On the convergence of an improved and adaptive kinetic simulated annealing

Michael Choi

The Chinese University of Hong Kong, Shenzhen
School of Data Science (SDS)

October 2nd 2020

Introduction

- Our focus today is stochastic optimization, in particular simulated annealing algorithms based on Langevin diffusion and its variants. We will highlight connections with sampling throughout the talk.

Introduction

- Our focus today is stochastic optimization, in particular simulated annealing algorithms based on Langevin diffusion and its variants. We will highlight connections with sampling throughout the talk.
- We will talk about a method to accelerate kinetic simulated annealing.

Introduction

- Our focus today is stochastic optimization, in particular simulated annealing algorithms based on Langevin diffusion and its variants. We will highlight connections with sampling throughout the talk.
- We will talk about a method to accelerate kinetic simulated annealing.
- Reference: "On the convergence of an improved and adaptive kinetic simulated annealing" arXiv:2009.00195v2

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be the target function to minimize.

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be the target function to minimize.
- Overdamped Langevin diffusion $(\mathcal{Z}_t)_{t \geq 0}$:

Definition (Overdamped Langevin)

The SDE of overdamped Langevin is given by

$$d\mathcal{Z}_t = -\nabla U(\mathcal{Z}_t)\, dt + \sqrt{2\epsilon_t} dB_t, \tag{1}$$

where $(B_t)_{t \geq 0}$ is the standard d-dimensional Brownian motion and $(\epsilon_t)_{t \geq 0}$ is the temperature or cooling schedule.

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be the target function to minimize.
- Overdamped Langevin diffusion $(\mathcal{Z}_t)_{t \geq 0}$:

Definition (Overdamped Langevin)

The SDE of overdamped Langevin is given by

$$d\mathcal{Z}_t = -\nabla U(\mathcal{Z}_t)\, dt + \sqrt{2\epsilon_t} dB_t, \tag{1}$$

where $(B_t)_{t \geq 0}$ is the standard d-dimensional Brownian motion and $(\epsilon_t)_{t \geq 0}$ is the temperature or cooling schedule.

- The instantaneous stationary distribution at time t is the Gibbs distribution

$$\mu^0_{\epsilon_t}(x) \propto e^{-\frac{1}{\epsilon_t} U(x)}.$$

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be the target function to minimize.
- Overdamped Langevin diffusion $(\mathcal{Z}_t)_{t\geq 0}$:

Definition (Overdamped Langevin)

The SDE of overdamped Langevin is given by

$$d\mathcal{Z}_t = -\nabla U(\mathcal{Z}_t)\, dt + \sqrt{2\epsilon_t} dB_t, \tag{1}$$

where $(B_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion and $(\epsilon_t)_{t\geq 0}$ is the temperature or cooling schedule.

- The instantaneous stationary distribution at time t is the Gibbs distribution

$$\mu^0_{\epsilon_t}(x) \propto e^{-\frac{1}{\epsilon_t}U(x)}.$$

- The overdamped Langevin diffusion is widely used in sampling, e.g. ULA, MALA...

Simulated annealing (SA)

- Convergence of SA depends on a constant E_* that is called the **critical height** or the hill-climbing constant.

Simulated annealing (SA)

- Convergence of SA depends on a constant E_* that is called the **critical height** or the hill-climbing constant.
-

$$E_* := \sup_{x,y\in\mathbb{R}^d} \inf_{\gamma\in\Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\},$$

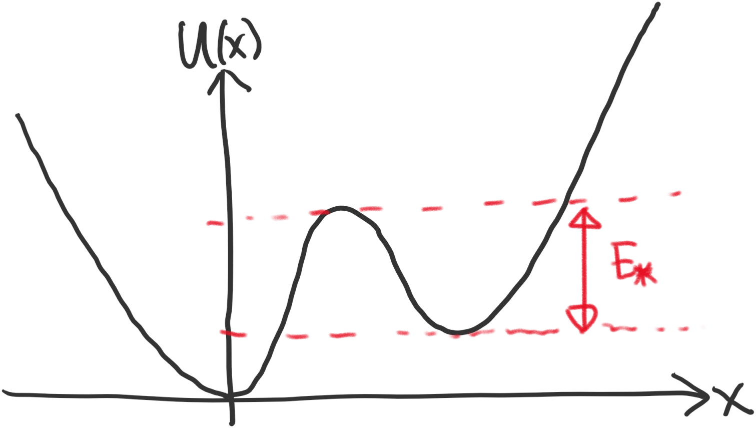
where for two points $x, y \in \mathbb{R}^d$, we write $\Gamma_{x,y}$ to be the set of C^1 parametric curves that start at x and end at y.

Simulated annealing (SA)

- Convergence of SA depends on a constant E_* that is called the **critical height** or the hill-climbing constant.
-
$$E_* := \sup_{x,y\in\mathbb{R}^d} \inf_{\gamma\in\Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\},$$

  where for two points $x, y \in \mathbb{R}^d$, we write $\Gamma_{x,y}$ to be the set of C^1 parametric curves that start at x and end at y.
- Intuitively speaking, E_* is the largest hill one need to climb starting from a local minimum to a fixed global minimum.

What is E_*?



Convergence of SA

Theorem (Convergence of SA (Chiang et al. '87, Holley et al. '89, Jacquot '92, Miclo '92 ...))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough}\, t, \tag{2}$$

where $E > E_*$, *for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(\mathcal{Z}_t) > \inf U + \delta\right) = 0.$$

Kinetic simulated annealing (KSA)

- Overdamped Langevin diffusion is used in SA, which is reversible w.r.t. the Gibbs distribution at each time t.

Kinetic simulated annealing (KSA)

- Overdamped Langevin diffusion is used in SA, which is reversible w.r.t. the Gibbs distribution at each time t.
- Underdamped/kinetic Langevin diffusion is used in KSA that incorporates the velocity or momentum variable.

Kinetic simulated annealing (KSA)

- Overdamped Langevin diffusion is used in SA, which is reversible w.r.t. the Gibbs distribution at each time t.
- Underdamped/kinetic Langevin diffusion is used in KSA that incorporates the velocity or momentum variable.
- As underdamped Langevin is in general non-reversible, this heuristic can hopefully improve the convergence.

Kinetic simulated annealing (KSA)

- Overdamped Langevin diffusion is used in SA, which is reversible w.r.t. the Gibbs distribution at each time t.
- Underdamped/kinetic Langevin diffusion is used in KSA that incorporates the velocity or momentum variable.
- As underdamped Langevin is in general non-reversible, this heuristic can hopefully improve the convergence.
- Non-reversible dynamics have been proposed to accelerate convergence in the context of sampling or optimization, e.g. Bierkens '16, Chen and Hwang '13, Diaconis et al. '00, Duncan et al. '16 '17, Hwang et al. '93 '05 ...

Kinetic simulated annealing (KSA)

- Underdamped Langevin diffusion $(\mathcal{X}_t, \mathcal{Y}_t)_{t \geq 0}$:

Definition (Underdamped Langevin)

The SDE of underdamped Langevin is given by

$$\begin{aligned} d\mathcal{X}_t &= \mathcal{Y}_t \, dt, \\ d\mathcal{Y}_t &= -\frac{1}{\epsilon_t}\mathcal{Y}_t \, dt - \nabla U(\mathcal{X}_t) \, dt + \sqrt{2} \, dB_t, \end{aligned}$$

where $(\mathcal{X}_t)_{t \geq 0}$ stands for the position and $(\mathcal{Y}_t)_{t \geq 0}$ is the velocity or momentum variable.

Kinetic simulated annealing (KSA)

- Underdamped Langevin diffusion $(\mathcal{X}_t, \mathcal{Y}_t)_{t \geq 0}$:

Definition (Underdamped Langevin)

The SDE of underdamped Langevin is given by

$$\begin{aligned} d\mathcal{X}_t &= \mathcal{Y}_t \, dt, \\ d\mathcal{Y}_t &= -\frac{1}{\epsilon_t} \mathcal{Y}_t \, dt - \nabla U(\mathcal{X}_t) \, dt + \sqrt{2} \, dB_t, \end{aligned}$$

where $(\mathcal{X}_t)_{t \geq 0}$ stands for the position and $(\mathcal{Y}_t)_{t \geq 0}$ is the velocity or momentum variable.

- The instantaneous stationary distribution at time t is the product distribution of the Gibbs distribution $\mu^0_{\epsilon_t}$ and the Gaussian distribution with mean 0 and variance ϵ_t:

$$\pi^0_{\epsilon_t}(x, y) \propto e^{-\frac{1}{\epsilon_t} U(x)} e^{-\frac{\|y\|^2}{2\epsilon_t}}.$$

Convergence of KSA

- Non-reversibility of underdamped Langevin imposes technical difficulties in analyzing the convergence of KSA.

Convergence of KSA

- Non-reversibility of underdamped Langevin imposes technical difficulties in analyzing the convergence of KSA.

Theorem (Convergence of KSA (Monmarché '18))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough}\, t,$$

where $E > E_*$, *for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(\mathcal{X}_t) > \inf U + \delta\right) = 0.$$

Improved simulated annealing (ISA)

- Many techniques have been developed in the literature to accelerate the convergence of Langevin diffusion, e.g. preconditioning (Li et al. '16), use of Lévy noise (Simsekli '17), generalized Langevin dynamics (Chak et al. '20), anti-symmetric perturbation of drift (Hwang et al. '93, Duncan et al. '17)...

Improved simulated annealing (ISA)

- Many techniques have been developed in the literature to accelerate the convergence of Langevin diffusion, e.g. preconditioning (Li et al. '16), use of Lévy noise (Simsekli '17), generalized Langevin dynamics (Chak et al. '20), anti-symmetric perturbation of drift (Hwang et al. '93, Duncan et al. '17)...
- In our talk today we will focus on a variant of overdamped Langevin diffusion with **state-dependent** diffusion coefficient, introduced by Fang et al. (SPA '97)

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t\geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

$$dZ_t = -\nabla U(Z_t)\,dt + \sqrt{2\left(f((U(Z_t)-c)_+) + \epsilon_t\right)}\,dB_t. \tag{3}$$

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t \geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

$$dZ_t = -\nabla U(Z_t)\, dt + \sqrt{2\left(f((U(Z_t) - c)_+) + \epsilon_t\right)}\, dB_t. \tag{3}$$

- Two parameters are introduced:
  - c: It is chosen such that $c > \inf U$
  - $f : \mathbb{R} \to \mathbb{R}^+$ twice-differentiable, non-negative, bounded and non-decreasing with $f(0) = f'(0) = f''(0) = 0$.

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t\geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

$$dZ_t = -\nabla U(Z_t)\,dt + \sqrt{2\left(f((U(Z_t)-c)_+) + \epsilon_t\right)}\,dB_t. \tag{3}$$

- Two parameters are introduced:
  - c: It is chosen such that $c > \inf U$
  - $f : \mathbb{R} \to \mathbb{R}^+$ twice-differentiable, non-negative, bounded and non-decreasing with $f(0) = f'(0) = f''(0) = 0$.
- The instantaneous stationary distribution at time t is

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+) + \epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+) + \epsilon_t}\,du\right)$$

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t \geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

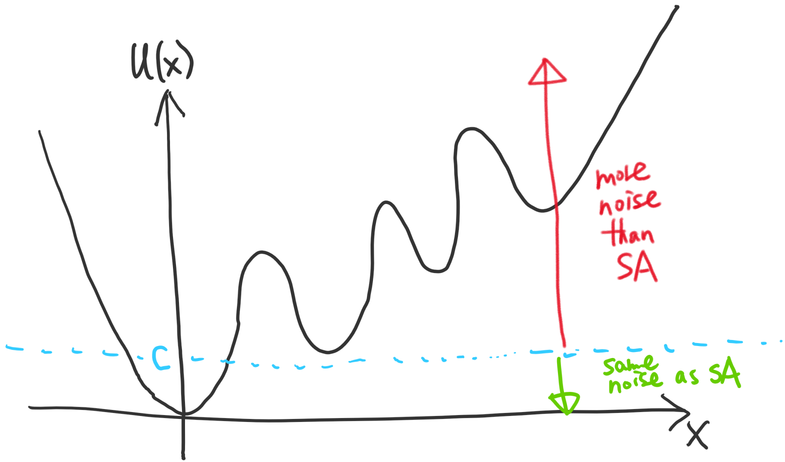
$$dZ_t = -\nabla U(Z_t)\,dt + \sqrt{2\left(f((U(Z_t) - c)_+) + \epsilon_t\right)}\,dB_t. \tag{3}$$

- Two parameters are introduced:
  - c: It is chosen such that $c > \inf U$
  - $f : \mathbb{R} \to \mathbb{R}^+$ twice-differentiable, non-negative, bounded and non-decreasing with $f(0) = f'(0) = f''(0) = 0$.
- The instantaneous stationary distribution at time t is

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x) - c)_+) + \epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u - c)_+) + \epsilon_t}\,du\right)$$

- If $f = 0$, then $\sqrt{2\left(f((U(Z_t) - c)_+) + \epsilon_t\right)} = \sqrt{2\epsilon_t}$, which reduces to the classical overdamped Langevin.

Idea of ISA



Convergence of ISA

- The idea of using state-dependent noise makes sense intuitively. However, is there convergence guarantee that this improved Langevin dynamics ISA converge faster?

Convergence of ISA

- The idea of using state-dependent noise makes sense intuitively. However, is there convergence guarantee that this improved Langevin dynamics ISA converge faster?
- Yes.

Convergence of ISA

- The idea of using state-dependent noise makes sense intuitively. However, is there convergence guarantee that this improved Langevin dynamics ISA converge faster?
- Yes.

Theorem (Convergence of ISA (Fang et al. '97))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough } t,$$

where $E > c_*$, *for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(Z_t) > \inf U + \delta\right) = 0.$$

Convergence of ISA

- The idea of using state-dependent noise makes sense intuitively. However, is there convergence guarantee that this improved Langevin dynamics ISA converge faster?
- Yes.

Theorem (Convergence of ISA (Fang et al. '97))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough } t,$$

where $E > c_*$, *for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(Z_t) > \inf U + \delta\right) = 0.$$

- Key ingredient in the proof: both the spectral gap and the log-Sobolev constant are of the order $\mathcal{O}\left(\exp\left\{\frac{c_*}{\epsilon_t}\right\}\right)$.

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y\in\mathbb{R}^d} \inf_{\gamma\in\Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

- The **clipped critical height** c_* is defined to be

$$c_* := \sup_{x,y\in\mathbb{R}^d} \inf_{\gamma\in\Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\wedge c\} - U(x)\wedge c - U(y)\wedge c + \inf U \right\}.$$

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

- The **clipped critical height** c_* is defined to be

$$c_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t)) \wedge c\} - U(x) \wedge c - U(y) \wedge c + \inf U \right\}.$$

- One way to understand c_*: pretend that we are minimizing $U \wedge c$ instead!

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

- The **clipped critical height** c_* is defined to be

$$c_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t)) \wedge c\} - U(x) \wedge c - U(y) \wedge c + \inf U \right\}.$$

- One way to understand c_*: pretend that we are minimizing $U \wedge c$ instead!
- We can show that the following two statements hold:
  - $c_* \leq E_*$
  - $c_* \leq c - \inf U$

Attempt #1: add state-dependent noise to the position

- Let's try casting the idea of state-dependent noise to kinetic simulated annealing.

Attempt #1: add state-dependent noise to the position

- Let's try casting the idea of state-dependent noise to kinetic simulated annealing.
- Attempt #1: add state-dependent noise to the position. Consider the following dynamics:

$$dX_t = Y_t\,dt + \sqrt{f((U(X_t) - c)_+)}\,dB_t,$$
$$dY_t = -\frac{1}{\epsilon_t} Y_t\,dt - \nabla U(X_t)\,dt + \sqrt{2}\,dB_t.$$

Attempt #1: add state-dependent noise to the position

- Let's try casting the idea of state-dependent noise to kinetic simulated annealing.
- Attempt #1: add state-dependent noise to the position. Consider the following dynamics:

$$dX_t = Y_t\,dt + \sqrt{f((U(X_t) - c)_+)}\,dB_t,$$
$$dY_t = -\frac{1}{\epsilon_t} Y_t\,dt - \nabla U(X_t)\,dt + \sqrt{2}\,dB_t.$$

- The above SDE is no longer degenerate: Brownian noise is added to both the position and momentum update.

Attempt #1: add state-dependent noise to the position

- Let's try casting the idea of state-dependent noise to kinetic simulated annealing.
- Attempt #1: add state-dependent noise to the position. Consider the following dynamics:

$$dX_t = Y_t\, dt + \sqrt{f((U(X_t) - c)_+)}\, dB_t,$$
$$dY_t = -\frac{1}{\epsilon_t} Y_t\, dt - \nabla U(X_t)\, dt + \sqrt{2}\, dB_t.$$

- The above SDE is no longer degenerate: Brownian noise is added to both the position and momentum update.
- The resulting instantaneous stationary distribution in x does not correspond to $\mu_{\epsilon_t}^f$

Attempt #1: add state-dependent noise to the position

- Let's try casting the idea of state-dependent noise to kinetic simulated annealing.
- Attempt #1: add state-dependent noise to the position. Consider the following dynamics:

$$dX_t = Y_t\,dt + \sqrt{f((U(X_t) - c)_+)}\,dB_t,$$
$$dY_t = -\frac{1}{\epsilon_t} Y_t\,dt - \nabla U(X_t)\,dt + \sqrt{2}\,dB_t.$$

- The above SDE is no longer degenerate: Brownian noise is added to both the position and momentum update.
- The resulting instantaneous stationary distribution in x does not correspond to $\mu^f_{\epsilon_t}$
- It seems adding state-dependent noise to the position is not the right direction...

Attempt #2: add state-dependent noise to the momentum

- Attempt #2: add state-dependent noise to the momentum. Consider the following dynamics:

$$dX_t = Y_t\,dt,$$
$$dY_t = -Y_t\,dt - \nabla U(X_t)\,dt + \sqrt{2(f((U(X_t) - c)_+) + \epsilon_t)}\,dB_t.$$

Attempt #2: add state-dependent noise to the momentum

- Attempt #2: add state-dependent noise to the momentum. Consider the following dynamics:

$$
\begin{aligned}
dX_t &= Y_t\, dt, \\
dY_t &= -Y_t\, dt - \nabla U(X_t)\, dt + \sqrt{2(f((U(X_t) - c)_+) + \epsilon_t)}\, dB_t.
\end{aligned}
$$

Attempt #2: add state-dependent noise to the momentum

- Attempt #2: add state-dependent noise to the momentum. Consider the following dynamics:

$$dX_t = Y_t\,dt,$$
$$dY_t = -Y_t\,dt - \nabla U(X_t)\,dt + \sqrt{2(f((U(X_t)-c)_+) + \epsilon_t)}\,dB_t.$$

- This changes the instantaneous stationary distribution in y, but not in x

Attempt #2: add state-dependent noise to the momentum

- Attempt #2: add state-dependent noise to the momentum. Consider the following dynamics:

$$dX_t = Y_t \, dt,$$
$$dY_t = -Y_t \, dt - \nabla U(X_t) \, dt + \sqrt{2(f((U(X_t) - c)_+) + \epsilon_t)} \, dB_t.$$

- This changes the instantaneous stationary distribution in y, but not in x
- It seems adding state-dependent noise to momentum is again not the right direction...

Attempt #3: change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

Attempt #3: change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

- Let's define H_{ϵ_t}:

$$H_\epsilon(x) := \int_{U_{min}}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon}\, du + \ln\left(f((U(x)-c)_+)+\epsilon\right).$$

so that

$$\mu^f_{\epsilon_t}(x) \propto e^{-H_{\epsilon_t}(x)}.$$

Attempt #3: change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

- Let's define H_{ϵ_t}:

$$H_\epsilon(x) := \int_{U_{min}}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon}\, du + \ln\left(f((U(x)-c)_+)+\epsilon\right).$$

  so that

$$\mu^f_{\epsilon_t}(x) \propto e^{-H_{\epsilon_t}(x)}.$$

- In SA,

$$\mu^0_{\epsilon_t}(x) \propto e^{-(1/\epsilon_t)U(x)}.$$

  We can understand as if the optimization landscape is modified from $(1/\epsilon_t)U(x)$ to $H_{\epsilon_t}(x)$.

Attempt #3: change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

- Let's define H_{ϵ_t}:

$$H_\epsilon(x) := \int_{U_{min}}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon}\, du + \ln\left(f((U(x)-c)_+)+\epsilon\right).$$

so that

$$\mu^f_{\epsilon_t}(x) \propto e^{-H_{\epsilon_t}(x)}.$$

- In SA,

$$\mu^0_{\epsilon_t}(x) \propto e^{-(1/\epsilon_t)U(x)}.$$

We can understand as if the optimization landscape is modified from $(1/\epsilon_t)U(x)$ to $H_{\epsilon_t}(x)$.

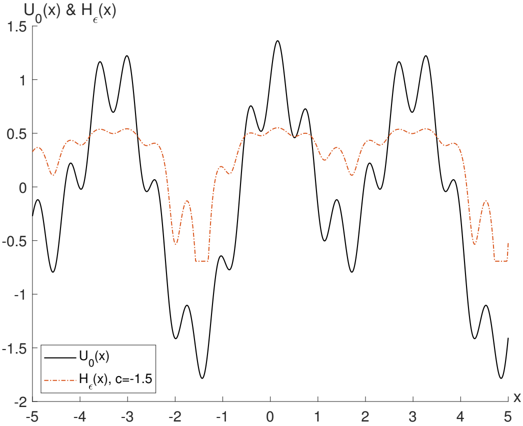
- The idea of state-dependent noise is embedded in the modified optimization landscape.

Idea of IKSA: landscape modification

- Consider the function

$$U_0(x) = \cos(2x) + \frac{1}{2}\sin(x) + \frac{1}{3}\sin(10x).$$

We take $\epsilon = 0.5$, $c = -1.5$ and $f = \arctan$.



Landscape modification in the wild

Image source: `https://kdlandscapingandsnowplowingbuffalo.com/renovation-landscape-modification/`

Improved kinetic simulated annealing (IKSA)

- Let's replace U by $\epsilon_t H_{\epsilon_t}$ in KSA and call the resulting dynamics IKSA.
- Improved kinetic Langevin diffusion $(X_t, Y_t)_{t \geq 0}$:

Definition (Improved kinetic Langevin)

The SDE of improved kinetic Langevin is given by

$$\begin{aligned} dX_t &= Y_t \, dt, \\ dY_t &= -\frac{1}{\epsilon_t} Y_t \, dt - \epsilon_t \nabla H_{\epsilon_t}(X_t) \, dt + \sqrt{2} \, dB_t. \end{aligned}$$

Improved kinetic simulated annealing (IKSA)

- Let's replace U by $\epsilon_t H_{\epsilon_t}$ in KSA and call the resulting dynamics IKSA.
- Improved kinetic Langevin diffusion $(X_t, Y_t)_{t \geq 0}$:

Definition (Improved kinetic Langevin)

The SDE of improved kinetic Langevin is given by

$$dX_t = Y_t\,dt,$$
$$dY_t = -\frac{1}{\epsilon_t} Y_t\,dt - \epsilon_t \nabla H_{\epsilon_t}(X_t)\,dt + \sqrt{2}\,dB_t.$$

- This method can be understood as **state-dependent preconditioning** of the gradient. While it is difficult to compute H_{ϵ_t}, luckily computing its gradient is feasible:

$$\nabla_x H_\epsilon = \frac{1 + f'((U(x) - c)_+)}{f((U(x) - c)_+) + \epsilon} \nabla_x U.$$

Note that H_ϵ and U share the same set of stationary points.

Improved kinetic simulated annealing (IKSA)

Definition (Improved kinetic Langevin)

The SDE of improved kinetic Langevin is given by

$$\begin{aligned} dX_t &= Y_t\, dt, \\ dY_t &= -\frac{1}{\epsilon_t} Y_t\, dt - \epsilon_t \nabla H_{\epsilon_t}(X_t)\, dt + \sqrt{2}\, dB_t. \end{aligned}$$

- The instantaneous stationary distribution at time t is the product distribution of $\mu^f_{\epsilon_t}$ and a Gaussian distribution with mean 0 and variance ϵ_t:

$$\pi^f_{\epsilon_t}(x, y) \propto \mu^f_{\epsilon_t}(x) e^{-\frac{\|y\|^2}{2\epsilon_t}} \propto e^{-H_{\epsilon_t}(x)} e^{-\frac{\|y\|^2}{2\epsilon_t}}.$$

Improved kinetic simulated annealing (IKSA)

Definition (Improved kinetic Langevin)

The SDE of improved kinetic Langevin is given by

$$\begin{aligned} dX_t &= Y_t\, dt, \\ dY_t &= -\frac{1}{\epsilon_t} Y_t\, dt - \epsilon_t \nabla H_{\epsilon_t}(X_t)\, dt + \sqrt{2}\, dB_t. \end{aligned}$$

- The instantaneous stationary distribution at time t is the product distribution of $\mu^f_{\epsilon_t}$ and a Gaussian distribution with mean 0 and variance ϵ_t:

$$\pi^f_{\epsilon_t}(x,y) \propto \mu^f_{\epsilon_t}(x) e^{-\frac{\|y\|^2}{2\epsilon_t}} \propto e^{-H_{\epsilon_t}(x)} e^{-\frac{\|y\|^2}{2\epsilon_t}}.$$

- If $f = 0$, then $\nabla U(X_t) = \epsilon_t \nabla H_{\epsilon_t}(X_t)$, which reduces to the classical kinetic Langevin.

Convergence of IKSA

- The idea of running kinetic simulated annealing on a modified landscape makes sense intuitively. However, are there results that prove this so-called improved kinetic Langevin dynamics IKSA converge faster?

Convergence of IKSA

- The idea of running kinetic simulated annealing on a modified landscape makes sense intuitively. However, are there results that prove this so-called improved kinetic Langevin dynamics IKSA converge faster?
- Yes.

Convergence of IKSA

- The idea of running kinetic simulated annealing on a modified landscape makes sense intuitively. However, are there results that prove this so-called improved kinetic Langevin dynamics IKSA converge faster?
- Yes.

Theorem (Convergence of IKSA (Choi '20))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough}\, t,$$

where $E > c_*$, for any $\delta > 0$ we have

$$\lim_{t \to \infty} \mathbb{P}\left(U(X_t) > \inf U + \delta\right) = 0.$$

Convergence of IKSA

- The idea of running kinetic simulated annealing on a modified landscape makes sense intuitively. However, are there results that prove this so-called improved kinetic Langevin dynamics IKSA converge faster?
- Yes.

Theorem (Convergence of IKSA (Choi '20))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough } t,$$

where $E > c_*$, for any $\delta > 0$ we have

$$\lim_{t \to \infty} \mathbb{P}\left(U(X_t) > \inf U + \delta\right) = 0.$$

- The proof relies on the framework introduced in Monmarché '18.

IAKSA

- The convergence of IKSA depends on the parameter $c > \inf U$. Ideally we want to choose c to be close to $\inf U$, but it can be hard to achieve in practice.

IAKSA

- The convergence of IKSA depends on the parameter $c > \inf U$. Ideally we want to choose c to be close to $\inf U$, but it can be hard to achieve in practice.
- To tune the parameter c, we use the running minimum generated by the algorithm on the fly by setting

$$c_t = \min_{0 \leq u \leq t} U(X_u).$$

IAKSA

- The convergence of IKSA depends on the parameter $c > \inf U$. Ideally we want to choose c to be close to $\inf U$, but it can be hard to achieve in practice.
- To tune the parameter c, we use the running minimum generated by the algorithm on the fly by setting

$$c_t = \min_{0 \leq u \leq t} U(X_u).$$

- $c_{*,t}$ is now time-dependent, and we have to choose $E > c_{*,t}$.

IAKSA

- The convergence of IKSA depends on the parameter $c > \inf U$. Ideally we want to choose c to be close to $\inf U$, but it can be hard to achieve in practice.
- To tune the parameter c, we use the running minimum generated by the algorithm on the fly by setting

$$c_t = \min_{0 \leq u \leq t} U(X_u).$$

- $c_{*,t}$ is now time-dependent, and we have to choose $E > c_{*,t}$.
- Picture to have in mind: the landscape is adaptively improving as the algorithm progresses.

IAKSA

- The convergence of IKSA depends on the parameter $c > \inf U$. Ideally we want to choose c to be close to $\inf U$, but it can be hard to achieve in practice.
- To tune the parameter c, we use the running minimum generated by the algorithm on the fly by setting

$$c_t = \min_{0 \leq u \leq t} U(X_u).$$

- $c_{*,t}$ is now time-dependent, and we have to choose $E > c_{*,t}$.
- Picture to have in mind: the landscape is adaptively improving as the algorithm progresses.
- The resulting diffusion is **non-Markovian**, and belongs to the class of **self-interacting diffusions**.

IAKSA

Theorem (Convergence of IAKSA (Choi '20))

Consider the dynamics

$$dX_t = Y_t\,dt,$$
$$dY_t = -\frac{1}{\epsilon_t} Y_t\,dt - \epsilon_t \nabla H_{\epsilon_t, c_t}(X_t)\,dt + \sqrt{2}\,dB_t.$$

where $c_t = \min_{0 \leq u \leq t} U(X_u)$. *Under the logarithmic cooling schedule of the form*

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough}\, t,$$

where $E > c_{*,t}$, *for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(X_t) > \inf U + \delta\right) = 0.$$

Numerical results

- We compare the following Langevin-based annealing algorithms on some standard global optimization benchmark functions:
  - IAKSA
  - IASA, i.e. ISA with the same f and c_t in IAKSA
  - KSA
  - SA
- We adopt the Euler-Maruyama discretization and use $f = \arctan$, suggested by Fang et al. '97.
- For further details on the parameters used, please refer to the paper.

Numerical results

- We plot $\log_{10} \mathbb{P}\left(\min_{v \leq t} U(X_v) > \inf U + \delta\right)$ or $\log_{10} \mathbb{P}\left(\min_{v \leq t} U(Z_v) > \inf U + \delta\right)$ against $\log_{10} t$, and similarly we plot $\log_{10} \mathbb{P}\left(U(X_t) > \inf U + \delta\right)$ or $\log_{10} \mathbb{P}\left(U(Z_t) > \inf U + \delta\right)$ against $\log_{10} t$. To compute these probabilities, we run 100 independent replicas and count the proportion of replicas for which $U(X_t) > \inf U + \delta$ or $\min_{v \leq t} U(X_v) > \inf U + \delta$.
- We inject the same sequence of Gaussian noise in each of the 100 replicas across all four annealing methods for fair comparison.

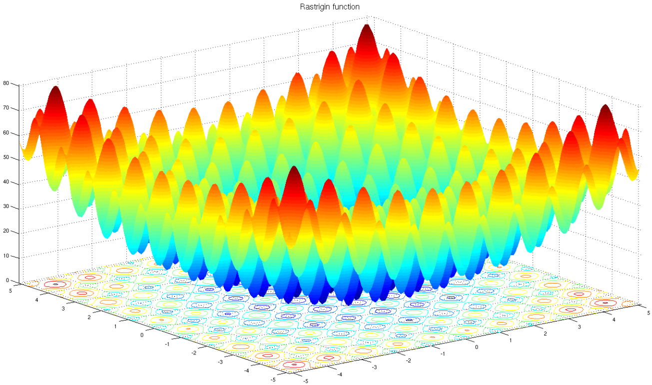
Rastrigin function

- The two-dimensional Rastrigin function:

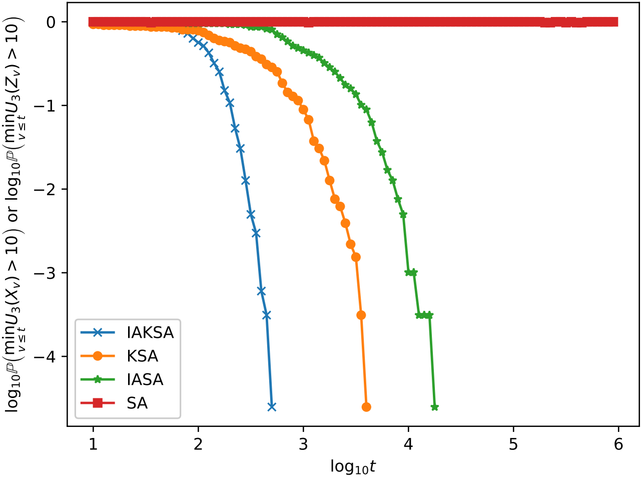
$$U_3(x_1, x_2) = 20 + \sum_{i=1}^{2} \left[x_i^2 - 10\cos(2\pi x_i)\right]$$

Image source: Wikipedia

`https://en.wikipedia.org/wiki/Rastrigin_function`



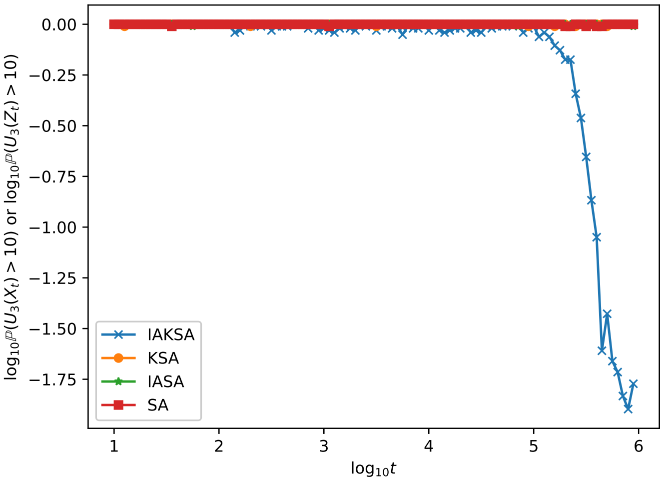
Rastrigin function



$\log_{10}\mathbb{P}\left(\min_{v\leq t} U_3(X_v) > \inf U + \delta\right)$ or

$\log_{10}\mathbb{P}\left(\min_{v\leq t} U_3(Z_v) > \inf U + \delta\right)$ against $\log_{10} t$

Rastrigin function



$$\log_{10} \mathbb{P}\left(U_3(X_t) > \inf U + \delta\right) \text{ or } \log_{10} \mathbb{P}\left(U_3(Z_t) > \inf U + \delta\right)$$

against $\log_{10} t$

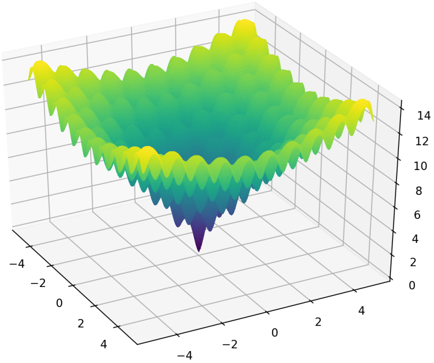
Ackley function

- The two-dimensional Ackley function:

$$U_1(x_1, x_2) = -20 \exp\left(-0.2\sqrt{\frac{1}{2}\sum_{i=1}^{2} x_i^2}\right) - \exp\left(\frac{1}{2}\sum_{i=1}^{2}\cos(2\pi x_i)\right) + 20 + e$$

Image source: PyPi

`https://pypi.org/project/landscapes/#ackley-function`



Ackley function

https://streamable.com/e/yeeftx

Use of state-dependent noise

- There seems to be very limited literature of state-dependent noise in stochastic optimization
- Some work that I am aware of: Fang et al. (SPA '97), Stuart and Mattingly (MPRF '02), Guo et al. '20
- This work hopes to promote the idea of state-dependent noise in sampling and optimization

Landscape modification

- **Stochastic** perspective: the use of state-dependent noise can be understood as a variance reduction technique

Landscape modification

- **Stochastic** perspective: the use of state-dependent noise can be understood as a variance reduction technique
- **Optimization** perspective: this is in some sense "equivalent" to changing the target function from U to $\epsilon_t H_{\epsilon_t}$

Landscape modification

- **Stochastic** perspective: the use of state-dependent noise can be understood as a variance reduction technique
- **Optimization** perspective: this is in some sense "equivalent" to changing the target function from U to $\epsilon_t H_{\epsilon_t}$
- In importance sampling we sample from alternative distribution for "better" sampling. In landscape modification we optimize an alternative function for "better" landscape.

Landscape modification

- **Stochastic** perspective: the use of state-dependent noise can be understood as a variance reduction technique
- **Optimization** perspective: this is in some sense "equivalent" to changing the target function from U to $\epsilon_t H_{\epsilon_t}$
- In importance sampling we sample from alternative distribution for "better" sampling. In landscape modification we optimize an alternative function for "better" landscape.
- Can other variance reduction techniques for Langevin diffusion give new landscape modification?

Landscape modification

- **Stochastic** perspective: the use of state-dependent noise can be understood as a variance reduction technique
- **Optimization** perspective: this is in some sense "equivalent" to changing the target function from U to $\epsilon_t H_{\epsilon_t}$
- In importance sampling we sample from alternative distribution for "better" sampling. In landscape modification we optimize an alternative function for "better" landscape.
- Can other variance reduction techniques for Langevin diffusion give new landscape modification?
- Conversely, can landscape modifcation gives new insights to variance reduction?

Image source: https://kdlandscapingandsnowplowingbuffalo.com/renovation-landscape-modification/

Thank you! Question(s)?