We thank the reviewers for the valuable time they have invested during this difficult period to review the paper and provide helpful suggestions for improving the manuscript. We also appreciate their complimentary comments, including "The paper presents novel theoretical results that are highly relevant for the machine learning community" (Reviewer 1), "The paper provides the first efficient robust batch learning algorithm for several fundamental learning problems" (Reviewer 2), and "The paper seems to be of high-quality. The results are impressive, non-trivial and interesting" (Reviewer 4). The remainder of this response mostly addresses suggestions and questions raised by Reviewers 1 and 3.

Both Reviewers 1 and 3 ask us to elaborate on the differences between [JO19] and this paper. The differences fall in two categories: technique, and applications. In terms of technique, [JO19] does not leverage the distribution structure, it simply uses all domain subsets as filters. It therefore requires a sample size linear in the domain size, which is prohibitive for large domains and impossible for the all-important infinite and continuous domains.

By contrast, this paper utilizes the distribution's structure, or even rough proximity to a structure, to identify a much smaller class of filters that as we show, suffices to address adversarial batches. This significant improvement, that as Reviewer 1 writes, requires a "non-trivial" combination of VC theory and the filtering framework, allows us to remove the sample complexity's dependence on the domain size and to greatly extend the reach of the filtering algorithm. We then apply it to derive (1) robust estimation for whole range of distributions, including infinite and even continuous, that hitherto could not be learned robustly, (2) robustness results for vital learning tasks, most notably classification, that were not addressed in [JO19]. Note also that: (1) our information theoretic results on classification assume only that the classifier's hypothesis has a finite VC dimension – the most common assumption in learning theory, and (2) our efficient classification algorithm applies to the fundamental problem of 1-d interval classifiers.

Reviewers 1 and 3 also ask related questions about how the results will hold if the distributions are unstructured (Reviewer 3) or may vary by a small amount from each other (Reviewer 1). If the distribution is completely unstructured, then as pointed out in lines 49-53 of the paper, the sample complexity grows linearly with the domain size, hence one cannot learn the type of distributions addressed in this paper. If the distribution can be approximated by a structured distribution then it is covered by the "opt density estimation framework" utilized in the paper, see e.g., lines 194-196.

Regarding Reviewer 1's specific question whether the technique also applies when the distributions underlying genuine batches differ from a common target distribution by a small TV distance, say $\eta > 0$. For simplicity, we presented the analysis for $\eta = 0$, but as noted in [JO19] for unstructured distributions, the filtering technique easily adapts to $\eta > 0$. For example, in density estimation the trivial empirical estimator achieves $\mathcal{O}(\eta+\beta)$ TV-error, or $\mathcal{O}(\beta)$ when $\eta = 0$. Even for binary alphabets, the lower bound is $\Omega(\eta+\beta/\sqrt{n})$, hence no algorithm can reduce the effect of the disparity between the batches and target distributions. Filtering reduces the effect of adversarial batches from $\mathcal{O}(\beta)$ to $\tilde{\mathcal{O}}(\beta/\sqrt{n})$. Since we cannot do anything sophisticated about η, the proof and algorithm easily extend to $\eta > 0$. For this reason we presented the simplest problem that captures the essence of the technique. We will add a similar explanation to the final version.

Reviewer 1 suggests that we elaborate on the relationship between filtering methods for Gaussian mean estimation derived e.g., in [DKK+16], and [JO19]. This relation was explained in [JO19]. Section 3 of this paper, mentions the many important contributions of [DKK+16, DKK+17, SCV17], the recent survey [DKK+19], and others, but for brevity does not repeat the explanation in [JO19]. To enhance the reader's understanding of the context, in the final version of the paper we will follow the reviewer's advice and expand this discussion and elaborate on the specific relation to [JO19].

Reviewer 1 similarly suggests that we elaborate on previous use of VC theory in structured density estimation (including [ADLS17]). Please note that Section 3 of the paper starts by stating that "The current results extend several long lines of work on estimating structured distributions, including [O'B16, Dia16, AM18, ADLS17]" and that we provide specific references to [ADLS17] in three additional locations in the main paper and several more times in the appendix. Also note that the previous applications of VC theory were for non-robust learning, hence somewhat different from the current application that requires several new ideas. For the reader's benefit we will follow the reviewer's advice and elaborate on the use of VC theory in density estimation in non-robust setting.

Reviewer 1 also suggested that we move some of the applications from the main paper to the appendix and some proofs from the appendix to the main paper. We fully sympathize with the reviewer's desire to see more hard proofs in the paper itself, but felt that one of the paper's main contributions is showing broad audiences that adversarial batches can be addressed efficiently for a large class of practical problems. We also note that Reviewer 4's response to question 2 seems to appreciate this information. We will try to accommodate Reviewer 1's request by including as much information about the proofs as we can in the extra page of the papers' final version.

Finally, Reviewer 3 asks about the time complexity of the paper's two efficient algorithms: learning piecewise polynomials, and interval classification. Both algorithms have very reasonable complexities. Learning t-piecewise, degree-d polynomial distributions takes $\mathcal{O}(m \cdot n^2(1 + t \cdot d \cdot \beta/\sqrt{n}))$ time, and t-interval classification takes $\mathcal{O}(m \cdot n^2(1 + t \cdot \beta/\sqrt{n}))$ time. Since there is a total of $m \cdot n$ samples, these complexities are not too high. We will mention these time complexities explicitly in the paper.