We thank the reviewers for their reviews and comments. We will incorporate your suggestions in the final version.

We analyze an extremely simple model (as explicitly discussed and motivated, e.g. on lines 59-61 and the broader impact section), but have reason to believe insights, behaviours and analysis methods will carry over to more complex and realistic models. The paper Woodworth et al. 2020 [30] already shows how phenomena in this simple model predicted by theory can be observed also for complex realistic networks. Insights about the extreme training accuracy needed to reach the limiting behaviour are already helping us and others understand apparent discrepancies regarding matrix factorization models. In addition, as we comment in lines 36-37, the effect of width is equivalent to that of initialization scale (more precisely: increasing width corresponds to effectively scaling initialization as sqrt(width) in matrix factorization / wide diagonal nets, see [30]). Therefore, our analysis is also relevant to more-overparameterized models, with width$\to \infty$, and the number of weights entering whenever we discuss init scale.

Normalization of the loss function (Reviewers #1 and #2): The loss is defined to be the average over all examples. Thus the interval of ϵ should be (0,1], this is a typo that does not cause difficulty with analysis and we will fix it.

For reviewer #1: - There is a typo in the inequality in line 146, we will fix it. Also, in eq (2),(3),(4) it should be $\forall n$, we will add it. All other minor typos will be fixed, thank you.
- Note that the algebraic form of the model appears in line 147.
- On the role of the number of observations: We analyze the linearly separable case. Overparameterization is a generic way of achieving separability, but once the data is separable, the number of observations does not play a direct role.
- Comparison to assumptions/setting of Chizat and Bach 2020 [6] / Lyu and Li 2020 [19]: These papers analyze more general models, as we discuss in Section 2. Our analytic results are only for the linear diagonal nets - which is a special case of the models in those papers. For this homogeneous model all the assumptions of previous works hold.

For reviewer #2: - The discrepancy between gradient flow and discrete gradient descent is indeed very interesting. We actually think that for large step-sizes one would exit the kernel regime much faster. Characterizing this is an important challenge, but seems difficult (both we, and others, have tried, and are still trying to study this). In this paper we focus on the training accuracy (ϵ), but we certainly are interested also in the effects of other hyperparameters, including, and perhaps most importantly, the step-size.
- Note that Q_μ is defined for general μ, and $\mu = \alpha^D$ is only for the square loss setting studied in Woodworth et al. [30] and takes different values in our Theorem 6. We will make it more clear.
- In line 285 it should be $\epsilon = \exp(-10000)$. We will fix that, thank you.

For reviewer #3: - Comparison with Ji and Telgarsky 2019 [16] and Lyu and Li 2020 [19]: References [16] and [19] analyze the asymptotic max-margin solution only for *infinite* training accuracy (ie infinite training time or zero training loss) and fixed initialization, i.e. the "rich regime", as discussed in the introduction, and in lines 125-140 in Section 2 (including Theorem 2, and footnote 3 where we discuss how [16] fits in). Section 2 is dedicated to discussing how this is only one extreme endpoint (the other, corresponding to the "kernel regime", occurs when the training accuracy is finite and the initialization grows to infinity). This is also visualized in Figure 1 and in the table above the figure. As discussed explicitly in lines 55-58, the submission fills the gaps between the two extremes and studies the transition between them. Moreover, [16]'s results are for fully connected linear networks where the rich regime and kernel regime both lead to L_2 max-margin, while for the diagonal network, the rich regime limit (L_1 or $L_{2/D}$) is very different from the kernel regime (L_2).
- Comparison with Woodworth et al. 2020 [30]: As discussed in the intro in lines 30-33 and in lines 119-125, [30] analyzes the transition for *regression with the squared loss*, and does not study training accuracy as a controlling hyper-parameter. The effect of the training accuracy is more important for classification problems since unlike regression, with exponential loss, no finite parameter can ever achieve zero training loss. Thus, as discussed in lines 40-45, the classification setting we study is substantially different, and in order to understand it we need to study the effect of training accuracy, which was not studied in [30].

For reviewer #4: - In eq (4) we take α to infinity only since we want to emphasize the importance of the order of limits when both α and time go to infinity (in contrast with eq. 2).
- Indeed, our theoretical results are asymptotic and obtained when *both* initialization and accuracy go to infinity at some relative rates. The case of finite initialization and infinite accuracy is already known in Lyu and Li 2020 [19], Nacson et al. 2019 [20], and so is the case of infinite initialization and finite accuracy Chizat et al. 2019 [7]. We fill in the gap by letting both go to infinity while maintaining some finite relationship between them. Non-asymptotic results for finite initialization and finite accuracy seem very difficult to obtain, but we see in figure 5(b) that the relationship predicted by our asymptotic results is already seen also for moderate finite α.
- The technical difficulty compared to [Woodworth et al. 2020] stems from the difference in KKT conditions. [Chizat et al 2019] derived only the kernel regime (large initialization and finite time).