Global Optimization Algorithm through High-Resolution Sampling

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

Benelux Meeting on Systems and Control
Egmond aan Zee, Netherlands, March 19th, 2025

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

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

Difficulties: Existence of local minimizers & saddle points.

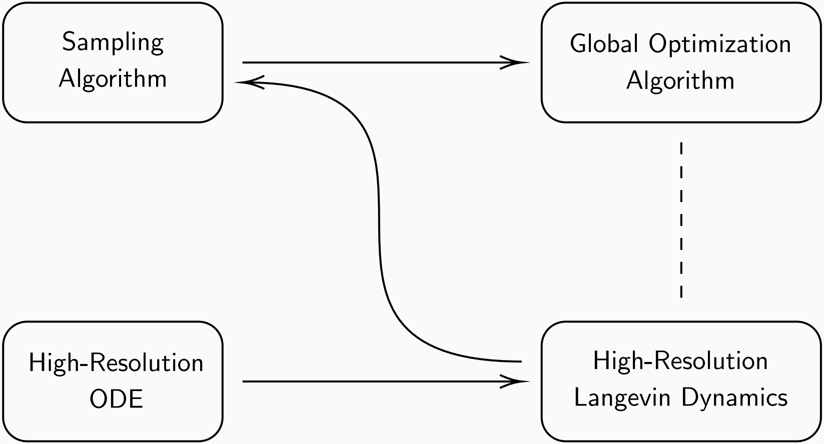
Approach:

- 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.

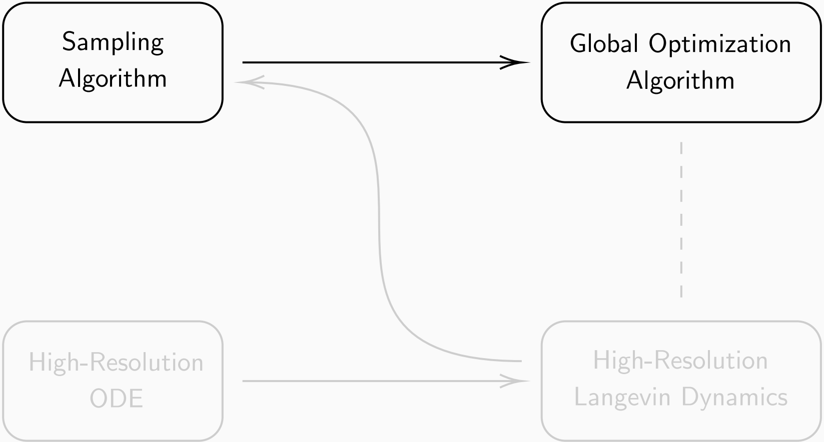
Main Assumptions:

- U has a finite number of global minimizers, with minimum value U^*,
- The measure $\boldsymbol{\mu}^a \propto \exp(-aU)$ exists and satisfies a growth condition.

Roadmap


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

Roadmap


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

Optimization through Sampling?

Define $\boldsymbol{\mu}^*$ to be a mixture of Dirac measures concentrated on the global minimizers of U (see Athreya and Hwang, 2010 for exact definition).

Theorem (Athreya and Hwang, 2010)

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

Convergence in the above is in the weak sense. Strong convergence (in KL divergence) with rates was later established in Hasenpflug, Rudolf, and Sprungk, 2024.

Intuitively:

$$\boxed{\operatorname{argmin}(U)} \Leftarrow \boxed{\boldsymbol{\mu}^*} \approx \boxed{\boldsymbol{\mu}^a \ (a \text{ large})} \approx \boxed{\tilde{\boldsymbol{\mu}}}$$

Question: How to choose and sample from $\tilde{\boldsymbol{\mu}}$?

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

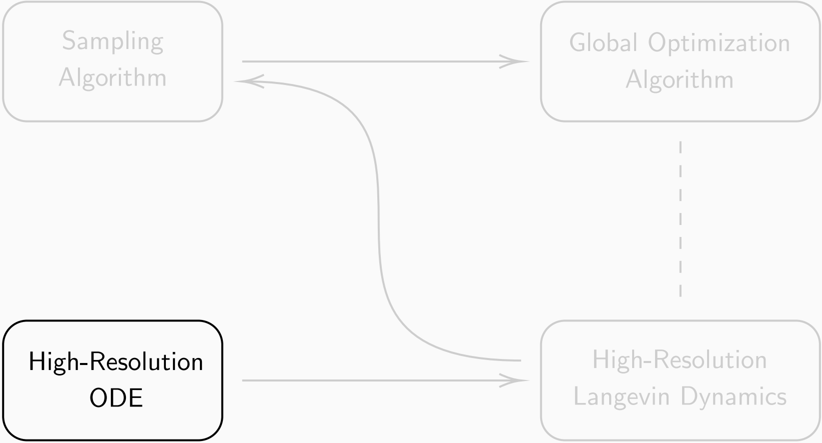
Theorem (Convergence of Global Optimization Algorithm)

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 $\tilde{X}$ given by Algorithm 1, 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
Algorithm
High-Resolution
ODE
High-Resolution
Langevin Dynamics

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:

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

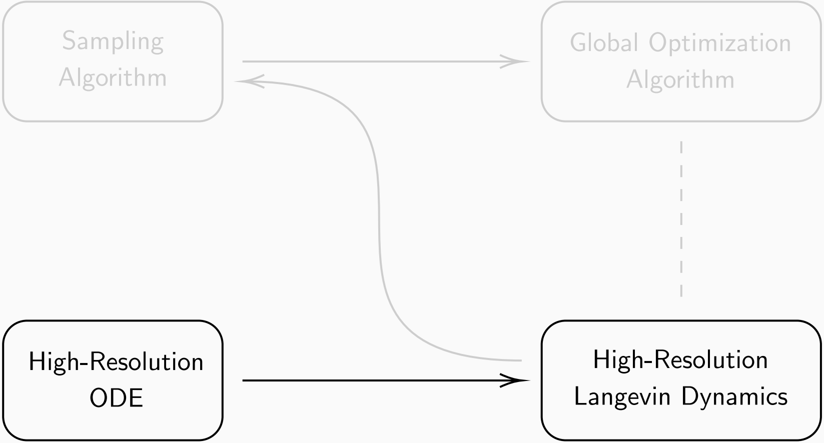
To capture acceleration behaviours, it has been proposed to study the **High-Resolution ODE**:

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

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

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

Roadmap


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

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:

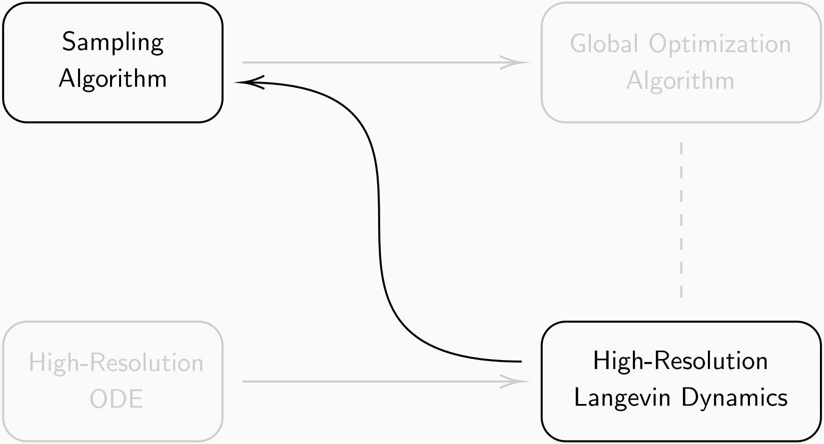
$$\begin{cases} \dot{x}(t) & = -\beta \nabla U(x(t)) + y(t) \\ \dot{y}(t) & = -\gamma \nabla U(x(t)) - \alpha y(t). \end{cases}$$

We consider a stochastic variant of it, namely

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

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

Roadmap


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

High-Resolution Langevin Dynamics

$$\text{argmin}(U) \Leftarrow \boldsymbol{\mu}^* \approx \boldsymbol{\mu}^a \ (a \text{ large})$$

$$\wr\wr$$

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}_{Kh} \ (K \text{ large}, h \text{ small}) \approx \boldsymbol{\mu}_t \ (t \text{ large})$$

Theorem (Convergence of High-Resolution Langevin)

Assume suitable parameter relations, and denote $\boldsymbol{\mu}_t = \mathcal{L}(X_t)$ the marginal law of the HRLD. Under weak assumptions;

1. $\text{KL}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}^a) \to 0$ at an exponential rate.
2. For a sufficiently small step size $h > 0$ and large number of iterations K, the law of the discretization of the HRLD, denoted by $(\tilde{X}_t, \tilde{Y}_t)$, satisfies $\text{KL}(\tilde{\boldsymbol{\mu}}_{Kh} \| \boldsymbol{\mu}^a) \leq \varepsilon$, for $\tilde{\boldsymbol{\mu}}_t = \mathcal{L}(\tilde{X}_t)$. This discretized process may be simulated.

Question: How do we simulate $(\tilde{X}_t)$ to sample from $\tilde{\boldsymbol{\mu}}_t$?

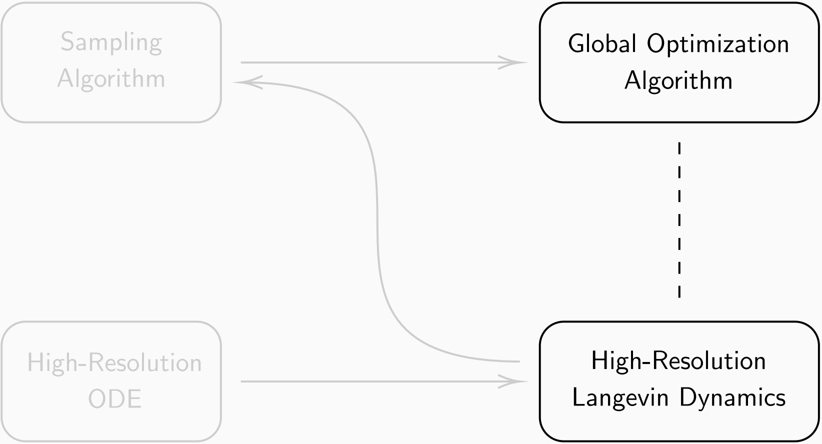
High-Resolution Langevin Algorithm

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

$$
\begin{aligned}
m_X &= \tilde{X}_{kh} - \beta h \nabla U(\tilde{X}_{kh}) + \frac{1 - e^{-\alpha h}}{\alpha} \tilde{Y}_{kh} - \frac{\gamma}{\alpha} \left(h - \frac{1 - e^{-\alpha h}}{\alpha} \right) \nabla U(\tilde{X}_{kh}) \\
m_Y &= e^{-\alpha h} \tilde{Y}_{kh} - \frac{\gamma}{\alpha} (1 - e^{-\alpha h}) \nabla U(\tilde{X}_{kh}) \\
\Sigma_{XX} &= \frac{\sigma_y^2}{\alpha^3} \left[2\alpha h - e^{-2\alpha h} + 4e^{-\alpha h} - 3 \right] \cdot I_d + 2\sigma_x^2 h \cdot I_d \\
\Sigma_{YY} &= \frac{\sigma_y^2 (1 - e^{-2\alpha h})}{\alpha} \cdot I_d, \quad \Sigma_{XY} = \Sigma_{YX} = \frac{\sigma_y^2 (1 - e^{-\alpha h})^2}{\alpha^2} \cdot I_d .
\end{aligned}
$$

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

Roadmap



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

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

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 $x^* = (0, \ldots, 0) \in \mathbb{R}^d$, with objective value 0. This function is highly multi-modal and satisfies our assumptions.

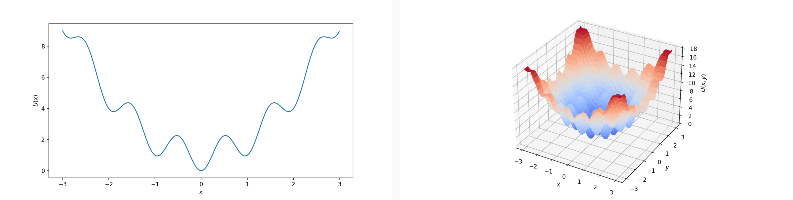
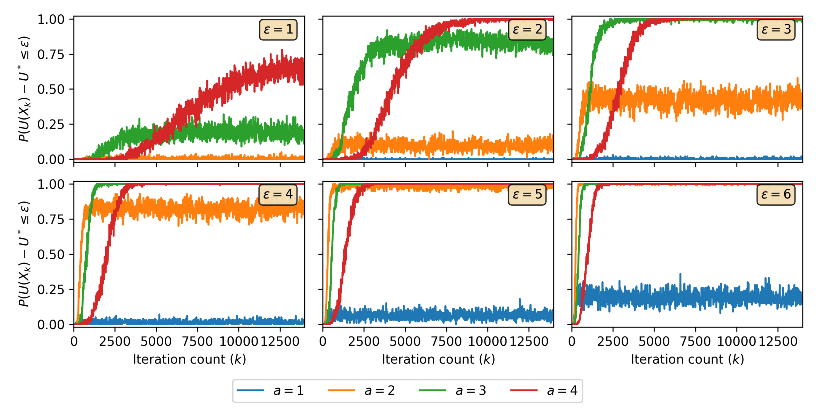


Figure 1: Rastrigin function for $d = 1$ and $d = 2$.

Empirical Probabilities

We set $d = 10$ and compute empirical probabilities over $M = 100$ runs.



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

Comparison to Guilmeau, Chouzenoux, and Elvira, 2021

For a fair comparison, we consider $K = 50$ and $K = 500$.

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

	SA	FSA	SMC	CSA	Ours[1]
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 $K = 500$.

[1] For well-chosen parameters

Conclusion

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!

References i

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

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.