Consensus-based Non-convex Optimization for High Dimensional Machine Learning Problems

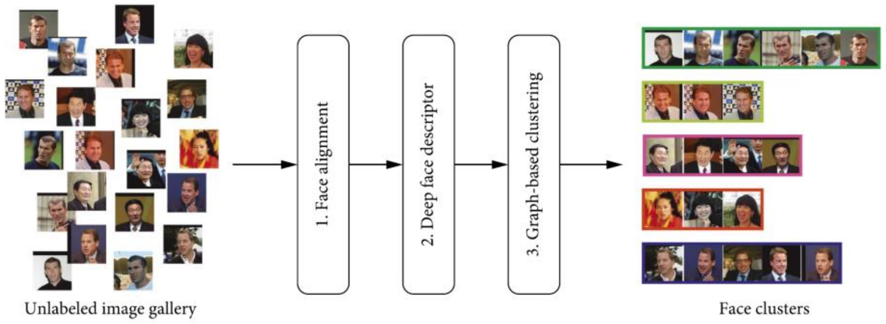
Shi Jin （金石）

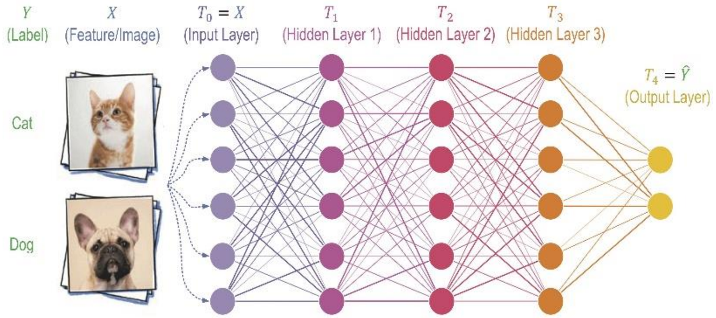
Institute of Natural Sciences and School of Mathematical Sciences

Shanghai Jiao Tong University, China

- In machine learning one seeks to find the global minimum of a loss function (non-convex, high dimensional)

$$\min_{x \in R^d} L(x) = \frac{1}{n}\sum_{i=1}^{n} l(f(x,\hat{x}_i),\hat{y}_i). \qquad x^* = \min_{x \in R^d} L(x)$$





Non-convex optimization is an NP-hard problem!

Gradient or Gradient free

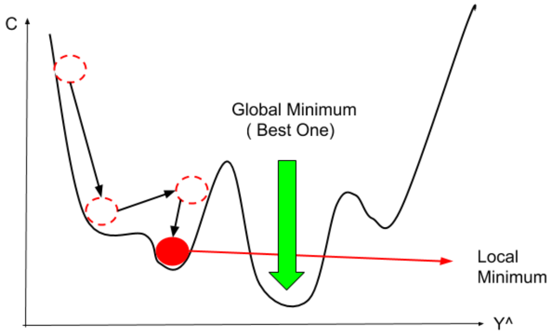
- The most popular method is the stochastic gradient descent method which needs to take the gradient (along a few randomly selected spatial directions at each iteration)
- Often the loss function is not a good function to take its gradient, or the function is known only in discrete set of data
- Alternative gradient-free numerical methods are of great interest

$$f(x^{k+1}) = f(x^k) + \nabla f(x^k)\cdot(x^{k+1} - x^k)$$

steepest descent: $x^{k+1} - x^k = -\nabla f(x^k)$

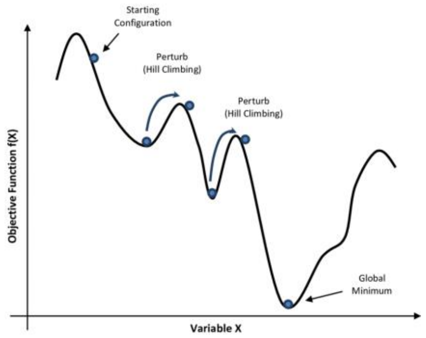
$$\Rightarrow f(x^{k+1}) = f(x^k) - \|\nabla f(x^k)\|^2$$

Gradient descent: $\boxed{x^{k+1} = x^k - \nabla f(x^k)}$

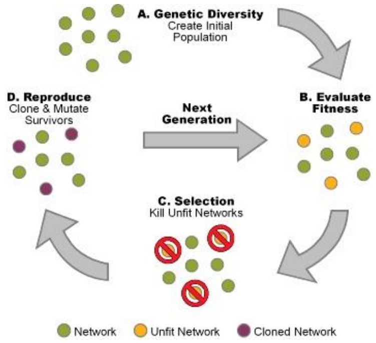


Gradient-free optimization methods: metaheuristics

- Simulated annealing: *Kirkpatrick ('83)*

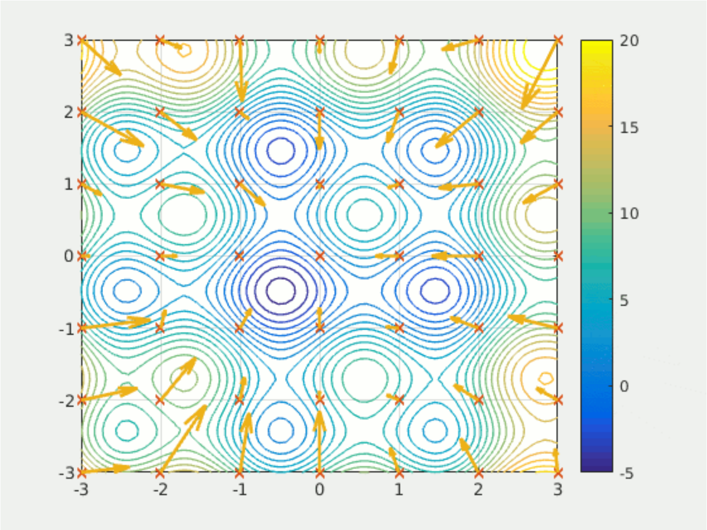


- Genetic algorithms: *Holland ('75)*



Swarming intelligence

a population of simple agents interacting with each other, and the collective behavior exhibits "intelligence" not known by individuals--better way to get out of local extrema compared to simulated annealing



Examples:

particle swarming optimization (PSO):

Kennedy, Eberhart and Shi ('95-'98)

ant colony optimization (ACO):

Moyson Manderick ('88)

artificial bee colony optimization (ABC):

Karaboga ('05)

About metaheuristics: from wikipedia

Most literature on metaheuristics is experimental in nature, describing empirical results based on computer experiments with the algorithms. ⋯ . While the field also features high-quality research, many of the publications have been of poor quality; flaws include vagueness, lack of conceptual elaboration, poor experiments, and ignorance of previous literature.[7]

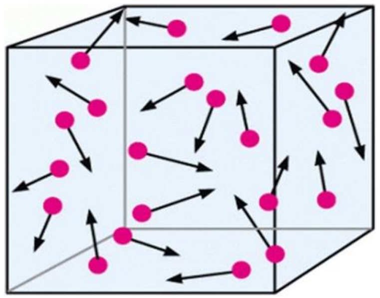
Sörensen, Kenneth (2015). "Metaheuristics—the metaphor exposed" . *International Transactions in Operational Research*. **22**: 3–18.

We develop an interacting particle system with proven convergence toward the global minimum for general non-convex, high dimensional functions

- Pinnau-Totzeck-Tse-Martin (M3AS '17)
- Carrillo-Choi-Totzeck-Tse (M3AS '18)

$$dX^j = -\lambda(X^j - \bar{x}^*)dt + \sigma|X^j - \bar{x}^*|dW^j \qquad j = 1, \cdots, N$$

$$\bar{x}^* = \frac{1}{\sum_j e^{-\beta L(X^j)}} \sum_j X^j e^{-\beta L(X^j)}$$



For sufficiently large β the particles form consensus--converge to the global minimum of L exponentially fast but the drift rate is dimension sensitive!

$$2\lambda > d\sigma^2$$

Laplace principle

for any probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$ compactly supported with $x_* \in \mathrm{supp}(\rho)$, then

$$\lim_{\beta\to\infty}\left(-\frac{1}{\beta}\log\left(\int_{\mathbb{R}^d} e^{-\beta L(x)}d\rho(x)\right)\right) = L(x^*) > 0. \tag{1.5}$$

Therefore, if L attains its minimum at a single point $x^* \in \mathrm{supp}(\rho)$, then the suitably normalized measure $e^{-\beta L(x)}\rho$ assigns most of its mass to a small region around x^* and hence we expect it approximates a Dirac distribution $\delta_{\bar{x}^*}$ for large $\beta \gg 1$. Consequently, the first moment of the normalized measure $e^{-\beta L(x)}\rho$, and thus, the discrete counterpart average $\bar{x}^*$, should provide a good estimate of the point at which the global minimum is attained, $x^* = \mathrm{argmin}\, L$.

Our improvement: a dimension-independent model!

(with J. Carrillo, Oxford; Lei Li, SJTU and Yuhua Zhu, Stanford)

- Use geometric Brownian motion

$$dX^j = -\lambda(X^j - \bar{x}^*)\,dt + \sigma \sum_{k=1}^{d} (X^j - \bar{x}^*)_k dW_k^j \vec{e}_k$$

- Random Batch to compute L:

$$\hat{L}^j(x) = \frac{1}{m}\sum_{i \in b} l_i,$$

where b is a random index subset of $\{1, \cdots, n\}$ containing m elements.

- Random Batch to evaluate : B randomly selected mini-batch

$$x_k^* = \frac{1}{\sum_{j \in B} \mu_j} \sum_{j \in B} X_k^j \mu_j, \quad \text{with} \quad \mu_j = e^{-\beta \hat{L}^j}$$

Heuristics:

Consider the case: $\bar{x}^* = a$

For particles to form a consensus:

PTTM model: $$\frac{d}{dt}\mathbb{E}(X-a)^2 = -2\lambda\mathbb{E}(X-a)^2 + \sigma^2\sum_{i=1}^{d}\mathbb{E}|X-a|^2 = (-2\lambda+\sigma^2 d)\mathbb{E}(X-a)^2$$

$$2\lambda > d\sigma^2.$$

Our model: $$\frac{d}{dt}\mathbb{E}(X-a)^2 = -2\lambda\mathbb{E}(X-a)^2 + \sigma^2\sum_{i=1}^{d}\mathbb{E}(X-a)_i^2 = (-2\lambda+\sigma^2)\mathbb{E}(X-a)^2$$

$$2\lambda > \sigma^2$$

our model is dimension insensitive!

Our improvement: a dimension-independent model!

(with J. Carrillo， Oxford; Lei Li, SJTU and Yuhua Zhu, Stanford)

- Use geometric Brownian motion

$$dX^j = -\lambda(X^j - \bar{x}^*)\,dt + \sigma \sum_{k=1}^{d} (X^j - \bar{x}^*)_k dW_k^j \vec{e}_k$$

- Random Batch to compute L:

$$\hat{L}^j(x) = \frac{1}{m} \sum_{i \in b} l_i,$$

where b is a random index subset of $\{1, \cdots, n\}$ containing m elements.

- Random Batch to evaluate : B randomly selected mini-batch

$$x_k^* = \frac{1}{\sum_{j \in B} \mu_j} \sum_{j \in B} X_k^j \mu_j, \quad \text{with} \quad \mu_j = e^{-\beta \hat{L}^j}$$

Convergence proof

Via mean-field limit: Carrillo-Choi- Totzeck-Tse; Carrillo-Jin-Li-Zhu

Formally, taking $N \to \infty$ in the model (2.2) with full batch (or alternatively, $\gamma \to 0$ and $N \to \infty$ in Algorithm 2.1 with full batch), the mean field limit of the model is formally given by the following stochastic differential equation for $X = X(t)$:

$$dX = -\lambda(X - \bar{x}^*)dt + \sigma \sum_{i=1}^{d} \vec{e}_i (X - \bar{x}^*)_i dW_i, \tag{3.1}$$

where

$$\bar{x}^* = \frac{\mathbb{E}(Xe^{-\beta L(X)})}{\mathbb{E}(e^{-\beta L(X)})}. \tag{3.2}$$

The law $\rho(\cdot, t)$ of the process $X(t)$ follows the nonlinear Fokker-Planck equation

$$\partial_t \rho = \lambda \nabla \cdot ((x - \bar{x}^*)\rho) + \frac{1}{2}\sigma^2 \sum_{i=1}^{d} \partial_{ii}((x - \bar{x}^*)_i^2 \rho) \qquad \bar{x}^* = \frac{\int_{\mathbb{R}^d} xe^{-\beta L(x)}\rho(x,t)\,dx}{\int_{\mathbb{R}^d} e^{-\beta L(x)}\rho(x,t)\,dx}$$

Convergence analysis for fully discrete particle systems (with Seung-yeal Ha, SNU,; Doheon Kim, KIAS)

Next, we consider time-discrete analogue of (1.1). For this, we set

$$h := \Delta t, \quad X_n := X(nh), \quad n = 0, 1, \cdots, \cdots .$$

Then the discrete scheme reads as follows:

$$(1.2) \qquad \begin{cases} X_{n+1}^i = X_n^i - \gamma(X_n^i - \bar{X}_n^*) - \sum_{l=1}^{d} (x_n^{i,l} - \bar{x}_n^{*,l}) \eta_n^l e_l, \quad n \geq 0, \ i = 1, \cdots, N, \\ \bar{X}_n^* = (x_n^{*,1}, \cdots, x_n^{*,d}) := \dfrac{\sum_{j=1}^N X_n^j e^{-\beta L(X_n^j)}}{\sum_{j=1}^N e^{-\beta L(X_n^j)}}, \end{cases}$$

where the random variables $\{\eta_n^l\}_{n,l}$ are i.i.d. with

$$(1.3) \qquad \mathbb{E}[\eta_n^l] = 0, \quad \mathbb{E}[|\eta_n^l|^2] = \zeta^2, \quad n = 1, \cdots, \quad l = 1, \cdots, d.$$

- Different γ and η_n^l correspond to different schemes (explicit, semi-implicit, exponential integrator, etc which leads to different numerical stability condition)

Euler-Maruyama method

- Model A: Consider the first-order Euler type discrete model in [14]:

$$X_{n+1}^i = X_n^i - \lambda h(X_n^i - \bar{X}_n^*) - \sum_{l=1}^{d}(x_n^{i,l} - \bar{x}_n^{*,l})\sigma\sqrt{h}Z_n^l e_l, \quad n \geq 0, \ i = 1, \cdots, N,$$

where the random variables $\{Z_n^l\}_{n,l}$ are i.i.d standard normal distributions, i.e. $Z_n^l \sim \mathcal{N}(0, 1^2)$. If we set

$$\gamma := \lambda h \quad \text{and} \quad \eta_n^l := \sigma\sqrt{h}Z_n^l. \tag{2.3}$$

Then, the above setting clearly satisfies the relations (2.2) with $\zeta = \sigma\sqrt{h}$.

A predictor-corrector method

- Model B: Consider a predictor-corrector type discrete model in [4].

$$
(2.4) \qquad \begin{cases} \hat{X}^i_n = \bar{X}^*_n + e^{-\lambda h}(X^i_n - \bar{X}^*_n), \\ X^i_{n+1} = \hat{X}^i_n - \sum_{l=1}^{d} (\hat{x}^{i,l}_n - \bar{x}^{*,l}_n)\sigma\sqrt{h}Z^l_n e_l, \quad n \geq 0, \ i = 1, \cdots, N. \end{cases}
$$

We substitute $(2.4)_1$ into $(2.4)_2$ and use an addition-subtraction trick to see that

$$
X^i_{n+1} = X^i_n - (1 - e^{-\lambda h})(X^i_n - \bar{X}^*_n) - \sum_{l=1}^{d} (x^{i,l}_n - \bar{x}^{*,l}_n) e^{-\lambda h}\sigma\sqrt{h}Z^l_n e_l, \quad n \geq 0, \ i = 1, \cdots, N.
$$

If we set

$$
(2.5) \qquad \gamma := 1 - e^{-\lambda h} \quad \text{and} \quad \eta^l_n := e^{-\lambda h}\sigma\sqrt{h}Z^l_n,
$$

then (2.4) reduces to the special case of (2.1) - (2.2) with $\zeta = e^{-\lambda h}\sigma\sqrt{h}$.

An exponential integrator method

$$\text{freeze } \bar{x}_k^* \text{ in a time-step interval.}$$

- Model C: Consider one of discrete optimization model proposed in [4]:

$$X_{n+1}^i = \bar{X}_n^* + \sum_{l=1}^d (x_n^{i,l} - \bar{x}_n^{*,l}) \left[\exp\left(-\left(\lambda + \frac{1}{2}\sigma^2 \right) h + \sigma\sqrt{h} Z_n^l \right) \right] e_l, \quad n \geq 0, \ i = 1, \cdots, N, \tag{2.6}$$

Again, the R.H.S. of (2.6) can be rewritten as

$$X_{n+1}^i = X_n^i - (1 - e^{-\lambda h})(X_n^i - \bar{X}_n^*) - \sum_{l=1}^d (x_n^{i,l} - \bar{x}_n^{*,l}) e^{-\lambda h} \left[\exp\left(-\frac{1}{2}\sigma^2 h + \sigma\sqrt{h} Z_n^l \right) - 1 \right] e_l.$$

We set

$$\gamma := 1 - e^{-\lambda h} \quad \text{and} \quad \eta_n^l := e^{-\lambda h} \left[\exp\left(-\frac{1}{2}\sigma^2 h + \sigma\sqrt{h} Z_n^l \right) - 1 \right]. \tag{2.7}$$

Then, we use the elementary facts [7]:

$$X \sim \text{Lognormal}(\alpha, \beta^2) \quad \Rightarrow \quad \mathbb{E}X = e^{\alpha + \frac{\beta^2}{2}} \quad \text{and} \quad \mathbb{E}X^2 = e^{2\alpha + 2\beta^2}$$

to see that (2.7) satisfies moment relations (2.2) with $\zeta = e^{-\lambda h}\sqrt{e^{\sigma^2 h} - 1}$.

- (Question A): Does the N-state ensemble $\{X_n^i\}$ exhibit a global consensus? i.e., does
$$X_n^i - X_n^j \to 0 \quad \text{as} \quad n \to \infty, \quad i,j = 1, \cdots, N \quad \text{in suitable sense?}$$
- (Question B): If the answer to the first problem is positive, then under what conditions on system parameters and initial data, does there exist a global consensus state X_∞ such that
$$X_n^i \to X_\infty \quad \text{for all } i, \quad \text{as } n \to \infty, \quad \text{such that } L(X_\infty) \sim \min_X L(X).$$

Emergence of global consensus:

Answer to question A

Theorem 2.1. *Let $\{\mathcal{X}_n\}$ be a solution process to (2.1). Then, the following three global consensus results hold.*

(1) Suppose that system parameters satisfy

$$|\gamma - 1| < 1 \quad \text{and} \quad 0 \le \zeta \le \infty.$$

Then, $\mathbb{E}[X_n^i - X_n^j]$ tends to zero asymptotically:

$$\lim_{n\to\infty} \mathbb{E}[X_n^i - X_n^j] = 0, \quad \forall\, i,j = 1,\cdots,N.$$

(2) Suppose that system parameters γ and ζ satisfy

$$(\gamma-1)^2 + \zeta^2 < 1$$

then, L^2 and almost-sure global consensus emerge asymptotically: for a.s. $\omega \in \Omega$,

$$\lim_{n\to\infty} \mathbb{E}|X_n^i - X_n^j|^2 = 0, \quad |x_n^{i,l} - x_n^{j,l}|^2 \le |x_0^{i,l} - x_0^{j,l}|^2 e^{-nY_n^l(\omega)}, \quad i,j = 1,\cdots,N, \; l = 1,\cdots,d,$$

where Y_n^l is a random variable satisfying

$$\lim_{n\to\infty} Y_n^l(\omega) = 1 - (\gamma-1)^2 - \zeta^2 > 0, \quad a.s.\ \omega \in \Omega, \quad l = 1,\cdots,d.$$

For the three specific numerical models:

Corollary 2.1. *The following assertions hold.*

(1) Suppose that system parameters satisfy

$$\lambda > \frac{\sigma^2}{2}, \quad 0 < h < \frac{2\lambda - \sigma^2}{\lambda^2},$$

then, Model A admits L^2 *and almost sure global consensus.*

(2) Suppose that system parameters satisfy

$$(1 + \sigma^2 h)e^{-2\lambda h} < 1,$$

then, Model B admits L^2 *and almost sure global consensus. .*

(3) Suppose that system parameters satisfy

$$\lambda > \frac{\sigma^2}{2}, \tag{2.13}$$

then, Model C admits L^2 *and almost sure global consensus, for any* $h > 0$.

- Remark: Models B and C are unconditional!

Main idea of proof:

- Previous work of Carrillo-Choi-Totzeck-Tse used L_2 norm of the particles thus obtain exponential growing term using Granwall inequality
- We estimate the diameter:

$$\mathcal{D}(\mathcal{X}_t) := \max_{1\le i,j\le N} |X_t^i - X_t^j| \qquad \mathcal{X}_t := (X_t^1, \cdots, X_t^N) \in \mathbb{R}^{Nd}$$

Convergence analysis and error estimates:
Answer to Question B

Consensus does not mean particles approach a fixed a common fixed state X_∞

- (Q1): What is a sufficient framework leading to the common asymptotic state:
$$X_n^i(\beta) \to X_\infty(\beta), \quad \text{as } n \to \infty \text{ for all } i = 1, \cdots, N?$$

- (Q2): If the above question is resolved, then how close is the asymptotic state X_∞ to the global minimum X_m of L if the latter exists?

Emergence of a common consensus

Theorem 3.1. *Suppose that system parameters satisfy*

$$(1-\gamma)^2 + \zeta^2 < 1,$$

and let $\{X_n^i\}_{1\le i\le N}$ *be a solution to* (2.1). *Then, there exists a common constant state* $X_\infty = (x_\infty^1, \cdots, x_\infty^d)$ *such that*

$$\lim_{n\to\infty} X_n^i = X_\infty \ a.s., \ 1 \le i \le N.$$

Error estimates

Some assumptions

- $(\mathcal{A}1)$: Let $L = L(x)$ be a C^2-objective function satisfying the following relations:

$$L_m := \min_{x \in \mathbb{R}^d} L(x) > 0 \quad \text{and} \quad C_L := \sup_{x \in \mathbb{R}^d} \|\nabla^2 L(x)\|_2 < \infty,$$

  where $\|\cdot\|_2$ denotes the spectral norm.

- $(\mathcal{A}2)$: Let X_* be the unique global minimum point of L in $\mathbb{R}^d$ satisfying the local convexity relation:

$$\det\left(\nabla^2 L(X_*)\right) > 0.$$

- $(\mathcal{A}3)$: Let X_{in} be a reference random variable with a law which is absolutely continuous with respect to the Lebesgue measure, and let f be the probability density function of X_{in} satisfying the following conditions:

$$f \text{ is compactly supported, continuous at } X_*, \text{ and } f(X_*) > 0.$$

Answer to question B

Theorem 3.2. *Suppose that the framework* $(\mathcal{A}1)-(\mathcal{A}3)$ *holds, and system parameters* β, γ, ζ *and the initial data* $\{X_0^i\}$ *satisfy*

$$\begin{aligned} &\beta > 0, \quad (\gamma-1)^2+\zeta^2<1, \quad X_0^i : i,i.d, \quad X_0^i \sim X_{in}, \\ &(1-\varepsilon)\mathbb{E}\Big[e^{-\beta L(X_{in})}\Big] \\ &\quad \geq \frac{2C_L\sqrt{\big(1+(1-\gamma)^2+\zeta^2\big)\big(\gamma^2+\zeta^2\big)}\beta e^{-\beta L_m}}{1-e^{-[1-(\gamma-1)^2-\zeta^2]}} \sum_{l=1}^{d}\left(\mathbb{E}\max_{1\le i\le N}(x_0^{i,l}-\bar{x}_0^l)^2\right), \end{aligned} \tag{3.3}$$

for some $0<\varepsilon<1$*. Then for a solution* $\{X_n^i\}_{1\le i\le N}$ *to* (1.1)*, one has the following error estimate:*

$$\left|\operatorname*{essinf}_{\omega\in\Omega} L(X_\infty) - L(X_*)\right| \le \frac{d}{2}\frac{\log\beta}{\beta} + E(\beta), \tag{3.4}$$

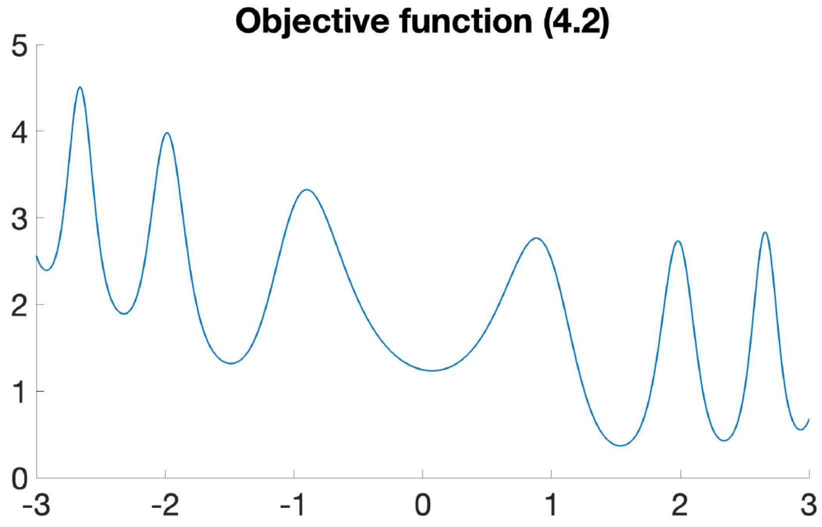
for some function $E(\beta)=\mathcal{O}\left(\frac{1}{\beta}\right)$.

- Remarks: 1) error proportional to $\frac{d}{2}\frac{\log\beta}{\beta}$
  2) initial data quite restrictive-close to X_* and support of initial distribution contains X_*

- Convergence analysis can even include Random Batch approximation: Ko-Ha-Jin-Kim (M3AS to appear)

An example

$$\ell(x, \hat{x}_i) = e^{\sin(2x^2)} + \frac{1}{10}(x - \hat{x}_i - \frac{\pi}{2})^2, \quad \hat{x}_i \sim N(0, 0.1)$$
$$L(x) = \frac{1}{n}\sum_i \ell(x, \hat{x}_i)$$



- SGD

$$x_{k+1} = x_k - \frac{1}{m}\sum_{i\in b_k} \nabla_x \ell(x_k, \hat{x}_i),$$

$$\gamma = 0.01, \quad m = 10^4, \quad n = 20$$

- CBO

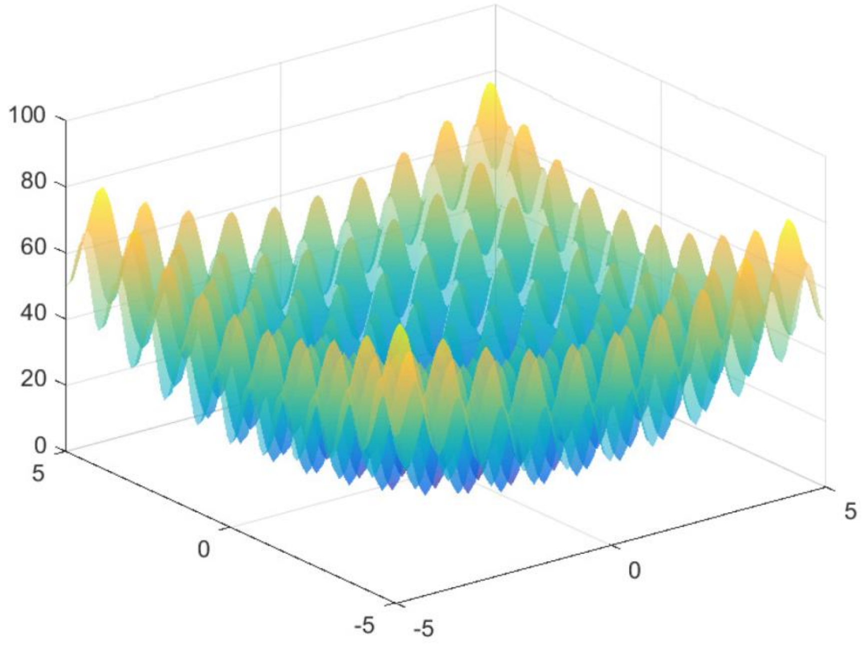
$$N = 100, \quad , M = 20, \quad \sigma = 5, \beta = 30,$$

	SGD	Algorithm 2.1
Success rate	18%	98%

Rastrigin function of 20 dimensions:

$$L(x) = \frac{1}{d}\sum_{i=1}^{d}\left[(x_i - B)^2 - 10\cos\left(2\pi(x_i - B)\right) + 10\right] + C,$$

$$B = \text{argmin } L(x), \quad C = \min L(x).$$



$d = 2, B = C = 0$

PTTM algorithm

TABLE 2. Rastrigin function in $d = 20$ with $\alpha = 30$.

x_*		N = 50	100	200
0	success rate	34.%	61.1%	62.2%
	$\frac{1}{d}\mathbb{E}[\|v_f(T) - x_*\|^2]$	$3.12e^{-1}$	$2.47e^{-1}$	$2.42e^{-1}$
1	success rate	34.5%	57.1%	61.6%
	$\frac{1}{d}\mathbb{E}[\|v_f(T) - x_*\|^2]$	$3.09e^{-1}$	$2.52e^{-1}$	$0.244e^{-1}$
2	success rate	35.5%	54.8%	62.4%
	$\frac{1}{d}\mathbb{E}[\|v_f(T) - x_*\|^2]$	$3.06e^{-1}$	$2.51e^{-1}$	$2.44e^{-1}$

our algorithm

Rastrigin function in $d = 20$ with $\alpha = 30$

	N = 50, M = 40 $\sigma = 5.15$	**N = 100, M = 70** $\sigma = 5.1$	**N = 200, M = 100** $\sigma = 5.1$
x* = 0, success rate	98%	99%	98%
x* = 0, $\frac{1}{d}\mathbb{E}\left[\|x_T^* - x^*\|^2\right]$	6.13E-04	5.03E-04	9.71E-04
x* = 1, success rate	98%	99%	95%
x* = 1, $\frac{1}{d}\mathbb{E}\left[\|x_T^* - x^*\|^2\right]$	1E-03	4.95E-04	3E-03
x* = 2, success rate	95%	100%	92%
x* = 2, $\frac{1}{d}\mathbb{E}\left[\|x_T^* - x^*\|^2\right]$	2.6E-03	8.06E-06	4E-03
Computing time saved	22.03%	30.11%	36.14%

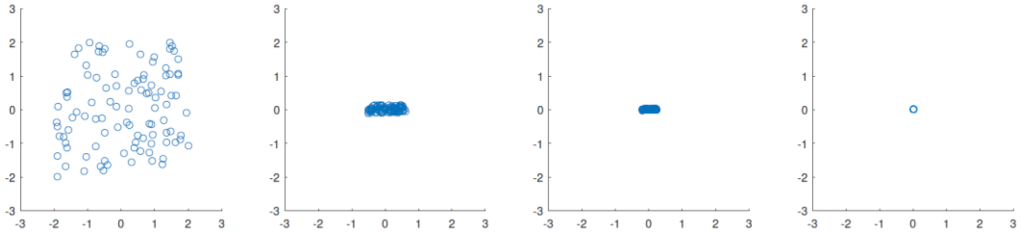


FIGURE 3. Temporal evolution of state configuration for $t = 0, 1, 2, 10$ $(\sigma = 1)$.

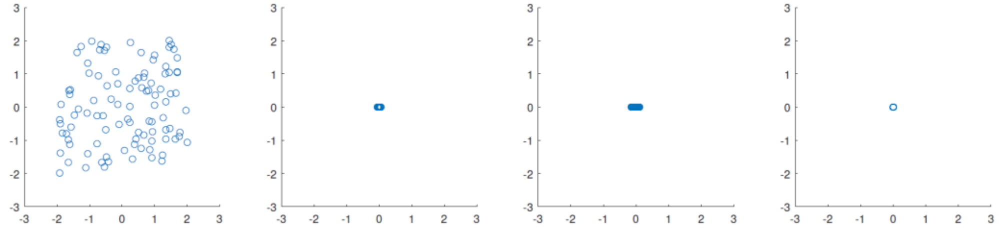


FIGURE 4. Temporal evolution of state configuration for $t = 0, 1, 2, 10$ $(\sigma = 2)$.

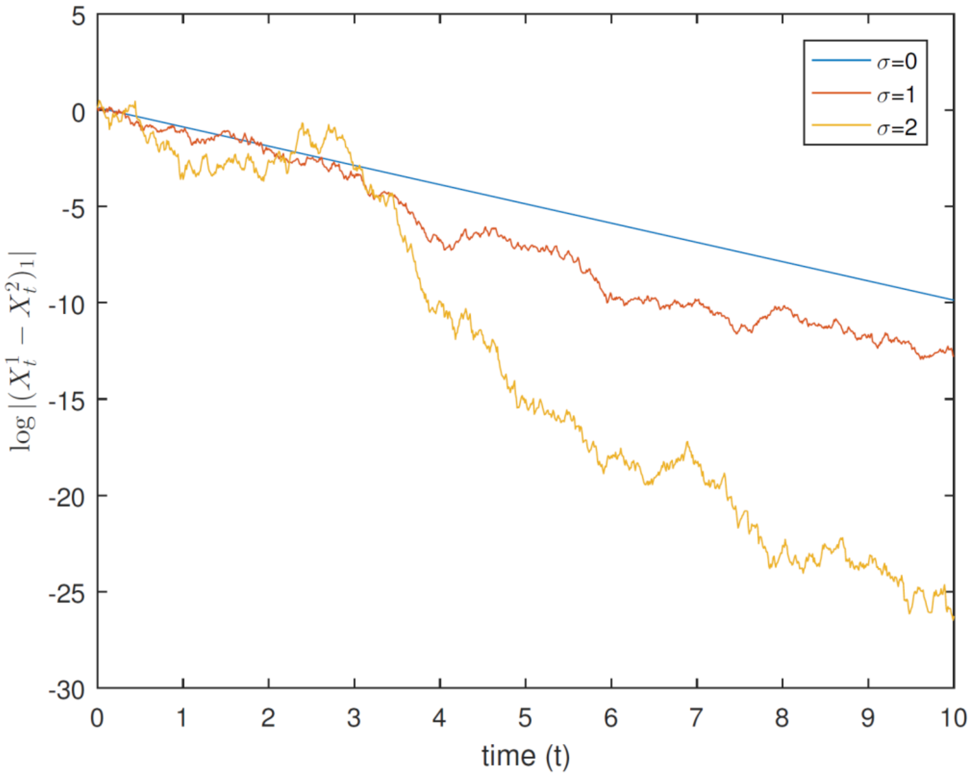


FIGURE 5. Graph of $\log|x_t^{1,1} - x_t^{2,1}|$ for $\sigma = 0, 1, 2$.

MNIST dataset

The MNIST data is a set of pictures for numbers from 0 to 9. The input data is a vector of dimension 728, it records the Grayscale of each pixel. We use the Neural Network without hidden layer to model this classification problem,

$$f(w,x) = a(ReLu(\theta x + B)), \quad w = (\theta, B),$$

where $x_j \in \mathbb{R}^{728}, \theta \in \mathbb{R}^{10\times 728}, B \in \mathbb{R}^{10}$. $ReLu(x) = x\mathbb{1}_{x\geq 0}$ is an activation function, while $a(x)$ is an activation function called *softmax*, which reads,

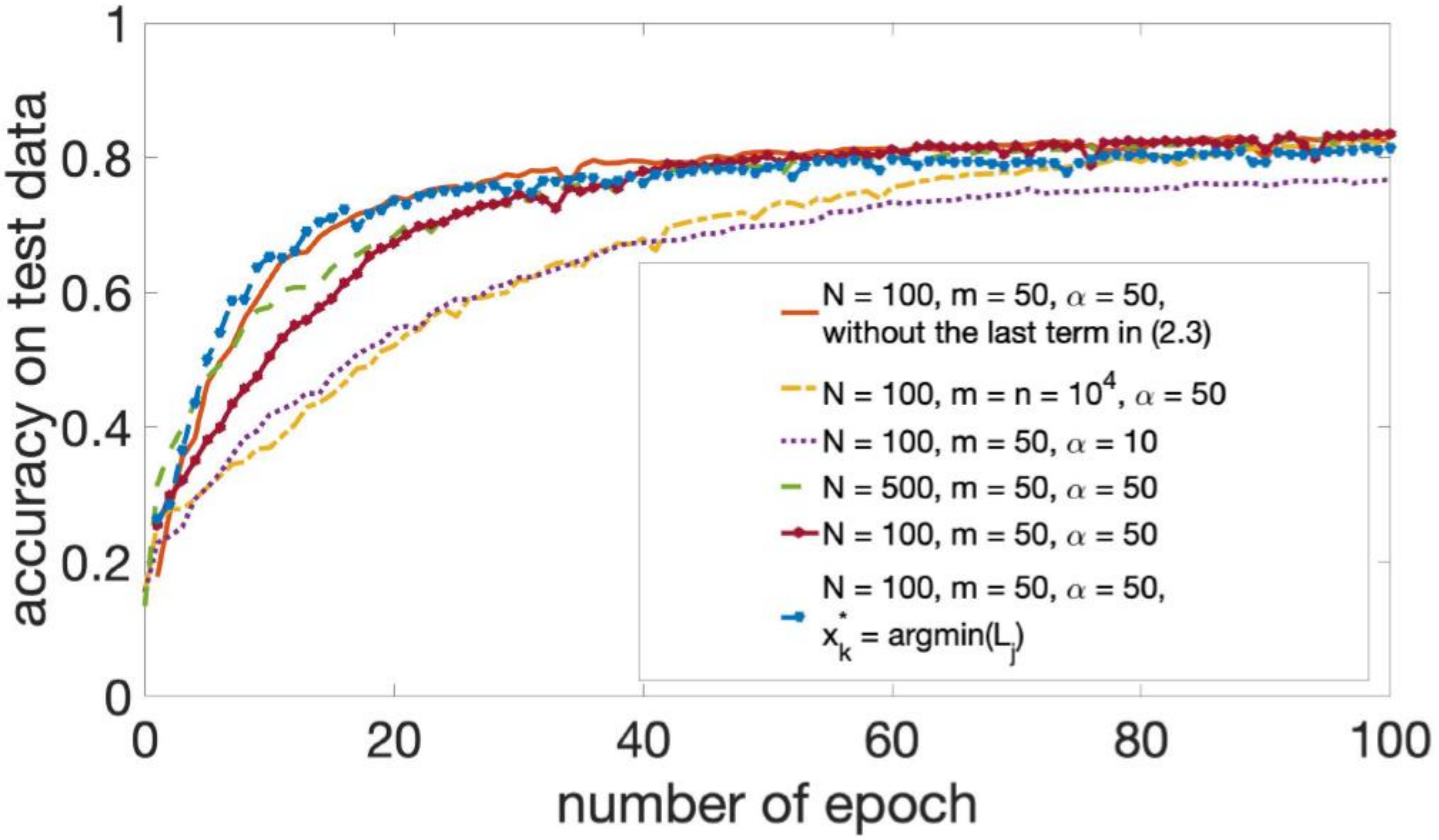
$$a(\mathbf{x}) = \frac{e^{x_j}}{\sum_j e^{x_j}}.$$

The objective function to be minimized is the following,

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} l(f(w,x_i),y_i), \quad l(f,y) = -\sum_{j=1}^{10} y_j \log(f_j), \tag{4.2}$$

where $y \in \{e_j\}_{j=1}^{10}$ is a vector of dimension 10 with only the j-th element 1.

$$N = 100, \quad M = 10, \quad n = 10^4, \quad m = 50, \quad \gamma = 0.1, \quad \sigma = \sqrt{0.1}, \quad \lambda = 1 \quad p = 10^4$$



N=1000

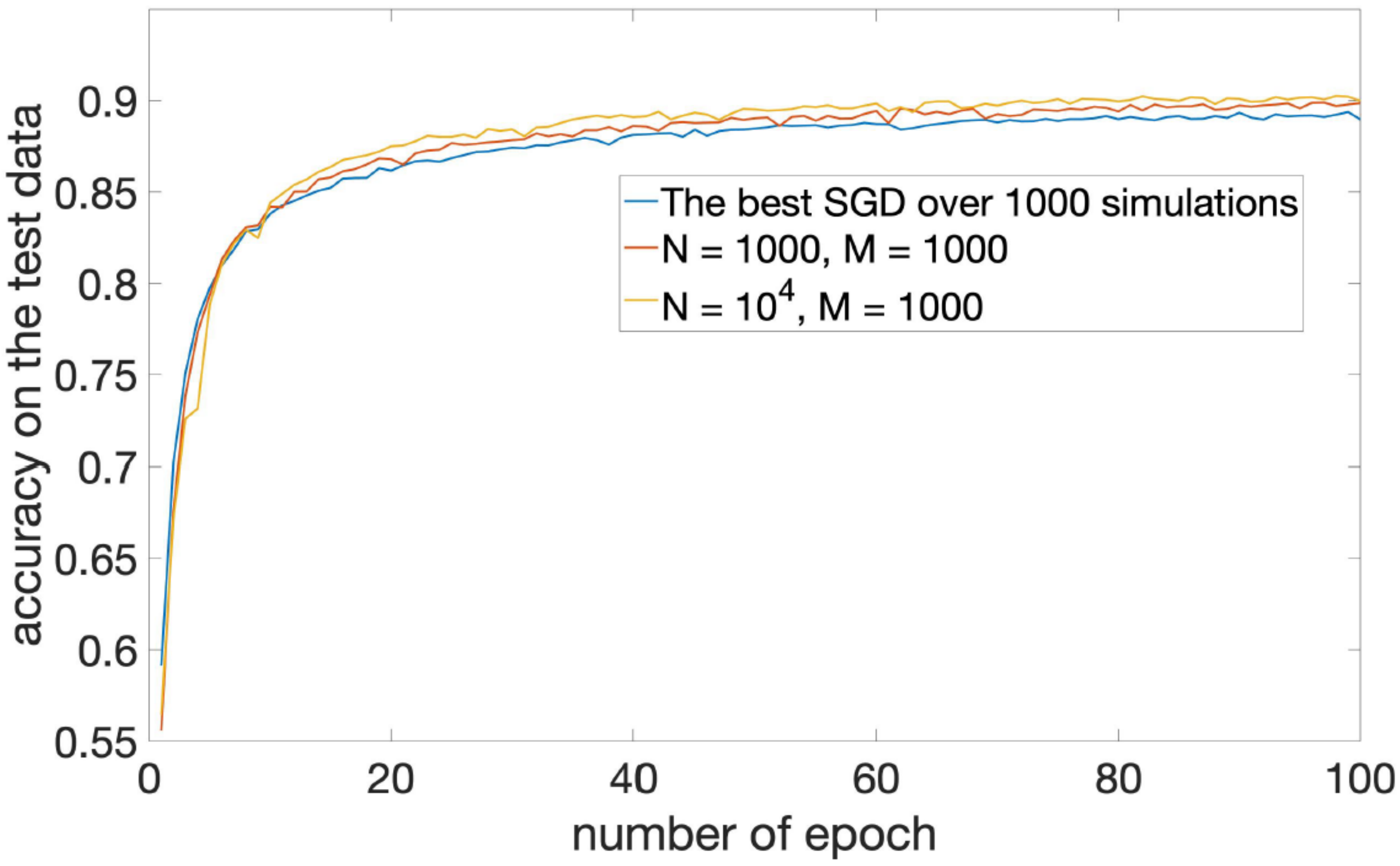


Figure 7: Comparison of our new CBO algorithm and SGD.

Some popular methods used in machine learning optimization

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$$

$$\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)$$

- GD

$$\theta^{t+1} = \theta^t - m^t,$$
$$m^t = \gamma m^{t-1} + \alpha \nabla_\theta \hat{f}(\theta^t)$$

- momentum

$$\theta^{t+1} = \theta^t - \gamma \frac{\hat{m}^t}{\sqrt{\hat{v}^t} + \epsilon},$$
$$m^t = \beta_1 m^{t-1} + (1-\beta_1)\nabla \hat{f}_\theta(\theta^t), \quad \hat{m}_t = \frac{m_t}{1-\beta_1^t},$$
$$v^t = \beta_2 v^{t-1} + (1-\beta_2)(\nabla_\theta \hat{f}(\theta^t))^2, \quad \hat{v}_t = \frac{v_t}{1-\beta_1^t},$$

- Adam

$0 < \beta_1, \beta_2 < 1$

CBO-adaptive momentum estimation method (CBO-Adam)
-- joint with *Jingrui Chen, Liyao Lv*

$$M^i_{t+1} = \beta_1 M^i_t + (1-\beta_1)(X^i_t - x^*) \qquad \hat{M}^i_{t+1} = M^i_{t+1}/(1-\beta_1^t);$$
$$V^i_{t+1} = \beta_2 V^i_t + (1-\beta_2)(X^i_t - x^*)^2 \qquad \hat{V}^i_{t+1} = V^i_{t+1}/(1-\beta_2^t);$$
$$X^i_{t+1} = X^i_t - \lambda \frac{\hat{M}^i_{t+1}}{\sqrt{\hat{V}^i_{t+1}+\epsilon}} + \sigma^t \sum_{k=1}^d \vec{e}_k z_i \quad z_i \text{ is a random variable.}$$

Linear stability shows that the dynamical system converges to the global equilibrium with the rate β_1 if

$$\frac{\mu-2}{\mu+1} < \beta_1 < \frac{\mu}{\mu+1} \qquad \mu = \frac{\lambda}{\epsilon}.$$

The Rastrigin function

d	N	M	Adam-CBO		N	M	Adam-CBO	
			$\mathcal{N}(0,1)$	$\mathcal{U}(-1,1)$			$\mathcal{N}(0,1)$	$\mathcal{U}(-1,1)$
100	1000	5	87%	39%	5000	5	100%	84%
100	1000	10	94%	60%	5000	10	100%	100%
100	1000	20	87%	49%	5000	20	100%	100%
100	1000	25	77%	53%	5000	25	100%	100%
100	1000	50	45%	8%	5000	50	100%	100%
100	1000	100	2%	0%	5000	100	100%	100%

TABLE 2. Comparison of success rates for different batch numbers when the dimension is 100, $\lambda = 0.1$, and $\sigma^t = 0.99^{\frac{t}{20}}$.

d	N	M	Adam-CBO	
			$\mathcal{N}(0,1)$	$\mathcal{U}(-1,1)$
1000	8000	50	92%	20%
1000	10000	50	100%	28%
1000	12000	50	100%	28%
1000	14000	50	100%	32%
1000	16000	50	100%	32%

TABLE 3. Comparison of success rates for different numbers of particles when the dimension is 1000, $\lambda = 0.1$, and $\sigma^t = 0.99^{\frac{t}{20}}$.

Solving PDEs with low regularity use Deep-Ritz (E and Yu)

$$\begin{cases} -\nabla \cdot (A(x)\nabla u) = -\sum_{i=1}^{d} \delta(x_i) & x \in \Omega = [-1,1]^d \\ u(x) = g(x) & x \in \partial\Omega \end{cases}$$

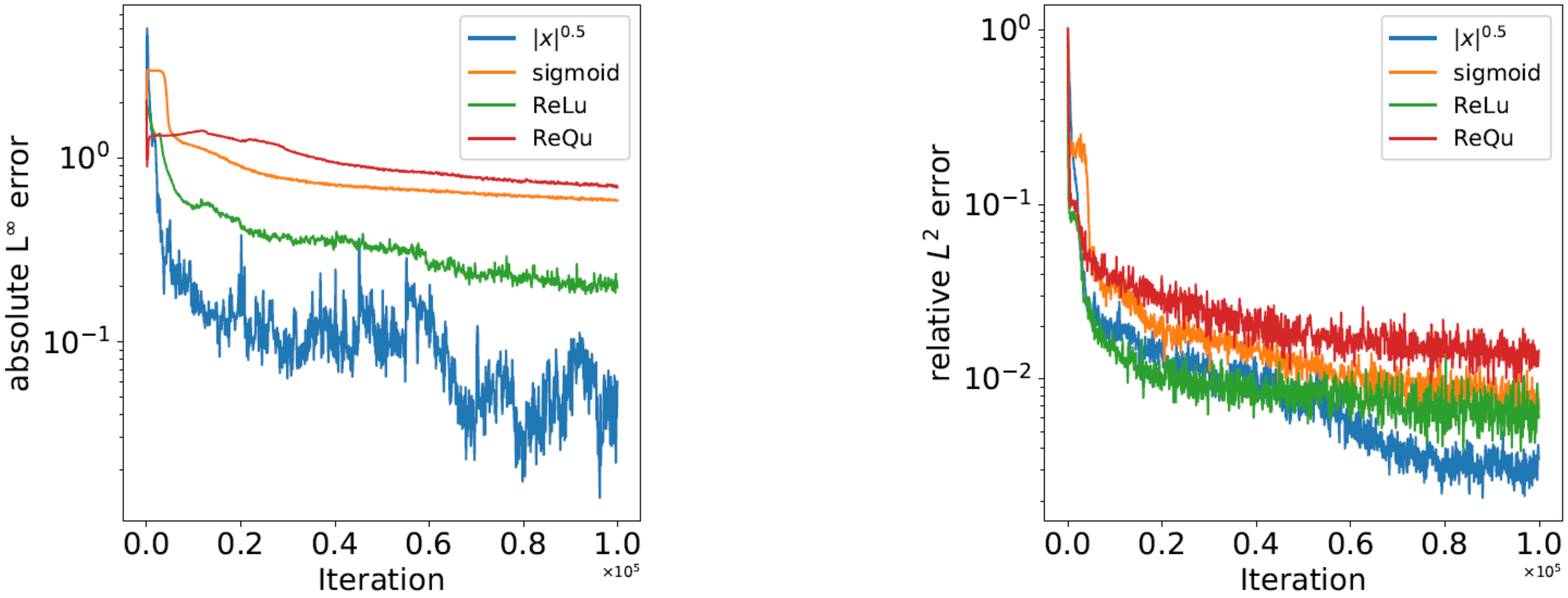
with

(38)
$$A(x) = \begin{bmatrix} (x_1^2)^{\frac{1}{4}} & & \\ & \ddots & \\ & & (x_d^2)^{\frac{1}{4}} \end{bmatrix}.$$

The exact solution $u(x) = \sum_{i=1}^{d} |x_i|^{\frac{1}{2}}$. One can see that the solution is only in $H^{1/2}(\Omega)$ and has singularities when evaluating its derivative at $x_i = 0$. The loss function in DRM reads as

(39)
$$I[u] = \int_\Omega \frac{1}{2}(\nabla u)^T A(x)\nabla u(x)\mathrm{d}x + \sum_{i=1}^{d}\int_{-1}^{1} \delta(x_i)u(x)\mathrm{d}x_i + \eta\int_{\partial\Omega}(u(x)-g(x))^2\mathrm{d}x,$$

where $\eta = 500$ is the penalty parameter for the boundary condition.



(a) L^∞ error

(b) L^2 error

FIGURE 6. Training process of Adam and Adam-CBO methods for (37) when the dimension is 4. (a) L^∞ error; (b) L^2 error.

Conclusions

- gradient-free consensus-based interacting particle systems are introduced for high dimensional non-convex optimization
- Rigorous mathematical convergence results for CBO provided for both the fully time-discrete particle system and (its mean-field limit) under dimension-independent conditions on the coefficients
- Initial data quite restrictive: close to global minimum
- Although the convergence rate does not depend on the dimension, the error does
- CBO-Adam works better in higher dimension but theory is lacking
- Further research include: mean-field limit;

  more computational tests and applications