The Mirror Langevin Algorithm Converges with Vanishing Bias

Andre Wibisono

Yale University

Joint work with Ruilin Li, Molei Tao, Santosh Vempala (Georgia Tech)

arXiv:2109.12077

Simons Workshop on Sampling Algorithms and Geometries on Probability Distributions

September 27, 2021

Plan

Mirror Langevin Algorithm

Continuous Time Dynamics

Main Result: Convergence of MLA with Vanishing Bias

Proof: Mean-Square Analysis

Sampling Problem

Goal: Sample from a probability distribution $\nu \propto e^{-f}$ on $\mathcal{X} \subseteq \mathbb{R}^d$

- Assume ν has density $\propto e^{-f}$ wrt Lebesgue measure dx on $\mathbb{R}^d$
- Assume $f : \mathcal{X} \to \mathbb{R}$ differentiable, can compute $\nabla f : \mathcal{X} \to \mathbb{R}^d$
- Nice theory for *log-concave sampling* (when $\mathcal{X}$ and f convex)
- Nice connection to optimization $\min_{x \in \mathcal{X}} f(x)$

Unconstrained Sampling with Langevin

To sample from $\nu \propto e^{-f}$ on $\mathcal{X} = \mathbb{R}^d$, we can use:

- In continuous time, the **Langevin Dynamics**:

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dW_t$$

  where W_t is the standard Brownian motion in $\mathbb{R}^d$

- In discrete time, the **Unadjusted Langevin Algorithm**:

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h}\, z_k$$

  where $h > 0$ is step size and $z_k \sim \mathcal{N}(0, I)$ is an independent Gaussian random variable in $\mathbb{R}^d$

"Distance" between Distributions

- **Wasserstein distance**

$$W_2(\rho,\nu) = \inf_{(X,Y)\sim\Pi(\rho,\nu)} \mathbb{E}[\|X-Y\|^2]^{\frac{1}{2}}$$

- **KL divergence (relative entropy)**

$$H_\nu(\rho) = \mathbb{E}_\nu\left[\frac{\rho}{\nu}\log\frac{\rho}{\nu}\right] = \int_{\mathcal{X}} \rho(x)\log\frac{\rho(x)}{\nu(x)}\,dx$$

- χ^2**-divergence**

$$\chi^2_\nu(\rho) = \mathrm{Var}_\nu\left(\frac{\rho}{\nu}\right) = \int_{\mathcal{X}} \frac{\rho(x)^2}{\nu(x)}dx - 1$$

Langevin Dynamics in Continuous Time

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dW_t$$

- Optimization interpretation as the **gradient flow** for minimizing KL divergence with W_2 metric in the space $\mathcal{P}$ of distributions over $\mathbb{R}^d$

  [Jordan, Kinderlehrer, & Otto. *The variational formulation of the Fokker–Planck equation.* SIAM Journal on Mathematical Analysis, 1998]

- If $\nu \propto e^{-f}$ is **strongly log-concave** ($\Leftrightarrow$ f strongly convex), then **LD** has exponential contraction in W_2 distance

- Exponential convergence when ν satisfies isoperimetry (log-Sobolev inequality (LSI) or Poincaré inequality)

  This yields a mixing time bound with $\log(1/\epsilon)$ dependence on error ϵ

Unadjusted Langevin Algorithm in Discrete Time

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h}\, z_k$$

- *Biased*: $x_k \sim \rho_k$ converges to $\nu_h \neq \nu$, bias scales with h.
  This yields a mixing time with poly$(1/\epsilon)$ dependence on error ϵ
- Can show mixing time in W_2 under strong convexity, smoothness

  [Dalalyan, *Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent*, COLT 2017]

  [Durmus & Moulines, *Nonasymptotic convergence analysis for the unadjusted Langevin algorithm*, Annals of Applied Probability, 2017]
- Can show mixing time in KL divergence under LSI, smoothness

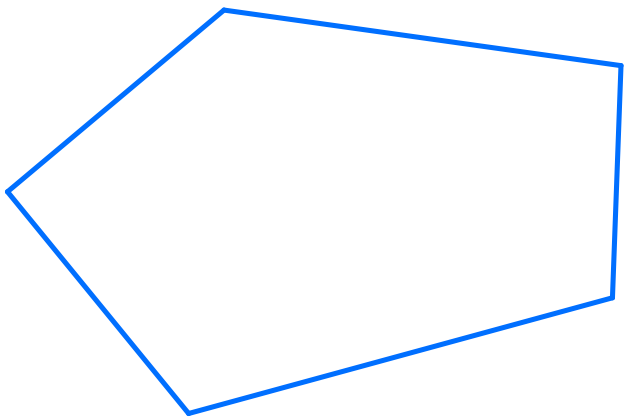
  [Vempala & Wibisono, *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, NeurIPS 2019]
- Can correct bias with accept-reject: Metropolis-Adjusted Langevin Algorithm (MALA) has mixing time $\log(1/\epsilon)$ in TV distance

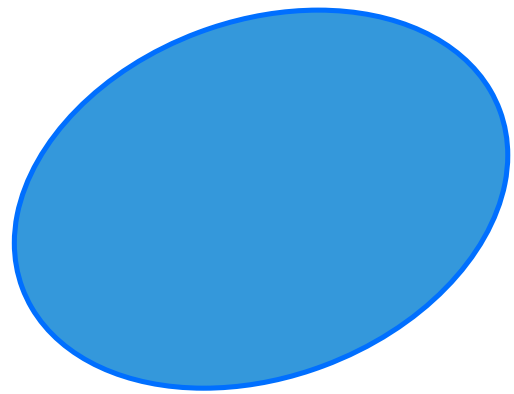
Constrained Sampling

Suppose we are in the constrained setting: $\mathcal{X} \subsetneq \mathbb{R}^d$

How to sample from $\nu \propto e^{-f}$ supported on $\mathcal{X}$?

- e.g. $\nu =$ uniform or Gaussian distribution on $\mathcal{X} =$ polytope





Constrained Sampling

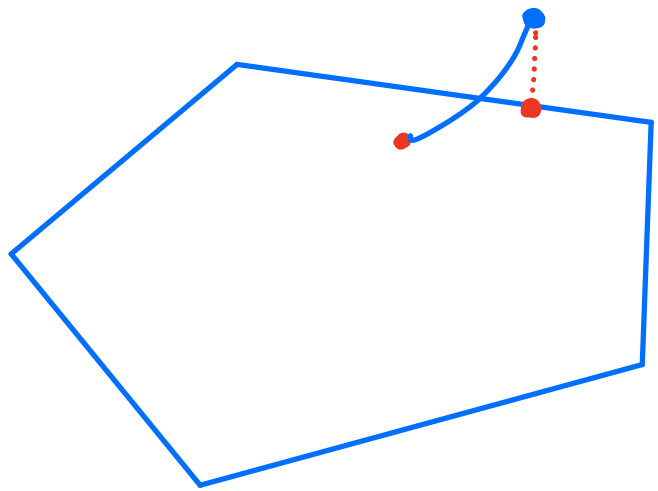
Some approaches:

1. **Projected Langevin Algorithm**: ULA + projection

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_k - h\nabla f(x_k) + \sqrt{2h}\, z_k\right)$$

   - Discretization of *reflected* Langevin dynamics
   - Mixing time bound in TV distance

     [Bubeck, Eldan, & Lehec, *Finite-Time Analysis of Projected Langevin Monte Carlo*, NeurIPS 2015]

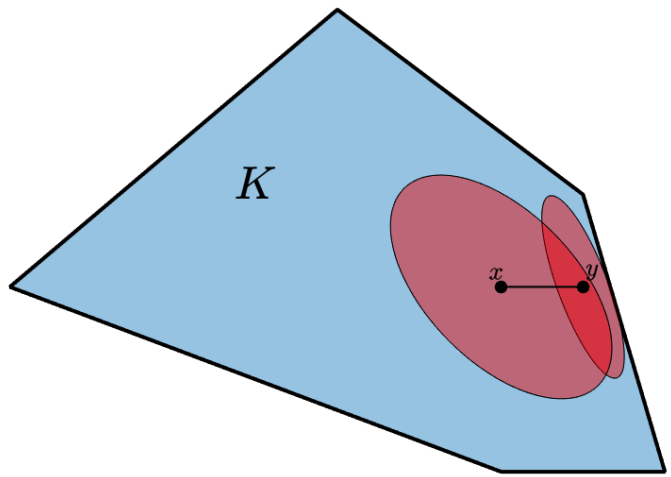


Constrained Sampling

2. **Dikin walk**: Ball walk with ellipsoid defined by the Hessian of the log-barrier function
   - Similar to *interior point methods* in optimization
   - Converges in $\tilde{O}(md)$ steps for $\nu =$ uniform on a polytope with m facets in d dimensions.

     [Kannan & Narayanan, *Random walks on polytopes and an affine interior point method for linear programming*, Mathematics of OR, 2012]
   - Converges in $\tilde{O}(d^2)$ using weighted barrier function

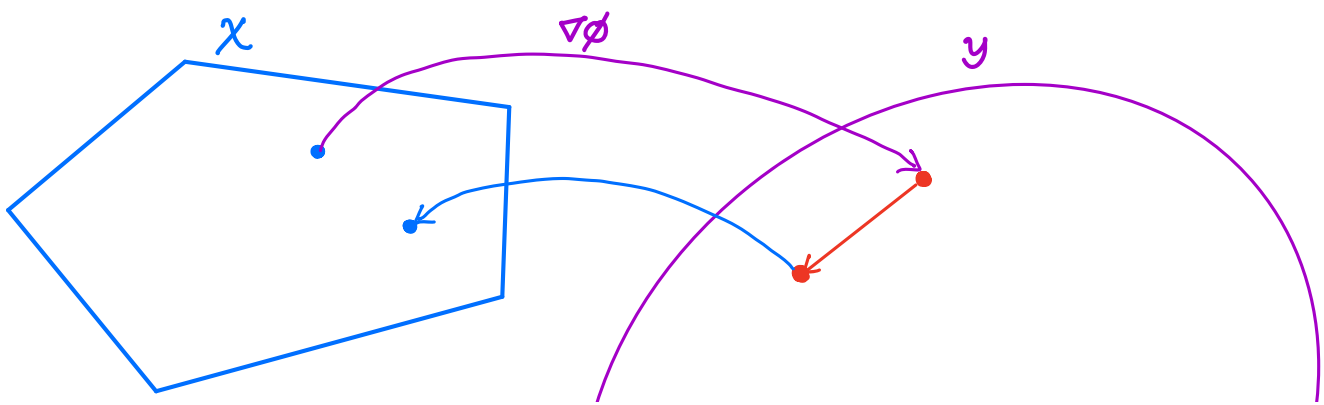
     [Laddha, Lee & Vempala, *Strong self-concordance and sampling*, STOC 2020]



Constrained Sampling

3. **Mirror Langevin Algorithm:** Use mirror map and apply Langevin update in dual space

   - Similar to *mirror descent* in optimization
   - Discretization of the **Mirror Langevin Dynamics** in continuous time
   - Proposed by [Zhang, Peyré, Fadili & Pereyra, *Wasserstein control of Mirror Langevin Monte Carlo*, COLT 2020]
   - See also related approach by [Hsieh, Kavis, Rolland & Cevher, *Mirrored Langevin Dynamics*, NeurIPS 2018]



Mirror Langevin Algorithm

The **Mirror Langevin Algorithm (MLA)** is:

$$x_{k+1} = \nabla\phi^*\left(\nabla\phi(x_k) - h\nabla f(x_k) + \sqrt{2h}\sqrt{\nabla^2\phi(x_k)}\, z_k\right)$$

with step size $h > 0$ and $z_k \sim \mathcal{N}(0, I)$ independent Gaussian

- $\phi\colon \mathcal{X} \to \mathbb{R}$ is a convex *Legendre function*, $\nabla\phi\colon \mathcal{X} \to \mathbb{R}^d$ is bijective, and $\nabla\phi^*(y) = \arg\max_{x\in\mathcal{X}} \langle x, y\rangle - \phi(x)$ where ϕ^* is dual function.
- Can write **MLA** in mirror descent form:

  $$x_{k+1} = \arg\min_{x\in\mathcal{X}} \left\{\langle h\nabla f(x_k) - \sqrt{2h}\sqrt{\nabla^2\phi(x_k)}\, z_k,\, x - x_k\rangle + D_\phi(x, x_k)\right\}$$

  where $D_\phi(x, x') = \phi(x) - \phi(x') - \langle\nabla\phi(x'), x - x'\rangle$ is Bregman divergence

Mirror Langevin Algorithm

Mirror Langevin Algorithm (MLA):

$$x_{k+1} = \nabla\phi^* \left(\nabla\phi(x_k) - h\nabla f(x_k) + \sqrt{2h}\sqrt{\nabla^2\phi(x_k)}\, z_k \right)$$

This is a discretization of the **Mirror Langevin Dynamics**:

$$\begin{cases} Y_t &= \nabla\phi(X_t) \\ dY_t &= -\nabla f(X_t)dt + \sqrt{2}\sqrt{\nabla^2\phi(X_t)}dW_t \end{cases}$$

Question: Why is this the correct dynamics to use?

Plan

Mirror Langevin Algorithm

Continuous Time Dynamics

Main Result: Convergence of MLA with Vanishing Bias

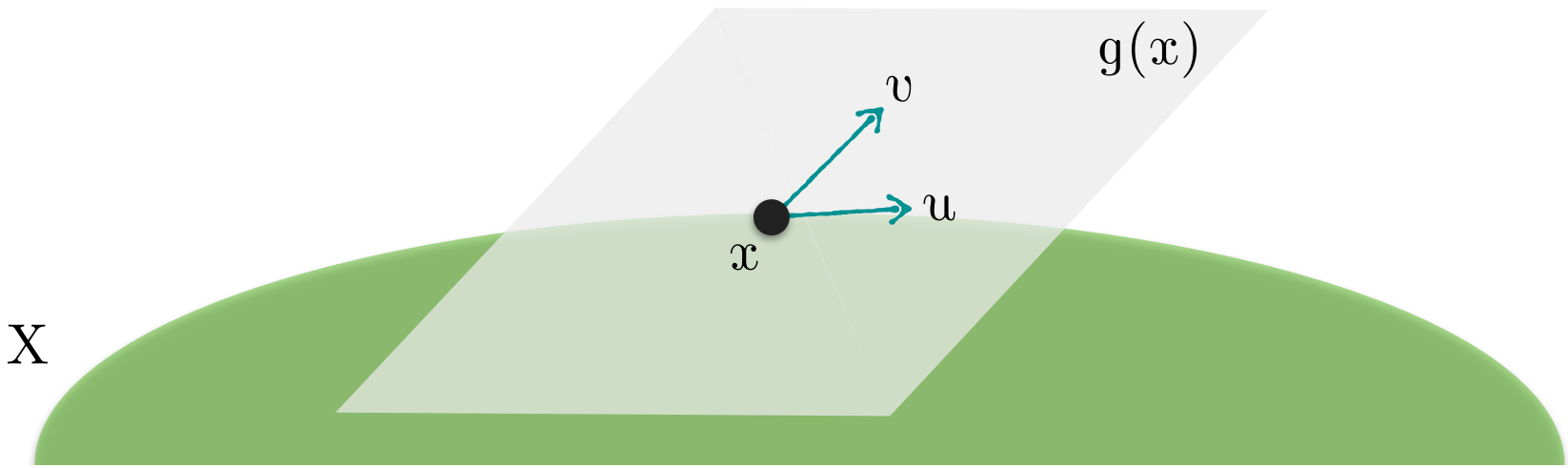
Proof: Mean-Square Analysis

Metric

Endow $\mathcal{X} \subseteq \mathbb{R}^d$ with a Riemannian metric g, as a matrix $\mathrm{g}(x) \succ 0$

$$\langle u, v \rangle_x := u^\top \mathrm{g}(x) v$$

- Euclidean metric: $\mathrm{g}(x) = I$
- Hessian metric: $\mathrm{g}(x) = \nabla^2 \phi(x)$ for some $\phi \colon \mathcal{X} \to \mathbb{R}$ convex

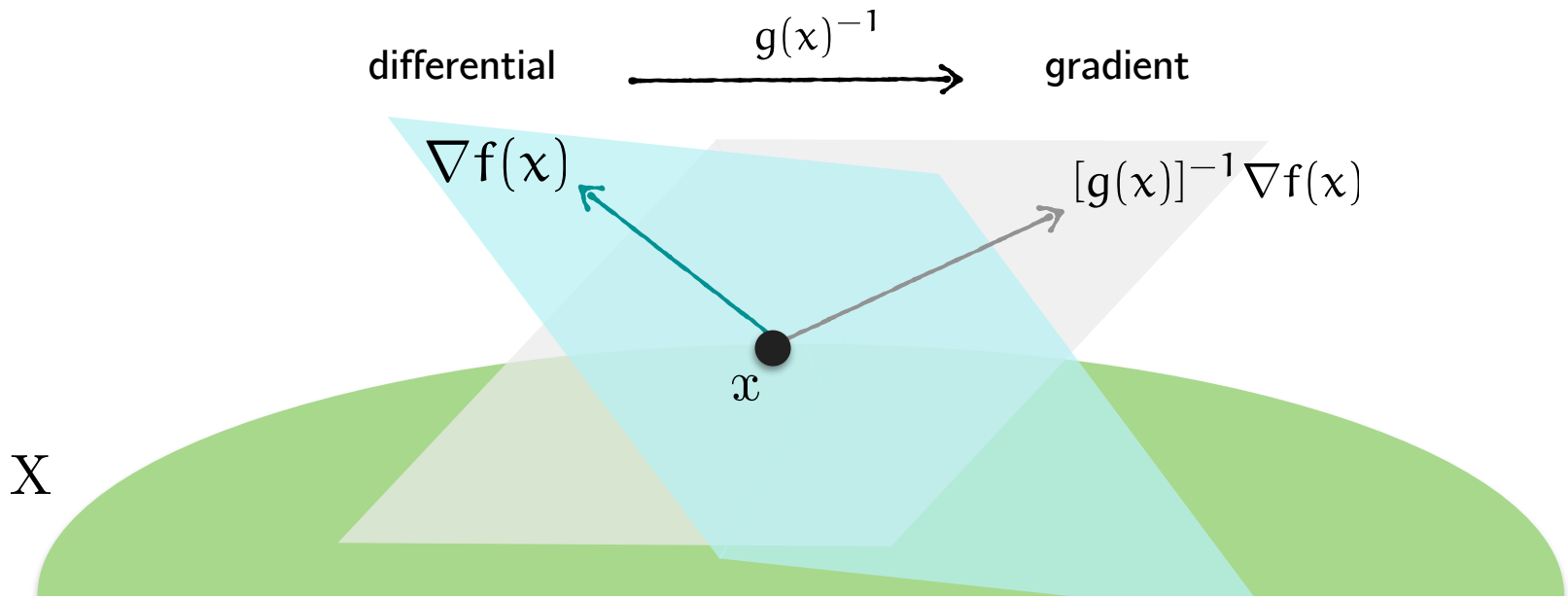


Review for Optimization

Suppose we want to $\min_{x \in \mathcal{X}} f(x)$ with metric g

Riemannian Gradient Flow (RGF):

$$\dot{X}_t = -g(X_t)^{-1} \nabla f(X_t)$$



Review for Optimization

Suppose we want to minimize $f : \mathcal{X} \to \mathbb{R}$ with metric g on $\mathcal{X}$

Riemannian Gradient Flow (RGF):

$$\dot{X}_t = \frac{d}{dt} X_t = -\mathrm{g}(X_t)^{-1} \nabla f(X_t)$$

- If f is *geodesically strongly convex* (strongly convex along geodesics), then RGF is exponentially contracting
- If f is *gradient dominated* ($\|\mathrm{grad}\, f(x)\|_x^2 \geq 2\alpha(f(x) - \min_{\mathcal{X}} f)$), then $f(X_t)$ converges exponentially fast:

$$f(X_t) - \min f \leq e^{-2\alpha t}(f(X_0) - \min f)$$

Review for Optimization

Suppose we use Hessian metric: $\mathrm{g}(x) = \nabla^2\phi(x) \succ 0$

RGF becomes the **Natural Gradient Flow (NGF)**:

$$\dot{X}_t = -\nabla^2\phi(X_t)^{-1}\nabla f(X_t)$$

- Discretizing **NGF** gives the *Natural Gradient Descent*:

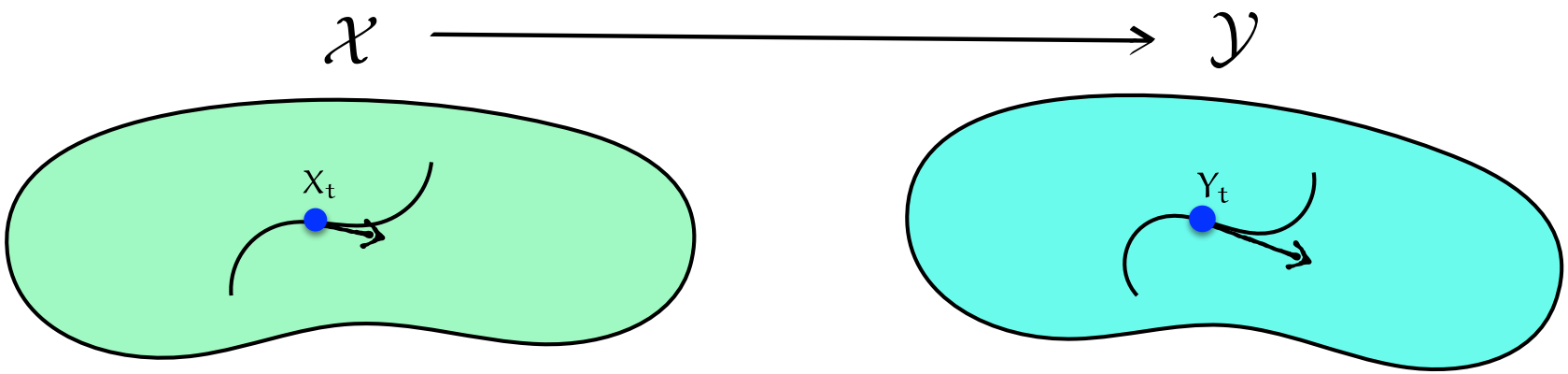
$$x_{k+1} = x_k - h\,\nabla^2\phi(x_k)^{-1}\nabla f(x_k)$$

[Amari, *Natural gradient works efficiently in learning*, Neural Computation, 1998]

Review for Optimization

Use the **mirror map** $\nabla\phi\colon \mathcal{X} \to \mathbb{R}^d$ to work in the *dual space*

$$y = \nabla\phi(x)$$



- Function $f\colon \mathcal{X} \to \mathbb{R}$ is pushed forward to the function $\tilde{f}\colon \mathbb{R}^d \to \mathbb{R}$

$$\tilde{f}(y) = f(\nabla\phi^*(y))$$

- Metric $\nabla^2\phi$ on $\mathcal{X}$ is pushed forward to the metric $\nabla^2\phi^*$ on $\mathbb{R}^d$

Review for Optimization

NGF: $\dot{X}_t = -\nabla^2\phi(X_t)^{-1}\nabla f(X_t)$

In dual space $Y_t = \nabla\phi(X_t)$, **NGF** becomes **Mirror Flow (MF)**:

$$\dot{Y}_t = -\nabla f(\nabla\phi^*(Y_t))$$

- This is also **NGF** for minimizing $\tilde{f} = f \circ \nabla\phi^*$ with metric $\nabla^2\phi^*$
- Discretizing **MF** gives the *Mirror Descent* algorithm:

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \left\{ \langle h\nabla f(x_k),\, x - x_k\rangle + D_\phi(x, x_k) \right\}$$

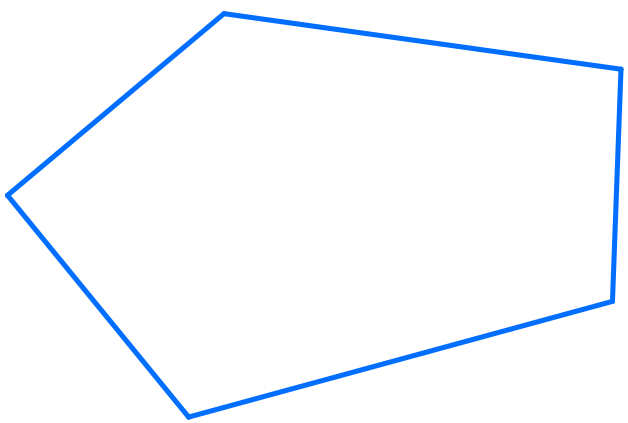
For Sampling

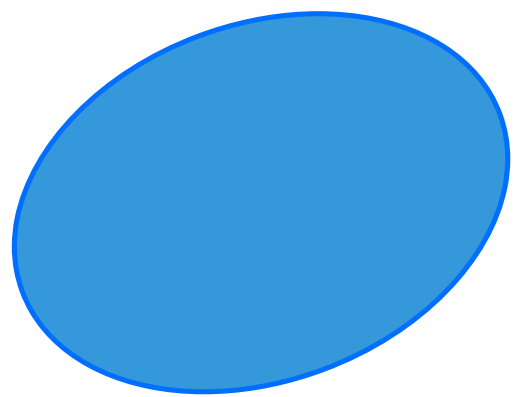
Suppose we want to sample from $\nu \propto e^{-f}$ supported on $\mathcal{X} \subseteq \mathbb{R}^d$

We endow $\mathcal{X}$ with a metric $\mathrm{g}(x) \succ 0$

- Assume $\mathrm{g}(x) \to \infty$ as $x \to \partial\mathcal{X}$, so process does not leave $\mathcal{X}$

What continuous-time dynamics to use?





Riemannian Langevin Dynamics

We want to sample from $\nu \propto e^{-f}$ on $\mathcal{X} \subseteq \mathbb{R}^d$ with metric g

Riemannian Langevin Dynamics (RLD):

$$dX_t = \left(\nabla \cdot (\mathrm{g}(X_t)^{-1}) - \mathrm{g}(X_t)^{-1} \nabla f(X_t)\right) dt + \sqrt{2}\sqrt{\mathrm{g}(X_t)^{-1}}\, dW_t$$

- See e.g. [Girolami & Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, Journal of the Royal Statistical Society: Series B, 2011]
- Euclidean case ($\mathrm{g}(x) = I$): This gives the Langevin Dynamics
- Stationary distribution is $\nu(x) \propto e^{-f(x)}$ (density with respect to dx)

Riemannian Langevin Dynamics

$$dX_t = \left(\nabla \cdot (\mathrm{g}(X_t)^{-1}) - \mathrm{g}(X_t)^{-1}\,\nabla f(X_t)\right) dt + \sqrt{2}\sqrt{\mathrm{g}(X_t)^{-1}}\, dW_t$$

- Density of $X_t \sim \rho_t$ follows the Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t\, \mathrm{g}^{-1}\, \nabla \log \frac{\rho_t}{\nu}\right)$$

- Optimization interpretation: **Gradient Flow** for minimizing relative entropy with respect to Wasserstein metric $W_{2,\mathrm{g}}$ on the space of probability distributions over $(\mathcal{X}, \mathrm{g})$
- Exponential convergence when ν satisfies isoperimetry w.r.t. g
  - Log-Sobolev inequality (LSI) $\Rightarrow$ in KL divergence
  - Poincaré inequality $\Rightarrow$ in χ^2-divergence

Mirror Langevin Dynamics

Suppose we use Hessian metric: $\mathrm{g}(x) = \nabla^2\phi(x) \succ 0$.

Mirror Langevin Dynamics (MLD):

$$dX_t = \left(\nabla \cdot (\nabla^2\phi(X_t)^{-1}) - \nabla^2\phi(X_t)^{-1}\,\nabla f(X_t)\right) dt + \sqrt{2}\sqrt{\nabla^2\phi(X_t)^{-1}}\, dW_t$$

In the dual space $Y_t = \nabla\phi(X_t)$, **MLD** becomes:

$$dY_t = -\nabla f(\nabla\phi^*(Y_t))\, dt + \sqrt{2\nabla^2\phi^*(Y_t)^{-1}}\, dW_t$$

- If ν satisfies LSI/Poincaré w.r.t. $\nabla^2\phi$ ("mirror Poincaré inequality"), then we have exponential convergence in KL or χ^2-divergence
- Studied by [Zhang et al. (2020)] and [Chewi et al. (2020)]
- Discretizing **MLD** in dual space gives Mirror Langevin Algorithm

Newton Langevin Dynamics

Let $\phi = f$. **Newton Langevin Dynamics (NLD)**:

$$dY_t = -Y_t\, dt + \sqrt{2}\sqrt{\nabla^2 f^*(Y_t)^{-1}}\, dW_t$$

- Remarkable property: Brascamp-Lieb inequality $\Rightarrow \nu \propto e^{-f}$ satisfies Poincaré with constant 1 w.r.t. metric $\nabla^2 f$, for any strictly convex f
- This implies χ^2-divergence converges exponentially fast with uniform rate for any strictly log-concave ν
- Studied by [Chewi et al. (2020)] and [Fathi (2019)]

Optimization: Newton Flow

Choose $\phi = f$. **Newton Flow:**

$$\dot{X}_t = -\nabla^2 f(X_t)^{-1} \nabla f(X_t)$$

In dual space $Y_t = \nabla \phi(X_t)$, this becomes:

$$\dot{Y}_t = -Y_t$$

Therefore, $\nabla f(X_t) = e^{-t} \nabla f(X_0)$

- Exponential convergence with uniform rate for *any* strictly convex f
- In discrete time, *Newton's Method* with self-concordance property

Plan

Mirror Langevin Algorithm

Continuous Time Dynamics

Main Result: Convergence of MLA with Vanishing Bias

Proof: Mean-Square Analysis

Mirror Langevin Algorithm: Previous Result

$$x_{k+1} = \nabla\phi^*\left(\nabla\phi(x_k) - h\nabla f(x_k) + \sqrt{2h}\sqrt{\nabla^2\phi(x_k)}\, z_k\right)$$

[Zhang et al. (2020)] showed a convergence analysis of MLA with *non-vanishing* bias (does not go to 0 with step size).

- Required assumption (among others):

  (A1) ϕ satisfies **modified self-concordance** with constant $\alpha > 0$:

$$\|\sqrt{\nabla^2\phi(x')} - \sqrt{\nabla^2\phi(x)}\|_{\mathrm{HS}} \leq \sqrt{\alpha}\|\nabla\phi(x') - \nabla\phi(x)\|_2$$

- Showed bias is $O(\sqrt{dh} + \sqrt{d\alpha})$
- Conjectured non-vanishing bias is unavoidable

Question: Do we need a better algorithm or a better analysis?

Mirror Langevin Algorithm: Alternative Discretization

Alternative discretization of **Mirror Langevin Dynamics**:

$$x_{k+\frac{1}{2}} = \nabla\phi^*\left(\nabla\phi(x_k) - h\nabla f(x_k)\right)$$
$$x_{k+1} = \nabla\phi^*(\tilde{X}_h)$$

where $d\tilde{X}_t = \sqrt{2\nabla^2\phi^*(\tilde{X}_t)^{-1}}\, dW_t$ from $\tilde{X}_0 = x_{k+\frac{1}{2}}$

- [Ahn & Chewi. *Efficient constrained sampling via the mirror-Langevin algorithm*. arXiv:2010.16212, 2020]
- Nice analysis with vanishing bias under self-concordance
- But requires solution of Brownian motion with changing covariance

Mirror Langevin Algorithm: Our Result

We study the basic **MLA:**

$$x_{k+1} = \nabla\phi^* \left(\nabla\phi(x_k) - h\nabla f(x_k) + \sqrt{2h\,\nabla^2\phi(x_k)}\, z_k \right)$$

- We show a convergence analysis of MLA with vanishing bias under a subset of the assumptions of [Zhang et al. (2020)].
- Proof technique uses the mean-square analysis framework.
- Will apply to the dual space $y_k = \nabla\phi(x_k)$:

$$y_{k+1} = y_k - h\nabla f(\nabla\phi^*(y_k)) + \sqrt{2h\,\nabla^2\phi^*(y_k)^{-1}}\, z_k$$

Mirror Langevin Algorithm: Assumptions

We assume:

(A1) ϕ satisfies **modified self-concordance** with constant $\alpha > 0$:

$$\|\sqrt{\nabla^2\phi(x')} - \sqrt{\nabla^2\phi(x)}\|_{\mathrm{HS}} \leq \sqrt{\alpha}\|\nabla\phi(x') - \nabla\phi(x)\|_2$$

(A2) f is M-**smooth** with respect to ϕ:

$$\|\nabla f(x') - \nabla f(x)\|_2 \leq M\|\nabla\phi(x') - \nabla\phi(x)\|_2$$

(A3) f is m-**strongly convex** with respect to ϕ:

$$\langle \nabla f(x') - \nabla f(x), \nabla\phi(x') - \nabla\phi(x)\rangle \geq m\|\nabla\phi(x') - \nabla\phi(x)\|_2^2.$$

These are a subset of the assumptions in [Zhang et al. (2020)]
(they also need bound on moment of $\nabla^2\phi$ and commutator of $\nabla^2 f, \nabla^2\phi$)

Mirror Langevin Algorithm: Assumptions

Equivalently for $A(y) = \sqrt{\nabla^2 \phi^*(y)^{-1}}$ and $g(y) = \nabla f(\nabla \phi^*(y))$

(A1) ϕ satisfies **modified self-concordance** $\Leftrightarrow$ A is $\sqrt{\alpha}$-Lipschitz

$$\|\sqrt{\nabla^2 \phi(x')} - \sqrt{\nabla^2 \phi(x)}\|_{\mathrm{HS}} \leq \sqrt{\alpha} \|\nabla \phi(x') - \nabla \phi(x)\|_2$$

$$\|A(y') - A(y)\|_{\mathrm{HS}} \leq \sqrt{\alpha} \|y' - y\|_2$$

(A2) f is M-**smooth** w.r.t. ϕ $\Leftrightarrow$ g is M-Lipschitz

$$\|\nabla f(x') - \nabla f(x)\|_2 \leq M \|\nabla \phi(x') - \nabla \phi(x)\|_2$$

$$\|g(y') - g(y)\|_2 \leq M \|y' - y\|_2$$

(A3) f is m-**strongly convex** w.r.t. ϕ $\Leftrightarrow$ g is m-monotone

$$\langle \nabla f(x') - \nabla f(x), \nabla \phi(x') - \nabla \phi(x) \rangle \geq m \|\nabla \phi(x') - \nabla \phi(x)\|_2^2$$

$$\langle g(y') - g(y), y' - y \rangle \geq m \|y' - y\|_2^2$$

Mirror Langevin Algorithm: Main Result

Theorem:[1] Assume **(A0)**, **(A1)**, **(A2)** with $\alpha < m/2$. If we run MLA in dual space with $h \leq h_{\max} = \mathcal{O}\left(\frac{(m-2\alpha)^2}{M^2(1+8\alpha)^2}\right)$ from $y_0 \sim \tilde{\rho}_0$, then

$$W_2(\tilde{\rho}_k, \tilde{\nu}) \leq e^{-(m-2\alpha)hk} W_2(\tilde{\rho}_0, \tilde{\nu}) + C_{\mathrm{MLA}}\sqrt{h}$$

where $C_{\mathrm{MLA}} = \mathcal{O}\left(\frac{M(1+8\alpha)\sqrt{d}}{m-2\alpha}\right)$, and $\tilde{\nu} = (\nabla\phi)_{\#}\nu$.

- Equivalently, MLA in the primal space satisfy

$$W_{2,\phi}(\rho_k, \nu) \leq e^{-(m-2\alpha)hk} W_{2,\phi}(\rho_0, \nu) + C_{\mathrm{MLA}}\sqrt{h}$$

  where $W_{2,\phi}(\rho,\nu)^2 = \inf_{(x,x')\sim\Pi(\rho,\nu)} \mathbb{E}[\|\nabla\phi(x) - \nabla\phi(x')\|^2]$

- This shows MLA has vanishing bias $O(\sqrt{dh})$

[1] [Li, Tao, Vempala, & Wibisono, (2021), Theorem 3.1]

Mirror Langevin Algorithm: Mixing Time

Corollary: Assume **(A0)**, **(A1)**, **(A2)** with $\alpha < m/2$.
To reach $W_2(\tilde{\rho}_k, \tilde{\nu}) \leq \epsilon$, can run MLA with $h = \epsilon^2 / C_{\mathrm{MLA}}^2$ for $k = \tau_{W_2}(\epsilon)$ iterations where

$$\tau_{W_2}(\epsilon) = \tilde{O}\left(\frac{C_{\mathrm{MLA}}^2}{(m - 2\alpha)\epsilon^2}\right) = \tilde{O}\left(\frac{d}{\epsilon^2}\right)$$

- Compare with ULA:

  f is smooth, strongly convex: bias $O(\sqrt{dh}) \Rightarrow$ mixing time $\tilde{O}(d/\epsilon^2)$

  f is also third-order smooth: bias $O(\sqrt{d}h) \Rightarrow$ mixing time $\tilde{O}(\sqrt{d}/\epsilon)$
  [Li et al. (2021)]: mean-square analysis

- MLA bias has $\sqrt{h}$ dependence due to multiplicative noise

Plan

Mirror Langevin Algorithm

Continuous Time Dynamics

Main Result: Convergence of MLA with Vanishing Bias

Proof: Mean-Square Analysis

Classical Mean-Square Analysis

How to bound the error of a discretization y_k of a diffusion Y_t?

$$e_k^2 = \mathbb{E}[\|y_k - Y_{\eta k}\|^2]$$

Classical Mean-Square Analysis:

- Studies how *local* integration error propagates to *global* error
- Assume one step $y_k \to y_{k+1}$ has bounded local error.
  Then can show global error e_k is bounded for $k \leq K$.
- Error bound only holds in *finite time*: constant $C \to \infty$ as $K \to \infty$
- [Milstein & Tretyakov, *Stochastic numerics for mathematical physics*, Springer, 2013]

Improved Mean-Square Analysis

- **Idea:** If the diffusion process is *contracting*, then can show the global error bound holds for *all* time

- Proposed and analyzed by [Li et al. (2019)] for speeding up ULA:

  [Xuechen Li, Denny Wu, Lester Mackey, & Murat Erdogdu. *Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond.* NeurIPS 2019]

- We will use a recent extension by [Li et al. (2021)] with a weaker requirement on the local errors:

  [Ruilin Li, Hongyuan Zha, & Molei Tao. *Sqrt(d) dimension dependence of Langevin Monte Carlo.* arXiv preprint arXiv:2109.03839, 2021]

Mean-Square Analysis: Ingredients

- Consider a continuous-time diffusion process $Y_t \in \mathbb{R}^d$, e.g. following an SDE:

$$dY_t = -g(Y_t)\,dt + \sqrt{2}A(Y_t)\,dW_t$$

  We assume $g\colon \mathbb{R}^d \to \mathbb{R}^d$ and $A\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are Lipschitz.

- Consider a discrete-time algorithm Alg_h that tries to simulate the solution Y_h at time $t = h$ from Y_0. We study the iterates

$$y_{k+1} = \mathrm{Alg}_h(y_k)$$

- We want to bound the error between the continuous-time process and the discrete-time algorithm:

$$e_k^2 = \mathbb{E}[\|Y_{hk} - y_k\|^2]$$

Mean-Square Analysis: Assumptions

We assume:

(M1) Diffusion is *contracting* with rate $\beta > 0$: there is $t_0 > 0$ such that for any two solutions Y_t, Y'_t with synchronous coupling,

$$\mathbb{E}[\|Y'_t - Y_t\|^2] \leq e^{-2\beta t}\, \mathbb{E}[\|Y'_0 - Y_0\|^2] \qquad \forall\ 0 \leq t < t_0$$

(M0) Since g and A are Lipschitz, have short-time deviation bound: $\exists\ t_0, C_0 > 0$ such that for any synchronous solutions Y_t, Y'_t,

$$\mathbb{E}[\|(Y'_t - Y'_0) - (Y_t - Y_0)\|^2] \leq C_0\, \mathbb{E}[\|Y'_0 - Y_0\|^2]\ t \quad \forall\ 0 < t \leq t_0$$

Algorithm and Local Error

From any Y_0, let Y_h be the solution to the diffusion process at time $t = h$, and let $\bar{Y}_1 = \text{Alg}_h(Y_0)$ the output of the algorithm

- We say Alg_h has **local weak error** of order p_1 if $\exists\ h_1, C_1, D_1 \geq 0$:

$$\|\mathbb{E}[Y_h - \bar{Y}_1]\| \leq \left(C_1 + D_1\sqrt{\mathbb{E}[\|Y_0\|^2}\right) h^{p_1} \quad \forall\ 0 < h \leq h_1$$

- We say Alg_h has **local strong error** of order p_2 if $\exists\ h_2, C_2, D_2 \geq 0$:

$$\mathbb{E}[\|Y_h - \bar{Y}_1\|^2] \leq \left(C_2^2 + D_2^2\mathbb{E}[\|Y_0\|^2\right) h^{2p_2} \quad \forall\ 0 < h \leq h_2$$

- $D_1 = D_2 = 0$: *Uniform bounds* [Li, Wu, Mackey, & Erdogdu (2019)]

Mean-Square Analysis: Assumptions

(M0) Diffusion has Lipschitz coefficients $\Rightarrow$ short-time deviation:

$$\mathbb{E}[\|(Y'_t - Y'_0) - (Y_t - Y_0)\|^2] \le C_0\, \mathbb{E}[\|Y'_0 - Y_0\|^2]\, t \quad \forall\ 0 < t \le t_0$$

(M1) Diffusion is *contracting* with rate $\beta > 0$:

$$\mathbb{E}[\|Y'_t - Y_t\|^2] \le e^{-2\beta t}\, \mathbb{E}[\|Y'_0 - Y_0\|^2] \qquad \forall\ 0 \le t < t_0$$

(M2) Algorithm has local weak error of order p_1 and local strong error of order p_2 with $\frac{1}{2} < p_2 \le p_1 - \frac{1}{2}$:

$$\|\mathbb{E}[Y_h - \bar{Y}_1]\| \le \left(C_1 + D_1 \sqrt{\mathbb{E}[\|Y_0\|^2}\right) h^{p_1} \quad \forall\ 0 < h \le h_1$$

$$\mathbb{E}[\|Y_h - \bar{Y}_1\|^2] \le \left(C_2^2 + D_2^2 \mathbb{E}[\|Y_0\|^2\right) h^{2p_2} \quad \forall\ 0 < h \le h_2$$

Mean-Square Analysis: Conclusion

Theorem:[2] Assume **(M0)**, **(M1)**, and **(M2)**. There is $h_{\max} > 0$ and $C > 0$ such that if we run Alg_h with $0 < h \leq h_{\max}$ from any $y_0 \sim \rho_0$, the global error remains bounded at *all* time:

$$\sqrt{\mathbb{E}[\|Y_{hk} - y_k\|^2]} \leq C h^{p_2 - \frac{1}{2}} \qquad \forall\ k \geq 0$$

Here $y_k = \mathrm{Alg}_h(y_{k-1})$ and Y_{hk} is diffusion solution from $Y_0 = y_0$.

- The constant is explicit:

$$C = \frac{1}{\sqrt{\beta}} \left(\frac{C_1 + C_0 C_2 + U(D_1 + C_0 D_2)}{\sqrt{\beta}} + C_2 + D_2 U \right)$$

where $U = \sqrt{\mathbb{E}[\|Y_0\|^2] + \mathbb{E}_\nu[\|Y\|^2]}$

[2] [Li, Zha, & Tao, (2021), Theorem 3.3]

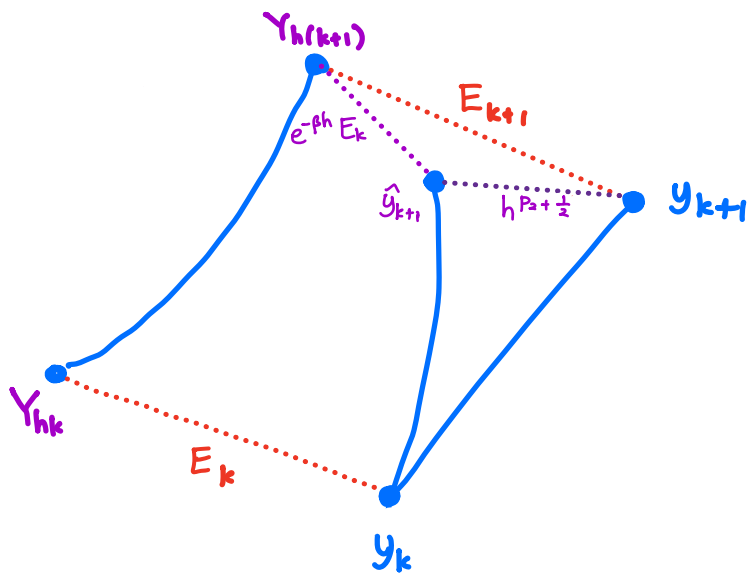
Mean-Square Analysis: Conclusion

Theorem:[3] Assume **(M0)**, **(M1)**, and **(M2)**.

$$\sqrt{\mathbb{E}[\|Y_{hk} - y_k\|^2]} \leq C h^{p_2 - \frac{1}{2}} \qquad \forall\ k \geq 0$$

Idea: Show $E_k = \sqrt{\mathbb{E}[\|Y_{hk} - y_k\|^2]}$ satisfies one-step recursion

$$E_{k+1}^2 \leq e^{-\beta h} E_k^2 + E_k h^{p_2 + \frac{1}{2}} + h^{2p_2}$$



[3] [Li, Zha, & Tao, (2021), Theorem 3.3]

Mean-Square Analysis: Conclusion

Corollary:[4] Assume **(M0)**, **(M1)**, and **(M2)**. From any $y_0 \sim \rho_0$, if we run Alg_h with $0 < h \leq h_{\max}$, then $y_k = \mathrm{Alg}_h(y_{k-1}) \sim \rho_k$ has:

$$W_2(\rho_k, \nu) \leq e^{-\beta k h} W_2(\rho_0, \nu) + C h^{p_2 - \frac{1}{2}} \qquad \forall\ k \geq 0$$

To reach $W_2(\rho_k, \nu) \leq \epsilon$, can run Alg_h with $h = (\epsilon / C)^{1/(p_2 - \frac{1}{2})}$ for $k \geq \tau(\epsilon)$ iterations, where

$$\tau(\epsilon) = \tilde{O}\left(\frac{C^{1/(p_2 - \frac{1}{2})}}{\beta \epsilon^{1/(p_2 - \frac{1}{2})})} \right)$$

[4] [Li, Zha, & Tao, (2021), Theorem 3.4]

Mean-Square Analysis of MLA

Dynamics for MLA is the **Mirror Langevin Dynamics**:

$$dY_t = -g(Y_t)\,dt + \sqrt{2}A(Y_t)\,dW_t$$

where $g(y) = \nabla f(\nabla \phi^*(y))$, $A(y) = \sqrt{\nabla^2 \phi^*(y)^{-1}}$.

Recall assumptions:

(A1) A is $\sqrt{\alpha}$-Lipschitz:

$$\|A(y') - A(y)\|_{\mathrm{HS}} \leq \sqrt{\alpha}\|y' - y\|_2$$

(A2) g is M-Lipschitz:

$$\|g(y') - g(y)\|_2 \leq M\|y' - y\|_2$$

(A3) g is m-monotone

$$\langle g(y') - g(y), y' - y\rangle \geq m\|y' - y\|_2^2$$

Mean-Square Analysis of MLA

Lemma: Assume **(A1)**, **(A2)**, and **(A3)** with $\alpha < \frac{m}{2}$. Then:

1. **MLD** is contractive with rate $\beta = m - 2\alpha$.
2. **MLD** has deviation bound:

$$\mathbb{E}[\|(Y_t' - Y_0') - (Y_t - Y_0)\|^2] \leq 4M\,\mathbb{E}[\|Y_0' - Y_0\|^2]\,t$$

for all $t \geq 0$

Contraction of MLD

Suppose we have two synchronous solutions Y'_t, Y_t of MLD:

$$dY'_t = -g(Y'_t)dt + \sqrt{2}A(Y'_t)dW_t$$
$$dY_t = -g(Y_t)dt + \sqrt{2}A(Y_t)dW_t.$$

Then

$$\begin{aligned}\frac{d}{dt}\mathbb{E}[\|Y'_t - Y_t\|^2] &= -2\mathbb{E}[\langle g(Y'_t) - g(Y_t), Y'_t - Y_t\rangle] + 4\mathbb{E}[\|A(Y'_t) - A(Y_t)\|^2_{\mathrm{HS}}] \\ &\leq -2(m - 2\alpha)\mathbb{E}[\|Y'_t - Y_t\|^2_2].\end{aligned}$$

Exponential contraction if $m > 2\alpha$:

$$\mathbb{E}[\|Y'_t - Y_t\|^2] \leq e^{-2(m-2\alpha)t}\mathbb{E}[\|Y'_0 - Y_0\|^2]$$

- Shows modified self-concordance ($\Leftrightarrow$ A Lipschitz) is natural for mean-square analysis
- In general threshold on α is needed for convergence of SDE, e.g. Geometric Brownian Motion: $dY_t = -Y_t\,dt + \sqrt{2\alpha}\,Y_t\,dW_t$

Mean-Square Analysis of MLA

Lemma: Assume **(A1)**, **(A2)**, and **(A3)** with $\alpha < \frac{m}{2}$. Then MLA with maximum step size $h_1 = \frac{1}{M^2+8\alpha}$ has:

1. Local weak error of order $p_1 \geq \frac{3}{2}$ with constants

$$C_1 = M\sqrt{(1+8\alpha)}C' \qquad D_1 = M\sqrt{1+8\alpha}$$

2. Local strong error of order $p_2 = 1$ with constants

$$C_2 = (1+8\alpha)C' \qquad D_2 = (1+8\alpha)$$

where $C' = (\|y^*\| + \|A(y^*)\|_{\mathrm{HS}} + \frac{1}{M}\|g(y^*)\|) = O(\sqrt{d})$, with $y^* = \nabla\phi(x^*)$ and $x^* = \arg\min_{x\in\mathcal{X}} f(x)$

Mean-Square Analysis of MLA

Proof of Main Theorem:

Assume **(A1)**, **(A2)**, and **(A3)** with $\alpha < \frac{m}{2}$.

Then MLA with maximum step size $h_{\max} = \mathcal{O}\left(\frac{(m-2\alpha)^2}{M^2(1+8\alpha)^2}\right)$ satisfies mean-square assumptions with $p_1 \geq \frac{3}{2}$ and $p_2 = 1$.

Therefore:

$$W_2(\tilde{\rho}_k, \tilde{\nu}) \leq e^{-(m-2\alpha)hk} W_2(\tilde{\rho}_0, \tilde{\nu}) + C_{\mathrm{MLA}}\sqrt{h}$$

where $C_{\mathrm{MLA}} = O\left(\frac{M(1+8\alpha)\sqrt{d}}{m-2\alpha}\right) = O(\sqrt{d})$ and $\tilde{\nu} = (\nabla\phi)_{\#}\nu$.

□

Discussion

We show MLA converges in W_2 with vanishing bias $O(\sqrt{dh})$
$\Rightarrow$ Mixing time bound in W_2 distance is $\tilde{O}(d/\epsilon^2)$

Limitations and questions:

1. Result is in W_2 distance w.r.t. Euclidean metric in $\mathcal{Y} = \mathbb{R}^d$. This is isometric to W_2 distance w.r.t *squared* Hessian metric $(\nabla^2\phi)^2$ in $\mathcal{X}$.

    - Want result in W_2 distance w.r.t. Hessian metric $\nabla^2\phi^*$ in $\mathcal{Y}$ (isometric to W_2 distance w.r.t. Hessian metric $\nabla^2\phi$ in $\mathcal{X}$).
    - Or want result in KL divergence, χ^2-divergence

Discussion

2. Result assumes relative smoothness and strong convexity of f w.r.t. ϕ, requires $m\nabla^2\phi(x) \preceq \nabla^2 f(x) \preceq M\nabla^2\phi(x)$

   - Since $\nabla^2\phi(x) \to \infty$, also need $\nabla^2 f(x) \to \infty$ as $x \to \partial\mathcal{X}$
   - Can we have result under weaker condition?

     e.g. to sample from uniform ($f = 0$) or Gaussian distribution ($f =$ quadratic) on polytope with $\phi =$ log-barrier function

Discussion

3. Result assumes ϕ satisfies modified self-concordance (MSC) with parameter $\alpha < m/2$

    - In 1 dimension, self-concordance (SC) is equivalent to MSC; but in $d > 1$ dimensions they are different.
    - In particular, SC is *affine-invariant*; MSC is not.
    - Can find example where MSC constant is arbitrarily large.
    - Want to have analysis of MLA with the more natural SC.

Example: Log-Barrier

Let $\mathcal{X}$ be polytope

$$\mathcal{X} = \{x \in \mathbb{R}^d : a_i^\top x \geq b_i \quad \forall\, i = 1, \ldots, m\}$$

for some $a_i \in \mathbb{R}^d$ with $\|a_i\| = 1$, and $b_i \in \mathbb{R}$. Let ϕ be log-barrier:

$$\phi(x) = -\sum_{i=1}^{m} \log(a_i^\top x - b_i)$$

- Recall Self-Concordance constant of ϕ is 2.
- Can show Modified Self-Concordance constant α scales as:

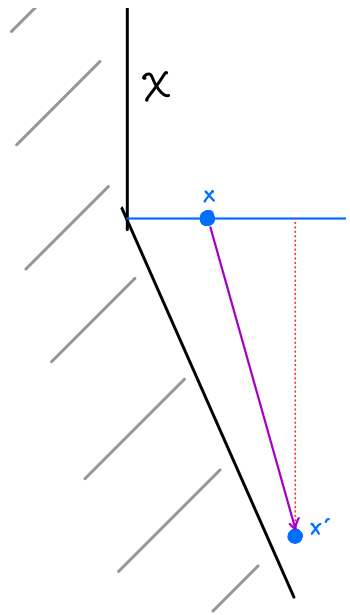
$$\alpha \approx \frac{1}{\sigma_{\min}(A)^2}$$

where $\sigma_{\min}(A)$ is smallest singular value of the constraint matrix $A = (a_1 \cdots a_m)$.

Example: Log-Barrier

Consider polyhedron in $d = 2$ for small $\epsilon > 0$:

$$\mathcal{X} = \{x = (x_1, x_2) \colon x_1 \geq 0, \sqrt{1-\epsilon^2}x_1 + \epsilon x_2 \geq 0\}$$



- $\sigma_{\min}(A) = \sqrt{1-\sqrt{1-\epsilon^2}} \approx \epsilon/\sqrt{2}$
- Can choose $x = (1, 0),\ x' = (2, -\frac{4}{\epsilon})$ such that

$$\alpha \geq \frac{\|\sqrt{\nabla^2\phi(x')} - \sqrt{\nabla^2\phi(x)}\|_{\mathrm{HS}}^2}{\|\nabla\phi(x') - \nabla\phi(x)\|_2^2} = \frac{1}{\sigma_{\min}(A)^2} \approx \frac{2}{\epsilon^2}$$

Discussion

4. Mean-square analysis requires continuous-time process is exponentially contracting

    - Can we have analysis when the process is converging?
    - We have this e.g. for ULA in KL divergence under LSI

      [Vempala & Wibisono, *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, NeurIPS 2019]

Mirror Langevin Algorithm Converges with Vanishing Bias

Ruilin Li, Molei Tao, Santosh Vempala, Andre Wibisono

arXiv:2109.12077

Thank You!

Questions?