A Statistical Framework for Differential Privacy[1]

Larry Wasserman[*‡] Shuheng Zhou[†]

[*]Department of Statistics
[‡]Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

[†]Seminar für Statistik
ETH Zürich, CH 8092

October 2, 2009

One goal of statistical privacy research is to construct a data release mechanism that protects individual privacy while preserving information content. An example is a *random mechanism* that takes an input database X and outputs a random database Z according to a distribution $Q_n(\cdot|X)$. *Differential privacy* is a particular privacy requirement developed by computer scientists in which $Q_n(\cdot|X)$ is required to be insensitive to changes in one data point in X. This makes it difficult to infer from Z whether a given individual is in the original database X. We consider differential privacy from a statistical perspective. We consider several data release mechanisms that satisfy the differential privacy requirement. We show that it is useful to compare these schemes by computing the rate of convergence of distributions and densities constructed from the released data. We study a general privacy method, called the exponential mechanism, introduced by McSherry and Talwar (2007). We show that the accuracy of this method is intimately linked to the rate at which the probability that the empirical distribution concentrates in a small ball around the true distribution.

1 Introduction

One goal of data privacy research is to derive a mechanism that takes an input database X and releases a transformed database Z such that individual privacy is protected yet information content is preserved. This is known as disclosure limitation. In this paper we will consider various methods

[1] We thank Avrim Blum, Katrina Ligett, Steve Fienberg, Alessandro Rinaldo and Yuval Nardi for many helpful discussions. We thank Wenbo Li and Mikhail Lifshits for helpful pointers and discussions on small ball probabilities. We thank the Associate Editor and three referees for a plethora of comments that led to improvements in the paper. Research supported by NSF grant CCF-0625879, a Google research grant and a grant from Carnegie Mellon's Cylab. The second author is also partially supported by the Swiss National Science Foundation (SNF) Grant 20PA21-120050/1.

for producing a transformed database Z and we will study the accuracy of inferences from Z under various loss functions.

There are numerous approaches to this problem. The literature is vast and includes papers from computer science, statistics and other fields. The terminology also varies considerably. We will use the terms "disclosure limitation" and "privacy guarantee" interchangeably.

Disclosure limitation methods include clustering (Sweeney, 2002, Aggarwal et al., 2006), ℓ-diversity (Machanavajjhala et al., 2006), t-closeness (Li et al., 2007), data swapping (Fienberg and McIntyre, 2004), matrix masking (Ting et al., 2008), cryptographic approaches (Pinkas, 2002, Feigenbaum et al., 2006), data perturbation (Evfimievski et al., 2004, Kim and Winkler, 2003, Warner, 1965, Fienberg et al., 1998) and distributed database methods (Fienberg et al., 2007, Sanil et al., 2004). Statistical references on disclosure risk and limitation include Duncan and Lambert (1986, 1989), Duncan and Pearson (1991), Reiter (2005). We refer to Reiter (2005) and Sanil et al. (2004) for further references.

One approach to defining a privacy guarantee that has received much attention in the computer science literature is known as *differential privacy* (Dwork et al., 2006, Dwork, 2006). There is a large body of work on this topic including, for example, Dinur and Nissim (2003), Dwork and Nissim (2004), Blum et al. (2005), Dwork et al. (2007), Nissim et al. (2007), Barak et al. (2007), McSherry and Talwar (2007), Blum et al. (2008), Kasiviswanathan et al. (2008). Blum et al. (2008) gives a machine learning approach to inference under differential privacy constraints and to some extent our results are inspired by that paper. Smith (2008) shows how to provide efficient point estimators while preserving differential privacy. He constructs estimators for parametric models with mean squared error $(1 + o(1))/(nI(\theta))$ where $I(\theta)$ is the Fisher information. Machanavajjhala et al. (2008) consider privacy for histograms by sampling from the posterior distribution of the cell probabilities. We discuss Machanavajjhala et al. (2008) further in Section 4. After submitting the first draft of this paper, new work has appeared on differential privacy that is also statistical in nature, namely, Ghosh et al. (2009), Dwork and Lei (2009), Dwork et al. (2009), Feldman et al. (2009).

The goals of this paper are to explain differential privacy in statistical language, to show how to compare different privacy mechanisms by computing the rate of convergence of distributions and densities based on the released data Z, and to study a general privacy method, called the exponen-

tial mechanism, due to McSherry and Talwar (2007). We show that the accuracy of this method is intimately linked to the rate at which the probability that the empirical distribution concentrates in a small ball around the true distribution. These so called "small ball probabilities" are well-studied in probability theory. To the best of our knowledge, this is the first time a connection has been made between differential privacy and small ball probabilities. We need to make two disclaimers. First, the goal of our paper is to investigate differential privacy. We will not attempt to review all approaches to privacy or to compare differential privacy with other approaches. Such an undertaking is beyond the scope of this paper. Second, we focus only on statistical properties here. We shall not concern ourselves in this paper with computational efficiency.

In Section 2 we define differential privacy and provide motivation for the definition. In Section 3 we discuss conditions that ensure that a privacy mechanism preserves information. In Section 4 we consider two histogram based methods. In Section 5 and 6, we examine another method known as the exponential mechanism. Section 7 contains a small simulation study and Section 8 contains concluding remarks. All technical proofs appear in Section 9.

1.1 Summary of Results

We consider several different data release mechanisms that satisfy differential privacy. We evaluate the utility of these mechanisms by evaluating the rate at which $d(P, P_Z)$ goes to 0, where P is the distribution of the data $X \in \mathcal{X}$, P_Z is the empirical distribution of the released data Z, and d is some distance between distributions. This gives an informative way to compare data release mechanisms. In more detail, we consider the Kolmogorov-Smirnov (KS) distance: $\sup_{x\in\mathcal{X}} |F(x) - \widehat{F}_Z(x)|$, where F, $\widehat{F}_Z$ denote the cumulative distribution function (cdf) corresponding to P and the empirical distribution function corresponding to P_Z, respectively. We also consider the squared L_2 distance: $\int (p(x) - \widehat{p}_Z)^2$, where $\widehat{p}_Z$ is a density estimator based on Z. Our results are summarized in the following tables, where n denotes the sample size.

The first table concerns the case where the data are in $\mathbb{R}^r$ and the density p of P is Lipschitz. Also reported are the minimax rates of convergence for density estimators in KS and in squared L_2 distances. We see that the accuracy depends both on the data releasing mechanism and the distance

function d. The results are from Sections 4 and 5 of the paper. (The exponential mechanism under L_2 distance is marked NA but is in the second table in case $r = 1$. We note that the rate for KS distance for perturbed histogram is $\sqrt{\log n/n}$ for $r = 1$.)

Distance	Data Release mechanism			minimax rate
	smoothed histogram	perturbed histogram	exponential mechanism	
L_2	$n^{-2/(2r+3)}$	$n^{-2/(2+r)}$	NA	$n^{-2/(2+r)}$
Kolmogorov-Smirnov	$\sqrt{\log n} \times n^{-2/(6+r)}$	$\log n \times n^{-2/(2+r)}$	$n^{-1/3}$	$n^{-1/2}$

The next table summarizes the results for the case where the dimension of X is $r = 1$ and the density p is assumed to be in a Sobolev space of order γ. We only consider the squared L_2 distance between the true density p and the estimated density $\widehat{p}_Z$ in this case. The results are from Section 6 of the paper.

	exponential mechanism	perturbed orthogonal series estimator	minimax rate
L_2	$n^{-\gamma/(2\gamma+1)}$	$n^{-2\gamma/(2\gamma+1)}$	$n^{-2\gamma/(2\gamma+1)}$

Our results show that, in general, privacy schemes seem not to yield minimax rates. Two exceptions are perturbation methods evaluated under L_2 loss which do yield minimax rates. An open question is whether the slower than minimax rates are intrinsic to the privacy methods. It is possible, for example, that our rates are not tight. This question could be answered by establishing lower bounds on these rates. We consider this an important topic for future research.

2 Differential Privacy

Let $X_1, \ldots, X_n$ be a random sample (independent and identically distributed) of size n from a distribution P where $X_i \in \mathcal{X}$. To be concrete, we shall assume that $\mathcal{X} \equiv [0,1]^r = [0,1] \times [0,1] \times \cdots \times [0,1]$ for some integer $r \geq 1$. Extensions to more general sample spaces are certainly possible but we focus on this sample space to avoid unnecessary technicalities. (In particular, it

is difficult to extend differential privacy to unbounded domains.) Let μ denote Lebesgue measure and let $p = dP/d\mu$ if the density exists. We call $X = (X_1, \ldots, X_n)$ a database. Note that $X \in \mathcal{X}^n = [0,1]^r \times \cdots \times [0,1]^r$. We focus on mechanisms that take a database X as input and output a sanitized database $Z = (Z_1, \ldots, Z_k) \in \mathcal{X}^k$ for public release. In general, Z need not be the same size as X. For some schemes, we shall see that large k can lead to low privacy and high accuracy while while small k can lead to high privacy and low accuracy. We will let $k \equiv k(n)$ change with n. Hence, any asymptotic statements involving n increasing will also allow k to change as well.

A *data release mechanism* $Q_n(\cdot|X)$ is a conditional distribution for $Z = (Z_1, \ldots, Z_k)$ given X. Thus, $Q_n(B|X = x)$ is the probability that the output database Z is in a set $B \in \mathcal{B}$ given that the input database is x, where $\mathcal{B}$ are the measurable subsets of $\mathcal{X}^k$. We call $Z = (Z_1, \ldots, Z_k)$ a *sanitized database*. Schematically:

$$\text{input database } X = (X_1, \ldots, X_n) \xrightarrow[\text{sanitize}]{Q_n(Z|X)} \text{output database } Z = (Z_1, \ldots, Z_k).$$

The marginal distribution of the output database Z induced by P and Q_n is $M_n(B) = \int Q_n(B|X = x) dP^n(x)$ where P^n is the n-fold product measure of P.

Example 2.1. *A simple example to help the reader have a concrete example in mind is adding noise. In this case, $Z = (Z_1, \ldots, Z_n)$ where $Z_i = X_i + \epsilon_i$ and $\epsilon_1, \ldots, \epsilon_n$ are mean 0 independent observations drawn from some known distribution H with density h. Hence Q_n has density $q_n(z_1, \ldots, z_n|x_1, \ldots, x_n) = \prod_{i=1}^n h(z_i - x_i)$.*

Definition 2.2. *Given two databases $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$, let $\delta(X, Y)$ denote the Hamming distance between X and Y: $\delta(X, Y) = \#\Big\{i : \ X_i \neq Y_i\Big\}$.*

A general data release mechanism is the *exponential mechanism* (McSherry and Talwar, 2007) which is defined as follows. Let $\xi : \mathcal{X}^n \times \mathcal{X}^k :\to [0, \infty)$ be any function. Each such ξ defines a different exponential mechanism. Let

$$\Delta \equiv \Delta_{n,k} = \sup_{\substack{x,y \in \mathcal{X}^n \\ \delta(x,y)=1}} \sup_{z \in \mathcal{X}^k} |\xi(x, z) - \xi(y, z)|, \tag{1}$$

that is, $\Delta_{n,k}$ is the maximum change to ξ caused by altering a single entry in x. Finally, let $(Z_1, \ldots, Z_k)$ be a random vector drawn from the density

$$h(z|x) = \frac{\exp\left(-\frac{\alpha\xi(x,z)}{2\Delta_{n,k}}\right)}{\int_{\mathcal{X}^k} \exp\left(-\frac{\alpha\xi(x,s)}{2\Delta_{n,k}}\right) ds} \tag{2}$$

where $\alpha \geq 0$, $z = (z_1, \ldots, z_k)$ and $x = (x_1, \ldots, x_n)$. In this case, Q_n has density $h(z|x)$. We'll discuss the exponential mechanism in more detail later.

There are many definitions of privacy but in this paper we focus on the following definition due to Dwork et al. (2006) and Dwork (2006).

Definition 2.3. *Let $\alpha \geq 0$. We say that Q_n satisfies* α-differential privacy *if*

$$\sup_{\substack{x,y \in \mathcal{X}^n \\ \delta(x,y)=1}} \sup_{B \in \mathcal{B}} \frac{Q_n(B|X=x)}{Q_n(B|X=y)} \leq e^{\alpha} \tag{3}$$

where $\mathcal{B}$ are the measurable sets on $\mathcal{X}^k$. The ratio is interpreted to be 1 whenever the numerator and denominator are both 0.

The definition of differential privacy is based on ratios of probabilities. It is crucial to measure closeness by ratios of probabilities since that protects rare cases which have small probability under Q_n. In particular, if changing one entry in the database X cannot change the probability distribution $Q_n(\cdot|X = x)$ very much, then we can claim that a single individual cannot guess whether he is in the original database or not. The closer e^{α} is to 1, the stronger privacy guarantee is. Thus, one typically chooses α close to 0. See Dwork et al. (2006) for more discussion on these points. Indeed, suppose that two subjects each believe that one of them is in the original database. Given Z and full knowledge of P and Q_n can they test who is in X? The answer is given in the following result. (In this result, we drop the assumption that the user does not know Q_n.)

Theorem 2.4. *Suppose that Z is obtained from a data release mechanism that satisfies α-differential privacy. Any level γ test which is a function of Z, P and Q_n of $H_0 : X_i = s$ versus $H_1 : X_i = t$ has power bounded above by γe^{α}.*

Thus, if Q_n satisfies differential privacy then it is virtually impossible to test the hypothesis that either of the two subjects is in the database since the power of such a test is nearly equal to its level. A similar calculation shows that if one does a Bayes test between H_0 and H_1 then the Bayes factor is always between $e^{-2\alpha}$ and $e^{2\alpha}$. For more detail on the motivation for the definition as well as consequences, see Dwork et al. (2006), Dwork (2006), Ganta et al. (2008), Rastogi et al. (2009).

The following result, which is proved in McSherry and Talwar (2007) (Theorem 6), shows that the exponential mechanism always preserves differential privacy.

Theorem 2.5. (McSherry and Talwar, 2007) *The exponential mechanism satisfies the α-differential privacy.*

To conclude this section we record a few useful facts. Let $T(X, R)$ be a function of X and some auxiliary random variable R which is independent of X. After including this auxiliary random variable we define differential privacy as before. Specifically, $T(X, R)$ satisfies differential privacy if for all B, and all x, x' with $\delta(x, x') = 1$ we have that $\mathbb{P}(T(X, R) \in B|X = x) \leq e^{\alpha}\mathbb{P}(T(X, R) \in B|X = x')$. The third part is Proposition 1 from Dwork et al. (2006).

Lemma 2.6. *We have the following:*

1. *If $T(X, R)$ satisfies differential privacy then $U = h(T(X, R))$ also satisfies differential privacy for any measurable function h.*

2. *Suppose that g is a density function constructed from a random vector $T(X, R)$ that satisfies differential privacy. Let $Z = (Z_1, \ldots, Z_k)$ be k iid draws from g. This defines a mechanism $Q_n(B|X) = \mathbb{P}(Z \in B|X)$. Then Q_n satisfies differential privacy for any k.*

3. *(Proposition 1 from Dwork et al. (2006).) Let $f(x)$ be a function of $x = (x_1, \ldots, x_n)$ and define $S(f) = \sup_{x,x':\delta(x,x')=1} \|f(x) - f(x')\|_1$ where $\|a\|_1 = \sum_j |a_j|$. Let R have density $g(r) \propto e^{-\alpha|r|/S(f)}$. Then $T(X, R) = f(X) + R$ satisfies differential privacy.*

3 Informative Mechanisms

A challenge in privacy theory is to find Q_n that satisfies differential privacy and yet yields datasets Z that preserve information. Informally, a mechanism is informative if it is possible to make precise inferences from the released data $Z_1, \ldots, Z_k$. Whether or not a mechanism is informative will depend on the goals of the inference. From a statistical perspective, we would like to infer P or functionals of P from Z. Blum et al. (2008) show that the probability content of some classes of intervals can be estimated accurately while preserving privacy. Their results motivated the current paper. We will assume throughout that the user has access to the sanitized data Z but not the mechanism Q_n. The question of how a data analyst can use knowledge of Q_n to improve inferences is left to future work.

There are many ways to measure the information in Z. One way is through distribution functions. Let F denote the cumulative distribution function (cdf) on $\mathcal{X}$ corresponding to P. Thus $F(x) = P(X \in (-\infty, x_1] \times \cdots \times (-\infty, x_r])$ where $x = (x_1, \ldots, x_r)$. Let $\widehat{F} \equiv \widehat{F}_X$ denote the empirical distribution function corresponding to X and similarly let $\widehat{F}_Z$ denote the empirical distribution function corresponding to Z. Let ρ denote any distance measure on distribution functions.

Definition 3.1. *Q_n is consistent with respect to ρ if $\rho(F, \widehat{F}_Z) \xrightarrow{P} 0$. Q_n is ϵ_n-informative if $\rho(F, \widehat{F}_Z) = O_P(\epsilon_n)$.*

An alternative to requiring $\rho(F, \widehat{F}_Z)$ to be small is to require $\rho(\widehat{F}, \widehat{F}_Z)$ to be small. Or one could require $Q_n(\rho(\widehat{F}, \widehat{F}_Z) > \epsilon | X = x)$ be small for all x as in Blum et al. (2008). These requirements are similar. Indeed, suppose ρ satisfies the triangle inequality and that $\widehat{F}$ is consistent in the ρ distance, that is, $\rho(\widehat{F}, F) \xrightarrow{P} 0$. Assume further that $\rho(\widehat{F}, F) = O_P(\epsilon_n)$. Then $\rho(F, \widehat{F}_Z) = O_P(\epsilon_n)$ implies that

$$\rho(\widehat{F}, \widehat{F}_Z) \leq \rho(\widehat{F}, F) + \rho(F, \widehat{F}_Z) = O_P(\epsilon_n);$$

Similarly, $\rho(\widehat{F}, \widehat{F}_Z) = O_P(\epsilon_n)$ implies that $\rho(F, \widehat{F}_Z) = O_P(\epsilon_n)$.

Let $\mathbb{E}_{P,Q_n}$ denote the expectation under the joint distribution defined by P^n and Q_n. Sometimes we write $\mathbb{E}$ when there is no ambiguity. Similarly, we use $\mathbb{P}$ to denote the marginal probability

under P^n and Q_n: $\mathbb{P}(A) = \int_A dQ_n(z_1, \ldots, z_k | x_1, \ldots, x_n) dP(x_1) \cdots dP(x_n)$ for $A \in \mathcal{X}^k$.

There are many possible choices for ρ. We shall mainly focus on the Kolmogorov-Smirnov (KS) distance $\rho(F, G) = \sup_x |F(x) - G(x)|$ and the squared L_2 distance $\rho(F, G) = \int (f(x) - g(x))^2 dx$ where $f = dF/d\mu$ and $g = dG/d\mu$. However, our results can be carried over to other distances as well.

Before proceeding let us note that we will need some assumptions on F otherwise we cannot have a consistent scheme as shown in the following theorem. The following result — essentially a re-expression of a result in Blum et al. (2008) in our framework — makes this clear.

Theorem 3.2. *Suppose that Q_n satisfies differential privacy and that $\rho(F, G) = \sup_x |F(x) - G(x)|$. Let F be a point mass distribution. Thus $F(y) = I(y \geq x)$ for some point $x \in [0, 1]$. Then $\widehat{F}_Z$ is inconsistent, that is, there is a $\delta > 0$ such that $\liminf_{n \to \infty} P^n(\rho(F, \widehat{F}_Z) > \delta) > 0$.*

4 Sampling From a Histogram

The goal of this section is to give two concrete, simple data release methods that achieve differential privacy. The idea is to draw a random sample from histogram. The first scheme draws observations from a smoothed histogram. The second scheme draws observations from a randomly perturbed histogram. We use the histogram for its familiarity and simplicity and because it is used in applications of differential privacy. We will see that the histogram has to be carefully constructed to ensure differential privacy. We then compare the two schemes by studying the accuracy of the inferences from the released data. We will see that the accuracy depends both on how the histogram is constructed and on what measure of accuracy we use.

Let $L > 0$ be a constant and suppose that $p = dP/d\mu \in \mathcal{P}$ where

$$\mathcal{P} = \left\{ p : \ |p(x) - p(y)| \leq L||x - y|| \right\} \tag{4}$$

is the class of Lipschitz functions. We assume throughout this section that $p \in \mathcal{P}$. The minimax rate of convergence for density estimators in squared L_2 distance for $\mathcal{P}$ is $n^{-2/(2+r)}$ (Scott, 1992).

Let $h = h_n$ be a binwidth such that $0 < h < 1$ and such that $m = 1/h^r$ is an integer. Partition $\mathcal{X}$ into m bins $\{B_1, \ldots, B_m\}$ where each bin B_j is a cube with sides of length h. Let $I(\cdot)$ denote the indicator function. Let $\widehat{f}_m$ denote the corresponding histogram estimator on $\mathcal{X}$, namely,

$$\widehat{f}_m(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h^r} I(x \in B_j)$$

where $\widehat{p}_j = C_j/n$ and $C_j = \sum_{i=1}^{n} I(X_i \in B_j)$ is the number of observations in B_j. Recall that $\widehat{f}_m$ is a consistent estimator of p if $h = h_n \to 0$ and $nh_n^r \to \infty$. Also, the optimal choice of $m = m_n$ for L_2 error under $\mathcal{P}$ is $m_n \asymp n^{r/(2+r)}$, in which case $\int (p - \widehat{f}_m)^2 = O_P(n^{-2/(2+r)})$ (Scott, 1992). Here, $a_n \asymp b_n$ means that both a_n/b_n and b_n/a_n are bounded for large n.

4.1 Sampling from a Smoothed Histogram

The first method for generating released data Z from a histogram while achieving differential privacy proceeds as follows. Recall that the sample space is $[0,1]^r$. Fix a constant $0 < \delta < 1$ and define the smoothed histogram

$$\widehat{f}_{m,\delta}(x) = (1-\delta)\widehat{f}_m(x) + \delta. \tag{5}$$

Theorem 4.1. *Let $Z = (Z_1, \ldots, Z_k)$ where $Z_1, \ldots, Z_k$ are k iid draws from $\widehat{f}_{m,\delta}(x)$. If*

$$k \log\left(\frac{(1-\delta)m}{n\delta} + 1\right) \leq \alpha \tag{6}$$

then α-differential privacy holds.

Note that for $\delta \to 0$ and $\frac{m}{n\delta} \to 0$, $\log\left(\frac{(1-\delta)m}{n\delta} + 1\right) = \frac{m}{n\delta}(1 + o(1)) \approx \frac{m}{n\delta}$. Thus (6) is approximately the same as requiring

$$\frac{mk}{\delta} \leq n\alpha. \tag{7}$$

Equation (7) shows an interesting tradeoff between m, k and δ. We note that sampling from the usual histogram corresponding to $\delta = 0$ does not preserve differential privacy.

Now we consider how to choose m, k, δ to minimize $\mathbb{E}(\rho(F, \widehat{F}_Z))$ while satisfying (6). Here, $\mathbb{E}$ is the expectation under the randomness due to sampling from P and due to the privacy mechanism Q_n. Thus, for any measurable function h,

$$\mathbb{E}(h(Z)) = \int \int h(z_1, \ldots, z_k) dQ_n(z_1, \ldots, z_k | x_1, \ldots, x_n) dP(x_1) \cdots dP(x_n).$$

Now we give a result that shows how accurate the inferences are in the KS distance using the smoothed histogram sampling scheme.

Theorem 4.2. *Suppose that $Z_1, \ldots, Z_k$ are drawn as described in the previous theorem. Suppose* (4) *holds. Let ρ be the KS distance. Then choosing $m \asymp n^{r/(6+r)}$, $k \asymp m^{4/r} = n^{4/(6+r)}$ and $\delta = (mk/n\alpha)$ minimizes $\mathbb{E}\rho(F, \widehat{F}_Z)$ subject to* (6)*. In this case, $\mathbb{E}\rho(F, \widehat{F}_Z) = O\left(\frac{\sqrt{\log n}}{n^{2/(6+r)}}\right)$.*

In this case we see that we have consistency since $\rho(F, \widehat{F}_Z) = o_P(1)$ but the rate is slower than the minimax rate of convergence for density estimators in KS distance, which is $n^{-1/2}$. Now let $\widehat{q}_j = \#\{Z_i \in B_j\}/k$ and

$$\rho(F, \widehat{F}_Z) = \int (p(x) - \widehat{f}_Z(x))^2 dx, \text{ where } \widehat{f}_Z(x) = h^{-r} \sum_{j=1}^{m} \widehat{q}_j I(x \in B_j). \tag{8}$$

Theorem 4.3. *Assume the conditions of the previous theorem. Let ρ be the squared L_2 distance as defined in (8). Then choosing*

$$m \asymp n^{r/(2r+3)}, \quad k \asymp n^{(r+2)/(2r+3)}, \quad \delta \asymp n^{-1/(r+3)}$$

minimizes $\mathbb{E}\rho(F, \widehat{F}_Z)$ subject to (6)*. In this case, $\mathbb{E}\rho(F, \widehat{F}_Z) = O(n^{-2/(2r+3)})$.*

Again, we have consistency but the rate is slower than the minimax rate which is $n^{-2/(2+r)}$. (Scott, 1992)

4.2 Sampling From a Perturbed Histogram

The second method, which we call the sampling from a perturbed histogram, is due to Dwork et. al. (2006). Recall that C_j is the number of observations in bin B_j. Let $D_j = C_j + \nu_j$ where $\nu_1, \ldots, \nu_m$ are independent, identically distributed draws from a Laplace density with mean 0 and variance $8/\alpha^2$. Thus the density of ν_j is $g(\nu) = (\alpha/4)e^{-|\nu|\alpha/2}$. Dwork et. al. (2006) show that releasing $D = (D_1, \ldots, D_m)$ preserves differential privacy. However, our goal is to release a database $Z = (Z_1, \ldots, Z_k)$ rather than just a set of counts. Now define

$$\widetilde{D}_j = \max\{D_j, 0\} \text{ and } \widehat{q}_j = \widetilde{D}_j / \sum_s \widetilde{D}_s.$$

Since D preserves differential privacy, it follows from Lemma 2.6 that $(\widehat{q}_1, \ldots, \widehat{q}_m)$ also preserve differential privacy; Moreover, any sample $Z = (Z_1, \ldots, Z_k)$ from $\widetilde{f}(x) = h^{-r} \sum_{j=1}^m \widehat{q}_j I(x \in B_j)$ preserve differential privacy for any k.

Theorem 4.4. *Let $Z = (Z_1, \ldots, Z_k)$ be drawn from $\widetilde{f}(x) = h^{-r} \sum_{j=1}^m \widehat{q}_j I(x \in B_j)$. Assume that there exists a constant $1 \leq C < \infty$ such that $\sup_x p(x) = C$.*
(1) Let ρ be the L_2 distance and $\widehat{f}_Z$ be as defined in (8). Let $m \asymp n^{r/(2+r)}$ and let $k \geq n$. Then we have $\mathbb{E}\rho(F, \widehat{F}_Z) = O(n^{-2/(2+r)})$.
(2) Let ρ be the KS distance. Let $m \asymp n^{r/(2+r)}$. Then $\mathbb{E}\rho(F, \widehat{F}_Z) = O\left(\min\left(\frac{\log n}{n^{2/(2+r)}}, \sqrt{\frac{\log n}{n}}\right)\right)$.

Hence, this method achieves the minimax rate of convergence in L_2 while the first data release method does not. This suggests that the perturbation method is preferable for the L_2 distance. The perturbation method does not achieve the minimax rate of convergence in KS distance; in fact, the exponential mechanism based method achieves a better rate as we shown in Section 5 (Theorem 5.4). We examine this method numerically in Section 7.

Another approach to histograms is given by Machanavajjhala et al. (2008). They put a Dirichlet $(a_1, \ldots, a_m)$ prior on the cell probabilities $p_1, \ldots, p_m$ where $p_j = \mathbb{P}(X_i \in B_j)$. The corresponding posterior is Dirichlet $(a_1 + C_1, \ldots, a_m + C_m)$. Next they draw $q = (q_1, \ldots, q_m)$ from the posterior and finally they sample new cell counts $D = (D_1, \ldots, D_m)$ from a Multinomial (k, q). Thus, the

distribution of D given X is

$$\mathbb{P}(D = d|X) = \frac{\prod_{j=1}^{m} \Gamma(d_j + a_j + C_j)}{\Gamma(k + n + \sum_j a_j)}.$$

They show that differential privacy requires $a_j + C_j \geq k/(e^\alpha - 1)$ for all j. If we take $a_1 = a_2 = \cdots = a_m$ then this is similar to the first histogram-based data release method we discussed in this section. They also suggest a weakened version of differential privacy.

5 Exponential Mechanism

In this section we will consider the exponential mechanism in some detail. We'll derive some general results about accuracy and apply the method to the mean, and to density estimation. Specifically, we will show the following for exponential mechanisms:

1. Choosing the size k of the released database is delicate. Taking k too large compromises privacy. Taking k too small compromises accuracy.

2. The accuracy of the exponential scheme can be bounded by a simple formula. This formula has a term that measures how likely it is for a distribution based on sample size k, to be in a small ball around the true distribution. In probability theory, this is known as a small ball probability.

3. The formula can be applied to several examples such as the KS distance, the mean, and nonparametric density estimation using orthogonal series. In each case we can use our results to choose k and to find the rate of convergence of an estimator based on the sanitized data.

In light of Theorem 3.2, we know that some assumptions are needed on P. We shall assume throughout this section that P has a bounded density p; note that this is a weaker condition than (4).

Recall the exponential mechanism. We draw the vector $Z = (Z_1, \ldots, Z_k)$ from $h(z|x)$ where

$$\begin{aligned} h(z|x) &= \frac{g_x(z)}{\int_{[0,1]^k} g_x(s)ds}, \quad \text{where } g_x(z) = \exp\left(-\frac{\alpha\, \rho(\widehat{F}_x, \widehat{F}_z)}{2\Delta_{n,k}}\right) \quad \text{and} \qquad (9) \\ \Delta &\equiv \Delta_{n,k} = \sup_{\substack{x,y\in\mathcal{X}^n \\ \delta(x,y)=1}} \sup_{z\in\mathcal{X}^k} |\rho(\widehat{F}_x, \widehat{F}_z) - \rho(\widehat{F}_y, \widehat{F}_z)|. \end{aligned}$$

Lemma 5.1. *For KS distance $\Delta_{n,k} \leq \frac{1}{n}$.*

This framework is used in Blum et al. (2008). For the rest of this section, assume that $Z = (Z_1, \ldots, Z_k)$ are drawn from an exponential mechanism Q_n.

Definition 5.2. *Let F denote the cumulative distribution function on $\mathcal{X}$ corresponding to P. Let $\widehat{G}$ denote the empirical cdf from a sample of size k from P, and let*

$$R(k, \epsilon) = P^k(\rho(F, \widehat{G}) \leq \epsilon).$$

$R(k, \epsilon)$ is called the small ball probability associated with ρ.

The following theorem bounds the accuracy of the estimator from the sanitized data by a simple formula involving the small ball probability.

Theorem 5.3. *Assume that P has a bounded density p, and that there exists $\epsilon_n \to 0$ such that*

$$\mathbb{P}\left(\rho(F, \widehat{F}_X) > \frac{\epsilon_n}{16}\right) = O\left(\frac{1}{n^c}\right) \qquad (10)$$

for some $c > 1$. Further suppose that ρ satisfies the triangle inequality. Let $Z = (Z_1, \ldots, Z_k)$ be drawn from $g_x(z)$ given in (9)*. Then,*

$$\mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n\right) \leq \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon_n}{16\Delta}\right)}{R(k, \epsilon_n/2)} + O\left(\frac{1}{n^c}\right). \qquad (11)$$

Thus, if we can choose $k = k_n$ in such a way that the right hand side of (11) goes to 0, then the mechanism is consistent. We now show some examples that satisfy these conditions and we show how to choose k_n.

5.1 The KS Distance

Theorem 5.4. *Suppose that P has a bounded density p and let $B := \log \sup_x p(x) > 0$. Let $Z = (Z_1, \ldots, Z_k)$ be drawn from $g_x(z)$ given in (9) with ρ being the KS distance. By requiring that $k_n \asymp \left(\frac{3\alpha}{B}\right)^{2/3} n^{2/3}$, we have for $\epsilon_n = 2\left(\frac{B}{3\alpha}\right)^{1/3} n^{-1/3}$, and for ρ being the KS distance,*

$$\rho(F, \widehat{F}_Z) = O_P\left(\epsilon_n\right). \tag{12}$$

Note that $\rho(F, \widehat{F}_Z)$ converges to 0 at a slower rate than $\rho(F, \widehat{F}_X)$. We thus see that the rate after sanitization is $n^{-1/3}$ which is slower than the optimal rate of $n^{-1/2}$. It is an open question whether this rate can be improved.

5.2 The Mean

It is interesting to consider what happens when $\rho(F, \widehat{F}_Z) = ||\mu - \overline{Z}||^2$ where $\mu = \int x dP(x)$ and $\overline{Z}$ is the sample mean of Z. In this case $\Delta \leq r/n$. Thus, $h(u|x) \approx e^{-n||\overline{X}-\overline{Z}||^2/(2\alpha)}$ so, approximately, $Z_1, \ldots, Z_k \sim N(\overline{X}, k\alpha/n)$. Indeed, it suffices to take $k = 1$ in this case since then $\overline{Z} = \overline{X} + O_P(1/\sqrt{n})$. Thus $\overline{Z}$ converges at the same rate as $\overline{X}$. This is not surprising: preserving a single piece of information requires a database of size $k = 1$.

6 Orthogonal Series Density Estimation

In this section, we develop an exponential scheme based on density estimation and we compare it to the perturbation approach. For simplicity we take $r = 1$. Let $\{1, \psi_1, \psi_2, \ldots, \}$ be an orthonormal basis for $L_2(0,1) = \{f : \int_0^1 f^2(x)dx < \infty\}$ and assume that $p \in L_2(0,1)$. Hence

$$p(x) \;=\; 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x) \;\text{ where }\; \beta_j = \int_0^1 \psi_j(x) p(x) dx.$$

We assume that the basis functions are uniformly bounded so that

$$c_0 \equiv \sup_j \sup_x |\psi_j(x)| < \infty. \tag{13}$$

Let $\mathcal{B}(\gamma, C)$ denote the Sobolev ellipsoid

$$\mathcal{B}(\gamma, C) = \left\{ \beta = (\beta_1, \beta_2, \ldots) : \sum_{j=1}^{\infty} \beta_j^2 j^{2\gamma} \leq C^2 \right\}$$

where $\gamma > 1/2$. Let

$$\mathcal{P}(\gamma, C) = \left\{ p(x) = 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x) : \ \beta \in \mathcal{B}(\gamma, C) \right\}.$$

The minimax rate of convergence in L_2 norm for $\mathcal{P}(\gamma, C)$ is $n^{-2\gamma/(2\gamma+1)}$ (Efromovich, 1999). Thus

$$\inf_{\widehat{p}} \sup_{P \in \mathcal{P}(\gamma, C)} E \int (\widehat{p}(x) - p(x))^2 dx \geq c_1 n^{-2\gamma/(2\gamma+1)}$$

for some $c_1 > 0$. This rate is achieved by the estimator

$$\widehat{p}(x) = 1 + \sum_{j=1}^{m_n} \widehat{\beta}_j \psi_j(x) \tag{14}$$

where $m_n = n^{1/(2\gamma+1)}$ and $\widehat{\beta}_j = n^{-1} \sum_{i=1}^{n} \psi_j(X_i)$. See Efromovich (1999).

For a function $u \in L_2(0,1)$, let us define $\|u\|_{\ell_2} = \left(\int_0^1 |u(x)|^2 dx \right)^{1/2}$, which is a norm on $L_2(0,1)$. Now consider an exponential mechanism based on

$$\xi(X, Z) = \left(\int \left(\widehat{p}(x) - \widehat{p}^*(x) \right)^2 dx \right)^{1/2} := \|\widehat{p} - \widehat{p}^*\|_{\ell_2} \qquad \text{where} \tag{15}$$

$$\widehat{p}^*(x) = 1 + \sum_{j=1}^{m_k} \widehat{\beta}_j^* \psi_j(x), \ \text{ for } \ m_k = k^{\frac{1}{2\gamma+1}} \ \text{ and } \ \widehat{\beta}_j^* = k^{-1} \sum_{i=1}^{k} \psi_j(Z_i). \tag{16}$$

Lemma 6.1. *Under the above scheme we have* $\Delta \leq \frac{2c_0^2 m_n}{n}$ *for* c_0 *as defined in* (13)*. Hence,*

$$g(z|x) = \exp\left(-\frac{\alpha \left\|\widehat{p}^* - \widehat{p}\right\|_{\ell_2}}{\Delta}\right) \leq \exp\left(-\frac{\alpha n \left\|\widehat{p}^* - \widehat{p}\right\|_{\ell_2}}{2c_0^2 m_n}\right) \quad \textit{almost surely.} \tag{17}$$

Theorem 6.2. *Let* $Z = (Z_1, \ldots, Z_k)$ *be drawn from* $g_x(z)$ *given in (17). Assume that* $\gamma > 1$*. If we choose* $k \asymp \sqrt{n}$ *then*

$$\rho^2(p, \widehat{p}^*) = O_P\left(n^{-\frac{\gamma}{2\gamma+1}}\right).$$

We conclude that the sanitized estimator converges at a slower rate than the minimax rate. Now we compare this to the perturbation approach. Let $Z = (Z_1, \ldots, Z_k)$ be an iid sample from

$$\widehat{q}(x) = 1 + \sum_{j=1}^{m_n} (\widehat{\beta}_j + \nu_j)\psi_j(x)$$

where $\nu_1, \ldots, \nu_m$ are iid draws from a Laplace distribution with density $g(\nu) = (n\alpha/(2c_0 m))e^{-n\alpha|\nu|/(c_0 m)}$. Thus, i the notation of 2.6, $R = (\nu_1, \ldots, \nu_m)$. It follows from Lemma 2.6 that, for any k, this preserves differential privacy. If $\widehat{q}(x) < 0$ for any x then we replace $\widehat{q}$ by $\widehat{q}(x)I(\widehat{q}(x) > 0)/\int \widehat{q}(s)I(\widehat{q}(s) > 0)ds$ as in Hall and Murison (1993).

Theorem 6.3. *Let* $Z = (Z_1, \ldots, Z_k)$ *be drawn from* $\widehat{q}$*. Assume that* $\gamma > 1$*. If we choose* $k \geq n$*, then*

$$\rho^2(p, \widehat{p}_Z) = O_P\left(n^{-\frac{2\gamma}{2\gamma+1}}\right)$$

where $\widehat{p}_Z$ *is the orthogonal series density estimator based on* Z*.*

Hence, again, the perturbation technique achieves the minimax rate of convergence and so appears to be superior to the exponential mechanism. We do not know if this is because the exponential mechanism is inherently less accurate, or if our bounds for the exponential mechanism are not tight enough.

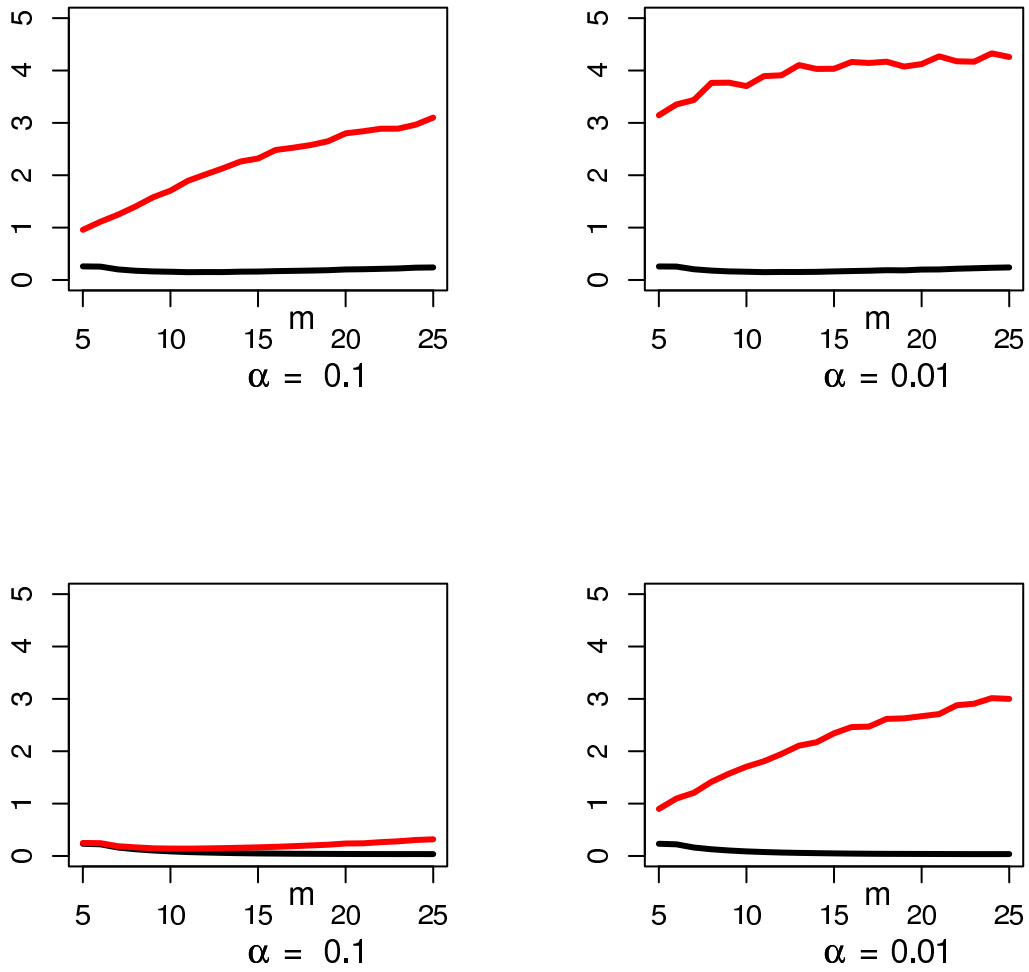

Figure 1: Top two plots $n = 100$. Bottom two plots $n = 1,000$. Each plot shows the mean integrated squared error of the histogram. The lower line is from the histogram based on the original data. The upper line is based on the perturbed histogram.

7 Example

Here we consider a small simulation study to see the effect of perturbation on accuracy. We focus on the histogram perturbation method with $r = 1$. We take the true density of X to be a Beta(10,10) density. We considered sample sizes $n = 100$ and $n = 1,000$ and privacy levels $\alpha = 0.1$, and $\alpha = 0.01$. We take ρ to be squared error distance. Figure 1 shows the results of 1,000 simulations for various numbers of bins m.

As expected, smaller values of α induce a larger information loss which manifests itself as a

larger mean squared error. Despite the fact that the perturbed histogram achieves the minimax rate, the error is substantially inflated by the perturbation. This means that the constants in the risk are important, not just the rate. Also, the risk of the sanitized histograms is much more sensitive to the choice of the number of cells than the original histogram is.

We repeated the simulations with a bimodal density, namely, $p(x)$ being an equal mixture of a Beta(10,3) density and Beta(3,10) density. The results turned out to be nearly identical to those above.

8 Conclusion

Differential privacy is an important type of privacy guarantee when releasing data. Our goal has been to present the idea in statistical language and then to show that loss functions based on distributions and densities can be useful for comparing privacy mechanisms.

We have seen that sampling from a histogram leads to differential privacy as long as either the histogram is shifted away from 0 by a factor δ or if the cells are perturbed appropriately. The latter method achieves a faster rate of convergence in L_2 distance. But, the simulation showed that the risk can nonetheless be quite large. This suggests that more work is needed to get precise finite sample risk bounds. Also, the choice of the smoothing parameter (number of cells in the histogram) has a larger effect on the sanitized histogram than on the original histogram.

We also studied the exponential mechanism. Here we derived a formula for assessing the accuracy of the method. The formula involves small ball probabilities. As far as we know, the connection between differential privacy and small ball probabilities has not been observed before.

Minimaxity is desirable for any statistical procedure. We have seen that in some cases the minimax rate is achieved and in some cases it is not. We do not yet have a complete minimax theory for differential privacy and this is the focus of our current work. We close with some open questions.

1. When is it possible for $\rho(F, \widehat{F}_Z)$ to have the same rate as $\rho(F, \widehat{F}_X)$?

2. When adaptive minimax methods are used, such as adapting to γ in Section 6 or when using

wavelet estimation methods, is some form of adaptivity preserved after sanitization?

3. Many statistical methods involve some sort of risk minimization. A example is choosing a bandwidth by cross-validation. What is the effect of sanitization on these procedures?

4. Are there other, better methods of sanitization that preserve differential privacy?

9 Proofs

9.1 Proof of Theorem 2.4

Without loss of generality take $i = 1$. Let $M_0(B) = \int Q(B|s, x_2, \ldots, x_n)\ dP(x_2, \ldots, x_n)$ and $M_1(B) = \int Q(B|t, x_2, \ldots, x_n) dP(x_2, \ldots, x_n)$. By the Neyman-Pearson lemma, the highest power test is to reject H_0 when $U > u$ where $U(z) = (dM_1/dM_0)(z)$ and u is chosen so that $\int I(U(z) > u) dM_0(z) \leq \gamma$. Since $(s, x_2, \ldots, x_n)$ and $(t, x_2, \ldots, x_n)$ differ in only one coordinate, $M_1(B) \leq e^\alpha M_0(B)$ and so the power is $M_1(U > u) \leq e^\alpha M_0(U > u) \leq \gamma e^\alpha$. $\square$

9.2 Proof of Lemma 2.6

For the first part simply note that $\mathbb{P}(h(T(X, R)) \in B|X = x) = \mathbb{P}(T(X, R) \in h^{-1}(B)|X = x) \leq e^\alpha \mathbb{P}(T(X, R) \in h^{-1}(B)|X = x') = e^\alpha \mathbb{P}(h(T(X, R)) \in B|X = x')$.

For the second part, let $Z = (Z_1, \ldots, Z_k)$ and note that Z is independent of X given $T(X, R)$. Let H be the distribution of $T(X, R)$. Hence,

$$\begin{aligned}
\mathbb{P}(Z \in B|X = x) &= \int \mathbb{P}(Z \in B|X = x, T = t) dH(t|X = x) dt \\
&= \int \mathbb{P}(Z \in B|T = t) dH(t|X = x) dt \\
&= \int \mathbb{P}(Z \in B|T = t) \frac{dH(t|X = x)}{dH(t|X = x')} dH(t|X = x') \\
&\leq e^\alpha \int \mathbb{P}(Z \in B|T = t) dH(t|X = x') \\
&= e^\alpha \mathbb{P}(Z \in B|X = x').
\end{aligned}$$

9.3 Proof of Theorem 3.2

Our proof is adapted from an argument given in Theorem 5.1. of Blum et al. (2008). Let $r = 1$ so that $\mathcal{X} = [0, 1]$. Let $P = \delta_0$ where δ_0 denotes a point mass at 0. Then $P^n(X = X_{(0)}) = 1$ where $X_{(0)} \equiv \{0, \ldots, 0\}$. Assume that Q_n is consistent. Since $F(0) = 1$, it follows that for any $\delta > 0$, $\mathbb{P}(\widehat{F}_Z(0) > 1 - \delta) \to 1$. But since $\mathbb{P}(\cdot) = \mathbb{E}_P Q_n(\cdot|X)$ and since $P^n(X = X_{(0)}) = 1$, this implies that $Q_n(\widehat{F}_Z(0) > 1 - \delta | X = X_{(0)}) \to 1$.

Let $v > 0$ be any point in $[0, 1]$ such that $Q_n(Z = v|X = X_{(0)}) = 0$. Let $X_{(1)} = \{v, 0, \ldots, 0\}$, $X_{(2)} = \{v, v, 0, \ldots, 0\}$, ..., $X_{(n)} = \{v, v, \ldots, v\}$. By assumption, $Q_n(Z = X_{(j)}|X = X_{(0)}) = 0$ for all $j \geq 1$. Differential privacy implies that $Q_n(Z = X_{(j)}|X = X_{(1)}) = 0$ for all $j \geq 1$. Applying differential privacy again implies that $Q_n(Z = X_{(j)}|X = X_{(2)}) = 0$ for all $j \geq 1$. Continuing this way, we conclude that $Q_n(Z = X_{(j)}|X = X_{(n)}) = 0$ for all $j \geq 1$.

Next let $P = \delta_v$. Arguing as before, we know that $Q_n(\widehat{F}_Z(v) < 1 - \delta | X = X_{(n)}) \to 0$. And since $F(v-) = 0$ we also have that $Q_n(\widehat{F}_Z(v-) > \delta | X = X_{(n)}) \to 0$. Here, $F(v-) = \lim_{i \to \infty} F(v_i)$ where $v_1 < v_2 < \ldots$ and $v_i \to v$. Hence, for $j/n > 1 - \delta$, $Q_n(Z = X_{(j)}|X = X_{(n)}) > 0$ which is a contradiction. □

9.4 Proof of Theorem 4.1

Suppose that X differs from Y in at most one observation. Let $\widehat{f}$ denote the perturbed histogram $\widehat{f}_{m,\delta}$ based on X and let $\widehat{g}_{m,\delta}$ denote the histogram based on Y, such that X and Y differ in one entry. We also use $\widehat{p}_j(X)$ and $\widehat{p}_j(Y)$ for cell proportions. Note that $|\widehat{p}_j(X) - \widehat{p}_j(Y)| < 1/n$ by definition. It is clear that the maximum density ratio for a single draw x_i, or all i, occurs in one bin B_j. Now consider $\mathbf{x} = (x_1, \ldots, x_i)$ such that for all $i = 1, \ldots, k$, we have $x_i \in B_j \subset [0, 1]^r$ and the following bounds.

1. Let $\widehat{p}_j(Y) = 0$; then in order to maximize $\widehat{f}(\mathbf{x})/\widehat{g}(\mathbf{x})$, we let $\widehat{p}_j(X) = 1/n$ and obtain

$$\frac{\widehat{f}(\mathbf{x})}{\widehat{g}(\mathbf{x})} = \prod_{i=1}^{k} \frac{\widehat{f}_{m,\delta}(x_i)}{\widehat{g}_{m,\delta}(x_i)} \leq \left(\frac{(1-\delta)m(1/n) + \delta}{\delta} \right)^k = \left(\frac{(1-\delta)m}{n\delta} + 1 \right)^k;$$

2. Otherwise, we let $\widehat{p}_j(Y) \geq 1/n$, (as by definition of $\widehat{p}_j$, it takes z/n for non-negative integers z) and let $\widehat{p}_j(X) = \widehat{p}_j(Y) \pm 1/n$. Now it is clear that in order to maximize the density ratio at x, we may need to reverse the role of X and Y,

$$\begin{aligned}
\max\left(\frac{\widehat{g}(\mathbf{x})}{\widehat{f}(\mathbf{x})}, \frac{\widehat{f}(\mathbf{x})}{\widehat{g}(\mathbf{x})}\right) &\leq \max\left(\left(\frac{(1-\delta)m\widehat{p}_j + \delta}{(1-\delta)m(\widehat{p}_j - (1/n)) + \delta}\right)^k, \left(\frac{(1-\delta)m(\widehat{p}_j + 1/n) + \delta}{(1-\delta)m\widehat{p}_j + \delta}\right)^k\right), \\
&\leq \max\left(\frac{(1-\delta)m(1/n)}{(1-\delta)m(\widehat{p}_j - (1/n)) + \delta} + 1\right)^k \\
&\leq \left(\frac{(1-\delta)m}{n\delta} + 1\right)^k,
\end{aligned}$$

where the maximum is achieved when $\widehat{p}_j(Y) = 1/n$ and $\widehat{p}_j(X) = 0$, given a fixed set of parameters m, n, δ.

Thus we have

$$\sup_{\mathbf{x}\in([0,1]^r,\ldots,[0,1]^r)} \frac{\widehat{f}(\mathbf{x})}{\widehat{g}(\mathbf{x})} \leq \left(\frac{(1-\delta)m}{n\delta} + 1\right)^k,$$

and the theorem holds. □

9.5 Proof of Theorem 4.2

Recall that $\widehat{F}_Z$ denotes the empirical distribution function corresponding to $Z = (Z_1, \ldots, Z_k)$, where $Z_i \in [0,1]^r$ for all i are i.i.d. draws from density function $\widehat{f}_{m,\delta}(x)$ as in (5) given $X = (X_1, \ldots, X_n)$. Let U denote the uniform cdf on $[0,1]^r$. Given $X = (X_1, \ldots, X_n)$ drawn from a distribution whose cdf is F, let $\widehat{f}_m$ denote the histogram estimator on X and let $\widehat{F}_m(x) = \int_0^x \widehat{f}_m(s)ds$ and $\widehat{F}_{m,\delta}(x) = (1-\delta)\widehat{F}_m(x) + \delta U(x)$. Define $F_m(x) = \mathbb{E}(\widehat{F}_m(x))$ and $\bar{f}_m(x) = \mathbb{E}(\widehat{f}_m(x))$.

The Vapnik-Chervonenkis dimension of the class of sets of the form $\{(-\infty, x_1] \times \cdots \times (-\infty, x_r]$ is r and so by the standard Vapnik-Chervonenkis bound, we have for $\epsilon > 0$ that

$$\mathbb{P}\left(\sup_{t\in[0,1]^r} |\widehat{F}_X(t) - F(t)| > \epsilon\right) \leq 8n^r \exp\left\{-\frac{n\epsilon^2}{32}\right\} \leq \exp\left\{-\frac{n\epsilon^2}{64}\right\} \tag{18}$$

for large n. Hence, $\mathbb{E}\sup_{t\in[0,1]^r} |\widehat{F}_X(t) - F(t)| = O\left(\sqrt{\frac{r\log n}{n}}\right)$. Given X, we have $Z_1, \ldots, Z_k \sim$

$\widehat{F}_{m,\delta}$ and so $\mathbb{E}\sup_{[0,1]^r}|\widehat{F}_Z(t)-\widehat{F}_{m,\delta}(t)| = O\left(\sqrt{\frac{r\log k}{k}}\right)$. Thus,

$$\begin{aligned}
\mathbb{E}\sup_{x\in[0,1]^r}\left|\widehat{F}_Z(x)-F(x)\right| &\leq \mathbb{E}\sup_x|\widehat{F}_Z(x)-\widehat{F}_{m,\delta}(x)|+\mathbb{E}\sup_x|\widehat{F}_{m,\delta}(x)-F(x)| \\
&\leq \mathbb{E}\sup_x|\widehat{F}_Z(x)-\widehat{F}_{m,\delta}(x)|+\mathbb{E}\sup_x|\widehat{F}_m(x)-F(x)|+\delta \\
&\leq \mathbb{E}\sup_x|\widehat{F}_Z(x)-\widehat{F}_{m,\delta}(x)|+\mathbb{E}\sup_x|\widehat{F}_m(x)-F(x)|+\delta \\
&= O\left(\sqrt{\frac{r\log k}{k}}\right)+\mathbb{E}\sup_x|\widehat{F}_m(x)-F(x)|+\delta.
\end{aligned}$$

By the triangle inequality, we have for all $x\in[0,1]^r$,

$$\left|\widehat{F}_m(x)-F(x)\right| \leq \left|\widehat{F}_m(x)-F_m(x)\right|+\left|F_m(x)-F(x)\right|,$$

and hence

$$\begin{aligned}
\mathbb{E}\sup_{x\in[0,1]^r}\left|\widehat{F}_m(x)-F(x)|\right| &\leq \mathbb{E}\sup_{x\in[0,1]^r}\left|\widehat{F}_m(x)-F_m(x)\right|+\mathbb{E}\sup_{x\in[0,1]^r}|F_m(x)-F(x)| \\
&= O\left(\sqrt{\frac{r\log n}{n}}\right)+\mathbb{E}\sup_{x\in[0,1]^r}|F_m(x)-F(x)| \qquad (19)
\end{aligned}$$

where the last step follows from the VC bound as in (18) for $F_m(x)$.

Next we bound $\sup_{x\in[0,1]^r}|F_m(x)-F(x)|$. Now $F(x)=P(A)$ where $A=\{(s_1,\ldots,s_r):s_i\leq x_i, i=1,\ldots,r\}$. If $x=(j_1h,\ldots,j_rh)$ for some integers $j_1,\ldots,j_r$ then $F(x)-F_m(x)=0$. For x not of this form, let $\widetilde{x}=(j_1h,\ldots,j_rh)$ where $j_i=\lfloor x_i/h\rfloor$. Let $R=\{(s_1,\ldots,s_r):\ s_i\leq \widetilde{x}_i, i=1,\ldots,r\}$. So

$$\begin{aligned}
F(x)-F_m(x) &= P(A)-P_m(A)=P(R)-P_m(R)+P(A\setminus R)-P_m(A\setminus R) \\
&= P(A\setminus R)-P_m(A\setminus R) \qquad (20)
\end{aligned}$$

where $P_m(B)=\int_B dF_m(u)$ and the set $A\setminus R$ intersects at most rh/h^r number of cubes in $\{B_1,\ldots,B_m\}$, given that $\mathrm{Vol}(A\setminus R)\leq 1-(1-h)^r\leq rh$. Now by the Lipschitz condition (4),

we have $\sup_{x\in[0,1]^r} |p(x) - \bar{f}_m(x)| \leq Lh\sqrt{r}$ and

$$
\begin{aligned}
&|P(A \setminus R) - P_m(A \setminus R)| \\
&\quad \leq \text{ number of cubes intersecting}(A \setminus R) \times \text{maximum density discrepancy} \times \text{volume of cube} \\
&\quad \leq (rh/h^r) \cdot (Lh\sqrt{r}) \cdot h^r \leq Lr^{3/2}m^{-2/r}. \qquad (21)
\end{aligned}
$$

Thus we have by (19), (20) and (21)

$$\mathbb{E}\sup_x |\widehat{F}_m(x) - F(x)| = O\left(\sqrt{\frac{r\log n}{n}}\right) + Lr^{3/2}m^{-2/r}. \qquad (22)$$

Hence,

$$\mathbb{E}\sup_x |\widehat{F}_Z(x) - F(x)| = O\left(\sqrt{\frac{r\log k}{k}}\right) + O\left(\sqrt{\frac{r\log n}{n}}\right) + Lr^{3/2}m^{-2/r} + \delta.$$

Set $m \asymp n^{r/(6+r)}$, $k \asymp m^{4/r} = n^{4/(6+r)}$ and $\delta = (mk/n\alpha)$ we get for all n large enough, $\mathbb{E}\sup_x |\widehat{F}_Z(x) - F(x)| = O\left(\frac{\sqrt{\log n}}{n^{2/(6+r)}}\right)$. $\square$

9.6 Proof of Theorem 4.3

Let $\widehat{f}_Z$ be the histogram based on Z as in (8). Then

$$(\widehat{f}_Z(u) - p(u))^2 \preceq (1-\delta)^2(p(u) - \widehat{f}_m(u))^2 + \delta^2(p(u)-1)^2 + (\widehat{f}_{m,\delta}(u) - \widehat{f}_Z)^2$$

where $\preceq$ means less than, up to constants. Hence,

$$\mathbb{E}\int (\widehat{f}_Z(u) - p(u))^2 du \preceq R_m + \delta^2 + \mathbb{E}\int (\widehat{f}_{m,\delta}(u) - \widehat{f}_Z(u))^2 du$$

where R_m is the usual L_2 risk of a histogram under the Lipschitz condition (4), namely, $m^{-2/r} + m/n$. Conditional on X, $\widehat{f}_Z$ is an unbiased estimate of $\widehat{f}_m$ with integrated variance m/k. So,

$$\mathbb{E}\int(\widehat{f}_Z(u) - p(u))^2 du \preceq m^{-2/r} + \frac{m}{n} + \delta^2 + \frac{m}{k}.$$

Minimizing this, subject to (6) yields

$$m \asymp n^{r/(2r+3)}, k \asymp n^{(r+2)/(2r+3)}, \delta \asymp n^{-1/(2r+3)}$$

which yields $\mathbb{E}\int(\widehat{f}_Z(u) - p(u))^2 du = O(n^{-2/(2r+3)})$. $\square$

9.7 Proof of Theorem 4.4

(1) Note that $p - \widehat{f}_Z = p - \widetilde{f} + \widetilde{f} - \widehat{f}_Z = p - \widetilde{f} + O_P\left(\frac{m}{k}\right)$. When $k \geq n$, the latter error is lower order than the other terms and may be ignored. Now,

$$p(x) - \widetilde{f}(x) = p(x) - \widehat{f}_m(x) + \widehat{f}_m(x) - \widetilde{f}(x).$$

Thus

$$\int(p(x) - \widetilde{f}(x))^2 dx \preceq \int(p(x) - \widehat{f}_m(x))^2 dx + \int(\widehat{f}_m(x) - \widetilde{f}(x))^2 dx.$$

The expected value of the first term is the usual risk, namely, $O(m^{-2/r} + m/n)$.

For the second term, we proceed as follows. Let $\widehat{p}_j = C_j/n$ and

$$\widehat{q}_j = \frac{(C_j + \nu_j)_+}{\sum_{s=1}^m (C_s + \nu_s)_+}.$$

We claim that

$$\max_j |\widehat{q}_j - \widehat{p}_j| = O\left(\frac{\log m}{n}\right)$$

almost surely, for all large n. We have

$$\widehat{q}_j = \frac{(C_j + \nu_j)_+}{n} \left(\frac{n}{\sum_{s=1}^m (C_s + \nu_s)_+} \right) = \frac{(C_j + \nu_j)_+}{n} \frac{1}{R_n}$$

where $R_n = (\sum_{s=1}^m (C_s + \nu_s)_+)/n$. Now

$$\widehat{p}_j - \frac{|\nu_j|}{n} \leq \widehat{p}_j + \frac{\nu_j}{n} = \frac{(C_j + \nu_j)}{n} \leq \frac{(C_j + \nu_j)_+}{n} \leq \widehat{p}_j + \frac{|\nu_j|}{n}.$$

Therefore,

$$\left| \frac{(C_j + \nu_j)_+}{n} - \widehat{p}_j \right| \leq \frac{|\nu_j|}{n} \leq \frac{M}{n}$$

where $M = \max\{|\nu_1|, \ldots, |\nu_m|\}$. Let $A > 0$. The density for ν_j has the form $f(\nu) = (\beta/2)e^{-\beta|\nu|}$. So,

$$\mathbb{P}(M > A \log m) \leq m\mathbb{P}(|\nu_j| > A \log m) = \beta m \int_{A \log m}^{\infty} e^{-\beta|\nu|} d\nu = \frac{1}{m^{A\beta - 1}}.$$

By choosing A large enough we have that $M < A \log m$ a.s. for large n, by the Borel-Cantelli lemma. Therefore,

$$\left| \frac{(C_j + \nu_j)_+}{n} - \widehat{p}_j \right| \leq \frac{\log m}{n}$$

Now we bound R_n. We have

$$1 - \frac{\sum_s |\nu_s|}{n} \leq 1 + \frac{\sum_s \nu_s}{n} \leq R_n = \frac{\sum_{s=1}^m (C_s + \nu_s)_+}{n} \leq 1 + \frac{\sum_s |\nu_s|}{n}$$

so that

$$|R_n - 1| \leq \frac{\sum_s |\nu_s|}{n} \leq \frac{Mm}{n} = O\left(\frac{m \log m}{n} \right) \quad a.s.$$

Therefore, $1/R_n = (1 + O(m \log m/n))$ and thus

$$\begin{aligned} \widehat{q}_j &= \left(\widehat{p}_j + O\left(\frac{\log m}{n} \right) \right) \left(1 + O\left(\frac{m \log m}{n} \right) \right) \\ &= \widehat{p}_j + \widehat{p}_j \, O\left(\frac{m \log m}{n} \right) + O\left(\frac{\log m}{n} \right) + O\left(\frac{m(\log m)^2}{n^2} \right). \end{aligned}$$

Next we claim that $\widehat{p}_j = O(1/m)$ a.s. To see this, note that $p_j \leq C/m$, by definition of C: $1 \leq C = \sup_x p(x) < \infty$. Hence, by Bernstein's inequality,

$$\begin{aligned}\mathbb{P}\left(\widehat{p}_j > \frac{2C}{m}\right) &= \mathbb{P}\left(\widehat{p}_j - p_j > \frac{2C}{m} - p_j\right) \leq \exp\left\{-\frac{1}{2}\frac{n((2C/m)-p_j)^2}{p_j + \frac{1}{3}((2C/m)-p_j)}\right\} \\ &\leq \exp\left\{-\frac{1}{2}\frac{nC^2/m^2}{(4C/3m)}\right\} = e^{-3nC/(8m)} \leq \frac{1}{n^2}\end{aligned}$$

for all $n \geq 16m\log n/3C$; Thus $\widehat{p}_j = O(1/m)$ a.s. for all large n. Thus, $\widehat{q}_j - \widehat{p}_j = O(\log m/n)$ almost surely for all large n. Hence,

$$\mathbb{E}\int (\widehat{f}_m(x) - \widetilde{f}(x))^2 dx = O\left(\frac{m\log m}{n}\right)^2.$$

So the risk is

$$O\left(m^{-2/r} + \frac{m}{n} + \left(\frac{m\log m}{n}\right)^2\right) = O\left(m^{-2/r} + \frac{m}{n}\right),$$

for $n \geq m\log^2 m$. This is the usual risk. Hence, we can choose $m \asymp n^{r/(2+r)}$ to achieve risk $n^{-2/(2+r)}$ for all n large enough.

(2) Let $\widehat{F}_m$ be the cdf based on the original histogram and let $\widetilde{F}_m$ be the cdf based on the perturbed histogram. We have

$$\begin{aligned}\mathbb{E}\sup_x |F(x) - \widehat{F}_Z(x)| &\leq \mathbb{E}\sup_x |F(x) - \widehat{F}_m(x)| + \mathbb{E}\sup_x |\widehat{F}_m(x) - \widetilde{F}_m(x)| + \mathbb{E}\sup_x |\widetilde{F}_m(x) - \widehat{F}_Z(x)| \\ &\leq \mathbb{E}\sup_x |F(x) - \widehat{F}_m(x)| + \mathbb{E}\sup_x |\widehat{F}_m(x) - \widetilde{F}_m(x)| + O\left(\sqrt{\frac{r\log k}{k}}\right).\end{aligned}$$

Since we may take k as large as we like, we can make the last term arbitrarily small. From (22),

$$\mathbb{E}\sup_x |F(x) - \widehat{F}_m(x)| = O\left(\sqrt{\frac{r\log n}{n}}\right) + Lr^{3/2}m^{-2/r}.$$

Let $\widehat{f}(x) = h^{-r}\sum_{j=1}^m \widehat{p}_j I(x \in B_j)$ and Let $\widetilde{f}(x) = h^{-r}\sum_{j=1}^m \widehat{q}_j I(x \in B_j)$. Let $x' = (u_1 h, \ldots, u_r h)$ where $u_i = \lceil x_i/h \rceil, \forall i = 1, \ldots, r$. Recall that $B_1, \ldots, B_m$ are the m bins of

$\mathcal{X}$ with sides of length of h. Let B_x denote the cube with the left-most corner being 0 and the right-most corner being x. Then for all x, we have

$$\begin{aligned}\left|\widehat{F}_m(x) - \widetilde{F}_m(x)\right| &= \left|\int_0^x \widehat{f}(s) - \widetilde{f}(s)ds\right| \leq \int_0^x \left|\widehat{f}(s) - \widetilde{f}(s)\right| ds \\ &\leq \int_0^{x'} \left|\widehat{f}(s) - \widetilde{f}(s)\right| ds \\ &= \sum_{\ell: B_\ell \subseteq B_{x'}} |\widehat{p}_\ell - \widehat{q}_\ell| \leq \sum_{\ell=1}^m |\widehat{p}_\ell - \widehat{q}_\ell|\end{aligned}$$

where we use the fact that there are at most m cubes. Hence,

$$\mathbb{E} \sup_{x\in[0,1]^r} |\widehat{F}_m(x) - \widetilde{F}_m(x)| \leq \frac{m \log m}{n}$$

where we use the fact that $\max_j |\widehat{p}_j - \widehat{q}_j| = O(\log m/n)$ a.s. So,

$$\mathbb{E} \sup_x |F(x) - \widehat{F}_Z(x)| = O\left(\sqrt{\frac{r \log n}{n}}\right) + Lr^{3/2} m^{-2/r} + O\left(\frac{m \log m}{n}\right).$$

Setting $m \asymp n^{r/(2+r)}$ yields

$$\mathbb{E} \sup_x |F(x) - \widehat{F}_Z(x)| = O\left(\min\left(\frac{\log n}{n^{2/(2+r)}}, \sqrt{\frac{\log n}{n}}\right)\right)$$

Hence for $r = 1$, the rate is $O\left(\sqrt{\frac{\log n}{n}}\right)$. For $r \geq 2$, the rate is dominated by the first term inside $O()$, and hence the rate is $O\left(\log n \times n^{-2/(2+r)}\right)$. □

9.8 Proof of Theorem 5.3

Let $B_\epsilon = \left\{u = (u_1, \ldots, u_k) : \ \rho(F, \widehat{F}_u) \leq \epsilon\right\}$ where $\widehat{F}_u$ is the empirical distribution based on $u = (u_1, \ldots, u_k) \in \mathcal{X}^k$. Also, let $A_n = \{\rho(\widehat{F}_X, F) \leq \epsilon_n/16\}$. For notational simplicity set

$\Delta = \Delta_{n,k}$. Then

$$\begin{aligned}
\mathbb{P}\left(\rho(F,\widehat{F}_Z) > \epsilon_n\right) &= \mathbb{P}\left(\rho(F,\widehat{F}_Z) > \epsilon_n, A_n\right) + \mathbb{P}\left(\rho(F,\widehat{F}_Z) > \epsilon_n, A_n^c\right) \\
&\leq \mathbb{P}\left(\rho(F,\widehat{F}_Z) > \epsilon_n, A_n\right) + \mathbb{P}\left(A_n^c\right) \\
&= \mathbb{P}\left(\rho(F,\widehat{F}_Z) > \epsilon_n, A_n\right) + O\left(\frac{1}{n^c}\right). \qquad (23)
\end{aligned}$$

By the triangle inequality $\rho(\widehat{F}_u, \widehat{F}_X) \geq \rho(\widehat{F}_u, F) - \rho(\widehat{F}_X, F)$. Then,

$$\begin{aligned}
\int_{B_\epsilon^c} g_x(u)du &= \int_{B_\epsilon^c} \exp\left(\frac{-\alpha\rho(\widehat{F}_X,\widehat{F}_u)}{2\Delta}\right)du \\
&\leq \int_{B_\epsilon^c} \exp\left(\frac{-\alpha(\rho(\widehat{F}_u,F) - \rho(\widehat{F}_X,F))}{2\Delta}\right)du \\
&= \exp\left(\frac{\alpha\rho(\widehat{F}_X,F)}{2\Delta}\right)\int_{B_\epsilon^c} \exp\left(\frac{-\alpha\rho(\widehat{F}_u,F)}{2\Delta}\right)du \\
&\leq \exp\left(\frac{\alpha\rho(\widehat{F}_X,F)}{2\Delta}\right)\exp\left(\frac{-\alpha\epsilon}{2\Delta}\right)\int_{B_\epsilon^c} du \\
&\leq \exp\left(\frac{\alpha\rho(\widehat{F}_X,F)}{2\Delta}\right)\exp\left(\frac{-\alpha\epsilon}{2\Delta}\right).
\end{aligned}$$

By the triangle inequality, we also have $\rho(\widehat{F}_u, \widehat{F}_X) \leq \rho(\widehat{F}_u, F) + \rho(\widehat{F}_X, F)$ and

$$\begin{aligned}
\int g_x(u)du &\geq \int_{B_{\epsilon/2}} g_x(u)du = \int_{B_{\epsilon/2}} \exp\left(\frac{-\alpha\rho(\widehat{F}_X,\widehat{F}_u)}{2\Delta}\right)du \\
&\geq \exp\left(\frac{-\alpha\rho(\widehat{F}_X,F)}{2\Delta}\right)\int_{B_{\epsilon/2}} \exp\left(\frac{-\alpha\rho(F,\widehat{F}_u)}{2\Delta}\right)du \\
&\geq \exp\left(\frac{-\alpha\rho(\widehat{F}_X,F)}{2\Delta}\right)\exp\left(\frac{-\alpha\epsilon}{4\Delta}\right)\int_{B_{\epsilon/2}} du \\
&= \exp\left(\frac{-2\alpha\rho(\widehat{F}_X,F) - \alpha\epsilon}{4\Delta}\right)\int_{B_{\epsilon/2}} \frac{p(u_1)\cdots p(u_k)}{p(u_1)\cdots p(u_k)}du \\
&\geq \frac{\exp\left(\frac{-2\alpha\rho(\widehat{F}_X,F)-\alpha\epsilon}{4\Delta}\right)}{(\sup_x p(x))^k}\mathbb{P}\left(\rho(F,\widehat{G}) \leq \epsilon/2\right)
\end{aligned}$$

where $\widehat{G}$ is the empirical cdf from a sample of size k drawn from P. Thus we have

$$\int_{B_\epsilon^c} h(u|x)du \;\leq\; \frac{(\sup_x p(x))^k \exp\left(\frac{\alpha\rho(\widehat{F}_X,F)}{\Delta}\right)\exp\left(\frac{-\alpha\epsilon}{4\Delta}\right)}{\mathbb{P}\left(\rho(F,\widehat{G})\leq\epsilon/2\right)}.$$

Thus, from (23),

$$\begin{aligned}\mathbb{P}\left(\rho(F,\widehat{F}_Z)>\epsilon\right) &\leq \mathbb{P}\left(\rho(\widehat{F}_X,F)\geq\frac{\epsilon}{16}\right)+\frac{(\sup_x p(x))^k\exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}\left(\rho(F,\widehat{G})\leq\epsilon/2\right)}\\ &= \frac{(\sup_x p(x))^k\exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}\left(\rho(F,\widehat{G})\leq\epsilon/2\right)}+O\left(\frac{1}{n^c}\right).\end{aligned}$$

Thus the theorem holds. □

9.9 Proof of Lemma 5.1

Proof of Lemma 5.1. We start with KS, By the triangle inequality, we have for all $z\in\mathcal{X}^k$ and for all $x,y\in\mathcal{X}^n$,

$$\left|\rho(\widehat{F}_x,\widehat{F}_z)-\rho(\widehat{F}_y,\widehat{F}_z)\right| \;\leq\; \rho(\widehat{F}_x,\widehat{F}_y).$$

Notice that changing one entry in x will change $\widehat{F}_x(t)$ by at most $\frac{1}{n}$ at any t by definition, that is,

$$\sup_{t\in[0,1]^r}|\widehat{F}_x(t)-\widehat{F}_y(t)| \;=\; \frac{1}{n}.$$

Thus the conclusion holds for the KS-distance. □

9.10 Proof of Theorem 5.4

We need the following small ball result; see Li and Shao (2001).

Theorem 9.1. *Let $r\geq 3$, and $\{X_t, t\in[0,1]^r\}$ be the Brownian sheet. Then there exists $0<C_r<$*

∞ *such that for all* $0 < \epsilon \leq 1$,

$$\log \mathbb{P}\left(\sup_{t\in[0,1]^r} |X_t| \leq \epsilon\right) \geq -C_r \epsilon^{-2} \log^{2r-1}(1/\epsilon)$$

where C_r *depends only on* r*. The same bound holds for a Brownian bridge.*

Proof of theorem 5.4. The Vapnik-Chervonenkis dimension of the class of sets of the form $\{(-\infty, x_1] \times \cdots \times (-\infty, x_r]$ is r and so by the standard Vapnik-Chervonenkis bound, we have for ϵ_n, k_n as specified in the theorem statement,

$$\begin{aligned}
\mathbb{P}\left(\sup_{[0,1]^r} |\widehat{F}_X(t) - F(t)| > \frac{\epsilon_n}{16}\right) &\leq 8n^r \exp\left\{-\frac{n(\epsilon_n/16)^2}{32}\right\} \\
&\leq 8\exp\left\{-c_5\left(\frac{B}{3\alpha}\right)^{2/3} n^{1/3} + r\log n\right\} \\
&= 8\exp\left\{-c_6\sqrt{k_n}\left(\frac{B}{3\alpha}\right) + c_7 r\log k_n\right\} \\
&= 8\exp\left\{-C_2\sqrt{k_n}\left(\frac{B}{3\alpha}\right)\right\} \qquad (24)
\end{aligned}$$

for some constants $c_5, c_6, c_7, C_2 > 0$ for n large enough. Thus (10) holds. Now we compute the small ball probability. Note that $\sqrt{k}(\widehat{F}_k - F)$ converges to a Brownian bridge B_k on $[0,1]^r$. More precisely, from Csörgő and Révész (1975) there exist a sequence of Brownian bridges B_k such that

$$\sup_t |\sqrt{k}(\widehat{F}_k - F)(t) - B_k(t)| = O\left(\frac{(\log k)^{3/2}}{k^\gamma}\right) \quad \text{a.s.} \qquad (25)$$

where $\gamma = 1/(2(r+1))$. It is clear that the RHS of (25) is $o(1)$ a.s. given a fixed r. Hence we have for $k = k_n$ and ϵ_n as chosen in the theorem statement, and for all $\epsilon \geq \epsilon_n$, it holds that

$$\begin{aligned}
\log \mathbb{P}(\sup_t |\widehat{F}_Z(t) - F(t)| \leq \epsilon/2) &= \log \mathbb{P}(\sup_t \sqrt{k}|\widehat{F}_Z(t) - F(t)| \leq \sqrt{k}\epsilon/2) \\
&\geq \log \mathbb{P}\left(\sup_t |B_k(t)| \leq \sqrt{k}\epsilon - O\left(k^{-\gamma}(\log k)^{3/2}\right)\right) \qquad (26) \\
&\geq \log \mathbb{P}\left(\sup_t |B_k(t)| \leq \frac{\sqrt{k}\epsilon}{4}\right) \qquad (27)
\end{aligned}$$

for all large n, where (26) follows from (25) and (27) holds given that $\sqrt{k}\epsilon \geq \sqrt{k_n}\epsilon_n \geq c$ for some constant $c > 1/2$ due to our choice of k_n and ϵ_n. Also, $\Delta \leq 1/n$ for KS distance. Hence, by Theorem 5.3 and (24), we have for $B = \log \sup_x p(x) > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n\right) \\
&\leq \; C_0 \exp\left\{-n\left(\frac{3\alpha\epsilon_n}{16} - \frac{Bk_n}{n} - \frac{C_1|\log(\sqrt{k_n}\epsilon_n/4)|^{2r-1}}{nk_n\epsilon_n^2}\right)\right\} + 8\exp\left\{-C_2\frac{B\sqrt{k_n}}{3\alpha}\right\} \\
&\leq \; C_0 \exp(-C_3 Bk_n/2) + 8\exp\left\{-C_2\left(\frac{B}{3\alpha}\right)\sqrt{k_n}\right\} \to 0 \qquad (28)
\end{aligned}$$

for some constants C_0, C_1, C_2 and C_3, where (28) holds when we take w.l.o.g. $k_n = \frac{1}{16}\left(\frac{3\alpha}{B}\right)^{2/3} n^{2/3}$ and $\epsilon_n \geq 2\left(\frac{B}{3\alpha}\right)^{1/3} n^{-1/3}$, given that $\epsilon_n \geq 2\left(\frac{B}{3\alpha}\right)^{1/3} n^{-1/3} = \frac{32k_nB}{3n\alpha}$ and hence $\frac{3\alpha\epsilon_n}{16} \geq \frac{2Bk_n}{n}$. Thus the result follows. □

Remark 9.2. *The constants taken in the proof are arbitrary; indeed, when we take $k_n = C_4\left(\frac{3\alpha}{B}\right)^{2/3} n^{2/3}$ and $\epsilon_n = 32C_4\left(\frac{B}{3\alpha}\right)^{1/3} n^{-1/3}$ with some constant $C_4 \geq 1/16$,* (28) *will hold with slightly different constants C_2, C_3. For k_n and ϵ_n as chosen above, it holds that $\sqrt{k_n}\epsilon_n \asymp 1$.*

9.11 Proofs for Lemma 6.1 and Theorem 6.2

Throughout this section, we let $\widehat{p}_X$ denote the estimator as defined in (14), which is based on a sample of size n drawn independently from F; Similarly, we let $\widehat{p}_k$ denote the same estimator based on an i.i.d. sample $(Y_1, \ldots, Y_k)$ of size k drawn from F, with $m_k = k^{1/(2\gamma+1)}$ replacing m_n and $\widehat{\beta}_j = k^{-1}\sum_{i=1}^k \psi_j(Y_i)$ in (14). We let $\widehat{p}_Z$ denote the estimator as in (16), based on an i.i.d. sample $Z = (Z_1, \ldots, Z_k)$ of size k drawn from $g_x(z)$ as in (17).

Proof of Lemma 6.1. Without loss of generality, let $X = (x, X_2, \ldots, X_n)$ and $Y = (y, X_2, \ldots, X_n)$ so that $\delta(X, Y) = 1$ and let $Z \in \mathcal{X}^k$. Recall that

$$\begin{aligned}
\xi(X, Z) &= \left(\int \left(\widehat{p}_X(x) - \widehat{p}_Z(x)\right)^2 dx\right)^{1/2}, \\
\xi(Y, Z) &= \left(\int \left(\widehat{p}_Y(x) - \widehat{p}_Z(x)\right)^2 dx\right)^{1/2}.
\end{aligned}$$

In particular, let us define $u = \widehat{p}_X - \widehat{p}_Z$ and $v = \widehat{p}_Y - \widehat{p}_Z$ and thus

$$\begin{aligned} |\xi(X,Z) - \xi(Y,Z)| &= \left| \left(\int \left(\widehat{p}_X(x) - \widehat{p}_Z(x)\right)^2 dx \right)^{1/2} - \left(\int \left(\widehat{p}_Y(x) - \widehat{p}_Z(x)\right)^2 dx \right)^{1/2} \right| \\ &= \left| \|u\|_{\ell_2} - \|v\|_{\ell_2} \right| \leq \|u - v\|_{\ell_2} \\ &= \|\widehat{p}_X - \widehat{p}_Z - (\widehat{p}_Y - \widehat{p}_Z)\|_{\ell_2} = \|\widehat{p}_X - \widehat{p}_Y\|_{\ell_2} \leq \frac{2c_0^2 m_n}{n}, \end{aligned}$$

where the first inequality is due to the triangle inequality for the $\|.\|_{\ell_2}$ and the last step is due to

$$\begin{aligned} |\widehat{p}_X(x) - \widehat{p}_Y(x)| &= \frac{1}{n} \left| \sum_{j=1}^{m_n} \left(\sum_{i=1}^{n} \psi_j(X_i) - \sum_{i=1}^{n} \psi_j(Y_i) \right) \psi_j(x) \right| \\ &= \frac{1}{n} \left| \sum_{j=1}^{m_n} \left(\psi_j(X_1) - \psi_j(Y_1) \right) \psi_j(x) \right| \\ &\leq \frac{1}{n} \sum_{j=1}^{m_n} (|\psi_j(X_1)| + |\psi_j(Y_1)|)|\psi_j(x)| \leq \frac{2c_0^2 m_n}{n}. \end{aligned}$$

Hence $\Delta \leq \frac{2c_0^2 m_n}{n}$. □

Proof of Theorem 6.2. For $u = (u_1, \ldots, u_k) \in \mathcal{X}^k$, we let

$$\widehat{p}_u(x) = 1 + \sum_{j=1}^{m_k} \widehat{\beta}_j \psi_j(x),$$

where $m_k = k^{\frac{1}{2\gamma+1}}$ and $\widehat{\beta}_j = k^{-1} \sum_{i=1}^{k} \psi_j(u_i)$.

Let $\widehat{F}_u$ be the empirical distribution based on u. Our proof follows that of Theorem 5.3, with

$$\rho(F, \widehat{F}_u) = \|p - \widehat{p}_u\|_{\ell_2} \text{ and } \rho(F_X, \widehat{F}_u) = \|\widehat{p}_X - \widehat{p}_u\|_{\ell_2}$$

as defined in (15) for $X = (X_1, \ldots, X_n)$. Now

$$B_\epsilon = \Big\{ u = (u_1, \ldots, u_k) : \ \|p - \widehat{p}_u\|_{\ell_2} < \epsilon \Big\}.$$

Thus the corresponding triangle inequalities that we use to replace that in Theorem 5.3 are:

$$\begin{aligned} \|\widehat{p}_u - \widehat{p}_X\|_{\ell_2} &\geq \|\widehat{p}_u - p\|_{\ell_2} - \|\widehat{p}_X - p\|_{\ell_2} \text{ and} \\ \|\widehat{p}_u - \widehat{p}_X\|_{\ell_2} &\leq \|\widehat{p}_u - p\|_{\ell_2} + \|p - \widehat{p}_X\|_{\ell_2}. \end{aligned}$$

Standard risk calculations show that (10) holds for some $c > 0$ with $\rho(F, \widehat{F}_X)$ being replaced with $\|\widehat{p}_X - p\|_{\ell_2}$. That is, by Markov's inequality,

$$\mathbb{P}(\|\widehat{p}_X - p\|_{\ell_2} > \epsilon) \leq \frac{\mathbb{E}\,\|\widehat{p}_X - p\|_{\ell_2}^2}{\epsilon^2}$$

and (10) follows from the polynomial decay of the mean squared error $\mathbb{E}||\widehat{p}_X - p||^2$. Thus, from (23), for $\widehat{p}_Z = \widehat{p}^*$ as in (16),

$$\begin{aligned} \mathbb{P}\left(\|p - \widehat{p}_Z\|_{\ell_2}) > \epsilon\right) &\leq \mathbb{P}\left(\|\widehat{p}_X - p\|_{\ell_2} \geq \frac{\epsilon}{16}\right) + \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}\left(\|p - \widehat{p}_k\|_{\ell_2} \leq \epsilon/2\right)} \\ &= \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}\left(\|p - \widehat{p}_k\|_{\ell_2} \leq \epsilon/2\right)} + O\left(\frac{1}{n^c}\right). \end{aligned}$$

We need to compute the small ball probability. Recall that $\widehat{p}_k$ denote the estimator based on a sample of size k. By Parseval's relation,

$$\int (p(x) - \widehat{p}_k(x))^2 dx = \sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2 + \sum_{m_k+1}^{\infty} \beta_j^2 \leq \sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2 + ck^{-2\gamma/(2\gamma+1)}.$$

Let $U_i = (\psi_1(X_i) - \beta_1, \ldots, \psi_{m_k}(X_i) - \beta_{m_k})^T$ and $Y_i = \Sigma_k^{-1/2} U_i$ where Σ_k is the covariance matrix of U_i. Hence, Y_i has mean 0 and identity covariance matrix. Let λ_k denote the largest eigenvalue of Σ_k. From Lemma 9.3 below, $\lambda = \limsup_{k\to\infty} \lambda_k < \infty$. Let $Q = \sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2$ and let $S = k^{-1/2} \sum_{i=1}^{k} Y_i$. Then, for all large k, and any $\delta > 0$,

$$\mathbb{P}(Q \leq \delta^2) = \mathbb{P}(S^T \Sigma_k S \leq k\delta^2) \geq \mathbb{P}\left(S^T S \leq \frac{k\delta^2}{\lambda_k}\right) \geq \mathbb{P}\left(S^T S \leq \frac{k\delta^2}{2\lambda}\right).$$

From Theorem 1.1 of Bentkus (2003) we have that

$$\sup_c \left|\mathbb{P}\left(S^T S \le c\right) - \mathbb{P}(\chi^2_{m_k} \le c)\right| = O\left(\sqrt{\frac{m_k^3}{k}}\right) = O\left(k^{-(\gamma-1)/(2\gamma+1)}\right).$$

Next we use the fact (see Rohde and Duembgen (2008) for example) that $\mathbb{P}(\chi^2_m \le m + a) \ge 1 - e^{-a^2/(4(m+a))}$. Let $k = \sqrt{n}$, $\epsilon_n = c_1 n^{-\gamma/(2\gamma+1)}$ where $c_1 \ge 4(2\lambda+1)(C^2+1)$

$$a = \frac{k(\epsilon_n/4 - C^2 k^{-2\gamma/(2\gamma+1)})}{2\lambda} - m_k \ge (C^2+1)n^{1/2(2\gamma+1)} - m_k \ge C^2 m_k,$$

since $m_k = k^{\frac{1}{2\gamma+1}} = n^{1/2(2\gamma+1)}$. We see that for all large k

$$\begin{aligned}
\mathbb{P}\left(\|p - \widehat{p}_k\|_{\ell_2} \le \frac{\sqrt{\epsilon_n}}{2}\right) &= \mathbb{P}\left(\int (p(x) - \widehat{p}_k(x))^2 dx \le \frac{\epsilon_n}{4}\right) \\
&\ge \mathbb{P}\left(\sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2 \le \frac{\epsilon_n}{4} - C^2 k^{-2\gamma/(2\gamma+1)}\right) \\
&= \mathbb{P}\left(\chi^2_{m_k} \le \frac{k(\epsilon_n/4 - C^2 k^{-2\gamma/(2\gamma+1)})}{2\lambda}\right) - O\left(k^{-(\gamma-1)/(2\gamma+1)}\right) \\
&\ge 1 - \exp\left(\frac{-a^2}{4(m_k+a)}\right) - O\left(k^{-(\gamma-1)/(2\gamma+1)}\right) \\
&\ge \frac{1}{2} - O\left(k^{-(\gamma-1)/(2\gamma+1)}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}\left(\|p - \widehat{p}_Z\|_{\ell_2}) > \sqrt{\epsilon_n}\right) &\le \mathbb{P}\left(\|\widehat{p}_X - p\|_{\ell_2} \ge \frac{\sqrt{\epsilon_n}}{16}\right) + \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\sqrt{\epsilon_n}}{16\Delta}\right)}{\mathbb{P}\left(\|p - \widehat{p}_k\|_{\ell_2} \le \sqrt{\epsilon_n}/2\right)} \\
&= \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\sqrt{\epsilon_n}}{16\Delta}\right)}{\mathbb{P}\left(\|p - \widehat{p}_k\|_{\ell_2} \le \sqrt{\epsilon_n}/2\right)} + O\left(\frac{1}{n^c}\right) \\
&\le \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha n\sqrt{\epsilon_n}}{32 c_0^2 m_n}\right)}{\mathbb{P}\left(\|p - \widehat{p}_k\|_{\ell_2} \le \sqrt{\epsilon_n}/2\right)} + O\left(\frac{1}{n^c}\right)
\end{aligned}$$

and so for $\gamma > 1$,

$$\begin{aligned}
\mathbb{P}(\int (\widehat{p}_Z - p)^2 \leq \epsilon_n) &\leq c_2 \exp\left(k \log \sup_x p(x)\right) \exp\left(\frac{-3\sqrt{c_1}\alpha n}{n^{1/(2\gamma+1)} n^{\gamma/2(2\gamma+1)}}\right) \\
&= c_2 \exp\left(n^{1/2} \log \sup_x p(x) - \alpha c_3 n^{\left(\frac{3\gamma}{2(2\gamma+1)}\right)}\right) \\
&= c_2 \exp\left(-\alpha c_4 n^{\left(\frac{3\gamma}{2(2\gamma+1)}\right)}\right) \to 0,
\end{aligned}$$

as $n \to \infty$ since $\frac{3\gamma}{2(2\gamma+1)} > 1/2$, where c_2, c_3, c_4 are some constants. Hence the theorem holds. □

Lemma 9.3. *Let* $\lambda = \limsup_{k\to\infty} \lambda_k$. *Then* $\lambda < \infty$.

Proof. Recall that the orthonormal basis is $\psi_0, \psi_1, \ldots,$ where $\psi_0 = 1$ and $\psi_j(x) = \sqrt{2}\cos(\pi j x)$. Also $p(x) = 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x)$ and $\sum_j \beta_j^2 j^{2\gamma} < \infty$. Note that $\sum_{j=1}^{\infty} |\beta_j|^k = O(1)$ for $k \geq 1$; see Efromovich (1999). Note that Σ_k is the covariance matrix of $\widehat{\beta}$ times n. We will use the standard identities $\cos^2(u) = (1 + \cos(2u))/2$ and $\cos(u)\cos(v) = \frac{\cos(u-v)+\cos(u+v)}{2}$. It follows that $\psi_j^2(x) = 1 + \frac{1}{\sqrt{2}}\psi_{2j}(x)$ and $\psi_j(x)\psi_k(x) = \frac{\psi_{j-k}(x)+\psi_{j+k}(x)}{\sqrt{2}}$. Now $\mathbb{E}(\widehat{\beta}_j) = \beta_j$. And

$$n\mathrm{Var}(\widehat{\beta}_j) = \mathrm{Var}(\psi_j(X)) = \int \psi_j^2(x)p(x)dx - \beta_j^2.$$

Now $\int \psi_j^2(x)p(x)dx = \int \psi_j^2(x)(1 + \sum_{\ell=1}^{\infty} \beta_\ell \psi_\ell(x))dx = 1 + \sum_{\ell=1}^{\infty} \beta_\ell \int \psi_\ell(x)\psi_j^2(x)dx = 1 + \frac{1}{2}\sum_{\ell=1}^{\infty} \beta_\ell \int \psi_\ell(x)\left(1 + \frac{\psi_{2j}(x)}{\sqrt{2}}\right)dx = 1 + \frac{\beta_{2j}}{\sqrt{2}}$. Thus, $\Sigma_{jj} = 1 + \frac{\beta_{2j}}{\sqrt{2}} - \beta_j^2$. Now consider $j \neq k$.

Then

$$\begin{aligned}
\mathbb{E}(\psi_j(X)\psi_k(X)) &= \int \psi_j(x)\psi_k(x)p(x)dx \\
&= \sum_\ell \beta_\ell \int \psi_j(x)\psi_k(x)dx \\
&= \beta_j \int \psi_j^2(x)\psi_k(x)dx + \beta_k \int \psi_k^2(x)\psi_j(x)dx + \sum_{\ell\neq j,k} \beta_\ell \int \psi_j(x)\psi_k(x)\psi_\ell(x)dx \\
&= \frac{\beta_j}{\sqrt{2}} \int \psi_{2j}(x)\psi_k(x)dx + \frac{\beta_k}{\sqrt{2}} \int \psi_{2k}(x)\psi_j(x)dx \\
&\quad + \frac{1}{\sqrt{2}} \sum_{\ell\neq j,k} \beta_\ell \int (\psi_{j-k}(x) + \psi_{j+k}(x))\psi_\ell(x) \\
&= \frac{\beta_j}{\sqrt{2}} I(2j = k) + \frac{\beta_k}{\sqrt{2}} I(2k = j) \\
&\quad + \frac{\beta_\ell}{\sqrt{2}} I(\ell = |j-k| \ \&\ j \neq 2k) + \frac{\beta_\ell}{\sqrt{2}} I(\ell = j+k) \\
&= \frac{\beta_k}{\sqrt{2}} I(2k = j) + \frac{\beta_{|j-k|}}{\sqrt{2}} I(j \neq 2k) + \frac{\beta_{j+k}}{\sqrt{2}} \\
&= \frac{\beta_{|j-k|}}{\sqrt{2}} + \frac{\beta_{j+k}}{\sqrt{2}},
\end{aligned}$$

where we used the fact that $\psi_{-j}(x) = \psi_j(x)$ for all $j = 1, 2, \ldots$ and $\int \psi_j(x)dx = 0$ for all $j > 0$. So, we have for all $j \in \{1, \ldots, p\}$,

$$\begin{aligned}
\sum_{k=1}^{p} |\Sigma_{jk}| &= |\Sigma_{jj}| + \sum_{j\neq k} \left| \frac{\beta_{|j-k|}}{\sqrt{2}} + \frac{\beta_{j+k}}{\sqrt{2}} - \beta_j\beta_k \right| \\
&\leq 1 + \left| \frac{\beta_{2j}}{\sqrt{2}} \right| + |\beta_j| \sum_k |\beta_k| + \Sigma_{j\neq k} \left| \frac{\beta_{|j-k|}}{\sqrt{2}} \right| + \left| \frac{\beta_{j+k}}{\sqrt{2}} \right| \\
&\leq 1 + \left| \frac{\beta_{2j}}{\sqrt{2}} \right| + (|\beta_j| + \sqrt{2}) \sum_{k=1}^{\infty} |\beta_k| \\
&= O(1).
\end{aligned}$$

Hence, $\text{limsup}_{k\to\infty} \lambda_{\max}(\Sigma_k) \leq \|\Sigma_k\|_\infty = O(1)$ and the lemma holds. □

9.12 Proof of Theorem 6.3

The proof is similar to the proof of Theorem 4.4, so we provide a short outline. In particular, the effect of truncation can be shown to be negligible as in the proof of Theorem 4.4. We have $p - \widehat{p}_Z = p - \widehat{q} + \widehat{q} - \widehat{p}_Z = p - \widehat{q} + O_P(m/k)$ and the latter term is negligible for $k \geq n$. Now $p - \widehat{q} = p - \widehat{p} + \widehat{p} - \widehat{q}$. The term $p - \widehat{p}$ is the usual error term and contributes $O(n^{-2\gamma/(2\gamma+1)})$ to the risk. For the second term, $\int (\widehat{p} - \widehat{q})^2 = \sum_{j=1}^{m} \nu_j^2 = O_P(m/n) = O_P(n^{-2\gamma/(2\gamma+1)})$. □

References

AGGARWAL, G., FEDER, T., KENTHAPADI, K., KHULLER, S., PANIGRAHY, R., THOMAS, D. and ZHU, A. (2006). Achieving anonymity via clustering. *Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 153–162.

BARAK, B., CHAUDHURI, K., DWORK, C., KALE, S., MCSHERRY, F. and TALWAR, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 273–282.

BENTKUS, V. (2003). On the dependence of the berryesseen bound on dimension. *Journal of Statistical Planning and Inference* **113** 385–402.

BLUM, A., DWORK, C., MCSHERRY, F. and NISSIM, K. (2005). Practical privacy: the SuLQ framework. *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 128–138.

BLUM, A., LIGETT, K. and ROTH, A. (2008). A Learning Theory Approach to Non-Interactive Database Privacy. *Proceedings of the 40th annual ACM symposium on Theory of computing* 609–618.

CSÖRGŐ, M. and RÉVÉSZ, P. (1975). A new method to prove strassen type laws of invariance principle. II. *Probability Theory and Related Fields* 261–269.

DINUR, I. and NISSIM, K. (2003). Revealing information while preserving privacy. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 202–210.

DUNCAN, G. and LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* 10–28.

DUNCAN, G. and LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* 207–217.

DUNCAN, G. and PEARSON, R. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science* **6** 219–232.

DWORK, C. (2006). Differential privacy. *33rd International Colloquium on Automata, Languages and Programming* 1–12.

DWORK, C. and LEI, J. (2009). Differential privacy and robust statistics. *Proceedings of the 41st ACM Symposium on Theory of Computing* 371–380.

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. *Proceedings of the 3rd Theory of Cryptography Conference* 265–284.

DWORK, C., MCSHERRY, F. and TALWAR, K. (2007). The price of privacy and the limits of LP decoding. *Proceedings of the 39th annual ACM symposium on Theory of computing* 85–94.

DWORK, C., NAOR, M., REINGOLD, O., ROTHBLUM, G. and VADHAN, S. (2009). On the complexity of differentially private data release. *Proceedings of the 41st ACM Symposium on Theory of Computing* 381–390.

DWORK, C. and NISSIM, K. (2004). Privacy-preserving datamining on vertically partitioned databases. *Proceedings of the 24th Annual International CryptologyConference –CRYPTO* 528–544.

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer-Verlag.

EVFIMIEVSKI, A., SRIKANT, R., AGRAWAL, R. and GEHRKE, J. (2004). Privacy preserving mining of association rules. *Information Systems* **29** 343 – 364.

FEIGENBAUM, J., ISHAI, Y., MALKIN, T., NISSIM, K., STRAUSS, M. J. and WRIGHT, R. N. (2006). Secure multiparty computation of approximations. *ACM Trans. Algorithms* **2** 435–472.

FELDMAN, D., FIAT, A., KAPLAN, H. and NISSIM, K. (2009). Private coresets. *Proceedings of the 41st ACM Symposium on Theory of Computing* 361–370.

FIENBERG, S. and MCINTYRE, J. (2004). Data Swapping: Variations on a Theme by Dalenius and Reiss. *Privacy in Statistical Databases* **3050** 14–29.

FIENBERG, S. E., KARR, A. F., NARDI, Y. and SLAVKOVIC, A. (2007). Secure logistic regression with distributed databases. *Bulletin of the ISI* .

FIENBERG, S. E., MAKOV, U. E. and STEELE, R. J. (1998). Disclosure limitation using perturbation and related methods for categorial data (with discussion). *Journal of Official Statistics* **14** 485–511.

GANTA, S., KASIVISWANATHAN, S. and SMITH, A. (2008). Composition attacks and auxiliary information in data privacy. *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 265–273CoRR abs/0803.0032: (2008).

GHOSH, A., ROUGHGARDEN, T. and SUNDARARAJAN, M. (2009). Universally utility-maximizing privacy mechanisms. *Proceedings of the 41st ACM Symposium on Theory of Computing* 351–360.

HALL, P. and MURISON, R. D. (1993). Correcting the negativity of high-order kernel density estimators. *Journal of Multivariate Analysis* **47** 103–122.

KASIVISWANATHAN, S., LEE, H., NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2008). What Can We Learn Privately? *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science* 531–540.

KIM, J. J. and WINKLER, W. E. (2003). Multiplicative noise for masking continuous data. Tech. rep., Statistical Research Division, US Bureau of the Census, Washington D.C.

LI, N., LI, T. and VENKATASUBRAMANIAN, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *Proceedings of the 23rd International Conference on Data Engineering* 106–115.

LI, W. and SHAO, Q.-M. (2001). Gaussian processes: Inequalities, small ball probabilities and applications. In *STOCHASTIC PROCESSES: THEORY AND METHODS. Handbook of Statistics* (C. Rao and D. Shanbhag, eds.), vol. 19. Elsevier, 533–598.

MACHANAVAJJHALA, A., GEHRKE, J., KIFER, D. and VENKITASUBRAMANIAM, M. (2006). ℓ-diversity: Privacy beyond kappa-anonymity. *Proceedings of the 22nd International Conference on Data Engineering* 24.

MACHANAVAJJHALA, A., KIFER, D., ABOWD, J., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets Practice on the Map. *Proceedings of the 24th International Conference on Data Engineering* 277–286.

MCSHERRY, F. and TALWAR, K. (2007). Mechanism Design via Differential Privacy. *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science* 94–103.

NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2007). Smooth sensitivity and sampling in private data analysis. *Proceedings of the 39th annual ACM annual ACM symposium on Theory of computing* 75–84.

PINKAS, B. (2002). Cryptographic techniques for privacy-preserving data mining. *ACM SIGKDD Explorations Newsletter* **4**.

RASTOGI, V., HAY, M., MIKLAU, G. and SUCIU, D. (2009). Relationship privacy: Output perturbation for queries with joins. *Proceedings of the Twenty-Eigth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009* 107–116.

REITER, J. (2005). Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association* **100** 1103 – 1113.

ROHDE, A. and DUEMBGEN, L. (2008). Confidence sets for the optimal approximating model - bridging a gap between adaptive point estimation and confidence regions. *arXiv:0802.3276v2 [math.ST]* .

SANIL, A. P., KARR, A., LIN, X. and REITER, J. P. (2004). Privacy preserving regression modelling via distributed computation. *Proceedings of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 677–682.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

SMITH, A. (2008). Efficient, differentially private point estimators. ArXiv:0809.4794v1.

SWEENEY, L. (2002). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10** 557–579.

TING, D., FIENBERG, S. E. and TROTTINI, M. (2008). Random orthogonal matrix masking methodology for microdata release. *Int. J. of Information and Computer Security* **2** 86–105.

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