An exact first-order algorithm for decentralized consensus optimization

Xinwei Sun(1301110047), Jingru Zhang(1301110029)

December 22, 2014

Decentralized Consensus Optimization

$$\min_{x \in \mathbb{R}^p} \bar{f}(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \qquad f_i : \mathbb{R}^p \to \mathbb{R}. \tag{1}$$

EXTRA method

- use a fixed step size independent of the network size.
- convergence rate $O(\frac{1}{k})$ for general convex objection with Lipschitz differential.
- linear convergence rate for (restricted) strongly convex.

Previous Methods

- existing first-order decentralized methods: (sub)gradient,(sub)gradient-push,fast (sub)gradient,dual averaging.
- more restrictive assumptions and worse convergence rate.
- using a fixed step size,do not converge to a solution $x^\star$,just in its neighborhood.

Notation

- $x_{(i)} \in \mathbb{R}^p$:local copy of the global variable x.
- $x_{(i)}^k$:its value at iteration k.
- $\mathbf{f}(\mathbf{x}) \triangleq \sum_{i=1}^n f_i(x_{(i)})$:an aggregate objective function of the local variables
- $$\mathbf{x} \triangleq \begin{pmatrix} -x_{(1)}^T- \\ -x_{(2)}^T- \\ \vdots \\ -x_{(n)}^T- \end{pmatrix} \in \mathbb{R}^{n\times p}$$

Notation

- $$\nabla \mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} -\nabla^T f_1(x_{(1)})- \\ -\nabla^T f_2(x_{(2)})- \\ \vdots \\ -\nabla^T f_n(x_{(n)})- \end{pmatrix} \in \mathbb{R}^{n \times p}.$$
- $\mathbf{x}$ is consensual if all of its rows are identical,i.e., $x_{(1)} = \cdots = x_{(n)}$.
- $\|A\|_G \triangleq \sqrt{trace(A^T GA)}$:G-matrix norm.
- $null\{A\} \triangleq \{x \in \mathbb{R}^n | Ax = 0\}$:null space of A.
- $span\{A\} \triangleq \{y \in \mathbb{R}^m | y = Ax, \forall x \in \mathbb{R}^n\}$:linear span of all the columns of A.

Algorithm 1 (DGD)

$$x_{(i)}^{k+1} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{k} - \alpha^{k} \nabla f_i(x_{(i)}^{k}), \textit{ for agent } i = 1, \cdots, n. \tag{2}$$

matrix version

$$\mathbf{x}^{k+1} = Wx^{k} - \alpha^{k} \nabla \mathbf{f}(\mathbf{x^k}). \tag{3}$$

$x_{(i)}^{k} \in \mathbb{R}^{p}$ is the local copy of x held by agent i at iteration k,
$W = [w_{ij}] \in \mathbb{R}^{n \times n}$ is a symmetric mixing matrix
satisfying $null\{I - W\} = span\{\mathbf{1}\}$.If two agents i and j are neither neighbors nor identical,then $w_{ij} = 0$,
$\sigma_{max}(W - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathbf{T}}) < 1$,
$\alpha^{k} > 0$ is a step size for iteration k.

Dilemma of DGD

- converge slowly to an exact solution with a sequence of diminishing step sizes.
- converge faster with a fixed step size but stall at an inaccurate solution ($O(\alpha)$-neighbourhood of a solution).

The Cause of Inexact Convergence With a Fixed Step Size

let $\mathbf{x}^{\infty}$ be the limit of $\mathbf{x}^{k}$. Take the limit over k and get

$$\mathbf{x}^{\infty} = \mathbf{W}\mathbf{x}^{\infty} - \alpha\nabla\mathbf{f}(\mathbf{x}^{\infty}). \tag{4}$$

When α is fixed and nonzero, assuming the consensus of x^{∞} (namely, it has identical rows $x_{(i)}^{\infty}$) will mean $\mathrm{x}^{\infty} = W\mathrm{x}^{\infty}$, as a result of $W\mathbf{1} = \mathbf{1}$, and thus $\nabla\mathrm{f}(\mathrm{x}^{\infty}) = \mathbf{0}$, which is equivalent to $\nabla f_i(x_{(i)}^{\infty}) = 0$, $\forall i$, i.e., the same point $x_{(i)}^{\infty}$ simultaneously minimizes f_i for all agents i. This is impossible in general and is different from our objective to find a point that minimizes $\sum_{i=1}^{n} f_i$.

Development of EXTRA

Consider the DGD update at iterations k+1 and k as follows

$$\mathbf{x}^{k+2} = \mathbf{W}\mathbf{x}^{k+1} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}), \tag{5}$$

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k}), \tag{6}$$

$$\tilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}, \text{ the choice of } \tilde{\mathbf{W}} \text{ will be generalized later.} \tag{7}$$

Subtracting the above two iterations of DGD,we get the update formula of EXTRA:

$$\mathbf{x}^{k+2} - \mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k+1} - \tilde{\mathbf{W}}\mathbf{x}^{k} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \alpha\nabla\mathbf{f}(\mathbf{x}^{k}). \tag{8}$$

Conclusion 1

Provided that $null\{\mathbf{I}-\mathbf{W}\} = span\{\mathbf{1}\}, \tilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}, \mathbf{1}^{\mathbf{T}}(\mathbf{W}-\tilde{\mathbf{W}}) = \mathbf{0}$ *and the continuity of* ∇f, *if a sequence following EXTRA converges to a point* $x^\star$, *then* $x^\star$ *is consensual and any of its identical row vectors solves problem.*

The Algorithm EXTRA

Algorithm 2 (EXTRA)

Choose $\alpha > 0$ *and mixing matrices* $\mathbf{W} \in \mathbb{R}^{n \times n}$ *and* $\tilde{\mathbf{W}} \in \mathbb{R}^{n \times n}$
Pick any $\mathbf{x}^0 \in \mathbb{R}^{n \times p}$
1.$\mathbf{x}^{\mathbf{1}} \leftarrow \mathbf{W}\mathbf{x}^{\mathbf{0}} - \alpha \nabla \mathbf{f}(\mathbf{x}^{\mathbf{0}})$;
2.for $k = 0, 1, \cdots$ *do*
$\mathbf{x}^{\mathbf{k+2}} \leftarrow (\mathbf{I} + \mathbf{W})\mathbf{x}^{\mathbf{k+1}} - \tilde{\mathbf{W}}\mathbf{x}^{\mathbf{k}} - \alpha[\nabla f(x^{k+1}) - \nabla f(x^k)]$;
end for

Assumptions on The Mixing Matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$

Assumption 1

(Mixing matrix). Consider a connected network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consisting of a set of agents $\mathcal{V} = \{1, 2, \cdots, n\}$ and a set of undirected edges $\mathcal{E}$. The mixing matrices $W = [w_{ij}] \in R^{n\times n}$ satisfy

1. *(Decentralized property) If $i \neq j$ and $(i, j) \notin \mathcal{E}$, then $w_{ij} = \tilde{w}_{ij} = 0$.*
2. *(Symmetry) $W = W^\top$, $\tilde{W} = \tilde{W}^\top$.*
3. *(Null space property) $null\{W - \tilde{W}\} = span\{1\}$, $null\{I - \tilde{W}\} \supseteq span\{1\}$.*
4. *(Spectral property) $\tilde{W} \succ 0$ and $\frac{I+W}{2} \succcurlyeq \tilde{W} \succcurlyeq W$.*

In fact,EXTRA can use the same $\mathbf{W}$ used in DGD and simply take $\tilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}$

Mixing Matrices

The mixing matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ diffuse information throughout the network, which can significantly affect performance.We will verify this point in the numerical experiments!!!

$\mathbf{W}$ can be chosen through following methods:

- symmetric doubly stochastic matrix: $\mathbf{W} = \mathbf{W}^{\mathbf{T}}, \mathbf{W1} = \mathbf{1}$ and $w_{ij} \geq 0$.
- Laplacian-based constant edge weight matrix
- Metropolis constant edge weight matrix
- symmetric fastest distributed linear averaging(FDLA) matrix

$\tilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}$ is found to be very efficient.

Due to the update formula

$$\mathbf{x^{k+2}} - \mathbf{x^{k+1}} = \mathbf{Wx^{k+1}} - \tilde{\mathbf{W}}\mathbf{x^{k}} - \alpha\nabla \mathbf{f}(\mathbf{x^{k+1}}) + \alpha\nabla \mathbf{f}(\mathbf{x^{k}}),$$

We add the iterations k together and obtain

$$\mathbf{x^{k+1}} = \mathbf{Wx^{k}} - \alpha\nabla f(x^{k}) + \sum_{t=0}^{k-1}(\mathbf{W} - \tilde{\mathbf{W}})\mathbf{x}^{t}, \qquad k = 0, 1, \cdots \tag{9}$$

An EXTRA update is a DGD update with a cumulative correction term.The role of the cumulative term $\sum_{t=0}^{k-1}(\mathbf{W} - \tilde{\mathbf{W}})\mathbf{x^{t}}$ is to neutralize $-\alpha\nabla f(x^{k})$ in $(span\{\mathbf{1}\})^{\perp}$

EXTRA

Preliminaries

Lemma. 1

Assume null$\{I - W\} = span\{1\}$. If

$$x^\star = \begin{bmatrix} -- & x_{(1)}^{\star\top} & -- \\ -- & x_{(2)}^{\star\top} & -- \\ & \vdots & \\ -- & x_{(n)}^{\star\top} & -- \end{bmatrix} \tag{10}$$

satisfies conditions:
1. $x^\star = Wx^\star \quad (consensus)$,
2. $1^\top \nabla f(x^\star) = 0 \quad (optimality)$,

then $x^\star = x_{(i)}^\star$, for any i, is a solution to consensus optimization problem.

EXTRA

Lemma. 2

Given mixing matrices W and $\tilde{W}$, define $U = (\tilde{W} - W)^{\frac{1}{2}}$ by letting $U \triangleq VS^{\frac{1}{2}}V^{\top} \in R^{n \times n}$ where $VSV^{\top} = \tilde{W} - W$ is the econmical-form singular value decomposition. Then, under someassumptions, $x^{\star}$ is consensual and $x^{\star}_{(1)} = x^{\star}_{(2)} = \cdots = x^{\star}_{(n)}$ is optimal to optimization problem if and only if there exists $q^{\star} = Up$ for some $p \in R^{n \times p}$ such that

$$Uq^{\star} + \alpha \nabla f(x^{\star}) = 0 \tag{11}$$

$$Ux^{\star} = 0 \tag{12}$$

EXTRA

Theorem 1

Under assumptions $1-3$, *If* α *satisfies* $0 < \alpha < \frac{2\lambda_{min}(\tilde{W})}{L_f}$, *then*

$$\|z^k - z^\star\|_G^2 - \|z^{k+1} - z^\star\|_G^2 \geq \zeta \|z^k - z^{k+1}\|_G^2, k = 0, 1, ..., \tag{13}$$

where $\zeta = 1 - \frac{\alpha L_f}{2\lambda_{min}(\tilde{W})}$.

$z^k = \begin{bmatrix} q^k \\ x^k \end{bmatrix}$. $z^k = \begin{bmatrix} q^k \\ x^k \end{bmatrix}$. $G = \begin{bmatrix} I & 0 \\ 0 & \tilde{W} \end{bmatrix}$.

EXTRA

The key steps in the proof of above theorem.

$$(I + W - 2\tilde{W})x^\star = 0 \tag{14}$$

$$\frac{2\alpha}{L_f}\|\nabla f(x^k) - \nabla f(x^\star)\|_F^2 \le 2\alpha\langle x^k - x^\star, \nabla f(x^k) - \nabla f(x^\star)\rangle \tag{15}$$

$$\langle x^{k+1} - x^\star, U(q^\star - q^{k+1})\rangle = \langle U(x^{k+1} - x^\star), q^\star - q^{k+1}\rangle = \langle Ux^{k+1}, q^\star - q^{k+1}\rangle \tag{16}$$

$$\langle x^{k+1} - x^\star, \tilde{W}(q^\star - q^{k+1})\rangle = \langle \tilde{W}(x^{k+1} - x^\star), q^\star - q^{k+1}\rangle \tag{17}$$

EXTRA

The key steps in the proof of above theorem

$$\|z^k - z^\star\|_G^2 - \|z^{k+1} - z^\star\|_G^2 - \|z^k - z^{k+1}\|_G^2 - 2\|x^{k+1} - x^\star\|_{I+W-2\tilde{W}}^2 + \frac{\alpha L_f}{2}\|x^k - x^{k+1}\|_F^2 \geq 0 \quad (18)$$

$$\|z^k - z^\star\|_G^2 - \|z^{k+1} - z^\star\|_G^2 - \|z^k - z^{k+1}\|_G^2 + \frac{\alpha L_f}{2}\|x^k - x^{k+1}\|_F^2 \geq 0 \quad (19)$$

$$\|z^k - z^\star\|_G^2 - \|z^{k+1} - z^\star\|_G^2 \geq \|z^k - z^{k+1}\|_{G'}^2 \quad (20)$$

$$\|z^k - z^{k+1}\|_{G'}^2 \geq \zeta\|z^k - z^{k+1}\|_G^2 \quad (21)$$

EXTRA

From the analysis above, we can give these thee necessary assumptions:

Assumption 2

(Mixing matrix). Consider a connected network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ *consisting of a set of agents* $\mathcal{V} = \{1, 2, \cdots, n\}$ *and a set of undirected edges* $\mathcal{E}$*. The mixing matrices* $W = [w_{ij}] \in R^{n \times n}$ *satisfy*

1. *(Decentralized property) If* $i \neq j$ *and* $(i, j) \notin \mathcal{E}$*, then* $w_{ij} = \tilde{w}_{ij} = 0$.
2. *(Symmetry)* $W = W^\top$, $\tilde{W} = \tilde{W}^\top$.
3. *(Null space property)* $null\{W - \tilde{W}\} = span\{1\}$, $null\{I - \tilde{W}\} \supseteq span\{1\}$.
4. *(Spectral property)* $\tilde{W} \succ 0$ *and* $\frac{I+W}{2} \succcurlyeq \tilde{W} \succcurlyeq W$.

EXTRA

Assumption 3

(Convex objective with Lipschitz continuous gradient) Objective functions f_i are proper closed convex and Lipschitz differentiable:

$$\|\nabla f_i(x_a) - \nabla f_i(x_b)\|_2 \leq L_{f_i}\|x_a - x_b\|_2, \forall x_a, x_b \in R^p,$$

where $L_{f_i} \geq 0$ are constant.

Assumption 4

(Solution existence) The optimization problem has a nonempty set of optimal solutions: $\mathcal{X}^\star \neq \emptyset$.

EXTRA

From the theorem above, we can get:

Theorem 2

In the same setting of Theorem 3, the following rates hold:
(1) Running-average progress:

$$\frac{1}{k}\sum_{t=1}^{k}\|z^t - z^{t+1}\|_G^2 = O(\tfrac{1}{k});$$

(2) Running-best progress:

$$min_{t\leq k}\{\|z^t - z^{t+1}\|_G^2\} = o(\tfrac{1}{k});$$

(3) Running-average optimality residuals:

$$\frac{1}{k}\sum_{t=1}^{k}\|Uq^t + \alpha\nabla f^{x^t}\|_{\tilde{W}}^2 = O(\tfrac{1}{k}) \text{ and } \frac{1}{k}\sum_{t=1}^{k}\|Ux^t\|_F^2 = O(\tfrac{1}{k});$$

(4) Running-best optimality residuals:

$$min_{t\leq k}\{\|Uq^t + \alpha\nabla f^{x^t}\|_{\tilde{W}}^2 = o(\tfrac{1}{k}) \text{ and } min_{t\leq k}\|Ux^t\|_F^2 = o(\tfrac{1}{k});$$

EXTRA

Theorem 3

If $g(x) \triangleq f(x) + \frac{1}{4\alpha}\|x\|^2_{\tilde{W}-W}$ is restricted strongly convex with respect to $x^\star$ with constant $\mu_g > 0$, then with proper step size $\alpha < \frac{2\mu_g \lambda_{min}(\tilde{W})}{L_f^2}$, there exists $\delta > 0$ such that the sequence $\{z^k\}$ generated by EXTRA satisfies

$$\|z^k - z^\star\|^2_G \geq (1+\delta)\|z^{k+1} - z^\star\|^2_G. \tag{22}$$

EXTRA

Denote μ_f is the constant of the restrictedly convex function $f(x)$, then in the proof of the theorem above,

$$\mu_g = min\{\mu_f - 2L_f\gamma, \frac{\tilde{\lambda}_{min}(\tilde{W} - W)}{2\alpha(1 + \frac{1}{\gamma^2})}\} \tag{23}$$

So we set

$$\alpha = \frac{\tilde{\lambda}_{min}(\tilde{W} - W)}{2(1 + \frac{1}{\gamma^2})(\mu_f - 2L_f\gamma)} = O(\frac{\mu_g}{L_f^2}) \tag{24}$$

The paper said that in this case if we set $\alpha = O(\frac{1}{L_f})$, the algorithm still converges and become faster, but it remains an open question to prove linear convergence under this step.

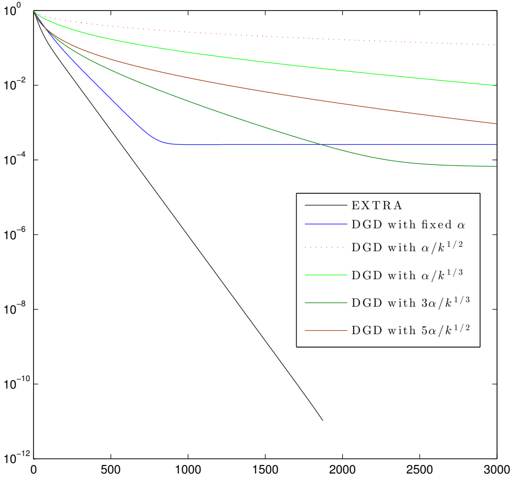
EXTRA

Experiment set:
We consider solving problem

$$minimize_x \bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\|M_{(i)}x - y_{(i)}\|_2^2 \tag{25}$$

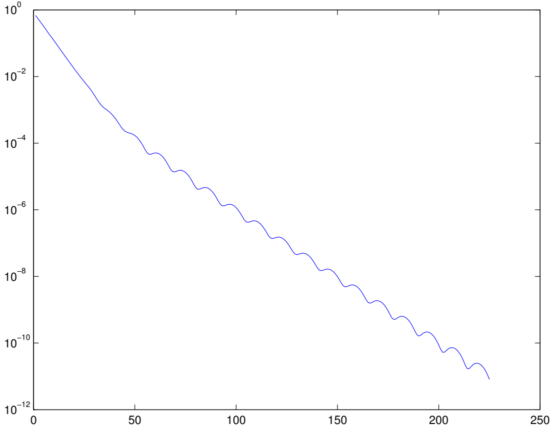
Here $y_{(i)} = M_{(i)}x + e_{(i)}$, where $y_{(i)} \in R^{m_i}$ and $M_{(i)} \in R^{m_i \times p}$ are measured data, $x \in R^p$ is unknown signal, and $e_{(i)} \in R^{m_i}$ is unknown noise. In this set, $n = 10$, $m_i = 1$, $\forall i$, $p = 5$. Data $M_{(i)}$ and $e_{(i)}$, $\forall i$, are generated following the standard normal equation. The algorithm starts from $x_{(i)}^0 = 0$, $\forall i$, and we set $\|x - x_i^0\| = 300$.

EXTRA


10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
0
500
1000
1500
2000
2500
3000
EXTRA
DGD with fixed α
DGD with $\alpha/k^{1/2}$
DGD with $\alpha/k^{1/3}$
DGD with $3\alpha/k^{1/3}$
DGD with $5\alpha/k^{1/2}$

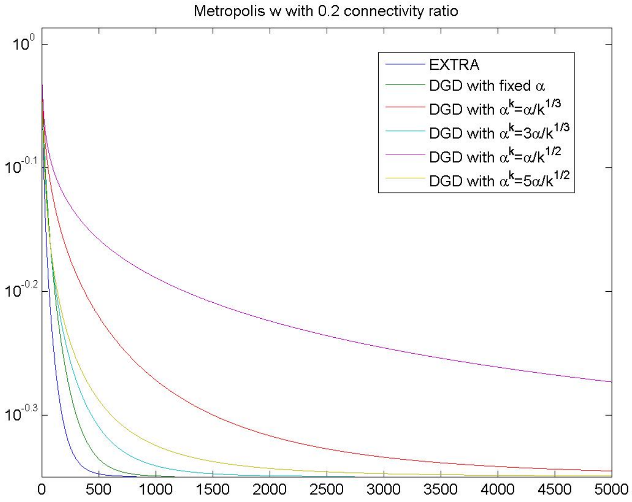
EXTRA

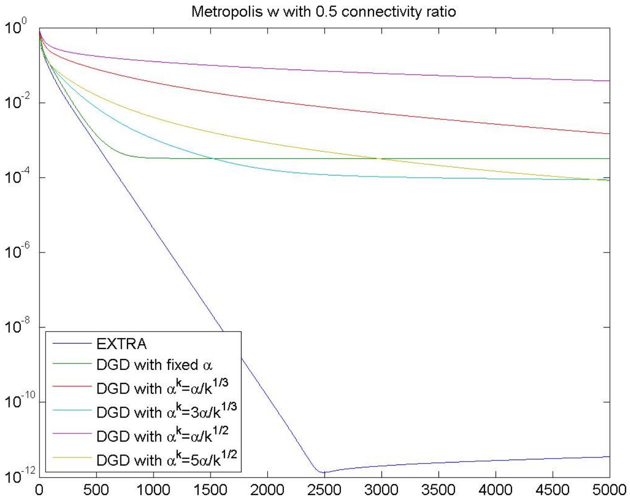
We adjust $M_{(i)}$, $\forall i$, to 20 to make $f_{(i)}(x)$ strongly convex. All other parameters are the same. And we set $\alpha = \frac{1}{2L_f}$.

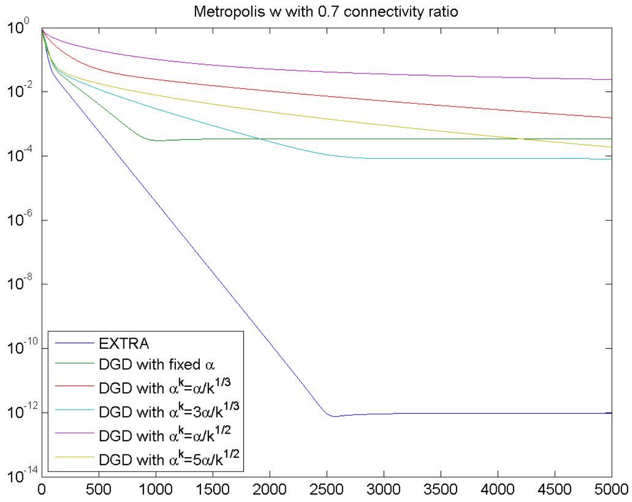


Other Mixing Matrices

We have tried other methods to choose mixing matrices $\mathbf{W}$ .The result of using metropolis method with different connectivity ratio $r = 0.2, 0.5, 0.7$ as follows. We can see the significant effect from mixing matrices.


Metropolis w with 0.2 connectivity ratio
EXTRA
DGD with fixed α
DGD with αk=α/k1/3
DGD with αk=3α/k1/3
DGD with αk=α/k1/2
DGD with αk=5α/k1/2
10^0
10^-0.1
10^-0.2
10^-0.3
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000


Metropolis w with 0.5 connectivity ratio
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
EXTRA
DGD with fixed α
DGD with $\alpha^k=\alpha/k^{1/3}$
DGD with $\alpha^k=3\alpha/k^{1/3}$
DGD with $\alpha^k=\alpha/k^{1/2}$
DGD with $\alpha^k=5\alpha/k^{1/2}$


Metropolis w with 0.7 connectivity ratio
10^0
10^-2
10^-4
10^-6
10^-8
10^-10
10^-12
10^-14
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
EXTRA
DGD with fixed α
DGD with α^k=α/k^1/3
DGD with α^k=3α/k^1/3
DGD with α^k=α/k^1/2
DGD with α^k=5α/k^1/2