Langevin Dynamics for sampling and global optimization

Kirill Neklyudov

DeepBayes

Goals of this talk

- Introduction to the Langevin dynamics
- Derive basics (1st half)
- Outline some important results (2nd half)
- Recommend some literature

Langevin Equation

Ito Stochastic Differential Equation (SDE):

$$\boxed{dX(t)} = -\underbrace{\nabla U(X(t))}_{\text{Force}} dt + \sigma \underbrace{\boxed{dBt}}_{\text{Random fluctuations}}$$

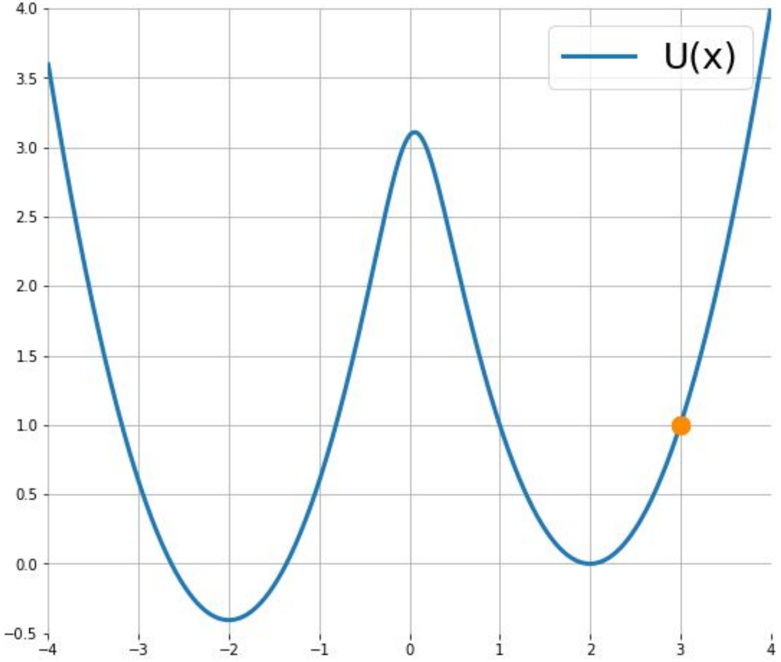
Discrete approximation:

$$\boxed{X_{t+1} - X_t} = -dt\nabla U(X_t) + \sigma \boxed{\sqrt{dt}\mathcal{N}(0,1)}$$

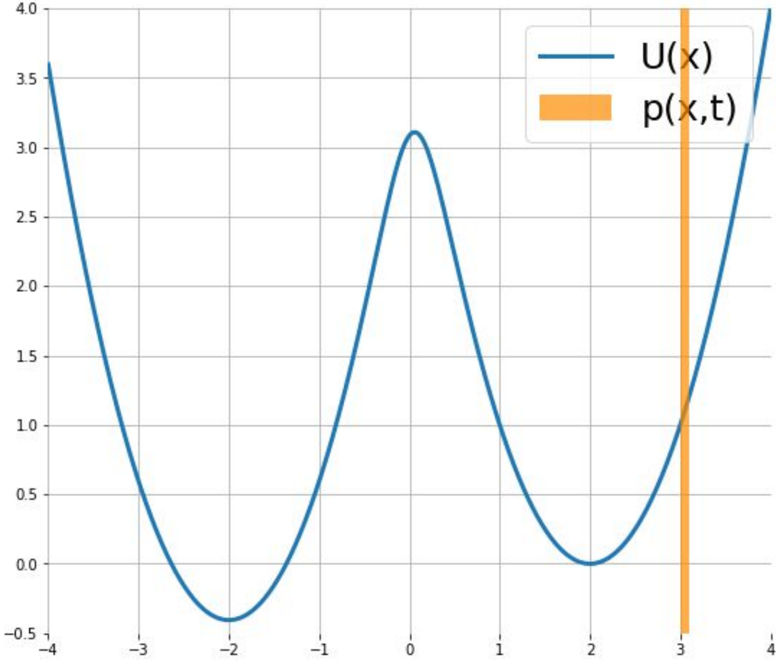
$$W_{t+1} - W_t = -\varepsilon \nabla \mathcal{L}(W_t) + \sigma\sqrt{\varepsilon}\mathcal{N}(0,1)$$

1-d simulation

Langevin equation



Fokker-Planck equation



Derivation of the Fokker-Planck equation

Langevin equation:

$$dX(t) = -\nabla U(X(t))dt + \sigma dBt$$

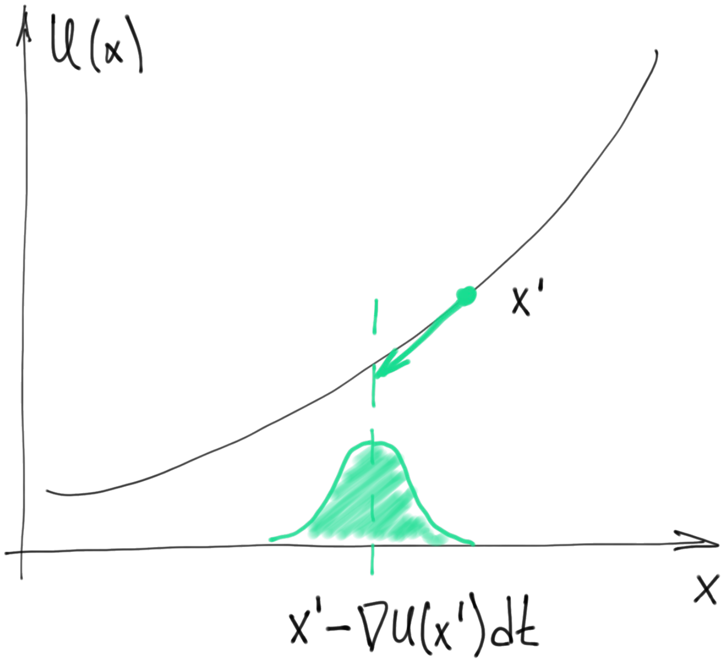
Increments of the Brownian motion:

$$dB_t \sim \mathcal{N}(0, dt \cdot I)$$

Consider a small increment of $X(t)$:

$$x - x' = -dt\nabla U(x') + \mathcal{N}(0, \sigma^2 dt)$$

$$x \sim \mathcal{N}\left(x' - \nabla U(x')dt, \sigma^2 dt\right)$$



Derivation of the Fokker-Planck equation

Unknown density!

Density of particle distribution:

$$p(x,t) = \int dx' p(x,t|x',t-dt)p(x',t-dt)$$

$$p(x,t|x',t-dt) = \frac{1}{(2\pi\sigma^2 dt)^{n/2}} \exp\left(\frac{-(\overbrace{x' - x - \nabla U(x')dt}^{y})^2}{2\sigma^2 dt} \right)$$

Using the change of variables formula, we obtain:

$$p(x,t) = \int dy \left| \frac{\partial x'}{\partial y} \right| \mathcal{N}(y|0,\sigma^2 dt \cdot I)p(x'(y),t-dt)$$

Changing variables

The change of variables:

$$y = x' - x - \nabla U(x')dt$$

$$\left|\frac{\partial x'}{\partial y}\right| = ? \qquad x'(y) = ?$$

We can't invert this

$$y = x' - x - \left(\nabla U(x) + \frac{\partial \nabla U(x)}{\partial x}(x' - x)dt + o(x' - x)\right)dt$$

$$\left(I - \frac{\partial \nabla U(x)}{\partial x}dt\right)x' = y + x + \nabla U(x)dt - \frac{\partial \nabla U(x)}{\partial x}xdt + o(dt)$$

$$x' = \left(I - \frac{\partial \nabla U(x)}{\partial x}dt\right)^{-1}\left(y + x + \nabla U(x)dt - \frac{\partial \nabla U(x)}{\partial x}xdt + o(dt)\right)$$

Changing variables

$$x' = \left(I - \frac{\partial \nabla U(x)}{\partial x} dt\right)^{-1} \left(y + x + \nabla U(x) dt - \frac{\partial \nabla U(x)}{\partial x} x dt + o(dt)\right)$$

$$= \left(\boxed{I} + \boxed{\frac{\partial \nabla U(x)}{\partial x} dt} + o(dt)\right) \left(\boxed{y + x + \nabla U(x) dt - \frac{\partial \nabla U(x)}{\partial x} x dt + o(dt)}\right)$$

$$(1 - A)^{-1} = 1 + A + A^2 + A^3 + \ldots$$

$$= y + x + \nabla U(x) dt - \frac{\partial \nabla U(x)}{\partial x} x dt + \frac{\partial \nabla U(x)}{\partial x} y dt + \frac{\partial \nabla U(x)}{\partial x} x dt + o(dt)$$

$$= x + y + \nabla U(x) dt + \frac{\partial \nabla U(x)}{\partial x} y dt + o(dt)$$

Changing variables

From the previous slide:

$$x' = x + y + \nabla U(x)dt + \frac{\partial \nabla U(x)}{\partial x}\boxed{ydt} + o(dt)$$

$$ydt = (x' - x - \nabla U(x)dt)dt$$

$$x - x' = -dt\nabla U(x') + \mathcal{N}(0, \sigma^2 dt)$$

Equation for the increment
(Langevin equation)

$$ydt = dt\underbrace{\sqrt{dt}\mathcal{N}(0, \sigma^2)}_{dB_t} = o(dt)$$

Note that

$$y = \sqrt{dt}\mathcal{N}(0, \sigma^2) \neq o(dt)$$ since $$\lim_{dt \to 0} \frac{\sqrt{dt}\mathcal{N}(0, \sigma^2)}{dt} = \infty$$

Changing variables

Finally!

$$x' = x + y + \nabla U(x)dt + o(dt)$$

With a little more efforts (homework):

$$\left|\frac{\partial x'}{\partial y}\right| = 1 + \mathrm{div}\nabla U(x)dt + o(dt)$$

Reminder: $\mathrm{div}\overrightarrow{f}(\overrightarrow{x}) = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \ldots + \frac{\partial f_n}{\partial x_n}$

Derivation of the Fokker-Planck equation

$$p(x,t) = \int dy \left|\frac{\partial x'}{\partial y}\right| \mathcal{N}(y|0,\sigma^2 dt \cdot I)\, p(x'(y), t-dt)$$

Increment of the density

$$\left.\begin{aligned} x' &= x + y + \nabla U(x)dt + o(dt) \\ \left|\frac{\partial x'}{\partial y}\right| &= 1 + \mathrm{div}\nabla U(x)dt + o(dt) \end{aligned}\right\}$$

Change of variables

$$p(x,t) = (1 + \mathrm{div}\nabla U(x)dt)\,\overset{???}{\boxed{\mathbb{E}_y \left[p(y + x + \nabla U(x)dt, t - dt) \right]}}$$

$$y \sim \mathcal{N}(0, \sigma^2 dt \cdot I)$$

Derivation of the Fokker-Planck equation

$$\mathbb{E}_y\Big[p(y + x + \nabla U(x)dt, t - dt)\Big] = \mathbb{E}_y\Big[$$

0th order: $p(x,t)+$

1st order: $+\nabla_x p(x,t)(y + \nabla U(x)dt) + \frac{\partial}{\partial t}p(x,t)(-dt)+$

2nd order: $+\frac{1}{2}(y + \nabla U(x)dt)^T \frac{\partial^2 p(x,t)}{\partial x^2}(y + \nabla U(x)dt)\Big]$

Taking the expectation

0th order: $p(x,t)+$

$$\mathbb{E}_y p(x,t) = p(x,t)$$

1st order: $+\nabla_x p(x,t)(y+\nabla U(x)dt)+\frac{\partial}{\partial t}p(x,t)(-dt)+$

$$\mathbb{E}_y\left[\nabla_x p(x,t)^T(y+\nabla U(x)dt)\right] = \nabla_x p(x,t)^T\mathbb{E}_y[y] + dt\nabla_x p(x,t)^T\nabla U(x)$$

$$y \sim \mathcal{N}(0,\sigma^2 dt)$$

$$= 0 + dt\nabla_x p(x,t)^T\nabla U(x)$$

$$\mathbb{E}_y\left[dt\frac{\partial}{\partial t}p(x,t)\right] = dt\frac{\partial}{\partial t}p(x,t)$$

Taking the expectation (2nd order)

$$\mathbb{E}_y\left[(y + \nabla U(x)dt)^T \frac{\partial^2 p(x,t)}{\partial x^2}(y + \nabla U(x)dt)\right] =$$

$$= \mathbb{E}_y\left[y^T \frac{\partial^2 p(x,t)}{\partial x^2} y + 2dt\nabla U(x)^T \frac{\partial^2 p(x,t)}{\partial x^2} y + \boxed{dt^2 \nabla U(x)^T \frac{\partial^2 p(x,t)}{\partial x^2} \nabla U(x)}\right] =$$

$$= \mathbb{E}_y\left[\sum_{i,j}\left(\frac{\partial^2 p(x,t)}{\partial x^2}\right)_{ij} y_i y_j\right] + 2dt\nabla U(x)^T \frac{\partial^2 p(x,t)}{\partial x^2} \underbrace{\boxed{\mathbb{E}_y y}}_{=0} + \boxed{o(dt)} =$$

$$= \underbrace{\boxed{\sum_{i=j}\left(\frac{\partial^2 p(x,t)}{\partial x^2}\right)_{ii}}}_{=\Delta p(x,t)} \underbrace{\boxed{\mathbb{E}_y\left[y_i^2\right]}}_{=\sigma^2 dt} + \sum_{i \neq j}\left(\frac{\partial^2 p(x,t)}{\partial x^2}\right)_{ij} \underbrace{\boxed{\mathbb{E}_y\left[y_i y_j\right]}}_{=0} + o(dt) =$$

Derivation of the Fokker-Planck equation

$$p(x,t) = (1 + \mathrm{div}\nabla U(x)dt)\mathbb{E}_y\left[p(y + x + \nabla U(x)dt, t - dt)\right]$$

Increment of the density

$$p(x,t) = (\boxed{1} + \boxed{\mathrm{div}\nabla U(x)dt})\Big(p(x,t) + dt\nabla_x p(x,t)^T\nabla U(x) - \\ - dt\frac{\partial}{\partial t}p(x,t) + \frac{1}{2}\sigma^2 dt\Delta p(x,t) + o(dt)\Big)$$

Taylor series

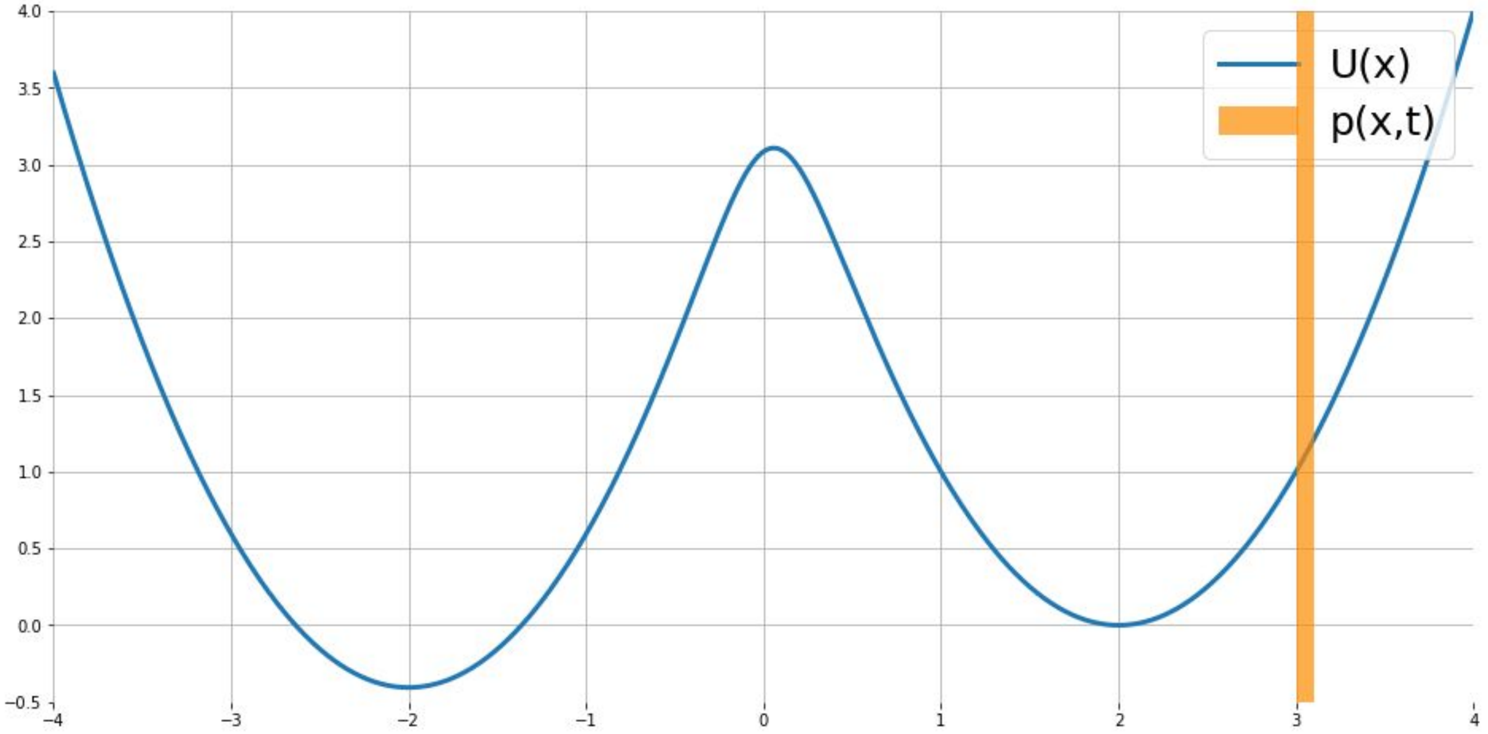
$$\cancel{p(x,t)} = \boxed{\cancel{p(x,t)} + dt\nabla_x p(x,t)^T\nabla U(x) - dt\frac{\partial}{\partial t}p(x,t) + \frac{1}{2}\sigma^2 dt\Delta p(x,t)} + \\ \boxed{+ p(x,t)\mathrm{div}\nabla U(x)dt + o(dt)}$$

$$\frac{\partial}{\partial t}p(x,t) = \nabla_x p(x,t)^T\nabla U(x) + p(x,t)\mathrm{div}\nabla U(x) + \frac{1}{2}\sigma^2\Delta p(x,t) + \cancelto{0}{\frac{o(dt)}{dt}}$$

5 min break

Fokker-Planck equation

$$\frac{\partial}{\partial t} p(x,t) = \nabla_x p(x,t)^T \nabla U(x) + p(x,t) \operatorname{div} \nabla U(x) + \frac{1}{2}\sigma^2 \Delta p(x,t)$$



Stationary distribution of the Langevin dynamics

$$\frac{\partial}{\partial t}p(x,t) = \nabla_x p(x,t)^T \nabla U(x) + p(x,t)\mathrm{div}\nabla U(x) + \frac{1}{2}\sigma^2 \Delta p(x,t)$$

$p(x,t) = \widehat{p}(x)$ (density does not change anymore)

Let $p_G(x) = \frac{1}{Z}\exp\left(-\frac{U(x)}{T}\right), \quad Z = \int dx \exp\left(-\frac{U(x)}{T}\right)$ Gibbs distribution

With a little efforts (homework), we obtain:

$0 = \nabla \widehat{p}(x)^T \nabla U(x) + \widehat{p}(x)\mathrm{div}\nabla U(x) + \frac{1}{2}\sigma^2 \Delta \widehat{p}(x)$ when $T = \frac{\sigma^2}{2}$

Sampling via the Langevin dynamics

$$dX(t) = -\nabla U(X(t))dt + \sigma dBt$$ Langevin equation

Particles have the stationary distribution:

$$p_G(x) = \frac{1}{Z}\exp\left(-\frac{2U(x)}{\sigma^2}\right), \quad Z = \int dx \exp\left(-\frac{2U(x)}{\sigma^2}\right)$$ Gibbs distribution

We want to sample from $p(x) = \dfrac{\widehat{p}(x)}{Z'}$

$$U(x) = -\log p(x), \quad \sigma = \sqrt{2}$$

$$p_G(x) = \frac{1}{Z}\exp\left(-\frac{-2\log p(x)}{2}\right) = p(x), \quad Z = \int dx \exp(\log p(x)) = 1$$

Note! $\nabla U(x) = -\nabla \log p(x) = -\nabla \log \widehat{p}(x) - \nabla \cancel{\log Z'}$

Sampling via the Langevin dynamics

Stochastic Differential Equation for sampling:

$$dX(t) = \nabla \log p(X(t))dt + \sqrt{2}dB_t$$

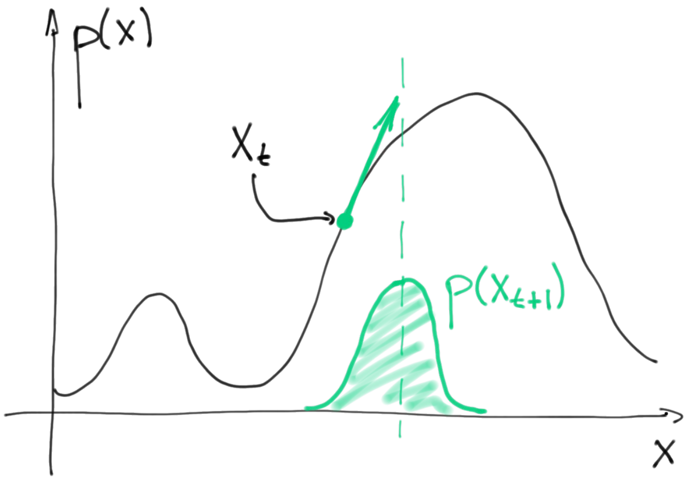
Discrete approximation:

$$X_{t+1} = X_t + dt\nabla \log p(X_t) + \mathcal{N}(0, 2dt)$$

More popular way:

$$X_{t+1} = X_t + \frac{\varepsilon}{2}\nabla \log p(X_t) + \mathcal{N}(0, \varepsilon)$$

$$X_t \sim p(x), \quad \forall\, t > t_\infty$$



Langevin dynamics for the Bayesian inference

Predictive distribution

We need samples

$$p(y|D_{\text{train}}) = \mathbb{E}_{p(\theta|D_{\text{train}})} p(y|\theta) \simeq \frac{1}{K}\sum_{i=1}^{K} p(y|\theta_i), \quad \boxed{\theta_i \sim p(\theta|D_{\text{train}})}$$

$$d\theta(t) = \nabla \log p(\theta(t)|D_{\text{train}})dt + \sqrt{2}dB_t$$

Langevin equation

???

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2}\nabla \log p(\theta_t|D_{\text{train}}) + \mathcal{N}(0,\varepsilon)$$

Discrete approximation

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2}\nabla_\theta\left(\sum_{i=1}^{N} \log p(\theta_t|(x_i, y_i)) + \log p(\theta_t)\right) + \mathcal{N}(0,\varepsilon)$$

???

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2}\nabla_\theta\left(\frac{N}{B}\sum_{k=1}^{B} \log p(\theta_t|(x_{i_k}, y_{i_k})) + \log p(\theta_t)\right) + \mathcal{N}(0,\varepsilon)$$

Borkar, Mitter, 1999

Consider a SDE:

Stationary distribution

$$dX(t) = h(X(t))dt + \sigma dBt \qquad X(t) \sim p_\sigma(x), \quad t > t_\infty$$

Discrete approximation:

Stationary distribution

$$X_{k+1} = X_k + \varepsilon(h(X_k) + M_k) + \mathcal{N}(0, \sigma^2 \varepsilon) \qquad X_k \sim \widehat{p}_\sigma(x), \quad k > k_\infty$$

$$\mathbb{E} M_k = 0, \quad \forall k$$

Theorem

$$\forall\, \delta > 0, \exists\, \varepsilon : \mathrm{KL}(p_\sigma(x) || \widehat{p}_\sigma(x)) < \delta$$

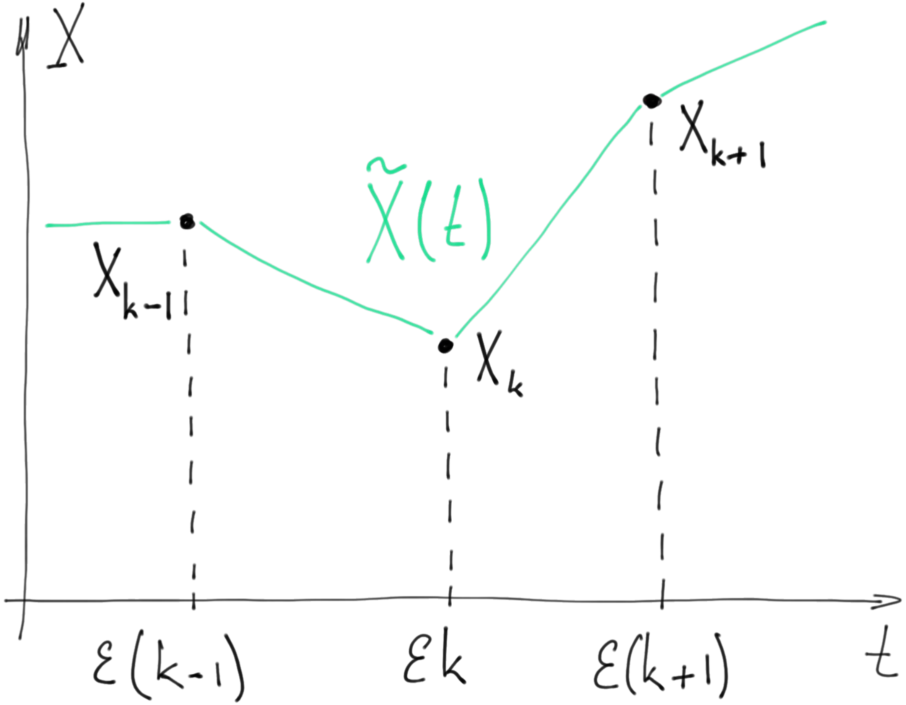
Sketch of the proof

$$\widetilde{X}(t) = X(0) + \int_0^t \left(h(\widetilde{X}(\lfloor s \rfloor_\varepsilon) + \xi_s \right) ds + \sigma \widetilde{B}(t)$$

$$\lfloor s \rfloor_\varepsilon = k\varepsilon, \quad \text{if} \ \ s \in [k\varepsilon, (k+1)\varepsilon)$$

$$\xi_s = M_k, \quad \text{if} \ \ s \in [k\varepsilon, (k+1)\varepsilon)$$

$$\widetilde{B}((k+1)\varepsilon) - \widetilde{B}(k\varepsilon) = \mathcal{N}(0, \varepsilon)$$



Lemma

$$\forall t \ \ \mathbb{E}\left[\|X(t) - \widetilde{X}(t)\|^2 \right] \to 0, \quad \text{as} \quad \varepsilon \to 0$$

What happened to the noise?

$$dX(t) = h(X(t))dt + \sigma dBt$$

Original dynamics

$$X(t) = X(0) + \int_0^t h(X(s))ds + \sigma B(t)$$

The same but integrated

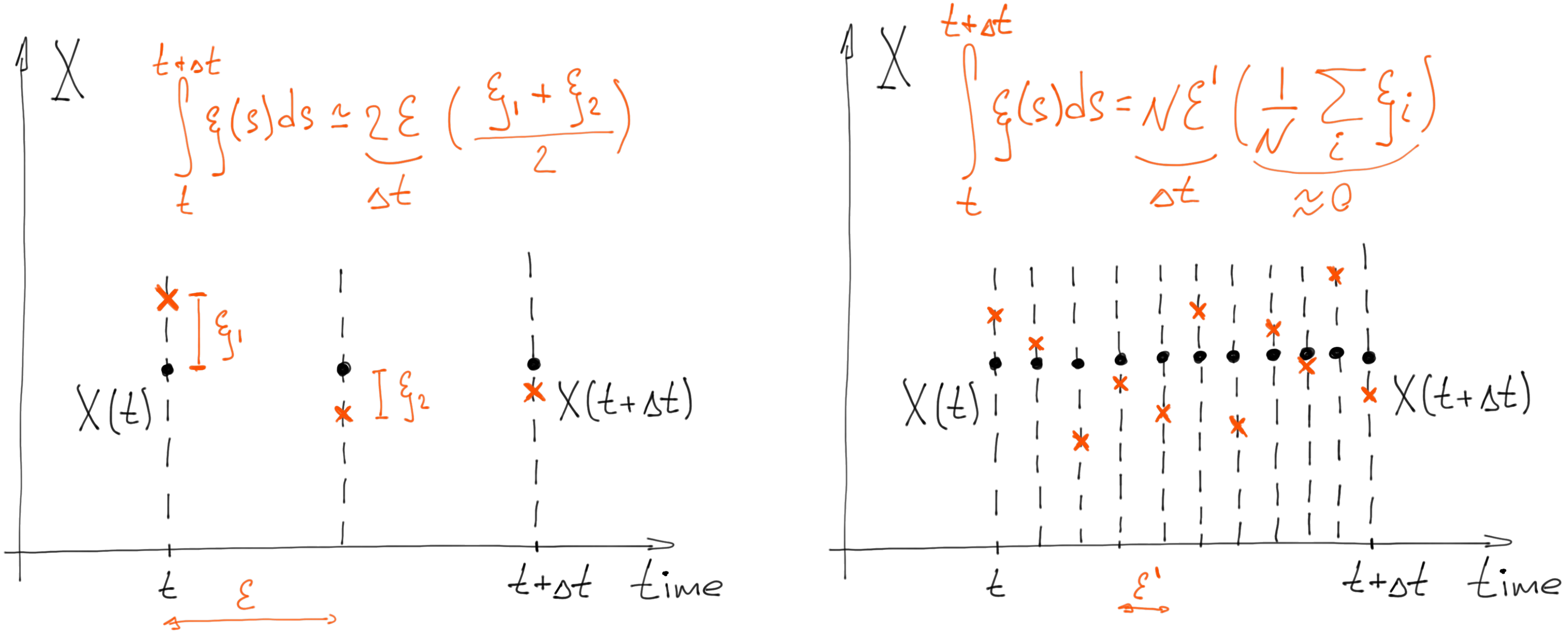
$$\widetilde{X}(t) = X(0) + \int_0^t \left(h(\widetilde{X}(\lfloor s \rfloor_\varepsilon) + \boxed{\xi_s} \right) ds + \sigma \widetilde{B}(t)$$

Our approximation

Free speed-up?

Lemma says $X(t) = \widetilde{X}(t), \quad \text{when} \quad \varepsilon \to 0$

What happened to the noise?



Computational efforts are hidden here

$$X(t) = \widetilde{X}(t), \quad \text{when } \boxed{\varepsilon \to 0}$$

Sketch of the proof

Stationary distribution

$$X(t) \sim p_\sigma(x), \quad t > t_\infty$$

Stationary distribution

$$X_k \sim \widehat{p}_\sigma(x), \quad k > k_\infty$$

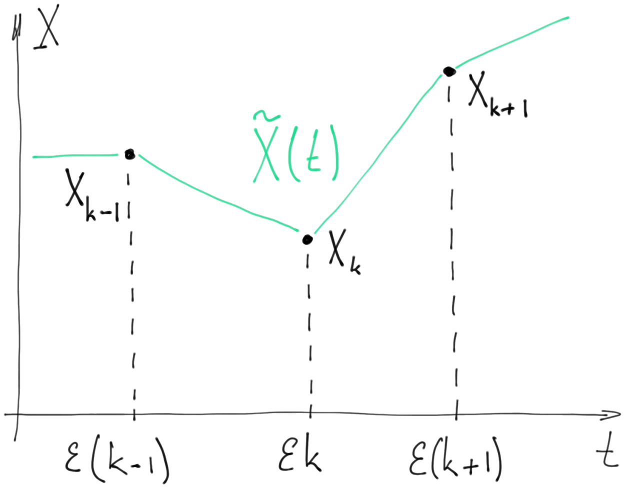
Lemma 1

$$\forall t \;\; \mathbb{E}\left[\|X(t) - \widetilde{X}(t)\|^2\right] \to 0, \quad \text{as} \quad \varepsilon \to 0$$

Lemma 2

$$X(t) \sim p(x,t)$$

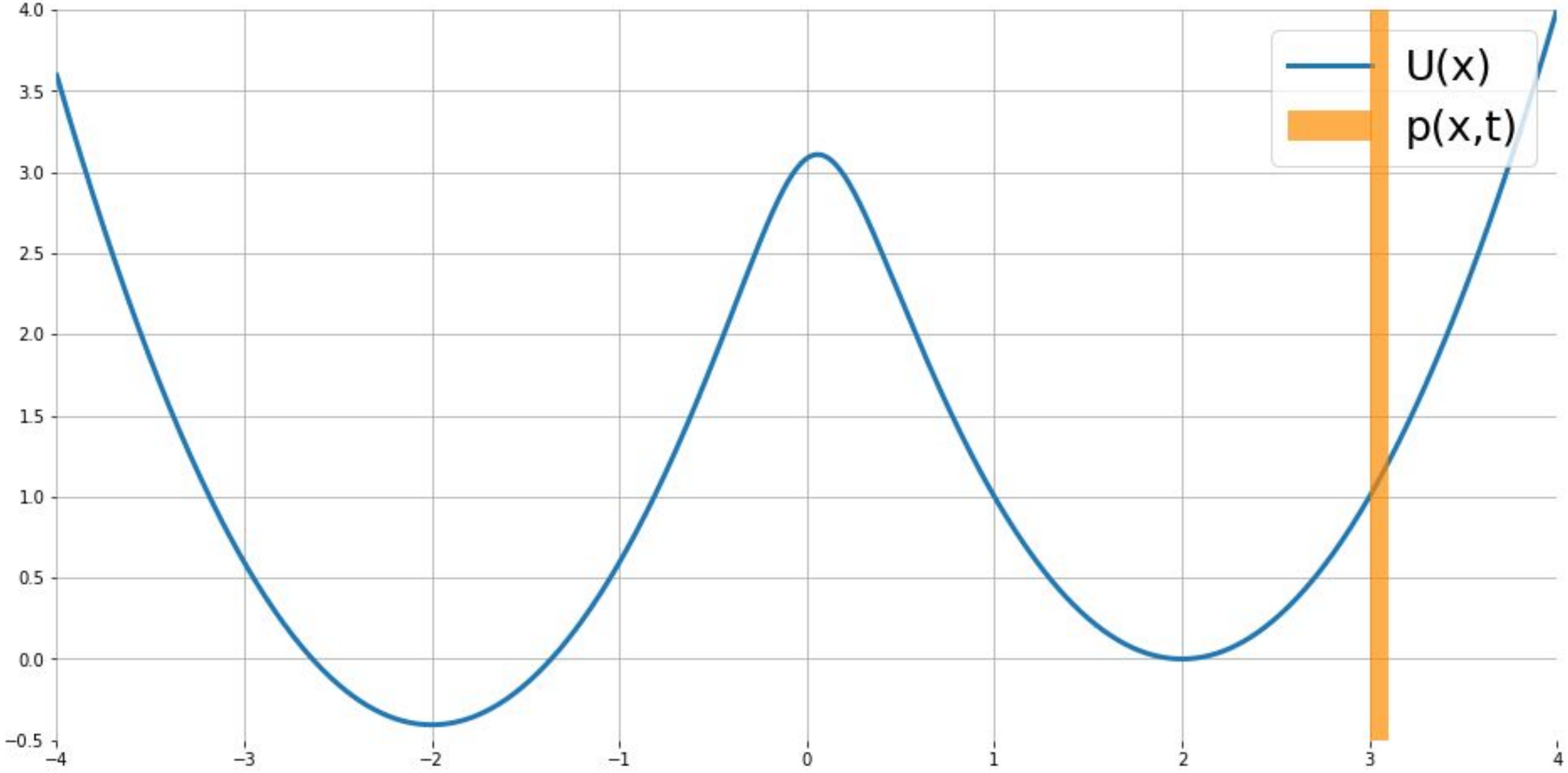
$$\mathrm{KL}(p_\sigma(x)||p(x,t)) \text{ is strictly decreasing in } t$$



Theorem

$$\forall\, \delta > 0, \exists\, \varepsilon : \mathrm{KL}(p_\sigma(x)||\widehat{p}_\sigma(x)) < \delta$$

Global optimization



Temperature annealing

$$dX(t) = -\nabla U(X(t))dt + \sigma dBt$$ Langevin equation

Particles have the stationary distribution:

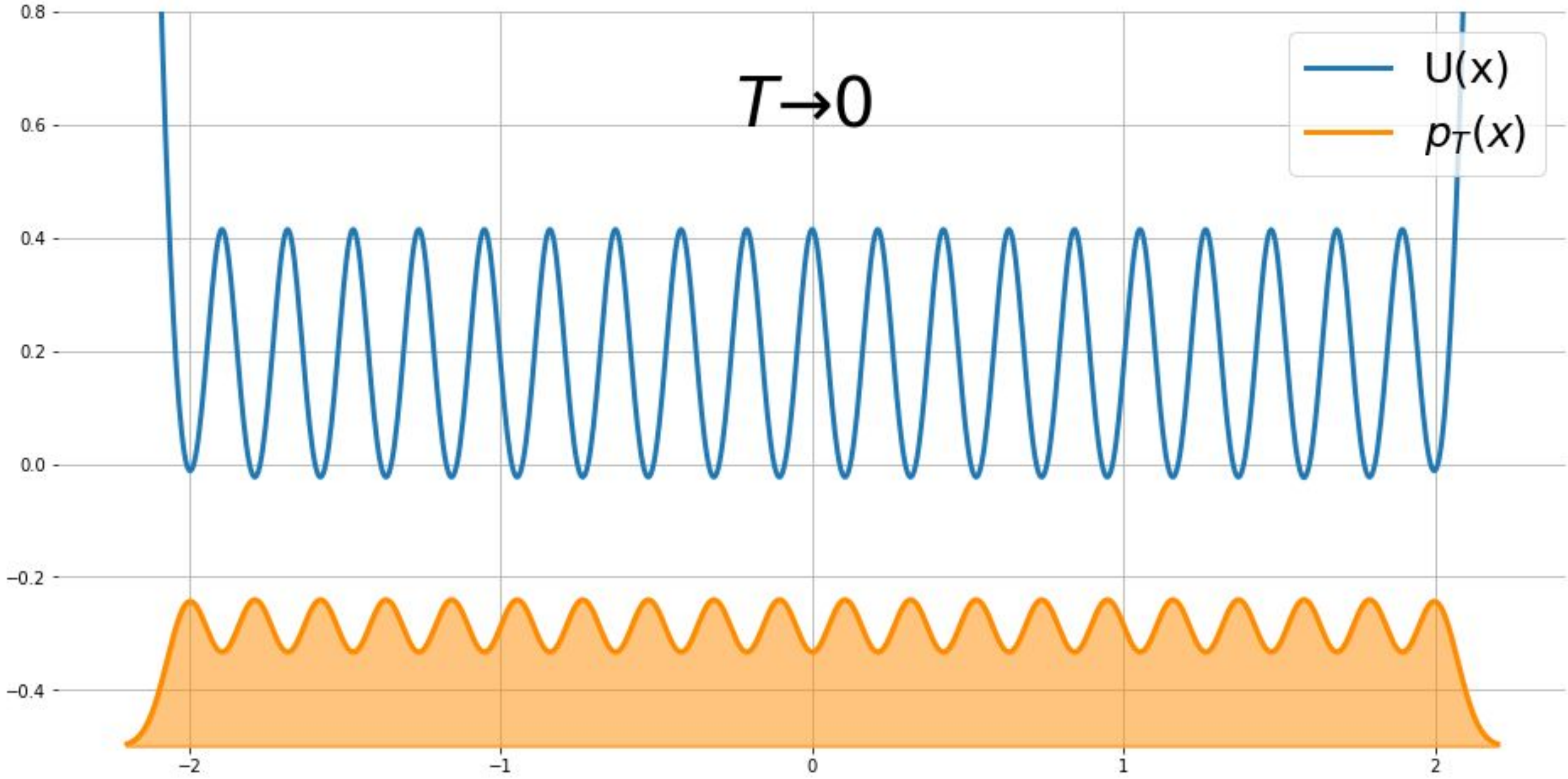
$$p_T(x) = \frac{1}{Z}\exp\left(-\frac{U(x)}{T}\right), \quad Z = \int dx \exp\left(-\frac{U(x)}{T}\right)$$ Gibbs distribution $T = \frac{\sigma^2}{2}$

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2T}dB_t$$

$$p_T(x) = \exp\left(-\frac{U(x)}{T}\right) \xrightarrow[T\to 0]{\mathcal{D}} \pi(x)$$

Concentrates on the global minima

Annealing example



SIAM J. CONTROL AND OPTIMIZATION
Vol. 25, No. 3, May 1987

015

DIFFUSION FOR GLOBAL OPTIMIZATION IN $\mathbb{R}^n$*

TZUU-SHUH CHIANG†, CHII-RUEY HWANG† AND SHUENN-JYI SHEU†

$$\boxed{dX(t) = -\nabla U(X(t))dt + \sqrt{2T(t)}dB_t \quad \text{Langevin equation}}$$

???

$X(t) \sim p(x,t)$ Distribution of particles

$$p_T(x) = \exp\left(-\frac{U(x)}{T}\right) \xrightarrow[T\to 0]{\mathcal{D}} \pi(x)$$

Theorem:

Annealing schedule

$$T(t) = \frac{c}{\log t}$$

$$p(x,t) \xrightarrow[t\to\infty]{P} \pi(x)$$

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 999–1018, September 1991

001

RECURSIVE STOCHASTIC ALGORITHMS FOR GLOBAL OPTIMIZATION IN $\mathbb{R}^d$*

SAUL B. GELFAND† AND SANJOY K. MITTER‡

$$X_{k+1} = X_k - \varepsilon_k(\nabla U(X_k) + M_k) + \sqrt{2T_k}\mathcal{N}(0,1) \qquad \mathbb{E}M_k = 0, \quad \forall k$$

$X_k \sim p(x,k)$ Distribution of particles

$$p_T(x) = \exp\left(-\frac{U(x)}{T}\right) \xrightarrow[T\to 0]{\mathcal{D}} \pi(x)$$

Theorem:

Annealing schedule

$$\varepsilon_k = \frac{c_1}{k} \qquad T_k = \frac{c_2}{k\log\log k} \qquad p(x,k) \xrightarrow[t\to\infty]{P} \pi(x)$$

Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis

Maxim Raginsky[*] Alexander Rakhlin[†] Matus Telgarsky[‡]

Continuous process

$$X(t) \sim p(x, t)$$

Discrete approximation

$$X_k \sim p(x, k)$$

Stationary distribution

$$p_T(x) = \frac{1}{Z} \exp\left(- \frac{U(x)}{T} \right)$$

Discretization error **Convergence speed**

$$W_2\Big(p(x,k), p_T(x)\Big) = W_2\Big(p(x,k), p(x,t)\Big) + W_2\Big(p(x,t), p_T(x)\Big)$$

$$W_2\Big(p(x,k), p_T(x)\Big) = O\Big((C + \varepsilon^{1/4})\varepsilon k\Big) + O\Big(\exp(-\varepsilon k)\Big)$$

Further reading

C.W. Gardiner

Handbook of Stochastic Methods

for Physics, Chemistry and the Natural Sciences

Second Edition
With 29 Figures

Springer

Bernt Øksendal

Stochastic Differential Equations

An Introduction with Applications
Fifth Edition, Corrected Printing
Springer-Verlag Heidelberg New York

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

The end