Robust Density Estimation under Besov IPM Losses

Ananya Uppal
Department of Mathematical Sciences
Carnegie Mellon University
auppal@andrew.cmu.edu

Shashank Singh
Machine Learning Department
Carnegie Mellon University
shashanksi@google.com

Barnabás Póczos
Machine Learning Department
Carnegie Mellon University
bapoczos@cs.cmu.edu

Abstract

We study minimax convergence rates of nonparametric density estimation under the Huber contamination model, in which a proportion of the data comes from an unknown outlier distribution. We provide the first results for this problem under a large family of losses, called Besov integral probability metrics (IPMs), that include the $\mathcal{L}^p$, Wasserstein, Kolmogorov-Smirnov, Cramer-von Mises, and other commonly used metrics. Under a range of smoothness assumptions on the population and outlier distributions, we show that a re-scaled thresholding wavelet estimator converges at minimax optimal rates under a wide variety of losses and also exhibits optimal dependence on the contamination proportion. We also provide a purely data-dependent extension of the estimator that adapts to both an unknown contamination proportion and the unknown smoothness of the true density. Finally, based on connections recently shown between density estimation under IPM losses and generative adversarial networks (GANs), we show that certain GAN architectures are robustly minimax optimal.

1 Introduction

In many settings, observed data contains not only samples from the distribution of interest, but also a small proportion of outlier samples. Because these outliers can exhibit arbitrary, unpredictable behavior, they can be difficult to detect or to explicitly account for. This has inspired a large body of work on *robust statistics*, which seeks statistical methods for which the error introduced by a small proportion of arbitrary outlier samples can be controlled.

The majority of work in robust statistics has focused on providing guarantees under the Huber ϵ-contamination model [Huber, 1965]. Under this model, data is assumed to be observed from a mixture distribution $(1-\epsilon)P + \epsilon G$, where P is an unknown population distribution of interest, G is an unknown outlier distribution, and $\epsilon \in [0, 1)$ is the "contamination proportion" of outlier samples. Equivalently, this models the misspecified case in which data are drawn from a small perturbation by $\epsilon(G - P)$ of the target distribution P of interest. The goal is then to develop methods whose performance degrades as little as possible when ϵ is non-negligible.

The present paper studies nonparametric density estimation under this model. Specifically, given independent and identically distributed samples from the mixture $(1-\epsilon)P + \epsilon G$, we characterize minimax optimal convergence rates for estimating P. Prior work on this problem has assumed P has a Hölder continuous density p and has provided minimax rates under total variation loss [Chen et al., 2018] or for estimating $p(x)$ at a point x [Liu and Gao, 2017]. In the present paper, in addition

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

to considering a much wider range of smoothness conditions (characterized by p lying in a Besov space), we provide results under a large family of losses called integral probability metrics (IPMs);

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim P} f(X) - \mathop{\mathbb{E}}_{X \sim Q} f(X) \right|, \tag{1}$$

where P and Q are probability distributions and $\mathcal{F}$ is a "discriminator class" of bounded Borel functions. As shown in several recent papers [Liu et al., 2017, Liang, 2018, Singh et al., 2018, Uppal et al., 2019], IPMs play a central role not only in nonparametric statistical theory and empirical process theory, but also in the theory of generative adversarial networks (GANs). Hence, this work advances not only basic statistical theory but also our understanding of the robustness properties of GANs.

In this paper, we specifically discuss the case of Besov IPMs, in which $\mathcal{F}$ is a Besov space (see Section 2.1). In classical statistical problems, Besov IPMs provide a unified formulation of a wide variety of distances, including $\mathcal{L}^p$ [Wasserman, 2006, Tsybakov, 2009], Sobolev [Mroueh et al., 2017, Leoni, 2017], maximum mean discrepancy (MMD; [Tolstikhin et al., 2017])/energy [Székely et al., 2007, Ramdas et al., 2017], Wasserstein/Kantorovich-Rubinstein [Kantorovich and Rubinstein, 1958, Villani, 2008], Kolmogorov-Smirnov [Kolmogorov, 1933, Smirnov, 1948], and Dudley metrics [Dudley, 1972, Abbasnejad et al., 2018]. Hence, as we detail in Section 4.3, our bounds for robust nonparametric density estimation apply under many of these losses. More recently, it has been shown that generative adversarial networks (GANs) can be cast in terms of IPMs, such that convergence rates for density estimation under IPM losses imply convergence rates for certain GAN architectures [Liang, 2018, Singh et al., 2018, Uppal et al., 2019]. Thus, as we show in Section 5, our results imply the first robustness results for GANs in the Huber model.

In addition to showing rates in the classical Huber model, which avoids assumptions on the outlier distribution G, we consider how rates change under additional assumptions on G. Specifically, we show faster convergence rates are possible under the assumption that G has a bounded density g, but that these rates are not further improved by additional smoothness assumptions on g.

Finally, we overcome a technical limitation of recent work studying density estimation under Besov IPMs losses. Namely, the estimators used in past work rely on the unrealistic assumption that the practitioner knows the Besov space in which the true density lies. This paper provides the first convergence rates for a purely data-dependent density estimator under Besov IPMs, as well as the first nonparametric convergence guarantees for a fully data-dependent GAN architecture.

1.1 Paper Organization

The rest of this paper is organized as follows. Section 2 formally states the problem we study and defines essential notation. Section 3 discusses related work in nonparametric density estimation. Section 4.1 contains minimax rates under the classical "unstructured" Huber contamination model, while Section 4.2 studies how these rates change when additional assumptions are made on the contamination distribution. Section 4.3 develops our general results from Sections 4.1 and 4.2 into concrete minimax convergence rates for important special cases. Finally, Section 5 applies our theoretical results to bound the error of perfectly optimized GANs in the presence of contaminated data. All theoretical results are proven in the Appendix.

2 Formal Problem Statement

We now formally state the problems studied in this paper. Let p be a density of interest and g be the contamination density such that $X_1, \ldots, X_n \sim (1-\epsilon)p + \epsilon g$ are n IID samples. We wish to use these samples to estimate p. We consider two qualitatively different types of contamination, as follows.

In the "unstructured" or Huber contamination setting, we assume that p lies in some regularity class $\mathcal{F}_g$, but g may be any compactly supported density. In particular, we assume that the data is generated from a density living in the set $\mathcal{M}(\epsilon, \mathcal{F}_g) = \{(1-\epsilon)p + \epsilon g : p \in \mathcal{F}_g, g \text{ has compact support}\}$. We then wish to bound the minimax risk of estimating p under an IPM loss $d_{\mathcal{F}_d}$; i.e., the quantity

$$\mathcal{R}(n, \epsilon, \mathcal{F}_g, \mathcal{F}_d) = \inf_{\widehat{p}_n} \sup_{f \in \mathcal{M}(\epsilon, \mathcal{F}_g)} \mathop{\mathbb{E}}_{f} \left[d_{\mathcal{F}_d}(p, \widehat{p}_n) \right] \tag{2}$$

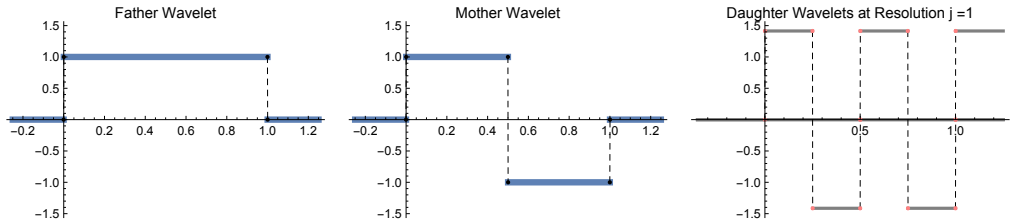

Figure 1: Father, Mother, and first few Daughter elements of the Haar Wavelet Basis.

where the infimum is taken over all estimators $\widehat{p}_n$.

In the "structured" contamination setting, we additionally assume that the contamination density g lives in a smoothness class $\mathcal{F}_c$. The data is generated by a density in $\mathcal{M}(\epsilon, \mathcal{F}_g, \mathcal{F}_c) = \{(1-\epsilon)p + \epsilon g : p \in \mathcal{F}_g, g \in \mathcal{F}_c\}$ and we seek to bound the minimax risk

$$\mathcal{R}(n, \epsilon, \mathcal{F}_g, \mathcal{F}_c, \mathcal{F}_d) = \inf_{\widehat{p}_n} \sup_{f \in \mathcal{M}(\epsilon, \mathcal{F}_g, \mathcal{F}_c)} \mathop{\mathbb{E}}_{f} \left[d_{\mathcal{F}_d}(\widehat{p}_n, p) \right]. \tag{3}$$

In the following section, we provide notation to formalize the spaces $\mathcal{F}_g, \mathcal{F}_c$ and $\mathcal{F}_d$ that we consider.

2.1 Set up and Notation

For non-negative real sequences $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$, $a_n \lesssim b_n$ indicates $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$, and $a_n \asymp b_n$ indicates $a_n \lesssim b_n \lesssim a_n$. For $q \in [1, \infty]$, $q' := \frac{q}{q-1}$ denotes the Hölder conjugate of q (with $1' = \infty, \infty' = 1$). $\mathcal{L}^q(\mathbb{R}^D)$ (resp. l^q) denotes the set of functions f (resp. sequences a) with $\|f\|_q := \left(\int |f(x)|^q \, dx\right)^{1/q} < \infty$ (resp. $\|a\|_{l^q} := \left(\sum_{n\in\mathbb{N}} |a_n|^q\right)^{1/q} < \infty$).

We now define the family of Besov spaces studied in this paper. Besov spaces generalize Hölder and Sobolev spaces and are defined using wavelet bases. As opposed to the Fourier basis, wavelet bases provide a localization in space as well as frequency which helps express spatially inhomogeneous smoothness.

A wavelet basis is formally defined by the mother ($\psi(x)$) and father($\phi(x)$) wavelets. The basis consists of two parts; first, the set of translations of the father and mother wavelets i.e.

$$\Phi = \{\phi(x-k) : k \in \mathbb{Z}^d\} \tag{4}$$

$$\Psi = \{\psi_\epsilon(x-k) : k \in \mathbb{Z}^d, \epsilon \in \{0,1\}^D\}, \tag{5}$$

second, the set of daughter wavelets, i.e.,

$$\Psi_j = \{2^{Dj/2}\psi_\epsilon(2^{Dj}x - k) : k \in \mathbb{Z}^d, \epsilon \in \{0,1\}^D\}. \tag{6}$$

Then the union $\Phi \cup \Psi \cup (\bigcup_{j\geq 0})\Psi_j$ is an orthonormal basis for $\mathcal{L}^2(\mathbb{R}^D)$.

We defer the technical assumptions on the mother and father wavelet to the appendix. Instead for intuition, we illustrate in Figure 1 the few terms of the best-known wavelet basis of $\mathcal{L}^2(\mathbb{R})$, the Haar wavelet basis.

In higher dimensions, the wavelet basis is defined using the tensor product of wavelets in dimension 1. For details, see, Härdle et al. [2012] and Meyer [1992].

To effectively express smooth functions we will require r-regular (r-regularity is precisely defined in the appendix) wavelets. We assume throughout our work that ϕ and ψ are compactly supported r-regular wavelets. We now formally define a Besov space.

Definition 1 (Besov Space)**.** *Given an r-regular wavelet basis of $\mathcal{L}^2(\mathbb{R}^D)$, let $0 \leq \sigma < r$, and $p, q \in [1, \infty]$. Then the* Besov space *$B^\sigma_{p,q}(\mathbb{R}^D)$ is defined as the set of functions $f : \mathbb{R}^D \to \mathbb{R}$ that satisfy*

$$\|f\|_{B^\sigma_{p,q}} := \|\alpha\|_{l^p} + \left\| \left\{ 2^{j(\sigma + D(1/2 - 1/p))} \|\beta_j\|_{l^p} \right\}_{j\in\mathbb{N}} \right\|_{l^q} < \infty \tag{7}$$

where α is the set of vectors $\{\alpha_\phi\}_{\phi\in\Phi}$ where $\alpha_\phi := \int_{\mathbb{R}^D} f(x)\phi(x)dx$ and β_j is the set of vectors $\{\beta_\psi\}_{\psi\in\Psi_j}$, where $\beta_\psi := \int_{\mathbb{R}^D} f(x)\psi(x)dx$.

The quantity $\|f\|_{B^{\sigma}_{p,q}}$ is called the *Besov norm of* f. For any $L > 0$, we write $B^{\sigma}_{p,q}(L)$ to denote the closed Besov ball $B^{\sigma}_{p,q}(L) = \{f \in B^{\sigma}_{p,q} : \|f\|_{B^{\sigma}_{p,q}} \leq L\}$. When the constant L is unimportant (e.g., for *rates* of convergence), $B^{\sigma}_{p,q}$ denotes a ball $B^{\sigma}_{p,q}(L)$ of finite but arbitrary radius L. We provide well-known examples from the rich class of resulting spaces in Section 4.3.

We now define "linear (distribution) estimators", a commonly used sub-class of distribution estimators:

Definition 2 (Linear Estimator)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. An estimate $\widehat{P}$ of P is said to be* linear *if there exist functions $T_i(X_i, \cdot) : \mathcal{F} \to \mathbb{R}$ such that for all measurable $A \in \mathcal{F}$, $\widehat{P}(A) = \sum_{i=1}^{n} T_i(X_i, A)$.*

Common examples of linear estimators are the empirical distribution, the kernel density estimator and the linear wavelet series estimator considered in this paper.

3 Related Work

This paper extends recent results in both non-parametric density estimation and robust estimation. We now summarize the results of the most relevant papers, namely those of Uppal et al. [2019], Chen et al. [2016], and Liu and Gao [2017].

3.1 Nonparametric Density Estimation under Besov IPM Losses

Uppal et al. [2019] studied the estimation of a density lying in a Besov space $B^{\sigma_g}_{p_g,q_g}$ under Besov IPM loss $d_{B^{\sigma_d}_{p_d,q_d}}$ with uncontaminated data. As shorthand, we will write $\mathcal{R}(n, \mathcal{F}_g, \mathcal{F}_d) = \mathcal{R}(n, 0, \mathcal{F}_g, \mathcal{F}_d)$ and $\mathcal{R}(n, \mathcal{F}_g, \mathcal{F}_c, \mathcal{F}_d) = \mathcal{R}(n, 0, \mathcal{F}_g, \mathcal{F}_c, \mathcal{F}_d)$ to denote the corresponding uncontaminated rates derived by Uppal et al. [2019]. They used the wavelet thresholding estimator, proposed in Donoho et al. [1996], to derive a minimax convergence rate of the form

$$\mathcal{R}\left(n, \mathcal{F}_g, \mathcal{F}_d\right) = n^{-1/2} + n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + n^{-\frac{\sigma_g+\sigma_d-D/p_g+D/p'_d}{2\sigma_g+D\left(1-2/p_g\right)}}, \tag{8}$$

(omitting polylog factors in n). Extending a classical result of Donoho et al. [1996], they also showed that, if the estimator is restricted to be linear (in the sense of Def. 2), then the minimax rate slows to

$$\mathcal{R}_L\left(n, \mathcal{F}_g, \mathcal{F}_d\right) = n^{-1/2} + n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + n^{-\frac{\sigma_g+\sigma_d-D/p_g+D/p'_d}{2\sigma_g+D(1-2/p_g+2/p'_d)}}. \tag{9}$$

The first two terms in (8) & (9) are identical and the third term in (9) is slower. In particular, when $p'_d > p_g$ and $\sigma_d < D/2$, linear estimators are strictly sub-optimal, while the wavelet thresholding estimator converges at the optimal rate. The present paper extends this work in two directions.

First, we study how the minimax risk of estimating the data density p changes when the observed data are contaminated by a proportion ϵ of outliers from a (potentially adversarially chosen) contamination distribution g. We show that, in most cases, wavelet thresholding estimators remain minimax optimal under both structured and unstructured contamination settings. Moreover, for $p'_d \leq p_g$ linear wavelet estimators are minimax optimal under the structured contamination setting and the unstructured contamination setting if the IPM is generated by a smooth enough class of functions ($\sigma_d \geq D/p_d$).

Second, noting that the estimators of Uppal et al. [2019] rely on knowledge of the smoothness parameter σ_g of the true density, we consider the more realistic case where σ_g is unknown. We develop a fully data-dependent variant of the wavelet thresholding estimator from Uppal et al. [2019] that is minimax optimal for all σ_g under structured contamination.

Finally, Uppal et al. [2019] also applied their results to bound the risk of a particular generative adversarial network (GAN) architecture. They show that the GAN is able to learn Besov densities at the minimax optimal rate. In this paper, we show that the same GAN architecture continues to be minimax optimal in the presence of outliers, and that, with minor modifications, it can do so without knowledge of the smoothness σ_g of the true density.

3.2 Nonparametric Density Estimation with Huber Contamination

Chen et al. [2016] give a unified study of a large class of robust nonparametric estimation problems under the total variation loss. In the particular case of estimating a σ_g-Hölder continuous density,

their results imply a minimax convergence rate of $n^{-\frac{\sigma_g}{2\sigma_g+1}} + \epsilon$, matching our results (theorem 3) for total variation loss. The results of Chen et al. [2016] are quite specific to total variation loss, whereas, we provide results for a range of loss functions as well as densities of varying smoothness. Moreover, the estimator studied by Chen et al. [2016] is not computable in practice. It involves solving a testing problem between all pairs of points in a total variation cover of the hypothesis class in which the true density is assumed to lie. In contrast, our upper bounds rely on a simple thresholded wavelet series estimator, which can be computed in linear time (in the sample size n) with a fast wavelet transform.

Liu and Gao [2017] studied 1-dimensional density estimation at a point x (i.e., estimating $p(x)$ instead of the entire density p) for Hölder smoothness densities under the Huber ϵ-contamination model. In the case of unstructured contamination (arbitrary G), Liu and Gao [2017] derived a minimax rate of

$$n^{-\frac{\sigma_0}{2\sigma_0+1}} + \epsilon^{\frac{\sigma_0}{\sigma_0+1}} \tag{10}$$

in root-mean-squared error. With the caveats that we study estimation of the entire density p rather than a single point $p(x)$ and assume that G has a density g, this corresponds to our setting when $p_g = q_g = \infty$, and $D = 1$. Our results (equation 18) imply an upper bound on the rate of

$$n^{-\frac{\sigma_0}{2\sigma_0+1}} + \epsilon^{\frac{\sigma_0}{\sigma_0+(1-1/p)}} \tag{11}$$

under the $\mathcal{L}^p$ loss. Interestingly, this suggests that estimating a density at a point under RMSE is harder than estimating an entire density under $\mathcal{L}^2$ loss, and is, in fact, as hard as estimation under $\mathcal{L}^\infty$ (sup-norm) loss. While initially perhaps surprising, this makes sense if one thinks of rates under $\mathcal{L}^\infty$ loss as being the rate of estimating the density at the worst-case point over the sample space, which may be the point x at which Liu and Gao [2017] estimate $p(x)$; under minimax analysis, these become similar.

We generalize these rates to (a) dimension $D > 1$, (b) densities p lying in Besov spaces $B^{\sigma_g}_{p_g,q_g}$, and (c) a wide variety of losses parametrized by Besov IPMs ($B^{\sigma_d}_{p_d,q_d}$).

Liu and Gao [2017] also study the case of structured contamination, in which g is assumed to be σ_c-Hölder continuous. Because they study estimation at a point, their results depend on an additional parameter, denoted m, which bounds the value of the contamination density g at the target point (i.e., $g(x) \leq m$). They derive a minimax rate of

$$n^{-\frac{\sigma_g}{2\sigma_g+1}} + \epsilon \min\{1, m\} + n^{-\frac{\sigma_c}{2\sigma_c+1}} \epsilon^{-\frac{\sigma_c}{2\sigma_c+1}}. \tag{12}$$

This rate contains a term depending only on n that is identical to the minimax rate in the uncontaminated case, a term depending only on ϵ, and a third "mixed" term. Notably, one can show that this mixed term $n^{-\frac{\sigma_c}{2\sigma_c+1}} \epsilon^{-\frac{\sigma_c}{2\sigma_c+1}}$ is always dominated by $n^{-\frac{\sigma_g}{2\sigma_g+D}} + \epsilon$, and so, unless $m \to 0$ as $n \to \infty$, the mixed term is negligible. In this paper, because we study estimation of the entire density p, the role of the parameter m is played by $M := \|g\|_\infty$. Since g is assumed to be a density with bounded support, we cannot have $M \to 0$; thus, in our results, the mixed term does not appear. Aside from this distinction, our results (Theorem 5) again generalize the results of Liu and Gao [2017] to higher dimensions, other Besov classes of densities, and new IPM losses.

Finally, we mention two early papers on robust nonparametric density estimation by Kim and Scott [2012] and Vandermeulen and Scott [2013]. These papers introduced variants of kernel density estimation based on M-estimation, for which they demonstrated robustness to arbitrary contamination using influence functions. These estimators are more complex than the scaled series estimates we consider, in that they non-uniformly weight the kernels centered at different sample points. While they also showed $\mathcal{L}^1$ consistency of these estimators, they did not provide rates of convergence, and so it is not clear when these estimators are minimax optimal.

4 Minimax Rates

Here we give our main minimax bounds. First, we state the estimators used for the upper bounds.

Estimators: To illustrate the upper bounds we consider two estimators that have been widely studied in the uncontaminated setting (see Donoho et al. [1996], Uppal et al. [2019]) namely the wavelet thresholding estimator and the linear wavelet estimator. All bounds provided here are tight up to polylog factors of n and $1/\epsilon$.

For any $j_1 \geq j_0 \geq 0$ the wavelet thresholding estimator is defined as

$$\widehat{p}_n = \sum_{\phi \in \Phi} \widehat{\alpha}_\phi \phi + \sum_{j=0}^{j_0} \sum_{\psi \in \Psi_j} \widehat{\beta}_\psi \psi + \sum_{j=j_0}^{j_1} \sum_{\psi \in \Psi_j} \widetilde{\beta}_\psi \psi \tag{13}$$

where $\widehat{\alpha}_\phi = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$ and $\widehat{\beta}_\psi = \frac{1}{n}\sum_{i=1}^n \psi(X_i)$ and coefficients of some of the wavelets with higher resolution (i.e., $j \in [j_0, j_1]$) are hard-thresholded: $\widetilde{\beta}_\psi = \widehat{\beta}_\psi 1_{\widehat{\beta}_\psi \geq t}$ for threshold $t = c\sqrt{j/n}$, where c is a constant.

The linear wavelet estimator is simply $\widehat{p}_n$ with only linear terms (i.e., $j_0 = j_1$). Here j_0, j_1 correspond to smoothing parameters which we carefully choose to provide upper bounds on the risk. In the sequel, let $\mathcal{F}_g = B^{\sigma_g}_{p_g,q_g}(L_g)$ and $\mathcal{F}_d = B^{\sigma_d}_{p_d,q_d}(L_d)$ be Besov spaces.

4.1 Unstructured Contamination

In this section we consider the density estimation problem under Huber's ϵ contamination model; i.e. we have no structural assumptions on the contamination. Let $X_1, \ldots, X_n \overset{IID}{\sim} (1-\epsilon)p + \epsilon g$, where p is the true density and g is any compactly supported probability density. We provide bounds on the minimax risk of estimating the density p. We let

$$\mathcal{M}(\epsilon, \mathcal{F}_g) = \{(1-\epsilon)p + \epsilon g : p \in \mathcal{F}_g, g \text{ has compact support}\} \tag{14}$$

and bound the minimax risk

$$\mathcal{R}(n, \epsilon, \mathcal{F}_g, \mathcal{F}_d) = \inf_{\widehat{p}} \sup_{f \in \mathcal{M}(\epsilon, \mathcal{F}_g)} \mathop{\mathbb{E}}_{f} d_{\mathcal{F}_d}(\widehat{p}, p) \tag{15}$$

where the infimum is taken over all estimators $\widehat{p}_n$ constructed from the n IID samples.

We first present our results for what Uppal et al. [2019] called the "Sparse" regime $p'_d \geq p_g$, in which the worst-case error is caused by large "spikes" in small regions of the sample space. Within this "Sparse" regime, we are able to derive minimax convergence rates for all Besov spaces $\mathcal{F}_g = B^{\sigma_g}_{p_g,q_g}(L_g)$ and $\mathcal{F}_d = B^{\sigma_d}_{p_d,q_d}(L_d)$. Surprisingly, we find that linear and nonlinear estimators have identical, rate-optimal dependence on the contamination proportion ϵ in this setting. Consequently, if ϵ is sufficiently large, then the difference in asymptotic rate between linear and nonlinear estimators vanishes. We first show the minimax rate in this setting that is achieved by a scaled version wavelet thresholding estimator i.e. $\frac{1}{(1-\epsilon)}\widehat{p}_n$. The proof is provided in section B.1 of the appendix.

Theorem 3. ***(Minimax Rate, Sparse Case)*** *Let $r > \sigma_g > D/p_g$ and $p'_d \geq p_g$. Then,*

$$\mathcal{R}\left(n, \epsilon, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) \sim \mathcal{R}\left(n, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) + \epsilon + \epsilon^{\frac{\sigma_g + \sigma_d + D/p'_d - D/p_g}{\sigma_g - D/p_g + D}} \tag{16}$$

On the other hand linear estimators are only able to achieve the following asymptotic rate. The proof is provided in section B.2 of the appendix.

Theorem 4. ***(Linear Minimax Rate, Sparse Case)*** *Let $r > \sigma_g > D/p_g$ and $p'_d \geq p_g$. Then,*

$$\mathcal{R}_L\left(n, \epsilon, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) \sim \mathcal{R}_L\left(n, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) + \epsilon + \epsilon^{\frac{\sigma_g + \sigma_d + D/p'_d - D/p_g}{\sigma_g - D/p_g + D}} \tag{17}$$

As is expected, the sub-optimality of linear estimators referred to in section 3, extends to the contaminated setting when contamination ϵ is small. However, if the contamination ϵ is large the distinction between linear and non-linear estimators disappears. More specifically, if ϵ is large enough then both estimators converge at the same rate of $\epsilon + \epsilon^{\frac{\sigma_g + \sigma_d + D/p'_d - D/p_g}{\sigma_g - D/p_g + D}}$.

Bounds for the regime $p'_d \leq p_g$: We note that the lower bounds that constitute the minimax rates above hold for all values of $p_g, p'_d \geq 1$. Furthermore, the linear wavelet estimator implies an upper bound (shown in section B.3 of the appendix) on the risk in the dense regime. Together, this gives, for all $r > \sigma_g > D/p_g$ and $p'_d \leq p_g$,

$$\Delta(n) + \epsilon + \epsilon^{\frac{\sigma_g + \sigma_d + D/p'_d - D/p_g}{\sigma_g - D/p_g + D}} \leq \mathcal{R}\left(n, \epsilon, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) \leq \Delta(n) + \epsilon + \epsilon^{\frac{\sigma_g + \sigma_d}{\sigma_g + D/p_d}} \tag{18}$$

where $\Delta(n) = \mathcal{R}\left(n, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right)$.

One can check that, when the discriminator is sufficiently smooth (specifically, $\sigma_d \geq D/p_d$), the term $\epsilon^{\frac{\sigma_g+\sigma_d}{\sigma_g+D/p_d}}$ on the right-hand side of Eq. (18) is dominated by ϵ; hence, the lower and upper bounds in Eq. (18) match and the thresholding wavelet estimator is minimax rate-optimal. When $\sigma_d < D/p_d$, a gap remains between our lower and upper bounds, and we do not know whether the thresholding wavelet estimator is optimal. The sample mean is generally well-known to be sensitive to outliers in the data, and a large amount of recent work [Lugosi and Mendelson, 2016, Lerasle et al., 2018, Minsker et al., 2019, Diakonikolas et al., 2019] has proposed estimators that might be better predictors of the mean in the case of contamination by outliers. Since the linear and thresholding wavelet estimators are both functions of the empirical means $\widehat{\beta}_\psi$ of the wavelet basis functions, we conjecture that a density estimator based on a better estimate of the wavelet mean β^p_ψ might be able to converge at a faster rate as $\epsilon \to 0$. We leave this investigation for future work.

4.2 Structured Contamination

In the previous section, we analyzed minimax rates without any assumptions on the outlier distribution. In certain settings, this may be an overly pessimistic contamination model, and the outlier distribution may in fact be somewhat well-behaved. In this section, we study the effects of assuming the contamination distribution G has a density g that is either bounded or smooth. Our results show that assuming boundedness of g improves the dependence of the minimax rate on ϵ to order $\asymp \epsilon$, but assuming additional smoothness of g does not further improve rates.

As described in Section 2, in this setting we consider a more general form of the minimax risk:

$$\mathcal{R}(n, \epsilon, \mathcal{F}_g, \mathcal{F}_c, \mathcal{F}_d) = \inf_{\widehat{p}} \sup_{f \in \mathcal{M}(\epsilon, \mathcal{F}_g, \mathcal{F}_c)} \mathbb{E}_f[d_{\mathcal{F}_d}(\widehat{p}, p)] \tag{19}$$

The additional parameter $\mathcal{F}_c$ denotes the class of allowed contamination distributions.

We provide the following asymptotic rate for the above minimax risk that is achieved by an adaptive wavelet thresholding estimator with $2^{j_0} = n^{\frac{1}{2r+D}}$ and $2^{j_1} = (n/\log n)^{1/D}$. Recall here that r is the regularity of the wavelets. Thus, for any $\sigma_g < r$, this estimator does not require the knowledge of σ_g.

Theorem 5 (Minimax Rate under Structured Contamination)**.** *Let $\sigma_g \geq D/p_g$, $\sigma_c > D/p_c$ and $\epsilon \leq 1/2$. Then, up to poly logarithmic factors of n,*

$$\mathcal{R}\left(n, \epsilon, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_c}_{p_c,q_c}, B^{\sigma_d}_{p_d,q_d}\right) \asymp \mathcal{R}\left(n, \epsilon, B^{\sigma_g}_{p_g,q_g}, \mathcal{L}^\infty, B^{\sigma_d}_{p_d,q_d}\right) \asymp \mathcal{R}\left(n, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) + \epsilon \tag{20}$$

The right-most term is simply ϵ plus the rate in the absence of contamination. The left two terms are the rates when the contamination density lies, respectively, in the Besov space $B^{\sigma_c}_{p_c,q_c}$ and the space $\mathcal{L}^\infty$ of essentially bounded densities. In particular, these rates are identical when $\sigma_c > D/p_c$. One can check (see Lemma 10 in the Appendix) that, if $\sigma_c > D/p_c$, then $B^{\sigma_c}_{p_c,q_c} \subseteq \mathcal{L}^\infty$. Hence, Theorem 5 shows that assuming boundedness of the contamination density improves the dependence on ϵ (compared to unstructured rates from the previous section), but that additional smoothness assumptions do not help.

In section B.1 of the appendix we first provided a proof of the upper bound using the classical wavelet thresholding estimator and then show the optimality of the adaptive version in section B.4.

4.3 Examples

Here, we summarize the implications of our main results for robust density estimation in a few specific examples, allowing us to directly compare with previous results.

The case $p_d = q_d = \infty$, includes, as examples, the total variation loss $d_{B^0_{p_d,q_d}}$ ($\sigma_d = 0$, Rudin [2006]) and the Wasserstein (a.k.a., Kantorovich-Rubinstein or earthmover) loss $d_{B^1_{p_d,q_d}}$ ($\sigma_d = 1$ [Villani, 2008]). Under these losses, the wavelet thresholding estimator is robustly minimax optimal, in both the arbitrary and structured contamination settings (note that here $\sigma_d \geq D/p_d = 0$). In particular, in the case of unstructured contamination, this generalizes the results of Chen et al. [2016] for total variation loss to a range of other losses and smoothness assumptions on p.

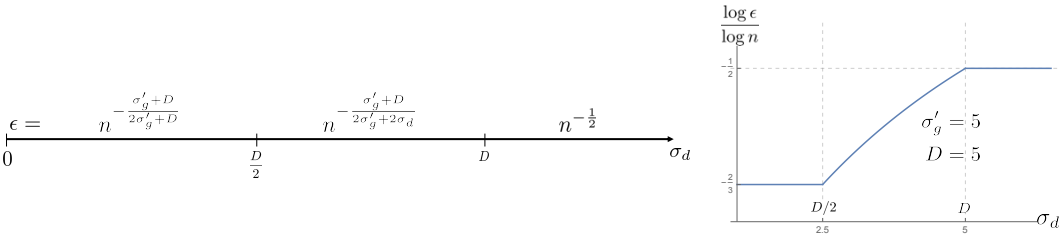

Figure 2: Asymptotic breakdown point as a function of σ_d, in the case $p_d = 1$; this includes as special cases the $\mathcal{L}^\infty$ and Kolmogorov-Smirnov losses.

Analogously, in the case $p_g = q_g = \infty$, the data distribution is itself σ_g-Hölder continuous, since the Besov space $B^{\sigma_g}_{p_g,q_g} = \mathcal{C}^{\sigma_g}$ is equivalent to the space of σ_g-Hölder continuous functions. In this setting, the linear wavelet estimator is robustly minimax optimal under any Besov IPM loss when contamination is structured, or under sufficiently smooth Besov IPM losses (with $\sigma_d \geq D/p_d$) when the contamination is unstructured.

One can also use our results to calculate the sensitivity of a given estimator to the proportion ϵ of outlier samples. In the terminology of robust statistics, this is quantified by the "asymptotic breakdown point" (i.e., the maximum proportion ϵ of outlier samples such that the estimator can still converge at the uncontaminated optimal rate). Figure 2 illustrates the asymptotic breakdown point, in the case $p_d = 1$, as a function of the discriminator smoothness σ_d. For sufficiently smooth losses (large σ_d, the estimator can tolerate a large number ($O(\sqrt{n})$) of arbitrary outliers before performance begins to degrade, whereas, for stronger losses (smaller σ_d), the estimator becomes more sensitive to outliers.

5 Robustness of Generative Adversarial Networks

Singh et al. [2018] showed (in their Theorem 9) that the problems of generating novel samples from a training density (also called "implicit generative modeling" [Mohamed and Lakshminarayanan, 2016]) and of estimating the training density are equivalent in terms of statistical minimax rates. Based on this result, and an oracle inequality of Liang [2018], several recent works [Liu et al., 2017, Liang, 2018, Singh et al., 2018, Uppal et al., 2019] have studied a statistical formulation of GANs as a distribution estimate based on empirical risk minimization (ERM) under an IPM loss. This formulation is as follows. Given a GAN with a discriminator neural network N_d encoding functions in $\mathcal{F}$ and a generator neural network N_g encoding distributions in $\mathcal{P}$, the GAN generator can be viewed as the distribution $\widehat{P}$ satisfying:

$$\widehat{P} = \inf_{P \in \mathcal{P}} d_{\mathcal{F}}(P, \widetilde{P}_n) \tag{21}$$

While $\widetilde{P}_n$ can be taken to be the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, these theoretical works have shown that convergence rates can be improved by applying regularization (e.g., in the form of smoothing the empirical distribution), consistent with the "instance noise trick" [Sønderby et al., 2016], a technique that is popular in practical GAN training and is mathematically equivalent to kernel smoothing. Here, we extend these results to the contamination setting and show that the wavelet thresholding estimator can be used to construct a GAN estimate that is robustly minimax optimal.

Suzuki [2018] (in section 3) showed that there is a fully connected ReLU network with depth at most logarithmic in $1/\delta$ and other size parameters at most polynomial in $1/\delta$ that can δ-approximate any sufficiently smooth Besov function class (e.g. $B^{\sigma_g}_{p_g,q_g}$ with $\sigma_g \geq D/p_g$). This was used in Uppal et al. [2019] to show that, for large enough network sizes, the perfectly optimized GAN estimate (of the form of Eq. (21)) converges at the same rate as the estimator $\widehat{p}$ used to generate it. So if we let the approximation error of the generator and discriminator network be at most the convergence rate (from Theorem 3 or Theorem 5) of the wavelet thresholding estimator then there is a GAN estimate $\widehat{P}$ that converges at the same rate and is therefore robustly minimax optimal. In particular, we have the following corollary:

Corollary 6. *Given a Besov density class $B^{\sigma_g}_{p_g,q_g}$ with $\sigma_g > D/p_g$ and discriminator class $B^{\sigma_d}_{p_d,q_d}$ with $\sigma_d > D/p_d$, there is a GAN estimate $\widehat{P}$ with discriminator and generator networks of depth at most logarithmic in n, or $1/\delta$ and other size parameters at most polynomial in n or $1/\delta$ such that*

$$\sup_{p \in \mathcal{M}\left(\epsilon, B^{\sigma_g}_{p_g,q_g}\right)} \mathbb{E}\left[d_{B^{\sigma_d}_{p_d,q_d}}(\widehat{p}, p)\right] \leq \delta + \mathcal{R}\left(n, B^{\sigma_g}_{p_g,q_g}, B^{\sigma_d}_{p_d,q_d}\right) + \epsilon \tag{22}$$

Since Besov spaces of compactly supported densities are nested ($B^{\sigma}_{p,q} \subseteq B^{\sigma'}_{p,q}$ for all $\sigma' \leq \sigma$), to approximate $B^{\sigma}_{p,q}$ for any $\sigma \geq r_0$ it is sufficient to approximate $B^{r_0}_{p,q}$. We can use this approximation network along with the adaptive wavelet thresholding estimator to construct a GAN estimate of the form of Eq. (21). Then under structured contamination this GAN estimate is minimax optimal for any density of smoothness $\sigma_g \in [r_0, r]$ and does not require explicit knowledge of σ_g. Thus, it is adaptive.

6 Conclusion

In this paper, we studied a variant of nonparametric density estimation in which a proportion of the data are contaminated by random outliers. For this problem, we provided bounds on the risks of both linear and nonlinear wavelet estimators, as well as general minimax rates. The main conclusions of our study are as follows:

1. The classical wavelet thresholding estimator originally proposed by Donoho et al. [1996], which is widely known to be optimal for uncontaminated nonparametric density estimation, continues to be, in many settings, minimax optimal in the presence of contamination.

2. Imposing a simple structural assumption, such as bounded contamination, can significantly alter how contamination affects estimation risk. At the same time, additional smoothness assumptions have no effect. This contrasts from the case of estimating a density at a point, as studied by Liu and Gao [2017] where the minimax rates get better with smoothness of the contamination density.

3. Linear estimators, exhibit optimal dependence on the contamination proportion, despite having sub-optimal risk with respect to the sample size. Hence, the difference between linear and nonlinear models diminishes in the presence of significant contamination.

4. For sufficiently smooth density and discriminator class, a fully-connected GAN architecture with ReLU activations can learn the distribution of the training data at the optimal rate, both (a) in the presence of contamination and (b) when the true smoothness of the density is not known.

Our results both extend recent results on nonparametric density estimation under IPM losses [Liang, 2018, Singh et al., 2018, Uppal et al., 2019] to the contaminated and adaptive settings and expand the study of nonparametric density estimation under contamination [Chen et al., 2016, Liu and Gao, 2017] to Besov densities and IPM losses.

Broader Impact

Since this work is of a theoretical nature, it is unlikely to disadvantage anyone or otherwise have significant negative consequences. One of the main contributions of this paper is to quantify the potential effects of misspecification biases on density estimation. Hence, the results in this paper may help researchers understand the potential effects of misspecification biases that can arise when invalid assumptions are made about the nature of the data generating process.

Acknowledgments and Disclosure of Funding

The authors thank anonymous reviewers for the feedback on improving this paper. The authors declare no competing interests. This work was supported by National Science Foundation award number DGE1745016, a grant from JPMorgan Chase Bank, and a grant from the Lockheed Martin Corporation.

References

Ehsan Abbasnejad, Javen Shi, and Anton van den Hengel. Deep Lipschitz networks and Dudley GANs, 2018. URL `https://openreview.net/forum?id=rkw-jlb0W`.

Mengjie Chen, Chao Gao, Zhao Ren, et al. A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

RM Dudley. Speeds of metric probability convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 22(4):323–332, 1972.

Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

Leonid Vasilevich Kantorovich and Gennady S Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

JooSeuk Kim and Clayton D Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13(Sep):2529–2565, 2012.

Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, and Guillaume Lecué. Monk–outlier-robust mean embedding estimation by median-of-means. *arXiv preprint arXiv:1802.04784*, 2018.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Haoyang Liu and Chao Gao. Density estimation with contaminated data: Minimax rates and theory of adaptation. *arXiv preprint arXiv:1712.07801*, 2017.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5551–5559, 2017.

Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.

Yves Meyer. *Wavelets and operators*, volume 1. Cambridge university press, 1992.

Stanislav Minsker et al. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. *arXiv preprint arXiv:1711.04894*, 2017.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Haskell P. Rosenthal. On the subspaces of $l^p(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill education, 2006.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabas Poczos. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems 31*, pages 10246–10257, 2018.

Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Alexandre B Tsybakov. *Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats*. Springer Series in Statistics. Springer, New York, 2009.

Ananya Uppal, Shashank Singh, and Barnaás Póczos. Nonparametric density estimation under Besov IPM losses. *arXiv preprint arXiv:1902.03511*, 2019.

Robert Vandermeulen and Clayton Scott. Consistency of robust kernel density estimators. In *Conference on Learning Theory*, pages 568–591, 2013.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

A Set up

Besov spaces rely on the notion of an r-regular multi-resolution approximation (MRA) of $\mathcal{L}_2(\mathbb{R}^D)$. In particular, the father wavelet of the wavelet basis used to define Besov spaces generates an MRA of $\mathcal{L}_2(\mathbb{R}^D)$.

The goal of a MRA is to efficiently approximate spatially varying smoothness. Härdle et al. [2012] explains it as a formulation that makes mathematically precise the intuitive idea of partitioning the domain and applying Fourier analysis to each piece.

Here we formally define an r-regular multi-resolution approximation.

Definition 7. *A* multiresolution approximation (MRA) *of $\mathcal{L}^2(\mathbb{R}^D)$ is a nested sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed linear subspaces of $L^2(\mathbb{R}^D)$ such that:*

1. $\bigcap_{j=-\infty}^{\infty} V_j = \{0\}$, and $\bigcup_{j=-\infty}^{\infty} V_j$ is dense in $\mathcal{L}^2(\mathbb{R}^D)$.
2. For every $f \in \mathcal{L}^2(\mathbb{R}^D)$ and $k \in \mathbb{Z}^D$, $f(x) \in V_0$ if and only if $f(x-k) \in V_0$.
3. For every $f \in \mathcal{L}^2(\mathbb{R}^D)$ and $j \in \mathbb{Z}$, $f(x) \in V_j$ if and only if $f(2x) \in V_{j+1}$
4. There is a "father wavelet" such that $\phi \in V_0$, $\{\phi(x-k) : k \in \mathbb{Z}^D\}$ is an orthonormal basis of $V_0 \subset \mathcal{L}^2(\mathbb{R}^D)$.

Given a father wavelet that generates a multi-resolution approximation, there exist "mother" wavelets with the following properties.

Lemma 8 ([Meyer, 1992], Section 3.9)**.** *Let $\{V_j\}_{j\in\mathbb{Z}}$ be an MRA of $\mathcal{L}^2(\mathbb{R}^D)$ with father wavelet ϕ. , and let W_j be the orthogonal complement of V_j in V_{j+1}. Then, for $E = \{0,1\}^D \setminus (0,\ldots,0)$, there exist "mother wavelets" $\{\psi_\epsilon\}_{\epsilon\in E}$ such that*

1. *ψ_ϵ is rapidly decreasing for every multi-index α with $|\alpha| \le r$ and every $\epsilon \in E$.*
2. *The set $\{\psi(x-k)\}_{\epsilon\in E, k\in\mathbb{Z}}$ is an orthonormal basis of W_j.*
3. *For all α with $|\alpha| \le r$ and $\epsilon \in E$, $\int x^\alpha \psi_\epsilon(x)dx = 0$.*

Moreover, $\{2^{Dj/2}\psi_\epsilon(2^j x - k) : \epsilon \in E, k \in \mathbb{Z}^D\} \cup \{2^{Dj/2}\phi(2^j x - k) : k \in \mathbb{Z}^D\}$ is an orthonormal basis of $V_j \subseteq \mathcal{L}^2(\mathbb{R}^D)$.

The r-regularity of the mother wavelet as described part 3 of the above lemma determines the r-regularity of the wavelet basis.

B Upper Bounds

B.1 Non-Linear Rate

In this section we provide proofs of the upper bounds stated above under both structured and unstructured contamination i.e. theorems 3 and 5. We will use a scaled version of the wavelet thresholding estimator to demonstrate these results. The proofs follow along the same lines as those of the uncontaminated version except the usual bias-variance trade-off now has an additional term; the misspecification error.

In particular, the bound on the bias remains unchanged. Moreover, we show that for resolutions small enough the variance can be bounded by the same term as before. This is straightforward for the variance of the linear terms but somewhat involved for that of the non-linear terms. So, we derive the bound for the non-linear terms at the very end.

There is a qualitative difference between the misspecification error under the structured and unstructured settings. When the contamination density is bounded, the misspecification error is simple bounded by the contamination proportion ϵ; in the unstructured setting, this error depends on the number of terms considered in the estimator.

We now provide the formal proof. Let

$$\mathcal{P} = \{p : p \ge 0, \|p\|_{\mathcal{L}^1} = 1, \text{supp}(p) \subseteq [-T, T]\}$$

denote the set of densities that are supported on the interval $[-T, T]$. We have assumed that our discriminator and generator classes are, respectively,

$$\mathcal{F}_d = \{f : \|f\|_{p_d,q_d}^{\sigma_d} \leq L_d\}$$
$$\text{and} \quad \mathcal{F}_g = \{p : \|p\|_{p_g,q_g}^{\sigma_g} \leq L_g\} \cap \mathcal{P}.$$

Let for any density function p

$$\alpha_\phi^p = \mathop{\mathbb{E}}_{X\sim p}[\phi(X)]$$
$$\text{and} \quad \beta_\psi^p = \mathop{\mathbb{E}}_{X\sim p}[\psi(X)].$$

Since $p \in \mathcal{F}_g$, we have that

$$p = \sum_{\phi\in\Phi} \alpha_\phi^p \phi + \sum_{j\geq 0}\sum_{\psi\in\Psi_j} \beta_\psi^p \psi,$$

where the convergence is in the L_p norm.

For the unstructured setting we merely assume that the contamination density is compactly supported on $[-T, T]$. Under the structured contamination setting, we additionally assume that the contamination density g is essentially bounded i.e. $\mathcal{F}_c = \mathcal{L}^\infty(L_c)$ (where L_c is a uniform bound on the $\mathcal{L}_\infty$ norm of any $g \in \mathcal{F}_c$).

We first show that it is enough to consider the "sparse" case (so called by Donoho et al. [1996]) characterized by $p_d' \geq p_g$ by the following lemma.

Lemma 9. *For $p_d' \leq p_g$ and compactly supported densities $p, q \in \mathcal{L}_{p_g} \subseteq \mathcal{L}_{p_d'}$ we have that,*

$$d_{\mathcal{B}_{p_d,q_d}^{\sigma_d}}(p, q) \leq d_{\mathcal{B}_{p_g',q_d}^{\sigma_d}}(p, q).$$

Proof. Suppose p, q are compactly supported on $[-T, T]$ then it is enough to show that

$$\mathcal{B}_{p_d,q_d}^{\sigma_d}(T) \subseteq \mathcal{B}_{p_g',q_d}^{\sigma_d}(T).$$

Using the fact that $p_d \geq p_g'$, this is clear since,

$$2^{j(\sigma_d + D/2 - D/p_g')} \|\beta_j\|_{p_g'} \leq 2^{j(\sigma_d + D/2 - D/p_d)} \|\beta_j\|_{p_d}$$

by using the simple fact that for a 2^{Dj}-dimensional vector x, $\|x\|_{p_g'} \leq 2^{Dj(1/p_g' - 1/p_d)} \|x\|_{p_d}$. □

Let $\widehat{p}_n$ be the wavelet thresholding estimator of p introduced by Donoho et al. [1996];

$$\widehat{p}_n = \sum_{\phi\in\Phi} \widehat{\alpha}_\phi \phi + \sum_{j=0}^{j_0}\sum_{\psi\in\Psi_j} \widehat{\beta}_\psi \psi + \sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j} \widetilde{\beta}_\psi \psi$$

where we threshold the higher resolution terms i.e.

$$\alpha_\phi^p = \mathop{\mathbb{E}}_{X\sim p}[\phi(X)] \qquad \widehat{\alpha}_\phi = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$$
$$\beta_\psi^p = \mathop{\mathbb{E}}_{X\sim p}[\psi(X)] \qquad \widehat{\beta}_\psi = \frac{1}{n}\sum_{i=1}^n \psi(X_i)$$
$$\widetilde{\beta}_\psi = \widehat{\beta}_\psi \mathbf{1}_{\{\widehat{\beta}_\psi > t\}}$$

with threshold $t = K\sqrt{j/n}$, where K is a constant to be specified later, and

$$2^{j_0} = \sqrt{n}^{\frac{1}{\sigma_g + D/2}}$$
$$2^{j_1} = \sqrt{n}^{\frac{1}{\sigma_g + D/2 - D/p_g}} \wedge \epsilon^{-\frac{1}{\sigma_g + D - D/p_g}}$$

We will use a scaled version of this estimator i.e. $\frac{1}{1-\epsilon}\widehat{p}_n$.

We decompose the risk of the above estimator as follows. At each resolution $\widehat{\alpha}_\phi$ or $\widehat{\beta}_\psi$ is an unbiased estimate of the co-efficient of the contaminated density $(1-\epsilon)p + \epsilon g$. So, by the triangle inequality, we can decompose the error as

$$
\begin{aligned}
&\mathbb{E}\, d_{\mathcal{F}}\left(\frac{\widehat{p}_n}{1-\epsilon}, p\right) \\
&\quad \leq \frac{1}{1-\epsilon}\,\mathbb{E}\, d_{\mathcal{F}}\left(\sum_{\phi\in\Phi}\widehat{\alpha}_\phi\phi, \sum_{\phi\in\Phi}(\alpha^p_\phi + \epsilon\alpha^g_\phi)\phi\right) && (23)\\
&\quad + \frac{1}{1-\epsilon} d_{\mathcal{F}}\left(\sum_{j=0}^{j_0}\sum_{\psi\in\Psi_j}\widehat{\beta}_\psi\psi, \sum_{j=0}^{j_0}\sum_{\psi\in\Psi_j}(\beta^p_\psi + \epsilon\beta^g_\psi)\psi\right) && (24)\\
&\quad + \frac{1}{1-\epsilon} d_{\mathcal{F}}\left(\sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j}\widetilde{\beta}_\psi\psi, \sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j}(\beta^p_\psi + \epsilon\beta^g_\psi)\psi\right) && (25)\\
&\quad + \frac{1}{1-\epsilon} d_{\mathcal{F}}\left(\sum_{\phi\in\Phi}\alpha^p_\phi\phi + \sum_{j=0}^{j_1}\sum_{\psi\in\Psi_j}\beta^p_\psi\psi, p\right) && (26)\\
&\quad + \epsilon d_{\mathcal{F}}\left(\sum_{\phi\in\Phi}\alpha^g_\phi\phi + \sum_{j=0}^{j_1}\sum_{\psi\in\Psi_j}\beta^g_\psi\psi, 0\right) && (27)
\end{aligned}
$$

where the first two terms constitute the error of the linear terms, the third term is the error of the non-linear terms, the fourth term is the bias and the last term is the misspecification error, respectively.

We will use the following upper bounds on the bias and variance of a linear wavelet estimator (when $j_0 = j_1$ above) from Appendix C of Uppal et al. [2019].

First we see that under Besov IPMs, if the moments of the wavelet co-efficients of the density don't grow too fast with the resolution then the variance of the linear wavelet estimator can be conveniently bounded.

Lemma 10. *(**Variance**) Let $X_1, \ldots, X_n \sim p$ where p is compactly supported and $\mathcal{F}_d = B^{\sigma_d}_{p_d,q_d}$. If $\mathbb{E}_p\,|\psi(X)|^{p'_d} \leq c_{p'_d} 2^{Dj(p'_d/2-1)}$ for all $\psi \in \Psi_j$, then the variance of a linear wavelet estimator $\widehat{p}_n$ with j_0 terms i.e.*

$$\widehat{p}_n = \sum_{\phi\in\Phi}\widehat{\alpha}_\phi\phi + \sum_{j=0}^{j_0}\sum_{\psi\in\Psi_j}\widehat{\beta}_\psi\psi$$

is bounded by

$$d_{\mathcal{F}_d}(\widehat{p}_n, \mathbb{E}[\widehat{p}_n]) \leq c\left(\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}}\right)$$

where $c = c_{p'_d}\left(\mathbb{E}_p\,|\psi(X)|^2\right)^{1/2}$ is a constant.

Note here that we do not need the density to lie in a Besov space but to simply have the given bound on the moments of its wavelet coefficients. However, for a bound on the bias provided below we need the full power of the Besov space.

Lemma 11. *(**Bias**) Let $X_1, \ldots, X_n \sim p$ where $p \in B^{\sigma_g}_{p_g,q_g}$ is compactly supported and $\sigma_g \geq D/p_g$, $\mathcal{F}_d = B^{\sigma_d}_{p_d,q_d}$. Then the bias of a linear wavelet estimator $\widehat{p}$ with j_0 terms is bounded by*

$$d_{\mathcal{F}_d}(p, \mathop{\mathbb{E}}_{p}[\widehat{p}_n]) \leq c 2^{-j_0(\sigma_d+\sigma_g-(D/p_g-D/p'_d)_+)}$$

where $c = L_d L_g$ is a constant.

We will also need the following bound on a density living in a Besov space.

Lemma 12. (Upper Bound on Smooth Besov Spaces) *Let $f \in B^{\sigma_g}_{p_g,q_g}$ where $\sigma_g > D/p_g$ then*

$$\|f\|_\infty \leq 4A\,\|\psi\|_\infty\, L_g(1 - 2^{(\sigma_g - D/p_g)q'_g})^{-1/q'_g}$$

This lemma implies that sufficiently smooth Besov spaces $B^{\sigma_g}_{p_g,q_q}$ are uniformly bounded.

We are now ready to provide upper bounds on the risk in both the structured and unstructured setting whenever $p'_d \geq p_g$.

Under both the structured and unstructured contamination setting we can immediately bound the bias term using lemma 11 (since this is just the bias of a linear wavelet estimator with j_1 terms) by

$$2^{-j_1(\sigma_g+\sigma_d+D/p'_d-D/p_g)}$$

Since, $\sigma_g > D/p_g$ we know that by lemma 12, $\|p\|_\infty < \infty$. Therefore, for any $\psi \in \Psi_j$,

$$\mathop{\mathbb{E}}_{(1-\epsilon)p+\epsilon g}\left[|\psi(X)|^{p'_d}\right] \leq (1-\epsilon)\|p\|_\infty 2^{Dj(p'_d/2-1)} + \epsilon \mathop{\mathbb{E}}_{g}\left[|\psi(X)|^{p'_d}\right]$$

When contamination is structured i.e. $\|g\|_\infty < \infty$ we have

$$\epsilon \mathop{\mathbb{E}}_{g}\left[|\psi(X)|^{p'_d}\right] \leq \epsilon \|g\|_\infty 2^{Dj(p'_d/2-1)}$$

and when the contamination density is not bounded above we have,

$$\epsilon \mathop{\mathbb{E}}_{g}|\psi(X)|^{p'_d} \leq \epsilon 2^{Djp'_d/2}$$

Since in this case, $2^{Dj_1} \leq \epsilon^{-\frac{D}{\sigma_g+D-D/p_g}} \leq 1/\epsilon$, $\epsilon \leq 2^{-Dj}$ for all $j \leq j_1$ the term above is always smaller than $2^{Dj(p'_d/2-1)}$.

So, under both cases, we have,

$$\mathop{\mathbb{E}}_{(1-\epsilon)p+\epsilon g}|\psi(X)|^{p'_d} \leq c2^{Dj(p'_d/2-1)} \tag{28}$$

We can now use lemma 10 to bound the variance for both cases as

$$\mathbb{E}\, d_{\mathcal{F}}(\widehat{p}_n, \mathop{\mathbb{E}}_{(1-\epsilon)p+\epsilon g}[\widehat{p}_n]) \leq c\left(\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}}\right)$$

We can also similarly bound the mis-specification error as this is simply the misspecification error of a linear wavelet estimator with j_1 terms. We then have an upper bound of

$$\epsilon 2^{j_1(D/p_d-\sigma_d)}.$$

We now bound the misspecification error i.e.

$$\frac{\epsilon}{1-\epsilon} d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{g}[\widehat{p}_n], 0\right)$$

We will use the following lemmas proven by Uppal et al. [2019] to first reduce the expression of the above distance to one in terms of wavelet coefficients (of g) only.

Lemma 13. *Let p, q be compactly supported probability densities and $\mathcal{F}_d = B^{\sigma_d}_{p_d,q_d}$, s.t. either $p, q \in L_{p'_d}$ or $\sigma_d > D/p_d$, then $d_{\mathcal{F}_d}(p,q) =$*

$$\sup_{f\in\mathcal{F}_d}\left|\sum_{\phi\in\Phi}\alpha^f_\phi\left(\alpha^p_\phi - \alpha^q_\phi\right) + \sum_{j\geq 0}\sum_{\psi\in\Psi_j}\beta^f_\psi\left(\beta^p_\psi - \beta^q_\psi\right)\right|$$

where for $f \in \mathcal{F}_d$

$$f = \sum_{\phi\in\Phi}\alpha^f_\phi\phi + \sum_{j\geq 0}\sum_{\psi\in\Psi_j}\beta^f_\psi\psi$$

Lemma 14. *Let $n_1, n_2 \in \mathbb{N}\cup\{\infty\}$ and η be any sequence of numbers. Then*

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\sup_{f\in\mathcal{F}_d}\sum_{j=n_1}^{n_2}\sum_{\psi\in\Psi_j}\gamma^f_\psi\eta_\psi \leq L_d\sum_{j=n_1}^{n_2}2^{-j\sigma'_d}\left(\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\sum_{\psi\in\Psi_j}|\eta_\psi|^{p'_d}\right)^{1/p'_d}$$

where $\sigma'_d = \sigma_d + D/2 - D/p_d$. Note that the above is true also if $\gamma = \alpha^f$ and $n_1 = n_2 = 0$.

Applying the lemmas above we have for any contamination density g,

$$\epsilon d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{g}[\widehat{p}_n], 0\right) \leq c\epsilon\left(\left\|\alpha^g\right\|_{p_d'} + \sum_{j=0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)}\left\|\beta_j^g\right\|_{p_d'}\right) \tag{29}$$

where for all $\phi \in \Phi$ and $\psi \in \Psi_j$, $\alpha_\phi^g = \int \phi(x)g(x)$ and $\beta_\psi^g = \int \psi(x)g(x)$.

When the contamination is structured we have the following upper bound on the wavelet coefficients

$$|\beta_\psi^g| = \left|\int \psi(x)g(x)d(x)\right| \leq \|\psi_\epsilon\|_\infty \|g\|_\infty 2^{-Dj/2} \implies \left\|\beta_\psi^g\right\|_{p_d'} \leq c2^{Dj(1/p_d' - 1/2)} \tag{30}$$

where $\psi \in \Psi_j$ and $|\alpha_\phi^g| \leq \|\phi\|_\infty \|g\|_\infty$. Thus implying the following bound on 29

$$c\epsilon\left(1 + \sum_{j=0}^{j_0} 2^{-j(\sigma_d + D/2 - D/p_d)} 2^{Dj/p_d'} 2^{-Dj/2}\right) \leq c\epsilon\left(1 + \sum_{j=0}^{j_0} 2^{-j\sigma_d}\right) \leq c\epsilon.$$

When the contamination is unstructured, by convexity we have,

$$\left\|\beta_j^g\right\|_{p_d'} = \left(\sum_{\psi\in\Psi_j} |\mathop{\mathbb{E}}_{g}\psi(X)|^{p_d'}\right)^{1/p_d'} \leq \left(\sum_{\psi\in\Psi_j} \mathop{\mathbb{E}}_{g}|\psi(X)|^{p_d'}\right)^{1/p_d'} \leq \left(\mathop{\mathbb{E}}_{g}\sum_{\psi\in\Psi_j}|\psi(X)|^{p_d'}\right)^{1/p_d'}$$

and we can interchange the expectation and sum in the last step because g is compactly supported which implies there are only finitely many non-zero terms to sum. The compactness of the wavelets implies only finitely many wavelets overlap at a point. So we have,

$$\int\left(\sum_{\psi\in\Psi_j}|\psi(x)|^{p_d'}\right)g(x)dx \leq c2^{Djp_d'/2} \implies \left\|\beta_j^g\right\|_{p_d'} \leq c2^{Dj/2} \tag{31}$$

where c might depend on the dimension. So we obtain the bound, (where the α term is bounded in the same way by a constant)

$$\begin{aligned}\epsilon d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{G}[\widehat{p}_n], 0\right) &\leq c\epsilon\left(1 + \sum_{j=0}^{j_0} 2^{-j(\sigma_d + D/2 - D/p_d)}2^{Dj/2}\right) = c\epsilon\left(1 + \sum_{j=0}^{j_0} 2^{j(D/p_d - \sigma_d)}\right) \\ &\leq c\epsilon\left(1 + 2^{j_0(D/p_d - \sigma_d)}\right)\end{aligned}$$

So, it only remains to bound the risk of the non-linear terms i.e.

$$d_{\mathcal{F}}\left(\sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j}\widetilde{\beta}_\psi\psi, \sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j}(\beta_\psi^p + \epsilon\beta_\psi^g)\psi\right)$$

From lemmas 13 and 14 we will upper bound the following:

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)}\left(\sum_{\psi\in\Psi_j}\mathbb{E}\left|(1-\epsilon)\beta_\psi^p + \epsilon\beta_\psi^g - \widetilde{\beta}_\psi\right|^{p_d'}\right)^{1/p_d'}$$

We will need the following moment and large deviation bounds from Uppal et al. [2019]:

Lemma 15. *(Moment Bounds) Let $X_1, \ldots, X_n \sim p$, $m \geq 1$ s.t. there is a constant c with $\mathbb{E}_p|\psi(X)|^m \leq c2^{Dj(m/2-1)}$ for all $\psi \in \Psi_j$. Let*

$$\gamma_\psi^p = \mathbb{E}[\psi(X)],$$
$$\widehat{\gamma}_\psi = \frac{1}{n}\sum_{i=1}^n \psi(X_i),$$

Then for all j s.t. $2^{Dj} \in \mathcal{O}(n)$,

$$\mathbb{E}[|\widehat{\gamma}_\psi - \gamma_\psi|^m] \leq cn^{-m/2}.$$

where $c = c_m\left(\mathbb{E}_p|\psi(X)|^2\right)^{m/2}$ is a constant.

Lemma 16. *(Large Deviations) Let $X_1, \ldots, X_n \sim p$ such that for a constant c, $\mathbb{E}_p |\psi(X)|^2 \leq c$ for $\psi \in \Psi_j$. Let*

$$\gamma_\psi^p = \mathbb{E}[\psi(X)],$$
$$\widehat{\gamma}_\psi = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i),$$

Let $l = \sqrt{j/n}$ and $\gamma > 0$, then, for all j s.t. $2^{Dj} \in o(n)$, we have,

$$\Pr(|\widehat{\gamma}_\psi - \gamma_\psi| > (K/2)l) \leq 2 \times 2^{-\gamma n l^2}$$

where K large enough such that

$$\frac{K^2}{8(c + \|\psi_\epsilon\|_\infty (K/3))} > \log 2\gamma$$

Both moment and large deviation bounds from above hold for all j s.t. $\mathbb{E}_p |\psi(X)|^m \leq c 2^{Dj(m/2-1)}$ which we have shown to hold for all $j \leq j_1$ (see equation 28) under both cases.

We now provide a general lemma bounding the non-linear term that we will also use when we provide a bound on the risk of the adaptive estimator.

Lemma 17. *Let $X_1, \ldots, X_n \sim p$ where p is compactly supported and $\mathcal{F}_d = B^{\sigma_d}_{p_d,q_d}$. If $\mathbb{E}_p |\psi(X)|^{p'_d} \leq c_{p'_d} 2^{Dj(p'_d/2-1)}$, then the risk of the non-linear terms of the wavelet thresholding estimator defined above i.e.*

$$d_{\mathcal{F}}\left(\sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j} \widetilde{\beta}_\psi \psi, \sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j} (\beta_\psi^p + \epsilon\beta_\psi^g)\psi\right)$$

is bounded by

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \frac{\left\|\beta_j^p\right\|_s^{s/p'_d} + \epsilon^{s/p'_d}\left\|\beta_j^g\right\|_s^{s/p'_d}}{\sqrt{n}^{1-s/p'_d}}. \tag{32}$$

for any $s \in [p_g, p'_d]$.

Proof. We follow the procedure of Donoho et al. [1996] and Uppal et al. [2019] and break up the term into different cases. The first two of which correspond to the situation where the empirical estimate and the true value of the co-efficient are far apart. Similar to the uncontaminated case, using the large deviation bounds above we show that the probability of this happening is negligible.

This leaves us with two cases to consider: when the estimate $\widehat{\beta}_\psi$ and the true coefficient are either both small or both large. We show that both of these cases reduce to the same term which we then bound using the properties of Besov spaces and the compactness of all densities considered.

1. Let A be the set of $\psi \in \Psi_j$ s.t. $\widehat{\beta}_\psi > t$ and $(1-\epsilon)\beta_\psi^p + \epsilon\beta_\psi^g < t/2$ and $r \geq 1/p'_d$ then by Hölder's inequality,

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \times \left(\sum_{\psi\in\Psi_j} \mathbb{E}\,|\beta_\psi^p - \widetilde{\beta}_\psi|^{p'_d} 1_A\right)^{1/p'_d}$$
$$\leq \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \times \left(\sum_{\psi\in\Psi_j} (\mathbb{E}\,|\beta_\psi^p - \widetilde{\beta}_\psi|^{p'_d r})^{1/r} \Pr(A)^{1/r'}\right)^{1/p'_d}.$$

   Using the large deviation and moment bound we get an upper bound,

$$\sum_{j=j_0}^{j_1} c 2^{-j(\sigma_d + D/2 - D/p_d)} \left(2^{Dj} n^{-p'_d/2} 2^{-j\gamma/r'}\right)^{1/p'_d}$$
$$\leq c n^{-1/2} 2^{-j_0(\sigma_d - D/2 + \gamma/p'_d r')}$$

   which is negligible compared to the linear term for large enough γ.

2. Let B be the set of $\psi \in \Psi_j$ s.t. $\widehat{\beta}_\psi < t$ and $(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi > 2t$ then same as above

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left\|\beta^p_j + \epsilon\beta^g_j\right\|_{p'_d} 2^{-\gamma j/p'_d}$$
$$\leq 2^{-j_0(\sigma_d+\sigma'_g+\gamma)} + \epsilon \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} 2^{-\gamma j/p'_d} \left\|\beta^g_j\right\|_{p'_d}$$

which is negligible compared to the bias term and the misspecification error for large enough γ.

In other words, for the upper bounds of the first two cases we have chosen γ (which in turn determines the value of the constant K for the threshold $t = K\sqrt{j/n}$) to be large enough so that the exponent of 2^j in the upper bound of these two terms is negative. This enables us to upper bound the geometric series (as a sum of j) by a constant multiple of the first term.

3. Let C be the set of $\psi \in \Psi_j$ s.t. $\widehat{\beta}_\psi > t$ and $(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi > t/2$ then for any $p_g \leq s \leq p'_d$,

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \times \left(\sum_{\psi\in C} \mathbb{E}\left|(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi - \widetilde{\beta}_\psi\right|^{p'_d}\right)^{1/p'_d}$$
$$\leq \sum_{j=j_0}^{j_1} C 2^{-j(\sigma_d+D/2-D/p_d)} \sqrt{j}^{s/p'_d} \times \frac{\left\|(1-\epsilon)\beta^p_j + \epsilon\beta^g_j\right\|_s^{s/p'_d}}{\sqrt{n}^{1-s/p'_d}}$$

where we have used the moment bound and the lower bound on $(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi$.

4. Let E be the set of $\psi \in \Psi_j$ s.t. $\widehat{\beta}_\psi < t$ and $(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi < 2t$ then for any $p_g \leq s \leq p'_d$:

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left(\sum_{\psi\in E} |\beta^p_\psi|^{p'_d}\right)^{1/p'_d}$$
$$\leq \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \times \left(\sum_{\psi\in\Psi_j} |(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi|^s (2t)^{p'_d-s}\right)^{1/p'_d}$$
$$= \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \sqrt{j}^{s/p'_d} \times \frac{\left\|(1-\epsilon)\beta^p_j + \epsilon\beta^g_j\right\|_s^{s/p'_d}}{\sqrt{n}^{1-s/p'_d}}$$

where we have used the upper bound on $(1-\epsilon)\beta^p_\psi + \epsilon\beta^g_\psi$.

By applying Jensen's inequality we can show that both 3 and 4 above are bounded, for any $s \in [p_g, p'_d]$, by the following (where we omit the $\sqrt{j}$ term since it only contributes a factor of polylog of n or ϵ to the upper bound),

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{\left\|\beta^p_j\right\|_s^{s/p'_d} + \epsilon^{s/p'_d}\left\|\beta^g_j\right\|_s^{s/p'_d}}{\sqrt{n}^{1-s/p'_d}}.$$

□

We can now bound 32 by

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{\left\|\beta^p_j\right\|_{p_g}^{s/p'_d} + \epsilon^{s/p'_d}\left\|\beta^g_j\right\|_{p_g}^{s/p'_d}}{\sqrt{n}^{1-s/p'_d}}.$$

Let A be the set of j s.t. $\left\|\beta_j^p\right\|_{p_g} \geq \epsilon \left\|\beta_j^g\right\|_{p_g}$ and $B = [j_0, j_1] \setminus A$. Then the above is upper bounded by

$$\begin{aligned}
&\sum_{j\in A} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{\left\|\beta_j^p\right\|_{p_g}^{p_g/p_d'}}{\sqrt{n}^{1-p_g/p_d'}} + \epsilon \sum_{j\in B} 2^{-j(\sigma_d+D/2-D/p_d)} \left\|\beta_j^g\right\|_{p_g} \\
&\quad \leq \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{\left\|\beta_j^p\right\|_{p_g}^{p_g/p_d'}}{\sqrt{n}^{1-p_g/p_d'}} + \epsilon \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left\|\beta_j^g\right\|_{p_g} \\
&\quad \leq n^{1/2(p_g/p_d'-1)} 2^{-j_m((\sigma_g+D/2)p_g/p_d'+\sigma_d-D/2)} + \epsilon \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left\|\beta_j^g\right\|_{p_g}
\end{aligned}$$

where the second term is bounded by the misspecification error.

So for both the structured and unstructured contamination setting the bound on all terms except the misspecification error is the same. In particular, we have, for the structured setting,

$$\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}} + 2^{-j_0(\sigma_d+\sigma_g-D/p_g+D/p_d')} + n^{1/2(p_g/p_d'-1)} 2^{-j_m((\sigma_g+D/2)p_g/p_d'+\sigma_d-D/2)} + \epsilon$$

which gives,

$$\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + n^{-\frac{\sigma_g+\sigma_d-D/p_g+D/p_d'}{2\sigma_g+D-2D/p_g}} + \epsilon$$

In contrast for the unstructured setting we have,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}} + 2^{-j_1(\sigma_d+\sigma_g-D/p_g+D/p_d')} + n^{1/2(p_g/p_d'-1)} 2^{-j_m((\sigma_g+D/2)p_g/p_d'+\sigma_d-D/2)} \\
&\epsilon 2^{Dj_1(D/p_d-\sigma_d)} + \epsilon
\end{aligned}$$

At the given values of j_0 and j_1 this gives us an upper bound of

$$\begin{aligned}
&\frac{1}{\sqrt{n}} + \sqrt{n}^{-\frac{\sigma_g+\sigma_d}{\sigma_g+D/2}} + \sqrt{n}^{-\frac{\sigma_g+\sigma_d+D/p_d'-D/p_g}{\sigma_g+D/2-D/p_g}} + \\
&\epsilon + \epsilon^{\frac{\sigma_g+\sigma_d+D/p_d'-D/p_g}{\sigma_g+D-D/p_g}}.
\end{aligned}$$

We note that the above proof implicitly assumes that $p_d' < \infty$ or equivalently $p_d > 1$. We provide a bound for the case $p_d' = \infty$ in the next section since in this case it is sufficient to look at linear estimators.

B.2 Linear Rate

In this section we provide an upper bound on the risk of the linear wavelet estimator which is simply a non-linear estimator with the added constraint that $j_0 = j_1$ or without its non-linear terms. Again, in view of lemma 9 it is sufficient to consider the case $p_d' \geq p_g$.

We use the upper bounds on the components of the error of the non-linear wavelet estimator computed above with the additional constraint that $j_0 = j_1$. Therefore we have the following upper bounds along with the implied rate. In the unstructured setting, the upper bound is

$$\begin{aligned}
&\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}} + 2^{-j_0(\sigma_d+\sigma_g-D/p_g+D/p_d')} \\
&\quad \epsilon + \epsilon 2^{Dj_0(D/p_d-\sigma_d)}
\end{aligned}$$

which implies the rate

$$\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d+D/p_d'-D/p_g}{2\sigma_g+D-2D/p_g+2D/p_d'}} + \epsilon + \epsilon^{\frac{\sigma_g+\sigma_d+D/p_d'-D/p_g}{\sigma_g+D-D/p_g}}.$$

In the structured setting, the upper bound is,

$$\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}} + 2^{-j_0(\sigma_d+\sigma_g-D/p_g+D/p'_d)} + \epsilon$$

which implies the rate

$$\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d+D/p'_d-D/p_g}{2\sigma_g+D-2D/p_g+2D/p'_d}} + \epsilon.$$

For $p'_d = \infty$ the bounds on the bias or the misspecification error still hold i.e.

$$2^{-j_0(\sigma_g+\sigma_d-D/p_g)} + \epsilon + \epsilon 2^{j_0(D-\sigma_d)}$$

The variance bound is given by

$$\sum_{j=0}^{j_0} 2^{-j(\sigma_d+D/2-D/p_d)} \mathbb{E} \sup_{\psi\in\Psi_j} |\widehat{\beta}_\psi - \beta^p_\psi| \leq \frac{2^{j(D/2-\sigma_d)}}{\sqrt{n}}$$

This is shown using the lemma above bounding large deviations. We have,

$$\mathcal{P}\left(\sup_{\psi\in\Psi_j} |\widehat{\beta}_\psi - \beta^p_\psi| \geq K\sqrt{j/n}\right) \leq 2^{j(D-\gamma)}$$

which implies

$$\sum_{j=0}^{j_0} \frac{2^{j(2D-\sigma_d-\gamma)}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$$

for γ sufficiently large.

Now, with a choice of $2^{j_0} = n^{\frac{1}{2\sigma_g+D-2D/p_g}} \wedge \epsilon^{-\frac{1}{\sigma_g+D-D/p_g}}$ we have an upper bound of

$$\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d-D/p_g}{2\sigma_g+D-2D/p_g}} + \epsilon + \epsilon^{\frac{\sigma_g+\sigma_d-D/p_g}{\sigma_g+D-D/p_g}}$$

Similarly, when contamination is structured we have a bound of

$$\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d-D/p_g}{2\sigma_g+D-2D/p_g}} + \epsilon.$$

B.3 Dense case: $p'_d \leq p_g$

Here we provide a better upper bound on the risk when $p'_d \leq p_g$ using the linear estimator from above. In this case we obtain a better bound without using the monotonicity of the dual Besov norms from lemma 9. While most of the components of the proof are the same as in the non-linear case of Section B.1 the bound on the variance is a little more involved.

Proof. Given $X_1, \ldots, X_n \overset{IID}{\sim} (1-\epsilon)p + \epsilon g$. Let our estimator be $\widehat{p} = \frac{1}{1-\epsilon}\widehat{p}_n$ where $\widehat{p}_n$ is the linear wavelet estimator defined above with

$$2^{j_0} = n^{\frac{1}{2\sigma_g+D}} \wedge \epsilon^{-\frac{1}{\sigma_g+D/p_d}}$$

We use the same bias variance decomposition as in the proof of the non-linear estimator with the additional constraint that $j_0 = j_1$. Therefore we have no non-linear terms to bound. We have the following upper bound on the error.

$$\begin{aligned}\mathbb{E}\, d_{\mathcal{F}_d}(\widehat{p}, p) \leq \frac{1}{1-\epsilon}\mathbb{E}\, d_{\mathcal{F}}(\widehat{p}_n, \mathop{\mathbb{E}}_{(1-\epsilon)p+\epsilon g}[\widehat{p}_n]) \\ + d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{p}[\widehat{p}_n], p\right) + \frac{\epsilon}{1-\epsilon} d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{g}[\widehat{p}_n], 0\right)\end{aligned}$$

where the first term is the stochastic error or the variance, the second term is the bias and the third term is the misspecification error. Here again, the bound on the bias is unchanged. Moreover, the misspecification error can be bounded in the same way as in the proof of the sparse case above. We can also show here that the variance bound remains the same but this is not as straightforward since we don't just have lower resolution terms.

Using lemma 11 we get the same bound as before on the bias or the second term above i.e.

$$d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{p}[\widehat{p}_n], p\right) \leq c2^{-j_0(\sigma_d+\sigma_g-(D/p_g-D/p'_d)_+)}$$

Now we bound the variance or the first term. Let $\psi \in \Psi_j$ and

$$Y_i = \psi_(X_i) - \mathbb{E}[\psi(X)]$$

then for all $m \geq 1$, applying first the triangle inequality and then Jensen's inequality repeatedly we get

$$\begin{aligned}\mathbb{E}[|Y_i|^m] &\leq \mathbb{E}[(|\psi(X_i)| + |\,\mathbb{E}[\psi(X_i)]|)^m] \\ &\leq 2^{m-1}\left(\mathbb{E}[|\psi(X_i)|^m] + |\,\mathbb{E}[\psi(X_i)]|^m\right) \\ &\leq 2^m\,\mathbb{E}[|\psi(X_i)|^m].\end{aligned}$$

Therefore, by Rosenthal's inequality, i.e.,

Lemma 18. *(**Rosenthal's Inequality** (Rosenthal [1970])) Let $m \in \mathbb{R}$ and $Y_1, \ldots, Y_n$ be IID random variables with $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] \leq \sigma^2$. Then there is a constant c_m that depends only on m s.t.*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right|^m\right] \leq c_m\left(\frac{\sigma^m}{n^{m/2}} + \frac{\mathbb{E}\,|Y_1|^m}{n^{m-1}}1_{2<m<\infty}\right)$$

we have,

$$\begin{aligned}&\mathbb{E}[|\widehat{\beta}_\psi - (1-\epsilon)\beta_\psi^p - \epsilon\beta_\psi^g|^{p'_d}] \\ &\leq c_{p'_d}\left(\left(\mathbb{E}\,|\psi(X)|^2\right)^{p'_d/2} + \frac{\mathbb{E}[|\psi(X)|^{p'_d}]}{n^{p'_d-1}}1_{p'_d\geq 2}\right)\end{aligned}$$

where $c_{p'_d}$ is a constant that only depends on p'_d.

This implies that the variance is bounded by:

$$\begin{aligned}&\sum_{j=0}^{j_0} 2^{-j(\sigma_d+D/2-D/p_d)}\times \\ &\left(\sum_{\psi\in\Psi_j}\mathbb{E}\,|\widehat{\beta}_\psi - (1-\epsilon)\beta_\psi^p - \epsilon\beta_\psi^g|^{p'_d}\right)^{1/p'_d} \\ \leq &\sum_{j=0}^{j_0}\frac{2^{-j(\sigma_d+D/2-D/p_d)}}{\sqrt{n}}\times \\ &\left(\sum_{\psi\in\Psi_j}\left(\mathbb{E}\,|\psi(X)|^2\right)^{p'_d/2} + \frac{\mathbb{E}[|\psi(X_i)|^{p'_d}]}{n^{p'_d/2-1}}1_{p'_d\geq 2}\right)^{1/p'_d}\end{aligned}$$

Now we can bound each of the terms inside the brackets separately. The second term is bounded as

$$\begin{aligned}
&\sum_{\psi\in\Psi_j} \frac{\mathbb{E}[|\psi(X_i)|^{p'_d}]}{n^{(p'_d/2-1)}} 1_{p'_d\geq 2} \\
&\leq \sum_{\psi\in\Psi_j} \frac{(1-\epsilon)\,\mathbb{E}_p[|\psi(X_i)|^{p'_d}] + \epsilon\,\mathbb{E}_g[|\psi(X_i)|^{p'_d}]}{n^{(p'_d/2-1)}} 1_{p'_d\geq 2} \\
&\leq \sum_{\psi\in\Psi_j} \frac{2^{Dj}2^{Dj(p'_d/2-1)} + \epsilon 2^{Djp'_d/2}}{n^{(p'_d/2-1)}} 1_{p'_d\geq 2} \\
&\leq \sum_{\psi\in\Psi_j} \frac{2^{Djp'_d/2}}{n^{(p'_d/2-1)}} 1_{p'_d\geq 2} \leq 2^{Dj} 1_{p'_d\geq 2}
\end{aligned}$$

While the first term is bounded as

$$\begin{aligned}
&\sum_{\psi\in\Psi_j} \left(\mathbb{E}\,|\psi(X)|^2\right)^{p'_d/2} \\
&\leq \sum_{\psi\in\Psi_j} \left((1-\epsilon)\mathop{\mathbb{E}}_{p}|\psi(X)|^2 + \epsilon \mathop{\mathbb{E}}_{g}|\psi(X)|^2\right)^{p'_d/2} \\
&\leq \sum_{\psi\in\Psi_j} \left((1-\epsilon)\,\|p\|_\infty + \epsilon 2^{Dj} w_\psi\right)^{p'_d/2}
\end{aligned}$$

where w_ψ is $\int 1_{\text{supp}(\psi)} g(x)dx$. Since we know that at any point at most finitely many wavelets intersect $\sum w_\psi \leq c\int g(x)dx = c$.

For $p'_d \leq 2$ by Jensen's we have,

$$\begin{aligned}
&2^{Dj}\left(\frac{1}{2^{Dj}}\sum_{\psi\in\Psi_j}(1-\epsilon)\,\|p\|_\infty + \epsilon 2^{Dj} w_\psi\right)^{p'_d/2} \\
&\leq 2^{Dj}\,(c+\epsilon)^{p'_d/2} \leq 2^{Dj}
\end{aligned}$$

For $p'_d \geq 2$ again by Jensen's, we have,

$$\begin{aligned}
(1-\epsilon)2^{Dj} + (\epsilon 2^{Dj})^{p'_d/2}\,\|w\|_{p'_d/2}^{p'_d/2} \\
\leq 2^{Dj} + (\epsilon 2^{Dj})^{p'_d/2}
\end{aligned}$$

where we have used the fact that $\|w\|_{p'_d/2} \leq \|w\|_1$. Since $2^{Dj_0} \leq (1/\epsilon)^{\frac{D}{\sigma_g + D/p_d}}$, for every $j \leq j_0$, $\epsilon \leq 2^{-j(\sigma_g + D/p_d)}$. This implies

$$(\epsilon 2^{Dj})^{p'_d/2} \leq 2^{j(D/p'_d - \sigma_g)p'_d/2} = 2^{j(D/2-\sigma_g p'_d/2)} \leq 2^{Dj}$$

In conclusion, the sum of the variance terms at any resolution j (not too large) is bounded by $2^{Dj/p'_d}$. Therefore, we have an upper bound for the variance term, which is the same as usual, i.e.,

$$\begin{aligned}
&\sum_{j=0}^{j_0} 2^{-j(\sigma_d + D/2 - D/p_d)} n^{-1/2} 2^{Dj/p'_d} \\
&\leq \sum_{j=0}^{j_0} 2^{j(D/2-\sigma_d)} n^{-1/2} \\
&\leq \frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}}
\end{aligned}$$

It only remains to bound the last term, or the misspecification error, which we can bound in the same as the non-linear case of section B.1 i.e. we have,

$$\frac{\epsilon}{1-\epsilon} d_{\mathcal{F}}\left(\mathop{\mathbb{E}}_{g}[\widehat{p}_n], 0\right) \leq c\epsilon\left(1 + 2^{j_0(D/p_d - \sigma_d)}\right)$$

Therefore, our upper bound is $\lesssim$

$$\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + n^{-\frac{\sigma_g+\sigma_d+D/p'_d-D/p_g}{2\sigma_g-2D/p_g+2D/p'_d+D}} + \epsilon + \epsilon^{\frac{\sigma_g+\sigma_d}{\sigma_g+D/p_d}} + \epsilon^{\frac{\sigma_g+\sigma_d+D/p'_d-D/p_g}{\sigma_g-D/p_g+D}}$$

□

B.4 Adaptivity

We now provide a version of the thresholding wavelet estimator above that is, under the structured contamination setting, adaptive to both the contamination proportion ϵ and smoothness of the true density σ. This essentially follows from the argument provided by Donoho et al. [1996] except that we extend it to higher dimensions. We reproduce the proof here for completeness

Given lemma 9 we only consider the case $p'_d \geq p_g$.

We now construct the adaptive version of the thresholding wavelet estimator.

Firstly, we no longer use a scaled version of $\widehat{p}_n$ but the estimator $\widehat{p}_n$ itself. This makes it adaptive to the contamination proportion ϵ and we will show that this costs us only a constant factor in the asymptotic rate. Secondly, we follow Donoho et al. [1996] and pick the following values for the resolution levels j_0, j_1,

$$2^{j_0} = n^{\frac{1}{D+2r}}$$
$$2^{j_1} = \left(\frac{n}{\log n}\right)^{1/D}$$

where r is the regularity of the wavelets used to construct the MRA defined above. We can decompose the error as

$$\begin{aligned}
&\mathbb{E}\, d_{\mathcal{F}}\left(\widehat{p}_n, p\right) \\
&\quad \leq \mathbb{E}\, d_{\mathcal{F}}\left(\sum_{k\in\mathbb{Z}} \widehat{\alpha}_k \phi_k + \sum_{j=0}^{j_0}\sum_{\psi\in\Psi_j} \widehat{\beta}_\psi \psi, \sum_{k\in\mathbb{Z}} \alpha_k^p \phi_k^p + \sum_{j=0}^{j_0}\sum_{\psi\in\Psi_j} \beta_\psi^p \psi\right) \\
&\quad + d_{\mathcal{F}}\left(\sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j} \widetilde{\beta}_\psi \psi, \sum_{j=j_0}^{j_1}\sum_{\psi\in\Psi_j} \beta_\psi^p \psi\right) \\
&\quad + d_{\mathcal{F}}\left(\sum_{k\in\mathbb{Z}} \alpha_k^p \phi_k + \sum_{j=0}^{j_1}\sum_{\psi\in\Psi_j} \beta_\psi^p \psi, p\right) \\
&\quad + \epsilon d_{\mathcal{F}}\left(\sum_{k\in\mathbb{Z}} \alpha_k^g \phi_k + \sum_{j=0}^{j_1}\sum_{\psi\in\Psi_j} \beta_\psi^g \psi, 0\right) \\
&\quad + \epsilon d_{\mathcal{F}}\left(\sum_{k\in\mathbb{Z}} \alpha_k^p \phi_k + \sum_{j=0}^{j_1}\sum_{\psi\in\Psi_j} \beta_\psi^p \psi, 0\right)
\end{aligned}$$

where we have an extra term at end as opposed to the non-adaptive case above. Since the density p is bounded above this term is bounded by ϵ and hence by the misspecification error. The bound on the misspecification error does not change since it does not depend on the values of j_0 or j_1.

Now, since, $\sigma_g < r$ we know that the number of linear terms or j_0 has is smaller than above. Moreover, since $\sigma_g > D/p_g$ the number of non-linear terms j_1 is larger than above. Therefore, it is clear that the bias and the variance bounds hold as above. It only remains to bound the non-linear terms which from lemma 17 amounts to bounding

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{\left\|\beta_j^p\right\|_s^{s/p_d'} + \epsilon^{s/p_d'} \left\|\beta_j^g\right\|_s^{s/p_d'}}{\sqrt{n}^{1-s/p_d'}}. \tag{33}$$

which following the same procedure as above is bounded above by

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{\left\|\beta_j^p\right\|_s^{s/p_d'}}{\sqrt{n}^{1-s/p_d'}} + \epsilon \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} 2^{Dj(1/p_d'-1/2)}$$

where the second term is bounded by the misspecification error. Now the first term is the same as in the case of uncontaminated setting and thereby we can bound it in the same way.

When $(2\sigma_g + D)p_g \leq (D - 2\sigma_d)p_d'$ let $s = -p_d' \frac{2\sigma_d+D-2D/p_d}{2\sigma_g+D-2D/p_g}$. Note that

$$p_g \leq -p_d' \frac{2\sigma_d + D - 2D/p_d}{2\sigma_g + D - 2D/p_g} \leq p_d'$$

where the first inequality is equivalent to the condition above and the second is equivalent to $\sigma_g \geq -\sigma_d$. We have the following bound, when we pick ,

$$\sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \frac{2^{-j(\sigma_g+D/2-D/p_g)s/p_d'}}{\sqrt{n}^{1+\frac{2\sigma_d+D-2D/p_d}{2\sigma_g+D-2D/p_g}}}$$
$$\asymp n^{-\frac{\sigma_g+\sigma_d+D/p_d'-D/p_g}{2\sigma_g+D-2D/p_g}}$$

as desired (where we omit any $\log(n)$ terms).

Now, when $(2\sigma_g + D)p_g \geq (D - 2\sigma_d)p_d'$ the error of the non-linear terms is bounded by the error of the first non-linear term i.e. $j = j_0$. We can bound the error of the non-linear terms for all $j \geq j_0^*$ where j_0^* is the original non-adaptive threshold i.e. $2^{j_0} = n^{\frac{1}{2\sigma_g+D}}$. For the extra terms between j_0 and j_0^* we show that in this range the error of the non-linear terms cannot be worse that the linear error. The large deviation terms (1) and (2) above are negligible by the same argument as above. The terms (3) and (4) can be trivially bounded by

$$2^{-j(\sigma_d+D/2-D/p_d)} \frac{2^{Dj/p_d}}{\sqrt{n}} = \frac{1}{\sqrt{n}} 2^{j(D/2-\sigma_d)}$$

which is bounded by the linear rate.

C Lower Bounds

In this section we prove our lower bounds. We first provide lower bounds in the case of structured contamination since these also hold for the case of unstructured contamination. We then provide additional lower bounds that are specific to the unstructured case.

C.1 Structured Contamination

We assume here that G has a density that lives in a Besov space i.e. $\mathcal{F}_c = B_{p_c,q_c}^{\sigma_c}$.

Proof. We will use Fano's lemma to imply lower bounds here.

First we show that the lower bounds on the risk in the setting of no contamination also bound the risk in the contaminated setting. The key idea here is that if the set of densities chosen to provide bounds in the uncontaminated setting (when $\epsilon = 0$) are perturbations of a "nice" density p_0, then in

the contaminated setting we can choose our contamination density to be this nice density g_0. This will imply that the contamination does not affect the samples i.e. the samples are generated merely from the perturbation (since $(1-\epsilon)(g_0+p_\tau)+\epsilon g_0 = g_0+(1-\epsilon)p_\tau$ where p_τ is some perturbation).

We first state Fano's lemma.

Lemma 19. *(***Fano's Lemma***; Simplified Form of Theorem 2.5 of Tsybakov [2009]) Fix a family $\mathcal{P}$ of distributions over a sample space $\mathcal{X}$ and fix a pseudo-metric $\rho : \mathcal{P}\times\mathcal{P}\to[0,\infty]$ over $\mathcal{P}$. Suppose there exists a set $T\subseteq\mathcal{P}$ such that there is a $p_0\in T$ with $p\ll p_0\ \forall p\in T$ and*

$$s := \inf_{p,p'\in T}\rho(p,p') > 0 \quad , \quad \sup_{p\in T} D_{KL}(p,p_0) \le \frac{\log|T|}{16},$$

where $D_{KL}:\mathcal{P}\times\mathcal{P}\to[0,\infty]$ denotes Kullback-Leibler divergence. Then,

$$\inf_{\widehat{p}}\sup_{p\in\mathcal{P}}\mathbb{E}\left[\rho(p,\widehat{p})\right] \ge \frac{s}{16}$$

where the inf *is taken over all estimators $\widehat{p}$.*

Now we choose our set of densities as,

$$p = p_0 \qquad\qquad p_\tau = p_0 + \frac{1}{(1-\epsilon)}c_g f_\tau$$
$$g = p_0 \qquad\qquad g_\tau = p_0.$$

where $p_0 + c_g f_\tau \in \mathcal{F}_g$ for every τ. Notice that the KL divergence remains unchanged from the uncontaminated setting,

$$\begin{aligned} &KL((1-\epsilon)p_0+\epsilon p_0,(1-\epsilon)(p_0+\frac{1}{1-\epsilon}c_g g_\tau)+\epsilon p_0) \\ &\qquad = KL(p_0,p_0+c_g f_\tau)\end{aligned}$$

i.e. the KL divergence doesn't depend on the existence of contamination. Neither does $d_{\mathcal{F}_d}(p_\tau,p_{\tau'})$. Since, $1-\epsilon\in[1/2,1]$ we can treat it as a constant and only write c_g henceforth. Therefore we are essentially in the case of no contamination i.e. if there exist densities p,p_τ indexed by τ such that they satisfy the assumptions of Fano's lemma then the conditions of Fano's lemma are also satisfied for $(1-\epsilon)p+\epsilon g$, $(1-\epsilon)p_\tau+\epsilon g_\tau$. Moreover, the distance we want to bound i.e. $d_{\mathcal{F}_d}(p_\tau,p_{\tau'})$ does not depend on the contamination either. Therefore, we have a lower bound here that is the same as the one in the setting with no contamination.

We note that the densities used to prove the lower bound in Uppal et al. [2019] are exactly of this form (see section B of the appendix) (Uppal et al. [2019] study the uncontaminated version of this problem). Therefore, their lower bound (see Theorem 4) is implied here i.e.

$$C\left(\frac{1}{\sqrt{n}} + n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + \left(\frac{\log n}{n}\right)^{\frac{\sigma_g+\sigma_d-D/p_g+D/p'_d}{2\sigma_g+D-2D/p_g}}\right)$$

Second, we consider the case where we "move" the perturbation so that the samples are generated from the same density. In particular, we first perturb the contamination and then we move this perturbation to the true density i.e.

$$p = p_0 \qquad\qquad \widetilde{p} = p_0 + \frac{\epsilon}{1-\epsilon}c\psi_\epsilon$$
$$g = g_0 + c\psi_\epsilon \qquad\qquad \widetilde{g} = g_0$$

Then the KL divergence between the densities that generate the samples is zero since they are the same ($(1-\epsilon)p+\epsilon g$ is the same in both cases). It is easy to see that $\widetilde{p}, g$ both live in the respective density classes i.e. $\mathcal{F}_g,\mathcal{F}_c$ for a small enough constant c. Using Le Cam's two point argument i.e.

Lemma 20. *(Le Cam (see section 2.3 of Tsybakov [2009])) Let P_1, P_2 be two probability measures on $\mathcal{X}$ s.t. $d(P_1,P_2)=s$. If $KL(P_1,P_2)\le\alpha<\infty$ then, for any $\widehat{P}$*

$$\mathop{\mathbb{E}}_{P_i}[d(\widehat{P},P_i)] \ge \frac{s}{8}e^{-\alpha}$$

we have a lower bound that is the distance between p_0 and $p_0 + \frac{\epsilon}{1-\epsilon}\psi_\epsilon$ i.e.

$$d_{\mathcal{F}_d}(p_0, p_0 + \frac{\epsilon}{1-\epsilon}\psi_\epsilon) = \epsilon.$$

□

This section provided lower bounds on the risk that are minimax in the structured contamination setting. We now provide additional bounds that hold when we have no structural assumptions on the contamination.

C.2 Unstructured Contamination

Here we assume only that g is a compactly supported probability density. We will pick a single perturbation of a "nice" density and use this to construct the contamination densities in such a way that the data is generated from the same density. Hence, the KL divergence between the data generating densities will be zero. Then, as before, we can apply Le Cam's two point argument to bound the risk.

C.2.1 Sparse or Lower Smoothness Case

Let $p = g_0$, $\widetilde{p} = g_0 + c_g\psi_0$ for some $\psi_0 \in \Psi_j$. Now we can pick densities $g, \widetilde{g}$ such that

$$(1-\epsilon)p + \epsilon g = (1-\epsilon)\widetilde{p} + \epsilon\widetilde{g}$$

if and only if

$$g - \widetilde{g} = \frac{(1-\epsilon)}{\epsilon}(\widetilde{p} - p)$$

integrates to zero and its L_1 norm is ≤ 2 (see Lemma 6.6 of Liu and Gao). For $\widetilde{p}$ to be a density in $\mathcal{F}_g$ we need

$$c_g \leq c\min(2^{-Dj/2}, 2^{-j(\sigma_g + D/2 - D/p_g)})$$

Since $\sigma_g \geq D/p_g$, we let $c_g = 2^{-j(\sigma_g + D/2 - D/p_g)}$. From the above constraint on the L_1 of $g - \widetilde{g}$ norm we need

$$\frac{c_g}{\epsilon}2^{-Dj/2}\left\|\psi\right\|_\infty \leq 2$$

This is equivalent to

$$2^{-j(\sigma_g + D - D/p_g)} \leq c\epsilon$$

where c is a constant. We pick $2^j = \epsilon^{-\frac{1}{\sigma_g + D - D/p_g}}$. We also choose a simple discriminator i.e.

$$\Omega_d = \{c_d\psi_0\}$$

where $c_d = 2^{-j(\sigma_d + D/2 - D/p_d)}$ so that $\Omega_d \subseteq \mathcal{F}_d$. Then, by 20 the minimax risk is lower bounded by

$$\begin{aligned} d_{\mathcal{F}_d}(p, \widetilde{p}) &\geq d_{\Omega_d}(p, \widetilde{p}) \\ &\gtrsim c_g c_d \\ &= c2^{-j(\sigma_g + \sigma_d + D - D/p_g - D/p_d)} \\ &= \epsilon^{\frac{\sigma_g + \sigma_d + D/p'_d - D/p_g}{\sigma_g + D - D/p_g}}. \end{aligned}$$