Long-time convergence of a consensus-based optimization method

Victor Priser[1], Pascal Bianchi[1] and Radu-Alexandru Dragomir[1]

[1]Télécom Paris —LTCI, France

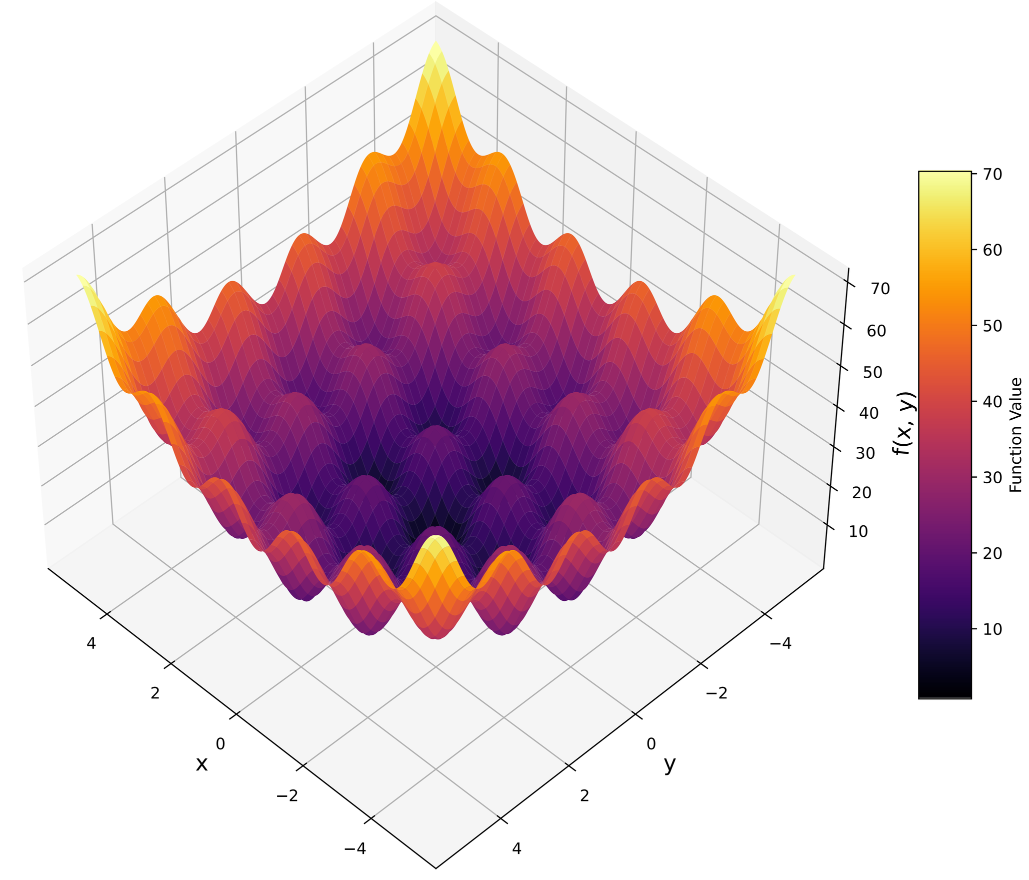
This research was funded by the chair DSAIDIS (Data Science & Artificial Intelligence for Digitalized Industry & Services)

Objective

We aim to find the *minimizer* of a function f, that is, to solve the problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

where f is a *non-convex function* with *unknown gradient*.



Existing solution

Algorithm: Generate particles to reach the minimizer of f.

- Simulated Annealing:
  - Generate new particles with noise of decreasing variance over time.
- Genetic and Evolutionary Algorithms:
  - Combine two particles to generate offspring with lower f value.
- Bayesian Optimization:
  - Build a random surrogate function $\tilde{f}$ fitted to f and sample the next particle by minimizing an acquisition criterion on $\tilde{f}$.
- Particle Swarm Optimization (PSO):
  - Make particles interact to cooperatively explore and find the minimizer of f.

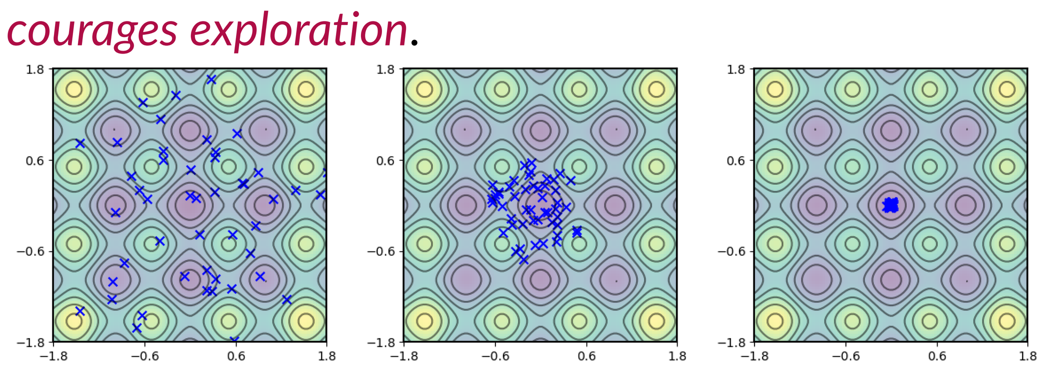
Consensus Based Optimization

- The Consensus-Based Optimization (CBO) algorithm is a *simplified version of Particle Swarm Optimization* (PSO).
- The algorithm is *initialized* with n particles $(X_0^1, \ldots, X_0^n)$.
- At each step $k+1$, the n particles $(X_{k+1}^1, \ldots, X_{k+1}^n)$ are *attracted* toward the particle with the lowest function value: $\arg\min_{x \in \{X_k^1, \ldots, X_k^n\}} f(x)$.

The *dynamics* of CBO are given , for $i \in \{1, \ldots, n\}$:

$$X_{k+1}^i = X_k^i + \eta\Big(\arg\min_{x \in \{X_k^1, \ldots, X_k^n\}} f(x) - X_k^i\Big) + \text{Noise},$$

where $\eta > 0$ is a *learning rate* and the *noise term encourages exploration*.



Laplace Principle: $\frac{\sum_{i=1}^n X_k^i e^{-\alpha f(X_k^i)}}{\sum_{i=1}^n e^{-\alpha f(X_k^i)}} \underset{\alpha \to \infty}{\longrightarrow} \arg\min_{x \in \{X_k^1, \ldots, X_k^n\}} f(x)$

Vanilla version of CBO

For every $i \in \{1, \ldots, n\}$

$$X_{k+1}^i = \eta\left(\frac{\sum_{i=1}^n X_k^i e^{-\alpha f(X_k^i)}}{\sum_{i=1}^n e^{-\alpha f(X_k^i)}} - X_k^i\right) + \text{Noise}$$

Our Algorithm

Empirical measure of CBO at time k

$$\mu_k^n := \frac{1}{n}\sum_{i=1}^n \delta_{X_k^i}$$

The *consensus* term $C_\alpha(\mu) := \frac{\int x e^{-\alpha f(x)}\,\mathrm{d}\mu(x)}{\int e^{-\alpha f(x)}\,\mathrm{d}\mu(x)}$.

Our algorithm

For every $i \in \{1, \ldots, n\}$

$$X_{k+1}^i = X_k^i + \eta(\mathrm{clip}_R(C_\alpha(\mu_k^n)) - X_k^i) + \sqrt{2\eta\frac{\gamma}{\alpha}}\xi_k^i$$

$(\xi_k^i)_{k,i} \sim_{i.i.d.} \mathcal{N}(0, I_d)$ and $\mathrm{clip}_R(x) := \frac{x}{\|x\|}\min(\|x\|, R)$.

- γ is a parameter of the algorithm that is assumed to be *sufficiently large*.
- The type of noise introduced here is *novel*.
- *Others* consider a noise term of the form $\sqrt{\eta}\,\|C_\alpha(\mu_k^n) - X_k^i\|\,\xi_k^i$ ([CCTT18, FKR24]).

Main result

W_2: 2-Wasserstein distance.

Theorem

For every $i \in \{1, \ldots, n\}$

$$\inf_{\|x - \arg\min f\| \le \frac{C}{\sqrt{\alpha}}} W_2(\mathcal{L}(X_k^i), \mathcal{N}(x, \frac{\gamma}{\alpha}))$$
$$\le C\left(\frac{1}{\sqrt{n}} + \sqrt{\eta} + (1+c)^{-k}\right).$$

The particles *converge*, with high probability, to a *neighborhood of radius $\mathcal{O}\left(\frac{1}{\sqrt{\alpha}}\right)$ around the minimizer* $\arg\min f$, for large n and k, and small η.

Literature review

Convergence of CBO:

	Low value of k	Large value of k
$n = \infty$	[CCTT18]	[CCTT18]
$n < \infty$	[FKR24] with $k \le \log(n)$	Our paper

Propagation of chaos

Our algorithm is the Euler scheme applied to the following Stochastic Differential Equation (SDE):

$$\mathrm{d}X_t^i = \left(\mathrm{clip}_R\left(C_\alpha(\mu_t^n)\right) - X_t^i\right)\mathrm{d}t + \sqrt{\frac{2\gamma}{\alpha}}\,\mathrm{d}B_t^i,$$

where B_t^i are n independent Brownian motions.

Propagation of chaos

When $n \gg 1$, particles $X_k^1, \ldots, X_k^n$ become i.i.d. and by the law of large number:

$$\mu_k^n \simeq \mathcal{L}(X_t^i)$$

The particles of the CBO algorithm are approximated by the *McKean–Vlasov stochastic differential equation (SDE)*

$$\mathrm{d}X_t = (\mathrm{clip}_R\left(C_\alpha(\mathcal{L}(X_t))\right) - X_t)\,\mathrm{d}t + \sqrt{\frac{2\gamma}{\alpha}}\,\mathrm{d}B_t,$$

In other words, assuming $Z \sim \mathcal{N}(0, I_d)$, we have

$$X_t = x_t + X_0 e^{-t} + \sqrt{1 - e^{-2t}}\sqrt{\frac{\gamma}{\alpha}}\,Z,$$

Mean field Equation

$$\dot{x}_t = \mathrm{clip}_R\left(C_\alpha(\mathcal{L}(X_t)) - x_t\right) \quad \text{and} \quad x_0 = 0$$

Convergence of the mean field equation

For large t,

$$\mathcal{L}(X_t) \simeq \mathcal{N}(x_t, \frac{\gamma}{\alpha}I_d)$$

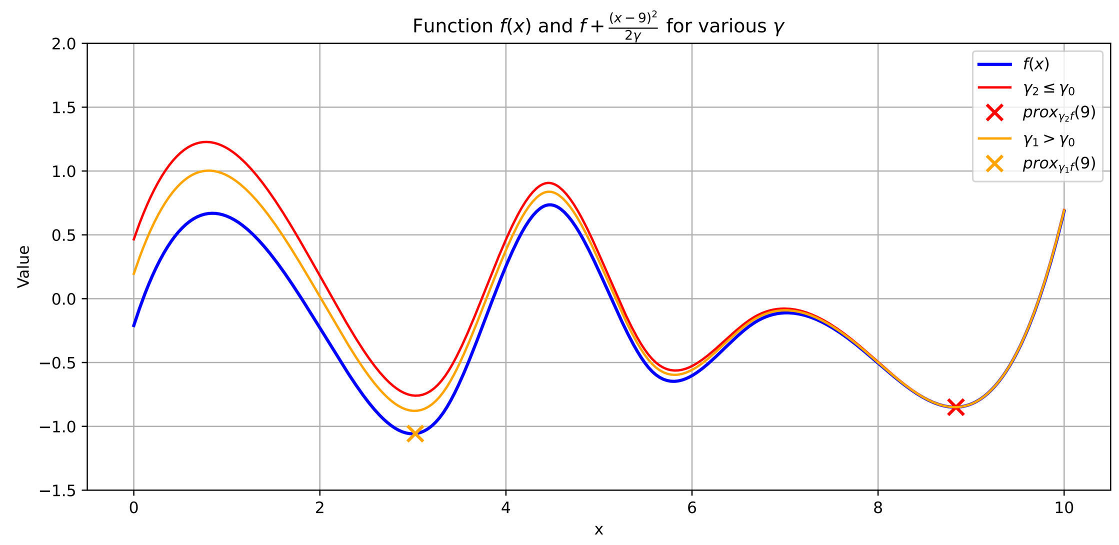
Laplace principle

$$C_\alpha(\rho_t) \simeq \frac{\int x \exp\left(-\alpha\left(f(x) + \frac{\|x - x_t\|^2}{2\gamma}\right)\right)\mathrm{d}x}{\int \exp\left(-\alpha\left(f(x) + \frac{\|x - x_t\|^2}{2\gamma}\right)\right)\mathrm{d}x}$$
$$\underset{\alpha \to \infty}{\longrightarrow} \underbrace{\arg\min_{x \in \mathbb{R}^d}\left(f(x) + \frac{\|x - x_t\|^2}{2\gamma}\right)}_{\mathrm{prox}_{\gamma f}(x_t)}$$

There exists a *perturbation term* ε_t such that $\mathbb{E}[\|\varepsilon_t\|] \ll 1$, and the dynamics satisfy

$$\dot{x}_t = \mathrm{prox}_{\gamma f}(x_t) - x_t + \varepsilon_t.$$

Problem: The operator $\mathrm{prox}_{\gamma f}$ may not be well-defined when f is *non-convex*.



There exists:

- δ such that f is convex on $B(\arg\min f, \delta)$.
- $\gamma_0(\delta, K)$ s.t. for $\gamma > \gamma_0(\delta, K)$

$$\forall x \in B(\arg\min f, K), \quad \mathrm{prox}_{\gamma f}(x) \in B(x, \delta)$$

$\bar{f}$: *convex function* s.t. $f = \bar{f}$ on $B(\arg\min f, \delta)$ then for $\gamma > \gamma_0(\delta, K)$

$$\forall x \in B(\arg\min f, K), \quad \mathrm{prox}_{\gamma f}(x) = \mathrm{prox}_{\gamma \bar{f}}(x).$$

Stability

For every $t > 0$,

$$\|x_t\| \le R.$$

Then, for $\gamma > \gamma_0(\delta, R)$,

$$\dot{x}_t = \mathrm{prox}_{\gamma \bar{f}}(x_t) - x_t + \varepsilon_t.$$

If $x_t \to x_\infty$ as $t \to \infty$, and in the noiseless case $(\varepsilon_t = 0)$, it follows that

$$\mathrm{prox}_{\gamma \bar{f}}(x_\infty) = x_\infty \implies \nabla \bar{f}(x_\infty) = 0.$$

Since *$\bar{f}$ is convex*, we conclude that

$$x_\infty = \arg\min \bar{f} = \arg\min f.$$

Bibliography

References

[CCTT18] J. A Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. An analytical framework for consensus-based global optimization method. *Mathematical Models and Methods in Applied Sciences*, 28(06):1037–1066, 2018.

[FKR24] M. Fornasier, T. Klock, and K. Riedl. Consensus-based optimization methods converge globally. *SIAM Journal on Optimization*, 34(3):2973–3004, September 2024.