Reply to Reviewer 2, 3 and 4

Novelty of the analysis: Besides keeping the estimator of $\nabla f(\mathbf{x}_k, \mathbf{y}_k)$ sufficiently accurate like SARAH/SPIDER for convex/nonconvex minimization, the minimax problem also requires $\mathbf{y}_k$ to be close to $\mathbf{y}^*(\mathbf{x}_k)$, which leads to a more challenging analysis. Our analysis is based on a recursive relationship between $\Delta_k = \mathbb{E}[\|\mathbf{v}_k - \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)\|_2^2 + \|\mathbf{u}_k - \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)\|_2^2]$ and $\delta_k = \mathbb{E}\,\|\mathcal{G}_{\lambda,\mathbf{y}}(\mathbf{x}_k, \mathbf{y}_k)\|_2^2$ (defined in line 364 of Appendix B) as in Lemma 13 (Corollary 2, line 418, Appendix B), which guarantees both Δ_k and δ_k to be smaller than $\mathcal{O}(\kappa^{-2}\varepsilon^2)$ (line 434-435, Appendix B). This is a nontrivial ingredient of the analysis. To achieve the desired convergence rate, all of the stepsizes, mini-batch sizes, number of the inner iterations, and the orders of Δ_k and δ_k must be balanced carefully. In comparison, SARAH/SPIDER for convex/nonconvex minimization only needs to consider the estimator of gradient, which does not involve the extra complexity in minimax problems.

Reply to Reviewer 2 and 4

Algorithm is complicated: As mentioned in the setting of experiments (line 589, Appendix F), we can select $q = m = \lceil n/S_2 \rceil$ heuristically and the empirical result show it performs well in practice. We agree that it is worth to see weather there exists a simpler variant of SREDA which also holds the theoretical guarantee.

Reply to Reviewer 3 and 5

Optimal dependency on ε: The lower bound means any stochastic first-order algorithm requires at least $\mathcal{O}(\varepsilon^{-3})$ calls of stochastic first-order oracle $(\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}; \xi), \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}; \xi))$ to find ε-stationary point of $\Phi(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ (line 93, Definition 1). As we mentioned in line 180-185, we can consider the special case of minimax problem whose objective function has the form $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + h(\mathbf{y})$ where g is possibly nonconvex and h is strongly-concave, which leads to minimizing on $\mathbf{x}$ and maximizing on $\mathbf{y}$ are independently. Consequently, finding $\mathcal{O}(\varepsilon)$-stationary point of the corresponding $\Phi(\mathbf{x})$ can be reduced to finding $\mathcal{O}(\varepsilon)$-stationary point of nonconvex function $g(\mathbf{x})$, which is based on the stochastic first order-oracle $\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}; \xi) = \nabla g(\mathbf{x}; \xi)$ (this equality holds for any $\mathbf{y}$ and we also have $\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \xi) = \nabla g(\mathbf{x}; \xi)$). Hence, the analysis of stochastic nonconvex miminization problem [5] based on $\nabla g(\mathbf{x}; \xi)$ can directly lead to the $\mathcal{O}(\varepsilon^{-3})$ lower bound for our minimax problem.

Reply to Reviewer 2

Relations to compositional optimization: We thank the reviewer for pointing out this valuable reference (arXiv:1908.11468). We are happy to cite this paper and compare it with SREDA. Both this work and SREDA use variance reduction to address nonconvex multi-level (two-level) optimization problem, however, their settings are quite different. We can reformulate our minimax problem (2) as compositional problem:

$$\min_{\mathbf{x},\mathbf{y}} f_2(f_1(\mathbf{x}, \mathbf{y})) + \Psi(\mathbf{y}), \tag{a}$$

where $f_1(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}))$, $f_2(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y})$ and $\Psi(\mathbf{y})$ is the indicator function of $\mathcal{Y}$. The two-level nested-SPIDER (arXiv:1908.11468) requires to access the stochastic gradient of f_1 and f_2 to solve the compositional problem. For the two-level formulation (a) of our minimax problem, it is not natural to access the (stochastic) gradient of f_1 as an oracle since we can not provide the closed form of $\arg\max_{\mathbf{y}\in\mathbb{R}} f(\mathbf{x}, \mathbf{y})$ and its (stochastic) gradient in general. A more reasonable way is to solve it by accessing the (stochastic) gradient of f like SREDA.

Reply to Reviewer 3

1. Problem formulation in experiments: Due to the space limitation, we gave the problem formulation in Appendix F. We are happy to follow the reviewer's suggestion and include it in the main text if the paper is accepted.

2. Minor remarks: Thanks for the suggestion. We will simplify the presentation of Theorem 2 and cite the papers about non-stochastic algorithms for minimax problems.

Reply to Reviewer 4

1 & 2. Please see "Reply to Reviewer 2, 3 and 4" and "Reply to Reviewer 2 and 4" above.

3. We thank the reviewer for pointing out this valuable reference, which is complementary to our work. We will cite it and compare to SREDA. First, the convergence result of epoch-GDA is based on the measure of nearly ε-stationary point because it also considers an additional constraint on $\mathbf{x}$, while SREDA is based on ε-stationary point. Second, the stochastic first-order oracle complexity of SREDA depends on $\mathcal{O}(\kappa^3\varepsilon^{-3})$, while epoch-GDA depends on $\tilde{\mathcal{O}}(\kappa^2\varepsilon^{-4})$.

Reply to Reviewer 5

Presentation of Algorithm 4: We thank the reviewer for this suggestion, and will add the arguments in the expression of the function ConcaveMaximizer.