We would like to thank the reviewers for their careful consideration of our paper and their positive feedback. Below we address the comments and questions asked by the reviewers.

Reviewer #1: The existence of an efficient proper learner with the same accuracy guarantee is left as an open problem.

The equivalence between Sloan's "malicious misclassification noise" model and the Massart model is well-known in the literature, see, e.g., the introduction of [ABHU15].

By the definition of the Massart model, the *only* assumption on $\eta(x)$ is that $\eta(x) \leq \eta < 1/2$ for all x in the domain.

The Boolean-valued setting is an important special case that captures the difficulty of the problem of learning halfspaces. In particular, Sloan's open problem was explicitly phrased for Boolean disjunctions, a *very* special case of Boolean halfspaces. (This can also be found in Avrim Blum's FOCS 2003 tutorial cited and linked from our paper.)

We chose to emphasize the "distribution-independent" aspect in the title to clarify the distinction with the previous references that learn halfspaces with Massart noise *under the uniform distribution on the unit sphere*, e.g., [ABHU15].

We agree that emphasizing which arguments hold generally and which depend on the Massart noise assumption is useful for providing intuition and we will revise accordingly. The main place where the Massart noise assumption is being used is to show that a vector $\mathbf{w}$ with negative loss $L(\mathbf{w})$ exists (Lemmas 2.3 and A.1). The remaining of the arguments and in particular Lemma 2.5 hold more generally, assuming we can find a vector with negative loss.

Reviewer #3: The complexity of our algorithm is polynomial in the dimension d, the error ϵ, and the bit complexity b of the examples. The reviewer commented on the dependence on b.

In terms of *computational complexity*, a polynomial dependence on b exists in *all known learning algorithms for halfspaces* — even in the realizable (noiseless) case. In fact, it is well-known that removing such a dependence on b in the runtime (even for the realizable case) amounts to developing a *strongly polynomial* time algorithm for general linear programs — a major open problem in theoretical computer science. (This is stated in lines 66-67 of our paper.)

In terms of *sample complexity*, even for the special case of Random Classification Noise all known algorithms — including [BFKV97, Coh97] – have a polynomial dependence on b as well. The reason is that the outlier removal lemma of [BFKV97, DV04a], used as a preprocessing step to create a margin condition, requires this many samples.

Statement of Open Problem: As we explain in the introduction of our paper, several authors have posed related versions of the open problem we study. Sloan's original open problem [Slo88, Slo92] asks whether there is an efficient learning algorithm in the Massart noise model for Boolean disjunctions — i.e., OR functions on $\{0,1\}^d$ — *a very special case* of Boolean halfspaces. (Note that $b = d$ when the domain is the Boolean hypercube.) As pointed out in Avrim Blum's FOCS'03 tutorial [Blu03] (lines 48-54 of our paper), and additional personal communication with him, even the *weak learning* version of this problem remained open. Cohen [Coh97] asked whether there is an efficient learning algorithm for halfspaces with Massart noise. For the important setting of Boolean halfspaces, i.e., halfspaces on $\{0,1\}^d$ — already a broad generalization of Sloan's open problem — our algorithm has $\mathrm{poly}(d/\epsilon)$ sample complexity and runtime.

Estimation in Line 3 of Alg 1: Indeed, checking the termination condition requires estimating the probability which can be done via sampling. The number of samples required, i.e., $O(\frac{1}{\epsilon^2}\log(1/\gamma\epsilon))$, is much smaller than the number of samples needed in every iteration. We will make a note of that in the revised version of our paper.

Conversion of SGD guarantees to high probability: Given that our loss function L is bounded in $[-1,1]$, we can obtain high probability guarantees of Lemma 2.4 by running SGD multiple times. Here is the simple argument in more detail, which we will include in the revision: At a single run, Markov's inequality for the nonnegative random variable $L(\bar{\mathbf{w}}) - L(\mathbf{w}^*)$ gives:

$$\mathbf{Pr}[L(\bar{\mathbf{w}}) - L(\mathbf{w}^*) \leq 2(\mathbf{E}[L(\bar{\mathbf{w}})] - L(\mathbf{w}^*))] \geq 1/2.$$

From the guarantee that $\mathbf{E}[L(\bar{\mathbf{w}})] \leq L(\mathbf{w}^*) + \epsilon$, this means that with probability at least $\frac{1}{2}$, we can find a vector $\bar{\mathbf{w}}$ with loss at most $L(\mathbf{w}^*) + 2\epsilon$. Running SGD $O(\log(1/\delta))$ times, there exists such a vector with probability at least $1-\delta$. Identifying such a good vector requires estimating the loss within ϵ for all the returned vectors, which requires $\tilde{O}(\log(1/\delta)/\epsilon^2)$ samples in total for all vectors, as the loss is bounded in $[-1,1]$. Thus, the total sample complexity is at most $\tilde{O}(\log(1/\delta)/\epsilon^2)$ to get the guarantee with probability at least $1-\delta$.

Typo at the end of Lemma 2.5: Indeed there is a typo in the last displayed equation, which we will rephrase to make the proof cleaner. The integral from $\bar{T}$ to 1 is lower-bounded by $-\lambda\mathbf{Pr}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq \bar{T}]$ (second to last displayed equation) and is also upper-bounded by $L(\mathbf{w})/2$, as the integral from 0 to $\bar{T}$ is non-negative. This directly yields that $\mathbf{Pr}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq \bar{T}] \geq |L(\mathbf{w})|/2\lambda$, as required to complete the proof.

Comment 6: We only need that λ is sufficiently close to η to get a meaningful bound. For large values of λ, the statement still trivially holds (but is vacuous). *Comments 7 and 8:* We will rephrase for clarity. *Comment 9:* Theta will be removed.