# Continuous and Discrete-time Accelerated Stochastic Mirror Descent for Strongly Convex Functions

**Pan Xu** [*1] **Tianhao Wang** [*2] **Quanquan Gu** [1]

## Abstract

We provide a second-order stochastic differential equation (SDE), which characterizes the continuous-time dynamics of accelerated stochastic mirror descent (ASMD) for strongly convex functions. This SDE plays a central role in designing new discrete-time ASMD algorithms via numerical discretization and providing neat analyses of their convergence rates based on Lyapunov functions. Our results suggest that the only existing ASMD algorithm, namely, AC-SA proposed in Ghadimi & Lan (2012) is one instance of its kind, and we can derive new instances of ASMD with fewer tuning parameters. This sheds light on revisiting accelerated stochastic optimization through the lens of SDEs, which can lead to a better understanding as well as new simpler algorithms of acceleration in stochastic optimization. Numerical experiments on both synthetic and real data support our theory.

## 1. Introduction

We study the following constrained optimization problem

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is $\mu$-strongly convex for some constant $\mu>0$ and $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$. Compared with general convexity, the strong convexity property often leads to great improvements in convergence rates of various algorithms and appears in enormous machine learning problems (Shalev-Shwartz & Kakade, 2009) due to the ubiquitousness of strongly convex loss functions such as log loss, square loss and strongly convex regularizers such as $\ell_2$ norm and Kullback-Leibler (KL) divergence.

When the constraint set $\mathcal{X}$ is endowed with a Bregman divergence (Bregman, 1967), (1.1) can be solved by mirror descent (MD) (Nemirovski, 1979; Nemirovski et al., 1983):

$$\begin{aligned}\mathbf{y}_{k+1} &= \nabla h(\mathbf{x}_k) - \eta_k \nabla f(\mathbf{x}_k),\\ \mathbf{x}_{k+1} &= \nabla h^*(\mathbf{y}_{k+1}),\end{aligned} \tag{1.2}$$

where $\eta_k>0$ is the step size, $h:\mathcal{X}\to\mathbb{R}$ is the distance generating function and $h^*$ is its conjugate. When $h(\cdot)=1/2\|\cdot\|_2^2$, the MD algorithm reduces to projected gradient descent (Luenberger et al., 1984), where $\nabla h^*$ is the Euclidean projection onto $\mathcal{X}$. In many machine learning applications with large scale and high dimensional data, the gradient of $f$ can be expensive to compute, and therefore stochastic gradient is used alternatively, which gives rise to stochastic mirror descent (SMD) (Nemirovski et al., 1983):

$$\begin{aligned}\mathbf{y}_{k+1} &= \nabla h(\mathbf{x}_k) - \eta_k G(\mathbf{x}_k;\xi_k),\\ \mathbf{x}_{k+1} &= \nabla h^*(\mathbf{y}_{k+1}),\end{aligned} \tag{1.3}$$

where $G(\mathbf{x};\xi)$ is the stochastic gradient indexed by random variable $\xi$. The stochastic gradient is often assumed to be an unbiased estimator of the full gradient, i.e., $\mathbb{E}[G(\mathbf{x};\xi)|\mathbf{x}]=\nabla f(\mathbf{x})$, and have bounded variance. If $f$ is $\mu$-strongly convex, 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 $O(M^2\log k/(\mu k))$ (Nesterov, 2009), where $\|\nabla f(\mathbf{x})\|\le M, \forall \mathbf{x}\in\mathcal{X}$ for a constant $M$.

If the objective function $f$ is additionally smooth, many convex optimization algorithms can be accelerated (Polyak, 1964; Nesterov, 1983; 2005; 2013) as well as SMD. Specifically, Ghadimi & Lan (2012) proposed a generic algorithmic framework to accelerate SMD for strongly convex and smooth functions, i.e., the accelerated stochastic approximation (AC-SA) algorithm with the following update:[1]:

$$\begin{aligned}\mathbf{x}_{k+1}^{\mathrm{md}} &= \frac{(1-\alpha_k)(\mu+\gamma_k)}{\gamma_k+(1-\alpha_k^2)\mu}\mathbf{x}_k^{\mathrm{ag}} + \frac{\alpha_k[(1-\alpha_k)\mu+\gamma_k]}{\gamma_k+(1-\alpha_k^2)\mu}\mathbf{x}_k,\\ \nabla h(\mathbf{x}_{k+1}) &= \frac{\alpha_k\mu}{\mu+\gamma_k}\nabla h(\mathbf{x}_{k+1}^{\mathrm{md}}) + \frac{(1-\alpha_k)\mu+\gamma_k}{\mu+\gamma_k}\nabla h(\mathbf{x}_k)\\ &\quad - \alpha_k/(\mu+\gamma_k)G(\mathbf{x}_{k+1}^{\mathrm{md}};\xi_{k+1}),\\ \mathbf{x}_{k+1}^{\mathrm{ag}} &= \alpha_k\mathbf{x}_{k+1} + (1-\alpha_k)\mathbf{x}_k^{\mathrm{ag}},\end{aligned} \tag{1.4}$$

[*]Equal contribution [1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA [2]School of Mathematical Sciences, University of Science and technology of China, Hefei, Anhui, China. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

*Proceedings of the $35^{th}$ International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. 

[1]Here we present the AC-SA algorithm in a slightly different but equivalent form as the original AC-SA algorithm in Ghadimi & Lan (2012), in order to make a clear comparison with our proposed continuous-time dynamics and new algorithms.

where $\alpha_k, \gamma_k > 0$ are step sizes, $\mu$ is the strong convexity constant of $f$, and $h$ is the strongly convex distance generating function associated with the Bregman divergence. Theoretically, for an $L$-smooth and $\mu$-strongly convex function, AC-SA achieves a nearly optimal convergence rate $O(L/k^2 + \sigma^2/(\mu k))$ in terms of expected function value gap, where $\sigma^2$ is the variance of the stochastic gradient. Compared with the non-accelerated SMD (1.3), the AC-SA algorithm enjoys an accelerated convergence rate in the sense that when the variance of the stochastic gradient vanishes, i.e., $\sigma = 0$, the convergence rate of AC-SA reduces to a faster rate $O(L/k^2)$. Despite the nearly optimal convergence rate, the update formulas in AC-SA are rather complicated and lack of intuition. As far as we know, this is the only algorithm of accelerated SMD in the literature.

In this work, we aim to better understand the acceleration of SMD and propose simpler algorithms of accelerated SMD for strongly convex functions. Inspired by an emerging body of research (Raginsky & Bouvrie, 2012; Su et al., 2014; Krichene et al., 2015; Mertikopoulos & Staudigl, 2016; Wilson et al., 2016; Wibisono et al., 2016; Krichene & Bartlett, 2017; Xu et al., 2018), which interpret existing accelerated deterministic and stochastic optimization algorithms by viewing them as Euler discretization of some continuous-time dynamics defined by ordinary differential equations (ODEs) or stochastic differential equations (SDEs), we propose a novel SDE-based interpretation of accelerated SMD for strongly convex functions. We derive several new discrete-time accelerated SMD algorithms from the proposed SDE. We provide a unified analysis of the convergence rate for both continuous-time dynamics and the discrete-time algorithms using Lyapunov functions. Thorough experiments corroborate our theory.

**Our Contributions:** Our key contributions are two-fold, which are summarized as follows.

- To the best of our knowledge, this is the first study, which proposes a stochastic differential equation-based interpretation for accelerated stochastic mirror descent when the objective function is strongly convex. We take a Lyapunov function based approach to prove that the convergence rate of the solution trajectories of the SDE is $O(1/t^2 + \sigma^2/t^{1-2q})$, where the variance of the stochastic noise is bounded by $O(\sigma^2 t^{2q})$ for constant $q < 1/2$.
- We invent several new accelerated algorithms of SMD via discretizing the proposed SDE using various Euler discretization schemes, and extend the Lyapunov function-based analysis for the continuous-time dynamics to the convergence analysis of the proposed discrete-time algorithms. Our analysis shows that these new algorithms achieve the same nearly optimal[2] convergence rate $O(L/k^2 + \sigma^2/(\mu k))$ for strongly convex and smooth functions as in Ghadimi & Lan (2012).

It is worth noting that while our study is focused on stochastic mirror descent, it sheds light on revisiting other stochastic optimization algorithms such as stochastic regularized dual averaging (Xiao, 2010; Chen et al., 2012) through the lens of SDEs, which can potentially lead to a better understanding and/or new simpler algorithms.

The remainder of this paper is organized as follows: In Section 2 we review the related work. In Section 3 we present the novel continuous-time dynamics for ASMD. We provide the convergence analysis of the continuous-time dynamics in Section 4. We design new algorithms for ASMD and prove their convergence rates in Section 5. Numerical experiments on both synthetic and real datasets are demonstrated in Section 6. Finally we conclude the paper in Section 7.

**Notation** Upper case letter $\boldsymbol{X}_t$ denotes a continuous-time curve vector, where $t \in \mathbb{T} \subseteq \mathbb{R}^+$ is the time index. $\dot{\boldsymbol{X}}_t$ with an over-dot denotes the time derivative of $\boldsymbol{X}_t$. Lower case letter $\mathbf{x}_k$ denotes 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 $\|\mathbf{x}\|_* = \sup_{\|\mathbf{y}\| \le 1} \langle \mathbf{x}, \mathbf{y} \rangle$. We use $\|\mathbf{A}\|_{\text{op}}$ to denote the operator norm (a.k.a., spectral norm) of matrix $\mathbf{A}$.

## 2. Related Work

Due to the great success in speeding up the convergence of gradient-based convex optimization algorithms such as gradient descent (Polyak, 1963) and mirror descent (Nemirovski, 1979; Nemirovski et al., 1983) for smooth functions, acceleration methods have received everlasting attention in convex optimization and led to various accelerated algorithms such as Polyak's heavy ball (Polyak, 1964), Nesterov's accelerated gradient descent (AGD) (Nesterov, 1983; 2005; 2013) and accelerated mirror descent (Nemirovski et al., 2009). The phenomenon of acceleration also plays an important role in stochastic convex optimization methods, such as accelerated stochastic gradient descent (Hu et al., 2009), accelerated stochastic mirror descent (SMD) (Lan, 2012; Ghadimi & Lan, 2012; 2013) and accelerated stochastic regularized dual averaging (SRDA) (Chen et al., 2012). Despite of the great success of acceleration methods in optimization, researchers have struggled in understanding the underlying mechanisms and their proofs. For example, the estimate sequences based proof for Nesterov's AGD is often considered as merely an "algebra trick" (Juditsky, 2013).

[2] Note that we can improve the convergence rate of our algorithms to be optimal, i.e., $O\big(\exp(-\sqrt{L/\mu}k) + \sigma^2/(\mu k)\big)$, using the same multi-stage shrinkage procedure technique proposed in Ghadimi & Lan (2013). However, the resulting algorithms are very complicated and are beyond the focus of this work.

To gain a better understanding of acceleration methods, there have emerged several lines of research that try to interpret Nesterov's accelerated gradient descent and accelerated mirror descent from various perspectives (Allen-Zhu & Orecchia, 2014; Su et al., 2014; Flammarion & Bach, 2015; Lin et al., 2015; Bubeck et al., 2015; Lan & Zhou, 2015; Lessard et al., 2016; Hu & Lessard, 2017; Scieur et al., 2017), among which the ODE based continuous-time dynamics is one of the very elegant interpretations. The discrete-time algorithms are closely connected with the continuous-time dynamics defined by ODE in the sense that they are nearly equivalent when the discretization step is sufficiently small. This ODE based analysis has also been extended to understanding the acceleration of mirror descent (Krichene et al., 2015; Wibisono et al., 2016; Wilson et al., 2016; Diakonikolas & Orecchia, 2017). Furthermore, Krichene et al. (2015); Wilson et al. (2016) proposed to discretize the continuous-time ODE and derive new discrete-time algorithms that enjoy the optimal convergence rates.

For general convex objective functions, Raginsky & Bouvrie (2012) proposed to study the continuous-time variant of stochastic mirror descent using stochastic differential equations (SDEs) and proved that the expected function gap is bounded by a quantity proportional to the variance of the stochastic gradient. Mertikopoulos & Staudigl (2016) proposed a modified SDE to describe SMD and proved almost-sure convergence of the function value along the solution trajectories of SDE to the minimal function value under an assumption that the stochastic variance vanishes with time. More recently, Krichene & Bartlett (2017) proposed a second-order SDE for accelerated SMD, and proved that the expected function value at the continuous-time iterative converges to the minimal function value at $O(1/t^{1/2})$ rate, where $t$ is the continuous time index. Xu et al. (2018) proposed a slightly different second-order SDE with refined rate $O(1/t^2 + 1/t^{1/2})$ to characterize accelerated SMD. However, all the studies along this line are limited to general convex functions, and are not applicable to the strongly convex functions. The continuous-time dynamics of accelerated stochastic mirror descent for strongly convex functions is still missing in the literature.

## 3. Continuous-time Dynamics for Accelerated SMD

We start with some preliminary definitions and propositions, and then we present the continuous-time dynamics for accelerated stochastic mirror descent from a Lagrangian mechanics (Wibisono et al., 2016) perspective.

### 3.1. Preliminaries

We first define the Bregman divergence associated with distance generating function $h$ and a constrained set $\mathcal{X}$.

**Definition 3.1.** The Bregman divergence is defined as

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

on a set $\mathcal{X} \subset \mathbb{R}^d$, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, where $h : \mathcal{X} \to \mathbb{R}$ is $\mu_h$-strongly convex, continuous and non-negative, and is also called distance-generating function.

It is easy to see that $D_h(\mathbf{x}, \mathbf{x}') \geq 0$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, due to the convexity of $h$.

**Definition 3.2.** The conjugate function of function $h : \mathcal{X} \to \mathbb{R}$ 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 $E$ which is a superset of $\mathcal{X}$.

In particular, if $h$ is strongly convex, it holds that

$$\nabla h^*(\mathbf{y}) = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{y}, \mathbf{x} \rangle - h(\mathbf{x}), \quad \text{for } \mathbf{y} \in E^*. \quad (3.2)$$

where $\nabla h^*$ is the gradient of $h^*$, and is called as mirror map, which maps the point in the dual space $E^*$ back to $\mathcal{X}$. The mirror map $\nabla h^*$ plays a central role in mirror descent as well as in deriving the continuous-time dynamics for MD.

Then we lay down the definitions of strong convexity and smoothness of $f$.

**Definition 3.3.** A function $f$ is $\mu$-strongly convex with respect to some $h$, if there exists a constant $\mu > 0$, such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, it satisfies

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \mu D_h(\mathbf{x}, \mathbf{y}).$$

**Definition 3.4.** A function $f$ is $L$-smooth, if there exists a constant $L > 0$, such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, it satisfies

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L \|\mathbf{x} - \mathbf{y}\|,$$

where $\|\cdot\|_*$ is the dual norm.

It is worth emphasizing that smoothness is essential for accelerating the convergence of convex optimization methods (Nemirovski et al., 1983; Nesterov, 2013).

### 3.2. SDE-based Dynamics

Now we are going to derive a continuous-time dynamics for accelerated stochastic mirror descent by viewing the optimizing process as a physical process. Specifically, we define a mechanical system associated with optimization problem (1.1), where $\boldsymbol{X} \in \mathcal{X}$ and $\boldsymbol{V} \in E$ denote the position and velocity of a particle respectively. The Bregman Lagrangian (Wibisono et al., 2016) of this physical system is defined as

$$\widetilde{\mathcal{L}}(\boldsymbol{X}, \boldsymbol{V}, t) = e^{\alpha_t + \gamma_t} \big( D_h(\boldsymbol{X} + e^{-\alpha_t} \boldsymbol{V}, \boldsymbol{X}) - e^{\beta_t} f(\boldsymbol{X}) \big), \quad (3.3)$$

where $\alpha_t, \beta_t, \gamma_t$ are arbitrary scaling functions that are continuously differentiable, $D_h(\boldsymbol{X} + e^{-\alpha_t}\boldsymbol{V}, \boldsymbol{X})$ is the kinetic function, and $f(\boldsymbol{X})$ is the potential function. This Bregman Lagrangian is first proposed by Wibisono et al. (2016) for characterizing deterministic acceleration methods, and later adopted by Xu et al. (2018) to analyze accelerated SMD. However, the Bregman Lagrangian in (3.3) is defined for general convex functions. When $f$ is $\mu$-strongly convex, we propose the following extended Bregman Lagrangian:

$$\begin{aligned}&\mathcal{L}(\boldsymbol{X}, \boldsymbol{V}, t)\\ &= e^{\alpha_t+\beta_t+\gamma_t}\big(\mu D_h(\boldsymbol{X} + e^{-\alpha_t}\boldsymbol{V}, \boldsymbol{X}) - f(\boldsymbol{X})\big)\end{aligned} \tag{3.4}$$

For the action functional $J(\boldsymbol{X}) = \int_{\mathbb{T}} \mathcal{L}(\boldsymbol{X}_t, \dot{\boldsymbol{X}}_t, t)\mathrm{d}t$ defined on curves $\{\boldsymbol{X}_t\}$, Hamilton's principle (Bailey, 1982), a.k.a., principle of least action, states that minimizing the action functional $J(\boldsymbol{X})$ requires the curve $\boldsymbol{X}_t$ to satisfy the following Euler-Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{X}}_t}(\boldsymbol{X}_t, \dot{\boldsymbol{X}}_t, t)\right\} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{X}_t}(\boldsymbol{X}, \dot{\boldsymbol{X}}_t, t). \tag{3.5}$$

Submit the Bregman Lagrangian (3.4) into the Euler-Lagrange equation (3.5), we have the following minimizing trajectory for $J(\boldsymbol{X})$

$$\begin{aligned}&e^{-\alpha_t}\frac{\mathrm{d}}{\mathrm{d}t}\nabla h\big(\boldsymbol{X}_t + e^{-\alpha_t}\dot{\boldsymbol{X}}_t\big)\\ &= -\big(\nabla h\big(\boldsymbol{X}_t + e^{-\alpha_t}\dot{\boldsymbol{X}}_t\big) - \nabla h(\boldsymbol{X}_t)\big) - \frac{1}{\mu}\nabla f(\boldsymbol{X}_t)\\ &\quad + (\dot{\beta}_t + \dot{\gamma}_t)e^{-\alpha_t}\big(\nabla h(\boldsymbol{X}_t) - \nabla h\big(\boldsymbol{X}_t + e^{-\alpha_t}\dot{\boldsymbol{X}}_t\big)\big).\end{aligned} \tag{3.6}$$

For the simplicity of presentation, we adopt the following ideal scaling condition which ensures the stability of (3.5):

$$\dot{\beta}_t = e^{\alpha_t}, \quad \dot{\gamma}_t = -e^{\alpha_t}. \tag{3.7}$$

Note that our ideal scaling condition is different from that of Wibisono et al. (2016) since our Bregman Lagrangian (3.4) is designed for strong convex function $f$, which differs from (3.3) defined by Wibisono et al. (2016). Substituting the scaling condition into (3.6), we obtain the following second-order ordinary differential equation whose solution minimizing the action functional $J(\boldsymbol{X})$

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\nabla h\big(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t\big)\\ &= -\dot{\beta}_t(\nabla f(\boldsymbol{X}_t)/\mu + \nabla h\big(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t\big) - \nabla h(\boldsymbol{X}_t)).\end{aligned}$$

Up to now, we have derived a continuous-time dynamics with full gradient $\nabla f(\boldsymbol{X}_t)$. In order to account for the randomness of stochastic gradient used in stochastic mirror descent, we add a Brownian motion term into the above second-order ODE and obtain the following second-order Itô stochastic differential equation (SDE) (Øksendal, 2003)

$$\begin{aligned}&\mathrm{d}\nabla h(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t)\\ &= -\dot{\beta}_t\left(\frac{1}{\mu}\nabla f(\boldsymbol{X}_t) + \nabla h(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t)\right)\mathrm{d}t\\ &\quad + \dot{\beta}_t\nabla h(\boldsymbol{X}_t)\mathrm{d}t - \dot{\beta}_t\sqrt{\delta}\boldsymbol{\sigma}(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t,\end{aligned} \tag{3.8}$$

where $\boldsymbol{B}_t \in \mathbb{R}^d$ is the standard Brownian motion, $\boldsymbol{\sigma}(\boldsymbol{X}_t, t) \in \mathbb{R}^{d\times d}$ is called the diffusion coefficient matrix and $\delta > 0$ is a constant. For the ease of demonstration, we further introduce another continuous curve $\{\boldsymbol{Y}_t\}$ in the dual space $E^*$ which is defined as $\boldsymbol{Y}_t = \nabla h(\boldsymbol{X}_t + 1/\dot{\beta}_t\dot{\boldsymbol{X}}_t)$. Since $\nabla h^*(\nabla h(\mathbf{x})) = \mathbf{x}$ for $\mathbf{x} \in \mathcal{X}$ when $h$ is strongly convex (Banerjee et al., 2005), we can rewrite the second-order SDE in (3.8) as the following system of first-order SDEs:

$$\mathrm{d}\boldsymbol{X}_t = \dot{\beta}_t(\nabla h^*(\boldsymbol{Y}_t) - \boldsymbol{X}_t)\mathrm{d}t, \tag{3.9a}$$

$$\begin{aligned}\mathrm{d}\boldsymbol{Y}_t = -\dot{\beta}_t\Bigg(&\frac{1}{\mu}\nabla f(\boldsymbol{X}_t)\mathrm{d}t + (\boldsymbol{Y}_t - \nabla h(\boldsymbol{X}_t))\mathrm{d}t\\ &+ \frac{\sqrt{\delta}\boldsymbol{\sigma}(\boldsymbol{X}_t, t)}{\mu}\mathrm{d}\boldsymbol{B}_t\Bigg).\end{aligned} \tag{3.9b}$$

We can see that the continuous-time curve $\boldsymbol{X}_t$ is updated in the primal space according to (3.9a), and the continuous-time curve $\boldsymbol{Y}_t$ is updated in the dual space of $\mathcal{X}$ according to (3.9b). This is analogous to the primal update and dual update in the discrete-time stochastic mirror descent. The diffusion term $-\sqrt{\delta}\dot{\beta}_t\boldsymbol{\sigma}(\boldsymbol{X}_t, t)/\mu$ accounts for the randomness introduced by stochastic gradient and corresponds the standard deviation of the stochastic gradient. It is also interesting to compare the above continuous-time dynamics for strongly convex functions with that for general convex functions as in Krichene & Bartlett (2017); Xu et al. (2018). The continuous-time dynamics for general convex functions has an extra sensitivity parameter, which does not appear in the Bregman Lagrangian but is artificially introduced to ensure the convergence of the dynamics. In contrast, such kind of sensitivity parameter is not needed in our continuous-time dynamics (3.9) for strongly convex functions, and makes the dynamics more concise and intuitive.

## 4. Convergence Analysis of the Continuous-time Dynamics

In this section, we provide a Lyapunov function based convergence analysis of our proposed continuous-time dynamics in (3.9). The high-level road map of our proof follows from the previous work on the continuous-time dynamics of accelerated SMD for general convex functions (Krichene & Bartlett, 2017; Xu et al., 2018). In detail, we define the Lyapunov function as follows

$$\mathcal{E}_t = e^{\beta_t}\big(f(\boldsymbol{X}_t) - f(\mathbf{x}^*) + \mu D_h(\mathbf{x}^*, \nabla h^*(\boldsymbol{Y}_t)\big). \tag{4.1}$$

We emphasize that the above Lyapunov function is unique to our continuous-time dynamics (3.9), and it is different from that in Xu et al. (2018) since we do not need the scaling parameter for the Bregman divergence. (4.1) is also different from the Lyapunov function used in Wilson et al. (2016) due to the fact that they are dealing with the deterministic problem and that their Lyapunov function for strongly convex functions is in the Euclidean space.

The following theorem characterizes the convergence rate of the solution of the continuous-time dynamics (3.9).

**Theorem 4.1.** Suppose $f$ is $\mu$-strongly convex with respect to a distance generating function $h$ and assume $h$ is $\mu_h$-strongly convex. Then the dynamics (3.9) for accelerated stochastic mirror descent satisfies

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] \leq e^{-\beta_t}\mathcal{E}_0 + e^{-\beta_t}\mathbb{E}\bigg[\int_0^t \frac{\delta\dot{\beta}_r^2 e^{\beta_r}}{2\mu}\operatorname{tr}\big(\boldsymbol{\sigma}_r^\top \nabla^2 h^*(\boldsymbol{Y}_r)\boldsymbol{\sigma}_r\big)\mathrm{d}r\bigg],$$

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

In Theorem 4.1 we assume that $h$ is $\mu_h$-strongly convex, which is a widely imposed assumption in the literature of mirror descent (Nesterov, 1983; Raginsky & Bouvrie, 2012; Lan, 2012; Ghadimi & Lan, 2012; Chen et al., 2012). By the property of conjugate function (Banerjee et al., 2005), we know that equivalently $h^*$ is $1/\mu_h$-smooth, i.e., $\|\nabla^2 h^*(\mathbf{y})\|_{\text{op}} \leq 1/\mu_h$ for all $\mathbf{y} \in E^*$.

**Remark 4.2.** If the diffusion coefficient $\boldsymbol{\sigma}_t$ satisfies $\|\boldsymbol{\sigma}_t\|_{\text{op}} \leq \sigma t^q$ for some constants $\sigma > 0$ and $q < 1/2$, and we choose $\beta_t = 2\log t$, we obtain the convergence rate

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

When $0 < q < 1/2$, the variance of SDE (3.9) increases with time $t$, while we can still guarantee the convergence of its solution to the minimizer of $f$. When $q = 0$, i.e., the variance of the stochastic gradient is uniformly bounded by a constant $\sigma^2$, we further obtain the convergence rate $O(1/t^2 + \sigma^2/(\mu t))$, which matches the near optimal convergence rate for accelerated SMD with strongly convex objective functions (Ghadimi & Lan, 2012).

**Remark 4.3.** It is interesting to compare the convergence rate of the continuous-time dynamics for strongly convex functions with the convergence rate for general convex function. More specifically, under the assumption that $\|\boldsymbol{\sigma}_t\| \leq \sigma t^q$ for some constants $\sigma > 0$ and $q < 1/2$, the convergence rate in Theorem 4.1 is faster by a factor of $\sigma^2/t^{1/2-q}$ than the rate $O(1/t^2 + \sigma^2/(\mu t^{1/2-q}))$ of the continuous-times dynamics of accelerated SMD for general convex functions, derived in Xu et al. (2018).

## 5. New ASMD Algorithms

In this section, we design various novel accelerated stochastic mirror descent algorithms for strongly convex objective functions based on Euler discretization of the SDE in (3.9). We also extend the Lyapunov function based analysis in Section 4 to analyze the convergence of the proposed discrete-time algorithms.

More specifically, we use implicit and explicit Euler discretization schemes (Kloeden & Platen, 1992) of differential equations and their composition to derive several discrete algorithms from (3.9). Let $\delta$ 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 explicit (forward) Euler discretization for the time derivatives $\dot{\boldsymbol{X}}_t$ and $\dot{\boldsymbol{Y}}_t$ are defined as

$$(\mathbf{x}_{k+1} - \mathbf{x}_k)/\delta \approx \dot{\boldsymbol{X}}_t, \quad (\mathbf{y}_{k+1} - \mathbf{y}_k)/\delta \approx \dot{\boldsymbol{Y}}_t, \tag{5.1}$$

and the implicit (backward) Euler discretization for the time derivatives $\dot{\boldsymbol{X}}_t$ and $\dot{\boldsymbol{Y}}_t$ are defined as

$$(\mathbf{x}_k - \mathbf{x}_{k-1})/\delta \approx \dot{\boldsymbol{X}}_t, \quad (\mathbf{y}_k - \mathbf{y}_{k-1})/\delta \approx \dot{\boldsymbol{Y}}_t. \tag{5.2}$$

For the scaling parameters, we choose $A_k = e^{\beta_t}$ and the discretizations are as follows

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

Similar discretization schemes are also used in recent work (Wilson et al., 2016; Xu et al., 2018).

### 5.1. Implicit Euler Discretization

We first show that the implicit discretization of (3.9) gives rise to an instance of accelerated SMD algorithm, which is able to obtain the desired convergence rate for strongly convex and smooth functions.

By the implicit Euler discretization of $\dot{\boldsymbol{X}}_t$ and $\dot{\boldsymbol{Y}}_t$ in (5.2), the continuous dynamics (3.9) can be discretized as follows

$$\nabla h^*(\mathbf{y}_{k+1}) = \mathbf{x}_{k+1} + 1/\tau_k(\mathbf{x}_{k+1} - \mathbf{x}_k), \tag{5.3a}$$

$$\mathbf{y}_{k+1} - \mathbf{y}_k = -\tau_k\big[G(\mathbf{x}_{k+1}; \xi_{k+1})/\mu + \mathbf{y}_{k+1} - \nabla h(\mathbf{x}_{k+1})\big], \tag{5.3b}$$

where $\tau_k = (A_{k+1} - A_k)/A_k$ and $G(\mathbf{x}_{k+1}; \xi_{k+1})$ is the stochastic gradient. It is worth noting that (5.3) actually gives the optimality condition for a discrete-time algorithm, which is depicted in Algorithm 1. We use the notation $\widetilde{f}(\mathbf{x}; \xi)$ to denote the stochastic objective function for random vector $\xi$ and $\mathbf{x} \in \mathbb{R}^d$. The stochastic gradient is then $G(\mathbf{x}; \xi) = \nabla \widetilde{f}(\mathbf{x}; \xi)$. Line 4 in Algorithm 1 is immediately from equation (5.3a). Furthermore, it is easy to verify that the optimality condition of the minimization problem in Line 5 of Algorithm 1 is identical to equation (5.3b) by calculating the gradient of the objective in the this minimization problem with respect to $\mathbf{x}$ and setting it to be zero.

To analyze the convergence of Algorithm 1, analogous to the continuous-time Lyapunov function in (4.1), we define the discrete-time Lyapunov function as follows

$$\mathcal{E}_k = A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*) + \mu D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))). \tag{5.4}$$

**Algorithm 1** Accelerated Stochastic Mirror Descent

1: **Input:** $A_0 = 1, \mu$.
2: **for** $k = 1$ to $K$ **do**
3: $A_k = k(k+1)/2$, $\tau_k = (A_{k+1} - A_k)/A_k$.
4: $\nabla h^*(\mathbf{y}_{k+1}) = \mathbf{x}_{k+1} + \frac{1}{\tau_k}(\mathbf{x}_{k+1} - \mathbf{x}_k)$.
5: $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}} \big\{\widetilde{f}(\mathbf{x}; \xi_k) + \mu D_h(\mathbf{u}, \nabla h^*(\mathbf{y}_k)) - \mu D_h(\mathbf{x}, \nabla h^*(\mathbf{y}_k))\big\}$, where $\mathbf{u} = \mathbf{x} + \frac{1}{\tau_k}(\mathbf{x} - \mathbf{x}_k)$.
6: **end for**

Now we present the convergence rate of our discretized Algorithm 1 in the following theorem.

**Theorem 5.1.** Suppose $f$ is $\mu$-strongly convex with respect to a distance generating function $h$, which is $\mu_h$-strongly convex. Assume that $\mathbb{E}[\|G(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2 | \mathbf{x}] \le \sigma^2$ for some constant $\sigma$. If choosing $A_k = k(k+1)/2$ and $A_0 = 1$, the output of Algorithm 1 satisfies

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

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

**Remark 5.2.** Theorem 5.1 implies $O(1/k^2 + \sigma^2/(\mu k))$ convergence rate for Algorithm 1, which matches the nearly optimal rate for stochastic optimization in Ghadimi & Lan (2012). When the variance of the stochastic gradient vanishes, i.e., $\sigma = 0$, the convergence rate is improved to $O(1/k^2)$. In addition, the nearly optimal rate in Theorem 5.1 only requires the strong convexity of $f$ and distance generating function $h$, which is consistent with the theoretic result for the continuous dynamics. Note that the Algorithm 1 can also be viewed as an accelerated proximal point algorithm (Güler, 1992).

Similar to the analysis of the continuous-time dynamics, we also provide a proof sketch of Theorem 5.1 based on the discrete-time Lyapunov function (5.4). The complete proof can be found in the supplementary materials.

*Proof Sketch of Theorem 5.1.* Recall the Lyapunov function $\mathcal{E}_k$ in (5.4). Direct calculation on the subtraction of $\mathcal{E}_{i+1}$ and $\mathcal{E}_i$ and simplifying the result yields

$$\mathbb{E}[\mathcal{E}_{i+1} - \mathcal{E}_i] \le \frac{(A_{i+1} - A_i)^2}{A_i} \frac{\sigma^2}{2\mu\mu_h}.$$

By taking telescope sum from $i = 0$ to $i = k-1$, we obtain

$$\mathbb{E}[\mathcal{E}_k] \le \mathcal{E}_0 + \frac{\sigma^2}{2\mu\mu_h} \sum_{i=0}^{k-1} \frac{(A_{i+1} - A_i)^2}{A_i}.$$

Choosing $A_k = k(k+1)/2$ and $A_0 = 1$, we have $\sum_{i=0}^{k-1}(A_{i+1} - A_i)^2/A_i = \sum_{i=1}^{k-1}(i+1)/i \le 2k$. Thus by the definition of $\mathcal{E}_k$ we have

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

which completes the proof. □

### 5.2. Hybrid Euler Discretization

Despite its close connection with the continuous dynamics and its optimal convergence rate, Algorithm 1 involves an implicit update step (Line 5) which is pretty hard to solve in practice. To design a more practical algorithm for accelerated SMD, we now choose to combine the explicit discretization of $\dot{\boldsymbol{Y}}_t$ in (5.1) and the implicit discretization of $\dot{\boldsymbol{X}}_t$ in (5.2) to obtain the following optimality condition

$$\nabla h^*(\mathbf{y}_k) = \mathbf{x}_{k+1} + 1/\tau_k(\mathbf{x}_{k+1} - \mathbf{x}_k), \tag{5.5a}$$
$$\mathbf{y}_{k+1} - \mathbf{y}_k = -\tau_k\big[G(\mathbf{x}_{k+1}; \xi_{k+1})/\mu + \mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\big], \tag{5.5b}$$

where $\tau_k = (A_{k+1} - A_k)/A_k$. Under this optimality condition, a new discrete-time algorithm can be derived, which is displayed in Algorithm 2.

**Algorithm 2** Accelerated Stochastic Mirror Descent (ASMD)

1: **Input:** $A_0 = 1, \mu$.
2: **for** $k = 1$ to $K$ **do**
3: $A_k = k(k+1)/2$, $\tau_k = (A_{k+1} - A_k)/A_k$.
4: $\mathbf{x}_{k+1} = \frac{\tau_k}{1+\tau_k}\nabla h^*(\mathbf{y}_k) + \frac{1}{1+\tau_k}\mathbf{x}_k$.
5: $\mathbf{y}_{k+1} = (1 - \tau_k)\mathbf{y}_k - \tau_k/\mu\big(G(\mathbf{x}_{k+1}; \xi_{k+1}) - \mu\nabla h(\mathbf{x}_{k+1})\big)$.
6: **end for**

Under the same discrete-time Lyapunov function (5.4), we have the following convergence result for Algorithm 2.

**Theorem 5.3.** Suppose $f$ is $\mu$-strongly convex and $L$-smooth. Suppose the distance generating function $h$ is $\mu_h$-strongly convex and $1-$smooth. Further assume that $\mathbb{E}[\|G(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2 | \mathbf{x}] \le \sigma^2$ for some constant $\sigma$. If choosing $A_k = k(k+1)/2$ and $A_0 = 1$, the output of Algorithm 2 is bounded by

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \frac{2\mathcal{E}_0}{k(k+1)} + \frac{\mu_h B_0^2 + C B_0 \mu\sqrt{\mu_h^3 M_{h,\mathcal{X}}}}{\mu_h^2 \mu(k+1)},$$

where $B_0 = (L+\mu)\sqrt{2M_{h,\mathcal{X}}/\mu_h} + \sigma + \|\nabla f(\mathbf{x}^*)\|_*$, $C > 0$ is an absolute constant and $M_{h,\mathcal{X}} = \sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} D_h(\mathbf{x}, \mathbf{x}')$.

In Theorem 5.3, we assume $h$ is $L_h$-smooth, which means the Bregman divergence $D_h(\cdot, \cdot)$ grows quadratically. This assumption is also made in the analysis of AC-SA in Ghadimi & Lan (2012). To ease the presentation, we assume w.l.o.g. $L_h = 1$; otherwise we can scale $h$ to be $h' = h/L_h$ which is $\mu_h/L_h$-strongly convex and 1-smooth.

**Remark 5.4.** Theorem 5.3 suggests that the convergence rate of Algorithm 2 is $O(1/k^2 + (\|\nabla f(\mathbf{x}^*)\|_* + M_{h,\mathcal{X}} + \sigma^2)/k)$, which matches the nearly optimal convergence rate of stochastic optimization for strongly convex functions (Ghadimi & Lan, 2012). Nevertheless, when the variance of the stochastic gradient vanishes, i.e, $\sigma = 0$, this convergence rate cannot be improved to the faster rate $O(1/k^2)$ since the extra term with $M_{h,\mathcal{X}}$ and $\|\nabla f(\mathbf{x}^*)\|_*$ cannot be zero.

### 5.3. Discretization with an Additional Sequence

In order to design a practical algorithm that both attains the nearly optimal convergence rate and get rid of the additional assumption on $M_{h,\mathcal{X}}$, we introduce an additional updating sequence to ensure that we do not move too far away from the prox-center. This calibration idea has also been used in Lan (2012); Ghadimi & Lan (2012); Allen-Zhu & Orecchia (2014), but with different additional sequences. The optimality condition is then given by

$$\nabla h^*(\mathbf{y}_k) = \mathbf{x}_k + 1/\tau_k'(\mathbf{z}_{k+1} - \mathbf{x}_k) \tag{5.6a}$$

$$\mathbf{y}_{k+1} - \mathbf{y}_k = -\tau_k'[G(\mathbf{z}_{k+1};\xi_{k+1})/\mu + \mathbf{y}_k - \nabla h(\mathbf{z}_{k+1})], \tag{5.6b}$$

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \left\{\langle G(\mathbf{z}_{k+1};\xi_{k+1}), \mathbf{x}\rangle + M_k D_h(\mathbf{z}_{k+1}, \mathbf{x})\right\}, \tag{5.6c}$$

where $\tau_k' = (A_{k+1} - A_k)/A_{k+1}$ and $M_k = \mu\mu_h A_{k+1}A_k/(2(A_{k+1} - A_k)^2)$. The proposed algorithm, named **ASMD3**, is derived from the optimality condition (5.6) and displayed in Algorithm 3.

**Algorithm 3** ASMD with Additional Sequences (ASMD3)

1: **Input:** $\mu, L, \mu_h$.
2: **for** $k = 1$ to $K$ **do**
3: $A_k = \mu\mu_h^2(k+1)(k+2)/(4L) + 1$, $\tau_k' = (A_{k+1} - A_k)/A_{k+1}$, and $M_k = \mu\mu_h A_{k+1}A_k/(2(A_{k+1} - A_k)^2)$.
4: $\mathbf{z}_{k+1} = \tau_k'\nabla h^*(\mathbf{y}_k) + (1-\tau_k')\mathbf{x}_k$.
5: $\mathbf{y}_{k+1} = (1-\tau_k')\mathbf{y}_k - \tau_k'(G(\mathbf{z}_{k+1};\xi_{k+1})/\mu - \nabla h(\mathbf{z}_{k+1}))$.
6: $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \left\{\langle G(\mathbf{z}_{k+1};\xi_{k+1}), \mathbf{x}\rangle + M_k D_h(\mathbf{z}_{k+1}, \mathbf{x})\right\}$.
7: **end for**

Although Algorithm 3 has an additional sequence in the update formula compared with Algorithms 1 and 2, we can still use the same Lyapunov function (5.4) to analyze its convergence rate.

**Theorem 5.5.** Under the same conditions as in Theorem 5.3, by choosing $A_k = \mu\mu_h^2(k+1)(k+2)/(4L) + 1$, and $M_k = \mu\mu_h A_{k+1}A_k/(2(A_{k+1} - A_k)^2)$, the output of Algorithm 3 satisfies

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \frac{4LA_0\mathcal{E}_0 + 4\sigma^2 k}{\mu\mu_h^2(k+1)(k+2)},$$

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

**Remark 5.6.** Theorem 5.5 suggests that the convergence rate of **ASMD3** is $O(L/k^2 + \sigma^2/(\mu k))$ which matches the nearly optimal rate in Ghadimi & Lan (2012) and is also able to reduce to the faster rate $O(L/k^2)$ when the variance of the stochastic gradient vanishes.

**Remark 5.7.** An interesting fact is that the first two sequences in (1.4) of the AC-SA algorithm bear some similarity with (5.6a) and (5.6b) of Algorithm 3. However, there is only one tuning parameter, i.e., $A_k$ in our algorithm. In contrast, AC-SA has two independent parameters $\gamma_k$ and $\alpha_k$ that need to be tuned together in practice.Note that the first two sequences in Algorithm 3 correspond to the straightforward discretization of continuous-time dynamics (3.9). In this sense, the AC-SA algorithm in Ghadimi & Lan (2012) can also be viewed as a special discretization of our proposed continuous-time dynamics (3.9), yet with more parameters, and a different additional sequence for calibration.

## 6. Numerical Experiments

In this section, we carry out numerical experiments to support the theoretical guarantees of our proposed algorithms.

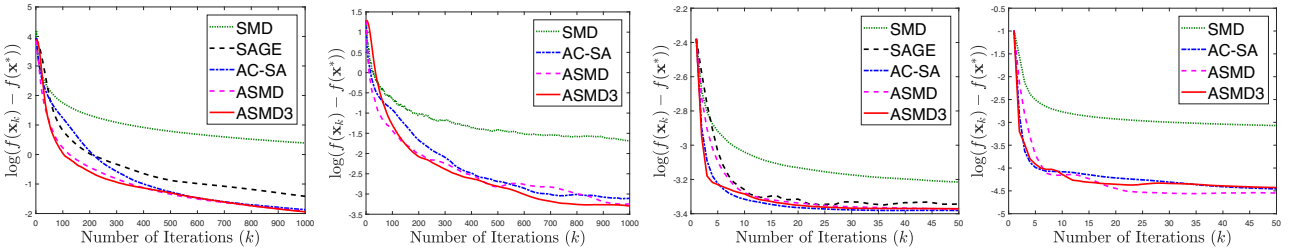
### 6.1. Compared Algorithms and Setup

We consider constrained stochastic optimization problems with strongly convex objective functions, and compare our Algorithm 2 (**ASMD**) and Algorithm 3 (**ASMD3**) with stochastic mirror descent (**SMD**), stochastic accelerated gradient (**SAGE**) (Hu et al., 2009) and accelerated stochastic approximation (**AC-SA**) (Ghadimi & Lan, 2012). Note that SAGE is not a mirror descent algorithm and we compare with it only when the constraint set is a Euclidean norm ball.

Two strongly convex loss functions are studied in our experiment: (1) $\ell_2$-regularized least square loss (a.k.a., ridge regression loss) $f(\mathbf{x}) = 1/(2n)\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^n$; and (2) $\ell_2$-regularized logistic regression loss $f(\mathbf{x}) = 1/n\sum_{i=1}^n \big( - y_i\mathbf{A}_{i*}\mathbf{x} + \log\big(1 + e^{\mathbf{A}_{i*}\mathbf{x}}\big)\big) + \lambda\|\mathbf{x}\|^2$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{x} \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ and $\mathbf{A}_{i*}$ is the $i$-th row of $\mathbf{A}$. We consider the following two distance generating function $h(\cdot)$ and the corresponding constraint set $\mathcal{X}$. Both of them are widely used in mirror descent based algorithms (Krichene et al., 2015; Krichene & Bartlett, 2017; Xu et al., 2018) because the mirror map $\nabla h^*$ has a closed-form expression.

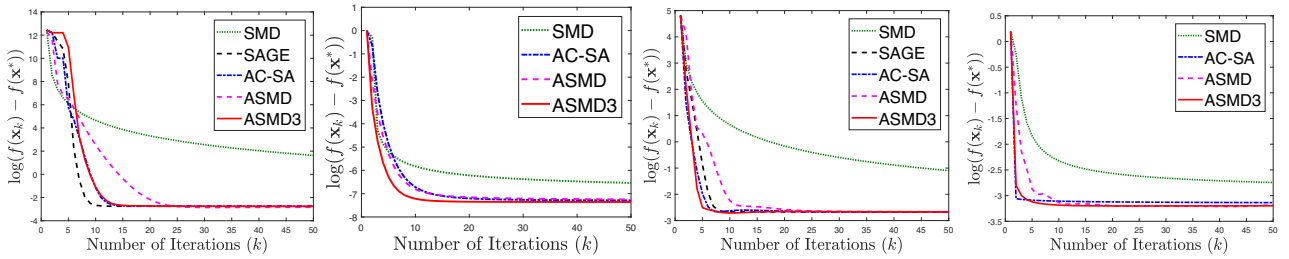
**Squared Euclidean norm**: $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$ and $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \le R\}$ for some constant $R > 0$.

**Negative entropy**: $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$ and $\mathcal{X} = \{\mathbf{x} : \sum_{i=1}^d x_i = 1, x_i \ge 0 \text{ for } i = 1, \ldots, d\}$.




(a) $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$ (b) $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$ (c) $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$ (d) $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$

*Figure 1.* Logarithmic averaged function value gap over 50 repetitions under different loss and distance generating functions on synthetic data. (a) and (b) use the ridge regression objective function; (c) and (d) use the logistic regression objective function.




(a) $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$ (b) $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$ (c) $h(\mathbf{x}) = 1/2\|\mathbf{x}\|_2^2$ (d) $h(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$

*Figure 2.* Logarithmic averaged function value gap over 50 repetitions under different loss and distance generating functions. (a) and (b) use the ridge regression objective function on **E2006-tfidf** dataset; (c) and (d) use the logistic regression objective function on **a9a** dataset.

### 6.2. Experiments on Synthetic Data

For synthetic data, we set $n = 200, d = 100$. Each entry of the design matrix $\mathbf{A}$ is randomly generated in $[0, 1]$ and the regression vector $\mathbf{x}^*$ is randomly generated from $N(\mathbf{0}, \mathbf{I}_{d\times d})$. For the ridge regression problem, the response vector $\mathbf{y}$ is generated by $\mathbf{y} = \mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_{n\times n})$ is the noise vector. For the logistic regression problem, we generate $y_i$ according to $y_i = 1$ with probability $1/(1 + \exp(\mathbf{A}_{i*}\mathbf{x}^*))$ and $y_i = 0$ otherwise for all $i$. In our experiment, we use a mini-batch of sample to compute the stochastic gradient with batch size 15. The radius of Euclidean ball $\mathcal{X}$ is set to be $R = 12$. Step sizes of all algorithms are tuned by grid search. The logarithmic function value gap averaged over 50 repetitions are shown in Figure 1. In detail, Figures 1(a) and 1(b) are the results of squared Euclidean norm distance and negative entropy distance for ridge regression respectively. We can see that the performance of our algorithms ASMD and ASMD3 is comparable or even better than AC-SA and SAGE, and much better than SMD. Similar conclusion can be drawn from Figures 1(c) and 1(d) for logistic regression. This is well aligned with our theoretical analysis.

### 6.3. Experiments on Real Data

We also conduct experiments on real-world datasets to demonstrate the performance of our algorithms. In particular, we use **E2006-tfidf** dataset (Kogan et al., 2009) for the ridge regression problem, where $n = 16087, d = 50000$. We use **a9a** dataset (Chang & Lin, 2011) for the logistic regression problem, where $n = 32561, d = 123$. In the experiment, we set batch size $b = \lceil 0.2\%n \rceil$ for both regression and classification. Step sizes of all algorithms are tuned via grid search. The results are shown in Figure 2. Figures 2(a) and 2(b) show the results for ridge regression on **E2006-tfidf** and Figures 2(c) and 2(d) show the results for logistic classification on **a9a**. All the experiment results demonstrate that the proposed ASMD and ASMD3 achieve comparable convergence speed as AC-SA and are much faster than SMD. This is consistent with our theory as well.

## 7. Conclusions

We propose a novel SDE based dynamics to characterize accelerated stochastic mirror descent (ASMD) for strongly convex functions from a variational perspective. We design several new algorithms for ASMD via different Euler discretization schemes of the continuous-time dynamics. We provide simple and unified Lyapunov function based analysis to prove the convergence rates of both the continuous-time dynamics and the discrete-time algorithms. Thorough experiments corroborate our theory.

## 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. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Bailey, C. Hamilton's principle and the calculus of variations. *Acta Mechanica*, 44(1-2):49–57, 1982.

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

Bregman, L. M. 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(3): 200–217, 1967.

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

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

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

Chen, X., Lin, Q., and Pena, J. Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems*, pp. 395–403, 2012.

Chlebus, E. An approximate formula for a partial sum of the divergent p-series. *Applied Mathematics Letters*, 22 (5):732–737, 2009.

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

Flammarion, N. and Bach, F. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pp. 658–695, 2015.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

Güler, O. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

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

Hu, C., Pan, W., and Kwok, J. T. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, pp. 781–789, 2009.

Itô, K. Stochastic integral. *Proc. Imp. Acad.*, 20(8):519–524, 1944. doi: 10.3792/pia/1195572786. URL https://doi.org/10.3792/pia/1195572786.

Juditsky, A. Convex optimization ii: Algorithms. *Lecture notes*, 2013.

Kloeden, P. E. and Platen, E. Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics*, 66(1):283–314, 1992.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280. Association for Computational Linguistics, 2009.

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

Krichene, W., Bayen, A., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pp. 2845–2853, 2015.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

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

Lessard, L., Recht, B., and Packard, A. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3384–3392, 2015.

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

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

Nemirovski, A. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979.

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

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pp. 372–376, 1983.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

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

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

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

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

Raginsky, M. and Bouvrie, J. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 6793–6800. IEEE, 2012.

Scieur, D., Roulet, V., Bach, F., and d'Aspremont, A. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, pp. 1109–1118, 2017.

Shalev-Shwartz, S. and Kakade, S. M. Mind the duality gap: Logarithmic regret algorithms for online optimization. In *Advances in Neural Information Processing Systems*, pp. 1457–1464, 2009.

Su, W., Boyd, S., and Candes, E. A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pp. 2510–2518, 2014.

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

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

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

Xu, P., Wang, T., and Gu, Q. Accelerated stochastic mirror descent: From continuous-time dynamics to discrete-time algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1087–1096, 2018.