MATH-IMS Applied Mathematics Colloquium Series

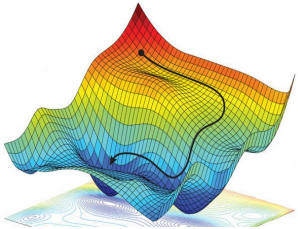
On the mean-field limit of stochastic particle optimization methods

Lorenzo Pareschi

Department of Mathematics and Computer Science, University of Ferrara, Italy

Chinese University of Hong Kong - September 24, 2021

Machine learning and nonconvex optimization



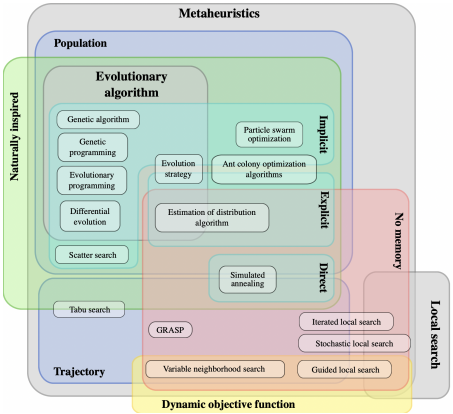
Machine learning mainly concerns nonlinear parametric algorithms; parameters are optimized towards several tasks (feature selection, dimensionality reduction, clustering, classification, regression, ...);

- The nonlinearity and nonconvex data misfits or penalizations/regularizations make training phase a nonconvex optimization;
- Large amount of parameters make the optimization high dimensional and quite hard;
- First order methods, e.g. (stochastic) gradient descent, are preferred both for speed and scalability and for being generally considered capable of escaping the trap of critical points;
- Drawbacks and limitations: objective function not differentiable, explosion or vanishing of gradients (Feedforward neural networks), lack of guarantees of global convergence etc.

Metaheuristics

Metaheuristics orchestrate an interaction between local improvement procedures and global/high level strategies, and combine random and deterministic decisions to escape from local optima and perform a robust search of the solution space.

- Simplex Heuristics (1965)
- Evolutionary Programming (1968)
- Metropolis-Hastings (1970)
- Simulated Annealing (1983)
- Genetic Algorithms (1992)
- Ant Colony Optimization (2005)
- Particle Swarm Optimization (2007)
- ...



⇒ Despite tremendous empirical success lack of a robust mathematical theory, i.e. mathematical formulation and convergence to global minimizers.

Stochastic particle optimization methods

We consider the optimization problem

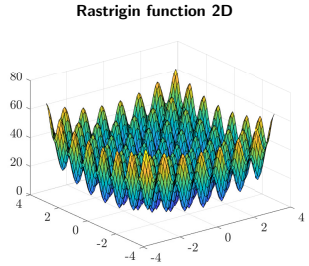
$$x^* \in \text{argmin}_{x \in \mathbb{R}^d} \mathcal{F}(x)\,,$$

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

The notion of stochastic optimization by particles pertains to different methods:

- Stochastic particle optimization sampling (SPOS) (Gradient-based, SG-MCMC,....)
- Particle swarm optimization (PSO)
- Consensus based optimization (CBO)

In the sequel we will focus on stochastic particle optimization methods which are gradient-free and based on metaheuristics.

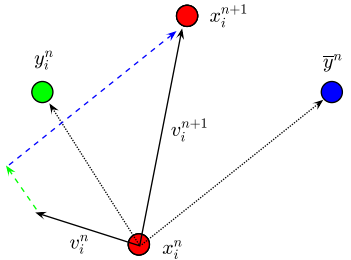


Part I:
Mean field Particle Swarm Optimization

The PSO method

Particle swarm optimization (PSO) is a metaheuristic optimizer that 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 algorithm[1]:

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



Local best and global best influence

- $\bar{y}^n$ is the global best position given by $\operatorname{argmin}(\mathcal{F}(x_1^n), \dots, \mathcal{F}(x_N^n), \mathcal{F}(\bar{y}^{n-1}))$;
- 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.

[1] J. Kennedy, R. Eberhart. *Proc. ICEC'95, 1995*; J. Kennedy. *Proc. ICEC'97, 1997*; J. Kennedy. *Springer, 2010*

The CBO method

A different approach to minimization is through consensus-based optimization (CBO) methods based on the evolution of N particles with positions $X_t^i \in \mathbb{R}^d$ according to the first order SDEs[2]:

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

- B_t^i denote independent Brownian motions;
- $D(X_t) = |X_t| I_d$ or $D(X_t) = \text{diag}\{(X_t)_1, (X_t)_2, \dots, (X_t)_d\}$;
- $\bar{X}_t^\alpha = \frac{1}{\sum_i \omega_{\mathcal{F}}^\alpha(X_t^i)} \sum_i X_t^i \omega_{\mathcal{F}}^\alpha(X_t^i)$ where $\omega_{\mathcal{F}}^\alpha(X_t^i) = \exp(-\alpha \mathcal{F}(X_t))$ and with this choice[3], for $\alpha \gg 1$, by Laplace's principle we have $\bar{X}^\alpha \approx \text{argmin}(\mathcal{F}(X_t^1), \dots, \mathcal{F}(X_t^N))$;
- $\lambda > 0$ and $\sigma \geq 0$ are drift parameter and noise parameter respectively;

[2] R. Pinnau, C. Totzeck, O. Tse and S. Martin. *M3AS, 2017*; J.A. Carrillo, Y.-P. Choi, C. Totzeck and O. Tse. *M3AS, 2018*; J.A. Carrillo, S. Jin, L. Li and Y. Zhu. *ESAIM.COCV, 2020*; C. Totzeck and M.-T. Wolfram. *MBE, 2020.*

[3] Here $\exp(-\alpha \mathcal{F}(x))$ is the Gibbs distribution corresponding to $\mathcal{F}(x)$. The larger the value of α, the larger the weight of the normalized Gibbs measure for the particle is on the minimum value of the cost function $\mathcal{F}(x)$.

CBO method and mean-field limit

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

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

and the CBO dynamics is approximated by solutions of the non-linear Fokker–Planck equation

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

Under suitable assumptions on λ, σ and α, independently of the dimension d, the variance $V(t) \to 0$ exponentially fast and the expectation $\mathbb{E}[\bar{X}_t^\alpha] \to \tilde{x}$. When α is sufficiently large and $\mathcal{F}$ has a unique global minimizer, together with some reasonable assumptions on $\mathcal{F}$, one can show that $\tilde{x} \approx x^*$ the global minimum[a].

[a] J.A. Carrillo, S. Jin, L. Li and Y. Zhu. *ESAIM.COCV, 2020*

Stochastic Differential PSO method (SD-PSO)

The obtain a differential formulation of the PSO method, one of the difficulties consists in the presence of particle memory. For this purpose we observe that the local best y_i^{n+1} can be written as a time update process

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

and then the PSO method can be generalized to the time discrete formalism

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

where by choosing

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

and $\Delta t = 1$, $\nu = 1/2$ we recover 'exactly' the classical PSO algorithm.

Stochastic Differential PSO method (SD-PSO)

The system can be understood as a discretization of the following system of second order SDEs[4]:

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

- $B_t^{k,i}, \quad k = 1, 2$ 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 $\alpha \gg 1$, $\bar{Y}^\alpha \approx \text{argmin}(\mathcal{F}(Y_t^1), \ldots, \mathcal{F}(Y_t^N))$.

[4] S.Grassi and L. Pareschi. *M3AS, 2021.*

Mean-field PSO limit (MF-PSO)

The derivation of the mean field description of the SD-PSO system is obtained by assuming for $N \gg 1$ 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)

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

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

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

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

Mean-field PSO limit (MF-PSO)

Assumptions

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

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

$$\mathcal{F}(x) - \underline{\mathcal{F}} \leq C_u(1 + |x|^2) \text{ for all } x \in \mathbb{R}^d;$$

(3) $\mathcal{F}$ *has quadratic growth at infinity. Namely, there exist constants* C_l, $M > 0$ *such that*

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

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

Under Assumptions (1)-(3) $\{(X_t^{i,N}, Y_t^{i,N}, V_t^{i,N})_{t\in[0,T]}\}_{i=1}^N$ is the unique solution to the SD-PSO system. Then, the limit f of the sequence of the empirical measure f^N exists. Moreover, f is the unique weak solution to MF-PSO equation[a]. Additionally, under Assumption (4) and in absence of memory effects, convergence to the global minimum has been proved[b].

[a] H.Huang. *Applied Mathematics Letters, 2021;*

[b] H.Huang and J. Qiu preprint '21; S.Grassi, L. Pareschi, H.Huang and J.Qiu. *IMS Lecture Notes, 2021.*;

From PSO to CBO: small inertia limit

We consider the MF-PSO system with $m = \varepsilon \ll 1$. Introducing the local Maxwellian with unitary mass and zero momentum

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

where $\Sigma(x_i,y_i,t)^2 = \sigma_1^2(x_i-y_i)^2 + \sigma_2^2(x_i - Y_i^\alpha(\rho))$, we can write

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

The r.h.s. is of order $\frac{1}{\varepsilon^2}$, and for small values of $\varepsilon \ll 1$ we have

$$f(x,y,v,t) \approx \rho(x,y,t)\mathcal{M}_\varepsilon(x,y,v,t).$$

From PSO to CBO: small inertia limit

Now considering $\rho u = \int_{\mathbb{R}^d} f(x,y,v,t) v dv$ and taking the first two moments of f, we get

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) + \nabla_y \cdot \left(\nu(x-y) S^\beta(x,y)\rho\right) = 0$$
$$\frac{\partial \rho u}{\partial t} + \int_{\mathbb{R}^d} v(v \cdot \nabla_x f) = -\frac{1-\varepsilon}{\varepsilon}\rho u + \frac{1}{\varepsilon}\left(\lambda_1(y-x) + \lambda_2(Y^\alpha(\rho) - x)\right)\rho.$$

and applying the equilibrium assumption $f = \rho \mathcal{M}_\varepsilon$, we have

$$\begin{aligned}\int_{\mathbb{R}^d} v(v \cdot \nabla_x(\rho(x,y,t)\mathcal{M}_\varepsilon(x,y,v,t))) &= \sum_{j=1}^d \frac{\partial}{\partial x_j}\left(\rho(x,y,t)\int_{\mathbb{R}^d} v_j(v_j \mathcal{M}_\varepsilon(x,y,v,t))dv\right) \\ &= \frac{\partial}{\partial x_j}\left(\rho(x,y,t)\int_{\mathbb{R}} v_j^2 M_\varepsilon(x_i,y_i,v_i,t)dv_j\right) \\ &= \frac{1}{2\varepsilon}\frac{\partial}{\partial x_j}\left(\rho(x,y,t)\Sigma(x_i,y_i,t)^2\right)\end{aligned}$$

From PSO to CBO: small inertia limit

thanks to this, we obtain a mean-field PSO with momentum[5]:

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) + \nabla_y \cdot \left(\nu(x-y) S^\beta(x,y)\rho\right) = 0$$
$$\frac{\partial (\rho u)_i}{\partial t} + \frac{1}{2\varepsilon}\frac{\partial}{\partial x_i}\left(\rho \cdot \Sigma(x_i, y_i, t)^2\right) = -\frac{1-\varepsilon}{\varepsilon}(\rho u)_i + \frac{1}{\varepsilon}\left(\lambda_1(y_i - x_i) + \lambda_2(Y_i^\alpha(\rho) - x_i)\right)\rho.$$

For $\varepsilon \to 0$ we get a mean-field CBO system with memory effects[6]:

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

Rigorous results on the small inertia limit have been proved recently[7].

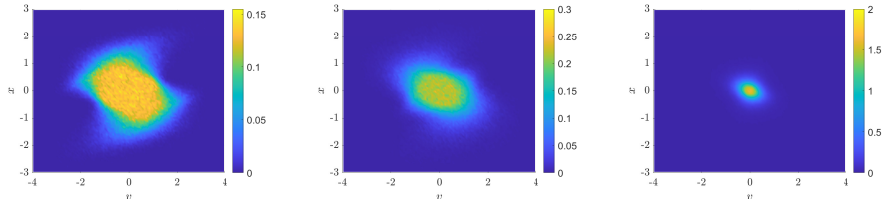
[5] See also the related work J.Chen, S.Jin and L. Lyu. *arXiv:2012.04827, 2021*

[6] S.Grassi and L. Pareschi. *M3AS, 2021.*

[7] C. Cipriani, H. Huang and J. Qiu. *arXiv:2104.06939, 2021.*; S.Grassi, H.Huang, L.Pareschi and J.Qiu. *IMS Lecture Notes, 2021.*

Validation of the mean-field limit

Ackley function in $d = 1$ with global best dynamics:

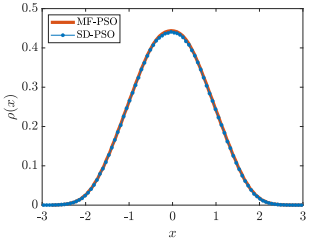


$t = 0.5$ $t = 1$ $t = 3$

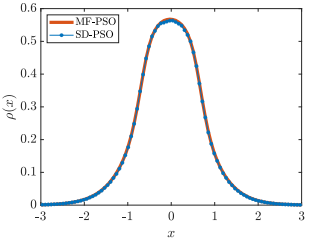
First row: solution of the SDE's system. Second row: solution of the mean-field limit solved with finite difference schemes. We used 5×10^5 particles, a grid size of $Nx = 90$, $Nv = 120$ and initialized f_0 as a uniform distribution on a restricted domain centered in the origin.

Validation of the mean-field limit

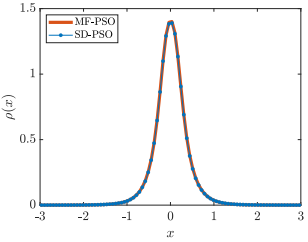
Ackley function in $d = 1$ with global best dynamics:



$t = 0.5$



$t = 1$

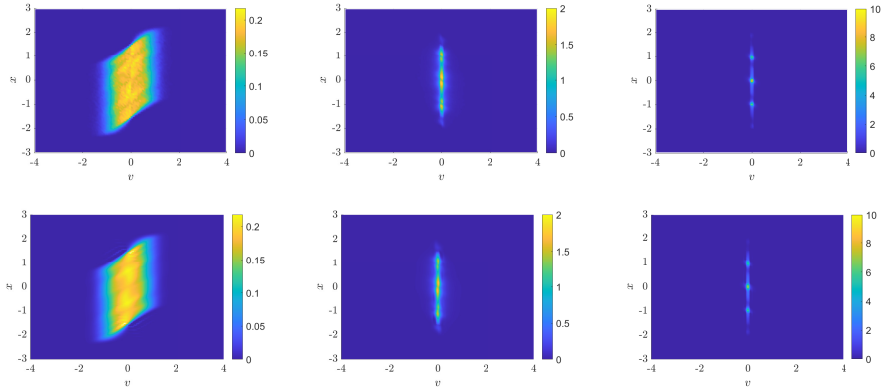


$t = 3$

Comparison between the marginal of the particle solution and $\rho(x,t) = \int_{\mathbb{R}} f(x,v,t)dv$.

Validation of the mean-field limit

Ackley function in $d = 1$ with local best dynamics:

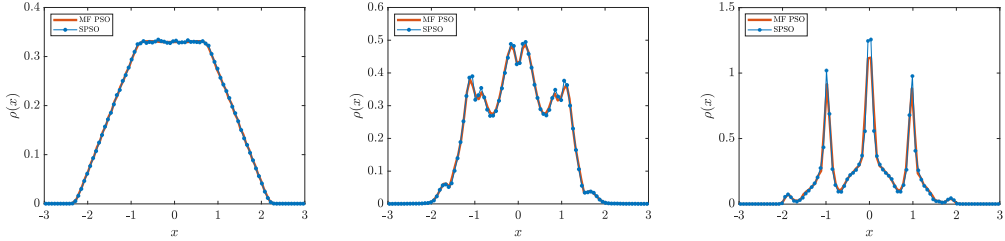


$t = 0.5$ $t = 3$ $t = 6$

First row: solution of the SDE's system. Second row: solution of the mean-field limit solved with finite difference schemes. We used 5×10^5 particles, a grid size of $Nx = 90$, $Ny = 90$, $Nv = 120$ and initialized f_0 as a uniform distribution on a restricted domain centered in the origin.

Validation of the mean-field limit

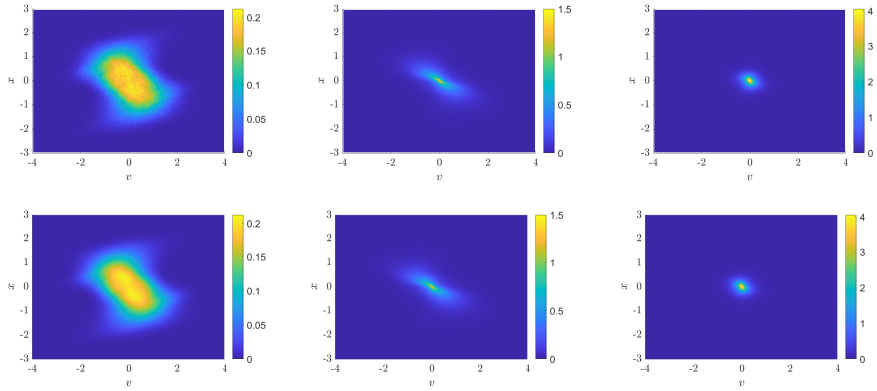
Ackley function in $d = 1$ with local best dynamics:



Comparison between the marginal of the particle solution and $\rho(x,t) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y,v,t)dvdy$.

Validation of the mean-field limit

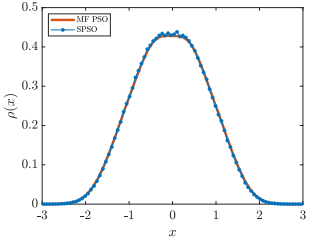
Ackley function in $d = 1$ with global and local best dynamics:



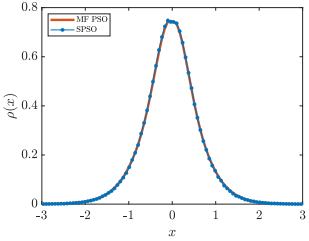
First row: solution of the SDE's system obtained. Second row: solution of the mean-field limit solved with finite difference schemes. We used 5×10^5 particles, a grid size of $Nx = 90$, $Ny = 90$, $Nv = 120$ and initialized f_0 as a uniform distribution on a restricted domain centered in the origin.

Validation of the mean-field limit - PSO

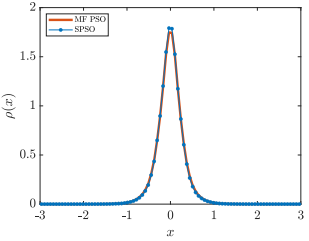
Ackley function in $d = 1$ with global and local best dynamics:



$t = 0.5$



$t = 1$



$t = 3$

Comparison between the marginal of the particle solution and $\rho(x,t) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y,v,t)dvdy$.

Results for Rastrigin $d = 20$

Rastrigin			Case without memory				Case with memory		
m		σ	$N = 50$	$N = 100$	$N = 200$	σ_2	$N = 50$	$N = 100$	$N = 200$
0.00	Rate	9.0	100.0%	100.0%	100.0%	11.0	100.0%	100.0%	100.0%
	Error		1.19e-04	1.11e-04	9.68e-05		6.83e-04	4.70e-04	4.69e-04
	n_{iter}		10000.0	10000.0	9912.4		10000.0	9878.2	3290.2
0.01	Rate	7.0	100.0%	100.0%	100.0%	9.0	100.0%	100.0%	100.0%
	Error		9.74e-05	2.01e-05	1.62e-05		8.60e-04	8.56e-04	8.81e-04
	n_{iter}		10000.0	6899.2	2060.1		9939.5	7012.2	5422.1
0.05	Rate	3.5	37.0%	74.0%	94.0%	4.5	100.0%	100.0%	100.0%
	Error		4.27e-04	1.26e-04	1.14e-04		1.15e-03	6.67e-04	6.54e-04
	n_{iter}		8233.2	7814.0	7326.6		9978.0	7657.6	5639.7
0.10	Rate	2.0	1.0%	5.5%	29.5%	3.0	80.8%	96.8%	100.0%
	Error		2.00e-04	1.28e-04	1.11e-04		2.94e-03	8.96e-04	8.24e-04
	n_{iter}		6155.4	6221.9	6214.3		9661.5	8676.5	7331.8

SD-PSO with and without memory for $\lambda_1 = \sigma_1 = 0$, $\lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$ and $\alpha = 5 \times 10^4$.

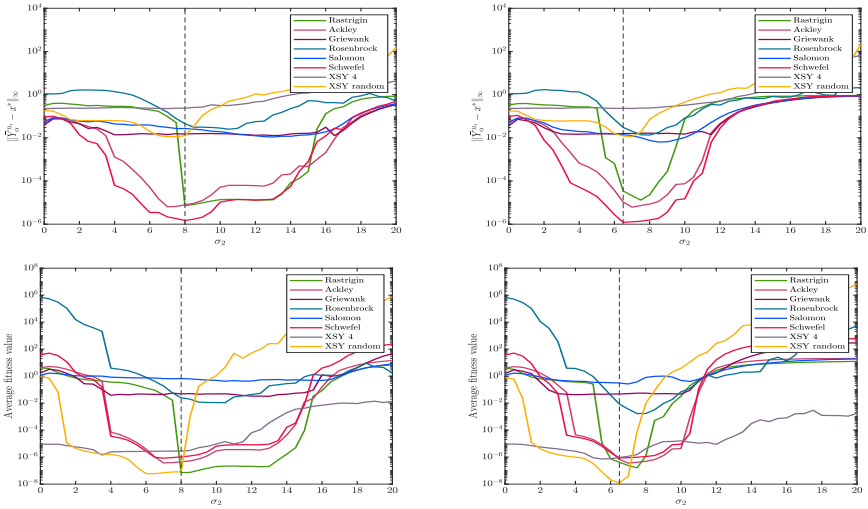
Results for some benchmark functions $d = 20$

		Case $\xi = 0$, $\sigma_2 = 8.0$			Case $\xi = 0.25$, $\sigma_2 = 6.5$		
		$N = 50$	$N = 100$	$N = 200$	$N = 50$	$N = 100$	$N = 200$
Griewank	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	2.28e-02	2.24e-02	2.19e-02	2.27e-02	2.16e-02	2.24e-02
	$\mathcal{F}_{avg}$	5.57e-02	5.21e-02	4.26e-02	5.25e-02	4.93e-02	2.28e-02
	n_{iter}	1010.8	861.6	761.7	1006.3	734.7	626.6
Rastrigin	Rate	34.0%	70.7%	95.0%	9.0%	26.4%	42.0%
	Error	1.78e-05	1.89e-05	2.05e-05	3.01e-05	3.12e-05	3.03e-05
	$\mathcal{F}_{avg}$	9.32e-08	9.68e-08	9.95e-08	2.41e-07	2.58e-07	2.44e-07
	n_{iter}	1308.5	1122.9	970.5	1631.0	1483.0	1334.8
Rosenbrock	Rate	49.3%	84.7%	100.0%	87.3%	100.0%	100.0%
	Error	2.60e-02	3.44e-02	1.08e-02	4.87e-02	3.32e-02	6.92e-03
	$\mathcal{F}_{avg}$	8.58e-02	1.25e-02	9.30e-03	2.12e-02	8.01e-03	3.23e-04
	n_{iter}	8009.3	8392.8	7358.0	9669.8	9553.8	7925.7
Schwefel 2.20	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	2.11e-05	1.73e-06	7.32e-07	3.65e-06	1.63e-06	1.09e-06
	$\mathcal{F}_{avg}$	2.93e-03	4.99e-04	2.18e-04	5.14e-05	2.46e-05	8.01e-06
	n_{iter}	865.9	749.8	668.3	863.2	747.0	665.8
Salomon	Rate	84.7%	98.7%	100.0%	100.0%	100.0%	100.0%
	Error	8.94e-02	6.45e-02	4.99e-02	3.72e-02	3.21e-02	2.75e-02
	$\mathcal{F}_{avg}$	8.96e-01	6.66e-01	5.24e-01	3.83e-01	3.21e-01	2.75e-01
	n_{iter}	1749.3	1657.9	1631.9	2193.7	1749.7	1138.2
XSY random	Rate	90.0%	99.3%	100.0%	100.0%	100.0%	100.0%
	Error	4.11e-02	2.26e-02	1.14e-02	2.45e-02	1.67e-02	1.66e-02
	$\mathcal{F}_{avg}$	5.64e-07	9.60e-08	6.06e-08	9.75e-09	7.26e-09	4.56e-09
	n_{iter}	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0
XSY 4	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	1.09e+00	9.85e-01	9.70e-01	8.56e-01	8.19e-01	7.97e-01
	$\mathcal{F}_{avg}$	2.88e-05	2.57e-05	7.44e-05	1.69e-07	1.42e-07	1.41e-07
	n_{iter}	9682.5	9018.1	8861.6	10000.0	10000.0	10000.0

SD-PSO with memory ($m = 0$) for $\lambda_1 = \xi \cdot \lambda_2$, $\sigma_1 = \xi \cdot \sigma_2$, $\lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$, $\alpha = 5 \times 10^4$.

Parameter selection for benchmark functions $d = 20$

$\xi = 0$ $\xi = 0.25$



SD-PSO with memory ($m = 0$) for λ_1 and σ_1 given by $\lambda_1 = \xi \cdot \lambda_2$ and $\sigma_1 = \xi \cdot \sigma_2$ with $\xi \in [0, 1]$.

Algorithmic improvements

- Evaluate $\bar{Y}_n^{\alpha,\mathcal{F}}$ on batches J_b of $N_b < N$ particles[8]

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

- 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

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

Binary interaction algorithms

A case of particular interest is when we restrict to batches of $N_b = 2$ particles interacting $(X_i^n, X_j^n) \to (X_i^{n+1}, X_j^{n+1})$ accordingly to[9]

$$
\begin{aligned}
X_i^{n+1} &= X_i^n + \lambda_1(X_{\beta,\mathcal{F}}(X_i^n, X_j^n) - X_i^n) + \lambda_2(X_{\alpha,\mathcal{F}}^n - X_i^n) + \sigma_1 D_1(X_i^n, X_j^n)\xi_1^n + \sigma_2 D_2(X_i^n)\xi_2^n \\
X_j^{n+1} &= X_j^n + \lambda_1(X_{\beta,\mathcal{F}}(X_j^n, X_i^n) - X_j^n) + \lambda_2(X_{\alpha,\mathcal{F}}^n - X_j^n) + \sigma_1 D_1(X_j^n, X_i^n)\theta_1^n + \sigma_2 D_2(X_j^n)\theta_2^n
\end{aligned}
$$

where $v_{\beta,\mathcal{F}}(X_i^n, X_j^n)$, $\beta > 0$, is the microscopic estimate of the best position

$$
X_{\beta,\mathcal{F}}(X_i^n, X_j^n) = \frac{\omega_\beta^{\mathcal{F}}(X_i^n) X_i^n + \omega_\beta^{\mathcal{F}}(X_j^n) X_j^n}{\omega_\beta^{\mathcal{F}}(X_i^n) + \omega_\beta^{\mathcal{F}}(X_j^n)}, \qquad \omega_\beta^{\mathcal{F}}(X) := e^{-\beta \mathcal{F}(X)},
$$

and $X_{\alpha,\mathcal{F}}^n$, $\alpha > 0$, is the macroscopic collective estimate

$$
X_{\alpha,\mathcal{F}}^n = \frac{\sum_{i=1}^N X_i^n \omega_\alpha^{\mathcal{F}}(X_i^n)}{\sum_{i=1}^N \omega_\alpha^{\mathcal{F}}(X_i^n)}, \qquad \omega_\alpha^{\mathcal{F}}(X) := e^{-\alpha \mathcal{F}(X)}
$$

with $\xi_k^n, \theta_k^n \in \mathbb{R}^d$, $k = 1, 2$ random vectors with mean 0 and variance 1 and

$$
D_1(X_i^n, X_j^n) = \operatorname{diag}\left\{(X_{\beta,\mathcal{F}}(X_i^n, X_j^n) - X)_h\right\}, \quad D_2(X) = \operatorname{diag}\left\{(X_{\alpha,\mathcal{F}}^n - X)_h\right\}, \quad h = 1, \dots, d.
$$

[9] A. Benfenati, G. Borghi, L. Pareschi. *arXiv:2105.02695, 2021*

Non local mean field PSO

The mathematical description of the above process for large numbers of interacting particles can resort on a Boltzmann type description and on the related mean field approximation.
Considering only microscopic best estimate ($\lambda_2 = \sigma_2 = 0$) we obtain the mean-field CBO dynamic

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

The explicit expression of the diffusion term are given below for the isotropic case

$$\int_{\mathbb{R}^d} D_{ii}^2(x, x_*) f(x_*, t)\, dx_* = \sum_{j=1}^{d} \int_{\mathbb{R}^d} \gamma_\beta^{\mathcal{F}}(x, x_*)^2 (x_{*,j} - x_j)^2 f(x_*, t)\, dx_*$$

and the anisotropic one

$$\int_{\mathbb{R}^d} D_{ii}^2(x, x_*) f(x_*, t)\, dx_* = \int_{\mathbb{R}^d} \gamma_\beta^{\mathcal{F}}(x, x_*)^2 (x_{*,i} - x_i)^2 f(x_*, t)\, dx_*.$$

In contrast to classical CBO method, both alignment as well as diffusion processes are nonlocal.

Convergence to global minimum

In this case, under suitable assumptions on the function $\mathcal{F}$, one can prove[10]

Theorem (Benfenati, Borghi, Pareschi)

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

$$\mu := \frac{\lambda}{C_{\beta,\mathcal{F}}} - 2\sigma^2\kappa > 0$$

$$\nu := \frac{4(\lambda c_1 + \sigma^2\kappa c_2)\beta e^{-\beta\mathcal{F}}}{\mu \|\omega_\beta^{\mathcal{F}}\|_{L^1(f_0)}} \max\{V(0)^{\frac{1}{2}}, V(0)\} < \frac{1}{2}$$

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

$$\mathcal{F}(\tilde{x}) \leq \inf_{x \in \mathbb{R}^d} \mathcal{F}(x) + r(\beta) + \frac{\log 2}{\beta}$$

where, if a miminizer x^* of $\mathcal{F}$ belongs to supp(f_0), then
$r(\beta) := -\frac{1}{\beta} \log \|\omega_\beta^{\mathcal{F}}\|_{L^1(f_0)} - \inf_{x \in \mathbb{R}^d} \mathcal{F}(x) \longrightarrow 0$ *as $\beta \to \infty$ thanks to the Laplace principle.*

[10] A. Benfenati, G. Borghi, L. Pareschi. *arXiv:2105.02695, 2021*

Application to a machine learning problem

We apply the binary CBO technique to a classical problem of Machine Learning: the scope is to recognize digital numbers contained in images of the MNIST dataset, by using a shallow network

$$f(x; W, b) = \mathrm{softmax}\left(\mathrm{ReLU}\left(Wx + b\right)\right)$$

where $x \in \mathbb{R}^{784}, W \in \mathbb{R}^{10\times 784}$, $b \in \mathbb{R}^{10}$. Moreover

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

being ReLU the well–known Rectified Linear Unit function. The training of the shallow network consists in minimizing the following function

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

where X is the training dataset, whose images are vectorized ($\mathbb{R}^{28\times 28} \to \mathbb{R}^{784}$) and stacked column–wise. The function ℓ is the cross entropy.

MNIST dataset - SGD and binary CBO

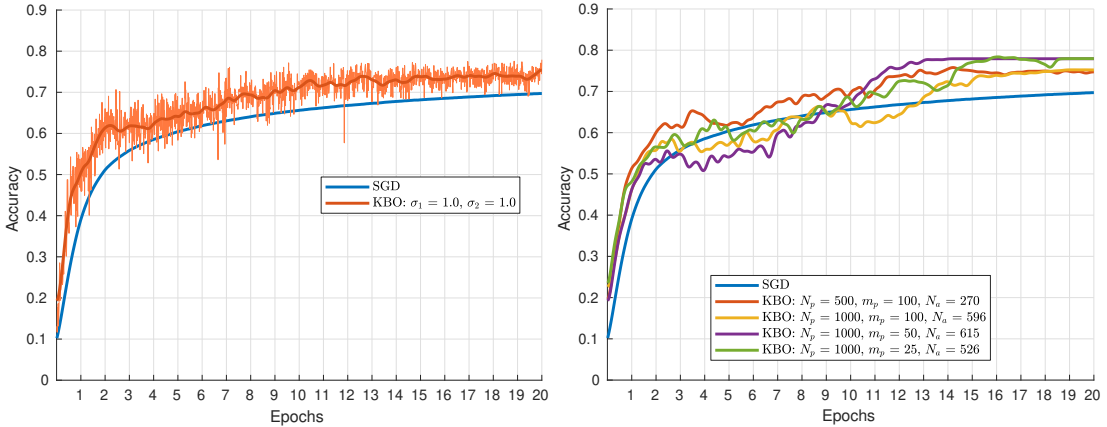
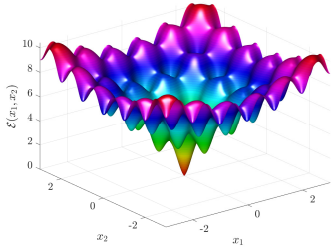


Figure: Performance comparison among SGD and binary CBO. The line referring to SGD shows the average over 500 simulations. The orange line refer to the CBO where both microscopic and macroscopic estimate are employed. The plot on the left depicts the performance of the CBO approach using $N_p = 500$ without any particle reduction strategy (the solid line is a smooth representation of the shaded one), while the plot on the right refers to particle reduction with $\mu = 0.1$ with different choices for particle numbers N_p and particles' batch m_p. The average number of particles is denoted by N_a.

Part II:
Consensus Based Optimization on hypersurfaces

Stochastic particle optimization on hypersurfaces

- Constrained problems on hypersurfaces are ubiquitous in the natural sciences, engineering or computer science.
- A vast class of optimization problems can be reduced to constrained optimizations over the sphere where the vector of particles has a unitary norm.
- For example a variety of nonlinear optimization problems on a sphere need to be performed over the surface of the Earth in geophysics, climate modeling, or global navigation.
- Other applications in signal processing and machine learning, for example the phase retrieval problem and the robust subspace detection.



Consensus based optimization on the sphere

Motivated by these aspects, we consider

$$x^* \in \arg\min_{x \in \mathbb{S}^{d-1}} \mathcal{F}(x)\,,$$

where $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ is a given continuous cost function.
System of N interacting particles $\{(X_t^i)_{t\geq 0}\}_{i=1,\dots,N}$ following a Kuramoto-Vicsek consensus based optimization (KV-CBO) dynamic in Itô's form

$$dX_t^i = \underbrace{\lambda P(X_t^i)\bar{X}_t^\alpha dt}_{\text{alignment}} + \underbrace{\sigma|X_t^i - \bar{X}_t^\alpha|P(X_t^i)dB_t^i}_{\text{exploration}} - \underbrace{\frac{\sigma^2}{2}(X_t^i - \bar{X}_t^\alpha)^2\frac{(d-1)X_t^i}{|X_t^i|^2}dt}_{\text{constraint}}\,,$$

where $\lambda, \sigma > 0$, $P(x) = I - \dfrac{xx^T}{|x|^2}$, and

$$x^* \approx \bar{X}_t^\alpha(\rho_t^N) = \frac{\sum_{j=1}^N X_t^j \omega_\alpha^{\mathcal{F}}(X_t^j)}{\sum_{j=1}^N \omega_\alpha^{\mathcal{F}}(X_t^j)}\,, \qquad \omega_\alpha^{\mathcal{F}}(x) := e^{-\alpha\mathcal{F}(x)}\,.$$

Simple numerical implementation (Euler-Maruyama)

Algorithm:

- Inputs: Δt, σ, α, d, N, n_T and the function $\mathcal{F}(\cdot)$
- Generate X_0^i, $i = 1, \dots, N$ sample vectors uniformly on $\mathbb{S}^{d-1}$;
- For $n = 0$ to n_T
  - Generate ΔB_n^i independent normal random vectors $N(0, \Delta t)$;
  - Compute $\bar{X}_n^{\alpha,\mathcal{F}}$;
  - $\tilde{X}_{n+1}^i \leftarrow$

    $$X_n^i + \Delta t P(X_n^i)\bar{X}_n^{\alpha,\mathcal{F}} + \sigma |X_n^i - \bar{X}_n^{\alpha,\mathcal{F}}| P(X_n^i)\Delta B_n^i - \Delta t \frac{\sigma^2}{2}(X_n^i - \bar{X}_n^{\alpha,\mathcal{F}})^2 (d-1) X_n^i,$$

  - $X_{n+1}^i \leftarrow \tilde{X}_{n+1}^i / |\tilde{X}_{n+1}^i|$, $i = 1, \dots, N$;

Ackley function on the sphere centered minimum

Ackley function on the sphere shifted minimum

Rastrigin function on the sphere centered minimum

Rastrigin function on the sphere shifted minimum

Ackley function in $d = 20$

x^*		$N = 50$ $N_b = 40$	$N = 100$ $N_b = 70$	$N = 200$ $N_b = 100$
$(0, \dots, 0, 1)^T$	Rate	100%	100%	100%
	Error	$2.24118e-08$	$1.3364e-09$	$3.51083e-09$
$(d^{-1/2}, \dots, d^{-1/2})^T$	Rate	98%	99%	100%
	Error	$1.15704e-06$	$1.476e-09$	$5.09216e-09$

Table I $\alpha = 5 \times 10^4$, $\sigma = 0.3$, $\Delta t = 0.05$, $N_{\min} = 10$ and $T = 100$

x^*		$N_0 = 100$ $N_b = 70$	$N_0 = 200$ $N_b = 100$	$N_0 = 400$ $N_b = 150$
$(0, \dots, 0, 1)^T$ $\mu = 0.3$	Rate	100%	100%	100%
	Error	$1.20639e-07$	$3.73419e-08$	$2.24362e-08$
	N_{avg}	21.6	38.7	71.4
$(d^{-1/2}, \dots, d^{-1/2})^T$ $\mu = 0.2$	Rate	100%	100%	100%
	Error	$1.34745e-06$	$2.02787e-08$	$8.06536e-09$
	N_{avg}	27.3	53.1	103.0

Table II $\alpha = 5 \times 10^4$, $\sigma = 0.3$, $\Delta t = 0.05$, $N_{\min} = 10$ and $T = 100$

Main theoretical result

Assumptions[11]:

1. $\mathcal{F} \in C^2(\mathbb{S}^{d-1})$;
2. For any $x \in \mathbb{S}^{d-1}$ there exists a minimizer $x^* \in \mathbb{S}^{d-1}$ of $\mathcal{F}$ (which may depend on x) such that it holds (*inverse continuity)*

$$|x - x^*| \leq C_0 |\mathcal{F}(x) - \underline{\mathcal{F}}|^\beta ,$$

where β, C_0 are some positive constants.

Theorem (Fornasier,Hui,P.,Sünnen)

For all $\varepsilon > 0$, assume that the intital datum and parameters are well-prepared for a time horizon $T^* > 0$ and parameter $\alpha > 0$ large enough. Then

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X^i_{\Delta t, n_{T^*}} - x^*\right|^2\right] \lesssim \underbrace{(\Delta t)^{2m}}_{Discretization\,error} + \underbrace{N^{-2/d}}_{Mean-field\,limit} + \underbrace{\varepsilon^2}_{Laplace\,principle} .$$

[11] Regularity of $\mathcal{F}$ is required to ensure formal well-posedness and for the large-time behavior analysis, but it is not necessary in numerical implementations. Requirement 2 is a bit more technical and needs to be verified, depending on the specific application

Fundamental steps of the proof I

- Well-posedness of the multiparticle SDE system;
- Well-posedness of the an auxiliary self-consistent nonlinear SDE satisfying

$$d\overline{X}_t = \lambda P(\overline{X}_t)\bar{X}_t^{\alpha}(\rho_t)dt + \sigma|\overline{X}_t - \bar{X}_t^{\alpha}(\rho_t)|P(\overline{X}_t)dB_t \\ - \frac{(d-1)\sigma^2}{2}(\overline{X}_t - \bar{X}_t^{\alpha}(\rho_t))^2\frac{\overline{X}_t}{|\overline{X}_t|^2}dt\,,$$

  with the initial data $\overline{X}_0$ distributed according to $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\rho_t = \mathrm{law}(\overline{\mathrm{X}}_{\mathrm{t}})$;;
- Proof that ρ_t solves the mean-field KV-CBO equation on the sphere and uniqueness:

$$\partial_t\rho_t = \lambda\nabla_{\mathbb{S}^{d-1}}\cdot((\langle\bar{X}_{\alpha,\mathcal{F}}(\rho_t),x\rangle x - \bar{X}_t^{\alpha}(\rho_t))\rho_t) + \frac{\sigma^2}{2}\Delta_{\mathbb{S}^{d-1}}(|x-\bar{X}_t^{\alpha}(\rho_t)|^2\rho_t),$$

with the initial data $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$. Here $\rho = \rho(t,x) \in \mathcal{P}(\mathbb{S}^{d-1})$ is a Borel probability measure on $\mathbb{S}^{d-1}$ and

$$\bar{X}_t^{\alpha}(\rho_t) = \frac{\int_{\mathbb{S}^{d-1}} x\,\omega_{\alpha}^{\mathcal{F}}(x)\,d\rho_t}{\int_{\mathbb{S}^{d-1}}\omega_{\alpha}^{\mathcal{F}}(x)\,d\rho_t}.$$

Fundamental steps of the proof II

- Mean-field limit:
$$\sup_{t\in[0,T]} \mathbb{E}\left[W_2^2(\rho_t^N, \rho_t)\right] \lesssim N^{-2/d} \to 0 \text{ as } N \to \infty\,.$$
where W_2 is the 2-Wasserstein distance.
- Regularity of solutions: for $\rho_0 \in L^2(\mathbb{S}^{d-1})$
$$\rho \in L^\infty([0,T]; L^2(\mathbb{S}^{d-1})) \cap L^2([0,T]; H^1(\mathbb{S}^{d-1})).$$
- Large time behavior: $\varepsilon > 0$ and assume that the intital datum and parameters are well-prepared for a time horizon $T^* > 0$ and parameter $\alpha^* > 0$ large enough. Then $E(\rho_{T^*})$ well approximates a minimizer x^* of $\mathcal{F}$,
$$|E(\rho_{T^*}) - x^*| \leq \varepsilon,$$
where $E(\rho_{T^*}) = \int_{\mathbb{S}^{d-1}} x d\rho_{T^*}$

Concluding step

Considering the time discretization error, the mean-field limit, and the asymptotic analysis, we conclude[12]

$$\begin{aligned}\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X^i_{\Delta t, n_{T^*}} - x^*\right|^2\right] &\lesssim \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X^i_{\Delta t, n_{T^*}} - X^i_{T^*}\right|^2\right] \\ &+ \mathbb{E}[W_2^2(\rho^N_{T^*}, \rho_{T^*})] + |E(\rho_{T^*}) - x^*|^2 \\ &\lesssim (\Delta t)^{2m} + N^{-2/d} + \varepsilon^2 \,.\end{aligned}$$

[12] The Euler-Maruyama scheme converges strongly with order $m = 1/2$.

Application: Phase retrieval problem

- Recently there has been growing interest in recovering an input vector $z^* \in \mathbb{R}^d$ from quadratic measurements
$$y_i = |\langle z^*, a_i \rangle|^2 + w_i, \quad i = 1, ..., M$$
where w_i is adversarial noise, and a_i are a set of known vectors.
- Phase retrieval problems arise in many areas of optics, where the detector can only measure the magnitude of the received optical wave. Important applications of phase retrieval include X-ray crystallography, transmission electron microscopy and coherent diffractive imaging.
- Several algorithms have been devised based on different principles, such as alternating projections, lifting and convex relaxation, and simple gradient descent for empirical risk minimization:
$$\mathcal{F}(z) = \sum_{i=1}^{M} \left| |\langle z, a_i \rangle|^2 - y_i \right|^2 .$$

Phase retrieval problem: comparison with state of the art

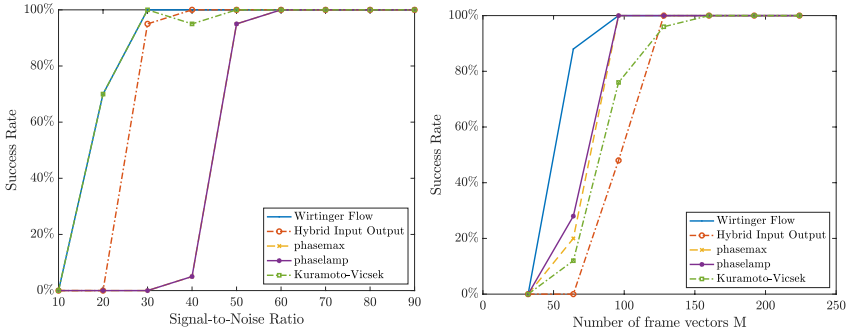


Figure: We used $\sigma = 0.2, \Delta t = 0.1, N = 10^4$ and chose the parameter α adaptively. The results are averaged 25 times.

Left: Success rate in terms of the Signal-to-Noise Ratio in dimension $d = 32$ for $M = 4d$ Gaussian vectors. The green dashed curve representing KV-CBO is exactly superimposed with the light blue curve of the Wirtinger Flow[13].

Right: Phase transitions for different Gaussian vectors M in dimension $d = 32$.

[13] E.J Candes, X. Li, M. Soltanolkotabi IEEE Tran. Inf. Theo. 2015

Application: Robust subspace detection

Cloud of points $\mathcal{Q} = \{x^{(i)} \in \mathbb{R}^d : i = 1, ..., M\}$ in an Euclidean space with $d \gg 1$. Robust subspace detection: minimization of the energy

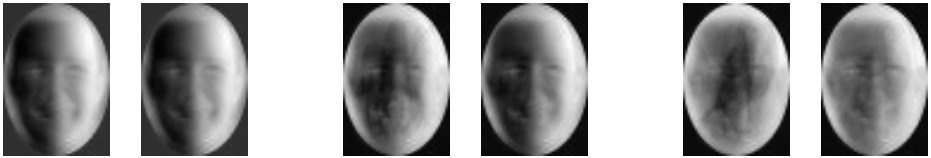
$$\mathcal{F}_p(x) := \sum_{i=1}^{M} |x^{(i)} - \langle x^{(i)}, x\rangle x|^p = \sum_{i=1}^{M} \left(|x^{(i)}|^2 - |\langle x^{(i)}, x\rangle|^2\right)^{p/2}, \quad x \in \mathbb{S}^{d-1},$$

for $0 < p \ll 2$.

Figure: Samples from the *10K US Adult Faces Database* and one outlier.

We chose a subset of $M = 421$ gray scale images of size 64×45 from the 10K US Adult Faces Database, which yields a point cloud in $d = 2880$.

Robust subspace detection: computation of eigenfaces



(a) (b) (c) (d) (e) (f)

Figure: $\lambda = 1, \sigma = 0.019, \Delta t = 0.25, T = 25000$, $N = 5000$ and $N_{min} = 150$ for (b) and (d). For (f) we used $N = 50000$ and $N_{min} = 5000$.

1. Eigenface for the point cloud of faces without outliers computed by SVD (a), and the KV-CBO method (b).
2. Eigenface for point cloud with 6 outliers computed by SVD (c), and the KV-CBO method with $p = 1$ (d).
3. Eigenface for point cloud with 12 outliers computed by SVD (e), and the KV-CBO method with $p = 0.5$ (f).

PSNR: 61.4214 for (a) and (b),
15.9764 for (a) and (c), 20.7344 for (a) and (d),
12.3109 for (a) and (e) and 14.2892 for (a) and (f).

Some future research directions

- Further testing of the CBO system with memory (and momentum);
- Rigorous convergence to global minimum in presence of memory
- Testing adaptive inertia and determination of optimal sets of parameters;
- Extensions of the methods to multi-objective problems;
- Extensions to constrained optimization problems.
- ...

Collaborators:

A. Benfenati (U. Milan), G. Borghi (U. Aachen & U. Ferrara), S. Grassi (U. Ferrara), M. Fornasier (TU Munich), P. Sünnen (TU Munich)
H. Huang (U. Calgary), J. Qiu (U. Calgary)

References

- Sara Grassi, Lorenzo Pareschi, From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit, Math. Mod. Meth. App. Sci. 31(8):1625–1657, 2021, arXiv:2012.05613, 2020
- Sara Grassi, Hui Huang, Lorenzo Pareschi, Jinniao Qiu, Mean-field particle swarm optimization, to appear in Modeling and Simulation for Collective Dynamics, IMS Lecture Note Series, World Scientific, preprint arXiv:2108.00393, 2021
- Alessandro Benfenati, Giacomo Borghi, Lorenzo Pareschi, Binary interaction methods for high dimensional global optimization and machine learning, preprint arXiv:2105.02695, 2021
- Massimo Fornasier, Hui Huang, Lorenzo Pareschi, Philippe Sünnen, Anisotropic Diffusion in Consensus-based Optimization on the Sphere, SIAM J. Opt. to appear arXiv:2104.00420, 2021
- Massimo Fornasier, Hui Huang, Lorenzo Pareschi, Philippe Sünnen, Consensus-based Optimization on the Sphere: Convergence to Global Minimizers and Machine Learning, J. Machine Learning Research to appear, arXiv:2001.11988, 2020
- Massimo Fornasier, Hui Huang, Lorenzo Pareschi, Philippe Sünnen, Consensus-Based Optimization on Hypersurfaces: Well-Posedness and Mean-Field Limit, Math. Mod. Meth. Appl. Sci. 30(14):2725-2751, 2020, arXiv:2001.11994, 2020