We are grateful to the reviewers for their insightful reviews and feedback. We have incorporated fixes to simple issues such as typos and missing references and do not address those issues here.

Reviewer 1 pointed out that some of the works on Massart noise get to error $OPT + \epsilon$, under Massart Noise. We thank the reviewer for this correction and will update the introduction as the reviewer suggested, as well as add citations to the earlier works in the Massart noise model.

We thank the Reviewer 2 for pointing out the issue with defining subderivative without convexity. We will address this and change it to a suitable generalized subdifferential.

Reviewer 2 is correct in pointing out that using the validation set to validate the pancake property is subtle. We were indeed imagining access to a clean validation set, and will clarify this in the revision.

Reviewer 3 points out that the margin separable assumption may be limiting. We will discuss this more in the revised version: the margin is needed only for the inliers, and one can treat all points violating the margin assumption as outliers from the point of view of the analysis. In particular, this implies that the results hold as long as the fraction of (true outliers + inliers violating the margin assumption) is suitably small.

Reviewer 3 asked for intuition on the HerSumNorm in Theorem 13. We will add more intuition in the full version on how it shows up in the proof. Our goal for Theorem 13 was to use the weakest condition under which we could give a simple proof. For specific noise and data models, we can bound the HerSumNorm using standard techniques, as we do in the Supplementary material.

Reviwer 3's question on Kernel methods is a very interesting one. While our Theorem 13 would extend to Kernel spaces, our current approach to proving that for a random sample D, the pair (D, μ) satisfies the dense pancakes condition (in the Supplementary material) requires $\Theta(d)$ samples. While we can use random projections to $\Theta(1/\gamma^2)$ dimensions and apply the algorithm and the Theorem there, extending these results to the usual SVM with Kernels is a very compelling open question.

Reviewer 3 asks if the Dense Pancakes condition is novel. To our knowledge, this condition has not explicitly appeared in any previous work, though it has likely shown up implicitly as a step in similar results under strong distributional assumptions. We thank the reviewer for their other suggestions for improvement and will revise the paper accordingly.

Reviewer 4 asks if assumption 4 is necessary or can be relaxed to accommodate smooth losses such as the logistic loss. This is a very interesting suggestion, to which we can give two responses. The first is that one can define a (huberized version of the) loss $\max(f(y\mathbf{w}^\top\mathbf{x}), f(1))$ for f being the logistic loss. Such a loss will satisfy the conditions and be close enough to the logistic loss for many purposes. The second response is that the proof can likely be extended to the actual logistic loss by adding an additional assumption on the HerSumNorm of the inliers. The place where we use assumption 4 is to derive the bound on the contribution from I_2 in (8). Suppose instead of condition (4) we had an upper bound, say $\frac{L}{(1+e)} < \frac{L}{3}$ on $|f'(x)|$ for $x \geq 1$. Then we could prove an upper bound of $\frac{L}{3} \cdot LinSumNorm(I_2) \leq \frac{L}{3} \cdot HerSumNorm(I)$ instead of 0 in equation (8). This would allow us to prove a variant of Thm 13 with an additional condition. This version will still imply a version of Thm 1 with slightly worse constants but now holding also for the logistic loss. We believe this additional complexity does not belong in the main theorem, but given the importance of the logistic loss, it would make sense to add such a statement in the Supplementary material, and we will do so in the revision. We thank the reviewer for this insightful question.