now, for a $\Delta = (\Delta_1, \ldots, \Delta_m) \in \{1,2\}^m$, construct the distribution $p^{\Delta}_{X,Z}(x,z)$ which has the same marginal distribution on Z as $p_{X,Z}(x,z)$, i.e., p_Z, and the conditional distributions defined as follows: for $z \in C_j$ we have

$$p^{\Delta}_{X|Z}(\cdot|z) = p_{\Delta_j}(\cdot).$$

We know from Lemma 7.1 that $\mathrm{TV}(U([0,1]^{d_X}), p_1) = \mathrm{TV}(U([0,1]^{d_X}), p_2) \geq c_{\mathrm{TV}}$ where c_{TV} is some positive constant, so it follows that

$$\int\int |p_{X|Z}(x|z) - p^{\Delta}_{X|Z}(x|z)| p_Z(z) dx dz = \int 2\,\mathrm{TV}(U([0,1]^{d_X}), p_1) p_Z(z) dz \geq 2c_{\mathrm{TV}} > 0. \tag{7.1}$$

Next the proof will emulate a classical reduction scheme from estimation to a testing problem. This reduction is similar to the one described in Section 2.2 of [22], yet there are differences hence we provide full details. For brevity note that our loss function is

$$\|\widehat{p}_{X|Z} p_Z - q_{X|Z} p_Z\|_1 = \int\int |\widehat{p}_{X|Z}(x|z) - q_{X|Z}(x|z)| p_Z(z) dx dz.$$

By Markov's inequality, we have that:

$$\mathbb{E}\|\widehat{p}_{X|Z} p_Z - q_{X|Z} p_Z\|_1 \geq c_{\mathrm{TV}} \mathbb{P}\Big(\|\widehat{p}_{X|Z} p_Z - q_{X|Z} p_Z\|_1 \geq c_{\mathrm{TV}} \Big). \tag{7.2}$$

Define the finite class of distributions $\mathcal{C}(\beta, W_1) = \{p_{X|Z}, \{p^{\Delta}_{X|Z}\}_{\Delta \in \{1,2\}^m}\}$, and for ease of notation enumerate the elements of $\mathcal{C}(\beta, W_1)$ by $p_0 = p_{X|Z}$, $p_{\Delta+1} = p^{\Delta}_{X|Z}$ where with a slight abuse of notation we refer to Δ as the integer with binary representation Δ. Thus the cardinality of the set $|\mathcal{C}(\beta, W_1)| = 2^m + 1$. In light of (7.2), it follows that in order to lower bound the quantity

$$\inf_{\widehat{p}_{X|Z}} \max_{i \in \{0\} \cup [2^m]} \mathbb{E}_{p_i p_Z} \|\widehat{p}_{X|Z} p_Z - p_i p_Z\|_1,$$

it suffices to control

$$\inf_{\widehat{p}_{X|Z}} \max_{i \in \{0\} \cup [2^m]} \mathbb{P}_{p_i p_Z} \Big(\|\widehat{p}_{X|Z} p_Z - p_i p_Z\|_1 \geq c_{\mathrm{TV}} \Big),$$

where we are indexing the expectation and the probability to stress that the distribution of the data is generated under the distribution $p_i p_Z$ (importantly note that we have n i.i.d. observations from $p_i p_Z$). Next we notice that

$$\mathbb{P}_{p_0 p_Z} \Big(\|\widehat{p}_{X|Z} p_Z - p_0 p_Z\|_1 \geq c_{\mathrm{TV}} \Big) \geq \mathbb{P}_{p_0 p_Z}(\psi^* \neq 0),$$

and

$$\mathbb{P}_{p_i p_Z} \Big(\|\widehat{p}_{X|Z} p_Z - p_i p_Z\|_1 \geq c_{\mathrm{TV}} \Big) \geq \mathbb{P}_{p_i p_Z}(\psi^* = 0),$$

where

$$\psi^* = \operatorname*{argmin}_{0\le i\le 2^m} \|\widehat{p}_{X|Z}p_Z - p_i p_Z\|_1,$$

and the above two inequalities follow by the triangle inequality and (7.1). We conclude that

$$\begin{aligned}\inf_{\widehat{p}_{X|Z}} \max_{i\in\{0\}\cup[2^m]} \mathbb{P}_{p_i p_Z}\left(\|\widehat{p}_{X|Z}p_Z - p_i p_Z\|_1 \ge c_{\mathrm{TV}}\right)\\ \ge \inf_{\psi} \max(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \max_{i\in[2^m]} \mathbb{P}_{p_i p_Z}(\psi = 0)),\end{aligned}$$

where the inf is taken over all measurable test functions with values in the set $\{0\}\cup[2^m]$. Using the fact that the max is bigger than the average we have

$$\begin{aligned}&\inf_{\widehat{p}_{X|Z}} \max_{i\in\{0\}\cup[2^m]} \mathbb{P}_{p_i p_Z}\left(\|\widehat{p}_{X|Z}p_Z - p_i p_Z\|_1 \ge c_{\mathrm{TV}}\right)\\ &\quad\ge \inf_{\psi} \max(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \max_{i\in[2^m]} \mathbb{P}_{p_i p_Z}(\psi = 0))\\ &\quad\ge \inf_{\psi} \max\left(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \frac{1}{2^m}\sum_{i\in[2^m]} \mathbb{P}_{p_i p_Z}(\psi = 0)\right).\end{aligned} \tag{7.3}$$

Suppose we are able to show that the TV distance between these distributions can be made arbitrarily small, i.e. for any $\epsilon > 0$ we can ensure that,

$$\mathrm{TV}\left(\mathbb{P}_{p_0 p_Z}, \frac{1}{2^m}\sum_{i\in[2^m]} \mathbb{P}_{p_i p_Z}\right) \le \epsilon, \tag{7.4}$$

then as a consequence we obtain that,

$$\begin{aligned}&\inf_{\psi} \max\left(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \frac{1}{2^m}\sum_{i\in[2^m]} \mathbb{P}_{p_i p_Z}(\psi = 0)\right)\\ &\quad\ge \inf_{\psi} \max\left(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \mathbb{P}_{p_0 p_Z}(\psi = 0) - \epsilon\right)\\ &\quad\ge \frac{1}{2} - \epsilon.\end{aligned}$$

Hence we conclude following (7.2) that,

$$\inf_{\widehat{p}_{X|Z}} \max_{i\in\{0\}\cup[2^m]} \mathbb{E}_{p_i p_Z}\|\widehat{p}_{X|Z}p_Z - p_i p_Z\|_1 \ge \frac{c_{\mathrm{TV}}}{2} - c_{\mathrm{TV}}\epsilon.$$

It remains to specify the choice of the constant c in our definition of p_1. When $\beta > 1$ or when $W_1 \le 2/d_X^\beta$ we can choose $c = 0$ to obtain the lower bound of $1/36d_X$. Otherwise, we must choose c large enough to ensure that $2(1-c)/d_X^\beta \le W_1$, i.e. we choose $c = 1 - (W_1 d_X^\beta)/2$ to obtain the lower bound of $\frac{W_1^2 d_X^{(2\beta-1)}}{144}$ as claimed, completing the proof of the theorem.

Finally, we prove the total variation bound in (7.4). Note that,

$$\mathrm{TV}\left(\mathbb{P}_{p_0 p_Z}, \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}\right) = \sup_{A\in\Sigma}\left|\mathbb{P}_{p_0 p_Z}(A) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A)\right|,$$

where Σ is the Borel σ-field. Let B be the event where at least two points Z_i for $i \in [n]$ belong to the same bin C_k for some k. The complement B^c is therefore the event where each point Z_i for $i \in [n]$ falls into its own bin. For any event A we have

$$\begin{aligned}
\left|\mathbb{P}_{p_0 p_Z}(A) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A)\right| &\le \left|\mathbb{P}_{p_0 p_Z}(A\cap B^c) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A\cap B^c)\right| \\
&\quad + \left|\mathbb{P}_{p_0 p_Z}(A\cap B) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A\cap B)\right| \\
&\le \left|\mathbb{P}_{p_0 p_Z}(A\cap B^c) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A\cap B^c)\right| \\
&\quad + \left|\mathbb{P}_{p_0 p_Z}(A\cap B)\right| + \left|\frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A\cap B)\right| \\
&\le \left|\mathbb{P}_{p_0 p_Z}(A\cap B^c) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A\cap B^c)\right| + 2\mathbb{P}_{p_0 p_Z}(B),
\end{aligned}$$

where in the last inequality we used the fact that $\mathbb{P}_{p_0 p_Z}(B) = \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(B)$ since the two distributions have the same marginal distribution on Z. Now by the definition of our distributions p_i we know that the mixture distribution $\frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}$ assigns the same measure to the set $A \cap B^c$ as the distribution $\mathbb{P}_{p_0 p_Z}$, and therefore

$$\left|\mathbb{P}_{p_0 p_Z}(A) - \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(A)\right| \le 2\mathbb{P}_{p_0 p_Z}(B).$$

Due to the definitions of the sets C_i we have that $\mathbb{P}_{p_0 p_Z}(B) = \frac{m^n - m(m-1)\ldots(m-n+1)}{m^n} = O(\frac{1}{m})$, for a sufficiently large m. It follows that

$$\mathrm{TV}\left(\mathbb{P}_{p_0 p_Z}, \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}\right) \le O(\frac{1}{m}).$$

Thus going back to (7.3), we have

$$\begin{aligned}
&\inf_{\psi}\max\left(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \frac{1}{2^m}\sum_{i\in[2^m]}\mathbb{P}_{p_i p_Z}(\psi = 0)\right) \\
&\ge \inf_{\psi}\max\left(\mathbb{P}_{p_0 p_Z}(\psi\neq 0), \mathbb{P}_{p_0 p_Z}(\psi = 0) - O(\frac{1}{m})\right) \\
&\ge \frac{1}{2} - O\left(\frac{1}{m}\right).
\end{aligned}$$

Taking m large enough completes the proof.

7.2 Proof of Theorem 3.5

For each $A_{\bar{j}}$, define the corresponding estimate

$$\widehat{p}_{X,\bar{j}}(x) := \frac{\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) K(\frac{X_i - x}{h})}{h^{d_X} \sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})}.$$

Using the triangle inequality we have

$$\begin{aligned}
&\mathbb{E} \int\int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \\
&\leq \int\int |\widetilde{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz + \mathbb{E} \int\int |\widehat{p}_{X|Z}(x|z) - \widetilde{p}_{X|Z}(x|z)| p_Z(z) dx dz, \qquad (7.5)
\end{aligned}$$

where

$$\widetilde{p}_{X|Z}(x|z) = \sum_{\bar{j}} \mathbb{1}\left(z \in A_{\bar{j}}\right) \mathbb{E}\left[\frac{\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) K(\frac{X_i - x}{h})}{h^{d_X} \sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})}\right] = \sum_{\bar{j}} \mathbb{1}\left(z \in A_{\bar{j}}\right) \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)],$$

and

$$\widehat{p}_{X|Z}(x|z) = \sum_{\bar{j}} \mathbb{1}\left(z \in A_{\bar{j}}\right) \frac{\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) K(\frac{X_i - x}{h})}{h^{d_X} \sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})} = \sum_{\bar{j}} \mathbb{1}\left(z \in A_{\bar{j}}\right) \widehat{p}_{X,\bar{j}}(x).$$

We proceed to bound the two terms of (7.5) separately.

Bounding the first term of (7.5):

By Lemma B.1 in Appendix B.1 we know that

$$\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)] = h^{-d_X} \mathbb{E}\left[K\left(\frac{X - x}{h}\right) \middle| Z \in A_{\bar{j}}\right] (1 - \mathbb{P}(Z \in A_{\bar{j}}^c)^n).$$

So we have

$$\begin{aligned}
&\int\int |\widetilde{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \\
=&\int\int \left| \sum_{\bar{j}} \mathbb{1}(z \in A_{\bar{j}}) \left(\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)] - p_{X|Z}(x|z)\right) \right| p_Z(z) dx dz \\
\leq&\sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left| \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)] - p_{X|Z}(x|z) \right| p_Z(z) dx dz \\
=&\sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left| h^{-d_X} \mathbb{E}\left[K\left(\frac{X - x}{h}\right) \middle| Z \in A_{\bar{j}}\right] (1 - \mathbb{P}(Z \in A_{\bar{j}}^c)^n) \right. \\
&\qquad \left. - p_{X|Z}(x|z)(1 - \mathbb{P}(Z \in A_{\bar{j}}^c)^n) - p_{X|Z}(x|z) \mathbb{P}(Z \in A_{\bar{j}}^c)^n \right| p_Z(z) dx dz
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left| (1 - \mathbb{P}(Z \in A_{\bar{j}}^c)^n) \left(h^{-d_X} \mathbb{E}\left[K\left(\frac{X-x}{h} \right) \middle| Z \in A_{\bar{j}} \right] - p_{X|Z}(x|z) \right) \right| p_Z(z) dx dz \\
&\quad + \sum_{\bar{j}} \int_{A_{\bar{j}}} \int p_{X|Z}(x|z) \mathbb{P}(Z \in A_{\bar{j}}^c)^n p_Z(z) dx dz \\
&\leq \sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left| h^{-d_X} \mathbb{E}\left[K\left(\frac{X-x}{h} \right) \middle| Z \in A_{\bar{j}} \right] - p_{X|Z}(x|z) \right| p_Z(z) dx dz \\
&\quad + \sum_{\bar{j}} \int_{A_{\bar{j}}} \int p_{X|Z}(x|z) \mathbb{P}(Z \in A_{\bar{j}}^c)^n p_Z(z) dx dz. \qquad (7.6)
\end{aligned}
$$

We consider the two terms separately. To upper bound the first term, we know from Lemma B.2 in Appendix B.1 that

$$
\left| h^{-d_X} \mathbb{E}\left[K\left(\frac{X-x}{h} \right) \middle| Z \in A_{\bar{j}} \right] - p_{X|Z}(x|z \in A_{\bar{j}}) \right| \leq C h^{\beta},
$$

for some constant C. Then applying the triangle inequality we have

$$
\begin{aligned}
&\sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left| h^{-d_X} \mathbb{E}\left[K\left(\frac{X-x}{h} \right) \middle| Z \in A_{\bar{j}} \right] - p_{X|Z}(x|z) \right| p_Z(z) dx dz \\
\leq &\sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left(\left| p_{X|Z}(x|z) - p_{X|Z}(x|z \in A_{\bar{j}}) \right| \right. \\
&\qquad \left. + \left| h^{-d_X} \mathbb{E}\left[K\left(\frac{X-x}{h} \right) \middle| Z \in A_{\bar{j}} \right] - p_{X|Z}(x|z \in A_{\bar{j}}) \right| \right) p_Z(z) dx dz \\
\leq &\sum_{\bar{j}} \int_{A_{\bar{j}}} \int \left| p_{X|Z}(x|z) - \int_{A_{\bar{j}}} p_{X|Z}(x|z') \frac{p_Z(z')}{\mathbb{P}(Z \in A_{\bar{j}})} dz' \right| p_Z(z) dx dz + C h^{\beta} \\
\leq &\sum_{\bar{j}} \int_{A_{\bar{j}}} \int_{A_{\bar{j}}} \int \left| p_{X|Z}(x|z) - p_{X|Z}(x|z') \right| dx p_Z(z) \frac{p_Z(z')}{\mathbb{P}(Z \in A_{\bar{j}})} dz' dz + C h^{\beta} \\
= &\sum_{\bar{j}} \int_{A_{\bar{j}}} \int_{A_{\bar{j}}} \| p_{X|Z=z} - p_{X|Z=\boldsymbol{z'}} \|_1 p_Z(z) \frac{p_Z(z')}{\mathbb{P}(Z \in A_{\bar{j}})} dz' dz + C h^{\beta} \\
\leq &\sum_{\bar{j}} \int_{A_{\bar{j}}} \int_{A_{\bar{j}}} W_2 \| z - z' \|_1^{\gamma} p_Z(z) \frac{p_Z(z')}{\mathbb{P}(Z \in A_{\bar{j}})} dz' dz + C h^{\beta} \\
\leq &\sum_{\bar{j}} \int_{A_{\bar{j}}} \int_{A_{\bar{j}}} W_2 \left(\frac{d_Z}{m} \right)^{\gamma} p_Z(z) \frac{p_Z(z')}{\mathbb{P}(Z \in A_{\bar{j}})} dz' dz + C h^{\beta} \\
= &\sum_{\bar{j}} \int_{A_{\bar{j}}} W_2 \left(\frac{d_Z}{m} \right)^{\gamma} p_Z(z) dz + C h^{\beta} \\
= &W_2 \left(\frac{d_Z}{m} \right)^{\gamma} + C h^{\beta},
\end{aligned}
$$

where W_2 and C are constants.

Now we upper bound the second term in equation (7.6). Notice that it reduces to:

$$\sum_{\bar{j}} \int_{A_{\bar{j}}} \int p_{X|Z}(x|z)\mathbb{P}(Z \in A_{\bar{j}}^c)^n p_Z(z)dxdz = \sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n,$$

where $p_{\bar{j}} = \mathbb{P}(Z \in A_{\bar{j}})$.

Lemma 7.2. *We have*

$$\sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n \leq \left(1 - \frac{1}{m^{d_Z}}\right)^n. \tag{7.7}$$

The proof of Lemma 7.2 relies on Lagrange multipliers and is deferred to Appendix B.1. Finally we have an upper bound for the entirety of the first term of (7.5):

$$\int\int |\widetilde{p}_{X|Z}(x|z) - p_{X|Z}(x|z)|p_Z(z)dxdz \leq W_2\left(\frac{d_Z}{m}\right)^\gamma + Ch^\beta + \left(1 - \frac{1}{m^{d_Z}}\right)^n.$$

Bounding the second term of (7.5):

Recall that both $\widetilde{p}_{X|Z}(x|z)$ and $\widehat{p}_{X|Z}(x|z)$ are of the form $\sum_{\bar{j}} \mathbb{1}\left(z \in A_{\bar{j}}\right) M$ where the first part $\sum_{\bar{j}} \mathbb{1}\left(z \in A_{\bar{j}}\right)$ depends on z while the second part M is independent of z. Then we can rewrite the integral as

$$\begin{aligned}
&\mathbb{E}\int\int |\widehat{p}_{X|Z}(x|z) - \widetilde{p}_{X|Z}(x|z)|p_Z(z)dxdz \\
\leq &\mathbb{E}\left[\sum_{\bar{j}} \int\int \mathbb{1}\left(z \in A_{\bar{j}}\right) \left|\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right| p_Z(z)dxdz\right] \\
= &\mathbb{E}\left[\sum_{\bar{j}} \int_{A_{\bar{j}}} p_Z(z)dz \int \left|\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right| dx\right] \\
= &\mathbb{E}\left[\sum_{\bar{j}} \int \mathbb{P}(Z \in A_{\bar{j}}) \left|\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right| dx\right] \\
= &\sum_{\bar{j}} \int \mathbb{P}(Z \in A_{\bar{j}}) \mathbb{E}\left[\left|\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right|\right] dx.
\end{aligned}$$

We can further bound this term by first bounding the inner expression. By Jensen's inequality we have

$$\mathbb{E}\left[\left|\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right|\right] \leq \sqrt{\mathbb{E}\left[\left(\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right)^2\right]} = \sqrt{\mathrm{var}[\widehat{p}_{X,\bar{j}}(x)]}.$$

But by Lemma B.3 in Appendix B.1 we know this variance is upper bounded as,

$$\mathrm{var}[\widehat{p}_{X,\bar{j}}(x)] \leq \frac{A}{nh^{d_X}\mathbb{P}(Z \in A_{\bar{j}})} + B\mathbb{P}(Z \in A_{\bar{j}}^c)^n,$$

for some constants A, B. Notice that this bound is independent of x, so substituting back into the second term we have

$$\begin{aligned}
&\sum_{\bar{j}} \int \mathbb{P}(Z \in A_{\bar{j}}) \mathbb{E}\left[\left|\widehat{p}_{X,\bar{j}}(x) - \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)]\right|\right] dx \\
\leq & K \sum_{\bar{j}} \int \mathbb{P}(Z \in A_{\bar{j}}) \sqrt{\frac{1}{nh^{d_X}\mathbb{P}(Z \in A_{\bar{j}})} + \mathbb{P}(Z \in A_{\bar{j}}^c)^n} dx \\
= & K \sum_{\bar{j}} \sqrt{\mathbb{P}(Z \in A_{\bar{j}})} \sqrt{\frac{1}{nh^{d_X}} + \mathbb{P}(Z \in A_{\bar{j}}) \mathbb{P}(Z \in A_{\bar{j}}^c)^n},
\end{aligned}$$

where K is a constant. Once again denote $p_{\bar{j}} = \mathbb{P}(Z \in A_{\bar{j}})$. Then by Cauchy-Schwarz we have

$$\begin{aligned}
K \sum_{\bar{j}} \sqrt{p_{\bar{j}}} \sqrt{\frac{1}{nh^{d_X}} + p_{\bar{j}}(1-p_{\bar{j}})^n} &\leq K \sqrt{\left(\sum_{\bar{j}} p_{\bar{j}}\right)\left(\sum_{\bar{j}} \frac{1}{nh^{d_X}} + \sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n\right)} \\
&= K \sqrt{\frac{m^{d_Z}}{nh^{d_X}} + \sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n} \\
&\leq K \sqrt{\frac{m^{d_Z}}{nh^{d_X}}} + K\left(1 - \frac{1}{m^{d_Z}}\right)^{\frac{n}{2}},
\end{aligned}$$

where the last step follows from (7.7), since we already proved that $\sum_{\bar{j}} p_{\bar{j}}\left(1-p_{\bar{j}}\right)^n \leq (1 - \frac{1}{m^{d_Z}})^n$.

So we have shown that the second term is upper bounded as

$$\mathbb{E} \int \int |\widehat{p}_{X|Z}(x|z) - \widetilde{p}_{X|Z}(x|z)| p_Z(z) dx dz \leq K \sqrt{\frac{m^{d_Z}}{nh^{d_X}}} + K\left(1 - \frac{1}{m^{d_Z}}\right)^{\frac{n}{2}}.$$

Combining the terms:

Combining the bounds for the two terms, we have found an upper bound to the loss function:

$$\begin{aligned}
&\mathbb{E} \int \int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)| p_Z(z) dx dz \\
&\quad \leq W_2 \left(\frac{d_Z}{m}\right)^{\gamma} + Ch^{\beta} + \left(1 - \frac{1}{m^{d_Z}}\right)^n + K \sqrt{\frac{m^{d_Z}}{nh^{d_X}}} + K\left(1 - \frac{1}{m^{d_Z}}\right)^{\frac{n}{2}},
\end{aligned}$$

where W_2, C and K are constants. This yields the optimal parameter values $h \asymp n^{\frac{-1}{d_X + d_Z \beta\gamma^{-1} + 2\beta}}$ and $m \asymp n^{\frac{\beta}{d_X + d_Z \beta\gamma^{-1} + 2\beta}}$, and a rate of $n^{\frac{-1}{\beta^{-1} d_X + \gamma^{-1} d_Z + 2}}$. Notice that by our selection of h, m, we have $(1 - \frac{1}{m^{d_Z}})^n \leq (1 - \frac{1}{m^{d_Z}})^{\frac{n}{2}} \leq \exp(\frac{-n}{2m^{d_Z}})$, both of which are negligible compared to other terms for sufficiently large n, and thus can be ignored.

7.3 Proof of Theorem 4.1

We first choose a "bump" function h supported on $[0,1]$ which is an infinitely differentiable function, and satisfies the conditions that $\int h(x)dx = 0$, $\int h^2(x)dx = 1$, and for which $\int |h(x)|dx$ is a non-zero constant.

We construct a collection of densities by setting the conditional distributions $p_{X|Z}$ to be the uniform density perturbed by bumps of an appropriate resolution. Formally, we bin $[0,1]$ into m bins for Z and r bins for X. We then define the following conditional density functions:

$$p^{\Delta}_{X|Z}(x|z) = 1 + \sum_{\bar{i}} \sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{i_k,r}(x_k) \prod_{k\in[d_Z]} h_{j_k,m}(z_k),$$

where we recall the shorthands $\bar{i} \in [r]^{d_X}, \bar{j} \in [m]^{d_Z}$ (r, m are integers chosen later), and $\Delta_{\bar{i},\bar{j}} \in \{\pm 1\}$. We further define

$$\begin{aligned} h_{i_k,r}(x_k) &= \rho\sqrt{r}h(rx_k - i_k + 1), \\ h_{j_k,m}(z_k) &= \rho\sqrt{m}h(mz_k - j_k + 1), \end{aligned}$$

where ρ is a positive constant which we will choose appropriately. The support of $h_{i_k,r}(x_k)$ is $x_k \in [\frac{i_k-1}{r}, \frac{i_k}{r}]$, and the support of $h_{j_k,m}(z_k)$ is $z_k \in [\frac{j_k-1}{m}, \frac{j_k}{m}]$. The supports of these bumps are disjoint for different values of i_k or j_k.

The following lemma develops some important properties of the perturbed densities p^{Δ}.

Lemma 7.3. *1. Suppose we ensure that,*

$$\rho^d r^{d_X/2} m^{d_Z/2} \|h\|_\infty^d \le \frac{1}{2}, \tag{7.8}$$

then p^{Δ} is a valid density.

2. Suppose we ensure that,

$$2\rho^d r^{d_X/2} m^{d_Z/2} r^{\beta} \|h\|_\infty^{d_Z} \left(\left[2 \prod_{k\in[d_X]} \|h^{(\alpha_k)}\|_\infty \right] \vee \left[\sqrt{d_X} \prod_{k\in[d_X]} \left[\|h^{(\alpha_k+1)}\|_\infty \vee \|h^{(\alpha_k)}\|_\infty \right] \right] \right) \le W_1, \tag{7.9}$$

then p^{Δ} satisfies the Hölder smoothness condition.

3. Finally, if we ensure that,

$$2\rho^d r^{\frac{d_X}{2}} m^{\frac{d_Z}{2}} m^{\gamma} \|h\|_\infty^{d_Z - 1} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) \le W_2, \tag{7.10}$$

then p^{Δ} satisfies the TV smoothness condition.

Our lower bound will follow as a consequence of Fano's inequality [23]. In order to apply Fano's inequality it will be useful to bound the KL divergence between a pair of densities p^{Δ} and $p^{\Delta'}$, and to show that we can construct a collection of sufficiently large cardinality which are well-separated in the loss function (1.1).

Lemma 7.4. *1. Suppose that the condition in (7.8) holds. For a given pairs of densities p^{Δ} and $p^{\Delta'}$ the KL divergence can be bounded as,*

$$KL(p^{\Delta}, p^{\Delta'}) \leq 8\|h\|_{\infty}\rho^{2d}r^{d_X}m^{d_Z}.$$

2. There is a subset T of densities p^{Δ} such that, $|T| \geq 2^{r^{d_X}m^{d_Z}/8}$, and furthermore for any pair of densities p^{Δ} and $p^{\Delta'}$ in T there is an absolute constant $C > 0$ such that,

$$\int\int |p^{\Delta}_{X|Z}(x|z) - p^{\Delta'}_{X|Z}(x|z)|p_Z(z)dxdz \geq \frac{1}{4}\|h\|_1^d p_{min}\rho^d r^{d_X/2}m^{d_Z/2}.$$

Ignoring constants in the remainder of the proof we now describe our choice of (ρ, r, m). We select $\rho^d \asymp 1/\sqrt{n}$, $r \asymp n^{\beta^{-1}/(d_X/\beta+d_Z/\gamma+2)}$ and $m \asymp n^{\gamma^{-1}/(d_X/\beta+d_Z/\gamma+2)}$ (with appropriately small constants) and observe that each of the conditions of Lemma 7.3 are satisfied.

The proof of the theorem follows from a straightforward application of Fano's inequality (see for instance [22, Theorem 2.7]). In rough terms, we apply Fano's inequality to the collection of distributions T. Provided that we can show that the average pairwise KL divergence between the n-sample product distributions is at most some small constant times $\log|T|$ we obtain a lower bound on the loss (1.1) of any estimator which scales as $\rho^d r^{d_X/2}m^{d_Z/2} \asymp n^{-1/(d_X/\beta+d_Z/\gamma+1)}$. We note that,

$$\mathrm{KL}((p^{\Delta})^n, (p^{\Delta'})^n) = n\mathrm{KL}(p^{\Delta}, p^{\Delta'}) \lesssim n\rho^{2d}r^{d_X}m^{d_Z} \lesssim r^{d_X}m^{d_Z} \lesssim \log|T|,$$

as desired, completing the proof of the theorem.

7.3.1 Proof of Lemma 7.3

We prove each of the three claims in turn.

Proof of Claim (1): We now verify that $p^{\Delta}_{X|Z}(x|z)$ is a density function. Consider the integral

$$\begin{aligned}
&\int\int \left(1 + \sum_{\bar{i}}\sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{i_k,r}(x_k) \prod_{k\in[d_Z]} h_{j_k,m}(z_k)\right) dxdz \\
=&1 + \sum_{\bar{i}}\sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \int h_{i_1,r}(x_1)dx_1 ... \int h_{j_{d_Z},m}(z_{d_Z})dz_{d_Z} \\
=&1 + \sum_{\bar{i}}\sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \left(\rho r^{-\frac{1}{2}} \int h(u)du\right) ... \left(\rho m^{-\frac{1}{2}} \int h(u)du\right) \\
=&1.
\end{aligned}$$

And under the additional assumption

$$|\Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{i_k,r}(x_k) \prod_{k\in[d_Z]} h_{j_k,m}(z_k)| \leq \rho^d r^{d_X/2}m^{d_Z/2}\|h\|_{\infty}^d \leq \frac{1}{2},$$

the function $p^{\Delta}_{X|Z}(x|z)$ is always positive. So it is indeed a density function.

Proof of Claim (2): Now we verify that $p^{\Delta}_{X|Z}(x|z)$ indeed satisfies the Hölder smoothness assumption. Since the L_2 norm is always smaller than or equal to the L_1 norm, it suffices to show

$$|D^{\alpha}p^{\Delta}_{X|Z}(x|z) - D^{\alpha}p^{\Delta}_{X|Z}(x'|z)| \leq W_1\|x-x'\|_2^{\beta-\ell}.$$

$p^{\Delta}_{X|Z}(x|z)$ is infinitely differentiable since h is infinitely differentiable, and so it is $\ell = \lfloor\beta\rfloor$ times differentiable. Now we want to show the second requirement. Without loss of generality let $(\alpha_1, ..., \alpha_{d_X})$ be fixed, and suppose we are given arbitrary x, x', z.

Since z is fixed and $\prod_{k\in[d_Z]} h_{j_k,m}(z_k)$ have disjoint support, the only non-zero one is $\prod_{k\in[d_Z]} h_{j_k^*,m}(z_k)$ for the bins $\bar{j}^* = (j_1^*, ..., j_{d_Z}^*)$.

So we have

$$\begin{aligned}
&|D^{\alpha}p^{\Delta}_{X|Z}(x|z) - D^{\alpha}p^{\Delta}_{X|Z}(x'|z)| \\
=&\Big|\sum_{\bar{i}}\sum_{\bar{j}}\Delta_{\bar{i},\bar{j}}\left(\prod_{k\in[d_X]} h^{(\alpha_k)}_{i_k,r}(x_k)\prod_{k\in[d_Z]} h_{j_k,m}(z_k) - \prod_{k\in[d_X]} h^{(\alpha_k)}_{i_k,r}(x'_k)\prod_{k\in[d_Z]} h_{j_k,m}(z_k)\right)\Big| \\
=&\rho^d r^{d_X/2} m^{d_Z/2} r^{\ell}\Big|\sum_{\bar{i}}\Delta_{\bar{i},\bar{j}^*}\Big(\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - i_k + 1)\prod_{k\in[d_Z]} h(mz_k - j_k^* + 1) \\
&\qquad\qquad - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - i_k + 1)\prod_{k\in[d_Z]} h(mz_k - j_k^* + 1)\Big)\Big| \\
\leq&\rho^d r^{d_X/2} m^{d_Z/2} r^{\ell}\|h\|_{\infty}^{d_Z}\Big|\sum_{\bar{i}}\Delta_{\bar{i},\bar{j}^*}\left(\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - i_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - i_k + 1)\right)\Big|.
\end{aligned}$$

Notice that in the above summation, there are at most two non-zero terms, as $\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - i_k + 1)$ have disjoint supports. Let $\bar{a} = (a_1, ..., a_{d_X})$ be such that $\forall k \in [d_x], x_k \in [\frac{a_k-1}{r}, \frac{a_k}{r}]$, and let $\bar{b} = (b_1, ..., b_{d_Z})$ be such that $\forall k \in [d_X], x'_k \in [\frac{b_k-1}{r}, \frac{b_k}{r}]$, which correspond to the two non-zero terms respectively. Then we have

$$\begin{aligned}
&|D^{\alpha}p^{\Delta}_{X|Z}(x|z) - D^{\alpha}p^{\Delta}_{X|Z}(x'|z)| \\
\leq&\rho^d r^{d_X/2} m^{d_Z/2} r^{\ell}\|h\|_{\infty}^{d_Z}\Big[|\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - a_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - a_k + 1)| \\
&\qquad + |\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - b_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - b_k + 1)|\Big].
\end{aligned}$$

We can further bound the term within the square brackets. We will find two upper bounds and use the minimum between the two. Firstly we have

$$\begin{aligned}
&|\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - a_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - a_k + 1)| \\
&\qquad + |\prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - b_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - b_k + 1)| \\
\leq&4\prod_{k\in[d_X]}\|h^{(\alpha_k)}\|_{\infty} := \mu_1.
\end{aligned}$$

Secondly, using the identity $|f(x)-f(x')| \le \sup_y \|\nabla f(y)\|_2 \|x-x'\|_2$, where ∇ is the gradient and we take $f(x) = \prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - i_k + 1)$, we have another upper bound. Here we have

$$
\begin{aligned}
&\sup_y \|\nabla f(y)\|_2 \\
=&\sup_y \sqrt{\sum_{l\in[d_X]} \left(r h^{(\alpha_l+1)}(rx_l - i_l + 1) \prod_{k\in[d_X],k\neq l} h^{(\alpha_k)}(rx_k - i_k + 1) \right)^2} \\
\le& r \sqrt{\sum_{l\in[d_X]} \left(\|h^{(\alpha_l+1)}\|_\infty \prod_{k\in[d_X],k\neq l} \|h^{(\alpha_k)}\|_\infty \right)^2} \\
\le& r\sqrt{d_X} \prod_{k\in[d_X]} \left[\|h^{(\alpha_k+1)}\|_\infty \vee \|h^{(\alpha_k)}\|_\infty \right].
\end{aligned}
$$

This identity gives us the upper bound

$$
\begin{aligned}
&| \prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - a_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - a_k + 1)| \\
&\quad + | \prod_{k\in[d_X]} h^{(\alpha_k)}(rx_k - b_k + 1) - \prod_{k\in[d_X]} h^{(\alpha_k)}(rx'_k - b_k + 1)| \\
\le& 2r\sqrt{d_X} \prod_{k\in[d_X]} \left[\|h^{(\alpha_k+1)}\|_\infty \vee \|h^{(\alpha_k)}\|_\infty \right] \|x-x'\|_2 := \mu_2 r \|x-x'\|_2.
\end{aligned}
$$

Taking the minimum of these two upper bounds gives a tighter upper bound. Let $\wedge$ denote the minimum between two terms and $\vee$ the maximum. Using the properties $(ab \wedge cd) \le (a\vee c)(b\wedge d), a,b,c,d>0$ and $(1\wedge u) \le u^\alpha$ for $u>0, 0<\alpha\le 1$, we have

$$
\begin{aligned}
&|D^\alpha p^\Delta_{X|Z}(x|z) - D^\alpha p^\Delta_{X|Z}(x'|z)| \\
\le& \rho^d r^{d_X/2} m^{d_Z/2} r^\ell \|h\|_\infty^{d_Z} \left[\mu_1 \wedge (r\mu_2 \|x-x'\|_2) \right] \\
\le& \rho^d r^{d_X/2} m^{d_Z/2} r^\ell \|h\|_\infty^{d_Z} (\mu_1 \vee \mu_2)(1 \wedge r\|x-x'\|_2) \\
\le& \rho^d r^{d_X/2} m^{d_Z/2} r^\ell \|h\|_\infty^{d_Z} (\mu_1 \vee \mu_2) r^{\beta-\ell} \|x-x'\|_2^{\beta-\ell} \\
\le& W_1 \|x-x'\|_2^{\beta-\ell},
\end{aligned} \tag{7.11}
$$

provided we ensure that $\rho^d r^{d_X/2} m^{d_Z/2} r^\beta (\mu_1 \vee \mu_2) \le W_1$, which is indeed the case. So $p^\Delta_{X|Z}(x|z)$ satisfies the Hölder smoothness condition.

Proof of Claim (3): Now we show that $p^\Delta_{X|Z}(x|z)$ also satisfies the TV smoothness assumption. We have

$$
\int |p^\Delta_{X|Z}(x|z) - p^\Delta_{X|Z}(x|z')| dx \le \int \sum_{\bar{i}} \sum_{\bar{j}} \left| \prod_{k\in[d_X]} h_{i_k,r}(x_k) \right| \left| \prod_{k\in[d_Z]} h_{j_k,m}(z_k) - \prod_{k\in[d_Z]} h_{j_k,m}(z'_k) \right| dx. \tag{7.12}
$$

Recall that $\prod_{k\in[d_Z]} h_{j_k,m}(z_k)$ have disjoint supports, so there are at most two non-zero terms within the summation. Suppose $\forall k \in [d_Z], z_k \in [\frac{j_k^*-1}{m}, \frac{j_k^*}{m}]$ for some specific j_k^* while $\forall k \in [d_Z], z'_k \in [\frac{j_k'^*-1}{m}, \frac{j_k'^*}{m}]$ for some specific $j_k'^*$, corresponding to the two non-zero terms. Then

$$\begin{aligned}
&\sum_{\bar{j}} \left| \prod_{k\in[d_Z]} h_{j_k,m}(z_k) - \prod_{k\in[d_Z]} h_{j_k,m}(z'_k) \right| \\
\leq &\left| \prod_{k\in[d_Z]} h_{j_k^*,m}(z_k) - \prod_{k\in[d_Z]} h_{j_k^*,m}(z'_k) \right| + \left| \prod_{k\in[d_Z]} h_{j_k'^*,m}(z_k) - \prod_{k\in[d_Z]} h_{j_k'^*,m}(z'_k) \right|.
\end{aligned}$$

We upper bound the first term and note that an identical upper bound holds for the second term. Using a similar approach to how we showed Hölder smoothness in (7.11), and by telescoping, we have

$$\begin{aligned}
&\left| \prod_{k\in[d_Z]} h_{j_k^*,m}(z_k) - \prod_{k\in[d_Z]} h_{j_k^*,m}(z'_k) \right| \\
\leq &\sum_{k\in[d_Z]} (\sqrt{m}\rho)^{d_Z-1} \|h\|_\infty^{d_Z-1} |h_{j_k^*,m}(z_k) - h_{j_k^*,m}(z'_k)| \\
\leq &\sum_{k\in[d_Z]} (\sqrt{m}\rho)^{d_Z-1} \|h\|_\infty^{d_Z-1} \rho\sqrt{m} \left(2\|h\|_\infty \wedge \|h'\|_\infty m|z_k - z'_k|\right) \\
\leq &\sum_{k\in[d_Z]} (\sqrt{m}\rho)^{d_Z-1} \|h\|_\infty^{d_Z-1} \rho\sqrt{m} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) \left(1 \wedge m|z_k - z'_k|\right) \\
\leq &\sum_{k\in[d_Z]} (\sqrt{m}\rho)^{d_Z-1} \|h\|_\infty^{d_Z-1} \rho\sqrt{m} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) m^\gamma |z_k - z'_k|^\gamma \\
\leq &\rho^{d_Z} m^{\frac{d_Z}{2}} m^\gamma \|h\|_\infty^{d_Z-1} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) \|z - z'\|_1^\gamma.
\end{aligned}$$

Substituting this result back in (7.12) gives

$$\begin{aligned}
&\int |p^\Delta_{X|Z}(x|z) - p^\Delta_{X|Z}(x|z')| dx \\
= &\sum_{\bar{j}} \left| \prod_{k\in[d_Z]} h_{j_k,m}(z_k) - \prod_{k\in[d_Z]} h_{j_k,m}(z'_k) \right| \sum_{\bar{i}} \int_0^1 \left| \prod_{k\in[d_X]} h_{i_k,r}(x_k) \right| dx \\
\leq &2\rho^{d_Z} m^{\frac{d_Z}{2}} m^\gamma \|h\|_\infty^{d_Z-1} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) \|z - z'\|_1^\gamma \sum_{\bar{i}} \rho^{d_X} r^{\frac{-d_X}{2}} \prod_{k\in[d_X]} \left(\int_0^1 |h(u)| du \right) \\
= &2\rho^d r^{\frac{d_X}{2}} m^{\frac{d_Z}{2}} m^\gamma \|h\|_\infty^{d_Z-1} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) \|z - z'\|_1^\gamma \\
\leq &W_2 \|z - z'\|_1^\gamma
\end{aligned}$$

provided we ensure that $2\rho^d r^{\frac{d_X}{2}} m^{\frac{d_Z}{2}} m^\gamma \|h\|_\infty^{d_Z-1} \left(2\|h\|_\infty \vee \|h'\|_\infty\right) \leq W_2$. This is indeed the case, and we obtain that $p^\Delta_{X|Z}(x|z)$ indeed satisfies the TV smoothness assumption.

7.3.2 Proof of Lemma 7.4

We prove each of the two claims in turn.

Proof of Claim (1): Recall that in constructing p^{Δ} and $p^{\Delta'}$ we do not perturb the marginal distribution over Z. As a consequence the KL divergence between p^{Δ} and $p^{\Delta'}$ can be written as:

$$d_{\mathrm{KL}}(p^{\Delta}, p^{\Delta'}) = \mathbb{E}_z d_{\mathrm{KL}}(p^{\Delta}_{X|Z}(x|z), p^{\Delta'}_{X|Z}(x|z)),$$

and we focus on upper bounding the KL divergence between the conditional densities.

$$\begin{aligned} d_{\mathrm{KL}}(p^{\Delta}_{X|Z}(x|z), p^{\Delta'}_{X|Z}(x|z)) \leq & d_{\chi^2}(p^{\Delta}_{X|Z}(x|z), p^{\Delta'}_{X|Z}(x|z)) \\ &= \int p^{\Delta'}_{X|Z}(x|z) \left(\frac{p^{\Delta}_{X|Z}(x|z)}{p^{\Delta'}_{X|Z}(x|z)} - 1 \right)^2 dx \\ &= \int \frac{(p^{\Delta}_{X|Z}(x|z) - p^{\Delta'}_{X|Z}(x|z))^2}{p^{\Delta'}_{X|Z}(x|z)} dx. \end{aligned}$$

Recall that by the condition in (7.8) we have that $|\Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{i_k,r}(x_k) \prod_{k\in[d_Z]} h_{j_k,m}(z_k)| \leq \rho^d r^{d_X/2} m^{d_Z/2} \|h\|_\infty^d \leq \frac{1}{2}$ which implies $p^{\Delta'}_{X|Z}(x|z) \geq \frac{1}{2}$. So we have

$$\begin{aligned} & d_{\mathrm{KL}}(p^{\Delta}_{X|Z}(x|z), p^{\Delta'}_{X|Z}(x|z)) \\ \leq & 2 \int \left(\sum_{\bar{i}} \sum_{\bar{j}} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}) \prod_{k\in[d_X]} h_{i_k,r}(x_k) \prod_{k\in[d_Z]} h_{j_k,m}(z_k) \right)^2 dx \\ \overset{\text{(i)}}{=} & 2 \sum_{\bar{i}} \sum_{\bar{j}} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}})^2 \prod_{k\in[d_Z]} h^2_{j_k,m}(z_k) \int \prod_{k\in[d_X]} h^2_{i_k,r}(x_k) dx, \\ \leq & 8\rho^{2d_X} r^{d_X} \sum_{\bar{j}} \prod_{k\in[d_Z]} h^2_{j_k,m}(z_k), \end{aligned}$$

where for (i) we note that the cross terms in expanding the square correspond to disjoint bumps and are 0. As a consequence we obtain that,

$$\begin{aligned} d_{\mathrm{KL}}(p^{\Delta}, p^{\Delta'}) &\leq 8\rho^{2d_X} r^{d_X} \int \sum_{\bar{j}} \prod_{k\in[d_Z]} h^2_{j_k,m}(z_k) p_Z(z) dz \\ &\leq 8\rho^{2d_X} r^{d_X} (\rho\sqrt{m})^{2d_Z} \|h\|_\infty \times \\ &\qquad \sum_{\bar{j}} \left[\prod_{k\in[d_Z]} \mathbb{1}\left(z_k \in [\frac{\bar{j}_k - 1}{m}, \frac{\bar{j}_k}{m}] \right) \right] p_Z\left([\frac{\bar{j}_1 - 1}{m}, \frac{\bar{j}_1}{m}] \times \cdots \times [\frac{\bar{j}_{d_Z} - 1}{m}, \frac{\bar{j}_{d_Z}}{m}] \right) \\ &= 8\|h\|_\infty \rho^{2d} r^{d_X} m^{d_Z}. \end{aligned}$$

Proof of Claim (2): Given that the marginal density of Z is lower bounded i.e. $p_Z(z) \geq p_{\min} > 0$, it suffices to instead ensure that for some absolute constant $C > 0$ we have that,

$$\int \int |p^{\Delta}_{X|Z}(x|z) - p^{\Delta'}_{X|Z}(x|z)| dx dz \geq C\rho^d r^{d_X/2} m^{d_Z/2}.$$

Using the Varshamov-Gilbert construction [22, Lemma 2.9] we know that there exist $N = 2^{r^{d_X} m^{d_Z}/8}$ vectors Δ on the hypercube $\{\pm 1\}^{r^{d_X} m^{d_Z}}$ such that $d_H(\Delta, \Delta') \geq r^{d_X} m^{d_Z}/8$ for each Δ, Δ' in that

set, where $d_H(\Delta, \Delta') = \frac{1}{2}\sum_{\bar{i}}\sum_{\bar{j}} |\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}|$ is the Hamming distance. Then

$$\begin{aligned}
&\int\int |p^{\Delta}_{X|Z}(x|z) - p^{\Delta'}_{X|Z}(x|z)|dxdz \\
=&\sum_{\bar{i}}\sum_{\bar{j}} |\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}| \int\int \left| \prod_{k\in[d_X]} h_{i_k,r}(x_k) \prod_{k\in[d_Z]} h_{j_k,m}(z_k) \right| dxdz \\
\geq&\frac{r^{d_X} m^{d_Z}}{4} \frac{\rho^d}{\sqrt{r^{d_X} m^{d_Z}}} \left(\int_0^1 |h(u)|du \right)^d \\
\geq&C\rho^d \sqrt{r^{d_X} m^{d_Z}}
\end{aligned}$$

as claimed, where $C = \|h\|_1^d/4$.

7.4 Proof of Theorem 6.1

We denote by Δ_1, Δ_2 the following,

$$\begin{aligned}
\Delta_1 &= \sup_{A\in\mathcal{A}} \left| \mathbb{P}_n(A) - \int_A p \right|, \\
\Delta_2 &= \sup_{A\in\mathcal{A}} \sup_{j\in[N]} \left| \int_A \widetilde{f}_j - \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \widehat{f}_j(x|Z_i)dx \right|.
\end{aligned}$$

The following lemma is a simple consequence of Hoeffding's inequality, and gives high-probability upper bounds on the above quantities:

Lemma 7.5. *With probability at least* $1-\delta$,

$$\begin{aligned}
\Delta_1 &\leq \sqrt{\frac{\log(2N/\delta)}{n}} \\
\Delta_2 &\leq \sqrt{\frac{3\log(2N/\delta)}{2n}}.
\end{aligned}$$

Taking this result as given we complete the proof, before returning to prove it. Let us denote by $\breve{f}$ the minimizer $\operatorname{argmin}_{\widehat{f}_j : j\in[N]} \int_z \|\widehat{f}_j - p(x|z)\|_1 p_Z(z)dz$, then we can write:

$$\int_z \|\psi(x|z) - p(x|z)\|_1 p_Z(z)dz \leq \underbrace{\int_z \|\psi(x|z) - \breve{f}(x|z)\|_1 p_Z(z)dz}_{T} + \int_z \|\breve{f}(x|z) - p(x|z)\|_1 p_Z(z)dz.$$

Abusing notation slightly in the remainder of the proof we identify $\breve{f}$ and ψ with their corresponding

oracle joint densities $\breve{f}(x|z)p_Z(z)$ and $\psi(x|z)p_Z(z)$. We note that,

$$\begin{aligned}
T &\leq 2 \sup_{A\in\mathcal{A}} \left| \int_A \psi - \int_A \breve{f} \right| \\
&\leq 2 \sup_{A\in\mathcal{A}} \left[\left| \int_A \psi - \mathbb{P}_n(A) \right| + \left| \int_A \breve{f} - \mathbb{P}_n(A) \right| \right] \\
&\leq 2 \sup_{A\in\mathcal{A}} \left[\left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \psi(x|Z_i)dx - \mathbb{P}_n(A) \right| + \left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \psi(x|Z_i)dx - \int_A \psi \right| \right. \\
&\left. + \left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \breve{f}(x|Z_i)dx - \mathbb{P}_n(A) \right| + \left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \breve{f}(x|Z_i)dx - \int_A \breve{f} \right| \right] \\
&\leq 4\Delta_2 + 4 \sup_{A\in\mathcal{A}} \left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \breve{f}(x|Z_i)dx - \mathbb{P}_n(A) \right|,
\end{aligned}$$

where in the final inequality we use the definition of the minimum distance estimator, and of Δ_2. We then note that,

$$\begin{aligned}
4 \sup_{A\in\mathcal{A}} \left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \breve{f}(x|Z_i)dx - \mathbb{P}_n(A) \right| &\leq 4\Delta_1 + 4 \sup_{A\in\mathcal{A}} \left| \frac{1}{n}\sum_{i=1}^n \int_{A^{Z_i}} \breve{f}(x|Z_i)dx - \int_A p \right| \\
&\leq 4\Delta_1 + 4\Delta_2 + 4 \sup_{A\in\mathcal{A}} \left| \breve{f}(A) - \int_A p \right| \\
&\leq 4\Delta_1 + 4\Delta_2 + 2 \int_z \|\breve{f}(x|z) - p(x|z)\|_1 p_Z(z)dz,
\end{aligned}$$

and putting these bounds together with the bounds in Lemma 7.5 we obtain our claimed result.

Proof of Lemma 7.5: We note that for fixed $A \in \mathcal{A}$ (and a fixed index $j \in [N]$) we are simply bounding the deviation of a sum of bounded (by 1) random variables from their mean. This is straightforward for the terms appearing in the definition of Δ_1. For the terms appearing in Δ_2 we note that,

$$\widetilde{f}_j(A) = \int_A \widetilde{f}(x|z)p_Z(z)dxdz = \mathbb{E}_Z \left[\int_{A^Z} f(x|Z)dx \right].$$

The result then follows by combining the Hoeffding bound with the union bound, noting that $\mathcal{A}$ has cardinality at most N^2.

7.5 Proof of Theorem 6.2

The proof is a straightforward application of Theorem 6.1. Let us denote $\mathcal{D}_1$ the half of the sample on which we construct our density estimates and $\mathcal{D}_2$ the half on which we run the selection procedure. By Theorem 6.1, setting $\delta = 1/n$, we obtain that conditioning on the first half of the sample, with probability at least $1 - 1/n$ we select ψ such that,

$$\int_z \|\psi(x|z) - p(x|z)\|_1 p_Z(z)dz \lesssim \min_{j\in[N]} \int_z \|\widehat{f}_j - p(x|z)\|_1 p_Z(z)dz + \sqrt{\frac{\log n}{n}}.$$

Let us denote by E the event on which this guarantee holds, and denote by $j^* \in [N]$ a density estimate constructed with (nearly) optimal choices of the tuning parameters. The expected error (the expectation taken over all samples), can be bounded as:

$$\mathbb{E}\left[\int_z \|\psi(x|z) - p(x|z)\|_1 p_Z(z)dz\right] \leq \mathbb{E}_{\mathcal{D}_1,\mathcal{D}_2}\left[\int_z \|\psi(x|z) - p(x|z)\|_1 p_Z(z)dz|E\right] + \frac{2}{n},$$

by noting that the error is always atmost 2 since both ψ and p are valid densities (and the L_1-loss is upper bounded by 2 for densities). Finally, we note that,

$$\begin{aligned}
\mathbb{E}\left[\int_z \|\psi(x|z) - p(x|z)\|_1 p_Z(z)dz\right] &\lesssim \mathbb{E}_{\mathcal{D}_1}\left[\min_{j\in[N]} \int_z \|\widehat{f}_j - p(x|z)\|_1 p_Z(z)dz + \sqrt{\frac{\log n}{n}}\right] + \frac{2}{n}, \\
&\lesssim \mathbb{E}_{\mathcal{D}_1}\left[\int_z \|\widehat{f}_{j^*} - p(x|z)\|_1 p_Z(z)dz\right] + \sqrt{\frac{\log n}{n}}, \\
&\lesssim n^{\frac{-1}{\beta^{-1}d_X + \gamma^{-1}d_Z + 2}} + \sqrt{\frac{\log n}{n}} \\
&\lesssim n^{\frac{-1}{\beta^{-1}d_X + \gamma^{-1}d_Z + 2}},
\end{aligned}$$

where the last inequality follows by noting that for any finite β or γ the rate of conditional density estimation is strictly slower than $\mathcal{O}(\sqrt{(\log n)/n})$.

8 Discussion

In this paper we looked at the problem of conditional density estimation under a weighted absolute value loss function. We first demonstrated that if one imposes smoothness only on the conditional densities $p_{X|Z}(x|z)$ with respect to x, conditional density estimation is impossible in a minimax sense regardless of the marginal density p_Z (which may even be known to the statistician). We then derived the minimax rate of estimation and showed an adaptive estimator which achieves the rate without knowledge of the smoothness parameters.

An interesting question that we intend to investigate in our future work is to generalize our results to an L_p loss function:

$$\int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)|^p p_Z(z)dz,$$

for some $p \geq 1$. We anticipate that such a modification will require a smoothness assumption stronger than TV smoothness. It will be interesting to see whether one can show that TV smoothness is not sufficient to analyze the L_p loss function for $p > 1$. In addition we are interested in quantifying higher order TV smoothness and studying the problem of conditional density estimation for such densities.

9 Acknowledgements

The authors are grateful to Larry Wasserman for helpful discussions. SB was partially supported by NSF grants DMS-1713003 and CCF-1763734.

References

[1] Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *Journal of machine learning research*, 7(Jul):1231–1264, 2006.

[2] Damir Filipović, Lane P Hughston, and Andrea Macrina. Conditional density models for asset pricing. *International Journal of Theoretical and Applied Finance*, 15(01):1250002, 2012.

[3] David M Bashtannyk and Rob J Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298, 2001.

[4] Rafael Izbicki and Ann B Lee. Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4): 1297–1316, 2016.

[5] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media, 2001.

[6] Luc. Devroye and Lazlo. Gyorfi. *Nonparametric Density Estimation: The L1 View*. Wiley Interscience Series in Discrete Mathematics. Wiley, 1985.

[7] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, 47(4):1893 – 1927, 2019.

[8] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727 – 749, 2018.

[9] Sam Efromovich. Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6):2504–2535, 2007.

[10] Murray Rosenblatt. Conditional probability density and regression estimators. *Multivariate analysis II*, 25:31, 1969.

[11] Rob J Hyndman, David M Bashtannyk, and Gary K Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996.

[12] Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

[13] Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

[14] Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

[15] Peter Hall, Qiwei Yao, et al. Approximating conditional distribution functions using dimension reduction. *The Annals of statistics*, 33(3):1404–1421, 2005.

[16] Peter Hall, Rodney CL Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94(445):154–163, 1999.

[17] Sam Efromovich. Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association*, 105(490):761–774, 2010.

[18] Gaëlle Chagny. Warped bases for conditional density estimation. *Mathematical Methods of Statistics*, 22(4):253–282, 2013.

[19] Domagoj Ćevid, Loris Michel, Nicolai Meinshausen, and Peter Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *arXiv preprint arXiv:2005.14458*, 2020.

[20] Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.

[21] Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *arXiv preprint arXiv:2001.03039*, 2020.

[22] Alexandre B. Tsybakov. *Introduction To Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

[23] Bin Yu. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY, 1997.

[24] O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512 – 2546, 1997.

A Additional Technical Results

A.1 Proof of Lemma 7.1

We begin by verifying the first and second claims. Notice that p_1 belongs to the Hölder class with any smoothness value β. For the $\beta \leq 1$ cases, p_1 is Hölder smooth with constant $\frac{2(1-c)}{d_X^\beta}$; furthermore linear functions are Hölder smooth for any $\beta > 1$ and any $W_1 > 0$. Now we consider p_2. We note that $p_1 \leq 2$. It immediately follows that if we define $p_2 = 2 - p_1$ then p_2 is a valid density in the same Hölder class as p_1, and $\frac{1}{2}p_1 + \frac{1}{2}p_2 = U([0,1]^{d_X})$.

Finally we examine the TV distance

$$\mathrm{TV}(U([0,1]^{d_X}), p_1) = \mathrm{TV}(U([0,1]^{d_X}), p_2) = \frac{1}{2}\int \left| 1 - \left(\alpha \sum_{i \in [d_x]} \frac{x_i}{d_X} + c \right) \right| dx.$$

We will lower bound this from below:

$$\begin{aligned}
\mathrm{TV}(U([0,1]^{d_X}), p_1) &\geq \frac{1}{2}\int 3 \left(\frac{\left| 1 - \left(\alpha \left(\sum_{i \in [d_x]} \frac{x_i}{d_X} \right) + c \right) \right|}{3} \right)^2 dx \\
&= \frac{1}{6}\int \left(1 - 2\alpha \sum_{i \in [d_x]} \frac{x_i}{d_X} - 2c + \alpha^2 \left(\sum_{i \in [d_x]} \frac{x_i}{d_X} \right)^2 + 2\alpha c \sum_{i \in [d_x]} \frac{x_i}{d_X} + c^2 \right) dx \\
&= \frac{1}{6}\left((1-c)^2 - (1-c)\alpha + \alpha^2 (\frac{1}{3d_X} + \frac{d_X - 1}{4d_X}) \right),
\end{aligned}$$

where the first inequality holds since $\frac{\left| 1 - \left(\alpha \left(\sum_{i \in [d_x]} \frac{x_i}{d_X} \right) + c \right) \right|}{3} \leq 1$, and we used $\int \sum_{i \in [d_X]} \frac{x_i}{d_X} dx = \frac{1}{2}$ and $\int (\sum_{i \in [d_X]} \frac{x_i}{d_X})^2 dx = \frac{\int x_1^2 dx_1}{d_X} + \frac{d_X - 1}{d_X} \int x_1 dx_1 \int x_2 dx_2 = \frac{1}{3d_X} + \frac{d_X - 1}{4d_X}$. Substituting in $\alpha = 2(1-c)$ from previous observations we get

$$\begin{aligned}
\mathrm{TV}(U([0,1]^{d_X}), p_1) &\geq \frac{1}{6}\left((1-c)^2 - 2(1-c)^2 + 4(1-c)^2 (\frac{1}{3d_X} + \frac{d_X - 1}{4d_X}) \right) \\
&= \frac{1}{6}\left((1-c)^2 \left(\frac{4}{3d_X} + \frac{d_X - 1}{d_X} - 1 \right) \right) \\
&= \frac{(1-c)^2}{18 d_X} \\
&\geq 0.
\end{aligned}$$

B Proofs of Section 3

Proof of Lemma 3.4. It is easy to see that the construction in [22, Proposition 1.3] provides kernels satisfying

$$\int |K_i(u)| |u|^\kappa du < \infty,$$

for any fixed $\kappa \geq 0$. This is so since by construction $K_i(u)$ are Legendre polynomials supported on $[-1,1]$. In addition they also satisfy $\int K_i^2(u)du < \infty$. Furthermore since each kernel is of order ℓ we have that for any non-negative integer $k \leq \ell$: $\int K_i(u)u^k du = 0$. The statement of the lemma follows immediately by combining these three properties.

□

Proof of Lemma 3.6. Suppose that

$$\int |p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)|dx \leq \epsilon_n(z),$$

and $\widehat{p}_{X|Z=z} \not\equiv 0$. Now consider the set $S = \{x \mid \widehat{p}_{X|Z}(x|z) \geq 0\}$. Observe that

$$\int_S |p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)|dx \leq \int |p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)|dx \leq \epsilon_n(z).$$

Since on the set S^c we have $|p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)| = p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)$, and we know $p_{X|Z}(x|z) \geq 0$, we can conclude by the above inequality that

$$\int_{S^c} -\widehat{p}_{X|Z}(x|z)dx \leq \epsilon_n(z).$$

In addition, since $\int \widehat{p}_{X|Z}(x|z)dx = 1$, we have

$$C = \int (\widehat{p}_{X|Z}(x|z))_+ dx \geq 1 - \epsilon_n(z).$$

By the triangle inequality we also have

$$C \leq \int |\widehat{p}_{X|Z}(x|z)|dx \leq \int |\widehat{p}_{X|Z}(x|z) - p_{X|Z}(x|z)|dx + 1 \leq 1 + \epsilon_n(z).$$

Finally, we know the following holds

$$\int |p_{X|Z}(x|z) - (\widehat{p}_{X|Z}(x|z))_+|dx \leq \int |p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)|dx \leq \epsilon_n(z).$$

Combining the above results gives us

$$\begin{aligned}\int |p_{X|Z}(x|z) - C^{-1}(\widehat{p}_{X|Z}(x|z))_+|dx &\leq \int |p_{X|Z}(x|z) - (\widehat{p}_{X|Z}(x|z))_+|dx + \int \frac{|1-C|}{C}(\widehat{p}_{X|Z}(x|z))_+|dx \\ &\leq \epsilon_n(z) + |1-C| \\ &\leq 2\epsilon_n(z).\end{aligned}$$

Next, notice that when $\widehat{p}_{X|Z=z} \equiv 0$ then we have $\int |p_{X|Z}(x|z) - \widehat{p}_{X|Z}(x|z)|dx = 1$, whereas,

$$\int |p_{X|Z}(x|z) - \bar{p}_{X|Z}(x|z)|dx \leq 2,$$

hence the same bound as above applies. Finally integrating the bound over z completes the proof.

□

B.1 Lemmas of Section 7.2

Proof of Lemma 7.2. Let

$$\sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n$$

be the objective function that we try to maximize. We will use the Karush-Kuhn-Tucker (KKT) conditions and we will subject the objective to the constraints $p_{\bar{j}} \geq 0$ for all $\bar{j}$ and $\sum_{\bar{j}} p_{\bar{j}} = 1$. We introduce the KKT multipliers $\lambda_{\bar{j}} \leq 0$ for all $\bar{j}$, and μ corresponding to the constraints respectively. Then taking the derivative with respect to some $\bar{j}$ we have the conditions

$$(1-p_{\bar{j}})^n - np_{\bar{j}}(1-p_{\bar{j}})^{n-1} - \lambda_{\bar{j}}p_{\bar{j}} + \mu = 0$$

and by complementary slackness $\lambda_{\bar{j}}p_{\bar{j}} = 0$ for all $\bar{j}$.

Let $S = \{\bar{j} \mid p_{\bar{j}} \neq 0\}$, which means from the conditions that on the set S, all $\lambda_{\bar{j}} = 0$ and $(1-p_{\bar{j}})^n - np_{\bar{j}}(1-p_{\bar{j}})^{n-1} = -\mu$. We can write this as $f(x) = (1-(n+1)x)(1-x)^{n-1}$ where $x = p_{\bar{j}}$. Clearly f is decreasing on $[0, \frac{1}{n+1}]$ and $f(\frac{1}{n+1}) = 0$. Since $|S| \leq m^{d_Z} \ll n+1$ (by our construction of m) and $\sum_{\bar{j}} p_{\bar{j}} = 1$, there exists $\bar{k} \in S$ such that $p_{\bar{k}} \geq \frac{1}{|S|} \geq \frac{1}{m^{d_Z}} \gg \frac{1}{n+1}$. But $(1-(n+1)p_{\bar{k}})(1-p_{\bar{k}})^{n-1} < 0$, so it follows that $\mu > 0$. Now observe $f'(x) = n(1-x)^{n-2}((n+1)x-2)$. This shows that on the interval $[\frac{1}{n+1}, 1]$, f changes from decreasing to increasing exactly once, at the point $\frac{2}{n+1} > \frac{1}{n+1}$. This implies that the equations $f(x) = -\mu$ for some $\mu > 0$ and $x \in [\frac{1}{n+1}, 1]$ can have at most two solutions.

Then simply divide $S = S_1 \cup S_2$, where all $p_{\bar{j}}$ on S_1 are equal to some v, and all $p_{\bar{j}}$ on S_2 are equal to some w, such that $|S_1|v + |S_2|w = 1, v, w \in [\frac{1}{n+1}, 1]$ and $f(v) = f(w) > 0$. Substituting this in, our objective function becomes

$$\sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n = |S_1|v(1-v)^n + |S_2|w(1-w)^n.$$

But the function $(1-x)^n$ is convex, so by Jensen's inequality it follows that

$$|S_1|v(1-v)^n + |S_2|w(1-w)^n \leq (1-|S_1|v^2 - |S_2|w^2)^n,$$

which is maximized when we minimize $|S_1|v^2 + |S_2|w^2$ under the constraint $|S_1|v + |S_2|w = 1$. This is done by rearranging to get $v = \frac{1-|S_2|w}{|S_1|}$, so the minimum is achieved at $v = w = \frac{1}{|S_1|+|S_2|}$. This result shows that $p_{\bar{j}}$ must have the same value of $\frac{1}{|S|}$ for all $\bar{j}$, which is what one would intuitively expect. Our objective function then becomes

$$\sum_{\bar{j}} p_{\bar{j}}(1-p_{\bar{j}})^n = |S|\frac{1}{|S|}\left(1-\frac{1}{|S|}\right)^n = \left(1-\frac{1}{|S|}\right)^n.$$

In order to maximize this, we want $|S|$ to be as large as possible, which in this case is m^{d_Z}. This completes the proof. □

Lemma B.1. *Under the same assumptions as Theorem 3.5, the estimator*

$$\widehat{p}_{X,\bar{j}}(x) = \frac{\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})K(\frac{X_i - x}{h})}{h^{d_X}\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})}$$

for some $\bar{j} \in [m]^{d_Z}$ *has the expected value*

$$\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)] = h^{-d_X}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\middle| Z \in A_{\bar{j}}\right](1-\mathbb{P}(Z \in A_{\bar{j}}^c)^n).$$

Proof of Lemma B.1. Using the law of total expectation we have

$$\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)] = \sum_{S\in 2^{[n]}} \mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c]\mathbb{P}(Z \in A_{\bar{j}})^{|S|}\mathbb{P}(Z \in A_{\bar{j}}^c)^{n-|S|},$$

where the condition in the conditional expectation means that

$$\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})K\left(\frac{X_i-x}{h}\right) = \sum_{i\in S} K\left(\frac{X_i-x}{h}\right)$$

and that

$$\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) = |S|.$$

Then the conditional expectation can be rewritten as:

$$\begin{aligned}
&\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c] \\
&= \mathbb{E}\left[\frac{\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})K\left(\frac{X_i-x}{h}\right)}{h^{d_X}\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})}\middle| Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c\right] \\
&= \mathbb{E}\left[\frac{\sum_{i\in S} K\left(\frac{X_i-x}{h}\right)}{h^{d_X}|S|}\middle| Z_i \in A_{\bar{j}}, i \in S\right] \\
&= \frac{1}{h^{d_X}|S|}\sum_{i\in S}\mathbb{E}\left[K\left(\frac{X_i-x}{h}\right)\middle| Z_i \in A_{\bar{j}}\right] \\
&= h^{-d_X}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\middle| Z \in A_{\bar{j}}\right].
\end{aligned}$$

Now notice that

$$\sum_{S\in 2^{[n]}} \mathbb{P}(Z \in A_{\bar{j}})^{|S|}\mathbb{P}(Z \in A_{\bar{j}}^c)^{n-|S|} = 1$$

and there is one special case when S is the empty set $\emptyset$, where by definition the estimator $\widehat{p}_{X,\bar{j}}(x) = \frac{0}{0} := 0$. This occurs when $|S| = 0$ with a corresponding probability of $\mathbb{P}(Z \in A_{\bar{j}}^c)^n$. So we subtract this probability, giving:

$$\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)] = h^{-d_X}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)|Z \in A_{\bar{j}}\right](1-\mathbb{P}(Z \in A_{\bar{j}}^c)^n)$$

as desired. □

Lemma B.2. *Under the same assumptions as Theorem 3.5, for some* $\bar{j} \in [m]^{d_Z}$ *we have that:*

$$\left|h^{-d_X}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)|Z \in A_{\bar{j}}\right] - p_{X|Z}(x|z \in A_{\bar{j}})\right| \leq Ch^{\beta}$$

for some constant C.

Proof of Lemma B.2. By assumption that K is a kernel of order ℓ, it follows that:

$$\begin{aligned}
&\left|h^{-d_X}\int K\left(\frac{y-x}{h}\right)p_{X|Z}(y|z\in A_{\bar{j}})dy - p_{X|Z}(x|z\in A_{\bar{j}})\right| \\
=&\left|\int K(u)p_{X|Z}(x+uh|z\in A_{\bar{j}})du - p_{X|Z}(x|z\in A_{\bar{j}})\right| \\
=&\left|\int K(u)\big[p_{X|Z}(x+uh|z\in A_{\bar{j}}) - p_{X|Z}(x|z\in A_{\bar{j}})\big]du\right|.
\end{aligned}$$

But by Lemma B.5 $p_{X|Z}(x+uh|z\in A_{\bar{j}})$ is ℓ times differentiable. Then the Taylor series is:

$$p_{X|Z}(x+uh|z\in A_{\bar{j}}) = \sum_{\|\alpha\|_1<\ell}\frac{D^\alpha p_{X|Z}(x|z\in A_{\bar{j}})}{\alpha!}(uh)^\alpha + \sum_{\|\alpha\|_1=\ell}\frac{D^\alpha p_{X|Z}(x+\tau uh|z\in A_{\bar{j}})}{\alpha!}u^\alpha h^\ell,$$

where $\|\alpha\|_1=\ell$ and $\tau\in[0,1]$. Substituting this back in cancels out the first summation and $p_{X|Z}(x|z\in A_{\bar{j}})$ (since the kernel is of order ℓ), giving:

$$\begin{aligned}
&\left|\int K(u)[p_{X|Z}(x+uh|z\in A_{\bar{j}}) - p_{X|Z}(x|z\in A_{\bar{j}})]du\right| \\
=&\left|\int K(u)\sum_{\|\alpha\|_1=\ell}\frac{D^\alpha p_{X|Z}(x+\tau uh|z\in A_{\bar{j}})}{\alpha!}u^\alpha h^\ell du\right| \\
=&\left|\int K(u)\sum_{\|\alpha\|_1=\ell}\frac{D^\alpha p_{X|Z}(x+\tau uh|z\in A_{\bar{j}})}{\alpha!}u^\alpha h^\ell du - \int K(u)\sum_{\|\alpha\|_1=\ell}\frac{D^\alpha p_{X|Z}(x|z\in A_{\bar{j}})}{\alpha!}u^\alpha h^\ell du\right| \\
\le&\int |K(u)|\frac{|u^\alpha h^\ell|}{\alpha!}\sum_{\|\alpha\|_1=\ell}\left|D^\alpha p_{X|Z}(x+\tau uh|z\in A_{\bar{j}}) - D^\alpha p_{X|Z}(x|z\in A_{\bar{j}})\right|du \\
\le&\int |K(u)|\frac{|u^\alpha h^\ell|}{\alpha!}\sum_{\|\alpha\|_1=\ell}W_1\|\tau uh\|_1^{\beta-\ell}du \\
=&h^\beta|\tau|^{\beta-\ell}\frac{W_1}{\alpha!}\sum_{\|\alpha\|_1=\ell}\int |K(u)|\cdot|u^\alpha|\cdot\|u\|_1^{\beta-\ell}du.
\end{aligned}$$

In the above, we used the fact that $p_{X|Z}(x|z\in A_{\bar{j}})$ also follows the Hölder smoothness condition because of Lemma B.5.

Since in $\mathbb{R}^d$ all norms are equivalent, we know that for any $q\ge 1$, there exists C_q such that $(\sum_i |y_i|^q)^{\frac{1}{q}} \le C_q\sum_i|y_i| = \|y\|_1$. Now let $y_i = |u_i|^q$ where $q=(\beta-\ell)^{-1}$. It follows that

$$\|u\|_1^{\beta-\ell} = \left(\sum_i |u_i|\right)^{\beta-\ell} \le C_{(\beta-\ell)^{-1}}\sum_i |u_i|^{\beta-\ell}$$

for some constant $C_{\beta-\ell}$. Then

$$\sum_{\|\alpha\|_1=\ell}|u^\alpha|\cdot\|u\|_1^{\beta-\ell} \le C_{(\beta-\ell)^{-1}}\sum_{\|\alpha\|_1=\ell}\sum_i |u^\alpha||u_i|^{\beta-\ell} \lesssim \sum_{\|\alpha\|_1=\beta}|u^\alpha|,$$

where in the last inequality the constant may depend on $\beta - \ell$, and the dimension d_X. Since we are assuming $\int |K(u)||u^\alpha| du < \infty$ for all $\|\alpha\|_1 \le \beta, \alpha \in \mathbb{R}_+^{d_X}$, this means that

$$\left| \int K(u)[p_{X|Z}(x + uh|z \in A_{\bar{j}}) - p_{X|Z}(x|z \in A_{\bar{j}})]du \right| \le Ch^\beta |\tau|^{\beta-\ell} \le Ch^\beta.$$

□

Lemma B.3. *Under the same assumptions as Theorem 3.5, the estimator*

$$\widehat{p}_{X,\bar{j}}(x) = \frac{\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) K(\frac{X_i - x}{h})}{h^{d_X} \sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}})}$$

for some $\bar{j} \in [m]^{d_Z}$ has its variance upper bounded as

$$\mathrm{var}[\widehat{p}_{X,\bar{j}}(x)] \le \frac{C}{nh^{d_X}\mathbb{P}(Z \in A_{\bar{j}})} + K\mathbb{P}(Z \in A_{\bar{j}}^c)^n.$$

where C, K are constants.

Proof of Lemma B.3. By the law of total variance we have:

$$\begin{aligned}\mathrm{var}(\widehat{p}_{X,\bar{j}}(x)) =& \mathbb{E}[\mathrm{var}(\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c)] \\ &+ \mathrm{var}_S[\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c]]. \end{aligned} \tag{B.1}$$

We proceed to bound the two terms separately.

Bounding the first term

The conditions in the expectation means that

$$\sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) K\left(\frac{X_i - x}{h}\right) = \sum_{i\in S} K\left(\frac{X_i - x}{h}\right), \quad \sum_{i\in[n]} \mathbb{1}(Z_i \in A_{\bar{j}}) = |S|.$$

We first consider the variance term inside the expectation:

$$\begin{aligned}&\mathrm{var}\left(\widehat{p}_{X,\bar{j}}(x)\middle| Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c\right) \\ =& \mathrm{var}\left(\frac{\sum_{i\in S} K(\frac{X_i - x}{h})}{h^{d_X}|S|}\middle| Z_i \in A_{\bar{j}}, i \in S\right) \\ =& \frac{1}{h^{2d_X}|S|^2} \sum_{i\in S} \mathrm{var}\left(K\left(\frac{X_i - x}{h}\right)\middle| Z_i \in A_{\bar{j}}\right) \\ \le& \frac{1}{h^{2d_X}|S|^2} \sum_{i\in S} \mathbb{E}\left[K^2\left(\frac{X_i - x}{h}\right)\middle| Z_i \in A_{\bar{j}}\right].\end{aligned}$$

Then when $S \neq \emptyset$ we have

$$\begin{aligned}
&\operatorname{var}(\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c) \\
\leq &\frac{1}{h^{2d_X}|S|}\mathbb{E}\left[K^2\left(\frac{X-x}{h}\right)\middle|Z \in A_{\bar{j}}\right] \\
\leq &\frac{2}{h^{2d_X}(|S|+1)}\mathbb{E}\left[K^2\left(\frac{X-x}{h}\right)\middle|Z \in A_{\bar{j}}\right],
\end{aligned}$$

where the last step makes a small sacrifice in the tightness of the bound in order to allow a very useful identity to be applied later. As for the other case when $S = \emptyset$, by definition the estimator $\widehat{p}_{X,\bar{j}}(x) = 0$, so the above still holds.

We proceed to bound the expectation $\mathbb{E}\big[K^2\big(\frac{X-x}{h}\big)\big|Z \in A_{\bar{j}}\big]$. First consider K^*, a bounded kernel of order ℓ, not necessarily equal to K. Now notice that $p_{X|Z}(x|z \in A_{\bar{j}}) \leq p_{\max} < \infty$. This can be proven by applying lemma B.2 and setting $h = 1$ to get

$$\left|\int K^*(y-x)\, p_{X|Z}(y|z \in A_{\bar{j}})dy - p_{X|Z}(x|z \in A_{\bar{j}})\right| \leq C.$$

It follows that

$$p_{X|Z}(x|z \in A_{\bar{j}}) \leq C + \int |K^*(y-x)\,|p_{X|Z}(y|z \in A_{\bar{j}})dy \leq C + K^*_{\max} < \infty,$$

where $K^*_{\max} = \sup_{u \in \mathbb{R}^{d_X}} |K^*(u)|$.

Now we have

$$\mathbb{E}\left[K^2\left(\frac{X-x}{h}\right)\middle|Z \in A_{\bar{j}}\right] = \int K^2\left(\frac{y-x}{h}\right)p_{X|Z}(y|z \in A_{\bar{j}})dy \leq p_{\max}h^{d_X}\int K^2(u)du \leq Ch^{d_X}, \tag{B.2}$$

where $C \geq p_{\max}\int K^2(u)du$ is a constant, since both the conditional density and $\int K^2(u)du$ are upper bounded.

Substituting this back into the first term of (B.1) we get

$$\begin{aligned}
&\mathbb{E}[\operatorname{var}(\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c)] \\
\leq &\mathbb{E}\left[\frac{2}{h^{2d_X}(|S|+1)}Ch^{d_X}\right] \\
= &\frac{2C}{h^{d_X}}\mathbb{E}\left[\frac{1}{|S|+1}\right].
\end{aligned}$$

By Lemma B.4 we have

$$\mathbb{E}\left[\frac{1}{|S|+1}\right] = \frac{1-\mathbb{P}(Z \in A_{\bar{j}}^c)^{n+1}}{(n+1)\mathbb{P}(Z \in A_{\bar{j}})} \leq \frac{1}{(n+1)\mathbb{P}(Z \in A_{\bar{j}})}.$$

Then the first term is bounded by

$$\mathbb{E}[\operatorname{var}(\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c)] \leq \frac{2C}{h^{d_X}(n+1)\mathbb{P}(Z \in A_{\bar{j}})} \asymp \frac{C}{nh^{d_X}\mathbb{P}(Z \in A_{\bar{j}})}$$

for some constant C.

Bounding the second term

Now we bound the second term from equation (B.1), and start by examining the inner expectation. When $S \neq \varnothing$

$$\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c] = h^{-d_X}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\middle|Z \in A_{\bar{j}}\right]$$

and otherwise when $S = \varnothing$

$$\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c] = 0.$$

Then

$$\begin{aligned}
&\text{var}_S[\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c]] \\
&= \left(h^{-d_X}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\middle|Z \in A_{\bar{j}}\right]\right)^2 \mathbb{P}(Z \in A_{\bar{j}}^c)^n(1-\mathbb{P}(Z \in A_{\bar{j}}^c)^n).
\end{aligned}$$

Since $\int K^2(u)du < \infty$ is upper bounded, it follows that $\int |K(u)|du \leq \sqrt{\int K^2(u)du} < \infty$. Therefore by the same logic as how we arrived at the bound in equation (B.2), we have $\mathbb{E}[K(\frac{X-x}{h})|Z \in A_{\bar{j}}] \leq C'h^{d_X}$ for some constant C'. Then we can simply bound the whole expression as:

$$\begin{aligned}
&\text{var}_S[\mathbb{E}[\widehat{p}_{X,\bar{j}}(x)|Z_i \in A_{\bar{j}}, i \in S, Z_i \in A_{\bar{j}}^c, i \in S^c]] \\
\leq& C'^2\mathbb{P}(Z \in A_{\bar{j}}^c)^n(1-\mathbb{P}(Z \in A_{\bar{j}}^c)^n) \\
\leq& K\mathbb{P}(Z \in A_{\bar{j}}^c)^n
\end{aligned}$$

for some constant K.

Combining the terms

For the variance (B.1) we have split it into two terms and upper bounded them individually. It follows that

$$\text{var}[\widehat{p}_{X,\bar{j}}(x)] \leq \frac{C}{nh^{d_X}\mathbb{P}(Z \in A_{\bar{j}})} + K\mathbb{P}(Z \in A_{\bar{j}}^c)^n$$

for some constants C, K as desired. □

Lemma B.4. *We have the identity*

$$\mathbb{E}\left[\frac{1}{|S|+1}\right] = \frac{1-\mathbb{P}(Z \in A_{\bar{j}}^c)^{n+1}}{(n+1)\mathbb{P}(Z \in A_{\bar{j}})}.$$

Proof. By definition of $|S|$, it can be regarded as a binomial distribution $|S| \sim Bin(n, p)$ where $p = \mathbb{P}(Z \in A_{\bar{j}})$. We also set $q = 1 - p = \mathbb{P}(Z \in A_{\bar{j}}^c)$. Then

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{|S|+1}\right] &= \sum_{k=0}^{n} \frac{1}{k+1}\mathbb{P}(|S| = k) \\
&= \sum_{k=0}^{n} \frac{1}{k+1}\frac{n!}{k!(n-k)!}p^k q^{n-k} \\
&= \frac{1}{(n+1)p}\sum_{k=0}^{n}\binom{n+1}{k+1}p^{k+1}q^{n-k} \\
&= \frac{1}{(n+1)p}[(\sum_{t=0}^{n+1}\binom{n+1}{t}p^t q^{(n+1)-t}) - q^{n+1}] \\
&= \frac{1}{(n+1)p}[(p+q)^{n+1} - q^{n+1}] \quad \text{by the binomial theorem} \\
&= \frac{1}{(n+1)p}[1 - q^{n+1}] \\
&= \frac{1 - \mathbb{P}(Z \in A_{\bar{j}}^c)^{n+1}}{(n+1)\mathbb{P}(Z \in A_{\bar{j}})}.
\end{aligned}$$

□

Lemma B.5. *Given the Hölder smoothness condition in Definition 2.1 with some smoothness β, the conditional density $p_{X|Z}(x|z \in A_{\bar{j}})$ also satisfies the same property. That is, it is $\ell = \lfloor\beta\rfloor$ times differentiable and satisfies*

$$\sup_{\alpha} |D^\alpha p_{X|Z}(x|z \in A_{\bar{j}}) - D^\alpha p_{X|Z}(x'|z \in A_{\bar{j}})| \leq W_1 \|x - x'\|_1^{\beta-\ell}$$

for all α such that $\|\alpha\|_1 = \ell$, $\alpha \in \mathbb{N}_0^{d_X}$, where $\alpha = (\alpha_1, ..., \alpha_{d_X})$

Proof. We first show that $p_{X|Z}(x|z \in A_{\bar{k}})$ is ℓ times differentiable. The Leibniz integral rule in higher dimensions allows switching the order of derivative and integration as follows

$$\frac{\partial}{\partial x_i}\left(\int_a^b f(x,z)dz\right) = \int_a^b \frac{\partial}{\partial x_i} f(x,z)dz,$$

where all elements of a, b are bounded. In context, let a, b be the lower and upper bound vector of $A_{\bar{j}}$ and let $f(x,z) = p_{X|Z}(x|z)\frac{p_Z(z)}{\mathbb{P}(Z \in A_{\bar{j}})}$. Then by the Leibniz integral rule, we have

$$D^\alpha p_{X|Z}(x|z \in A_{\bar{j}}) = \int_{A_{\bar{j}}} D^\alpha p_{X|Z}(x|z)\frac{p_Z(z)}{\mathbb{P}(Z \in A_{\bar{j}})}dz, \quad \|\alpha\|_1 = i, \text{ for } 1 \leq i \leq \ell,$$

thus proving that $p_{X|Z}(x|z \in A_{\bar{j}})$ is ℓ times differentiable.

Now, for an arbitrary α such that $\|\alpha\|_1 = \ell$, applying the Leibniz integral rule gives:

$$\begin{aligned}
&|D^\alpha p_{X|Z}(x|z \in A_{\bar{j}}) - D^\alpha p_{X|Z}(x'|z \in A_{\bar{j}})| \\
\leq &\int_{A_{\bar{j}}} \left| D^\alpha p_{X|Z}(x|z) - D^\alpha p_{X|Z}(x'|z) \right| \frac{p_Z(z)}{\mathbb{P}(Z \in A_{\bar{j}})} dz \\
\leq &\int_{A_{\bar{j}}} W_1 \|x - x'\|_1^{\beta-\ell} \frac{p_Z(z)}{\mathbb{P}(Z \in A_{\bar{j}})} dz \\
= &W_1 \|x - x'\|_1^{\beta-\ell}.
\end{aligned}$$

So $p_{X|Z}(x|z \in A_{\bar{j}})$ indeed follows the Hölder smoothness condition. □

C Proofs of Section 5

Proof of Theorem 5.1. We first show that $p_{X|Z}(x|z) = \frac{g(x,z)}{\int g(x,z)dx}$ is Hölder smooth in x for any fixed z. Notice that the denominator $\int g(x,z)dx \geq a \cdot \mu([0,1]^{d_X}) = a > 0$ is lower bounded by some constant. Then to show that $p_{X|Z}(x|z)$ is Hölder smooth it suffices to show that $g(x,z)$ is Hölder smooth in x. But we already required this in the assumptions, where taking all partial derivatives with respect to x satisfies the Hölder condition $\sup_\alpha |D^\alpha g(x,z) - D^\alpha g(x',z)| \leq C\|x - x'\|_1^{\beta-\ell}$.

Now we show that $p_{X|Z}(x|z)$ is Lipschitz smooth in z by showing that its derivative is bounded. Without loss of generality take the partial derivative with respect to some $z_i \in Z$, then we have:

$$\sup_{x,z} \left| \frac{\partial}{\partial z_i} p_{X|Z}(x|z) \right| = \sup_{x,z} \left| \frac{\frac{\partial}{\partial z_i} g(x,z) \cdot \int g(x,z)dx - \int \frac{\partial}{\partial z_i} g(x,z)dx \cdot g(x,z)}{(\int g(x,z)dx)^2} \right|,$$

where we used the Leibniz integral rule to change the order of the derivative and the integral. But $\int g(x,z)dx \geq a$, thus the denominator $(\int g(x,z)dx)^2 \geq a^2$ is lower bounded by some positive constant. Then we just need to show that the numerator is bounded. We have:

$$\begin{aligned}
&\sup_{x,z} \left| \frac{\partial}{\partial z_i} g(x,z) \cdot \int g(x,z)dx - \int \frac{\partial}{\partial z_i} g(x,z)dx \cdot g(x,z) \right| \\
\leq &\sup_{x,z} \left| \frac{\partial}{\partial z_i} g(x,z) \cdot \int g(x,z)dx \right| + \sup_{x,z} \left| \int \frac{\partial}{\partial z_i} g(x,z)dx \cdot g(x,z) \right|.
\end{aligned}$$

Recall that $g(x,z)$ is ℓ times differentiable, where $\ell = \lfloor \beta \rfloor \geq 1$. It follows that $g(x,z)$ and its derivatives up to order ℓ are continuous. Furthermore, since it is defined on a compact space $[0,1]^d$, $g(x,z)$ and $\frac{\partial}{\partial z_i} g(x,z)$ are bounded. Then each term in the expression above are bounded, and thus the numerator is bounded. So we have shown that the first derivative $\sup_{x,z} |\frac{\partial}{\partial z_i} p_{X|Z}(x|z)| \leq K$ is bounded by some constant, and therefore $p_{X|Z}$ is Lipschitz smooth in z. Substituting this result into the TV smoothness condition (see Definition 3.1) we have:

$$\begin{aligned}
\|p_{X|Z=z} - p_{X|Z=z'}\|_1 &= \int |p_{X|Z}(x|z) - p_{X|Z}(x|z')| dx \\
&\leq \int_0^1 K\|z - z'\|_1 dx \\
&= K\|z - z'\|_1
\end{aligned}$$

as desired. So $p_{X|Z}$ is indeed Hölder smooth and TV smooth (hence it is also γ-TV smooth). □

Proof of Theorem 5.2. We will first show that the function $p_{X|Z}(x|z)$ is Hölder smooth. To see this note that $\ell = \lfloor \beta \rfloor = 0$ so we do not take partial derivatives, and

$$|p_{X|Z}(x|z) - p_{X|Z}(x'|z)| = \frac{|\exp(g(x,z)) - \exp(g(x',z))|}{\int \exp(g(x,z))dx} \leq \exp(M)|\exp(g(x,z)) - \exp(g(x',z))|.$$

Next let $g = g(x,z)$ and $g' = g(x',z)$ for brevity. We have

$$|e^g - e^{g'}| \leq |g - g'| \sum_{k=1}^{\infty} \frac{\sum_{i=0}^{k-1} |g|^i |g'|^{(k-1-i)}}{k!} \leq |g - g'| \exp(M) \leq C \exp(M) \|x - x'\|_1^\beta,$$

and we conclude that

$$|p_{X|Z}(x|z) - p_{X|Z}(x'|z)| \leq C \exp(2M) \|x - x'\|_1^\beta.$$

Next we will control the quantity $\|p_{X|Z=z} - p_{X|Z=z'}\|_1$. To see this we note that

$$\begin{aligned}
\|p_{X|Z=z} - p_{X|Z=z'}\|_1 &= \int |p_{X|Z}(x|z) - p_{X|Z}(x|z')|dx \\
&= \int \left(\frac{\max(p_{X|Z}(x|z), p_{X|Z}(x|z'))}{\min(p_{X|Z}(x|z), p_{X|Z}(x|z'))} - 1 \right) \min(p_{X|Z}(x|z), p_{X|Z}(x|z'))dx \\
&\leq \int \left(\frac{\max(p_{X|Z}(x|z), p_{X|Z}(x|z'))}{\min(p_{X|Z}(x|z), p_{X|Z}(x|z'))} - 1 \right) p_{X|Z}(x|z)dx.
\end{aligned}$$

Suppose now that the function $\log p_{X|Z}(x|z)$ is Hölder with constants K and γ in z then the above can be bounded as

$$\begin{aligned}
\|p_{X|Z=z} - p_{X|Z=z'}\|_1 &\leq \int (\exp(K\|z - z'\|_1^\gamma) - 1) p_{X|Z}(x|z)dx \\
&= K\|z - z'\|_1^\gamma + \sum_{k \geq 2} (K\|z - z'\|_1^\gamma)^k / k! \\
&\leq K\|z - z'\|_1^\gamma + K\|z - z'\|_1^\gamma \sum_{k \geq 2} (K d_Z)^{k-1} / k! \\
&= K\|z - z'\|_1^\gamma + K\|z - z'\|_1^\gamma (e^{K d_Z} - 1 - K d_Z)/(K d_Z) \\
&= B\|z - z'\|_1^\gamma,
\end{aligned}$$

where $B = K(1 + (e^{K d_Z} - 1 - K d_Z)/(K d_Z))$. It remains to show that $\log p_{X|Z}(x|z)$ is Hölder with constants K and γ. Consider the difference

$$\begin{aligned}
\log p_{X|Z}(x|z) - \log p_{X|Z}(x|z') &= g(x,z) - g(x,z') - \log \frac{\int \exp(g(x,z))dx}{\int \exp(g(x,z'))dx} \\
&\leq C\|z - z'\|_1^\gamma - \log \frac{\int \exp(g(x,z) - g(x,z')) \exp(g(x,z'))dx}{\int \exp(g(x,z'))dx} \\
&\leq C\|z - z'\|_1^\gamma + \frac{\int (g(x,z') - g(x,z)) \exp(g(x,z'))dx}{\int \exp(g(x,z'))dx} \\
&\leq 2C\|z - z'\|_1^\gamma,
\end{aligned}$$

where we used Jensen's inequality in the next to last inequality. Reversing the roles of z and z' we complete the proof.

□