Walkman: A Communication-Efficient Random-Walk Algorithm for Decentralized Optimization

Xianghui Mao⋄ Kun Yuan* Yubin Hu⋄
Yuantao Gu⋄ Ali H. Sayed† Wotao Yin‡

⋄Tsinghua EE  *UCLA ECE  †EPFL Engineering  ‡UCLA Math
1. Decentralized optimization

2. The Walkman method

3. Convergence

4. Communication analysis

5. Simulation results
1. Decentralized optimization

2. The Walkman method

3. Convergence

4. Communication analysis

5. Simulation results
Decentralized optimization

- Consider a decentralized optimization problem over a network \((V, E)\)

\[
\min_{x \in \mathbb{R}^p} \quad r(x) + \frac{1}{n} \sum_{i=1}^{n} f_i(x),
\]

where \(n\) is the number of nodes

- Node \(i\) has access to \(f_i(x)\). All nodes can access \(r(x)\).

- Both \(f_i(x)\) and \(r(x)\) can be non-convex
Gossip-based approaches

- Agent communicates with all, or a random subset, of direct neighbors
- Prior methods: DGD[1], diffusion[2], D-ADMM[3, 4], EXTRA[5], PG-EXTRA[6], DIGing[7], Exact diffusion[8], NIDS[9] ...
- Convergence rates are comparable to standard centralized optimization.
- Every agent communicates ⇒ per-iteration comm. cost at $O(n) – O(n^2)$.
Random-walk approaches

A walker moves $x$ through the network and updates it. $x^k$ is the $k$th value.

Agent $i$ receiving $x$ will update it with local (sub)gradient $\nabla f_i$.

$O(1)$ communication per iteration.


Figure: A random walk $(1, 6, 9, 1, 2, 6, 5, ...)$
Proposed method: Walkman

- Walkman is a random-walk (RW) algorithm
- Exact convergence with fixed step-size; much faster than existing RWs
- Convergence guarantee established for non-convex and convex scenarios
- More communication efficient than state-of-the-art methods
- Can escape from saddle points on tested non-convex problems
Walkman communication efficiency

- Comm. complexity for various algorithms for decentralized least squares

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Communication Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walkman (proposed)</td>
<td>$O\left( \ln \left( \frac{1}{\epsilon} \right) \cdot \frac{n \ln^3(n)}{(1-\lambda_2(P))^2} \right)$</td>
</tr>
<tr>
<td>D-ADMM [14]</td>
<td>$O\left( \ln \left( \frac{1}{\epsilon} \right) \cdot \left( \frac{m}{(1-\lambda_2(P))^{1/2}} \right) \right)$</td>
</tr>
<tr>
<td>EXTRA [5]</td>
<td>$O\left( \ln \left( \frac{1}{\epsilon} \right) \cdot \left( \frac{m}{1-\lambda_2(P)} \right) \right)$</td>
</tr>
<tr>
<td>Exact diffusion [8]</td>
<td>$O\left( \ln \left( \frac{1}{\epsilon} \right) \cdot \left( \frac{m}{1-\lambda_2(P)} \right) \right)$</td>
</tr>
</tbody>
</table>

- Walkman is most communication efficient when

$$\lambda_2(P) \leq 1 - \frac{1}{m^{2/3}}$$

$\lambda_2(P)$ is a measure of network connectivity, and $m$ is the number of edges.
1. Decentralized optimization

2. The Walkman method

3. Convergence

4. Communication analysis

5. Simulation results
Problem reformulation

- Recall the problem

\[
\minimize_{x \in \mathbb{R}^p} r(x) + \frac{1}{n} \sum_{i=1}^{n} f_i(x),
\]

- Create local variables \( y_i \) and make then all equal to \( x \).

- Defining \( y = \text{col}\{y_1, y_2, \cdots, y_n\} \in \mathbb{R}^{np} \) and \( F(y) = \sum_{i=1}^{n} f_i(y_i) \), we have

\[
\minimize_{x, y} r(x) + \frac{1}{n} F(y),
\]

subject to \( 1 \otimes x - y = 0, \quad (2) \)

where \( 1 = [1 \ 1 \ \ldots \ 1]^T \in \mathbb{R}^n \) and \( \otimes \) is the Kronecker product

- The above two problems are equivalent.
Standard ADMM

- The augmented Lagrangian function of problem (2) is

$$L_\beta (x, y; z) = r(x) + \frac{1}{n} \left( F(y) + \langle z, 1 \otimes x - y \rangle + \frac{\beta}{2} \| 1 \otimes x - y \|^2 \right),$$

where $z \in \mathbb{R}^{np}$ is the dual variable (Lagrange multiplier)

- The standard ADMM to solve (2) is

$$\bar{x}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} (y_i^k - \frac{z_i^k}{\beta}),$$

$$x^{k+1} = \text{prox}_{\frac{1}{\beta} r}(\bar{x}^{k+1}),$$

$$y_i^{k+1} = \text{prox}_{\frac{1}{\beta} f_i} \left( x_i^{k+1} + \frac{z_i^k}{\beta} \right), \quad \forall i \in V$$

$$z_i^{k+1} = z_i^k + \beta (x_i^{k+1} - y_i^{k+1}), \quad \forall i \in V$$

- Step 1 uses a reduce operation, implementable in a distributed $1$-to-$N$ setting but not in our decentralized setting
• To update $x$ with only one $y_i$ at a time.

• To decentralize the computation of $\bar{x}^{k+1}$, we propose

$$\bar{x}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} (y_i^k - \frac{z_i^k}{\beta}),$$

$$x^{k+1} = \text{prox}_{\frac{1}{\beta} r}(\bar{x}^{k+1}),$$

$$y_i^{k+1} = \begin{cases} \text{prox}_{\frac{1}{\beta} f_i}(x^{k+1} + \frac{z_i^k}{\beta}), & i = i_k, \\ y_i^k, & \text{otherwise}, \end{cases}$$

$$z_i^{k+1} = \begin{cases} z_i^k + \beta(x^{k+1} - y_i^{k+1}), & i = i_k \\ z_i^k, & \text{otherwise}. \end{cases}$$

• A walker will carry $\bar{x}$ while visiting a sequence of nodes
- Recall: among \( \left\{ y_1 - \frac{z_1}{\beta}, \ldots, y_n - \frac{z_n}{\beta} \right\} \), only \( y_i - \frac{z_{ik}}{\beta} \) is updated.

- The computation of \( \bar{x}^{k+2} \) is equivalent to

\[
\bar{x}^{k+2} = \bar{x}^{k+1} + \frac{1}{n} \left( y_{ik}^{k+1} - \frac{z_{ik}^{k+1}}{\beta} \right) - \frac{1}{n} \left( y_{ik}^{k} - \frac{z_{ik}^{k}}{\beta} \right)
\]  

(7)

Such computation can be conducted locally.
• It is expensive to solve subproblem

\[
\text{minimize } f_i(y) + \frac{\beta}{2} \|y - (x^{k+1} + \frac{z_i^k}{\beta})\|^2
\]  \hspace{1cm} (8)

• When (8) is not easy to solve, we can linearize (8) and update \( y_i \) cheaply

\[
y_i^{k+1} = \begin{cases} 
  x^{k+1} + \frac{1}{\beta} z_i^k - \frac{1}{\beta} \nabla f_i(y_i^k), & i = i_k \\
  y_i^k, & \text{otherwise.}
\end{cases}
\]  \hspace{1cm} (9)
A walker carries $\bar{x}^k$ around the network

Each local variable $y^k_i$ is expected to converge to $x^*$

The node activation is Markovian: node $i_{k+1}$ must be the neighbor of $i_k$. 

---

**Algorithm 1: Walkman**

**Initialization:** initialize $\sum_{i=1}^n (y^0_i - z^0_i / \beta) = 0$;

**Repeat** for $k = 0, 1, 2, \ldots$ until convergence

agent $i_k$ do:

- update $x^{k+1}$ according to (4);
- update $y_{i_k}^{k+1}$ according to (5) or (9);
- update $z_{i_k}^{k+1}$ according to (6);
- update $\bar{x}^{k+2}$ according to (7);
- send $\bar{x}^{k+2}$ via edge $(i_k, i_{k+1})$ to agent $i_{k+1}$;

**End**
Outline

1. Decentralized optimization

2. The Walkman method

3. Convergence

4. Communication analysis

5. Simulation results
Assumptions

Assumption (A1: Random walk)

Random walk \((i_k)_{k \geq 0}, i_k \in V\) forms an irreducible, aperiodic Markov chain with transition probability matrix \(P\) and stationary distribution \(\pi\).

This guarantees each agent to be visited for infinitely many times.

Assumption (A2: Coerciveness)

The objective function \(r(x) + \frac{1}{n} \sum_{i=1}^{n} f_i(x)\), is bounded from below over \(\mathbb{R}^p\) and is coercive over \(\mathbb{R}^p\), that is, \(r(x^k) + \frac{1}{n} \sum_{i=1}^{n} f_i(x^k) \to \infty\) for any sequence \(x^k \in \mathbb{R}^p\) and \(\|x^k\| \to \infty\).

There exists a bounded minimal function value.

The boundedness of \(x^k\) implies the boundedness of the function value.
**Assumptions**

<table>
<thead>
<tr>
<th>Assumption (A3: $f_i$ smoothness)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each $f_i(x)$ is $L$-Lipschitz differentiable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assumption (A4: $r$ is semiconvex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function $r(x)$ is $\gamma$-semiconvex, that is, $r(\cdot) + \frac{\gamma}{2} | \cdot |^2$ is convex.</td>
</tr>
</tbody>
</table>
Convergence property

Theorem

Under A1-A4, for $\beta > \max\{\gamma, 2L + 2\}$ (resp. $\beta > \max\{\gamma, 2L^2 + L + 2\}$), it holds that any limit point $(x^*, y^*, z^*)$ of the sequence $(x^k, y^k, z^k)$ generated by Walkman with $\text{prox}_{f_i}$ (resp. $\nabla f_i$) satisfies: $x^* = x_i^* = y_i^*$, $i = 1, \ldots, n$, where $x^*$ is a stationary point of (1), with probability 1, that is,

$$\Pr \left( 0 \in \partial r(x^*) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^*) \right) = 1.$$ 

If the objective of (1) is convex, then $x^*$ is a minimizer.

Implication: Walkman almost surely converges to a stationary point.
Convergence rate

- We examine the convergence rate for decentralized least squares

\[
\text{minimize } x \quad \frac{1}{2n} \sum_{i=1}^{n} \|A_i x - b_i\|^2
\]

This is a special case for problem (1) where \( r = 0 \).

- Node \( i \) possesses \( A_i \) and \( b_i \)

- We need the **mixing time** to characterize the convergence rate
Mixing time

- For $\delta > 0$, mixing time [16, Chapter 11] is defined as the smallest integer $\tau(\delta)$ such that

$$\left| \left[ P^{\tau(\delta)} \right]_{ij} - \pi_j \right| \leq \delta \pi_*, \quad \forall i, j \in V. \quad (10)$$

where $\pi_* := \min_{i \in V} \pi_i$

- After $\tau(\delta)$, each agent $j$ will be visited with probability $\approx \pi_j$.

- Inequality (10) is guaranteed when

$$\tau(\delta) := \left\lceil \frac{1}{1 - \lambda_2(P)} \ln \frac{\sqrt{2}}{\delta \pi_*} \right\rceil \quad (11)$$

where $\lambda_2(P) := \sup \left\{ \| f^T P \| / \| f \| : f^T 1 = 0, f \in \mathbb{R}^n \right\}$. 
Convergence rate

**Theorem**

Under A1, for $\beta > 2\sigma_{\text{max}}^* + 2$ with $\sigma_{\text{max}}^* := \max_i \sigma_{\text{max}}(A_i^T A_i)$, we have linear convergence:

$$\mathbb{E}\|y^t - y^*\|^2 \leq \left(1 + \frac{n(1 - \delta)\pi_*\nu}{4\beta^2\tau(\delta)}\right)^{-\left\lfloor \frac{t}{\tau(\delta)} \right\rfloor} C_0, \quad \forall \ t \geq 0,$$

where $\nu = \frac{(n-1)(\beta-\sigma_{\text{max}}^*)}{n^2}$, and $C_0$ is a constant only dependent on $A_1, \cdots, A_n, b_1, \cdots, b_n$, and $\beta$.

Quantity $\tau(\delta)$ behaves as the iteration numbers in an epoch.

Walkman solves the least squares problem at a linear convergence rate.
Outline

1. Decentralized optimization

2. The Walkman method

3. Convergence

4. Communication analysis

5. Simulation results
Communication complexity

- For simplicity, assume $P$ is a symmetric real matrix, modeling a gossip matrix of undirected graph.

- Communication complexity of Walkman:

\[
O\left(\frac{\ln \left(\frac{n}{\epsilon}\right)}{\ln \left(1 + \frac{(1 - \delta)\pi^*}{\tau(\delta)}\right)}\cdot \frac{\tau(\delta)}{\text{comm. per epoch}}\right)
\]

\[
\text{epoch numbers}
\]

- Substitute $\tau(\delta)$ with (11)

\[
O\left(\frac{\ln \left(\frac{n}{\epsilon}\right)}{\ln \left(1 + \frac{1 - \lambda_2(P)}{2n \ln(2n)}\right)}\cdot \frac{\ln(n)}{1 - \lambda_2(P)}\right)
\]

\[
\text{epoch numbers}
\]

\[
\text{comm. per epoch}
\]

The number of edge $m$ is not explicitly involved.
### Communication comparison

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Communication Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walkman (proposed)</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) \cdot \frac{n \ln^3(n)}{(1-\lambda_2(P))^2} \right)$</td>
</tr>
<tr>
<td>D-ADMM [14]</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) \cdot \left( \frac{m}{(1-\lambda_2(P))^{1/2}} \right) \right)$</td>
</tr>
<tr>
<td>EXTRA [5]</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) \cdot \left( \frac{m}{1-\lambda_2(P)} \right) \right)$</td>
</tr>
<tr>
<td>Exact diffusion [8]</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) \cdot \left( \frac{m}{1-\lambda_2(P)} \right) \right)$</td>
</tr>
</tbody>
</table>

- Walkman is more communication-efficient when:

  
  $$
  \lambda_2(P) \leq 1 - \frac{n^{2/3} \ln^2(n)}{m^{2/3}} \approx 1 - \left( \frac{n}{m} \right)^{2/3},
  $$

  where the approximation holds for $\ln(n) \ll n$ and with $\ln(n)$ ignored.

- When the graph is moderately well connected, Walkman is more communication-efficient.
Communication comparison on complete graph

- $m = O(n^2), \lambda_2(P) = 0$

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Communication Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walkman (proposed)</td>
<td>$O\left(\ln\left(\frac{1}{\epsilon}\right) n \ln^3(n)\right)$</td>
</tr>
<tr>
<td>D-ADMM [14]</td>
<td>$O\left(\ln\left(\frac{1}{\epsilon}\right) n^2\right)$</td>
</tr>
<tr>
<td>EXTRA [5]</td>
<td>$O\left(\ln\left(\frac{1}{\epsilon}\right) n^2\right)$</td>
</tr>
<tr>
<td>Exact diffusion [8]</td>
<td>$O\left(\ln\left(\frac{1}{\epsilon}\right) n^2\right)$</td>
</tr>
</tbody>
</table>
Communication comparison on random graph

- Random Graphs [17], $G(n, p)$: an $n$-node graph is generated with each edge populating independently with probability $p \in (0, 1)$. $P_{i,j} = \frac{1}{d_{\text{max}}}$ if node $i$ and $j$ are connected, and 0 otherwise, where $d_{\text{max}}$ is the maximum degree. $P_{i,i} = 1 - \sum_{j \neq i} P_{i,j}$.

- $\mathbb{E}[m] = \frac{p(n^2-n)}{2} = O(n^2)$, $1 - \lambda_2(P) = O(1)$.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Communication Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walkman (proposed)</td>
<td>$O\left(\ln \left(\frac{1}{\varepsilon}\right) n \ln^3(n)\right)$</td>
</tr>
<tr>
<td>D-ADMM [14]</td>
<td>$O\left(\ln \left(\frac{1}{\varepsilon}\right) n^2\right)$</td>
</tr>
<tr>
<td>EXTRA [5]</td>
<td>$O\left(\ln \left(\frac{1}{\varepsilon}\right) n^2\right)$</td>
</tr>
<tr>
<td>Exact diffusion [8]</td>
<td>$O\left(\ln \left(\frac{1}{\varepsilon}\right) n^2\right)$</td>
</tr>
</tbody>
</table>
Communication comparison on cycle graph

- $m = O(n), 1 - \lambda_2(P) = O\left(\frac{1}{n^2}\right),$

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Communication Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walkman (proposed)</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) n^5 \ln^3(n) \right)$</td>
</tr>
<tr>
<td>D-ADMM [14]</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) n^2 \right)$</td>
</tr>
<tr>
<td>EXTRA [5]</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) n^3 \right)$</td>
</tr>
<tr>
<td>Exact diffusion [8]</td>
<td>$O \left( \ln \left( \frac{1}{\epsilon} \right) n^3 \right)$</td>
</tr>
</tbody>
</table>
Outline

1. Decentralized optimization
2. The Walkman method
3. Convergence
4. Communication analysis
5. Simulation results
Convex Problem: Least Squares

\[
\text{minimize} \quad \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \| A_i x - b_i \|^2,
\]

- \( n = 50 \) nodes are uniformly placed in a \( 30 \times 30 \) square; \( r = 15 \).
- \( A_i \in \mathbb{R}^{5 \times 10}, \ x \in \mathbb{R}^{10} \) and \( b_i := A_i x_0 + v_i \).
Convex Problem: Sparsity Inducing Logistic Regression

\[
\text{minimize} \quad \lambda \|x\|_1 + \frac{1}{nb} \sum_{i=1}^{n} \sum_{j=1}^{b} \log \left( 1 + \exp \left( -y_{ij} v_{ij}^T x \right) \right),
\]

- \( v_{ij} \in \mathbb{R}^5 \) is the feature and \( y_{ij} \in \{-1, +1\} \) is the label; \( b = 10 \)
- \( v_{ij} \sim \mathcal{N}(0, 1) \)
Nonconvex Problem: Nonnegative Principal Component Analysis

\[
\begin{align*}
\text{minimize} \quad & x \in \{ x : \| x \| \leq 1, x_i \geq 0, \forall i \in \{1, \ldots, p\} \} \\
\text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^{n} -x^T \left( \frac{1}{b} \sum_{j=1}^{b} y_{ij} y_{ij}^T \right) x
\end{align*}
\]

- Walkman escapes from saddle point and reaches lower function value.
Summary

- Walkman converges exactly to the stationary point with fixed step-size
- Walkman is communication efficient than state-of-the-art methods
- Random-walk algorithms have great potential to save communications

**Acknowledgements:** NSF, NSFC

References:


