Accelerated and Sparse Algorithms for Personalized PageRank

David Martínez-Rubio
joint work with Elias Wirth and Sebastian Pokutta (art by DALL·E 3)

Technische Universität Berlin, Zuse Institute Berlin

ZUSE INSTITUTE BERLIN

What PageRank is about

- Rank all websites on the public internet.

What PageRank is about

- Rank all websites on the public internet.

But also...

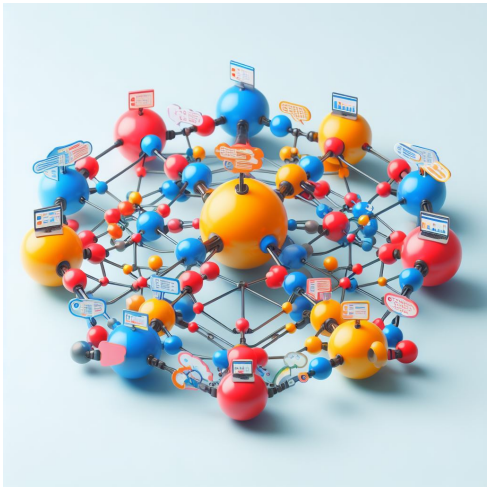
What PageRank is about

- Rank all websites on the public internet.

But also...

- **Cluster computation:** large intraconnectivity, low connectivity with the rest.

- Other data analyses on large graphs.



arXiv > cs > arXiv:1407.5107

Computer Science > Social and Information Networks

[Submitted on 18 Jul 2014]

PageRank beyond the Web

David F. Gleich

The Random Walker

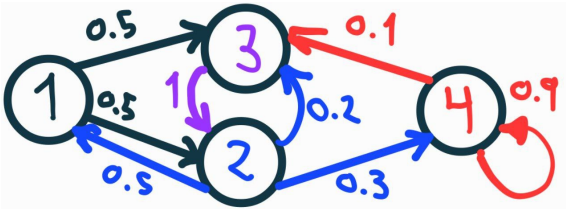
The Random Walker

- In PageRank websites are sorted by importance.
- Rank $\overset{prop.}{\propto}$ visit frequency of a random surfer on the internet.
- The random surfer moves uniformly at random from one page to the next one following hyperlinks.
- This is almost perfect. But we need to fix one issue.

Markov Chains

$$Q = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.2 & 0.3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

$$x_{t+1} = x_t^\top Q.$$

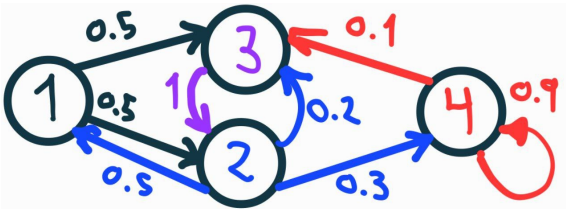


- The future is random, and only depends on the past through the current state.

Markov Chains

$$Q = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.2 & 0.3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

$$x_{t+1} = x_t^\top Q.$$

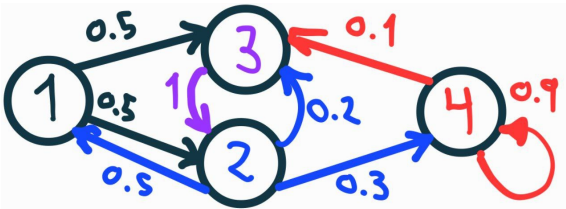


- The future is random, and only depends on the past through the current state.
- **Hitting time** h_{ij}: Expectation on the minimum time to reach j from i.

Markov Chains

$$Q = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.2 & 0.3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

$$x_{t+1} = x_t^\top Q.$$

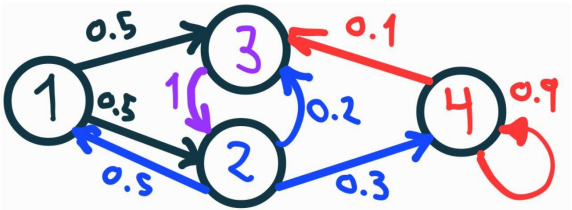


- The future is random, and only depends on the past through the current state.
- **Hitting time** h_{ij}: Expectation on the minimum time to reach j from i.
- **Stationary distribution** π: A distribution for which $\pi^\top Q = \pi$.

Markov Chains

$$Q = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.2 & 0.3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

$$x_{t+1} = x_t^\top Q.$$



- The future is random, and only depends on the past through the current state.
- **Hitting time** h_{ij}: Expectation on the minimum time to reach j from i.
- **Stationary distribution** π: A distribution for which $\pi^\top Q = \pi$.

Fundamental Theorem of Markov Chains:

A finite $\overbrace{\text{strongly connected}}^{\text{irreducible}}$ Markov chain has a unique stationary distribution π, where $\pi_j = 1/h_{j,j}$, for each state j.

[Perron-Frobenius theorem ; If also aperiodic, then $\lim_{n\to\infty} x_0^\top Q^n = \pi$; Sampling ; MCMC]

Teleportation distribution: Making everything strongly connected

- Problem: The internet graph is not strongly connected.

Teleportation distribution: Making everything strongly connected

- Problem: The internet graph is not strongly connected.
- Define a new Markov chain: with probability α we jump to a random node with distr. s. O/w we run the walk.
- Its transition matrix is $Q = (1-\alpha)AD^{-1} + \alpha \mathbb{1} s^\top$.
- Google originally used uniform $s = \mathbb{1}/n$ and $\alpha \approx 0.15$.

Teleportation distribution: Making everything strongly connected

- Problem: The internet graph is not strongly connected.
- Define a new Markov chain: with probability α we jump to a random node with distr. s. O/w we run the walk.
- Its transition matrix is $Q = (1-\alpha)AD^{-1} + \alpha \mathbb{1} s^\top$.
- Google originally used uniform $s = \mathbb{1}/n$ and $\alpha \approx 0.15$.

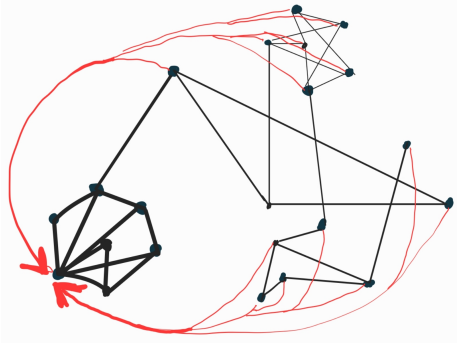
- PageRank Problem:

For Q stochastic and irreducible, compute $\pi \in \Delta^n$ such that $\pi^\top Q = \pi$.

- It is an eigenvector problem, a Principal Component Analysis (PCA) problem.
- But it has structure.

Local clustering with **Personalized** PageRank

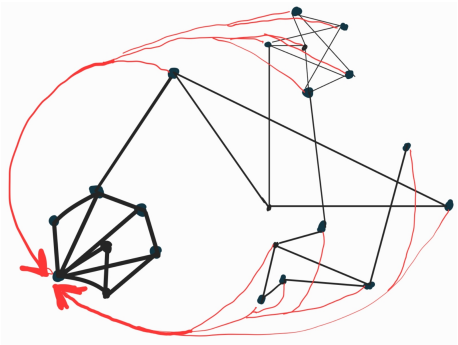
- If for undirected graphs our teleportation distr. is e_i, we can find a local cluster around node i.



Local clustering with **Personalized** PageRank

- If for undirected graphs our teleportation distr. is e_i, we can find a local cluster around node i.
- A is symmetric, so $x^\top(1-\alpha)AD^{-1} + \alpha e_i = x$ for $x \in \Delta^n$ can be cast as the quadratic minimization:

$$\min_{x \geq 0} \left\{ \frac{1}{2} x^\top x - \frac{1}{2} x^\top ((1-\alpha)AD^{-1})x - \alpha x^\top e_i \right\}.$$



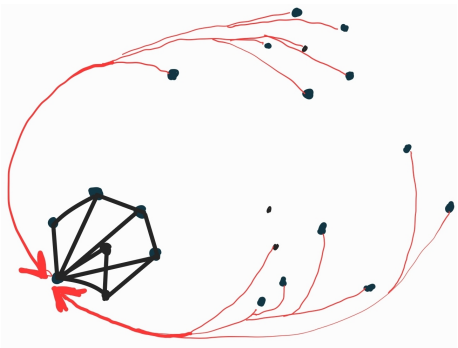
Local clustering with **Personalized** PageRank

- If for undirected graphs our teleportation distr. is e_i, we can find a local cluster around node i.
- A is symmetric, so $x^\top(1-\alpha)AD^{-1} + \alpha e_i = x$ for $x \in \Delta^n$ can be cast as the quadratic minimization:

$$\min_{x \geq 0} \left\{ \frac{1}{2} x^\top x - \frac{1}{2} x^\top ((1-\alpha)AD^{-1})x - \alpha x^\top e_i \right\}.$$

- Adding ℓ_1 regularization makes the solution sparse:

$$\min_{x \geq 0} \left\{ \frac{1}{2} x^\top x - \frac{1}{2} x^\top (1-\alpha)AD^{-1}x - \alpha x^\top e_i + \rho \|x\|_1 \right\}.$$



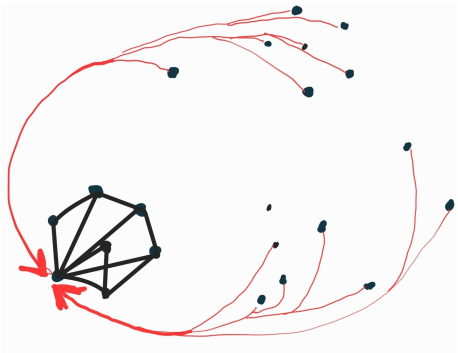
Local clustering with **Personalized** PageRank

- If for undirected graphs our teleportation distr. is e_i, we can find a local cluster around node i.
- A is symmetric, so $x^\top(1-\alpha)AD^{-1} + \alpha e_i = x$ for $x \in \Delta^n$ can be cast as the quadratic minimization:

$$\min_{x \geq 0} \left\{ \frac{1}{2} x^\top x - \frac{1}{2} x^\top ((1-\alpha)AD^{-1})x - \alpha x^\top e_i \right\}.$$

- Adding ℓ_1 regularization makes the solution sparse:

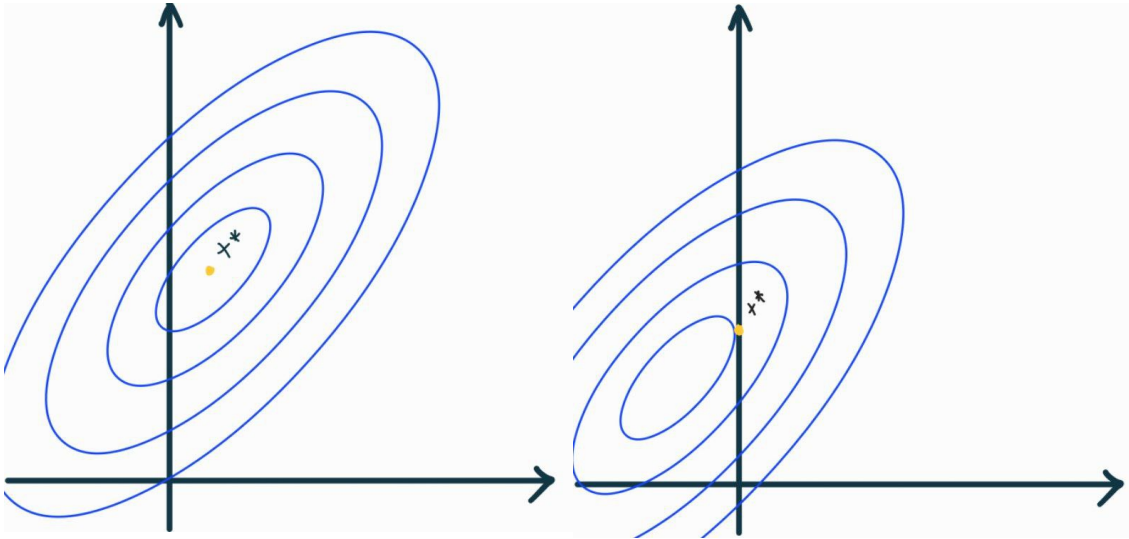
$$\min_{x \geq 0} \left\{ \frac{1}{2} x^\top x - \frac{1}{2} x^\top (1-\alpha)AD^{-1}x - \alpha x^\top e_i + \rho \|x\|_1 \right\}.$$



Massage the problem and you obtain, for symmetric Q s.t. $0 \prec \alpha \cdot I \preccurlyeq Q \preccurlyeq I$ and $Q_{ij} \leq 0$ for $i \neq j$:

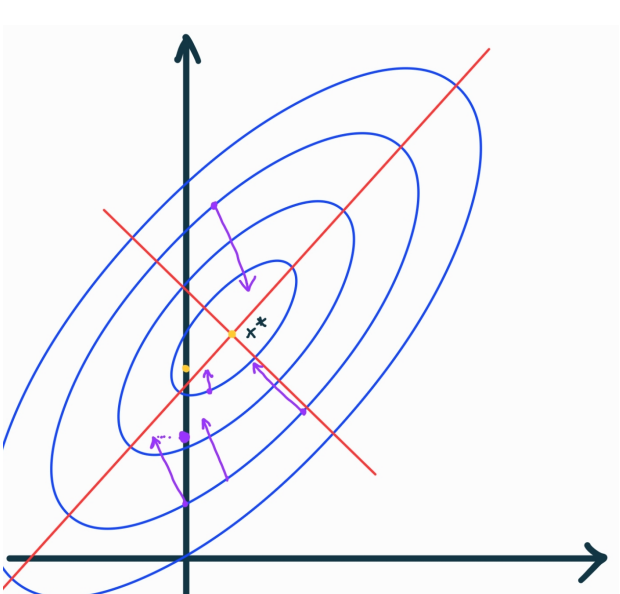
$$\min_{x \in \mathbb{R}^n_{\geq \mathbf{0}}} \{g(\mathrm{x}) \stackrel{\text{def}}{=} \frac{1}{2}\langle Q\mathrm{x}, \mathrm{x}\rangle - \langle \mathrm{b}, \mathrm{x}\rangle\}.$$

ℓ_1-regularization induces sparsity



Gradient Descent

- Gradient descent with step-size $\leq 1/\lambda_{\max}$ stays in the current eigenvector's orthant.
- We can show that GD restricted to the vertical axis is always coordinatewise lower than its minimizer.
- We can discover good coordinates one by one!



Coordinate discovery by approximate optimization

1. Because $Q_{ij} \leq 0$ for $i \neq j$, for $y = x + \Delta e_i$, we have $\forall j \neq i$:
   - $\nabla_j g(y) \leq \nabla_j g(x)$ if $\Delta > 0$
   - $\nabla_j g(y) \geq \nabla_j g(x)$ otherwise.

Coordinate discovery by approximate optimization

1. Because $Q_{ij} \leq 0$ for $i \neq j$, for $y = x + \Delta e_i$, we have $\forall j \neq i$:
   - $\nabla_j g(y) \leq \nabla_j g(x)$ if $\Delta > 0$
   - $\nabla_j g(y) \geq \nabla_j g(x)$ otherwise.
2. Recall, $\nabla_i g(x^{(*,C^{(t)})}) < 0$ only if i is good. So by 1., for $x \in C^{(t)}$ s.t. $x \leq x^{(*,C^{(t)})}$, new coordinates i can only satisfy $\nabla_i g(x) < 0$ if they are good.

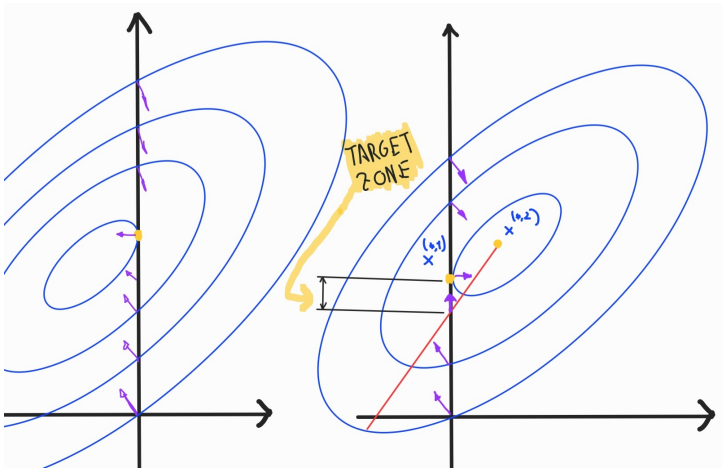


Figure: A negative coordinate gradient for a point $x \leq x^{(*,C^{(t)})}$ implies the coordinate is good, but not necessarily if $x \not\leq x^{(*,C^{(t)})}$.

Coordinate discovery by approximate optimization

1. Because $Q_{ij} \leq 0$ for $i \neq j$, for $\mathrm{y} = \mathrm{x} + \Delta \mathrm{e}_i$, we have $\forall j \neq i$:
   - $\nabla_j g(\mathrm{y}) \leq \nabla_j g(\mathrm{x})$ if $\Delta > 0$
   - $\nabla_j g(\mathrm{y}) \geq \nabla_j g(\mathrm{x})$ otherwise.
2. Recall, $\nabla_i g(\mathrm{x}^{(*, C^{(t)})}) < 0$ only if i is good. So by 1., for $x \in C^{(t)}$ s.t. $x \leq \mathrm{x}^{(*, C^{(t)})}$, new coordinates i can only satisfy $\nabla_i g(x) < 0$ if they are good.
3. **Strategy**: Get close to $\mathrm{x}^{(*, C^{(t)})}$ with Proj. AGD and then move slightly towards $\mathbb{0}$ to be $\leq \mathrm{x}^{(*, C^{(t)})}$.

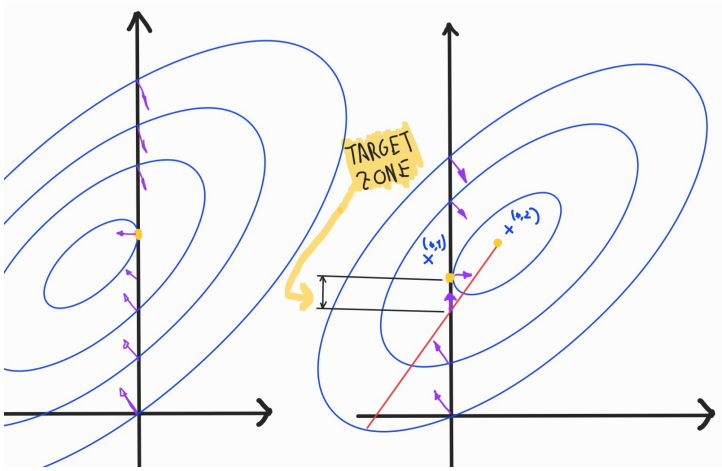


Figure: A negative coordinate gradient for a point $x \leq \mathrm{x}^{(*, C^{(t)})}$ implies the coordinate is good, but not necessarily if $x \not\leq \mathrm{x}^{(*, C^{(t)})}$.

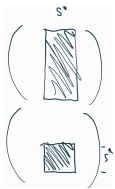
Results and comparison

- The Hessian of g is Q, satisfying $\mu I \preccurlyeq Q \preccurlyeq LI$, its condition number is L/μ.
- $\mathcal{S}^* \overset{\text{def}}{=} \text{supp}(x^*)$, $\text{vol}(\mathcal{S}^*) \overset{\text{def}}{=} \text{nnz}(Q_{:,\mathcal{S}^*})$ and $\widetilde{\text{vol}}(\mathcal{S}^*) \overset{\text{def}}{=} \text{nnz}(Q_{\mathcal{S}^*,\mathcal{S}^*})$.
- For ℓ_1-regularized personalized PageRank, it is $\text{vol}(\mathcal{S}^*) \leq \frac{1}{\rho} + |\mathcal{S}^*|$ [FRS+19].

Method	Time complexity	Space complexity
`ISTA` [FRS+19]	$\widetilde{\mathcal{O}}(\text{vol}(\mathcal{S}^*)\frac{L}{\mu})$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
`CDPR` **(Ours)**	$\mathcal{O}(\lvert\mathcal{S}^*\rvert^3 + \lvert\mathcal{S}^*\rvert\text{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert^2)$
`ASPR` **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\widetilde{\text{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\mu}} + \lvert\mathcal{S}^*\rvert\text{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$

$\text{vol}(\mathcal{S}^*) \overset{\text{def}}{=}$

$\widetilde{\text{vol}}(\mathcal{S}^*) \overset{\text{def}}{=}$



An exact algorithm: Conjugate Directions for PageRank (CDPR)

- Start at $x^{(0)} = \mathbb{0}$.

An exact algorithm: Conjugate Directions for PageRank (CDPR)

- Start at $x^{(0)} = \mathbb{0}$.

- For $t > 0$, define the set of new good coordinates $N^{(t)} \stackrel{\text{def}}{=} \{i \in [n] \mid \nabla_i g(x^{(t)}) < 0\}$ and select $i \in N^{(t)}$, $u^{(t)} \stackrel{\text{def}}{=} \nabla_i g(x^{(t)})\, e_i$.

An exact algorithm: Conjugate Directions for PageRank (CDPR)

- Start at $\mathrm{x}^{(0)} = \mathbb{0}$.
- For $t > 0$, define the set of new good coordinates $N^{(t)} \stackrel{\text{def}}{=} \{i \in [n] \mid \nabla_i g(\mathrm{x}^{(t)}) < 0\}$ and select $i \in N^{(t)}$, $\mathrm{u}^{(t)} \stackrel{\text{def}}{=} \nabla_i g(\mathrm{x}^{(t)})\, \mathrm{e}_i$.
- Compute direction $\mathrm{d}^{(t)}$ from $\mathrm{u}^{(t)}$ by Q-Gram-Schmidt using all previous (sparse) directions so $\langle \mathrm{d}^{(t)}, Q\mathrm{d}^{(k)} \rangle = 0$ for all $k < t$.

An exact algorithm: Conjugate Directions for PageRank (CDPR)

- Start at $\mathsf{x}^{(0)} = \mathbb{0}$.
- For $t > 0$, define the set of new good coordinates $N^{(t)} \stackrel{\text{def}}{=} \{i \in [n] \mid \nabla_i g(\mathsf{x}^{(t)}) < 0\}$ and select $i \in N^{(t)}$, $\mathsf{u}^{(t)} \stackrel{\text{def}}{=} \nabla_i g(\mathsf{x}^{(t)})\, \mathsf{e}_i$.
- Compute direction $\mathsf{d}^{(t)}$ from $\mathsf{u}^{(t)}$ by Q-Gram-Schmidt using all previous (sparse) directions so $\langle \mathsf{d}^{(t)}, Q\mathsf{d}^{(k)} \rangle = 0$ for all $k < t$.
- Optimize on the line $\mathsf{x}^{(t+1)} \leftarrow \arg\min_{\eta^{(t)}} \{\mathsf{x}^{(t)} + \eta^{(t)} \mathsf{d}^{(t)}\}$. It is $\mathsf{x}^{(t+1)} = \mathsf{x}^{(*, C^{(t)})}$.

An exact algorithm: Conjugate Directions for PageRank (CDPR)

- Start at $\mathsf{x}^{(0)} = \mathbb{0}$.
- For $t > 0$, define the set of new good coordinates $N^{(t)} \stackrel{\text{def}}{=} \{i \in [n] \mid \nabla_i g(\mathsf{x}^{(t)}) < 0\}$ and select $i \in N^{(t)}$, $\mathsf{u}^{(t)} \stackrel{\text{def}}{=} \nabla_i g(\mathsf{x}^{(t)})\, \mathsf{e}_i$.
- Compute direction $\mathsf{d}^{(t)}$ from $\mathsf{u}^{(t)}$ by Q-Gram-Schmidt using all previous (sparse) directions so $\langle \mathsf{d}^{(t)}, Q\mathsf{d}^{(k)} \rangle = 0$ for all $k < t$.
- Optimize on the line $\mathsf{x}^{(t+1)} \leftarrow \arg\min_{\eta^{(t)}} \{\mathsf{x}^{(t)} + \eta^{(t)} \mathsf{d}^{(t)}\}$. It is $\mathsf{x}^{(t+1)} = \mathsf{x}^{(*, C^{(t)})}$.
- Time complexity $\mathcal{O}(|\mathcal{S}^*|^3 + |\mathcal{S}^*|\mathrm{vol}(\mathcal{S}^*))$ and space complexity $\mathcal{O}(|\mathcal{S}^*|^2)$.

Accelerated and Sparse PageRank (ASPR) algorithm

- **Lemma**. Let $\bar{x}^{(t+1)}$ be an $\varepsilon \cdot \frac{\mu^2}{2(1+|S^{(t)}|)L^2}$ minimizer in $C^{(t)}$. Define $x^{(t+1)} \leftarrow \mathrm{Proj}_{\mathbb{R}^n_{\geq 0}}(\bar{x}^{(t+1)} - \delta_t \mathbb{1})$ for $\delta_t = \sqrt{\frac{\varepsilon\mu}{(1+|S^{(t)}|)L^2}}$. Then, $x^{(t+1)} \leq x^{(*,C^{(t)})}$ and $x^{(t+1)}$ is a global ε-minimizer or there is i s.t. $\nabla_i g(x^{(t+1)}) < 0$, so we expand the current set of good coordinates $S^{(t)}$.

Accelerated and Sparse PageRank (ASPR) algorithm

- **Lemma**. Let $\bar{x}^{(t+1)}$ be an $\varepsilon \cdot \frac{\mu^2}{2(1+|S^{(t)}|)L^2}$ minimizer in $C^{(t)}$. Define $x^{(t+1)} \leftarrow \mathrm{Proj}_{\mathbb{R}^n_{\geq 0}}(\bar{x}^{(t+1)} - \delta_t \mathbb{1})$ for $\delta_t = \sqrt{\frac{\varepsilon\mu}{(1+|S^{(t)}|)L^2}}$. Then, $x^{(t+1)} \leq x^{(*,C^{(t)})}$ and $x^{(t+1)}$ is a global ε-minimizer or there is i s.t. $\nabla_i g(x^{(t+1)}) < 0$, so we expand the current set of good coordinates $S^{(t)}$.

- **Intuition**. $x^{(t+1)}$ is almost optimal in $C^{(t)}$, so if its global gap is $> \varepsilon$ then 1 step of GD makes more progress than what it is possible in $C^{(t)}$. $\implies \exists i \notin S^{(t)}$ s.t. $\nabla_i g(x^{(t+1)}) < 0$.

Accelerated and Sparse PageRank (ASPR) algorithm

- **Lemma**. Let $\bar{x}^{(t+1)}$ be an $\varepsilon \cdot \frac{\mu^2}{2(1+|S^{(t)}|)L^2}$ minimizer in $C^{(t)}$. Define $x^{(t+1)} \leftarrow \mathrm{Proj}_{\mathbb{R}^n_{\geq 0}}(\bar{x}^{(t+1)} - \delta_t \mathbb{1})$ for $\delta_t = \sqrt{\frac{\varepsilon\mu}{(1+|S^{(t)}|)L^2}}$. Then, $x^{(t+1)} \leq x^{(*,C^{(t)})}$ and $x^{(t+1)}$ is a global ε-minimizer or there is i s.t. $\nabla_i g(x^{(t+1)}) < 0$, so we expand the current set of good coordinates $S^{(t)}$.

- **Intuition**. $x^{(t+1)}$ is almost optimal in $C^{(t)}$, so if its global gap is $> \varepsilon$ then 1 step of GD makes more progress than what it is possible in $C^{(t)}$. $\implies \exists i \notin S^{(t)}$ s.t. $\nabla_i g(x^{(t+1)}) < 0$.

- Subproblem optimization only needs gradients in $C^{(t)}$, costing $\mathcal{O}(\widetilde{\mathrm{vol}}(\mathcal{S}^*))$ each. And one full gradient is used at the end of each stage to find new good coordinates, costing $\mathcal{O}(\mathrm{vol}(\mathcal{S}^*))$. It is done at most $|\mathcal{S}^*|$ times.

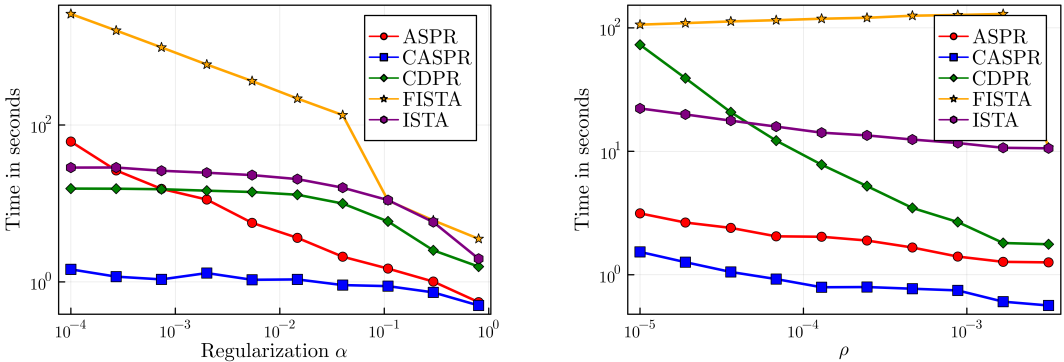
- Time complexity $\widetilde{\mathcal{O}}(|\mathcal{S}^*|\widetilde{\mathrm{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\mu}} + |\mathcal{S}^*|\mathrm{vol}(\mathcal{S}^*))$ and space complexity $\mathcal{O}(|\mathcal{S}^*|)$.

Comparisons and other results

Method	Time complexity	Space complexity
ISTA [FRS+19]	$\widetilde{\mathcal{O}}(\mathrm{vol}(\mathcal{S}^*)\frac{L}{\mu})$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
CDPR **(Ours)**	$\mathcal{O}(\lvert\mathcal{S}^*\rvert^3 + \lvert\mathcal{S}^*\rvert\mathrm{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert^2)$
ASPR **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\widetilde{\mathrm{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\mu}} + \lvert\mathcal{S}^*\rvert\mathrm{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
CASPR **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\widetilde{\mathrm{vol}}(\mathcal{S}^*)\min\left\{\sqrt{\frac{L}{\mu}}, \lvert\mathcal{S}^*\rvert\right\} + \lvert\mathcal{S}^*\rvert\mathrm{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
LASPR **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\mathrm{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$

Experiments

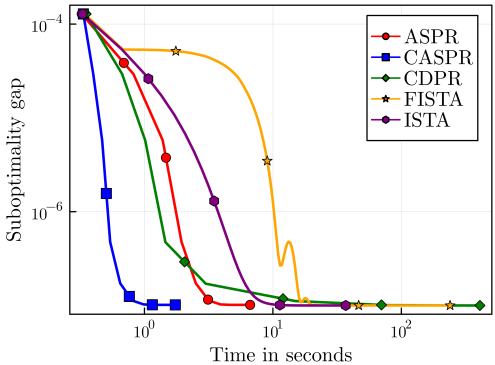
Results from a Standford Network Analysis Project graph with 3.7M nodes and 16.5M edges.

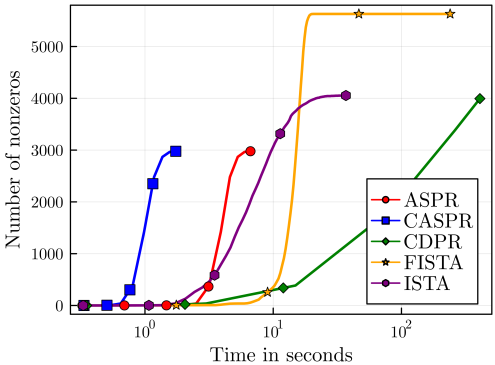


Left: Time taken to optimize to 10^{-6} accuracy, while fixing $\rho = 10^{-4}$ and varying the regularization α.
Right: Time taken to optimize to 10^{-6} accuracy, while fixing $\alpha = 0.05$ and varying ρ.

Experiments

Results from a Standford Network Analysis Project graph with 3.7M nodes and 16.5M edges.



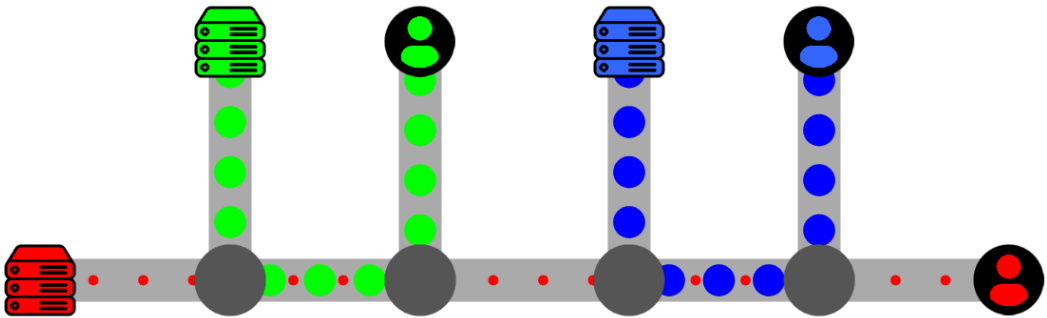


Left: Gap versus time.

Right: Number of non-zeros of the iterates with time. We obtain greater sparsity. This is due to the algorithms optimizing in the space of currently known good coordinates before adding new ones.

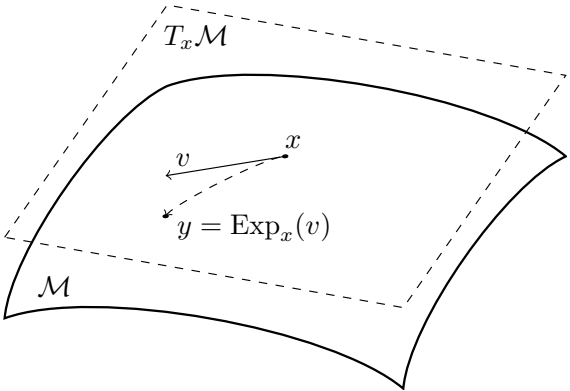
Other research: Packing Proportional Fairness

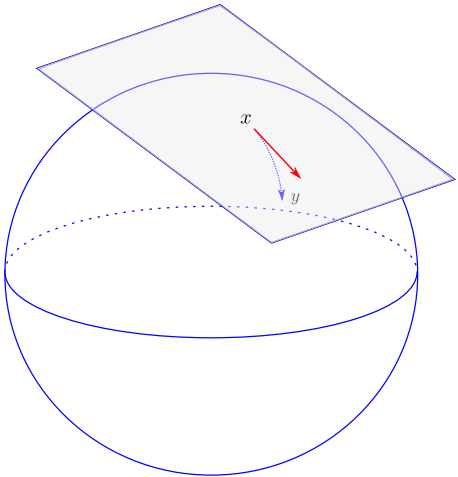
- Pairs server-user in a shared network with limited link capacities.
- How much flow should each pair receive, while satisfying proportional fairness axioms?



Other research: Riemannian Optimization in Machine Learning

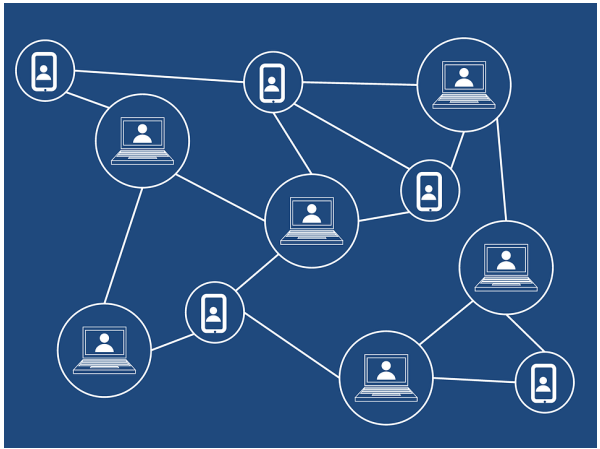
- How to approximate $\min_{x \in \mathcal{M}} f(x)$ for a Riemannian manifold $\mathcal{M}$?
- Mixture of Gaussians, operator scaling, dictionary learning, low-rank matrix completion, RNNs...

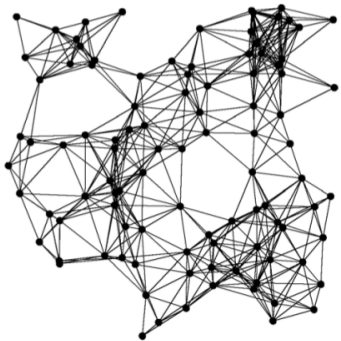




Other research: Decentralized Cooperative Stochastic Bandits

- Decentralized network, each node can only communicate to their neighbors.
- They are facing the same stochastic multi-armed bandit problem. How to behave and share information to minimize regret?





Thank you!

Questions?