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
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
- …
Stochastic Gradient Descent

SGD update:

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

Unbiased estimator of the gradient:

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

Convergence rate

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

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

- convex & bounded gradient
- bounded gradient: \[\|G(x; \xi)\|_2 \leq G\]
- strongly convex & bounded gradient
- \(\mu\)-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]

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

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

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

**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**

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
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]
  • 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

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

\[ dX_t = (-0.5X_t + \sin(0.01t))dt + 0.2dB_t \]
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**

[Raginsky & Bouvrie, 2012]
[Metrikopoulos & Staudigl, 2016]

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

**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**

[Krichene & Bartlett, 2017]

\[ dZ_t = -\eta_t [\nabla f(X_t) dt + \sigma(X_t, t) dB_t] \]
\[ dX_t = a_t [\nabla h^*(Z_t/s_t) - X_t] dt, \]
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>![Checkmark]</td>
<td>![Checkmark]</td>
<td></td>
<td>![Checkmark]</td>
<td></td>
</tr>
<tr>
<td>(Mertikopoulos and Staudigl, 2016)</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td></td>
</tr>
<tr>
<td>(Krichene &amp; Bartlett, 2017)</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
</tr>
<tr>
<td>This talk</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
<td>![Checkmark]</td>
</tr>
</tbody>
</table>
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]

\[ \mathcal{L}(X, V, t) = e^{\alpha t + \gamma t} \left( D_{h}(X + e^{-\alpha t}V, X) - e^{\beta tf}(X) \right) \]
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

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

continuous-time dynamics of AMD

[Wibisono et al, 2016]
Continuous-time Dynamics of SMD: General Convex Functions

Add a Brownian motion?

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

Does not converge

So we introduce an extra shrinkage parameter

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

This is the continuous-time dynamics of accelerated SMD for general convex function
Convergence Rate of Continuous-time Dynamics: General Convex Functions

Stochastic differential equation (SDE):

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

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

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

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

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

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

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

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 parameters
- \( B_t \in \mathbb{R}^d \) 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]
Discretization of SDE

Continuous-Time to Discrete-time sequence

Forward (Explicit) Euler Discretization

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

Backward (Implicit) Euler Discretization

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

SDEs for general convex functions:

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

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

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

\[ \text{Optimal rate } [\text{Ghadimi & Lan, 2012}] \]

\[ \text{Implicit update } \]
New Discrete-time Algorithm (ASMD)

SDEs for general convex functions:

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

Hybrid discretization

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

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:

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

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:

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

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

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

- **Optimal rate** [Ghadimi & Lan, 2012]
- **Implicit update**
New Discrete-time Algorithm (ASMD)

SDEs for strongly convex functions:

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

Hybrid discretization

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

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:

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

Explicit discretization with additional sequence

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

Convergence rate

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

- Optimal rate [Ghadimi & Lan, 2012]
- Explicit (practical) algorithm
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:**
  \[
  h(x) = \sum_{i=1}^d x_i \log x_i
  \]
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


