We thank all the reviewers for their thoughtful feedback. We are happy to see that they agree with us on the importance of understanding online learning and differential privacy from the perspective of smoothed analysis. We will incorporate their feedback in the final version of the paper.

Below, we address the questions of **Reviewer $\#3$** regarding the definition of smoothness.

Regarding the smoothness assumption (1), while we assume that $1/\sigma$ is a constant in T, we can use values of σ that depend on the dimension of the space to account for the volume of high dimensional domains [1]. Since our regret bounds are only logarithmic in $1/\sigma$, they gracefully scale with the dimension of the space (which in the case of halfspaces, polynomial thresholds, etc, is bounded as a function of the VC dimension). In the context of the example mentioned in the review, note that $\sigma = c2^{-d}$ is only a weak smoothness assumption and still leads to a regret bound of $\tilde{O}\left(\sqrt{Td \cdot \mathrm{VCDim}\left(\mathcal{F}\right)}\right)$ against non-adaptive adversaries. Similar bounds hold against adaptive adversaries when the ϵ-bracketing number grows comparably to the size of a classical ϵ-cover, which as we argued in the paper is the case for many classes including halfspaces, polynomial thresholds, and convex polytopes.

Next we address (2) and compare our definition to Speilman and Teng's definition of smoothness.

Let us start by highlighting the similarities between our model of smoothness and Speilman and Teng's. Both of these models upper bound the maximum density within a domain. Our smoothness assumption does this directly by bounding the density by $1/\sigma$ times that of the uniform distribution. On the other hand, Speilman and Teng's model does this indirectly. To illustrate this, let $\mathcal{X}$ be the unit ball in $\mathbb{R}^d$ and take a Gaussian convolution with variance $O(1/d)$. For ease of presentation, clip this Gaussian convolution at a ball of radius 2. [2] It is not hard to see that the maximum density achieved anywhere in this ball is at most approximately $O(d)^{d/2}$, which is within an $\exp(d)$ factor of the uniform density on the ball of radius 2 (or even the unit ball) as required by our model of smoothness. Again, because of the logarithmic dependence on the value of $1/\sigma$, this leads to a regret bound of $\tilde{O}(\sqrt{Td \cdot \mathrm{VCDim}\left(\mathcal{F}\right)})$, against a non-adaptive smooth adversary. Note that, we have not optimized these parameters and indeed better translation between the two models of smoothness is possible when you choose the clipping region more carefully. This shows that the smoothness framework of Speilman and Teng is captured by the smoothness framework we use in this paper (within reasonable parameters).

We believe there are additional advantages in using our model of smoothness when formalizing learnability in a general sense. One advantage is that our smoothing model treats combinatorial domains (say $[n]$ or graphs on n vertices) and geometric domains (say $D \subseteq \mathbb{R}^d$) in a unified manner and allows us to deliver results most meaningful to the analysis of learnability in presence of some smoothness without getting bogged down with domain-specific definitions of smoothness. Another advantage of our model — which is specifically important in machine learning — is that it naturally allows the adversary to have arbitrary correlations between the labels and the instances as long as the marginal on the instances is a “smooth" distribution. This can be handled by other models, albeit with more awkwardness in separating how an instance is generated by random shifts but its label is generated exactly by the adversary.

We agree that these are important discussion points and we will incorporate a summary of them in the final version of the paper.

[1] Our results also extend to values of $1/\sigma$ that are subexponential in T

[2] Variance of $1/d$ is chosen so that the noise is of the same order as size of the domain, i.e., Gaussian distribution with variance $1/d$ is close to a uniform distribution over the unit ball. This clipping is done for ease of presentation and to ensure that no matter what the original distribution over $\mathcal{X}$ was, the Gaussian convolution has most of its density within the ball of radius 2. Qualitatively similar results can be achieved even without this clipping or with clipping to $\mathcal{X}$ which would be necessary if we restrict the adversary to only use instances in $\mathcal{X}$.