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:

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

where $0 \prec \alpha \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.

Problem

Problem:

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

where $0 \prec \alpha \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$. For ℓ_1-regularized personalized PageRank, it is

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

where $\alpha, \rho > 0$, $\mathcal{L} \stackrel{\text{def}}{=} I - D^{-1/2}AD^{-1/2}$ is the symmetric normalized Laplacian matrix, which satisfies $0 \prec \mathcal{L} \preccurlyeq 2I$.

Problem

Problem:

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

where $0 \prec \alpha \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$. For ℓ_1-regularized personalized PageRank, it is

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

where $\alpha, \rho > 0$, $\mathcal{L} \stackrel{\text{def}}{=} I - D^{-1/2}AD^{-1/2}$ is the symmetric normalized Laplacian matrix, which satisfies $0 \prec \mathcal{L} \preccurlyeq 2I$.

The problem comes from the personalized PageRank problem

$$\min\{f(x) \stackrel{\text{def}}{=} \langle x, Qx\rangle - \alpha\langle D^{-1/2}s, x\rangle\},$$

by adding the ℓ_1 regularization $+\alpha\rho\|D^{1/2}x\|_1$ and noticing that the minimizer is in $\mathbb{R}_{\geq 0}$, for $\rho > 0$.

The personalized PageRank vector is the solution to the system

$$x = (1-\alpha)Wx + \alpha s = ((1-\alpha)W + \alpha s\mathbb{1}^T)x,$$

where $W = (I + AD^{-1})/2$ and $s \in \Delta^n$ is a distribution over the nodes.

Results and comparison

- The Hessian of g is Q, satisfying $\alpha I \preccurlyeq Q \preccurlyeq LI$, its condition number is L/α.
- $\mathcal{S}^* \stackrel{\text{def}}{=} \mathrm{supp}(x^*)$, $\mathrm{vol}(\mathcal{S}^*) \stackrel{\text{def}}{=} \mathrm{nnz}(Q_{:,\mathcal{S}^*})$ 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 $\mathrm{vol}(\mathcal{S}^*) \leq \frac{1}{\rho} + |\mathcal{S}^*|$.

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

- Problem:

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

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 $\mathrm{x}^* \stackrel{\text{def}}{=} \arg\min_{\mathrm{x} \in \mathbb{R}^n_{\geq \mathbf{0}}} g(\mathrm{x})$.

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 $i \in S$.
- $C \stackrel{\text{def}}{=} \mathrm{span}(\{\mathrm{e}_i \mid i \in S\}) \cap \mathbb{R}^n_{\geq 0}$.
- $\mathrm{x}^{(*,C)} \stackrel{\text{def}}{=} \arg\min_{\mathrm{x}\in C} g(\mathrm{x})$ 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 $S \subseteq \mathcal{S}^*$.

Proof of 1.: $\bar{g} \stackrel{\text{def}}{=} g$ restricted to $\mathrm{span}(\{\mathrm{e}_i \mid i \in S\})$. 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:

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

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

A geometric lemma

Suppose:

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

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 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 $\mathrm{x}^{(*,C)}$ and set of indices $S \cup \{i \mid \nabla_i g(\mathrm{x}^{(*,C)}) < 0\}$ 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 $C^{(t)} \stackrel{\text{def}}{=} \mathrm{span}(\{e_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, 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 $C^{(t)} \stackrel{\text{def}}{=} \mathrm{span}(\{\mathrm{e}_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, where $S^{(t)}$ is the set of currently known good coordinates.

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

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 $C^{(t)} \stackrel{\text{def}}{=} \mathrm{span}(\{e_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, where $S^{(t)}$ is the set of currently known good coordinates.
- At the minimizer $x^{(*,t+1)} \stackrel{\text{def}}{=} x^{(*,C^{(t)})}$, we are optimal ($x^{(*,t+1)} = x^*$) or we have $\nabla_i g(x^{(*,t+1)}) < 0$ only if i is good and new, i.e., only if $i \in \mathcal{S}^* \setminus S^{(t)}$.
- 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 ($\texttt{CDPR}$)

- 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 $x^{(0)} = \mathbb{0}$.
- 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 $k < t$.

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 $x^{(0)} = \mathbb{0}$.
- 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 $x^{(t+1)} = x^{(*, C^{(t)})}$.
- 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 $y = x - \Delta e_i$, we have $\forall j \neq i$: $\nabla_j g(y) \geq \nabla_j g(x)$ if $\Delta > 0$ and $\nabla_j g(y) \leq \nabla_j g(x)$ otherwise.

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 $i \in \mathcal{S}^* \setminus S^{(t)}$. So by 1., for $x \in C^{(t)}$ s.t. $x \leq \mathrm{x}^{(*, C^{(t)})}$ it is
   $\nabla_i g(x) < 0$ only if $i \in \mathcal{S}^*$:
   We can detect new coordinates!

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 $i \in \mathcal{S}^* \setminus S^{(t)}$. So by 1., for $x \in C^{(t)}$ s.t. $x \leq \mathrm{x}^{(*,C^{(t)})}$ it is $\nabla_i g(x) < 0$ only if $i \in \mathcal{S}^*$:
   We can detect new coordinates!

3. To ensure there exists such an $i \in \mathcal{S}^* \setminus S^{(t)}$, get close to $\mathrm{x}^{(*,C^{(t)})}$ from below: optimize using accelerated projected gradient descent (APGD) to get close to $\mathrm{x}^{(*,C^{(t)})}$ and then move slightly towards $\mathbb{0}$ to be $\leq \mathrm{x}^{(*,C^{(t)})}$.

An inexact algorithm: Accelerated and Sparse PageRank ($\texttt{ASPR}$)

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 $i \in \mathcal{S}^* \setminus S^{(t)}$. So by 1., for $x \in C^{(t)}$ s.t. $x \leq \mathrm{x}^{(*, C^{(t)})}$ it is $\nabla_i g(x) < 0$ only if $i \in \mathcal{S}^*$:
   We can detect new coordinates!

3. To ensure there exists such an $i \in \mathcal{S}^* \setminus S^{(t)}$, get close to $\mathrm{x}^{(*, C^{(t)})}$ from below: optimize using accelerated projected gradient descent ($\texttt{APGD}$) to get close to $\mathrm{x}^{(*, C^{(t)})}$ and then move slightly towards $\mathbb{0}$ to be $\leq \mathrm{x}^{(*, C^{(t)})}$.

4. **Lemma**. Optimizing with accuracy $\hat{\varepsilon}_t = \varepsilon \cdot \frac{\alpha^2}{2(1+|S^{(t)}|)L^2}$ to get $\bar{\mathrm{x}}^{(t+1)}$ and reducing $\mathrm{x}^{(t+1)} \leftarrow \max\{\mathbb{0}, \bar{\mathrm{x}}^{(t+1)} - \delta_t \mathbb{1}\}$ for $\delta_t = \sqrt{\frac{\varepsilon\alpha}{(1+|S^{(t)}|L^2)}}$, we either expand $S^{(t)}$ using 2. with $\mathrm{x}^{(t+1)}$, or $\mathrm{x}^{(t+1)}$ is an ε-minimizer.

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 $i \in \mathcal{S}^* \setminus S^{(t)}$. So by 1., for $x \in C^{(t)}$ s.t. $x \leq \mathrm{x}^{(*,C^{(t)})}$ it is $\nabla_i g(x) < 0$ only if $i \in \mathcal{S}^*$:
   We can detect new coordinates!

3. To ensure there exists such an $i \in \mathcal{S}^* \setminus S^{(t)}$, get close to $\mathrm{x}^{(*,C^{(t)})}$ from below: optimize using accelerated projected gradient descent (APGD) to get close to $\mathrm{x}^{(*,C^{(t)})}$ and then move slightly towards $\mathbb{0}$ to be $\leq \mathrm{x}^{(*,C^{(t)})}$.

4. **Lemma**. Optimizing with accuracy $\hat{\varepsilon}_t = \varepsilon \cdot \frac{\alpha^2}{2(1+|S^{(t)}|)L^2}$ to get $\bar{\mathrm{x}}^{(t+1)}$ and reducing $\mathrm{x}^{(t+1)} \leftarrow \max\{\mathbb{0}, \bar{\mathrm{x}}^{(t+1)} - \delta_t \mathbb{1}\}$ for $\delta_t = \sqrt{\frac{\varepsilon\alpha}{(1+|S^{(t)}|L^2)}}$, we either expand $S^{(t)}$ using 2. with $\mathrm{x}^{(t+1)}$, or $\mathrm{x}^{(t+1)}$ is an ε-minimizer.

5. The lemma above is proven by showing that if the global gap is $> \varepsilon$, then one step of gradient descent reduces the function value more than what it can be reduced in $C^{(t)}$.

Accelerated and Sparse PageRank (ASPR) algorithm

Theorem

The iterates of ASPR satisfy:

1. $x_i^{(*,t)} > 0$ if and only if $i \in S^{(t-1)}$. Also, $\nabla_i g(\mathrm{x}^{(*,t)}) = 0$ if $i \in S^{(t-1)}$.
2. It is $\mathrm{x}^{(t)} \leq \mathrm{x}^{(*,t)} \leq \mathrm{x}^*$ and $\mathrm{x}^{(*,t-1)} \leq \mathrm{x}^{(*,t)}$.
3. $S^{(t-1)} \subsetneq S^{(t)} \stackrel{\text{def}}{=} S^{(t-1)} \cup \{i \in [n] \mid \nabla_i g(\mathrm{x}^{(t)}) < 0\} \subseteq \mathcal{S}^*$, or $\mathrm{x}^{(t)}$ is an ε-minimizer of g.

- APGD only needs gradients restricted to $C^{(t)}$, costing $\mathcal{O}(\widetilde{\mathrm{vol}}(\mathcal{S}^*))$ each. Then, it uses a full gradient to find the new good coordinates, costing $\mathcal{O}(\mathrm{vol}(\mathcal{S}^*))$. It is done at most $|\mathcal{S}^*|$ times.
- All new good coordinates are incorporated to $S^{(t)}$ unlike for CDPR.
- Time complexity $\widetilde{\mathcal{O}}(|\mathcal{S}^*|\widetilde{\mathrm{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\alpha}} + |\mathcal{S}^*|\mathrm{vol}(\mathcal{S}^*))$ and space complexity $\mathcal{O}(|\mathcal{S}^*|)$.

Variants

- **Lemma.** $S \subseteq [n]$. If x is s.t. $x_j = 0$ if $j \notin S$ and $\nabla_j g(x) \leq 0$ if $j \in S$, then for any $i \notin S$ s.t. $\nabla_i g(x) < 0$, it is $i \in \mathcal{S}^*$.

- **Variant:** During `APGD`'s execution, one can compute the full gradient from time to time to check the condition, expand $S^{(t)}$, and restart.

Variants

- **Lemma.** $S \subseteq [n]$. If x is s.t. $x_j = 0$ if $j \notin S$ and $\nabla_j g(x) \leq 0$ if $j \in S$, then for any $i \notin S$ s.t. $\nabla_i g(x) < 0$, it is $i \in \mathcal{S}^*$.

- **Variant:** During `APGD`'s execution, one can compute the full gradient from time to time to check the condition, expand $S^{(t)}$, and restart.

- Interestingly, if we compute a full gradient at every iteration, `ASPR` is not better than `CDPR`, up to constants and logs, in the regime in which ISTA is not better, up to constants and logs.

Variants

- **Lemma.** $S \subseteq [n]$. If x is s.t. $x_j = 0$ if $j \notin S$ and $\nabla_j g(x) \leq 0$ if $j \in S$, then for any $i \notin S$ s.t. $\nabla_i g(x) < 0$, it is $i \in \mathcal{S}^*$.

- **Variant:** During APGD's execution, one can compute the full gradient from time to time to check the condition, expand $S^{(t)}$, and restart.

- Interestingly, if we compute a full gradient at every iteration, ASPR is not better than CDPR, up to constants and logs, in the regime in which ISTA is not better, up to constants and logs.

- **Lemma.** If we observe $\nabla_j g(x) \leq 0$ for all $i \in S^{(t)}$, it is $x \leq \mathrm{x}^{(*,t+1)} \leq \mathrm{x}^*$.

- **Variant:** With such an x, we can update the feasible set: $C \leftarrow C \cap \{y \mid y \geq x\}$.

Variants

- **Lemma.** $S \subseteq [n]$. If x is s.t. $x_j = 0$ if $j \notin S$ and $\nabla_j g(x) \leq 0$ if $j \in S$, then for any $i \notin S$ s.t. $\nabla_i g(x) < 0$, it is $i \in \mathcal{S}^*$.

- **Variant:** During $\mathtt{APGD}$'s execution, one can compute the full gradient from time to time to check the condition, expand $S^{(t)}$, and restart.

- Interestingly, if we compute a full gradient at every iteration, $\mathtt{ASPR}$ is not better than $\mathtt{CDPR}$, up to constants and logs, in the regime in which ISTA is not better, up to constants and logs.

- **Lemma.** If we observe $\nabla_j g(x) \leq 0$ for all $i \in S^{(t)}$, it is $x \leq \mathrm{x}^{(*,t+1)} \leq \mathrm{x}^*$.

- **Variant:** With such an x, we can update the feasible set: $C \leftarrow C \cap \{y \mid y \geq x\}$.

- **Variant:** Using the (unconstrained) Conjugate Gradients algorithm (CG) instead of $\mathtt{APGD}$, the guarantee improves. And we can forgo the knowledge of the strong convexity constant α.

Comparisons

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