SHANG++: Robust Stochastic Acceleration under Multiplicative Noise

Anonymous authors
Paper under double-blind review

Abstract

Training with multiplicative noise scaling (MNS) is often destabilized by momentum methods such as Nesterov's acceleration, as gradient noise can overwhelm the signal. A new method, SHANG++, is introduced to achieve fast convergence while remaining robust under MNS. With only one-shot hyperparameter tuning, SHANG++ consistently reaches accuracy within 1% of the noise-free setting across convex problems and deep networks. In experiments, it outperforms existing accelerated methods in both robustness and efficiency, demonstrating strong performance with minimal parameter sensitivity.

1 Introduction

Empirical Risk Minimization (ERM) is central to modern large-scale machine learning, including deep neural networks and reinforcement learning (Hastie et al., 2009). It is formulated as

$$\min_{x\in\mathbb{R}^d} f(x,X,Y), \quad f(x,X,Y)=\frac{1}{N}\sum_{i=1}^{N} f(x,X_i,Y_i)=\frac{1}{N}\sum_{i=1}^{N} f_i(x), \tag{1.1}$$

where $\{(X_i,Y_i)\}_{i=1}^N$ is a large dataset ($N\gg 1$), and $f_i(x)$ is the loss for the i-th sample. Efficiently computing the minimizer $x^*=\arg\min_x f(x)$ is critical for training large models.

Exact gradient evaluation is expensive, so Stochastic Gradient Descent (SGD) uses mini-batches:

$$g(x)=\frac{1}{M}\sum_{i\in B}\nabla f_i(x), \tag{1.2}$$

where $B\subset\{1,\dots,N\}$ is a random batch of size M. SGD slows down when the condition number κ of f is large. Momentum methods such as Heavy Ball (HB) (Polyak, 1964) and Nesterov accelerated gradient (NAG) (Nesterov, 1983) are widely used to accelerate convergence. In training deep neural networks, Adam (Adaptive Moment Estimation) (Kingma & Ba, 2015) is a widely used optimization algorithm that combines momentum and adaptive step sizes for fast and stable convergence.

The mini-batch estimator $g(x)$ reduces the cost of computing $\nabla f(x)$ but introduces noise. In regimes such as small-batch training or highly over-parameterized models, the variance can scale with and even dominate the signal $\|\nabla f(x)\|^2$. This effect is modeled by the multiplicative-noise scaling (MNS) condition (Wu et al., 2019; 2022; Gupta et al., 2024).

Definition 1.1 (Multiplicative Noise Scaling (MNS)). The stochastic gradient estimator $g(x)$ satisfies the MNS condition if there exists $\sigma\geq 0$ such that

$$\mathbb{E}\left[\|g(x)-\nabla f(x)\|^2\right]\leq\sigma^2\|\nabla f(x)\|^2. \tag{1.3}$$

Momentum methods are highly sensitive to stochastic noise (Devolder et al., 2014; Aujol & Dossal, 2015; Liu et al., 2018), and stability depends critically on parameter choices (Kidambi et al., 2018; Liu & Belkin, 2020; Assran & Rabbat, 2020; Ganesh et al., 2023). Gupta et al. (2024) showed that under MNS with $\sigma\geq 1$, NAG fails to converge in both strongly convex and convex settings.

To address this, several corrections have been developed. Vaswani et al. (2019) introduced a four-parameter NAG variant and proved convergence rate $\left(1-(1+\sigma^2)^{-1}\sqrt{\mu/L}\right)^k$ in the strongly convex case, where L/μ is the condition number of f, and $\mathcal{O}(1/k^2)$ in the convex case. Liu & Belkin (2020) proposed the Mass method with three parameters and a correction term, though acceleration was shown only for over-parameterized linear models. Gupta et al. (2024) later proposed AGNES, a three-parameter extension of NAG with the same guarantees as Vaswani et al. (2019). More recently, Hermant et al. (2025) introduced SNAG, a four-parameter variant that attains the same rates with a mild parameter adjustment.

From the viewpoint of provable convergence in convex settings, these algorithms are competitive. Yet our deep-learning experiments show that they often lose acceleration under high noise and can perform worse than SGD even with recommended hyperparameters (see Section 3). For instance, on CIFAR-100 with ResNet-50 and batch size 50, SGD reaches 58.326% test accuracy, while AGNES achieves only 42.82%. With a further reduction in batch size, both AGNES and SNAG oscillate heavily with large performance swings, requiring extra hyperparameter tuning.

Motivated by this gap, our goal is not only to design another accelerated method, but to develop a complementary approach that (i) retains optimal theoretical guarantees, (ii) reduces tuning effort, and (iii) improves stability. Our contributions emphasize simplicity (fewer parameters), provable acceleration with explicit noise dependence, and robust empirical behavior.

1. Section 2 presents SHANG++, a stochastic extension of HNAG (Chen & Luo, 2021) for robust convergence under multiplicative noise, sharpening existing guarantees with minimal hyperparameter complexity. SHANG++ achieves accelerated rates of $\mathcal{O}(1/k^2)$ in convex settings and the fastest known rate $\left(1+\frac{2}{1+\sigma^2}\sqrt{\mu/(L-\mu)}\right)^{-k}$ for quadratic strongly convex problems with multiplicative noise.
2. Section 3 validates SHANG++ on convex optimization, image classification, and generative modeling (on benchmark datasets MNIST, CIFAR-10, CIFAR-100). SHANG++ matches or improves upon NAG, SNAG, AGNES, and Adam, with clear advantages under high multiplicative noise.
3. Section 3 tests robustness to multiplicative noise. At realistic noise levels ($\sigma \leq 0.5$), SHANG++ maintains near noise-free accuracy (within 1% degradation), supporting our theory. These results show that stability can be achieved with fewer parameters and a simpler design, improving earlier corrections such as AGNES and SNAG.

Notation. Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. The Bregman divergence of f between $x, y \in \mathbb{R}^d$ is

$$D_f(y, x) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

The function f is μ-strongly convex if for some $\mu > 0$,

$$D_f(y, x) \geq \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

It is L-smooth, for some $L > 0$, if its gradient is L-Lipschitz:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \forall x, y \in \mathbb{R}^d.$$

Let $\mathcal{S}_{\mu,L}$ be the class of all differentiable functions that are both μ-strongly convex and L-smooth. For $f \in \mathcal{S}_{\mu,L}$, the Bregman divergence satisfies

$$\frac{\mu}{2}\|x - y\|^2 \leq D_f(x, y) \leq \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d, \tag{1.4}$$

Parameters μ and L are treated as known hyperparameters for the given problem. Their adaptivity is beyond the scope of this work.

Limitation. Current convergence guarantees hold only for convex objectives under multiplicative noise scaling and do not extend directly to general non-convex landscapes.

Although SHANG++ reduces tuning complexity through one-shot, non-adaptive hyperparameters, its performance may still depend on accurate estimates of smoothness constants (e.g., L, μ). In highly non-convex settings or under very high noise, the one-shot strategy may require refinement.

2 Stochastic Hessian-driven Accelerated Nesterov Gradient

Our method is inspired by the second-order dynamical system introduced in Chen & Luo (2021), known as the Hessian-driven Nesterov Accelerated Gradient (HNAG) flow:

$$\gamma x'' + (\gamma + \mu)x' + \beta\gamma\nabla^2 f(x)x' + (1 + \mu\beta + \gamma\beta')\nabla f(x) = 0, \tag{2.1}$$

where $\beta > 0$ is any continuously differentiable function on $[0, \infty)$ and γ is a time-scaling factor. This second-order ODE can be equivalently reformulated as the first-order system:

$$x' = v - x - \beta\nabla f(x), \qquad v' = \frac{\mu}{\gamma}(x - v) - \frac{1}{\gamma}\nabla f(x), \qquad \gamma' = \mu - \gamma, \tag{2.2}$$

which removes the explicit dependence on $\nabla^2 f(x)$.

Methods. Discretizing (2.2) via a Gauss–Seidel–type scheme, adding an extra term $-m(x_{k+1} - x_k)$ to the x-update, and replacing $\nabla f(x_k)$ with an unbiased estimator $g(x_k)$ yield the Stochastic Hessian-driven Nesterov Accelerated Gradient (SHANG++) method:

$$\begin{cases} \dfrac{x_{k+1} - x_k}{\alpha_k} = v_k - x_{k+1} - m(x_{k+1} - x_k) - \beta_k g(x_k), \\ \dfrac{v_{k+1} - v_k}{\alpha_k} = \dfrac{\mu}{\gamma_k}(x_{k+1} - v_{k+1}) - \dfrac{1}{\gamma_k} g(x_{k+1}), \\ \dfrac{\gamma_{k+1} - \gamma_k}{\alpha_k} = \mu - \gamma_{k+1}, \end{cases} \tag{2.3}$$

where $\alpha_k > 0$ is the step size, $m \geq 0$ controls the extra noise-damping term, and $\beta_k > 0$ depends on α_k and γ_k, typically scaling as $\frac{\alpha_k}{\gamma_k/(1+\sigma^2)}$.

If the damping term is absorbed into the left-hand side, the x-update becomes

$$\frac{x_{k+1} - x_k}{\tilde{\alpha}_k} = v_k - x_{k+1} - \beta_k g(x_k), \tag{2.4}$$

where $\tilde{\alpha}_k = \frac{\alpha_k}{1+m\alpha_k} \leq \alpha_k$.

SHANG++ can thus be interpreted as a modified discretization of the HNAG flow with a reduced step size $\tilde{\alpha}_k$. The case $m = 0$ recovers SHANG, a direct stochastic extension of HNAG. The "++" indicates two improvements: faster theoretical convergence and greater robustness to noise.

With the parameter choices specified in Theorem 2.1 for the strongly convex case $f \in \mathcal{S}_{\mu,L}$, and in Theorem 2.2 for $\mu = 0$, faster convergence guarantees can be established.

SHANG++ for Strongly Convex Minimization. Let $f \in \mathcal{S}_{\mu,L}$ with $0 < \mu < L < \infty$. Define the auxiliary function

$$f_{-\mu}(x) = f(x) - \tfrac{\mu}{2}\|x - x^*\|^2.$$

Clearly, $\nabla f_{-\mu}(x^*) = 0$. Since $f \in \mathcal{S}_{\mu,L}$, it follows that $f_{-\mu} \in \mathcal{S}_{0,L-\mu}$. Let $g_{-\mu}(x_k) := g(x_k) - \mu(x_k - x^*)$ denote a stochastic estimate of $\nabla f_{-\mu}(x_k)$. As no randomness is introduced in the shift, the MNS condition

$$\mathbb{E}\big[\|g_{-\mu}(x_k) - \nabla f_{-\mu}(x_k)\|^2\big] \leq \sigma^2\|\nabla f_{-\mu}(x_k)\|^2$$

still holds provided (1.3) holds.

Setting $\gamma = \mu$ and $m = \beta\mu$, and substituting $g_{-\mu}(x_k)$ and $x_k^+ := x_k - \alpha\beta g_{-\mu}(x_k)$ into (2.3) yields

$$\begin{aligned} \frac{x_{k+1} - x_k^+}{\alpha} &= v_k - x_{k+1} - \beta\mu(x_{k+1} - x^*), \\ \frac{v_{k+1} - v_k}{\alpha} &= x^* - v_{k+1} - \tfrac{1}{\mu} g_{-\mu}(x_{k+1}). \end{aligned} \tag{2.5}$$

Schemes (2.5) and (2.3) generate the same sequences $(x_k, v_k)_0^\infty$; the explicit appearance of x^* is only for analysis and does not affect the algorithm itself.

Theorem 2.1. Let $f \in \mathcal{S}_{\mu,L}$. Given $x_0^+ = v_0 = x_0$, suppose (x_k, v_k) are generated by (2.5) with $g(x_k)$ defined in (1.2) and MNS (1.3) holds. If the step size satisfies $0 < \alpha \le \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L-\mu}}$ and $\beta = \frac{\alpha}{\mu/(1+\sigma^2)}$, then

$$\mathbb{E}\Big[f_{-\mu}(x_k^+) - f_{-\mu}(x^*) + \frac{\mu}{2}\|v_k - x^*\|^2\Big] \le (1 + \alpha + (1+\sigma^2)\alpha^2)^{-k}\big(f(x_0) - f(x^*)\big).$$

If f is quadratic, a sharper rate holds:

$$\mathbb{E}\Big[f_{-\mu}(x_k^+) - f_{-\mu}(x^*) + \frac{\mu}{2}\|v_k - x^*\|^2\Big] \le (1 + 2\alpha + \alpha^2)^{-k}\big(f(x_0) - f(x^*)\big).$$

Proof. We give an outline of the proof and refer to the Appendix C.1 for the full details.

Let $z_k^+ = (x_k^+, v_k)$ and define the Lyapunov function

$$\mathcal{E}(z_k^+) = f_{-\mu}(x_k^+) - f_{-\mu}(x^*) + \frac{\mu}{2}\|v_k - x^*\|^2. \tag{2.6}$$

Given (x_k, v_k) and $g(x_k)$, the quantities x_k^+ and x_{k+1} are deterministic, while randomness is introduced through $g(x_{k+1})$ and consequently affects (x_{k+1}^+, v_{k+1}). The expectation $\mathbb{E}$ is with respect to the randomness in $g(x_{k+1})$.

First of all, we have the sufficient decay of SGD for $x_{k+1}^+ := x_{k+1} - \alpha\beta g_{-\mu}(x_{k+1})$: if $\alpha\beta = \frac{\alpha^2}{\mu/(1+\sigma^2)} \le \frac{1}{(1+\sigma^2)(L-\mu)}$, which is equvialent to $\alpha \le \frac{1}{(1+\sigma^2)}\sqrt{\mu/(L-\mu)}$, then

$$\mathbb{E}\left[f_{-\mu}(x_{k+1}^+) - f_{-\mu}(x_{k+1})\right] \le -\frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2 = -\frac{(1+\sigma^2)\alpha^2}{2\mu}\|\nabla f_{-\mu}(x_{k+1})\|^2. \tag{2.7}$$

Then by the definition of Bregmann divergence:

$$\mathcal{E}(z_{k+1}) - \mathcal{E}(z_k^+) = \langle \nabla\mathcal{E}(z_{k+1}), z_{k+1} - z_k^+\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1}). \tag{2.8}$$

Expanding the first term and using the update in (2.5) gives

$$\begin{aligned} &-(1+\beta\mu)\alpha\langle \nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x^*), x_{k+1} - x^*\rangle - \alpha\mu\|v_{k+1} - x^*\|^2 \\ &+ \alpha\langle g_{-\mu}(x_{k+1}), v_k - v_{k+1}\rangle + \langle \nabla f_{-\mu}(x_{k+1}) - g_{-\mu}(x_{k+1}), v_k - x^*\rangle. \end{aligned} \tag{2.9}$$

The first two terms can be bounded by $-(1+\beta\mu)\alpha\mathcal{E}(z_{k+1})$ by using $\langle \nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x^*), x_{k+1} - x^*\rangle = D_{f_{-\mu}}(x_{k+1}, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})$. After taking the expectation $\mathbb{E}(\langle \nabla f_{-\mu}(x_{k+1}) - g_{-\mu}(x_{k+1}), v_k - x^*\rangle) = 0$. The most difficult term is the expectation of the cross term $\mathbb{E}\left[\langle g_{-\mu}(x_{k+1}), v_k - v_{k+1}\rangle\right]$, as both $g_{-\mu}(x_{k+1})$ and v_{k+1} are random variables. We use the identity $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ to obtain

$$\begin{aligned} \alpha\langle g_{-\mu}(x_{k+1}), v_k - v_{k+1}\rangle &= \langle \frac{\alpha}{\sqrt{\mu}} g_{-\mu}(x_{k+1}), \sqrt{\mu}(v_k - v_{k+1})\rangle \\ &= \frac{\alpha^2}{2\mu}\|g_{-\mu}(x_{k+1})\|^2 + \frac{\mu}{2}\|v_k - v_{k+1}\|^2 - \frac{\alpha^2\mu}{2}\|v_{k+1} - x^*\|^2. \end{aligned}$$

where the term involving $v_{k+1} - x^*$ follows from $\frac{\alpha}{\sqrt{\mu}} g_{-\mu}(x_{k+1}) - \sqrt{\mu}(v_k - v_{k+1}) = \alpha\sqrt{\mu}\big(\frac{1}{\mu} g_{-\mu}(x_{k+1}) - \frac{v_k - v_{k+1}}{\alpha}\big) = \alpha\sqrt{\mu}\big(x^* - v_{k+1}\big)$ by the update of v_{k+1}. Taking expectations termwise and applying the MNS condition to the first term yields the positive gradient contribution $\frac{\alpha^2(1+\sigma^2)}{2\mu}\|\nabla f_{-\mu}(x_{k+1})\|^2$, which can be canceled by the negative term in (2.7). The positive $\frac{\mu}{2}\|v_k - v_{k+1}\|^2$ is canceled by $-\frac{\mu}{2}\|v_k - v_{k+1}\|^2$ contained in $-D_{\mathcal{E}}(z_k^+, z_{k+1})$.

Using $\beta\mu = (1+\sigma^2)\alpha$, we obtain

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+)\right] - \mathcal{E}(z_k^+) \le \mathbb{E}\left[-(1 + (1+\sigma^2)\alpha)\alpha\mathcal{E}(z_{k+1}^+)\right].$$

Moving $\mathcal{E}(z_{k+1}^+)$ to the left-hand side yields the desired result.

When f is quadratic, the Bregman divergence is symmetric, $D_f(x_{k+1}, x^*) = D_f(x^*, x_{k+1})$, and the extra negative terms $-\beta\mu\alpha D_{f_{-\mu}}(x^*, x_{k+1}) - \frac{\alpha^2\mu}{2}\|v_{k+1} - x^*\|^2 \le -\alpha^2\mathcal{E}(z_{k+1})$, which sharpens the constant to $1 + 2\alpha + \alpha^2$. □

When $\sigma = 0$, SHANG++ reduces to the deterministic HNAG++ method of Chen & Xu (2025). As σ grows, convergence slows but acceleration is preserved. While Gupta et al. (2024) interpret noise as inflating smoothness to $(1+\sigma^2)^2 L$, our analysis shows it perturbs both smoothness and curvature, giving $L_\sigma = (1+\sigma^2)L$ and $\mu_\sigma = \mu/(1+\sigma^2)$. The noise-damping term in SHANG++ further reduces L_σ to $(1+\sigma^2)(L-\mu)$, explaining its stronger stability.

Quadratic Loss Consider a special case of problem (1.1): the quadratic loss with Tikhonov regularization (also known as weight decay), which is widely used in regression tasks. The objective takes the form

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}(x^\top X_i - Y_i)^2 + \frac{\lambda}{2}\|x\|_2^2 = \frac{1}{N}\|X^\top x - Y\|_2^2 + \frac{\lambda}{2}\|x\|_2^2, \tag{2.10}$$

where $\frac{1}{N}\sum_{i=1}^{N}(x^\top X_i - Y_i)^2$ is the empirical quadratic loss and $\frac{\lambda}{2}\|x\|_2^2$ is the regularizer with $\lambda > 0$. The Tikhonov regularizer ensures that the objective is λ–strongly convex with smoothness constant $(L+\lambda)$. Under multiplicative noise scaling, setting $\alpha = \sqrt{\mu_\sigma/L_\sigma} = \frac{1}{1+\sigma^2}\sqrt{\lambda/L}$ yields the accelerated convergence rate $1 - 2\alpha = 1 - 2\sqrt{\mu_\sigma/L_\sigma}$ in the leading term.

Batching. Gradient noise can be reduced by increasing the mini-batch size M in (1.2). If σ_1^2 is the MNS constant for $M = 1$, then $\sigma_M^2 = \sigma_1^2/M$. Another approach is to average K independent gradient estimators, $g^K = \frac{1}{K}\sum_{i=1}^{K} g_i$, which gives an effective MNS constant of σ^2/K. Both strategies reduce noise at the cost of higher computation, and a straightforward analysis shows that averaging multiple estimates can accelerate convergence to some extent.

Variance decay under MNS. Beyond the expectation bound, we show geometric variance decay of the Lyapunov energy. Specifically, by Theorem D.1,

$$\mathrm{Var}\left(f_{-\mu}(x_k^+) - f_{-\mu}(x^*) + \frac{\mu}{2}\|v_k - x^*\|^2\right) \le (f(x_0) - f(x^*))^2(r^2 + K_2)^k.$$

A sufficient (practically verifiable) condition is $K_2 < 1 - r^2$, where $r = (1+\alpha+\alpha^2)^{-1}$ is the decay rate in Theorem 2.1 and K_2 collects the fluctuation constants. This holds, for example, in low-condition regime, with a damped stepsize $\alpha \leftarrow \delta\alpha$ $(0 < \delta \le 1)$ or with a minibatch of larger M (or K independent multiple estimates). Complete proofs and the explicit expressions of related constants are provided in Appendix D.

SHANG++ Method for Convex Minimization Recall the modified step size $\tilde{\alpha}_k = \frac{\alpha_k}{1+m\alpha_k}$. To facilitate analysis, we define an auxiliary time-scaling variable $\tilde{\gamma}_k = \frac{\gamma_k}{1+m\alpha_k}$. Setting $\alpha_k = \frac{2}{k+1}$ and $\gamma_k/(1+\sigma^2) = \alpha_k\tilde{\alpha}_k L_\sigma$, for any fixed $m \ge 0$, we obtain:

$$\frac{\tilde{\gamma}_{k+1} - \tilde{\gamma}_k}{\tilde{\alpha}_k} = -(1 + \frac{1}{2(k+1+2m)})\tilde{\gamma}_{k+1} \le -\tilde{\gamma}_{k+1} \tag{2.11}$$

Replacing the x-update in (2.3) with the equivalent modified discretization (2.4) and combining it with (2.11) yields the following convergence result.

Theorem 2.2. Let $f \in \mathcal{S}_{0,L}$. Suppose that (x_k, v_k) are generated by the time-stepping scheme (2.3). $g(x_k)$ defined in (1.2) and MNS holds. Given $x_0^+ = v_0 = x_0, m \ge 0$, choose the step size $\alpha_k = \frac{2}{k+1}$, $\gamma_k/(1+\sigma^2) = \alpha_k\tilde{\alpha}_k L_\sigma$ and $\beta_k = \frac{\alpha_k}{\gamma_k/(1+\sigma^2)}$, we have

$$\mathbb{E}\left[f(x_{k+1}^+) - f(x^*) + \frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1} - x^*\|^2\right] \le \frac{(1+2m)(2+2m)}{(k+2+2m)(k+3+2m)}\mathcal{E}(z_0;\tilde{\gamma}_0) = \mathcal{O}(\frac{L_\sigma}{k^2})$$

Proof. We provide a proof sketch; the full proof appears in Appendix C.2. Define $x_k^+ = x_k - \tilde{\alpha}_k\beta_k g(x_k)$ and Lyapunov function

$$\mathcal{E}(z_k^+;\tilde{\gamma}_k) = f(x_k^+) - f(x^*) + \frac{\tilde{\gamma}_k}{2}\|v_k - x^*\|^2 \tag{2.12}$$

where $\tilde{\gamma}_k = \frac{\gamma_k}{1+m\alpha_k}$. Using $\gamma_k/(1+\sigma^2) = \alpha_k\tilde{\alpha}_k L_\sigma \Leftrightarrow \tilde{\gamma}_k/(1+\sigma^2) = \tilde{\alpha}_k^2 L_\sigma$ and the L-smoothness of f to obtain the upper bound of $\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] - \mathcal{E}(z_k^+;\tilde{\gamma}_k)$.

$$\mathbb{E}\left[\mathcal{E}(z_{k+1};\tilde{\gamma}_k) - \mathcal{E}(z_k^+;\tilde{\gamma}_k) - \frac{\tilde{\alpha}_k^2(1+\sigma^2)}{2\tilde{\gamma}_k}\|\nabla f(x_{k+1})\|^2 + \frac{\tilde{\gamma}_{k+1}-\tilde{\gamma}_k}{2}\|v_{k+1}-x^*\|^2\right] \quad (2.13)$$

Using (2.11), the last term less than $-\frac{\tilde{\alpha}_k\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2$. Then expaning the difference $\mathbb{E}\left[\mathcal{E}(z_{k+1};\tilde{\gamma}_k) - \mathcal{E}(z_k^+;\tilde{\gamma}_k)\right]$ and using the updates and $\alpha_k\tilde{\gamma}_k/\gamma_k = \tilde{\alpha}_k$ yield

$$\mathbb{E}\left[-\tilde{\alpha}_k\langle\nabla f(x_{k+1}) - \nabla f(x^*), x_{k+1}-x^*\rangle + \tilde{\alpha}_k\langle g(x_{k+1}), v_k - v_{k+1}\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\tilde{\gamma}_k)\right] \quad (2.14)$$

For the cross term $\mathbb{E}\left[\tilde{\alpha}_k\langle g(x_{k+1}), v_k - v_{k+1}\rangle\right]$, by Cauchy-Schwarz and Young's inequality,

$$\mathbb{E}\left[\tilde{\alpha}_k\langle g(x_{k+1}), v_k - v_{k+1}\rangle\right] \le \mathbb{E}\left[\frac{\tilde{\alpha}_k^2(1+\sigma^2)}{2\tilde{\gamma}_k}\|\nabla f(x_{k+1})\|^2 + \frac{\tilde{\gamma}_k}{2}\|v_k - v_{k+1}\|^2\right] \quad (2.15)$$

which are canceled respectively by the negative gradient term $-\frac{\tilde{\alpha}_k^2(1+\sigma^2)}{2\tilde{\gamma}_k}\|\nabla f(x_{k+1})\|^2$ and by $-D_{\mathcal{E}}(z_k^+, z_{k+1};\tilde{\gamma}_k)$. Putting everything together to obtain

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] - \mathcal{E}(z_k^+;\tilde{\gamma}_k) \le -\tilde{\alpha}_k\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] \quad (2.16)$$

Rearranging and substituting $\tilde{\alpha}_k = \frac{\alpha_k}{1+m\alpha_k} = \frac{2}{k+1+2m}$ yield the claimed rate. □

We compare the parameters

$$\text{(SHANG)} \quad \frac{\gamma_k}{1+\sigma^2} = \alpha_k^2 L_\sigma, \qquad \text{(SHANG++)} \quad \frac{\gamma_k}{1+\sigma^2} = \alpha_k\tilde{\alpha}_k L_\sigma = \alpha_k^2 \cdot \frac{L_\sigma}{1+m\alpha_k},$$

which reduces the effective Lipschitz constant from L_σ to $\frac{L_\sigma}{1+m\alpha_k}$. The noise-damping term offsets part of the σ^2–induced amplification, improving stability by slowing down the effective rate. Our experiments suggest that choosing m in the range $[0, 1.5]$ provides a good trade-off.

3 Numerical Experiments

We design our experiments to validate the theoretical alignment, scalability, and robustness of SHANG++ and SHANG ($m = 0$).

Throughout this section, NAG refers to the stochastic version of Nesterov's accelerated gradient (Nesterov, 1983) by replacing $\nabla f(x)$ by $g(x)$. While SNAG refers to the method in (Hermant et al., 2025), which can be treat as an alternative discretization of the HNAG flow (Appendix E). The stability of SNAG can be also explained with our theoretical analysis.

Convex optimization We first consider the family of objective functions from Gupta et al. (2024):

$$f_d : \mathbb{R} \to \mathbb{R}, \qquad f_d(x) = \begin{cases} |x|^d, & |x| < 1, \\ 1 + d(|x|-1), & \text{else}, \end{cases}$$

for $d \ge 2$, with gradient estimators $g(x) = (1+\sigma Z)\nabla f(x)$, where $Z \sim \mathcal{N}(0, I_d)$ is a standard normal random variable. The functions f_d belong to $\mathcal{S}_{0,L}$ with $L = d(d-1)$.

We compare SHANG and SHANG++ with SGD, NAG, AGNES (Gupta et al., 2024), and SNAG (Hermant et al., 2025) under $\sigma \in \{0, 10, 50\}$ and $d \in \{4, 16\}$. The parameters used follow their optimal choices for the convex case. All simulations are initialized at $x_0 = 1$, and expectations are averaged over 200 independent runs. See Appendix A.1 for the full experimental setup, hyperparameter choices, and results.

In Figure 3.1, SHANG and SHANG++ remain stable as σ increases, while NAG diverges at large noise. SHANG outperforms classical momentum methods, and SHANG++ further accelerates convergence, showing that its noise-damping term improves both rates and stability. These results confirm robustness with minimal tuning and preserved acceleration even under high noise.

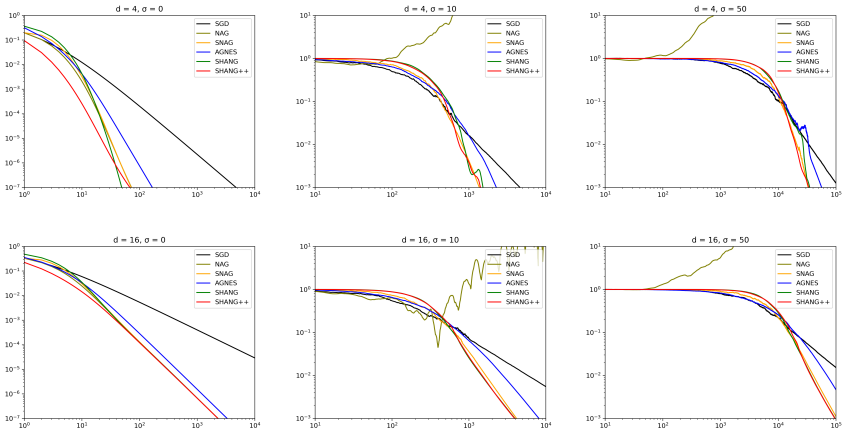

Figure 3.1: Performance of different algorithms under varying noise levels.

For deep learning tasks, we adopt SHANG++ with three explicit hyperparameters (α, γ, m), with $\mu = 0$ and $\beta = \alpha/\gamma$, summarized in Algorithm 1, where v is updated first by index shifting.

Algorithm 1: SHANG++ for Deep Learning

Input: Objective function f, initial point x_0, step size α, time scaling factor γ, noise-damping m, , iteration horizon T.

$k \leftarrow 1$, $v_0 \leftarrow x_0$, $x_1 \leftarrow x_0$, $\tilde{\alpha} \leftarrow \frac{\alpha}{1+m\alpha}$

while $k \leq T$ do

$g_k \leftarrow \frac{1}{M} \sum_{i \in B} \nabla f_i(x_k)$ // stochastic gradient estimate

$v_k \leftarrow v_{k-1} - \frac{\alpha}{\gamma} g_k$

$x_{k+1} \leftarrow \frac{1}{1+\tilde{\alpha}} x_k + \frac{\tilde{\alpha}}{1+\tilde{\alpha}} v_k - \frac{\tilde{\alpha}}{1+\tilde{\alpha}} \frac{\alpha}{\gamma} g_k$

$k \leftarrow k + 1$

end

return x_T

Classification Tasks on MNIST, CIFAR-10 and CIFAR-100 We benchmark SHANG, SHANG++, Adam (Kingma & Ba, 2015), SNAG, AGNES, NAG, SHB (SGD with momentum), and plain SGD on three tasks: training LeNet-5 on MNIST (LeCun et al., 1998), ResNet-34 (He et al., 2016) on CIFAR-10 (Krizhevsky, 2009), and ResNet-50 on CIFAR-100. Each model is trained for 50 epochs, and results are reported as mean ± s.d. over five random seeds.

For hyperparameter selection, SHANG and SHANG++ used $\alpha = 0.5$ with γ chosen from grids: $\{1, 1.5, 2\}$ for LeNet-5, $\{5, 10\}$ for ResNet-34, and $\{10, 15\}$ for ResNet-50. SHANG++ fixed $m = 1.5$. AGNES followed defaults $(\eta, \alpha, m) = (0.01, 0.001, 0.99)$; SNAG used (η, β) with $\eta \in \{0.5, \ldots, 0.001\}$, $\beta \in \{0.7, 0.8, 0.9, 0.99\}$, where $(0.05, 0.9)$ performed best, consistent with prior CIFAR work. Other baselines used $\eta = 0.001$ and momentum 0.99 when applicable. After 25 epochs, all baseline learning rates (including AGNES's correction) were decayed by 0.1, while γ was doubled for our methods. Full details are in Appendix A.2.

Figure 3.2 shows ResNet-34/50 training and test losses on CIFAR-10/100. SHANG and SHANG++ deliver competitive or superior performance to non-adaptive baselines. Batch size strongly affects gradient variance: smaller batches increase noise, larger batches reduce it. At 256, all methods are stable and gaps narrow; at 50, NAG, SNAG, and AGNES oscillate with wider bands (AGNES also plateaus higher). In contrast, SHANG and SHANG++ achieve the lowest losses with tight bands across seeds. Adam remains competitive in accuracy but shows noisier test loss. Table 3.1 further summarizes results: SHANG and

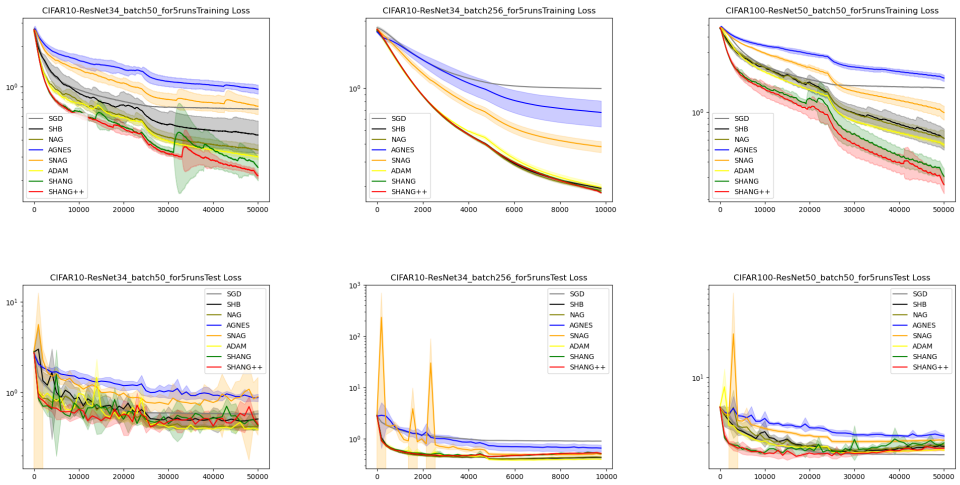

Figure 3.2: Training and test loss (log scale, running average with decay 0.99) on CIFAR-10 with ResNet-34 (batch sizes 50 and 256) (Left and Middle Column) and CIFAR-100 with ResNet-50 (batch size 50) (Right Column).

SHANG++ often match or surpass AGNES and SNAG, while clearly improving over SGD and NAG.

Table 3.1: Test accuracy of SGD, NAG, Adam, AGNES, SHANG, and SHANG++ on MNIST (LeNet-5), CIFAR-10 (ResNet-34), and CIFAR-100 (ResNet-50). Here b is batch size.

	SGD	NAG	Adam	AGNES	SNAG	SHANG	SHANG++
LeNet-5	91.068	98.906	99.072	98.876	99.07	99.064	**99.112**
($b = 50$)	±0.113	±0.082	±0.071	±0.093	±0.085	±0.018	±0.026
ResNet-34	79.908	86.428	87.378	70.492	77.654	87.15	**87.398**
($b = 50$)	±0.114	±0.805	±0.26	±2.511	±2.7	±0.824	±0.502
ResNet-34	68.49	87.614	**88.226**	77.84	84.5	86.67	86.572
($b = 256$)	±0.192	±0.291	±0.106	±3.696	±0.92	±0.13	±0.169
ResNet-50	58.326	57.658	59.872	42.82	49.514	63.306	**65.018**
($b = 50$)	±0.506	±1.443	±0.614	±1.239	±1.559	±0.934	±1.254

Robustness to Multiplicative Gradient Noise Our theory predicts that time-scale coupling (α, γ) in SHANG and (α, γ, m) in SHANG++ mitigates multiplicative gradient noise. To test this, we fix one hyperparameter configuration per optimizer and evaluate across $\sigma \in \{0, 0.05, 0.1, 0.2, 0.5\}$. The effective noise is higher than nominal σ, since minibatch SGD adds sampling noise. This one-shot protocol isolates each optimizer's robustness without re-tuning. All experiments use CIFAR-10 with ResNet-34, batch size 50, the same settings as subsection 3, trained for 100 epochs and averaged over three seeds. Final validation error at epoch 100 is reported; full setup and hyperparameters are in Appendix A.4.

Figure 3.3 shows mean final validation error under varying noise, and Table 3.2 reports relative degradation $\Delta(\sigma) = (\mathbb{E}(\sigma) - \mathbb{E}(0))/\mathbb{E}(0)$, where $\mathbb{E}(\sigma)$ is the mean Top-1 error at noise level σ (averaged over three seeds).

1. At $\sigma = 0$, SHANG and SHANG++ reach 15.9%, outperforming SNAG (17.5%) and AGNES (20.5%).
2. At $\sigma = 0.1$, SHANG improves slightly (-0.3 pt), SHANG++ is nearly unchanged (-0.1 pt), SNAG improves marginally (-0.4 pt), while AGNES worsens (+3.3 pt).

3. At $\sigma = 0.5$, SHANG and SHANG++ remain near 16%, while SNAG rises to 17.6% and AGNES drifts to 23.2% (≈13.5% relative increase).

These results align with our Lyapunov analysis: time-scale coupling (α, γ, m) suppresses σ^2 amplification, ensuring stable performance without re-tuning. SNAG is stable but less accurate, while AGNES is most sensitive to noise.

Table 3.2: Relative change in final Top-1 error compared with $\sigma = 0$ (lower is better; negative values indicate improvement). Values are averaged over three seeds.

Method	Relative degradation $\Delta(\%)$ at σ			
	0.05	0.1	0.2	0.5
SHANG	−2.5	−2.1	−1.0	−0.2
SHANG++	+3.4	−0.6	−2.1	−0.9
AGNES	−14.4	+16.0	+14.6	+13.5
SNAG	−2.0	−2.1	−5.0	0.7

Figure 3.3: Validation error under varying multiplicative noise level σ. Lower is better.

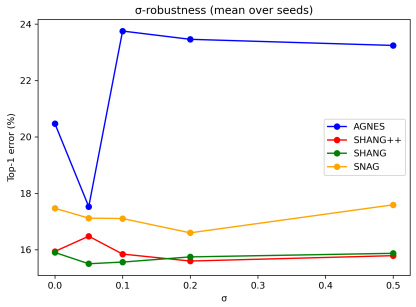

Image Reconstruction with Small Batch Size We further evaluate our algorithms on a generative task of image reconstruction with small-batch training, using a lightweight U-Net (Ronneberger et al., 2015) on CIFAR-10 with batch size 5. SHANG and SHANG++ are compared against SNAG, AGNES, NAG, SGD, SHB, and Adam, with full experimental details provided in the appendix A.5. Figure 3.4 shows training and test losses. Adam

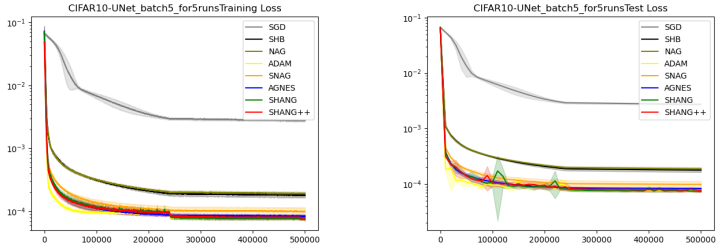

Figure 3.4: Training and test loss (log scale, running average with decay 0.99) on CIFAR-10 using U-Net with batch size 5.

achieves the lowest loss due to its adaptive learning rate, but both SHANG and SHANG++ outperform all other non-adaptive methods. In particular, SHANG++ shows stable and efficient training even in this high-noise regime, highlighting its practical robustness. We also conduct a comparative hyperparameter study; full settings and results are given in Appendix A.6.

4 Conclusion

We presented SHANG++, an accelerated first-order stochastic optimizer for robust and simple training under multiplicative noise. Theoretically, it retains the optimal $\mathcal{O}(1/k^2)$ rate in convex settings and achieves the fastest known acceleration under MNS for quadratic problems. Empirically, across convex tasks, image classification, and generative reconstruction, one-shot hyperparameter choices sustain near noise-free accuracy (within 1% for $\sigma \leq 0.5$). Compared with NAG, SNAG, AGNES, and Adam, SHANG++ shows greater stability in small-batch or high-noise regimes while delivering competitive or improved accuracy, making it a practical optimizer for large-scale noisy training.

References

Mahmoud Assran and Michael Rabbat. On the convergence of nesterov's accelerated gradient method in stochastic settings. In Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR, 2020.

Jean-François Aujol and Charles Dossal. Stability of over-relaxations for the forward-backward algorithm, application to fista. SIAM Journal on Optimization, 25(4):2408–2433, 2015.

G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM Journal on Optimization, 3(3):538–543, 1993.

Long Chen and Hao Luo. A unified convergence analysis of first-order convex optimization methods via strong lyapunov functions, 2021.

Long Chen and Zeyi Xu. Hnag++: A super-fast accelerated gradient method for convex optimization, 2025. Under preparation.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 146:37–75, 2014.

Swetha Ganesh, Rohan Deb, Gugan Thoppe, and Amarjit Budhiraja. Does momentum help in stochastic optimization? a sample complexity analysis. In Uncertainty in Artificial Intelligence (UAI), pp. 602–612. PMLR, 2023.

Kanan Gupta, Jonathan W. Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

Julien Hermant, Marien Renaud, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Gradient correlation is a key ingredient to accelerate SGD with momentum. In International Conference on Learning Representations (ICLR), 2025.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In International Conference on Learning Representations (ICLR), 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

Achim Klenke. Probability Theory: A Comprehensive Course. Universitext. Springer, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. In International Conference on Learning Representations (ICLR), 2020.

Tianyi Liu, Zhehui Chen, Enlu Zhou, and Tuo Zhao. Toward deeper understanding of nonconvex stochastic optimization with momentum using diffusion approximations, 2018.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. Soviet Mathematics Doklady, 27:372–376, 1983.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, 1964.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1195–1204. PMLR, 2019.

Lei Wu, Mingze Wang, and Weijie J. Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Xiaoxia Wu, Simon S. Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network, 2019.

LLM usage

In preparing this manuscript, large language models (LLMs) were employed exclusively to assist with language-related tasks, such as improving readability, grammar, and style. The models were not used for research ideation, development of methods, data analysis, or interpretation of results. All scientific content, including problem formulation, theoretical analysis, and experimental validation, was conceived, executed, and verified entirely by the authors. The authors bear full responsibility for the accuracy and integrity of the manuscript.

Ethics statement

This work is purely theoretical and algorithmic, focusing on convex optimization methods. It does not involve human subjects, sensitive data, or applications that raise ethical concerns related to privacy, security, fairness, or potential harm. All experiments are based on publicly available datasets or synthetic data generated by standard procedures. The authors believe that this work fully adheres to the ICLR Code of Ethics.

Reproducibility statement

We have taken several measures to ensure the reproducibility of our results. All theoretical assumptions are explicitly stated, and complete proofs are provided in the appendix. For the experimental evaluation, we describe the setup, parameter choices, and baselines in detail in the main text. The source code for our algorithms and experiments are available as supplementary materials. Together, these resources should allow others to reproduce and verify our theoretical and empirical findings.

A Supplement of Experiments

Here are some experimental setup and results that are not presented in the main text.

A.1 Supplement of the convex experiment

For the convex example in Section 3, we compare SHANG and SHANG++ with SGD, NAG, AGNES, and SNAG under $\sigma \in \{0, 10, 50\}$ and $d \in \{4, 16\}$. The parameters used follow their optimal choices for the convex case. For SHANG, $\alpha_k = \frac{2}{k+1}$, $\gamma_k = \alpha_k^2 L(1+\sigma^2)^2$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$; For SHANG++, $\alpha_k = \frac{2}{k+1}$, $m = 1.5$, $\gamma_k = \frac{\alpha_k^2}{1+m\alpha_k}(1+\sigma^2)^2 L$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$; For AGNES, we adopted the best-performing parameters reported by the authors

for this problem: learning rate $\eta = \frac{1}{L(1+2\sigma^2)}$, correction step size $\alpha = \frac{\eta}{1+\sigma^2}$, and momentum $m_k = \frac{k}{k+5}$. For SNAG, we use $s = \frac{1}{L(1+\sigma^2)}$, $\eta_k = \frac{1}{L(1+\sigma^2)^2}\frac{k+1}{2}$, $\beta = 1$, $\alpha_k = \frac{k^2/(k+1)}{2+(k^2/(k+1))}$. For NAG, we used a learning rate of $\frac{1}{L(1+\sigma^2)}$ and momentum parameter of $\frac{k}{k+3}$. SGD was also run with a learning rate of $\frac{1}{L(1+\sigma^2)}$. All hyperparameter notations match those used in the original publications; note, however, that symbol meanings may vary across algorithms (e.g., α denotes the discretization step size in SHANG, while in AGNES it refers to the correction step size). All simulations are initialized at $x_0 = 1$, and expectations are averaged over 200 independent runs.

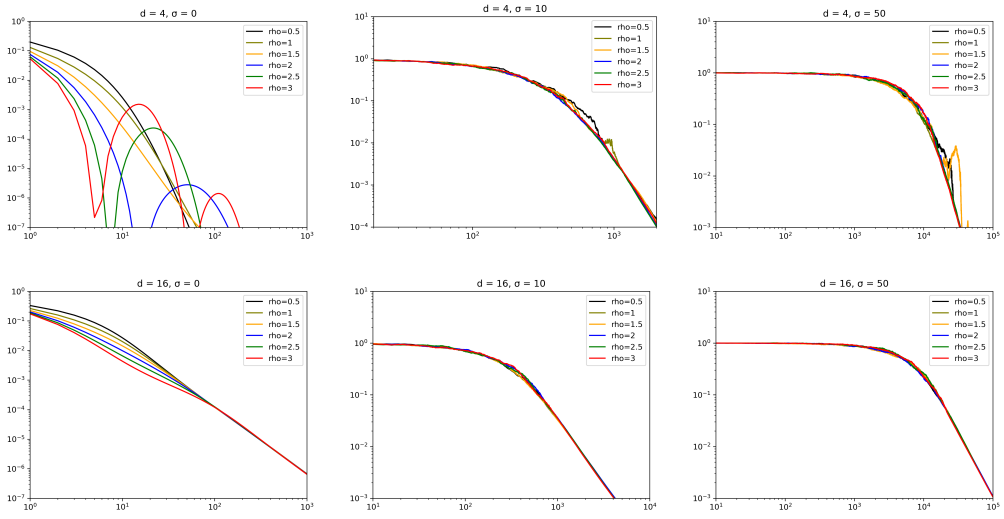

Figure A.1: Log-log plots of $\mathbb{E}[f_d(x_k)]$ for SHANG++ using $m = 0.5$ (black), $m = 1$ (olive), $m = 1.5$ (orange), $m = 2$ (blue), $m = 2.5$ (green), $m = 3$ (red) with $d = 4$ (Top Row) and $d = 16$ (Bottom Row), under noise levels $\sigma = 0$ (Left Column), $\sigma = 10$ (Middle Column) and $\sigma = 50$ (Right Column). From the figures, it can be observed that $m \leq 1.5$ provides a good choice.

Figure A.1 highlights SHANG++'s stability across m: values $m \leq 1.5$ consistently yield strong performance. Our theoretical variance-decay predictions directly manifest in practice.

A.2 Supplement of Classification Tasks

Setup. We benchmark SHANG, SHANG++, Adam, SNAG, AGNES, NAG, SHB (or SGD with momentum) and SGD on the following tasks: training LeNet-5 on the MNIST dataset, training ResNet-34 on the CIFAR-10 image dataset and training ResNet-50 on the CIFAR-100 dataset with standard data augmentation (normalization, random crop, and random flip). All models have pretrain set to True. For each dataset, we run all algorithms for 50 epochs with batch size 50 and report averages over five trials. After 25 epochs, the learning rates for all baseline methods (excluding SHANG and SHANG++) are decayed by a factor of 0.1; AGNES's correction step size is similarly reduced. For our methods, the time-scaling factor γ is doubled after 25 epochs.

For hyperparameter selection, our two methods were evaluated under three settings: $\alpha = 0.5$ with $\gamma \in \{1, 1.5, 2\}$ for LeNet-5, $\gamma \in \{5, 10\}$ for ResNet-34 and $\gamma \in \{10, 15\}$ for ResNet-50. For SHANG++, we fixed $m = 1.5$. AGNES employed the default parameter configuration recommended by its authors, $(\eta, \alpha, m) = (0.01, 0.001, 0.99)$, which has demonstrated strong performance across various tasks. For SNAG, we adopt the two-parameter variant (η, β) proposed by the original authors for machine-learning tasks. Hyperparameters are selected via a grid search, learning rate $\eta \in \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ and momentum $\beta \in \{0.7, 0.8, 0.9, 0.99\}$. Among these, $(\eta, \beta) = (0.05, 0.9)$ yields the best performance, which coincides with the parameter choice recommended by the original authors for training CNNs

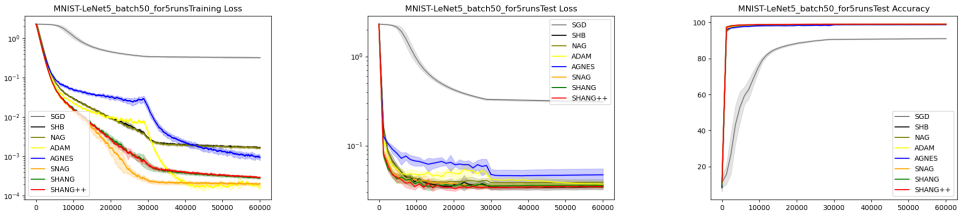

Figure A.2: Training loss (log scale) (left), test loss (log scale) (middle) as a running average with decay rate 0.99, and test accuracy (right) on the MNIST dataset using LeNet-5 trained with batch size 50. The compared methods include SGD (gray), SHB (black), NAG (olive), AGNES (blue), SNAG (orange), Adam (yellow), SHANG (green) and SHANG++ (red). In SHANG, $(\alpha, \gamma) = (0.5, 2)$ and in SHANG++, $(\alpha, \gamma, m) = (0.5, 2, 1.5)$.

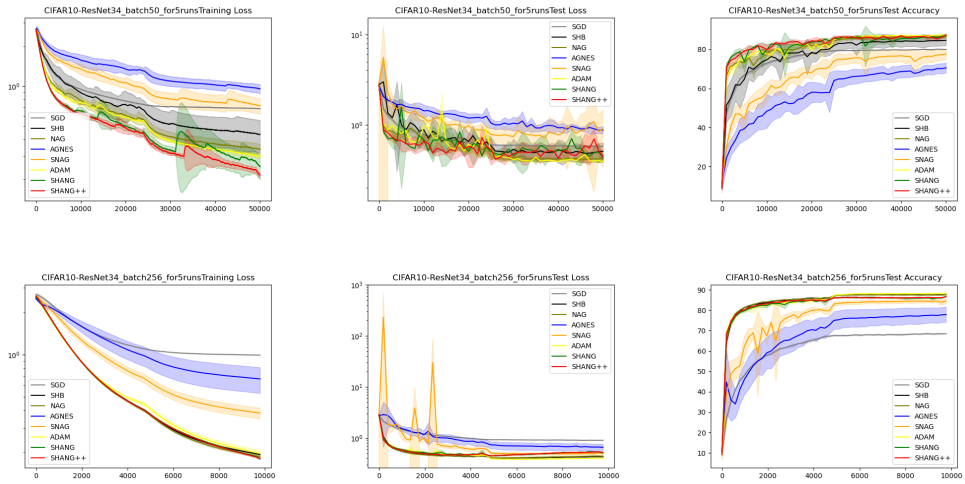

Figure A.3: Training loss (log scale) (left), test loss (log scale) (middle) as a running average with decay rate 0.99, and test accuracy (right) on the CIFAR-10 dataset using ResNet-34 trained with batch sizes 50 (Top Row) and 256 (Bottom Row). The compared methods include SGD (gray), SHB (black), NAG (olive), AGNES (blue), SNAG (orange), Adam (yellow), SHANG (green) and SHANG++ (red). For the choice of γ in SHANG and SHANG++, $\gamma = 10$.

on the CIFAR dataset. All other baseline algorithms used a fixed learning rate of $\eta = 0.001$; for those involving momentum, the momentum coefficient was set to 0.99.

Results. Figures A.2, A.3, A.5, and A.4 depict the evolution of training/test loss and test accuracy across datasets. Overall, SHANG and SHANG++ achieve competitive or superior performance compared with non-adaptive baselines.

A.3 Batch-Size Scaling on CIFAR-10 (ResNet-34)

To further assess the robustness of our algorithms to stochastic gradient noise, we evaluate all methods on CIFAR-10 with ResNet-34 under two batch-size settings: 50 and 256. Smaller batches introduce higher gradient variance, whereas larger batches reduce the noise level. Importantly, all hyperparameters are kept fixed across batch sizes to isolate the effect of noise on algorithmic performance.

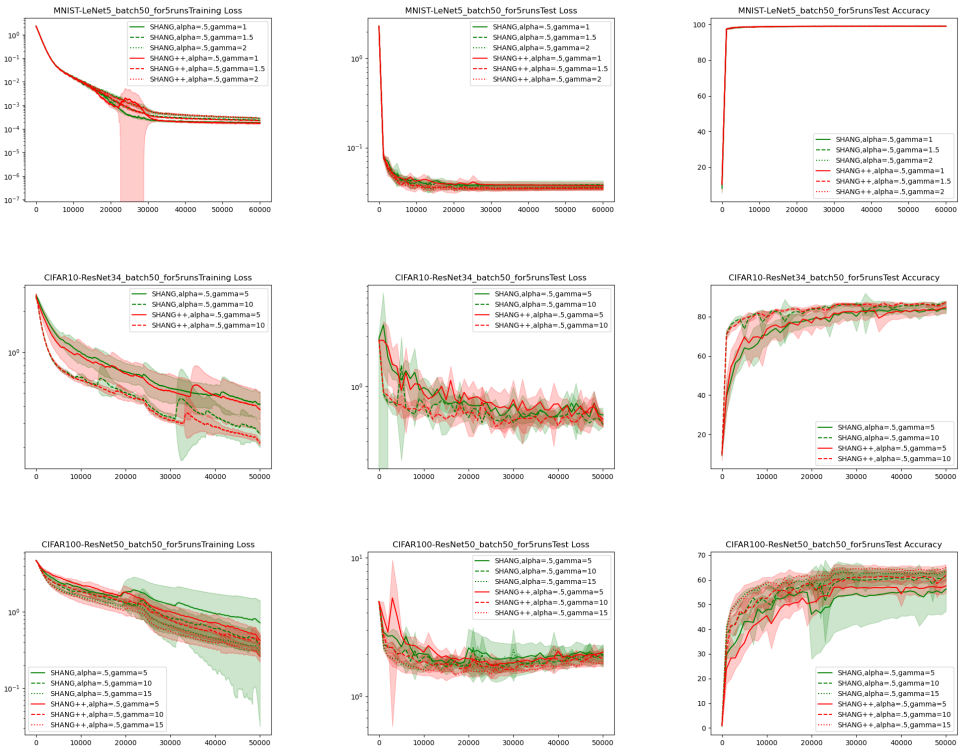

Figure A.4: Training loss (log scale) (Left Column), test loss (log scale) (Middle Column) as a running average with decay rate 0.99, and test accuracy (Right Column) on the MNIST dataset using LeNet-5 (Top Row), CIFAR-10 dataset using ResNet-34 (Middle Row) and CIFAR-100 dataset using ResNet-50 (Bottom Row) trained with batch size 50. The compared methods include SHANG (green) and SHANG++ (red) under different parameter choices.

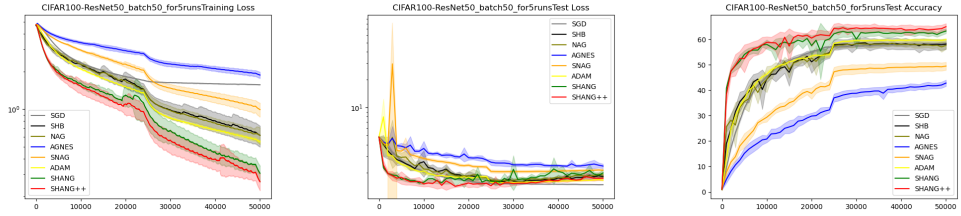

Figure A.5: Training loss (log scale) (left), test loss (log scale) (middle) as a running average with decay rate 0.99, and test accuracy (right) on the CIFAR-100 dataset using ResNet-50 trained with batch size 50. The compared methods include SGD (gray), SHB (black), NAG (olive), AGNES (blue), SNAG (orange), Adam (yellow), SHANG (green) and SHANG++ (red). For the choice of γ in SHANG and SHANG++, $\gamma = 15$.

Setup. All data augmentation and experiments setting follows Appendix A.2. Hyperparameters are held fixed across batch sizes: for SHANG/SHANG++ we use $(\alpha, \gamma) = (0.5, 10)/(\alpha, \gamma, m) = (0.5, 10, 1.5)$, and all baselines reuse their best settings from Appendix 3. No re-tuning is performed when switching the batch size.

Results. Figure A.3 shows the training/test dynamics.

- Small batch (50). Classical momentum variants (NAG, SNAG) and AGNES exhibit larger oscillations and wider variance bands; AGNES also shows a higher error plateau. In contrast, SHANG/SHANG++ produce the lowest losses among non-adaptive methods and maintain narrow shaded regions, indicating markedly improved stability across seeds. Adam remains competitive in accuracy but with higher variance in test loss.
- Large batch (256). The gap between methods narrows: all optimizers become more stable and the curves cluster. SHANG/SHANG++ continue to match the best-performing baselines while preserving smooth convergence.

Robustness to multiplicative noise translates into tangible benefits in the small-batch regime: with a single, fixed hyperparameterization ($\alpha = 0.5, \gamma = 10, m = 1.5$), SHANG/SHANG++ achieve stable training and strong test accuracy without re-tuning, whereas competing momentum methods are more sensitive (larger variance, higher plateaus). As batch size increases, all methods stabilize and the performance gap diminishes, consistent with the noise-abatement expected from larger batches.

A.4 Supplement of Robustness to Multiplicative Gradient Noise

All runs use an identical experimental setup: CIFAR-10 dataset, ResNet-34, batch size 50, trained for 100 epochs, and averaged over three random seeds. Note that the actual gradient noise level experienced by the optimizer is higher than the nominal σ, because minibatch stochastic gradient descent inherently introduces sampling noise. The multiplicative noise we introduce,

$$g(x_k) = (1 + \sigma\mathcal{N}(0, I_d))\nabla f(x_k),$$

is therefore imposed on top of this intrinsic minibatch stochasticity. We record the final validation error at epoch 100.

Discussion. The empirical trends align with our Lyapunov analysis: coupling the time scales (α, γ, m) suppresses the σ^2 amplification and yields stable behavior across noise levels without retuning. SNAG—while reasonably stable—does not match the consistently low error of SHANG/SHANG++, and AGNES is the most sensitive to increased multiplicative noise.

A.5 Supplement of Image Reconstruction

We further evaluate our algorithms on a generative task—image reconstruction with small-batch training, which introduces substantial gradient noise. Specifically, we train a lightweight U-Net (Ronneberger et al., 2015) (base channels $32 \to 64 \to 128$, with bilinear up-sampling and feature concatenation) on CIFAR-10 using batch size 5. We compare SHANG ($\alpha = 0.5, \gamma = 0.5$) and SHANG++ ($\alpha = 0.5, \gamma = 0.5, m = 1$) against SNAG, AGNES, NAG, SGD, SHB, and Adam. All other experimental settings follow those in earlier sections.

A.6 Hyperparameter comparison

To identify optimal hyperparameter configurations for our stochastic algorithms, we perform grid searches over $\alpha \in (0.005, 0.1)$ and $\gamma \in (0.5, 30)$ on MNIST and CIFAR-10 (Figures A.6). For SHANG++, we additionally vary $m \in (0.5, 3)$ while keeping $\alpha = 0.5$ fixed. Results indicate that: (1) $\alpha = 0.5$ and $m = 1.5$ are generally effective across tasks; (2) Smaller γ values work well for LeNet-5, while larger γ are preferred for deeper networks like ResNet-34; (3) SHANG++ exhibits low sensitivity to m in practice, with performance remaining stable across tested values. These findings confirm the practical usability and tuning simplicity of our methods.

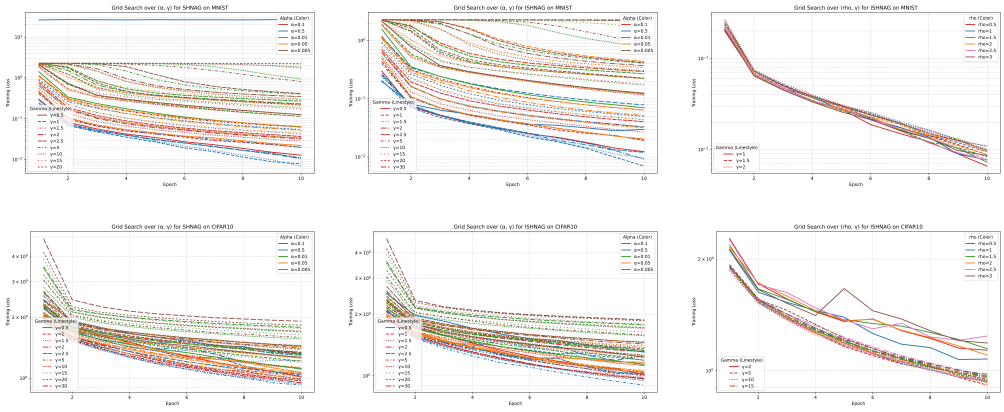

Figure A.6: Training loss (log scale) on the MNIST dataset using LeNet-5 (Top row) and CIFAR-10 dataset using ResNet-34 (Bottom row) trained with batch size 50. The plots show results for SHANG (left) and SHANG++ (middle) under different combinations of hyperparameters $\alpha \in \{0.1, 0.5, 0.01, 0.05, 0.005\}$ (different color) and $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 5, 10, 15, 20\}$ (different line style). The left two figures show that $\alpha = 0.5$ and $\gamma \in \{1, 1.5, 2\}$ are relatively good parameter choices. The rightmost plot illustrates the performance of the ISHNAG method under different combinations of $\gamma \in \{1, 1.5, 2\}$ (on MNIST dataset), $\gamma \in \{2, 5, 10, 15\}$ (on CIFAR-10 dataset) and $m \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ with α fixed at 0.5. The differences among various m values are minor for this task. In practice, we typically choose $m = 1.5$.

B SHANG

B.1 model

Applying a Gauss-Seidel-type scheme to discretize HNAG flow (2.2) and replace the deterministic gradient $\nabla f(x_k)$ with its unbiased stochastic estimate $g(x_k)$, we can obtain the Stochastic Hessian-driven Nesterov Accelerated Gradient (SHANG) method:

$$\begin{aligned} \frac{x_{k+1} - x_k}{\alpha_k} &= v_k - x_{k+1} - \beta_k g(x_k) \\ \frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(x_{k+1} - v_{k+1}) - \frac{1}{\gamma_k} g(x_{k+1}) \\ \frac{\gamma_{k+1} - \gamma_k}{\alpha_k} &= \mu - \gamma_{k+1} \end{aligned} \tag{B.1}$$

In the strongly convex case, we fix $\gamma = \mu$ and use a constant step size α; in general case, we set $\mu = 0$ and allow both α_k and γ_k to vary. The coupling $\beta_k > 0$ depends on (α_k, γ_k) and typically scales as $(1+\sigma^2)\alpha_k/\gamma_k$. Consequently, SHANG reduces to a two-parameter scheme (α, β) in the strongly convex regime and a three-parameter scheme (α, γ, β) otherwise. For practical tuning, tying β to α and γ via $\beta = \alpha/\gamma$ yields an effective two-parameter (α, γ) algorithm. The SHANG method for deep learning tasks is described in Algorithm 2.

Algorithm 2: SHANG for Deep Learning

Input: Objective function f, initial point x_0, stepsize α, time scaling factor γ, iteration horizon T.

$n \leftarrow 0, \quad v_0 \leftarrow x_0, \quad x_1 \leftarrow x_0$

while $k < T$ do

  $g_k \leftarrow \nabla f(x_k)$ // gradient estimate

  $v_k = v_{k-1} - \frac{\alpha}{\gamma} g_k$

  $x_{k+1} = \frac{1}{1+\alpha} x_k + \frac{\alpha}{1+\alpha} v_k - \frac{\alpha}{1+\alpha}\frac{\alpha}{\gamma} g_k$

  $k \leftarrow k+1$

end

return x_T

Observe that SHANG is the $m = 0$ special case of SHANG++. Table B.1 summarizes the theoretical convergence complexities and the number of tunable parameters required by leading stochastic optimization methods under multiplicative noise. As shown, SHANG and SHANG++ achieve optimal theoretical guarantees while significantly reducing hyper-parameter complexity.

Table B.1: Assume f is L-smooth and $g(x)$ satisfies the multiplicative noise scaling (MNS) condition (see Definition 1.1) with constant $\sigma \geq 0$. This table summarizes the iteration complexity of leading first-order stochastic optimization algorithms under optimal parameter settings to reach ε-precision.

Algorithm	Convex	Strongly Convex
SGD (Hermant et al., 2025)	$(1+\sigma^2)\frac{L}{\varepsilon}$	$(1+\sigma^2)\frac{L}{\mu}\log(\frac{1}{\varepsilon})$
NAG (Gupta et al., 2024)	$\sqrt{\frac{1+\sigma^2}{1-\sigma^2}}\sqrt{\frac{L}{\varepsilon}}$	$\sqrt{\frac{1+\sigma^2}{1-\sigma^2}}\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})$
AGNES (Gupta et al., 2024)	$\sqrt{\frac{L(1+2\sigma^2)(1+\sigma^2)}{\varepsilon}}$	$(1+\sigma^2)\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})$
SNAG (Hermant et al., 2025)	$(1+\sigma^2)\sqrt{\frac{L}{\varepsilon}}$	$(1+\sigma^2)\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})$
SHANG (Our Algorithm 1)	$(1+\sigma^2)\sqrt{\frac{L}{\varepsilon}}$	$(\mathbf{1}+\sigma^{\mathbf{2}})\sqrt{\frac{\mathbf{L}}{\mu}}\log(\frac{\mathbf{1}}{\varepsilon})$
SHANG++ (Our Algorithm 2)	$(1+\sigma^2)\sqrt{\frac{L}{\varepsilon}}$	$(\mathbf{1}+\sigma^{\mathbf{2}})\sqrt{\frac{\mathbf{L}}{\mu}-\mathbf{1}}(\mathbf{1}+\sqrt{\frac{\mu}{\mathbf{L}-\mu}})^{-\mathbf{1}}\log(\frac{\mathbf{1}}{\varepsilon})$
SHANG++ for quadratic f (Our Algorithm 2)	$(1+\sigma^2)\sqrt{\frac{L}{\varepsilon}}$	$\frac{\mathbf{1}+\sigma^{\mathbf{2}}}{\mathbf{2}}\sqrt{\frac{\mathbf{L}}{\mu}-\mathbf{1}}(\mathbf{1}+\frac{\mathbf{1}}{\mathbf{2}(\mathbf{1}+\sigma^{\mathbf{2}})}\sqrt{\frac{\mu}{\mathbf{L}-\mu}})^{-\mathbf{1}}\log(\frac{\mathbf{1}}{\varepsilon})$

B.2 Convergence Analysis for SHANG

Define the discrete Lyapunov function

$$\mathcal{E}(z_k^+;\gamma_k) = f(x_k^+) - f(x^*) + \frac{\gamma_k}{2}\|v_k - x^*\|^2 \tag{B.2}$$

where $z_k^+ = (x_k^+, v_k)$, $z_k = (x_k, v_k)$ and $z^* = (x^*, x^*)$. The following theorem establishes a decay bound for $\mathbb{E}\left[\mathcal{E}(z_k^+;\gamma_k)\right]$.

Theorem B.1. Let $f \in \mathcal{S}_{\mu,L}$, (x_k, v_k) be generated by SHANG (B.1). $x_k^+ = x_k - \alpha_k\beta_k g(x_k)$ is an auxiliary variable. Assume $g(x)$ (defined in (1.2)) satisfies the MNS condition with constant σ. Given $x_0^+ = v_0 = x_0$,

(1) When $0 < \mu < L < \infty$, choose step size $0 < \alpha \leq \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L}}$ and $\beta = \frac{(1+\sigma^2)\alpha}{\mu}$, we have

$$\mathbb{E}\left[f(x_{k+1}^+) - f(x^*) + \frac{\mu}{2}\|v_{k+1} - x^*\|^2\right] \leq (1+\alpha)^{-(k+1)}\mathcal{E}_0^\mu$$

(2) When $\mu = 0$, choose $\alpha_k = \frac{2}{k+1}$, $\gamma_k = \alpha_k^2(1+\sigma^2)^2 L$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$, we have

$$\mathbb{E}\left[f(x_{k+1}^+) - f(x^*) + \frac{\gamma_{k+1}}{2}\|v_{k+1} - x^*\|^2\right] \leq \frac{2}{(k+2)(k+3)}\mathcal{E}_0^{\gamma_0} = \mathcal{O}(\frac{1}{k^2})$$

where $\mathcal{E}_0^\mu = f(x_0) - f(x^*) + \frac{\mu}{2}\|x_0 - x^*\|^2$ and $\mathcal{E}_0^{\gamma_0} = f(x_0) - f(x^*) + \frac{\gamma_0}{2}\|x_0 - x^*\|^2$.

When $\sigma = 0$, SHANG reduces to the deterministic HNAG method analyzed in Chen & Luo (2021).

Before presenting the proof of Theorem B.1, we first establish several auxiliary lemmas, beginning with one that relies on conditional expectations under the MNS assumption.

Lemma B.1. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_{k\geq 0}, \mathbb{P})$ be a complete probability space with filtration $\{\mathcal{F}_k\}_{k\geq 0}$. Suppose x_k is generated by SHANG/SHANG++, $g(x_k)$ denotes the stochastic estimator of $\nabla f(x_k)$, then the following statements hold

1. $\mathbb{E}\left[g(x_k) \mid \mathcal{F}_k\right] = \nabla f(x_k)$.
2. $\mathbb{E}\left[\|g(x_k) - \nabla f(x_k)\|^2\right] \leq \sigma^2\|\nabla f(x_k)\|^2$.
3. $\mathbb{E}\left[\langle g(x_k), \nabla f(x_k)\rangle\right] = \|\nabla f(x_k)\|^2$
4. $\mathbb{E}\left[\|g(x_k)\|^2\right] \leq (1+\sigma^2)\|\nabla f(x_k)\|^2$

Proof of Lemma B.1. First and second claim. This follows from Fubini's theorem.

Third claim. For the third result, we observe that since f is a deterministic function, $\nabla f(x_k)$ is $\mathcal{F}_k$-measurable, then, by the Theorem 8.14 in Klenke (2013), we have

$$\mathbb{E}\left[\langle g(x_k), \nabla f(x_k)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle g(x_k), \nabla f(x_k)\rangle \mid \mathcal{F}_k\right]\right] = \mathbb{E}\left[\langle \mathbb{E}\left[g(x_k) \mid \mathcal{F}_k\right], \nabla f(x_k)\rangle\right] = \mathbb{E}\left[\|\nabla f(x_k)\|^2\right]$$

Fourth claim. For the fourth result, using the previous results, we have

$$\begin{aligned}
\mathbb{E}\left[\|g(x_k)\|^2\right] &= \mathbb{E}\left[\|g(x_k) - \nabla f(x_k)\|^2 + 2\langle g(x_k), \nabla f(x_k)\rangle - \|\nabla f(x_k)\|^2\right] \\
&= \mathbb{E}\left[\|g(x_k) - \nabla f(x_k)\|^2\right] + \mathbb{E}\left[2\langle g(x_k), \nabla f(x_k)\rangle\right] - \|\nabla f(x_k)\|^2 \\
&\leq \sigma^2\|\nabla f(x_k)\|^2 + 2\|\nabla f(x_k)\|^2 - \|\nabla f(x_k)\|^2 \\
&= (1+\sigma^2)\|\nabla f(x_k)\|^2
\end{aligned}$$

□

Under the MNS assumption, this setup of auxiliary variable x^+ yields the following descent lemma for smooth objectives.

Lemma B.2. Suppose that $x_k^+ = x_k - \eta g(x_k)$, $f \in \mathcal{C}_L^{1,1}$. Given $0 < \eta \leq \frac{1}{L(1+\sigma^2)}$, we have

$$\mathbb{E}\left[f(x_k^+) - f(x^*)\right] \leq f(x_k) - f(x^*) - \frac{\eta}{2}\|\nabla f(x_k)\|^2$$

Proof of Lemma B.2. Using the L-smoothness of the function f:

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d \tag{B.3}$$

and Lemma B.1, under the assumption of $0 < \eta \leq \frac{1}{L(1+\sigma^2)}$, we can obtain the desired result

$$\begin{aligned}
\mathbb{E}\left[f(x_k^+)\right] &\leq \mathbb{E}\left[f(x_k) - \langle \eta g(x_k), \nabla f(x_k)\rangle + \frac{L}{2}\|\eta g(x_k)\|^2\right] \\
&= f(x_k) - \mathbb{E}\left[\langle \eta g(x_k), \nabla f(x_k)\rangle\right] + \mathbb{E}\left[\frac{L}{2}\|\eta g(x_k)\|^2\right] \\
&\leq f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L\eta^2(1+\sigma^2)}{2}\|\nabla f(x_k)\|^2 \\
&= f(x_k) - \eta(1 - \frac{L(1+\sigma^2)\eta}{2})\|\nabla f(x_k)\|^2 \\
&\leq f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2
\end{aligned}$$

□

Define an auxiliary variable $x_k^+ = x_k - \alpha_k\beta_k g(x_k)$, substitue it into (Eq.B.1) yield:

$$\begin{aligned}
\frac{x_{k+1} - x_k^+}{\alpha_k} &= v_k - x_{k+1} \\
\frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(x_{k+1} - v_{k+1}) - \frac{1}{\gamma_k}g(x_{k+1}) \\
\frac{\gamma_{k+1} - \gamma_k}{\alpha_k} &= \mu - \gamma_{k+1}
\end{aligned} \tag{B.4}$$

The next lemma controls the decay of $\mathbb{E}\left[\mathcal{E}(z_{k+1}^+; \gamma_{k+1})\right]$.

Lemma B.3. Let $f \in \mathcal{S}_{\mu,L}$ with $0 \leq \mu < L < \infty$, Lyapunov function $\mathcal{E}$ is defined by (B.2). Given (v_k, x_k^+), (x_{k+1}, v_{k+1}) are generated by (B.4) and $x_{k+1}^+ = x_{k+1} - \alpha_{k+1}\beta_{k+1}g(x_{k+1})$. Assume $0 < \alpha_{k+1}\beta_{k+1} = \alpha_k\beta_k \leq \frac{1}{L(1+\sigma^2)}$, we have

$$\begin{aligned}
&(1+\alpha_k)\mathbb{E}\left[\mathcal{E}(z_{k+1}^+; \gamma_{k+1})\right] \\
&\leq \mathcal{E}(z_k^+; \gamma_k) + \mathbb{E}\left[\frac{1}{2}(\frac{\alpha_k^2(1+\sigma^2)}{\gamma_k} - (1+\alpha_k)\alpha_k\beta_k)\|\nabla f(x_{k+1})\|^2 - \frac{\alpha_k\mu}{2}\|x_{k+1} - v_{k+1}\|^2 - D_f(x_k^+, x_{k+1})\right]
\end{aligned}$$

proof of Lemma B.3. By Lemma B.2, if $0 < \alpha_k\beta_k = \alpha_{k+1}\beta_{k+1} \leq \frac{1}{L(1+\sigma^2)}$, we obtain the one-step decrease

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+; \gamma_{k+1})\right] - \mathcal{E}(z_k^+; \gamma_k) \\
&\leq \mathbb{E}\left[\mathcal{E}(z_{k+1}; \gamma_{k+1}) - \mathcal{E}(z_k^+; \gamma_k) - \frac{\alpha_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right] \\
&= \mathbb{E}\left[\mathcal{E}(z_{k+1}; \gamma_k) - \mathcal{E}(z_k^+; \gamma_k) + \frac{\gamma_{k+1} - \gamma_k}{2}\|v_{k+1} - x^*\|^2 - \frac{\alpha_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right]
\end{aligned} \tag{B.5}$$

Applying the Bregman divergence identity Chen & Teboulle (1993):

$$\langle \nabla f(y) - \nabla f(x), y - z\rangle = D_f(z, y) + D_f(y, x) - D_f(z, x) \quad \forall, x, y, z \in \mathbb{R}^d \tag{B.6}$$

together with the representation $\mathcal{E}(z;\gamma) = D_{\mathcal{E}}(z, z^*; \gamma)$ and the update rules into (B.5), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\gamma_{k+1})\right] - \mathcal{E}(z_k^+;\gamma_k) \\
&\leq \mathbb{E}\left[\langle \nabla\mathcal{E}(z_{k+1};\gamma_k), z_{k+1} - z_k^+\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\gamma_k) + \frac{\gamma_{k+1}-\gamma_k}{2}\|v_{k+1}-x^*\|^2 - \frac{\alpha_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right] \\
&\leq \mathbb{E}\left[\langle \nabla f(x_{k+1}) - \nabla f(x^*), x_{k+1} - x_k^+\rangle + \gamma_k\langle v_{k+1}-x^*, v_{k+1}-v_k\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\gamma_k)\right. \\
&\qquad \left. + \frac{\alpha_k(\mu-\gamma_{k+1})}{2}\|v_{k+1}-x^*\|^2 - \frac{\alpha_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right] \\
&= \mathbb{E}\left[-\alpha_k\langle \nabla f(x_{k+1}) - \nabla f(x^*), x_{k+1}-x^*\rangle + \alpha_k\langle \nabla f(x_{k+1}), v_k - x^*\rangle - \alpha_k\langle g(x_{k+1}), v_{k+1}-x^*\rangle\right. \\
&\qquad + \alpha_k\mu\langle v_{k+1}-x^*, x_{k+1}-v_{k+1}\rangle + \frac{\alpha_k(\mu-\gamma_{k+1})}{2}\|v_{k+1}-x^*\|^2 - D_{\mathcal{E}}(z_k^+, z_{k+1};\gamma_k) \\
&\qquad \left. - \frac{\alpha_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right]
\end{aligned} \tag{B.7}$$

By the definition of the Bregman divergence and the μ-strong convexity of f, we have

$$\begin{aligned}
\langle \nabla f(x_{k+1}) - \nabla f(x^*), x_{k+1} - x^*\rangle &= D_f(x_{k+1}, x^*) + D_f(x^*, x_{k+1}) \\
&\geq D_f(x_{k+1}, x^*) + \frac{\mu}{2}\|x_{k+1}-x^*\|^2
\end{aligned} \tag{B.8}$$

and

$$\alpha_k\mu\langle v_{k+1}-x^*, x_{k+1}-v_{k+1}\rangle = \frac{\alpha_k\mu}{2}(\|x_{k+1}-x^*\|^2 - \|x_{k+1}-v_{k+1}\|^2 - \|v_{k+1}-x^*\|^2) \tag{B.9}$$

We denote $\mathcal{F}_{k+1} = \sigma(x_0, \cdots, x_{k+1})$ the σ-algebra generated by the $k+1$ first interates $\{x_i\}_{i=1}^{k+1}$ generated by SHANG. Since f is a deterministic function, $v_k - x^*$ is $\mathcal{F}_{k+1}$-measurable, then

$$\begin{aligned}
\mathbb{E}\left[\langle g(x_{k+1}), v_k - x^*\rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle g(x_{k+1}), v_k - x^*\rangle \mid \mathcal{F}_{k+1}\right]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}\left[g(x_{k+1}) \mid \mathcal{F}_{k+1}\right], v_k - x^*\rangle\right] \\
&= \mathbb{E}\left[\langle \nabla f(x_{k+1}), v_k - x^*\rangle\right]
\end{aligned}$$

Now, we apply this result in reverse, and using Young Inequality, Cauchy-Schwarz Inequality to obtain

$$\begin{aligned}
&\mathbb{E}\left[\alpha_k\langle \nabla f(x_{k+1}), v_k - x^*\rangle - \alpha_k\langle g(x_{k+1}), v_{k+1}-x^*\rangle\right] \\
&= \mathbb{E}\left[\alpha_k\langle g(x_{k+1}), v_k - v_{k+1}\rangle\right] \\
&\leq \mathbb{E}\left[\frac{\alpha_k^2}{2\gamma_k}\|g(x_{k+1})\|^2 + \frac{\gamma_k}{2}\|v_k - v_{k+1}\|^2\right] \\
&\leq \mathbb{E}\left[\frac{\alpha_k^2(1+\sigma^2)}{2\gamma_k}\|\nabla f(x_{k+1})\|^2 + \frac{\gamma_k}{2}\|v_k - v_{k+1}\|^2\right]
\end{aligned} \tag{B.10}$$

In addition, using the identity of squares (for v) and Bregman divergence indentity (B.6) (for x^+), we have the component form of

$$D_{\mathcal{E}}(z_k^+, z_{k+1};\gamma_k) = D_f(x_k^+, x_{k+1}) + \frac{\gamma_k}{2}\|v_k - v_{k+1}\|^2 \tag{B.11}$$

Substituting (B.8-B.11) back into (B.7), we can obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\gamma_{k+1})\right] - \mathcal{E}(z_k^+;\gamma_k) \\
&\leq \mathbb{E}\left[-\alpha_k D_f(x_{k+1}, x^*) + \frac{1}{2}\left(\frac{\alpha_k^2(1+\sigma^2)}{\gamma_k} - \alpha_k\beta_k\right)\|\nabla f(x_{k+1})\|^2\right. \\
&\qquad \left. - \frac{\alpha_k\gamma_{k+1}}{2}\|v_{k+1}-x^*\|^2 - \frac{\alpha_k\mu}{2}\|x_{k+1}-v_{k+1}\|^2 - D_f(x_k^+, x_{k+1})\right] \\
&\leq \mathbb{E}\left[-\alpha_k D_f(x_{k+1}^+, x^*) + \frac{1}{2}\left(\frac{\alpha_k^2(1+\sigma^2)}{\gamma_k} - (1+\alpha_k)\alpha_k\beta_k\right)\|\nabla f(x_{k+1})\|^2\right. \\
&\qquad \left. - \frac{\alpha_k\gamma_{k+1}}{2}\|v_{k+1}-x^*\|^2 - \frac{\alpha_k\mu}{2}\|x_{k+1}-v_{k+1}\|^2 - D_f(x_k^+, x_{k+1})\right]
\end{aligned} \tag{B.12}$$

By moving $\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\gamma_{k+1}) = D_f(x_{k+1}^+, x^*) + \frac{\gamma_{k+1}}{2}\|v_{k+1} - x^*\|^2\right]$ to the left side of the inequality to obtain the desired result. □

Now we begin to prove Theorem B.1.

Proof. (1). When $0 < \mu < L < \infty$, set $\gamma = \mu$. By Lemma B.3, if $\alpha\beta \leq \frac{1}{(1+\sigma^2)L}$, we have

$$\begin{aligned}
&(1+\alpha)\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] \\
&\leq \mathcal{E}(z_k^+;\mu) + \mathbb{E}\left[\frac{1}{2}\left(\frac{\alpha^2(1+\sigma^2)}{\mu} - (1+\alpha)\alpha\beta\right)\|\nabla f(x_{k+1})\|^2 - \frac{\alpha\mu}{2}\|x_{k+1} - v_{k+1}\|^2 - D_f(x_k^+, x_{k+1})\right]
\end{aligned} \tag{B.13}$$

Assume $\alpha\beta = \frac{(1+\sigma^2)\alpha^2}{\mu} \leq \frac{1}{(1+\sigma^2)L}$, i.e., the step size satisfies $0 < \alpha \leq \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L}}$ to ensure that all the coefficients of the terms on the right side of the inequality, except for $\mathcal{E}(z_k^+;\mu)$, are non-positive. Thus,

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] \leq (1+\alpha)^{-1}\mathcal{E}(z_k^+;\mu) \leq (1+\alpha)^{-(k+1)}\mathcal{E}(z_0;\mu) \tag{B.14}$$

(2). When $\mu = 0$. Assume $\alpha_k = \frac{2}{k+1}$, $\gamma_k = \alpha_k^2(1+\sigma^2)^2L$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$. Using Lemma B.3 to obtain

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\gamma_{k+1})\right] \leq \frac{k+1}{k+3}\mathcal{E}(z_k^+;\gamma_k) \leq \frac{2}{(k+2)(k+3)}\mathcal{E}(z_0;\gamma_0) \tag{B.15}$$

□

Corollary B.1. Under the setting of Theorem B.1, SHANG achieves an ε-precision solution within the following number of iterations:

(1) When $\mu = 0$, with $\alpha_k = \frac{2}{k+1}$, $\gamma_k = \alpha_k^2(1+\sigma^2)^2L$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$,

$$k \geq \sqrt{\frac{2(f(x_0) - f(x^*) + 2(1+\sigma^2)^2L\|x_0 - x^*\|^2)}{\varepsilon}}$$

(2) When $0 < \mu < L < \infty$, with $\alpha = \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L}}$ and $\beta = \frac{(1+\sigma^2)\alpha}{\mu}$,

$$k \geq (1+\sigma^2)\sqrt{\frac{L}{\mu}}\log\left(\frac{f(x_0) - f(x^*) + \frac{\mu}{2}\|x_0 - x^*\|^2}{\varepsilon}\right).$$

Corollary B.2. In the setting of Theorem B.1, $f(x_k^+) \overset{a.s.}{\to} f(x^*)$.

proof of Corollary B.2. We assume that all the conditions of Theorem B.1 have been met, we have

$$\mathbb{E}\left[\mid f(x_k^+) - f(x^*) \mid\right] = \mathbb{E}\left[f(x_k^+) - f(x^*)\right] \leq Cq^k$$

holds for some positive constant C. Here $0 < q < 1$ is the decay factor. In fact, $q = (1 + \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L}})^{-1}$ in strongly convex cases and $q = \frac{2}{(k+2)(k+3)}$ in convex cases. Since

$$\begin{aligned}
\mathbb{P}\left(\lim_{k\to\infty} f(x_k^+) \neq f(x^*)\right) &= \mathbb{P}\left(\lim_{k\to\infty}\sup \mid f(x_k^+) - f(x^*) \mid > 0\right) \\
&= \mathbb{P}\left(\bigcup_{n=1}^{\infty}\lim_{k\to\infty}\sup \mid f(x_k^+) - f(x^*) \mid > \frac{1}{n}\right) \\
&\leq \sum_{n=1}^{\infty}\mathbb{P}\left(\lim_{k\to\infty}\sup \mid f(x_k^+) - f(x^*) \mid > \frac{1}{n}\right)
\end{aligned}$$

For any $\varepsilon = \frac{1}{n} > 0$ and for any $N \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{P}\left(\lim_{k\to\infty} \sup \mid f(x_k^+) - f(x^*) \mid > \varepsilon\right) &\leq \mathbb{P}\left(\exists k \geq N \quad \text{s.t.} \mid f(x_k^+) - f(x^*) \mid > \varepsilon\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{\infty} \{\mid f(x_k^+) - f(x^*) \mid > \varepsilon\}\right) \\
&\leq \sum_{k=1}^{\infty} \mathbb{P}\left(\mid f(x_k^+) - f(x^*) \mid > \varepsilon\right) \\
&\leq \sum_{k=1}^{\infty} \frac{\mathbb{E}\left[\mid f(x_k^+) - f(x^*) \mid\right]}{\varepsilon} \\
&\leq \frac{C}{\varepsilon} \sum_{k=1}^{\infty} q^k
\end{aligned}$$

where in the penultimate step we use Markov's inequality. The right-hand side of the inequality above represents an infinite series, and by leveraging the convergence of this series, we can conclude that the right-hand side can be made arbitrarily small. Consequently, the left-hand side of the inequality must be zero, implying that $\mathbb{P}\left(\lim_{k\to\infty} f(x_k^+) \neq f(x^*)\right) = 0$. Therefore, $f(x_k^+)$ converges in probability to $f(x^*)$. □

C SHANG++

C.1 Proof of Theorem 2.1

Setting $\gamma = \mu$ and $m = \beta\mu$, and substituting $g_{-\mu}(x_k)$ and $x_k^+ := x_k - \alpha\beta g_{-\mu}(x_k)$ into (2.3) yield:

$$\begin{aligned}
\frac{x_{k+1} - x_k^+}{\alpha} &= v_k - x_{k+1} - \beta\mu(x_{k+1} - x^*) \\
\frac{v_{k+1} - v_k}{\alpha} &= x^* - v_{k+1} - \frac{1}{\mu} g_{-\mu}(x_{k+1}) \\
x_{k+1}^+ &= x_{k+1} - \alpha\beta g_{-\mu}(x_{k+1})
\end{aligned} \tag{C.1}$$

We note that schemes (2.5) and (2.3) generate the same sequences x_k and v_k; the appearance of x^* does not affect the algorithm itself. The form (2.5) is introduced purely for theoretical analysis.

Define the discrete Lyapunov function

$$\mathcal{E}(z_k^+; \mu) = f_{-\mu}(x_k^+) - f_{-\mu}(x^*) + \frac{\mu}{2}\|v_k - x^*\|^2 \tag{C.2}$$

The next lemma controls the decay of $\mathbb{E}\left[\mathcal{E}(z_{k+1}^+; \mu)\right]$.

Lemma C.1. Let $f \in \mathcal{S}_{\mu,L}$ with $0 < \mu < L < \infty$, then $f_{-\mu} \in \mathcal{S}_{0,L-\mu}$. Lyapunov function $\mathcal{E}$ is defined by (C.2). Given (x_k^+, x_k, v_k), $(x_{k+1}^+, x_{k+1}, v_{k+1})$ are generated by (C.1). Assume $0 < \alpha\beta \leq \frac{1}{(L-\mu)(1+\sigma^2)}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+; \mu)\right] - \mathcal{E}(z_k^+; \mu) \\
&\leq \mathbb{E}\left[-(1+\beta\mu)\alpha\big(D_{f_{-\mu}}(x_{k+1}^+, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})\big) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1} - x^*\|^2\right. \\
&\qquad \left. + \frac{1}{2}\big(\frac{\alpha^2(1+\sigma^2)}{\mu} - (1 + (1+\beta\mu)\alpha)\alpha\beta\big)\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right]
\end{aligned}$$

Moreover, if f is quadratic, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[-2(1+\beta\mu)\alpha D_{f_{-\mu}}(x_{k+1}^+, x^*) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1}-x^*\|^2\right. \\
&\qquad \left. + \frac{1}{2}\Big(\frac{\alpha^2(1+\sigma^2)}{\mu} - (1+2(1+\beta\mu)\alpha)\alpha\beta\Big)\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right]
\end{aligned}$$

proof of Lemma C.1. By Lemma B.2, if $0 < \alpha\beta \le \frac{1}{(L-\mu)(1+\sigma^2)}$, we obtain the one-step decrease

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[\mathcal{E}(z_{k+1};\mu) - \mathcal{E}(z_k^+;\mu) - \frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2\right]
\end{aligned} \tag{C.3}$$

Expand it yields

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[\langle\nabla\mathcal{E}(z_{k+1};\mu), z_{k+1}-z_k^+\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\mu) - \frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2\right] \\
&\le \mathbb{E}\left[\langle\nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x^*), x_{k+1}-x_k^+\rangle + \mu\langle v_{k+1}-x^*, v_{k+1}-v_k\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\mu)\right. \\
&\qquad \left. - \frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2\right] \\
&= \mathbb{E}\left[-(1+\beta\mu)\alpha\langle\nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x^*), x_{k+1}-x^*\rangle - \alpha\mu\|v_{k+1}-x^*\|^2 - D_{\mathcal{E}}(z_k^+, z_{k+1};\mu)\right. \\
&\qquad \left. + \alpha\langle\nabla f_{-\mu}(x_{k+1}), v_k - x^*\rangle - \alpha\langle g_{-\mu}(x_{k+1}), v_{k+1}-x^*\rangle - \frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2\right] \\
&= \mathbb{E}\left[-(1+\beta\mu)\alpha\big(D_{f_{-\mu}}(x_{k+1}, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})\big) - \alpha\mu\|v_{k+1}-x^*\|^2 - D_{\mathcal{E}}(z_k^+, z_{k+1};\mu)\right. \\
&\qquad \left. + \alpha\langle\nabla g_{-\mu}(x_{k+1}), v_k - v_{k+1}\rangle - \frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2\right] \\
&\le \mathbb{E}\left[-(1+\beta\mu)\alpha\big(D_{f_{-\mu}}(x_{k+1}^+, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})\big) - \alpha\mu\|v_{k+1}-x^*\|^2 - D_{\mathcal{E}}(z_k^+, z_{k+1};\mu)\right. \\
&\qquad \left. + \alpha\langle\nabla g_{-\mu}(x_{k+1}), v_k - v_{k+1}\rangle - (1+(1+\beta\mu)\alpha)\frac{\alpha\beta}{2}\|\nabla f_{-\mu}(x_{k+1})\|^2\right]
\end{aligned} \tag{C.4}$$

Using the update for v_{k+1} and Lemma B.1 yields the following bound.

$$\begin{aligned}
&\mathbb{E}\left[\alpha\langle g_{-\mu}(x_{k+1}), v_k - v_{k+1}\rangle\right] \\
&= \mathbb{E}\left[\langle\frac{\alpha}{\sqrt{\mu}}g_{-\mu}(x_{k+1}), \sqrt{\mu}(v_k - v_{k+1})\rangle\right] \\
&= \mathbb{E}\left[\frac{1}{2}\|\frac{\alpha}{\sqrt{\mu}}g_{-\mu}(x_{k+1})\|^2 + \frac{1}{2}\|\sqrt{\mu}(v_k - v_{k+1})\|^2 - \frac{1}{2}\|\frac{\alpha}{\sqrt{\mu}}g_{-\mu}(x_{k+1}) - \sqrt{\mu}(v_k - v_{k+1})\|^2\right] \\
&= \mathbb{E}\left[\frac{\alpha^2}{2\mu}\|g_{-\mu}(x_{k+1})\|^2 + \frac{\mu}{2}\|v_k - v_{k+1}\|^2 - \frac{\alpha^2\mu}{2}\|\frac{1}{\mu}g_{-\mu}(x_{k+1}) - \frac{v_k - v_{k+1}}{\alpha}\|^2\right] \\
&\le \mathbb{E}\left[\frac{\alpha^2(1+\sigma^2)}{2\mu}\|\nabla f_{-\mu}(x_{k+1})\|^2 + \frac{\mu}{2}\|v_k - v_{k+1}\|^2 - \frac{\alpha^2\mu}{2}\|v_{k+1}-x^*\|^2\right]
\end{aligned} \tag{C.5}$$

Substituting (B.11) and (C.5) into (C.4) to obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[-(1+\beta\mu)\alpha\big(D_{f_{-\mu}}(x_{k+1}^+, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})\big) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1}-x^*\|^2\right. \\
&\qquad \left. + \frac{1}{2}\Big(\frac{\alpha^2(1+\sigma^2)}{\mu} - (1+(1+\beta\mu)\alpha)\alpha\beta\Big)\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right]
\end{aligned} \tag{C.6}$$

Moreover, if f is quadratic, we have

$$\langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), x - x^* \rangle = D_{f_{-\mu}}(x, x^*) + D_{f_{-\mu}}(x^*, x) = 2D_{f_{-\mu}}(x, x^*) \tag{C.7}$$

Substituting (C.7) back into (C.6), we have the following sharper decay bound:

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[-2(1+\beta\mu)\alpha D_{f_{-\mu}}(x_{k+1}^+, x^*) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1} - x^*\|^2 \right. \\
&\quad \left. + \frac{1}{2}\Big(\frac{\alpha^2(1+\sigma^2)}{\mu} - (1 + 2(1+\beta\mu)\alpha)\alpha\beta\Big)\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right]
\end{aligned} \tag{C.8}$$

By moving $\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right]$ to the left side of the inequality to obtain the desired result. □

Now we begin to prove Theorem 2.1.

Proof. By Lemma C.1, Assume $\beta = \frac{(1+\sigma^2)\alpha}{\mu}$ and $\alpha\beta \le \frac{1}{(1+\sigma^2)(L-\mu)}$, i.e., $0 < \alpha \le \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L-\mu}}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[-(1+\alpha+\sigma^2\alpha)\alpha\big(D_{f_{-\mu}}(x_{k+1}^+, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})\big) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1} - x^*\|^2 \right. \\
&\quad \left. - \frac{(1+\alpha+\sigma^2\alpha)\alpha}{2(1+\sigma^2)(L-\mu)}\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right] \\
&= \mathbb{E}\left[-(1+(1+\sigma^2)\alpha)\alpha\mathcal{E}(z_{k+1}^+) - \frac{\alpha^2\mu}{2}\|v_{k+1} - x^*\|^2 + \frac{\alpha\mu((1+\sigma^2)\alpha - 1)}{2}\|v_{k+1} - x^*\|^2 \right. \\
&\quad \left. -(1+(1+\sigma^2)\alpha)\alpha D_{f_{-\mu}}(x^*, x_{k+1}) - \frac{(1+\alpha+\sigma^2\alpha)\alpha}{2(1+\sigma^2)(L-\mu)}\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right]
\end{aligned} \tag{C.9}$$

Since $\alpha \le \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L-\mu}}$, then $(1+\sigma^2)\alpha \le \sqrt{\frac{\mu}{L-\mu}} \le 1$. Thus,

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] \le (1+\alpha+(1+\sigma^2)\alpha^2)^{-1}\mathcal{E}(z_k^+;\mu) \le (1+\alpha+(1+\sigma^2)\alpha^2)^{-(k+1)}\mathcal{E}(z_0;\mu) \tag{C.10}$$

Moreover, if f is quadratic, using Lemma C.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] - \mathcal{E}(z_k^+;\mu) \\
&\le \mathbb{E}\left[-(1+\alpha)\alpha\big(D_{f_{-\mu}}(x_{k+1}, x^*) + D_{f_{-\mu}}(x^*, x_{k+1})\big) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1} - x^*\|^2 \right. \\
&\quad \left. - \frac{(1+\alpha+\sigma^2\alpha)\alpha}{2(1+\sigma^2)(L-\mu)}\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right] \\
&\le \mathbb{E}\left[-(2\alpha+\alpha^2)D_{f_{-\mu}}(x_{k+1}^+, x^*) - \alpha^2 D_{f_{-\mu}}(x_{k+1}^+, x^*) - \frac{\alpha\mu(2+\alpha)}{2}\|v_{k+1} - x^*\|^2 \right. \\
&\quad \left. - \frac{2\alpha+\alpha^2}{2(1+\sigma^2)(L-\mu)}\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right] \\
&= \mathbb{E}\left[-(2\alpha+\alpha^2)\mathcal{E}(z_{k+1}^+) - \alpha^2 D_{f_{-\mu}}(x_{k+1}^+, x^*) - \frac{2\alpha+\alpha^2}{2(1+\sigma^2)(L-\mu)}\|\nabla f_{-\mu}(x_{k+1})\|^2 - D_{f_{-\mu}}(x_k^+, x_{k+1})\right]
\end{aligned} \tag{C.11}$$

Thus,

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\mu)\right] \le (1+2\alpha+\alpha^2)^{-1}\mathcal{E}(z_k^+;\mu) \le (1+2\alpha+\alpha^2)^{-(k+1)}\mathcal{E}(z_0;\mu) \tag{C.12}$$

□

Corollary C.1. Under the setting of Theorem 2.1, choose $\alpha = \frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L-\mu}}$ and $\beta = \frac{(1+\sigma^2)\alpha}{\mu}$, SHANG++ guarantees an ε-precision solution within the following number of iterations:

$$k \geq (1+\sigma^2)\sqrt{\frac{L}{\mu}-1}\left(1+\sqrt{\frac{\mu}{L-\mu}}\right)^{-1}\log\left(\frac{f(x_0)-f(x^*)}{\varepsilon}\right)$$

If f is quadratic,

$$k \geq \frac{1+\sigma^2}{2}\sqrt{\frac{L}{\mu}-1}\left(1+\frac{1}{2(1+\sigma^2)}\sqrt{\frac{\mu}{L-\mu}}\right)^{-1}\log\left(\frac{f(x_0)-f(x^*)}{\varepsilon}\right)$$

C.2 Proof of Theorem 2.2

To facilitate analysis, we define an auxiliary time-scaling factor $\tilde{\gamma}_k = \frac{\gamma_k}{1+m\alpha_k}$. For any $m \geq 0$, setting $\alpha_k = \frac{2}{k+1}$, $\tilde{\alpha}_k = \frac{\alpha_k}{1+m\alpha_k} = \frac{2}{k+1+2m}$ and $\gamma_k = \alpha_k\tilde{\alpha}_k(1+\sigma^2)^2L$, we have

$$\begin{aligned}
\frac{\tilde{\gamma}_{k+1}-\tilde{\gamma}_k}{\tilde{\alpha}_k} &= \frac{1+m\alpha_k}{\alpha_k}\Big(\frac{\alpha_{k+1}^2(1+\sigma^2)^2L}{(1+m\alpha_{k+1})^2}-\frac{\alpha_k^2(1+\sigma^2)^2L}{(1+m\alpha_k)^2}\Big)\\
&= \frac{k+1+2m}{2}\Big(\frac{4(1+\sigma^2)^2L}{(k+2+2m)^2}-\frac{4(1+\sigma^2)^2L}{(k+1+2m)^2}\Big)\\
&= \frac{k+1+2m}{2}\Big(1-\frac{(k+2+2m)^2}{(k+1+2m)^2}\Big)\tilde{\gamma}_{k+1}\\
&= -\Big(1+\frac{1}{2(k+1+2m)}\Big)\tilde{\gamma}_{k+1}\\
&\leq -\tilde{\gamma}_{k+1}
\end{aligned} \tag{C.13}$$

Define $x_k^+ = x_k - \tilde{\alpha}_k\beta_k g(x_k)$, we can obtain the following equivalent form of SHANG++ for convex problems:

$$\begin{aligned}
\frac{x_{k+1}-x_k^+}{\tilde{\alpha}_k} &= v_k - x_{k+1}\\
\frac{v_{k+1}-v_k}{\alpha_k} &= -\frac{1}{\gamma_k}g(x_{k+1})\\
\frac{\tilde{\gamma}_{k+1}-\tilde{\gamma}_k}{\tilde{\alpha}_k} &\leq -\tilde{\gamma}_{k+1}
\end{aligned} \tag{C.14}$$

Denote the discrete Lyapunov function by

$$\mathcal{E}(z_k^+;\tilde{\gamma}_k) = f(x_k^+) - f(x^*) + \frac{\tilde{\gamma}_k}{2}\|v_k - x^*\|^2 \tag{C.15}$$

The following Lemma establishes a decay bound for $\mathbb{E}\left[\mathcal{E}(z_k^+;\tilde{\gamma}_k)\right]$.

Lemma C.2. Let $f \in \mathcal{S}_{0,L}$, Lyapunov function $\mathcal{E}$ is defined by (C.15). Given (x_k, v_k, x_k^+), (x_{k+1}, v_{k+1}) are generated by (C.14) and $x_{k+1}^+ = x_{k+1} - \tilde{\alpha}_k\beta_k g(x_{k+1})$. Assume $0 < \tilde{\alpha}_k\beta_k = \tilde{\alpha}_{k+1}\beta_{k+1} \leq \frac{1}{L(1+\sigma^2)}$, we have

$$\begin{aligned}
&(1+\tilde{\alpha}_k)\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right]\\
&\leq \mathcal{E}(z_k^+;\tilde{\gamma}_k) + \mathbb{E}\left[-\tilde{\alpha}_k D_f(x^*,x_{k+1}) - D_f(x_k^+,x_{k+1}) + \frac{1}{2}\Big(\frac{\tilde{\alpha}_k^2(1+\sigma^2)}{\tilde{\gamma}_k} - (1+\tilde{\alpha}_k)\tilde{\alpha}_k\beta_k\Big)\|\nabla f(x_{k+1})\|^2\right]
\end{aligned}$$

proof of Lemma C.2. By Lemma B.2, if $0 < \tilde{\alpha}_k\beta_k = \tilde{\alpha}_{k+1}\beta_{k+1} \leq \frac{1}{L(1+\sigma^2)}$, we obtain the one-step decrease

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] - \mathcal{E}(z_k^+;\tilde{\gamma}_k) \\
&\leq \mathbb{E}\left[\mathcal{E}(z_{k+1};\tilde{\gamma}_{k+1}) - \mathcal{E}(z_k^+;\tilde{\gamma}_k) - \frac{\tilde{\alpha}_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right] \\
&= \mathbb{E}\left[\mathcal{E}(z_{k+1};\tilde{\gamma}_k) - \mathcal{E}(z_k^+;\tilde{\gamma}_k) + \frac{\tilde{\gamma}_{k+1}-\tilde{\gamma}_k}{2}\|v_{k+1}-x^*\|^2 - \frac{\tilde{\alpha}_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right]
\end{aligned} \tag{C.16}$$

Expand the above equation and use the update to obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] - \mathcal{E}(z_k^+;\tilde{\gamma}_k) \\
&\leq \mathbb{E}\left[\langle\nabla\mathcal{E}(z_{k+1};\tilde{\gamma}_k), z_{k+1}-z_k^+\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\tilde{\gamma}_k) + \frac{\tilde{\gamma}_{k+1}-\tilde{\gamma}_k}{2}\|v_{k+1}-x^*\|^2 - \frac{\tilde{\alpha}_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right] \\
&\leq \mathbb{E}\left[\langle\nabla f(x_{k+1}) - \nabla f(x^*), x_{k+1}-x_k^+\rangle + \tilde{\gamma}_k\langle v_{k+1}-x^*, v_{k+1}-v_k\rangle - D_{\mathcal{E}}(z_k^+, z_{k+1};\tilde{\gamma}_k)\right. \\
&\qquad \left. -\frac{\tilde{\alpha}_k\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 - \frac{\tilde{\alpha}_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right] \\
&= \mathbb{E}\left[-\tilde{\alpha}_k\langle\nabla f(x_{k+1}) - \nabla f(x^*), x_{k+1}-x^*\rangle + \tilde{\alpha}_k\langle\nabla f(x_{k+1}), v_k - x^*\rangle - \frac{\alpha_k\tilde{\gamma}_k}{\gamma_k}\langle g(x_{k+1}), v_{k+1}-x^*\rangle\right. \\
&\qquad \left. -\frac{\tilde{\alpha}_k\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 - D_{\mathcal{E}}(z_k^+, z_{k+1};\tilde{\gamma}_k) - \frac{\tilde{\alpha}_k\beta_k}{2}\|\nabla f(x_{k+1})\|^2\right]
\end{aligned} \tag{C.17}$$

Using Young Inequality, Cauchy-Schwarz Inequality and $\frac{\alpha_k\tilde{\gamma}_k}{\gamma_k} = \tilde{\alpha}_k$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\alpha}_k\langle\nabla f(x_{k+1}), v_k - x^*\rangle - \frac{\alpha_k\tilde{\gamma}_k}{\gamma_k}\langle g(x_{k+1}), v_{k+1}-x^*\rangle\right] \\
&= \mathbb{E}\left[\tilde{\alpha}_k\langle g(x_{k+1}), v_k - v_{k+1}\rangle\right] \\
&\leq \mathbb{E}\left[\frac{\tilde{\alpha}_k^2}{2\tilde{\gamma}_k}\|g(x_{k+1})\|^2 + \frac{\tilde{\gamma}_k}{2}\|v_k - v_{k+1}\|^2\right] \\
&\leq \mathbb{E}\left[\frac{\tilde{\alpha}_k^2(1+\sigma^2)}{2\tilde{\gamma}_k}\|\nabla f(x_{k+1})\|^2 + \frac{\tilde{\gamma}_k}{2}\|v_k - v_{k+1}\|^2\right]
\end{aligned} \tag{C.18}$$

Substituting (B.11) and (C.18) back into (C.17), we can obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] - \mathcal{E}(z_k^+;\tilde{\gamma}_k) \\
&\leq \mathbb{E}\left[-\tilde{\alpha}_k D_f(x_{k+1}, x^*) - \tilde{\alpha}_k D_f(x^*, x_{k+1}) + \frac{1}{2}\left(\frac{\tilde{\alpha}_k^2(1+\sigma^2)}{\tilde{\gamma}_k} - \tilde{\alpha}_k\beta_k\right)\|\nabla f(x_{k+1})\|^2\right. \\
&\qquad \left. -\frac{\tilde{\alpha}_k\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 - D_f(x_k^+, x_{k+1})\right] \\
&\leq \mathbb{E}\left[-\tilde{\alpha}_k\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1}) - \tilde{\alpha}_k D_f(x^*, x_{k+1}) + \frac{1}{2}\left(\frac{\tilde{\alpha}_k^2(1+\sigma^2)}{\tilde{\gamma}_k} - (1+\tilde{\alpha}_k)\tilde{\alpha}_k\beta_k\right)\|\nabla f(x_{k+1})\|^2\right. \\
&\qquad \left. -D_f(x_k^+, x_{k+1})\right]
\end{aligned} \tag{C.19}$$

By moving $\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right]$ to the left side of the inequality to obtain the desired result. □

Now we prove the theorem 2.2.

Proof. Assume $\alpha_k = \frac{2}{k+1}$, $\gamma_k = \alpha_k\tilde{\alpha}_k(1+\sigma^2)^2L$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$. Then

$$\tilde{\alpha}_k\beta_k = \frac{(1+\sigma^2)\tilde{\alpha}_k\alpha_k}{\gamma_k} = \frac{(1+\sigma^2)\tilde{\alpha}_k^2}{\tilde{\gamma}_k} \tag{C.20}$$

Using Lemma C.2 to obtain

$$\mathbb{E}\left[\mathcal{E}(z_{k+1}^+;\tilde{\gamma}_{k+1})\right] \le (1+\tilde{\alpha}_k)^{-1}\mathcal{E}(z_k^+;\tilde{\gamma}_k) \le \Pi_{i=0}^k(1+\tilde{\alpha}_i)^{-1}\mathcal{E}(z_0^+;\tilde{\gamma}_0) \tag{C.21}$$

Since $\tilde{\alpha}_k = \frac{2}{k+1+2m}$, then

$$\Pi_{i=0}^k(1+\tilde{\alpha}_i)^{-1} = \Pi_{i=0}^k \frac{i+1+2m}{i+3+2m} = \frac{(1+2m)(2+2m)}{(k+3+2m)(k+2+2m)}$$

□

Corollary C.2. Under the setting of Theorem 2.2, choose $m \ge 0$, $\alpha_k = \frac{2}{k+1}$, $\tilde{\alpha}_k = \frac{\alpha_k}{1+m\alpha_k}$, $\gamma_k = \alpha_k\tilde{\alpha}_k(1+\sigma^2)^2L$ and $\beta_k = \frac{(1+\sigma^2)\alpha_k}{\gamma_k}$, SHANG++ guarantees to reach an ε-precision at the following interations:

$$k \ge \sqrt{(1+2m)(2+2m)(f(x_0)-f(x^*)+\frac{2(1+\sigma^2)^2L}{(1+2m)^2}\|x_0-x^*\|^2)/\varepsilon}$$

Corollary C.3. Under the setting of Theorem 2.2, $f(x_k^+) \overset{a.s.}{\to} f(x^*)$.

D Variance Decay Analysis

We study the variance decay of the Lyapunov energy (B.2)

$$\mathcal{E}_k := \mathcal{E}(z_k^+;\tilde{\gamma}_k) = f(x_k^+) - f(x^*) + \frac{\tilde{\gamma}_k}{2}\|v_k - x^*\|^2$$

under the unified stochastic model of SHANG and SHANG++. Throughout we work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the post-update filtration $\mathcal{F}_k := \sigma(x_0, v_0, \zeta_0, \dots, \zeta_k)$, where each ζ_k collects the randomness used to form the stochastic gradient at step k. We write $g_k := g(x_k, \zeta_k)$ and $g_{k+1} := g(x_{k+1}, \zeta_{k+1})$.

Assumptions. We make the following standard assumptions.

- A1. Smooth convexity. $f \in \mathcal{S}_{\mu,L}$ with $0 \le \mu < L < \infty$.
- A2. Unbiasedness at the query point. $\mathbb{E}[g_{k+1} \mid \mathcal{F}_k] = \nabla f(x_{k+1})$. Equivalently, with $\xi_{k+1} := g_{k+1} - \nabla f(x_{k+1})$, $\mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0$.
- A3. Multiplicative noise scaling (MNS). $\mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k] \le \sigma^2\|\nabla f(x_{k+1})\|^2$.
- A4. Bounded conditional kurtosis. There exists $\chi \ge 1$ such that $\mathbb{E}[\|\xi_{k+1}\|^4 \mid \mathcal{F}_k] \le \chi\big(\mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k]\big)^2$ (e.g., $\chi = 3$ for Gaussian noise).

Unified stochastic model. The updates for SHANG/SHANG++ can be written as

$$\begin{aligned} x_k^+ &= x_k - \tilde{\alpha}_k\beta_k\, g_k \\ \frac{x_{k+1}-x_k^+}{\tilde{\alpha}_k} &= v_k - x_{k+1} \\ \frac{v_{k+1}-v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(x_{k+1}-v_{k+1}) - \frac{1}{\gamma_k}g_{k+1} \\ \frac{\gamma_{k+1}-\gamma_k}{\alpha_k} &= \mu - \gamma_{k+1}. \end{aligned} \tag{D.1}$$

where $\alpha_k > 0$, $\gamma_k > 0$, and we introduce $\tilde{\alpha}_k = \frac{\alpha_k}{1+m\alpha_k}$ and $\tilde{\gamma}_k = \frac{\gamma_k}{1+m\alpha_k}$ with $m \ge 0$. Equivalently (and crucial for variance analysis), (x_{k+1}^+, v_{k+1}) are affine in the fresh gradient g_{k+1} while x_{k+1} depends only on past randomness:

$$\begin{aligned} x_{k+1}^+ &= \frac{1}{1+\tilde{\alpha}_k}x_k^+ + \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}v_k - \tilde{\alpha}_{k+1}\beta_{k+1}g_{k+1} \;=\; x_{k+1} - \tilde{\alpha}_{k+1}\beta_{k+1}g_{k+1}, \\ v_{k+1} &= \frac{\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}x_k^+ + \Big(\frac{\gamma_k}{\gamma_k+\alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}\Big)v_k - \frac{\alpha_k}{\gamma_k+\alpha_k\mu}g_{k+1} \\ \gamma_{k+1} &= \frac{\alpha_k}{1+\alpha_k}\mu + \frac{1}{1+\alpha_k}\gamma_k \end{aligned} \tag{D.2}$$

By the filtration choice, x_{k+1} is $\mathcal{F}_k$-measurable and g_{k+1} uses fresh randomness ζ_{k+1}; hence with $\xi_{k+1} := g_{k+1} - \nabla f(x_{k+1})$ we have $\mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0$. This linear structure will allow us to bound the one-step fluctuation $\mathcal{E}_{k+1} - \mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k]$ and to propagate variance.

Lemma D.1 (One-step fluctuation). There exist explicit constants $A_k, B_k, C_k \geq 0$ (functions of $\alpha_k, \tilde{\alpha}_k, \tilde{\gamma}_k, \mu, L$) such that, with ξ_{k+1},

$$\left|\mathcal{E}_{k+1} - \mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k]\right| \leq A_k\sqrt{\mathcal{E}_k}\|\xi_{k+1}\| + B_k\|\xi_{k+1}\|^2 + C_k\mathcal{E}_k$$

and

$$\begin{aligned}
A_k &= \left(B_x(1+B_xL)\sqrt{2Lc_1} + B_v\tilde{\gamma}_{k+1}(c_2 + B_v\sqrt{2Lc_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L)})\right) \\
B_k &= \frac{LB_x^2 + \tilde{\gamma}_{k+1}B_v^2}{2} \\
C_k &= (LB_x^2 + \tilde{\gamma}_{k+1}B_v^2)Lc_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L)\sigma^2
\end{aligned}$$

where $c_1 = \max\{\frac{1}{1+\tilde{\alpha}_k}, \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}\frac{L}{\tilde{\gamma}_k}\}$, $c_2 = \max\{\frac{\alpha_k\sqrt{2\mu}}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}, \left(\frac{\gamma_k}{\gamma_k+\alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}\right)\sqrt{\frac{2}{\tilde{\gamma}_k}}\}$ when $\mu > 0$ and $c_2 = \sqrt{\frac{2}{\tilde{\gamma}_k}}$ when $\mu = 0$. $B_x = \tilde{\alpha}_{k+1}\beta_{k+1}$ and $B_v = \frac{\alpha_k}{\gamma_k+\alpha_k\mu}$.

proof of Lemma D.1. Using $\xi_{k+1} := g_{k+1} - \nabla f(x_{k+1})$, we can rewrite the updates of x_{k+1}^+ and v_{k+1} as

$$\begin{aligned}
x_{k+1}^+ &= U_k - \tilde{\alpha}_{k+1}\beta_{k+1}\nabla f(x_{k+1}) - \tilde{\alpha}_{k+1}\beta_{k+1}\xi_{k+1} = \hat{U}_k - B_x\xi_{k+1} \\
v_{k+1} &= V_k - \frac{\alpha_k}{\gamma_k+\alpha_k\mu}\nabla f(x_{k+1}) - \frac{\alpha_k}{\gamma_k+\alpha_k\mu}\xi_{k+1} = \hat{V}_k - B_v\xi_{k+1}
\end{aligned} \tag{D.3}$$

where $U_k = \frac{1}{1+\tilde{\alpha}_k}x_k^+ + \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}v_k$, $V_k = \frac{\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}x_k^+ + \left(\frac{\gamma_k}{\gamma_k+\alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}\right)v_k$, $\hat{U}_k = U_k - B_x\nabla f(x_{k+1})$ and $\hat{V}_k = V_k - B_v\nabla f(x_{k+1})$. $B_x = \tilde{\alpha}_{k+1}\beta_{k+1}$ and $B_v = \frac{\alpha_k}{\gamma_k+\alpha_k\mu}$ are positive constants. It should be noted that $U_k, \hat{U}_k, V_k$ and $\hat{V}_k$ are measurable with respect to $\mathcal{F}_k$.

Let's first focus on the left part of $\mathcal{E}_{k+1}$. Expanding $f(x_{k+1}^+) = f(\hat{U}_k - B_x\xi_{k+1})$ at point $\hat{U}_k$ using Taylor series gives

$$f(\hat{U}_k - B_x\xi_{k+1}) = f(\hat{U}_k) - \langle\nabla f(\hat{U}_k), B_x\xi_{k+1}\rangle + r(\hat{U}_k, \xi_{k+1}) \tag{D.4}$$

where

$$|\, r(\hat{U}_k, \xi_{k+1})\,| = \left|\int_0^1 \langle\nabla f(\hat{U}_k - tB_x\xi_{k+1}) - \nabla f(\hat{U}_k), -B_x\xi_{k+1}\rangle dt\right| \leq \frac{L}{2}\|B_x\xi_{k+1}\|^2 = \frac{LB_x^2}{2}\|\xi_{k+1}\|^2 \tag{D.5}$$

Then

$$\begin{aligned}
&|\, f(x_{k+1}^+) - f(x^*) - \mathbb{E}\left[f(x_{k+1}^+) - f(x^*) \mid \mathcal{F}_k\right]\,| \\
=&|\, f(\hat{U}_k - B_x\xi_{k+1}) - f(x^*) - \mathbb{E}\left[f(\hat{U}_k - B_x\xi_{k+1}) - f(x^*) \mid \mathcal{F}_k\right]\,| \\
=&|\, -\langle\nabla f(\hat{U}_k), B_x\xi_{k+1}\rangle + r(\hat{U}_k, \xi_{k+1}) - \mathbb{E}\left[r(\hat{U}_k, \xi_{k+1}) \mid \mathcal{F}_k\right]\,| \\
\leq\ & B_x\|\nabla f(\hat{U}_k)\| \cdot \|\xi_{k+1}\| + \frac{LB_x^2}{2}\|\xi_{k+1}\|^2 + \frac{LB_x^2}{2}\mathbb{E}\left[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\right]
\end{aligned} \tag{D.6}$$

where the last step uses Cauchy-Schwarz inequality and (D.5).

Since $\hat{U}_k = U_k - B_x\nabla f(x_{k+1}) = x_{k+1} - B_x\nabla f(x_{k+1})$ and $x_{k+1} = \frac{1}{1+\tilde{\alpha}_k}x_k^+ + \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}v_k$, by triangle inequality and smooth convexity of f, we have

$$\begin{aligned}
\|\nabla f(\hat{U}_k)\| &\le \|\nabla f(\hat{U}_k) - \nabla f(x_{k+1})\| + \|\nabla f(x_{k+1})\| \\
&\le L\|\hat{U}_k - x_{k+1}\| + \|\nabla f(x_{k+1})\| \\
&= (1 + B_xL)\|\nabla f(x_{k+1})\| \\
&\le (1 + B_xL)\sqrt{2L}\sqrt{f(x_{k+1}) - f(x^*)} \\
&\le (1 + B_xL)\sqrt{2L}\sqrt{\frac{1}{1+\tilde{\alpha}_k}(f(x_k^+) - f(x^*)) + \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}(f(v_k) - f(x^*))} \\
&\le (1 + B_xL)\sqrt{2L}\sqrt{\frac{1}{1+\tilde{\alpha}_k}(f(x_k^+) - f(x^*)) + \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}\frac{L}{2}\|v_k - x^*\|^2} \\
&\le (1 + B_xL)\sqrt{2Lc_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L)}\sqrt{\mathcal{E}_k}
\end{aligned} \tag{D.7}$$

where $c_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L) = \max\{\frac{1}{1+\tilde{\alpha}_k}, \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}\frac{L}{\tilde{\gamma}_k}\}$.

On the other hand,

$$\frac{LB_x^2}{2}\mathbb{E}\left[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\right] \le \frac{LB_x^2\sigma^2}{2}\|\nabla f(x_{k+1})\|^2 \le L^2B_x^2\sigma^2c_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L)\mathcal{E}_k \tag{D.8}$$

Substituting (D.7) and (D.8) back into (D.6), we have

$$\begin{aligned}
&| f(x_{k+1}^+) - f(x^*) - \mathbb{E}\left[f(x_{k+1}^+) - f(x^*) \mid \mathcal{F}_k\right] | \\
&\le B_x(1 + B_xL)\sqrt{2Lc_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L)}\sqrt{\mathcal{E}_k}\|\xi_{k+1}\| + \frac{LB_x^2}{2}\|\xi_{k+1}\|^2 + L^2B_x^2\sigma^2c_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L))\mathcal{E}_k
\end{aligned} \tag{D.9}$$

For the middle part of $\mathcal{E}_{k+1}$, since

$$\frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1} - x^*\|^2 = \frac{\tilde{\gamma}_{k+1}}{2}\|\hat{V}_k - x^*\|^2 + \frac{\tilde{\gamma}_{k+1}B_v^2}{2}\|\xi_{k+1}\|^2 - \tilde{\gamma}_{k+1}\langle\hat{V}_k - x^*, B_v\xi_{k+1}\rangle, \tag{D.10}$$

we have

$$\begin{aligned}
&| \frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1} - x^*\|^2 - \mathbb{E}\left[\frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] | \\
&=| -\tilde{\gamma}_{k+1}\langle\hat{V}_k - x^*, B_v\xi_{k+1}\rangle + \frac{\tilde{\gamma}_{k+1}B_v^2}{2}\big(\|\xi_{k+1}\|^2 - \mathbb{E}\left[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\right]\big) | \\
&\le B_v\tilde{\gamma}_{k+1}\|\hat{V}_k - x^*\| \cdot \|\xi_{k+1}\| + \frac{\tilde{\gamma}_{k+1}B_v^2}{2}\|\xi_{k+1}\|^2 + \frac{\tilde{\gamma}_{k+1}B_v^2}{2}\mathbb{E}\left[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\right]
\end{aligned} \tag{D.11}$$

Using triangle inequality and convexity of $\|\cdot\|$, we have

$$\begin{aligned}
&\|\hat{V}_k - x^*\| \\
&= \|V_k - x^* - B_v\nabla f(x_{k+1})\| \\
&\le \|\frac{\alpha_k\mu}{(\gamma_k + \alpha_k\mu)(1+\tilde{\alpha}_k)}x_k^+ + \Big(\frac{\gamma_k}{\gamma_k + \alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k + \alpha_k\mu)(1+\tilde{\alpha}_k)}\Big)v_k - x^*\| + B_v\|\nabla f(x_{k+1})\| \\
&\le \frac{\alpha_k\mu}{(\gamma_k + \alpha_k\mu)(1+\tilde{\alpha}_k)}\|x_k^+ - x^*\| + \Big(\frac{\gamma_k}{\gamma_k + \alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k + \alpha_k\mu)(1+\tilde{\alpha}_k)}\Big)\|v_k - x^*\| + B_v\|\nabla f(x_{k+1})\| \\
&\le \frac{\alpha_k\mu}{(\gamma_k + \alpha_k\mu)(1+\tilde{\alpha}_k)}\|x_k^+ - x^*\| + \Big(\frac{\gamma_k}{\gamma_k + \alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k + \alpha_k\mu)(1+\tilde{\alpha}_k)}\Big)\|v_k - x^*\| \\
&\quad + B_v\sqrt{2Lc_1(\tilde{\alpha}_k, \tilde{\gamma}_k, L)}\sqrt{\mathcal{E}_k}
\end{aligned} \tag{D.12}$$

Next, we will consider two cases.

Case 1: $\mu > 0$. Using the strong convexity of f, we have

$$\begin{aligned}
&\|\hat{V}_k - x^*\| \\
&\leq \frac{\alpha_k\sqrt{2\mu}}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}\sqrt{f(x_k^+)-f(x^*)} + \Big(\frac{\gamma_k}{\gamma_k+\alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}\Big)\sqrt{\frac{2}{\tilde{\gamma}_k}}\sqrt{\frac{\tilde{\gamma}_k}{2}}\|v_k - x^*\| \\
&\quad + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)}\sqrt{\mathcal{E}_k} \\
&\leq \big(c_2(\tilde{\alpha},\mu,\gamma_k) + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)}\big)\sqrt{\mathcal{E}_k}
\end{aligned} \tag{D.13}$$

where $c_2(\tilde{\alpha},\mu,\gamma_k) = \max\{\frac{\alpha_k\sqrt{2\mu}}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}, \Big(\frac{\gamma_k}{\gamma_k+\alpha_k\mu} + \frac{\tilde{\alpha}_k\alpha_k\mu}{(\gamma_k+\alpha_k\mu)(1+\tilde{\alpha}_k)}\Big)\sqrt{\frac{2}{\tilde{\gamma}_k}}\}$. Thus,

$$\begin{aligned}
&|\ \frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 - \mathbb{E}\left[\frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right]\ | \\
&\leq B_v\tilde{\gamma}_{k+1}\big(c_2(\tilde{\alpha},\mu,\gamma_k) + B_v\sqrt{2Lc_1(\tilde{\alpha},\mu,L)}\big)\sqrt{\mathcal{E}_k}\|\xi_{k+1}\| + \frac{\tilde{\gamma}_{k+1}B_v^2}{2}\|\xi_{k+1}\|^2 + \tilde{\gamma}_{k+1}B_v^2L\sigma^2c_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)\mathcal{E}_k
\end{aligned} \tag{D.14}$$

Combining (D.9) and (D.14), we have

$$\begin{aligned}
&|\ \mathcal{E}_{k+1} - \mathbb{E}\left[\mathcal{E}_{k+1} \mid \mathcal{F}_k\right]\ | \\
&\leq \big(B_x(1+B_xL)\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)} + B_v\tilde{\gamma}_{k+1}(c_2(\tilde{\alpha},\mu,\gamma_k) + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)})\big)\sqrt{\mathcal{E}_k}\|\xi_{k+1}\| \\
&\quad + \frac{LB_x^2+\tilde{\gamma}_{k+1}B_v^2}{2}\|\xi_{k+1}\|^2 + (LB_x^2+\tilde{\gamma}_{k+1}B_v^2)Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)\sigma^2\mathcal{E}_k
\end{aligned} \tag{D.15}$$

Case 2: $\mu = 0$.

$$\begin{aligned}
\|\hat{V}_k - x^*\| &\leq \|v_k - x^*\| + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)}\sqrt{\mathcal{E}_k} \\
&\leq (\sqrt{\frac{2}{\tilde{\gamma}_k}} + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)})\sqrt{\mathcal{E}_k}
\end{aligned} \tag{D.16}$$

Thus,

$$\begin{aligned}
&|\ \frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 - \mathbb{E}\left[\frac{\tilde{\gamma}_{k+1}}{2}\|v_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right]\ | \\
&\leq B_v\tilde{\gamma}_{k+1}(\sqrt{\frac{2}{\tilde{\gamma}_k}} + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)})\sqrt{\mathcal{E}_k}\|\xi_{k+1}\| + \frac{\tilde{\gamma}_{k+1}B_v^2}{2}\|\xi_{k+1}\|^2 + \tilde{\gamma}_{k+1}B_v^2L\sigma^2c_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)\mathcal{E}_k
\end{aligned} \tag{D.17}$$

Combining (D.9) and (D.17), we have

$$\begin{aligned}
&|\ \mathcal{E}_{k+1} - \mathbb{E}\left[\mathcal{E}_{k+1} \mid \mathcal{F}_k\right]\ | \\
&\leq \big(B_x(1+B_xL)\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)} + B_v\tilde{\gamma}_{k+1}(\sqrt{\frac{2}{\tilde{\gamma}_k}} + B_v\sqrt{2Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)})\big)\sqrt{\mathcal{E}_k}\|\xi_{k+1}\| \\
&\quad + \frac{LB_x^2+\tilde{\gamma}_{k+1}B_v^2}{2}\|\xi_{k+1}\|^2 + (LB_x^2+\tilde{\gamma}_{k+1}B_v^2)Lc_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)\sigma^2\mathcal{E}_k
\end{aligned} \tag{D.18}$$

□

Proposition D.1 (Conditional variance bound). Let $S_k := 2L\sigma^2c_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L)$ with $c_1(\tilde{\alpha}_k,\tilde{\gamma}_k,L) = \max\{\frac{1}{1+\tilde{\alpha}_k}, \frac{\tilde{\alpha}_k}{1+\tilde{\alpha}_k}\frac{L}{\tilde{\gamma}_k}\}$. Under assumptions (A2)–(A4) and the setting of Lemma D.1 (In particular, stepsizes and hence A_k, B_k, C_k, S_k are $\mathcal{F}_k$-measurable),

$$\mathrm{Var}(\mathcal{E}_{k+1} \mid \mathcal{F}_k) \leq K_{2,k}\mathcal{E}_k^2, \qquad K_{2,k} = 3\big(A_k^2S_k + \chi B_k^2S_k^2 + C_k^2\big)$$

proof of Proposition D.1. By the definition of conditional variance,

$$\mathrm{Var}(\mathcal{E}_{k+1} \mid \mathcal{F}_k) = \mathbb{E}\big[\left(\mathcal{E}_{k+1} - \mathbb{E}\left[\mathcal{E}_{k+1} \mid \mathcal{F}_k\right]\right)^2 \mid \mathcal{F}_k\big] \tag{D.19}$$

From Lemma D.1 and inequality $(x+y+z)^2 \le 3(x^2+y^2+z^2)$,

$$(\mathcal{E}_{k+1} - \mathbb{E}\left[\mathcal{E}_{k+1} \mid \mathcal{F}_k\right])^2 \le 3\big(A_k^2\mathcal{E}_k\|\xi_{k+1}\|^2 + B_k^2\|\xi_{k+1}\|^4 + C_k^2\mathcal{E}_k^2\big) \tag{D.20}$$

Since A_k, B_k, C_k and $\mathcal{E}_k$ are all measurable with respect to the σ-algebra $\mathcal{F}_k$. Using assumptions (A2-A4) yields

$$\mathbb{E}\left[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\right] \le \sigma^2\|\nabla f(x_{k+1})\|^2 \le 2L\sigma^2 c_1\mathcal{E}_k = S_k\mathcal{E}_k \tag{D.21}$$

and

$$\mathbb{E}\left[\|\xi_{k+1}\|^4 \mid \mathcal{F}_k\right] \le \chi\big(\mathbb{E}\left[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\right]\big)^2 \le \chi S_k^2\mathcal{E}_k^2 \tag{D.22}$$

Taking $\mathbb{E}\left[\cdot \mid \mathcal{F}_k\right]$ in the previous inequality gives

$$\mathrm{Var}(\mathcal{E}_{k+1} \mid \mathcal{F}_k) \le 3(A_k^2 S_k + \chi B_k^2 S_k^2 + C_k^2)\mathcal{E}_k^2 \tag{D.23}$$

□

Theorem D.1 (Geometric variance decay). Assume the drift inequality (from the expectation analysis)

$$\mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k] \le q\mathcal{E}_k \qquad \text{for some } q \in (0,1), \tag{D.24}$$

and assumptions (A2)–(A4) hold. Let $K_{2,k}$ be given in Proposition D.1 and suppose $K_2 := \sup_k K_{2,k} < 1 - q^2$ satisfied. Then with $\theta := q^2 + K_2 \in (0,1)$, for all $k \ge 0$, given initial $\mathcal{E}_0$,

$$\mathrm{Var}(\mathcal{E}_{k+1}) \le \mathcal{E}_0^2\theta^{k+1}$$

Proof. By the law of total variance and Proposition D.1,

$$\mathrm{Var}(\mathcal{E}_{k+1}) = \mathbb{E}\big[Var(\mathcal{E}_{k+1} \mid \mathcal{F}_k)\big] + \mathrm{Var}\left(\mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k]\right) \le K_2\mathbb{E}[\mathcal{E}_k^2] + q^2\,\mathrm{Var}(\mathcal{E}_k). \tag{D.25}$$

Since $\mathbb{E}[\mathcal{E}_k^2] = \mathrm{Var}(\mathcal{E}_k) + \big(\mathbb{E}[\mathcal{E}_k]\big)^2$ and (Eq.(D.24)), we get

$$\mathrm{Var}(\mathcal{E}_{k+1}) \le (K_2 + q^2)\,\mathrm{Var}(\mathcal{E}_k) + K_2\big(\mathbb{E}[\mathcal{E}_k]\big)^2 \le (K_2 + q^2)\,\mathrm{Var}(\mathcal{E}_k) + K_2(\mathbb{E}[\mathcal{E}_0])^2 q^{2k} \tag{D.26}$$

Solving this linear recursion yields

$$\mathrm{Var}(\mathcal{E}_{k+1}) \le (K_2+q^2)^{k+1}\,\mathrm{Var}(\mathcal{E}_0) + K_2(\mathbb{E}[\mathcal{E}_0])^2\sum_{j=0}^{k}(K_2+q^2)^{k-j}q^{2j} \le (K_2+q^2)^{k+1}(\mathrm{Var}(\mathcal{E}_0)+(\mathbb{E}[\mathcal{E}_0])^2) \tag{D.27}$$

Since $\mathcal{E}_0$ is given by the initial point $x_0 = v_0$, it is a constant ,then $\mathrm{Var}(\mathcal{E}_0) = 0$ and $\mathbb{E}[\mathcal{E}_0] = \mathcal{E}_0$. □

Corollary D.1 (Upper bound of $K_{2,k}$ in strongly convex setting). Define $\kappa = \frac{L}{\mu}$ is the condition number of f. Under the setting of Theorem B.1-2.1 and Assumptions (A1)-(A4), with $K_{2,k} = 3(A_k^2 S_k + \chi B_k^2 S_k^2 + C_k^2)$ defined above, we have the explicit upper bound

(1) For SHANG,

$$K_2 \le \begin{cases} 12a_0^2\sigma^2((3+\sigma^2)a_0+1)^2 + \ 12(\chi+1)a_0^4\sigma^4 & \alpha\kappa \le 1 \\ 12a_0^3\sigma^2\sqrt{\kappa}(1+(3+\sigma^2)a_0^{\frac{3}{2}}\kappa^{\frac{1}{4}})^2 + 12(\chi+1)\,a_0^6\,\sigma^4\,\kappa & \alpha\kappa \ge 1 \end{cases}$$

where

(2) For SHANG++,

$$K_2 \le \begin{cases} 12a_0^2\sigma^2\dfrac{\kappa}{\kappa-1}(1+(3+\sigma^2)a_0\sqrt{\dfrac{\kappa}{\kappa-1}})^2 + 12(\chi+1)a_0^4\sigma^4(\dfrac{\kappa}{\kappa-1})^2 & \alpha\kappa \le 1+(1+\sigma^2)\alpha^2 \\ 12a_0^3\sigma^2\dfrac{\kappa^2}{(\kappa-1)^{\frac{3}{2}}}(1+(3+\sigma^2)a_0^{\frac{3}{2}}\dfrac{\kappa}{(\kappa-1)^{\frac{3}{4}}})^2 + 12(\chi+1)a_0^6\sigma^4\kappa(\dfrac{\kappa}{\kappa-1})^3 & \alpha\kappa \ge 1+(1+\sigma^2)\alpha^2 \end{cases}$$

Proof. Case 1: SHANG. When $m = 0$, scheme (D.1) is algorithm SHANG. From Theorem B.1, when $\gamma = \mu$, $\alpha = \frac{1}{(1+\sigma^2)\sqrt{\kappa}}$ and $\beta = \frac{(1+\sigma^2)\alpha}{\mu}$, we have

$$\mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k] \le (1+\alpha)^{-1}\mathcal{E}_k = q\mathcal{E}_k \tag{D.28}$$

and

$$\begin{aligned}
A = A_k &= \frac{\alpha^2}{\mu}\big(1+\sigma^2+(1+\sigma^2)^2\alpha^2\kappa+\frac{1}{(1+\alpha)^2}\big)\sqrt{2Lc_1}+\frac{\alpha}{1+\alpha}c_2\\
B = B_k &= \frac{\alpha^2}{2\mu}((1+\sigma^2)^2\alpha^2\kappa+\frac{1}{(1+\alpha)^2})\\
C = C_k &= \frac{\alpha^2}{\mu}((1+\sigma^2)^2\alpha^2\kappa+\frac{1}{(1+\alpha)^2})L\sigma^2c_1\\
S = S_k &= 2L\sigma^2c_1
\end{aligned}$$

where $c_1 = \max\{\frac{1}{1+\alpha}, \frac{\alpha}{1+\alpha}\kappa\}$ and $c_2 = \frac{1+\alpha+\alpha^2}{(1+\alpha)^2}\sqrt{\frac{2}{\mu}}$.

(1): Assume $\alpha\kappa \le 1$, i.e., $\kappa \le (1+\sigma^2)^2$, so that $c_1 = \frac{1}{1+\alpha}$.

Since $c_1 = \frac{1}{1+\alpha}$ and $\alpha^2\kappa = \frac{1}{(1+\sigma^2)^2} \le 1$, we bound each term in K_2.

For the B^2S^2 term, using $B = \frac{\alpha^2}{2\mu}((1+\sigma^2)^2\alpha^2\kappa + \frac{1}{(1+\alpha)^2})$,

$$\begin{aligned}
B^2S^2 &= \Big[\frac{\alpha^2}{2\mu}\Big((1+\sigma^2)^2\alpha^2\kappa+\tfrac{1}{(1+\alpha)^2}\Big)\Big]^2\cdot(2\mu\kappa\sigma^2c_1)^2\\
&= \alpha^4\kappa^2\sigma^4c_1^2\Big((1+\sigma^2)^2\alpha^2\kappa+\tfrac{1}{(1+\alpha)^2}\Big)^2 \;\le\; 4\,a_0^4\,\sigma^4,
\end{aligned} \tag{D.29}$$

where we used $c_1 \le 1$ and $\alpha^4\kappa^2 = \frac{1}{(1+\sigma^2)^4}$. We denote $a_0 = \frac{1}{1+\sigma^2}$. Hence $3\chi B^2S^2 \le 12\chi a_0^4\sigma^4$.

For the C^2 term, note $C = 2BLc_1\sigma^2$ implies $C^2 = B^2S^2$. Hence

$$C^2 \le 4\,a_0^4\,\sigma^4, \tag{D.30}$$

so $3C^2 \le 12\,a_0^4\,\sigma^4$.

For the A^2S term, splitting $A = A_1 + A_2$ with

$$A_1 := \frac{\alpha^2}{\mu}\Big((1+\sigma^2)+(1+\sigma^2)^2\alpha^2\kappa+\tfrac{1}{(1+\alpha)^2}\Big)\sqrt{2Lc_1}, \qquad A_2 := \frac{\alpha}{1+\alpha}\,c_2,$$

For A_2, since $c_2 = \frac{1+\alpha+\alpha^2}{(1+\alpha)^2}\sqrt{2/\mu} \le \sqrt{2/\mu}$,

$$A_2^2S = \frac{\alpha^2}{(1+\alpha)^2}\,c_2^2\cdot 2\mu\kappa\sigma^2c_1 \le 4\,\kappa\sigma^2c_1\cdot\frac{\alpha^2}{(1+\alpha)^2} = 4\,\sigma^2\cdot\frac{\alpha^2\kappa}{(1+\alpha)^3} \le 4a_0^2\sigma^2 \tag{D.31}$$

For A_1, using $c_1 = \frac{1}{1+\alpha}$ and $\alpha^2\kappa = a_0^2 \le 1$,

$$\begin{aligned}
A_1^2S &= \Big[\frac{\alpha^2}{\mu}\sqrt{2Lc_1}\Big]^2\Big((1+\sigma^2)+(1+\sigma^2)^2\alpha^2\kappa+\tfrac{1}{(1+\alpha)^2}\Big)^2\cdot 2L\sigma^2c_1\\
&= 4\,\alpha^4\kappa^2c_1^2\,\sigma^2\Big((1+\sigma^2)^2+1+\tfrac{1}{(1+\alpha)^2}\Big)^2 \;\le\; 4\,a_0^4\,\sigma^2\cdot(3+\sigma^2)^2 = 4(3+\sigma^2)^2a_0^4\sigma^2
\end{aligned} \tag{D.32}$$

Therefore, using $(x+y)^2 \le (1+\tau)x^2 + (1+1/\tau)y^2$ with $\tau = \sqrt{A_2^2S/A_1^2S}$:

$$3A^2S \le 3(\sqrt{A_1^2S}+\sqrt{A_2^2S})^2 \le 3(2(3+\sigma^2)a_0^2\sigma + 2a_0\sigma)^2 = 12a_0^2\sigma^2((3+\sigma^2)a_0+1)^2 \tag{D.33}$$

Combining (D.29)-(D.33), we have

$$K_2 \le 12a_0^2\sigma^2((3+\sigma^2)a_0+1)^2 + \;12(\chi+1)a_0^4\sigma^4 \tag{D.34}$$

(2): Assume $\alpha\kappa \geq 1$, i.e., $\kappa \geq (1+\sigma^2)^2$, so that $c_1 = \frac{\alpha}{1+\alpha}\kappa$.

For the B^2S^2 and C^2 terms. We have

$$B^2S^2 = \frac{\alpha^4\kappa^2\sigma^4}{(1+\alpha)^2}\Big((1+\sigma^2)^2\alpha^2\kappa + \tfrac{1}{(1+\alpha)^2}\Big)^2 \;\leq\; \frac{\alpha^6\kappa^4\sigma^4}{(1+\alpha)^2}\cdot 4 \;\leq\; 4\,a_0^6\,\sigma^4\,\kappa \tag{D.35}$$

Hence

$$3\big(\chi B^2S^2 + C^2\big) \;\leq\; 12(\chi+1)\,a_0^6\,\sigma^4\,\kappa \tag{D.36}$$

For the A^2S term,

$$A_2^2S = \frac{\alpha^2}{(1+\alpha)^2}c_2^2\cdot 2L\sigma^2c_1 \;\leq\; \frac{\alpha^2}{(1+\alpha)^2}\cdot\frac{2}{\mu}\cdot 2\mu\kappa\sigma^2c_1 = 4\,\sigma^2\,\frac{\alpha^3}{(1+\alpha)^3}\,\kappa^2 \;\leq\; 4\,a_0^3\,\sigma^2\sqrt{\kappa} \tag{D.37}$$

Moreover,

$$A_1^2S = \frac{4\,\alpha^6\sigma^2\,\kappa^4}{(1+\alpha)^2}\Big((1+\sigma^2)+(1+\sigma^2)^2\alpha^2\kappa+\tfrac{1}{(1+\alpha)^2}\Big)^2 \leq \frac{4\,\alpha^6\sigma^2\,\kappa^4}{(1+\alpha)^2}\cdot(3+\sigma^2)^2 \;\leq\; 4(3+\sigma^2)^2\,a_0^6\,\sigma^2\,\kappa. \tag{D.38}$$

Combining (D.37) and (D.38),

$$3A^2S \;\leq 3(\sqrt{A_1^2S}+\sqrt{A_2^2S})^2 \leq 12a_0^3\sigma^2\sqrt{\kappa}(1+(3+\sigma^2)a_0^{\frac{3}{2}}\kappa^{\frac{1}{4}})^2 \tag{D.39}$$

Adding (D.35) - (D.39) yields

$$K_2 \leq 12a_0^2\sigma^2\sqrt{\kappa}(1+(3+\sigma^2)a_0^{\frac{3}{2}}\kappa^{\frac{1}{4}})^2 + 12(\chi+1)\,a_0^6\,\sigma^4\,\kappa \tag{D.40}$$

Case 2: SHANG++. When $m = \beta\mu = (1+\sigma^2)\alpha$, scheme (D.1) is algorithm SHANG++. From Theorem 2.1, when $\gamma = \mu$, $\alpha = \frac{1}{(1+\sigma^2)\sqrt{\kappa-1}}$ and $\beta = \frac{(1+\sigma^2)\alpha}{\mu}$, we have

$$\mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k] \leq (1+\alpha+\alpha^2)^{-1}\mathcal{E}_k = q\mathcal{E}_k \tag{D.41}$$

and

$$\begin{aligned}
A = A_k &= \frac{\alpha^2}{\mu}\big(\frac{1+\sigma^2}{1+(1+\sigma^2)\alpha^2} + \frac{(1+\sigma^2)^2\alpha^2\kappa}{(1+(1+\sigma^2)\alpha^2)^2} + \frac{1}{(1+\alpha)^2}\big)\sqrt{2Lc_1} + \frac{\alpha}{1+\alpha}c_2\\
B = B_k &= \frac{\alpha^2}{2\mu}\big(\frac{(1+\sigma^2)^2\alpha^2\kappa}{(1+(1+\sigma^2)\alpha^2)^2} + \frac{1}{(1+\alpha)^2}\big)\\
C = C_k &= \frac{\alpha^2}{\mu}\big(\frac{(1+\sigma^2)^2\alpha^2\kappa}{(1+(1+\sigma^2)\alpha^2)^2} + \frac{1}{(1+\alpha)^2}\big)L\sigma^2c_1\\
S = S_k &= 2L\sigma^2c_1
\end{aligned}$$

where $c_1 = \max\{\frac{1+(1+\sigma^2)\alpha^2}{1+\alpha+(1+\sigma^2)\alpha^2}, \frac{\alpha}{1+\alpha+(1+\sigma^2)\alpha^2}\kappa\}$ and $c_2 = \frac{1+\alpha+(2+\sigma^2)\alpha^2}{(1+\alpha)(1+\alpha+(1+\sigma^2)\alpha^2)}\sqrt{\frac{2}{\mu}}$.

Similar to the derivation of SHANG, we have

(1): Assume $\alpha\kappa \leq 1+(1+\sigma^2)\alpha^2$.

$$K_2 \leq 12a_0^2\sigma^2\frac{\kappa}{\kappa-1}(1+(3+\sigma^2)a_0\sqrt{\frac{\kappa}{\kappa-1}})^2 + 12(\chi+1)a_0^4\sigma^4(\frac{\kappa}{\kappa-1})^2 \tag{D.42}$$

(2): Assume $\alpha\kappa \geq 1+(1+\sigma^2)\alpha^2$.

$$K_2 \leq 12a_0^3\sigma^2\frac{\kappa^2}{(\kappa-1)^{\frac{3}{2}}}(1+(3+\sigma^2)a_0^{\frac{3}{2}}\frac{\kappa}{(\kappa-1)^{\frac{3}{4}}})^2 + 12(\chi+1)a_0^6\sigma^4\kappa(\frac{\kappa}{\kappa-1})^3 \tag{D.43}$$

□

When does variance decay hold? By Theorem D.1, geometric variance decay

$$\mathrm{Var}(\mathcal{E}_k) \;\le\; \mathcal{E}_0^2(q^2+K_2)^k$$

follows whenever $K_2 < 1-q^2$, where $q=(1+\alpha)^{-1}$ for SHANG and $q=(1+\alpha+\alpha^2)^{-1}$ for SHANG++. The bounds in Corollary D.1 make this condition directly checkable as a function of the condition number $\kappa = L/\mu$, the noise level σ^2 via $a_0=(1+\sigma^2)^{-1}$, and the stepsize α:

- In the low-condition regime (the branch with smaller c_1), K_2 scales like $\mathcal{O}\big(a_0^2\sigma^2\big) + \mathcal{O}\big(a_0^4\sigma^4\big)$ (for SHANG++ with a mild factor $(\kappa/(\kappa-1))^{\text{powers}}$), whereas $1-q^2 = \Theta(\alpha) = \Theta(a_0/\sqrt{\kappa})$.
- In the high-condition regime (the branch with larger c_1), the leading term is $K_2 = \mathcal{O}\big(a_0^3\sigma^2\sqrt{\kappa}\big) + \mathcal{O}\big(a_0^6\sigma^4\kappa\big)$ (again with the expected $(\kappa/(\kappa-1))$ corrections for SHANG++), while still $1-q^2=\Theta(a_0/\sqrt{\kappa})$.

Thus, for fixed κ, smaller noise (larger a_0) and moderate stepsizes make $K_2 < 1-q^2$ easier to satisfy; for large κ, the $O(\sqrt{\kappa})$ factor in the leading term of K_2 becomes the main bottleneck.

How to enforce the condition in practice. Two standard knobs guarantee $K_2 < 1-q^2$ without fine tuning:

1. Stepsize damping. Replace α by $\beta\,\alpha$ with $\beta \in (0,1]$. Then the leading term in K_2 scales like $\mathcal{O}(\beta^3)$, whereas $1-q^2$ scales like $\mathcal{O}(\beta)$ (both for SHANG and SHANG++); hence there exists $\beta_0 = \beta_0(\kappa,\sigma^2,\chi) \in (0,1]$ such that $K_2 < 1-q^2$ for all $\beta \le \beta_0$.
2. Mini-batching or averaging multiple independent estimates. Replacing σ^2 by σ^2/M reduces the leading term in K_2 by a factor $1/M$ while leaving $1-q^2$ essentially unchanged; the explicit constants in the corollary yield simple batch-size thresholds (e.g., $M \gtrsim \sigma^2\sqrt{\kappa}$ up to the displayed constants). Section 2 also notes that averaging multiple independent estimates does not incur additional computational costs.

E SNAG as a Discretization of the HNAG Flow

Under the multiplicative noise assumption, one of the most recent first-order stochastic methods designed to overcome the divergence of NAG and accelerate SGD is the Stochastic Nesterov Accelerated Gradient (SNAG) method (Hermant et al., 2025). Its iteration reads:

$$\begin{aligned} x_{k+1} &= \hat{\alpha}_{k+1}x_k + (1-\hat{\alpha}_{k+1})v_{k+1} - \hat{\alpha}_{k+1}s\,g(x_k),\\ v_{k+1} &= \hat{\beta}v_k + (1-\hat{\beta})x_k - \eta_k g(x_k), \end{aligned} \tag{E.1}$$

where $g(x_k)$ is a stochastic gradient estimator, and $\hat{\alpha}_{k+1}$, s, $\hat{\beta}$, and η_k are parameters.

By reparameterizing as

$$\hat{\alpha}_{k+1} = \frac{1}{1+\alpha_{k+1}}, \quad s = \alpha_{k+1}\beta_{k+1}, \quad \hat{\beta} = \frac{1}{1+\frac{\alpha_{k+1}\mu}{\gamma_{k+1}}}, \quad \eta_k = \frac{1}{1+\frac{\alpha_{k+1}\mu}{\gamma_{k+1}}}\frac{\alpha_{k+1}}{\gamma_{k+1}}, \tag{E.2}$$

the SNAG scheme (E.1) becomes equivalent to the following update:

$$\begin{aligned} \frac{x_{k+1}-x_k}{\alpha_{k+1}} &= v_{k+1} - x_{k+1} - \beta_{k+1}g(x_k),\\ \frac{v_{k+1}-v_k}{\alpha_{k+1}} &= \frac{\mu}{\gamma_{k+1}}(x_k - v_{k+1}) - \frac{1}{\gamma_{k+1}}g(x_k),\\ \frac{\gamma_{k+1}-\gamma_k}{\alpha_{k+1}} &\le \mu - \gamma_{k+1}. \end{aligned} \tag{E.3}$$

Hence, SNAG can be interpreted as a new discretization of the HNAG flow (2.2).

Parameter choices. For convex objectives $f \in \mathcal{S}^{1,1}_{0,L}$, Hermant et al. (2025) shows that the optimal parameters are

$$s = \frac{1}{L(1+\sigma^2)}, \quad \eta_k = \frac{k+1}{2L(1+\sigma^2)^2}, \quad \hat{\beta} = 1, \quad \hat{\alpha}_k = \frac{\frac{k^2}{k+1}}{2+\frac{k^2}{k+1}}.$$

This leads to

$$\alpha_{k+1} = \frac{2}{k+1-\frac{k+1}{k+2}}, \quad \alpha_{k+1}\beta_{k+1} = \frac{1}{L(1+\sigma^2)}, \quad \gamma_{k+1} = \alpha_{k+1}\tfrac{2}{k+1}(1+\sigma^2)^2 L.$$

For strongly convex objectives $f \in \mathcal{S}_{\mu,L}$, the optimal parameters become

$$s = \frac{1}{L(1+\sigma^2)}, \quad \eta_k = \eta = \frac{1}{(1+\sigma^2)\sqrt{\mu L}}, \quad \hat{\beta} = 1 - \frac{1}{1+\sigma^2}\sqrt{\tfrac{\mu}{L}}, \quad \hat{\alpha}_k = \hat{\alpha} = \frac{1}{1+\frac{1}{1+\sigma^2}\sqrt{\frac{\mu}{L}}}.$$

Consequently,

$$\alpha = \frac{1}{1+\sigma^2}\sqrt{\tfrac{\mu}{L}}, \quad \alpha\beta = \frac{1}{L(1+\sigma^2)}, \quad \gamma = \mu(1-\alpha).$$

The condition $\gamma = \mu(1-\alpha)$ indicates that, in the strongly convex case, the update for v is more accurately viewed as applying a rescaled step size $\tilde{\alpha} = \frac{\alpha}{1-\alpha}$ to the v–dynamics of the HNAG flow:

$$\frac{v_{k+1} - v_k}{\tilde{\alpha}} = x_k - v_{k+1} - \frac{1}{\mu} g(x_k).$$

In summary, the above parameter rearrangements confirm that the optimal choices in SNAG are consistent with those obtained from various discretization schemes of the HNAG flow, see Chen & Luo (2021) for details.