We thank the reviewers for their consideration of our paper and their positive feedback. We start with a summary of our contributions that we believe addresses several reviewer comments. The main contribution of our work is a strong separation between convex and non-convex surrogates for the problem of learning halfspaces with adversarial label noise. On the negative side, we prove that optimizing *any* convex surrogate objective leads to significantly suboptimal error guarantees. On the positive side, we show that *any* stationary point of a carefully chosen non-convex objective suffices to obtain nearly best possible error guarantees.

We emphasize that our lower bound result for convex surrogates is a significant contribution of our work, which we believe is of independent interest. In particular, it applies to most common methods used in practice for classification like logistic regression and hinge-loss minimization. We show that such **convex optimization approaches are provably suboptimal and simple non-convex methods outperform them**. Based on the reviews, we realize that this message was lost in our write-up and we will highlight it in the final version.

Our positive structural result (i.e., that any stationary point of our non-convex surrogate works) provides a novel understanding of the underlying learning problem from an optimization point of view. An immediate corollary is that vanilla projected SGD (or *any* other method that converges to a stationary point) efficiently solves our learning problem to near-optimal accuracy. As a result, we also obtain the following implications for free: (1) We give an algorithm with near-linear sample complexity in the dimension d (which is information-theoretically optimal) that runs in sample-linear time. Moreover, our algorithm works with a single pass over the data and has minimal memory requirements (we just need to store one sample at each step). (2) Our algorithm succeeds for a broad class of distributions, including ones for which no polynomial-time algorithm was previously known. Designing noise-tolerant algorithms under natural distributional assumptions was stated as an open problem in Section 5 of [ABL17].

Comparison to [DKTZ20]. While the presentation/structure of our technical section resembles [DKTZ20], the proof of our positive structural result is very different from that in [DKTZ20]. In terms of similarities, both works establish that any stationary point of a certain non-convex surrogate works for the corresponding learning problem. Moreover, both works use a parametric class of functions that are smooth surrogates of the 0-1 loss with a smoothing parameter σ.

Here we handle adversarial label noise which is a much more challenging noise model than the bounded/Massart model in [DKTZ20]. The worst-case strategy of the adversary (in the adversarial label model) differs from that in the Massart model. In particular, one needs to understand how the adversary can distribute the corrupted labels in order to bound their effect and as a result, our analysis departs significantly from that of [DKTZ20]. Our key technical contribution is that there exists a value of the smoothing parameter σ that works. In [DKTZ20], it was shown that any sufficiently small value of σ suffices, which does not hold in the agnostic model. Here we have to carefully choose the correct value of σ. Finally, we remark that to obtain our structural result (Lemma 3.2) we inherently require different distributional assumptions compared to [DKTZ20] (that include log-concave and certain heavy-tailed distributions).

Reviewer 1 Our work studies the standard ("passive") PAC learning model with adversarial label noise. A discussion/comparison of label complexity (aka "active learning") is orthogonal to the focus of our submission. In the standard PAC model with adversarial label noise, our algorithm is simpler, more sample-efficient and significantly faster than the localization-based method introduced in [ABL17]. [ABL17] establish polynomial upper bounds on the sample complexity and runtime of their algorithm, but the degree of the polynomial is not specified. As explained in the preceding discussion, our approach has some high-level similarities with [DKTZ20], but our main technical contribution (Lemma 3.2) requires a conceptually different proof and different set of assumptions. In particular, there appears to be no black-box way to reduce our setting to the Massart setting.

Reviewer 2 We agree that using this notation in lines 419-422 is confusing. We will fix this in the final version.

Reviewer 3 We note that in the adversarial label noise model, the parameter OPT (i.e., the fraction of corrupted labels), or a reasonable upper bound, is typically known a priori to the algorithm. Previous algorithms (and in particular [ABL17]) operate under this assumption. Under such an assumption, our algorithm does not need to "guess" OPT and amounts to a simple SGD. Moreover, even with this additional guessing step, the sample complexity and runtime of our algorithm remains near-optimal.

Reviewer 4 While we study linear classification, our learning problem is non-convex due to the noise in the labels. A broad range of prior works have used convex surrogates of the zero-one loss to address this non-convexity. In this work, we show that convex surrogates inherently fail with adversarial label noise. In contrast, we design a simple non-convex surrogate that we show leads to near-optimal accuracy. The intuition is that in our setting the non-convex landscape is well-behaved in the sense that *any* stationary point suffices. We show that (1) one should not use convex surrogates to learn linear classifiers in the presence of noise, and (2) one can replace convex losses with a simple non-convex loss and get much better classification error. Regarding (1): At a high level, convex loss functions will assign large weights to samples that are far from the origin. An adversary can take advantage of these points, flip their labels, and make the error of the convex minimizer large.