Accelerated and Sparse Algorithms for Approximate Personalized PageRank

David Martínez-Rubio
joint work with Elias Wirth, Sebastian Pokutta

Technische Universität Berlin, Zuse Institute Berlin

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$AD^{-1}x = x.$$

Stationary distribution of random walk.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$\frac{1}{2}(I + AD^{-1})x = x.$$

Stationary distribution of lazy random walk.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$\left((1-\alpha)\frac{1}{2}(I + AD^{-1}) + \alpha s \mathbb{1}^T\right) x = x.$$

Add teleportation distribution (ensures uniqueness if the resulting graph is strongly connected).

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$(1-\alpha)\frac{1}{2}(I + AD^{-1})x + \alpha s = x.$$

Use x is a distribution.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$0 = y^T \left(\alpha I + \frac{1-\alpha}{2} \mathcal{L} \right) y - \alpha y^T \left(D^{-1/2} \mathrm{s} \right)$$

Reformulate as (approximately) solving a quadratic problem. Reparametrize $x = D^{1/2} y$.

$$Q \stackrel{\text{def}}{=} \alpha I + \frac{1-\alpha}{2} \mathcal{L} \qquad \text{and} \qquad \mathrm{b} \stackrel{\text{def}}{=} \alpha \left(D^{-1/2} \mathrm{s} \right)$$

where $\mathcal{L} \stackrel{\text{def}}{=} I - D^{-1/2} A D^{-1/2}$ is G's symmetric normalized Laplacian, and is $0 \preccurlyeq \mathcal{L} \preccurlyeq 2I$.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$y^T \left(\alpha I + \frac{1-\alpha}{2}\mathcal{L}\right) y - \alpha y^T \left(D^{-1/2}\mathrm{s}\right) + \alpha\rho\|D^{1/2}y\|_1$$

Add ℓ_1-regularization to induce sparsity.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$y^T \left(\alpha I + \frac{1-\alpha}{2}\mathcal{L}\right) y - \alpha y^T \left(D^{-1/2}\mathrm{s} - \rho D^{1/2}\right)$$

Use $y \in \mathbb{R}_{\geq 0}$ and simplify.

$$Q \stackrel{\text{def}}{=} \alpha I + \frac{1-\alpha}{2}\mathcal{L} \qquad \text{and} \qquad \mathrm{b} \stackrel{\text{def}}{=} \alpha \left(D^{-1/2}\mathrm{s} - \rho D^{1/2}\mathbb{1}\right)$$

where $\alpha, \rho > 0$, $\mathcal{L} \stackrel{\text{def}}{=} I - D^{-1/2}AD^{-1/2}$ is G's symmetric normalized Laplacian, and is $0 \preccurlyeq \mathcal{L} \preccurlyeq 2I$.

Problem

Problem:

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

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.
It includes ℓ_1-regularized personalized undirected PageRank on graph G, used in local clustering.

$$y^T \left(\alpha I + \frac{1-\alpha}{2} \mathcal{L} \right) y - \alpha y^T \left(D^{-1/2}\mathrm{s} - \rho D^{1/2} \right)$$

Use $y \in \mathbb{R}_{\geq 0}$ and simplify.

$$Q \stackrel{\text{def}}{=} \alpha I + \frac{1-\alpha}{2}\mathcal{L} \qquad \text{and} \qquad \mathrm{b} \stackrel{\text{def}}{=} \alpha \left(D^{-1/2}\mathrm{s} - \rho D^{1/2} \mathbb{1} \right)$$

where $\alpha, \rho > 0$, $\mathcal{L} \stackrel{\text{def}}{=} I - D^{-1/2} A D^{-1/2}$ is G's symmetric normalized Laplacian, and is $0 \preccurlyeq \mathcal{L} \preccurlyeq 2I$.

COLT 2022 Open Problem: Can we solve this in an accelerated way without depending on the size of the graph?

Results and comparison

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

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)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\widetilde{\mathrm{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\mu}} + \lvert\mathcal{S}^*\rvert\mathrm{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$

A geometric lemma

Suppose:

- $x^{(0)} \in \mathbb{R}^n_{\geq 0}$ and $S \subseteq [n]$ s.t. $x_i^{(0)} = 0$ if $i \notin S$ and $\nabla_i g(x^{(0)}) \leq 0$ if $i \in S$.
- $C \stackrel{\text{def}}{=} \text{span}(\{e_i \mid i \in S\}) \cap \mathbb{R}^n_{\geq 0}$.
- $x^{(*,C)} \stackrel{\text{def}}{=} \arg\min_{x \in C} g(x)$ and $x^* \stackrel{\text{def}}{=} \arg\min_{x \in \mathbb{R}^n_{\geq 0}} g(x)$.

A geometric lemma

Suppose:

- $\mathrm{x}^{(0)} \in \mathbb{R}^n_{\geq 0}$ and $S \subseteq [n]$ s.t. $x_i^{(0)} = 0$ if $i \notin S$ and $\nabla_i g(\mathrm{x}^{(0)}) \leq 0$ if $i \in S$.
- $C \stackrel{\text{def}}{=} \mathrm{span}(\{\mathrm{e}_i \mid i \in S\}) \cap \mathbb{R}^n_{\geq 0}$.
- $\mathrm{x}^{(*,C)} \stackrel{\text{def}}{=} \arg\min_{\mathrm{x} \in C} g(\mathrm{x})$ and $\mathrm{x}^* \stackrel{\text{def}}{=} \arg\min_{\mathrm{x} \in \mathbb{R}^n_{\geq 0}} g(\mathrm{x})$.

Then:

1. It holds that $\mathrm{x}^{(0)} \leq \mathrm{x}^{(*,C)}$ and $\nabla_i g(\mathrm{x}^{(*,C)}) = 0$ for all $i \in S$.
2. If for $i \in S$, we have $x_i^{(0)} > 0$ or $\nabla_i g(\mathrm{x}^{(0)}) < 0$, then $x_i^{(*,C)} > 0$.
3. If $x_i^{(*,C)} > 0$ for all $i \in S$, we have $\mathrm{x}^{(*,C)} \leq \mathrm{x}^*$ and therefore $S \subseteq \mathcal{S}^*$.

A geometric lemma

Suppose:

- $\mathrm{x}^{(0)} \in \mathbb{R}^n_{\geq 0}$ and $S \subseteq [n]$ s.t. $x_i^{(0)} = 0$ if $i \notin S$ and $\nabla_i g(\mathrm{x}^{(0)}) \leq 0$ if $i \in S$.
- $C \stackrel{\text{def}}{=} \operatorname{span}(\{\mathrm{e}_i \mid i \in S\}) \cap \mathbb{R}^n_{\geq 0}$.
- $\mathrm{x}^{(*,C)} \stackrel{\text{def}}{=} \arg\min_{\mathrm{x} \in C} g(\mathrm{x})$ and $\mathrm{x}^* \stackrel{\text{def}}{=} \arg\min_{\mathrm{x} \in \mathbb{R}^n_{\geq \mathbf{0}}} g(\mathrm{x})$.

Then:

1. It holds that $\mathrm{x}^{(0)} \leq \mathrm{x}^{(*,C)}$ and $\nabla_i g(\mathrm{x}^{(*,C)}) = 0$ for all $i \in S$.
2. If for $i \in S$, we have $x_i^{(0)} > 0$ or $\nabla_i g(\mathrm{x}^{(0)}) < 0$, then $x_i^{(*,C)} > 0$.
3. If $x_i^{(*,C)} > 0$ for all $i \in S$, we have $\mathrm{x}^{(*,C)} \leq \mathrm{x}^*$ and therefore $S \subseteq \mathcal{S}^*$.

Proof of 1.: $\bar{g} \stackrel{\text{def}}{=} g$ restricted to $\operatorname{span}(\{\mathrm{e}_i \mid i \in S\})$. Let $\{\mathrm{x}^{(t)}\}_{t=0}^{\infty}$ be the iterates of $\mathtt{PGD}(C, \mathrm{x}^{(0)}, \bar{g})$.
We start with $\nabla \bar{g}(\mathrm{x}^{(0)}) \leq 0$. By induction:

$$\mathrm{x}^{(t+1)} = \mathrm{x}^{(t)} - \underbrace{1/L \nabla \bar{g}(\mathrm{x}^{(t)})}_{\leq 0} \geq \mathrm{x}^{(t)} \text{ and } \nabla \bar{g}(\mathrm{x}^{(t+1)}) = \underbrace{\nabla \bar{g}(\mathrm{x}^{(t)})}_{\leq 0} \cdot \underbrace{(I - 1/L Q_{S,S})}_{\geq 0} \leq 0,$$

$\mathrm{x}^{(t)} \to \mathrm{x}^{(*,C)}$, $\nabla \bar{g}(\mathrm{x}^{(t)}) \to \nabla \bar{g}(\mathrm{x}^{(*,C)})$ (so ≤ 0, and by optimality it is ≥ 0.)

A geometric lemma

Suppose:

- $x^{(0)} \in \mathbb{R}^n_{\geq 0}$ and $S \subseteq [n]$ s.t. $x_i^{(0)} = 0$ if $i \notin S$ and $\nabla_i g(x^{(0)}) \leq 0$ if $i \in S$.
- $C \stackrel{\text{def}}{=} \operatorname{span}(\{e_i \mid i \in S\}) \cap \mathbb{R}^n_{\geq 0}$.
- $x^{(*,C)} \stackrel{\text{def}}{=} \arg\min_{x \in C} g(x)$ and $x^* \stackrel{\text{def}}{=} \arg\min_{x \in \mathbb{R}^n_{\geq 0}} g(x)$.

Then:

1. It holds that $x^{(0)} \leq x^{(*,C)}$ and $\nabla_i g(x^{(*,C)}) = 0$ for all $i \in S$.
2. If for $i \in S$, we have $x_i^{(0)} > 0$ or $\nabla_i g(x^{(0)}) < 0$, then $x_i^{(*,C)} > 0$.
3. If $x_i^{(*,C)} > 0$ for all $i \in S$, we have $x^{(*,C)} \leq x^*$ and therefore $S \subseteq \mathcal{S}^*$.

Proof of 2.: We have that $x_i^{(1)} > 0$ by the assumption on $x_i^{(0)}$ and the PGD update rule. By the monotonicity of iterates in the proof of 1., we obtain the result.

Proof of 3.: Sketch: Apply 1. and 2. to the initial point $x^{(*,C)}$ and set of indices $S \cup \{i \mid \nabla_i g(x^{(*,C)}) < 0\}$ and then again and so on until you get to x^*.

Algorithmic scheme

- **Definition:** i is a good coordinate iff $i \in \mathcal{S}^*$. Otherwise it is bad.

Algorithmic scheme

- **Definition:** i is a good coordinate iff $i \in \mathcal{S}^*$. Otherwise it is bad.

- **Idea for an algorithm:** discover good coordinates sequentially, by optimizing in the subspace $C^{(t)} \stackrel{\text{def}}{=} \operatorname{span}(\{e_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, where $S^{(t)}$ is the set of currently known good coordinates.

Algorithmic scheme

- **Definition:** i is a good coordinate iff $i \in \mathcal{S}^*$. Otherwise it is bad.

- **Idea for an algorithm:** discover good coordinates sequentially, by optimizing in the subspace $C^{(t)} \stackrel{\text{def}}{=} \operatorname{span}(\{e_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, where $S^{(t)}$ is the set of currently known good coordinates.

- By the geometric lemma, at the minimizer $x^{(*,t+1)} \stackrel{\text{def}}{=} x^{(*,C^{(t)})}$ we have $\nabla_i g(x^{(*,t+1)}) < 0$ only if i is good and new, i.e., only if $i \in \mathcal{S}^* \setminus S^{(t)}$.

Algorithmic scheme

- **Definition:** i is a good coordinate iff $i \in \mathcal{S}^*$. Otherwise it is bad.

- **Idea for an algorithm:** discover good coordinates sequentially, by optimizing in the subspace $C^{(t)} \stackrel{\text{def}}{=} \operatorname{span}(\{\mathrm{e}_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, where $S^{(t)}$ is the set of currently known good coordinates.

- By the geometric lemma, at the minimizer $\mathrm{x}^{(*,t+1)} \stackrel{\text{def}}{=} \mathrm{x}^{(*,C^{(t)})}$ we have $\nabla_i g(\mathrm{x}^{(*,t+1)}) < 0$ only if i is good and new, i.e., only if $i \in \mathcal{S}^* \setminus S^{(t)}$.

- An approximate version of this holds, after overcoming some technicalities.

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 $x^{(0)} = \mathbb{0}$.
- 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 $i \in N^{(t)}$, $u^{(t)} \stackrel{\text{def}}{=} \nabla_i g(x^{(t)})\, e_i$.
- Compute direction $d^{(t)}$ from $u^{(t)}$ by Q-Gram-Schmidt using all previous (sparse) directions so $\langle d^{(t)}, Qd^{(k)} \rangle = 0$ for all $k < t$.
- Optimize on the line $x^{(t+1)} \leftarrow \arg\min_{\eta^{(t)}} \{x^{(t)} + \eta^{(t)} d^{(t)}\}$. It is $x^{(t+1)} = x^{(*, C^{(t)})}$.

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 $k < t$.
- Optimize on the line $\mathrm{x}^{(t+1)} \leftarrow \arg\min_{\eta^{(t)}} \{\mathrm{x}^{(t)} + \eta^{(t)} \mathrm{d}^{(t)}\}$. It is $\mathrm{x}^{(t+1)} = \mathrm{x}^{(*, C^{(t)})}$.
- Time complexity $\mathcal{O}(|\mathcal{S}^*|^3 + |\mathcal{S}^*| \mathrm{vol}(\mathcal{S}^*))$ and space complexity $\mathcal{O}(|\mathcal{S}^*|^2)$.

An inexact algorithm: Accelerated and Sparse PageRank (ASPR)

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

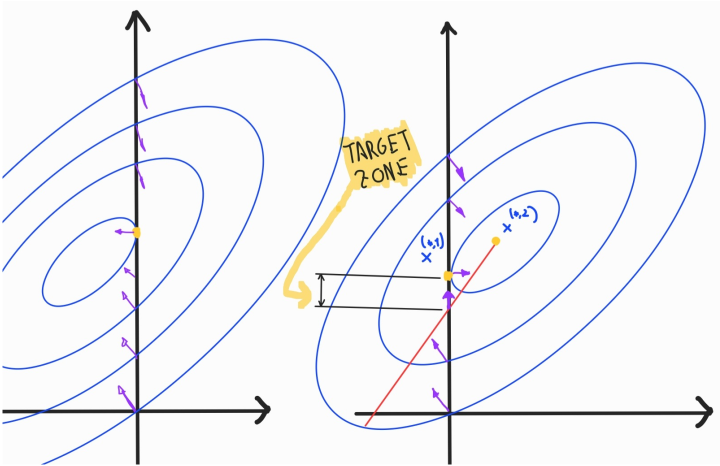


Figure: A negative coordinate gradient for a point $x \leq \mathrm{x}^{(*,C^{(t)})}$ implies the coordinate is good, but not necessarily if $x \not\leq \mathrm{x}^{(*,C^{(t)})}$.

An inexact algorithm: Accelerated and Sparse PageRank (ASPR)

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

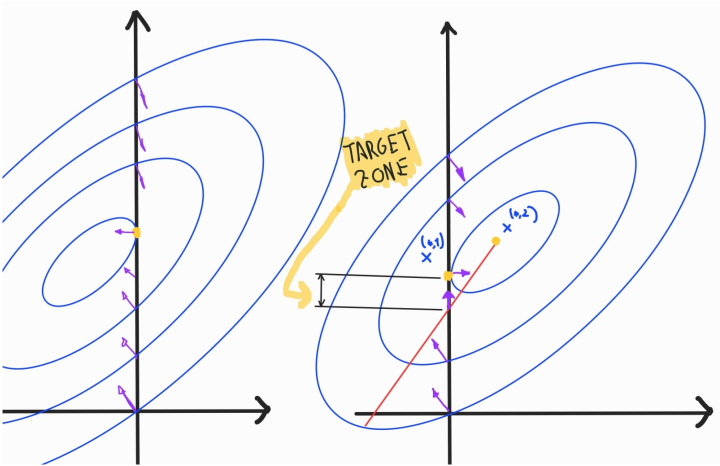


Figure: A negative coordinate gradient for a point $x \leq \mathrm{x}^{(*,C^{(t)})}$ implies the coordinate is good, but not necessarily if $x \not\leq \mathrm{x}^{(*,C^{(t)})}$.

An inexact algorithm: Accelerated and Sparse PageRank (ASPR)

1. Because $Q_{ij} \leq 0$ for $i \neq j$, for $\mathrm{y} = \mathrm{x} - \Delta \mathrm{e}_i$, we have $\forall j \neq i$: $\nabla_j g(\mathrm{y}) \geq \nabla_j g(\mathrm{x})$ if $\Delta > 0$ and $\nabla_j g(\mathrm{y}) \leq \nabla_j g(\mathrm{x})$ 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 $\leq \mathrm{x}^{(*,C^{(t)})}$.

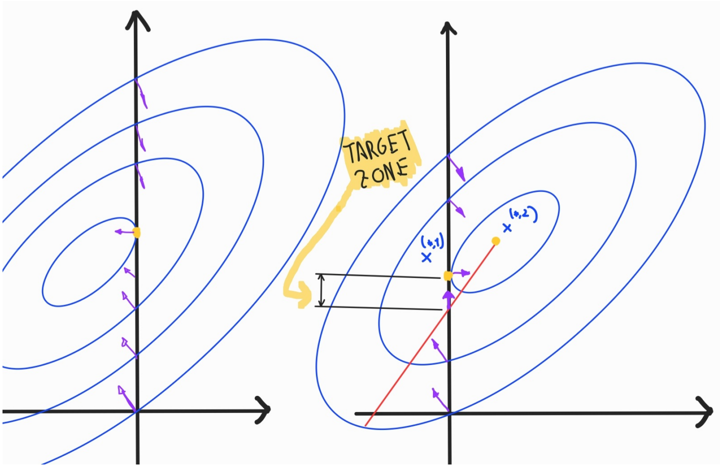


Figure: A negative coordinate gradient for a point $x \leq \mathrm{x}^{(*,C^{(t)})}$ implies the coordinate is good, but not necessarily if $x \not\leq \mathrm{x}^{(*,C^{(t)})}$.

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

Accelerated and Sparse PageRank (ASPR) algorithm

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

- **Intuition**. $\mathrm{x}^{(t+1)}$ 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. $\nabla_i g(\mathrm{x}^{(t+1)}) < 0$.

- 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.

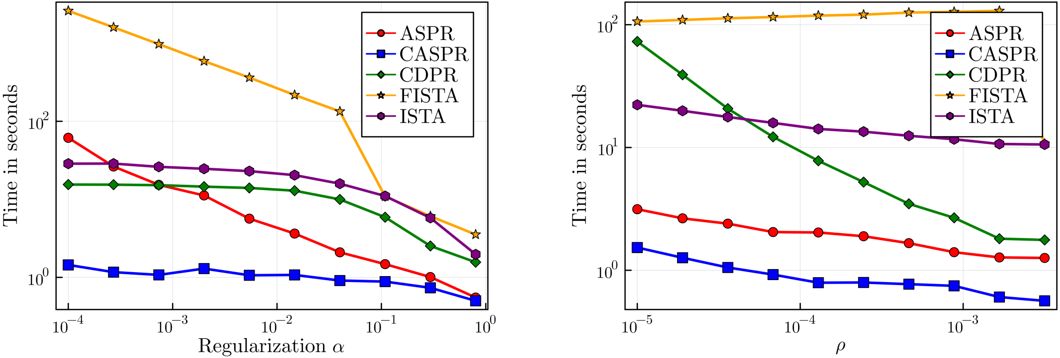
- Time complexity $\widetilde{\mathcal{O}}(|\mathcal{S}^*|\widetilde{\mathrm{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\mu}} + |\mathcal{S}^*|\mathrm{vol}(\mathcal{S}^*))$ and space complexity $\mathcal{O}(|\mathcal{S}^*|)$.

Comparisons and other results

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

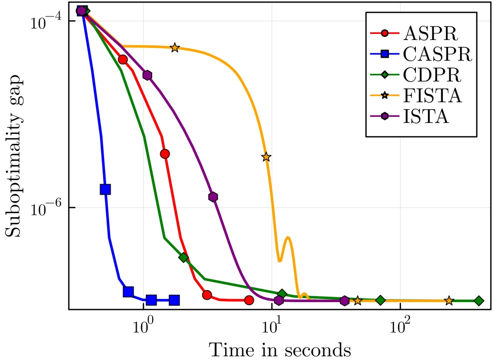
Experiments

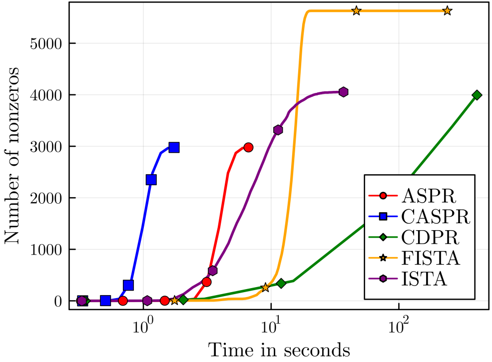
Results from a Standford Network Analysis Project graph with 3.7M nodes and 16.5M edges.



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.

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.

Thank you!