Acceleration and Averaging in Stochastic Descent Dynamics

Walid Krichene
Google Research
walidk@google.com

Peter Bartlett
U.C. Berkeley
peter@berkeley.edu

Introduction

Setting

- Smooth convex minimization $\min_{x \in \mathcal{X}} f(x)$
  f is convex, ∇f is Lipschitz, and $\mathcal{X}$ is compact.
- Continuous-time dynamics.

Contributions

- Analysis of stochastic dynamics
- Different regimes: persistent noise and vanishing noise
- Acceleration: faster convergence in vanishing noise regime

Stochastic Accelerated Mirror descent

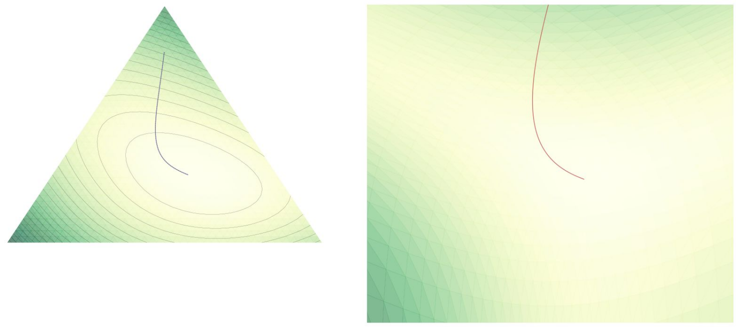
Mirror descent [1]

$$\begin{cases} \dot z(t) = -\eta(t) \nabla f(x(t)) \\ x(t) = \nabla \psi^*(z(t)) \end{cases}$$

Lipschitz mirror map $\nabla \psi^*$ maps dual space $\mathbb{R}^n$ to $\mathcal{X}$

Lyapunov functions: Bregman divergence

$$D_\psi(x^\star, X(t)) = D_{\psi^*}(Z(t), z^\star)$$

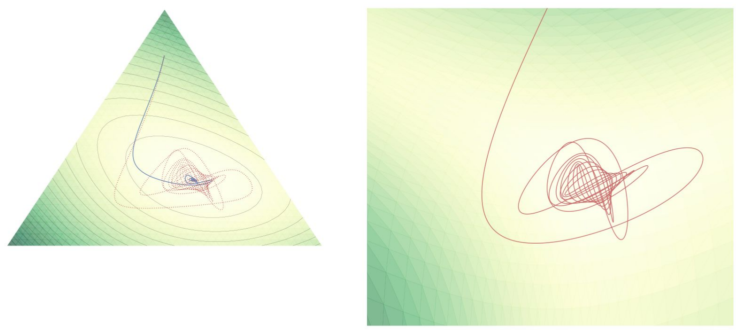


Accelerated Mirror descent [2]

$$\begin{cases} \dot z(t) = -\eta(t) \nabla f(x(t)) \\ x(t) = \dfrac{\int_{t_0}^t w(\tau) \nabla \psi^*(z(\tau)) d\tau}{\int_{t_0}^t w(\tau) d\tau} \end{cases}$$

Averaging in the primal space

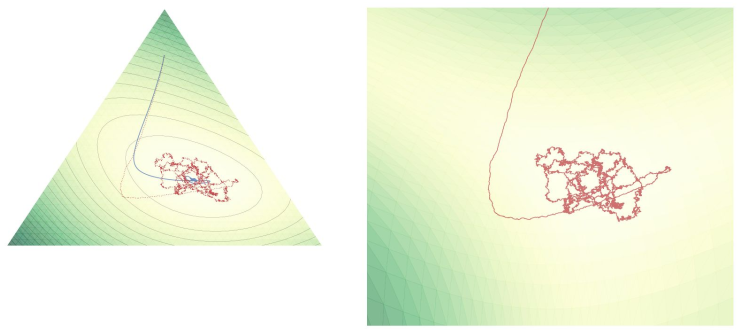
Special case: Nesterov's ODE [3] $(\eta(t) = w(t) = t)$



Stochastic Accelerated Mirror descent

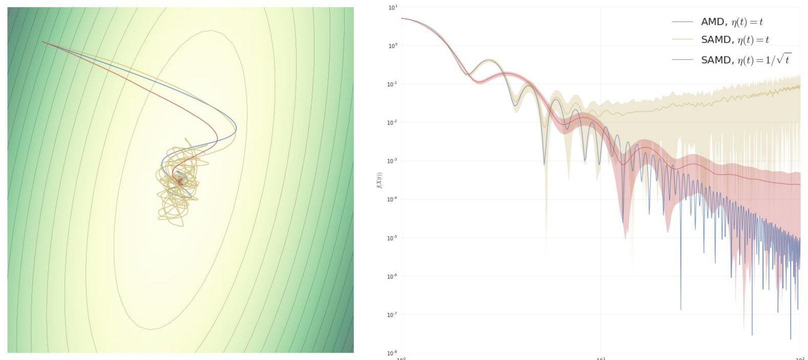
$$\begin{cases} dZ(t) = -\eta(t) \left[\nabla f(X(t)) + \sigma(X(t), t) dB(t)\right] \\ X(t) = \dfrac{\int_{t_0}^t w(\tau) \nabla \psi^*(Z(\tau)) d\tau}{\int_{t_0}^t w(\tau) d\tau} \end{cases}$$

- $B(t)$: Brownian motion
- $\sigma(x, t)$: volatility matrix, with $\sup_{x \in \mathcal{X}} \|\sigma(x,t)\sigma(x,t)^T\|_i \le \sigma_*^2(t)$

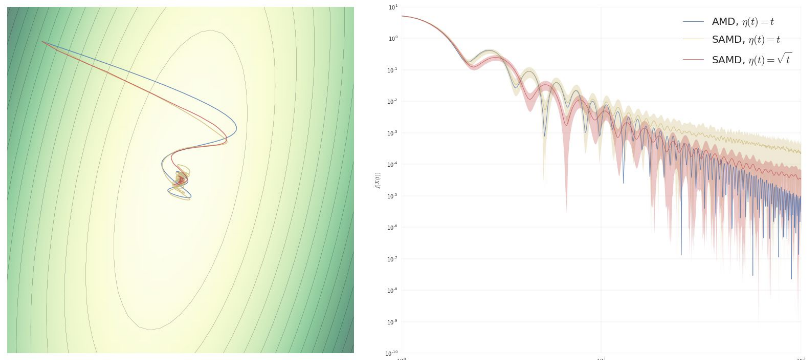


Averaging weights

Accelerated dynamics may diverge in the presence of noise, without proper weight rectification.



Persistent noise $\sigma_*(t) = 1$



Vanishing noise $\sigma_*(t) = t^{-\frac{1}{2}}$

Analysis and convergence rates

Deterministic ($\sigma_* = 0$)

$$E(t) = \left[\int_{t_0}^t \eta\right][f(x(t)) - f(x^\star)] + D_{\psi^*}(z(t), z^\star)$$

E is Lyapunov function

$$f(x(t)) - f(x^\star) \le \frac{E(t)}{\int_{t_0}^t \eta} \le \frac{E(0)}{\int_{t_0}^t \eta}$$

Stochastic

$$E(t) = \left[\int_{t_0}^t \eta\right][f(X(t)) - f(x^\star)] + D_{\psi^*}(Z(t), z^\star)$$

By Itô's formula

$$dE \le 0 - \underbrace{\langle \Delta, \eta\sigma dB\rangle}_{\text{Volatility}} + \underbrace{\frac{1}{2}\mathrm{tr}(\eta\sigma^T \nabla^2 \psi^*(Z)\sigma\eta)dt}_{\text{Itô correction term}}$$

$$\mathbb{E}[f(X(t))] - f(x^\star) \le \frac{E(0) + \frac{nL_{\psi^*}}{2}\int_{t_0}^t \eta^2 \sigma_*^2}{\int_{t_0}^t \eta}$$

We also show a.s. asymptotic rate

$$f(X(t)) - f(x^\star) \le \mathcal{O}\left(\frac{nb(t) + \sqrt{b(t) \log\log b(t)}}{\int_{t_0}^t \eta}\right) \ a.s. \text{ where } b(t) = \int_{t_0}^t \eta^2 \sigma_*^2$$

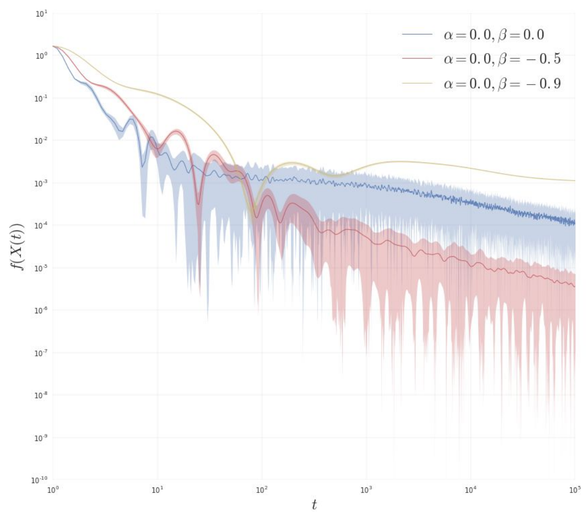
Choosing the weights in the vanishing noise regime

Given $\sigma_*(t) = t^\alpha$,

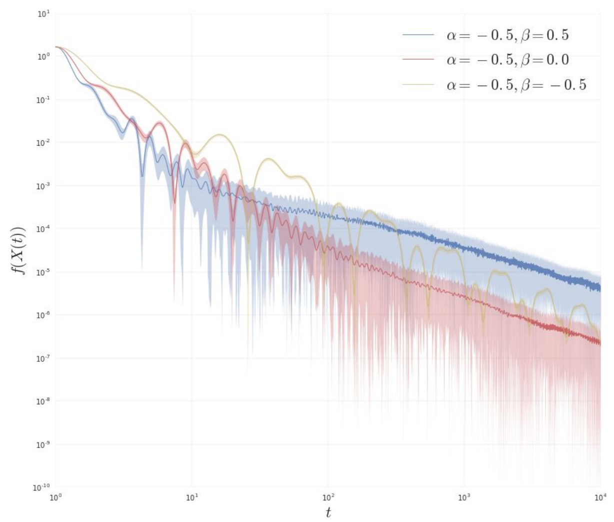
choose $\eta(t) = t^\beta, \ \beta = -\alpha - \frac{1}{2}$

Resulting rate

$$\mathbb{E}[f(X(t)) - f(x^\star)] \le \mathcal{O}(t^{\alpha - \frac{1}{2}})$$



Persistent noise $\sigma_*(t) = 1$



Vanishing noise $\sigma_*(t) = t^{-\frac{1}{2}}$

Additional results, open questions

Results

- A.S. convergence if $\eta(t)\sigma_*(t) = o(1/\sqrt{\log t})$ and $\int_{t_0}^t \eta^2\sigma_*^2 = o\left(\int_{t_0}^t \eta\right)$
  Show $(X(t), Z(t))$ is an Asymptotic Pseudo Trajectory of $(x(t), z(t))$.
- Deterministic dynamics: scaling time => arbitrarily fast convergence.
  Stochastic dynamics: scaling time => scales covariation.
- More general family of stochastic dynamics: time-varying sensitivity.

Open questions

- Discretization
- Faster rates for strongly convex functions
- Multiplicative models of noise:
  $\sup_{x \in \mathcal{X}} \|\sigma(x)\sigma(x)^T\|_i$ scales with $\|\nabla f(x)\|_*^2$

[1] Nemirovski and Yudin. Problems Complexity and Method Efficiency in Optimization. Wiley-Interscience series in discrete mathematics. Wiley, 1983.
[2] W. Krichene, A. Bayen and P. Bartlett. Accelerated Mirror Descent in Continuous and Discrete Time. NIPS 2015.
[3] W. Su, S. Boyd and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. NIPS 2014.