We are grateful to all the reviewers, from whose critical reviews we have learned so much.

Prop. 3.1 explains that it is far beyond a rare exception that an AdaGrad scheme has an asymptotic direction independent of the initial conditions. It negates the claim "the implicit bias of AdaGrad does indeed depend on the initial conditions, including initialization and step size" on Page 8, Gunasekar, Suriya, et al. [2018a]. We think this is where the significance of Prop. 3.1 lies in, though it is derived under certain stringent assumption as mentioned by Reviewer ♯1.

However, in the statement of Prop. 3.1, as pointed out by Reviewer ♯4, the term "absolutely continuous distribution" was misused. It should be revised as "distribution whose density function is nonzero almost everywhere".

The arguments between Lemma 3.4 and Lemma 3.5 are neither a rigorous proof nor a sketch. They are devoted to introducing certain objects and notations to be used in the rest of the paper, as well as some intuitions.

$\boldsymbol{\beta}(t) - \mathbf{1} \to \mathbf{0}$ is obvious, for $\boldsymbol{h}(t) \to \boldsymbol{h}_\infty$ and $\boldsymbol{\beta}(t) = \boldsymbol{h}_\infty^{-1} \odot \boldsymbol{h}(t) \to \mathbf{1}$ (see the formula above Line 93).

Some intuitions are provided in Line 94-98 and Line 106-124. We were very lucky to find that the *induced form* is equivalent to the primary Adagrad scheme in studying the existence of the asymptotic direction. If one replaces $\boldsymbol{\beta}(t)$ in the induced form with $\mathbf{1}$, the limit of $\boldsymbol{\beta}(t)$, then a GD scheme is obtained. Therefore we expected that the induced form and the GD scheme have similar asymptotic behaviors. Thus the orthogonal decomposition in Line 116 emerges, and we found that when the increment of the induced form is decomposed in this way, the projection in the asymptotic direction of the GD scheme eventually becomes overwhelming. The accumulated effects determine the whole trend of the induced form.

We are sorry to admit that there were a handful of minor errors in our paper. As noticed by Reviewer ♯4, the usage of the symbol $y_n \in \{-1, 1\}$ might cause some confusion. In fact, from Line 75 to Line 140, we rewrote the loss $\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^N l\left(y_n \boldsymbol{w}^T \boldsymbol{x}_n\right)$ as $\sum_{n=1}^N l\left(\boldsymbol{w}^T \boldsymbol{x}_n\right)$ by redefining $y_n \boldsymbol{x}_n$ as $\boldsymbol{x}_n$ to simplify notation. In Line 145 and Example 3.1, we resumed the usage of y_n to be in conformance with the setting of "separable data". Furthermore, as pointed out by Reviewer ♯3, the formula in Line 43 on Page 4 of Appendix is inaccurate. It should be written as

$$\boldsymbol{\delta}(t)^T \widehat{\boldsymbol{u}} = -\eta \nabla \mathcal{L}_{ind}\left(\boldsymbol{v}(t)\right)^T \widehat{\boldsymbol{u}} = -\eta \sum_{n=1}^N l'\left(\boldsymbol{v}^T \boldsymbol{\xi}_n\right) \boldsymbol{\xi}_n^T \widehat{\boldsymbol{u}} \geq -\eta \sum_{n=1}^N l'\left(\boldsymbol{v}^T \boldsymbol{\xi}_n\right) > 0.$$

These errors mainly arose from the switching between different versions of the draft. Fortunately, they did not shake the foundation of our reasoning at all.

It may be helpful to point out that our theory can be set up without the leverage of the results of Soudry et al. [2018]. In fact, our geometric estimation approach introduced in this paper can be used to give a much shorter proof for the existence of the asymptotic direction of GD iterates, though without any convergence rate derived. Since our approach does not rely on convergence rates, it may be applied to more cases where the learning rate η is not a constant.

We did a few of numerical simulations to verify and illustrate our theory. We should have organized the paper better to include some of the examples. Here is one of them.

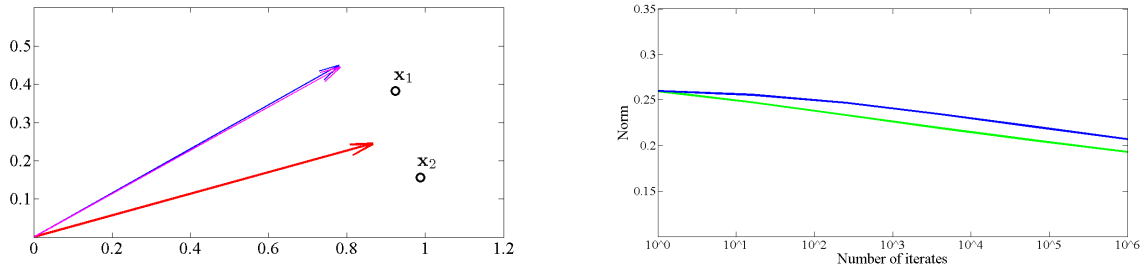

Figure 1: **Left.** $\mathbf{x}_1 = \left(\cos\frac{\pi}{8}, \sin\frac{\pi}{8}\right)$ and $\mathbf{x}_2 = \left(\cos\frac{\pi}{20}, \sin\frac{\pi}{20}\right)$ are two support vectors. The red arrow denotes the direction of the max-margin separator. The blue and magenta arrows denote the asymptotic directions of AdaGrad iterates with $\eta = 0.1$ and 0.5, respectively. The small angle between the two illustrates that the asymptotic direction may depend on η. However, all the asymptotic directions apparently diverge from the max-margin separator.
Right. The blue and green curves plot $\left\|\mathbf{w}(t)/\|\mathbf{w}(t)\| - \mathbf{d}_A\right\|$ vs. the number of the iterates with $\eta = 0.1$ and 0.5, respectively, where $\mathbf{d}_A$ is the asymptotic direction of AdaGrad iterates. It can be observed that the directions of AdaGrad iterates slowly converge to $\mathbf{d}_A$. This aligns with Theorem 3.2.