We thank all reviewers for their comments and acknowledgement of our contribution. All comments are *very* useful and will be addressed in greater details in the revised version. Below we address each reviewer's comments separately.

Response to Reviewer 1:

GDRS for GANs. We thank the reviewer for recognizing our potential contribution to solving GANs, and will follow up with more in-depth studies on the topic. Here, we reiterate that micro-macro modeling is also a very general problem setting, as suggested by multiple examples in the introduction and numerical experiment sections.

Symmetry in the micro tasks. While the individual SIR model is not directly powered by micro-features such as county population, we did provide asymmetry by aligning the model's output for each county with the first day of a reported infection within that county (normalized by county population), which serves as an implicit feature. We apologize and will eliminate such confusion in the revised version.

Computation advantage of GDRS for large and small N. The reviewer raised a very good point. Indeed, applying full gradient descent to solve the empirical approximation of CSO,

$$\min_{\theta\in\Theta_h} \tfrac{1}{M}\sum_{i=1}^{M} \ell\left(\tfrac{1}{N}\sum_{j=1}^{N} h_\theta(\xi_i, x_{ij}), \bar{y}_i\right), \tag{1}$$

has the *same per-iteration computation* complexity as applying GDRS to solve the empirical approximation of the minimax reformulation:

$$\min_{\theta\in\Theta_h} \max_{\lambda\in\Lambda_u} \Phi(\theta,\lambda) := \tfrac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} \left\{h_\theta(\xi_i, x_{ij}) u_\lambda(\xi_i, \bar{y}_i) - \ell^*(u_\lambda(\xi_i, \bar{y}_i))\right\}. \tag{2}$$

However, using same amount of MN samples and without assuming strong convexity conditions, the generalization bound of (2) is of order $\mathcal{O}(1/\sqrt{MN})$, while the generalization bound of (1) is of order $\mathcal{O}(1/\sqrt{N} + 1/M)$ (Hu et al.(2019)). This is an important reason why we decided to pursue the empirical approximation of the minimax reformulation rather than the original CSO in the first place. We will add this clarification in the revised version.

Grid search for SIR model. Using fine-grained brute-force search will likely achieve an optimal solution; however, the computational cost will be very high given the expensive cost of calling a differential equation solver for evaluating each solution candidate. Our gradient-based method is much more efficient but only finds a stationary point. We are afraid that such a comparison may not be fair.

Response to Reviewer 2:

Convergence. The reviewer is absolutely correct that the overall complexity depends on the dimension. Our claim that "if $m = \log_{1-q(\epsilon)} \epsilon$, then GDRS converges to a neighborhood of the stationary point at rate $\mathcal{O}(\log T/\sqrt{T})$, which matches the best-known-rate of projected gradient method for nonconvex minimization" is trying to emphasize that when m is sufficiently large, i.e., *with high per-iteration cost*, the *iteration complexity* of GDRS becomes the same as projected gradient method. We will modify the sentence to avoid any confusion.

Novelty/Relation to [18]. GDRS was proposed originally in [18] in 1983, but as reviewer pointed out, it is rarely known to the machine learning community, which motivated us to speak it out for many potential applications, not limited to macro-learning. Although we adopted the exact original form of GDRS, in this paper we extended the asymptotic analysis of [18] to nonconvex-nonconcave objectives, and more importantly, we provided the first non-asymptotic convergence analysis with explicit dependence on m and other factors.

Response to Reviewer 3:

Novelty/Relation to [18]. Please see the second point in response to Reviewer 2.

Tuning hyperparameters. We agree with the reviewer and will report tuning in the revised version. Here, ν only appears in theoretical analysis and does not need tuning. In practice, we kept m small to get fast run time.

Response to Reviewer 4:

What does ".06" error rate correspond to? ".06" is the *average error* between the observed infection numbers and the estimated infection numbers over a period of 128 days on the testing counties.

Response to Reviewer 5:

Elusive assumption and theorem statement. We agree with the reviewer that fractional derivatives deserve more detailed treatment in the main text, and we will provide examples illustrating the spirit of the theorem.

Impact of the distribution Q. This is a very good suggestion. We will provide analysis for specific distributions, accompanied by numerical illustrations.