We thank the reviewers for taking the time to carefully read the paper and their constructive comments. We see a common concern on the lack of discussion on the limitations/possible extensions of our methods, which we discuss below and will give more details in the paper:

(1) Proximal/constrained setting: As pointed out by Reviewer#1, currently, the proposed methods do not work in this setting as they rely on the "co-coercivity" property (we will correct the terminology). We may consider dropping the negative gradient norms in their Lyapunov functions, which leads to less tight formulations and slower rates (as suggested by Reviewer#4) but also makes them less reliant on that property. We think this might be feasible.

(2) Prior knowledge of strong convexity constant μ: This methodology requires a known μ since even if it is applied to a non-accelerated method, the parameter choice is always related to μ.

(3) Non-strongly convex case: After submission, we discovered that when $\mu = 0$ in the framework of G-TM, it is possible to adopt a variable parameter setting (the benefit of allowing variable choices) that leads to the $O(1/K^2)$ rate. The special part is that, since the Lyapunov function becomes $T_k = a_k \cdot \big(f(y_k) - f(x^\star) - \frac{1}{2L}\|\nabla f(y_k)\|^2\big) + \frac{L}{4}\|z_k - x^\star\|^2$, one step GD at y_{K-1} is critical to obtain a convergence guarantee. We suspect that this scheme is equivalent to the optimized gradient method (Kim and Fessler, 2016), which could answer some of the questions raised in that work.

(4) BS-SGD is a promising direction, which requires considerable efforts (Moulines and Bach, 2011).

To Reviewer#1 Thank you for your detailed comments. Please also see the revision plan to Reviewer#2.

[Comparison with TMM] We admit that the claimed "redundant parameters" problem of TMM is a bit artificial and will make the following two changes to improve the comparison: (i) we will add a table comparing the parameters of NAG, TMM and G-TM (optimal tuning), and provide the guarantee of TMM (Eq.(11) in [7]) in Section 3.1; (ii) we will change "redundant parameters" to "a general scheme" that allows variable parameter choices and is easier to extend (see extension (3) above). About the flawed guarantee, thanks for pointing out the intermediate inequality. Actually, in our derivation of the original TMM (we started from Eq.(11) in [7]), it is possible to obtain an initial constant that is roughly equal to that of NAG (i.e., C_0^{NAG}) when κ is large. This can also be seen in Figure 2a where TMM hits the upper bound of NAG in the first iteration. We didn't discuss this because we thought G-TM already resolves this issue.

[About the proofs] For a Lyapunov function $T_k = a_k h(y_k) + b_k\|z_k - x^\star\|^2$, our strategy is to first build contractions for $h(y_k)$ (Lemma 3) and $\|z_k - x^\star\|^2$ (Lemma 1), and then sum them up and eliminate residual terms. Comparing with the proof in [7], the key difference is that we point out Lemma 1, which allows shifted stochastic gradient and reads as a classic inequality whose usage has been well studied. We may instead say our proofs are more extensible.

[Co-coercivity] Thanks for the detailed clarification. We will revise related sentences following existing literature.

[Appendix F] Thank you for suggesting the existing works. We will mention that Appendix F is just for completeness.

[Typos/minor comments] We will fix the typos and try to improve the unclear parts.

To Reviewer#2 Thank you for pointing out the presentation issues. Here is our revision plan.

We will rewrite Section 2 as a formal introduction to the shifting theory (or high level ideas). Since our objective is to minimize h ⇔ choosing a family of Lyapunov function that only involves h, a critical issue is that we cannot even compute its gradient $\nabla h(x) = \nabla f(x) - \mu(x - x^\star)$. We figured out that in some simple cases, a change of "perspective" is enough to access this gradient information. Take GD: $x_{k+1} = x_k - \eta\nabla f(x_k)$ as an example (which we will include as a motivating example). We can rewrite the update as $x_{k+1} - x^\star = (1 - \eta\mu)(x_k - x^\star) - \eta\nabla h(x_k)$, and thus

$$\|x_{k+1} - x^\star\|^2 = (1-\eta\mu)^2\|x_k - x^\star\|^2 \underbrace{-2\eta(1-\eta\mu)\langle\nabla h(x_k), x_k - x^\star\rangle + \eta^2\|\nabla h(x_k)\|^2}_{\leq 0 \text{ if } \eta = 2/(L+\mu)\text{, which is based on the co-coercivity of } \nabla h.},$$

which is just the one-line proof of GD in the textbook (Theorem 2.1.15, [27]) but looks more structured in our opinion. However, this change of "perspective" is too abstract for more complicated schemes. We thus encode this idea into Lemmas 1 and 2 with some template updating rules, which serve as instantiations of the shifted gradient oracle. Then, we can directly choose various gradient estimators for h (GD, SVRG, SAGA, ...), and by applying Lemma 1, we obtain a practical updating rule together with a classic inequality whose usage has been well studied. We will also rewrite Lemma 1 to clarify this usage. Now we have enough building blocks to migrate existing schemes to the shifted objective (since we can query its gradient oracle through Lemma 1). G-TM is basically NAG migrated to the shifted objective. Technically speaking, the most important techniques in NAG (in our opinion) are Lemma 3 for f and the standard mirror descent lemma. G-TM was derived by having a shifted version of Lemma 3 for h and the shifted mirror descent lemma. We can also regard G-TM as a more "aggressive" parameter setting of NAG to get some insight, just like GD with $\frac{2}{L+\mu}$ and $\frac{1}{L}$ learning rates. BS-SVRG/SAGA were derived by "replacing" the shifted gradient in G-TM as shifted SVRG/SAGA estimator (through Lemma 1). We will add descriptions in each section to make their derivations clearer.

To Reviewer#3 Thank you for your appreciation of our work. We will try to make the experiments more comprehensive.

To Reviewer#4 Thank you for your comments and suggestions. We will improve our paper following your suggestions, and also hope that the revision suggested by Reviewer#2 may make our ideas clearer.

References: (Kim and Fessler, 2016) Optimized first-order methods for smooth convex minimization. In *Math. Program.*.

(Moulines and Bach, 2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NeurIPS*.