Optimization and sampling by linear kinetic equations

Lorenzo Pareschi

School of Mathematical and Computer Sciences and Maxwell Institute for Mathematical Sciences
Heriot-Watt University, Edinburgh, UK

Department of Mathematics and Computer Science
University of Ferrara, IT

Interacting Particle Systems: Analysis, Control, Learning and Computation

ICERM, May 6-10, 2024

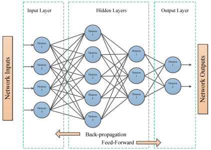
MAXWELL INSTITUTE FOR MATHEMATICAL SCIENCES

The Wolfson* Foundation
THE ROYAL SOCIETY

HERIOT WATT UNIVERSITY

Motivations

- High-dimensional (nonconvex) optimization problems are pervasive in many fields, particularly in cutting-edge areas such as machine learning, signal/image processing and optimal control.



Training neural networks

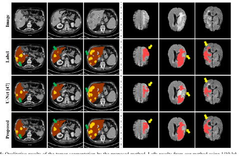
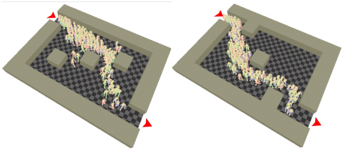


Image
Label
U-Net [47]
Proposed
5: Qualitative results of the tumor segmentation by the proposed method. Left: results from our method using 1/10 lab

Computer assisted tomography



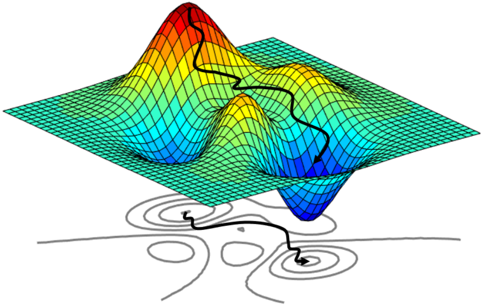
Crowd evacuation control

- Optimization and sampling methods often draw from similar probabilistic principles and techniques, making them interconnected in various computational contexts.
- Stochastic gradient descent (SGD) methods are are efficient, scalable, and adept at avoiding critical points. They are closely linked to sampling via Langevin dynamics.
- Metaheuristic algorithms gained popularity for their broad applicability and minimal assumptions on optimization/sampling problems.

Metaheuristics

Metaheuristic algorithms, often nature-inspired, combine random and deterministic moves with local and global strategies to escape local minima and perform a robust search of the solution.

- Metropolis-Hastings (1953,1970)
- Simplex Heuristics (1965)
- Evolutionary Programming (1966)
- Genetic Algorithms (GA) (1975)
- Simulated Annealing (SA) (1983)
- Particle Swarm Optimization (PSO) (1995)
- Ant Colony Optimization (ACO) (1997)
- ...



⇒ Despite the significant empirical success, most results are experimental in nature and lack a rigorous mathematical foundation.

Classical metaheuristics

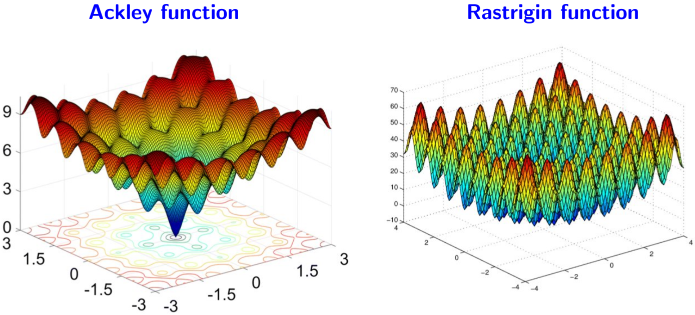
Consider the optimization problem

$$x^* \in \text{argmin}_{x \in \mathbb{R}} \mathcal{F}(x)\,,$$

$\mathcal{F}(x) : \mathbb{R}^d \to \mathbb{R}$ is a (non convex, high dimensional, possibly non smooth) cost function.

Algorithm	**Feature**
Simulated Annealing (SA)	Generates a single point X^n at each iteration. The sequence of points approaches an optimal solution.
Genetic Algorithm (GA)	Generates a population of points X_i^n at each iteration. The fittest evolve towards an optimal solution.
Particle Swarm Optimization (PSO)	Generates a swarm of points (X_i^n, V_i^n) at each iteration. The swarm moves towards an optimal solution.

Metaheuristics optimization in action


Ackley function

Rastrigin function

Examples of swarm-based optimization processes

CBO methods: a PDEs perspective on metaheuristcs

Consensus-based optimization (CBO) considers the evolution of N particles $X_t^i \in \mathbb{R}^d$ according to[1]:

$$\boxed{dX_t^i = \underbrace{-\lambda(X_t^i - \bar{X}_t^\alpha)dt}_{\text{alignment}} + \underbrace{\sigma D(X_t^i - \bar{X}_t^\alpha)dB_t^i}_{\text{exploration}},}$$

where $\lambda > 0$, $\sigma > 0$, $D(X_t) = |X_t| I_d$ (isotropic) or $D(X_t) = \text{diag}\left\{(X_t)_1, \dots, (X_t)_d\right\}$ (anisotropic)

$$\bar{X}_t^\alpha = \frac{1}{\sum_i e^{-\alpha\mathcal{F}(X_t^i)}} \sum_i X_t^i e^{-\alpha\mathcal{F}(X_t^i)} \xrightarrow[\alpha\to+\infty]{} \text{argmin}(\mathcal{F}(X_t^1), \dots, \mathcal{F}(X_t^N)) \text{ (Laplace principle)}$$

The behavior for $N \gg 1$ is obtained by assuming that the (X_t^i), $i = 1, \dots, N$ are i.i.d. with the same distribution $\rho(x,t)$ (propagation of chaos assumption) satisfying the Fokker–Planck equation

$$\boxed{\partial_t \rho = \nabla_x \cdot \lambda(x - \bar{x}^\alpha(\rho))\rho(t) + \frac{\sigma^2}{2} \sum_{j=1}^d \partial_{jj}((x - \bar{x}^\alpha(\rho))_j^2 \rho(t))}$$

[1] Pinnau, Totzeck, Tse, Martin '17; Carrillo, Choi, Totzeck, Tse '18; Carrillo, Jin, Li, Zhu '20; Fornasier, Huang, Sünnen, Pareschi 21; Carrillo, Hoffmann, Stuart, Vaes '22; Borghi, Herty, Pareschi '23; . . .

Questions arising

- Can we extend the concepts and analysis of CBO to other widely used metaheuristic algorithms?
- Can this approach lead to the design of new, more efficient and mathematically explainable algorithms?
- Could this approach enhance our understanding of the relationship between metaheuristics and gradient-based methods?

Outline

1. Motivations

2. Optimization by linear kinetic equations
   - Simulated annealing
   - Convergence to equilibrium
   - Mean-field Langevin limit
   - Generalizations

3. Concluding remarks

Simulated Annealing

N. Metropolis

Starting from a random trial point $X^0 \in \mathbb{R}^d$ and a control temperature T^0, the simulated annealing (SA) algorithm can be summarized as[b]

1. Move the current point
$$\tilde{X}^{n+1} = X^n + \sigma^n \xi$$
where $\xi \sim U(-1,1)^d$ and $\sigma^n > 0$ depends on T^n. Typically $\sigma^n \sim \sqrt{T^n}$.

2. If $\tilde{X}^{n+1}$ is better than the current point $\mathcal{F}(\tilde{X}^{n+1}) < \mathcal{F}(X^n)$, it becomes the next point. If $\tilde{X}^{n+1}$ is worse $\mathcal{F}(\tilde{X}^{n+1}) \geq \mathcal{F}(X^n)$ it is accepted with probability $e^{-\frac{\mathcal{F}(\tilde{X}^{n+1})-\mathcal{F}(X^n)}{T^n}}$.

3. The algorithm systematically lowers the temperature, accordingly to a law of the type
$$T^{n+1} = \lambda^{n+1} T_0, \qquad \lambda^n \in (0,1),$$
where $T_0 > 0$ is a given initial temperature. A classical choice is $\lambda^n = 1/\ln(n+2)$.

⇒ For a fixed T the algorithm corresponds to Metropolis-Hasting sampling from the Boltzmann-Gibbs probability density $Ce^{-\frac{\mathcal{F}(x)}{T}}$.

[b] Metropolis et al. '53; Kirkpatrick, Gelatt, Vecchi '83

Simulated annealing and Langevin dynamics

Consider the stochastic differential process[2]

$$dX_t = -\nabla_x \mathcal{F}(X_t)dt + \sqrt{2T}dB_t,$$

referred to as Langevin equation. It can be understood as the limit for small learning rates of a stochastic gradient descent (SGD) method.

The process is refereed to as continuous simulated annealing since its mean field description

$$\boxed{\frac{\partial f}{\partial t}(x,t) = \nabla_x \cdot (\nabla_x \mathcal{F}(x) f(x,t)) + T\Delta_{xx} f(x,t),}$$

where $f(x,t)$ is the probability density to have a trial point in position $x \in \mathbb{R}^d$ at time $t > 0$, admits as stationary state the Boltzmann-Gibbs distribution

$$f^\infty_{\mathcal{F}}(x) = Ce^{\frac{-\mathcal{F}(x)}{T}}.$$

[2] Geman, Hwang '86; Hwang et al '87; Locatelli '00; Monmarché '18; Chizat '22

Annealing process

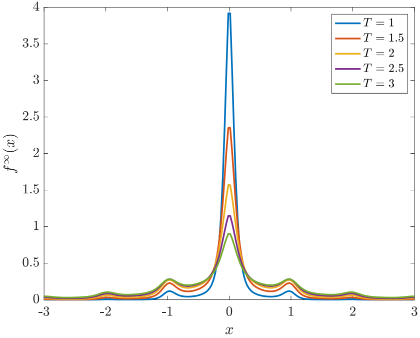
By the Laplace principle

$$\lim_{T\to 0} -T\log\left(\int_{\mathbb{R}^d} g(x)e^{\frac{-\mathcal{F}(x)}{T}}\,dx\right) = \inf_{x\in\mathrm{supp}(g)} \mathcal{F}(x),$$

where $g(x)$ is a pdf in $\mathbb{R}^d$. For $T \ll 1$, the equilibrium state concentrates on global minima of $\mathcal{F}(x)$

$$f^{\infty}_{\mathcal{F}}(x) \to \delta(x - x^*).$$

Time to reach equilibrium increases exponentially with $1/T$!



Slowly decreasing $T(t)$ so that the solution approaches $f^{\infty}_{\mathcal{F}}(x)$ at a faster rate and concentrates on minima asymptotically. For $T(t) \sim 1/\log(2+t)$ it converges weakly to the set of global minima[3].

- $\Rightarrow$ It requires the gradient evaluation, in contrast with the gradient-free nature of SA algorithm.
- $\Rightarrow$ Derivation of the SDE Langevin diffusion from Metropolis-Hasting[4].

[3] Hajek '88

[4] Roberts, Gelman, Gilks '97; Roberts, Rosenthal '01; Pillai, Stuart, Thiéry '14

Optimization by linear kinetic equations

After introducing the probability density $f(x,t)$, we can write the evolution equation[5]

$$\frac{\partial f(x,t)}{\partial t} = \mathcal{L}_{\mathcal{F}}(f(x,t))$$
$$\mathcal{L}_{\mathcal{F}}(f(x,t)) = \underbrace{\langle B_{\mathcal{F}}(x' \to x) f(x',t)}_{\text{gain}} - \underbrace{B_{\mathcal{F}}(x \to x') f(x,t) \rangle}_{\text{loss}}$$

where $\langle \cdot \rangle = \mathbb{E}_{\xi}[\cdot]$ denotes the expectation with respect to the selection probability $p(\xi)$, $\xi \in \mathbb{R}^d$,

$$x' = x + \sigma(t)\xi,$$

is the new trial-point position, and

$$B_{\mathcal{F}}(x \to x') = \min\left\{1, \frac{f^{\infty}_{\mathcal{F}}(x')}{f^{\infty}_{\mathcal{F}}(x)}\right\} = \begin{cases} 1, & \mathcal{F}(x') < \mathcal{F}(x), \\ \dfrac{f^{\infty}_{\mathcal{F}}(x')}{f^{\infty}_{\mathcal{F}}(x)}, & \mathcal{F}(x') \geq \mathcal{F}(x), \end{cases}$$

is the transition probability from $x \to x'$.

[5] Kolokoltsov '10; Pareschi, Toscani '13

Proposition

The Gibbs distribution $f^\infty_{\mathcal{F}}(x)$ satisfies $\mathcal{L}_{\mathcal{F}}(f^\infty_{\mathcal{F}}(x)) = 0$, $\forall\, x \in \mathbb{R}^d$.

For a symmetric selection probability we have the weak form

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^d} f(x,t)\phi(x)\,dx = \left\langle \int_{\mathbb{R}^d} B_{\mathcal{F}}(x \to x')(\phi(x') - \phi(x)) f(x,t)\,dx \right\rangle.$$

The above equation can be written as a classical linear Boltzmann equation[6]

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^d} f(x,t)\phi(x)\,dx = \left\langle \int_{\mathbb{R}^d} \beta_{\mathcal{F}}(x \to x')(\phi(x') - \phi(x)) f(x,t) f^\infty_{\mathcal{F}}(x')\,dx \right\rangle,$$

where $\beta_{\mathcal{F}}(x \to x') \geq 0$ is now a symmetric collision kernel

$$\beta_{\mathcal{F}}(x \to x') = \begin{cases} \frac{1}{f^\infty_{\mathcal{F}}(x')}, & \mathcal{F}(x') < \mathcal{F}(x) \\ \frac{1}{f^\infty_{\mathcal{F}}(x)}, & \mathcal{F}(x') \geq \mathcal{F}(x). \end{cases}$$

[6] Bisi, Canizo, Lods '15, '19; Toscani, Spiga '04; Michel, Mischler, Perthame '05

Entropies and steady states

Theorem

For any convex function $\Phi(x)$, *we have*

$$H_\Phi(f|f^\infty_{\mathcal{F}}) = \int_{\mathbb{R}^d} f^\infty_{\mathcal{F}}(x)\Phi\left(\frac{f(x,t)}{f^\infty_{\mathcal{F}}(x)}\right)dx \qquad \Longrightarrow \qquad \frac{dH_\Phi(f|f^\infty_{\mathcal{F}})}{dt} = -I_{\mathcal{F}}[f] \leq 0,$$

where for $h(x,y) = (x-y)(\Phi'(x) - \Phi'(y)) \geq 0$

$$I_{\mathcal{F}}[f] = \frac{1}{2}\left\langle \int_{\mathbb{R}^d} B_{\mathcal{F}}(x \to x')f^\infty_{\mathcal{F}}(x)\,h\left(\frac{f(x',t)}{f^\infty_{\mathcal{F}}(x')}, \frac{f(x,t)}{f^\infty_{\mathcal{F}}(x)}\right)dx\right\rangle$$

In the case $\Phi(x) = x\log(x) - x + 1$ we have the Shannon-Boltzmann entropy $H(f|f^\infty_{\mathcal{F}})$ for which a modified logarithmic Sobolev inequality[7]

$$I_{\mathcal{F}}[f] \geq \lambda H(f|f^\infty_{\mathcal{F}}) \Rightarrow H(f|f^\infty_{\mathcal{F}}) \leq H(f_0|f^\infty_{\mathcal{F}})e^{-\lambda t},$$

thanks to the Csiszár–Kullback inequality implies the convergence in $L_1(\mathbb{R}^d)$ of $f(x,t)$ to $f^\infty_{\mathcal{F}}(x)$.

[7] Holley, Strook '88; Miclo '92; Trouvé '96; Carlen, Carvalho '04; Toscani, Villani '99; Matthes, Toscani '12; Desvillettes, Mouhot, Villani '11

Annealing and long time behavior

In the general case where $T = T(t)$ we must take into account the normalization constant

$$\phi(x) = \log\left(\frac{f(x,t)}{f^\infty_{\mathcal{F}}(x,t)}\right) = \log(f(x,t)) + \frac{\mathcal{F}(x)}{T(t)} - \log\left(C(t)\right)$$

to get

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^d} f(x,t)\log\left(\frac{f(x,t)}{f^\infty_{\mathcal{F}}(x,t)}\right)\,dx &= \int_{\mathbb{R}^d}\frac{\partial f(x,t)}{\partial t}\left(\log(f(x,t)) + \frac{\mathcal{F}(x)}{T(t)} - \log\left(C(t)\right)\right)dx\\ &\quad -\frac{T'(t)}{T^2(t)}\int_{\mathbb{R}^d}\mathcal{F}(x)\left(f(x,t) - f^\infty_{\mathcal{F}}(x,t)\right)\,dx\end{aligned}$$

This requires $T'(t) = o(T^2(t))$ as $T(t)\to 0$. For example if $T(t)\approx 1/t$ we get $T'(t)/T(t)^2 \approx 1$ whereas for $T(t)\approx 1/\log(t)$ we get $T'(t)/T(t)^2 \approx 1/t$ and the quantity can be bounded

$$\frac{dH(f|f^\infty_{\mathcal{F}})}{dt} \le -\lambda H(f|f^\infty_{\mathcal{F}}) + \frac{c}{t}\|\mathcal{F}\|_\infty \|f - f^\infty_{\mathcal{F}}\|_1 \le -\left(\lambda - \frac{c}{t}\|\mathcal{F}\|_\infty\right) H(f|f^\infty_{\mathcal{F}}).$$

$\Rightarrow$ By Laplace principle, as $T(t)\to 0$ the equilibrium $f^\infty_{\mathcal{F}}(x,t)$ concentrates on the global minimum x^*, then also $f(x,t)$ concentrates on x^* and the solution converges to the global minimum[8].

[8] Borghi, Pareschi '24

From SA to Langevin: mean-field scaling

Let us observe that the weak form of the kinetic equation can be reformulated as follows

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^d} f(x,t)\phi(x)\,dx = \left\langle \int_{\mathbb{R}^d} (\phi(x') - \phi(x)) f(x,t)\,dx \right\rangle \\ - \left\langle \int_{\mathbb{R}^d} \left(1 - \frac{f^\infty_{\mathcal{F}}(x')}{f^\infty_{\mathcal{F}}(x)}\right) \Psi(\mathcal{F}(x') \geq \mathcal{F}(x))(\phi(x') - \phi(x)) f(x,t)\,dx \right\rangle .$$

By analogy with the grazing collision limit of the Boltzmann equation, we consider the scaling[9]

$$t \to t/\varepsilon, \quad \sigma(t) \to \sqrt{\varepsilon}\sigma(t),$$

and write for small values of $\varepsilon \ll 1$

$$\phi(x') = \phi(x) + (x'-x)\cdot\nabla_x\phi(x) + \frac{1}{2}\sum_{i,j=1}^{d}(x'_i - x_i)(x'_j - x_j)\frac{\partial^2\phi(x)}{\partial x_i \partial x_j} + O(\varepsilon^{3/2})$$

$$f^\infty_{\mathcal{F}}(x') = f^\infty_{\mathcal{F}}(x) - (x'-x)\cdot\frac{1}{T(t)}(\nabla_x\mathcal{F}(x))f^\infty_{\mathcal{F}}(x) + O(\varepsilon).$$

[9] Desvillettes '92; Villani '98; Pareschi, Toscani '13

Assuming $p(\xi)$ with mean 0 and identity covariance matrix $\Sigma = I_d$

$$\int_{\mathbb{R}^d} p(\xi)\xi_i\xi_j \, d\xi = \delta_{ij},$$

where δ_{ij} is the Kronecker delta, we formally have

$$\begin{aligned}\frac{\partial}{\partial t}\int_{\mathbb{R}^d} f(x,t)\phi(x)\,dx = {} & \frac{\sigma(t)^2}{2}\sum_{i=1}^{d}\int_{\mathbb{R}^d}\frac{\partial^2\phi(x)}{\partial x_i^2} f(x,t)\,dx \\ & - \frac{\sigma(t)^2}{2T(t)}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p(\xi)\xi\cdot\nabla_x\mathcal{F}(x)\xi\cdot\nabla_x\phi(x) f(x,t)\,d\xi\,dx.\end{aligned}$$

Taking $2T(t) = \sigma^2(t)$, we can revert to the original variables to recover the Langevin dynamics

$$\boxed{\frac{\partial f(x,t)}{\partial t} = \nabla_x\cdot\left(\nabla_x\mathcal{F}(x)f(x,t)\right) + T(t)\Delta_{xx}f(x,t).}$$

Variantions on the theme, improvements, generalizations

- Maxwellian SA.
  If $\tilde{X}^{n+1}$ is worse than X^n we interpolate with a weight proportional to the Gibbs' measure, thus avoiding acceptance/rejection[10].
- Entropy controlled SA.
  A time evolution of a temperature distribution is considered aimed at minimizing the entropy to speed up convergence of standard simulated annealing[11].
- Parallel tempering SA.
  Samples have independent temperatures, so that $f = f(x, T, t)$, which can be modified along the dynamic in order to lead low temperature samples to the global minima[12].
- Sampling.
  The ideas can be generalized to the Metropolis-Hasting sampling algorithm. The main difference lies in the transition probability which defines the kernel in the kinetic equation[13].

[10] Pareschi '24

[11] Herty, Pareschi, Zanella '24

[12] Blondeel, Pareschi '24

[13] Borghi, Pareschi '24

Maxwellian SA

We can formulate a simulated annealing-type process avoiding the acceptance-rejection dynamic.

1. We start from the trial point
2. Then, we define

$$\tilde{X}^{n+1} = X^n + \sigma^n \xi.$$

$$X^{n+1} = \begin{cases} \tilde{X}^{n+1} & \text{if } \mathcal{F}(\tilde{X}^{n+1}) - \mathcal{F}(X^n) < 0 \\ X^n + e^{-\frac{\mathcal{F}(\tilde{X}^{n+1}) - \mathcal{F}(X^n)}{T^n}} (\tilde{X}^{n+1} - X^n) & \text{if } \mathcal{F}(\tilde{X}^{n+1}) - \mathcal{F}(X^n) \geq 0. \end{cases}$$

Thus, if $\tilde{X}^{n+1}$ is worse than X^n we interpolate with a weight proportional to the Gibbs' measure.

In a continuous setting we have the update rule

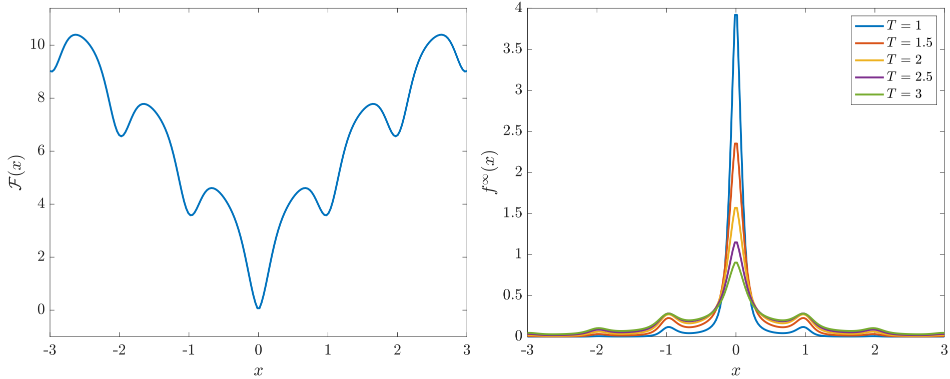
$$x' = x + B_{\mathcal{F}}(x \to x + \sigma(t)\xi)\sigma(t)\xi, \qquad B_{\mathcal{F}}(x \to x + \sigma(t)\xi) = \min\left\{1, \frac{f^{\infty}_{\mathcal{F}}(x + \sigma(t)\xi)}{f^{\infty}_{\mathcal{F}}(x)}\right\}.$$

The corresponding kinetic equation has the form of a Maxwell model and can be written as

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} f(x,t)\phi(x)\,dx = \left\langle \int_{\mathbb{R}^d} (\phi(x') - \phi(x)) f(x,t)\,dx \right\rangle.$$

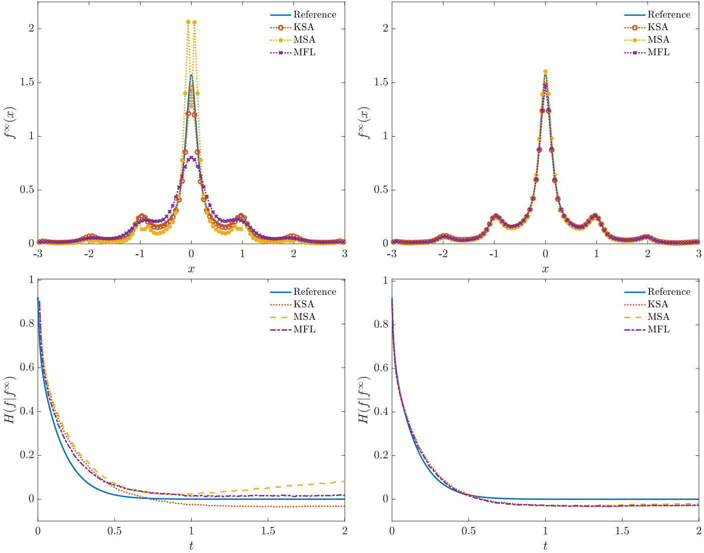
⇒ It is possible to show that the mean-field scaling yields again the Langevin dynamics.

The prototype Ackley function



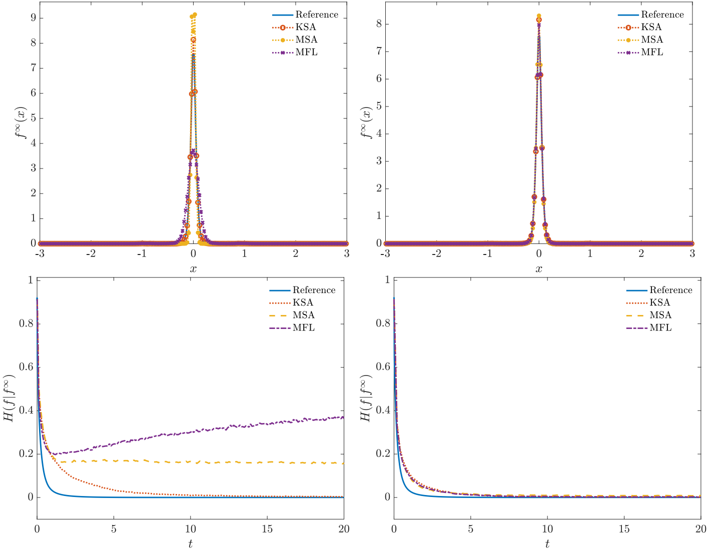
The prototype Ackley function (left) and the corresponding steady states (right) given by the Boltzmann-Gibbs measure for various values of the control temperature.

The prototype Ackley function: fixed temperature $T = 2$



Probability density (top) and relative entropy (bottom) for $\varepsilon = 0.01$ (left) and $\varepsilon = 0.0001$ (right).

The prototype Ackley function: annealing $T(t) = 2\log(2)/\log(2+t)$



Probability density (top) and relative entropy (bottom) for $\varepsilon = 0.01$ (left) and $\varepsilon = 0.0001$ (right).

Entropy controlled SA

We consider the following system of kinetic equations in weak form

$$\frac{\partial}{\partial_t}\int_{\mathbb{R}^d} f(x,t)\varphi(x)dx$$
$$= \frac{1}{2}\mathbb{E}_\xi\left[\int_{\mathbb{R}^d}(\varphi(x')-\varphi(x))(B_{\mathcal{F}}(x\to x')f(x,t)-B_{\mathcal{F}}(x'\to x)f(x',t))dx\right]$$
$$\frac{\partial}{\partial_t}\int_{\mathbb{R}_+} g(T,t)\varphi(T)dT = \mathbb{E}_\eta\left[\int_{\mathbb{R}_+}\varphi(T')-\varphi(T)g(T,t)dT\right]$$

where

$$x' = x + \sqrt{2\mathcal{D}[g]}\xi.$$

The term $\mathcal{D}[g] = \mathcal{D}[g](t) \geq 0$ depends on $g(T,t)$ and

$$T' = T - \lambda[f]T + \sqrt{\kappa(T)}\eta,$$

with $\lambda = \lambda[f] \in [0,1]$ a control parameter which depends on $f(x,t)$, and η a random variable such that $\mathbb{E}[\eta] = 0$, $\mathbb{E}_\eta[\eta^2] = 2\sigma^2 < +\infty$ and is weighted by the function $\kappa(\cdot) \geq 0$.

Mean-field entropy control

Taking $\mathcal{D}[g]$ as the mean value

$$\mathcal{D}[g](t) = \int_{\mathbb{R}_+} T g(T,t) dT,$$

one can show that

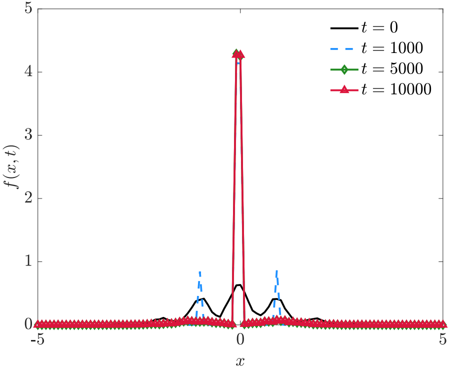
$$\frac{d}{dt} H(f|f^\infty_{\mathcal{F}})(t) = -I_H(f|f^\infty_{\mathcal{F}}) - \frac{\lambda[f](t)}{\mathcal{D}^2[g](t)} \int_{\mathbb{R}^d} \mathcal{F}(x)(f^\infty_{\mathcal{F}}(x,t) - f(x,t))dx,$$

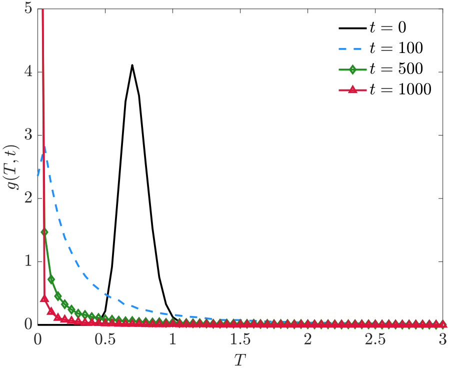
where

$$I_H(f|f^\infty_{\mathcal{F}})(t) = \int_{\mathbb{R}^d} \mathcal{D}[g](t) f(x,t) \nabla_x \log \frac{f(x,t)}{f^\infty_{\mathcal{F}}(x,t)} dx$$

Thus one can choose $\lambda[f](t)$ to speed up the convergence rate of the algorithm.

Rastrigin $d = 1$





Parallel tempering SA

In parallel tempering (PT) a collection of particles X_i^n with different temperatures T_i^n is considered. Adjacent temperatures i and j are then swapped with probability[14]

$$\exp\left[\frac{\left(\frac{1}{T_i^n}-\frac{1}{T_j^n}\right)(\mathcal{F}(X_i^{n+1})-\mathcal{F}(X_j^{n+1}))}{\bar{T}}\right],$$

where $\bar{T}$ acts as a global temperature. This is needed to control the acceptance ratio.

A kinetic model embedding SA and PT for $f = f(x,T,t)$ can be derived in the form

$$\boxed{\frac{\partial f}{\partial t} = \mathcal{L}_{\mathcal{F}}(f) + \mu J_{\mathcal{F}}(f,f)}$$

where $J_{\mathcal{F}}(f,f)$ is a Boltzmann-type operator modeling the binary particle interactions by temperature exchanges and μ is a scaling factor.

[14] Swendsen, Wang '86; Geyer '91; Marinari, Parisi '92

The weak form of this operator reads

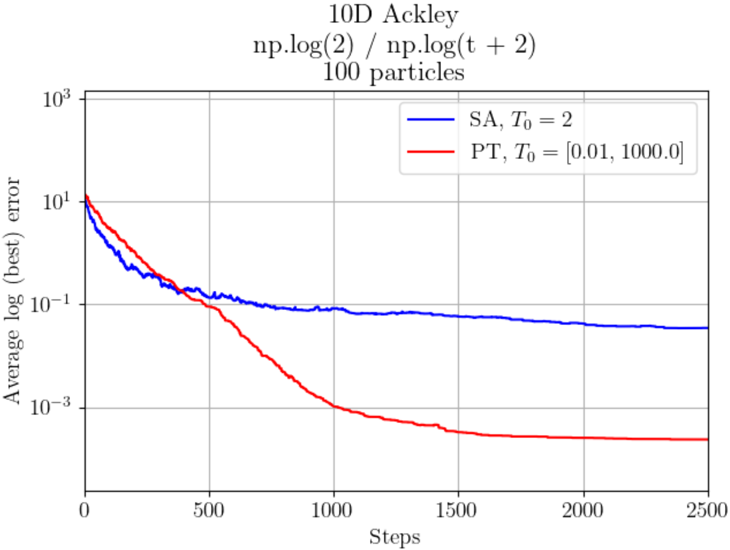
$$\int_{\mathbb{R}_+} J_{\mathcal{F}}(f,f)\phi(T)\,dT\,dx = \int_{\mathbb{R}} C_{\mathcal{F}}(x,x_*,T,T_*)(\phi(T')-\phi(T))f(x_*,T_*)f(x,T)\,dT\,dT_*\,dx\,dx_*,$$

where

$$C_{\mathcal{F}}(x,x_*,T,T_*) = \Psi(|T-T_*|<\Delta)\exp\left[\frac{\left(\frac{1}{T}-\frac{1}{T_*}\right)(\mathcal{F}(x)-\mathcal{F}(x_*))}{\bar{T}}\right]$$

with $\Psi(\cdot)$ the indicator function, $\Delta > 0$ and $T' = \eta T + (1-\eta)T_*$.

Error behavior SA vs SA+PT $\mu = 1/3$, $d = 10$



Generalizations to sampling

The above ideas can be extended to the general Metropolis-Hasting sampling algorithm.

Let $M(x)$ be a function that is proportional to the desired probability density function $f^\infty(x)$, namely, $M(x)/M(y) = f^\infty(x)/f^\infty(y)$ for $x, y \in \mathbb{R}^d$.

The kinetic formalism used in the simulated annealing case applies also to the Metropolis-Hasting process where the main difference lies in the transition probability that reads

$$B_M(x \to x') = \begin{cases} 1, & p(x|x')M(x') > p(x'|x)M(x) \\ \dfrac{p(x|x')M(x')}{p(x'|x)M(x)}, & p(x|x')M(x') < p(x'|x)M(x), \end{cases}$$

where x' is generated from a given proposal density $p(x'|x)$. The most common choices are the uniform or the normal distributions centered in x with a given variance σ.

Concluding remarks

- A kinetic/mean-field description of stochastic particle optimization methods may pave the way to a mathematical foundation of metaheuristic algorithms for global optimization.
- This entails new difficulties as we have to deal with concepts such as memory or other heuristic rules that can be very difficult to translate into differential form.
- The resulting PDEs are studied using classical trend to equilibrium tools (entropy inequalities, Wasserstain distance, asymptotic limits, . . .), enabling the design of more efficient algorithms.
- Several open problems concerning the limit as $N \to \infty$, the behavior for a finite number of particles, the dependence on the hyper-parameters, the rates of convergence . . .

Collaborators:

A. Benfenati (Milano), G. Borghi (Aachen & Ferrara), S. Grassi (Ferrara), M. Herty (Aachen), F. Blondeel (Leuven & Ferrara), M. Fornasier (Munich), P. Sünnen (Munich), H. Huang (Graz), J. Qiu (Calgary), M. Zanella (Pavia)