We thank the reviewers for their useful comments, which help us improve the clarity and quality of our paper. We address these comments point by point. The paper will also be revised accordingly, with additional details given there.

***Reviewer 1** suggested us to provide a short "proof sketch".* We will present a brief summary of our proof strategy. It generally follows the classical recipe for establishing the weak convergence of interactive particle systems. Our proof sketch will also provide a roadmap to the SM, which contains all the technical details of the proof.

***Reviewer 1** asked why we choose $\log\cosh(x)$ as the regularization function $H(x)$ in example 1.* In fact, any convex function with its minimum reached at zero would be fine, for example $H(x) = |x|$ or x^2. The function $H(x) = \log\cos(x)$ is just a convenient special case since its derivative $H'(x) = \tanh(x)$ is smooth and bounded.

***Reviewer 1** asked what "a sufficiently strong regularizer" means in example 1.* It means to use a sufficiently large λ. In our experiments, $\lambda > 1$ is sufficient. In such cases, the regularized variables, e.g. $\boldsymbol{w}^\top\boldsymbol{w} - 1$ and each diagonal term of $\boldsymbol{V}^\top\boldsymbol{V} - \boldsymbol{I}$ in (4) are restricted to around 0 during the training process.

***Reviewer 1**'s additional comments:*
- We will emphasize that $P_{\tilde{\mathbf{c}}}$ is fixed in our model. Meanwhile, a learnable $P_{\tilde{\mathbf{c}}}$ is indeed an interesting idea to explore.
- By "analyzing the long-time behavior', we meant the investigation of the local stability of the ODE in Section 4. We will clarify this point.
- The matrix $\boldsymbol{M}(t)$ is $(2d+1)\times(2d+1)$, where d is the number of features in the true data model (1).
- Thank you for catching the various typos and other miscellaneous issues. We will fix these and post the code.

***Reviewer 2** suggested avoiding using a single matrix M for the macroscopic states.* We agree with the reviewer that $\boldsymbol{M}$ is simply a concatenation of different macroscopic states $\boldsymbol{P}$, $\boldsymbol{q}$, $\boldsymbol{S}$, $\boldsymbol{r}$ and z, all of which have concrete meanings. However, we believe that there is indeed some benefit in introducing $\boldsymbol{M}$, which serves to streamline later presentations (e.g. those in Sec. 3.1). In the revised paper, we will start our discussions by first defining $\boldsymbol{P}$, $\boldsymbol{q}$, $\boldsymbol{S}$, $\boldsymbol{r}$ and z as the actual macroscopic states we study. We will then introduce $\boldsymbol{M}$ and emphasize that it is just a convenient and compact notation.

***Reviewer 2** remarked that the notions of the macroscopic and microscopic states are standard in physics.* It is indeed a standard assumption in statistical physics that the macroscopic states of large systems tend to converge to deterministic values due to self-averaging. However, we note that the mean-field regime in our work was not considered in previous theoretical analysis of GAN. For example, a series of recent work [11-16] considers a different scaling regime where the learning rate goes to zero but the system dimension n stays fixed. In that regime, the microscopic dynamics are deterministic even with the presence of the microscopic noise. In contrast, we study the regime where the learning rate is fixed but the dimension $n \to \infty$. This setting allows us to quantify the effect of training noise in the learning dynamics. Thus, we believe our work provides new and valuable insights into the theoretical understanding of GAN.

***Reviewer 2** suggested adding more discussions on the learning dynamics.* In the revised paper, we will use the one extra page available to us to provide more discussions on the hierarchical interactions of the macroscopic variables. Specifically, we will consider the special setting where the ratio of the generator's and the discriminator's learning rates $\tilde{\tau}/\tau \to 0$. This allows us to further simplify the ODE in (8) and to show that the learning process will be a combination of two dynamics with two different time scales. The discriminator's process is associated with the faster time scale, whereas the generator's process relates to the slower one.

***Reviewer 3** suggested us to extend the discussion of the phase diagram to $d \geq 2$ cases.* After the initial submission we have done some additional work on understanding the phase diagram for cases when $d \geq 2$. Accordingly, we will expand the current discussions on line 45–50 in the SM. Specifically, we will present local stability analysis about the phases of Info-1 and Noninfo-1 for $d = 2$. The analytical characterizations of the other phases are much more challenging when $d \geq 2$. Instead of showing analytical results, we will characterize these phases numerically by directly computing the eigenvalues of the Jacobian of the limiting ODEs.

***Reviewer 3** asked how one can orthonormalize the columns of a general $\boldsymbol{U}$.* Suppose $\boldsymbol{U}$ in the data model (1) is not orthogonal. We can always rewrite the product $\boldsymbol{U}\boldsymbol{c}$ as $(\boldsymbol{U}\boldsymbol{R})(\boldsymbol{R}^{-1}\boldsymbol{c})$, where $\boldsymbol{c}$ is the original random variable in the feature space, and $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ is a matrix that orthogonalizes and normalizes the columns of $\boldsymbol{U}$. We can then study an equivalent system where the new feature vector is $\boldsymbol{R}^{-1}\boldsymbol{c}$.

***Reviewer 3** asked how one can generalize our analysis to handle non-linear generators.* One possible extension is to add a non-linear function, *e.g.* ReLU, to the linear data model (1) and the generator (2). Our analysis technique can be extended to handle this case, where one can still obtain a differential equation in the scaling limit. Another possible extension is a mixture of Gaussian model, where we need to treat $P_{\tilde{\mathbf{c}}}$ as a learnable distribution rather than a fixed one. (This was also mentioned by Reviewer 1.) We will comment on these possible extensions in the revised paper.

***Reviewer 3** pointed out several typos and presentation issues.* Thank you. We will correct them in the revised version. By "heuristic derivation", we meant that certain steps presented in that section were not fully rigorous as we directly discard higher-order terms without any justification, in order to highlight the main ideas. In Section S-IV in the SM, we rigorously justify these steps by providing bounds on those terms.