Convergence of Consensus-Based Optimization (CBO) with Random Batch Interactions

Dongnam Ko

The Catholic University of Korea

Joint work with Seung-Yeal Ha, Shi Jin, and Doheon Kim

FAU DCN-AvH Seminar,
28 June, 2024

Overview

Table of Contents

consensus-based optimization (CBO)

CBO: an evolutionary type of gradient-free optimization algorithms to find the minimum of a given cost function.

Similar methods; Genetic algorithm, Particle Swarm Optimization, Wolf-pack optimization, Ant Colony Optimization, etc.:

1. First, **spread the particles** into the domain.
2. Second, **evaluate current values** from particles' positions.
3. Third, process time-**evolution toward the possible minimum** positions.

For a given $L(x)$, we want $x_i(t)$ to approach $x_* := \operatorname{argmin}_{x \in \mathbb{R}^d} L(x)$.

consensus-based optimization (CBO)

SIMULATIONS:

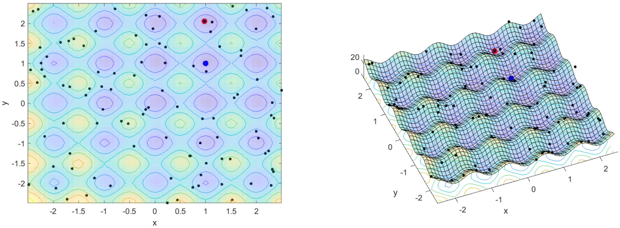

Figure: Initial particle distribution and the Rastrigin cost function

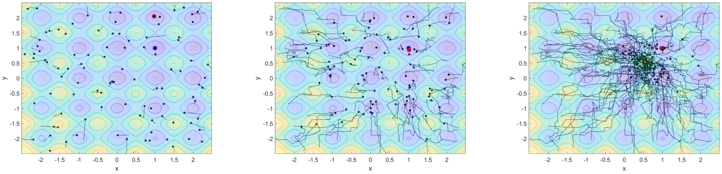

Figure: Particle distribution at (left) $t = 2$, (middle) $t = 10$, (right) $t = 50$

consensus-based optimization (CBO)

Therefore, each particle **explores the domain**:

- x_i = i-th agent's guess for $\mathrm{argmin}_{x\in\mathbb{R}^d} L(x)$
- Iterate on $t \in \mathbb{N}$,

$$x_i(t+1) = x_i(t) + \text{(interactions)} + \text{(random walks)}, \quad i = 1, \dots, N$$

Three questions:

1. **[Consensus]** $x_i(t) - x_j(t)$ decays to zero.
2. **[Convergence]** all $x_i(t)$ converge to its limit $x_i(\infty)$.
3. **[Optimality]** $x_i(\infty) \approx \mathrm{argmin}_{x\in\mathbb{R}^d} L(x)$ for some i.

consensus-based optimization (CBO)

CBO is a kind of simplied version of multi-point optimization methods.

Algorithm [K.–Ha–Jin–Kim 2022] based on [Carrillo–Jin–Li–Zhu 2021]

$$X^i_{(t+1)} = X^i_t + \gamma(\bar{X}^{i,*}_t - X^i_t) + \mathrm{diag}(\eta^{i,1}_t, \dots, \eta^{i,d}_t)(\bar{X}^{i,*}_t - X^i_t),$$

$$\gamma > 0, \quad \eta^{i,\ell}_t \sim \mathcal{N}(0, \sqrt{\zeta}) \quad \text{for each} \quad i, \ell, t \quad \text{and}$$

$$\bar{X}^{i,*}_t := \mathrm{argmin}_{x \in \{X^j_t | j \in N_i(t)\}} L(x), \quad N_i(t) \subset \{1, 2, \dots, N\}.$$

This time-evolution corresponds to a continuous stochastic dynamics:

$$dX^i_t = \lambda(\bar{X}^{i,*}_t - X^i_t)dt + \sigma\, \mathrm{diag}(\bar{X}^{i,*}_t - X^i_t)dW^i_t,$$

which is surely **gradient-free**.

Main result: Consensus of a general CBO algorithm

For the CBO algorithm

$$X^i_{(t+1)} = X^i_t + \gamma(\bar{X}^{i,*}_t - X^i_t) + \mathrm{diag}(\eta^{i,1}_t, \dots, \eta^{i,d}_t)(\bar{X}^{i,*}_t - X^i_t),$$

we assume $\bar{X}^{i,*}_t$ is an arbitrary convex combination among X^j_t.

Theorem (K.–Ha–Jin–Kim 2022)

For sufficiently small $\mathrm{Var}(\eta^{i,l}_t), l = 1, \dots, N$,
(1) for some positive random variable ε,

$$\mathbb{E}\max_{i,j} \|X^i_t - X^j_t\| = \mathcal{O}(e^{-\varepsilon t}), \quad t \to \infty.$$

(2) the following holds almost surely: for some positive random ε,

$$\max_{i,j} \|X^i_t - X^j_t\| = \mathcal{O}(e^{-\varepsilon t}), \quad t \to \infty.$$

Numerical simulations; case 1

- A common way to set the target point $\bar{X}_t^{i,*}$ of the i-th particle is to set the best point among the whole particles:

$$\bar{X}_t^{i,*} := argmin_{x\in\{X_t^j\}} L(x), \quad \text{i.e.,} \quad L(X_t^{i,*}) = \min_{j=1,\dots,N} L(X_t^j).$$

- The simulation $(N = 100, \gamma = 0.01, \zeta = 0.5)$ on **Rastrigin functions** suggests the following probabilities to find the minimum: In 2D and 3D, nearly 100%, however, not exceeding 50% for high dimensional problems.

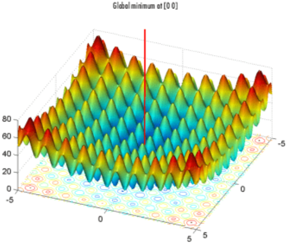

Figure: The landscape of Rastrigin function in d=2 (2D).

Table of Contents

Random Batch Method (RBM)

Suggestion on raising success rate: consider complex interaction structure, such as RBM (in [Carrillo–Jin–Li–Zhu 2021]).

Random Batch Method [Jin–Li–Liu, 2020] suggests a random sampling on the interactions at each time-step.

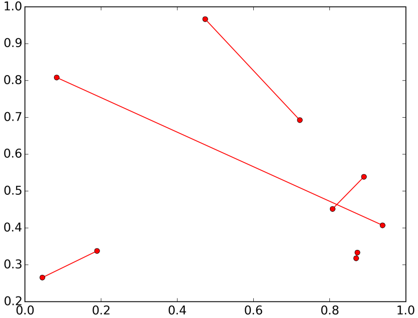

Figure: Random pairing from 10 particles ($P = 2$) from $10 \times 9/2$ interactions

Network structure of RBM

From RBM, we the interaction network becomes sparse:

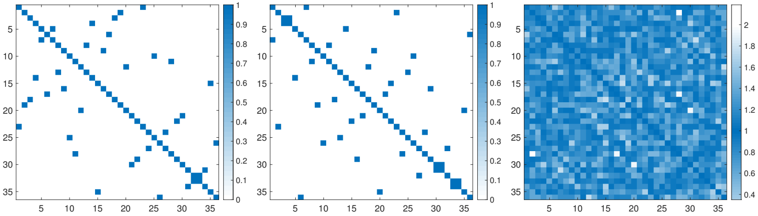

Figure: Color representation of the adjacency matrix.
(Left two) Examples of random networks from RBM ($P = 2, N = 36$).
(Right) An average of independent 400 matrices.

Convergence theorem [Jin–Li–Liu, 2020]

Following the law of large numbers, the solution of RBM converges to the original (all-to-all) dynamics as the time-discretization goes zero.

Numerical simulations; case RBM

The simulation result shows that if P gets smaller,
then the success rate grows but the cost of computation also grows.

Success rate	Full batch ($P = 100$)	$P = 50$	$P = 10$
d = 2	1.000	1.000	1.000
d = 3	0.988	0.983	0.998
d = 4	0.798	0.920	0.988
d = 5	0.712	0.658	0.931
d = 6	0.513	0.655	0.880
d = 7	0.388	0.464	0.854
d = 8	0.264	0.389	0.832
d = 9	0.170	0.323	0.868
d = 10	0.117	0.274	0.886

Figure: Success rates from 1000 simulations.

Numerical simulations; case RBM

The accuracy of algorithm directly depends on the computation time.

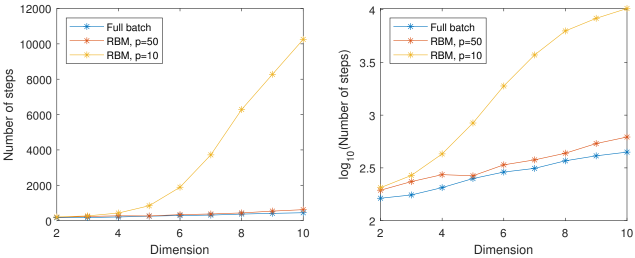

Figure: Average number of steps until stopping criterion holds.

The stopping criterion is made with the change of positions,

$$\sum_{i=1}^{N} |x_{n+1}^i - x_n^i|^2 < 10^{-3}.$$

consensus-based optimization (CBO)

Algorithm [K.–Ha–Jin–Kim 2022] based on [Carrillo–Jin–Li–Zhu 2021]

$$X^i_{(t+1)} = X^i_t + \gamma(\bar{X}^{i,*}_t - X^i_t) + \mathrm{diag}(\eta^{i,1}_t, \dots, \eta^{i,d}_t)(\bar{X}^{i,*}_t - X^i_t),$$

$$\gamma > 0, \quad \eta^{i,\ell}_t \sim \mathcal{N}(0, \sqrt{\zeta}) \quad \text{for each} \quad i, \ell, t \quad \text{and}$$

$$\bar{X}^{i,*}_t := \mathrm{argmin}_{x \in \{X^j_t | j \in N_i(t)\}} L(x), \quad N_i(t) \text{ determined by RBM.}$$

It corresponds to a continuous SDE:

$$dX^i_t = \lambda(\bar{X}^{i,*}_t - X^i_t)dt + \sigma\,\mathrm{diag}(\bar{X}^{i,*}_t - X^i_t)dW^i_t.$$

Here $\bar{X}^{i,*}_t$ is the minimum guess from the batch of the i-th particle.
Objectives: here we want to guarantee the termination of the algorithm by showing convergence of the particles: $X^i_t \to X_\infty$ as $t \to \infty$.

Table of Contents

Previous works

Proposal of algorithms & Analysis of the convergence

- [Askari-Sichani–Jalili 2013]: suggestion of the first CBO
- [Pinnau–Totzeck–Tse–Martin 2017]: suggestion of CBO with noise

- [Carrillo–Choi–Totzeck–Tse 2018]: convergence of the kinetic CBO
- [Ha–Jin–Kim 2020,2021]: convergence of a simple case
- [Carrillo–Jin–Li–Zhu 2021]: suggestion of CBO with RBM
- [Fornasier—Klock—Riedl 2021]: optimality of kinetic CBO

- [K.–Ha–Jin–Kim 2022]; current talk.
- [Byeon–Ha–Won 2024 arXiv]; further study on convergence condition

Two examples in Literature

The exploration of CBO algorithm has two parts of randomness.

Algorithm with interaction network [Askari-Sichani–Jalili 2013]

$$x_i(t+1) = x_i(t) + \gamma(\bar{x}_i^*(t) - x_i(t)), \quad \bar{x}_i^*(t) = \operatorname{argmin}_{x_k(t):k\in N_i(t)} L(\cdot)$$

Algorithm for noisy trajectory [Pinnau–Totzeck–Tse–Martin 2017]

$$dX_t^i = \lambda(\bar{X}_t^* - X_t^i)dt + \sigma|\bar{X}_t^* - X_t^i|dW_t^i,$$

with

$$\bar{X}_t^* := \frac{1}{\sum_{j=1}^N e^{-\beta L(X_t^j)}} \sum_{j=1}^N e^{-\beta L(X_t^j)} X_t^j.$$

The second $\bar{X}_t^*$ is from the **Laplace principle**, which converges to the argument minimum as $\beta \to \infty$. This is an observable in F-P dynamics.

Algorithm in [Pinnau–Totzeck–Tse–Martin 2017]

Algorithm [Pinnau–Totzeck–Tse–Martin 2017]

$$dX_t^i = \lambda(\bar{X}_t^* - X_t^i)dt + \sigma|\bar{X}_t^* - X_t^i|dW_t^i,$$

with

$$\bar{X}_t^* := \frac{1}{\sum_{j=1}^N e^{-\beta L(X_t^j)}} \sum_{j=1}^N e^{-\beta L(X_t^j)} X_t^j.$$

We can formally send $N \to \infty$ to get

$$\bar{X}_t^* \to \frac{1}{\int_{\mathbb{R}^d} e^{-\beta L(x)} d\rho_t} \int_{\mathbb{R}^d} e^{-\beta L(x)} x d\rho_t, \quad \rho_t : \text{prob. measure of } X_t^i.$$

If L has a unique minimizer x_* in the support of ρ_t, then

$$m[\rho_t] := \frac{1}{\int_{\mathbb{R}^d} e^{-\beta L(x)} d\rho_t} \int_{\mathbb{R}^d} e^{-\beta L(x)} x d\rho_t \to x_* \quad \text{as} \quad \beta \to \infty.$$

Mean-field limit to the kinetic dynamics

From the mean-field limit process, the dynamics of

$$dX_t^i = \lambda(\bar{X}_t^* - X_t^i)dt + \sigma|\bar{X}_t^* - X_t^i|dW_t^i$$

becomes dynamics of the density $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ as a Fokker-Planck equation:

$$\partial_t \rho_t = \lambda \nabla \cdot ((x - m[\rho_t])\rho_t) + \frac{\sigma^2}{2}\Delta(|x - m[\rho_t]|^2 \rho_t).$$

Theorem (Convergence) [Pinnau–Totzeck–Tse–Martin 2017]

If λ is large enough (compared to d, σ^2, and $e^{-\beta}$), then $\mathbb{E}(\rho_t)$ converges and

$$\mathrm{Var}(\rho_t) = O(e^{-ct}), \quad t \to \infty.$$

Idea: $\frac{d}{dt}\mathrm{Var}(\rho_t) = -2\lambda\,\mathrm{Var}(\rho_t) + (d\sigma^2/2)\int (x - m[\rho_t])^2 d\rho_t.$

Limitations of kinetic approach

CBO with RBM converges to the kinetic equation as $\tau \to 0$ and $N \to \infty$. However, kinetic equation limit has two significant limitation.

1. **Dynamics ignores network structure**:
   Interactions are always all-to-all.

2. **Sample space covers the whole space**:
   We already know the whole data; including global minimum.
   [Fornasier–Klock–Riedl, 2021]

We should get back to the discrete-time dynamics in order to analyze the performance of algorithms.

Table of Contents

Main result: Consensus of a general CBO algorithm

For the CBO algorithm

$$X^i_{(t+1)} = X^i_t + \gamma(\bar{X}^{i,*}_t - X^i_t) + \mathrm{diag}(\eta^{i,1}_t, \dots, \eta^{i,d}_t)(\bar{X}^{i,*}_t - X^i_t),$$

we assume $\bar{X}^{i,*}_t$ is in the convex hull among X^i_t.

Theorem (K.–Ha–Jin–Kim 2022)

For sufficiently small (not depending on d) $\zeta := \mathrm{Var}(\eta^{i,l}_t)$,
(1) for some positive random variable ε,

$$\mathbb{E} \max_{i,j} \|X^i_t - X^j_t\| = \mathcal{O}(e^{-\varepsilon t}), \quad t \to \infty.$$

(2) the following holds almost surely: for some positive random ε,

$$\max_{i,j} \|X^i_t - X^j_t\| = \mathcal{O}(e^{-\varepsilon t}), \quad t \to \infty.$$

Algorithm in [Askari-Sichani–Jalili 2013]

We may **rewrite the dynamics** without noise in [Askari-Sichani–Jalili 2013],

$$x_i(t+1) = x_i(t) + \gamma(\bar{x}_i^*(t) - x_i(t)), \quad \bar{x}_i^*(t) = \text{argmin}_{x_k(t):k\in N_i(t)} L(\cdot),$$

as in the **matrix form**:

$$X(t+1) = A(t)X(t).$$

For example, if there are 4 particles and the third is the $\bar{x}_i^*(t)$ for all i:

$$A(t) = \begin{bmatrix} 1-\gamma & 0 & \gamma & 0 \\ 0 & 1-\gamma & \gamma & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \gamma & 1-\gamma \end{bmatrix}.$$

In this case, we have

$$|x_i(t+1) - x_j(t+1)| \le (1-\gamma)|x_i(t) - x_j(t)|.$$

Ergodicity coefficient

Question: **How can we prove the consensus**?
Answer: Analysis on the **diameter** as the variance in the kinetic equation.

$$\mathcal{D}(x) := \max_{i,j} |x_i - x_j|.$$

For a stochastic matrix, define ergodicity coefficient as

$$\alpha(A) := \min_{i,j} \sum_k \min\{a_{ik}, a_{jk}\} \in [0, 1].$$

Note: $\alpha(A) = 1 \Leftrightarrow$ all rows of A are identical.

Proposition [Askari-Sichani–Jalili 2013] from [Markov 1906]

$$\mathcal{D}(Ax) \leq (1 - \alpha(A))\mathcal{D}(x).$$

Problem 1; the case of ergodicity zero

For example, the stochastic matrix may look like

$$A = \begin{bmatrix} 1-\gamma & \gamma & 0 & 0 \\ 0 & 1-\gamma & \gamma & 0 \\ 0 & 0 & 1-\gamma & \gamma \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{such as} \quad A = \begin{bmatrix} 0.9 & 0.1 & 0 & 0 \\ 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Since this is a block matrix, the ergodicity constant should be zero:

$$\alpha(A) := \min_{i,j} \sum_k \min\{a_{ik}, a_{jk}\} = 0. \quad (i = 1, j = 3)$$

However, the network is connected; if we consider A^2 and A^3,

$$A^2 = \begin{bmatrix} 0.81 & 0.18 & 0.01 & 0 \\ 0 & 0.81 & 0.18 & 0.01 \\ 0 & 0 & 0.81 & 0.19 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 0.729 & 0.243 & 0.027 & 0.001 \\ 0 & 0.729 & 0.243 & 0.028 \\ 0 & 0 & 0.729 & 0.271 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Therefore, $\alpha(A^3) \neq 0$ since $i = 1$ and $j = 4$ has length 3 connectivity.

Analytical result in [Askari-Sichani–Jalili 2013]

Therefore, if the network is connect so that

$$\lim_{n\to\infty} \alpha(A^n) = 1,$$

then $\mathcal{D}(A^n x) \leq (1-\alpha(A^n))\mathcal{D}(x) \to 0$.
It is convenient to check the convergence to 1 as follows. Define

$$A((t,s]) := A(t-1)A(t-2)\dots A(s) \qquad (t > s).$$

Proposition [Askari-Sichani–Jalili 2013]

Assume that there exists $0 = t_0 < t_1 < t_2 < \dots$ satisfying

$$\sum_{i=1}^{\infty} \alpha(A((t_i, t_{i-1}])) = \infty.$$

Then $\lim_{n\to\infty} \alpha(A^n) = 1$ and $\mathcal{D}(X(t)) \to 0$ as $t \to \infty$.

Problem 2; connectivity of RBM

The problem occurs when $\bar{x}_i^* \neq \bar{x}_j^*$.

For example, the stochastic matrix may look like

$$A(t) = \begin{bmatrix} 1-\gamma & \gamma & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-\gamma & \gamma \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Since this is a block matrix, the ergodicity constant should be zero:

$$\alpha(A) := \min_{i,j} \sum_k \min\{a_{ik}, a_{jk}\} = 0. \quad (i = 1, j = 3)$$

Therefore, the random network needs connectivity.

Lemma: positive ergodicity of a network

Assume that for any i and j, suppose that the information of i at time t is observed by j at time $t + m$. Then, $\alpha(A((s + m, s])) \geq \gamma(1 - \gamma)^m$.

Consensus with RBM

Now we may prove that the diameter decays almost surely.

First, from **ergodicity** argument, we have

$$\mathcal{D}(X(t_{km})) \le \exp\Big(-\gamma(1-\gamma)^m \sum_{s=1}^{k} \mathcal{G}_{s,m} \Big) \mathcal{D}(X(0)),$$

where $\mathcal{G}_{s,m}$ is 1 if all particles are connected by length m, and 0 otherwise. (This is a random variable with positive probability to be 1)

Note that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{s=1}^{k} \mathcal{G}_{s,m} = \mathbb{E}[\mathcal{G}_{s,m}] = p_m > 0.$$

Therefore, we conclude the decay of the diameter

$$\mathcal{D}(X(t_{km})) \le \exp(-\Lambda(m,k)k)\mathcal{D}(X(0)),$$

$$\lim_{k\to\infty} \Lambda(m,k) = \gamma(1-\gamma)^m p_m.$$

In other words, the algorithm satisfy consensus.

Ergodicity with noise?

Now we consider noise;

$$X^i_{(t+1)} = X^i_t + \gamma(\bar{X}^{i,*}_t - X^i_t) + \mathrm{diag}(\eta^{i,1}_t, \dots, \eta^{i,d}_t)(\bar{X}^{i,*}_t - X^i_t),$$

Then, the network structure

$$A(t) = \begin{bmatrix} 1-\gamma & \gamma & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-\gamma & \gamma \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

becomes

$$A_\varepsilon(t) = \begin{bmatrix} 1-\gamma+\varepsilon & \gamma-\varepsilon & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-\gamma+\varepsilon & \gamma-\varepsilon \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then, A is stochastic but can have negative ergodicity.

Table of Contents

Summary and remaining questions

Remarks

1. The convergence of CBO holds with random batch method.
2. **This is the first result on the convergence of CBO with individual random motion on particles.**
3. No performance result yet on the global minimizer.

Thank you very much