Adaptive Extragradient Methods for Root-finding Problems under Relaxed Assumptions

Yang Luo
UNC Chapel Hill

Michael J. O'Neill
UNC Chapel Hill

Abstract

We develop a new class of self-tuning algorithms to solve a root-finding problem involving a Lipschitz continuous operator, with applications in convex optimization, minimax saddle point problems and variational inequalities. Our methods are adaptive to the unknown, problem specific parameters, such as the Lipschitz constant and the variance of the stochastic operator. Unlike prior work, our approach does not rely on restrictive assumptions, such as a bounded domain, boundedness of the operator or a light-tailed distribution. We prove a $\tilde{\mathcal{O}}(N^{-1/2})$ average-iterate convergence rate of the restricted merit function under an affine noise assumption, matching the optimal rate up to log factors. In addition, we improve the convergence rate to $\mathcal{O}(N^{-1})$ under a strong growth condition, characterizing the field of cutting-edge machine learning models and matching the optimal rate for the *deterministic regime*. Finally, we illustrate the effectiveness of the proposed algorithms through numerical experiments on saddle point problems. Our results suggest that the adaptive step sizes automatically take advantage of the structure of the noise and observe improved convergence in certain settings, such as when the strong growth condition holds. To the best of our knowledge, this is the first method for root-finding problems under mild assumptions that adapts to the structure of the noise to obtain an improved convergence rate.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Root-finding problems encapsulate a wide array of optimization tasks including convex minimization, convex-concave saddle-point problems, and finding a Nash equilibrium in multi-player games Juditsky et al. (2011); Nemirovski (2004). Given a single valued operator $F : \mathbb{R}^p \to \mathbb{R}^p$, we are interested in finding x^* such that

$$F(x^*) = 0. \tag{1}$$

Throughout, we assume that the solution set S^* of (1) is nonempty. In convex minimization, F is simple the gradient of the function to be minimized, while for convex-concave saddle point problems

$$\min_{x\in\mathbb{R}^m} \max_{y\in\mathbb{R}^q} \quad \mathcal{L}(x, y), \tag{2}$$

the operator can be defined as

$$F(z) := (\nabla_x \mathcal{L}(x, y), -\nabla_y \mathcal{L}(x, y)),$$

where z denotes (x, y). Problems of this form have a long history of study, though interest has been reignited in recent years due to the advent of certain modern machine learning models. Of particular interest is training generative adversarial networks (GANs), for which the optimization task is formulated as a zero-sum game Gidel et al. (2019); Goodfellow et al. (2014). Other recent applications include robust adversarial reinforcement learning Pinto et al. (2017), actor-critic methods Pfau and Vinyals (2016), and adversarial example gamesBose et al. (2020).

The large-scale nature of modern machine learning applications pose unique challenges for solving (1), such as the need for efficient algorithms that only require stochastic access to F, many of which have been proposed in recent years Chavdarova et al. (2019),Diakonikolas et al. (2021),Gidel et al. (2019),Hsieh et al. (2020),Juditsky et al. (2011),Mertikopoulos et al. (2019),Yadav et al. (2018). However, these methods typically require very carefully chosen stepsize sequences based on unknown problem parameters such as the smoothness and noise level. In order to avoid this issue,

building upon the success of adaptive gradient methods for nonconvex minimization Duchi et al. (2011); McMahan and Streeter (2010), a number of works have proposed different adaptive gradient methods for (1), see Antonakopoulos et al. (2021); Bach and Levy (2019); Ene and Nguyen (2022); Hsieh et al. (2022); Liu et al. (2020). Unfortunately, the analysis of these methods rely upon restrictive assumptions, such as a uniform bound on the norm of F, which do not hold even in simple cases such as unconstrained, quadratic minimization and bilinear saddle point problems. In addition, none of these methods are known to adapt to the level of noise present, unlike adaptive gradient methods for unconstrained minimization, which exhibit superior convergence when the stochastic gradients have additional properties, such as obeying a strong growth condition Faw et al. (2022); Wang et al. (2023).

In this work, we develop a new adaptive gradient framework to solve root-finding problems involving a Lipschitz continuous operator which automatically adopts to the smoothness and noise level of the problem while eliminating the restrictive assumptions utilized in prior work.

1.1 Preliminaries

Throughout, we impose the following assumptions:

Assumption 1.1. *The operator F in (1) satisfies two out of the following three conditions:*

(a) Minty variational condition, i.e.

$$\langle F(x), x - x^* \rangle \geq 0, \quad \forall x^* \in \mathcal{S}^*, x \in \mathrm{dom}(F).$$

(a') Monotonicity, i.e.

$$\langle F(x) - F(y), x - y \rangle \geq 0, \quad \forall x, y \in \mathrm{dom}(F).$$

(b) L-Lipschitz continuity ($L > 0$), i.e.

$$\|F(x) - F(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathrm{dom}(F).$$

Remark 1.1. *Assumption (a) is a relaxation of monotonicity. It covers all monotone operators, but is also verified for a wide range of non-monotone problems, see references Facchinei and Pang (2003); Iusem et al. (2017); Kannan and Shanbhag (2019); Liu et al. (2020). Assumption (b) is essential for studying gradient-based methods.*

Connection to Variational Inequalities. An astute reader may notice that all of the example applications can be formulated as variational inequality problems due to the connection with root-finding problems. Consider the classical *Stampacchia* variational inequality problem (SVI), defined as:

$$\text{find } x^* \in \mathbb{R}^p, \text{ i.e., } \langle F(x^*), x - x^* \rangle \geq 0, \forall x \in \mathbb{R}^p. \quad (3)$$

The solution to (3) is considered as a strong solution to variational inequality problems. The weak (dual) formulation of (3) is defined as:

$$\text{find } x^* \in \mathbb{R}^p, \text{ i.e., } \langle F(x), x - x^* \rangle \geq 0, \forall x \in \mathbb{R}^p, \quad (4)$$

which is also known as *Minty* variational inequality problem. If F is monotone, then (3) and (4) are equivalent, leading to a root-finding problem.

Characterization of Solutions. Our goal is to design and analyze algorithms which generate approximate solutions of (1) within a tolerance $\epsilon \geq 0$. Specifically, we consider two optimality criteria. From the perspective of root-finding problems, we aim to find an ϵ-approximate solution, x_ϵ^*, defined as:

Definition 1.1. *Given a tolerance $\epsilon > 0$, we say that x_ϵ^* is an ϵ-solution of (1) if*

$$\|F(x_\epsilon^*)\| \leq \epsilon. \quad (5)$$

On the other hand, from the perspective of variational inequalities, (1) reduces to a Minty variational inequality problem (see (4)) under monotonicity of F. Following previous work Ene and Nguyen (2022); Hsieh et al. (2022); Nesterov (2007), when the operator F is monotone, we analyze the convergence of our algorithms via a restricted merit function, defined as:

Definition 1.2. *For any fixed value $D > 0$, we define the set $D_{x^0} := \{x \in \mathbb{R}^p : \|x - x^0\| \leq D\}$, then the restricted merit function for any $y \in \mathbb{R}^p$, $Err_D(y)$ is defined as:*

$$Err_D(y) := \sup_{x \in D_{x^0}} \{\langle y - x, F(x) \rangle\}. \quad (6)$$

Here we refer to Lemma 1 of Nesterov (2007) as a justification of the restricted merit function for analyzing convergence of (4) solutions.

Lemma 1.1 (Lemma 1 Nesterov (2007))**.** *Let D be any fixed positive value. The function Err_D is well-defined and convex on $\mathbb{R}^p$. For any $x \in \mathbb{R}^p$ such that $\|x - x^0\| \leq D$, we have $Err_D(x) \geq 0$. If x^* is a weak solution and $\|x^* - x^0\| \leq D$, then $Err_D(x^*) = 0$. Moreover, if $Err_D(\bar{x}) = 0$ for some $\bar{x} \in \mathbb{R}^p$ with $\|\bar{x} - x^0\| < D$, then $\bar{x}$ is a weak solution.*

Thus under Assumption 1.1 (a') and (b), our goal is to find a solution $x_\epsilon^* \in \mathbb{R}^p$ such that

$$Err_D(x_\epsilon^*) \leq \epsilon.$$

Notation. Throughout the paper, we use $\|\cdot\|$ to denote the Euclidean norm. When discussing the complexity of algorithms, we use $\mathcal{O}$ to denote the standard big-O notation and $\tilde{\mathcal{O}}$ to denote the big-O notation excluding logarithmic factors.

1.2 Related work and contributions

Adaptive methods. One of the main components in our framework is the adaptive gradient method (AdaGrad), introduced in Duchi et al. (2011); McMahan and Streeter (2010). Unlike stochastic gradient descent (SGD), AdaGrad adjusts the learning rate for each parameter based on its previous gradients, eliminating the need to manually tune hyperparameters. Later, Ward et al. (2020) popularized a single stepsize version of AdaGrad known as AdaGrad-Norm for smooth and non-convex minimization problems. The update rule, for any $\eta > 0$ and $b_0 > 0$, is as follows:

$$x^{k+1} = x^k - \frac{\eta}{b_{k+1}}\tilde{F}(x^k), \quad b_{k+1}^2 = b_k^2 + \|\tilde{F}(x^k)\|^2, \tag{7}$$

where $\tilde{F}$ denotes a stochastic gradient of the function being minimized. AdaGrad-Norm inherits the auto-tuning property of adaptive schemes and has a worst-case complexity of $\tilde{O}(N^{-1/2})$ in terms of the norm of the gradient squared for non-convex problems. However, the convergence guarantees originally required the gradient $F(x)$ to be uniformly bounded, i.e., $\|F(x)\| \leq M$ for all x. This requirement excludes even simple quadratic functions, raising concerns about its practical applicability. Subsequent work relaxed this assumption through a variety of approaches Faw et al. (2022); Wang et al. (2023). In addition, Liu et al. (2023) also prove a number of convergence results under various conditions, including last iterate convergence for a variant of AdaGrad-Norm. However, these methods are designed only for (non)convex minimization, while our work aims to address a much broader class of problems, including variational inequalities and saddle point problems.

Extragradient-type methods. Regarding algorithms for tackling root-finding problems, one-step gradient schemes are known to diverge unless we impose strong assumptions, such as co-coercivity of the operator. To overcome this challenge, extragradient (EG) methods, introduced by Korpelevich (1976), are a popular approach due to their stability and convergence properties when dealing with root finding problems and variational inequalities. In the unconstrained, deterministic case, the extragradient scheme is given as follows:

$$\begin{aligned} \text{Extrapolation step:} \quad \bar{x}^k &= x^k - \alpha_k F(x^k), \\ \text{Updated step:} \quad x^{k+1} &= x^k - \gamma_k F(\bar{x}^k). \end{aligned} \tag{8}$$

This extrapolation step is often perceived as a first-order approximation to the implicit method given by $x^{k+1} = x^k - \gamma_k F(x^{k+1})$, which has superior convergence properties and is known to be more stable than to explicit methods. To ensure a successful approximation and better convergence performance of EG, the assumption of L-Lipschitz continuity of F is crucial, as highlighted in Gorbunov et al. (2022a); Mishchenko et al. (2020). In general, deterministic EG can achieve both averaged and last iterate $\mathcal{O}(1/N)$ convergence rates under monotonicity and L-Lipschitz continuity assumptions Antonakopoulos et al. (2021); Cai et al. (2022); Gorbunov et al. (2022b).

Stochastic Extragradient schemes. In terms of stochastic extragradient schemes (SEG), two main approaches for generating the stochastic oracles exist: independent-sampling SEG (I-SEG) and same-sampling SEG (S-SEG). The convergence properties of S-SEG usually rely on the almost surely L-Lipschitz and monotonicity of the stochastic estimator $F(x, \xi)$ Mishchenko et al. (2020); Gorbunov et al. (2022a). However, despite the almost surely L-Lipschitz continuity assumption of F, the iterates of I-SEG rarely converge. This shortcoming was observed in the stochastic Mirror-Prox scheme proposed by Juditsky et al. (2011), which even fails on simple bilinear saddle point games. To address this issue, various approaches to reduce variance were introduced in the works of Chavdarova et al. (2019),Diakonikolas et al. (2021),Hsieh et al. (2020). In particular, Diakonikolas et al. (2021) continuously increase the batch size to reduce the variance introduced by the stochastic estimators $\tilde{F}$, which may be computationally intensive. Additionally, Chavdarova et al. (2019) proposed a stochastic extragradient method with variance reduction (SVRE), achieving a linear convergence rate but under strong monotonicity. In contrast, Hsieh et al. (2020) addressed this non-convergent issue by proposing a double stepsize EG (DSEG) with $\alpha_k \geq \gamma_k > 0$. In DSEG, the operator was no longer required to be strongly monotone, but it needed to satisfy an error condition, which is one of the settings under which we prove convergence of our algorithms.

Adaptive Extragradient methods. In the deterministic setting, there is a significant body of literature focused on adaptive extragradient algorithms, for solving variational inequalities, such as Antonakopoulos et al. (2019), Antonakopoulos et al. (2020), Böhm (2022), Bot et al. (2023), and Malitsky (2020). These works develop extragradient-style algorithms with adaptive step sizes under various assumptions, however, all require deterministic access to the operator F. Extensions to the stochastic case are of interest in these works, as Antonakopoulos et al. (2019) and Malitsky

(2020) suggest the stochastic setting as one direction of future work while Antonakopoulos et al. (2020) test their algorithm on a stochastic problem successfully, but do not provide convergence guarantees.

Adaptive stochastic Extragradient methods. Many recent works have focused on generalizing adaptive gradient methods to root finding problems, beginning with Bach and Levy (2019). In this work, the authors propose a stochastic Mirror-Prox algorithm using AdaGrad-style stepsizes under the condition that the dual norm of the stochastic oracles is uniformly bounded almost surely and the domain is bounded. In Liu et al. (2020), an Optimistic AdaGrad scheme is proposed and convergence is proven under similarly restrictive assumptions. More recently, Hsieh et al. (2022) devised optimistic methods using AdaGrad-Norm techniques for multi-player games. Their convergence results require a bounded operator and an almost surely bounded stochastic operator. In another line of work, Ene and Nguyen (2022) relaxes the almost surely bounded stochastic estimator assumptions to a uniformly bounded noise assumption but still relied on the bounded operator assumption, which is highly restrictive.

Contributions. We propose novel adaptive extragradient methods in both the deterministic and stochastic settings, named AdaEG-D and AdaEG-S respectively. Both algorithms are designed to solve (1), covering a wide range of optimization problems. We first prove a $\mathcal{O}(1/N)$ convergence rate of the averaged squared norm operator for AdaEG-D. Next, we introduce AdaEG-S for the stochastic setting. Under an affine noise assumption, we establish a $\tilde{\mathcal{O}}(N^{-1/2})$ convergence rate of the restricted merit function, matching the optimal rate for the stochastic setting, up to log factors, and a $\tilde{\mathcal{O}}(N^{-1/4})$ convergence rate of the averaged squared norm operator. In addition, we improve the convergence rate of these two distinct goals to match the optimal *deterministic* rate when the noise exhibits certain structure, such as the strong growth condition that is known to occur for over-parameterized machine learning models Vaswani et al. (2019). Finally, we implement our proposed algorithms and apply them to min-max saddle point problems. The experiments illustrate its effectiveness and auto-tuning properties.

Outline. Section 2 provides the convergence properties of AdaEG-D. Section 3 analyzes the convergence results of AdaEG-S as well as demonstrates the main techniques used to address the difficulties of moving from a deterministic to a stochastic scheme. Section 4 illustrates the effectiveness of our proposed schemes through saddle point problems. Section 5 summarizes this work and outlines possible future directions.

2 DETERMINISTIC ALGORITHM

Prior to considering the stochastic setting, we propose an adaptive extragradient method in the deterministic setting to solve the root-finding problem (1) under Assumptions 1.1 (a) and (b). In AdaEG-D, we initialize the hyper-parameters η, b_0 with positive values, and update step sizes of (8) as below:

$$\alpha_k = \gamma_k = \frac{\eta}{b_{k+1}}, \quad b_{k+1}^4 = b_k^4 + \|F(x^k)\|^2, \quad \forall k \geq 0.$$

The full algorithm is given in Algorithm 1. To motivate our choice of step sizes, we present the following descent lemma.

Algorithm 1 Adaptive Extragradient Norm-D

Initialize: $k = 0, N > 0, \eta > 0, \bar{b}_0, b_0 = \sqrt{\bar{b}_0}$.
while $k < N$: **do**

$$\begin{cases} b_{k+1}^4 & := b_k^4 + \|F(x^k)\|^2, \\ \bar{x}^k & := x^k - \dfrac{\eta}{b_{k+1}} F(x^k), \\ x^{k+1} & := x^k - \dfrac{\eta}{b_{k+1}} F(\bar{x}^k), \end{cases} \qquad \text{(AdaEG-D)}$$

end while

Lemma 2.1 (Descent Lemma for AdaEG-D)**.** *Let Assumptions 1.1 (a) and (b) hold and let x^k be generated by Algorithm 1. Then, for any $x^* \in S^*$ and for all $k \geq 0$:*

$$\begin{aligned} \|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 \\ - \tfrac{\eta^2}{b_{k+1}^2}\|F(x^k)\|^2 + \tfrac{\eta^4 L^2}{b_{k+1}^4}\|F(x^k)\|^2. \end{aligned} \tag{9}$$

Derivation of the step size. By selecting step sizes b_{k+1} such that $\eta^2 L^2 < b_{k+1}^2, \forall k$, we can immediately obtain a sub-linear convergence rate of the averaged iterates. However, this selection of step sizes relies on prior knowledge of the Lipschitz constant L. Our goal is to design a step size that is independent of L while achieving the same convergence rates. To this end, we require that our step size sequence $\{b_k\}_{k\geq 0}$ satisfies the following two conditions:

(1) $\sum_{k=0}^{N-1} \eta^4 L^2 \|F(x^k)\|^2 / b_{k+1}^4$ can be bounded by a relatively insignificant constant.

(2) The sequence $\{b_k\}_{k\geq 0}$ is non-decreasing and can be bounded from above by a constant.

Combining these two conditions with the results of standard technical lemmas for adaptive gradient methods,[1]

[1]See the supplementary material for details.

we construct the step size sequence such that

$$b_{k+1}^4 = b_k^4 + \|F(x^k)\|^2,$$

and obtain the following complexity result.

Theorem 2.2. *Let Assumptions 1.1 (a) and (b) hold and let x^k be generated by Algorithm 1. Then, for any $x^* \in S^*$ and for all $N \geq 1$:*

$$\frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2 \leq \mathcal{O}\left(\frac{\|x^0 - x^*\|^2 + \eta^2 L^2}{N}\right). \quad (10)$$

Remark 2.1. *The right side of* (10) *is simply a constant on the order $\mathcal{O}(1/N)$ so that by the identity,*

$$\min_{k=0,1,\ldots,N-1}\|F(x^k)\|^2 \leq \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2,$$

we obtain a $\mathcal{O}(1/N)$ best-iterate convergence rate, matching the optimal rate Ouyang and Xu (2021).

3 STOCHASTIC ALGORITHM

In this section, we introduce an adaptive stochastic EG algorithm. Let $\tilde{F}$ denote the stochastic estimate of F, satisfying the following assumptions at each iteration.

Assumption 3.1 (Unbiased stochastic operator)**.**

$$\mathbb{E}[\tilde{F}(x^k) \mid \mathcal{F}^k] = F(x^k), \quad \mathbb{E}[\tilde{F}(\bar{x}^k) \mid \bar{\mathcal{F}}^k] = F(\bar{x}^k).$$

Assumption 3.2 (Affine variance)**.**

$$\mathbb{E}[\|\tilde{F}(x^k) - F(x^k)\|^2 \mid \mathcal{F}^k] \leq \sigma_0^2 + \sigma_1^2\|F(x^k)\|^2,$$
$$\mathbb{E}[\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^2 \mid \bar{\mathcal{F}}^k] \leq \sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2.$$

Here, $\mathcal{F}^k$ and $\bar{\mathcal{F}}^k$ denote the natural filtrations, including all the randomness up to the construction of points x^k and $\bar{x}^k$. By definition, $\mathcal{F}^k \subset \bar{\mathcal{F}}^k \subset \mathcal{F}^{k+1}$. Additionally, σ_0, σ_1 are non-negative values.

Strong growth condition. The strong growth condition refers to Assumption 3.2 with $\sigma_0 = 0, \sigma_1 > 0$. We also refer it as the decaying noise setting.

Our full algorithm is presented in Algorithm 2. Unlike in the deterministic setting, we utilize two step sizes of different orders. Due to their construction, Algorithm 2 takes more aggressive extrapolation steps than update steps. This strategy is known to improve the convergence properties and enables last iterate convergence of stochastic EG under certain conditions Hsieh et al. (2020). Importantly, this will allow us to prove convergence in terms of the average norm of the true operator of the iterates, which we use to bound the adaptive step sizes in Algorithm 2. The need for a bound on the adaptive step sizes is essentially the root cause of the restrictive assumptions in previous work; by proving a bound based on the design of the algorithm, we are able to remove these restrictive assumptions.

Algorithm 2 Adaptive Extragradient Norm-S

Initialize: $k = 0, x^0 \in \mathbb{R}^p, N > 0, \eta > 0, \bar{b}_0 > 0, b_0 = \sqrt{\bar{b}_0}$.
while $k < N$: **do**

$$\begin{cases} b_{k+1}^4 := b_k^4 + \|\tilde{F}(x^k)\|^2, \quad \alpha_k = \dfrac{\eta}{b_{k+1}} \\ \bar{x}^k \quad := x^k - \alpha_k\tilde{F}(x^k), \\ \bar{b}_{k+1}^2 := \bar{b}_k^2 + \|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2, \quad \gamma_k = \dfrac{\eta}{\bar{b}_{k+1}} \\ x^{k+1} := x^k - \gamma_k\tilde{F}(\bar{x}^k), \end{cases} \quad (11)$$

end while

To develop the intuition behind our step size scheme, we first consider the following lemma.

Lemma 3.1. *Let Assumptions 1.1 (a) and (b), 3.1, and 3.2 hold and let x^k be generated by Algorithm 2. Let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation. Then, for any $x^* \in S^*$ and for all $k \geq 0$:*

$$\begin{aligned} \mathbb{E}_k[\|x^{k+1} - x^*\|^2] \leq \|x^k - x^*\|^2 + \mathbb{E}_k\big[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2\big], \\ +\mathbb{E}_k\big[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2 - 2\gamma_k\langle\bar{x}^k - x^*, F(\bar{x}^k)\rangle\big], \\ +\mathbb{E}_k\big[-2\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\big], \\ +\mathbb{E}_k\big[-2\gamma_k\alpha_k\langle\tilde{F}(x^k), F(x^k)\rangle\big]. \end{aligned} \quad (12)$$

Splitting step size selection. The summation of the two positive terms, $\sum_{k=0}^{N-1}\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2$ and $\sum_{k=0}^{N-1}\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2$, on the right hand side of (12) are effectively "noise" terms and need to be controlled by the step sizes. Motivated by standard technical results for adaptive gradient methods, we impose the following two conditions on the step sizes to bound these summations by logarithmic quantities:

$$\bar{b}_{k+1}^2 \sim \sum_{i=0}^{k}\|\tilde{F}(\bar{x}^k)\|^2, \quad b_{k+1}^2\bar{b}_{k+1} \sim \sum_{i=0}^{k}\|\tilde{F}(x^k)\|^2.$$

These two conditions directly motivate our choice of step sizes. As can be seen in the first condition, we require

$$\bar{b}_{k+1} \sim \sqrt{\sum_{i=0}^{k}\|\tilde{F}(\bar{x}^k)\|^2},$$

while needing

$$b_{k+1} \sim \sqrt[4]{\sum_{i=0}^{k}\|\tilde{F}(x^k)\|^2},$$

which explains why the exponents differ for b_{k+1} and $\bar{b}_{k+1}$, as well as why the $\bar{b}_{k+1}$ includes terms involving both $\tilde{F}(x^k)$ and $\tilde{F}(\bar{x}^k)$.

Remark 3.1. *We wish to stress here that the choice to use the current stochastic oracles, while at odds with many prior works Bach and Levy (2019); Ene and Nguyen (2022); Hsieh et al. (2022), is intentional, as it is the most commonly used approach in practical methods and is key to relaxing assumptions such as a bounded domain or bounded operator. This is due to the fact that the analysis of adaptive stochastic algorithms all rely certain technical results similar to those which we employ. However, when the step sizes are not updated based on the current oracles, these lemma require an a-priori bound on the norm of the stochastic oracles (see, e.g. Lemma 10 of Bach and Levy (2019)), leading to more restrictive assumptions.*

Now, we discuss certain challenges in deriving the convergence results of Algorithm 2 as well as present our main techniques to address these issues.

De-correlated stepsizes. Since the step sizes γ_k and α_k are updated according to the current stochastic oracles $\tilde{F}(x^k)$ and $\tilde{F}(\bar{x}^k)$, we cannot directly take the expectation of the inner products in (12), i.e,

$$\mathbb{E}_k\left[-2\gamma_k\alpha_k\langle\tilde{F}(x^k), F(x^k)\rangle\right] \neq -2\gamma_k\alpha_k\|F(x^k)\|^2.$$

To proceed, we utilize the pre-iteration step sizes $\gamma_{k-1}, \alpha_{k-1}$ to resolve this correlation issue. Then,

$$\begin{aligned}&\mathbb{E}_k\left[-2\gamma_k\alpha_k\langle\tilde{F}(x^k), F(x^k)\rangle\right],\\ &\leq 2\mathbb{E}_k[|\gamma_k\alpha_k - \gamma_{k-1}\alpha_{k-1}||\langle\tilde{F}(x^k), F(x^k)\rangle|]\\ &- 2\gamma_{k-1}\alpha_{k-1}\|F(x^k)\|^2,\end{aligned}$$

we obtain the negative term $-2\gamma_{k-1}\alpha_{k-1}\|F(x^k)\|^2$ on the right side which drives convergence and enables us to bound the error term:

$$2\mathbb{E}_k[|\gamma_k\alpha_k - \gamma_{k-1}\alpha_{k-1}||\langle\tilde{F}(x^k), F(x^k)\rangle|].$$

Now we focus on the term

$$\mathbb{E}_k\left[-2\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right].$$

If we repeat the de-correlated step size technique here, as a byproduct, we obtain an extra error term proportional to $\|x^k - x^*\|^2$, which unfortunately prevents telescoping of the terms

$$\mathbb{E}_k\left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right].$$

In order to overcome this difficulty, we propose three different solutions.

Error Bound Condition: First, we consider the case where F satisfies:

Assumption 3.3 (Error bound condition)**.**

$$\operatorname{dist}(x, S^*)^2 \leq q\|F(x)\|^2, \quad \forall x \in \operatorname{dom}(F). \tag{13}$$

As a result, the extra error term involving $\|x^k - x^*\|^2$ can be cancelled out with the previously derived $-2\gamma_{k-1}\alpha_{k-1}\|F(x^k)\|^2$.

Star-strongly Monotone Operator: Alternatively, we can impose the star-strongly monotone condition on F, i.e.

Assumption 3.4 (Star-strongly monotone)**.**

$$\langle F(x), x - x^*\rangle \geq m\|x - x^*\|^2, \quad \forall x, y \in \operatorname{dom}(F), \tag{14}$$

which is a relaxation of m-strong monotonicity and can be seen as an extension of quasi-strong convexity assumption, which has been used in the minimization setting in Gower et al. (2019) and Necoara et al. (2019). In addition,Loizou et al. (2021), Beznosikov et al. (2023), and Zhang et al. (2023)have analyzed stochastic gradient descent ascent methods for variational inequalities under Assumption 3.4, without adaptive stepsize schemes. Extragradient schemes have also been proposed for stochastic variational inequalities obeying Assumption 3.4, including a clipped extragradient (and clipped gradient descent ascent) method in Gorbunov et al. (2022c) as well as a past extragradient method in Choudhury et al. (2023).

As a result of this assumption, we can bound the additional error term with $\mathbb{E}_k\left[-2\gamma_k\langle\bar{x}^k - x^*, F(\bar{x}^k)\rangle\right]$.

Scaled Sub-Weibull Noise: Our final approach is inspired by Liu et al. (2023). We re-scale the term

$$\mathbb{E}_k[\|x^{k+1} - x^*\|^2]$$

by a carefully chosen parameter:

$$\beta_{k+1} = \frac{\bar{b}_{k+1}^{1/2}}{2\bar{b}_{k+1}^{1/2} - \bar{b}_0^{1/2}}.$$

In addition, we impose the scaled sub-Weibull noise assumption on the stochastic oracle in the updated step, given as:

Assumption 3.5 (Scaled Sub-Weibull)**.**

$$\mathbb{E}\left[\exp\left(\frac{\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\sqrt{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}}\right)^{1/\theta}\middle|\bar{\mathcal{F}}_k\right] \leq e, \tag{15}$$

for some $\theta > 0, \sigma_0, \sigma_1 \geq 0$.

Remark 3.2. *Note that Assumption 3.5 implies Assumption 3.2 (up to a constant factor) and is more general when compared with the assumptions used in other adaptive extragradient methods Ene and Nguyen (2022); Hsieh et al. (2022); Liu et al. (2023) since we allow both σ_0, σ_1 to be non-negative. In addition, this condition only applies to the stochastic gradients in the update step.*

As a result, in each of these three cases, we can derive a descent lemma, which is fundamental to our analysis and enables us to derive an upper bound on the quantity

$$\mathbb{E}\left(\frac{\sum_{k=0}^{N-1}\|F(x^k)\|^2}{\bar{b}_k^{3/2}}\right)$$

in terms of the expected value of the adaptive step sizes, which we present in the following lemma.

Lemma 3.2 (Informal)**.** *Let Assumptions 1.1 (a) and (b), 3.1, and 3.2 hold and let either Assumption 3.3, Assumption 3.4, or Assumption 3.5*[2] *hold as well. Then, for any $N \geq 1$, the following holds for the iterates of Algorithm 2:*

$$\mathbb{E}\left(\frac{\sum_{k=0}^{N-1}\|F(x^k)\|^2}{\bar{b}_k^{3/2}}\right) \leq C_1 + C_2 \ln \mathbb{E}\sqrt{\bar{b}_N}, \quad (16)$$

and for any $x^* \in S^*$:

$$\mathbb{E}\left[\|x^N - x^*\|^2\right] \leq M_1 + M_2 \ln \mathbb{E}\sqrt{\bar{b}_N} \quad (17)$$

$$\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \leq M_1 + M_2 \ln \mathbb{E}\sqrt{\bar{b}_N},$$

where $C_1, C_2, M_1, M_2 > 0$ are problem specific constants and dependent on which set of assumptions hold.

Bounding the stochastic step sizes. It should be clear that the two adaptive step size denominators, $\bar{b}_k$ and b_k, are monotonically increasing with k, which raises concerns about the possibility of these terms becoming too large too quickly and thus leading to short stepsizes and slow convergence of Algorithm 2. Previous work addressed similar concerns by imposing more restrictive assumptions either on the operator's norm or the noise. In contrast, we derive a closed-form upper bound for the step size $\bar{b}_k$ at each iteration $k \geq 1$. To construct this bound, we employ a "divide and conquer" technique, first proposed for adaptive gradient methods by Wang et al. (2023), in conjunction with the result of Lemma 3.2. Specifically, we divide the iterations based on the set $\{\|F(x)\|^2 \geq \bar{\sigma}^2\}$, where

$$\bar{\sigma}^2 = \left(2 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)\frac{\sigma_0^2}{1+\sigma_1^2}.$$

From the design of the step size $\bar{b}_N^2$, Assumption 3.2, and the result of Lemma 3.2, we are able to find a reasonable bound on $\mathbb{E}[\bar{b}_N^{1/2}]$ on any iteration satisfying $\|F(x^k)\|^2 \geq \bar{\sigma}^2$. On the other hand, for iterations satisfying $\|F(x^k)\|^2 \leq \bar{\sigma}^2$, an upper bound follows directly by Assumption 3.2. Combining these two cases yields an upper bound on the expected step size denominator of the form

$$\mathbb{E}\left[\bar{b}_N^{1/2}\right] \leq 2D_1 + 8D_2 \ln(e + D_2),$$

where $D_1 = \mathcal{O}(\bar{\sigma}^{1/2}N^{1/4})$, and $D_2 = \mathcal{O}(1)$.

Remark 3.3. *The growth rate of the step size denominator, $\bar{b}_N$, is determined by the order of D_1. When $\sigma_0 = 0$ and thus $\bar{\sigma} = 0$, D_1 becomes a constant, allowing us to explicitly lower bound the step size by a constant. Otherwise, $\bar{b}_N$ grows at a rate of $\mathcal{O}(N^{1/4})$. The first case, where $\sigma_0 = 0$, occurs when the strong growth condition holds, which has been observed for over-parameterized deep neural networks Vaswani et al. (2019). In this setting, the overall convergence rate of Algorithm 2 improves significantly.*

With this bound, we are able to prove our first convergence result, in terms of the average operator norm.

Theorem 3.3 (Informal)**.** *Let Assumptions 1.1 (a) and (b), 3.1, and 3.2 hold and let either Assumption 3.3, Assumption 3.4, or Assumption 3.5 hold as well. Then, for any $N \geq 1$, the following holds for the iterates of Algorithm 2, with probability at least $1 - \delta$:*

$$\min_{k=0,\ldots,N-1}\|F(x^k)\|^2 \leq \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2 \leq \frac{c_N}{\delta^4},$$

where

$$c_N = \mathcal{O}\left(\frac{D_1^3 \ln D_1}{N}\right).$$

We note that this result holds with high probability, which is a common consequence of using an adaptive stepsize scheme.

Finally, in order to obtain convergence in terms of the restricted merit function, we need to convert our analysis from a "descent" style of analysis such as in Lemma 3.1 to a "regret" style, where the stepsize γ_k is divided through prior to taking expectations. We achieve this without additional assumptions (beyond monotonicity) by repeated use of the three different bounds developed in Lemma 3.2, leading to our final convergence result.

[2] This result holds with high probability when Assumption 3.5 holds.

Theorem 3.4 (Informal). *Let Assumptions 1.1 (a') and (b), 3.1, and 3.2 hold and let either Assumption 3.3*[3]*, Assumption 3.4, or Assumption 3.5 hold as well. Then, for any $N \geq 1$, the following holds for the iterates of Algorithm 2, with probability at least $1-\delta$:*

$$\mathbb{E}\left[Err_D(y^N)\right] \leq \frac{c_N}{\delta^2}, \tag{18}$$

where

$$y^N = \tfrac{1}{N}\sum_{k=0}^{N-1} \bar{x}^k, \quad c_N = \mathcal{O}\left(\frac{D_1^2 \ln D_1}{N}\right).$$

Remark 3.4. *Depending on the noise settings, we obtain different rates of convergence. Under Assumption 3.2, when $\sigma_0 > 0$, $D_1 = \mathcal{O}(N^{1/4})$, leading to*

$$\mathbb{E}\left[Err_D(y^N)\right] = \tilde{\mathcal{O}}\Big(\frac{1}{N^{1/2}}\Big),$$
$$\min_{k=0,\ldots,N-1} \|F(x^k)\|^2 = \tilde{\mathcal{O}}\Big(\frac{1}{N^{1/4}}\Big),$$

which matches the optimal complexity for the restricted merit function (up to log factors) in the stochastic setting. When $\sigma_0 = 0$, then D_1 is simply a constant, which improves the complexity results to

$$\mathbb{E}\left[Err_D(y^N)\right] = \mathcal{O}\Big(\frac{1}{N}\Big),$$
$$\min_{k=0,\ldots,N-1} \|F(x^k)\|^2 = \mathcal{O}\Big(\frac{1}{N}\Big),$$

which matches the optimal complexity for the restricted merit function in the deterministic regime.

4 NUMERICAL EXPERIMENTS

We compare the performance of Algorithm 2 with the Double Stepsize Extragradient (DSEG) method of Hsieh et al. (2020) on minimax games. In addition, due to considering different noise regimes, we also include a fixed stepsize variant of the DSEG method. The fixed stepsize version is often known as Extragradient+ (EG+) Diakonikolas et al. (2021).

Bilinear saddle point problem. We first consider the bi-linear game setup from Hsieh et al. (2020) defined as below:

$$\min_{\theta\in\mathbb{R}^d}\max_{\phi\in\mathbb{R}^d} \mathcal{L}(\theta,\phi) = \theta^T C \phi,$$

where $C \in \mathbb{R}^{d\times d}$ is a positive definite matrix and $d = 50$.

[3]When Assumption 3.3 holds, we also require that S^* is a singleton set to prove the result of Theorem 3.4

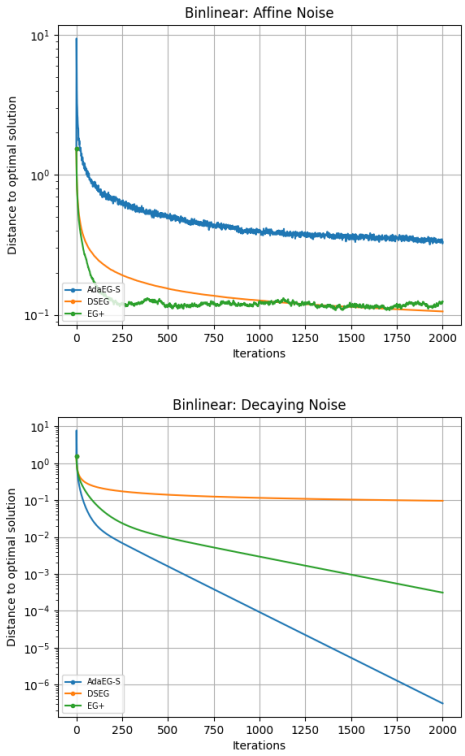


Figure 1: Distance of last iterate to the optimal solution on a bilinear game.

Strongly convex-concave game. Additionaly, we consider a strongly convex-concave saddle point problem set up from Hsieh et al. (2020), defined as:

$$\min_{\theta\in\mathbb{R}^d}\max_{\phi\in\mathbb{R}^d} \mathcal{L}(\theta,\phi) =$$
$$(\theta^T A_2 \theta)^2 + 2\theta^T A_1 \theta + 4\theta^T C\phi - 2\phi^T B_1 \phi - (\phi^T B_2 \phi)^2,$$

where A_1, A_2, B_1, B_2 are 50×50 positive definite matrices. Full details of the experiments can be found in the supplementary material.[4]

Results. We plot the mean over 20 runs for each scenario. Figures 1 and 2 compare the distance of the iterates x^k to the optimal solution while Figures 3 and 4 plot the distance of the average iterate to the optimal solution. In general, when comparing EG+ and DSEG, DSEG appears to perform best in the affine noise setting ($\sigma_0 > 0$) while EG+ is most effective in the decaying noise setting ($\sigma_0 = 0$). In contrast, AdaEG-S is able to adapt to the type of noise effectively and is competitive in all regimes, especially when considering the distance of the average iterate to the solution. Moreover, when the noise is decaying, AdaEG-S demonstrates much faster convergence compared to the

[4]The code is available at: `https://github.com/MichaelJONeill/Adaptive-Extragradient-Methods`

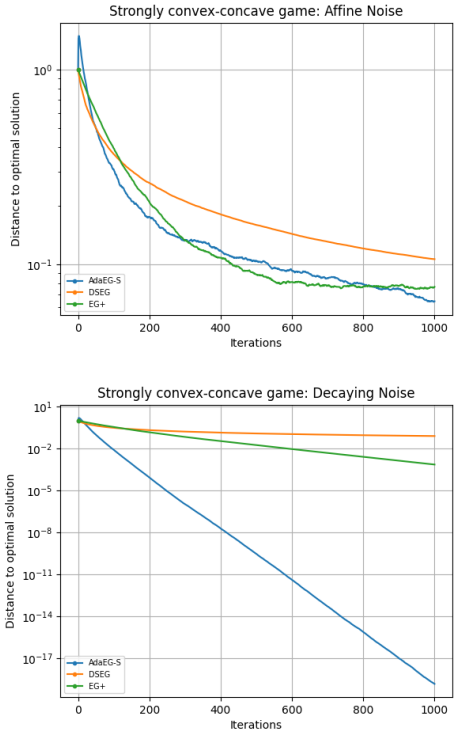


Figure 2: Distance of last iterate to the optimal solution on a strongly convex-concave game.

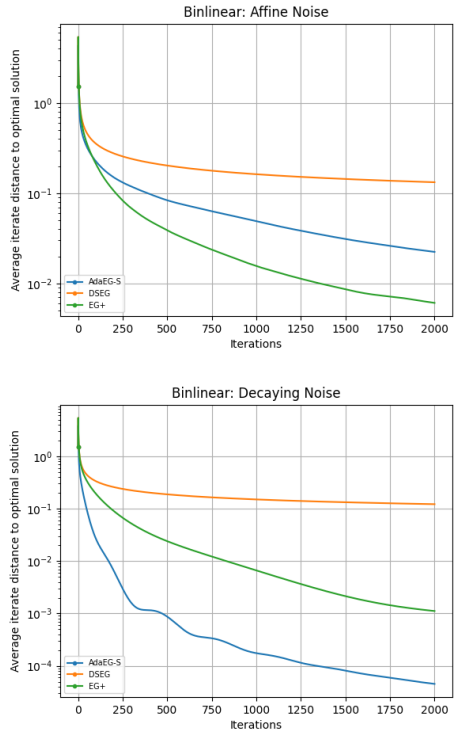


Figure 3: Distance of average iterate to the optimal solution on a bilinear game.

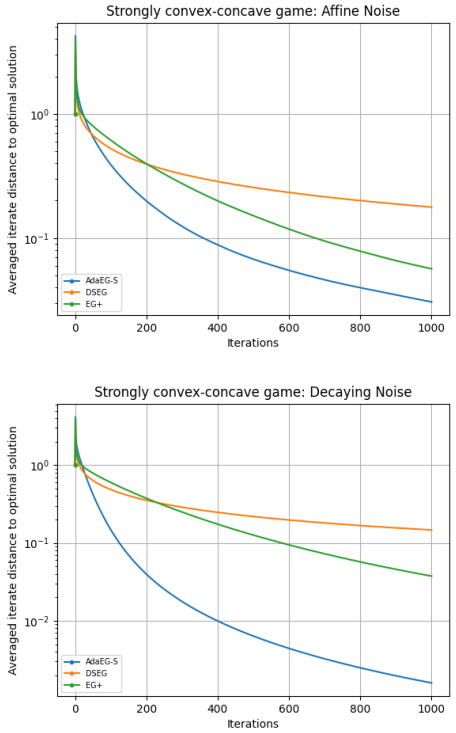


Figure 4: Distance of average iterate to the optimal solution on a bilinear game.

other two algorithms and appears to exhibit linear last iterate convergence on strongly convex-concave minimax problem. While EG+ also appears to obtain linear convergence in this case, it is at a significantly slower rate. We view this as confirmation that AdaEG-S is able to naturally adapt to the noise present in the problem. In addition, EG+ and DSEG required problem specific parameter tuning, while the same hyperparameters were used for all problems in AdaEG-S, again demonstrating its adaptive nature. A more thorough investigation of the sensitivity of AdaEG-S to its hyperparameters as well as some additional plots are provided in Appendix D.

5 CONCLUDING REMARKS

In this paper, we propose novel adaptive extragradient methods and demonstrate their convergence under mild conditions in both the deterministic and the stochastic settings. Furthermore, we prove that our algorithm adapts to the nature of the noise, exhibiting faster convergence when the strong growth condition holds. Intriguingly, our experimental results suggest last-iterate convergence behavior when this growth condition holds, which is a direction for future investigation.

References

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1laEnA5Ym`.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and Will Hamilton. Adversarial example games. *Advances in neural information processing systems*, 33: 8921–8934, 2020.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg8jjC9KQ`.

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Skj8Kag0Z`.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Kimon Antonakopoulos, Veronica E. Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR 2021-9th International Conference on Learning Representations*, pages 1–28, 2021.

Francis Bach and Kfir Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on learning theory*, pages 164–194. PMLR, 2019.

Alina Ene and Huy Nguyen. Adaptive and universal algorithms for variational inequalities with optimal convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6559–6567, 2022.

Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation. *Advances in Neural Information Processing Systems*, 35:6544–6556, 2022.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJxIm0VtwH`.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pages 313–355. PMLR, 2022.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR, 2023.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. On the convergence of adagrad (norm) on $\mathbb{R}^d$: Beyond convexity, non-asymptotic rate and acceleration. In *International Conference on Learning Representations*. International Conference on Learning Representations, 2023.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022a.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022b.

Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. *Advances in Neural Information Processing Systems*, 32, 2019.

Kimon Antonakopoulos, E Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. *arXiv preprint arXiv:2010.12100*, 2020.

Axel Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. *arXiv preprint arXiv:2201.12247*, 2022.

Radu I Bot, Michael Sedlmayer, and Phan Tu Vuong. A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to gans. *Journal of Machine Learning Research*, 24(8): 1–37, 2023.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International conference on artificial intelligence and statistics*, pages 172–235. PMLR, 2023.

Siqi Zhang, Sayantan Choudhury, Sebastian U Stich, and Nicolas Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. *arXiv preprint arXiv:2306.05100*, 2023.

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechenskii, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *Advances in Neural Information Processing Systems*, 35:31319–31332, 2022c.

Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. *Advances in Neural Information Processing Systems*, 36:64918–64956, 2023.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A TECHNICAL LEMMAS AND PRELIMINARIES

Lemma A.1. *Consider the sequence $\{a_i\}_{i\geq 0}$ of non-negative numbers and $a_0 > 0$. Then for any $N \geq 1$,*

$$\begin{aligned} &\textstyle\sum_{k=1}^{N-1} \frac{a_k}{\sum_{i=0}^{k} a_i} \leq \ln \sum_{k=0}^{N-1} a_i - \ln a_0, \\ &\textstyle\sum_{k=1}^{N-1} \frac{a_k}{\left(\sum_{i=0}^{k} a_i\right)^{3/2}} \leq \frac{2}{\sqrt{a_0}} \end{aligned} \tag{19}$$

Proof. This follows directly from (Wang et al., 2023, Lemma 10). □

Lemma A.2 (Wang et al. (2023))**.** *Given any constants $A > 0, B > 0$, and variable x, if*

$$x \leq A + B\ln(x),$$

then we have

$$x \leq 2A + 8B\ln(e + B).$$

Proof. For simplicity, let us define function $f(x) := 2B\ln(x) - x, \forall x$. Then from its derivative $f'(x) = \frac{2B}{x} - 1$, we know that if $x > 2B$, f is decreasing. By defining the point $\bar{x} = 8B\ln(e + B)$, it easy to show that $\bar{x} > 2B$ and $f(\bar{x}) < 0$. Thus for any $x > \bar{x}$, $f(x) < 0$, that is $B\ln(x) < \frac{x}{2}$. Plugging this identity into $x \leq A + B\ln(x)$, we obtain $x < 2A$ in the case when $x > \bar{x}$. Altogether, we can conclude that $x \leq 2A + \bar{x} = 2A + 8B\ln(e + B)$ as desired. This result was originally given without proof in Wang et al. (2023). □

In the main body of the paper, we consider three cases under which we are able to prove our main convergence results. In this appendix, we slightly modify one of these conditions to include additional constant factors, which were previously suppressed in order to make the presentation more clear. In particular, throughout the supplementary material, we work with the follow assumption when considering the case of sub-Weibull noise.

Assumption A.1 (Scaled Sub-Weibull with Constant Factors)**.**

$$\mathbb{E}\left[\exp\left(\left(\frac{\sqrt{\Gamma(2\theta+1)e}\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\sqrt{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}}\right)^{1/\theta}\right) \middle| \bar{\mathcal{F}}_k\right] \leq e, \tag{20}$$

for some $\theta > 0, \sigma_0 \geq 0$, and $\sigma_1 \geq 0$.

We include these constant factors due to the following result.

Lemma A.3. *Let Assumption A.1 hold. Then, for all k,*

$$\mathbb{E}[\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^2 | \bar{\mathcal{F}}_k] \leq \sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2. \tag{21}$$

Proof. The proof follows directly by the proof of (Liu et al., 2023, Lemma A.3). □

From this result, we can see that Assumption A.1 implies that the second condition of Assumption 3.2 holds, motivating the addition of the constants in Assumption A.1.

In addition, when this assumption holds, we have the following result.

Lemma A.4. *Let Assumption A.1 hold and let $E(\delta)$ be the event that*

$$\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^2 \leq \frac{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}{\Gamma(2\theta+1)e}\log^{2\theta}\left(\frac{Ne}{\delta}\right) \tag{22}$$

holds for all $k = 0, \ldots, N-1$. Then,

$$\mathbb{P}[E(\delta)] \geq 1 - \delta. \tag{23}$$

Proof. Consider any $k \in \{0, \ldots, N-1\}$. Then,

$$
\begin{aligned}
&\mathbb{P}\left[\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^2 > \frac{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}{\Gamma(2\theta+1)e}\log^{2\theta}\left(\frac{Ne}{\delta}\right)\right] \\
&= \mathbb{P}\left[\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^{\frac{1}{\theta}} > \left(\frac{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}{\Gamma(2\theta+1)e}\right)^{\frac{1}{2\theta}}\log\left(\frac{Ne}{\delta}\right)\right] \\
&= \mathbb{P}\left[\exp\left(\left(\frac{\sqrt{\Gamma(2\theta+1)e}\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\sqrt{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}}\right)^{1/\theta}\right) > \exp\left(\log\left(\frac{Ne}{\delta}\right)\right)\right] \\
&\leq \exp\left(-\log\left(\frac{Ne}{\delta}\right)\right)\mathbb{E}\left[\exp\left(\left(\frac{\sqrt{\Gamma(2\theta+1)e}\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\sqrt{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}}\right)^{1/\theta}\right)\right] \\
&= \exp\left(\log\left(\frac{\delta}{Ne}\right)\right)\mathbb{E}\left[\mathbb{E}\left[\exp\left(\left(\frac{\sqrt{\Gamma(2\theta+1)e}\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\sqrt{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}}\right)^{1/\theta}\right)\Big|\bar{\mathcal{F}}_k\right]\right] \\
&\leq \frac{\delta e}{Ne} = \frac{\delta}{N},
\end{aligned}
$$

where the first inequality is due to Markov's inequality. Therefore,

$$
\begin{aligned}
\mathbb{P}[E(\delta)] &= 1 - \sum_{k=0}^{N-1}\mathbb{P}\left[\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^2 > \frac{\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2}{\Gamma(2\theta+1)e}\log^{2\theta}\left(\frac{Ne}{\delta}\right)\right] \\
&\geq 1 - \sum_{k=0}^{N-1}\frac{\delta}{N} \\
&= 1 - \delta.
\end{aligned}
$$

□

B CONVERGENCE PROOFS FOR DETERMINISTIC ALGORITHM

In this section, we provide the missing proofs of the results presented in Section 2.

Lemma B.1 (Proof of Lemma 2.1)**.** *Let Assumptions 1.1 (a) and (b) hold and let $\{x^k\}$ be generated by Algorithm 1. Then for any $x^* \in S^*$ and $k \geq 0$,*

$$
\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{\eta^2}{b_{k+1}^2}\left(1 - \frac{\eta^2 L^2}{b_{k+1}^2}\right)\|F(x^k)\|^2. \tag{24}
$$

Proof. From the definition of x^{k+1} in Algorithm 1, the Minty variational condition and $L-$ Lipschitz continuity of F, we have:

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - \frac{\eta}{b_{k+1}} F(\bar{x}^k) - x^*\|^2 \\
&= \|x^k - x^*\|^2 + \frac{\eta^2}{b_{k+1}^2}\|F(\bar{x}^k)\|^2 - 2\langle x^k - x^*, \frac{\eta}{b_{k+1}} F(\bar{x}^k)\rangle \\
&= \|x^k - x^*\|^2 + \frac{\eta^2}{b_{k+1}^2}\|F(\bar{x}^k)\|^2 - 2\langle x^k - \bar{x}^k + \bar{x}^k - x^*, \frac{\eta}{b_{k+1}} F(\bar{x}^k)\rangle \\
&\leq \|x^k - x^*\|^2 + \frac{\eta^2}{b_{k+1}^2}\|F(\bar{x}^k)\|^2 - 2\langle x^k - \bar{x}^k, \frac{\eta}{b_{k+1}} F(\bar{x}^k)\rangle \\
&= \|x^k - x^*\|^2 + \frac{\eta^2}{b_{k+1}^2}\|F(\bar{x}^k)\|^2 - 2\langle \frac{\eta}{b_{k+1}} F(x^k), \frac{\eta}{b_{k+1}} F(\bar{x}^k)\rangle \\
&= \|x^k - x^*\|^2 + \frac{\eta^2}{b_{k+1}^2}\|F(x^k) - F(\bar{x}^k)\|^2 - \frac{\eta^2}{b_{k+1}^2}\|F(x^k)\|^2 \\
&\leq \|x^k - x^*\|^2 - \frac{\eta^2}{b_{k+1}^2}(1 - \frac{\eta^2 L^2}{b_{k+1}^2})\|F(x^k)\|^2. \qquad (25)
\end{aligned}$$

□

Now, we present a lemma about the behavior of the stepsize denominator sequence, $\{b_k\}$.

Lemma B.2. *Under the assumptions of Lemma B.1, for any $N \geq 1$,*

$$\sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_{k+1}^2} \leq \frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 \left(1 + \ln b_N^4\right).$$

In addition,

$$b_N^2 \leq 2C_1 + 8C_2 \ln(e + C_2),$$

where

$$C_1 := b_0^2 + \frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2, \qquad C_2 := 2\eta^2 L^2.$$

Proof. Summing the result of B.1 from $k = 0, \ldots, N-1$, we have

$$\begin{aligned}
\eta^2 \sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_{k+1}^2} &\leq \sum_{k=0}^{N-1} \left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right] + \eta^4 L^2 \sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_{k+1}^4} \qquad (26) \\
&\leq \|x^0 - x^*\|^2 + \eta^4 L^2 \left(1 + \ln b_N^4\right),
\end{aligned}$$

where the last inequality follows by Lemma A.1. Dividing both sides by η^2,

$$\sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_{k+1}^2} \leq \frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 \left(1 + \ln b_N^4\right), \qquad (27)$$

which proves the first result.

Next, we utilize this inequality to bound the step size denominator b_N^2. From the construction of b_N^4, we have

$$\begin{aligned} b_N^2 &= \frac{b_0^4 + \sum_{k=0}^{N-1} \|F(x^k)\|^2}{b_N^2} \\ &\leq b_0^2 + \sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_{k+1}^2} \\ &\leq b_0^2 + \frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 \left(1 + \ln b_N^4\right) \\ &= b_0^2 + \frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 + 2\eta^2 L^2 \ln b_N^2 \\ &= C_1 + C_2 \ln b_N^2. \end{aligned} \tag{28}$$

Applying Lemma A.2 yields the second result. □

Theorem B.3 (Proof of Theorem 2.2)**.** *Let Assumptions 1.1 (a) and (b) hold and let $\{x^k\}$ be generated by Algorithm 1. Then for any $x^* \in S^*$ and $N \geq 1$,*

$$\min_{k=0,1,\ldots,N-1} \|F(x^k)\|^2 \leq \frac{2C_1 + 8C_2 \ln(2 + C_2)}{N} \left[\frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 C_3\right], \tag{29}$$

where C_1 and C_2 are defined in Lemma B.2 and

$$C_3 := (1 + 2\ln(2C_1 + 8C_2 \ln(2 + C_2)))$$

.

Proof. Since the step size sequence $\{b_k\}_{k\geq 0}$ is non-decreasing in k, we have

$$\sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_N^2} \leq \sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_{k+1}^2}.$$

From the result of Lemma B.2, it follows that

$$\begin{aligned} \sum_{k=0}^{N-1} \|F(x^k)\|^2 &\leq b_N^2 \left[\frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 \left(1 + 2\ln b_N^2\right)\right], \\ &\leq (2C_1 + 8C_2 \ln(2 + C_2)) \left[\frac{\|x^0 - x^*\|^2}{\eta^2} + \eta^2 L^2 C_3\right]. \end{aligned} \tag{30}$$

Dividing both sides by N, and using

$$\min_{k=0,1,\ldots,N-1} \|F(x^k)\|^2 \leq \frac{1}{N} \sum_{k=0}^{N-1} \|F(x^k)\|^2$$

proves the result. □

C CONVERGENCE PROOFS FOR THE STOCHASTIC ALGORITHM

Now, we present our proofs of the convergence results presented in Section 3. We recall here that we denote the conditional expectation as $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$.

Lemma C.1 (Proof of Lemma 3.1)**.** *Let Assumption 1.1 (b) hold and let $\{x^k\}$ be generated by Algorithm 2. Then, for any $x^* \in S^*$ and $k \geq 0$,*

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1}-x^*\|^2\right] \quad \leq \quad & \|x^k-x^*\|^2+\mathbb{E}_k\left[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2\right]+\mathbb{E}_k\left[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2\right], \\ & +\mathbb{E}_k\left[-2\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle\right]+\mathbb{E}_k\left[-2\gamma_k\alpha_k\langle\tilde{F}(x^k),F(x^k)\rangle\right], \\ & +\mathbb{E}_k\left[-2\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle\right].\end{aligned} \tag{31}$$

Proof. From the definition of x^{k+1} in Algorithm 2,

$$\begin{aligned}&\|x^{k+1}-x^*\|^2 \\ &=\|x^k-x^*\|^2+\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2-2\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)\rangle, \\ &=\|x^k-x^*\|^2+\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2-2\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle-2\gamma_k\langle x^k-x^*,F(\bar{x}^k)\rangle.\end{aligned} \tag{32}$$

We process the last inner product as follows,

$$\begin{aligned}-2\gamma_k\langle x^k-x^*,F(\bar{x}^k)\rangle &= -2\gamma_k\langle x^k-\bar{x}^k,F(\bar{x}^k)\rangle-2\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle, \\ &= -2\alpha_k\gamma_k\langle\tilde{F}(x^k),F(\bar{x}^k)\rangle-2\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle, \\ &= -2\alpha_k\gamma_k\langle\tilde{F}(x^k),F(\bar{x}^k)-F(x^k)+F(x^k)\rangle-2\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle, \\ &\leq 2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2-2\alpha_k\gamma_k\langle\tilde{F}(x^k),F(x^k)\rangle-2\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle,\end{aligned} \tag{33}$$

here we use the L-Lipschitz continuity of F to derive the inequality. Combining this inequality with (32), and taking the conditional expectation proves the result. □

C.1 General Descent Lemma

Prior to consider the three cases individually, we prove a general descent lemma. This lemma has two cases, one which includes the scaling sequence $\{\beta_k\}$ (defined below) and one which does not.

Lemma C.2 (General Descent Lemma)**.** *Let Assumptions 1.1 (a) and (b) hold and let $\{x^k\}$ be generated by Algorithm 2. Then for any $x^* \in S^*$, the following holds:*

$$\begin{aligned}&\mathbb{E}_k[\|x^{k+1}-x^*\|^2] \\ &\leq \|x^k-x^*\|^2+\mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2]-\mathbb{E}_k[2\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle] \\ &-\mathbb{E}_k[2\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle]+\mathbb{E}_k[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2]-\mathbb{E}_k[2\alpha_k\gamma_k\|F(x^k)\|^2] \\ &+\mathbb{E}_k[2\alpha_k\gamma_k\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle].\end{aligned} \tag{34}$$

In addition, defining

$$\beta_{k+1} := \frac{\bar{b}_{k+1}^{1/2}}{2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2}}, \tag{35}$$

then,

$$\begin{aligned}&\mathbb{E}_k[\beta_{k+1}\|x^{k+1}-x^*\|^2] \\ &\leq \mathbb{E}_k[\beta_{k+1}\|x^k-x^*\|^2]+\mathbb{E}_k[\beta_{k+1}\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2]-\mathbb{E}_k[2\beta_{k+1}\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle] \\ &-\mathbb{E}_k[2\beta_{k+1}\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle]+\mathbb{E}_k[2L\beta_{k+1}\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2] \\ &-\mathbb{E}_k[2\beta_{k+1}\alpha_k\gamma_k\|F(x^k)\|^2]+\mathbb{E}_k[2\beta_{k+1}\alpha_k\gamma_k\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle],\end{aligned} \tag{36}$$

holds as well.

Proof. From the update rule for x^{k+1}, for any $k \geq 0$,

$$\begin{aligned}&\mathbb{E}_k[\|x^{k+1}-x^*\|^2] \\ &\quad=\mathbb{E}_k[\|x^k-\gamma_k\tilde{F}(\bar{x}^k)-x^*\|^2] \\ &\quad=\mathbb{E}_k[\|x^k-x^*\|^2]+\mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2]-\mathbb{E}_k[2\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)\rangle].\end{aligned} \tag{37}$$

We process the last inner product as follows,

$$\begin{aligned}
&-\mathbb{E}_k[2\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k)\rangle] \\
&\quad = -2\mathbb{E}_k[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k) + F(\bar{x}^k)\rangle] \\
&\quad = -2\mathbb{E}_k[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle] - 2\mathbb{E}_k[\gamma_k\langle x^k - x^*, F(\bar{x}^k)\rangle] \\
&\quad = -2\mathbb{E}_k[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle] - 2\mathbb{E}_k[\gamma_k\langle x^k - \bar{x}^k, F(\bar{x}^k)\rangle] \\
&\qquad -2\mathbb{E}_k[\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle].
\end{aligned} \tag{38}$$

Combining equations (37) and (38), we obtain

$$\begin{aligned}
&\mathbb{E}_k[\|x^{k+1} - x^*\|^2] \\
&\leq \mathbb{E}_k\|x^k - x^*\|^2 + \mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2] - 2\mathbb{E}_k[\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle] \\
&\quad -2\mathbb{E}_k[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle] - 2\mathbb{E}_k[\gamma_k\langle x^k - \bar{x}^k, F(\bar{x}^k)\rangle].
\end{aligned} \tag{39}$$

By applying the update rule rule $\bar{x}^k$ and $L-$Lipschitz continuity of F, we have

$$\begin{aligned}
&-2\mathbb{E}_k[\gamma_k\langle x^k - \bar{x}^k, F(\bar{x}^k)\rangle] \\
&= -2\mathbb{E}_k[\gamma_k\langle x^k - \bar{x}^k, F(\bar{x}^k) - F(x^k) + F(x^k)\rangle] \\
&\leq 2L\mathbb{E}_k[\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2] - 2\mathbb{E}_k[\gamma_k\alpha_k\langle \tilde{F}(x^k), F(x^k)\rangle] \\
&= 2L\mathbb{E}_k[\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2] - 2\mathbb{E}_k[\gamma_k\alpha_k\langle \tilde{F}(x^k) - F(x^k), F(x^k)\rangle] \\
&\quad -2\mathbb{E}_k[\gamma_k\alpha_k\|F(x^k)\|^2].
\end{aligned} \tag{40}$$

By plugging inequality (40) into (39), we obtain the first result (34). The second result follows by a nearly identical argument. □

Given this result, we proceed by bounding the three main error terms present in the results of Lemma C.2 one at a time. In order to bound these error terms, we make heavy use of Young's inequality, i.e.

$$2ab \leq \frac{1}{\rho}a^2 + \rho b^2, \quad \forall \rho > 0 \tag{41}$$

as well as

$$a^2 - b^2 = 2a(a-b) - (a-b)^2 \leq 2(a-b)a. \tag{42}$$

In addition, we also make repeated use of Hölder's inequality:

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \tag{43}$$

C.2 Bounding Error Term I

In order to bound the first error term, we consider each case individually.

Lemma C.3 (Error Term I under the Error Bound Condition.)**.** *Let Assumptions 1.1 (a), (b), 3.1, 3.2 and 3.3 with $\sigma_1 = 0$ in Assumption 3.2. Then, for any $k \geq 0$ and $x^* \in S^*$:*

$$\begin{aligned}
&-\mathbb{E}_k\left[2\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\
&\leq \frac{\alpha_{k-1}\gamma_{k-1}}{7q}\|x^k - x^*\|^2 + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right],
\end{aligned} \tag{44}$$

where $q > 0$ is the constant in the error bound condition.

Proof. By introducing the de-correlated stepsize γ_{k-1} and applying Assumption 3.1, we have:

$$\begin{aligned} -2\mathbb{E}_k\left[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] &= -2\mathbb{E}_k\left[(\gamma_k - \gamma_{k-1})\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\ &\leq 2\mathbb{E}_k\left[|\gamma_k - \gamma_{k-1}|\,\|x^k - x^*\|\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|\right]. \end{aligned} \tag{45}$$

Next, we bound the difference between the step sizes as:

$$\begin{aligned} |\gamma_k - \gamma_{k-1}| &= \eta\left|\frac{1}{\bar{b}_{k+1}} - \frac{1}{\bar{b}_k}\right| \\ &= \eta\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_k\bar{b}_{k+1}(\bar{b}_k + \bar{b}_{k+1})} \\ &\leq \eta\frac{\sqrt{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}}{\bar{b}_k(\bar{b}_k + \bar{b}_{k+1})}. \end{aligned} \tag{46}$$

Applying Young's inequality (41) with the choice of

$$a = \frac{\eta\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|\|x^k - x^*\|}{\bar{b}_k^{3/4}}, \quad b = \frac{\sqrt{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}}{\bar{b}_k^{1/4}(\bar{b}_k + \bar{b}_{k+1})}, \quad \rho = 7q\sigma_0^2$$

we have,

$$\begin{aligned} &2\mathbb{E}_k\left[\eta\frac{\sqrt{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}\|x^k - x^*\|\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\bar{b}_k(\bar{b}_k + \bar{b}_{k+1})}\right] \\ &\leq \mathbb{E}_k\left[\frac{\eta^2\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|^2\|x^k - x^*\|^2}{7q\sigma_0^2\bar{b}_k^{3/2}}\right] + \mathbb{E}_k\left[\frac{7q\sigma_0^2(\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2)}{\bar{b}_k^{1/2}(\bar{b}_k + \bar{b}_{k+1})^2}\right] \\ &\leq \frac{\eta^2\|x^k - x^*\|^2}{7q\bar{b}_k^{3/2}} + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right] \\ &\leq \frac{\eta^2\|x^k - x^*\|^2}{7q\bar{b}_k b_k} + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right], \end{aligned} \tag{47}$$

where we use the bounded variance assumption and the fact that $\bar{b}_k^{1/2} \geq b_k$ to derive the second and third inequalities. Combining results the from (45) to (47), we obtain:

$$\begin{aligned} &-\mathbb{E}_k\left[2\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\ &\leq \frac{\alpha_{k-1}\gamma_{k-1}}{7q}\|x^k - x^*\|^2 + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right]. \end{aligned} \tag{48}$$

□

Lemma C.4 (Error Term I under Star-strong Monotonicity.)**.** *Let Assumptions 1.1 (a), (b), 3.1, 3.2 and 3.4 hold and let* $\tilde{\mathcal{P}}_k := \dfrac{\|F(x^k)\|^2}{\bar{b}_{k+1}}$*. Then, for any* $k \geq 1$ *and* $x^* \in S^*$*,*

$$\begin{aligned} &-2\mathbb{E}_k\left[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\ &\leq \frac{m\eta\|x^k - x^*\|^2}{4\bar{b}_k} + \frac{4\eta\sigma_0^2}{m}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + \frac{8L^2\eta^3\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \\ &\quad + \frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\tilde{\mathcal{P}}_{k-1} - \tilde{\mathcal{P}}_k\right] + \frac{\eta^2\|F(x^k)\|^2}{4b_k\bar{b}_k} + \frac{256\eta^2L^2\sigma_1^4}{m^2b_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]. \end{aligned} \tag{49}$$

In addition, when $k = 0$,

$$\begin{aligned}&-2\mathbb{E}_0\left[\gamma_0\langle x^0 - x^*, \tilde{F}(\bar{x}^0) - F(\bar{x}^0)\rangle\right]\\&\leq \frac{m\eta\|x^0 - x^*\|^2}{4\bar{b}_0} + \frac{4\eta\sigma_0^2}{m}\mathbb{E}_0\left[\frac{1}{\bar{b}_0} - \frac{1}{\bar{b}_1}\right] + \frac{8L^2\eta^3\sigma_1^2}{m}\mathbb{E}_0\left[\frac{\|\tilde{F}(x^0)\|^2}{b_1^4}\right] + \frac{8\eta\sigma_1^2}{mb_0^2}\|F(x^0)\|^2.\end{aligned} \tag{50}$$

Proof. We introduce a new de-correlated step size $\hat{\gamma}_k$, defined by

$$\hat{\gamma}_k := \frac{\eta}{\hat{b}_k} = \frac{\eta}{\sqrt{\bar{b}_k^2 + \|\tilde{F}(x^k)\|^2}}.$$

Then, by the tower property of expectation and Assumption 3.1,

$$\begin{aligned}-2\mathbb{E}_k\left[\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] &= -2\mathbb{E}_k\left[(\gamma_k - \hat{\gamma}_k)\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right]\\&\leq 2\mathbb{E}_k\left[|\gamma_k - \hat{\gamma}_k|\,\|x^k - x^*\|\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|\right].\end{aligned} \tag{51}$$

Now, we can bound the difference between the stepsizes as follows,

$$\begin{aligned}|\gamma_k - \hat{\gamma}_k| &= \eta\left|\frac{1}{\hat{b}_k} - \frac{1}{\bar{b}_{k+1}}\right|\\&= \eta\frac{\|\tilde{F}(\bar{x}^k)\|^2}{\hat{b}_k\bar{b}_{k+1}(\hat{b}_k + \bar{b}_{k+1})}\\&\leq \eta\frac{\|\tilde{F}(\bar{x}^k)\|}{\hat{b}_k(\hat{b}_k + \bar{b}_{k+1})}.\end{aligned} \tag{52}$$

Then,

$$\begin{aligned}&2\mathbb{E}_k\left[\eta\frac{\|\tilde{F}(\bar{x}^k)\|\|x^k - x^*\|\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\hat{b}_k(\hat{b}_k + \bar{b}_{k+1})}\right]\\&= 2\mathbb{E}_k\left[\eta\frac{\|x^k - x^*\|}{\hat{b}_k}\mathbb{E}\left[\frac{\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|\|\tilde{F}(\bar{x}^k)\|}{\hat{b}_k + \bar{b}_{k+1}} \mid \bar{\mathcal{F}^k}\right]\right]\\&\overset{(i)}{\leq} \frac{\eta m}{4}\mathbb{E}_k\left[\frac{\|x^k - x^*\|^2}{\hat{b}_k}\right] + \mathbb{E}_k\left[\frac{4\eta}{m\hat{b}_k}\mathbb{E}\left[\frac{\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|\|\tilde{F}(\bar{x}^k)\|}{\hat{b}_k + \bar{b}_{k+1}} \mid \bar{\mathcal{F}^k}\right]^2\right]\\&\overset{(ii)}{\leq} \frac{\eta m}{4}\mathbb{E}_k\left[\frac{\|x^k - x^*\|^2}{\hat{b}_k}\right] + \mathbb{E}_k\left[\frac{4\eta}{m}\left(\sigma_0^2 + \sigma_1^2\|F(\bar{x}^k)\|^2\right)\left(\frac{1}{\hat{b}_k} - \frac{1}{\bar{b}_{k+1}}\right)\right]\\&\overset{(iii)}{\leq} \frac{\eta m\|x^k - x^*\|^2}{4\bar{b}_k} + \frac{4\eta\sigma_0^2}{m}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + \frac{4\eta\sigma_1^2}{m}\mathbb{E}_k\left[\|F(\bar{x}^k)\|^2\left(\frac{1}{\hat{b}_k} - \frac{1}{\bar{b}_{k+1}}\right)\right].\end{aligned} \tag{53}$$

To derive inequality (i), we apply Young's inequality (41) with the choice of

$$\rho = \frac{4\eta}{m} > 0, \quad a = \frac{\eta\|x^k - x^*\|}{\hat{b}_k^{1/2}}, \quad b = \frac{\|\tilde{F}(\bar{x}^k)\|\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|}{\hat{b}_k^{1/2}(\hat{b}_k + \bar{b}_{k+1})}.$$

Inequality (ii) comes from Hölder's inequality (43), the affine variance assumption, and the identity that

$$\frac{\|\tilde{F}(\bar{x}^k)\|^2}{\hat{b}_k\bar{b}_{k+1}(\hat{b}_k + \bar{b}_{k+1})} = \frac{1}{\hat{b}_k} - \frac{1}{\bar{b}_{k+1}},$$

while final inequality follows from the fact that $\hat{b}_k \geq \bar{b}_k$.

In addition,

$$\begin{aligned}
&\frac{4\eta\sigma_1^2}{m}\mathbb{E}_k\left[\|F(\bar{x}^k)\|^2\left(\frac{1}{\hat{b}_k}-\frac{1}{\bar{b}_{k+1}}\right)\right]\\
&\leq \frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)-F(x^k)\|^2}{\hat{b}_k}\right]+\frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\|F(x^k)\|^2\left(\frac{1}{\hat{b}_k}-\frac{1}{\bar{b}_{k+1}}\right)\right]\\
&\overset{(i)}{\leq} \frac{8L^2\eta^3\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right]+\frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\|F(x^k)\|^2\left(\frac{1}{\hat{b}_k}-\frac{1}{\bar{b}_{k+1}}\right)\right]\\
&\overset{(ii)}{\leq} \frac{8L^2\eta^3\eta\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right]+\frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\tilde{\mathcal{P}}_{k-1}-\tilde{\mathcal{P}}_k\right]\\
&\quad+\frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|F(x^k)\|^2-\|F(x^{k-1})\|^2}{\hat{b}_k}\right].
\end{aligned}\tag{54}$$

Here, (i) follows from the fact that $\hat{b}_k \geq b_{k+1}^2$ and (ii) comes from the definition of potential function $\mathcal{P}_k$ and $\hat{b}_k \geq \bar{b}_k$. Similarly, we apply Young's inequality (41) and (42) to handle the last term in (54),

$$\begin{aligned}
&\frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|F(x^k)\|^2-\|F(x^{k-1})\|^2}{\hat{b}_k}\right]\\
&\leq \frac{16\eta\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|F(x^k)\|\|F(x^k)-F(x^{k-1})\|}{\hat{b}_k}\right]\\
&= \frac{16\eta^2L\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|F(x^k)\|\|\tilde{F}(\bar{x}^{k-1})\|}{\hat{b}_k\bar{b}_k}\right]\\
&\leq \frac{\eta^2\|F(x^k)\|^2}{4\bar{b}_k^{3/2}}+\frac{256\eta^2L^2\sigma_1^4}{m^2}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^{1/2}\hat{b}_k^2}\right]\\
&\leq \frac{\eta^2\|F(x^k)\|^2}{4b_k\bar{b}_k}+\frac{256\eta^2L^2\sigma_1^4}{m^2b_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right],
\end{aligned}\tag{55}$$

where the last inequality comes from the fact that $\bar{b}_k \geq b_k^2, \hat{b}_k \geq \bar{b}_k$, and $\hat{b}_k \geq b_0^2$.

Combining the results from (53) to (55), we obtain:

$$\begin{aligned}
&-2\mathbb{E}_k\left[\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle\right]\\
&\leq \frac{m\eta\|x^k-x^*\|^2}{4\bar{b}_k}+\frac{4\eta\sigma_0^2}{m}\mathbb{E}_k\left[\frac{1}{\bar{b}_k}-\frac{1}{\bar{b}_{k+1}}\right]+\frac{8L^2\eta^3\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right]\\
&\quad+\frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\tilde{\mathcal{P}}_{k-1}-\tilde{\mathcal{P}}_k\right]+\frac{\eta^2\|F(x^k)\|^2}{4b_k\bar{b}_k}+\frac{256\eta^2L^2\sigma_1^4}{m^2b_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right].
\end{aligned}\tag{56}$$

The result when $k=0$ follows by terminating the proof after inequality (i) in (54). □

Lemma C.5 (Error Term I under Sub-Weibull Noise.)**.** *Let Assumptions 1.1 (a) and (b), 3.1, 3.2 and A.1 hold. Let β_k be defined in* (35)*, let $N \geq 1$, and let*

$$T_k := \frac{1}{\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})},\quad \bar{\mathcal{P}}_k := \frac{\|F(\bar{x}^k)\|^2}{\bar{b}_{k+1}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})},\quad \ell(N,\delta) := \frac{\log^{2\theta}\left(\frac{Ne}{\delta}\right)}{\Gamma(2\theta+1)e}.\tag{57}$$

Then, for any $x^* \in S^*$, with probability at least $1-\delta$:

$$\begin{aligned}
&-2\mathbb{E}_k\left[\beta_{k+1}\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\
&\leq \mathbb{E}_k\left[(\beta_k - \beta_{k+1})\left\|x^k - x^*\right\|^2\right] + \frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_k\left[T_{k-1} - T_k\right] \\
&\quad + \frac{4\eta^2\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\bar{\mathcal{P}}_{k-1} - \bar{\mathcal{P}}_k\right] + \frac{4\rho\eta^3 L\ell(N,\delta)\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{\bar{b}_k^2}\right] \\
&\quad + \frac{4\rho\eta^3 L\ell(N,\delta)\sigma_1^2}{\bar{b}_0^2}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{24\eta^5 L^3\ell(N,\delta)\sigma_1^2}{\rho\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \\
&\quad + \frac{24\eta^3 L\ell(N,\delta)\sigma_1^2}{\rho\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(x^k)\|^2}{\bar{b}_k^{1/2}b_k(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})}\right] + \frac{4\rho\eta^3 L\ell(N,\delta)\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right],
\end{aligned} \tag{58}$$

holds for all $k \in \{1,\ldots,N-1\}$, where $\rho > 0$ is a constant measurable to $\mathcal{F}^k$. In addition, we have,

$$\begin{aligned}
-2\mathbb{E}_0\left[\beta_1\gamma_0\langle x^0 - x^*, \tilde{F}(\bar{x}^0) - F(\bar{x}^0)\rangle\right] \quad \leq \quad & \mathbb{E}_k\left[(\beta_0 - \beta_1)\left\|x^0 - x^*\right\|^2\right] + \frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_0\left[T_{-1} - T_0\right] \\
& + 8\eta^2\ell(N,\delta)\sigma_1^2\left(\frac{L^2\eta^2}{\bar{b}_0} + \frac{\|F(x^0)\|^2}{\bar{b}_0^2}\right),
\end{aligned} \tag{59}$$

also holds with probability at least $1-\delta$.

Proof. Let $\hat{b}_k^2 := \bar{b}_k^2 + \|\tilde{F}(x^k)\|^2$, which is measurable to $\bar{\mathcal{F}}_k$. Then,

$$\begin{aligned}
&-\mathbb{E}_k\left[2\beta_{k+1}\gamma_k\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\
&= -\mathbb{E}_k\left[\frac{2\eta}{\bar{b}_{k+1}^{1/2}\left(2\bar{b}_{k+1}^{1/2} - \bar{b}_0^{1/2}\right)}\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\
&= -\mathbb{E}_k\left[\left(\frac{2\eta}{\bar{b}_{k+1}^{1/2}\left(2\bar{b}_{k+1}^{1/2} - \bar{b}_0^{1/2}\right)} - \frac{2\eta}{\hat{b}_k^{1/2}\left(2\hat{b}_k^{1/2} - \bar{b}_0^{1/2}\right)}\right)\langle x^k - x^*, \tilde{F}(\bar{x}^k) - F(\bar{x}^k)\rangle\right] \\
&\leq \mathbb{E}_k\left[\left|\frac{2\eta}{\bar{b}_{k+1}^{1/2}\left(2\bar{b}_{k+1}^{1/2} - \bar{b}_0^{1/2}\right)} - \frac{2\eta}{\hat{b}_k^{1/2}\left(2\hat{b}_k^{1/2} - \bar{b}_0^{1/2}\right)}\right|\left\|x^k - x^*\right\|\left\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\right\|\right] \\
&= \mathbb{E}_k\left[A_k\|x^k - x^*\|\|\tilde{F}(\bar{x}^k) - F(\bar{x}^k)\|\right],
\end{aligned} \tag{60}$$

where we define

$$A_k := \frac{2\eta}{\hat{b}_k^{1/2}\left(2\hat{b}_k^{1/2} - \bar{b}_0^{1/2}\right)} - \frac{2\eta}{\bar{b}_{k+1}^{1/2}\left(2\bar{b}_{k+1}^{1/2} - \bar{b}_0^{1/2}\right)}. \tag{61}$$

In addition, let

$$\lambda_k := \frac{2\bar{b}_k^{1/2}}{2\bar{b}_k^{1/2} - \bar{b}_0^{1/2}} - \frac{2\bar{b}_{k+1}^{1/2}}{2\bar{b}_{k+1}^{1/2} - \bar{b}_0^{1/2}} > 0. \tag{62}$$

Then, by applying Young's inequality with λ_k, we obtain:

$$\begin{aligned}
&\mathbb{E}_k\left[A_k\|x^k-x^*\|\|\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\|\right]\\
&\leq \mathbb{E}_k\left[\frac{\lambda_k}{2}\|x^k-x^*\|^2\right]+\mathbb{E}_k\left[\frac{1}{2\lambda_k}A_k^2\|\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\|^2\right]\\
&\leq \mathbb{E}_k\left[\frac{\lambda_k}{2}\|x^k-x^*\|^2\right]+\mathbb{E}_k\left[\frac{1}{2\lambda_k}A_k^2\left(\frac{\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2}{\Gamma(2\theta+1)e}\log^{2\theta}\left(\frac{Ne}{\delta}\right)\right)\right],
\end{aligned} \tag{63}$$

where the last inequality holds for all $k\in\{0,\ldots,N-1\}$ with probability at least $1-\delta$ by the sub-Weibull noise assumption and Lemma A.4.

Now, we simplify each term on the right side one by one,

$$\begin{aligned}
\mathbb{E}_k\left[\frac{\lambda_k}{2}\|x^k-x^*\|^2\right] &= \mathbb{E}_k\left[\left(\frac{\bar{b}_k^{1/2}}{2\bar{b}_k^{1/2}-\bar{b}_0^{1/2}}-\frac{\bar{b}_{k+1}^{1/2}}{2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2}}\right)\|x^k-x^*\|^2\right]\\
&= \mathbb{E}_k\left[(\beta_k-\beta_{k+1})\|x^k-x^*\|^2\right],
\end{aligned} \tag{64}$$

and

$$\begin{aligned}
&\mathbb{E}_k\left[\frac{1}{2\lambda_k}A_k^2\left[\frac{\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2}{\Gamma(2\theta+1)e}\log^{2\theta}\left(\frac{Ne}{\delta}\right)\right]\right]\\
&=\frac{\log^{2\theta}\left(\frac{Ne}{\delta}\right)}{2\Gamma(2\theta+1)e}\mathbb{E}_k\left[(\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2)\frac{A_k^2}{\lambda_k}\right]\\
&\leq\frac{\log^{2\theta}\left(\frac{Ne}{\delta}\right)}{\Gamma(2\theta+1)e\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{2\eta}{\hat{b}_k^{1/2}}\left(\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2\right)A_k\right]\\
&=\frac{4\eta^2\log^{2\theta}\left(\frac{Ne}{\delta}\right)}{\Gamma(2\theta+1)e\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2}{\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right].
\end{aligned} \tag{65}$$

The inequality comes from the relationship between $\beta_k-\beta_{k+1}$ and A_k defined in (61), which is:

$$\begin{aligned}
\frac{\lambda_k}{2A_k}=\frac{\beta_k-\beta_{k+1}}{A_k} &= \frac{\bar{b}_0^{1/2}\left(\bar{b}_{k+1}^{1/2}-\bar{b}_k^{1/2}\right)\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}\left(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2}\right)}{2\eta\left(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2}\right)\left(2\left(\bar{b}_{k+1}-\hat{b}_k\right)+\bar{b}_0^{1/2}\left(\hat{b}_k^{1/2}-\bar{b}_{k+1}^{1/2}\right)\right)}\\
&\overset{(i)}{\geq} \frac{\bar{b}_0^{1/2}\left(\bar{b}_{k+1}^{1/2}-\bar{b}_k^{1/2}\right)\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}}{4\eta\left(\bar{b}_{k+1}-\hat{b}_k\right)}\\
&= \frac{\bar{b}_0^{1/2}\left(\bar{b}_{k+1}^{1/2}-\bar{b}_k^{1/2}\right)\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}}{4\eta\left(\bar{b}_{k+1}^{1/2}-\hat{b}_k^{1/2}\right)\left(\bar{b}_{k+1}^{1/2}+\hat{b}_k^{1/2}\right)}\\
&\overset{(ii)}{\geq} \frac{\bar{b}_0^{1/2}\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}}{4\eta\left(\bar{b}_{k+1}^{1/2}+\hat{b}_k^{1/2}\right)}\\
&\overset{(iii)}{\geq} \frac{\bar{b}_0^{1/2}\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}}{8\eta\bar{b}_{k+1}^{1/2}}\\
&= \frac{\bar{b}_0^{1/2}\hat{b}_k^{1/2}}{8\eta},
\end{aligned} \tag{66}$$

where to derive inequality (i), we apply the fact that $\bar{b}_k^{1/2}\leq\hat{b}_k^{1/2}\leq\bar{b}_{k+1}^{1/2}$, the second inequality (ii) comes from the fact that $\hat{b}_k^{1/2}\geq\bar{b}_k^{1/2}$, and the last inequality (iii) is due to the identity that $\bar{b}_{k+1}^{1/2}+\hat{b}_k^{1/2}\leq 2\bar{b}_{k+1}^{1/2}$.

Combining results from (64) and (65) and recalling the definition of $\ell(N,\delta)$, we obtain that,

$$\begin{aligned}
&\mathbb{E}_k\left[-2\beta_{k+1}\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle\right],\\
\leq\ &\mathbb{E}_k\left[(\beta_k-\beta_{k+1})\|x^k-x^*\|^2\right]+\frac{4\eta^2\ell(N,\delta)}{\bar{b}_0^{1/2}}\mathbb{E}_k\left(\frac{\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{\sigma_0^2+\sigma_1^2\|F(\bar{x}^k)\|^2}{\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right)\\
\leq\ &\mathbb{E}_k\left[(\beta_k-\beta_{k+1})\|x^k-x^*\|^2\right]+\frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k^{1/2}(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{1}{\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right]\\
&+\frac{4\eta^2\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{\|F(\bar{x}^k)\|^2}{\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right],
\end{aligned}\tag{67}$$

in which the first two terms on the right-hand side gives us desirable telescoping sums. Now we need continue processing the last term from (67).

$$\begin{aligned}
&\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{\|F(\bar{x}^k)\|^2}{\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right]\\
&=\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^{k-1})\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{\|F(\bar{x}^k)\|^2}{\hat{b}_k^{1/2}\bar{b}_{k+1}^{1/2}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right]+\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2-\|F(\bar{x}^{k-1})\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\overset{(i)}{\leq}\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^{k-1})\|^2}{\bar{b}_k(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2})}-\frac{\|F(\bar{x}^k)\|^2}{\bar{b}_{k+1}(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2})}\right]+\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2-\|F(\bar{x}^{k-1})\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\overset{(ii)}{=}\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\bar{\mathcal{P}}_{k-1}-\bar{\mathcal{P}}_k\right]+\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2-\|F(\bar{x}^{k-1})\|^2}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\overset{(iii)}{\leq}\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\bar{\mathcal{P}}_{k-1}-\bar{\mathcal{P}}_k\right]+\frac{8\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|(\|F(\bar{x}^k)\|-\|F(\bar{x}^{k-1})\|)}{\hat{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\overset{(iv)}{\leq}\frac{4\eta^2\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\bar{\mathcal{P}}_{k-1}-\bar{\mathcal{P}}_k\right]+\frac{8\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|\|\tilde{F}(\bar{x}^{k-1})\|}{\hat{b}_k\bar{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}+\frac{\|F(\bar{x}^k)\|\|\tilde{F}(x^{k-1})\|}{\hat{b}_kb_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\qquad+\frac{8\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|\|\tilde{F}(x^k)\|}{\hat{b}_kb_{k+1}(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right],
\end{aligned}\tag{68}$$

here we use the identity $\bar{b}_k\leq\hat{b}_k\leq\bar{b}_{k+1}$ to derive inequality (i), equation (ii) follows from the definition of $\bar{\mathcal{P}}_k$, the inequality (iii) follows by (42), while the final inequality is from L-Lipschitz continuity of F, the triangle inequality and the fact that

$$\bar{x}^k-\bar{x}^{k-1}=\frac{\eta}{b_k}\tilde{F}(x^{k-1})-\frac{\eta}{\bar{b}_k}\tilde{F}(\bar{x}^{k-1})-\frac{\eta}{b_{k+1}}\tilde{F}(x^k).$$

To handle the last three terms from (68), we apply Young's inequality (41) with positive ρ for three times:

$$\begin{aligned}
&\frac{8\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|\|\tilde{F}(\bar{x}^{k-1})\|}{\hat{b}_k\bar{b}_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}+\frac{\|F(\bar{x}^k)\|\|\tilde{F}(x^{k-1})\|}{\hat{b}_kb_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}+\frac{\|F(\bar{x}^k)\|\|\tilde{F}(x^k)\|}{\hat{b}_kb_{k+1}(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\leq\frac{12\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2}{\rho\hat{b}_k^{1/2}b_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]+\frac{4\rho\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\hat{b}_k^{3/2}\bar{b}_k^{3/2}(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\quad+\frac{4\rho\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{\hat{b}_k^{3/2}b_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]+\frac{4\rho\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{\hat{b}_k^{3/2}b_{k+1}(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\leq\frac{12\eta^3L\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2}{\rho\hat{b}_k^{1/2}b_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]+\frac{4\rho\eta^3L\sigma_1^2}{\bar{b}_0^2}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&\quad+\frac{4\rho\eta^3L\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{\bar{b}_k^2}\right]+\frac{4\rho\eta^3L\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right],
\end{aligned}\tag{69}$$

where the final inequality follows due to the construction of step sizes $\bar{b}_k, b_k, \hat{b}_k$, so that $\hat{b}_k^2 \geq b_{k+1}^4$ and $\hat{b}_k \geq \bar{b}_k \geq b_k^2$. Next, by applying the triangle inequality and the L-Lipschitz continuity assumption of F, we have

$$\begin{aligned}
&\frac{12\eta^3 L\sigma_1^2}{\rho\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)\|^2}{\hat{b}_k^{1/2}b_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\leq \frac{24\eta^3 L\sigma_1^2}{\rho\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(\bar{x}^k)-F(x^k)\|^2}{\hat{b}_k^{1/2}b_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}+\frac{\|F(x^k)\|^2}{\hat{b}_k^{1/2}b_k(2\hat{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]\\
&\leq \frac{24\eta^3 L\sigma_1^2}{\rho\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\eta^2L^2\|\tilde{F}(x^k)\|^2}{b_{k+1}^4\bar{b}_0^{1/2}}+\frac{\|F(x^k)\|^2}{\bar{b}_k^{1/2}b_k(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2})}\right].
\end{aligned} \tag{70}$$

Plugging these results into (67), with T_k defined in (57), it follows that

$$\begin{aligned}
&\mathbb{E}_k\left[-2\beta_{k+1}\gamma_k\langle x^k-x^*,\tilde{F}(\bar{x}^k)-F(\bar{x}^k)\rangle\right]\\
&\leq \mathbb{E}_k\left[(\beta_k-\beta_{k+1})\|x^k-x^*\|^2\right]+\frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_k\left[T_{k-1}-T_k\right]\\
&\quad+\frac{4\eta^2\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\bar{\mathcal{P}}_{k-1}-\bar{\mathcal{P}}_k\right]+\frac{4\rho\eta^3L\ell(N,\delta)\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&\quad+\frac{4\rho\eta^3L\ell(N,\delta)\sigma_1^2}{\bar{b}_0^2}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]+\frac{24\eta^5L^3\ell(N,\delta)\sigma_1^2}{\rho\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right]\\
&\quad+\frac{24\eta^3L\ell(N,\delta)\sigma_1^2}{\rho\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(x^k)\|^2}{\bar{b}_k^{1/2}b_k(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2})}\right]+\frac{4\rho\eta^3L\ell(N,\delta)\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right].
\end{aligned} \tag{71}$$

For the case where $k=0$, by (67), we have

$$\begin{aligned}
&\mathbb{E}_0\left[-2\beta_1\gamma_0\langle x^0-x^*,\tilde{F}(\bar{x}^0)-F(\bar{x}^0)\rangle\right],\\
\leq\ &\mathbb{E}_k\left[(\beta_0-\beta_1)\|x^0-x^*\|^2\right]+\frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_0\left[\frac{1}{\bar{b}_0^{1/2}(2\bar{b}_0^{1/2}-\bar{b}_0^{1/2})}-\frac{1}{\bar{b}_1^{1/2}(2\bar{b}_1^{1/2}-\bar{b}_0^{1/2})}\right]\\
&+\frac{4\eta^2\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_0\left[\frac{\|F(\bar{x}^0)\|^2}{\hat{b}_0(2\hat{b}_0^{1/2}-\bar{b}_0^{1/2})}\right]\\
\leq\ &\mathbb{E}_k\left[(\beta_0-\beta_1)\|x^0-x^*\|^2\right]+\frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_0\left[\frac{1}{\bar{b}_0^{1/2}(2\bar{b}_0^{1/2}-\bar{b}_0^{1/2})}-\frac{1}{\bar{b}_1^{1/2}(2\bar{b}_1^{1/2}-\bar{b}_0^{1/2})}\right]\\
&+\frac{8\eta^2\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_0\left[\frac{L^2\eta^2\|\tilde{F}(x^0)\|^2}{\hat{b}_0^{3/2}b_1^2}+\frac{\|F(x^0)\|^2}{\hat{b}_0^{3/2}}\right]\\
\leq\ &\mathbb{E}_k\left[(\beta_0-\beta_1)\|x^0-x^*\|^2\right]+\frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_0\left[\frac{1}{\bar{b}_0^{1/2}(2\bar{b}_0^{1/2}-\bar{b}_0^{1/2})}-\frac{1}{\bar{b}_1^{1/2}(2\bar{b}_1^{1/2}-\bar{b}_0^{1/2})}\right]\\
&+8\eta^2\ell(N,\delta)\sigma_1^2\left(\frac{L^2\eta^2}{\bar{b}_0}+\frac{\|F(x^0)\|^2}{\bar{b}_0^2}\right),
\end{aligned} \tag{72}$$

where we used Lipschitz continuity, the definition of $\bar{x}^0$ and the construction of our stepsizes to derive the final two inequalities. □

C.3 Bounding Error Term II

Now, we focus our attention on bounding the second error term.

Lemma C.6. *Let Assumptions 1.1 (a), (b), 3.1, and 3.2 hold and let $\{x^k\}$ be generated by Algorithm 2. Let*

$$\mathcal{P}_k := \frac{\|F(x^k)\|^2}{b_{k+1}^2} \quad \textit{and} \quad \tilde{\mathcal{P}}_k := \frac{\|F(x^k)\|^2}{\bar{b}_{k+1}}.$$

Then, for all $k \geq 1$,

$$\begin{aligned}
&\mathbb{E}_k[2\alpha_k\gamma_k\langle\tilde{F}(x^k) - F(x^k), F(x^k)\rangle] \\
&\leq \frac{6\alpha_{k-1}\gamma_{k-1}}{\rho}\|F(x^k)\|^2 + \frac{\eta^2\sigma_0^2\rho}{b_0}\mathbb{E}_k\left[\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right] + \eta^2\sigma_1^2\rho\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right] \\
&+ \frac{2\eta^2\sigma_0^2\rho}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 2\eta^2\sigma_1^2\rho\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right] \\
&+ \frac{\eta^4L^2\sigma_1^4\rho^3}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{2\eta^4L^2\sigma_1^4\rho^3}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right].
\end{aligned} \tag{73}$$

where $\rho > 0$ is measurable to $\mathcal{F}_k$. Similarly,

$$\begin{aligned}
&\mathbb{E}_k[2\beta_{k+1}\alpha_k\gamma_k\langle\tilde{F}(x^k) - F(x^k), F(x^k)\rangle] \\
&\leq \frac{6\eta^2\|F(x^k)\|^2}{\lambda\bar{b}_k^{1/2}(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})b_k} + \frac{2\lambda\eta^2\sigma_0^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right] + 2\lambda\sigma_1^2\eta^2\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right] \\
&+ \frac{8L^2\lambda^3\sigma_1^4\eta^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{8\lambda\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 8\lambda\sigma_1^2\eta^2\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right] \\
&+ \frac{128L^2\lambda^3\sigma_1^4\eta^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right],
\end{aligned} \tag{74}$$

where $\lambda > 0$ is measurable to $\mathcal{F}_k$.

In addition, when $k = 0$,

$$\begin{aligned}
\mathbb{E}_0\left[2\alpha_0\gamma_0\langle\tilde{F}(x^0) - F(x^0), F(x^0)\rangle\right] &\leq \frac{3\eta^2}{\rho\bar{b}_0b_0}\|F(x^0)\|^2 + \frac{\eta^2\rho}{\bar{b}_0^{1/2}}\left(\sigma_0^2 + \sigma_1^2\|F(x^0)\|^2\right)\mathbb{E}_0\left[\frac{1}{b_0^2} - \frac{1}{b_1^2}\right] \\
&+ \frac{2\eta^2\rho}{b_0}\left(\sigma_0^2 + \sigma_1^2\|F(x^0)\|^2\right)\mathbb{E}_k\left[\frac{1}{\bar{b}_0} - \frac{1}{\bar{b}_1}\right].
\end{aligned} \tag{75}$$

where $\rho > 0$ is measurable to $\mathcal{F}_k$. Similarly,

$$\begin{aligned}
\mathbb{E}_0\left[2\beta_1\alpha_0\gamma_0\langle\tilde{F}(x^0) - F(x^0), F(x^0)\rangle\right] &\leq \frac{3\eta^2}{\lambda\bar{b}_0b_0}\|F(x^0)\|^2 + \frac{2\eta^2\lambda}{\bar{b}_0^{1/2}}\left(\sigma_0^2 + \sigma_1^2\|F(x^0)\|^2\right)\mathbb{E}_0\left[\frac{1}{b_0^2} - \frac{1}{b_1^2}\right] \\
&+ \frac{4\eta^2\lambda}{b_0}\left(\sigma_0^2 + \sigma_1^2\|F(x^0)\|^2\right)\mathbb{E}_k\left[\frac{1}{\bar{b}_0} - \frac{1}{\bar{b}_1}\right].
\end{aligned} \tag{76}$$

where $\lambda > 0$ is measurable to $\mathcal{F}_k$.

Proof. By introducing the decorrelated stepsize $\gamma_{k-1}\alpha_{k-1}$ and using Assumption 3.1, we have:

$$\begin{aligned}
&\mathbb{E}_k\left[2\gamma_k\alpha_k\langle\tilde{F}(x^k) - F(x^k), F(x^k)\rangle\right] \\
&= \mathbb{E}_k\left[2(\gamma_k\alpha_k - \gamma_{k-1}\alpha_{k-1})\langle\tilde{F}(x^k) - F(x^k), F(x^k)\rangle\right] \\
&\leq \mathbb{E}_k\left[2|\gamma_k\alpha_k - \gamma_{k-1}\alpha_{k-1}||\langle\tilde{F}(x^k) - F(x^k), F(x^k)\rangle|\right].
\end{aligned} \tag{77}$$

From the definition of stepsizes γ_k, α_k and by applying the triangle inequality, we have

$$|\gamma_k\alpha_k - \gamma_{k-1}\alpha_{k-1}| = \eta^2\left|\frac{1}{\bar{b}_{k+1}b_{k+1}} - \frac{1}{\bar{b}_kb_k}\right| \leq \eta^2\left|\frac{1}{\bar{b}_{k+1}b_{k+1}} - \frac{1}{\bar{b}_{k+1}b_k}\right| + \eta^2\left|\frac{1}{\bar{b}_{k+1}b_k} - \frac{1}{\bar{b}_kb_k}\right|.$$

First, we bound:

$$\begin{aligned}\left|\frac{1}{\bar{b}_{k+1}b_{k+1}}-\frac{1}{\bar{b}_{k+1}b_k}\right| &= \frac{1}{\bar{b}_{k+1}}\left|\frac{1}{b_{k+1}}-\frac{1}{b_k}\right| \\ &= \frac{1}{\bar{b}_{k+1}}\left|\frac{b_k^4-b_{k+1}^4}{b_{k+1}b_k(b_{k+1}+b_k)(b_{k+1}^2+b_k^2)}\right| \\ &= \frac{1}{\bar{b}_{k+1}}\left|\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}b_k(b_{k+1}+b_k)(b_{k+1}^2+b_k^2)}\right| \\ &\leq \frac{\|\tilde{F}(x^k)\|}{\bar{b}_{k+1}b_k(b_{k+1}^2+b_k^2)},\end{aligned} \tag{78}$$

where the last inequality comes from the fact that

$$b_{k+1}(b_{k+1}+b_k) \geq b_{k+1}^2 \geq \|\tilde{F}(x^k)\|.$$

Next,

$$\begin{aligned}\left|\frac{1}{\bar{b}_{k+1}b_k}-\frac{1}{\bar{b}_k b_k}\right| &= \frac{1}{b_k}\left|\frac{1}{\bar{b}_{k+1}}-\frac{1}{\bar{b}_k}\right| \\ &= \frac{1}{b_k}\left|\frac{\bar{b}_k^2-\bar{b}_{k+1}^2}{\bar{b}_{k+1}\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\right| \\ &= \frac{1}{b_k}\left|\frac{\|\tilde{F}(\bar{x}^k)\|^2+\|\tilde{F}(x^k)\|^2}{\bar{b}_{k+1}\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\right| \\ &\leq \frac{\|\tilde{F}(\bar{x}^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}+\frac{\|\tilde{F}(x^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)},\end{aligned} \tag{79}$$

where the last inequality comes from the fact that

$$\bar{b}_{k+1} \geq \|\tilde{F}(\bar{x}^k)\|, \quad \bar{b}_{k+1} \geq \|\tilde{F}(x^k)\|.$$

Combining results from (78) and (79), we obtain,

$$|\gamma_k\alpha_k-\gamma_{k-1}\alpha_{k-1}| \leq \eta^2\frac{\|\tilde{F}(x^k)\|}{\bar{b}_{k+1}b_k(b_{k+1}^2+b_k^2)}+\eta^2\frac{\|\tilde{F}(\bar{x}^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}+\eta^2\frac{\|\tilde{F}(x^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}. \tag{80}$$

Therefore, to bound $2\,|\gamma_k\alpha_k-\gamma_{k-1}\alpha_{k-1}|\,\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle$, we start by bounding

$$2\eta^2\frac{\|\tilde{F}(x^k)\|}{\bar{b}_{k+1}b_k(b_{k+1}^2+b_k^2)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle,$$

and then bounding

$$2\eta^2\frac{\|\tilde{F}(\bar{x}^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle,$$

and lastly bounding

$$2\eta^2\frac{\|\tilde{F}(x^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle.$$

Focusing on the first term, we process it as follows:

$$\begin{aligned}
&\mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(x^k)\|}{\bar{b}_{k+1}b_k(b_{k+1}^2+b_k^2)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle\right]\\
&\leq \mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(x^k)\|\|\tilde{F}(x^k)-F(x^k)\|\|F(x^k)\|}{\bar{b}_{k+1}b_k(b_{k+1}^2+b_k^2)}\right]\\
&\overset{(i)}{\leq} \mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(x^k)\|\|\tilde{F}(x^k)-F(x^k)\|\|F(x^k)\|}{\bar{b}_k b_k(b_{k+1}^2+b_k^2)}\right]\\
&= \frac{2\eta^2\|F(x^k)\|}{\bar{b}_k b_k}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|\|\tilde{F}(x^k)-F(x^k)\|}{b_{k+1}^2+b_k^2}\right]\\
&\overset{(ii)}{\leq} \frac{\eta^2}{\rho\bar{b}_k b_k}\|F(x^k)\|^2+\frac{\eta^2\rho}{\bar{b}_k b_k}\left(\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|\|\tilde{F}(x^k)-F(x^k)\|}{b_{k+1}^2+b_k^2}\right]\right)^2\\
&\overset{(iii)}{\leq} \frac{\eta^2}{\rho\bar{b}_k b_k}\|F(x^k)\|^2+\frac{\eta^2\rho}{\bar{b}_k b_k}\left(\mathbb{E}_k\left[\|\tilde{F}(x^k)-F(x^k)\|^2\right]\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2+b_k^2)^2}\right]\right).
\end{aligned}\tag{81}$$

Here, inequality (i) is from the non-decreasing property of sequence $\{\bar{b}_k\}$ and (ii) comes from (41). The last inequality is obtained from Hölder's Inequality (43). Now from the affine variance assumption that

$$\mathbb{E}_k\left[\|\tilde{F}(x^k)-F(x^k)\|^2\right]\leq\sigma_0^2+\sigma_1^2\|F(x^k)\|^2,$$

and thus we have

$$\begin{aligned}
&\frac{\eta^2\rho}{\bar{b}_k b_k}\left(\mathbb{E}_k\left[\|\tilde{F}(x^k)-F(x^k)\|^2\right]\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2+b_k^2)^2}\right]\right)\\
&\leq \frac{\eta^2\rho}{\bar{b}_k b_k}\left(\sigma_0^2+\sigma_1^2\|F(x^k)\|^2\right)\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2+b_k^2)^2}\right]\\
&\overset{(i)}{\leq} \frac{\eta^2\rho}{\bar{b}_k^{1/2}}\left(\sigma_0^2+\sigma_1^2\|F(x^k)\|^2\right)\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2+b_k^2)b_k^2b_{k+1}^2}\right]\\
&\overset{(ii)}{=} \frac{\eta^2\rho}{\bar{b}_k^{1/2}}\left(\sigma_0^2+\sigma_1^2\|F(x^k)\|^2\right)\mathbb{E}_k\left[\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right]\\
&\overset{(iii)}{=} \frac{\eta^2\sigma_0^2\rho}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right]+\frac{\eta^2\sigma_1^2\rho}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\mathcal{P}_{k-1}-\mathcal{P}_k\right]\\
&\quad+\mathbb{E}_k\left[\frac{\eta^2\sigma_1^2\rho}{\bar{b}_k^{1/2}}\frac{\|F(x^k)\|^2-\|F(x^{k-1})\|^2}{b_k^2}\right]\\
&\overset{(iv)}{\leq} \frac{\eta^2\sigma_0^2\rho}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right]+\frac{\eta^2\sigma_1^2\rho}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\mathcal{P}_{k-1}-\mathcal{P}_k\right]\\
&\quad+\mathbb{E}_k\left[\frac{2L\eta^3\sigma_1^2\rho}{\bar{b}_k^{1/2}}\frac{\|\tilde{F}(\bar{x}^{k-1})\|\|F(x^k)\|}{\bar{b}_k b_k^2}\right].
\end{aligned}\tag{82}$$

Here, we use the fact that $\bar{b}_k^{1/2}\geq b_k$ to derive inequality (i) and the equivalence that

$$\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2+b_k^2)^2b_k^2b_{k+1}^2}=\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2},$$

to obtain equation (ii) while equation (iii) is due to our construction of potential function $\mathcal{P}_k$. The last inequality

(iv) comes from the update rule for x^k and (42). To continue, we again apply Young's inequality,

$$\begin{aligned}\frac{2L\eta^3\sigma_1^2\rho}{\bar{b}_k^{1/2}}\frac{\|\tilde{F}(\bar{x}^{k-1})\|\|F(x^k)\|}{\bar{b}_k b_k^2} &\leq \frac{\eta^2\|F(x^k)\|^2}{\rho\bar{b}_k b_k}+\frac{\eta^4L^2\sigma_1^4\rho^3\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2 b_k^3}\\ &= \frac{\alpha_{k-1}\gamma_{k-1}\|F(x^k)\|^2}{\rho}+\frac{\eta^4L^2\sigma_1^4\rho^3\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2 b_k^3}.\end{aligned} \tag{83}$$

Our next step is to bound the term $2\eta^2\frac{\|\tilde{F}(\bar{x}^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle$ in a similar approach as before. We have:

$$\begin{aligned}&\mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(\bar{x}^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle\right]\\ &\leq \mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(\bar{x}^k)\|\|\tilde{F}(x^k)-F(x^k)\|\|F(x^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\right]\\ &= \frac{2\eta^2\|F(x^k)\|}{\bar{b}_k b_k}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^k)\|\|\tilde{F}(x^k)-F(x^k)\|}{(\bar{b}_{k+1}+\bar{b}_k)}\right]\\ &\overset{(i)}{\leq} \frac{\eta^2}{\rho\bar{b}_k b_k}\|F(x^k)\|^2+\frac{\eta^2\rho}{\bar{b}_k b_k}\left(\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^k)\|\|\tilde{F}(x^k)-F(x^k)\|}{(\bar{b}_{k+1}+\bar{b}_k)}\right]\right)^2\\ &\overset{(ii)}{\leq} \frac{\eta^2}{\rho\bar{b}_k b_k}\|F(x^k)\|^2+\frac{\eta^2\rho}{\bar{b}_k b_k}\left(\mathbb{E}_k\left[\|\tilde{F}(x^k)-F(x^k)\|^2\right]\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^k)\|^2}{(\bar{b}_{k+1}+\bar{b}_k)^2}\right]\right).\end{aligned} \tag{84}$$

Here, inequality (i) is from Young's inequality (41) and (ii) comes from Hölder's Inequality (43).

Now from the affine variance assumption, we have

$$\begin{aligned}&\frac{\eta^2\rho}{\bar{b}_k b_k}\left(\mathbb{E}_k\left[\|\tilde{F}(x^k)-F(x^k)\|^2\right]\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^k)\|^2}{(\bar{b}_{k+1}+\bar{b}_k)^2}\right]\right)\\ &\leq \frac{\eta^2\rho}{\bar{b}_k b_k}\left(\sigma_0^2+\sigma_1^2\|F(x^k)\|^2\right)\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^k)\|^2}{(\bar{b}_{k+1}+\bar{b}_k)^2}\right]\\ &\leq \frac{\eta^2\rho}{b_k}\left(\sigma_0^2+\sigma_1^2\|F(x^k)\|^2\right)\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^k)\|^2+\|\tilde{F}(x^k)\|^2}{\bar{b}_k\bar{b}_{k+1}(\bar{b}_{k+1}+\bar{b}_k)}\right]\\ &\overset{(i)}{=} \frac{\eta^2\rho}{b_k}\left(\sigma_0^2+\sigma_1^2\|F(x^k)\|^2\right)\mathbb{E}_k\left[\frac{1}{\bar{b}_k}-\frac{1}{\bar{b}_{k+1}}\right]\\ &\overset{(ii)}{=} \frac{\eta^2\sigma_0^2\rho}{b_k}\mathbb{E}_k\left[\frac{1}{\bar{b}_k}-\frac{1}{\bar{b}_{k+1}}\right]+\frac{\eta^2\sigma_1^2\rho}{b_k}\mathbb{E}_k\left[\tilde{\mathcal{P}}_{k-1}-\tilde{\mathcal{P}}_k\right]\\ &\quad+\frac{\eta^2\sigma_1^2\rho}{b_k}\frac{\|F(x^k)\|^2-\|F(x^{k-1})\|^2}{\bar{b}_k}\\ &\leq \frac{\eta^2\sigma_0^2\rho}{b_k}\mathbb{E}_k\left[\frac{1}{\bar{b}_k}-\frac{1}{\bar{b}_{k+1}}\right]+\frac{\eta^2\sigma_1^2\rho}{b_k}\mathbb{E}_k\left[\tilde{\mathcal{P}}_{k-1}-\tilde{\mathcal{P}}_k\right]\\ &\quad+\frac{2L\eta^3\sigma_1^2\rho}{b_k}\frac{\|\tilde{F}(\bar{x}^{k-1})\|\|F(x^k)\|}{\bar{b}_k^2}.\end{aligned} \tag{85}$$

We apply the equivalence that

$$\frac{\|\tilde{F}(\bar{x}^k)\|^2+\|\tilde{F}(x^k)\|^2}{\bar{b}_k\bar{b}_{k+1}(\bar{b}_{k+1}+\bar{b}_k)}=\frac{1}{\bar{b}_k}-\frac{1}{\bar{b}_{k+1}},$$

to obtain equation (i) and the equality (ii) is due to our construction of potential function $\tilde{\mathcal{P}}_k$. Similarly, the last inequality is due to (42) and L-Lipschitz continuity of F. Again, we apply (41) with $\rho_4 > 0$, then:

$$\frac{2L\eta^3\sigma_1^2\rho}{b_k}\frac{\|F(\bar{x}^{k-1})\|\|F(x^k)\|}{\bar{b}_k^2} \quad \leq \quad \frac{\eta^2\|F(x^k)\|^2}{\rho\bar{b}_k b_k} + \frac{\eta^4 L^2\sigma_1^4\rho^3\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}. \tag{86}$$

By applying the same process, we can bound the term $\mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(x^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle\right]$ as follows:

$$\begin{aligned}
&\mathbb{E}_k\left[2\eta^2\frac{\|\tilde{F}(x^k)\|}{b_k\bar{b}_k(\bar{b}_{k+1}+\bar{b}_k)}\langle\tilde{F}(x^k)-F(x^k),F(x^k)\rangle\right] \\
&\quad\leq \frac{2\alpha_{k-1}\gamma_{k-1}}{\rho}\|F(x^k)\|^2 + \frac{\eta^2\sigma_0^2\rho}{b_k}\mathbb{E}_k\left[\frac{1}{\bar{b}_k}-\frac{1}{\bar{b}_{k+1}}\right] + \frac{\eta^2\sigma_1^2\rho}{b_k}\mathbb{E}_k\left[\tilde{\mathcal{P}}_{k-1}-\tilde{\mathcal{P}}_k\right] \\
&\quad\quad+\mathbb{E}_k\left[\frac{\eta^4L^2\sigma_1^4\rho^3\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right].
\end{aligned} \tag{87}$$

Combining everything, we obtain the desired results. The second result follows by a nearly identical argument as the first result combined with the fact that $\beta_k \in [1/2, 1]$ for all k.

Now, for the first case where $k = 0$, we have by (80), (81), (ii) in (82), (84), and (i) (85) as well as handling the third term in a similar manner,

$$\begin{aligned}
2\mathbb{E}_0\left[|\gamma_0\alpha_0-\gamma_{-1}\alpha_{-1}|\|\tilde{F}(x^k)-F(x^k)\|\|F(x^k)\|\right] \leq{}& \frac{3\eta^2}{\rho\bar{b}_0 b_0}\|F(x^0)\|^2 + \frac{\eta^2\rho}{\bar{b}_0^{1/2}}\left(\sigma_0^2+\sigma_1^2\|F(x^0)\|^2\right)\mathbb{E}_0\left[\frac{1}{b_0^2}-\frac{1}{b_1^2}\right] \\
&+\frac{2\eta^2\rho}{b_0}\left(\sigma_0^2+\sigma_1^2\|F(x^0)\|^2\right)\mathbb{E}_k\left[\frac{1}{\bar{b}_0}-\frac{1}{\bar{b}_1}\right].
\end{aligned}$$

A similar approach can also prove the final case. □

C.4 Bounding Error Term III

Finally, we prove a bound on the third error term.

Lemma C.7. *Let Assumptions 1.1 (a) and (b) hold and let $\{x^k\}$ be generated by Algorithm 2. Let*

$$R_k := \frac{\|F(x^k)\|^2}{b_{k+1}\bar{b}_{k+1}}, \quad \textit{and} \quad \bar{R}_k := \frac{\|F(x^k)\|^2}{b_{k+1}\bar{b}_{k+1}^{1/2}\left(2\bar{b}_{k+1}^{1/2}-\bar{b}_0^{1/2}\right)}.$$

Then, for any $x^* \in S^*$, $\lambda > 0$ that is measurable to $\mathcal{F}_k$, and $k \geq 1$,

$$-\mathbb{E}_k\left[2\alpha_k\gamma_k\|F(x^k)\|^2\right] \leq \mathbb{E}_k\left[2\eta^2\left(R_{k-1}-R_k\right)+\frac{2\lambda L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right] + \frac{2\eta^2\|F(x^k)\|^2}{\lambda b_k\bar{b}_k} - \frac{2\eta^2\|F(x^k)\|^2}{b_k\bar{b}_k}, \tag{88}$$

and

$$\begin{aligned}
-\mathbb{E}_k\left[2\beta_{k+1}\alpha_k\gamma_k\|F(x^k)\|^2\right] \quad \leq{}& \mathbb{E}_k\left[2\eta^2\left(\bar{R}_{k-1}-\bar{R}_k\right)+\frac{2\lambda L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right] + \frac{2\eta^2\|F(x^k)\|^2}{\lambda b_k\bar{b}_k^{1/2}(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2})} \\
&-\frac{2\eta^2\|F(x^k)\|^2}{b_k\bar{b}_k^{1/2}(2\bar{b}_k^{1/2}-\bar{b}_0^{1/2})}.
\end{aligned} \tag{89}$$

In addition, when $k = 0$,

$$-\mathbb{E}_k\left[2\alpha_0\gamma_0\|F(x^0)\|^2\right] \leq \frac{2\eta^2\|F(x^0)\|^2}{b_0\bar{b}_0} - \frac{2\eta^2\|F(x^0)\|^2}{b_0\bar{b}_0}, \tag{90}$$

and

$$-\mathbb{E}_k\left[2\beta_1\alpha_0\gamma_0\|F(x^0)\|^2\right] \leq \frac{2\eta^2\|F(x^0)\|^2}{b_0\bar{b}_0} - \frac{2\eta^2\|F(x^0)\|^2}{b_0\bar{b}_0^{1/2}(2\bar{b}_0^{1/2}-\bar{b}_0^{1/2})}. \tag{91}$$

Proof. We introduce the de-correlated stepsize $\alpha_{k-1}\gamma_{k-1} = \dfrac{\eta^2}{b_k\bar{b}_k}$. Then we have,

$$\frac{-2\eta^2\|F(x^k)\|^2}{b_{k+1}\bar{b}_{k+1}} = \frac{-2\eta^2\|F(x^k)\|^2}{b_k\bar{b}_k} + 2\eta^2\left(\frac{\|F(x^k)\|^2}{b_k\bar{b}_k} - \frac{\|F(x^k)\|^2}{b_{k+1}\bar{b}_{k+1}}\right).$$

Similar to our earlier results, we would like to create a telescoping sum by utilizing the potential function R_k defined above. Thus,

$$\begin{aligned} &2\eta^2\left(\frac{\|F(x^k)\|^2}{b_k\bar{b}_k} - \frac{\|F(x^k)\|^2}{b_{k+1}\bar{b}_{k+1}}\right) \\ &= 2\eta^2\left(R_{k-1} - R_k + \frac{\|F(x^k)\|^2 - \|F(x^{k-1})\|^2}{b_k\bar{b}_k}\right). \end{aligned} \tag{92}$$

Utilizing (42), we obtain:

$$\begin{aligned} 2\eta^2\frac{\|F(x^k)\|^2 - \|F(x^{k-1})\|^2}{b_k\bar{b}_k} &\leq 4\eta^2\frac{(\|F(x^k)\| - \|F(x^{k-1})\|)\|F(x^k)\|}{b_k\bar{b}_k} \\ &\leq 4\eta^2\frac{\|F(x^k) - F(x^{k-1})\|\|F(x^k)\|}{b_k\bar{b}_k} \\ &\leq 4L\eta^2\frac{\|x^k - x^{k-1}\|\|F(x^k)\|}{b_k\bar{b}_k} \\ &= 4L\eta^3\frac{\|\tilde{F}(\bar{x}^{k-1})\|\|F(x^k)\|}{b_k\bar{b}_k^2}. \end{aligned} \tag{93}$$

We apply the triangle inequality and L-Lipschitz continuity to derive the above inequalities and the last equality is due to the update rule for x^k. Applying Young's inequality, (41), with $\lambda > 0$, we have

$$4L\eta^3\frac{\|\tilde{F}(\bar{x}^{k-1})\|\|F(x^k)\|}{b_k\bar{b}_k^2} \leq \frac{2\lambda L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k} + \frac{2\eta^2\|F(x^k)\|^2}{\lambda b_k\bar{b}_k}. \tag{94}$$

Plugging the results back to (92), we can derive the desirable bound. The proof for the second result follows by a nearly identical argument, so we omit it here. The results for when $k = 0$ follow trivially. □

C.4.1 Descent Lemmas

In this subsection, we present a proof of Lemma 3.2 by considering each case one at a time.

Lemma C.8 (Proof of Lemma 3.2 under Error Bound condition.)**.** *Let Assumptions 1.1 (a), (b), 3.1, 3.2, and 3.3 hold where $\sigma_1 = 0$ in Assumption 3.2 and let $\{x^k\}$ be generated by Algorithm 2. Then, for any $x^* \in S^*$ and $N \geq 1$,*

$$\begin{aligned} \mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\|F(x^k)\|^2}{b_k\bar{b}_k}\right] &\leq C_{1,EB} + C_{2,EB}\ln\mathbb{E}[\bar{b}_N^{1/2}], \\ \mathbb{E}[\text{dist}(x^N, S^*)^2] &\leq M_{1,EB} + M_{2,EB}\ln\mathbb{E}[\bar{b}_N^{1/2}], \end{aligned} \tag{95}$$

where

$$\begin{aligned}
M_{1,EB} &:= \|x^0 - x^*\|^2 + \frac{21\eta^2\sigma_0^2}{b_0^3} + \frac{(21\eta^2\sigma_1^2 + 4\eta^2)\|F(x^0)\|^2}{b_0^3} + M_{2,EB},\\
M_{2,EB} &:= \eta^2 + 2L\eta^3 + \frac{7\sigma_0^2 q^2}{b_0} + \frac{\left(1029\sigma_1^4 + 8\right)\eta^4 L^2}{b_0^3},\\
C_{1,EB} &:= \frac{2M_{1,EB}}{\eta^2},\\
C_{2,EB} &:= \frac{8M_{2,EB}}{\eta^2}.
\end{aligned}$$

In addition, if the solution set S^* is a singleton set, then,

$$\mathbb{E}\left[\sum_{k=0}^{N-1} 2\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \leq M_{1,EB} + M_{2,EB}\ln \mathbb{E}[\bar{b}_N^{1/2}]. \tag{96}$$

Proof. Combining inequalities from (34), (44), (73) and (88), and setting $\lambda = 4, \rho = 7$ we obtain the following for each $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}_k[\|x^{k+1} - x^*\|^2]\\
&\leq \|x^k - x^*\|^2 + \mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2] - \mathbb{E}_k[2\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle] + \mathbb{E}_k[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2]\\
&+\frac{\alpha_{k-1}\gamma_{k-1}}{7q}\|x^k - x^*\|^2 + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right]\\
&+\frac{7\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right] + 7\eta^2\sigma_1^2\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right]\\
&+\frac{14\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 14\eta^2\sigma_1^2\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right]\\
&+\frac{343\eta^4L^2\sigma_1^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{686\eta^4L^2\sigma_1^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&+\mathbb{E}_k\left[2\eta^2\left(R_{k-1} - R_k\right) + \frac{8L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right] - \frac{9}{14}\alpha_{k-1}\gamma_{k-1}\|F(x^k)\|^2.
\end{aligned} \tag{97}$$

Next, we apply the Minty variational inequality, to the third inner product term in the right side of (97), apply the error bound condition, Assumption 3.3, and take the infimum over S^* on both sides of (97) to derive the following,

$$\begin{aligned}
\mathbb{E}_k\left[\operatorname{dist}(x^{k+1}, S^*)^2\right] \leq\ & \operatorname{dist}(x^k, S^*)^2 + \mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2] + \mathbb{E}_k[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2]\\
&-\frac{\alpha_{k-1}\gamma_{k-1}}{2}\|F(x^k)\|^2 + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right]\\
&+\frac{7\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right] + 7\eta^2\sigma_1^2\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right]\\
&+\frac{14\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 14\eta^2\sigma_1^2\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right]\\
&+\frac{343\eta^4L^2\sigma_1^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{686\eta^4L^2\sigma_1^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&+\mathbb{E}_k\left[2\eta^2\left(R_{k-1} - R_k\right) + \frac{8L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right].
\end{aligned} \tag{98}$$

Here we apply the concavity of the infimum to derive the left hand side (98) and we employ the error bound condition to bound the term $\frac{\alpha_{k-1}\gamma_{k-1}}{7q}\text{dist}(x^k, S^*)^2$ by $\frac{\alpha_{k-1}\gamma_{k-1}}{7}\|F(x^k)\|^2$ to obtain the negative term $-\frac{\alpha_{k-1}\gamma_{k-1}}{2}\|F(x^k)\|^2$ in the right side of (98).

On the other hand, if the solution set S^* is a singleton set, then by moving the negative inner product term $-\mathbb{E}_k\left[2\gamma_k\langle\bar{x}^k - x^*, F(\bar{x}^k)\rangle\right]$ to the left side of (97) and by applying the error bound condition, we can obtain the following,

$$\begin{aligned}
&\mathbb{E}_k\left[2\gamma_k\langle\bar{x}^k - x^*, F(\bar{x}^k)\rangle\right]\\
&\leq \|x^k - x^*\|^2 - \mathbb{E}_k[\|x^{k+1} - x^*\|^2] + \mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2] + \mathbb{E}_k[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2]\\
&-\frac{\alpha_{k-1}\gamma_{k-1}}{2}\|F(x^k)\|^2 + \frac{7q\sigma_0^2}{\bar{b}_k^{1/2}}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2 + \|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}^2}\right]\\
&+\frac{7\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k^2} - \frac{1}{\bar{b}_{k+1}^2}\right] + 7\eta^2\sigma_1^2\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right]\\
&+\frac{14\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 14\eta^2\sigma_1^2\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right]\\
&+\frac{343\eta^4L^2\sigma_1^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{686\eta^4L^2\sigma_1^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&+\mathbb{E}_k\left[2\eta^2\left(R_{k-1} - R_k\right) + \frac{8L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right].
\end{aligned} \tag{99}$$

Now, since the sequences $\{b_k\}_{k\geq 0}, \{\bar{b}_k\}_{k\geq 0}$ are non-decreasing and $\gamma_k \leq \frac{\alpha_k^2}{\eta}, \forall k$, we take the total expectation of this inequality and sum for $k = 0, ..., N-1$, together with the results in Lemma A.1, we can find a bound on these terms. Specifically, by defining

$$\begin{aligned}
M_{1,EB} &:= \|x^0 - x^*\|^2 + \frac{21\eta^2\sigma_0^2}{b_0^3} + \frac{(21\eta^2\sigma_1^2 + 4\eta^2)\|F(x^0)\|^2}{b_0^3} + M_{2,EB},\\
M_{2,EB} &:= \eta^2 + 2L\eta^3 + \frac{7\sigma_0^2q^2}{b_0} + \frac{\left(1029\sigma_1^4 + 8\right)\eta^4L^2}{b_0^3},
\end{aligned}$$

for each $N \geq 1$, we derive,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\eta^2\|F(x^k)\|^2}{2b_k\bar{b}_k}\right] &\leq M_{1,EB} + M_{2,EB}\mathbb{E}\ln[\bar{b}_N^2],\\
\mathbb{E}\left[\text{dist}(x^N, S^*)^2\right] &\leq M_{1,EB} + M_{2,EB}\mathbb{E}\ln[\bar{b}_N^2],
\end{aligned} \tag{100}$$

and in the case when S^* is a singleton set, we have the following for each $N \geq 1$:

$$\mathbb{E}\left[\sum_{k=0}^{N-1} 2\gamma_k\langle\bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \leq M_{1,EB} + M_{2,EB}\mathbb{E}\ln[\bar{b}_N^2].$$

Applying Jensen's inequality on the concave function ln, and the construction of $C_{1,EB}, C_{2,EB}$, proves the result. □

Now, we consider the star-strongly monotone case.

Lemma C.9 (Proof of Lemma 3.2 under Star-strong Monotonicity.)**.** *Let Assumptions 1.1 (a), (b), 3.1, 3.2, and*

3.4 hold and let $\{x^k\}$ be generated by Algorithm 2. Then, for any $x^* \in S^*$ and $N \geq 1$,

$$\begin{aligned} \mathbb{E}\left[\sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_k \bar{b}_k}\right] &\leq C_{1,SM} + C_{2,SM} \ln \mathbb{E}[\bar{b}_N^{1/2}], \\ \mathbb{E}\left[\|x^N - x^*\|^2\right] &\leq M_{1,SM} + M_{2,SM} \ln \mathbb{E}[\bar{b}_N^{1/2}], \\ \mathbb{E}\left[\sum_{k=0}^{N-1} \gamma_k \langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] &\leq M_{1,SM} + M_{2,SM} \ln \mathbb{E}[\bar{b}_N^{1/2}], \end{aligned} \tag{101}$$

where

$$\begin{aligned} M_{1,SM} &:= \left(1 + \frac{m\eta}{b_0^2}\right)\|x^0 - x^*\|^2 + \frac{24\eta^2\sigma_0^2}{b_0^3} + \frac{4\eta\sigma_0^2}{mb_0^2} + \frac{(48\eta^2\sigma_1^2 + 4\eta^2)\|F(x^0)\|^2}{b_0^3} \\ &\quad + \frac{16\eta\sigma_1^2\|F(x^0)\|^2}{mb_0^2} + M_{2,SM}, \\ M_{2,SM} &:= \eta^2 + 2L\eta^3 + \frac{\left(1536\sigma_1^4 + 8\right)\eta^4 L^2}{b_0^3} + \frac{m\eta}{b_0^2} + \frac{256\eta^2 L^2\sigma_1^4}{m^2 b_0} + m\eta^3 + \frac{8L^2\eta^3\sigma_1^2}{m}, \\ C_{1,SM} &:= \frac{2M_{1,SM}}{\eta^2}, \\ C_{2,SM} &:= \frac{8M_{2,SM}}{\eta^2}. \end{aligned}$$

Proof. Combining inequalities from (34), (49), (73) and (88), we obtain the following for each $k \geq 1$,

$$\begin{aligned} &\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] \\ &\leq \|x^k - x^*\|^2 + \mathbb{E}_k[\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2] - \mathbb{E}_k[2\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle] + \mathbb{E}_k[2L\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2] \\ &+ \frac{m\eta\|x^k - x^*\|^2}{4\bar{b}_k} + \frac{4\eta\sigma_0^2}{m}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + \frac{8L^2\eta^3\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \\ &+ \frac{8\eta\sigma_1^2}{m}\mathbb{E}_k\left[\frac{\|F(x^{k-1})\|^2}{\bar{b}_k} - \frac{\|F(x^k)\|^2}{\bar{b}_{k+1}}\right] + \frac{\eta^2\|F(x^k)\|^2}{4b_k\bar{b}_k} + \frac{256\eta^2L^2\sigma_1^4}{m^2b_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] \\ &+ \frac{6\alpha_{k-1}\gamma_{k-1}}{\rho}\|F(x^k)\|^2 + \frac{\eta^2\sigma_0^2\rho}{b_0}\mathbb{E}_k\left[\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right] + \eta^2\sigma_1^2\rho\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right] \\ &+ \frac{2\eta^2\sigma_0^2\rho}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 2\eta^2\sigma_1^2\rho\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right] \\ &+ \frac{\eta^4L^2\sigma_1^4\rho^3}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{2\eta^4L^2\sigma_1^4\rho^3}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] \\ &+ \mathbb{E}_k\left[2\eta^2\left(R_{k-1} - R_k\right) + \frac{2\lambda L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right] + \frac{2\eta^2\|F(x^k)\|^2}{\lambda b_k\bar{b}_k} - \frac{2\eta^2\|F(x^k)\|^2}{b_k\bar{b}_k}. \end{aligned} \tag{102}$$

Now, since F satisfies star-strongly monotone condition, Assumption 3.4, we can derive that

$$-\mathbb{E}_k\left[\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \leq -\mathbb{E}_k[m\gamma_k\|\bar{x}^k - x^*\|^2] \leq -\mathbb{E}_k\left[\frac{m\gamma_k}{2}\|x^k - x^*\|^2\right] + \mathbb{E}_k[m\gamma_k\alpha_k^2\|\tilde{F}(x^k)\|^2].$$

Then, for any $k \geq 1$, we can handle the negative term as follows,

$$\begin{aligned}
-\mathbb{E}_k\Big[\frac{m\gamma_k}{2}\|x^k - x^*\|^2\Big] &= \mathbb{E}_k\Big[\frac{m(\gamma_{k-1}-\gamma_k)\|x^k - x^*\|^2}{2} - \frac{m\gamma_{k-1}\|x^k - x^*\|^2}{2}\Big] \\
&= \mathbb{E}_k\Big[\frac{m\gamma_{k-1}\|x^{k-1} - x^*\|}{2} - \frac{m\gamma_k\|x^k - x^*\|}{2} - \frac{m\gamma_{k-1}\|x^k - x^*\|^2}{2} \\
&\quad + \frac{m\gamma_{k-1}}{2}(\|x^k - x^*\|^2 - \|x^{k-1} - x^*\|^2)\Big] \\
&\leq \mathbb{E}_k\Big[\frac{m\gamma_{k-1}\|x^{k-1} - x^*\|}{2} - \frac{m\gamma_k\|x^k - x^*\|}{2} - \frac{m\gamma_{k-1}\|x^k - x^*\|^2}{2} \\
&\quad + m\gamma_{k-1}\|x^k - x^*\|\|x^{k-1} - x^k\|\Big] \\
&= \mathbb{E}_k\Big[\frac{m\gamma_{k-1}\|x^{k-1} - x^*\|}{2} - \frac{m\gamma_k\|x^k - x^*\|}{2} - \frac{m\gamma_{k-1}\|x^k - x^*\|^2}{2} \\
&\quad + m\gamma_{k-1}^2\|x^k - x^*\|\|\tilde{F}(\bar{x}^{k-1})\|\Big],
\end{aligned}$$

while when $k = 0$, we have $-\mathbb{E}_0\Big[\frac{m\gamma_0}{2}\|x^k - x^*\|^2\Big] \leq -\frac{m\gamma_{-1}}{2}\|x^0 - x^*\|^2 + -\mathbb{E}_0\Big[\frac{m\gamma_0}{2}\|x^0 - x^*\|^2\Big]$. Now we apply Young's inequality to bound the last product to get

$$m\gamma_{k-1}^2\|x^k - x^*\|\|\tilde{F}(\bar{x}^{k-1})\| \leq \frac{m\gamma_{k-1}\|x^k - x^*\|^2}{4} + m\gamma_{k-1}^3\|\tilde{F}(\bar{x}^{k-1})\|^2.$$

Combining all results above, we obtain

$$\begin{aligned}
-\mathbb{E}_k\left[\gamma_k\langle\bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \quad \leq \quad & \frac{m\gamma_{k-1}\|x^{k-1} - x^*\|}{2} - \mathbb{E}_k\left[\frac{m\gamma_k\|x^k - x^*\|}{2}\right] - \frac{m\gamma_{k-1}\|x^k - x^*\|^2}{4} \\
& + m\gamma_{k-1}^3\|\tilde{F}(\bar{x}^{k-1})\|^2 + \mathbb{E}_k\left[m\gamma_k\alpha_k^2\|\tilde{F}(x^k)\|^2\right].
\end{aligned} \tag{103}$$

By selecting $\rho = 8$ and $\lambda = 4$ and using the fact that both sequences $\{b_k\}_{k\geq 0}, \{\bar{b}_k\}_{k\geq 0}$ are non-decreasing as well as the fact that $\gamma_k \leq \frac{\alpha_k^2}{\eta}$, we combine (102) with (103), take the total expectation, and sum for $k = 0, ..., N-1$

to obtain,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\eta^2\|F(x^k)\|^2}{2b_k\bar{b}_k}\right]\\
&\leq\left(1+\frac{m\eta}{b_0^2}\right)\|x^0-x^*\|^2-\mathbb{E}\left[\|x^N-x^*\|^2\right]-\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle\right]\\
&+\frac{24\eta^2\sigma_0^2}{b_0^3}+\frac{4\eta\sigma_0^2}{mb_0^2}+\frac{(48\eta^2\sigma_1^2+4\eta^2)\|F(x^0)\|^2}{b_0^3}+\frac{16\eta\sigma_1^2\|F(x^0)\|^2}{mb_0^2}\\
&+\left(\eta^2+\frac{\left(1536\sigma_1^4+8\right)\eta^4L^2}{b_0^3}+\frac{m\eta^3}{b_0^2}+\frac{256\eta^2L^2\sigma_1^4}{m^2b_0}\right)\mathbb{E}\left[\sum_{k=1}^{N-1}\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&+\left(2L\eta^3+m\eta^3+\frac{8L^2\eta^3\sigma_1^2}{m}\right)\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \qquad (104)\\
&\leq\left(1+\frac{m\eta}{b_0^2}\right)\|x^0-x^*\|^2-\mathbb{E}\left[\|x^N-x^*\|^2\right]-\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle\right]\\
&+\frac{24\eta^2\sigma_0^2}{b_0^3}+\frac{4\eta\sigma_0^2}{mb_0^2}+\frac{(48\eta^2\sigma_1^2+4\eta^2)\|F(x^0)\|^2}{b_0^3}+\frac{16\eta\sigma_1^2\|F(x^0)\|^2}{mb_0^2}\\
&+\left(\eta^2+\frac{\left(1536\sigma_1^4+8\right)\eta^4L^2}{b_0^3}+\frac{m\eta^3}{b_0^2}+\frac{256\eta^2L^2\sigma_1^4}{m^2b_0}\right)\left(1+\mathbb{E}\ln[\bar{b}_N^2]\right)\\
&+\left(2L\eta^3+m\eta^3+\frac{8L^2\eta^3\sigma_1^2}{m}\right)\left(1+\mathbb{E}\ln[\bar{b}_N^2]\right),
\end{aligned}$$

where the second inequality is from the result of Lemma A.1. By defining

$$\begin{aligned}
M_{1,SM} &:= \left(1+\frac{m\eta}{b_0^2}\right)\|x^0-x^*\|^2+\frac{24\eta^2\sigma_0^2}{b_0^3}+\frac{4\eta\sigma_0^2}{mb_0^2}+\frac{(48\eta^2\sigma_1^2+4\eta^2)\|F(x^0)\|^2}{b_0^3}\\
&\quad+\frac{16\eta\sigma_1^2\|F(x^0)\|^2}{mb_0^2}+M_{2,SM},\\
M_{2,SM} &:= \eta^2+2L\eta^3+\frac{\left(1536\sigma_1^4+8\right)\eta^4L^2}{b_0^3}+\frac{m\eta}{b_0^2}+\frac{256\eta^2L^2\sigma_1^4}{m^2b_0}+m\eta^3+\frac{8L^2\eta^3\sigma_1^2}{m}.
\end{aligned}$$

Then, since F satisfies Minty variational condition, finally for each $N\geq 1$, we derive,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\eta^2\|F(x^k)\|^2}{2b_k\bar{b}_k}\right] &\leq M_{1,SM}+M_{2,SM}\mathbb{E}\ln[\bar{b}_N^2],\\
\mathbb{E}\left[\|x^N-x^*\|^2\right] &\leq M_{1,SM}+M_{2,SM}\mathbb{E}\ln[\bar{b}_N^2],\\
\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k\langle\bar{x}^k-x^*,F(\bar{x}^k)\rangle\right] &\leq M_{1,SM}+M_{2,SM}\mathbb{E}\ln[\bar{b}_N^2].
\end{aligned} \qquad (105)$$

Applying Jensen's inequality on the concave function ln, and the construction of $C_{1,SM},C_{2,SM}$, we prove the results as desired. □

Finally, we consider the case in which the sub-Weibull noise condition holds.

Lemma C.10 (Proof of Lemma 3.2 under Sub-Weibull Noise.)**.** *Let Assumptions 1.1 (a), (b), 3.1, 3.2, and A.1 hold and let $\{x^k\}$ be generated by Algorithm 2. Let*

$$\ell(N,\delta):=\frac{\log^{2\theta}\left(\frac{Ne}{\delta}\right)}{\Gamma(2\theta+1)e}.$$

Then, for any $x^* \in S^*$ and $N \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{N-1} \frac{\|F(x^k)\|^2}{b_k \bar{b}_k}\right] &\leq C_{1,SW}(N,\delta) + C_{2,SW}(N,\delta)\ln \mathbb{E}[\bar{b}_N^{1/2}], \\
\mathbb{E}\left[\beta_N \|x^N - x^*\|^2\right] &\leq M_{1,SW}(N,\delta) + M_{2,SW}(N,\delta)\ln \mathbb{E}[\bar{b}_N^{1/2}], \\
\mathbb{E}\left[\sum_{k=0}^{N-1} \gamma_k \langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] &\leq M_{1,SW}(N,\delta) + M_{2,SW}(N,\delta)\ln \mathbb{E}[\bar{b}_N^{1/2}],
\end{aligned} \tag{106}$$

where

$$\begin{aligned}
M_{1,SW}(N,\delta) &:= \beta_0\|x^0 - x^*\|^2 + \frac{4\eta^2\ell(N,\delta)}{b_0^4}\left(\sigma_0^2 + 3\sigma_1^2\|F(x^0)\|^2\right) + \left(128\sigma_1^2\eta^2 + 4\eta^2\right)\frac{\|F(x^0)\|^2}{b_0^3} \\
&\quad + \frac{208\eta^2\sigma_0^2}{b_0^3} + \frac{208\eta^2\sigma_0^2}{b_0^3} + \frac{8\eta^2\ell(N,\delta)\sigma_1^2L^2\eta^2}{b_0^2} + M_{2,SW}(N,\delta), \\
M_{2,SW}(\delta) &:= 2L\eta^3 + \frac{\eta^3L^2}{2b_0^{3/2}} + \frac{192\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^{5/2}} + \eta^2 + \frac{17*8^4L^2\eta^4\sigma_1^4}{b_0^3} + \frac{16L^2\eta^4}{b_0^3} \\
&\quad + \frac{384\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^3}, \\
C_{1,SW}(\delta) &:= \frac{4M_{1,SW}(\delta)}{\eta^2} \\
C_{2,SW}(\delta) &:= \frac{16M_{2,SW}(\delta)}{\eta^2}.
\end{aligned}$$

Proof. Combining inequalities from (36), (58), (74) and (89), we obtain the following for each $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}_k\left[\beta_{k+1}\|x^{k+1} - x^*\|^2\right] \\
&\leq \mathbb{E}_k\left[\beta_{k+1}\|x^k - x^*\|^2\right] + \mathbb{E}_k[\beta_{k+1}\gamma_k^2\|\tilde{F}(\bar{x}^k)\|^2] - \mathbb{E}_k[2\beta_{k+1}\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle] \\
&\quad + \mathbb{E}_k\left[(\beta_k - \beta_{k+1})\|x^k - x^*\|^2\right] + \frac{4\eta^2\ell(N,\delta)\sigma_0^2}{\bar{b}_0}\mathbb{E}_k\left[T_{k-1} - T_k\right] + \mathbb{E}_k[2L\beta_{k+1}\alpha_k^2\gamma_k\|\tilde{F}(x^k)\|^2] \\
&\quad + \frac{4\eta^2\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\bar{\mathcal{P}}_{k-1} - \bar{\mathcal{P}}_k\right] + \frac{4\rho\eta^3L\ell(N,\delta)\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{\bar{b}_k^2}\right] \\
&\quad + \frac{4\rho\eta^3L\ell(N,\delta)\sigma_1^2}{\bar{b}_0^2}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{24\eta^5L^3\ell(N,\delta)\sigma_1^2}{\rho\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \\
&\quad + \frac{24\eta^3L\ell(N,\delta)\sigma_1^2}{\rho\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{\|F(x^k)\|^2}{\bar{b}_k^{1/2}b_k(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})}\right] + \frac{4\rho\eta^3L\ell(N,\delta)\sigma_1^2}{\bar{b}_0}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \\
&\quad + \frac{6\eta^2\|F(x^k)\|^2}{\lambda\bar{b}_k^{1/2}(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})b_k} + \frac{2\lambda\eta^2\sigma_0^2}{\bar{b}_0^{1/2}}\mathbb{E}_k\left[\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right] + 2\lambda\sigma_1^2\eta^2\mathbb{E}_k\left[\frac{\mathcal{P}_{k-1}}{\bar{b}_k^{1/2}} - \frac{\mathcal{P}_k}{\bar{b}_{k+1}^{1/2}}\right] \\
&\quad + \frac{8L^2\lambda^3\sigma_1^4\eta^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \frac{8\lambda\eta^2\sigma_0^2}{b_0}\mathbb{E}_k\left[\frac{1}{\bar{b}_k} - \frac{1}{\bar{b}_{k+1}}\right] + 8\lambda\sigma_1^2\eta^2\mathbb{E}_k\left[\frac{\tilde{\mathcal{P}}_{k-1}}{b_k} - \frac{\tilde{\mathcal{P}}_k}{b_{k+1}}\right] \\
&\quad + \frac{128L^2\lambda^3\sigma_1^4\eta^4}{b_0^3}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + \mathbb{E}_k\left[2\eta^2\left(\bar{R}_{k-1} - \bar{R}_k\right) + \frac{2\lambda L^2\eta^4\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^3 b_k}\right] \\
&\quad + \frac{2\eta^2\|F(x^k)\|^2}{\lambda b_k\bar{b}_k^{1/2}(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})} - \frac{2\eta^2\|F(x^k)\|^2}{b_k\bar{b}_k^{1/2}(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})}.
\end{aligned} \tag{107}$$

By selecting $\lambda = 8$ and $\rho = \left(\frac{48\eta L\ell(N,\delta)\sigma_1^2}{\bar{b}_0^{1/2}}\right)$, and using the fact that both sequences $\{b_k\}_{k\geq 0}, \{\bar{b}_k\}_{k\geq 0}$ are

non-decreasing and that $\gamma_k \leq \dfrac{\alpha_k^2}{\eta}, 1/2 \leq \beta_{k+1} \leq 1, \forall k$, we take the total expectation of the above inequality and the sum from $k = 0, ..., N-1$ to obtain,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{N-1} \frac{\eta^2\|F(x^k)\|^2}{2b_k\bar{b}_k^{1/2}(2\bar{b}_k^{1/2} - \bar{b}_0^{1/2})}\right] \\
&\leq \beta_0\|x^0 - x^*\|^2 - \mathbb{E}\left[\beta_N\|x^{k+1} - x^*\|^2\right] - \mathbb{E}\left[\sum_{k=0}^{N-1} 2\beta_{k+1}\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \\
&+ \frac{4\eta^2\ell(N,\delta)}{b_0^4}\left(\sigma_0^2 + 3\sigma_1^2\|F(x^0)\|^2\right) + \left(128\sigma_1^2\eta^2 + 4\eta^2\right)\frac{\|F(x^0)\|^2}{b_0^3} + \frac{208\eta^2\sigma_0^2}{b_0^3} + \frac{8\eta^2\ell(N,\delta)\sigma_1^2L^2\eta^2}{b_0^2} \\
&+ \left(2L\eta^3 + \frac{\eta^3L^2}{2b_0^{3/2}} + \frac{192\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^{5/2}}\right)\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^4}\right] \\
&+ \left(\eta^2 + \frac{17*8^4L^2\eta^4\sigma_1^4}{b_0^3} + \frac{16L^2\eta^4}{b_0^3}\right)\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] \\
&+ \frac{384\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^3}\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2 + \|\tilde{F}(x^{k-1})\|^2}{\bar{b}_k^2}\right] \\
&\leq \beta_0\|x^0 - x^*\|^2 - \mathbb{E}\left[\beta_N\|x^{k+1} - x^*\|^2\right] - \mathbb{E}\left[\sum_{k=0}^{N-1} 2\beta_{k+1}\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] \\
&+ \frac{4\eta^2\ell(N,\delta)}{b_0^4}\left(\sigma_0^2 + 3\sigma_1^2\|F(x^0)\|^2\right) + \left(128\sigma_1^2\eta^2 + 4\eta^2\right)\frac{\|F(x^0)\|^2}{b_0^3} + \frac{208\eta^2\sigma_0^2}{b_0^3} + \frac{8\eta^2\ell(N,\delta)\sigma_1^2L^2\eta^2}{b_0^2} \\
&+ \left(2L\eta^3 + \frac{\eta^3L^2}{2b_0^{3/2}} + \frac{192\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^{5/2}}\right)(1 + \mathbb{E}\ln(\bar{b}_N^2)) \\
&+ \left(\eta^2 + \frac{17*8^4L^2\eta^4\sigma_1^4}{b_0^3} + \frac{16L^2\eta^4}{b_0^3}\right)(1 + \mathbb{E}\ln(\bar{b}_N^2)) \\
&+ \frac{384\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^3}(1 + \mathbb{E}\ln(\bar{b}_N^2)),
\end{aligned}
\tag{108}
$$

where the second inequality is from results in Lemma A.1. By defining

$$
\begin{aligned}
M_{1,SW}(N,\delta) &:= \beta_0\|x^0 - x^*\|^2 + \frac{4\eta^2\ell(N,\delta)}{b_0^4}\left(\sigma_0^2 + 3\sigma_1^2\|F(x^0)\|^2\right) + \left(128\sigma_1^2\eta^2 + 4\eta^2\right)\frac{\|F(x^0)\|^2}{b_0^3} \\
&\quad + \frac{208\eta^2\sigma_0^2}{b_0^3} + \frac{208\eta^2\sigma_0^2}{b_0^3} + \frac{8\eta^2\ell(N,\delta)\sigma_1^2L^2\eta^2}{b_0^2} + M_{2,SW}(N,\delta), \\
M_{2,SW}(N,\delta) &:= 2L\eta^3 + \frac{\eta^3L^2}{2b_0^{3/2}} + \frac{192\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^{5/2}} + \eta^2 + \frac{17*8^4L^2\eta^4\sigma_1^4}{b_0^3} + \frac{16L^2\eta^4}{b_0^3} \\
&\quad + \frac{384\eta^4L^2\ell(N,\delta)^2\sigma_1^4}{b_0^3}.
\end{aligned}
$$

Then, since F satisfies Minty variational condition, finally for each $N \geq 1$, we derive,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\eta^2\|F(x^k)\|^2}{4b_k\bar{b}_k}\right] &\leq M_{1,SW}(\delta) + M_{2,SW}(\delta)\mathbb{E}\ln[\bar{b}_N^2], \\
\mathbb{E}\left[\beta_N\|x^N - x^*\|^2\right] &\leq M_{1,SW}(\delta) + M_{2,SW}(\delta)\mathbb{E}\ln[\bar{b}_N^2], \\
\mathbb{E}\left[\sum_{k=0}^{N-1} 2\beta_{k+1}\gamma_k\langle \bar{x}^k - x^*, F(\bar{x}^k)\rangle\right] &\leq M_{1,SW}(\delta) + M_{2,SW}(\delta)\mathbb{E}\ln[\bar{b}_N^2].
\end{aligned}
\tag{109}
$$

Applying Jensen's inequality on the concave function ln, and the construction of $C_{1,SW}(N,\delta), C_{2,SW}(N,\delta)$, we prove the results as desired. □

C.5 Bounding the Stepsize

In this subsection, we provide a bound on the stepsizes in Algorithm 2, which will be instrumental in deriving our final convergence results.

Lemma C.11 (Bounding step size.)**.** *Let Assumptions 1.1 (a), (b), 3.1, and 3.2 hold and let $\{x^k\}$ be generated by Algorithm 2. Then for all $N \geq 1$, we consider three different cases:*

- *Error bound condition: Let Assumption 3.3 hold as well and assume that $\sigma_1 = 0$ in Assumption 3.2, then we can bound the step size, $b_N^{1/2}$, as:*

$$\begin{aligned}
\mathbb{E}\left[\bar{b}_N^{1/2}\right] &\leq 2D_{1,EB} + 8D_{2,EB}\ln(e + D_{2,EB}),\\
D_{1,EB} &= \left(\bar{b}_0^2 + N\left(4 + \frac{2\eta^2L^2}{b_0^2}\right)\bar{\sigma}^2\right)^{1/4} + \left(4 + \frac{2\eta^2L^2}{b_0^2}\right)C_{1,EB},\\
D_{2,EB} &= \left(4 + \frac{2\eta^2L^2}{b_0^2}\right)C_{2,EB},\\
\bar{\sigma}^2 &= \left(2 + \frac{2\eta^2L^2}{b_0^2}\right)\sigma_0^2.
\end{aligned} \tag{110}$$

- *Star-strongly monotone: Let Assumption 3.4 hold as well, then we can bound the step size, $b_N^{1/2}$, as:*

$$\begin{aligned}
\mathbb{E}\left[\bar{b}_N^{1/2}\right] &\leq 2D_{1,SM} + 8D_{2,SM}\ln(e + D_{2,SM}),\\
D_{1,SM} &= \left(\bar{b}_0^2 + N\left(4 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\bar{\sigma}^2\right)^{1/4}\\
&\quad + \left(4 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)C_{1,SM},\\
D_{2,SM} &= \left(4 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)C_{2,SM},\\
\bar{\sigma}^2 &= \left[2 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right]\frac{\sigma_0^2}{1+\sigma_1^2}.
\end{aligned} \tag{111}$$

- *Sub-Weibull noise: Let Assumption 3.5 hold as well, then we can bound the step size, $b_N^{1/2}$, as:*

$$\begin{aligned}
\mathbb{E}\left[\bar{b}_N^{1/2}\right] &\leq 2D_{1,SW}(\delta) + 8D_{2,SW}(\delta)\ln(e + D_{2,SW}(\delta)),\\
D_{1,SW}(\delta) &= \left(\bar{b}_0^2 + N\left(4 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\bar{\sigma}^2\right)^{1/4}\\
&\quad + \left(4 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)C_{1,SW}(\delta),\\
D_{2,SW}(\delta) &= \left(4 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)C_{2,SW}(\delta),\\
\bar{\sigma}^2 &= \left[2 + \frac{2\eta^2L^2(1+\sigma_1^2)}{b_0^2}\right]\frac{\sigma_0^2}{1+\sigma_1^2}.
\end{aligned} \tag{112}$$

Proof. We assume that

$$\sum_{k=0}^{N-1}\mathbb{E}\left[\frac{\|F(x^k)\|^2}{b_k\bar{b}_k}\right] \leq C_1 + C_2\ln\mathbb{E}\sqrt{\bar{b}_N}, \tag{113}$$

for some arbitrary constants C_1, C_2. Then the fact that $b_k \leq \bar{b}_k^{1/2}$, we obtain

$$\sum_{k=0}^{N-1} \mathbb{E}\left[\frac{\|F(x^k)\|^2}{\bar{b}_k^{3/2}}\right] \leq C_1 + C_2 \ln \mathbb{E}\sqrt{\bar{b}_N}. \tag{114}$$

In contrast to the previous work, where they assume bounded gradients, i.e., $\|F(x)\| \leq M, \forall x$, we develop a divide-and-conquer method to bound the step size $\{\bar{b}_N\}_{N\geq 1}$, inspired by the work in Wang et al. (2023).

In particular, we divide the iterations based on the set $\{\|F(x^k)\|^2 \geq \bar{\sigma}^2\}$, where

$$\bar{\sigma}^2 = \left(2 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)\frac{\sigma_0^2}{1+\sigma_1^2}.$$

Under the Lipschitz continuity and noise assumptions, we can upper bound the expectation of stochastic oracles in the form of $\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2$.

Given any $k \geq 0$, then

$$\begin{aligned}
&\mathbb{E}_k\left[\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right] \\
&\leq 2\sigma_0^2 + (1+\sigma_1^2)\|F(x^k)\|^2 + (1+\sigma_1^2)\mathbb{E}_k[\|F(\bar{x}^k)\|^2] \\
&\leq 2\sigma_0^2 + (1+\sigma_1^2)\|F(x^k)\|^2 + (1+\sigma_1^2)\mathbb{E}_k[\|F(\bar{x}^k) - F(x^k) + F(x^k)\|^2] \\
&\leq 2\sigma_0^2 + 3(1+\sigma_1^2)\|F(x^k)\|^2 + 2(1+\sigma_1^2)\mathbb{E}_k[\|F(\bar{x}^k) - F(x^k)\|^2] \\
&\leq 2\sigma_0^2 + 3(1+\sigma_1^2)\|F(x^k)\|^2 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\mathbb{E}_k[\|\tilde{F}(x^k)\|^2] \\
&\leq 2\sigma_0^2 + 3(1+\sigma_1^2)\|F(x^k)\|^2 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}(\sigma_0^2 + (1+\sigma_1^2)\|F(x^k)\|^2) \\
&\leq \left(2 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)\sigma_0^2 + \left(3 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\|F(x^k)\|^2.
\end{aligned} \tag{115}$$

We first consider the iterations where $\|F(x^k)\|^2 \geq \bar{\sigma}^2$, for which we can derive

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)\mathcal{I}\{\|F(x^k)\|^2 \geq \bar{\sigma}^2\}}{\bar{b}_N^{3/2}}\right] \\
&\leq \mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\mathbb{E}_k\left[\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right]\mathcal{I}\{\|F(x^k)\|^2 \geq \bar{\sigma}^2\}}{\bar{b}_k^{3/2}}\right] \\
&\leq \left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\mathbb{E}\left[\sum_{k=0}^{N-1}\frac{\|F(x^k)\|^2}{\bar{b}_k^{3/2}}\right] \\
&\leq \left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\left(C_1 + C_2 \ln \mathbb{E}\left[\sqrt{\bar{b}_N}\right]\right),
\end{aligned} \tag{116}$$

where the first inequality is from the non-decreasing property of sequence $\{\bar{b}_k\}_{k\geq 0}$ and the tower rule of expectation, the second is due to the results from (115), and the last one is from (113).

Next, we consider those iterations satisfying $\|F(x^k)\|^2 < \bar{\sigma}^2$, then

$$\begin{aligned}
&\mathbb{E}\left[\frac{\bar{b}_0^2 + \sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)\mathcal{I}\{\|F(x^k)\|^2 < \bar{\sigma}^2\}}{\bar{b}_N^{3/2}}\right] \\
&= \mathbb{E}\left[\frac{\bar{b}_0^2 + \sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)}{\bar{b}_N^{3/2}}\mathcal{I}\{\|F(x^k)\|^2 < \bar{\sigma}^2\}\right] \\
&= \mathbb{E}\left[\left(\bar{b}_0^2 + \sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)\right)^{1/4}\mathcal{I}\{\|F(x^k)\|^2 < \bar{\sigma}^2\}\right] \\
&\leq \mathbb{E}\left[\left(\bar{b}_0^2 + \sum_{k=0}^{N-1}\mathbb{E}_k\left[\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right]\right)^{1/4}\mathcal{I}\{\|F(x^k)\|^2 < \bar{\sigma}^2\}\right] \\
&\leq \left(\bar{b}_0^2 + N\left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\bar{\sigma}^2\right)^{1/4}.
\end{aligned} \tag{117}$$

Applying results from (116) and (117), we can bound the term $\mathbb{E}\left[\bar{b}_N^{1/2}\right]$ as following:

$$\begin{aligned}
\mathbb{E}\left[\bar{b}_N^{1/2}\right] &= \mathbb{E}\left[\frac{\bar{b}_N^2}{\bar{b}_N^{3/2}}\right] \\
&= \mathbb{E}\left[\frac{\bar{b}_0^2 + \sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)}{\bar{b}_N^{3/2}}\right] \\
&= \mathbb{E}\left[\frac{\bar{b}_0^2 + \sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)\mathcal{I}\{\|F(x^k)\|^2 < \bar{\sigma}^2\}}{\bar{b}_N^{3/2}}\right] \\
&\quad + \mathbb{E}\left[\frac{\sum_{k=0}^{N-1}\left(\|\tilde{F}(\bar{x}^k)\|^2 + \|\tilde{F}(x^k)\|^2\right)\mathcal{I}\{\|F(x^k)\|^2 \geq \bar{\sigma}^2\}}{\bar{b}_N^{3/2}}\right] \\
&\leq \left(\bar{b}_0^2 + N\left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\bar{\sigma}^2\right)^{1/4} \\
&\quad + \left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\left(C_1 + C_2 \ln \mathbb{E}\sqrt{\bar{b}_N}\right), \\
&:= D_1 + D_2 \ln \mathbb{E}\left[\bar{b}_N^{1/2}\right],
\end{aligned} \tag{118}$$

with the definitions of

$$\begin{aligned}
D_1 = &\left(\bar{b}_0^2 + N\left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)\bar{\sigma}^2\right)^{1/4} + \left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)C_1, \\
D_2 = &\left(4 + \frac{2\eta^2 L^2(1+\sigma_1^2)}{b_0^2}\right)(1+\sigma_1^2)C_2.
\end{aligned} \tag{119}$$

Using Lemma A.2, we can show that

$$\mathbb{E}\left[\bar{b}_N^{1/2}\right] \leq 2D_1 + 8D_2 \ln(e + D_2).$$

By applying the results of Lemmas C.8, C.9, and C.10, we can derive inequality (114) for our three different cases. □

Remark C.1. *In Lemma C.11, we explicitly bound the term* $\mathbb{E}\left[\bar{b}_N^{1/2}\right]$, *resulting in* $\mathbb{E}\left[\bar{b}_N^{1/2}\right] = \mathcal{O}(N^{1/4})$ *with* $\sigma_0 > 0, \sigma_1 \geq 0$. *In addition, if we choose* $\sigma_0 = 0, \sigma_1 > 0$, *then* $\bar{\sigma}^2 = 0$, *leading to a constant upper bound of* $\mathbb{E}\left[\bar{b}_N^{1/2}\right]$.

Following Lemma C.11, the next corollary provides an upper bound of $\bar{b}_N$ with high probability, which holds for our three different cases: error bound condition, star-strongly monotone and sub-Weibull noise assumption.

Corollary C.11.1. *With probability at least* $1-\delta$, *in all three cases, the sequence of* $\{\bar{b}_k\}$ *in Algorithm 2 satisfies for all* $N \geq 1$,

$$\bar{b}_N \leq \frac{c_N}{\delta^2}, \tag{120}$$

where

$$c_N := \mathcal{O}(N^{1/2}), \quad \textit{if} \quad \sigma_0, \sigma_1 > 0.$$

In addition, when $\sigma_0 = 0$, c_N *is simply a constant.*

Proof. For the ease of notation, let us define

$$a := \frac{2D_1 + 8D_2 \ln(e + D_2)}{\delta}, \quad \delta > 0.$$

Then, from the Markov Inequality, we have

$$\mathbb{P}\left(\bar{b}_N \geq a^2\right) = \mathbb{P}\left(\bar{b}_N^{1/2} \geq a\right) \leq \frac{\mathbb{E}\left[\bar{b}_N^{1/2}\right]}{a} \leq \delta. \tag{121}$$

Therefore, with probability $1-\delta$, that

$$\bar{b}_N \leq a^2 = \frac{(2D_1 + 8D_2 \ln(e + D_2))^2}{\delta^2}.$$

Defining

$$c_N := (2D_1 + 8D_2 \ln(e + D_2))^2, \tag{122}$$

we can obtain the desired results (120). By substituting different values of $D_i's$, we obtain the results for three different cases discussed in Lemma C.11. □

We note here that in the sub-Weibull noise case c_N is technically $\tilde{O}(N^{1/2})$ when $\sigma_0 > 0$ and $\tilde{O}(1)$ when $\sigma_0 = 0$, due to the $\ell(N, \delta)$ factors present in the result of Lemma C.10.

C.6 Convergence Results

We begin this subsection with our convergence result in terms the averaged operator norm squared.

Theorem C.12 (Proof of Theorem 3.3)**.** *Let Assumptions 1.1 (a), (b) 3.1, and 3.2 hold and let* $\{x^k\}$ *be generated by Algorithm 2. Then with probability at least* $1-\delta$, *we have the following:*

- *Error bound condition: Let Assumption 3.3 hold as well and assume that* $\sigma_1 = 0$ *in 3.2. Then,*

$$\frac{1}{N}\sum_{k=0}^{N-1} \|F(x^k)\|^2 \leq \frac{\kappa_{N,EB}^3 (C_{1,EB} + C_{2,EB}) \ln(\kappa_{N,EB})}{N\delta^4}, \tag{123}$$

  where

$$\kappa_{N,EB} \quad := \quad 2D_{1,EB} + 8D_{2,EB} \ln(e + D_{2,EB}) = \mathcal{O}(N^{1/4}).$$

- *Star-strongly monotone: Let Assumption 3.4 hold as well. Then,*

$$\frac{1}{N}\sum_{k=0}^{N-1} \|F(x^k)\|^2 \leq \frac{\kappa_{N,SM}^3 (C_{1,SM} + C_{2,SM}) \ln(\kappa_{N,SM})}{N\delta^4}, \tag{124}$$

  where

$$\kappa_{N,SM} \quad := \quad 2D_{1,SM} + 8D_{2,SM} \ln(e + D_{2,SM}).$$

- *Sub-Weibull noise: Let Assumption A.1 hold as well. Then,*

$$\frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2 \leq \frac{16\kappa_{N,SW}(\delta/2)^3(C_{1,SW}+C_{2,SW})\ln(\kappa_{N,SW}(\delta/2)}{N\delta^4}, \tag{125}$$

where

$$q_{N,SW}(\delta) \quad := \quad 2D_{1,SW}(\delta)+8D_{2,SW}(\delta)\ln(e+D_{2,SW}(\delta)).$$

Proof. Let us prove the result for error bound condition and the other two settings follow in a nearly same argument. By applying Hölder's Inequality

$$\left[\mathbb{E}\left|X\right|^4\right]^{1/4} \geq \frac{\mathbb{E}\left|XY\right|}{\left[\mathbb{E}\left|Y\right|^{4/3}\right]^{3/4}},$$

with the choice of

$$X = \left(\frac{\sum_{k=0}^{N-1}\|F(x^k)\|^2}{\bar{b}_N^{3/2}}\right)^{1/4}, \quad Y = (\bar{b}_N^{3/2})^{1/4},$$

from the results in Lemma C.8, we have

$$\begin{aligned}
\mathbb{E}\left[N^{-1}\sum_{k=0}^{N-1}\|F(x^k)\|^2\right]^{1/4} &= N^{-1/4}\mathbb{E}\left[\sum_{k=0}^{N-1}\|F(x^k)\|^2\right]^{1/4} \\
&\leq N^{-1/4}\left(\mathbb{E}\sqrt{\bar{b}_N}\right)^{3/4}\left[\mathbb{E}\frac{\sum_{k=0}^{N-1}\|F(x^k)\|^2}{\bar{b}_N^{3/2}}\right]^{1/4} \\
&\leq N^{-1/4}\left(\kappa_{N,EB}\right)^{3/4}\left(C_{1,EB}+C_{2,EB}\ln\left(q_{N,EB}\right)\right)^{1/4}.
\end{aligned} \tag{126}$$

Then applying Markov's Inequality, with probability at least $1-\delta$,

$$\frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2 \leq \frac{\kappa_{N,EB}^3(C_{1,EB}+C_{2,EB})\ln(\kappa_{N,EB})}{N\delta^4},$$

which finishes the proof.

The other two cases follow by similar arguments. The sub-Weibull case picks up additional constant factors due to needing to apply Markov's inequality with $\delta/2$ and taking the union bound of this event and the event present in Lemma C.10. □

Remark C.2. *From the construction of κ_N, it is clear to see that $\kappa_N = \tilde{\mathcal{O}}\left(N^{1/4}\right)$ when $\sigma_0 > 0, \sigma_1 \geq 0$, leading to*

$$\frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2 = \tilde{\mathcal{O}}\left(\frac{1}{N^{1/4}}\right);$$

while $\kappa_N = \mathcal{O}(1)$ under the scaled noise assumption, where $\sigma_0 = 0$, thus improving the rate to

$$\frac{1}{N}\sum_{k=0}^{N-1}\|F(x^k)\|^2 = \mathcal{O}(1/N).$$

Now, we turn our attention to convergence in terms of the restricted merit function (6).

Theorem C.13 (Proof of Theorem 3.4)**.** *Let Assumptions 1.1 (a'), (b), 3.1, and 3.2 hold and let $\{x^k\}$ be generated by Algorithm 2. Let D in (6) be sufficiently large such that $\|x^0 - x^*\| \leq D$ for some $x^* \in S^*$. Then with probability at least $1-\delta$, we have the following results:*

- *Error bound condition: Let Assumption 3.3 hold as well, assume that $\sigma_1 = 0$ in Assumption 3.2, and assume that the solution set S^* is a singleton. Then,*

$$\begin{aligned}
&Err_D(\mathbb{E}[y^N]) \\
&\leq N^{-1}\Bigg(\frac{c_{N,EB}}{\eta\delta^2}\left(M_{1,EB} + 4M_{2,EB}\ln\sqrt{c_{N,EB}}\right) \\
&+\frac{(3\eta+\eta(1+\sigma_1^2))c_{N,EB}}{\delta^2}\left(C_{1,EB} + C_{2,EB}\ln\sqrt{c_{N,EB}}\right) \\
&+\frac{c_{N,EB}\|x^0-w\|^2}{\eta\delta^2} + \frac{c_{N,EB}\eta}{\delta^2} + \frac{3\eta\sigma_0^2 c_{N,EB}}{2b_0^2\delta} + \frac{\eta c_{N,EB}^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\frac{\left(\eta/2+3\eta^2L\right)c_{N+1,EB}}{\delta^2} + \frac{c_{N,EB}\eta^3L^2\ln\sqrt{c_{N,EB}}}{2\delta^2 b_0^3}\Bigg),
\end{aligned} \tag{127}$$

where

$$\begin{aligned}
y^N &:= \textstyle\sum_{k=1}^N \frac{\bar{x}^k}{N}, \\
c_{N,EB} &:= (2D_{1,EB} + 8D_{2,EB}\ln(e+D_{2,EB}))^2 = \mathcal{O}(N^{1/2}).
\end{aligned}$$

- *Star-strongly monotone: Let Assumption 3.4 hold as well. Then,*

$$\begin{aligned}
&Err_D(\mathbb{E}[y^N]) \\
&\leq N^{-1}\Bigg(\frac{c_{N,SM}}{\eta\delta^2}\left(M_{1,SM} + 4M_{2,SM}\ln\sqrt{c_{N,SM}}\right) \\
&+\frac{(3\eta+\eta(1+\sigma_1^2))c_{N,SM}}{\delta^2}\left(C_{1,SM} + C_{2,SM}\ln\sqrt{c_{N,SM}}\right) \\
&+\frac{c_{N,SM}\|x^0-w\|^2}{\eta\delta^2} + \frac{c_{N,SM}\eta}{\delta^2} + \frac{3\eta\sigma_0^2 c_{N,SM}}{2b_0^2\delta} + \frac{\eta c_{N,SM}^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\frac{\left(\eta/2+3\eta^2L\right)c_{N+1,SM}}{\delta^2} + \frac{c_{N,SM}\eta^3L^2\ln\sqrt{c_{N,SM}}}{2\delta^2 b_0^3}\Bigg),
\end{aligned} \tag{128}$$

where

$$\begin{aligned}
y^N &:= \textstyle\sum_{k=1}^N \frac{\bar{x}^k}{N}, \\
c_{N,SM} &:= (2D_{1,SM} + 8D_{2,SM}\ln(e+D_{2,SM}))^2.
\end{aligned}$$

- *Sub-Weibull noise: Let Assumption A.1 hold as well. Then,*

$$\begin{aligned}
&Err_D(\mathbb{E}[y^N]) \\
&\leq N^{-1}\Bigg(4\frac{c_{N,SW}(\delta/2)}{\eta\delta^2}\left(M_{1,SW}(\delta/2) + 4M_{2,SW}(\delta/2)\ln\sqrt{c_{N,SW}(\delta/2)}\right) \\
&+\frac{4(3\eta+\eta(1+\sigma_1^2))c_{N,SW}(\delta/2)}{\delta^2}\left(C_{1,SW}(\delta/2) + C_{2,SW}(\delta/2)\ln\sqrt{c_{N,SW}(\delta/2)}\right) \\
&+\frac{4c_{N,SW}(\delta/2)\|x^0-w\|^2}{\eta\delta^2} + \frac{4c_{N,SW}\eta}{\delta^2} + \frac{6\eta\sigma_0^2 c_{N,SW}(\delta/2)}{2b_0^2\delta} + \frac{2\eta c_{N,SW}(\delta/2)^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\frac{4\left(\eta/2+3\eta^2L\right)c_{N+1,SW}(\delta/2)}{\delta^2} + \frac{2c_{N,SW}(\delta/2)\eta^3L^2\ln\sqrt{c_{N,SW}(\delta/2)}}{\delta^2 b_0^3}\Bigg),
\end{aligned} \tag{129}$$

where

$$\begin{aligned} y^N &:= \sum_{k=1}^N \frac{\bar{x}^k}{N}, \\ c_{N,SW}(\delta) &:= (2D_{1,SW}(\delta) + 8D_{2,SW}(\delta)\ln(e + D_{2,SW}(\delta)))^2. \end{aligned}$$

Proof. Here we demonstrate the proof for the case when error bound condition holds while the other two cases follow by a nearly identical argument. For any $w \in \mathbb{R}^p$, and from the updated rule of x^{k+1}, we derive the following for each $k \geq 0$,

$$\begin{aligned} &\|x^{k+1} - w\|^2 \\ &= \|x^k - \gamma_k \tilde{F}(\bar{x}^k) - w\|^2, \\ &= \|x^k - w\|^2 + \gamma_k^2 \|\tilde{F}(\bar{x}^k)\|^2 - 2\gamma_k \langle \tilde{F}(\bar{x}^k), x^k - w\rangle, \\ &= \|x^k - w\|^2 + \gamma_k^2 \|\tilde{F}(\bar{x}^k)\|^2 - 2\gamma_k \langle \tilde{F}(\bar{x}^k), x^k - \bar{x}^k\rangle - 2\gamma_k \langle \tilde{F}(\bar{x}^k), \bar{x}^k - w\rangle. \end{aligned} \tag{130}$$

Dividing both sides by γ_k, by the non-increasing nature of γ, and choosing the $x^* \in S^*$ such that $\|x^0 - x^*\| \leq D$,

$$\begin{aligned} &2\langle \tilde{F}(\bar{x}^k), \bar{x}^k - w\rangle \\ &= \frac{1}{\gamma_k}\|x^k - w\|^2 - \frac{1}{\gamma_k}\|x^{k+1} - w\|^2 + \gamma_k \|\tilde{F}(\bar{x}^k)\|^2 - 2\langle \tilde{F}(\bar{x}^k), x^k - \bar{x}^k\rangle, \\ &= \frac{1}{\gamma_{k-1}}\|x^k - w\|^2 - \frac{1}{\gamma_k}\|x^{k+1} - w\|^2 + \gamma_k \|\tilde{F}(\bar{x}^k)\|^2 - 2\alpha_k \langle \tilde{F}(\bar{x}^k), \tilde{F}(x^k)\rangle \\ &\quad + \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\|x^k - w\|^2, \\ &\leq \frac{1}{\gamma_{k-1}}\|x^k - w\|^2 - \frac{1}{\gamma_k}\|x^{k+1} - w\|^2 + \gamma_k \|\tilde{F}(\bar{x}^k)\|^2 - 2\alpha_k \langle \tilde{F}(\bar{x}^k), \tilde{F}(x^k)\rangle \\ &\quad + 2\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\|x^k - x^*\|^2 + 2\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\|x^* - w\|^2, \end{aligned} \tag{131}$$

in addition,

$$\begin{aligned} &-\frac{2}{\gamma_{k-1}}\|x^k - x^*\|^2 \\ &= -\frac{2}{\gamma_{k-1}}\|x^{k-1} - \gamma_{k-1}\tilde{F}(\bar{x}^{k-1}) - x^*\|^2, \\ &= -\frac{2}{\gamma_{k-1}}\|x^{k-1} - x^*\|^2 - 2\gamma_{k-1}\|\tilde{F}(\bar{x}^{k-1})\|^2 + 4\langle \tilde{F}(\bar{x}^{k-1}), \bar{x}^{k-1} - x^*\rangle \\ &\quad + 4\langle \tilde{F}(\bar{x}^{k-1}), x^{k-1} - \bar{x}^{k-1}\rangle. \end{aligned} \tag{132}$$

Combining results from (131) and (132), taking conditional expectation $\mathbb{E}_{k-1}[\cdot]$ on both sides and using the tower property of expectation, we obtain,

$$\begin{aligned} &2\mathbb{E}_{k-1}\left[\langle F(\bar{x}^k), \bar{x}^k - w\rangle\right] \\ &\leq \mathbb{E}_{k-1}\left[\frac{1}{\gamma_{k-1}}\|x^k - w\|^2 - \frac{1}{\gamma_k}\|x^{k+1} - w\|^2 + \frac{2}{\gamma_k}\|x^k - x^*\|^2 - \frac{2}{\gamma_{k-1}}\|x^{k-1} - x^*\|^2\right] \\ &+ \mathbb{E}_{k-1}\left[\gamma_k \|\tilde{F}(\bar{x}^k)\|^2 - 2\gamma_{k-1}\|\tilde{F}(\bar{x}^{k-1})\| + 4\langle F(\bar{x}^{k-1}), \bar{x}^{k-1} - x^*\rangle\right] \\ &+ \mathbb{E}_{k-1}\left[-2\alpha_k \langle F(\bar{x}^k), \tilde{F}(x^k)\rangle + 4\alpha_{k-1}\langle F(\bar{x}^{k-1}), \tilde{F}(x^{k-1})\rangle\right]. \end{aligned} \tag{133}$$

To proceed, we focus on the final two inner products,

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[-2\alpha_k\langle F(\bar{x}^k), \tilde{F}(x^k)\rangle\right] \\
&= \mathbb{E}_{k-1}\left[-2\alpha_k\langle F(\bar{x}^k) - F(x^k) + F(x^k), \tilde{F}(x^k)\rangle\right] \\
&\leq \mathbb{E}_{k-1}\left[2\alpha_k^2 L\|\tilde{F}(x^k)\|^2 + 2|\alpha_k - \alpha_{k-1}|\|F(x^k)\|\|\tilde{F}(x^k)\| - 2\alpha_{k-1}\|F(x^k)\|^2\right] \\
&\leq \mathbb{E}_{k-1}\left[2\alpha_k^2 L\|\tilde{F}(x^k)\|^2 + 2\eta\frac{\|F(x^k)\|\|\tilde{F}(x^k)\|^2}{b_k(b_{k+1}^2 + b_k^2)} - 2\alpha_{k-1}\|F(x^k)\|^2\right],
\end{aligned} \tag{134}$$

here we apply Lipschitz continuity of F and de-correlated step size α_{k-1} to derive the first inequality; and the second inequality comes from the following identity:

$$\begin{aligned}
|\alpha_k - \alpha_{k-1}| &= \eta\left|\frac{1}{b_{k+1}} - \frac{1}{b_k}\right| \\
&= \eta\left|\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}b_k(b_{k+1} + b_k)(b_{k+1}^2 + b_k^2)}\right| \\
&\leq \frac{\|\tilde{F}(x^k)\|}{b_k(b_{k+1}^2 + b_k^2)}.
\end{aligned}$$

Similar to the previous analysis, we apply Young's inequality and Hölder's inequality to derive the following,

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[2\eta\frac{\|F(x^k)\|\|\tilde{F}(x^k)\|^2}{b_k(b_{k+1}^2 + b_k^2)}\right] \\
&= \mathbb{E}_{k-1}\left[2\alpha_{k-1}\|F(x^k)\|\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2 + b_k^2)}\right]\right] \\
&\leq \mathbb{E}_{k-1}\left[\alpha_{k-1}\|F(x^k)\|^2 + \alpha_{k-1}\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^2 + b_k^2}\right]^2\right] \\
&\leq \mathbb{E}_{k-1}\left[\alpha_{k-1}\|F(x^k)\|^2 + \alpha_{k-1}\mathbb{E}_k\|\tilde{F}(x^k)\|^2\mathbb{E}_k\left[\frac{\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2 + b_k^2)^2}\right]\right] \\
&\leq \mathbb{E}_{k-1}\left[\alpha_{k-1}\|F(x^k)\|^2 + \eta[\sigma_0^2 + (1+\sigma_1^2)\|F(x^k)\|^2]\mathbb{E}_k\left[\frac{b_k\|\tilde{F}(x^k)\|^2}{(b_{k+1}^2 + b_k^2)b_k^2 b_{k+1}^2}\right]\right] \\
&\leq \mathbb{E}_{k-1}\Big[\alpha_{k-1}\|F(x^k)\|^2 + \eta\sigma_0^2\mathbb{E}_k\left[b_N\left(\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right)\right] \\
&\quad + \eta(1+\sigma_1^2)\|F(x^k)\|^2\mathbb{E}_k\left[b_N\left(\frac{1}{b_k^2} - \frac{1}{b_{k+1}^2}\right)\right]\Big].
\end{aligned}$$

Now, let E be the event that

$$\bar{b}_N \leq \frac{c_N}{\delta^2} \tag{135}$$

occurs. Then, by Corollary C.11.1, it follows that event E occurs with probability at least $1-\delta$ and $b_N \leq \bar{b}_N^{1/2} \leq$

$c_N^{1/2}/\delta$. Thus,

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[2\eta\frac{\|F(x^k)\|\|\tilde{F}(x^k)\|^2}{b_k(b_{k+1}^2+b_k^2)}\right]\\
&\leq \mathbb{E}_{k-1}\Big[\alpha_{k-1}\|F(x^k)\|^2+\frac{\eta\sigma_0^2c_N^{1/2}}{\delta}\mathbb{E}_k\left[\left(\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right)\right]\\
&\quad+\frac{\eta(1+\sigma_1^2)c_N^{1/2}}{\delta}\|F(x^k)\|^2\mathbb{E}_k\left[\left(\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right)\right]\Big].
\end{aligned}\tag{136}$$

Next, we apply Young's inequality with $\lambda=c_N^{1/2}(1+\sigma_1^2)/\delta$, and inequality (42) to process the last term,

$$\begin{aligned}
&\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\|F(x^k)\|^2\mathbb{E}_k\left[\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right]\\
&=\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\left(\mathbb{E}_k\left[\frac{\|F(x^{k-1})\|^2}{b_k^2}-\frac{\|F(x^k)\|^2}{b_{k+1}^2}\right]+\mathbb{E}_k\left[\frac{\|F(x^k)\|^2-\|F(x^{k-1})\|^2}{b_k^2}\right]\right)\\
&\leq\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\left(\mathbb{E}_k\left[\frac{\|F(x^{k-1})\|^2}{b_k^2}-\frac{\|F(x^k)\|^2}{b_{k+1}^2}\right]+\mathbb{E}_k\left[\frac{2\|F(x^k)\|\|F(x^k)-F(x^{k-1})\|}{b_k^2}\right]\right)\\
&\leq\eta\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\mathbb{E}_k\left[\frac{\|F(x^{k-1})\|^2}{b_k^2}-\frac{\|F(x^k)\|^2}{b_{k+1}^2}\right]\\
&\quad+\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\mathbb{E}_k\left[\frac{\|F(x^k)\|^2}{\lambda b_k}+\lambda L^2\eta^2\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2b_k^3}\right]\\
&\leq\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\mathbb{E}_k\left[\frac{\|F(x^{k-1})\|^2}{b_k^2}-\frac{\|F(x^k)\|^2}{b_{k+1}^2}\right]+\frac{c_N(1+\sigma_1^2)^2\eta^3L^2}{b_0^3\delta^2}\mathbb{E}_k\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&\quad+\alpha_{k-1}\mathbb{E}_k[\|F(x^k)\|^2],
\end{aligned}\tag{137}$$

combining results in (134), (136), (137), we finally obtain

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[-2\alpha_k\langle F(\bar{x}^k),\tilde{F}(x^k)\rangle\right]\\
&\leq 2\eta^2L\mathbb{E}_{k-1}\left[\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^2}\right]+\frac{\eta\sigma_0^2c_N^{1/2}}{\delta}\mathbb{E}_{k-1}\left[\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right],\\
&+\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\mathbb{E}_{k-1}\left[\frac{\|F(x^{k-1})\|^2}{b_k^2}-\frac{\|F(x^k)\|^2}{b_{k+1}^2}\right]+\frac{c_N(1+\sigma_1^2)^2\eta^3L^2}{b_0^3\delta^2}\mathbb{E}_{k-1}\left[\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right].
\end{aligned}\tag{138}$$

Next, we bound $\mathbb{E}_{k-1}\left[4\alpha_{k-1}\langle F(\bar{x}^{k-1}),\tilde{F}(x^{k-1})\rangle\right]$ by applying a similar approach.

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[4\alpha_{k-1}\langle F(\bar{x}^{k-1}),\tilde{F}(x^{k-1})\rangle\right]\\
&=\mathbb{E}_{k-1}\left[4\alpha_{k-1}\langle F(\bar{x}^{k-1})-F(x^{k-1})+F(x^{k-1}),\tilde{F}(x^{k-1})\rangle\right]\\
&\leq\mathbb{E}_{k-1}\left[4\alpha_{k-1}^2L\|\tilde{F}(x^{k-1})\|^2+4|\alpha_{k-1}-\alpha_{k-2}|\|F(x^{k-1})\|\|\tilde{F}(x^{k-1})\|+4\alpha_{k-2}\|F(x^{k-1})\|^2\right]\\
&\leq\mathbb{E}_{k-1}\left[4\alpha_{k-1}^2L\|\tilde{F}(x^{k-1})\|^2+4\eta\frac{\|F(x^{k-1})\|\|\tilde{F}(x^{k-1})\|^2}{b_{k-1}(b_{k-1}^2+b_k^2)}+4\alpha_{k-2}\|F(x^{k-1})\|^2\right].
\end{aligned}\tag{139}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[4\eta\frac{\|F(x^{k-1})\|\|\tilde{F}(x^{k-1})\|^2}{b_{k-1}(b_{k-1}^2+b_k^2)}\right]\\
&\leq 2\alpha_{k-2}\|F(x^{k-1})\|^2+2\alpha_{k-2}\mathbb{E}_{k-1}[\|\tilde{F}(x^{k-1})\|^2]\mathbb{E}_{k-1}\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{(b_{k-1}^2+b_k^2)^2}\right]\\
&\leq 2\alpha_{k-2}\|F(x^{k-1})\|^2+2\eta\mathbb{E}_{k-1}\left[\left(\sigma_0^2+(1+\sigma_1^2)\|F(x^{k-1})\|^2\right)\right]\mathbb{E}_{k-1}\left[\frac{b_{k-1}\|\tilde{F}(x^{k-1})\|^2}{(b_{k-1}^2+b_k^2)b_{k-1}^2b_k^2}\right]\\
&\leq 2\alpha_{k-2}\|F(x^{k-1})\|^2+2\eta\sigma_0^2\mathbb{E}_{k-1}\left[b_N\left(\frac{1}{b_{k-1}^2}-\frac{1}{b_k^2}\right)\right]+2\eta(1+\sigma_1^2)\mathbb{E}_{k-1}\left[\frac{\|F(x^{k-1})\|^2}{b_{k-1}}\right].
\end{aligned}\tag{140}$$

Combining this with (139), we get and noting that event E occurs with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[4\alpha_{k-1}\langle F(\bar{x}^{k-1}),\tilde{F}(x^{k-1})\rangle\right]\\
&\leq 4\eta^2L\mathbb{E}_{k-1}\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{b_{k-1}^2}\right]+2\eta\sigma_0^2\mathbb{E}_{k-1}\left[b_N\left(\frac{1}{b_{k-1}^2}-\frac{1}{b_k^2}\right)\right]\\
&+\left(6\eta+2\eta\sigma_1^2\right)\mathbb{E}_{k-1}\left[\frac{b_N\|F(x^{k-1})\|^2}{b_{k-1}}\right]\\
&\leq 4\eta^2L\mathbb{E}_{k-1}\left[\frac{\|\tilde{F}(x^{k-1})\|^2}{b_{k-1}^2}\right]+\frac{2\eta\sigma_0^2c_N^{1/2}}{\delta}\mathbb{E}_{k-1}\left[\frac{1}{b_{k-1}^2}-\frac{1}{b_k^2}\right]\\
&+(6\eta+2\eta\sigma_1^2)\mathbb{E}_{k-1}\left[\frac{\|F(x^{k-1})\|^2}{b_{k-1}}\right]
\end{aligned}\tag{141}$$

Therefore, combining (133), (138), and (141), taking the total expectation and summing from $k=1,\ldots,N$,

$$\begin{aligned}
&2\mathbb{E}\left[\textstyle\sum_{k=1}^N\langle F(\bar{x}^k),\bar{x}^k-w\rangle\right]\\
&\leq\mathbb{E}\left[\textstyle\sum_{k=1}^N\frac{1}{\gamma_{k-1}}\|x^k-w\|^2-\frac{1}{\gamma_k}\|x^{k+1}-w\|^2+\frac{2}{\gamma_k}\|x^k-x^*\|^2-\frac{2}{\gamma_{k-1}}\|x^{k-1}-x^*\|^2\right]\\
&+\eta\mathbb{E}\left[\textstyle\sum_{k=1}^N\frac{\|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}}\right]+4\mathbb{E}\left[\textstyle\sum_{k=1}^N\langle F(\bar{x}^{k-1}),\bar{x}^{k-1}-x^*\rangle\right]\\
&+2\eta^2L\mathbb{E}\left[\textstyle\sum_{k=1}^N\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^2}\right]+\frac{\eta\sigma_0^2c_N^{1/2}}{\delta}\mathbb{E}\left[\textstyle\sum_{k=1}^N\left(\frac{1}{b_k^2}-\frac{1}{b_{k+1}^2}\right)\right]\\
&+\frac{\eta c_N^{1/2}(1+\sigma_1^2)}{\delta}\mathbb{E}\left[\textstyle\sum_{k=1}^N\left(\frac{\|F(x^{k-1})\|^2}{b_k^2}-\frac{\|F(x^k)\|^2}{b_{k+1}^2}\right)\right]+\frac{c_N(1+\sigma_1^2)^2\eta^3L^2}{b_0^3\delta^2}\mathbb{E}\left[\textstyle\sum_{k=1}^N\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right]\\
&+4\eta^2L\mathbb{E}\left[\textstyle\sum_{k=1}^N\frac{\|\tilde{F}(x^{k-1})\|^2}{b_{k-1}^2}\right]+\frac{2\eta c_N^{1/2}\sigma_0^2}{\delta}\mathbb{E}\left[\textstyle\sum_{k=1}^N\left(\frac{1}{b_{k-1}^2}-\frac{1}{b_k^2}\right)\right]\\
&+\left(6\eta+2\eta\sigma_1^2\right)\mathbb{E}\left[\textstyle\sum_{k=1}^N\frac{\|F(x^{k-1})\|^2}{b_{k-1}}\right].
\end{aligned}\tag{142}$$

Since the step size sequences are non-decreasing over time, by Assumption 1.1 (a), we have

$$\begin{aligned}
&\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\langle F(\bar{x}^k), \bar{x}^k - w\rangle\right] \\
&\leq \mathbb{E}\left[\frac{\|x^N - x^*\|^2}{\gamma_{N-1}}\right] + 2\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\frac{\gamma_{k-1}}{\gamma_{N-1}}\langle F(\bar{x}^{k-1}), \bar{x}^{k-1} - x^*\rangle\right] \\
&\quad + \left(3\eta + \eta(1+\sigma_1^2)\right)\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\frac{\bar{b}_N\|F(x^{k-1})\|^2}{\bar{b}_{k-1}b_{k-1}}\right] \\
&+\mathbb{E}\left[\frac{\|x^0 - w\|^2}{\gamma_1}\right] + \mathbb{E}\left[\frac{\|x^1 - x^0\|^2}{\gamma_1}\right] + \frac{3\eta\sigma_0^2 c_N^{1/2}}{2b_0^2\delta} + \frac{\eta c_N^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\frac{\eta}{2}\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\frac{\|\tilde{F}(\bar{x}^k)\|^2}{\bar{b}_{k+1}}\right] + \eta^2 L\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\frac{\|\tilde{F}(x^k)\|^2}{b_{k+1}^2}\right] \\
&\quad + \frac{c_N(1+\sigma_1^2)^2\eta^3L^2}{2b_0^3\delta^2}\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\frac{\|\tilde{F}(\bar{x}^{k-1})\|^2}{\bar{b}_k^2}\right] + 2\eta^2 L\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\frac{\|\tilde{F}(x^{k-1})\|^2}{b_{k-1}^2}\right].
\end{aligned} \tag{143}$$

Note that in the case when S^* is a singleton set, we have $\|x^N - x^*\|^2 = \mathrm{dist}(x^N, S^*)^2$. Thus, whenever event E occurs, by the results of Lemma C.8 and Corollary C.11.1, we can bound the first three terms with order $\tilde{\mathcal{O}}(N^{1/2})$. In addition, applying Lemma A.1 and Corollary C.11.1, we can also bound the last four terms with order $\tilde{\mathcal{O}}(N^{1/2})$ as well. Specifically, under the error bound condition, we can derive the following with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}\left[\textstyle\sum_{k=1}^{N}\langle F(\bar{x}^k), \bar{x}^k - w\rangle\right] \\
&\leq \frac{c_N}{\eta\delta^2}\left(M_{1,EB} + 4M_{2,EB}\ln\mathbb{E}[\bar{b}_N^{1/2}]\right) + \frac{\left(3\eta + \eta(1+\sigma_1^2)\right)c_N}{\delta^2}\left(C_{1,EB} + C_{2,EB}\ln\mathbb{E}[\bar{b}_N^{1/2}]\right) \\
&+\frac{c_N\|x^0 - w\|^2}{\eta\delta^2} + \mathbb{E}\left[\frac{c_N\gamma_1^2\|\tilde{F}(\bar{x}^0)\|^2}{\eta\delta^2}\right] + \frac{3\eta\sigma_0^2 c_N^{1/2}}{2b_0^2\delta} + \frac{\eta c_N^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\left(\eta/2 + 3\eta^2 L\right)\mathbb{E}[\bar{b}_{N+1}] + \frac{c_N(1+\sigma_1^2)^2\eta^3L^2}{2b_0^3\delta^2}\ln\mathbb{E}[\bar{b}_N^{1/2}] \\
&\leq \frac{2c_N}{\eta\delta^2}\left(M_{1,EB} + 4M_{2,EB}\ln\sqrt{c_N} + D_{S^*}^2\right) \\
&+\frac{\left(3\eta + \eta(1+\sigma_1^2)\right)c_N}{\delta^2}\left(C_{1,EB} + C_{2,EB}\ln\sqrt{c_N}\right) \\
&+\frac{c_N\|x^0 - w\|^2}{\eta\delta^2} + \frac{c_N\eta}{\delta^2} + \frac{3\eta\sigma_0^2 c_N^{1/2}}{2b_0^2\delta} + \frac{\eta c_N^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\frac{\left(\eta/2 + 3\eta^2 L\right)c_{N+1}}{\delta^2} + \frac{c_N(1+\sigma_1^2)^2\eta^3L^2\ln\sqrt{c_N}}{2\delta^2 b_0^3} \\
&= \frac{2c_N}{\eta\delta^2}\left(M_{1,EB} + 4M_{2,EB}\ln\sqrt{c_N} + D_{S^*}^2\right) \\
&+\frac{\left(3\eta + \eta(1+\sigma_1^2)\right)c_N}{\delta^2}\left(C_{1,EB} + C_{2,EB}\ln\sqrt{c_N}\right) \\
&+\frac{c_N\|x^0 - w\|^2}{\eta\delta^2} + \frac{c_N\eta}{\delta^2} + \frac{3\eta\sigma_0^2 c_N^{1/2}}{2b_0^2\delta} + \frac{\eta c_N^{1/2}(1+\sigma_1^2)\|F(x^0)\|^2}{2b_0^2\delta} \\
&+\frac{\left(\eta/2 + 3\eta^2 L\right)c_{N+1}}{\delta^2} + \frac{c_N\eta^3L^2\ln\sqrt{c_N}}{2\delta^2 b_0^3},
\end{aligned} \tag{144}$$

where

$$c_N := 2D_{1,EB} + 8D_{2,EB}\ln(e + D_{2,EB}) = \mathcal{O}(N^{1/2}),$$

and the equality comes from the fact that $\sigma_1 = 0$ in the error bound condition. Lastly, we apply monotonicity of F, take supremum over all $w \in D_{x^0}$ and divide both sides by N to get:

$$Err_D(\mathbb{E}[y^N]) = \sup_{w \in D_{x^0}} \langle F(w), \mathbb{E}\left[\sum_{k=1}^{N} \frac{\bar{x}^k}{N}\right] - w\rangle = \tilde{\mathcal{O}}\left(\frac{1}{N^{1/2}}\right),$$

as desired.

The other two cases follow by similar arguments. The sub-Weibull case picks up additional constant factors due to needing to apply the results of Corollary C.11.1 with $\delta/2$ and taking the union bound of this event and the event present in Lemma C.10. □

Remark C.3. *Now we discuss the convergence rate in three different settings. Based on the results in Lemma C.11, we know that the order of $D_1's$ depends on the values of σ_0, σ_1 in the noise assumption. Under the error bound condition, we have $\sigma_1 = 0, \sigma_0 > 0$, leading to $D_{1,EB} = \mathcal{O}\left(N^{1/4}\right)$ and if the optimal solution is unique, $Err(\mathbb{E}[y_N]) = \tilde{\mathcal{O}}(\frac{1}{N^{1/2}})$. While in the other two settings, if both σ_0, σ_1 are positive, then $D_1's = \mathcal{O}(N^{1/4})$, achieving the same convergence rate as in error bound condition. On the other hand, if we consider the scaled noise setting, in which $\sigma_0 = 0$, then both $c_{N's}$ and $\ln \mathbb{E}\sqrt{\bar{b}_N}$ can be bounded by constants. Thus the convergence rate of the expected restricted merit function can be improved to $\mathcal{O}(1/N)$, which is optimal convergence rate for deterministic extragradient scheme and beyond the convergence rate for other adaptive extragradient-type methods.*

D EXPERIMENTS

D.1 Experimental Setup

Computing resources. All the experiments are implemented in Python 3, running on a laptop with 2.3 GHz Quad-Core Intel Core i5 and 16 GB memory.

Positive definite matrices. To generate the positive definite matrix C, we first generate a matrix U with entries uniformly distributed from $[0, 1]$, then we can obtain a semi-definite matrix UU^T. By adding a scaled identity matrix with a positive diagonal, we can derive a symmetric positive definite matrix C.

Stochastic estimators. We generate the stochastic estimate $\tilde{F}(\theta, \phi)$ by adding two different noise distributions Z:

- Gaussian noise $Z \sim N(0, \sigma I_d)$ with $\sigma = 0.5$
- scaled Gaussian noise $Z \sim N(0, \sigma I_d)$ with $\sigma = 0.01\|(\theta^t, \phi^t)\|^2$

Hyperparameters. The general setting of double step sizes in DSEG Hsieh et al. (2020) is given as

$$\alpha_k = \frac{\eta_1}{(k+b)^\alpha}, \quad \gamma_k = \frac{\eta_2}{(k+b)^\rho}.$$

For DSEG, we follow the hyperparameter settings in Hsieh et al. (2020) for bilinear games and set

$$\eta_1 = 1, \quad \eta_2 = 0.1, \quad b = 19, \quad \alpha = 0.1, \quad \rho = 0.9,$$

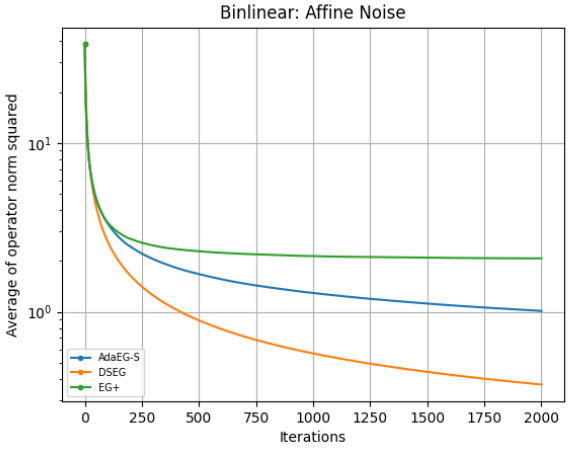
while for the strongly convex-concave game, we set

$$\eta_1 = 0.25, \quad \eta_2 = 0.15, \quad b = 19, \quad \alpha = 0.1, \quad \rho = 0.9,$$

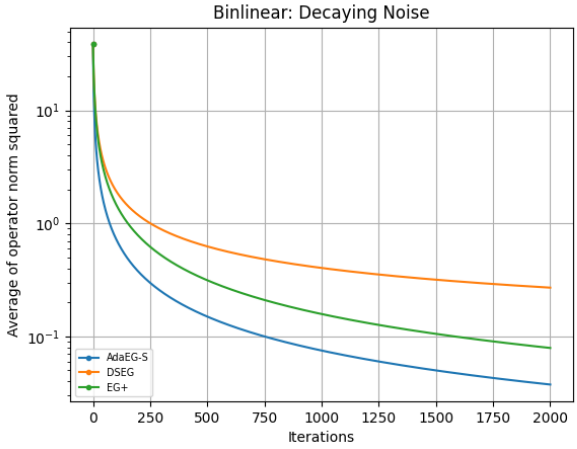
which was found via hyperparameter tuning.

For the fixed stepsize variant (which we refer to as EG+) Diakonikolas et al. (2021), we use the stepsizes

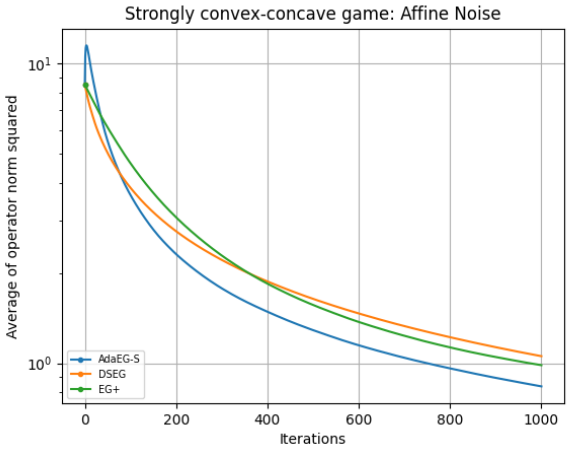
$$\alpha = \frac{\eta_1}{b^\alpha}, \quad \gamma_k = \frac{\eta_2}{b^\gamma}$$



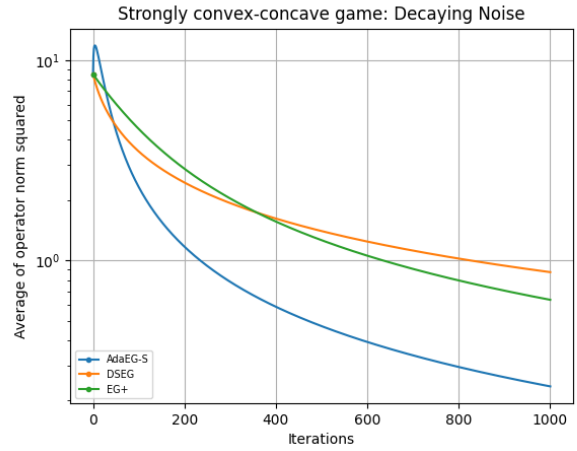
(a) Bilinear game under affine noise.



(b) Bilinear game under decaying noise.



(c) Strongly convex-concave game under affine noise.



(d) Strongly convex-concave game under decaying noise.

Figure 5: Averaged operator norm squared of the iterates.

with the settings of

$$\eta_1 = 1, \quad \eta_2 = 0.1, \quad b = 19, \quad \alpha = 0.1, \quad \rho = 0.9,$$

for bilinear games and

$$\eta_1 = 0.2, \quad \eta_2 = 0.05, \quad b = 19, \quad \alpha = 0.1, \quad \rho = 0.9,$$

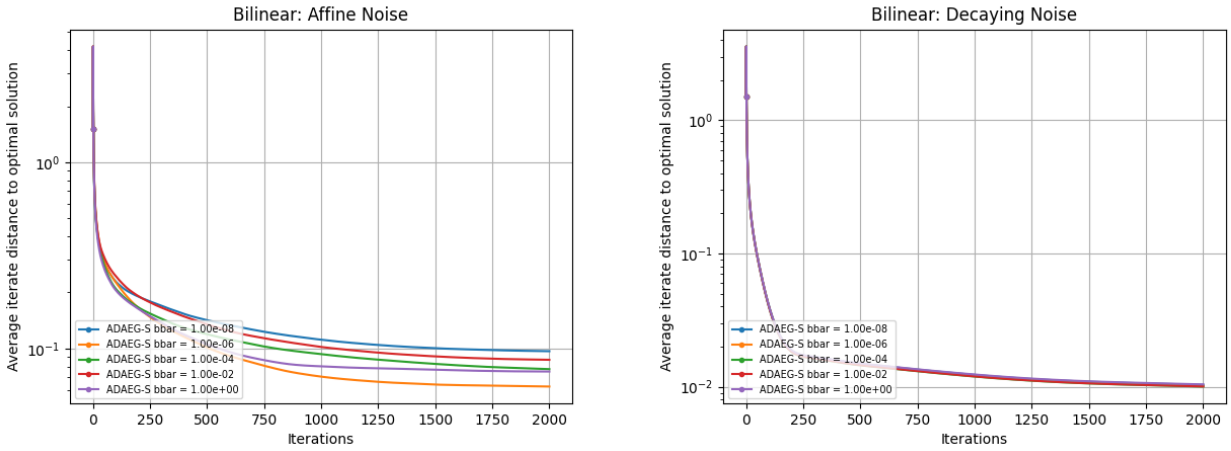
for strongly convex-concave games, which are settings that were found via hyperpameter tuning.

For every problem, we set $\eta = 1, \bar{b}_0 = 10^{-2}$ for our implementation of Algorithm 2.

Code. The code is available at: `https://github.com/MichaelJONeill/Adaptive-Extragradient-Methods`.

D.2 Additional Figures

Here, for completeness, we also plot the average of the operator norm squared in Figure 5 with respect to the experiments we detailed in the main body of the paper. As we can see, Algorithm 2 performs well in all settings and is particularly effective when the noise is decaying, though the DSEG method does outperform it in the case of affine noise on the bilinear game.



(a) Bilinear game under affine noise.

(b) Bilinear game under decaying noise.

Figure 6: Sensitivity of Algorithm 2 to $\bar{b}_0$. Values of $\bar{b}_0 = \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$.

D.3 Additional Experiments

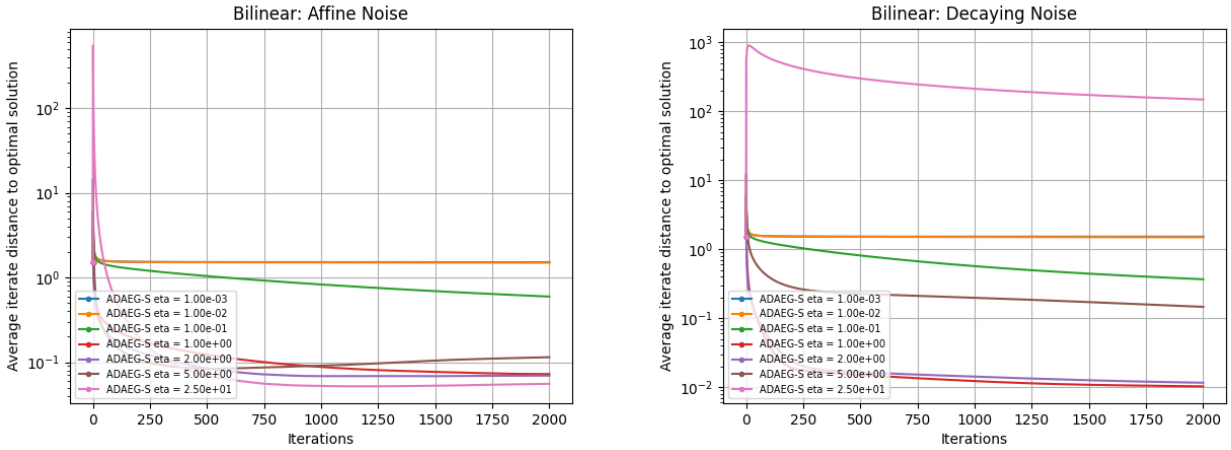
In addition to the experiments we performed in the main body of the paper, we also ran a number of experiments to investigate the sensitivity of Algorithm 2 to its hyperparameters on the bilinear saddle point problem. For brevity, we only provide plots in terms of the average iterate's distance to the solution. To begin, ran Algorithm 2 $\eta = 1$ and varied the values of $\bar{b}_0$ between 10^{-8} and 1. As we can see in Figure 6, the algorithm is relatively insensitive to this parameter, especially in the decaying noise setting. In addition, we may have observed superior performance in our experiments in the main body had we chosen a smaller value of $\bar{b}_0$.

Next, we investigated the sensitivity of Algorithm 2 to the parameter η. The results of these experiments, shown in Figure 7 shows that our proposed algorithm is significantly more sensitive to η than to $\bar{b}_0$. We see that smaller values of η (such as $\eta = 10^{-3}, 10^{-2}, 10^{-1}$) exhibit relatively slow convergence while larger values of η perform well. However, for the decaying noise case, when η becomes too large (such as 5 or 25), performance is significantly hindered. Thus, there appears to be a "sweet spot" around $\eta = 1$ or $\eta = 2$, where the algorithm performs well in both settings. Unfortunately, this does suggest that some tuning of η may be necessary to obtain practical performance, but now we show that the non-adaptive methods, DSEG Hsieh et al. (2020) and EG+ Diakonikolas et al. (2021) exhibit more sensitivity to their hyperparameters, which is suggestive that Algorithm 2 is more robust to its hyperparameters than other methods.

To see this sensitivity, we consider the DSEG method of Hsieh et al. (2020), where we will use the settings $\eta_2 = 0.1\eta_1$, $b = 19$, $\alpha = 0.1$, $\rho = 0.9$ and vary the value of $\eta_1 = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 2\}$ (we do not plot any values larger than $\eta_1 = 2$, as we observed divergence of the algorithm in this case). We plot the behavior of this algorithm in Figure 8. As we can see, this algorithm is also highly sensitive to the choice of η_1 and exhibits improved convergence as η_1 grows. However, as mentioned above, we observed divergence in cases where η_1 was too large (e.g. $\eta_1 = 5$), so there is a limit to how large this parameter can be taken.

Next, we consider the EG+ method of Diakonikolas et al. (2021), where we will use the settings $\eta_2 = 0.1\eta_1$, $b = 19$, $\alpha = 0.1$, $\rho = 0.9$ and vary the value of $\eta_1 = \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ (we do not plot any values larger than $\eta_1 = 1$, as we observed divergence of the algorithm in this case). We plot the behavior of this algorithm in Figure 9. As we can see, this algorithm is also highly sensitive to the choice of η_1 and exhibits improved convergence as η_1 grows. However, as mentioned above, we observed divergence in cases where η_1 was too large (e.g. $\eta_1 = 2$), so there is a limit to how large this parameter can be chosen.

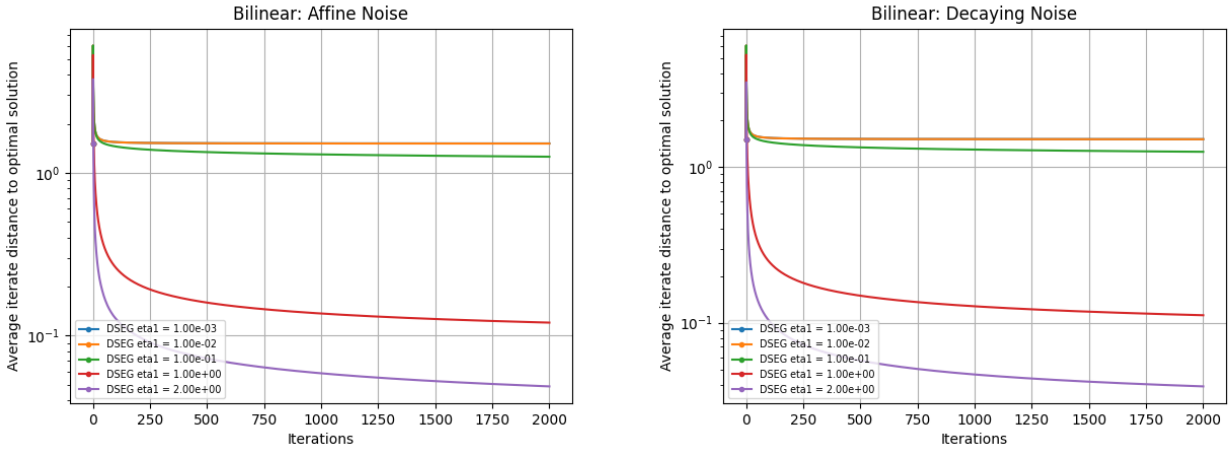
Finally, in order to compare more closely between the methods, we plot all three with values of their respective parameters that performed well in the previous experiments. For Algorithm 2, we choose $\eta = \{1, 2, 5\}$, for DSEG, we choose $\eta_1 = \{1, 2\}$ and for EG+ we choose $\eta_1 = 1$. As we can see, in the case of affine noise, EG+ outperforms Algorithm 2 when η_1 is chosen well. However, this algorithm is extremely sensitive to η_1, as we have observed before, given that it begins diverging when $\eta_1 = 2$ and has significantly impaired performance for $\eta_1 = 10^{-1}$.



(a) Bilinear game under affine noise.

(b) Bilinear game under scaled noise.

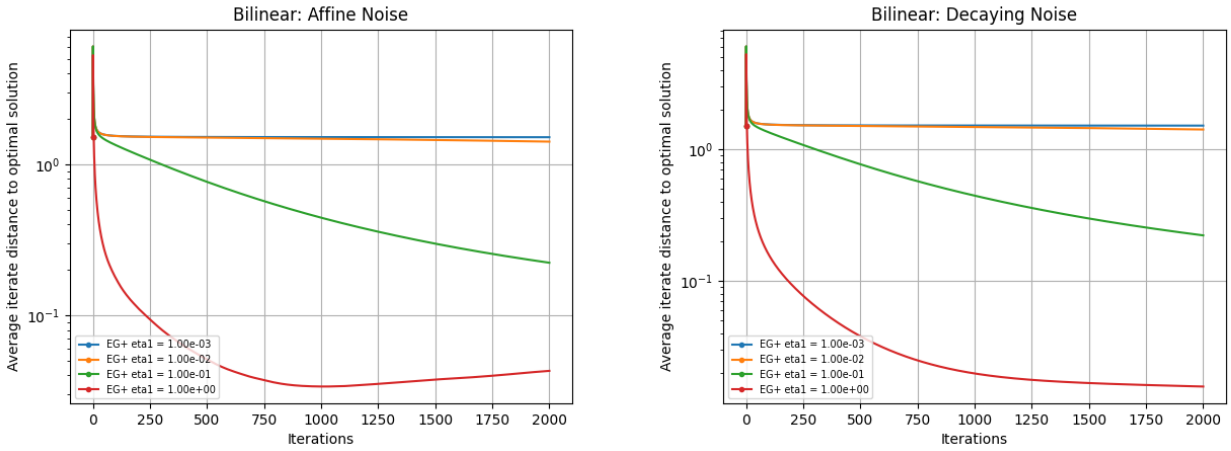
Figure 7: Sensitivity of Algorithm 2 to η. Values of $\eta = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 2, 5, 25\}$.



(a) Bilinear game under affine noise.

(b) Bilinear game under decaying noise.

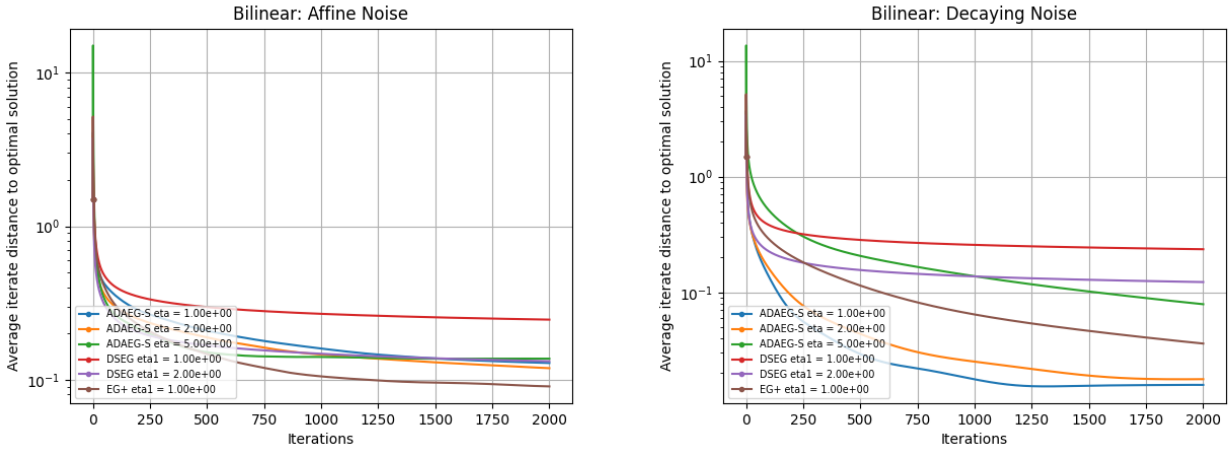
Figure 8: Sensitivity of DSEG of Hsieh et al. (2022) to η_1. Values of $\eta_1 = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 2\}$.



(a) Bilinear game under affine noise.

(b) Bilinear game under decaying noise.

Figure 9: Sensitivity of EG+ of Diakonikolas et al. (2021) to η_1. Values of $\eta_1 = \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$.



(a) Bilinear game under affine noise.

(b) Bilinear game under decaying noise.

Figure 10: Sensitivity of EG+ of Diakonikolas et al. (2021) to η_1. Values of $\eta_1 = \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$.

On the other hand, Algorithm 2 performs similarly for a much wider range of choices for η suggesting that it is an effective adaptive stepsize rule, that is less dependent on the parameters of the method. In addition, in the decaying noise case, Algorithm 2 outperforms all other methods for multiple choices of η. Finally, we wish to note that we never observed divergence of Algorithm 2 when η was chosen to be large, but could easily find values for which a large value of η_1 lead to divergence in the other algorithms.