We thank all the reviewers for their comments and suggestions.

Reviewers #1 & #4:

"More details regarding the implications for real-world data would be beneficial."
Due to space constraints we omitted a discussion about the “breakdown point” which may have practical implications. The “asymptotic breakdown point” is the maximum asymptotic proportion ϵ of outlier samples such that the estimator still converges at the uncontaminated optimal rate; i.e., the proportion of corruption the estimator can tolerate before performance degrades.

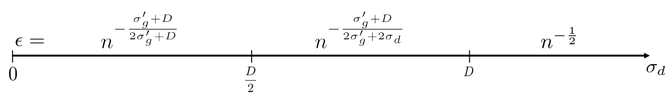

Figure 1: Asymptotic breakdown point when $p_d' = \infty$ and $\sigma_g' = \sigma_g - D/p_g$.

Figure 1 illustrates the asymptotic breakdown point for the case $p_d' = \infty$ as a function of σ_d; this includes as special cases the L_∞ loss and Kolmogorov-Smirnov metric.

In Figure 2 we pick $\sigma_g' = D = 5$ for the above setting to show how the breakdown point increases as a function of σ_d (smoothness of functions that define the IPM; larger smoothness implies weaker loss). So, estimation under a weaker loss may be more robust to contamination. We will add more discussion about how understanding breakdown points could be useful in the context of real-world data.

"A test on synthetic data for different contaminations would be ideal."
We felt it would be difficult for experiments on synthetic data to elucidate our results, due to their asymptotic nature. However, if the reviewers feel this would help clarify the significance of the paper we will add it for the camera ready version.

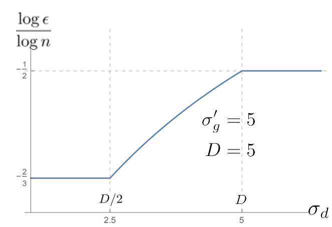

Figure 2: Exponent of n in breakdown point as a function of σ_d.

"It is not clear where the claim that a GAN with ‘ReLu activations can learn the distribution of the ...’ is shown."
In corollary 6 we extend the result of Uppal et. al. [2019] to show that a perfectly optimized GAN estimate (of the form of eq. (7)) with large enough fully-connected ReLU generator and discriminator networks converges (under Besov IPMs) to the true Besov distribution at the minimax optimal rate, both adaptively and in the presence of contamination. We will clarify this in the paper.

“Line 305 : “estimator of 4.2”, what is 4.2 referring to?”
We meant to refer to the wavelet thresholding estimator defined in section 4.

Reviewer #2:

"line 102, could the authors check the definition of the Haar wavelets scaled from the mother wavelet? There is no λ in the definition."
In dimension D there are $2^D - 1$ mother wavelets ψ_ϵ indexed by $\epsilon \in \{0,1\}^D \setminus (0,\ldots,0)$. For each of these mother wavelets, daughter wavelets $2^{Dj/2}\psi_\epsilon(2^{Dj}x - k)$ are constructed by translation (by integers k) and scaling (by 2^{-Dj} horizontally and $2^{Dj/2}$ vertically). At resolution j, such wavelets can be indexed by $\lambda = 2^{-j}k + 2^{-j-1}\epsilon$, as any such value of λ uniquely defines k and ϵ. For example, in dimension $D = 1$, $\lambda = 1.5$ implies that $j = 0$ and $k = 1$. We will clarify this re-indexing in the paper.

Reviewer #3:
"This rate, interestingly, is better than the rate achieved for density estimation at a point"
The rates differ primarily in the ϵ term i.e. $\epsilon^{-\frac{\sigma_0}{2\sigma_0+1}}$ vs $\epsilon^{-\frac{\sigma_0}{2\sigma_0+1/2}}$ (section 3.2) which is obtained when the misspecification error dominates the variance. The misspecification error $\epsilon d_{\mathcal{F}_d}(\mathbb{E}_g[\widehat{p}_n])$ (eq. (19)) is the error due to contamination. The point-wise misspecification error scales as the $\mathcal{L}^\infty$ error or $\epsilon 2^{Dj}$ with resolution j which is larger than the $\mathcal{L}^{p_d'}$ error which scales as $\epsilon 2^{Dj/p_d}$. Intuitively, this reflects the fact that while under both losses the “worst-case” contamination densities are “spikes” concentrated around single points (rather than, say, uniform over the domain), the $\mathcal{L}^\infty$ loss, as well as estimation at a point, are more sensitive to such contamination than, e.g., $\mathcal{L}^2$ loss. We will try to clarify this in the paper.

"the description of the wavelet estimator (in particular how it differed from that of Donoho)"
The wavelet estimator that we have used is similar to that of Donoho et. al. [1996]. We merely chose to estimate the linear terms (up to resolution j_0) with basis $(\cup_k \left(\phi_{0k} \cup \bigcup_{0 \le j \le j_0} \psi_{jk}\right))$ as opposed to $\cup_k \phi_{jk}$. These bases are equivalent. However, no prior work has studied the *robustness properties of this estimator in the presence of contamination*. As the reviewer suggested, we will add discussion of the novel aspects of the proofs to the main paper.