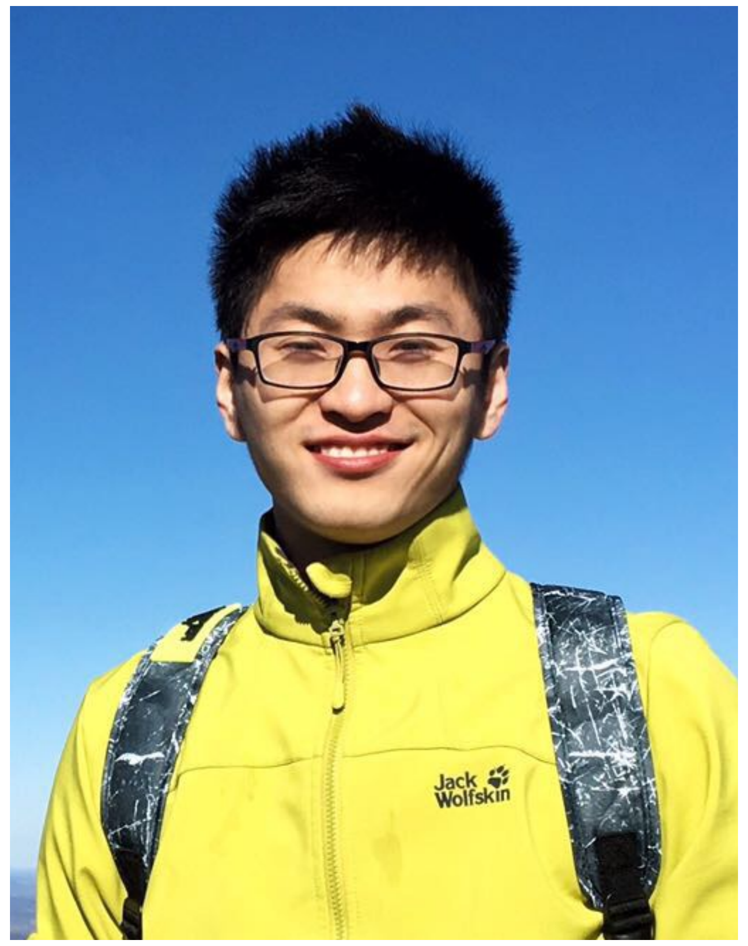
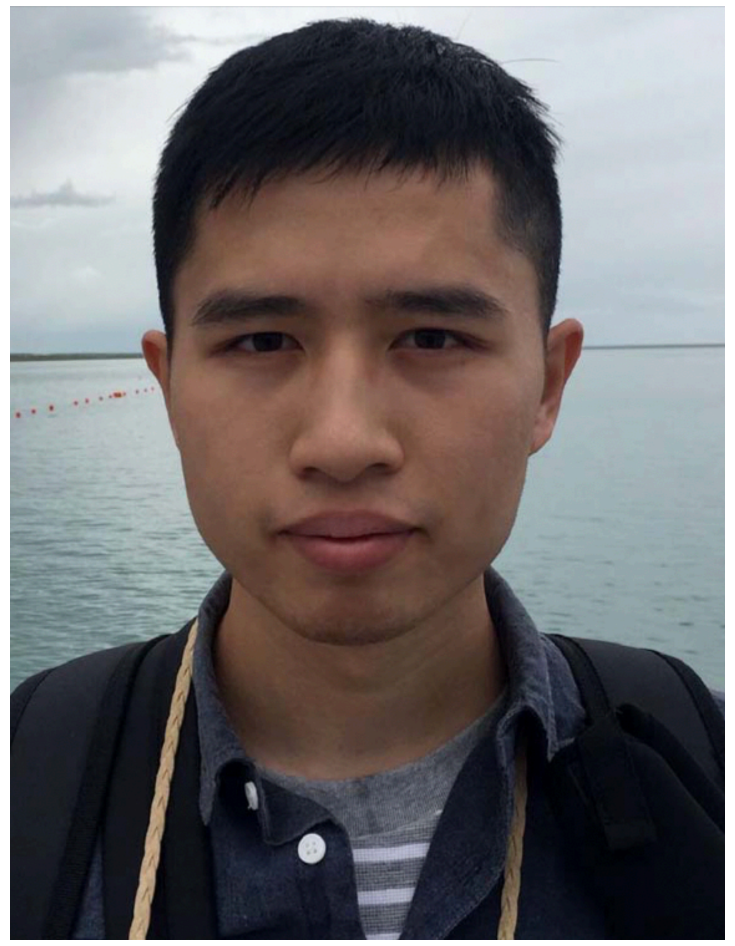
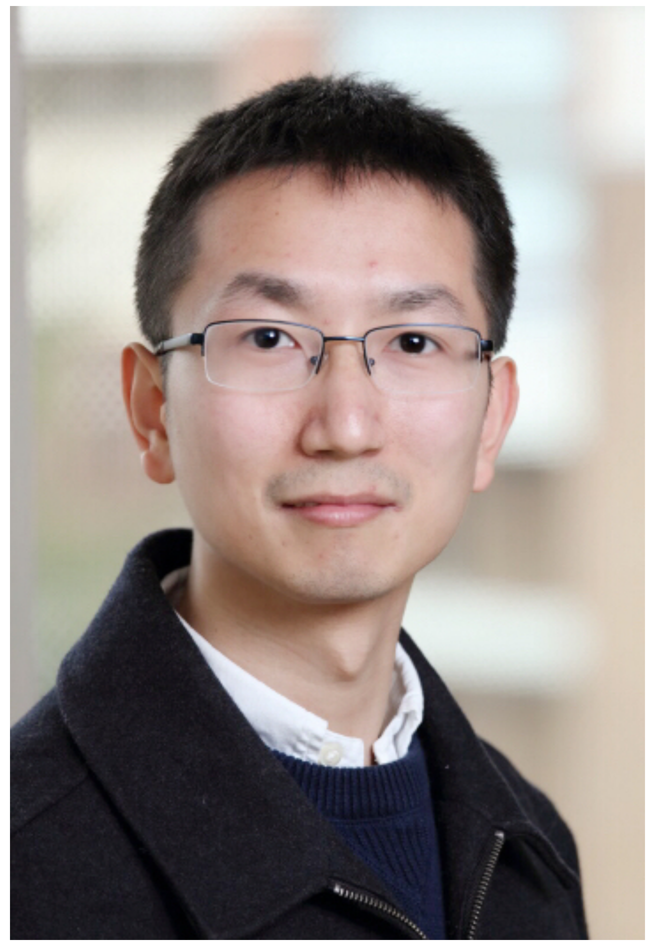
Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization

Pan Xu* — UCLA
Jinghui Chen* — UVa
Difan Zou — UCLA
Quanquan Gu — UCLA

UCLA

UNIVERSITY of VIRGINIA ENGINEERING

Langevin Dynamics

Langevin Dynamics:

$$dX(t) = \underbrace{\boxed{-\nabla F_n(X(t))dt}}_{\text{drift term}} + \underbrace{\boxed{\sqrt{2\beta^{-1}}dB(t)}}_{\text{diffusion term}},$$

- β: inverse temperature parameter
- $B(t)$: standard Brownian motion

Asymptotic property (Roberts & Tweedie, 1996): converges to a stationary distribution

$$\pi(d\mathbf{x}) \propto \exp(-\beta F_n(\mathbf{x}))$$

Implication: The stationary distribution concentrates on the global minima.

Gradient Langevin Dynamics

Langevin Dynamics:

$$d\boldsymbol{X}(t) = -\nabla F_n(\boldsymbol{X}(t))dt + \sqrt{2\beta^{-1}}d\boldsymbol{B}(t),$$

Gradient Langevin Dynamics (GLD, aka. Langevin Monte Carlo):

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta\nabla F_n(\boldsymbol{X}_k) + \sqrt{2\eta\beta^{-1}} \cdot \boldsymbol{\epsilon}_k,$$

- η is the step size
- $\boldsymbol{\epsilon}_k$ is a standard Gaussian random vector

Goal: bound the **Optimization Error**

$$\mathbb{E}[F_n(\boldsymbol{X}_k) - F_n(\mathbf{x}^*)] \qquad \mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}\, F_n(\mathbf{x})$$

Decomposition of Optimization Error

Goal: bound the **Optimization Error** $\mathbb{E}[F_n(\boldsymbol{X}_k) - F_n(\mathbf{x}^*)]$

Decomposition: (Raginsky et al., 2017)

$$X_k \longrightarrow X(t) \longrightarrow X^\pi \longrightarrow \mathbf{x}^*$$

$$\mathbb{E}[F_n(\boldsymbol{X}_k) - F_n(\boldsymbol{X}(k\eta))] \quad \mathbb{E}[F_n(\boldsymbol{X}(k\eta)) - F_n(\boldsymbol{X}^\pi)] \quad \mathbb{E}[F_n(\boldsymbol{X}^\pi) - F_n(\mathbf{x}^*)]$$

Discretization Error **Ergodicity** **Model Error**

Iteration complexity: $\mathbb{E}[F_n(\boldsymbol{X}_n) - F_n(\mathbf{x}^*)] \leq \epsilon + O\left(\frac{d}{\beta}\log\frac{\beta}{d}\right)$

$$k = \widetilde{O}\left(\frac{1}{\epsilon^4 \lambda^{*5}} \log^5 \frac{1}{\epsilon}\right)$$

Model Error

Novel Decomposition for Faster Rates

Goal: bound the **Optimization Error** $\mathbb{E}[F_n(\boldsymbol{X}_k) - F_n(\mathbf{x}^*)]$

Decomposition (this paper):

$$X_k \longrightarrow X^\mu \longrightarrow X^\pi \longrightarrow \mathbf{x}^*$$

$\mathbb{E}[F_n(\boldsymbol{X}_k) - F_n(\boldsymbol{X}^\mu)]$ $\mathbb{E}[F_n(\boldsymbol{X}^\mu) - F_n(\boldsymbol{X}^\pi)]$ $\mathbb{E}[F_n(\boldsymbol{X}^\pi) - F_n(\mathbf{x}^*)]$

Ergodicity (Discrete)

Distance between stationary distributions

Model Error

Iteration complexity: $\mathbb{E}[F_n(\boldsymbol{X}_n) - F_n(\mathbf{x}^*)] \le \epsilon + O\left(\frac{d}{\beta}\log\frac{\beta}{d}\right)$

$k = \widetilde{O}\left(\frac{1}{\epsilon^4\lambda^{*5}}\log^5\frac{1}{\epsilon}\right) \longrightarrow \boxed{k = \widetilde{O}\left(\frac{1}{\epsilon\lambda^*}\log\frac{1}{\epsilon}\right)}$

Model Error

Global Convergence of Variants of GLD

Stochastic Gradient Langevin Dynamics (SGLD):

$$\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k - \eta \nabla G(\boldsymbol{Y}_k) + \sqrt{2\eta\beta^{-1}} \cdot \boldsymbol{\epsilon}_k,$$

- unbiased stochastic gradient $\mathbb{E}[\nabla G(X)|X] = \nabla F_n(X)$

Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD):

$$\boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k - \eta \widetilde{\nabla}_k + \sqrt{2\eta\beta^{-1}} \cdot \boldsymbol{\epsilon}_k,$$

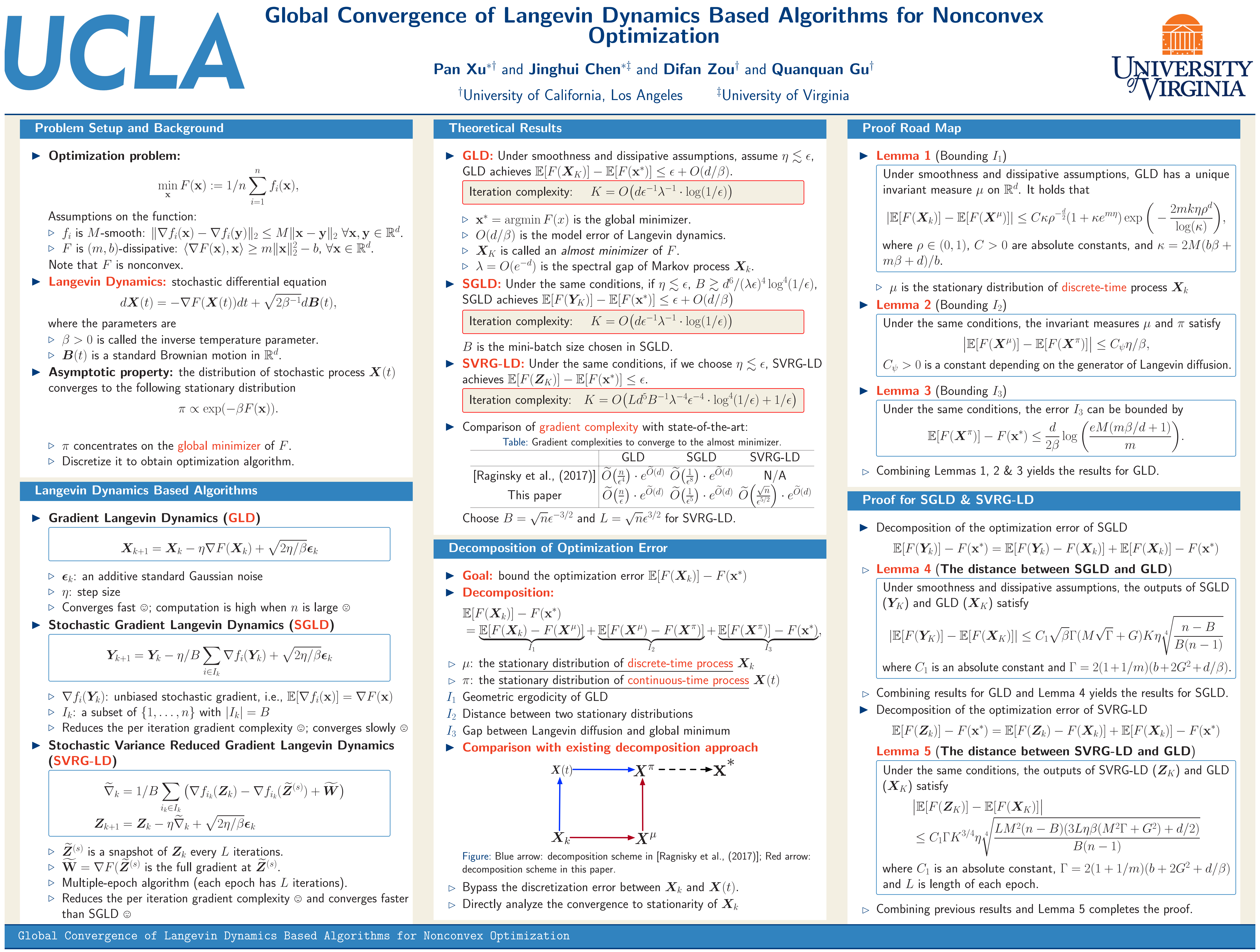
- semi-stochastic gradient $\widetilde{\nabla}_k = \nabla G_k(\boldsymbol{Z}_k) - \nabla G_k(\widetilde{\boldsymbol{Z}}^{(s)}) + \nabla F_n(\widetilde{\boldsymbol{Z}}^{(s)})$
- $\widetilde{Z}^{(s)}$ is a snapshot of Z_k, updated after every m iterations.

Thanks!

Poster session:

10:45 AM -- 12:45 PM

@Room 210 & 230 AB #46



UCLA

Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization

Pan Xu[*†] and **Jinghui Chen**[*‡] and **Difan Zou**[†] and **Quanquan Gu**[†]

[†]University of California, Los Angeles [‡]University of Virginia

UNIVERSITY of VIRGINIA

Problem Setup and Background

- **Optimization problem:**

$$\min_{\mathbf{x}} F(\mathbf{x}) := 1/n \sum_{i=1}^{n} f_i(\mathbf{x}),$$

Assumptions on the function:

- f_i is M-smooth: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \le M\|\mathbf{x}-\mathbf{y}\|_2 \ \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d$.
- F is (m,b)-dissipative: $\langle \nabla F(\mathbf{x}), \mathbf{x}\rangle \ge m\|\mathbf{x}\|_2^2 - b$, $\forall \mathbf{x} \in \mathbb{R}^d$.

Note that F is nonconvex.

- **Langevin Dynamics:** stochastic differential equation

$$d\boldsymbol{X}(t) = -\nabla F(\boldsymbol{X}(t))dt + \sqrt{2\beta^{-1}}d\boldsymbol{B}(t),$$

where the parameters are

- $\beta > 0$ is called the inverse temperature parameter.
- $\boldsymbol{B}(t)$ is a standard Brownian motion in $\mathbb{R}^d$.

- **Asymptotic property:** the distribution of stochastic process $\boldsymbol{X}(t)$ converges to the following stationary distribution

$$\pi \propto \exp(-\beta F(\mathbf{x})).$$

- π concentrates on the global minimizer of F.
- Discretize it to obtain optimization algorithm.

Langevin Dynamics Based Algorithms

- **Gradient Langevin Dynamics (GLD)**

$$\boxed{\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k) + \sqrt{2\eta/\beta}\boldsymbol{\epsilon}_k}$$

- $\boldsymbol{\epsilon}_k$: an additive standard Gaussian noise
- η: step size
- Converges fast ☺; computation is high when n is large ☹

- **Stochastic Gradient Langevin Dynamics (SGLD)**

$$\boxed{\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k - \eta/B \sum_{i\in I_k}\nabla f_i(\boldsymbol{Y}_k) + \sqrt{2\eta/\beta}\boldsymbol{\epsilon}_k}$$

- $\nabla f_i(\boldsymbol{Y}_k)$: unbiased stochastic gradient, i.e., $\mathbb{E}[\nabla f_i(\mathbf{x})] = \nabla F(\mathbf{x})$
- I_k: a subset of $\{1,\ldots,n\}$ with $|I_k| = B$
- Reduces the per iteration gradient complexity ☺; converges slowly ☹

- **Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD)**

$$\widetilde{\nabla}_k = 1/B \sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{Z}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{Z}}^{(s)}) + \widetilde{\boldsymbol{W}}\big)$$
$$\boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k - \eta\widetilde{\nabla}_k + \sqrt{2\eta/\beta}\boldsymbol{\epsilon}_k$$

- $\widetilde{\boldsymbol{Z}}^{(s)}$ is a snapshot of $\boldsymbol{Z}_k$ every L iterations.
- $\widetilde{\mathbf{W}} = \nabla F(\widetilde{\boldsymbol{Z}}^{(s)})$ is the full gradient at $\widetilde{\boldsymbol{Z}}^{(s)}$.
- Multiple-epoch algorithm (each epoch has L iterations).
- Reduces the per iteration gradient complexity ☺ and converges faster than SGLD ☺

Theoretical Results

- **GLD:** Under smoothness and dissipative assumptions, assume $\eta \lesssim \epsilon$, GLD achieves $\mathbb{E}[F(\boldsymbol{X}_K)] - \mathbb{E}[F(\mathbf{x}^*)] \le \epsilon + O(d/\beta)$.

Iteration complexity: $K = O\big(d\epsilon^{-1}\lambda^{-1}\cdot\log(1/\epsilon)\big)$

- $\mathbf{x}^* = \operatorname{argmin} F(x)$ is the global minimizer.
- $O(d/\beta)$ is the model error of Langevin dynamics.
- $\boldsymbol{X}_K$ is called an *almost minimizer* of F.
- $\lambda = O(e^{-d})$ is the spectral gap of Markov process $\boldsymbol{X}_k$.

- **SGLD:** Under the same conditions, if $\eta \lesssim \epsilon$, $B \gtrsim d^6/(\lambda\epsilon)^4\log^4(1/\epsilon)$, SGLD achieves $\mathbb{E}[F(\boldsymbol{Y}_K)] - \mathbb{E}[F(\mathbf{x}^*)] \le \epsilon + O(d/\beta)$

Iteration complexity: $K = O\big(d\epsilon^{-1}\lambda^{-1}\cdot\log(1/\epsilon)\big)$

B is the mini-batch size chosen in SGLD.

- **SVRG-LD:** Under the same conditions, if we choose $\eta \lesssim \epsilon$, SVRG-LD achieves $\mathbb{E}[F(\boldsymbol{Z}_K)] - \mathbb{E}[F(\mathbf{x}^*)] \le \epsilon$.

Iteration complexity: $K = O\big(Ld^5B^{-1}\lambda^{-4}\epsilon^{-4}\cdot\log^4(1/\epsilon) + 1/\epsilon\big)$

- Comparison of gradient complexity with state-of-the-art:

Table: Gradient complexities to converge to the almost minimizer.

	GLD	SGLD	SVRG-LD
[Raginsky et al., (2017)]	$\widetilde{O}\big(\frac{n}{\epsilon^4}\big)\cdot e^{\widetilde{O}(d)}$	$\widetilde{O}\big(\frac{1}{\epsilon^8}\big)\cdot e^{\widetilde{O}(d)}$	N/A
This paper	$\widetilde{O}\big(\frac{n}{\epsilon}\big)\cdot e^{\widetilde{O}(d)}$	$\widetilde{O}\big(\frac{1}{\epsilon^5}\big)\cdot e^{\widetilde{O}(d)}$	$\widetilde{O}\Big(\frac{\sqrt{n}}{\epsilon^{5/2}}\Big)\cdot e^{\widetilde{O}(d)}$

Choose $B = \sqrt{n}\epsilon^{-3/2}$ and $L = \sqrt{n}\epsilon^{3/2}$ for SVRG-LD.

Decomposition of Optimization Error

- **Goal:** bound the optimization error $\mathbb{E}[F(\boldsymbol{X}_k)] - F(\mathbf{x}^*)$
- **Decomposition:**

$$\mathbb{E}[F(\boldsymbol{X}_k)] - F(\mathbf{x}^*) = \underbrace{\mathbb{E}[F(\boldsymbol{X}_k) - F(\boldsymbol{X}^\mu)]}_{I_1} + \underbrace{\mathbb{E}[F(\boldsymbol{X}^\mu) - F(\boldsymbol{X}^\pi)]}_{I_2} + \underbrace{\mathbb{E}[F(\boldsymbol{X}^\pi)] - F(\mathbf{x}^*)}_{I_3},$$

- μ: the stationary distribution of discrete-time process $\boldsymbol{X}_k$
- π: the stationary distribution of continuous-time process $\boldsymbol{X}(t)$

I_1 Geometric ergodicity of GLD

I_2 Distance between two stationary distributions

I_3 Gap between Langevin diffusion and global minimum

- **Comparison with existing decomposition approach**

$$\begin{array}{ccccc} \boldsymbol{X}(t) & \longrightarrow & \boldsymbol{X}^\pi & \dashrightarrow & \mathbf{X}^* \\ \uparrow & & \uparrow & & \\ \boldsymbol{X}_k & \longrightarrow & \boldsymbol{X}^\mu & & \end{array}$$

Figure: Blue arrow: decomposition scheme in [Ragnisky et al., (2017)]; Red arrow: decomposition scheme in this paper.

- Bypass the discretization error between $\boldsymbol{X}_k$ and $\boldsymbol{X}(t)$.
- Directly analyze the convergence to stationarity of $\boldsymbol{X}_k$

Proof Road Map

- **Lemma 1** (Bounding I_1)

Under smoothness and dissipative assumptions, GLD has a unique invariant measure μ on $\mathbb{R}^d$. It holds that

$$|\mathbb{E}[F(\boldsymbol{X}_k)] - \mathbb{E}[F(\boldsymbol{X}^\mu)]| \le C\kappa\rho^{-\frac{d}{2}}(1+\kappa e^{m\eta})\exp\bigg(-\frac{2mk\eta\rho^d}{\log(\kappa)}\bigg),$$

where $\rho \in (0,1)$, $C > 0$ are absolute constants, and $\kappa = 2M(b\beta + m\beta + d)/b$.

- μ is the stationary distribution of discrete-time process $\boldsymbol{X}_k$

- **Lemma 2** (Bounding I_2)

Under the same conditions, the invariant measures μ and π satisfy

$$\big|\mathbb{E}[F(\boldsymbol{X}^\mu)] - \mathbb{E}[F(\boldsymbol{X}^\pi)]\big| \le C_\psi\eta/\beta,$$

$C_\psi > 0$ is a constant depending on the generator of Langevin diffusion.

- **Lemma 3** (Bounding I_3)

Under the same conditions, the error I_3 can be bounded by

$$\mathbb{E}[F(\boldsymbol{X}^\pi)] - F(\mathbf{x}^*) \le \frac{d}{2\beta}\log\bigg(\frac{eM(m\beta/d+1)}{m}\bigg).$$

- Combining Lemmas 1, 2 & 3 yields the results for GLD.

Proof for SGLD & SVRG-LD

- Decomposition of the optimization error of SGLD

$$\mathbb{E}[F(\boldsymbol{Y}_k)] - F(\mathbf{x}^*) = \mathbb{E}[F(\boldsymbol{Y}_k) - F(\boldsymbol{X}_k)] + \mathbb{E}[F(\boldsymbol{X}_k)] - F(\mathbf{x}^*)$$

- **Lemma 4 (The distance between SGLD and GLD)**

Under smoothness and dissipative assumptions, the outputs of SGLD $(\boldsymbol{Y}_K)$ and GLD $(\boldsymbol{X}_K)$ satisfy

$$|\mathbb{E}[F(\boldsymbol{Y}_K)] - \mathbb{E}[F(\boldsymbol{X}_K)]| \le C_1\sqrt{\beta}\Gamma(M\sqrt{\Gamma}+G)K\eta\sqrt[4]{\frac{n-B}{B(n-1)}}$$

where C_1 is an absolute constant and $\Gamma = 2(1+1/m)(b+2G^2+d/\beta)$.

- Combining results for GLD and Lemma 4 yields the results for SGLD.
- Decomposition of the optimization error of SVRG-LD

$$\mathbb{E}[F(\boldsymbol{Z}_k)] - F(\mathbf{x}^*) = \mathbb{E}[F(\boldsymbol{Z}_k) - F(\boldsymbol{X}_k)] + \mathbb{E}[F(\boldsymbol{X}_k)] - F(\mathbf{x}^*)$$

Lemma 5 (The distance between SVRG-LD and GLD)

Under the same conditions, the outputs of SVRG-LD $(\boldsymbol{Z}_K)$ and GLD $(\boldsymbol{X}_K)$ satisfy

$$\big|\mathbb{E}[F(\boldsymbol{Z}_K)] - \mathbb{E}[F(\boldsymbol{X}_K)]\big| \le C_1\Gamma K^{3/4}\eta\sqrt[4]{\frac{LM^2(n-B)(3L\eta\beta(M^2\Gamma+G^2)+d/2)}{B(n-1)}}$$

where C_1 is an absolute constant, $\Gamma = 2(1+1/m)(b+2G^2+d/\beta)$ and L is length of each epoch.

- Combining previous results and Lemma 5 completes the proof.