R1: Thanks for your positive evaluation! There are four parameters in both Algorithm 1 and Algorithm 2: L, γ, δ and σ. L is the Lipschitz constant of the operator, which is tuned in implementation or alternatively handled by adopting an extra parameter-free strategy. γ, δ are parameters of the strongly convex norm square $\frac{1}{2}\|\cdot\|^2$, which can be easily verified in practice. In line 133-134, we have shown the values of γ and δ for the p-norm $\frac{1}{2}\|\cdot\|_p^2$. σ is the constant in Assumption 3 ($\sigma = 0$) or Assumption 4 ($\sigma > 0$). In practice, the nonzero σ is often obtained by an explicit ℓ_2-norm regularization, so it can also be verified effectively. In camera-ready, we will include a paragraph to show the choice of these parameters. Regarding the lack of simulations, we will add numerical experiments to compare our algorithm with existing ones in camera-ready. Finally, the abbrv ODE is somewhat misleading indeed. We will use OptDE instead.

R2: Thanks for your positive feedback! Thanks also for pointing out the extra strength that we were not paying attention to. By Dang&Lan 2015 [8], it seems that such a strength also exists for extragradient-type methods. We believe it is a natural by-product from proving approximate strong solution guarantees, while both results in [Thm 2, 1] and [Thm.1, 2] are in terms of approximate weak solution guarantees. On the simulations front, per your suggestion, we will do experiments to validate the behavior of our algorithms, particularly for the last-iterate convergence.

Thanks for your insightful observation in terms of the definition of restricted strong merit function! Our definition is not an exact analog of Nesterov [30] indeed. However, as shown in page 329 of [30], the restricted merit function will only be informative when D satisfies $D \geq \|\boldsymbol{w}^* - \bar{\boldsymbol{w}}\|$, where $\boldsymbol{w}^*$ is the solution of a monotone problem. This is because, by Lemma 1 of [30], only under the condition $D \geq \|\boldsymbol{w}^* - \bar{\boldsymbol{w}}\|$ do we get the following: the solution that makes the restricted merit function 0 is the solution of the underlying monotone problem. Consequently, since only a large D is informative and the value of D only appears in the theoretical guarantee, we do not need to worry about what will happen in the case of small D: we can just pick a large enough D to make $\boldsymbol{w}^*$ contained in the set. In camera-ready, we will explicitly discuss this point.

Additionally, you totally got the main point in Section 4! For minimization problems, we can reduce the effect with a small step size (i.e., learning rate). However, in the nonmonotone setting (Assumption 3), possibly due to the lack of a Lyapunov function and the inability of performing averaging *simultaneously*, "one cannot obtain provable convergence rates by only decreasing the step size, whereas a large batch size is necessary"(line 265-268). We believe such a fact partly validates why we must use a large batch size in the training of GAN. Finally, we will change the title to "Optimistic Dual Extrapolation for a Class of Nonmonotone Variational Inequalities" to avoid the possibility of overclaim. Thanks for catching all the typos: consider all of them fixed.

R3: Thanks for your positive evaluation! In this paper, we are mainly concerned with computational complexity in finding an ϵ-accurate solution. If we hope to guarantee the convergence in the strict sense of last-iterate, the best possible rate will be $O(1/\epsilon^2)$ for extragradient (EG) even in the monotone setting [13]. Meanwhile, optimistic methods can be viewed as approximations of EG and the nonmonotone setting includes the monotone one as an instance. Thus we can not expect a better rate than $O(1/\epsilon^2)$ rate in the strict sense of last-iterate for optimistic methods in the nonmonotone setting. To avoid the $O(1/\epsilon^2)$ barrier, we relax the concept of last iterate convergence as follows: we only guarantee the convergence rate when the iterate $k \geq O(1/\epsilon)$. For the beginning $k \leq O(1/\epsilon)$ iterations, the last iterate may not necessarily converge. Additionally, going beyond "asymptotic convergence" (which only characterizes qualitative convergence when the number of iteration tends to ∞), we provide explicit finite-time convergence rates. Furthermore, in optimization, it is standard to treat the accuracy parameter ϵ as an input of an algorithm. The regularization trick in this paper depends on the specification of ϵ beforehand. Of course, developing algorithms that are agnostic to knowing ϵ is interesting (and significantly but beyond the scope of this paper) and we leave it for future research.

Thank you for your detailed writing suggestions! Following them, we will discuss more about the derivation of the VI problem, the properties of different gap functions and a comparison of the assumptions appearing in the literature. Meanwhile, we will move the discussion of natural residual earlier to make the result about Iusem et al. 2017 in Table 2 more clear. Thanks also for pointing out the relevant ICLR reference! In addition to the difference you mentioned, we also study the setting where *a strongly weak solution exists* while they did not. The results of this setting are significant as they allow us to obtain near-optimal approximate strong solution guarantees for the monotone setting. As mentioned in Remark 2, we consider the dual extrapolation approach because we can give a unified convergence analysis under Assumptions 3 and 4 using estimation sequence. Meanwhile, if there exists a regularizer, the lazy update can exploit the structure of regularizer better. For simplicity, we did not consider a composite term in the algorithm. However, it can be handled in the same way as the constrained set, if a certain efficient proximal operator exists for the composite term.

R4: Thanks for your positive feedback! We will reorganize the writing according to you and **R3**. We must use strongly convex norms, where the strong convexity is used to cancel certain errors in the convergence analysis; additionally, it also makes the solution of subproblems unique. Following your suggestion, we will define the distance generating kernels by h and make the gradient in terms of the function. We will correct all the typos you pointed out. The term "optimistic" is coined by Rakhlin and Sridharan [35] in the online learning context, which then has been used in a confusing way in the existing literature indeed. In our context, we do not "conservatively" compute a new gradient but instead reuse the computed past gradients for the extrapolation step, which is thus an "optimistic" procedure.