A consensus-based global optimization method for high dimensional machine learning problems

Yuhua Zhu

Joint work with Jose Carrilo, Shi Jin, Lei Li

Oct 22, 2019

Motivations

The Model and algorithm

Numerical experiments

Motivations

The Model and algorithm

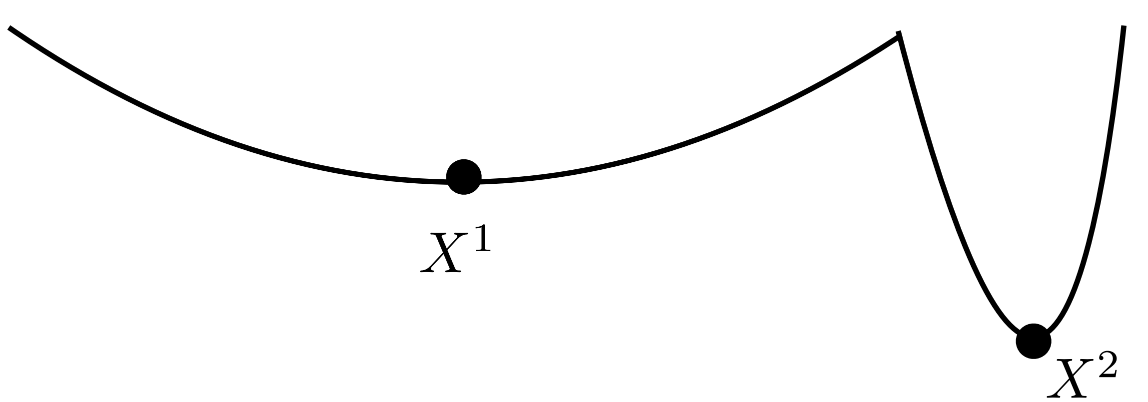
Numerical experiments

Goal: find $x^* = \text{argmin}_x \ L(x)$, $L(x)$ is a non-convex function.

For example: $L(x) = \frac{1}{n}\sum_i l_i(x)$

Why non-gradient method?

- Gradient is hard to calculate
- Objective function is non-smooth
- Flat local minimum



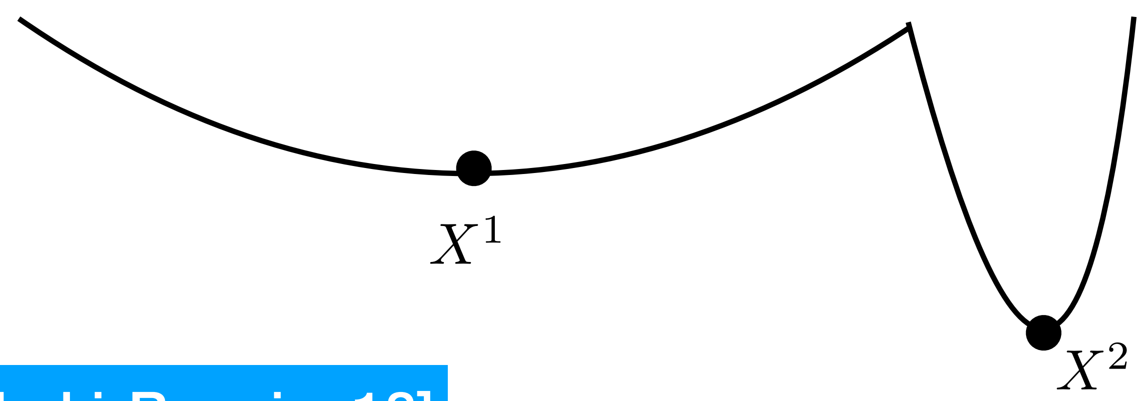
It is hard for gradient based method to escape from flat local minimum

$$\text{GD: } X'(t) = -\nabla L(X(t))$$

$$\text{SGD: } dX_t = -\nabla L(X_t) + \sqrt{\tfrac{1}{\beta}} dB_t$$

$$\text{p.d.f of SGD: } \partial_t p(t,x) = \nabla \cdot \left[\nabla L(x) p + \frac{2}{\beta} \nabla p \right]$$

$$p^{\infty}(x) = \tfrac{1}{Z} e^{-\frac{\beta}{2} L(x)}$$



[Z-Dai, 18], [Jastrzebski-Bengio, 18]

$$\frac{\mathbb{P}(\text{converge to } X^1)}{\mathbb{P}(\text{converge to } X^2)} = \sqrt{\frac{\det H_2}{\det H_1}} e^{\frac{\beta}{2}(L_2 - L_1)}$$

>1 **<1**

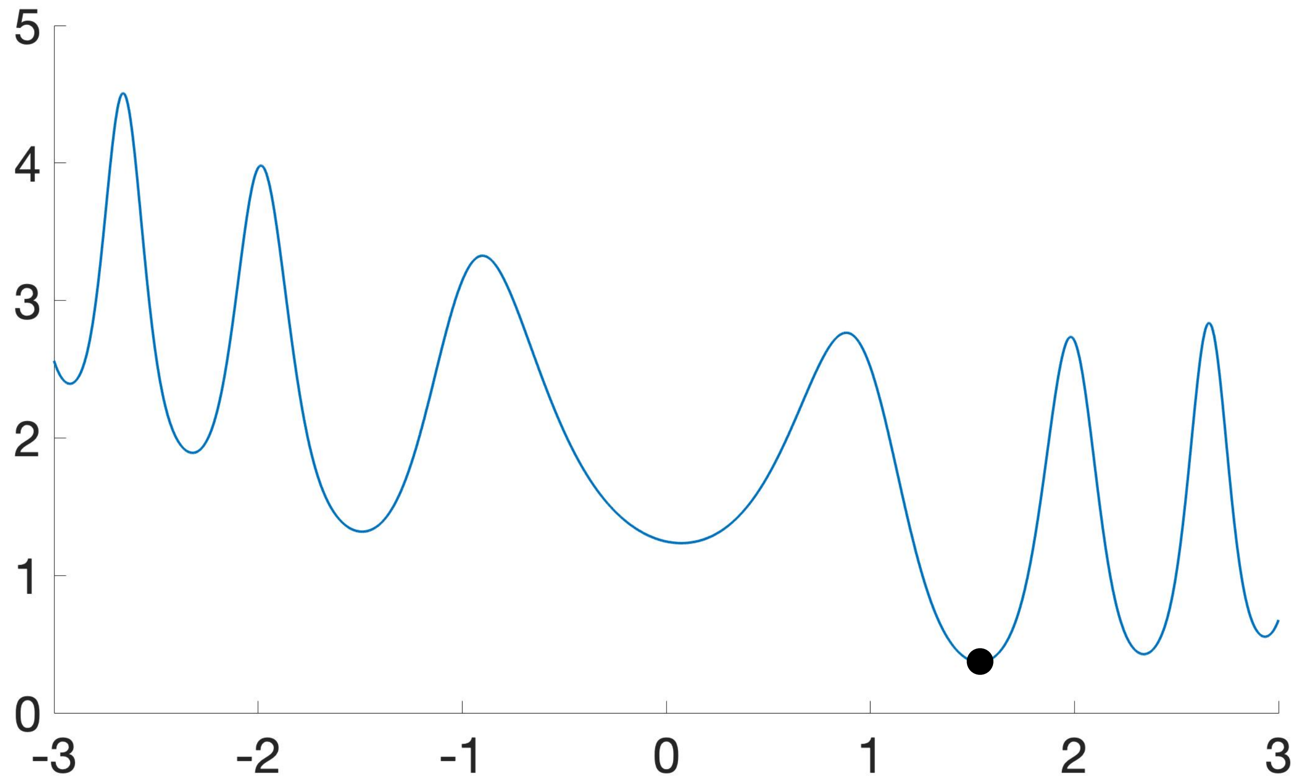
When $\dfrac{\sqrt{\det H_1}}{\sqrt{\det H_2}} < \dfrac{e^{\frac{\beta}{2} L_1}}{e^{\frac{\beta}{2} L_2}}$, SGD is more likely to converge to the flat local minimum.

It is hard for gradient based method to escape from flat local minimum

Example:

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

$$L(x) = \frac{1}{n}\sum_i \ell(x, \hat{x}_i)$$



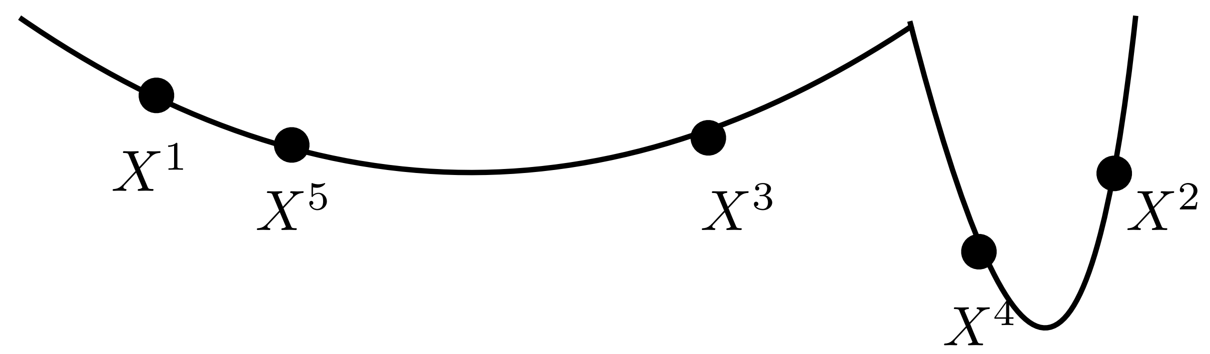
Success rate for SGD to find the correct global minimum is 18%

Motivations

The Model and algorithm

Numerical experiments

Related Work [Pinnau-Totzeck-Tse-Martin, 17]



For $j = 1, \cdots, N$

$$dX^j = -\lambda(X^j - \bar{x}^*)H^\epsilon(L(X^j) - L(\bar{x}^*))\,dt + \sigma|X^j - \bar{x}^*|dW^j$$

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

Relax to their weighted average, in the meantime, explore their surrounding environment.

Require $\lambda \sim O(d)$ to guarantee the convergence of the method

Bad for high-dimensional problems

First improvement

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

component-wise geometric Brownian motion

— Intuitively, now the diffusivity allows the particles to explore each dimension with different rate, so more possible to find the global minimum.

Previous model

Assume $x^* = a$ is a constant.

$$dX = -\lambda(X-a)\,dt + \sigma|X-a|dW^j$$

For each dimension i

$$d[(X)_i - (a)_i] = -\lambda[(X)_i - (a)_i]\,dt + \sigma|X-a|d(W^j)_i$$

By Ito's formula and then take expectation

$$d\mathbb{E}[(X)_i - (a)_i]^2 = -2\lambda\mathbb{E}[(X)_i - (a)_i]^2\,dt + \sigma^2\mathbb{E}|X-a|^2 dt$$

Sum over all dimension

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

$$\boxed{2\lambda > d\sigma^2}$$

[Carrillo-Choi-Totzeck-Tse, 18]

New model

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

$$d[(X)_i - (a)_i] = -\lambda[(X)_i - (a)_i]\,dt + \sigma[(X)_i - (a)_i]d(W^j)_i$$

$$d\mathbb{E}[(X)_i - (a)_i]^2 = -2\lambda\mathbb{E}[(X)_i - (a)_i]^2\,dt + \sigma^2[(X)_i - (a)_i]^2 dt$$

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

$$\boxed{2\lambda > \sigma^2}$$

Mean field limit of the continuous model

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

$$\Big\downarrow \; N \to \infty$$

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

$$\text{with } X^* = \frac{\mathbb{E}(X e^{-\beta L(X)})}{\mathbb{E}(e^{-\beta L(X)})}.$$

Theorem: [Carrilo-Jin-Li-Z, 19]

Under some condition on the initial distribution of X and λ, σ, $X(t) \to \tilde{x}$ exponentially fast and,

$$L(\tilde{x}) \le -\frac{1}{\beta} \log \mathbb{E} e^{-\beta L(X(0))} + \frac{\log 2}{\beta} \le L(x^*) + O(\beta^{-1})$$

The initial law of X

The largeness of β

Numerical method

A gradient-free optimization method

$$\text{Goal: find } x^* = \text{argmin}_x \; L(x) = \text{argmin}_x \; \frac{1}{n}\sum_i l_i(x)$$

Algorithm [Carrillo-Jin-Li-Z-19]

Initially, randomly generate N particles X^j, at each step we randomly update M particles.

O(nN)

- Calculate $\boxed{\cancel{L(X^j)}}$, $\boxed{\cancel{j = 1, \cdots, N.}}$

O(N)

$$\hat{L}(X^j) = \frac{1}{m}\sum_{i\in b} l_i(x), \; b \subset \{1. \cdots, n\}.$$

O(n)

only for $j \in B \subset \{1, \cdots, N\}, |B| = M$

O(M)

O(m)

O(1)

- Find a weighted average: $\bar{X}^* = \frac{1}{\boxed{\sum_{j=1}^N} \mu^j} \boxed{\sum_{j=1}^N} X^j \mu^j, \quad \mu^j = e^{-\beta \boxed{\cancel{L(X^j)}}}$

- Let X^j move towards X^* and and explore their neighbor at the same time.

$$X^j \leftarrow X^j - \lambda\gamma(X^j - \bar{X}^*) + \sigma\sqrt{\gamma}\sum_{i=1}^d \vec{e}_i \left(X^j - \bar{X}^*\right)_i z_i, \quad z_i \sim \mathcal{N}(0,1)$$

Motivations

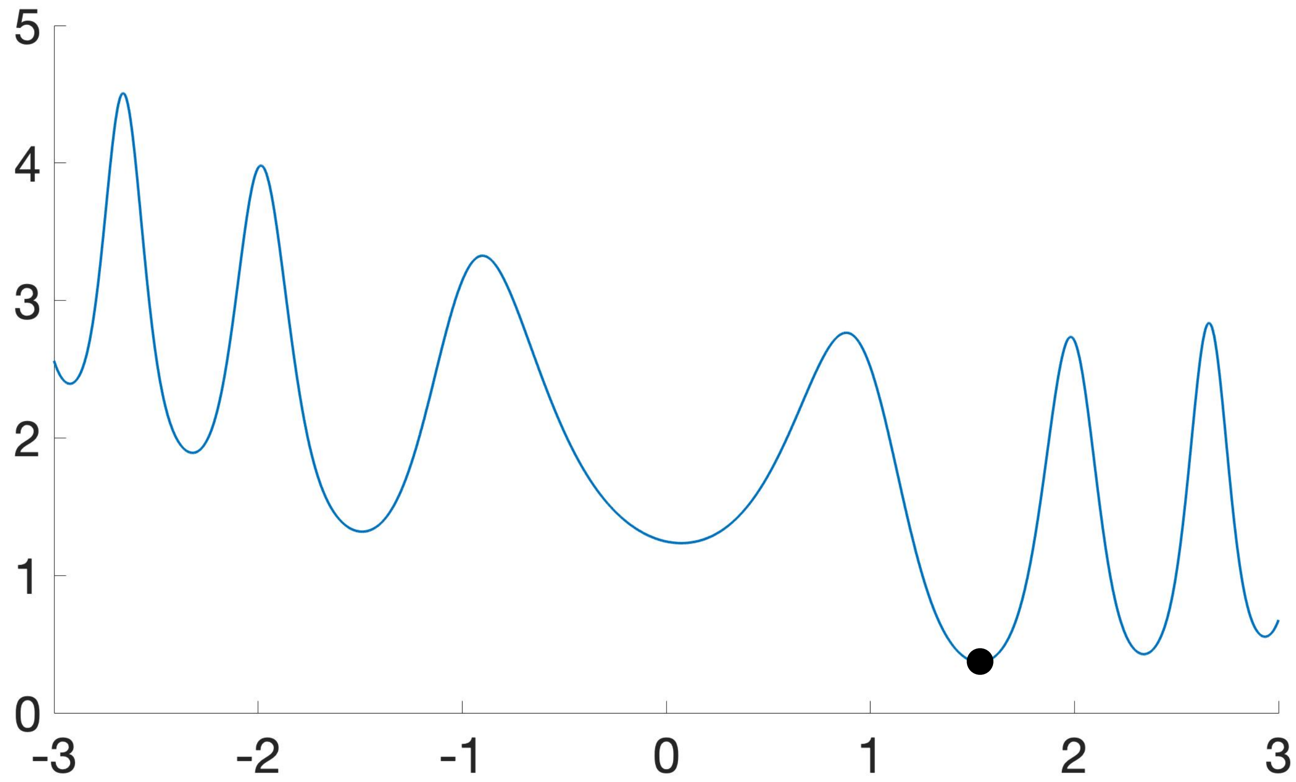
The Model and algorithm

Numerical experiments

Example:

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

$$L(x) = \frac{1}{n}\sum_i \ell(x, \hat{x}_i)$$



Success rate of our method is 98%!
(with N = 100, M = 20)

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

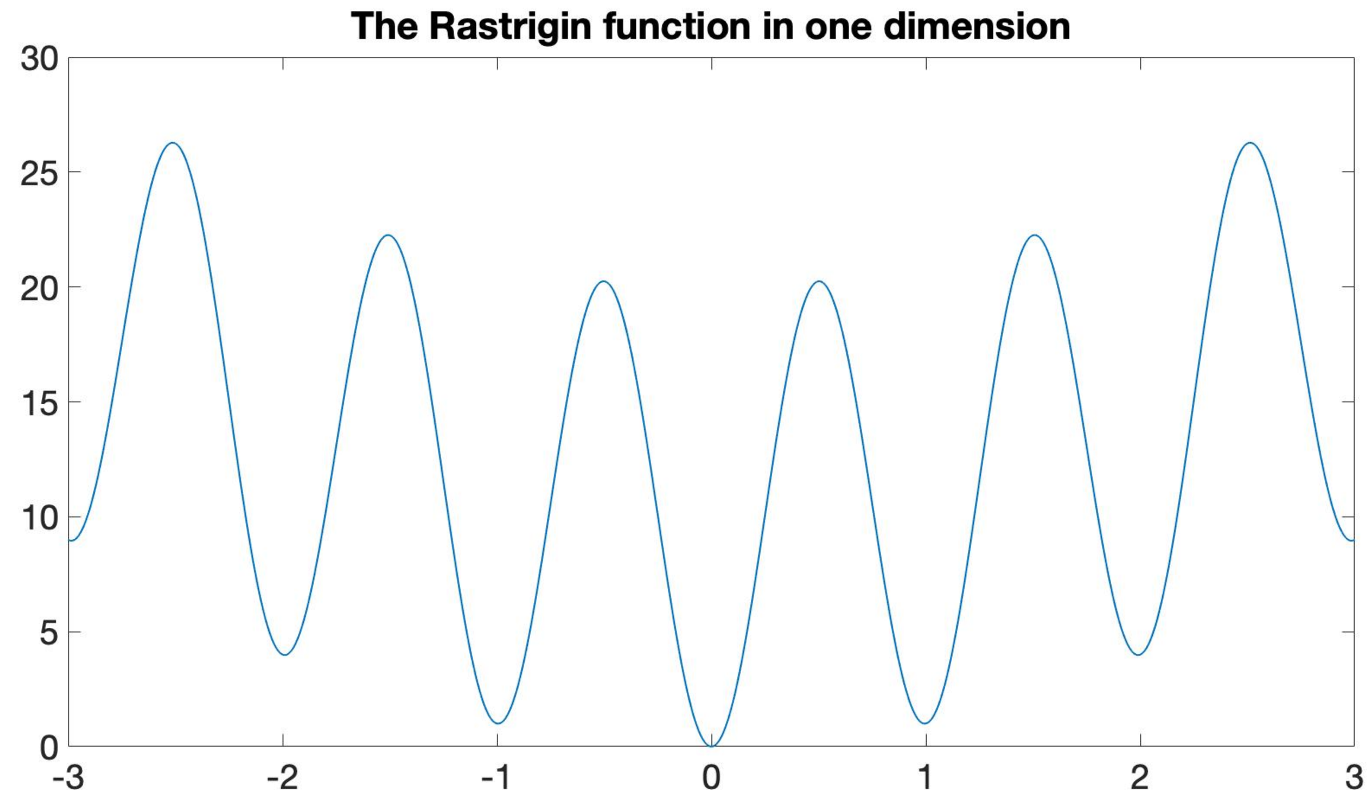


TABLE 2. Rastrigin function in $d = 20$ with $\alpha = 30$.

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%
	$\frac{1}{d}\mathbb{E}[\|v_f(T) - x_*\|^2]$	$3.06e^{-1}$	$2.51e^{-1}$	$2.44e^{-1}$

[Pinnau-Totzeck-Tse-Martin, 17]

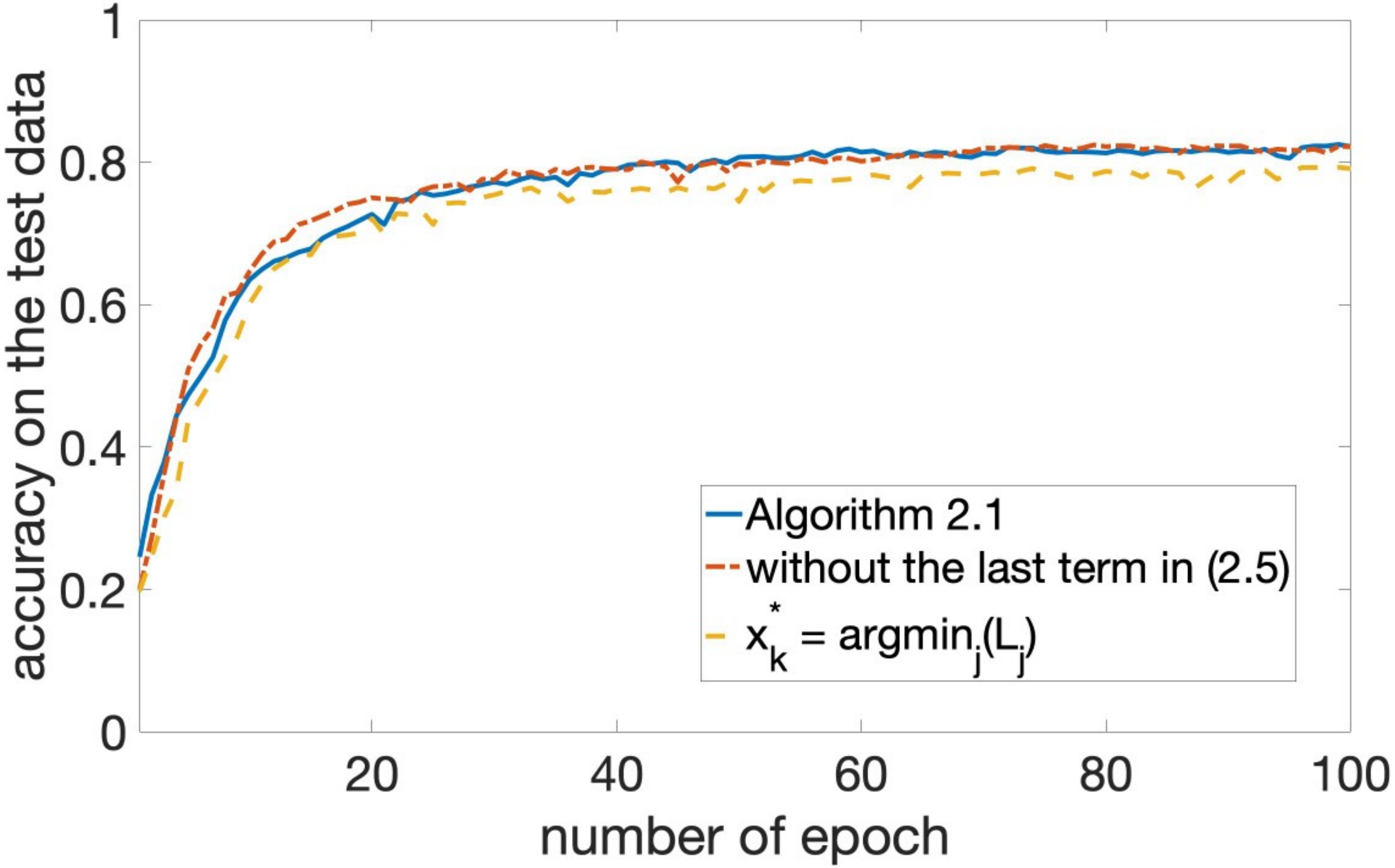
Rastrigin function in $d = 20$ with $\beta = 30$

	N = 50, M = 40 $\sigma = 5.1$	N = 100, M = 70 $\sigma = 5.1$	N = 200, M = 100 $\sigma = 5.1$
x* = 0, success rate	97%	99%	98%
x* = 0, $\frac{1}{d}\mathbb{E}\left[\|x_T^* - x^*\|^2\right]$	5.6E-03	5.03E-04	9.71E-04
x* = 1, success rate	94%	99%	95%
x* = 1, $\frac{1}{d}\mathbb{E}\left[\|x_T^* - x^*\|^2\right]$	3.9E-03	4.95E-04	3E-03
x* = 2, success rate	97%	100%	92%
x* = 2, $\frac{1}{d}\mathbb{E}\left[\|x_T^* - x^*\|^2\right]$	3.0E-03	8.06E-06	4E-03
Computing time saved	22.03%	30.11%	36.14%

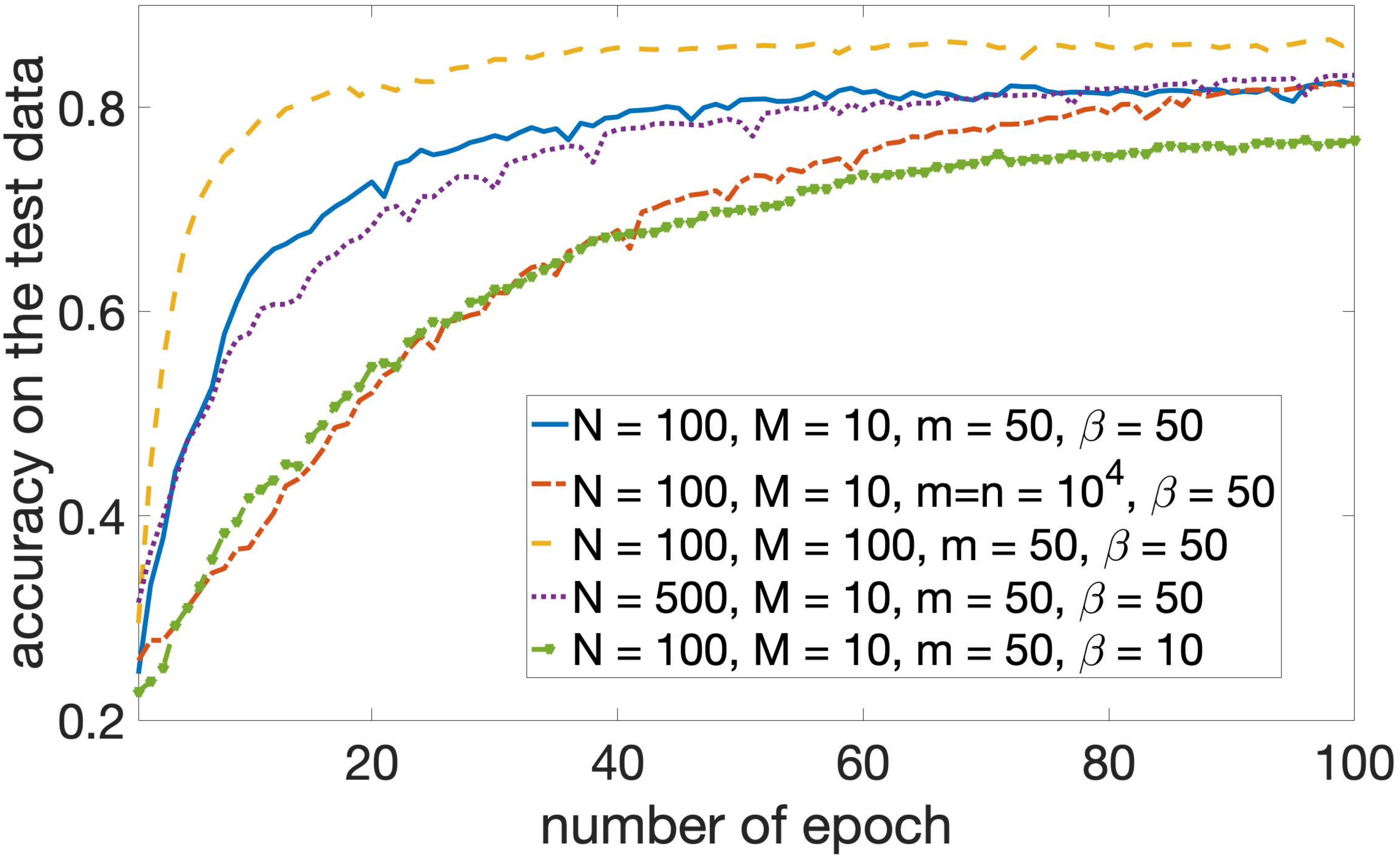
Learning MNIST data with two layer Neural Network

$$X \in \mathbb{R}^{7290}$$

$$\text{Only using } N = 100, M = 10$$



How parameters affect the performance



Future Directions

- Ongoing work: Constrained optimization problem
- How to choose all the parameters?
- Theory for the numerical method.

Thanks!