Accelerated Stochastic Mirror Descent: From Continuous-time Dynamics to Discrete-time Algorithms

Pan Xu[*]
Department of Computer Science
University of Virginia

Tianhao Wang[*]
School of Mathematical Sciences
Univ. of Science & Technology of China

Quanquan Gu
Department of Computer Science
University of Virginia

Abstract

We present a new framework to analyze accelerated stochastic mirror descent through the lens of continuous-time stochastic dynamic systems. It enables us to design new algorithms, and perform a unified and simple analysis of the convergence rates of these algorithms. More specifically, under this framework, we provide a Lyapunov function based analysis for the continuous-time stochastic dynamics, as well as several new discrete-time algorithms derived from the continuous-time dynamics. We show that for general convex objective functions, the derived discrete-time algorithms attain the optimal convergence rate. Empirical experiments corroborate our theory.

1 INTRODUCTION

We consider the constrained optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function and $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$. Denote by $\mathbf{x}^*$ the minimizer of (1.1), i.e., $f(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. Projected gradient descent (Luenberger et al., 1984) can be used to solve (1.1) if $\mathcal{X}$ is a simple constraint set. When the computation of the gradient of f is expensive, or it is not directly accessible, stochastic gradient $\nabla\widetilde{f}(\mathbf{x},\xi)$ can be used, where ξ is some random variable and one often assumes the stochastic gradient is an unbiased estimator of the full gradient, i.e., $\mathbb{E}[\nabla\widetilde{f}(\mathbf{x},\xi)|\mathbf{x}] = \nabla f(\mathbf{x})$. The projected stochastic gradient descent (SGD) takes the following update formula

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{X}}(\mathbf{x}_k - \eta_k\nabla\widetilde{f}(\mathbf{x}_k,\xi_k)), \tag{1.2}$$

where $\eta_k > 0$ is the step size and $\Pi_{\mathcal{X}}$ denotes the Euclidean projection onto $\mathcal{X}$. When the constraint set $\mathcal{X}$ is endowed with a Bregman divergence (Bregman, 1967) that is induced by a continuously differentiable and strongly convex distance generating function $h : \mathcal{X} \to \mathbb{R}$, the projected SGD algorithm can be generalized to the stochastic mirror descent (SMD) (Nemirovskii et al., 1983):

$$\begin{cases}\mathbf{y}_{k+1} = \nabla h(\mathbf{x}_k) - \eta_k\nabla\widetilde{f}(\mathbf{x}_k,\xi_k),\\ \mathbf{x}_{k+1} = \nabla h^*(\mathbf{y}_{k+1}),\end{cases} \tag{1.3}$$

where h^* is the conjugate function of h. It is easy to show that SGD is a special case of SMD when choosing $h(\cdot) = 1/2\|\cdot\|_2^2$. It is well known that the expected objective function value after k iterations of SMD (i.e., $\mathbb{E}[f(\mathbf{x}_k)]$) converges to the minimal value $f(\mathbf{x}^*)$ at an optimal rate of $O(G/\sqrt{k})$, where G is the Lipschitz constant of f, and k is the number of iterations.

In order to accelerate first-order stochastic optimization algorithms such as SGD and SMD, various momentum techniques have been proposed (Polyak, 1964; Lan, 2012; Nesterov, 2013). In particular, Lan (2012) proposed an accelerated stochastic approximation (AC-SA) algorithm to accelerate SMD. For L-smooth and general convex function, AC-SA achieves the optimal convergence rate $O(L/k^2+\sigma/\sqrt{k})$ in terms of expected function value gap, where σ^2 is the variance of stochastic gradient. Compared with the non-accelerated SMD (1.3), the acceleration part comes from $O(L/k^2)$ and when the variance of the stochastic gradient vanishes, i.e., $\sigma = 0$, the convergence rate of AC-SA reduces to $O(L/k^2)$, which is the optimal convergence rate in the deterministic setting. Despite the optimal convergence rate, the update formulas in these algorithms are often elusive and lack of intuitive interpretations.

[*] Equal Contribution

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

On the other hand, recent years have witnessed the emergence of a line of research which attempts to interpret the stochastic mirror descent from the perspective of continuous-time dynamics. To the best of our knowledge, Raginsky and Bouvrie (2012) is the first work that interprets stochastic mirror descent as the discretization of the following Itô stochastic differential equation (SDE) (Øksendal, 2003)

$$\begin{cases} \mathrm{d}\boldsymbol{Y}_t = -\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sigma \mathrm{d}\boldsymbol{B}_t, \\ \boldsymbol{X}_t = \nabla h^*(\boldsymbol{Y}_t), \end{cases} \tag{1.4}$$

where $-\nabla f(\boldsymbol{X}_t)\mathrm{d}t$ is called the drift term, and σ is called the diffusion term, which corresponds to the variance of the stochastic gradient. Later Mertikopoulos and Staudigl (2016) extended the above first-order SDE in (1.4) by replacing the constant diffusion term σ with a general matrix $\sigma(\boldsymbol{X}_t, t)$ and proved almost-sure convergence of the function value $f(\boldsymbol{X}_t)$ along the solution trajectories of SDE to the minimal function value $f(\mathbf{x}^*)$. Very recently, Krichene and Bartlett (2017) proposed the following second-order SDE to interpret accelerated stochastic mirror descent

$$\begin{cases} \mathrm{d}\boldsymbol{Y}_t = -\eta_t\big(\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t\big), \\ \mathrm{d}\boldsymbol{X}_t = a_t(\nabla h^*(\boldsymbol{Y}_t/s_t) - \boldsymbol{X}_t)\mathrm{d}t, \end{cases} \tag{1.5}$$

where η_t, a_t and s_t are scaling functions of t. They proved $O(1/t^{1/2-q})$ convergence rate of expected function value along the solution trajectories of the SDE to the minimal function value, under the assumption that $\|\sigma(\mathbf{x}, t)\| \leq t^q$ ($q < 1/2$). However, there is still a gap between the continuous-time dynamics and discrete-time algorithms for accelerated stochastic mirror descent. In particular, it remains unclear whether the continuous-time SDE can be used to design and analyze new discrete-time algorithms.

In this paper, we aim to bridge this gap by proposing a new SDE-based interpretation for accelerated stochastic mirror descent, and derive new discrete-time algorithms from this SDE. We provide a unified analysis based on Lyapunov function, which connects both continuous-time dynamics and discrete-time algorithms derived thereof. We also use numerical experiments to backup our theory.

Our Contributions: We summarize the key contributions of our work as follows.

- We propose a new stochastic differential equation-based interpretation for accelerated stochastic mirror descent, motivated by Lagrangian mechanics. We take a Lyapunov function approach to prove that the convergence rate of the solution trajectories of the SDE matches the optimal rate of accelerated stochastic mirror descent for general convex function.
- We derive several new accelerated algorithms of SMD via discretizing the proposed SDE using various Euler discretization schemes, and provide an analogous Lyapunov function-based analysis for the new algorithms, which largely resembles the proof in the continuous-time dynamics. We show that these new algorithms also achieve the optimal convergence rate of accelerated SMD for general convex and smooth functions.

It is worth highlighting that under our framework, the discrete-time algorithms are closely connected with the continuous-time dynamics in the sense that they are nearly equivalent when the discretization step is sufficiently small.

The remainder of this paper is organized as follows: in Section 2 we review related work We introduce the preliminaries in Section 3, and present the new continuous-time dynamics for accelerated SMD in Section 4. We show how we can discretize the continuous-time dynamics to invent new discrete-time algorithms in Section 5. We provide numerical experiments in Section 6 to validate our theory and conclude the paper with Section 7.

Notation We use upper case letters $\boldsymbol{X}_t$ to denote continuous-time curve vector, where $t \geq 0$ is the time index. $\dot{\boldsymbol{X}}_t$ with an over-dot denotes the derivative of $\boldsymbol{X}_t$ with time t. Lower case letters $\mathbf{x}_k$ denote the trajectory of a discrete-time algorithm, where $k = 0, 1, \ldots$ is the index of iteration number. For all $\mathbf{x} \in \mathbb{R}^d$, we fix a general norm $\|\mathbf{x}\|$ and its dual norm is given by $\|\mathbf{x}\|_* = \sup_{\|\mathbf{y}\| \leq 1}\langle \mathbf{x}, \mathbf{y}\rangle$.

2 RELATED WORK

There is a vast literature of deterministic optimization methods for convex optimization problems. The most widely used first-order deterministic optimization algorithms include gradient descent (Polyak, 1963), mirror descent (Nemirovskii et al., 1983), Nesterov's accelerated gradient method (Nesterov, 1983, 2005, 2013) and accelerated mirror descent (Nemirovski et al., 2009). While Nesterov's accelerated gradient method attains the optimal convergence rate for first-order black box model, it falls short of an intuitive interpretation. Recently, there is a series of work, which interprets Nesterov's accelerated gradient methods from different perspectives, including ordinary differential equation (ODE) (Su et al., 2014), control theory (Lessard et al., 2016; Hu and Lessard, 2017), linear coupling (Allen-Zhu and Orecchia, 2014), geometric interpretation (Bubeck et al., 2015) and game theory (Lan and Zhou, 2015), to name a few. Along these studies, the ODE-based interpretation (Su et al., 2014) is perhaps the most simple and elegant approach, which models the continuous-time limit of the accelerated methods using ODE and analyzes the stability of the resulting ODE by constructing a Lyapunov function (Su et al., 2014). Moreover, it has been extended to analyze the accelerated mirror descent (Krichene et al., 2015; Wibisono et al., 2016; Wilson et al., 2016; Diakonikolas and Orecchia, 2017). In particular, Wibisono and Wilson (2015); Wibisono et al. (2016) derived a class of ODE-based continuous dynamics for a family of accelerated optimization methods from a novel Lagrangian functional called Bregman Lagrangian. On the other hand, the ODE-based continuous-time dynamics are also able to provide guidance on designing new algorithms by carefully discretizing the ODEs. For example, Krichene et al. (2015); Wilson et al. (2016) rediscovered several discrete-time algorithms and derived a few new algorithms via different Euler discretization schemes for the ODE. Wilson et al. (2016) also found that not all discretization schemes lead to practical algorithms and/or achieve accelerated rates.

Due to its success in large-scale machine learning, stochastic first-order optimization method has also drawn a lot of research interest. Instead of using the gradient, stochastic first-order optimization methods use the stochastic gradient. Representative stochastic first-order optimization methods include stochastic gradient descent (SGD) (Robbins and Monro, 1951), stochastic mirror descent (SMD) (Nemirovski et al., 2009), and their accelerated variants (Hu et al., 2009; Lan, 2012; Ghadimi and Lan, 2012; Chen et al., 2012). Analogous to the ODE-based interpretation of deterministic optimization, stochastic optimization has an SDE-based interpretation. More specifically, Raginsky and Bouvrie (2012) studied the continuous-time dynamics of stochastic mirror descent using Itô's stochastic differential equations, where they showed that the solutions of the stochastic dynamics do not converge to the global minimizer due to the adverse effects brought by Brownian motion. Mertikopoulos and Staudigl (2016) extended the SDE in Raginsky and Bouvrie (2012) for stochastic mirror descent and proved almost-sure convergence of the solution trajectories for SMD. However, continous-time dynamics for accelerated stochastic mirror descent remain under-studied. To the best of our knowledge, Krichene and Bartlett (2017) is the only existing work that proposes a second-order stochastic dynamics for accelerated stochastic mirror descent and proves convergence rates of the function values along solution trajectories of SDE both in terms of almost-sure convergence and in expectation. Unlike the ODE-based interpretation for deterministic optimization, based on which quite a few discrete-time algorithms are rediscovered or invented, it is unclear whether we can derive new accelerated SMD algorithms from its SDE-based interpretation. This motivates our work.

3 PRELIMINARIES

3.1 Bregman Divergence

Mirror descent is based on Bregman divergence (Bregman, 1967). In detail, if the domain $\mathcal{X}$ is endowed with a convex and differentiable function $h : \mathcal{X} \to \mathbb{R}$, the Bregman divergence (Bregman, 1967) is defined as

$$D_h(\mathbf{x}, \mathbf{x}') = h(\mathbf{x}) - h(\mathbf{x}') - \langle \nabla h(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle, \quad (3.1)$$

for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. We always have $D_h(\mathbf{x}, \mathbf{x}') \geq 0$ due to the convexity of h. Following the assumptions in Ghadimi and Lan (2012), we assume that $D_h(\cdot, \cdot)$ grows quadratically, without loss of generality, with parameter 1, i.e.,

$$D_h(\mathbf{x}, \mathbf{x}') \leq \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2, \text{for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \quad (3.2)$$

A plain example of Bregman divergence is the Euclidean distance: let $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$, then $D(\mathbf{x}, \mathbf{x}') = 1/2\|\mathbf{x} - \mathbf{x}'\|_2^2$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Another example is the Kullback-Leibler (KL) divergence: let $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$ be the negative entropy, then $D(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d x_i \log(x_i/x_i')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, which is the Kullback-Leibler divergence. Here $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}_+^d : \sum_{i=1}^d x_i = 1\}$ is the unit simplex in $\mathbb{R}^d$.

Recall the SMD update (1.3), the iteration is mapped back to the primal space $\mathcal{X}$ via the mirror mapping ∇h^* after one step of gradient descent in the dual space. The mirror mapping is the gradient of the convex conjugate function h^*, which is defined as

$$h^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{y}, \mathbf{x} \rangle - h(\mathbf{x}), \quad \text{for } \mathbf{y} \in E^*,$$

where E^* is the dual space of E and $\mathcal{X} \subset E$.

3.2 Assumptions and Propositions

To ease the presentation, we lay down assumptions and some propositions of Bregman divergence that will be used in our analysis.

Assumption 3.1. *f is continuously differentiable and the gradient ∇f is Lipschitz continuous with parameter L_f, that is, for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L_f\|\mathbf{x} - \mathbf{y}\|. \quad (3.3)$$

This is also called L_f-smoothness of f.

Assumption 3.2. *h is μ_h-strongly convex with some constant $\mu_h > 0$, that is, for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*

$$h(\mathbf{x}) \geq h(\mathbf{y}) + \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu_h}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Assumption 3.2 is commonly made in the analysis of mirror descent algorithms (Lan, 2012; Ghadimi and Lan, 2012) and its continuous-time dynamics (Wilson et al., 2016; Krichene and Bartlett, 2017)

Assumption 3.3. *$\mathcal{X}$ is convex and compact. There is a constant $M_{h,\mathcal{X}} > 0$ such that*

$$M_{h,\mathcal{X}} = \sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} D_h(\mathbf{x}, \mathbf{x}').$$

This is a natural assumption imposed in constrained optimization problems (Lan, 2012).

When the distance generating function h is strongly convex, the following two standard results on conjugate functions hold.

Proposition 3.4. *Suppose Assumption 3.2 holds for h. Then its conjugate function h^* is $1/\mu_h$-smooth and*

$$\nabla h^*(\mathbf{y}) = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{y}, \mathbf{x} \rangle - h(\mathbf{x}), \quad \text{for } \mathbf{y} \in E^*,$$

Proposition 3.5. *Suppose Assumption 3.2 holds for h. For all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, we have*

$$\nabla h^*(\nabla h(\mathbf{x}))) = \mathbf{x}, \quad D_h(\mathbf{x}, \mathbf{x}') = D_{h^*}(\nabla h(\mathbf{x}'), \nabla h(\mathbf{x})).$$

We refer interested readers to Banerjee et al. (2005) for detailed discussions of these properties.

4 THE PROPOSED CONTINUOUS-TIME DYNAMICS

In this section, we present a continuous-time dynamics for accelerated stochastic mirror descent, and analyze its convergence. We start with deriving the dynamics from Bregman Lagrangian (Wibisono et al., 2016).

4.1 Continuous-time Dynamics for Accelerated Stochastic Mirror Descent

We borrow ideas from Wibisono et al. (2016) by viewing the optimizing process as a physical process. For the mechanical system associated with optimization problem (1.1), we use $\boldsymbol{X}_t$ and $\dot{\boldsymbol{X}}_t$ to denote the position and velocity respectively. We define the Bregman Lagrangian (Wibisono and Wilson, 2015) as a weighted sum of kinetic Lyapunov $D_h(\boldsymbol{X}_t + e^{-\alpha_t}\dot{\boldsymbol{X}}_t, \boldsymbol{X}_t)$ and potential Lyapunov $f(\boldsymbol{X}_t)$ as follows

$$\begin{aligned} &\mathcal{L}(\boldsymbol{X}_t, \dot{\boldsymbol{X}}_t, t) \\ &= e^{\alpha_t+\gamma_t}\big(D_h(\boldsymbol{X}_t + e^{-\alpha_t}\dot{\boldsymbol{X}}_t, \boldsymbol{X}_t) - e^{\beta_t} f(\boldsymbol{X}_t)\big), \end{aligned} \tag{4.1}$$

where $\alpha_t, \beta_t, \gamma_t$ are arbitrary scaling functions that are continuously differentiable with respect to t. Consider an action functional $J(\boldsymbol{X}) = \int_{\mathbb{T}} \mathcal{L}(\boldsymbol{X}_t, \dot{\boldsymbol{X}}_t, t)\mathrm{d}t$ which is defined on curves $\{\boldsymbol{X}_t\}_{t \in \mathbb{R}_+}$. By Hamilton's principle (or principle of least action), minimizing the action functional $J(\boldsymbol{X})$ requires that the curve $\boldsymbol{X}_t$ satisfies the following Euler-Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{X}}_t}(\boldsymbol{X}_t, \dot{\boldsymbol{X}}_t, t)\right\} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{X}_t}(\boldsymbol{X}_t, \dot{\boldsymbol{X}}_t, t). \tag{4.2}$$

To simplify the notation, we adopt the following ideal scaling conditions suggested by Wibisono et al. (2016), which are required for the stability of ODE (4.2):

$$\dot{\beta}_t = e^{\alpha_t}, \quad \dot{\gamma}_t = e^{\alpha_t}. \tag{4.3}$$

Submitting the Bregman Lagrangian $\mathcal{L}$ in (4.1) into the Euler-Lagrange equation (4.2), we obtain the following condition for $\boldsymbol{X}_t$ that minimizes $J(\boldsymbol{X})$

$$\mathrm{d}\nabla h\big(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t\big) = -\dot{\beta}_t e^{\beta_t}\nabla f(\boldsymbol{X}_t)\mathrm{d}t. \tag{4.4}$$

It is worth noting that (4.4) has been used to describe the continuous-time dynamics of many optimization problems in Wibisono et al. (2016). However, in stochastic optimization, one uses the stochastic gradient rather than gradient. To account for this, we add some random noise to the gradient of f. More specifically, to adapt the dynamics (4.4) to stochastic optimization, we add a Brownian motion term after the gradient $\nabla f(\boldsymbol{X}_t)$ to form the following Itô stochastic differential equation (SDE) (Øksendal, 2003)

$$\begin{aligned} &\mathrm{d}\nabla h\big(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t\big) \\ &\quad = -\dot{\beta}_t e^{\beta_t}\big(\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{\delta}\sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t\big), \end{aligned} \tag{4.5}$$

where $\boldsymbol{B}_t \in \mathbb{R}^d$ is the standard Brownian motion, $\sigma(\boldsymbol{X}_t, t) \in \mathbb{R}^{d\times d}$ and $\delta > 0$ is a constant. The term $\sqrt{\delta}\sigma(\boldsymbol{X}_t, t)$ is called the diffusion coefficient that accounts for the variance of the stochastic gradient.

Note that (4.5) is a second-order SDE. We define a variable $\boldsymbol{Y}_t$ in the dual space E^* as

$$\boldsymbol{Y}_t = \nabla h\big(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t\big). \tag{4.6}$$

Based on (4.5) and (4.6), and using Proposition 3.5, we obtain the following continuous-time dynamics for accelerated stochastic mirror descent

$$\begin{cases} \mathrm{d}\boldsymbol{X}_t = \dot{\beta}_t(\nabla h^*(\boldsymbol{Y}_t) - \boldsymbol{X}_t)\mathrm{d}t, & (4.7\text{a}) \\ \mathrm{d}\boldsymbol{Y}_t = -\dot{\beta}_t e^{\beta_t}\big(\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{\delta}\sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t\big). & (4.7\text{b}) \end{cases}$$

The solution trajectories of (4.7) do not converge to the minimizer of f due to the stochastic noise. So we modify it by introducing a shrinkage parameter $s_t > 0$ to decrease the adverse effect of the stochastic noise.

$$\begin{cases} \mathrm{d}\boldsymbol{X}_t = \dot{\beta}_t(\nabla h^*(\boldsymbol{Y}_t) - \boldsymbol{X}_t)\mathrm{d}t, & (4.8\text{a}) \\ \mathrm{d}\boldsymbol{Y}_t = -\dfrac{\dot{\beta}_t e^{\beta_t}}{s_t}\big(\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{\delta}\sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t\big), & (4.8\text{b}) \end{cases}$$

where $\boldsymbol{B}_t \in \mathbb{R}^d$ is the standard Brownian motion, $\beta_t, s_t > 0$ are scaling functions of time t and $\delta > 0$ is a constant. The idea of introducing shrinkage parameter s_t to offset the disadvantage brought by stochastic gradient is similar to that in Mertikopoulos and Staudigl (2016); Krichene and Bartlett (2017). This amount to using time-decaying step size in the discrete-time stochastic mirror descent algorithm to ensure its convergence.

4.2 Convergence Rate of the Continuous-time Dynamics

In this subsection, we analyze the convergence of our proposed continuous-time dynamics (4.8) for accelerated stochastic mirror descent. The following theorem spells out its convergence rate.

Theorem 4.1. Suppose f is convex and Assumptions 3.2 and 3.3 hold. If the diffusion coefficient in (4.8) satisfies $\|\sigma(\boldsymbol{X}_t,t)^\top\sigma(\boldsymbol{X}_t,t)\| \le t^{2q}$ for some $q < 1/2$, then it holds that

$$\begin{aligned}&\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)]\\ &\quad\le \frac{\mathcal{E}_0 + (s_t - s_0)M_{h,\mathcal{X}}}{e^{\beta_t}}\\ &\qquad + \frac{\mathbb{E}\big[\int_0^t \frac{\delta\dot{\beta}_r^2 e^{2\beta_r}}{s_r}\,\mathrm{tr}\left(\sigma_r^\top\nabla^2 h^*(\boldsymbol{Y}_r)\sigma_r\right)\mathrm{d}r\big]}{2e^{\beta_t}},\end{aligned} \tag{4.9}$$

where $\sigma_r = \sigma(\boldsymbol{X}_r, r)$ and $\mathcal{E}_0 = e^{\beta_0}(f(\boldsymbol{X}_0) - f(\mathbf{x}^*)) + s_t D_{h^*}(\boldsymbol{Y}_0, \nabla h(\mathbf{x}^*))$.

Note that the convergence of the continuous-time dynamics (4.8) does not require the smoothness of f.

Remark 4.2. If we choose $\beta_t = p\log t$ for some constant $p > 0$ and $s_t = t^\alpha$ for some $\alpha \in \mathbb{R}$ and note that $\|\sigma(\boldsymbol{X}_t,t)^\top\sigma(\boldsymbol{X}_t,t)\| \le t^{2q}$, we have the following convergence rate

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] = O\bigg(\frac{1 + t^\alpha + t^{2p-\alpha+2q-1}}{t^p}\bigg).$$

If we further choose $\alpha = p + q - 1/2$ and $p = 2$, then we obtain

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] = O\bigg(\frac{1}{t^2} + \frac{1}{t^{1/2-q}}\bigg),$$

which matches the optimal convergence rate for accelerated stochastic mirror descent algorithms (Lan, 2012), where the variance of stochastic gradient is bounded by a constant i.e., $q = 0$. In contrast, the convergence rate of the continuous dynamics (1.5) proposed by Krichene and Bartlett (2017) for accelerated stochastic mirror descent is $O(1/t^{1/2})$, which only matches the optimal rate up to the dominating term.

Remark 4.3. When the variance of stochastic gradient vanishes, i.e., $\sigma(\boldsymbol{X}, t) = 0$, and if we choose the shrinkage parameter $s_t = 1$ for all $t \in \mathbb{R}_+$, the convergence rate in (4.9) becomes $O(1/t^2)$, which matches the optimal convergence rate of deterministic accelerated mirror descent algorithms for general convex functions (Nesterov, 1983).

Here we provide a proof of Theorem 4.1 via constructing the following Lyapunov function for the stochastic dynamics system (4.8)

$$\mathcal{E}_t = e^{\beta_t}(f(\boldsymbol{X}_t) - f(\mathbf{x}^*)) + s_t D_{h^*}(\boldsymbol{Y}_t, \nabla h(\mathbf{x}^*)). \tag{4.10}$$

Proof of Theorem 4.1. Denote $\sigma_t = \sigma(\boldsymbol{X}_t, t)$, and applying Itô's Lemma to the Lyapunov function yields

$$\begin{aligned}\mathrm{d}\mathcal{E}_t = {}& \frac{\partial\mathcal{E}_t}{\partial t}\mathrm{d}t + \Big\langle\frac{\partial\mathcal{E}_t}{\partial\boldsymbol{X}_t}, \mathrm{d}\boldsymbol{X}_t\Big\rangle + \Big\langle\frac{\partial\mathcal{E}_t}{\partial\boldsymbol{Y}_t}, \mathrm{d}\boldsymbol{Y}_t\Big\rangle\\ &+ \frac{\delta\dot{\beta}_t^2 e^{2\beta_t}}{2s_t^2}\,\mathrm{tr}\left(\sigma_t^\top\frac{\partial^2\mathcal{E}_t}{\partial\boldsymbol{Y}_t^2}\sigma_t\right)\mathrm{d}t.\end{aligned} \tag{4.11}$$

By simple calculus, we have

$$\begin{aligned}&\frac{\partial\mathcal{E}_t}{\partial t} = \dot{\beta}_t e^{\beta_t}(f(\boldsymbol{X}_t) - f(\mathbf{x}^*)) + \dot{s}_t D_{h^*}(\boldsymbol{Y}_t, \nabla h(\mathbf{x}^*)),\\ &\frac{\partial\mathcal{E}_t}{\partial\boldsymbol{X}_t} = e^{\beta_t}\nabla f(\boldsymbol{X}_t), \quad \frac{\partial\mathcal{E}_t}{\partial\boldsymbol{Y}_t} = s_t(\nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*).\end{aligned}$$

Submitting the above calculations and dynamics (4.8) into (4.11) yields

$$\begin{aligned}\mathrm{d}\mathcal{E}_t &= \big[\dot{\beta}_t e^{\beta_t}(f(\boldsymbol{X}_t) - f(\mathbf{x}^*)) + \dot{s}_t D_{h^*}(\boldsymbol{Y}_t, \nabla h(\mathbf{x}^*))\big]\mathrm{d}t\\ &\quad + \dot{\beta}_t e^{\beta_t}\langle\nabla h^*(\boldsymbol{Y}_t) - \boldsymbol{X}_t, \nabla f(\boldsymbol{X}_t)\rangle\mathrm{d}t\\ &\quad - \dot{\beta}_t e^{\beta_t}\langle\nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sigma_t\mathrm{d}\boldsymbol{B}_t\rangle\\ &\quad + \frac{1}{2}\frac{\delta\dot{\beta}_t^2 e^{2\beta_t}}{s_t^2}\,\mathrm{tr}\left(\sigma_t^\top\frac{\partial^2\mathcal{E}_t}{\partial\boldsymbol{Y}_t^2}\sigma_t\right)\mathrm{d}t\\ &= \dot{\beta}_t e^{\beta_t}\big[f(\boldsymbol{X}_t) - f(\mathbf{x}^*) + \langle\mathbf{x}^* - \boldsymbol{X}_t, \nabla f(\boldsymbol{X}_t)\rangle\big]\mathrm{d}t\\ &\quad + \dot{s}_t D_{h^*}(\boldsymbol{Y}_t, \nabla h(\mathbf{x}^*))\mathrm{d}t\\ &\quad - \dot{\beta}_t e^{\beta_t}\langle\nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \sigma_t\mathrm{d}\boldsymbol{B}_t\rangle\\ &\quad + \mathbb{E}\bigg[\frac{\delta\dot{\beta}_t^2 e^{2\beta_t}}{2s_t}\,\mathrm{tr}\left(\sigma_t^\top\nabla^2 h^*(\boldsymbol{Y}_t)\sigma_t\right)\bigg]\mathrm{d}t\\ &\le \dot{s}_t M_{h,\mathcal{X}} + \frac{1}{2s_t}\delta\dot{\beta}_t^2 e^{2\beta_t}\,\mathrm{tr}\left(\sigma_t^\top\nabla^2 h^*(\boldsymbol{Y}_t)\sigma_t\right)\mathrm{d}t\\ &\quad - \dot{\beta}_t e^{\beta_t}\langle\nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \sigma_t\mathrm{d}\boldsymbol{B}_t\rangle,\end{aligned}$$

where the inequality follows from the convexity of f and Assumption 3.3 that $D_{h^*}(\boldsymbol{Y}_t, \nabla h(\mathbf{x}^*)) \le M_{h,\mathcal{X}}$. Integrating from 0 to t, we obtain

$$\begin{aligned}\mathcal{E}_t \le {}& \mathcal{E}_0 + (s_t - s_0)M_{h,\mathcal{X}}\\ &+ \frac{1}{2}\int_0^t \frac{\delta\dot{\beta}_r^2 e^{2\beta_r}}{s_r}\,\mathrm{tr}\left(\sigma_r^\top\nabla^2 h^*(\boldsymbol{Y}_r)\sigma_r\right)\mathrm{d}r\\ &- \int_0^t \dot{\beta}_r e^{\beta_r}\langle\nabla h^*(\boldsymbol{Y}_r) - \mathbf{x}^*, \sigma_r\mathrm{d}\boldsymbol{B}_r\rangle.\end{aligned}$$

By the definition of $\mathcal{E}_t$ and the fact that $D_h^* \geq 0$, we immediately get

$$\begin{aligned} & e^{\beta_t}(f(\boldsymbol{X}_t) - f(\mathbf{x}^*)) \\ & \leq \mathcal{E}_0 + (s_t - s_0)M_{h,\mathcal{X}} \\ & \quad + \int_0^t \frac{\delta \dot{\beta}_r^2 e^{2\beta_r}}{2s_r} \operatorname{tr}\left(\sigma_r^\top \nabla^2 h^*(\boldsymbol{Y}_r)\sigma_r\right)\mathrm{d}r \\ & \quad - \int_0^t \dot{\beta}_r e^{\beta_r} \langle \nabla h^*(\boldsymbol{Y}_r) - \mathbf{x}^*, \sigma_r \mathrm{d}\boldsymbol{B}_r \rangle. \end{aligned}$$

Taking expectation of both sides and using the martingale property of Itô integral, we obtain

$$\begin{aligned} & \mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] \\ & \leq \frac{\mathcal{E}_0 + (s_t - s_0)M_{h,\mathcal{X}}}{e^{\beta_t}} \\ & \quad + \frac{\mathbb{E}\left[\int_0^t \frac{\delta\dot{\beta}_r^2 e^{2\beta_r}}{s_r} \operatorname{tr}\left(\sigma_r^\top \nabla^2 h^*(\boldsymbol{Y}_r)\sigma_r\right)\mathrm{d}r\right]}{2e^{\beta_t}}. \end{aligned}$$

This completes the proof. □

5 THE PROPOSED DISCRETE-TIME ALGORITHMS

Our primary goal in proposing the continuous-time dynamics for accelerated stochastic mirror descent is to design new discrete-time algorithms. In this section, we provide several discretizations of (4.8). We will adapt the Lyapunov function-based analysis in previous section to the discrete-time algorithms and deliver very simple and intuitive proofs for the new algorithms. The high-level idea of this section is inspired by Wilson et al. (2016), which discussed various discretization schemes for the continuous-time dynamics of deterministic optimization.

We use Euler discretization schemes (Kloeden and Platen, 1992) of differential equations and its composition to derive several discrete algorithms of dynamic (4.8). Specifically, note that (4.8) is a system of two first-order differential equations, and thus we can compose explicit and implicit Euler discretizations in four different ways. Let δ be the time step and

$$\mathbf{x}_k = \boldsymbol{X}_t, \quad \mathbf{x}_{k+1} = \boldsymbol{X}_{t+\delta}, \quad \mathbf{y}_k = \boldsymbol{Y}_t, \quad \mathbf{y}_{k+1} = \boldsymbol{Y}_{t+\delta}.$$

The following approximations

$$(\mathbf{x}_{k+1} - \mathbf{x}_k)/\delta \approx \dot{\boldsymbol{X}}_t, \quad (\mathbf{y}_{k+1} - \mathbf{y}_k)/\delta \approx \dot{\boldsymbol{Y}}_t$$

are the explicit (forward) Euler discretization for the time derivatives of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, and

$$(\mathbf{x}_k - \mathbf{x}_{k-1})/\delta \approx \dot{\boldsymbol{X}}_t, \quad (\mathbf{y}_k - \mathbf{y}_{k-1})/\delta \approx \dot{\boldsymbol{Y}}_t$$

are the implicit (backward) Euler discretization for the time derivatives of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. For the scaling parameters, we choose $A_k = e^{\beta_t}$, $s_k = s_t$, and the discretizations as follows

$$(A_{k+1} - A_k)/\delta \approx \mathrm{d}e^{\beta_t}/\mathrm{d}t, \quad (A_{k+1} - A_k)/(\delta A_k) \approx \dot{\beta}_t.$$

5.1 Implicit Euler Discretization

Algorithm 1 Accelerated Stochastic Mirror Descent (Implicit)

```
Input: $A_0 = 1/2$, $\mathbf{x}_0 = \mathbf{y}_0 = \mathbf{0}$.
for $k = 0$ to $K$ do
    $A_{k+1} = (k+1)(k+2)/2$, $\tau_k = (A_{k+1} - A_k)/A_k$, $s_k = \sigma k^{3/2} + 1$
    $\nabla h^*(\mathbf{y}_{k+1}) = \mathbf{x}_{k+1} + \frac{1}{\tau_k}(\mathbf{x}_{k+1} - \mathbf{x}_k)$.
    $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \Big\{\widetilde{f}(\mathbf{x}, \xi_{k+1}) + \frac{s_k}{A_{k+1}} D_h(\mathbf{u}, \nabla h^*(\mathbf{y}_k))\Big\}$, and $\mathbf{u} = \mathbf{x} + \frac{1}{\tau_k}(\mathbf{x} - \mathbf{x}_k)$.
end for
```

The first discrete-time algorithm we derive is from implicit Euler discretization of dynamics (4.8). The discretization process is displayed in Algorithm 1. Here we use Borel function $\widetilde{f}(\mathbf{x}_k, \xi_k)$ to denote the noisy objective function, where $\{\xi_k\}_{k=0,1,\dots}$ is a sequence of independent random variables. Denote $\Delta(\mathbf{x}_k) = \nabla\widetilde{f}(\mathbf{x}_k, \xi_k) - \nabla f(\mathbf{x}_k)$ and we assume that

$$\mathbb{E}[\widetilde{f}(\mathbf{x}_k, \xi_k)|\mathbf{x}_k] = f(\mathbf{x}_k), \quad \mathbb{E}[\nabla\widetilde{f}(\mathbf{x}_k, \xi_k)|\mathbf{x}_k] = \nabla f(\mathbf{x}_k),$$
$$\mathbb{E}[\|\Delta(\mathbf{x}_k)\|_*^2|\mathbf{x}_k] \leq \sigma^2.$$

The optimality condition of Algorithm 1 is given by

$$\begin{cases} \nabla h^*(\mathbf{y}_{k+1}) = \mathbf{x}_{k+1} + \dfrac{A_k}{A_{k+1} - A_k}(\mathbf{x}_{k+1} - \mathbf{x}_k), & (5.1\text{a}) \\ \mathbf{y}_{k+1} - \mathbf{y}_k = -\dfrac{A_{k+1} - A_k}{s_k}\nabla\widetilde{f}(\mathbf{x}_{k+1}, \xi_{k+1}), & (5.1\text{b}) \end{cases}$$

where $\nabla\widetilde{f}(\mathbf{x}_{k+1}, \xi_{k+1})$ is the stochastic gradient.

Inspired by the Lyapunov function (4.10) for the continuous-time dynamics, we construct the following Lyapunov function to analyze the convergence of Algorithm 1.

$$\mathcal{E}_k = \mathbb{E}[A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + s_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))]. \tag{5.2}$$

Based on the above Lyapunov function, we can prove the following theorem, which states the convergence rate of Algorithm 1.

Theorem 5.1. Suppose f is convex. Under Assumptions 3.2 and 3.3, if we choose $A_k = k(k+1)/2$, $A_0 = 1/2$ and $s_k = \sigma k^{3/2} + 1$, the expected function value gap at $\mathbf{x}_k$ output by Algorithm 1 is bounded as

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq & \frac{2(\mathcal{E}_0 + M_{h,\mathcal{X}})}{k(k+1)} \\ & + \sigma\bigg(M_{h,\mathcal{X}} + \frac{1}{3\mu_h}\bigg)\frac{\sqrt{k+1}}{k}, \end{aligned}$$

where $\mathcal{E}_0 = A_0(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + s_0 D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_0))$.

Remark 5.2. Theorem 5.1 suggests that the convergence rate of Algorithm 1 is $O((1+M_{h,\mathcal{X}})/k^2 + M_{h,\mathcal{X}}\sigma/\sqrt{k})$, which matches the optimal rate of accelerated stochastic optimization for general convex and smooth functions (Lan, 2012). Note that when variance of the stochastic gradient vanishes, we can choose $s_k = 1$ for all k and obtain the optimal convergence rate of accelerated gradient descent for general convex and smooth functions $O(1/k^2)$.

Remark 5.3. The optimal convergence rate of Algorithm 1 does not require the smooth assumption on f, which is aligned with the analysis of continuous-time dynamics (4.8). In fact, Algorithm 1 can be seen as an accelerated proximal point algorithm whose optimal convergence rate is also $O(1/k^2)$ (Güler, 1992).

5.2 Hybrid Euler Discretization

Although Algorithm 1 is succinct and attains the optimal convergence rate, the implicit Euler discretizations of both stochastic processes make it hard to implement in practice due to the requirement of an exact minimization in each iteration (Step 5 in Algorithm 1). Nevertheless, the implicit discretization sheds great light on the connection between continuous-time dynamics and discrete-time algorithms. Now we derive a more practical algorithm combining implicit and explicit Euler discretizations of dynamics (4.8). The discretization process is displayed in Algorithm 2. The optimality condition of Algorithm 2 is given by

Algorithm 2 Accelerated Stochastic Mirror Descent (ASMD)

```
Input: $A_0 = s_0 = 1/2, \mathbf{x}_0 = \mathbf{y}_0 = \mathbf{0}$.
for $k = 0$ to $K$ do
    $A_{k+1} = (k+1)(k+2)/2$, $\tau_k = (A_{k+1} - A_k)/A_k$, $s_{k+1} = (k+1)^{3/2}$.
    $\mathbf{x}_{k+1} = \frac{\tau_k}{\tau_k+1}\nabla h^*(\mathbf{y}_k) + \frac{1}{\tau_k+1}\mathbf{x}_k$.
    $\mathbf{y}_{k+1} = \mathbf{y}_k - \frac{A_{k+1}-A_k}{s_k}\nabla\widetilde{f}(\mathbf{x}_{k+1}, \xi_{k+1})$.
end for
```

$$
\begin{cases}
\nabla h^*(\mathbf{y}_k) = \mathbf{x}_{k+1} + \dfrac{A_k}{A_{k+1}-A_k}(\mathbf{x}_{k+1} - \mathbf{x}_k), & \text{(5.3a)}\\[2ex]
\mathbf{y}_{k+1} - \mathbf{y}_k = -\dfrac{A_{k+1}-A_k}{s_k}\nabla\widetilde{f}(\mathbf{x}_{k+1}, \xi_{k+1}). & \text{(5.3b)}
\end{cases}
$$

Based on the same Lyapunov function (5.2) of Algorithm 1, we can prove the convergence rate of Algorithm 2.

Theorem 5.4. Suppose f is convex. Under Assumptions 3.1, 3.2 and 3.3, if we choose $A_k = k(k+1)/2$, $s_k = k^{3/2}$ and $A_0 = s_0 = 1/2$, the expected function value gap at $\mathbf{x}_k$ output by Algorithm 2 is bounded as

$$
\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \frac{2\mathcal{E}_0}{k(k+1)} + \frac{C_0\sqrt{k}}{\mu_h^2(k+1)},
$$

where $C_0 = 2((4L_f^2 + \mu_h^2)\mathcal{M}_{h,\mathcal{X}} + \mu_h\sigma^2 + 2\|\nabla f(\mathbf{x}^*)\|_*^2)$ and $\mathcal{E}_0 = A_0(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + s_0 D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_0))$.

Remark 5.5. Compared with Algorithm 1, Algorithm 2 is a much more practical algorithm and can be easily implemented. The convergence rate of Algorithm 2 is $O(M_{h,\mathcal{X}}/k^2 + (L_f^2+\sigma^2)/\sqrt{k})$. Although this matches the optimal rate of accelerated stochastic mirror descent (Lan, 2012), yet when the variance of the stochastic gradient vanishes, i.e., $\sigma = 0$, it is not reducible to the optimal convergence rate $O(1/k^2)$ for deterministic optimization.

5.3 Discretization with an Additional Sequence

Algorithm 3 Accelerated Stochastic Mirror Descent (ASMD3)

```
Input: $A_0 = 1/2$, $\mathbf{x}_0 = \mathbf{y}_0 = \mathbf{z}_0 = \mathbf{0}$.
for $k = 0$ to $K$ do
    $A_{k+1} = \mu_h^2(k+1)(k+2)/(4L_f)$, $s_k = \sigma/L_f(k+1)^{3/2} + 1$, $M_k = L_f(A_{k+1} - A_k)^2/(\mu_h^2 s_k A_{k+1})$.
    $\mathbf{z}_{k+1} = \frac{A_{k+1}-A_k}{A_{k+1}}\nabla h^*(\mathbf{y}_k) + \frac{A_k}{A_{k+1}}\mathbf{x}_k$.
    $\mathbf{y}_{k+1} = \mathbf{y}_k - \frac{A_{k+1}-A_k}{s_k}\nabla\widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1})$.
    $\mathbf{x}_{k+1} = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\Big\{\langle\nabla\widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \mathbf{x}\rangle + \frac{L_f}{M_k}D_h(\mathbf{z}_{k+1}, \mathbf{x})\Big\}$.
end for
```

As we showed in previous subsection, Algorithm 2 is not able to obtain the optimal rate for deterministic optimization when using the full gradient instead of stochastic gradient. In order to address this issue, we will modify the algorithm by adding an additional sequence in a similar way to linear coupling (Allen-Zhu and Orecchia, 2014). The new algorithm with three sequences is presented in Algorithm 3. Note that the additionally added sequence can be seen as a step of mirror descent. The optimality condition of Algorithm 3 is given by

$$
\begin{cases}
\mathbf{z}_{k+1} = \dfrac{A_{k+1}-A_k}{A_{k+1}}\nabla h^*(\mathbf{y}_k) + \dfrac{A_k}{A_{k+1}}\mathbf{x}_k, & \text{(5.4a)}\\[2ex]
\mathbf{y}_{k+1} - \mathbf{y}_k = -\dfrac{A_{k+1}-A_k}{s_k}\nabla\widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), & \text{(5.4b)}\\[2ex]
\mathbf{x}_{k+1} = \underset{\mathbf{x}\in\mathcal{X}}{\mathrm{argmin}}\Big\{\langle\nabla\widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \mathbf{x}\rangle & \text{(5.4c)}\\[1ex]
\qquad\qquad + \dfrac{L_f}{M_k}D_h(\mathbf{z}_{k+1}, \mathbf{x})\Big\}. &
\end{cases}
$$

In order to prove the convergence of Algorithm 3 with the additional sequence, we define a new Lyapunov function as $\mathcal{E}_k = \mathbb{E}[A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + s_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))]$, which is slightly different from that of previous algorithms. Then we can show the following result.

Theorem 5.6. Suppose f is convex. Under Assumptions 3.1, 3.2 and 3.3, if we choose $A_k = \mu_h^2 k(k+$

$1)/(4L_f)$, $s_k = \sigma/L_f(k+1)^{3/2} + 1$ and $M_k = L_f(A_{k+1} - A_k)^2/(\mu_h^2 s_k A_{k+1})$, the expected function value gap at $\mathbf{x}_k$ output by Algorithm 3 is bounded as

$$\begin{aligned}&\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)]\\ &\quad\le \frac{4L_f(\mathcal{E}_0 + M_{h,\mathcal{X}})}{\mu_h^2 k(k+1)} + \frac{\sigma(\mu_h^2 + 12M_{h,\mathcal{X}})\sqrt{k+1}}{3\mu_h^2 k},\end{aligned}$$

where $\mathcal{E}_0 = A_0(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + s_0 D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_0))$.

Remark 5.7. The convergence rate of Algorithm 3 is in the order of $O(L_f(1 + M_{h,\mathcal{X}})/k^2 + \sigma(1 + M_{h,\mathcal{X}})/\sqrt{k})$, which matches the optimal rate of accelerated stochastic mirror descent for general convex and smooth functions (Lan, 2012). More importantly, when the stochastic gradient reduces to the full gradient, namely, $\sigma = 0$, the $\sigma(1 + M_{h,\mathcal{X}})/\sqrt{k}$ term diminishes and the optimal convergence rate for stochastic optimization reduces to the optimal convergence rate $O(1/k^2)$ for deterministic optimization of general convex and smooth functions.

6 EXPERIMENTS

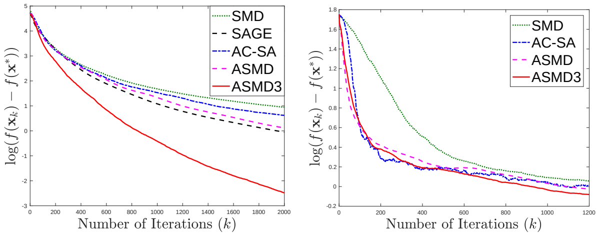
In this section, we conduct numerical experiments to verify the convergence rate of the proposed algorithms derived from the continuous-time dynamics. We compare our Algorithm 2 (**ASMD**) and Algorithm 3 (**ASMD3**) with stochastic mirror descent (**SMD**), accelerated stochastic approximation (**AC-SA**) (Lan, 2012) and stochastic accelerated gradient (**SAGE**) (Hu et al., 2009).

We apply the above optimization algorithms to the linear regression problem on a convex compact subset of $\mathbb{R}^d$.

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \textstyle\sum_{i=1}^n (\mathbf{A}_{i*}\mathbf{x} - y_i)^2, \tag{6.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\mathbf{A}_{i*}$ denotes the i-th row of $\mathbf{A} \in \mathbb{R}^{n\times d}$. We set $n = 100$, $d = 200$, and generated the design matrix $\mathbf{A}$ with entries following $N(0, 1)$. The response vector was generated by $\mathbf{y} = \mathbf{A}\mathbf{u}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_{n\times n})$ and $\mathbf{u}^* \sim N(\mathbf{0}, \mathbf{I}_{d\times d})$ were randomly generated. It is easy to verify that the objective function f is convex and L-smooth with $L = \|\mathbf{A}^\top\mathbf{A}\|_2$. We considered two settings of distance generating function h and the constrained set $\mathcal{X}$:

- Setting (1): $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$ is the squared Euclidean norm, and $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \le 2\|\mathbf{u}^*\|_2\}$. Then the mirror mapping $\nabla h^*(\mathbf{x}) = \mathrm{argmin}_{\mathbf{u}\in\mathcal{X}} \|\mathbf{x} - \mathbf{u}\|_2^2$ reduces to projection onto $\mathcal{X}$ and has a closed form solution.
- Setting (2): $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$ is the negative entropy and $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}_+^d : \sum_{i=1}^d x_i = 1\}$ is a simplex. In this case the mirror mapping also has a closed-form solution $[\nabla h^*(\mathbf{x})]_i = e^{x_i}/\sum_{i=1}^d e^{x_i}$ for all $i = 1, \ldots, d$ (Banerjee et al., 2005).



(a) Setting (1): $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$, and $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le 2\|\mathbf{u}^*\|_2\}$

(b) Setting (2): $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$, and $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}_+^d : \sum_{i=1}^d x_i = 1\}$

Figure 1: Logarithmic averaged function value gap over 50 repetitions for all methods under different settings.

Since (6.1) has a finite-sum structure, we simply choose the stochastic gradient by uniformly sampling i from $\{1, 2, \ldots, n\}$ with mini batch size 1. We plotted the function value gap $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ for each algorithm in Figure 1, where k is the number of iterations and $\mathbf{x}^*$ is the global minimizer of (6.1) solved by CVX program (Grant and Boyd, 2008). Figure 1(a) and Figure 1(b) show the averaged results over 50 repetitions for Setting (1) and Setting (2) respectively. Note that **SAGE** is not applicable to mirror descent. Experimental results in both settings demonstrate that our algorithms **ASMD** and **ASMD3** achieve comparable convergence rate as AC-SA, and converge much faster than SMD. This is well-aligned with our theory.

7 CONCLUSIONS

In this paper, we bridge the gap between continuous-time dynamics and discrete-time algorithms for accelerated stochastic mirror descent by presenting a variational analysis based on Bregman Lagrangian and Lyapunov functions. We not only propose a new continuous-time dynamics of accelerate stochastic mirror descent, but also derive several new algorithms from the continuous dynamics. Both the continuous-time dynamics and the discrete-time algorithms achieve the optimal convergence rate for general convex and smooth functions in the stochastic optimization setting.

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments. This research was sponsored in part by the National Science Foundation IIS-1618948 and IIS-1652539. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Allen-Zhu, Z. and Orecchia, L. (2014). Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537* .

Banerjee, A., Merugu, S., Dhillon, I. S. and Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research* **6** 1705–1749.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* **7** 200–217.

Bubeck, S., Lee, Y. T. and Singh, M. (2015). A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187* .

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

Chen, X., Lin, Q. and Pena, J. (2012). Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems*.

Chlebus, E. (2009). An approximate formula for a partial sum of the divergent p-series. *Applied Mathematics Letters* **22** 732–737.

Diakonikolas, J. and Orecchia, L. (2017). Accelerated extra-gradient descent: A novel accelerated first-order method. *arXiv preprint arXiv:1706.04680* .

Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization* **22** 1469–1492.

Grant, M. and Boyd, S. (2008). Cvx: Matlab software for disciplined convex programming.

Güler, O. (1992). New proximal point algorithms for convex minimization. *SIAM Journal on Optimization* **2** 649–664.

Hu, B. and Lessard, L. (2017). Dissipativity theory for nesterov's accelerated method. *arXiv preprint arXiv:1706.04381* .

Hu, C., Pan, W. and Kwok, J. T. (2009). Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*.

Kloeden, P. E. and Platen, E. (1992). Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics* **66** 283–314.

Krichene, W. and Bartlett, P. L. (2017). Acceleration and averaging in stochastic mirror descent dynamics. *arXiv preprint arXiv:1707.06219* .

Krichene, W., Bayen, A. and Bartlett, P. L. (2015). Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*.

Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming* **133** 365–397.

Lan, G. and Zhou, Y. (2015). An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000* .

Lessard, L., Recht, B. and Packard, A. (2016). Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization* **26** 57–95.

Luenberger, D. G., Ye, Y. et al. (1984). *Linear and nonlinear programming*, vol. 2. Springer.

Mertikopoulos, P. and Staudigl, M. (2016). On the convergence of gradient-like flows with noisy gradient input. *arXiv preprint arXiv:1611.06730* .

Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* **19** 1574–1609.

Nemirovskii, A., Yudin, D. B. and Dawson, E. R. (1983). Problem complexity and method efficiency in optimization .

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, vol. 27.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming* **103** 127–152.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media.

Øksendal, B. (2003). Stochastic differential equations. In *Stochastic differential equations*. Springer, 65–84.

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **3** 643–653.

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

Raginsky, M. and Bouvrie, J. (2012). Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *Decision* *and Control (CDC), 2012 IEEE 51st Annual Conference on*. IEEE.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Su, W., Boyd, S. and Candes, E. (2014). A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*.

Wibisono, A. and Wilson, A. C. (2015). On accelerated methods in optimization. *arXiv preprint arXiv:1509.03616* .

Wibisono, A., Wilson, A. C. and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences* 201614734.

Wilson, A. C., Recht, B. and Jordan, M. I. (2016). A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635* .

A Proof of Main Results for Discrete-time Algorithms

In this section, we provide the proofs for convergence rates of discrete-time algorithms derived from the stochastic dynamics (4.8) of mirror descent.

A.1 Proof of Convergence of Algorithm 1

Proof of Theorem 5.1. Let $\mathcal{E}_k = \mathbb{E}[A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + s_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)))]$. We have

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))] \\
&= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[s_k(-D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \langle \mathbf{y}_k - \mathbf{y}_{k+1}, \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1})\rangle)] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)(\langle \nabla \widetilde{f}(\mathbf{x}_{k+1}, \xi_{k+1}), \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1})\rangle) - s_k D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k))],
\end{aligned}$$

where the last inequality follows from (5.1). We further plug in the update of $\nabla h^*(\mathbf{y}_{k+1})$ in (5.1a) and obtain

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &\le \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)\langle \nabla f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1}\rangle - A_k\langle \nabla f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)\langle \Delta(\mathbf{x}_{k+1}), \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1})\rangle - s_k D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k))] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)(f(\mathbf{x}^*) - f(\mathbf{x}_{k+1})) + A_k(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)\langle \Delta(\mathbf{x}_{k+1}), \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1})\rangle - s_k D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k))] \\
&= \mathbb{E}[(A_{k+1} - A_k)\langle \Delta(\mathbf{x}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle - s_k D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k))] \\
&\quad + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})),
\end{aligned}$$

where the second inequality is due to the convexity of f and the last equality uses the fact that $\mathbf{y}_k$ is independent of $\Delta(\mathbf{x}_{k+1})$ and that $\mathbb{E}[\Delta(\mathbf{x}_{k+1})] = \mathbf{0}$. By Assumption 3.3 we have $D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) \le M_{h,\mathcal{X}}$. The strong convexity of h implies that $D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) \ge \mu_h/2\|\nabla h^*(\mathbf{y}_{k+1}) - \nabla h^*(\mathbf{y}_k)\|^2$. Hence, we have

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &\le \mathbb{E}\bigg[(A_{k+1} - A_k)\langle \Delta(\mathbf{x}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle - \frac{s_k\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2\bigg] \\
&\quad + \mathbb{E}[(s_{k+1} - s_k)M_{h,\mathcal{X}}] \\
&\le \mathbb{E}\bigg[(A_{k+1} - A_k)\|\Delta(\mathbf{x}_{k+1})\|_*\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\| - \frac{s_k\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2\bigg] \\
&\quad + \mathbb{E}[(s_{k+1} - s_k)M_{h,\mathcal{X}}] \\
&\le \frac{(A_{k+1} - A_k)^2\sigma^2}{2s_k\mu_h} + (s_{k+1} - s_k)M_{h,\mathcal{X}},
\end{aligned} \tag{A.1}$$

where the second inequality follows from Cauchy-Schwartz inequality and the last inequality is due to the simple inequality that $bx - ax^2/2 \le b^2/2a, \forall\ a \ge 0$ and that $\mathbb{E}[\|\Delta(\mathbf{x}_{k+1})\|_*^2] \le \sigma^2$. Summing (A.1) over 0 to $k-1$ gives

$$\mathcal{E}_k \le \mathcal{E}_0 + M_{h,\mathcal{X}}s_k + \frac{\sigma^2}{2\mu_h}\sum_{i=0}^{k-1}\frac{(A_{i+1} - A_i)^2}{s_i}. \tag{A.2}$$

To achieve the optimal rate we set $A_k = \frac{k(k+1)}{2}$ and $s_k = \sigma(k+1)^{3/2} + 1$ and we have

$$\begin{aligned}
\mathcal{E}_k &\le \mathcal{E}_0 + M_{h,\mathcal{X}}\Big(\sigma(k+1)^{3/2} + 1\Big) + \frac{\sigma}{2\mu_h}\sum_{i=0}^{k-1}\sqrt{i+1} \\
&\le \mathcal{E}_0 + M_{h,\mathcal{X}}\big(\sigma(k+1)^{3/2} + 1\big) + \frac{\sigma}{3\mu_h}(k+1)^{3/2},
\end{aligned}$$

where the second inequality follows from Lemma B.1. Plug in the definition of $\mathcal{E}_k$ and we have

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \frac{2(\mathcal{E}_0 + M_{h,\mathcal{X}})}{k(k+1)} + \bigg(M_{h,\mathcal{X}}\sigma + \frac{\sigma}{3\mu_h}\bigg)\frac{\sqrt{k+1}}{k}.$$

. □

A.2 Proof of Convergence of Algorithm 2

We first lay down the following technical lemma about the gradient bound of f in $\mathcal{X}$.

Lemma A.1. Under Assumptions 3.1, 3.2 and 3.3, we have

$$\|\nabla f(\mathbf{x})\|_* \leq \frac{L_f\sqrt{2M_{h,\mathcal{X}}}}{\sqrt{\mu_h}} + \|\nabla f(\mathbf{x}^*)\|_*.$$

Proof of Theorem 5.4. Let $\mathcal{E}_k = \mathbb{E}[A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + s_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))]$. Note that by three point identity we have

$$D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1})) = \langle \mathbf{y}_k - \mathbf{y}_{k+1}, \mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\rangle.$$

Therefore we have

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + s_k\mathbb{E}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))] \\
&= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + s_k\mathbb{E}[D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1})) + \langle \mathbf{y}_k - \mathbf{y}_{k+1}, \mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\rangle] \\
&\leq \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)(\langle \nabla \widetilde{f}(\mathbf{x}_{k+1}, \xi_{k+1}), \mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\rangle) + s_k D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1}))],
\end{aligned}$$

where the last inequality follows from (5.3b). We further plug in the update of $\nabla h^*(\mathbf{y}_k)$ in (5.3a) and obtain

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &\leq \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)\langle \nabla f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1}\rangle - A_k\langle \nabla f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)\langle \Delta(\mathbf{x}_{k+1}), \mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\rangle + s_k D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1}))] \\
&\leq \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) - A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*))] + \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)(f(\mathbf{x}^*) - f(\mathbf{x}_{k+1})) + A_k(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}))] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)\langle \Delta(\mathbf{x}_{k+1}), \mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\rangle + s_k D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1}))] \\
&= \mathbb{E}[(s_{k+1} - s_k)D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) + s_k D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1}))],
\end{aligned} \tag{A.3}$$

where the second inequality is due to the convexity of f and the last equality uses the fact that $\mathbf{y}_k$ is independent of $\Delta(\mathbf{x}_{k+1})$ and that $\mathbb{E}[\Delta(\mathbf{x}_{k+1})] = \mathbf{0}$. By Assumption 3.3 we have $D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) \leq M_{h,\mathcal{X}}$. Then by smoothness of h^*, we have

$$\begin{aligned}
D_h(\nabla h^*(\mathbf{y}_k), \nabla h^*(\mathbf{y}_{k+1})) &= D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k) \\
&\leq \frac{1}{2\mu_h}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_*^2 \\
&\leq \frac{8L_f^2 M_{h,\mathcal{X}} + 2\mu_h\sigma^2 + 4\|\nabla f(\mathbf{x}^*)\|_*^2}{\mu_h}\frac{(A_{k+1} - A_k)^2}{2\mu_h s_k^2},
\end{aligned} \tag{A.4}$$

where the last inequality is due to (5.3b) and Lemma A.1. Submit (A.4) back into (A.3) and we get the following bound

$$\mathcal{E}_{k+1} - \mathcal{E}_k \leq \mathbb{E}\bigg[M_{h,\mathcal{X}}(s_{k+1} - s_k) + \frac{4L_f^2 M_{h,\mathcal{X}} + \mu_h\sigma^2 + 2\|\nabla f(\mathbf{x}^*)\|_*^2}{\mu_h^2}\frac{(A_{k+1} - A_k)^2}{s_k}\bigg] \tag{A.5}$$

Therefore, by summing (A.5) over 0 to $k-1$, we have

$$\mathcal{E}_k \le \mathcal{E}_0 + M_{h,\mathcal{X}} s_k + \frac{4L_f^2 M_{h,\mathcal{X}} + \mu_h \sigma^2 + 2\|\nabla f(\mathbf{x}^*)\|_*^2}{\mu_h^2} \sum_{j=0}^{k-1} \frac{(A_{j+1}-A_j)^2}{s_j}.$$

If we choose $A_j = j(j+1)/2$ and $s_j = j^{3/2}$, then we have

$$\begin{aligned}
\mathcal{E}_k &\le \mathcal{E}_0 + M_{h,\mathcal{X}} k^{3/2} + \frac{4L_f^2 M_{h,\mathcal{X}} + \sigma^2 \mu_h + 2\|\nabla f(\mathbf{x}^*)\|_*^2}{\mu_h^2} \sum_{j=0}^{k-1} \sqrt{j} \\
&\le \mathcal{E}_0 + M_{h,\mathcal{X}} k^{3/2} + \frac{4L_f^2 M_{h,\mathcal{X}} + \sigma^2 \mu_h + 2\|\nabla f(\mathbf{x}^*)\|_*^2}{\mu_h^2} k^{3/2},
\end{aligned}$$

where the last inequality is due to Lemma B.1. By the definition of $\mathcal{E}_k$ and that $D_h(\cdot,\cdot) \ge 0$, we have $\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \mathcal{E}_k / A_k$. Finally, the convergence rate of Algorithm 2 is

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \frac{2\mathcal{E}_0}{k(k+1)} + \frac{2((4L_f^2 + \mu_h^2) M_{h,\mathcal{X}} + \sigma^2 \mu_h + 2\|\nabla f(\mathbf{x}^*)\|_*^2)\sqrt{k}}{\mu_h^2 (k+1)} = O\bigg(\frac{1}{k^2} + \frac{1}{\sqrt{k}}\bigg).$$

□

A.3 Proof of Convergence of Algorithm 3

Proof of Theorem 5.6. Consider the Lyapunov function $\mathcal{E}_k = \mathbb{E}[A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + s_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))]$.

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &= \mathbb{E}[A_{k+1} f(\mathbf{x}_{k+1}) - A_k f(\mathbf{x}_k) - (A_{k+1} - A_k) f(\mathbf{x}^*)] \\
&\quad + \mathbb{E}[s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))) + (s_{k+1} - s_k) D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{z}_{k+1})) + A_k(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_k)) + (A_{k+1} - A_k)(f(\mathbf{z}_{k+1}) - f(\mathbf{x}^*))] \\
&\quad + \mathbb{E}[s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))) + (s_{k+1} - s_k) D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{z}_{k+1})) + A_k \langle \nabla f(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_k \rangle + (s_{k+1} - s_k) M_{h,\mathcal{X}}] \\
&\quad + \mathbb{E}[s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))) + (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}^* \rangle] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{z}_{k+1})) + (A_{k+1} - A_k)(\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^* \rangle)] \\
&\quad + \mathbb{E}[s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))) + (s_{k+1} - s_k) M_{h,\mathcal{X}}],
\end{aligned} \tag{A.6}$$

where the first inequality follows from the convexity of f and Assumption 3.3, and the second inequality is due to (5.4a). We next bound the term $(A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^* \rangle$.

$$\begin{aligned}
&(A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^* \rangle \\
&= (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}) \rangle + (A_{k+1} - A_k)\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^* \rangle \\
&\quad - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^* \rangle \\
&= (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}) \rangle - s_k \langle \mathbf{y}_{k+1} - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^* \rangle \\
&\quad - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^* \rangle \\
&= (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}) \rangle - s_k D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) \\
&\quad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^* \rangle \\
&\le (A_{k+1} - A_k)\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}) \rangle - \frac{\mu_h s_k}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2 \\
&\quad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^* \rangle,
\end{aligned} \tag{A.7}$$

where the second equality is due to (5.4b), the third equality follows from three point identity, and the inequality follows from Assumption 3.2. Then we denote

$$\mathbf{w} = \frac{A_{k+1} - A_k}{A_{k+1}} \nabla h^*(\mathbf{y}_{k+1}) + \frac{A_k}{A_{k+1}} \mathbf{x}_k, \tag{A.8}$$

and obviously $\mathbf{w} \in \mathcal{X}$. Comparing this definition with (5.4a), we have $\mathbf{z}_{k+1} - \mathbf{w} = (A_{k+1} - A_k)/A_{k+1}(\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}))$, which implies

$$
\begin{aligned}
&(A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\le A_{k+1}\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \mathbf{z}_{k+1} - \mathbf{w}\rangle - \frac{\mu_h s_k A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{w}\|^2 \\
&\qquad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\le A_{k+1}\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \mathbf{z}_{k+1} - \mathbf{w}\rangle - \frac{\mu_h s_k A_{k+1}^2}{(A_{k+1} - A_k)^2} D_h(\mathbf{z}_{k+1}, \mathbf{w}) \\
&\qquad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\le A_{k+1}\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}, \xi_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \frac{\mu_h^2 s_k A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2 \\
&\qquad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&= A_{k+1}\langle \nabla f(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \frac{\mu_h^2 s_k A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2 + A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle \\
&\qquad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)((\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\le A_{k+1}(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_{k+1})) + \left(\frac{A_{k+1}L_f}{2} - \frac{\mu_h^2 s_k A_{k+1}^2}{2(A_{k+1} - A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2 + A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle \\
&\qquad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle,
\end{aligned}
$$

where the second inequality is due to the quadratic growth of $D_h(\cdot, \cdot)$, the third inequality is due to optimal condition of (5.4c) and Assumption 3.2. Here we choose M_k to be $M_k = L_f(A_{k+1} - A_k)^2/(\mu_h^2 s_k A_{k+1})$. Furthermore, we choose $A_k = \mu_h^2 k(k+1)/(4L_f)$ and thus

$$
\begin{aligned}
&(A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\le A_{k+1}(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_{k+1})) + \frac{A_{k+1}}{2}\left(L_f - \frac{L_f(k+2)}{k+1}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2 - \frac{\mu_h^2 (s_k - 1) A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2 \\
&\qquad + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\qquad + A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle \\
&\le A_{k+1}(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_{k+1})) + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))) - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\qquad + A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \frac{\mu_h^2 (s_k - 1) A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2.
\end{aligned}
$$

Take the expectation and note that $\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)$ are independent of $\Delta(\mathbf{z}_{k+1})$ and $\mathbb{E}[\Delta(\mathbf{z}_{k+1})] = \mathbf{0}$. We obtain

$$
\begin{aligned}
&\mathbb{E}[(A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_{k+1})) + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})))] \\
&\qquad + \mathbb{E}\left[A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \frac{\mu_h^2 (s_k - 1) A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2\right] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_{k+1})) + s_k(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})))] \\
&\qquad + \mathbb{E}\left[A_{k+1}\|\Delta(\mathbf{z}_{k+1})\|_* \cdot \|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\| - \frac{\mu_h^2 (s_k - 1) A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2\right],
\end{aligned}
\tag{A.9}
$$

where the last inequality comes from Cauchy-Schwartz inequality. Assume that $s_k > 1$, submitting (A.9) back into (A.6) yields

$$
\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &\le \mathbb{E}\left[A_{k+1}\|\Delta(\mathbf{z}_{k+1})\|_* \cdot \|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\| - \frac{\mu_h^2 (s_k - 1) A_{k+1}^2}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2\right] \\
&\qquad + \mathbb{E}[(s_{k+1} - s_k)M_{h,\mathcal{X}}] \\
&\le \frac{(A_{k+1} - A_k)^2\sigma^2}{2\mu_h^2 (s_k - 1)} + (s_{k+1} - s_k)M_{h,\mathcal{X}},
\end{aligned}
$$

where the second inequality follows the simple inequality that $bx - ax^2/2 \le b^2/2a,\ \forall a > 0$ and $\mathbb{E}[\|\Delta(\mathbf{x}_{k+1})\|_*^2] \le \sigma^2$. Sum up the above inequality from 0 to $k-1$ and we have

$$\mathcal{E}_k \le \mathcal{E}_0 + \frac{\sigma^2}{2\mu_h^2}\sum_{i=0}^{k-1}\frac{(A_{i+1}-A_i)^2}{s_i - 1} + s_k M_{h,\mathcal{X}}.$$

Plugging the choice $A_k = \mu_h k(k+1)/(4L_f)$ and setting $s_k = \sigma/L_f(k+1)^{3/2} + 1$, we have

$$\begin{aligned}\mathcal{E}_k &\le \mathcal{E}_0 + \frac{\mu_h^2\sigma}{8L_f}\sum_{i=0}^{k-1}\sqrt{i+1} + \frac{M_{h,\mathcal{X}}\sigma(k+1)^{3/2}}{L_f} + M_{h,\mathcal{X}}\\ &\le \mathcal{E}_0 + M_{h,\mathcal{X}} + \frac{\mu_h^2\sigma(k+1)^{3/2}}{12L_f} + \frac{M_{h,\mathcal{X}}\sigma(k+1)^{3/2}}{L_f},\end{aligned}$$

where the second inequality is due to Lemma B.1. By the definition of $\mathcal{E}_k$, we have

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \frac{4L_f(\mathcal{E}_0 + M_{h,\mathcal{X}})}{\mu_h^2 k(k+1)} + \frac{\sigma(\mu_h^2 + 12M_{h,\mathcal{X}})\sqrt{k+1}}{3\mu_h^2 k}.$$

□

B Technical Lemmas

In this section, we provide the proof of technical lemmas used in the proof of main theorems.

Proof of Lemma A.1. By smoothness of f we have

$$\|\nabla f(\mathbf{x})\|_* = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_* + \|\nabla f(\mathbf{x}^*)\|_* \le L_f\|\mathbf{x} - \mathbf{x}^*\| + \|\nabla f(\mathbf{x}^*)\|_*.$$

By Assumption 3.2 we have

$$D_h(\mathbf{x}, \mathbf{x}') = h(\mathbf{x}) - h(\mathbf{x}') - \langle\nabla h(\mathbf{x}'), \mathbf{x}'\rangle \ge \frac{\mu_h}{2}\|\mathbf{x} - \mathbf{x}'\|^2.$$

Combining the two inequalities above yields

$$\|\nabla f(\mathbf{x})\|_* \le \frac{\sqrt{2}L_f}{\sqrt{\mu_h}}\sqrt{D_h(\mathbf{x}, \mathbf{x}')} + \|\nabla f(\mathbf{x}^*)\|_* \le \frac{L_f\sqrt{2M_{h,\mathcal{X}}}}{\sqrt{\mu_h}} + \|\nabla f(\mathbf{x}^*)\|_*,$$

where the last inequality comes from Assumption 3.3. □

Lemma B.1. (Chlebus, 2009) For $p < 0$, the divergence rate of p-series is given by

$$1 + \frac{k^{1-p} - 1}{1-p} \le \sum_{j=1}^{k}\frac{1}{j^p} \le \frac{(k+1)^{1-p} - 1}{1-p}.$$

Lemma B.2 (The three point identity (Chen and Teboulle, 1993))**.** For any a, b that are interior points of dom h and $c \in$ dom h, we have

$$D_h(c, a) + D_h(a, b) - D_h(c, b) = \langle\nabla h(b) - \nabla h(a), c - a\rangle.$$