Kinetic equations in global optimization and applications

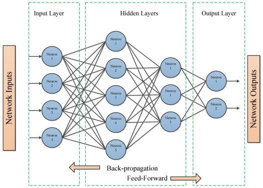
Lorenzo Pareschi

Maxwell Institute for Mathematical Sciences
& Department of Mathematics
Heriot Watt University, Edinburgh, UK

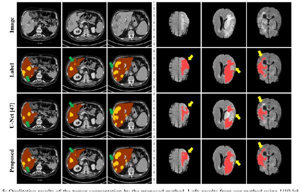
The Many Facets of Kinetic Theory, ICMS Edinburgh
September 23-27, 2024

Why this matters

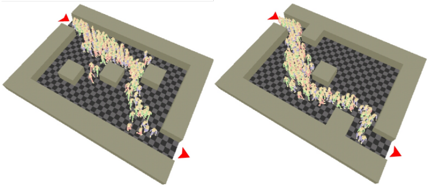
- High-dimensional (nonconvex) optimization problems are pervasive in many fields, particularly in cutting-edge areas such as machine learning, signal/image processing and optimal control.



Training neural networks



Computer assisted tomography



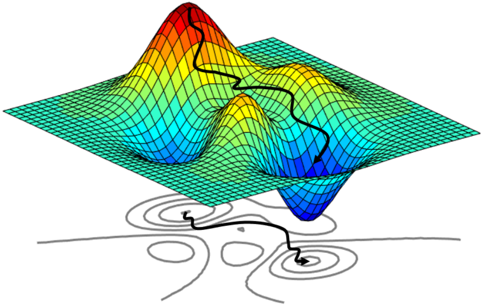
Crowd evacuation control

- Stochastic gradient descent-type methods (SGD, Adam, RMSProp, . . .), are favored for their efficiency, scalability, ability to evade critical points, and their solid mathematical foundations.
- Metaheuristic (gradient-free) algorithms gained popularity due to the minimal assumptions on the optimization problem, making them versatile and applicable to a wider range of problems.

Metaheuristic optimization

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.

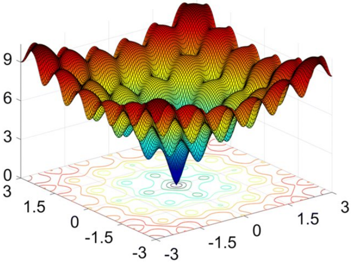
- Metropolis-Hastings (1953,1970)
- Simplex Heuristics (1965)
- Evolutionary Programming (1966)
- Genetic Algorithms (GA) (1975)
- Simulated Annealing (SA) (1983)
- Particle Swarm Optimization (PSO) (1995)
- Ant Colony Optimization (ACO) (1997)
- . . .



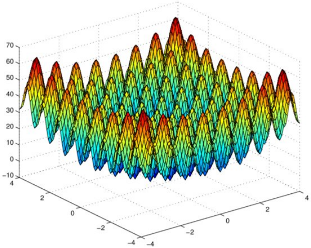
⇒ Despite the significant empirical success, most results are experimental in nature and lack a rigorous mathematical foundation.

Metaheuristics in action

Ackley function



Rastrigin function



Examples of swarm-based optimization processes

CBO methods: a mean-field perspective on metaheuristcs

Consider the optimization problem

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

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

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

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

where $\lambda > 0$ and $\sigma > 0$ characterize the alignment and exploration strength;

$D(X_t) = |X_t| I_d$ (isotropic) or $D(X_t) = \text{diag}\,\{(X_t)_1, (X_t)_2, \dots, (X_t)_d\}$ (anisotropic)

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

[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; . . .

CBO in action

Mean-field limit of CBO

The behavior of the CBO system for $N \gg 1$ is obtained by assuming that the (X_t^i), $i = 1, \dots, N$ are independent with the same distribution $\rho(x,t)$ (propagation of chaos assumption)

$$\rho_N(x,t) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_t^i) \approx \rho(x,t), \qquad \bar{X}_t^\alpha \approx \bar{x}^\alpha(\rho) = \frac{\int_{\mathbb{R}^d} x\, e^{-\alpha \mathcal{F}(x)} \rho(x,t)dx}{\int_{\mathbb{R}^d} e^{-\alpha \mathcal{F}(x)} \rho(x,t)dx}.$$

Under the propagation of chaos assumption, the update rule becomes independent of the index i and can be re-written as a mono-particle process.
In such a situation, the dynamics (anisotropic) is approximated by the Fokker–Planck equation[2]

$$\boxed{\partial_t \rho = \underbrace{\lambda \nabla_x \cdot (x - \bar{X}^\alpha(\rho))\rho}_{\text{transport}} + \underbrace{\frac{\sigma^2}{2} \Delta_x \left(\|x - \bar{X}^\alpha(\rho)\|_2^2 \rho \right)}_{\text{diffusion}},}$$

whereas in the anisotropic case the diffusion term is replaced by[3]

$$\frac{\sigma^2}{2} \sum_{j=1}^{d} \partial_{jj}((x - \bar{X}^\alpha(\rho))_j^2 \rho).$$

[2] Pinnau, Totzeck, Tse, Martin '17; Carrillo, Choi, Totzeck, Tse '18

Convergence to global minimum

Theorem (Carrillo, Choi, Totzeck, Tse '18)

If $\sigma^2 < 2\lambda/d$ and $\alpha \gg 1$, the variance $V(t) \to 0$ and the expectation $\mathbb{E}[\bar{X}_t^\alpha] \to \tilde{x}$. When $\alpha \to +\infty$ and $\mathcal{F}$ has a unique global minimizer, under reasonable assumptions on $\mathcal{F}$, we have $\tilde{x} \approx x^*$ the global minimum[4].

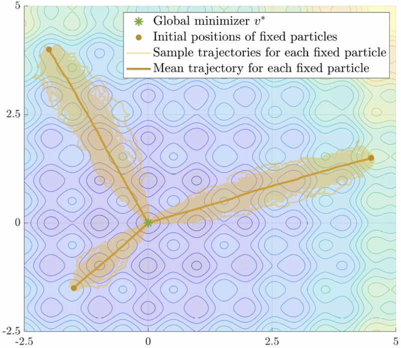
In the isotropic case the variance satisfies

$$\frac{dV(\rho)}{dt} = -\left(2\lambda - \sigma^2 d\right) V(\rho) + \frac{d\sigma^2}{2} \|\mathbb{E}(\rho) - \bar{X}^\alpha(\rho)\|_2^2.$$

The second term is controlled if $V(\rho_0)$ satisfies some boundedness assumptions. Next one shows that $\mathcal{F}(\tilde{x}) \approx \mathcal{F}(x^*)$ for $\alpha \gg 1$ and $\mathcal{F} \in C^2(\mathbb{R}^d)$ with boundedness assumptions on $\Delta_{xx}\mathcal{F}$. In the anisotropic case, the dimensional dependence on σ^2 is removed.

[4] Carrillo, Choi, Totzeck, Tse '18; Carrillo, Jin, Li, Zhu '20;

Convergence as a minimizer of the square distance from x^*



Individual agents follow, on average, the gradient flow of the map $x \mapsto \|x - x^*\|_2^2$
Consider the energy functional

$$\mathcal{V}(\rho) = \frac{1}{2}\int_{\mathbb{R}^d} \|x - x^*\|_2^2 \rho \, dx = \frac{1}{2} W_2^2(\rho, \delta_{x^*})$$

where W_2^2 is Wasserstein-2 distance.

Then $\mathcal{V}(\rho) \to 0$ simultaneously shows consensus formation and convergence of ρ to the Dirac delta δ_{x^*} with respect to the Wasserstein distance. This can be achieved under less restrictive conditions[5].

[5] M. Fornasier, T. Klock, K. Riedl '22

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?
- Are there any discernible links between CBO and established metaheuristics?
- Could this approach enhance our understanding of the relationship between metaheuristics and gradient-based methods?

Three tales of kinetic equations in global optimization

1. Simulated Annealing (SA) and linear kinetic equations

2. Genetic Algorithms (GA) and Boltzmann equations

3. Particle Swarm Optimization (PSO) and Vlasov-Fokker-Planck equations

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.

⇒ The algorithms are part of the Matlab Global Optimization Toolbox: `simulannealbnd`, `ga`, `particleswarm`.

Tale I:

Simulated Annealing and linear kinetic equations

There is a deep and useful connection between statistical mechanics (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and multivariate or combinatorial optimization (finding the minimum of a given function depending on many parameters) . . . This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods.

(S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by Simulated Annealing, Science, 1983)

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[6]

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

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

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

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

Annealing process

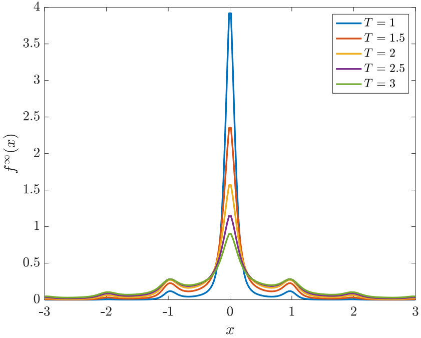
By the Laplace principle

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

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

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

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



Slowly decreasing $T(t)$ so that the solution approaches $f_{\mathcal{F}}^{\infty}(x)$ 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[7].

⇒ It requires the gradient evaluation, in contrast with the gradient-free nature of SA algorithm.

⇒ Derivation of the SDE Langevin diffusion from Metropolis-Hasting[8].

[7] Hajek '88

[8] Gelfand '87; 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[9]

$$\begin{aligned}\frac{\partial f(x,t)}{\partial t} &= \mathcal{L}_{\mathcal{F}}(f(x,t))\\ \mathcal{L}_{\mathcal{F}}(f(x,t)) &= \underbrace{\mathbb{E}\left[B_{\mathcal{F}}(x' \to x) f(x',t)\right]}_{\text{gain}} - \underbrace{\mathbb{E}\left[B_{\mathcal{F}}(x \to x')\right] f(x,t)}_{\text{loss}}\end{aligned}$$

where $\mathbb{E}[\cdot]$ 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 $x \to x'$.

[9] 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

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

The above equation can be written as a classical linear Boltzmann equation[10]

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

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

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

[10] 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 $h(x,y) = (x-y)(\Phi'(x) - \Phi'(y)) \geq 0$

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

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[11]

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

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

[11] 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

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^d} f(x,t)\log\left(\frac{f(x,t)}{f^\infty_{\mathcal{F}}(x,t)}\right)\,dx = &\int_{\mathbb{R}^d}\frac{\partial f(x,t)}{\partial t}\left(1+\log\left(\frac{f(x,t)}{f^\infty_{\mathcal{F}}(x,t)}\right)\right)\,dx\\ &-\frac{T'(t)}{T^2(t)}\int_{\mathbb{R}^d}\mathcal{F}(x)\left(f(x,t)-f^\infty_{\mathcal{F}}(x,t)\right)\,dx\end{aligned}$$

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

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

which, thanks to a Grönwall's Lemma-type argument, leads to the desired entropy decay for $t \gg 1$.
$\Rightarrow$ 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[12].

[12] 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

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

By analogy with the grazing collision limit of the Boltzmann equation, we consider the scaling[13]

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

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

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

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

[13] 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

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

Maxwellian simulated annealing

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

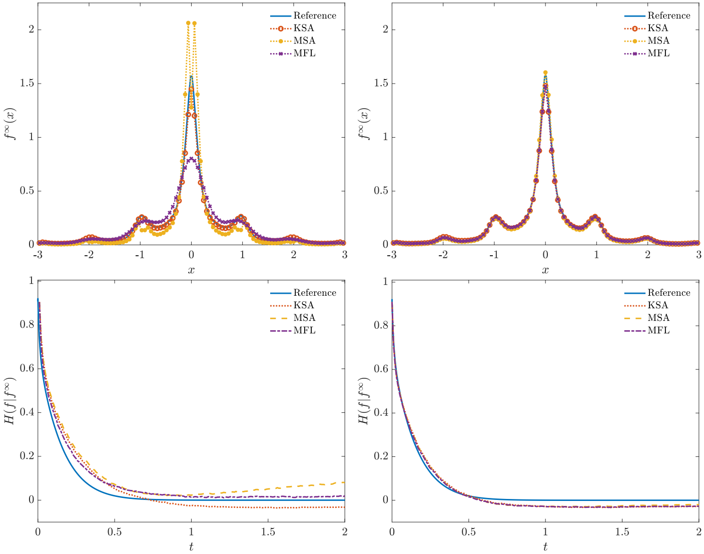
$$x' = x + B_{\mathcal{F}}(x \to x + \sigma(t)\xi)\sigma(t)\xi, \qquad B_{\mathcal{F}}(x \to x + \sigma(t)\xi) = \min\left\{1, \frac{f^\infty_{\mathcal{F}}(x + \sigma(t)\xi)}{f^\infty_{\mathcal{F}}(x)}\right\}.$$

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

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} f(x,t)\phi(x)\, dx = \mathbb{E}_\xi \left[\int_{\mathbb{R}^d} (\phi(x') - \phi(x)) f(x,t)\, dx \right].$$

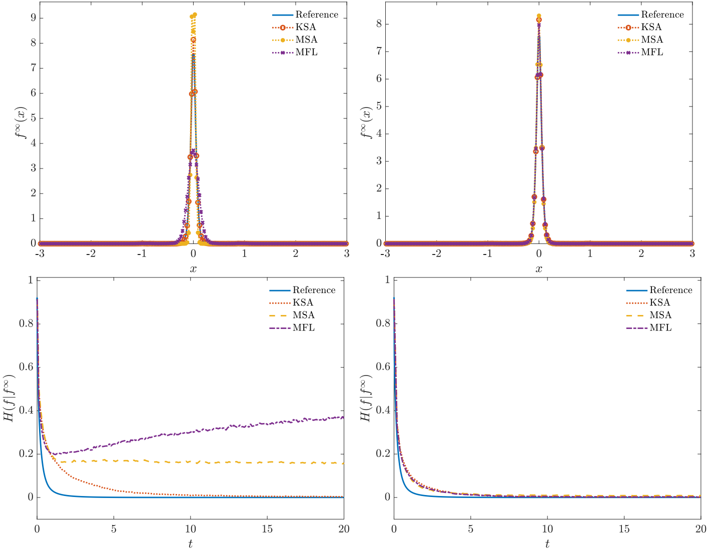
⇒ It is possible to show that the mean-field scaling yields again the Langevin dynamics.

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).

Generalizations and improvements

- 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.
- 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[14].
- Parallel tempering SA.
  Collective behavior of samples with different temperatures, so that $f = f(x, T, t)$, which learn along the dynamic how to lower the temperature and reach the global minima[15].

[14] Herty, Pareschi, Zanella '24

[15] Blondeel, Pareschi '24

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.

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, and η a random variable such that $\mathbb{E}_\eta[\eta] = 0$, $\mathbb{E}_\eta[\eta^2] = 2\sigma^2 < +\infty$, 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.

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[16]

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

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.

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

The weak form of this operator reads

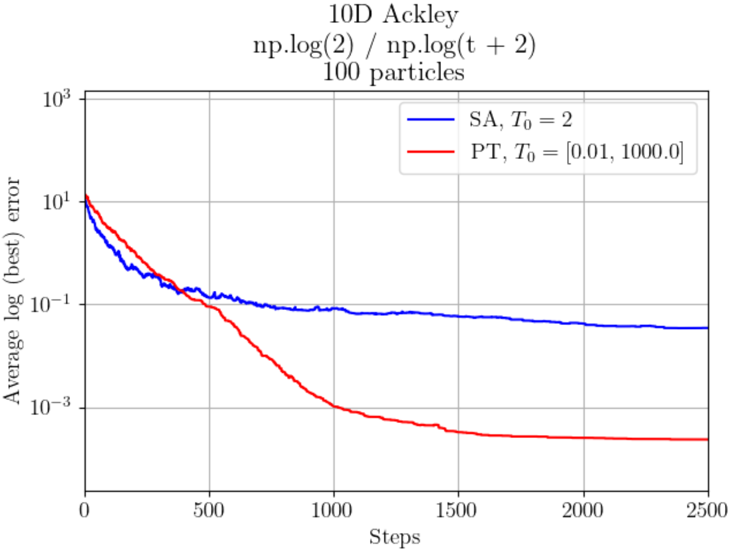
$$\int_{\mathbb{R}_+} J_{\mathcal{F}}(f,f)\phi(T)\,dT\,dx = \int_{\mathbb{R}} C_{\mathcal{F}}(x,x_*,T,T_*)(\phi(T_*)-\phi(T))f(x_*,T_*)f(x,T)\,dT\,dT_*\,dx\,dx_*,$$

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, $\Delta > 0$.

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



Tale II:

Genetic Algorithms and Boltzmann equations

. . . Finally, and quite important for future studies, genetic algorithms began to be seen as a theoretical tool for investigating the phenomena generated by complex adaptive systems - a collective designation for nonlinear defined systems designation systems by the interaction of large numbers of adaptive agents (economies, political systems, ecologies, immune systems, developing embryos, brains, and the like).

(John H. Holland, Adaptation in natural and artificial systems, MIT Press, 1992)

Genetic algorithms

J.H. Holland

Genetic algorithms iteratively evolve a population of points to find better solutions. Starting from $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$ the algorithm can be summarized as follows[c].

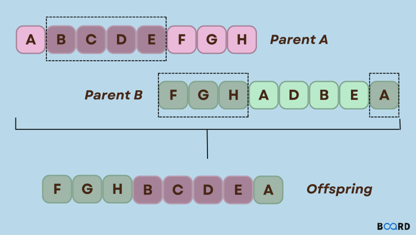
1. Selection: select a group of individuals that have better fitness in the current population, called parents, who contribute with their genes—the entries of their vectors—to their children.
2. Given a pair of parents, children are generated according to two main evolutionary dynamics:
    - Crossover: by combining the vectors of a pair of parents in different ways.
    - Mutation: by introducing random changes, or mutations, to a single parent.
3. The individuals which are directly passed to the next generation can be either chosen randomly or following a fitness-based mechanism (elite group).

[c] Holland '92; Touring '50.

Evolution through binary interactions

Let $\odot$ denote the component-wise (Hadamard) product.
Given a pair of parents (x, x_*) selected according to $\mathcal{B}_{\mathcal{F}}(x, x_*) \geq 0$, the process leading to an offspring x' is

$$x' = \underbrace{(1-\gamma) \odot x + \gamma \odot x_*}_{\text{crossover}} + \underbrace{\mathcal{D} \odot \xi}_{\text{mutation}}$$



where $\gamma \in [0,1]^d$ is a crossover vector, $\xi \sim p(\xi)$ a random vector in $\mathbb{R}^d$, with zero mean and identity covariance matrix $\Sigma = I_d$, and $\mathcal{D} \in \mathbb{R}^d$ is a mutation vector assumed time dependent.

We assume that the number of individuals is the same for all generations and that after the generating procedure is repeated for a fraction of selected individuals the remaining individuals are directly taken from the previous generation, eventually according to the elitist strategy.

A Boltzmann description of genetic algorithms

If parents are replaced by children, the mathematical description of the above process for large numbers of interacting particles can resort on a Boltzmann type equation that in weak-form reads[17]

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^d} f(x,t)\phi(x)\,dx = \mathbb{E}\left[\int_{\mathbb{R}^{2d}} \mathcal{B}_{\mathcal{F}}[f](x,x_*)\,(\phi'-\phi)\,f(x,t)f(x_*,t)\,dx\,dx_*\right]$$

where ϕ is a smooth function, $\phi' = \phi(x')$, $\phi'_* = \phi(x'_*)$, $\phi_* = \phi(x_*)$.

We use the notation $\mathbb{E}\left[g\right] = \int_{\mathbb{R}^d} g(\xi)p(\xi)\,d\xi$, to denote the mathematical expectation with respect to the random vector ξ entering the definitions of x'.

$\Rightarrow$ The only collision invariant is the total mass of the particles. We expect the momentum to drift towards the global minimum and, by suitable mutation choices, the variance to vanish as new generations are created.

[17] Borghi, Pareschi '23

Selection kernels

Popular selection methods are Fitness based selection using the Boltzmann-Gibbs measure

$$\mathcal{B}_{\mathcal{F}}[f](x,x_*) = \frac{e^{-\alpha(\mathcal{F}(x)+\mathcal{F}(x_*))}}{\left(\int e^{-\alpha\mathcal{F}(x)} f(x,t)\,dx\right)^2},$$

and Ranked based selection

$$\mathcal{B}_{\mathcal{F}}[f](x,x_*) = \frac{\int_{\mathcal{F}(x)\le\mathcal{F}(y)} f(y,t)\,dy\ \int_{\mathcal{F}(x_*)\le\mathcal{F}(y)} f(y,t)\,dy}{\left(\int \int_{\mathcal{F}(x)\le\mathcal{F}(y)} f(y,t)\,dy\ f(x,t)\,dx\right)^2}.$$

$\Rightarrow$ As in simulated annealing, we can embed the selection mechanism into the interaction rule and consider Maxwellian-type models in which the selection kernel does not appear.

Maxwell-type models for Boltzmann-Gibbs selection

Given the pair (x, x_*) we write the update rule[18]

$$\begin{aligned} x' &= x(1-\lambda\gamma_\alpha(x,x_*)) + \lambda\gamma_\alpha(x,x_*)x_* + \sigma D(x,x_*)\xi, \\ &= x(1-\lambda) + \lambda x^\alpha(x,x_*) + \sigma D(x,x_*)\xi, \end{aligned}$$

where $\lambda > 0$, $\sigma > 0$, and x^α is a local minimum estimate

$$\gamma_\alpha(x,x_*) = \frac{e^{-\alpha\mathcal{F}(x_*)}}{e^{-\alpha\mathcal{F}(x)} + e^{-\alpha\mathcal{F}(x_*)}}, \quad x^\alpha(x,x_*) = \frac{x e^{-\alpha\mathcal{F}(x)} + x_* e^{-\alpha\mathcal{F}(x_*)}}{e^{-\alpha\mathcal{F}(x)} + e^{-\alpha\mathcal{F}(x_*)}}$$

$$D(x,x_*) = \operatorname{diag}\left\{\gamma_\alpha(x,x_*)(x_*-x)_1, \dots, \gamma_\alpha(x,x_*)(x_*-x)_d\right\}.$$

The time evolution of the expected position $m(t)$ for $\phi(x) = x$ satisfies

$$\frac{dm(t)}{dt} = 2\lambda \int_{\mathbb{R}^{2d}} \gamma_\alpha(x,x_*) f(x,t) f(x_*,t) x_*\, dx_*\, dx - \lambda m(t),$$

[18] Benfenati, Borghi, Pareschi '22

Large time behavior: decay of the variance

Assumption (I)

Let us assume $\mathcal{F}(y)$ positive and for all $y \in \mathbb{R}^d$

$$\underline{\mathcal{F}} := \inf_{x\in\mathbb{R}^d} \mathcal{F}(x) \le \mathcal{F}(y) \le \sup_{x\in\mathbb{R}^d} \mathcal{F}(x) =: \overline{\mathcal{F}}\,.$$

Proposition

Under Assumption I if α is sufficiently large, for the variance $V(t)$ we have

$$\frac{dV(t)}{dt} \le -\left(\frac{\lambda}{C_\alpha} - \lambda^2 - \sigma^2\right) V(t)\,,$$

for all $t > 0$, where $C_\alpha := e^{\alpha(\overline{\mathcal{F}}-\underline{\mathcal{F}})}$.

Therefore, if $\sigma^2 < \dfrac{\lambda}{C_\alpha} - \lambda^2$ there exits $\tilde{x} \in \mathbb{R}^d$ s.t. $m(t) \to \tilde{x}$, $V(t) \to 0$ as $t \to \infty$.

Convergence to global minimum

Assumption (II)

$\mathcal{F} \in \mathcal{C}^2(\mathbb{R}^d)$ and there exist $c_1, c_2 > 0$ such that

1. $\sup\limits_{x\in\mathbb{R}^d} |\nabla \mathcal{F}(x)| \leq c_1$;
2. $\sup\limits_{x\in\mathbb{R}^d} \|\nabla^2 \mathcal{F}(x)\|_2 \leq c_2$.

Theorem

If the model parameters $\{\lambda, \sigma, \alpha\}$ and the initial data f_0 satisfy

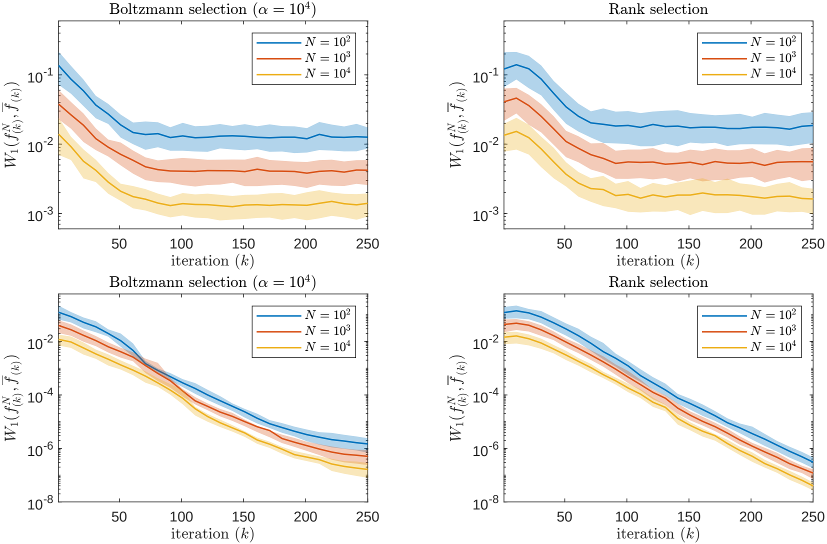
$$\mu := \frac{\lambda}{C_\alpha} - \lambda^2 - \sigma^2 > 0, \quad \nu := \frac{2(\sqrt{2}\lambda c_1 + (\lambda^2+\sigma^2)c_2)\alpha e^{-\alpha \underline{\mathcal{F}}}}{\mu \|e^{-\alpha \mathcal{F}}\|_{L^1(f_0)}} \max\{\sqrt{V(0)}, V(0)\} < \frac{1}{2}$$

then there exists $\tilde{x} \in \mathbb{R}^d$ such that $m(t) \longrightarrow \tilde{x}$ as $t \to \infty$.
Moreover, it holds

$$\mathcal{F}(\tilde{x}) \leq \underline{\mathcal{F}} + r(\alpha) + \frac{\log 2}{\alpha}$$

where, if $x^* \in$ supp(f_0), then $r(\alpha) = -\frac{1}{\alpha}\log \|e^{-\alpha\mathcal{F}}\|_{L^1(f_0)} - \underline{\mathcal{F}} \longrightarrow 0$ as $\alpha \to \infty$.

Evolution of Wasserstein distance



Boltzmann selection ($\alpha = 10^4$)

Rank selection

Boltzmann selection ($\alpha = 10^4$)

Rank selection

Constant mutation (top) vs decreasing mutation (bottom) strengths. Ackley function for $d = 1$.

Adding global information

If one considers the modified update rule based on global macroscopic information

$$x' = x(1-\lambda) + \lambda \bar{x}_\alpha(t) + \sigma D(x - \bar{x}_\alpha(t))\xi,$$

where now $\bar{x}_\alpha(t)$ is a CBO-type estimate of the global minimum

$$\bar{x}_\alpha(t) = \frac{\int_{\mathbb{R}^d} x \exp^{-\alpha \mathcal{F}(x)} f(x,t)\,dx}{\int_{\mathbb{R}^d} \exp^{-\alpha \mathcal{F}(x)} f(x,t)\,dx}, \quad D(x - \bar{x}_\alpha(t)) = \operatorname{diag}\left\{(\bar{x}_\alpha - x)_1, \dots, (\bar{x}_\alpha - x)_d\right\},$$

it is possible to show convergence to the global minimum under weaker conditions, which mitigates the limitations induced by large α for local microscopic information.

$\Rightarrow$ In practice we combine the two dynamics, local best alignment and global best alignment, giving rise to a method with four parameters $\lambda_1, \sigma_1, \lambda_2, \sigma_2$ that govern the intensity of the local and global effects, respectively.

From GA to CBO: mean-field scaling

If we introduce the following scaling for $\varepsilon \ll 1$

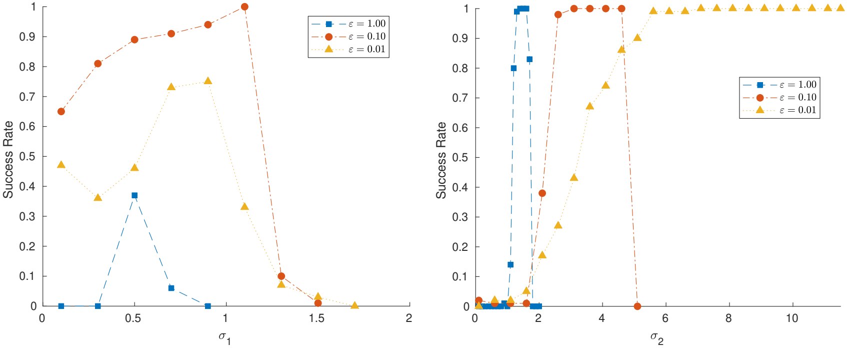
$$t \to t/\varepsilon, \qquad \lambda \to \lambda\varepsilon, \qquad \sigma \to \sigma\sqrt{\varepsilon}.$$

One formally shows that as $\varepsilon \to 0$, the microscopic dynamic lead to a modified mean-field CBO

$$\begin{aligned} \frac{\partial f(x,t)}{\partial t} &+ \lambda \nabla_x \left(f(x,t) \int_{\mathbb{R}^d} \gamma_\beta^{\mathcal{F}}(x,x_*)(x_* - x) f(x_*,t)\, dx_* \right) \\ &= \frac{\sigma^2}{2} \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2} \left(f(x,t) \int_{\mathbb{R}^d} \gamma_\beta^{\mathcal{F}}(x,x_*)^2 (x_{*,i} - x_i)^2 f(x_*,t)\, dx_* \right). \end{aligned}$$

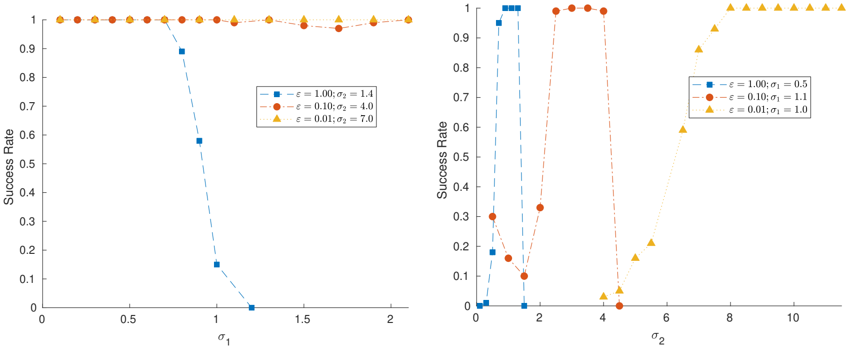
whereas the macroscopic dynamic leads to the the standard mean-field CBO.

A validation example



Minimization of Rastrigin function in $d = 20$ with $N = 200$ particles. Left image refers to the local best only ($\lambda_2 = \sigma_2 = 0$, $\lambda_1 = 1$), while the right one refers to the global best only ($\lambda_1 = \sigma_1 = 0$, $\lambda_2 = 1$). Simulation is successful if $\|x_{\alpha,\mathcal{F}} - v^*\| < 0.25$.

A validation example



Minimization of Rastrigin function in $d = 20$ with $N = 200$ particles. Left image refers to the optimal value of σ_2 for the global best, while the right one refers to the optimal value of σ_1 for the local best. Simulation is successful if $\|x_{\alpha,\mathcal{F}} - v^*\| < 0.25$.

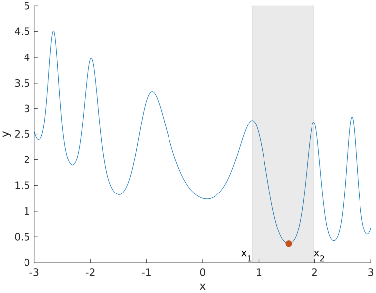
Comparison with Stochastic Gradient Descent

We want to minimize

$$L(x) = \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i)$$

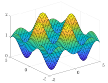
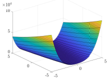
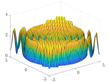
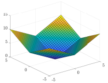
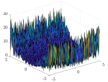
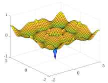
$$f(x,\xi_i) = \exp\left(\sin(2x^2)\right) + \frac{1}{10}\left(x - \xi_i - \frac{\pi}{2}\right)^2,$$

$\xi_i \sim \mathcal{N}(0, 0.01)$.



Method	ε	σ_1	σ_2	Success Rate
SGD	*	*	*	18.00%
KBO	1	0.5	0.5	98.50%
KBO	0.1	1.0	1.3	100.00%
KBO	0.01	1.0	6.5	98.70%

We fixed $N = 20$ and the maximum iterations number $N_t = 100$. Here $n = 10000$

Name	Function $\mathcal{F}(x)$	Range	x^*	$\mathcal{F}(x^*)$	Sketch in 2D
Griewank	$1+\sum_{i=1}^{d}\frac{(x_i)^2}{4000}-\prod_{i=1}^{d}\cos\left(\frac{x_i}{i}\right)$	$[-600,600]^d$	$(0,\dots,0)$	0	
Rosenbrock	$1-\cos\left(2\pi\sqrt{\sum_{i=1}^{d}(x_i)^2}\right)+0.1\sqrt{\sum_{i=1}^{d}(x_i)^2}$	$[-5,10]^d$	$(1,\dots,1)$	0	
Salomon	$1-\cos\left(2\pi\sqrt{\sum_{i=1}^{d}(x_i)^2}\right)+0.1\sqrt{\sum_{i=1}^{d}(x_i)^2}$	$[-100,100]^d$	$(0,\dots,0)$	0	
Schwefel 2.20	$\sum_{i=1}^{d}\lvert x_i\rvert$	$[-100,100]^d$	$(0,\dots,0)$	0	
XSY random	$\sum_{i=1}^{d}\eta_i\lvert x_i\rvert^i, \qquad \eta_i\sim\mathcal{U}(0,1)$	$[-5,5]^d$	$(0,\dots,0)$	0	
XSY 4	$\left(\sum_{i=1}^{d}\sin^2(x_i)-e^{-\sum_{i=1}^{d}(x_i)^2}\right)e^{-\sum_{i=1}^{d}\sin^2\sqrt{\lvert x_i\rvert}}$	$[-10,10]^d$	$(0,\dots,0)$	-1	

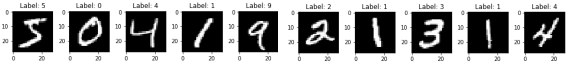
High dimensional benchmark functions ($d = 50$)

Function		$\delta = 0.25$	$\delta = 0.1$	Function		$\delta = 0.1$	$\delta = 0.25$
Salomon	SR	100%	100%	Rastrigin	SR	75%	84%
	Iters	6306	10000		Iters	3893	2320
	Error	9.64e-02	4.92e-02		Error	6.91e-01	2.23e-05
	Fval	0.96	0.49		Fval	0.25	8.95e-7
	N_a	133	215		N_a	182	804
Griewank	SR	100%	100%	Schwefel 2.22	SR	100%	100%
	Iters	2722	1696		Iters	2165	1631
	Error	9.22e-03	7.29e-03		Error	1.27e-03	1.49e-06
	Fval	2.49e-2	1.04e-2		Fval	0.27	6.9e-4
	N_a	258	985		N_a	335	1017
StyLank	SR	77%	100%	Schwefel 2.23	SR	100%	100%
	Iters	5923	2062		Iters	10000	10000
	Error	4.56e-03	4.70e-05		Error	4.53e-02	4.69e-02
	Fval	-1958.29	-1958.29		Fval	1e-5	3.74e-8
	N_a	132	874		N_a	75	215
Neg. Exp.	SR	100%	100%	Sphere	SR	100%	100%
	Iters	2517	1325		Iters	2368	1529
	Error	1.11e-03	1.40e-03		Error	1.02e-03	1.88e-04
	Fval	-1	-1		Fval	1.00e-5	9.35e-7
	N_a	271	1129		N_a	291	1051
Sum of Square	SR	100%	100%	Ackley	SR	100%	100%
	Iters	2788	1719		Iters	2701	1674
	Error	1.15e-03	2.96e-05		Error	1.69e-03	3.87e-06
	Fval	2.93e-3	1.02e-6		Fval	3.32e-2	7.00e-5
	N_a	252	966		N_a	259	994

Results averaged 100 times starting with $N = 2000$. Success Rate (SR), the Average number of iteration (Iters), the mean square error (Error) and the the average functions values (Fval) achieved on successful runs, and the average number of particles N_a used along the simulation. Simulation is successful if $\|x_{\alpha,\mathcal{F}} - v^*\| < \delta$.

Application to a machine learning problem

Recognize digital numbers contained in images of the MNIST dataset



of 28×28 images by using a shallow network $f(x; W, b) = \text{softmax}\left(\text{ReLU}\left(Wx + b\right)\right)$, where $x \in \mathbb{R}^{784}, W \in \mathbb{R}^{10\times 784}$, and a bias $b \in \mathbb{R}^{10}$. Moreover

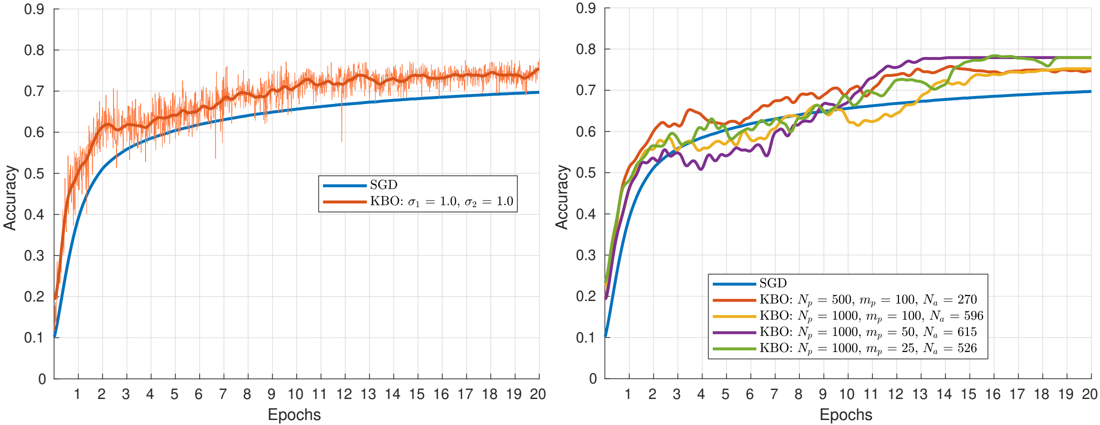
$$\text{softmax}(x) = \frac{e_i^x}{\sum_i e_i^x}, \quad \text{ReLU(x)} = \max(0, \text{x})$$

being ReLU the Rectified Linear Unit function. The training consists in minimizing the following function for $n = 10^4$

$$L(X, y; f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(X^{(i)}; W, b), y^i\right), \quad \ell(x, y) = -\sum_{i=1}^{10} y_i \log(x_i)$$

where X is the training dataset of vectorized images ($\mathbb{R}^{28\times 28} \to \mathbb{R}^{784}$) and the function ℓ is the cross entropy.

Recognize digital numbers: MNIST dataset



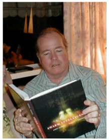
The plot on the left depicts the performance using $N = 500$ without any particle reduction strategy, while the plot on the right refers to particle reduction with different choices for particle numbers N and particles' batch m_p. The average number of particles is denoted by N_a. Here $\lambda_1 = \lambda_2 = 1$, $\sigma_1 = \sigma_2 = 1$, $\varepsilon = 0.1$, $\alpha = \beta = 5 \times 10^6$.

Tale III:
Particle Swarm Optimization and Vlasov-Fokker-Planck equations

Particle swarm optimization has roots in two main component methodologies. Perhaps more obvious are its ties to artificial-life in general, and to birds flocking, fish schooling and swarming theory in particular. It is also related, however, to evolutionary computation, and has ties to both genetic algorithms and evolutionary programming.

(J. Kennedy; R. C. Eberhart, Particle swarm optimization, IEEE Proceedings of ICNN'95, 1995)

Particle Swarm Optimization

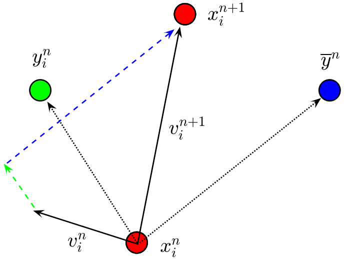


J. Kennedy

R.C. Eberhart

Particle swarm optimization (PSO) exploits the behavior of N particles with position $x_i \in \mathbb{R}^d$ and velocity $v_i \in \mathbb{R}^d$, $i = 1, \dots, N$ accordingly to[a]

$$
\begin{aligned}
x_i^{n+1} &= x_i^n + v_i^{n+1}, \\
v_i^{n+1} &= m v_i^n + \underbrace{\frac{c_1}{2}\left(y_i^n - x_i^n\right) + \frac{c_2}{2}\left(\bar{y}^n - x_i^n\right)}_{\text{alignment}} + \underbrace{\frac{c_1}{2} R_1^n \left(y_i^n - x_i^n\right) + \frac{c_2}{2} R_2^n \left(\bar{y}^n - x_i^n\right)}_{\text{exploration}}
\end{aligned}
$$



Local best and global best influence

- $\bar{y}^n$ is the global best position given by argmin$(\mathcal{F}(x_1^n), \dots, \mathcal{F}(x_N^n), \mathcal{F}(\bar{y}^{n-1}))$;
- y_i^n is the local best position;
- $m \in (0, 1]$ is the inertia weight;
- R_1^n, R_2^n are d-dimensional diagonal matrices of random numbers with distribution $\mathcal{U}(-1, 1)$;
- $c_1, c_2 \in \mathbb{R}$ are acceleration coefficients.

[a] Kennedy, Eberhart ’95; Kennedy ’97

A time discrete formulation of PSO

The obtain a differential formulation of PSO, a major difficulty consists in the presence of particle memory. To this aim we rewrite the local best y_i^{n+1} as

$$y_i^{n+1} = y_i^n + \frac{1}{2}\left(x_i^{n+1} - y_i^n\right)\left(1 + \text{sign}\left(\mathcal{F}(y_i^n) - \mathcal{F}(x_i^{n+1})\right)\right)$$

so that the PSO method can be generalized to the time discrete formalism

$$\begin{aligned} X_i^{n+1} &= X_i^n + \Delta t\, V_i^{n+1}, \\ Y_i^{n+1} &= Y_i^n + \nu\,\Delta t\left(X_i^{n+1} - Y_i^n\right)\left(1 + \text{sign}\left(\mathcal{F}(Y_i^n) - \mathcal{F}(X_i^{n+1})\right)\right), \\ \\ m\,V_i^{n+1} &= m\,V_i^n - (1-m)\,V_i^{n+1} + \lambda_1\,\Delta t\,(Y_i^n - X_i^n) + \lambda_2\,\Delta t\left(\bar{Y}^n - X_i^n\right) \\ &+ \sigma_1\sqrt{\Delta t}\,\tilde{R}_1^n(Y_i^n - X_i^n) + \sigma_2\sqrt{\Delta t}\,\tilde{R}_2^n D(\bar{Y}^n - X_i^n) \end{aligned}$$

- $\tilde{R}_k^n$, $k = 1, 2$ diagonal matrices of uniform random numbers with mean 0 and variance 1;
- $\lambda_k = \frac{c_k}{2}$, $\sigma_k = \frac{c_k}{2\sqrt{3}}$, $\quad k = 1, 2$; Classic PSO if $\Delta t = 1$, $\nu = 1/2$.

Stochastic Differential PSO (SD-PSO)

The system corresponds to a discretization of the following second order system of SDEs[19]:

$$
\begin{aligned}
dX_t^i &= V_t^i dt, \\
dY_t^i &= \underbrace{\nu \left(X_t^i - Y_t^i\right) S^\beta(X_t^i, Y_t^i) dt,}_{\text{memory effect}} \\
m dV_t^i &= -(1-m)V_t^i dt + \lambda_1 \left(Y_t^i - X_t^i\right) dt + \lambda_2 \left(\bar{Y}_t^\alpha - X_t^i\right) dt \\
&\quad + \sigma_1 D(Y_t^i - X_t^i) dB_t^{1,i} + \sigma_2 D(\bar{Y}_t^\alpha - X_t^i) dB_t^{2,i},
\end{aligned}
$$

- $B_t^{k,i}, \quad k = 1, 2$ denote independent Brownian motions;
- $D(Y_t) = \text{diag}\left\{(Y_t)_1, (Y_t)_2, \ldots, (Y_t)_d\right\}$;
- $S^\beta(x, y) = 1 + \tanh\left(\beta(\mathcal{F}(y) - \mathcal{F}(x))\right)$ is a sigmoid that for $\beta \gg 1$ approximates the $1 + \text{sign}(\cdot)$ function;
- $\bar{Y}_t^\alpha = \frac{1}{\sum_i \omega_{\mathcal{F}}^\alpha(Y_t^i)} \sum_i Y_t^i \omega_{\mathcal{F}}^\alpha(Y_t^i)$ where $\omega_{\mathcal{F}}^\alpha(Y_t^i) = \exp(-\alpha \mathcal{F}(Y_t))$ is a regularized global best. For the Laplace's principle, with this choice, for $\alpha \gg 1$, $\bar{Y}^\alpha \approx \text{argmin}(\mathcal{F}(Y_t^1), \ldots, \mathcal{F}(Y_t^N))$.

[19] Grassi, Pareschi '21

Mean-field limit of SD-PSO

The behavior for $N \gg 1$ is obtained by assuming that the triples (X_t^i, Y_t^i, V_t^i) are independent with the same distribution $f(x,y,v,t)$ (propagation of chaos assumption)

$$f_N(x,y,v,t) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_t^i)\delta(y - Y_t^i)\delta(v - V_t^i) \approx f(x,y,v,t).$$

$$\bar{Y}_t^\alpha \approx \bar{y}^\alpha(\rho) = \frac{\int_{\mathbb{R}^d} y\, \omega_{\mathcal{F}}^\alpha(y)\rho(y,t)dy}{\int_{\mathbb{R}^d} \omega_{\mathcal{F}}^\alpha(y)\rho(y,t)dy}, \quad \rho(y,t) = \int\int_{\mathbb{R}^d\times\mathbb{R}^d} f(x,y,v,t)dxdv.$$

Consequently, $f(x,y,v,t)$ is a weak solution of the Vlasov-Fokker-Plank equation:

$$\begin{aligned}\partial_t f \;+\; v\cdot\nabla_x f + \nabla_y\cdot\left(\nu(x-y)S^\beta(x,y)f\right) = \\ \nabla_v\cdot\Bigg(\frac{1-m}{m}vf + \frac{\lambda_1}{m}(x-y)f + \frac{\lambda_2}{m}(x-\bar{y}^\alpha(\rho))f \\ + \left(\frac{\sigma_2^2}{2m^2}D(x-\bar{y}^\alpha(\rho))^2 + \frac{\sigma_1^2}{2m^2}D(x-y)^2\right)\nabla_v f\Bigg)\end{aligned}$$

Convergence to global minimum

Assumption

(1) *There exists some constant* $L > 0$ *such* $|\mathcal{F}(x) - \mathcal{F}(y)| \leq L(1 + |x| + |y|)|x - y|$ *for all* $x, y \in \mathbb{R}^d$*;*

(2) $\mathcal{F}$ *is bounded from below with* $\underline{\mathcal{F}} := \inf \mathcal{F}$ *and there exists some constant* $C_l > 0$, $M > 0$ *such that*

$$\mathcal{F}(x) - \underline{\mathcal{F}} \geq C_l |x|^2 \text{ for all } |x| \geq M\,.$$

(3) $\mathcal{F} \in C^2(\mathbb{R}^d)$ *with* $\|\nabla^2 \mathcal{F}\|_\infty \leq c_{\mathcal{F}}$ *for some constant* $c_{\mathcal{F}} > 0$.

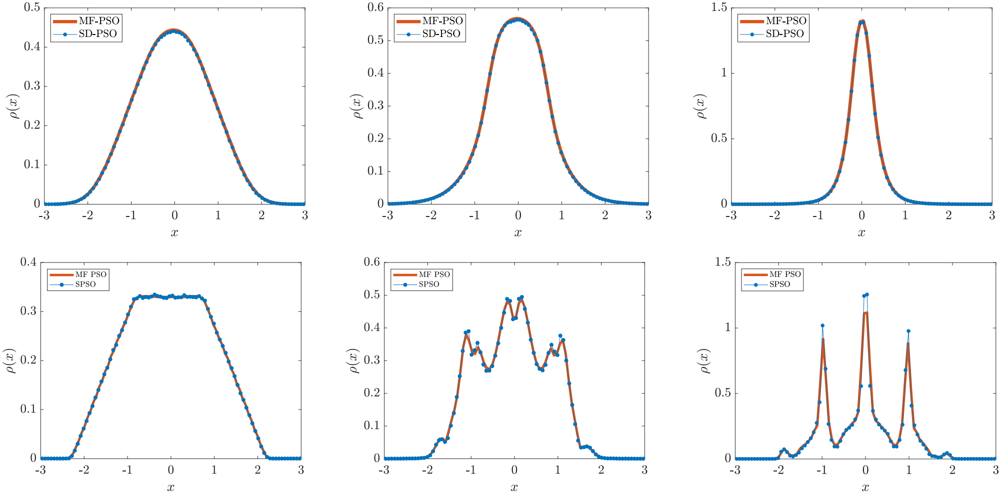
Theorem

Under Assumptions (1)-(2) the SD-PSO system admits a unique solution and the limit $f(x, v, t)$ *of the empirical measures* f^N *exists. Moreover,* f *is the unique weak solution to the Vlasov-Fokker-Planck equation describing the mean-field PSO limit. Under Assumption (3) convergence to the global minimum holds true*[20].

[20] Huang, Qiu and Riedl '23; Grassi, Pareschi, Huang and Qiu '21;

Validation of the mean-field limit

Ackley function in $d = 1$ with global (top) and local (bottom) best:



$t = 0.5$ $t = 3$ $t = 6$

From PSO to CBO: small inertia limit

We rescale the MF-PSO taking $m = \varepsilon \ll 1$ and define the local Maxwellian[21].

$$\mathcal{M}_\varepsilon(x,y,v,t) = \prod_{i=1}^{d} M_\varepsilon(x_i,y_i,v_i,t),\ \ M_\varepsilon(x_i,y_i,v_i,t) = \frac{\varepsilon^{1/2}}{\pi^{1/2}|\Sigma(x_i,y_i,t)|}\exp\left(-\frac{\varepsilon v_j^2}{\Sigma(x_i,y_i,t)^2}\right)$$

where $\Sigma(x_i,y_i,t)^2 = \sigma_1^2(x_i-y_i)^2 + \sigma_2^2(x_i - \bar{y}_i^\alpha(\rho))$, we can write

$$\begin{aligned}\partial_t f + v\cdot\nabla_x f \ + \ \nabla_y\cdot\Big(\nu(x-y)S^\beta(x,y)f\Big)&\\ +\frac{1}{\varepsilon}\nabla_v\cdot(\varepsilon v f + \lambda_1\varepsilon(y-x)f) + \lambda_2(\bar{y}^\alpha(\rho)-x)f&\\ = \frac{1}{2\varepsilon^2}\sum_{j=1}^{d}\Sigma(x_i,y_i,t)^2\frac{\partial}{\partial v_j}\left(f\frac{\partial}{\partial v_j}\log\left(\frac{f}{M_\varepsilon(x_i,y_i,v_i,t)}\right)\right)&,\end{aligned}$$

The r.h.s. is $O\left(\varepsilon^{-2}\right)$, and for $\varepsilon \ll 1$ we have $f(x,y,v,t) \approx \rho(x,y,t)\mathcal{M}_\varepsilon(x,y,v,t)$.

[21] Choi, Ha, Noh '13; Cipriani, Huang, Qiu '22; Grassi, Huang, Pareschi, Qiu '21

From PSO to CBO: small inertia limit

Thanks to this, we obtain a mean-field PSO with momentum[22]:

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) + \nabla_y \cdot \Big(\nu(x-y) S^\beta(x,y)\rho\Big) = 0$$

$$\frac{\partial (\rho u)_i}{\partial t} + \frac{1}{2\varepsilon}\frac{\partial}{\partial x_i}\left(\rho \cdot \Sigma(x_i,y_i,t)^2\right) = -\frac{1-\varepsilon}{\varepsilon}(\rho u)_i + \frac{1}{\varepsilon}\left(\lambda_1(y_i - x_i) + \lambda_2(\bar{y}_i^\alpha(\rho) - x_i)\right)\rho.$$

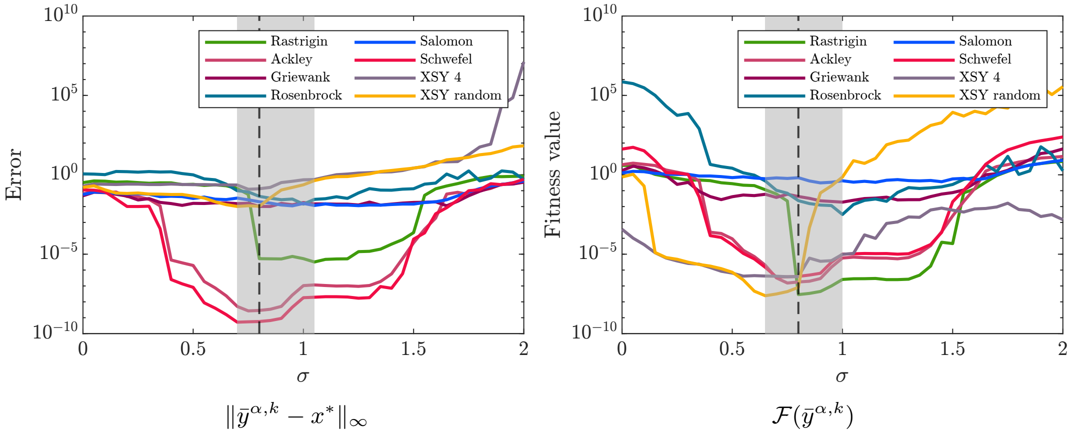
For $\varepsilon \to 0$ we get a mean-field CBO system with memory effects[23]:

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot \left(\lambda_1(y-x) + \lambda_2(\bar{y}^\alpha(\rho) - x)\right)\rho + \nabla_y \cdot \Big(\nu(x-y)S^\beta(x,y)\rho\Big)$$
$$= \frac{1}{2}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}\left(\rho\left(\sigma_1^2(x_j-y_j)^2 + \sigma_2^2(x_j - \bar{y}_j^\alpha(\rho))^2\right)\right).$$

[22] Chen, Jin, Lyu ’22

[23] Borghi, Grassi, Pareschi ’23

Hyper-parameters selection for benchmark functions $d = 20$



$\|\bar{y}^{\alpha,k} - x^*\|_\infty$

$\mathcal{F}(\bar{y}^{\alpha,k})$

Optimization on benchmark functions global best only and $m = \varepsilon = 0$. Error and fitness values for different σ. Here $N = 200$, $\Delta t = 1$, $\lambda = 0.01$, and α is adaptive following a SA strategy with $\alpha_0 = 10$.

		CBO ($\sigma = \sqrt{2}/2$)			CBO-ME ($\sigma = 0.8$)			Matlab - particleswarm			Matlab - particleswarm ($c_1 = 0$)		
		$N = 50$	$N = 100$	$N = 200$	$N = 50$	$N = 100$	$N = 200$	$N = 50$	$N = 100$	$N = 200$	$N = 50$	$N = 100$	$N = 200$
Ackley	Rate	99.8%	100.0%	100.0%	100.0%	100.0%	100.0%	17.1%	41.2%	54.3%	4.2%	16.2%	40.1%
	Error	4.22e-06	2.14e-06	3.55e-06	2.42e-06	1.89e-06	1.56e-06	6.17e-09	8.86e-11	2.01e-12	2.23e-08	1.80e-10	8.65e-13
	$\mathcal{F}$	1.18e-04	5.81e-05	7.30e-05	1.54e-04	4.96e-05	4.99e-05	6.24e-09	7.65e-11	1.94e-12	2.06e-08	1.70e-10	8.01e-13
	Iterations	912.3	718.1	623.2	977.2	703.3	622.2	501.8	424.2	341.3	502.3	421.2	321.2
Rastrigin	Rate	12.1%	34.3%	62.7%	23.2%	69.7%	89.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Error	1.28e-04	1.83e-04	2.34e-04	9.73e-05	1.27e-04	1.76e-04	-	-	-	-	-	-
	$\mathcal{F}$	4.51e-06	9.03e-06	1.46e-05	2.54e-06	4.31e-06	8.28e-06	-	-	-	-	-	-
	Iterations	1083.0	933.7	819.8	1007.6	922.5	769.9	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0
Rosenbrock	Rate	65.3%	86.7%	100.0%	70.1%	94.2%	100.0%	9.3%	22.6%	36.6%	46.7%	60.7%	76.7%
	Error	1.84e-02	2.43e-02	1.42e-02	3.60e-02	4.01e-02	1.82e-02	6.19e-04	2.56e-04	1.67e-04	4.44e-02	4.45e-02	4.46e-02
	$\mathcal{F}$	6.13e-03	7.57e-03	2.40e-03	1.26e-02	1.42e-02	2.65e-03	3.80e-02	3.76e-02	2.56e-02	2.56e-03	8.95e-04	3.71e-04
	Iterations	5773.2	5423.2	5233.1	5933.2	4956.2	4155.2	4822.2	3823.2	3026.3	5924.2	3834.1	2933.3
Schwefel 2.20	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	5.79e-06	8.23e-07	2.44e-07	8.42e-06	1.03e-06	2.76e-07	8.34e-10	1.97e-12	4.58e-14	1.68e-07	3.41e-10	8.03e-14
	$\mathcal{F}$	1.04e-03	2.15e-04	8.36e-05	1.50e-03	3.12e-04	9.37e-05	1.94e-09	6.36e-12	1.52e-13	2.44e-07	6.48e-10	2.46e-13
	Iterations	822.2	682.2	622.1	655.2	544.2	455.2	491.2	434.2	399.1	578.2	467.2	423.2
Salomon	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Error	3.12e-02	2.14e-02	1.87e-02	5.28e-02	4.49e-02	3.91e-02	-	-	-	-	-	-
	$\mathcal{F}$	3.14e-01	2.15e-01	1.88e-01	2.44e-01	1.86e-01	1.91e-01	-	-	-	-	-	-
	Iterations	10000.0	10000.0	10000.0	8872.2	9021.2	5356.5	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0
XSY random	Rate	52.3%	81.7%	92.6%	100.0%	100.0%	100.0%	3.2%	17.1%	31.2%	100.0%	100.0%	100.0%
	Error	2.64e-02	1.62e-02	9.80e-03	3.06e-02	1.86e-02	1.15e-02	2.25e-01	9.56e-02	8.42e-02	6.23e-02	5.12e-02	2.34e-02
	$\mathcal{F}$	6.95e-08	3.54e-08	2.13e-08	2.21e-06	4.85e-08	3.17e-08	3.35e-04	2.28e-04	1.34e-04	8.22e-04	4.11e-04	3.45e-04
	Iterations	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0
XSY 4	Rate	27.2%	89.3%	100.0%	25.2%	91.2%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Error	8.10e-01	7.12e-01	7.89e-01	8.01e-01	7.55e-01	6.17e-01	-	-	-	-	-	-
	$\mathcal{F}$	4.79e-07	3.78e-07	3.46e-07	1.58e-06	8.56e-07	5.43e-07	-	-	-	-	-	-
	Iterations	10000.0	10000.0	10000.0	9733.2	9531.1	8733.2	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0

Algorithmic improvements

- Evaluate $\bar{Y}_n^{\alpha,\mathcal{F}}$ on batches J_b of $N_b < N$ particles[24]

$$\bar{Y}_n^{\alpha,\mathcal{F}} \approx \frac{\sum_{i\in J_b} \omega_\alpha^{\mathcal{F}}(Y_n^i) Y_n^i}{\sum_{i\in J_b} \omega_\alpha^{\mathcal{F}}(Y_n^i)}.$$

- Discard particles in time accordingly to the variance Σ_n of the solution

$$N_{n+1} = \min\left\{ \left[\!\left[N_n \left(1 + \mu \left(\frac{\Sigma_{n+1} - \Sigma_n}{\Sigma_n} \right)\right) \right]\!\right], N_{\min} \right\}$$

- Decrease σ in time as in simulated annealing
- Increase α in time to achieve higher precision

[24] S. Jin, L. Li, J-G. Liu, JCP 2020 and SINUM 2021;

Speed up by particle reduction

		$\mu = 0$	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.5$
Rastrigin	Rate	100.0%	100.0%	100.0%	100.0%
	Error	9.22e-05	7.76e-05	3.54e-05	1.34e-05
	$\mathcal{F}$	2.90e-06	2.99e-06	1.45e-06	1.12e-06
	w_{iter}	1150.3	720.6	250.5	106.3
	CTS	-	39.2%	78.9%	92.3%
		$\mu = 0$	$\mu = 0.01$	$\mu = 0.02$	$\mu = 0.05$
Rosenbrock	Rate	100.0%	100.0%	99.4%	99.0%
	Error	2.12e-02	2.21e-02	1.78e-02	1.45e-02
	$\mathcal{F}$	4.22e-03	5.67e-03	4.12e-03	4.45e-03
	w_{iter}	3189.3	840.3	350.3	102.3
	CTS	-	75.3%	90.2%	92.4%

Table: Algorithm with particle reduction for different values of μ. The system is initialized with $N_0 = 200$ particles. Performance metric considered: success rate, error ($\|\bar{y}^{\alpha,k} - x^*\|_\infty$), fitness value $\mathcal{F}(\bar{y}^{\alpha,k})$, weighted iteration, and Computational Time Saved (CTS).

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 (Edinburgh), 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)

Codes:

`https://github.com/borghig/CBOswarm`