Convergence of the discrete consensus-based optimization algorithm with heterogeneous noises

Dongnam Ko

The Catholic University of Korea

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

Analysis and PDE seminar, 13 April, 2022

Overview

Table of Contents

consensus-based optimization (CBO)

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

The basic principles are the same as other aggregation (multi-point) methods; Ant Colony Optimization, Particle Swarm Optimization, Genetic algorithm, 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 $x_* := \mathrm{argmin}_{x\in\mathbb{R}^d} L(x)$.

consensus-based optimization (CBO)

Therefore, each particle **explores the domain** based on the values of other particles.

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

$$x_i(t+1) = x_i(t) + (\text{interaction with other } x_j(t)\text{'s}), \quad i = 1, \ldots, N$$

Then, we have the following 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.

In a common multi-point algorithm, these three conditions may not be satisfied, but works well in practical problems with high probability.

consensus-based optimization (CBO)

Our interests lies in the consensus and convergence of the following version of the CBO algorithm.

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 discrete time-evolution is based on the following 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**.

consensus-based optimization (CBO)

SIMULATIONS:

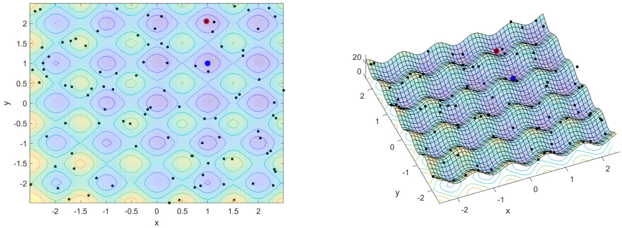


Figure: Initial particle distribution and the Rastrigin cost function

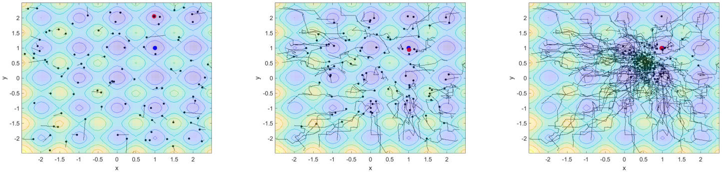
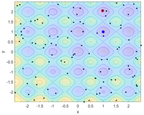
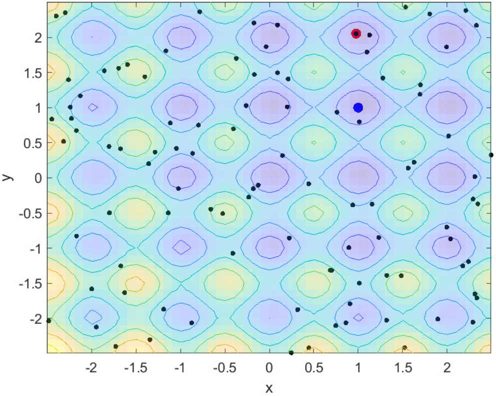


Figure: Particle distribution at (left) $t = 2$, (middle) $t = 10$, (right) $t = 50$

consensus-based optimization (CBO)





Previous works

Proposal of algorithms & Analysis of the convergence

- [Askari-Sichani–Jalili 2013] proposed and analyzed CBO without noise
- [Pinnau–Totzeck–Tse–Martin 2017] proposed CBO with noise

- [Carrillo–Choi–Totzeck–Tse 2018] analyzed the convergence of the kinetic CBO dynamics
- [Ha–Jin–Kim 2020,2021] analyzed the convergence of a simplified CBO algorithm.

- [Fornasier–Huang–Pareschi–Sünnen 2020] proposed CBO on hypersurfaces
- [Kim–Kang–Kim–Ha–Yang 2020] proposed CBO on the Stiefel manifold
- [Carrillo–Jin–Li–Zhu 2021] proposed CBO for high dimensional problems

Two examples in Literature

The exploration of CBO algorithm has two parts of randomness.

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

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

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 $\beta \to \infty$.

Table of Contents

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

Algorithm [Pinnau–Totzeck–Tse–Martin 2017]

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

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

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

Convergence for different noise

The same argument works with different multiplicative noise.

Algorithm for high dimension [Carrillo–Jin–Li–Zhu 2021]

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

Then, $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ satisfies the following Fokker-Planck equation:

$$\partial_t \rho_t = \lambda \nabla \cdot ((x - m[\rho_t])\rho_t) + \frac{\sigma^2}{2}\sum_{i=1}^{d} \partial_{ii}((x - m[\rho_t])_i^2 \rho_t).$$

Theorem (Convergence) [Carrillo–Jin–Li–Zhu 2021]

If λ is large enough (not depending on the dimension d), then $\mathbb{E}(\rho_t)$ converges and

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

Optimality of kinetic CBO

The **optimality** of the result also can be proved partially for a Fokker-Planck equation (actual proof is on SDE with a distributed initial data):

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

Theorem (Optimality) [Fornasier–Klock–Riedl, 2021]

Suppose that the cost function L is coercive (far-field) and the initial data ρ_0 is nonzero near the minimum point. If λ is large enough (compared to d, σ^2) and a tolerance constant ε is given, then large enough α satisfies

$$\int (x - x_*)^2 d\rho_t = O(e^{-ct}), \quad \text{until it is less then } \varepsilon.$$

Idea: if ρ_0 contains the minimum point, then $m[\rho_t] \sim x_*$ for $t > 0$.
It requires a quantitative estimate for Laplace principle.

Limitations of kinetic approach

The **kinetic equation** approximating the CBO model has two significant differences from the original discrete-time optimization algorithm.

1. **Lack of interaction network structure**:
All the individuals are considered as not-distinguishable particles. Therefore, we cannot consider 'local' communication between particles, but 'global' all-to-all interactions happen.

2. **Support of density functions covers the whole space**:
Since the density function becomes positive for the whole space, its searching space becomes the whole domain. It implies that we already evaluate all the cost values, so that the optimal point is already known.

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

Table of Contents

CBO with interaction network without noise

Before we proceed to the analysis of CBO with RBM, we consider the case without noise. This problem is related to the first CBO algorithm.

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

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

Until now, we considered **kinetic interpretation** of the CBO dynamics. From now on, we analyze **discrete-time** CBO algorithm itself.

Again, we have the following 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] $x_i(\infty) \approx \operatorname{argmin}_{x\in\mathbb{R}^d} L(x)$ for some i.

Algorithm in [Askari-Sichani–Jalili 2013]

We may **rewrite the dynamics** of [Askari-Sichani–Jalili 2013],

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

as in the **matrix form**:

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

Then, for small γ, $A(t)$ is a diagonal-dominant stochastic matrix (each entry≥ 0, each row sum$=1$), where the diagonal term is $(1-\gamma)$ and there is only one off-diagonal term in each column, γ.

We expect that, if $\bar{x}_i^*(t) = \bar{x}_j^*(t)$, then a kind of consensus works:

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

Ergodicity coefficient

Question: **How can we prove the consensus**?
Answer: The **Lyapunov approach** as in the kinetic equation.

For a real vector $x = (x_1, \dots, x_N) \in \mathbb{R}^N$, we may set

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

For a stochastic matrix (each entry≥ 0, each row sum$= 1$) $A = (a_{ij})$, we define its 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 [Markov 1906]

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

Ergodicity coefficient

Proposition [Markov 1906]

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

Different proofs can be found in Liturature, and we may verify it directly:

$$\begin{aligned}
\mathcal{D}(Az) &= \max_{i,j} \Big(\sum_k a_{ik} z_k - \sum_k a_{jk} z_k \Big) \\
&= \max_{i,j} \Big(\sum_k (a_{ik} - \min\{a_{ik}, a_{jk}\}) z_k - \sum_k (a_{jk} - \min\{a_{ik}, a_{jk}\}) z_k \Big) \\
&\leq \max_{i,j} \Big(1 - \sum_k \min\{a_{ik}, a_{jk}\}\Big) \Big(\max_k z_k - \min_k z_k \Big) \\
&= \Big(1 - \min_{i,j} \sum_k \min\{a_{ik}, a_{jk}\}\Big) \mathcal{D}(z) \\
&= (1 - \alpha(A))\mathcal{D}(x).
\end{aligned}$$

Ergodicity with interaction network

If we use the full network structure,

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

then the situation becomes simple.

For example, if there are 4 particles and the third is the smallest at t,

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

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

Clearly, $\alpha(A(t)) = \gamma > 0$ and the diameter decays to zero:

$$\mathcal{D}(A(t)\mathrm{x}) \leq (1-\gamma)\mathcal{D}(\mathrm{x}), \quad t \geq 0.$$

Ergodicity with random network

However, if $N_1(t) = N_2(t) = \{1, 2\}$ and $N_3(t) = N_4(t) = \{3, 4\}$, particles between $\{1, 2\}$ and $\{3, 4\}$ do not interact each other.

For example, the stochastic matrix looks 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 should mix the particles enough along time in order to make the diameter decay.

Analytical result in [Askari-Sichani–Jalili 2013]

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 $\mathcal{D}(X(t)) \to 0$ as $t \to \infty$.

Sketch of proof: From

$$X(t_n) = A((t_n, t_{n-1}]) \dots A((t_1, t_0])X(0)$$

we have

$$\begin{aligned}\mathcal{D}(X(t_n)) &= (1-\alpha(A((t_n,t_{n-1}])) \dots (1-\alpha(A((t_1,t_0])))\mathcal{D}(X(0)) \\ &\le \exp(-\alpha(A((t_n,t_{n-1}])) \dots \exp(-\alpha(A((t_1,t_0])))\mathcal{D}(X(0)).\end{aligned}$$

An example of random network: Random Batch Method

Another way to introduce randomness to the dynamics, is a random network.

Algorithm 2 [Carrillo–Jin–Li–Zhu 2021]

$$x_i(t+1) = x_i(t) + (\lambda I + \mathrm{diag}(noise))(\bar{x}_i^*(t) - x_i(t)),$$

$$\bar{x}_i^*(t) = \mathrm{argmin}_{x_k(t):k\in N_i(t)} L(\cdot) \quad \text{or} \quad \frac{1}{\sum_{j\in N_i(t)} e^{-\beta L(X_t^j)}} \sum_{j\in N_i(t)} e^{-\beta L(X_t^j)} X_t^j,$$

where the neighborhood $N_i(t)$ is determined by the graph $G(t)$, which is a disjoint union of P-vertex complete graphs, determined by a randomly chosen partition of $\{1, \ldots, N\}$ into P-element sets.

Such partition of a graph is called Random Batch Method (RBM) [Jin–Li–Liu, 2020].

Random Batch Method(RBM)

Remaining question: how we achieve the sufficient condition.

Random Batch Method [Jin–Li–Liu, 2020] suggests a simple way to generate a random network with mixing particles.

At each time n, we choose (randomly) a partition of $\{1,\dots,N\}$ with size P, $(2 \le P \le N)$: for $m = \lceil N/P \rceil$,

$$\{1,\dots,N\} = \mathcal{B}_1^n \cup \cdots \cup \mathcal{B}_m^n, \quad |\mathcal{B}_\ell^n| = P \quad \text{for} \quad \ell < m \quad \text{and} \quad |\mathcal{B}_m^n| \le P.$$

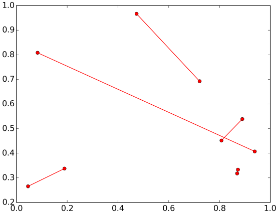


Figure: Pairing from 10 particles ($P = 2$)

random network and the ergodicity

From RBM, we expect that

$$A((s+m,s]) := A(s+m-1)A(s+m-2)\dots A(s) \qquad (m>0)$$

has a positive ergodicity with a high probability if m is large enough.

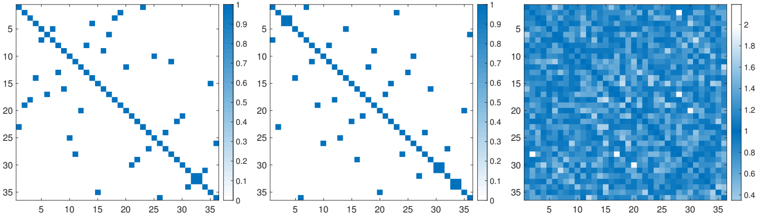


Figure: (Left two) examples of random networks from RBM ($P=2, N=36$), (Right) an average of independent 400 matrices.

Lemma: positive ergodicity of RBM

Assume that for any i and j, there exists one batch at a time $t \in [s, s+m)$ containing i and j. Then, $\alpha(A((s+m,s])) \geq \gamma(1-\gamma)^m$.

Consensus with RBM

Extending the result of [Askari-Sichani–Jalili 2013], we may prove that RBM also guarantees the consensus almost surely.

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

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

where $\mathcal{G}_{s,m}$ is a boolean random variable that becomes 1 if there exists one batch at a time $t \in [(s-1)m, sm)$ containing both i and j. Then,

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

Therefore, we conclude the decay of the diameter

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

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

In other words, the algorithm satisfy consensus.

Table of Contents

Stochastic dynamics; Geometric Brownian Motion

Geometric Brownian motion: When the noise differs by the state values, particle trajectory may converge to the equilibrium.

Let S_t be the process following

$$\begin{aligned} dS_t &= -\mu S_t dt + \sigma S_t dB_t, \quad \mu, \sigma \in \mathbb{R}, \\ &= S_t(-\mu dt + \sigma dB_t). \end{aligned}$$

Then, from dS_t/S_t, S_t has an explicit solution;

$$S_t = S_0 \exp\left(-(\mu + \sigma^2/2)t + \sigma B_t\right).$$

Therefore, as $t \to \infty$, S_t tends to zero.

“Multiplicative noise may result convergence $(\mu dt \gtrsim \sigma dB_t)$**."**

Main result: Consensus of a general CBO algorithm

Multiplicative noise, whole domain, linear reverting drift:
$\Rightarrow$ **almost sure convergence.**

1. There are many interacting particles.
2. The consensus, i.e., the decay of relative positions.
3. The discrete-time dynamics requires analysis of ergodicity.

Algorithm 1 [K.–Ha–Jin–Kim 2022]

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

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

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

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 of information at i:

$$\bar{X}^{i,*}_t := \sum_{j \in N_i(t)} f_{ij}(t) X^j_t : \text{ convex combination}, \quad \sum_{j \in N_i(t)} f_{ij}(t) = 1, \quad \forall i.$$

Theorem (K.–Ha–Jin–Kim 2022)

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

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

Ergodicity argument?

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

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 elements and negative ergodicity.

Ergodicity coefficient revisited

For a real vector $x = (x_1, \dots, x_N) \in \mathbb{R}^N$, define

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

For a matrix $P = (p_{ij})$ with "each entry$\in \mathbb{R}$, each row sum$= 1$", define its ergodicity coefficient as

$$\alpha(P) := \min_{i,j} \sum_k \min\{p_{ik}, p_{jk}\} \in (-\infty, 1].$$

Note: $\alpha_1(P) = 1 \Leftrightarrow$ all rows of P are same.

Proposition [Alpin–Gabassov 1976]

$$\mathcal{D}(Px) \leq (1 - \alpha(P))\mathcal{D}(x).$$

Reformulation of the algorithm

We may reformulate the dynamics

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

into the matrix form:

$$\mathrm{x}^l_{t+1} = \left[(1-\gamma)I_N + \gamma C_t - H^\ell_t(I_N - C_t)\right]\mathrm{x}^\ell_t, \quad 1 \le \ell \le d,\ t \ge 0,$$

with $C_t := (f_{ij}(X^1_t, \dots, X^N_t))_{1\le i,j\le N}$, $\mathrm{x}^\ell_t := (x^{1,\ell}_t, \dots, x^{N,\ell}_t)^\top$, and

$$H^\ell_n := \mathrm{diag}(\eta^{1,\ell}_n, \dots, \eta^{N,\ell}_n).$$

The basic idea is the same as dynamics without noise.

$$\mathcal{D}(\mathrm{x}^\ell_{n+1}) \le [1 - \alpha(\text{product of } n - n_0 + 1 \text{ matrices})]\,\mathcal{D}(\mathrm{x}^\ell_{n_0}) < \mathcal{D}(\mathrm{x}^\ell_{n_0}).$$

Proof of the Main Result

From the analysis without noise, we may set

$$A(t) := (1-\gamma) I_N + \gamma C_t,$$

then the ergodicity of A is the same as before:

$$\alpha\Big(A(s+m-1)A(s+m-2)\dots A(s)\Big) \\ \geq \alpha\left(\gamma(1-\gamma)^{m-1}\sum_{r=s}^{s+m-1} C_r\right) \quad \geq \gamma(1-\gamma)^{m-1}\mathcal{G}_{s,m}.$$

Again, $\mathcal{G}_{s,m}$ is 1 if the network is mixed enough, and 0 otherwise.

In the noisy case, the transition matrix is no more $A(t)$ but $A_\varepsilon(t)$,

$$A_\varepsilon(t) = \big[(1-\gamma)I_N + \gamma C_s - H_s^\ell(I_N - C_s)\big] =: A(s) + B(s).$$

Our strategy is to consider $B(s) := -H_s^\ell(I_N - C_s)$ as error terms.

Properties of ergodicity coefficient

Lemma (sub-additivity and worst-case estimation)

1. $\alpha(A+B) \geq \alpha(A) + \alpha(B)$,
2. $\alpha(B) \geq -\|B\|_{1,\infty}$, where $\|B\|_{1,\infty} := \max_{1\leq i\leq N} \sum_{j=1}^{N} |a_{ij}|$.

For example,

$$\begin{aligned} &\alpha\left[A(s) + B(s)\right] \\ &\quad \geq \alpha(A(s)) + \alpha(B(s)) \\ &\quad \geq \alpha(A(s)) - \|B\|_{1,\infty} \\ &\quad = \alpha(A(s)) - \| - H_s^\ell(I_N - C_s)\|_{1,\infty} \\ &\quad \geq \alpha(A(s)) - 2\|H_s^\ell\|_{1,\infty}. \end{aligned}$$

Therefore, small noise $\varepsilon(s) := 2\|H_s^\ell\|_{1,\infty}$ could be absorbed to $\alpha(A(s))$.

Proof of the Main Result

Note that $\alpha(A(s))$ is usually zero, so we consider a series as before. We expand the matrix multiplications to get

$$
\begin{aligned}
&\alpha\big([A(s+m-1)+B(s+m-1)][A(s+m-2)+B(s+m-2)]\dots[A(s)+B(s)]\big)\\
&\geq \gamma(1-\gamma)^{m-\mathbf{1}}\mathcal{G}_{s,m}-2\left[(1+\varepsilon(s+m-1))\cdots(1+\varepsilon(s))-1\right], \quad \varepsilon(s):=2\|H_s^\ell\|_{\mathbf{1},\infty}.
\end{aligned}
$$

Note that the error term vanishes as the noise strength goes zero,

$$
\max_{t,i,\ell}|\eta_t^{i,\ell}|^2 \to 0.
$$

Hence, if the **variance of noise** is small enough, then it seems that

$$
\alpha(\text{product of the matrices}) > 0.
$$

A rigorous proof needs the law of large numbers.

Proof of the Main result

For $n \in [km, (k+1)m)$, we conclude

$$
\begin{aligned}
\mathcal{D}(x_n^\ell) &\le \exp\Bigg[-\gamma(1-\gamma)^{m-1} \sum_{j=1}^{k} \mathcal{G}_{(j-1)m,jm} \\
&\quad + \sum_{j=1}^{k+1} 2\Big[\Big(1+2\|H_{jm-1}^\ell\|_\infty\Big)\cdots\Big(1+2\|H_{(j-1)m}^\ell\|_\infty\Big) - 1\Big]\Bigg]\mathcal{D}(x_0^\ell) \\
&= \exp\Bigg[-\gamma(1-\gamma)^{m-1} \left(\frac{1}{k}\sum_{j=1}^{k} \mathcal{G}_{(j-1)m,jm}\right)\cdot\frac{k}{n}\cdot n \\
&\quad + \left(\frac{1}{k+1}\sum_{j=1}^{k+1} 2\Big[\Big(1+2\|H_{jm-1}^\ell\|_\infty\Big)\cdots\Big(1+2\|H_{(j-1)m}^\ell\|_\infty\Big) - 1\Big]\right)\cdot\frac{k+1}{n}\cdot n\Bigg]\mathcal{D}(x_0^l).
\end{aligned}
$$

Now we can use the strong law of large numbers to estimate the average values in the decay constant to conclude the result:

$$\mathcal{D}(x_n^l) \text{ exponentially decays almost surely.}$$

Optimality argument for CBO algorithm

We have proved that the CBO algorithm terminates with exponential convergence speed.

Our remaining question is the **optimality** of the final result.

Unfortunately, we only have little information about it.

- In the kinetic level, the optimality is guaranteed if ρ_0 is nonzero around the global minimum x_*.
- However, this implies that we already evaluate the global minimum by the current candidate particles.
- Since $m[\rho_t] \sim x_*$ is critically used, this makes critical difficulty.

Monotonicity of CBO algorithm

A partial optimality result can be written when $\bar{X}_t^{i,*}$ is the argument minimum of L:

$$\begin{cases} X_{t+1}^i = X_t^i + (\gamma I + \mathrm{diag}(\eta_t^{i,1}, \ldots \eta_t^{i,d}))(\bar{X}_t^{i,*} - X_t^i), \quad i = 1, \cdots, N, \\ \bar{X}_t^{i,*} := X_t^k \quad \text{with } k = \min \mathrm{argmin}_{j \in [i]_t} L(X_t^j). \end{cases}$$

Proposition (Monotonicity of the optimum candidate)

For all $n \geq 0$, we have

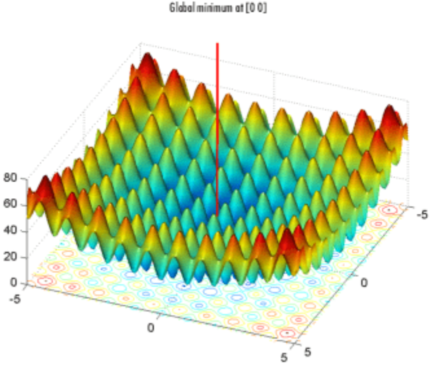
$$\min_{1 \leq j \leq N} L(X_{n+1}^j) \leq \min_{1 \leq j \leq N} L(X_n^j).$$

It guarantees that our guesses on minimizer get improved along time.

Numerical simulations

- In principle, the particles' initial positions should surround the global minimum.
- In practice, the simulation on Rastrigin functions suggests high probabilities to find the minimum, more than 88%:

$$N = 100, \gamma = 0.01, \zeta = 0.5, P = 10, d = 2, \dots, 10.$$



Numerical simulations

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

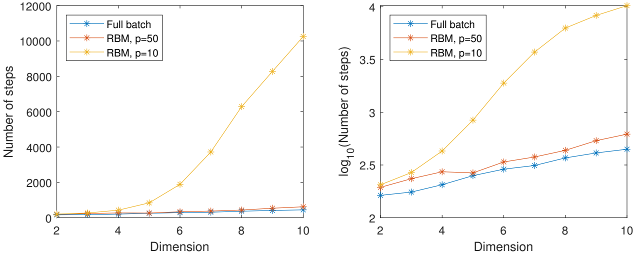


Figure: Average number of steps until stopping criterion holds.

The stopping criterion is made with the change of positions,

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

Table of Contents

Summary and remaining questions

Remarks

1. **The convergence of CBO algorithm holds with multiplicative and heterogeneous noise.**
2. The particles searches the minimizer along noisy sample paths in initial convex hull with randomized exploration direction.
3. No performance estimates to find the global minimizer.
4. The number of steps differ from the batch size P and dimension d, however, there is no clear explanation on it.

Future directions

1. Optimality of kinetic CBO dynamics when the initial data does not contains the global minimizer. (for example, an annulus)
2. Optimality of particle CBO dynamics in a simple situation, for example, in 1D.

Thank you very much