Consensus-based Global Optimization Method

Xufeng Cai

Zhiyuan College, Shanghai Jiao Tong University, China

Abstract. In optimization problems posed by machine learning and data science, the objective functions are usually not necessarily regular enough. For this issue, we study a new gradient-free method for global non-convex optimization, with the knowledge gleaned from the interacting particle system and flock model. In this report, we generally summarize the algorithm and relevant convergence analysis. Furthermore, we conduct numerical tests for classical numerical optimization problems.

Keywords: Particle System · Consensus State · Gradient-free Global Optimization.

1 Introduction

Nature and human societies offer many examples of self-organized behavior. Birds form flocks which fly in the same direction. Human crowds form parties to reach a consensus. The self-organized aspect of such systems is their dynamics, governed solely by interactions among its individuals or agents, which tend to cluster into colonies, flocks, parties, etc. [3]. In these models, N agents, each with vector of opinions quantified by $x_i \in R^d$, update their states based on the following equations:

$$\frac{d}{dt}\mathbf{x}_i = \alpha \sum_{j \neq i} a_{ij} \left(\mathbf{x}_j - \mathbf{x}_i\right) \quad , \quad a_{ij} = \frac{\phi_{ij}}{\sum_k \phi_{ik}} \tag{1}$$

Here ϕ is a scaled influence function acting on the "difference of states" $\mid x_i - x_j \mid$, and $\phi_{ij} := \phi(\mid x_j - x_i \mid)$. The metric $\mid \cdot \mid$ needs to be properly interpreted, adapted to the specific context of the problem at hand [2]. Under proper conditions, the particle system would relax towards a consensus state.

Based on the ideas of flocking models and interacting particle systems, can we apply such ideas to optimization by choosing proper influence functions? Consider sampling several initial parameters as "particles" to form a system to reach a consensus state, and such state may give out a good approximation for the global minimum.

In this report we consider an unconstrained optimization problem

$$x^* = \underset{x \in R^d}{\arg\min}\, L(x) \tag{2}$$

Here the problem may be in high dimensions, and the objective function $L(x)$ is not necessarily convex. Actually, in optimization problems that arise in machine learning and data science, the objective functions are not necessarily regular enough.

For these optimization problems, gradient-based methods have been dominating. However, those methods may counter several critical issues in various situations. For example, in machine learning problems, when neural networks get deeper, the gradient tends to explode or vanish. Also, gradient-based methods would easily trapped in some bad local minima. All of these motivate us to find a more general framework of gradient-free optimization method.

Here we introduce an improved Consensus-based Optimization (CBO) method [1], with the knowledge gleaned from interacting particle systems and flock models. Particularly, we sample different initial parameters of the objective function as particles, and describe their interactions based on the relative distances between certain particles and their weighted average. Inspired by statistical physics, we formulate an average weighted by the Gibbs distribution corresponding to the objective function. Under several conditions, it could be proven that sampling particles would relax towards their weighted average exponentially fast, and such a consensus state would provide a good approximation for the global minimum of the objective function.

2 Consensus-based Global Optimization Method

Considering the system of N particles, labeled as $X_j, j = 1, \cdots, N$, that tend to relax toward their weighted average, and meanwhile also undergo fluctuation with a multiplicative noise

$$dX^j = -\lambda \left(X^j - \bar{x}^*\right) dt + \sigma \sum_{k=1}^{d} \left(X^j - \bar{x}^*\right)_k dW_k^j \boldsymbol{e}_k \tag{3}$$

where $\bar{x}^*$ is the weighted average of the positions of the particles according to

$$\bar{x}^* = \frac{1}{\sum_{j=1}^{N} e^{-\beta L(x;j)}} \sum_{j=1}^{N} X^j e^{-\beta L\left(x^j\right)} \tag{4}$$

Here λ and σ are the drift rate and noise intensity respectively, $\left(X^j - \bar{x}^*\right)_k$ is the kth component of $X^j - \bar{x}^*$, $\{W_k\}_{k=1}^{d}$ are independent standard Brownian motions, and e_k is the unit vector along the kth dimension. Actually, we use the component-wise geometric Brownian motion for the noise term, which reduces the dimension dependence for the convergence estimates. The primary idea can be shown as

$$\frac{d}{dt} E|X-a|^2 = -2\lambda E|X-a|^2 + \sigma^2 \sum_{i=1}^{d} E(X-a)_i^2 = \left(-2\lambda + \sigma^2\right) E|X-a|^2 \tag{5}$$

Hence, we only need $2\lambda > \sigma^2$ for particles to concentrate, which is dimensional free.

Inspired by Statistical Mechanism, we choose the Gibbs distribution $e^{-\beta L(x)}$ corresponding to the objective function $L(x)$ as the weights for the average of particles. Indeed, we can regard β as the reciprocal of temperature, and the objective function $L(x)$ corresponds to a potential in which particles move along steepest descent modulated by Brownian noise with such chosen β. Hence, the normalized Gibbs measure would assign more its weight on the minimum of $L(x)$ if the temperature goes smaller. The formal quantitative illustration of this intuition is given by Laplace principle: for any probability measure $\rho \in \mathcal{P}(R^d)$ compactly supported with $x_* \in supp(\rho)$, then

$$\lim_{\beta \to \infty} \left(-\frac{1}{\beta} \log \left(\int_{R^d} e^{-\beta L(x)} d\rho(x) \right) \right) = L\left(x^*\right) > 0 \tag{6}$$

Hence, for large $\beta \gg 1$, the normalized measure $e^{-\beta L(x)}$ approximates a Dirac distribution $\delta_{\bar{x}^*}$, if $L(x)$ attains its minimum at a single point $\bar{x}^* \in supp(\rho)$. Thus, the first momentum of the normalized measure should provide a good estimate for the global minimum $x^* = \arg\min L$.

2.1 General Framework of the Algorithm

Here we use the random mini-batch method to calculate the objective function and update the weighted average. Such method could not only save computational costs and accelerating the optimizing process, but also provide extra random noise. The general framework of CBO algorithm is stated below:

Algorithm 1: Consensus-based Optimization Method

```
Result: the estimation $\bar{x}^*$ for the global minimum
Generate $\left\{X_0^j \in R^d\right\}_{j=1}^N$ according to the same distribution $\rho_0$. Set the
 remainder set $R_0$ to be empty;
while $\frac{1}{d}\left\|\Delta\bar{x}^*\right\|_2^2 \geq \epsilon$ do
    Pick batches $B_1^k, B_2^k, \cdots, B_q^k$ from the kth remainder set $\mathcal{R}_k$ and
     $\{1, 2, \cdots, N\}$, where $q = \left\lfloor \frac{N+|\mathcal{R}_k|}{M} \right\rfloor$.;
    Set the remainder set to be $\mathcal{R}_{k+1}$;
    Set $\theta = 0$;
    while $\theta \leq q$ do
        Calculate $L^j := L\left(X^j\right), \forall j \in B_\theta^k$;
        if $n \gg 1$ then
            Generate a random index subset $\mathrm{A}_\theta^k \subset \{1, \cdots, n\}$ with
             $\left|A_\theta^k\right| = m$, and approximate $L^j$ by
             $\hat{L}^j := \hat{L}_\theta^k\left(X^j\right) = \frac{1}{m}\sum_{i \in A_{\hat{\theta}}^k} \ell_i\left(X^j\right), \forall j \in B_\theta^k$;
        end
        $\bar{x}_{k,\theta}^* \leftarrow \frac{1}{\sum_{j \in B_\theta^k} \mu_j} \sum_{j \in B_\theta^k} X^j \mu_j$, with $\mu_j = e^{-\beta L^j}$ or $e^{-\beta \hat{L}^j}$;
        $\hat{X}^j \leftarrow X^j - e^{\lambda\gamma_{k,\theta}}\left(X^j - \bar{x}_{k,\theta}^*\right)$;
        $X^j \leftarrow \hat{X}^j + \sigma_{k,\theta}\sqrt{\gamma_{k,\theta}}\sum_{i=1}^d \boldsymbol{e}_i \left(X^j - \bar{x}_{k,\theta}^*\right)_i z_i, \quad z_i \sim \mathcal{N}(0,1)$;
        $\theta = \theta + 1$;
    end
end
```

Here, ϵ is the tolerance in the stopping criteria, M is the batch size for updating particles, and m is the batch size for calculating objective functions. For updating X^j in the algorithm, the first step is given by the exact solution of the stochastic differential equation $dX^j = -\lambda\left(X^j - x^*\right)dt$ from $t = k\gamma$ to $t = (k+1)\gamma$.

2.2 Mean-field Limit Analysis

In this section we first consider the models of mean-field limit ($N \to \infty$) of particle systems [1], and then we come to analyze N-particle system under time-continuous situation in the next section.

Time Continuous Model. With taking $N \to \infty$, we can obtain that

$$dX = -\lambda \left(X - \bar{x}^*\right) dt + \sigma \sum_{i=1}^{d} \boldsymbol{e}_i \left(X - \bar{x}^*\right)_i dW_i \tag{7}$$

with

$$\bar{x}^* = \frac{E\left(X e^{-\beta L(X)}\right)}{E\left(e^{-\beta L(X)}\right)} \tag{8}$$

The law $\rho(\cdot, t)$ of $X(t)$ in (7) and (8) follows the following Fokker-Planck equation

$$\partial_t \rho = \lambda \nabla \cdot \left(\left(x - \bar{x}^*\right) \rho\right) + \frac{1}{2}\sigma^2 \sum_{i=1}^{d} \partial_{ii} \left(\left(x - \bar{x}^*\right)_i^2 \rho\right) \tag{9}$$

with

$$\bar{x}^* = \frac{\int x e^{-\beta L(x)} \rho(x,t) dx}{\int e^{-\beta L(x)} \rho(x,t) dx} \tag{10}$$

Furthermore, we assume that

$$L_m := \inf L > 0, c_L := \max\left(\left\|\max_i \left|\partial_{ii} L\right|\right\|_\infty, \left\|r\left(\nabla^2 L\right)\right\|_\infty\right) < \infty,$$

where $\nabla^2 L$ is the Hessian of L, $r(\nabla^2 L)$ is the relevant spectral radius and $\partial_{ii} L$ is the diagonal element of the Hessian. Let

$$V(t) := E|X - EX|^2 \quad \text{and} \quad M_L(t) := E e^{-\beta L(X)},$$

then we can prove that, if

$$\begin{aligned} &\mu := 2\lambda - \tfrac{1}{2}\sigma^2 - \sigma^2 \tfrac{e^{-\beta L_m}}{M_L(0)} > 0, \\ &\nu := \tfrac{2V(0)}{\mu M_L^2(0)} \beta e^{-2\beta L_m} c_L \left(2\lambda + \sigma^2\right) \leq \tfrac{3}{4} \end{aligned} \tag{11}$$

then $V(t) \to 0$ exponentially fast, and there exists a $\tilde{x}$ such that $\bar{x}^*(t) \to \tilde{x}$ and $EX \to \tilde{x}$ exponentially fast. Moreover, it holds that

$$L(\tilde{x}) \leq -\frac{1}{\beta} \log M_L(0) - \frac{1}{2\beta} \log(1 - \nu), \tag{12}$$

thus

$$L(\tilde{x}) \leq L_m + O\left(\beta^{-1}\right), \beta \to \infty. \tag{13}$$

From the above estimation, we can see that apart from the assumption we made about λ, β, σ and the initial distribution of particles, how well $L(\tilde{x})$ approximates L_m also depends on how well $-\frac{1}{\beta} \log M_L(0)$ approximates L_m. Thus, the largeness of β and initialization of particle mean a lot for the approximation quality.

Semi-discrete Model. Here we fix γ and let $N \to \infty$. We use the time continuous stochastic differential equation to replace the discrete scheme at each iteration, and obtain the following scheme:

$$dX = -\lambda \left(X - \bar{x}_m^*\right) dt + \sigma \sum^{d} \left(X - \bar{x}_m^*\right)_i dW_i \boldsymbol{e}_i, \quad t \in I_m \tag{14}$$

where $I_m := [t_{m-1}, t_m)$, $t_m = m\gamma$ and

$$\bar{x}_m^* = \frac{\int x \exp(-\beta L(x)) d\rho\left(x, t_{m-1}\right)}{\int \exp(-\beta L(x)) d\rho\left(x, t_{m-1}\right)} \tag{15}$$

We let $\rho(\cdot, t_m)$ denotes the law of X at t_m, $m(t) := EX$ and $V(t) := E|X - EX|^2$. Then we can arrive at the results: if the weighted average sequence $\{\bar{x}_n^*\}$ is bounded and $2\lambda > \sigma^2$, then $m(t_n) \to \bar{m}$ and $V(t_n) \to 0$ exponentially fast. Thus, $\bar{x}_n^* \to \bar{m}$ and the law of X weakly converges to $\delta(x - \bar{m})$ in the dual of $C_b\left(R^d\right)$ with the supremum norm.

2.3 Convergence of CBO method

In this section, we successively answer questions "Does the N-particle system reach a global consensus?" and "under what condition the consensus state would provide a good approximation for the global minimum of the objective funcion?" [4].

We first rewrite (3) in the consensus form:

$$dX_t^i = \lambda \sum_{k=1}^{N} \psi_t^k \left(X_t^k - X_t^i\right) dt + \sigma \sum_{k=1}^{N} \sum_{l=1}^{d} \psi_t^k \left(x_t^{k,l} - x_t^{i,l}\right) dW_t^l e_l, \quad t > 0 \tag{16}$$

where $\psi_t^k := \psi^k(X, t)$ is the communication weight function

$$\psi_t^k := \frac{e^{-\beta L\left(X_t^k\right)}}{\sum_{l=1}^{N} e^{-\beta L\left(X_t^l\right)}}, \quad t \geq 0, \quad k = 1, \cdots, N. \tag{17}$$

Let the $(\cdot)_n$ denotes the state at discrete time $t = nh$, then we have

$$\begin{cases} X_{n+1}^i = X_n^i + \lambda h \sum_{k=1}^{N} \psi_n^k \left(X_n^k - X_n^i\right) + \sigma\sqrt{h} \sum_{k=1}^{N} \sum_{l=1}^{d} \psi_n^k \left(x_n^{k,l} - x_n^{i,l}\right) Z_n^l e_l, \\ \psi_n^k := \frac{e^{-\beta L\left(X_n^k\right)}}{\sum_{i=1}^{N} e^{-\beta L\left(X_n^i\right)}}, \quad i = 1, \cdots, N, \quad n \geq 0. \end{cases} \tag{18}$$

with $Z_n^l \sim \mathcal{N}(0, 1)$. Next we come to answer the above two questions in a general way, and readers can can refer to [4] for the detailed proof.

Global Consensus State: For this question we focus on the continuous situation, and the discrete one can be proved similarly. Consider the deterministic part in (16), we have

$$\frac{dX_t^i}{dt} = \lambda \sum_{k=1}^{N} \psi_t^k \left(X_t^k - X_t^i\right). \tag{19}$$

Thus, we can see that the time dependent convex hull generated by N particles is contractive. Also, the communication weights ψ_t^k are nonnegative and normalized to have a summation 1 for all $t \geq 0$, and they are independent of i in (16). Hence, we can have

$$\frac{d}{dt}\left(X_t^i - X_t^j\right) = -\lambda\left(X_t^i - X_t^j\right), \quad t > 0, \tag{20}$$

which has an analytic solution of

$$\left(X^i - X^j\right)(t) = e^{-\lambda t}\left(X_0^i - X_0^j\right), \quad t \geq 0. \tag{21}$$

Moreover, the maximal and minimal values of the component state of X_t^i are monotone so that they converge the same value. In this way, we can prove that the state X_t^i tends to a unique consensus state X_∞ independent of i for any initial state.

Adding the noise term with $\sigma > 0$, the convex hull spanned by X_t^i is not contractive anymore. By stochastic calculus, we can solve

$$d\left(x_t^{i,l} - x_t^{j,l}\right) = -\lambda\left(x_t^{i,l} - x_t^{j,l}\right)dt - \sigma\left(x_t^{i,l} - x_t^{j,l}\right)dW_t^l, \quad t > 0 \tag{22}$$

to obtain the exact solution

$$x_t^{i,l} - x_t^{j,l} = \left(x_0^{i,l} - x_0^{j,l}\right)\exp\left[-\left(\lambda + \frac{\sigma^2}{2}\right)t + \sigma W_t^l\right], \quad t \geq 0. \tag{23}$$

Hence, the relative difference would reach the a.s. convergence

$$\lim_{t\to\infty}\left|x_t^{i,l} - x_t^{j,l}\right| = 0, \qquad \text{a.s.} \tag{24}$$

Under the condition $2\lambda > \sigma^2$ to reduce the dimensional dependence, there exists a random vector X_∞ being the a.s. limit of X_t^i.

Approximation for the Global Minimum: We assume the particle system reaches a global consensus state. Under proper assumptions on λ, σ and initial data distribution $X^{initial}$, we have

$$\begin{cases} Ee^{-\beta L(X_\infty)} \geq \varepsilon E e^{-\beta L\left(X^{initial}\right)}, \quad or \\ \frac{1}{\beta}\log Ee^{-\beta L(X_\infty)} \leq -\frac{1}{\beta}\log Ee^{-\beta L\left(X^{initial}\right)} - \frac{1}{\beta}\log\varepsilon. \end{cases} \tag{25}$$

If the global minimizer X^* of L is contained in $supp(law(X^{initial}))$, then Laplaces principle yields the desired estimate

$$\operatorname*{ess\,inf}_{\omega\in\Omega} L\left(X^\infty(\omega)\right) \leq L_m + \mathcal{O}\left(\frac{1}{\beta}\right) \qquad \text{for} \qquad \omega \in \Omega, \beta \gg 1. \tag{26}$$

3 Numerical Experiment

We have implemented the CBO method for three different optimization problems: **Logistic Regression**, **Compressive Sensing** and **Filtered Back Projection**. Recently we've obtained reasonable results for Logistic Regression problem, and the other two stay tuned.

3.1 Logistic Regression with l_0 Regularization

Given a set of training data $(x_i,\, y_i)$,$i = 1, ..., N$, where the input $x_i \in R_n$, and the output $y_i \in \{1, -1\}$, we wish to find a proper classification rule. Here we consider the logistic regression model given by

$$Prob(b|\mathbf{a}) = \frac{1}{1 + exp(-b(\mathbf{u}^T\mathbf{a} + v))} \tag{27}$$

The conditional probability $Prob(b|a)$ is larger than 0.5 if $\mathbf{u}^T\mathbf{a} + v$ has the same sign as b,and less than 0.5 otherwise. Then we can derive the objective function based on the minimizing the average negative log-likelihood with l_0 regularization:

$$\min_{(u,v)\in X} \frac{1}{N}\sum_{i=1}^{N} log\Big(1 + exp(-y_i(\mathbf{u}^T\mathbf{x}_i + v))\Big) + \lambda \|u\|_0 \tag{28}$$

where $X = [-10^{10}, 10^{10}]^{n+1}$ and $\lambda = 0.00005$. We use Gisette Data Set for our numerical experiment, where the train set contains 6000 samples with n = 5000 and the test set contains 1000 samples. The stopping criteria is set to be

$$\frac{1}{n}\|\bar{x}_k - \bar{x}_{k-1}\|_2^2 \leq 10^{-8} \tag{29}$$

where $\bar{x}_k$ is the weighted average in CBO algorithm.

For this problem, We use the classification accuracy on test set for evaluation, and draw initial particles from the standard normal distribution. We set $N = 43200$, $M = 41400$, $n = 6000$, $m = 256$, $\gamma = 0.05$, $\sigma = \sqrt{1.5}$, $\lambda = 1$. Here the N is the number of total particles, M is the batch size used to update $\bar{x}$; n is the number of total training data, m is the batch size used to calculate the estimated objective function; γ, σ, λ are the learning rate, the noise rate and drift rate respectively.

Since the batch size for updating the weighted average is quite close to the total number of particles, I omit the remainder set while choose the random mini-batch for particles. We also only allow all the particles to do an independent Brownian motion with variance σ when $\bar{x}$ stops updating and then the algorithm repeats until stabilization. The criteria of adding BM is set to be

$$\frac{\|\bar{x}_k - \bar{x}_{k-1}\|}{\max\{1, \|\bar{x}_k\|\}} < 10^{-2} \tag{30}$$

The numerical results (fixed random seed 0) are shown below:

Table 1. Numerical results for Logistic Regression

Method	Loss	Accuracy
CBO	0.4568	0.9210

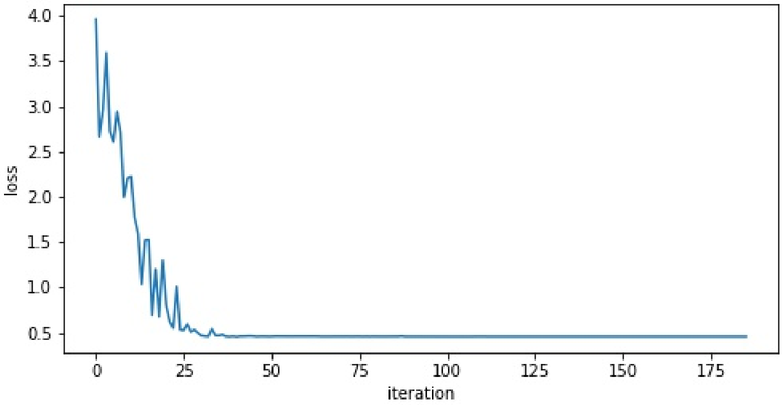

Fig. 1. The value of the objective function versus iteration during the process of CBO optimization

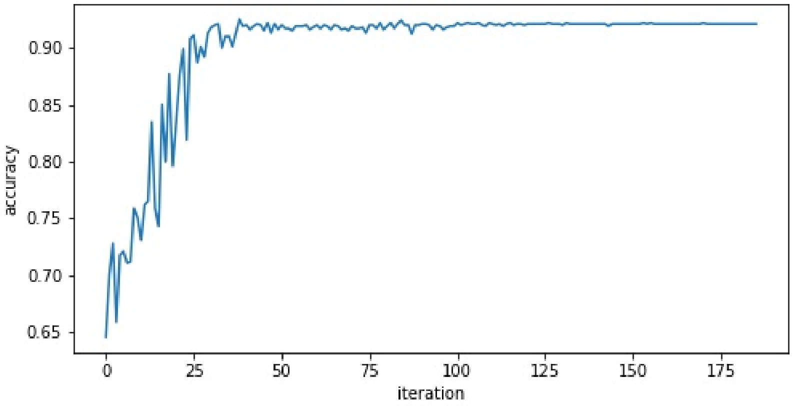

Fig. 2. The accuracy of prediction on the test set versus iteration during the process of CBO optimization

Here are more results for CBO with small amount of particles ($m = 256$, $\gamma = 0.05$, $\sigma = \sqrt{0.1}$, $\lambda = 1$):

Table 2. More numerical results for CBO with less particles

M	N	Accuracy
100	50	0.7430
150	100	0.7580
300	250	0.7950
500	450	0.8110
600	500	0.8340
700	500	0.8340
700	600	0.8320

We also make a comparison to the state-of-the-art algorithms for this problem referred in [5]. However, we have to mention that those methods first use FISTA to solve l_1 regularized logistic regression problem to find a starting point, while CBO still starts from the standard normal distribution directly.

Table 3. More numerical results for CBO with less particles

Method	Accuracy
PIHT	0.9710
IFB	0.9710
mAPG	0.9700
nmAPG	0.9720
EPIHT	0.9780
New PIHT	0.9760
CBO	0.9210

4 Further Directions

1. When calculating the weighted average, the weight is referred to be $exp(-\beta(x))$. If the value of the objective function is too large, the weight will numerically be **zero**. What I thought is to do a step of re-scale for the objective function, just like the activation layer in machine learning (such as sigmoid function). This can also lead to another question about determining the most efficient value interval of loss for CBO algorithm.
2. Another problem is the initialization of particles. Now I just all draw initial particles from standard normal distribution. For Compressive Sensing problem, I also tried to firstly solve l_1 regularized problem by FISTA, and then add noise (standard normal distribution) to initialize particles. But I find that CBO optimizes just a little for the starting point obtained by FISTA.

3. Since CBO only uses the values of the objective function, it cannot fully exploit the information of the objective function for some problems. For example, I found that CBO algorithm would somehow "ignore" the regularization term in the compressive sensing problem. Indeed CBO will minimize the parameters equally in dimension instead of increasing the sparsity. What we considering doing is to develop relevant algorithms based on corresponding specific algorithms for particular problems to preserve the structure of the solution during optimization.
4. CBO converges quite fast in iterations, but every iteration quite takes time when particles increase. What I have done to resolve this issue is to implement a parallel multiprocessing version of CBO, but how to deal with the remainder set while multiprocessing is still under consideration.

References

1. Carrillo, J.A.T., Jin, S., Li, L., Zhu, Y.: A consensus-based global optimization method for high dimensional machine learning problems (2019)
2. Motsch, S., Tadmor, E.: Heterophilious dynamics enhances consensus. SIAM Review **56**, 577–621 (2013)
3. Tadmor, E.: (November 02, 2015), `https://sinews.siam.org/Details-Page/mathematical-aspects-of-self-organized-dynamics`
4. Totzeck, C., Pinnau, R., Blauth, S., Schotthöfer, S.: A numerical comparison of consensusbased global optimization to other particlebased global optimization schemes (2018)
5. Zhang, X., Zhang, X.: A new proximal iterative hard thresholding method with extrapolation for l_0 minimization. Journal of Scientific Computing **79**(2), 809–826 (May 2019)