Global Optimization Algorithm through High-Resolution Sampling

Daniel Cortild, N. Oudjane, C. Delplancke, J. Peypouquet

PGMO Days, Paris, November 19th, 2024

Problem Statement

We consider minimization problems of the following form: Given a (possibly nonconvex) smooth potential $U\colon \mathbb{R}^d \to \mathbb{R}$, find

$$x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} U(x).$$

This framework does not include constrained problems!

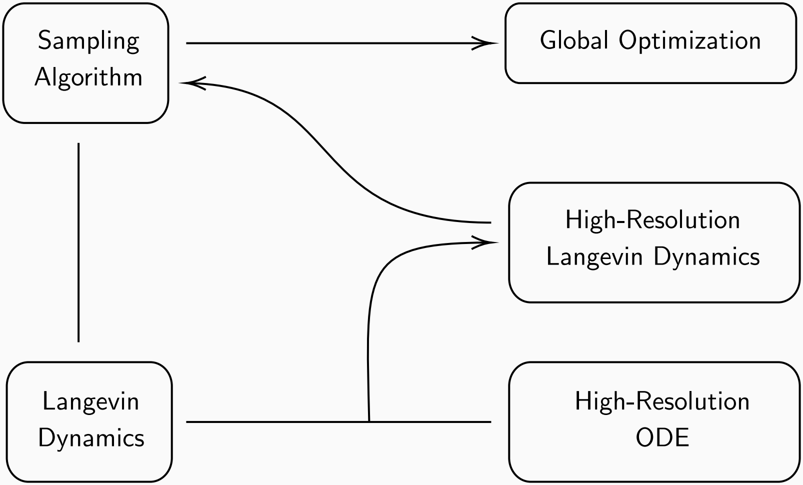
Approach:

- Build a probability distribution such that its samples are close to the global minimizers.
- Build an algorithm to sample, at least approximately, from that distribution.

Assumptions:

- U is twice differentiable and that ∇U is Lipschitz continuous,
- There exists an $a_0 > 0$ such that $\int_{\mathbb{R}^d} \exp(-a_0 U(x)) dx < +\infty$,
- The measure $\boldsymbol{\mu}^a \propto \exp(-aU)$ satisfies a log-Sobolev inequality,
- U has a finite number of global minimizers, with minimum value U^*.

Roadmap


Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

Some Notions

We will be working on the space of probability measures, which we denote $\mathcal{P}(\mathbb{R}^d)$.

Kullback-Leibler Divergence: For any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\mathbb{R}^d)$, we define

$$\mathrm{KL}(\boldsymbol{\nu}\|\boldsymbol{\mu}) = \mathbb{E}_{x\sim\boldsymbol{\nu}}\left[\log \frac{d\boldsymbol{\nu}}{d\boldsymbol{\mu}}(x)\right].$$

Relative Fischer Information: For any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\mathbb{R}^d)$, we define

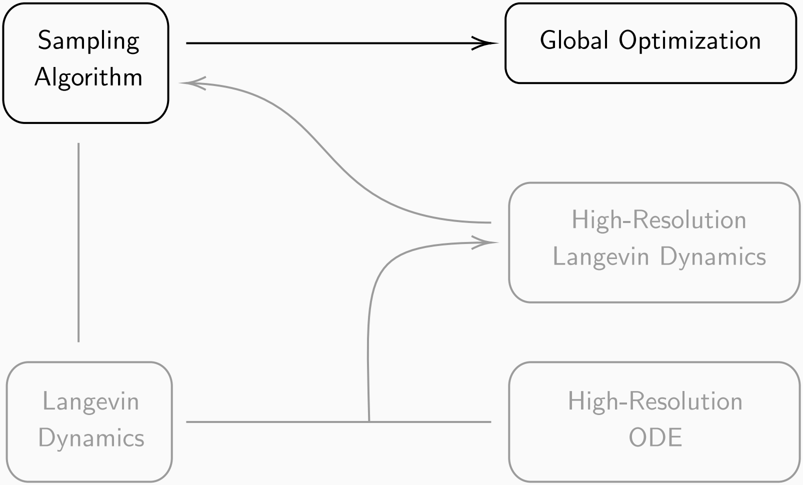
$$\mathrm{Fi}(\boldsymbol{\nu}\|\boldsymbol{\mu}) = \mathbb{E}_{x\sim\boldsymbol{\nu}}\left[\left\|\nabla \log \frac{d\boldsymbol{\nu}}{d\boldsymbol{\mu}}(x)\right\|^2\right].$$

Log-Sobolev Inequality We say $\boldsymbol{\mu}$ satisfies a log-Sobolev inequality if, for all $\boldsymbol{\nu} \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathrm{KL}(\boldsymbol{\nu}\|\boldsymbol{\mu}) \leq \frac{1}{2\rho}\mathrm{Fi}(\boldsymbol{\nu}\|\boldsymbol{\mu}).$$

This may be compared to a Polyak-Lojasiewicz inequality in $\mathbb{R}^d$.

Roadmap


Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

Optimization through Sampling?

Define $\boldsymbol{\mu}^*$ to be an appropriate mixture of Dirac measures concentrated on the global minimizers of U.

Theorem (Athreya and Hwang, 2010)

Let $\boldsymbol{\mu}^a \propto \exp(-aU)$. Then it holds that $\boldsymbol{\mu}^a \to \boldsymbol{\mu}^*$.

Convergence in the above is in the weak sense. Strong convergence was later established in Hasenpflug, Rudolf, and Sprungk, 2024.

Intuitively:

- Sampling from $\boldsymbol{\mu}^*$ gives us a global minimizer of U. However, we cannot sample from $\boldsymbol{\mu}^*$.
- By picking $a > 0$ sufficiently large, $\boldsymbol{\mu}^a$ is 'close' to $\boldsymbol{\mu}^*$. However, we also cannot sample from $\boldsymbol{\mu}^a$ directly.
- We can however design an algorithm that samples from some $\tilde{\boldsymbol{\mu}}$, that is 'close' to $\boldsymbol{\mu}^a$.
- Running this multiple times will prevent outliers.

Global Optimization Algorithm

Algorithm 1 Global Optimization Algorithm

Require: Oracle algorithm and suitable parameters.

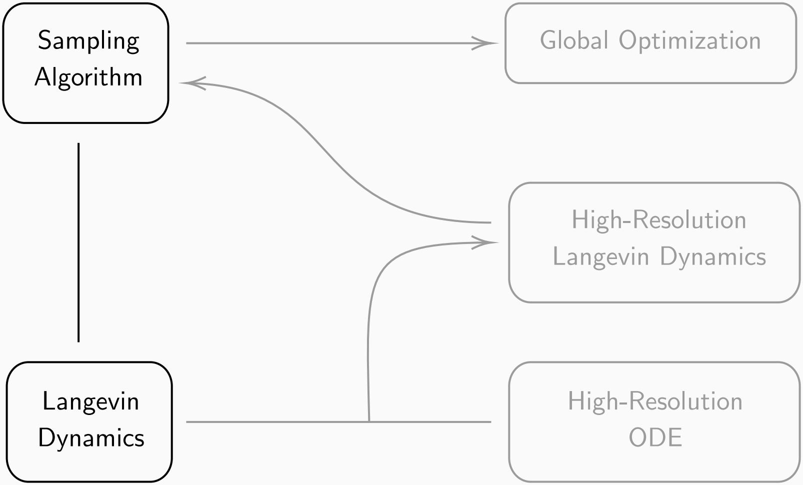
1: Generate N random i.i.d. samples $\tilde{X}^{(i)}$ according to oracle algorithm where $i = 1, \ldots, N$.
2: Define $\tilde{X} = \tilde{X}^{(I)}$ where $I = \operatorname{argmin}_{i=1\ldots,N} U(\tilde{X}^{(i)})$.

Theorem (Convergence of Global Optimization Algorithm)

Fix $\varepsilon > 0$. Suppose we can sample from a distribution $\tilde{\boldsymbol{\mu}}$ satisfying that $\mathrm{KL}(\tilde{\boldsymbol{\mu}} \| \boldsymbol{\mu}^a)$ is small.
Then we can guarantee, for $\tilde{X} \sim \tilde{\boldsymbol{\mu}}$, that $\mathbb{P}(U(\tilde{X}) - U^* \leq \varepsilon)$ is high.

Question: How do we ensure that $\mathrm{KL}(\tilde{\boldsymbol{\mu}} \| \boldsymbol{\mu}^a)$ is small?

Roadmap



Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

Sampling through Continuous Dynamics

Consider the stochastic differential equation (SDE), known as the **Langevin Dynamics**:

$$dX_t = -\gamma \nabla U(X_t)dt + \sqrt{2\gamma/a}dB_t,$$

where (B_t) is a standard d-dimensional Brownian motion.
It is known that

- (X_t) has a unique (strong) solution,
- if we denote by $\boldsymbol{\mu}_t = \mathcal{L}(X_t)$, one can show that $\boldsymbol{\mu}_t$ converges linearly to $\boldsymbol{\mu}^a \propto \exp(-aU)$ in KL divergence.

Conclusion: The Langevin dynamics is a good candidate to design a sampling algorithm!

Approximate Sampling

Langevin Dynamics:

$$dX_t = -\gamma \nabla U(X_t) dt + \sqrt{2\gamma / a} dB_t.$$

Issues:

- Even though $\boldsymbol{\mu}_t \to \boldsymbol{\mu}^a$, we cannot simulate the process for $t = \infty$.
- In fact, we cannot simulate $\boldsymbol{\mu}_t$ at all!

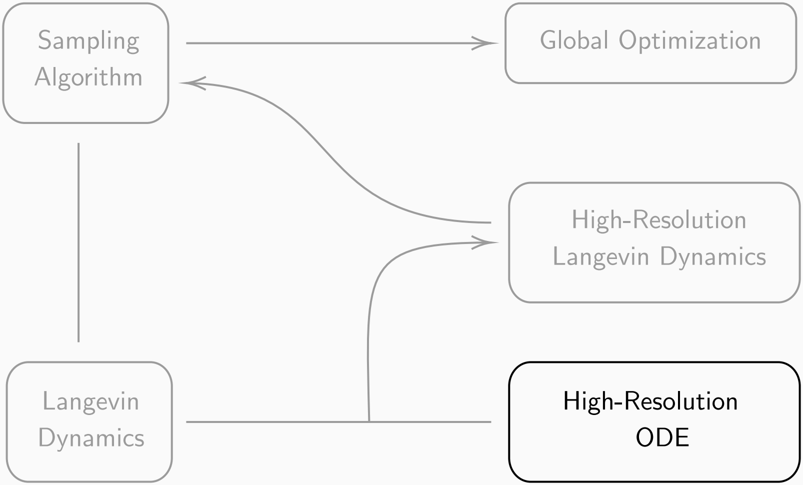
Solution: Discretize the SDE. For instance, the Euler-Maruyama discretization reads

$$X_{(k+1)h} - X_{kh} = -\gamma h \nabla U(X_{kh}) + \sqrt{2\gamma h / a} \xi_k,$$

where $\xi_k \sim \mathcal{N}(0, 1)$ are independent.

This process can be simulated by simulating Gaussians. One can prove convergence to $\boldsymbol{\mu}^a$ in KL as $h \to 0$, although without an explicit rate, see Vempala and Wibisono, 2019.

Roadmap



Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

Recent Deterministic Trends

Recent trends analyse continuous dynamics to gain insights into the discretized algorithms. For instance, Gradient Descent is a discretization of the Gradient Flow:

$$\dot{x}(t) = -\gamma \nabla f(x(t)) \quad \to \quad x_{k+1} = x_k - \gamma h \nabla f(x_k).$$

To capture acceleration behaviours, it has been proposed (Alvarez et al., 2002) to study the **High-Resolution ODE**:

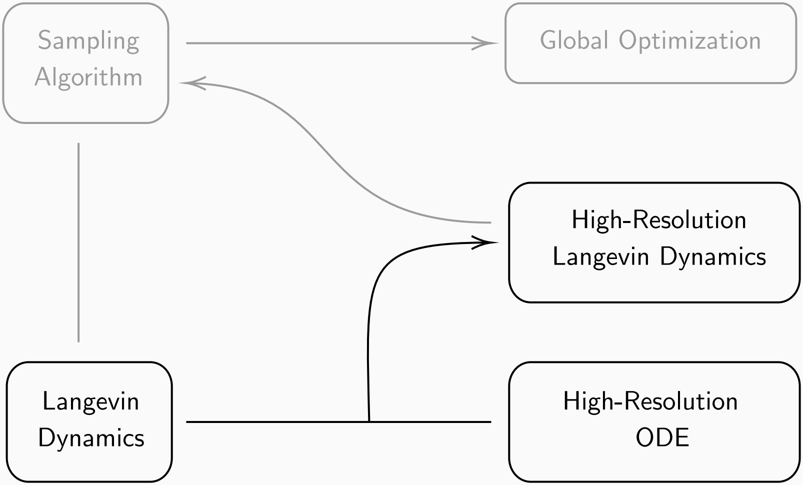
$$\ddot{x}(t) + \alpha \dot{x}(t) + \beta \nabla^2 U(x(t)) \dot{x}(t) + \gamma \nabla U(x(t)) = 0,$$

where $\alpha, \beta, \gamma > 0$. Equivalently, under a change of variables,

$$\begin{cases} \dot{x}(t) & = -\beta \nabla U(x(t)) + y(t) \\ \dot{y}(t) & = -\gamma \nabla U(x(t)) - \alpha y(t). \end{cases}$$

Discretizations have given rise to accelerated algorithms, see for instance Attouch et al., 2022.

Roadmap


Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

High-Resolution Langevin Dynamics

One can view the Langevin Dynamics as a stochastic variant of the Gradient Flow:

$$\dot{x}(t) = -\gamma \nabla U(x(t)) \quad \leftrightarrow \quad dX_t = -\gamma \nabla U(X_t)dt + \sqrt{2\gamma/a}dB_t.$$

Recall the High-Resolution ODE in first-order form:

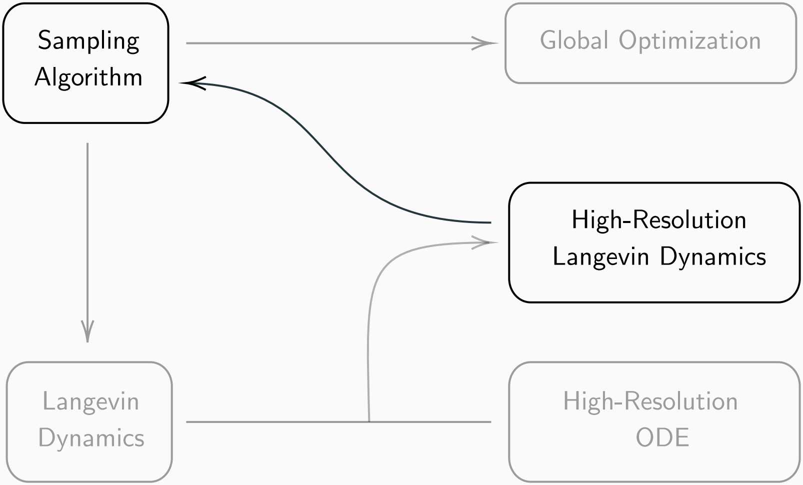
$$\begin{cases} \dot{x}(t) & = -\beta \nabla U(x(t)) + y(t) \\ \dot{y}(t) & = -\gamma \nabla U(x(t)) - \alpha y(t). \end{cases}$$

We consider a stochastic variant of it, namely

$$\begin{cases} dX_t = (-\beta \nabla U(X_t) + Y_t)dt + \sqrt{2\sigma_x^2}dB_t^x \\ dY_t = (-\gamma \nabla U(X_t) - \alpha Y_t)dt + \sqrt{2\sigma_y^2}dB_t^y. \end{cases}$$

We call these dynamics the **High-Resolution Langevin Dynamics**.

Roadmap



Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

High-Resolution Langevin Dynamics

We propose and study the **High-Resolution Langevin Dynamics**:

$$\begin{cases} dX_t = (-\beta \nabla U(X_t) + Y_t)dt + \sqrt{2\sigma_x^2} dB_t^x \\ dY_t = (-\gamma \nabla U(X_t) - \alpha Y_t)dt + \sqrt{2\sigma_y^2} dB_t^y, \end{cases} \tag{1}$$

Theorem (Convergence of High-Resolution Langevin)

Assume suitable parameter relations, and denote $\boldsymbol{\mu}_t = \mathcal{L}(X_t)$.

1. Under weak assumptions, (1) admits a weak solution (X_t, Y_t) such that $\boldsymbol{\mu}^a \propto \exp(-aU)$ is the invariant law of (X_t).
2. $\mathrm{KL}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}^a) \to 0$ at an exponential rate.
3. For a sufficiently small step size $h > 0$ and large number of iterations K, the discretization of System (1), denoted by $(\tilde{X}_t, \tilde{Y}_t)$, satisfies $\mathrm{KL}(\tilde{\boldsymbol{\mu}}_{Kh} \| \boldsymbol{\mu}^a) \leq \varepsilon$, for $\tilde{\boldsymbol{\mu}}_t = \mathcal{L}(\tilde{X}_t)$. This discretized process may be simulated.

High-Resolution Langevin Algorithm

1. Simulate $(\tilde{X}_0, \tilde{Y}_0) \sim \tilde{\boldsymbol{\mu}}_0$.

2. Iteratively generate $(\tilde{X}_{(k+1)h}, \tilde{Y}_{(k+1)h}) \sim \mathcal{N}(m, \Sigma)$ where

$$m_X = \tilde{X}_{kh} - \beta h \nabla U(\tilde{X}_{kh}) + \frac{1 - e^{-\alpha h}}{\alpha} \tilde{Y}_{kh} - \frac{\gamma}{\alpha} \left(h - \frac{1 - e^{-\alpha h}}{\alpha} \right) \nabla U(\tilde{X}_{kh})$$

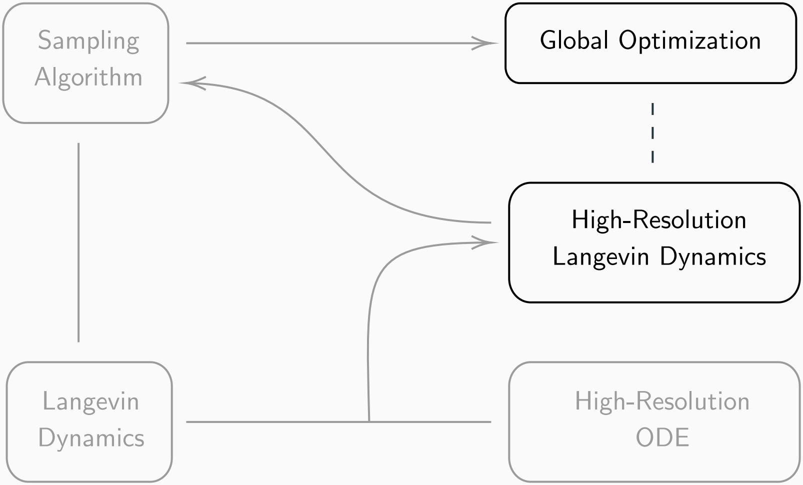
$$m_Y = e^{-\alpha h} \tilde{Y}_{kh} - \frac{\gamma}{\alpha} (1 - e^{-\alpha h}) \nabla U(\tilde{X}_{kh})$$

$$\Sigma_{XX} = \frac{\sigma_y^2}{\alpha^3} \left[2\alpha h - e^{-2\alpha h} + 4e^{-\alpha h} - 3 \right] \cdot I_d + 2\sigma_x^2 h \cdot I_d$$

$$\Sigma_{YY} = \frac{\sigma_y^2 (1 - e^{-2\alpha h})}{\alpha} \cdot I_d, \quad \Sigma_{XY} = \Sigma_{YX} = \frac{\sigma_y^2 (1 - e^{-\alpha h})^2}{\alpha^2} \cdot I_d.$$

3. Return $(\tilde{X}_{Kh}, \tilde{Y}_{Kh})$.

Roadmap



Sampling Algorithm
Global Optimization
High-Resolution Langevin Dynamics
Langevin Dynamics
High-Resolution ODE

Global Optimization through High-Resolution Sampling

Algorithm 2 Global Optimization through High-Resolution Sampling

Require: Suitable parameters and an initial distribution $\tilde{\boldsymbol{\mu}}_0$.

Ensure: Produce $\tilde{X}$ satisfying $\mathbb{P}(U(\tilde{X}) - U^* \leq \varepsilon) \geq 1 - \delta$.

```
for i = 1, ..., N do
    Simulate (X̃_0^(i), Ỹ_0^(i)) ~ μ̃_0.
    for k = 0, ..., K − 1 do
        Generate (X̃_{(k+1)h}^(i), Ỹ_{(k+1)h}^(i)) ~ N(m, Σ) with m, Σ as before.
    end for
end for
Define X̃ = X̃^(I) where I = argmin_{i=1...,N} U(X̃_{Kh}^(i)).
```

for $i = 1, \ldots, N$ **do**
- Simulate $(\tilde{X}_0^{(i)}, \tilde{Y}_0^{(i)}) \sim \tilde{\boldsymbol{\mu}}_0$.
- **for** $k = 0, \ldots, K-1$ **do**
  - Generate $(\tilde{X}_{(k+1)h}^{(i)}, \tilde{Y}_{(k+1)h}^{(i)}) \sim \mathcal{N}(m, \Sigma)$ with m, Σ as before.
- **end for**

end for

Define $\tilde{X} = \tilde{X}^{(I)}$ where $I = \mathrm{argmin}_{i=1\ldots,N}\, U(\tilde{X}_{Kh}^{(i)})$.

Numerical Results

Rastrigin Function

Consider the **Rastrigin function** $U\colon \mathbb{R}^d \to \mathbb{R}$ defined by

$$U(x) = d + \|x\|^2 - \sum_{i=1}^{d} \cos(2\pi x_i).$$

Its minimum is located in $x^* = (0, \dots, 0) \in \mathbb{R}^d$, with objective value 0. This function is highly multi-modal and satisfies a log-Sobolev inequality.

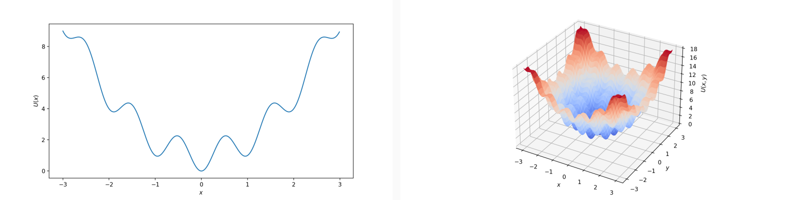
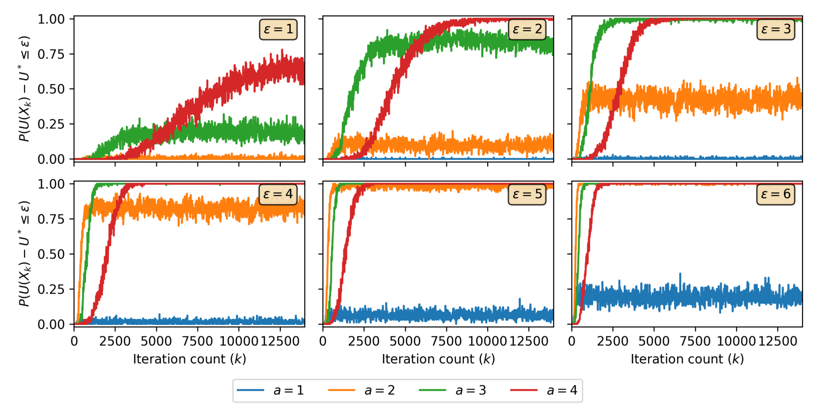


Figure 1: Rastrigin function for $d = 1$ and $d = 2$.

Selected Parameters

- **Problem Parameters:** We limit ourselves to $d = 10$.
- **Sampling Algorithm Parameters:** For a given a, we fix $\alpha = 1$, $\beta = 1$, $b = 10$, $\gamma = a/10$, $\sigma_x^2 = 1/a$ and $\sigma_y^2 = 0.1$. Moreover, we set the step-size $h = 0.01$.
- **Optimization Algorithm Parameters:** We set the sample count $N = 10$ and the iteration count $K = 14000$.
- **Initial Value:** We initialize our algorithm in $X_0 \sim \mathcal{N}(3 \cdot 1_d, 10 \cdot I_d)$.
- **Post-Processing Parameters:** We will compute empirical probabilities that $U(\tilde{X}_k) - U^* \leq \varepsilon$ over $M = 100$ runs.
- **Free Parameters (to be varied):** The threshold $\varepsilon > 0$ and the inverse temperature $a > 0$.

Empirical Probabilities



Observation: Small values of a converge faster, but to less accurate thresholds.

Comparison to Guilmeau, Chouzenoux, and Elvira, 2021

We modify some parameters for a fair comparison:

- Empirical probabilities are computed over $M = 50$ runs.
- Iteration count is now $K = 50$ or $K = 500$.
- Sample count is now $N = 250$.
- Initial distribution is now deterministically $\tilde{X}_0 = (1, \ldots, 1)^{10}$.

Denote by A_K and S_K the average and standard deviation over all runs after K iterations.

	SA	FSA	SMC	CSA	Ours
A_{50}	3.29	3.36	3.26	**3.23**	14.04
S_{50}	**0.425**	0.453	0.521	0.484	2.563
A_{500}	2.52	2.64	2.62	2.47	**0.38**
S_{500}	0.320	0.304	0.413	0.502	**0.101**

Conclusion: Our algorithm is slow for $K = 50$, but good for $K = 500$.

Further Research Directions:

- Optimal parameter selection (in algorithm and the balance between N and K).
- Development of a cooling scheme (online?).

Paper: Daniel Cortild, Claire Delplancke, Nadia Oudjane, and Juan Peypouquet (Oct. 2024). **Global Optimization Algorithm through High-Resolution Sampling.** arXiv:2410.13737

Thank you!

Alvarez, Felipe, Hedy Attouch, Jérôme Bolte, and Patrick Redont (2002). **"A second-order gradient-like dissipative dynamical system with Hessian-driven damping.: Application to optimization and mechanics".** In: *Journal de mathématiques pures et appliquées* 81.8. Publisher: Elsevier, pp. 747–779.

Athreya, Krishna B and Chii-Ruey Hwang (2010). **"Gibbs measures asymptotics".** In: *Sankhya* 72. Publisher: Springer, pp. 191–207.

Attouch, Hedy, Zaki Chbani, Jalal Fadili, and Hassan Riahi (2022). **"First-order optimization algorithms via inertial systems with Hessian driven damping".** In: *Mathematical Programming*. Publisher: Springer, pp. 1–43.

Guilmeau, Thomas, Emilie Chouzenoux, and Víctor Elvira (2021). **"Simulated Annealing: a Review and a New Scheme".** In: *2021 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 101–105.

Hasenpflug, Mareike, Daniel Rudolf, and Björn Sprungk (2024). **"Wasserstein convergence rates of increasingly concentrating probability measures".** In: *The Annals of Applied Probability* 34.3. Publisher: Institute of Mathematical Statistics, pp. 3320–3347.

Vempala, Santosh and Andre Wibisono (2019). **"Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices".** In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.