We thank the reviewers for their interest in the contributions of the paper and their detailed comments. We share their enthusiasm regarding our theoretical contributions: we find fascinating how stochasticity can hurt convergence in differentiable games and how variance reduction fixes it. We will revise our paper to reflect points raised in their reviews (including the introduction, as noted by R3).

**R1 & R3: Extension of Theorem 1 that satisfies Assumption 1.** We obtain a result similar to Theorem 1 that satisfies Assumption 1 by adding an $\ell_2$ penalty to Eq. 1, thus considering the following optimization problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \max_{\boldsymbol{\varphi} \in \mathbb{R}^d} \frac{\epsilon}{2}\|\boldsymbol{\theta}\|^2 - \frac{\epsilon}{2}\|\boldsymbol{\varphi}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}^\top \boldsymbol{A}_i \boldsymbol{\varphi} \tag{1}$$

We can follow the same proof technique as in §C.1 and get a similar result as L88 with additional $\eta\epsilon$ terms:

$$\mathbb{E}[N_{t+1}] = \left(1 - \tfrac{|I|}{n}(2\eta\epsilon - \eta^2(1+\epsilon^2)) + \tfrac{|I|^2}{n^2}(2\eta^2(\eta\epsilon - 1) + \eta^4)\right)\mathbb{E}[N_t] \underset{|I| \ll n}{\approx} \left(1 - \eta\tfrac{|I|}{n}(2\epsilon - \eta(1+\epsilon^2))\right)\mathbb{E}[N_t]\,.$$

Thus, for any step-size (roughly) larger than $2\epsilon$, the stochastic extragradient method diverges geometrically. However, the full batch method [Harker and Pang] and SVRE (Thm. 2) do converge for any step-size smaller than 1 (particularly for any step-size in $[2\epsilon, 1]$). This provides an example that satisfies Assumption 1 where the stochasticity breaks the properties of extragradient (a step-size around $\epsilon$ would lead to a much slower convergence rate than for SVRE).

**R1: Guarantee for the non-convex case.** Recently, [Lin, Jin, and Jordan, 2019] provide guarantees in the min-max setting when one of the two functions is non-convex and the other one is convex. Proving global convergence rate when both $\mathcal{L}_G$ and $\mathcal{L}_D$ are non-convex in the full batch setting remains an open question that highly interests the optimization community, but is outside the scope of this paper. As noticed by the reviewers, our goal was rather to study the theoretical impact of stochasticity in convex games (and empirically for GANs).

**R3: Differences between this work and Palaniappan & Bach, and novelty.** We point out some of the differences in lines 131–135, lines 164–174 and Table 1 of the paper. We agree that our algorithm may seem conceptually analogous to the one of Palaniappan & Bach (which combines gradient method with variance reduction) as SVRE combines extragradient with variance reduction. However, pointing out that stochasticity could be an important consideration for solving the training instabilities of GANs is novel and there was no algorithm for extragradient that does variance reduction. Also, it was not known whether extragradient would benefit (theoretically and practically) from variance reduction. Our analysis largely differs from the one of P & B since the original analysis of extragradient is completely different from the one of the gradient method. Precisely, the key point that allows for proving that the method has a convergence rate of the order of $\mu/L$ (which is significantly better than the one in P & B) holds in Eq. 53 and 54. Table 3 also compares SVRE with the existing standard methods. Regarding the practical contribution, we are excited that SVRE resolves partially the known GAN training instabilities as, to our knowledge, SVRE is the first constant step size method that works for non trivial datasets. Related works plug in *Adam* to make the algorithm work, which unfortunately does not work consistently across hyperparameters for GANs and diverges at some point (see Fig. 8).

**R3: Wall-clock time for BatchE.** Fig. 1 below shows the wall clock time on **MNIST**, for a fixed GPU. The trend is similar to Fig. 3a, where we used the number of mini-batches as a more portable comparison point (and standard in optimization) enabling better reproducibility (since it is both hardware and implementation-independent) of the results.

**R4: More analysis and intuition in the experimental evaluation part.** Lines 252–257 & 294–301 discuss the main points about the results from Tab. 2; while App. G.3 provides a more detailed discussion and additional experiments (moved to App. due to space constraints). In short, SVRE might do worse than EG-A because the latter has the benefit from adaptive step-sizes with Adam; developing a convergent adaptive step-size version of SVRE is an open problem.

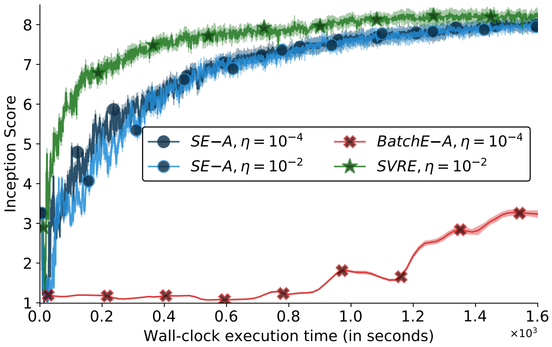

Figure 1: Wall–clock time on **MNIST**, using **Tesla V100-SXM2-16GB** GPUs (see Appendix F.2.1 for experimental setup). We used $time.perf\_counter()$ & $torch.cuda.synchronize()$ to syncronize the cuda execution–following the recommendation for PyTorch, see the following link: *https://discuss.pytorch.org/t/best-way-to-measure-timing/39496*.