Two Facets of Stochastic Optimization: Continuous-time Dynamics and Discrete-time Algorithms

Quanquan Gu

Department of Computer Science
University of California, Los Angeles

Joint work with Pan Xu and Tianhao Wang

ACC Workshop on Interplay between Control, Optimization, and Machine Learning

Optimization for Machine Learning

- Many machine learning methods can be formulated as an optimization problem

$$\min_{x \in \mathcal{X}} f(x)$$

  - $f : \mathbb{R}^d \to \mathbb{R}$ is a (strongly) convex function
  - $\mathcal{X} \subseteq \mathbb{R}^d$ is a constrained set

- Stochastic optimization plays a central role in large-scale machine learning
  - stochastic gradient descent
  - stochastic mirror descent
  - stochastic Langevin gradient descent
  - accelerated variants
  - …

Outline

- Stochastic Mirror Descent
- Understanding Acceleration in Optimization
- Continuous-time Dynamics for Accelerated Stochastic Mirror Descent
- Discretization of SDEs and New ASMD Algorithms
- Experiments

Stochastic Gradient Descent

SGD update:

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k - \eta_k G(x_k; \xi_k))$$

step size

stochastic gradient

Unbiased estimator of the gradient:

$$\mathbb{E}_{\xi_k}[G(x_k; \xi_k)] = \nabla f(x_k)$$

Convergence rate

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{G}{\sqrt{k}}\right)$$

convex & bounded gradient

bounded gradient: $\|G(x;\xi)\|_2 \leq G$

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{G^2 \log k}{\mu k}\right)$$

strongly convex & bounded gradient

μ**-strongly convex:** $f(x) \geq f(y) + \langle \nabla f(x), x - y \rangle + \mu/2\|x - y\|_2^2$

From Euclidean Space to Non-Euclidean Space: Stochastic Mirror Descent

Bregman divergence $D_h(x, z) := h(z) - h(x) - \langle \nabla h(x), z - x \rangle$

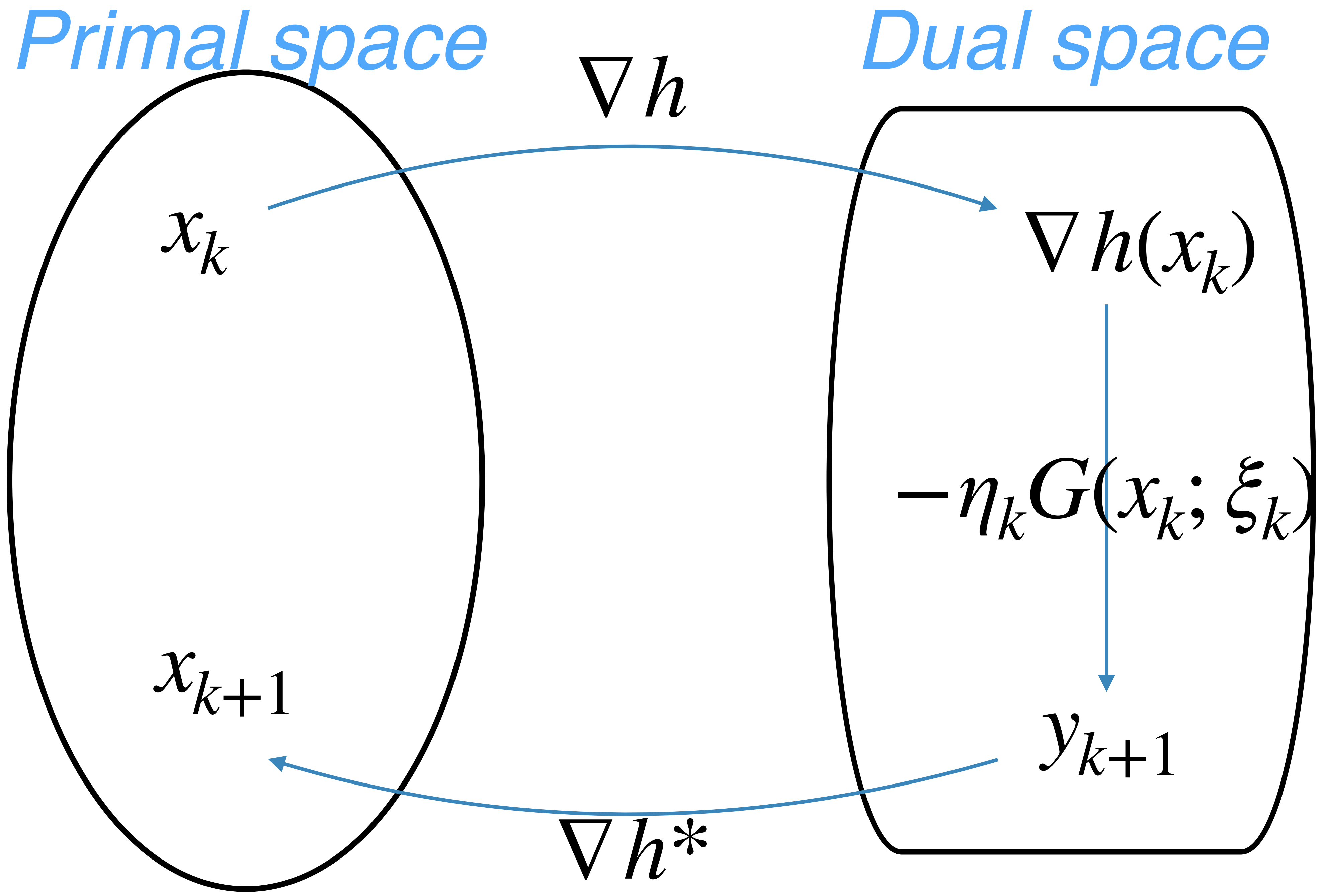
strongly convex distance generating function

Stochastic Mirror Descent (SMD) update:

$$y_{k+1} = \nabla h(x_k) - \eta_k G(x_k; \xi_k)$$

Descent method in the dual space

$$x_{k+1} = \nabla h^*(y_{k+1})$$



$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{G}{\sqrt{k}}\right)$$

convex & bounded gradient

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{G^2 \log k}{\mu k}\right)$$

strongly convex & bounded gradient

Accelerated Stochastic Mirror Descent

ASMD update [Lan, 2012; Saeed & Lan, 2012]

$$x_k^{md} = \beta_k^{-1} x_k + (1 - \beta_k^{-1}) x_k^{md}$$

$$x_{k+1} = \nabla h^*(\nabla h(x_k) - \gamma_k \nabla G(x_k^{md}; \xi_k))$$

$$x_{k+1}^{ag} = \beta_k^{-1} x_{k+1} + (1 - \beta_k^{-1}) x_k^{ag}$$

Hard to Interpret!

Convergence rate

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{L}{k^2} + \frac{\sigma}{\sqrt{k}}\right)$$ **convex & bounded gradient**

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{L}{k^2} + \frac{\sigma^2}{\mu k}\right)$$ **strongly convex & bounded gradient**

$$\mathbb{E}[\|G(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2^2] \leq \sigma^2$$

Smooth: $f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + L/2 \|\mathbf{x} - \mathbf{y}\|_2^2$

when $\sigma = 0$, it matches the optimal rate of deterministic mirror descent

We Want to …

- Better understand accelerated stochastic mirror descent
- Derive intuitive and simple accelerated stochastic mirror descent algorithms
- Deliver simple proof of the convergence rates

Outline

- Stochastic Mirror Descent
- Understanding Acceleration in Optimization
- Continuous-time Dynamics for Accelerated Stochastic Mirror Descent
- Discretization of SDEs and New ASMD Algorithms
- Experiments

Interpretations of Nesterov's AGD/AMD

- Ordinary Differential Equation interpretation
[Su et al, 2014] [Krichene et al, 2015] [Wibisono et al, 2016] [Wilson et al, 2016] [Diakonikolas & Orecchia, 2018]

- Other interpretations
    - Linear Matrix Inequality [Lessard et al, 2016]
    - Dissipativity Theory [Hu & Lessard, 2017]
    - Linear Coupling [Allen-Zhu & Orecchia, 2017]
    - Geometry [Bubeck et al, 2015]
    - Game theory [Lan & Zhou, 2018]

From ODE to SDE

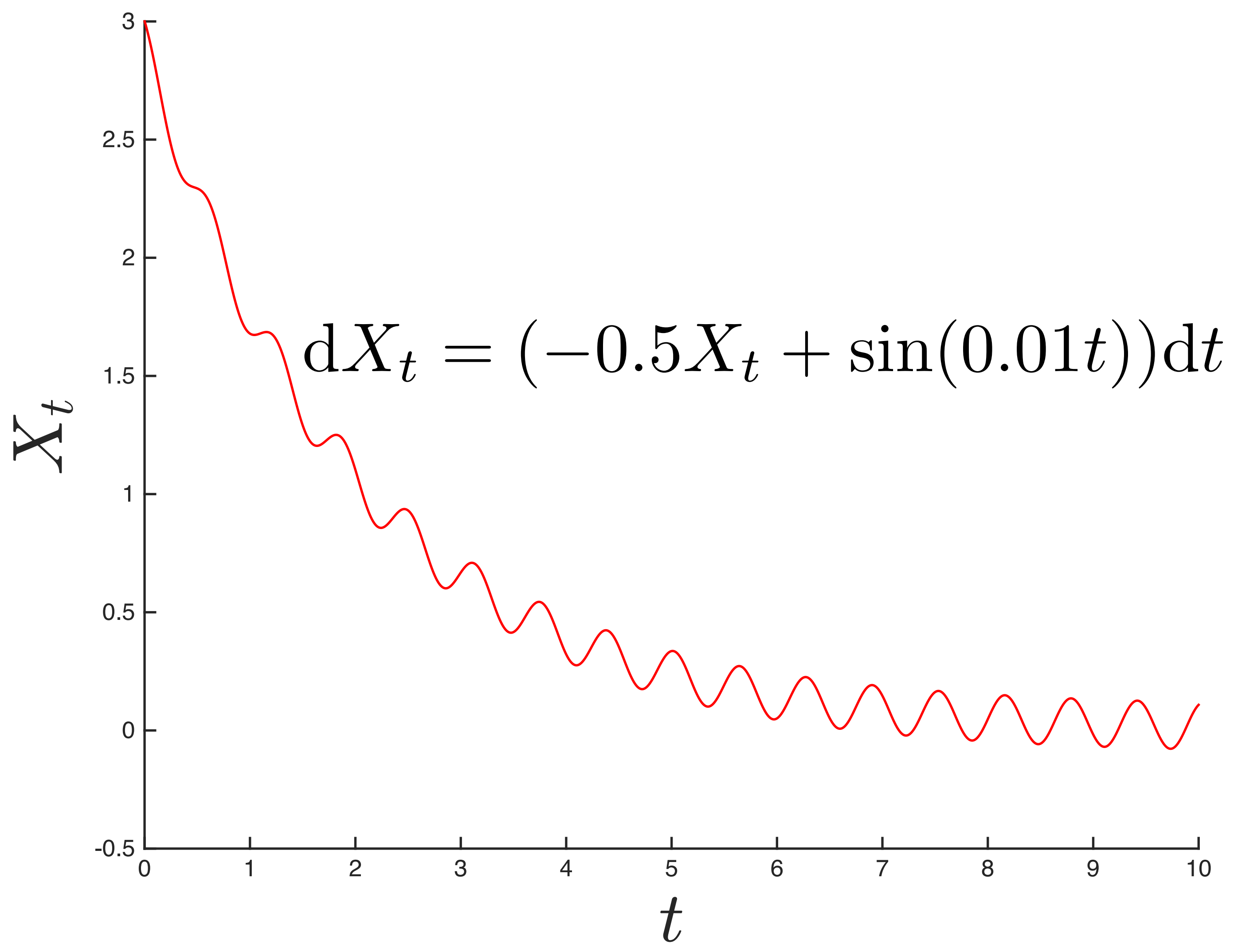
Ordinary Differential Equation

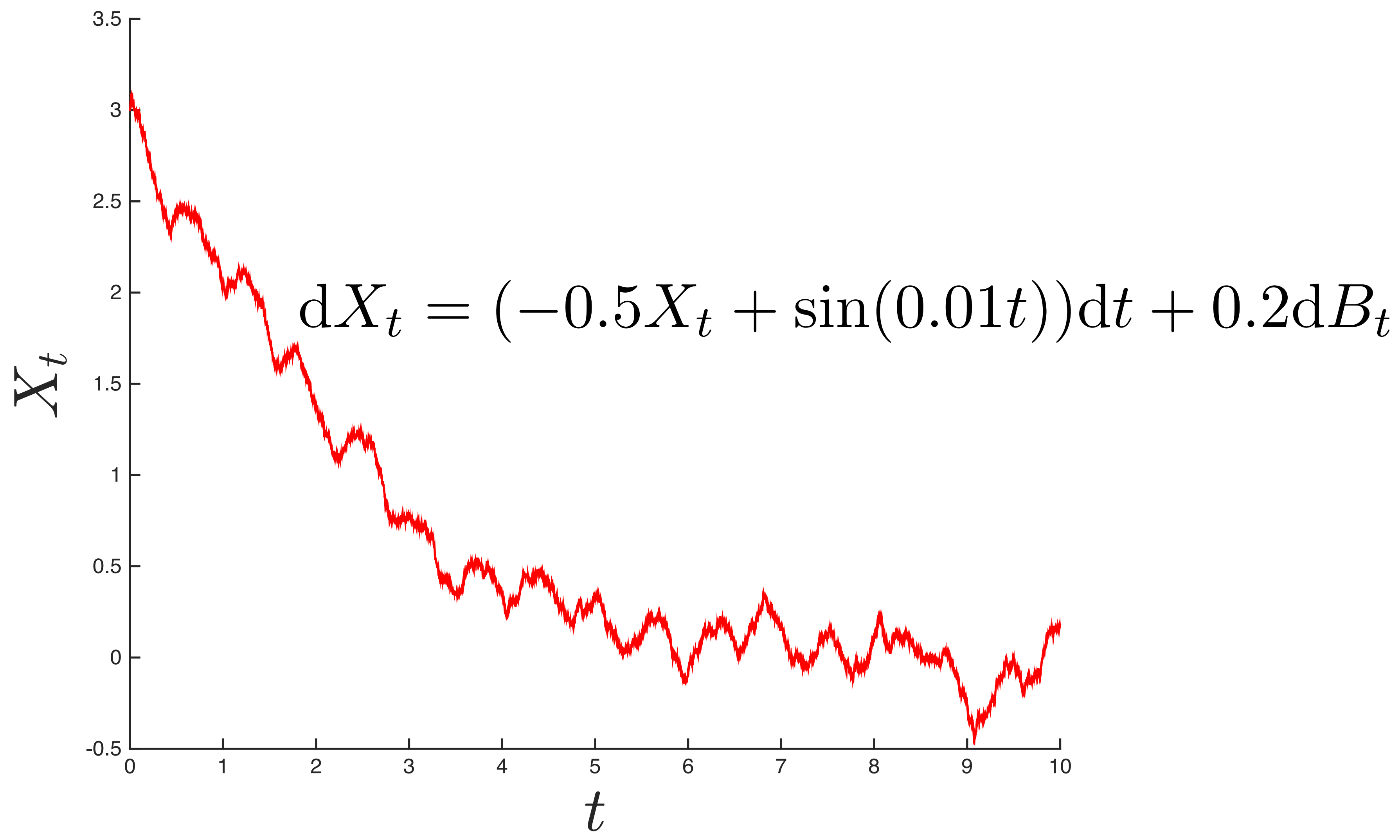
$$\mathrm{d}\boldsymbol{X}_t = u(\boldsymbol{X}_t, t)\mathrm{d}t$$

Stochastic Differential Equation

Brownian motion

$$\mathrm{d}\boldsymbol{X}_t = u(\boldsymbol{X}_t, t)\mathrm{d}t + \sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t$$





SDE Interpretations of Stochastic Optimization

Stochastic Gradient Descent

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

↔

Stochastic Gradient Flow

$$\mathrm{d}\boldsymbol{X}_t = -\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sigma \mathrm{d}\boldsymbol{B}_t$$

Stochastic Mirror Descent

$$\mathbf{y}_{k+1} = \nabla h(\mathbf{x}_k) - \eta_k \nabla \widetilde{f}(\mathbf{x}_k, \xi_k)$$

$$\mathbf{x}_{k+1} = \nabla h^*(\mathbf{y}_{k+1})$$

↔

Stochastic Mirror Flow

[Raginsky & Bouvrie, 2012]

[Metrikopoulos & Staudigl, 2016]

$$\mathrm{d}\nabla h(\boldsymbol{X}_t) = -\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sigma \mathrm{d}\boldsymbol{B}_t$$

Accelerated Stochastic Mirror Descent

$$\mathbf{y}_{k+1} = \nabla h(\mathbf{x}_k) - \eta_k \nabla \widetilde{f}(\mathbf{x}_k, \xi_k)$$

$$\mathbf{x}_{k+1} = \alpha_k \nabla h^*(\mathbf{y}_{k+1}) + (1 - \alpha_k)\mathbf{x}_k$$

↔

Accelerated Stochastic Mirror Flow

[Krichene & Bartlett, 2017]

$$\mathrm{d}\boldsymbol{Z}_t = -\eta_t[\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{B}_t]$$

$$\mathrm{d}\boldsymbol{X}_t = a_t[\nabla h^*(\boldsymbol{Z}_t/s_t) - \boldsymbol{X}_t]\mathrm{d}t,$$

Yet ...

	Convex	Strongly Convex	Acceleration	Converge	Discrete-time Algorithm
(Raginsky & Bouvrie, 2012)	✔				
(Mertikopoulos and Staudigl, 2016)	✔	✔		✔	
(Krichene & Bartlett, 2017)	✔		✔	✔	

Yet …

	Convex	Strongly Convex	Acceleration	Converge	Discrete-time Algorithm
(Raginsky & Bouvrie, 2012)	✓				
(Mertikopoulos and Staudigl, 2016)	✓	✓		✓	
(Krichene & Bartlett, 2017)	✓		✓	✓	
This talk	✓	✓	✓	✓	✓

Outline

- Stochastic Mirror Descent
- Understanding Acceleration in Optimization
- Continuous-time Dynamics for Accelerated Stochastic Mirror Descent
- Discretization of SDEs and New ASMD Algorithms
- Experiments

Lagrangian Mechanics Behind Optimization

- Optimization: Mechanical/Physical system with friction

- Undamped Lagrangian $\mathscr{L}(X, V, t) = \frac{1}{2}||V||^2 - f(X)$

kinetic energy   potential energy

- **Principle of Least Action**: real-world motion X_t minimize

$$J(X) = \int_{\mathbb{T}} \mathscr{L}(X_t, \dot{X}_t, t)dt$$

- Euler-Lagrange equation

$$\frac{d}{dt}\left\{\frac{\partial \mathscr{L}}{\partial \dot{X}_t}(X_t, \dot{X}_t, t)\right\} = \frac{\partial \mathscr{L}}{\partial X_t}(X_t, \dot{X}_t, t)$$

Bregman Lagrangian for Mirror Descent: General Convex Functions

- Damped Lagrangian

$$\mathcal{L}(X, V, t) = e^{\gamma_t}\left(\frac{1}{2}||V||^2 - f(X)\right)$$

- Solution to Euler-Lagrangian equation

$$\ddot{X}_t + \dot{\gamma}_t + \nabla f(X_t) = 0$$

- Damped Bregman Lagrangian [Wibisono et al, 2016]

$$\mathcal{L}(X, V, t) = e^{\alpha_t + \gamma_t}\left(D_h(X + e^{-\alpha_t}V, X) - e^{\beta_t}f(X)\right)$$

Continuous-time Dynamics of MD: General Convex Functions

By Euler-Lagrange equation, choosing $e^{\alpha_t} = \dot{\beta}_t, \quad \gamma_t = \beta_t$

$$\frac{d}{dt}\nabla h(X_t + 1/\dot{\beta}_t \dot{X}_t) = -\dot{\beta}_t e^{\beta_t} \nabla f(X_t)$$

a second order ODE

Define $Y_t = \nabla h(X_t + 1/\dot{\beta}_t \dot{X}_t)$ **and rewrite the ODE**

$$\begin{cases} dX_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)dt \\ dY_t = -\dot{\beta}_t e^{\beta_t} \nabla f(X_t) dt \end{cases}$$

continuous-time dynamics of AMD

[Wibisono et al, 2016]

Continuous-time Dynamics of SMD: General Convex Functions

Add a Brownian motion?

$$\begin{cases} dX_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)dt \\ dY_t = -\dot{\beta}_t e^{\beta_t} \nabla f(X_t)dt + \sqrt{\delta}\sigma(X_t, t)dB_t \end{cases}$$

Does not converge

Brownian motion: $\sqrt{\delta}\sigma(X_t, t)dB_t$

So we introduce an extra shrinkage parameter

$$\begin{cases} dX_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)dt \\ dY_t = -\frac{\dot{\beta}_t e^{\beta_t}}{s_t}(\nabla f(X_t)dt + \sqrt{\delta}\sigma(X_t, t)dB_t) \end{cases}$$

shrinkage parameter: s_t

This is the continuous-time dynamics of accelerated SMD for general convex function

Convergence Rate of Continuous-time Dynamics: General Convex Functions

Stochastic differential equation (SDE):

$$\begin{cases} dX_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)dt \\ dY_t = -\frac{\dot{\beta}_t e^{\beta_t}}{s_t}\left(\nabla f(X_t)dt + \sqrt{\delta}\sigma(X_t, t)dB_t\right) \end{cases}$$

drift term: $\nabla f(X_t)dt$

diffusion term: $\sqrt{\delta}\sigma(X_t, t)dB_t$

- $t > 0$ is time index
- δ, β_t, s_t are scaling parameter
- $B_t \in \mathbb{R}^d$ is the standard Brownian motion

Convergence of the proposed SDE:

$$\mathbb{E}[f(X_t) - f(x^*)] = O\left(\frac{1}{t^2} + \frac{\sigma^2}{t^{1/2-q}}\right)$$

- diffusion term $\|\sigma(X_t, t)\|_2 \le \sigma t^q$
- optimal convergence rate of ASMD $O\left(\frac{1}{k^2} + \frac{\sigma^2}{\sqrt{k}}\right)$

when $q = 0$, it matches optimal rate for stochastic mirror descent for general convex functions [Lan, 2012; Saeed & Lan, 2012]

Roadmap of the Proof

- **Lyapunov Function**

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

- **Step 1: bounding** $\mathrm{d}\mathcal{E}_t$

Rewrite the stochastic dynamics as the following SDE

$$\mathrm{d}\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{bmatrix} = \begin{bmatrix}\dot{\beta}_t(\nabla h^*(\boldsymbol{Y}_t) - \boldsymbol{X}_t)\\ -\dot{\beta}_t e^{\beta_t}\nabla f(\boldsymbol{X}_t)/s_t\end{bmatrix}\mathrm{d}t + \begin{bmatrix}\mathbf{0}\\ -\dot{\beta}_t e^{\beta_t}\sqrt{\delta}\sigma(\boldsymbol{X}_t, t)/s_t\end{bmatrix}\mathrm{d}\boldsymbol{B}_t.$$

Applying Itô's Lemma to $\mathcal{E}_t$ with respect to the above SDE yields

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

Roadmap of the Proof

- **Step 2: integrating and taking expectation**

$$\mathbb{E}[\mathcal{E}_t] \le \mathcal{E}_0 + (s_t - s_0)M_{h,\mathcal{X}} + \frac{1}{2}\mathbb{E}\left[\int_0^t \frac{\dot\beta_r^2 e^{2\beta_r}}{s_r}\mathrm{tr}\big(\sigma_r^\top \nabla^2 h^*(\mathbf{Y}_r)\sigma_r\big)\mathrm{d}r\right],$$

where $M_{h,\mathcal{X}}$ is the diameter of $\mathcal{X}$

- **Step 3: choosing parameters**

Plugging in parameters: $\beta_t = 2\log t$ and $s_t = t^{3/2+q}$

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] \le \frac{\mathbb{E}[\mathcal{E}_t]}{e^{\beta_t}} = O\left(\frac{1}{t^2} + \frac{1}{t^{1/2-q}}\right).$$

Bregman Lagrangian for Mirror Descent: Strongly Convex Functions

- Damped Lagrangian

$$\mathscr{L}(X, V, t) = e^{\gamma_t}\left(\frac{1}{2}||V||^2 - f(X)\right)$$

- Solution to Euler-Lagrangian equation

$$\ddot{X}_t + \dot{\gamma}_t + \nabla f(X_t) = 0$$

- Damped Bregman Lagrangian [Xu et al., 2018]

$$\mathscr{L}(X, V, t) = e^{\alpha_t + \beta_t + \gamma_t}\left(\mu D_h(X + e^{-\alpha_t}V, X) - f(X)\right)$$

Continuous-time Dynamics of MD: Strongly Convex Functions

By Euler-Lagrange equation, choosing $e^{\alpha_t} = \dot{\beta}_t, \quad \dot{\gamma}_t = -e^{\alpha_t}$

$$\frac{d}{dt}\nabla h(X_t + 1/\dot{\beta}_t \dot{X}_t) = -\dot{\beta}_t e^{\beta_t}(\nabla f(X_t)/\mu + \nabla h(X_t + 1/\dot{\beta}_t \dot{X}_t) - \nabla h(X_t))$$

a second order ODE

Define $Y_t = \nabla h(X_t + 1/\dot{\beta}_t \dot{X}_t)$ **and add Brownian motion to the ODE**

$$\begin{cases} dX_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)dt \\ \mathbf{d}Y_t = -\dot{\beta}_t\left(\frac{1}{\mu}\nabla f(X_t)\mathbf{d}t + (Y_t - \nabla h(X_t))\mathbf{d}t + \frac{\sqrt{\delta}\sigma(X_t, t)}{\mu}\mathbf{d}B_t\right) \end{cases}$$

This is the continuous-time dynamics of accelerated SMD for strongly convex function [Xu et al., 2018]

Convergence Rate of Continuous-time Dynamics: Strongly Convex Functions

Stochastic differential equation (SDE):

$$\begin{cases} dX_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)dt \\ \mathbf{d}Y_t = -\dot{\beta}_t\left(\underbrace{\frac{1}{\mu}\nabla f(X_t)\mathbf{d}t + (Y_t - \nabla h(X_t))\mathbf{d}t}_{\text{drift term}} + \underbrace{\frac{\sqrt{\delta}\sigma(X_t, t)}{\mu}\mathbf{d}B_t}_{\text{diffusion term}}\right) \end{cases}$$

- $t > 0$ is time index
- δ, β_t are scaling parameter
- $B_t \in \mathbb{R}^d$ is the standard Brownian motion

Convergence of the proposed SDE:

$$\mathbb{E}[f(X_t) - f(x^*)] = O\left(\frac{1}{t^2} + \frac{\sigma^2}{\mu t^{1-2q}}\right)$$

- **diffusion term** $\|\sigma(X_t, t)\|_2 \leq \sigma t^q$
- **optimal convergence rate of ASMD** $O\left(\frac{1}{k^2} + \frac{\sigma^2}{\mu k}\right)$

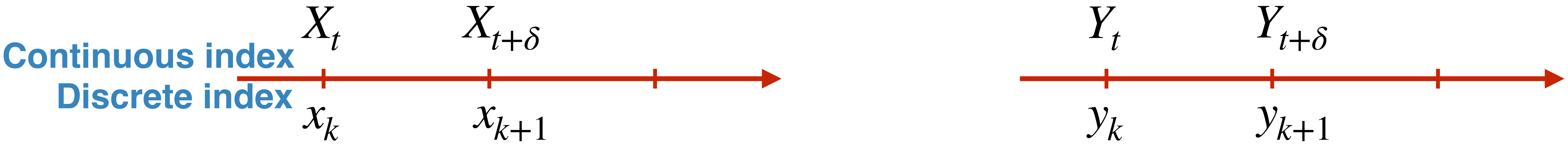
when $q = 0$, it matches optimal rate for stochastic mirror descent for general convex functions [Lan, 2012; Saeed & Lan, 2012]

Outline

- Stochastic Mirror Descent
- Understanding Acceleration in Optimization
- Continuous-time Dynamics for Accelerated Stochastic Mirror Descent
- Discretization of SDEs and New ASMD Algorithms
- Experiments

Discretization of SDE

Continuous-Time to Discrete-time sequence



Forward (Explicit) Euler Discretization

$$\frac{x_{k+1} - x_k}{\delta} \approx \dot{X}_t \qquad \frac{y_{k+1} - y_k}{\delta} \approx \dot{Y}_t$$

Backward (Implicit) Euler Discretization

$$\frac{x_k - x_{k-1}}{\delta} \approx \dot{X}_t \qquad \frac{y_k - y_{k-1}}{\delta} \approx \dot{Y}_t$$

New Discrete-time Algorithm (Implicit)

SDEs for general convex functions:

$$\mathrm{d}X_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)\mathrm{d}t$$
$$\mathrm{d}Y_t = -\frac{\dot{\beta}_t e^{\beta_t}}{s_t}(\nabla f(X_t)\mathrm{d}t + \sqrt{\delta}\sigma(X_t, t)\mathrm{d}B_t)$$

Implicit discretization

$$y_{k+1} - y_k = -\tau_k / s_k G(x_{k+1}; \xi_{k+1})$$
$$\nabla h^*(y_{k+1}) = x_{k+1} + 1/\tau_k(x_{k+1} - x_k)$$

Convergence rate

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{1}{k^2} + \frac{\sigma}{\sqrt{k}}\right)$$

- **Optimal rate** [Ghadimi & Lan, 2012] 👍
- **Implicit update** 👎

New Discrete-time Algorithm (ASMD)

SDEs for general convex functions:

$$\mathrm{d}X_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)\mathrm{d}t$$

$$\mathrm{d}Y_t = -\frac{\dot{\beta}_t e^{\beta_t}}{s_t}(\nabla f(X_t)\mathrm{d}t + \sqrt{\delta}\sigma(X_t, t)\mathrm{d}B_t)$$

Hybrid discretization

$$\nabla h^*(y_k) = x_{k+1} + 1/\tau_k(x_{k+1} - x_k)$$

$$y_{k+1} - y_k = -\tau_k/s_k G(x_{k+1}; \xi_{k+1})$$

Convergence rate

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{1}{k^2} + \frac{\sigma^2 + 1}{\sqrt{k}}\right)$$

- **Not optimal rate** [Ghadimi & Lan, 2012]
- **Explicit (practical) algorithm**

New Discrete-time Algorithm (ASMD3)

SDEs for general convex functions:

$$
\begin{aligned}
\mathrm{d}X_t &= \dot{\beta}_t(\nabla h^*(Y_t) - X_t)\mathrm{d}t \\
\mathrm{d}Y_t &= -\frac{\dot{\beta}_t e^{\beta_t}}{s_t}(\nabla f(X_t)\mathrm{d}t + \sqrt{\delta}\sigma(X_t, t)\mathrm{d}B_t)
\end{aligned}
$$

Explicit discretization with additional sequence

$$
\begin{aligned}
\nabla h^*(y_k) &= x_k + 1/\tau_k(z_{k+1} - x_k) \\
y_{k+1} - y_k &= -\tau_k/s_k G(z_{k+1}; \xi_{k+1}) \\
x_{k+1} &= \arg\min_{x \in \mathcal{X}} \{\langle G(z_{k+1}; \xi_{k+1}), x\rangle + M_k D_h(z_{k+1}, x)\}
\end{aligned}
$$

Convergence rate

$$
\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{1}{k^2} + \frac{\sigma^2}{\sqrt{k}}\right)
$$

- **Optimal rate** [Ghadimi & Lan, 2012] 👍
- **Explicit (practical) algorithm** 👍

New Discrete-time Algorithm (Implicit)

SDEs for strongly convex functions:

$$\mathrm{d}X_t = \dot{\beta}_t(\nabla h^*(Y_t) - X_t)\mathrm{d}t$$
$$\mathrm{d}Y_t = -\dot{\beta}_t\left(\frac{1}{\mu}\nabla f(X_t)\mathrm{d}t + (Y_t - \nabla h(X_t))\mathrm{d}t + \frac{\sqrt{\delta}\sigma(X_t, t)}{\mu}\mathrm{d}B_t\right)$$

Implicit discretization

$$y_{k+1} - y_k = -\tau_k\big(G(x_{k+1}; \xi_{k+1})/\mu + y_{k+1} - \nabla h(x_{k+1})\big)$$
$$\nabla h^*(y_{k+1}) = x_{k+1} + 1/\tau_k(x_{k+1} - x_k)$$

Convergence rate

$$\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{1}{k^2} + \frac{\sigma^2}{\mu k}\right)$$

- **Optimal rate** [Ghadimi & Lan, 2012] 👍
- **Implicit update** 👎

New Discrete-time Algorithm (ASMD)

SDEs for strongly convex functions:

$$
\begin{aligned}
\mathrm{d}X_t &= \dot{\beta}_t(\nabla h^*(Y_t) - X_t)\mathrm{d}t \\
\mathrm{d}Y_t &= -\dot{\beta}_t\left(\frac{1}{\mu}\nabla f(X_t)\mathrm{d}t + (Y_t - \nabla h(X_t))\mathrm{d}t + \frac{\sqrt{\delta}\sigma(X_t, t)}{\mu}\mathrm{d}B_t\right)
\end{aligned}
$$

Hybrid discretization

$$
\begin{aligned}
\nabla h^*(y_k) &= x_{k+1} + 1/\tau_k(x_{k+1} - x_k) \\
y_{k+1} - y_k &= -\tau_k\Big(G(x_{k+1}; \xi_{k+1})/\mu + y_{k+1} - \nabla h(x_{k+1})\Big)
\end{aligned}
$$

Convergence rate

$$
\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{1}{k^2} + \frac{\sigma^2 + 1}{\mu k}\right)
$$

- **Not optimal rate** [Ghadimi & Lan, 2012] 👎
- **Explicit (practical) algorithm** 👍

New Discrete-time Algorithm (ASMD3)

SDEs for strongly convex functions:

$$
\begin{aligned}
\mathrm{d}X_t &= \dot{\beta}_t(\nabla h^*(Y_t) - X_t)\mathrm{d}t \\
\mathrm{d}Y_t &= -\dot{\beta}_t\left(\frac{1}{\mu}\nabla f(X_t)\mathrm{d}t + (Y_t - \nabla h(X_t))\mathrm{d}t + \frac{\sqrt{\delta}\sigma(X_t, t)}{\mu}\mathrm{d}B_t\right)
\end{aligned}
$$

Explicit discretization with additional sequence

$$
\begin{aligned}
\nabla h^*(y_k) &= x_k + 1/\tau_k(z_{k+1} - x_k) \\
y_{k+1} - y_k &= -\tau_k\big(G(z_{k+1};\xi_{k+1})/\mu + y_k - \nabla h(z_{k+1})\big) \\
x_{k+1} &= \arg\min_{x\in\mathcal{X}}\{\langle G(z_{k+1};\xi_{k+1}), x\rangle + M_k D_h(z_{k+1}, x)\}
\end{aligned}
$$

Convergence rate

$$
\mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{1}{k^2} + \frac{\sigma^2}{\mu k}\right)
$$

- **Optimal rate** [Ghadimi & Lan, 2012]
- **Explicit (practical) algorithm**

Outline

- Stochastic Mirror Descent
- Understanding Acceleration in Optimization
- Continuous-time Dynamics for Accelerated Stochastic Mirror Descent
- Discretization of SDEs and New ASMD Algorithms
- Experiments

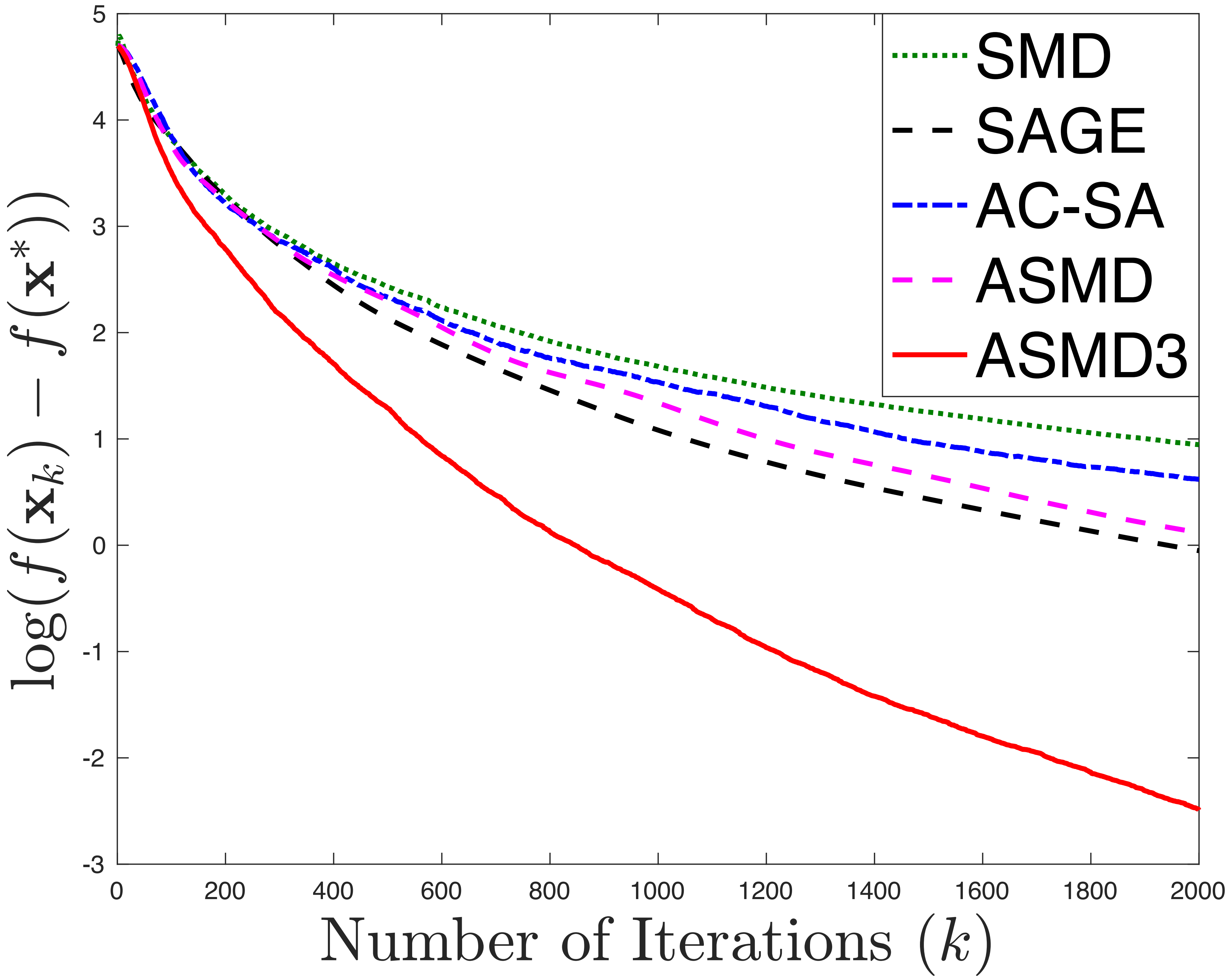
Experiment Results: General Convex Case

Baselines: **SMD, SAGE [Hu et al., 2009], AC-SA [Ghadimi & Lan, 2012]**

Optimization problem: $$\min_{x \in \mathcal{X}} \frac{1}{2n} \|Ax - y\|_2^2$$

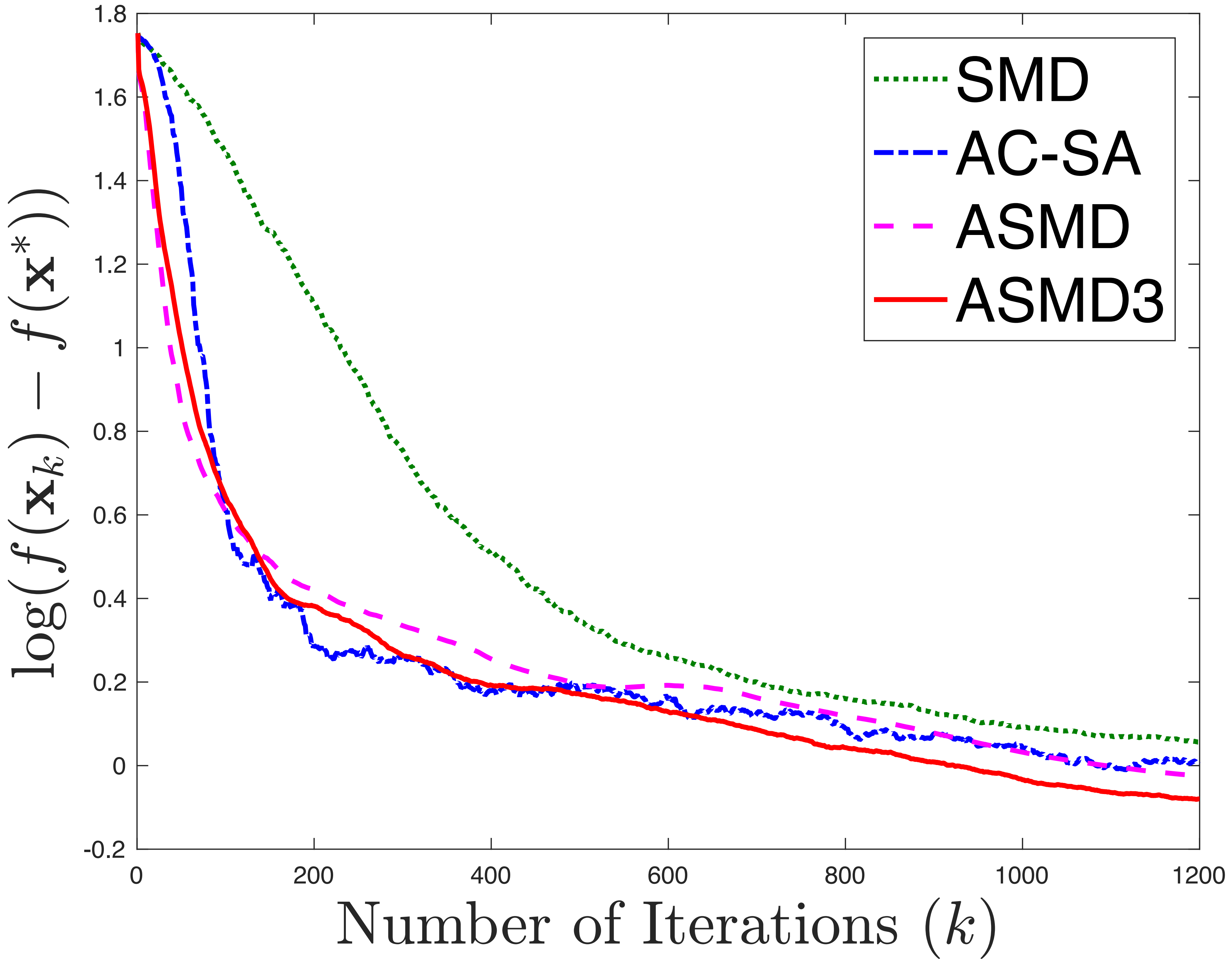
- **constrain set:** $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$
- **distance generating function:**

$$h(x) = \frac{1}{2}\|x\|_2^2$$



- **constrain set:** $\mathcal{X} = \left\{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\right\}$
- **distance generating function:**

$$h(x) = \sum_{i=1}^d x_i \log x_i$$



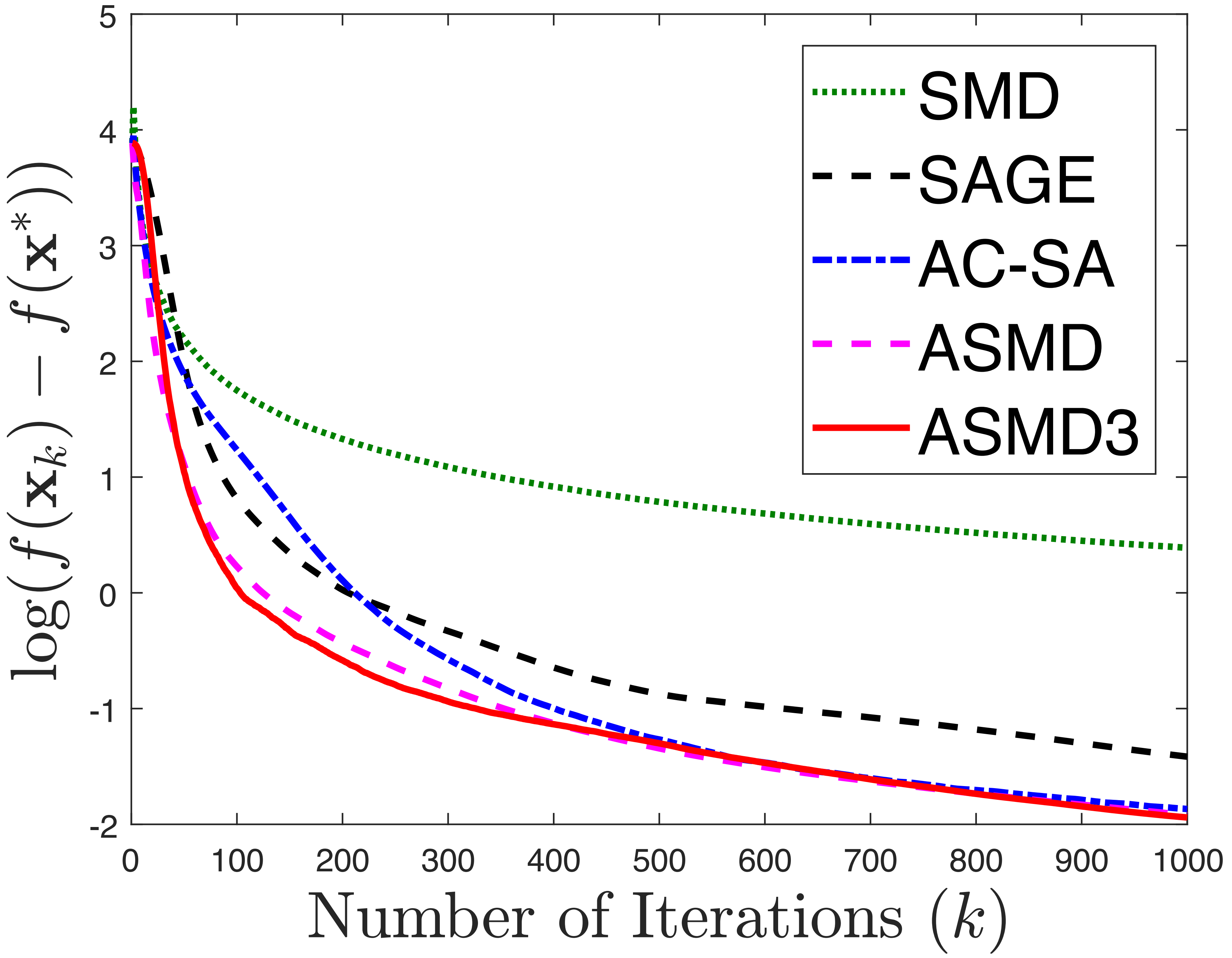
Experiment Results: Strongly Convex Case

Baselines: **SMD, SAGE [Hu et al., 2009], AC-SA [Ghadimi & Lan, 2012]**

Optimization problem: $$\min_{x \in \mathcal{X}} \frac{1}{2n} \|Ax - y\|_2^2 + \lambda \|x\|_2^2$$

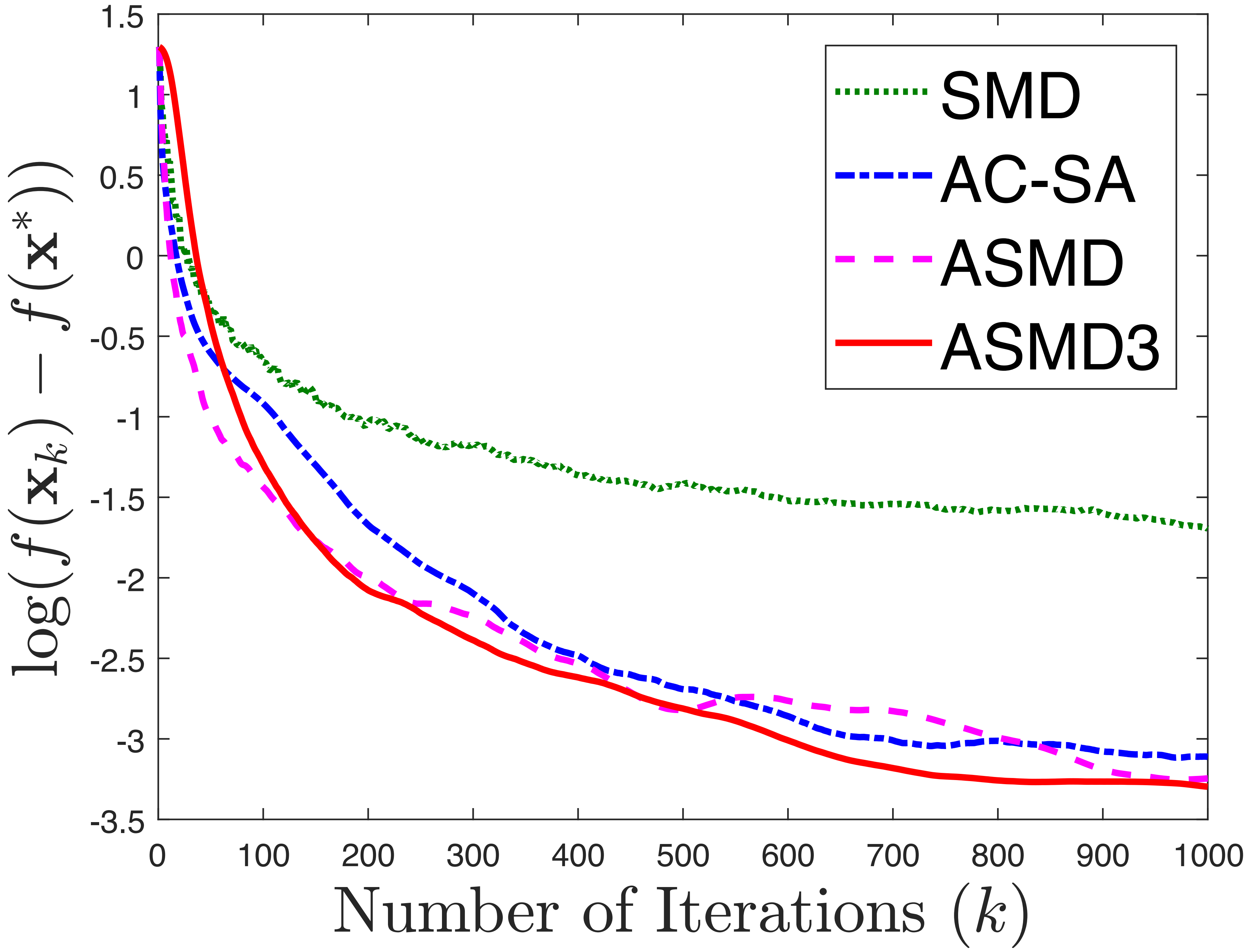
- **constrain set:** $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$
- **distance generating function:**

$$h(x) = \frac{1}{2}\|x\|_2^2$$



- **constrain set:** $\mathcal{X} = \left\{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\right\}$
- **distance generating function:**

$$h(x) = \sum_{i=1}^d x_i \log x_i$$



Take Away

Continuous-time dynamics can help us

- better understand stochastic optimization
- derive new discrete-time algorithms based on various discretization schemes
- deliver a unified and simple proof of convergence rates

Thank You

Reference

- Allen-Zhu, Z., & Orecchia, L. (2017). Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

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

- Diakonikolas J, Orecchia L. Accelerated Extra-Gradient Descent: A Novel Accelerated First-Order Method. In9th Innovations in Theoretical Computer Science Conference (ITCS 2018) 2018 Jan (Vol. 94).

- 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.

- Hu, B., & Lessard, L. (2017, August). Dissipativity theory for Nesterov's accelerated method. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 1549-1557). JMLR. org.

- 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.

Reference

- Diakonikolas, Jelena, and Lorenzo Orecchia. "Accelerated Extra-Gradient Descent: A Novel Accelerated First-Order Method." 9th Innovations in Theoretical Computer Science Conference (ITCS 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

- 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.

- Krichene, W. and Bartlett, P.L., 2017. Acceleration and averaging in stochastic descent dynamics. In Advances in Neural Information Processing Systems (pp. 6796-6806).

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

- Lan, G. and Zhou, Y., 2018. An optimal randomized incremental gradient method. Mathematical programming, 171(1-2), pp.167-215.

- 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.

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

Reference

- 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.

- 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.

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

- Xu, P., Wang, T. and Gu, Q., 2018, July. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In International Conference on Machine Learning (pp. 5488-5497).