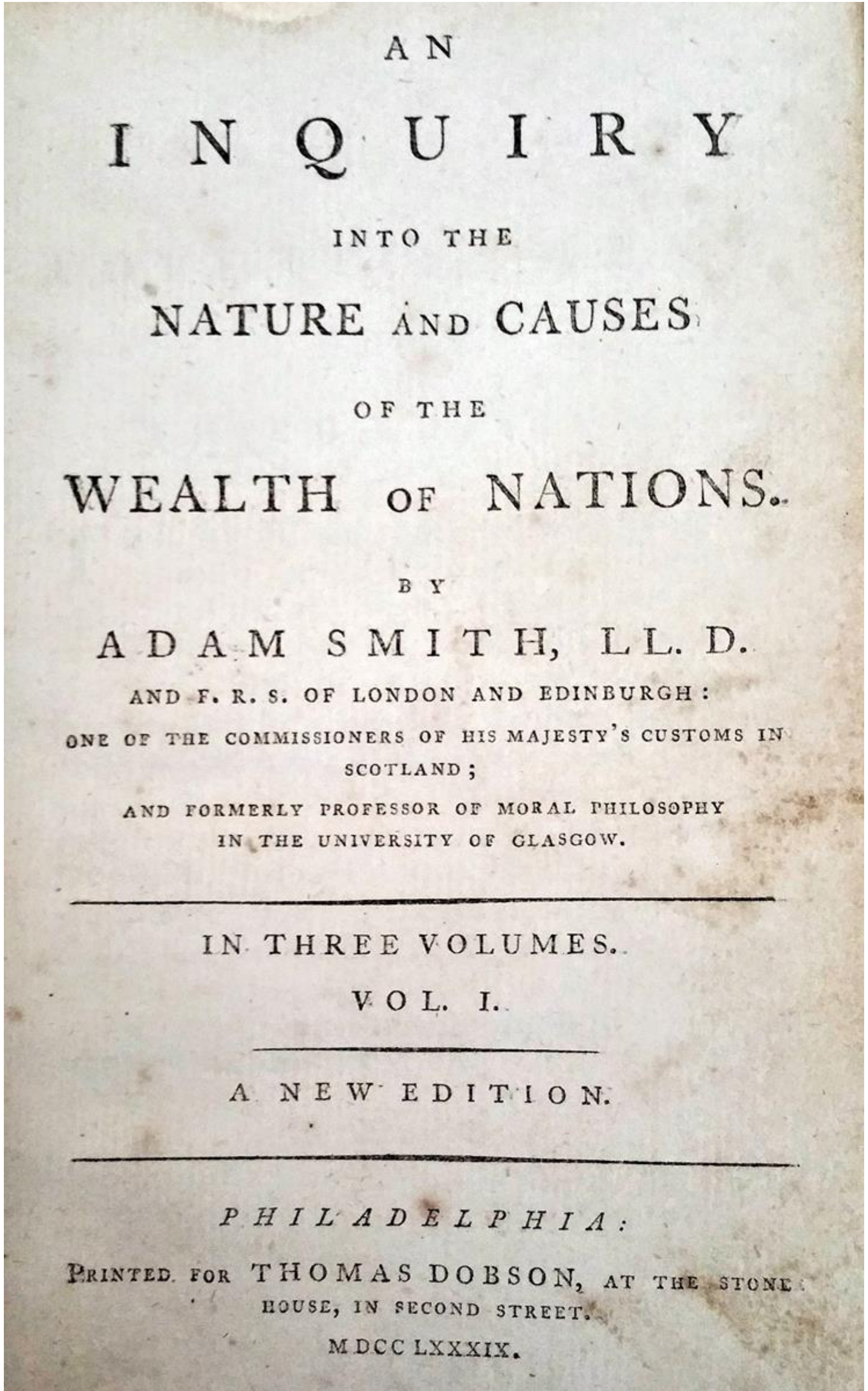
Replica Exchange for Non-Convex Optimization

Jing Dong

Columbia University

Joint work with Xin T. Tong at NUS


AN

INQUIRY

INTO THE

NATURE AND CAUSES

OF THE

WEALTH OF NATIONS.

BY

ADAM SMITH, LL. D.

AND F. R. S. OF LONDON AND EDINBURGH:

ONE OF THE COMMISSIONERS OF HIS MAJESTY'S CUSTOMS IN SCOTLAND;

AND FORMERLY PROFESSOR OF MORAL PHILOSOPHY IN THE UNIVERSITY OF GLASGOW.

IN THREE VOLUMES.

VOL. I.

A NEW EDITION.

PHILADELPHIA:

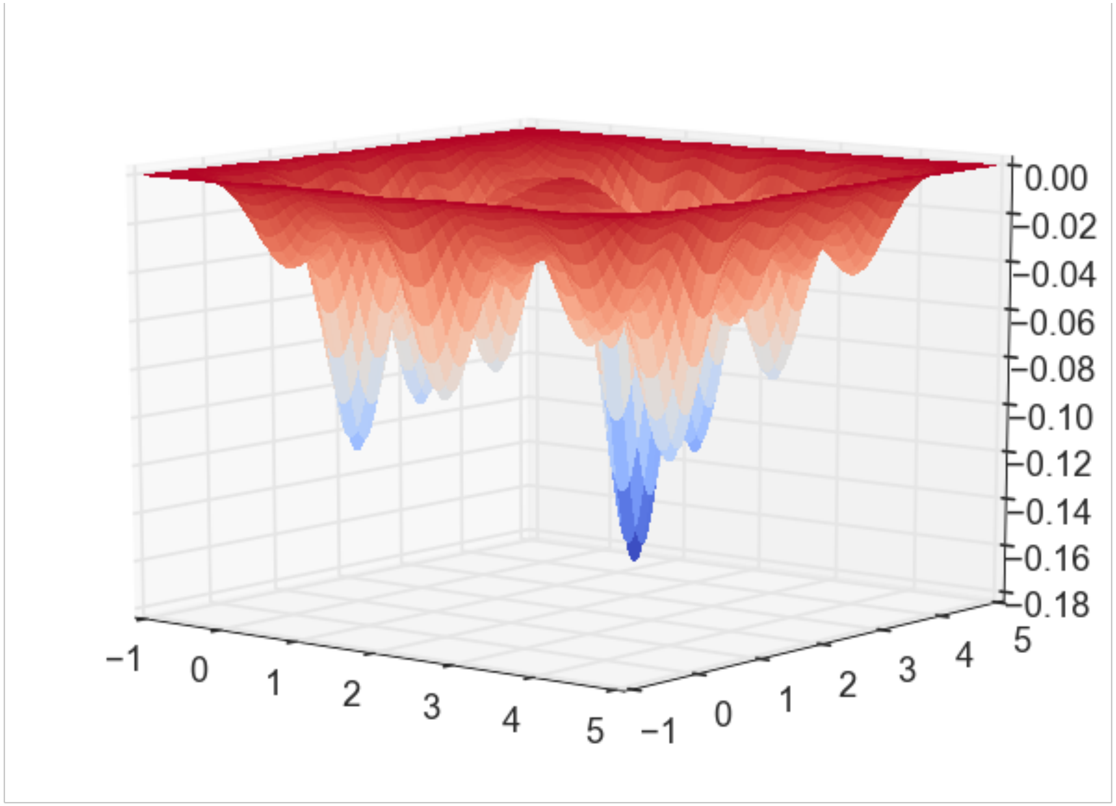
PRINTED FOR THOMAS DOBSON, AT THE STONE HOUSE, IN SECOND STREET.

MDCCLXXXIX.

Division of labor

Objective

$$\min_{x \in R^d} F(x)$$



Gradient Descent (GD)

$$X_{n+1} = X_n - h\nabla F(X_n)$$

When the step size h is properly chosen

- If F is convex

$$F(X_n) - F^* = O(1/n)$$

- If F is strongly convex, i.e., $F(y) - F(x) \geq \langle \nabla F(x), y - x \rangle + \frac{m}{2}\|y - x\|^2$

$$F(X_n) - F^* = O((1 - mh)^n)$$

Gradient Descent (GD)

$$X_{n+1} = X_n - h\nabla F(X_n)$$

When the step size h is properly chosen

- If F is convex

$$F(X_n) - F^* = O(1/n)$$

- If F is strongly convex, i.e., $F(y) - F(x) \geq \langle \nabla F(x), y - x \rangle + \frac{m}{2}\|y - x\|^2$

$$F(X_n) - F^* = O((1-mh)^n)$$

- However, if F is non-convex, X_n can be trapped in local minimums or saddle points

Langevin Dynamic (LD)

$$dY_t = -\nabla F(Y_t)dt + \sqrt{2\gamma}\, dB_t$$

Under suitable regularity conditions on F, Y_t has a stationary distribution

$$\pi(y) \propto \exp\left(-\frac{1}{\gamma}F(y)\right)$$

Langevin Dynamic (LD)

$$dY_t = -\nabla F(Y_t)dt + \sqrt{2\gamma}\, dB_t$$

Under suitable regularity conditions on F, Y_t has a stationary distribution

$$\pi(y) \propto \exp\left(-\frac{1}{\gamma}F(y)\right)$$

Let x_0 denote a local minimum of F, z_0 be the communicating saddle point, and τ_0 denote the time to "escape"

$$E_{x_0}[\tau_0] \sim \frac{Z_0}{(2\pi\gamma)^{d/2}} \frac{2\pi\gamma\sqrt{|\det(\nabla^2 F(z_0))|}}{|\lambda_1(z_0)|} \exp\left(\frac{F(z_0)-F(x_0)}{\gamma}\right)$$

Menz and Schlichting (2014)

Langevin Dynamic (LD)

$$dY_t = -\nabla F(Y_t)dt + \sqrt{2\gamma}\,dB_t$$

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h}\,Z_n$$

Langevin Dynamic (LD)

$$dY_t = -\nabla F(Y_t)dt + \sqrt{2\gamma}\, dB_t$$

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h} Z_n$$

If F is strongly convex, $E[F(Y_n)] - F^* = O(\exp(-mnh) + \gamma h)$

If γ is a constant, to achieve an ε accuracy, we need $h = O(\varepsilon)$

$$n = O(\varepsilon^{-1} \log(1/\varepsilon))$$

GD versus LD

Gradient Descent:

$$X_{n+1} = X_n - h\nabla F(X_n)$$

- Good at exploitation
- Terrible at exploration

Langevin Dynamics

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h} Z_n$$

- Good at exploration
- Inefficient at exploitation

GD versus LD

Gradient Descent:

$$X_{n+1} = X_n - h\nabla F(X_n)$$

- Good at exploitation
- Terrible at exploration

Langevin Dynamics

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h} Z_n$$

- Good at exploration
- Inefficient at exploitation

Can we enjoy the benefit of both?

Yes! We can let them "collaborate".

GDxLD

Algorithm 1: GDxLD: offline optimization

Input: Temperature γ, step size h, number of steps N, and initial X_0, Y_0.

for $n = 0$ *to* $N-1$ **do**

$X'_{n+1} = X_n - \nabla F(X_n)h$;

$Y'_{n+1} = Y_n - \nabla F(Y_n)h + \sqrt{2\gamma h} Z_n$, where $Z_n \sim N(0, I_d)$.;

if $F(Y'_{n+1}) < F(X'_{n+1})$ **then**

$(X_{n+1}, Y_{n+1}) = (Y'_{n+1}, X'_{n+1})$;

else

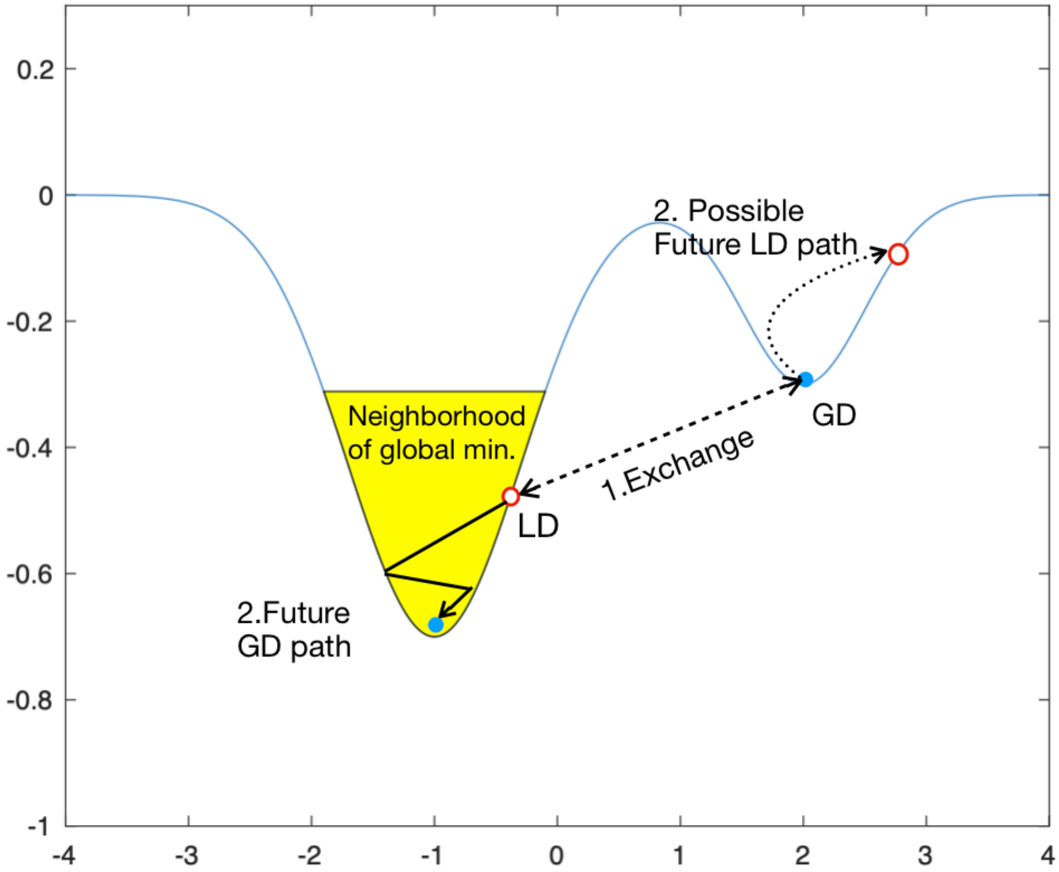
$(X_{n+1}, Y_{n+1}) = (X'_{n+1}, Y'_{n+1})$.

end

end

Output: X_N as an optimizer for F.

GDxLD


0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-4
-3
-2
-1
0
1
2
3
4
2. Possible
Future LD path
Neighborhood
of global min.
GD
1.Exchange
LD
2.Future
GD path

GDxLD

Assumption 1. The gradient is Lipschitz continuous.

$$\|\nabla F(x) - \nabla F(y)\| \leq L \|x - y\|$$

Assumption 2. The objective function is coercive.

$$-\langle \nabla F(x), x \rangle \leq -\lambda_0 \|x\|^2 + M_0$$

Assumption 3. There is a unique global minimum. The objective function is (strongly) convex in a neighborhood of the global minimum.

GDxLD

Assumption 1. The gradient is Lipschitz continuous.

$$\|\nabla F(x) - \nabla F(y)\| \le L \|x - y\|$$

Assumption 2. The objective function is coercive.

$$-\langle \nabla F(x), x \rangle \le -\lambda_0 \|x\|^2 + M_0$$

Assumption 3. There is a unique global minimum. The objective function is (strongly) convex in a neighborhood of the global minimum.

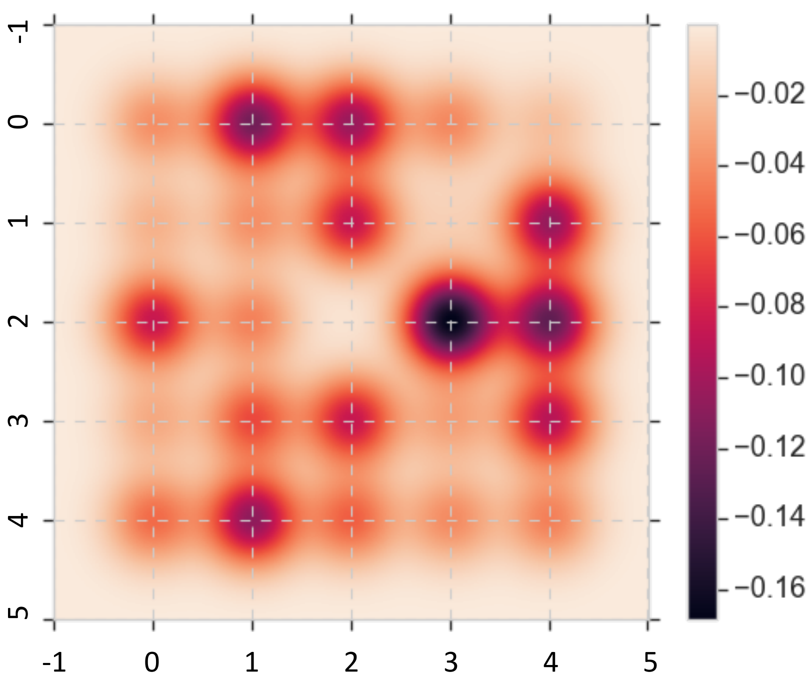
Under Assumptions 1,2,3, and $h<1/(2L)$, for any ε>0 and δ>0, there exists $N(\varepsilon,\delta) = O(\varepsilon^{-1}) + O(\log(1/\delta))$, such that for any $n>N(\varepsilon, \delta)$,

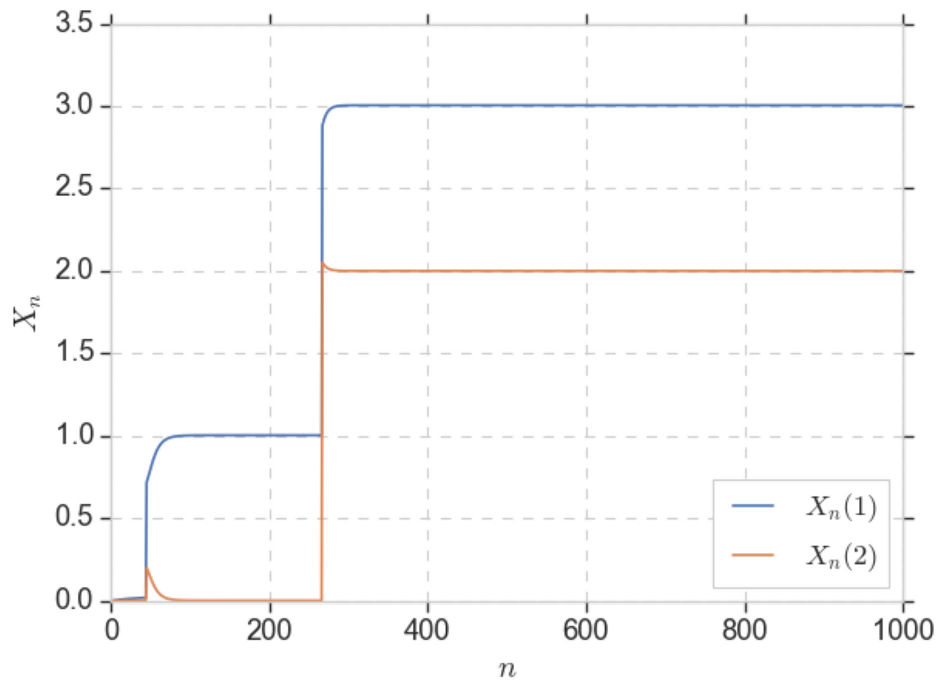
$$P(F(X_n) - F^* \le \varepsilon) \ge 1 - \delta.$$

If in addition, F is strongly convex in a neighborhood of X^* and $h<\min\{1/(2L), 1/m\}$,

$$N(\varepsilon,\delta) = O(\log(1/\varepsilon)) + O(\log(1/\delta)).$$

GDxLD

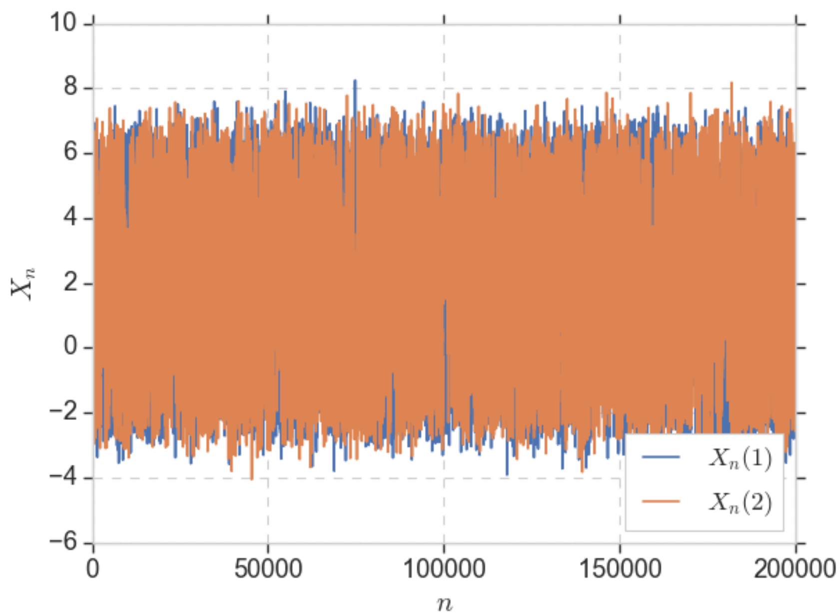




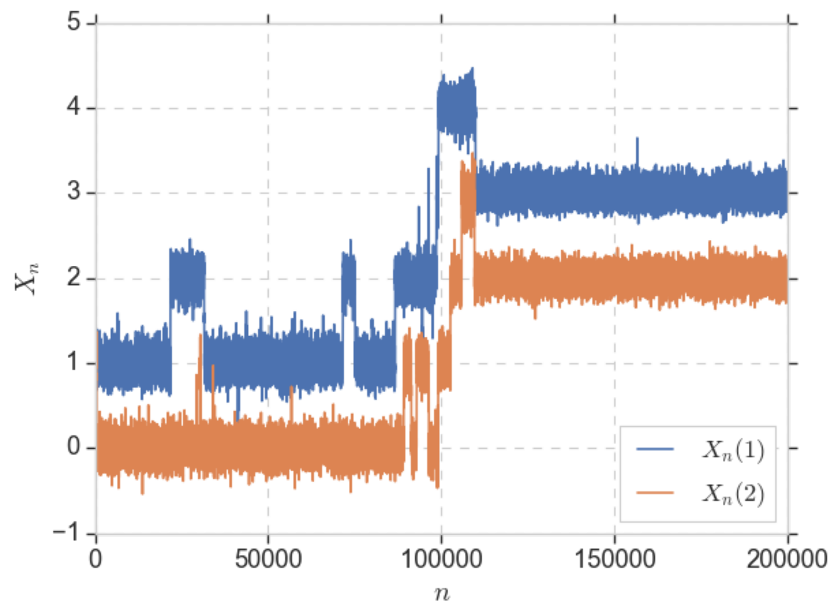
$\gamma = 1, h = 0.1$

LD

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h} Z_n$$



$\gamma = 1$



$\gamma = 0.01$

$h = 0.1$

Online Optimization with Stochastic Gradient

$$F(x) = E_S[f(x,S)]$$

$$\nabla F(x) = E_S[\nabla_x f(x,S)]$$

$$\hat{F}(X_n) = \frac{1}{B}\sum_{i=1}^{B} f(X_n, s_i), \quad \nabla\hat{F}(X_n) = \frac{1}{B}\sum_{i=1}^{B} \nabla_x f(X_n, \tilde{s}_i)$$

Online Optimization with Stochastic Gradient

$$F(x) = E_S[f(x,S)]$$
$$\nabla F(x) = E_S[\nabla_x f(x,S)]$$

$$\hat{F}(X_n) = \frac{1}{B}\sum_{i=1}^{B} f(X_n, s_i), \quad \nabla\hat{F}(X_n) = \frac{1}{B}\sum_{i=1}^{B} \nabla_x f(X_n, \tilde{s}_i)$$

Stochastic Gradient Descent (SGD)

$$X_{n+1} = X_n - h\nabla\hat{F}(X_n)$$

Stochastic Gradient Langevin Dynamics (SGLD)

$$Y_{n+1} = Y_n - h\nabla\hat{F}(Y_n) + \sqrt{2\gamma h}Z_n$$

SGDxSGLD

Algorithm 2: SGDxSGLD: online optimization

Input: Temperature γ, step size h, number of steps N, initial X_0, Y_0, estimation error parameter Θ (when using batch means, Θ is the batch size, it controls the accuracy of $\hat{F}_n$ and $\nabla \hat{F}_n$), threshold t_0, and exchange boundary $\hat{M}_v$.

for $n = 0$ *to* $N-1$ **do**

$X'_{n+1} = X_n - h\nabla \hat{F}_n(X_n)$;

$Y'_{n+1} = Y_n - h\nabla \hat{F}_n(Y_n) + \sqrt{2\gamma h} Z_n$, where $Z_n \sim N(0, I_d)$;

if $\hat{F}_n(Y'_{n+1}) < \hat{F}_n(X'_{n+1}) - t_0$, $\|X'_{n+1}\| \leq \hat{M}_V$, *and* $\|Y'_{n+1}\| \leq \hat{M}_V$ **then**

$(X_{n+1}, Y_{n+1}) = (Y'_{n+1}, X'_{n+1})$;

else

$(X_{n+1}, Y_{n+1}) = (X'_{n+1}, Y'_{n+1})$.

end

end

Output: X_N as an optimizer for F.

SGDxSGLD

Assumption 4. The estimation errors are sub-Gaussian.

SGDxSGLD

Assumption 4. The estimation errors are sub-Gaussian.

Under Assumptions 1,2,3, and 4, assuming F is strongly convex in a neighborhood of X^* and $h<\min\{1/(2L), 1/m\}$, for any ε>0 and δ>0, there exists

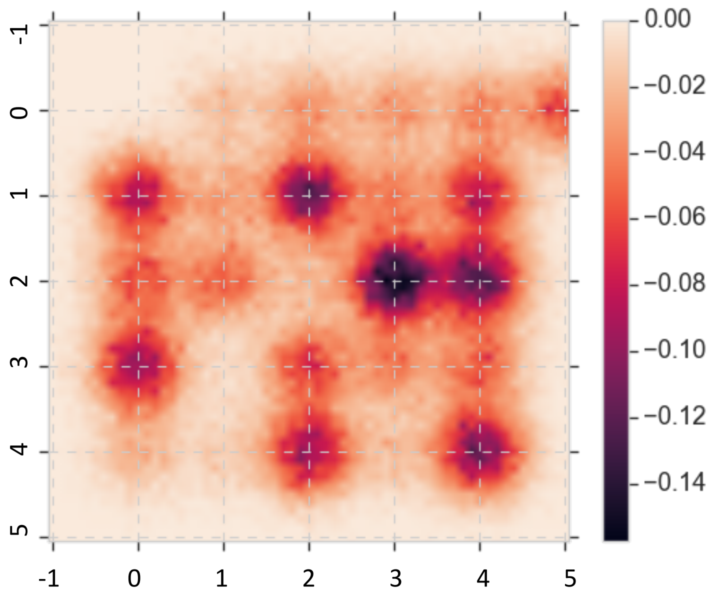
$$N(\varepsilon,\delta) = O(\log(1/\varepsilon)) + O(\log(1/\delta)),$$

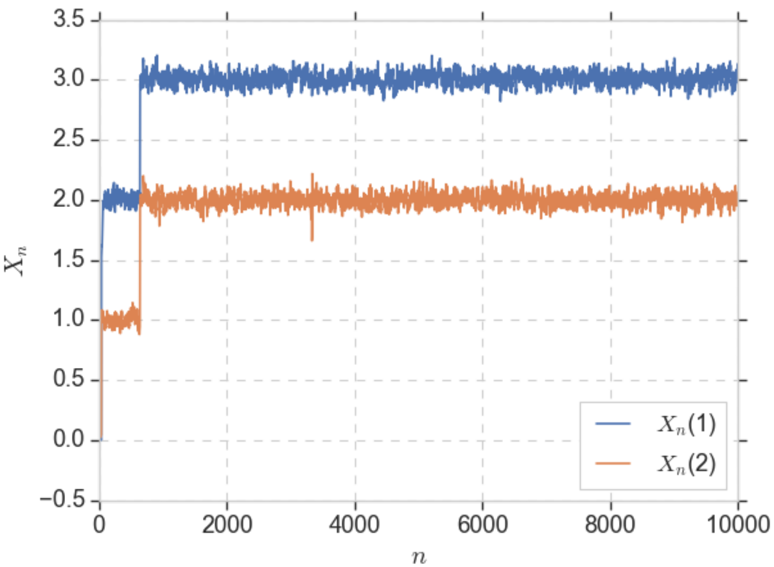
such that for any fixed $N>N(\varepsilon, \delta)$, setting $B = O((\varepsilon\delta)^{-1})$, we have

$$P(F(X_N) - F^* \leq \varepsilon) \geq 1-\delta.$$

If we hold δ and h fixed, then to achieve an ε accuracy, we need to set the number of iterations N=O(log(1/ε)) and the batch size B=O(1/ε). In this case, the total complexity is $O(\varepsilon^{-1}\log(1/\varepsilon))$.

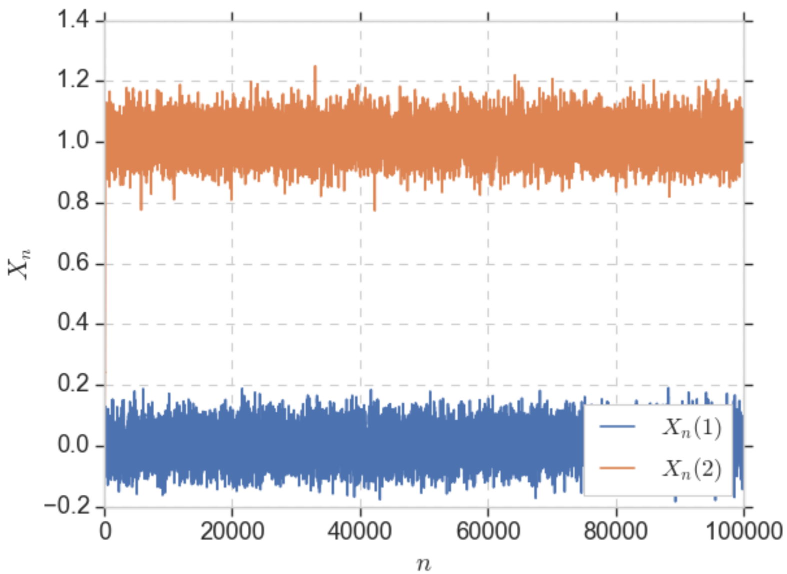
SGDxSGLD





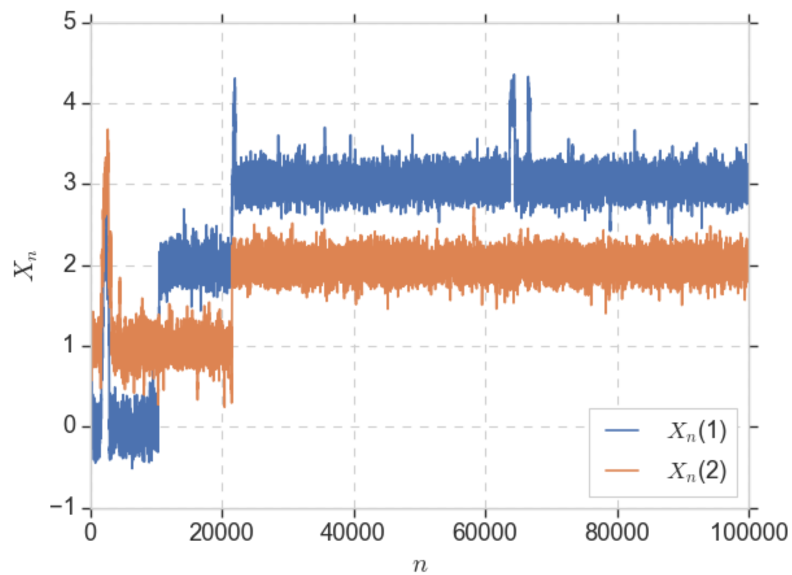
$$\gamma = 1, h = 0.1, B = 10^3, t_0 = 0.05, M = 5$$

SGD and SGLD



SGD

$h = 0.1, B = 10^3$



SGLD

$\gamma = 0.01, h = 0.1, B = 10^3$

Literature Review

Offline: $O(\log(1/\varepsilon))$,   Online: $O(\varepsilon^{-1}\log(1/\varepsilon))$

Finding second order stationary point (local minimums)

- Perturbed Gradient Descent (Jin et al 2017, Jin et al 2019)

$$X_{n+1} = X_n - h\left(\nabla F(X_n) + \frac{r}{\sqrt{d}} Z_n\right) \text{ where } Z_n \sim N(0, I)$$

  - Exact gradient: $O(\varepsilon^{-2})$
  - Stochastic gradient: $O(\varepsilon^{-4})$

- Natasha2 (Allen-Zhu 2017),
- Hessian information: cubic-regularization, trust region (Nestrov and Polyak 2006, Curtis et al 2014, Agarwal et al 2017, Fang et al 2019)

Better dependence on dimension.

Literature Review

Offline: $O(\log(1/\varepsilon))$, Online: $O(\varepsilon^{-1}\log(1/\varepsilon))$

Nonconvex optimization

- SGLD (Dalalyan 2017, Ragingsky et al 2017, Xu et al 2019)
  - Exact gradient: $O(\varepsilon^{-1})$
  - Stochastic gradient: $O(\varepsilon^{-5})$
- Underdamped Langevin dynamics (Cheng et al 2018, Gao et al 2019)

Dependence on the spectral gap

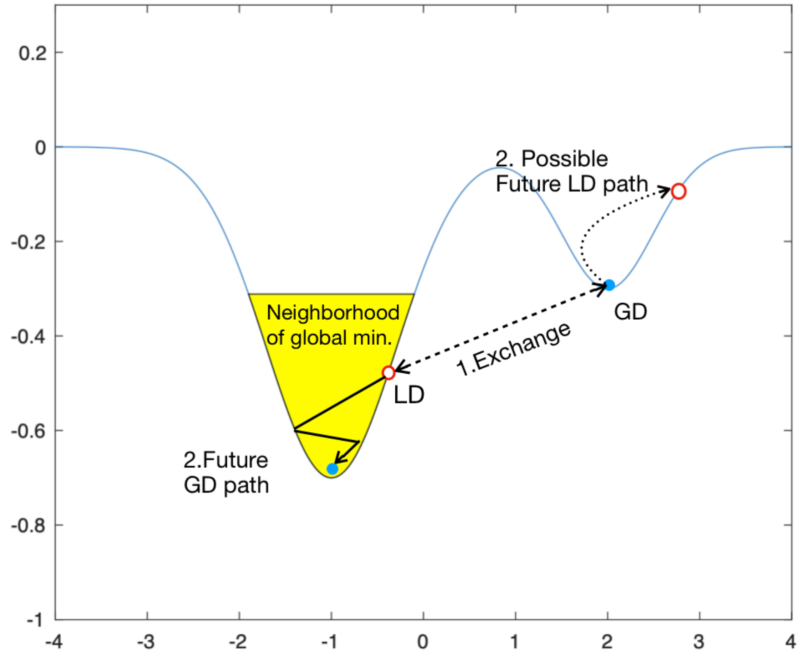
Connection to MCMC

- Replica-exchange Langevin dynamics (Dupuis et al 2012, Chen et al 2019)

Connection to simulated annealing (Gidas 1985, Woodard et al 2009)

Complexity Analysis

Assumption 3. X^* is a unique global minimum. The exists $r_0>0$, such that the sub-level set $B_0=\{x:F(x)\leq F(X^*)+r_0\}$ is radically convex with X^* being the center. F is convex in B_0.



Complexity Analysis

Step 1. There exists a large constant M such that Y visits the set $\{x : \| x - X^* \| \leq M\}$ "very often".

Step 2. During each visit to the set $\{x : \| x - X^* \| \leq M\}$, there is a positive probability that Y will visit B_0.

Step 3. Once Y is in B_0, X will be swapped there (if not there already). Then, the rest of the analysis follows standard gradient descent arguments.

Complexity Analysis: Step 1

$$\tau_k = \{n > \tau_{k-1} : F(Y_n) \leq R\}$$

For a properly chosen parameter η, V(x)=exp(ηF(x)) satisfies

$$E_n[V(Y'_{n+1})] \leq \exp\left(-\frac{1}{4}\eta h \lambda_0 F(Y_n) + \eta h C\right) V(Y_n)$$

$$E_n[V(X'_{n+1})] \leq \exp\left(-\frac{1}{4}\eta h \lambda_0 F(Y_n) + \eta h C\right) V(X_n)$$

Complexity Analysis: Step 1

$$\tau_k = \{n > \tau_{k-1} : F(Y_n) \le R\}$$

For a properly chosen parameter η, V(x)=exp(ηF(x)) satisfies

$$E_n[V(Y'_{n+1})] \le \exp\left(-\frac{1}{4}\eta h \lambda_0 F(Y_n) + \eta h C\right) V(Y_n)$$

$$E_n[V(X'_{n+1})] \le \exp\left(-\frac{1}{4}\eta h \lambda_0 F(Y_n) + \eta h C\right) V(X_n)$$

$\Downarrow$

For any K≥0,

$$E[\exp(\eta h C \tau_K)] \le \exp(K(2\eta h C + \eta R))(V(X_0) + V(Y_0))$$

Complexity Analysis: Step 2

$$\tau_k = \{n > \tau_{k-1} : F(Y_n) \le R\}$$

$$D = \max\{\| x - h\nabla F(x) \| : F(x) \le R\}$$

If $F(Y_n) \le R$, for any $r>0$, there exit an $\alpha(r,D)>0$, such that

$$P_n(\| Y'_{n+1} \| \le r) > \alpha(r,D)$$

A lower bound for $\alpha(r,D)$ is given by

$$\alpha(r,D) \ge \frac{S_d r^d}{(4\gamma h\pi)^{d/2}} \exp\left(-\frac{1}{2\gamma h}(D^2 + r^2) \right)$$

Complexity Analysis

Step 1. There exists a large constant M such that Y visits the set $\{x : \| x - X^* \| \leq M\}$ "very often".

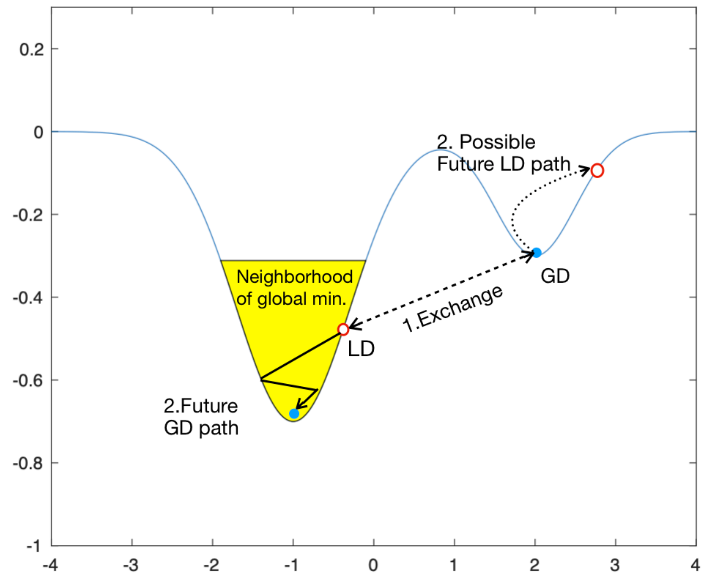
Step 2. During each visit to the set $\{x : \| x - X^* \| \leq M\}$, there is a positive probability that Y will visit B_0.

Step 3. Once Y is in B_0, X will be swapped there (if not there already). Then, the rest of the analysis follows standard gradient descent arguments.

Conclusion

$$X_{n+1} = X_n - h\nabla F(X_n)$$

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h} Z_n$$



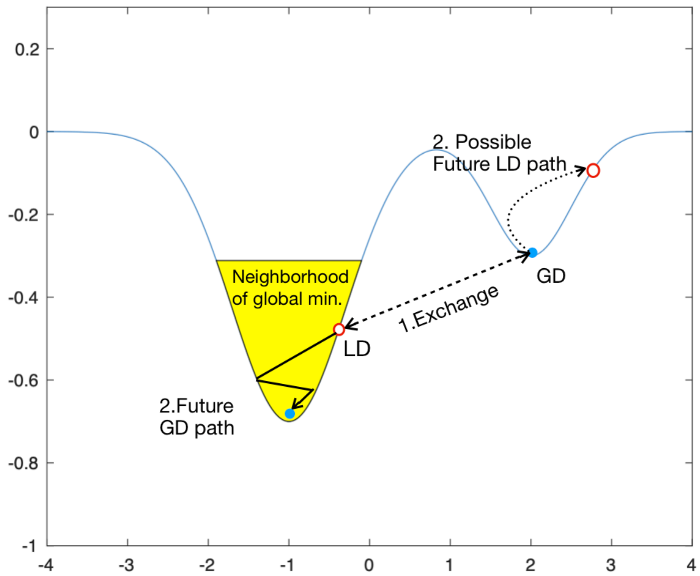
Offline: $O(\log(1/\varepsilon))$, Online: $O(\varepsilon^{-1}\log(1/\varepsilon))$

https://arxiv.org/pdf/2001.08356.pdf

Conclusion

$$X_{n+1} = X_n - h\nabla F(X_n)$$

$$Y_{n+1} = Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h}Z_n$$



Offline: $O(\log(1/\varepsilon))$, Online: $O(\varepsilon^{-1}\log(1/\varepsilon))$

https://arxiv.org/pdf/2001.08356.pdf

Thank you!