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:

\[ x_{k+1} = \Pi_x (x_k - \eta_k G(x_k; \xi_k)) \]

step size \hspace{0.5cm} stochastic gradient

Unbiased estimator of the gradient:

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

Convergence rate

\[ \mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{G}{\sqrt{k}}\right) \quad \text{convex & bounded gradient} \]

\[ \mathbb{E}[f(x_k) - f(x^*)] = O\left(\frac{G^2 \log k}{\mu k}\right) \quad \mu\text{-strongly convex & bounded gradient} \]

\[ \text{bounded gradient: } \|G(x; \xi)\|_2 \leq G \]

\[ \mu\text{-strongly convex: } f(x) \geq f(y) + \langle \nabla f(x), x - y \rangle + \mu/2 \|x - y\|_2^2 \]
From Euclidean Space to Non-Euclidean Space: Stochastic Mirror Descent

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

**Stochastic Mirror Descent (SMD) update:**
\[ y_{k+1} = \nabla h(x_k) - \eta_k G(x_k; \xi_k) \]
\[ x_{k+1} = \nabla h^*(y_{k+1}) \]

Descent method in the dual space

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

**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]**

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

\[
x_{k+1} = \nabla h^*(\nabla h(x_k) + \gamma \nabla G(x_k^{md}, \xi_k))
\]

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

**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

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

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

Hard to Interpret!
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

$$dX_t = u(X_t, t)dt$$

Stochastic Differential Equation

$$dX_t = u(X_t, t)dt + \sigma(X_t, t)dB_t$$

Brownian motion

Ordinary Differential Equation

$$dX_t = (-0.5X_t + \sin(0.01t))dt$$

Stochastic Differential Equation

$$dX_t = (-0.5X_t + \sin(0.01t))dt + 0.2dB_t$$
SDE Interpretations of Stochastic Optimization

**Stochastic Gradient Descent**

\[ x_{k+1} = x_k - \eta_k \nabla \tilde{f}(x_k, \xi_k) \]

**Stochastic Gradient Flow**

\[ dX_t = -\nabla f(X_t)dt + \sigma dB_t \]

**Stochastic Mirror Descent**

\[ y_{k+1} = \nabla h(x_k) - \eta_k \nabla \tilde{f}(x_k, \xi_k) \]
\[ x_{k+1} = \nabla h^*(y_{k+1}) \]

**Stochastic Mirror Flow**

\[ d\nabla h(X_t) = -\nabla f(X_t)dt + \sigma dB_t \]

**Accelerated Stochastic Mirror Descent**

\[ y_{k+1} = \nabla h(x_k) - \eta_k \nabla \tilde{f}(x_k, \xi_k) \]
\[ x_{k+1} = \alpha_k \nabla h^*(y_{k+1}) + (1 - \alpha_k)x_k \]

**Accelerated Stochastic Mirror Flow**

\[ dZ_t = -\eta_t[\nabla f(X_t)dt + \sigma(X_t, t)dB_t] \]
\[ dX_t = a_t[\nabla h^*(Z_t/s_t) - X_t]dt, \]
Yet ...

<table>
<thead>
<tr>
<th></th>
<th>Convex</th>
<th>Strongly Convex</th>
<th>Acceleration</th>
<th>Converge</th>
<th>Discrete-time Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Raginsky &amp; Bouvrie, 2012)</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Mertikopoulos and Staudigl, 2016)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>(Krichene &amp; Bartlett, 2017)</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>
Yet ...

<table>
<thead>
<tr>
<th></th>
<th>Convex</th>
<th>Strongly Convex</th>
<th>Acceleration</th>
<th>Converge</th>
<th>Discrete-time Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Raginsky &amp; Bouvrie, 2012)</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Mertikopoulos and Staudigl, 2016)</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>(Krichene &amp; Bartlett, 2017)</td>
<td>✅</td>
<td></td>
<td>✅</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>This talk</td>
<td>✅</td>
<td></td>
<td>✅</td>
<td>✅</td>
<td></td>
</tr>
</tbody>
</table>
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 \( \mathcal{L}(X, V, t) = \frac{1}{2} \| V \|^2 - f(X) \)

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

  \[
  J(X) = \int_T \mathcal{L}(X_t, \dot{X}_t, t) dt
  \]

- Euler-Lagrange equation

  \[
  \frac{d}{dt} \left\{ \frac{\partial \mathcal{L}}{\partial \dot{X}_t}(X_t, \dot{X}_t, t) \right\} = \frac{\partial \mathcal{L}}{\partial X_t}(X_t, \dot{X}_t, t)
  \]
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

\[ \ddot{X}_t + \dot{Y}_t + \nabla f(X_t) = 0 \]

- 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 $e^{\alpha_t} = \dot{\beta}_t$, $\gamma_t = \beta_t$

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

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?

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

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

So we introduce an extra shrinkage parameter

\[
\begin{align*}
\frac{dX_t}{dt} &= \dot{s}_t (\nabla h^*(Y_t) - X_t) dt \\
\frac{dY_t}{dt} &= -\dot{s}_t (\nabla f(X_t) dt + \sqrt{\delta} \sigma(X_t, t) dB_t)
\end{align*}
\]

Brownian motion

shrinkage parameter
Convergence Rate of Continuous-time Dynamics: General Convex Functions

Stochastic differential equation (SDE):

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

- \( t > 0 \) is time index
- \( \delta, \beta, s_t \) 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 \( \|\sigma(X_t, t)\|_2 \leq \sigma t^q \)
- 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

\[ E_t = e^{\beta_t}(f(X_t) - f(x^*)) + s_t D_h^*(Y_t, \nabla h(x^*)) \]

• Step 1: bounding \( dE_t \)

Rewrite the stochastic dynamics as the following SDE

\[
\begin{bmatrix}
\dot{X}_t \\
\dot{Y}_t
\end{bmatrix} = \begin{bmatrix}
\beta_t (\nabla h^*(Y_t) - X_t) \\
-\beta_t e^{\beta_t} \nabla f(X_t)/s_t
\end{bmatrix} dt + \begin{bmatrix}
0 \\
-\beta_t e^{\beta_t} \sqrt{\delta} \sigma(X_t,t)/s_t
\end{bmatrix} dB_t.
\]

Applying Itô’s Lemma to \( E_t \) with respect to the above SDE yields

\[
\begin{align*}
dE_t &= \frac{\partial E_t}{\partial t} dt + \left< \frac{\partial E_t}{\partial X_t}, dX_t \right> + \left< \frac{\partial E_t}{\partial Y_t}, dY_t \right> + \frac{\dot{\beta}_t^2 e^{2\beta_t}}{2s_t^2} \text{tr} \left( \sigma_t^T \frac{\partial^2 E_t}{\partial Y_t^2} \sigma_t \right) dt \\
&\leq \dot{s}_t M_{h,x} + \frac{1}{2s_t} \beta_t^2 e^{2\beta_t} \text{tr} \left( \sigma_t^T \nabla^2 h^*(Y_t) \sigma_t \right) dt - \beta_t e^{\beta_t} \left< \nabla h^*(Y_t) - x^*, \sigma_t dB_t \right>.
\end{align*}
\]
Roadmap of the Proof

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

$$\mathbb{E}[\mathcal{E}_t] \leq \mathcal{E}_0 + (s_t - s_0)M_{h,X} + \frac{1}{2}\mathbb{E}\left[\int_0^t \frac{\dot{\beta}_r^2 e^{2\beta_r}}{s_r} \text{tr}(\sigma_r^\top \nabla^2 h^*(Y_r)\sigma_r) \, dr\right],$$

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

• **Step 3: choosing parameters**

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

$$\mathbb{E}[f(X_t) - f(x^*)] \leq \frac{\mathbb{E}[\mathcal{E}_t]}{e^{\beta_t}} = O\left(\frac{1}{t^2} + \frac{1}{t^{1/2-q}}\right).$$
Bregman Lagrangian for Mirror Descent: Strongly Convex Functions

• Damped Lagrangian

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

• Solution to Euler-Lagrangian equation

\[ \dddot{X}_t + \dot{X}_t + \nabla f(X_t) = 0 \]

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

\[ \mathcal{L}(X, V, t) = e^{\alpha_t + \beta_t + \gamma_t} \left( \mu D_h(X + e^{-\alpha_t} V, X) - f(X) \right) \]
Continuous-time Dynamics of MD: Strongly Convex Functions

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

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

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

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

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):

\[
\begin{align*}
    dX_t &= \dot{\beta}_t (\nabla h^*(Y_t) - X_t) dt \\
    dY_t &= -\dot{\beta}_t \left( \frac{1}{\mu} \nabla f(X_t) dt + (Y_t - \nabla h(X_t)) dt + \frac{\sqrt{\delta} \sigma(X_t, t)}{\mu} dB_t \right)
\end{align*}
\]

- \( t > 0 \) is time index
- \( \delta, \beta_t \) 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

\[
\frac{x_{k+1} - x_k}{\delta} \approx \dot{X}_t
\]

\[
\frac{y_{k+1} - y_k}{\delta} \approx \dot{Y}_t
\]

Backward (Implicit) Euler Discretization

\[
\frac{x_k - x_{k-1}}{\delta} \approx \dot{X}_t
\]

\[
\frac{y_k - y_{k-1}}{\delta} \approx \dot{Y}_t
\]
New Discrete-time Algorithm (Implicit)

SDEs for general convex functions:

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

Implicit discretization

\[
\begin{align*}
y_{k+1} - y_k &= -\tau_k/s_k G(x_{k+1}; \xi_{k+1}) \\
\nabla h^*(y_{k+1}) &= x_{k+1} + 1/\tau_k (x_{k+1} - x_k)
\end{align*}
\]

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:

\[
\begin{align*}
\frac{dX_t}{dt} &= \hat{\beta}_t (\nabla h^*(Y_t) - X_t)dt \\
\frac{dY_t}{dt} &= -\frac{1}{s_t} (\nabla f(X_t)dt + \sqrt{\delta} \sigma(X_t, t)dB_t)
\end{align*}
\]

Hybrid discretization

\[
\begin{align*}
\nabla h^*(y_k) &= x_{k+1} + 1/\tau_k (x_{k+1} - x_k) \\
y_{k+1} - y_k &= -\tau_k/s_k G(x_{k+1}; \xi_{k+1})
\end{align*}
\]

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:

\[ \begin{align*}
    dX_t &= \dot{\beta}_t (\nabla h^*(Y_t) - X_t) dt \\
    dY_t &= -\frac{\dot{\beta}_t}{s_t} (\nabla f(X_t) dt + \sqrt{\delta} \sigma(X_t, t) dB_t)
\end{align*} \]

Explicit discretization with additional sequence

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

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

Implicit discretization
\[
\begin{align*}
y_{k+1} - y_k &= -\tau_k\left(G(x_{k+1}; \xi_{k+1})/\mu + y_{k+1} - \nabla h(x_{k+1})\right) \\
\nabla h^*(y_{k+1}) &= x_{k+1} + 1/\tau_k(x_{k+1} - x_k)
\end{align*}
\]

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

Optimal rate [Ghadimi & Lan, 2012]

Implicit update
New Discrete-time Algorithm (ASMD)

SDEs for strongly convex functions:
\[
\begin{align*}
\, & dX_t = \dot{\beta}_t (\nabla h^*(Y_t) - X_t) dt \\
\, & dY_t = -\dot{\beta}_t \left( \frac{1}{\mu} \nabla f(X_t) dt + (Y_t - \nabla h(X_t)) dt + \frac{\sqrt{\delta} \sigma(X_t, t)}{\mu} dB_t \right)
\end{align*}
\]

Hybrid discretization
\[
\nabla h^*(y_k) = x_{k+1} + \frac{1}{\tau_k} (x_{k+1} - x_k)
\]
\[
y_{k+1} - y_k = -\tau_k \left( G(x_{k+1}; \xi_{k+1})/\mu + y_{k+1} - \nabla h(x_{k+1}) \right)
\]

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:
\[
\begin{align*}
    dX_t &= \hat{\beta}_t (\nabla h^*(Y_t) - X_t)dt \\
    dY_t &= -\hat{\beta}_t \left( \frac{1}{\mu} \nabla f(X_t)dt + (Y_t - \nabla h(X_t))dt + \frac{\sqrt{\delta} \sigma(X_t, t)}{\mu} dB_t \right)
\end{align*}
\]

Explicit discretization with additional sequence
\[
\begin{align*}
    \nabla h^*(y_k) &= x_k + 1/\tau_k (z_{k+1} - x_k) \\
    y_{k+1} - y_k &= -\tau_k \left( G(z_{k+1}; \xi_{k+1})/\mu + y_k - \nabla h(z_{k+1}) \right) \\
    x_{k+1} &= \arg \min_{x \in \mathcal{X}} \left\{ \langle G(z_{k+1}; \xi_{k+1}), x \rangle + M_k D_h(z_{k+1}, x) \right\}
\end{align*}
\]

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:** \( \mathcal{X} = \{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_i \geq 0 \} \)
- **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:** \( \mathcal{X} = \{ x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0 \} \)
- **distance generating function:**
  \( h(x) = \sum_{i=1}^d x_i \log x_i \)

---

![Graph 1](image1.png)

- **Graph 1:** Log \( (f(x_k) - f(x^*)) \) vs. Number of Iterations (k)
  - SMD
  - SAGE
  - AC-SA
  - ASMD
  - ASMD3

![Graph 2](image2.png)

- **Graph 2:** Log \( (f(x_k) - f(x^*)) \) vs. Number of Iterations (k)
  - SMD
  - AC-SA
  - ASMD
  - ASMD3
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


Reference


Reference


