Dual Randomized Coordinate Descent Method for Solving a Class of Nonconvex Problems

Amir Beck

School of Mathematical Sciences, Tel Aviv University

Joint work with Marc Teboulle

One World Optimization Seminar, September 7, 2020

The Main Model

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\},$$

- $\mathbf{A} \in \mathbb{R}^{n\times d}$
- $f : \mathbb{R}^n \to (-\infty, \infty]$ proper, closed, **strongly convex**;
- $g : \mathbb{R}^d \to (-\infty, \infty]$ proper closed **convex** with a compact domain;
- $\text{dom}(g) \subseteq \text{dom}(h)$, where $h(\mathbf{x}) \equiv f(\mathbf{A}\mathbf{x})$.

convention: $\infty - \infty = -\infty$

The Main Model

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\},$$

- $\mathbf{A} \in \mathbb{R}^{n\times d}$
- $f : \mathbb{R}^n \to (-\infty, \infty]$ proper, closed, **strongly convex**;
- $g : \mathbb{R}^d \to (-\infty, \infty]$ proper closed **convex** with a compact domain;
- $\mathrm{dom}(g) \subseteq \mathrm{dom}(h)$, where $h(\mathbf{x}) \equiv f(\mathbf{A}\mathbf{x})$.

convention: $\infty - \infty = -\infty$

MAIN GOALS:

- improved optimality conditions
- develop randomized dual-based decomposition methods

Three PCA Prototype Problems

MODEL I: "standard PCA"

Given n points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find a normalized vector $\mathbf{x} \in \mathbb{R}^d$ for which the projected data $\mathbf{a}_1^T\mathbf{x}, \mathbf{a}_2^T\mathbf{x}, \ldots, \mathbf{a}_n^T\mathbf{x}$ has maximum variance

Three PCA Prototype Problems

MODEL I: "standard PCA"

Given n points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find a normalized vector $\mathbf{x} \in \mathbb{R}^d$ for which the projected data $\mathbf{a}_1^T\mathbf{x}, \mathbf{a}_2^T\mathbf{x}, \ldots, \mathbf{a}_n^T\mathbf{x}$ has maximum variance

- Under the assumption that $\sum_{i=1}^n \mathbf{a}_i = \mathbf{0}$, the problem is

$$\max_{\|\mathbf{x}\|_2=1} \frac{1}{2}\sum_{i=1}^n (\mathbf{a}_i^T\mathbf{x})^2.$$

- Denote $\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix} \in \mathbb{R}^{n\times d}$. Then the problem is

$$\text{(PCA)} \quad \max_{\|\mathbf{x}\|_2\le 1} \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2.$$

- Fits model (P) with $f(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ and $g = \delta_{B_2[\mathbf{0},1]}$.

Model II: Sparse PCA

Additional information: sought vector is sparse.

- [d'Aspremont et. al. 05']

$$\text{(SPCA)} \quad \max\{0.5\|\mathbf{A}\mathbf{x}\|_2^2 : \|\mathbf{x}\|_2 \le 1, \|\mathbf{x}\|_0 \le s\},$$

$\|\mathbf{x}\|_0 \equiv \#\{i : x_i \neq 0\}$, $s \le d$.
DOES NOT FIT MODEL (P) (feasible set nonconvex)

Model II: Sparse PCA

Additional information: sought vector is sparse.

- [d'Aspremont et. al. 05']

$$\text{(SPCA)} \quad \max\{0.5\|\mathbf{Ax}\|_2^2 : \|\mathbf{x}\|_2 \leq 1, \|\mathbf{x}\|_0 \leq s\},$$

  $\|\mathbf{x}\|_0 \equiv \#\{i : x_i \neq 0\}$, $s \leq d$.
  DOES NOT FIT MODEL (P) (feasible set nonconvex)
- BUT... equivalent to

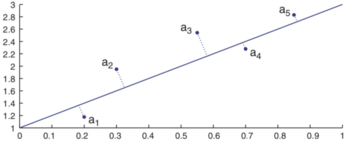
$$\max\{0.5\|\mathbf{Ax}\|_2^2 : \mathbf{x} \in \text{conv}(B_2[\mathbf{0}, 1] \cap C_s)\},$$

  $C_s = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s\}$.
- Fits model (P) with $f(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ and $g = \delta_{\text{conv}(B_2[\mathbf{0},1]\cap C_s)}$.

Model III: Square Root PCA

second interpretation of PCA (pearson 1901):

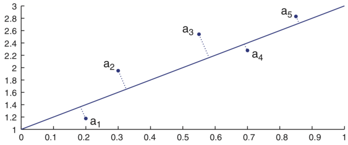
Given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find $\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1$ for which the sum of distances^2 of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ to $\mathrm{sp}(\mathbf{x})$ is minimal.



Model III: Square Root PCA

second interpretation of PCA (pearson 1901):

Given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find $\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1$ for which the sum of distances^2 of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ to $\mathrm{sp}(\mathbf{x})$ is minimal.



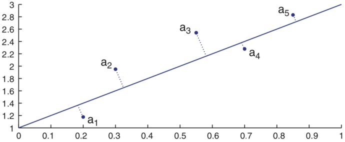
- $$\text{(PCA')} \qquad \min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2^2.$$

  SAME RESULT AS PCA!

Model III: Square Root PCA

second interpretation of PCA (pearson 1901):

Given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find $\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1$ for which the sum of distances^2 of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ to $\mathrm{sp}(\mathbf{x})$ is minimal.



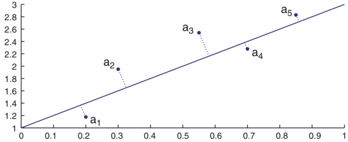
- $$\text{(PCA')} \qquad \min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2^2.$$

  SAME RESULT AS PCA!
- A robust version of (PCA'): $\min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2,$

Model III: Square Root PCA

second interpretation of PCA (pearson 1901):

Given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find $\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1$ for which the sum of distances^2 of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ to $\mathrm{sp}(\mathbf{x})$ is minimal.



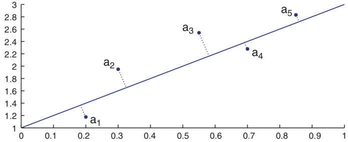
- $$\text{(PCA')} \quad \min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2^2.$$

  SAME RESULT AS PCA!
- A robust version of (PCA'): $\min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2$,
- $\Leftrightarrow \min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x}\rangle^2}$

Model III: Square Root PCA

second interpretation of PCA (pearson 1901):

Given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find $\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1$ for which the sum of distances^2 of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ to $\mathrm{sp}(\mathbf{x})$ is minimal.



- $$(\text{PCA'}) \qquad \min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2^2.$$

  SAME RESULT AS PCA!
- A robust version of (PCA'): $\min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \|\mathbf{a}_i - (\mathbf{a}_i^T\mathbf{x})\mathbf{x}\|_2$,
- $\Leftrightarrow \min_{\|\mathbf{x}\|_2=1} \sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x}\rangle^2}$
- $\Leftrightarrow \min_{\|\mathbf{x}\|_2\le 1} \sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x}\rangle^2}$ (by concavity)
  NOT SMOOTH OVER THE DOMAIN. We will consider a smooth approximation (a better reason in the sequal)

Model III: Square Root PCA

$$\text{(SRPCA)} \quad \max_{\|\mathbf{x}\|_2 \leq 1} -\sum_{i=1}^{n} \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x} \rangle^2 + \varepsilon^2}.$$

Fits model (P) with

$$f(\mathbf{z}) = \begin{cases} -\sum_{i=1}^{n} \sqrt{\|\mathbf{a}_i\|_2^2 + \varepsilon^2 - z_i^2} & |z_i| \leq \sqrt{\|\mathbf{a}_i\|_2^2 + \varepsilon^2}, \\ \infty & \text{else}, \end{cases} \quad g = \delta_{B_2[\mathbf{0},1]}.$$

Note: the inclusion $\text{dom}(g) \subseteq \text{dom}(f \circ \mathbf{A})$ holds.

Three PCA Models

name	$f(\mathbf{x})$	$g(\mathbf{x})$
PCA	$\frac{1}{2}\|\mathbf{x}\|_2^2$	$\delta_{B_2[\mathbf{0},1]}(\mathbf{x})$
SPCA	$\frac{1}{2}\|\mathbf{x}\|_2^2$	$\delta_{\mathrm{conv}(B_2[\mathbf{0},1]\cap C_s)}(\mathbf{x})$
SRPCA	$-\sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2-x_i^2}$ $(\lvert x_i\rvert \le \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2})$	$\delta_{B_2[\mathbf{0},1]}(\mathbf{x})$

The two options for f are strongly convex

Optimality Conditions

- Recall the main model: (P) $\max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$
- instance of **DC optimization** [review - Horst, Thoai '99]

General DC problem:

$$\max_{\mathbf{x}} s(\mathbf{x}) - t(\mathbf{x})$$

s, t - extended real-valued convex functions, $\mathrm{dom}(t) \subseteq \mathrm{dom}(s)$

Optimality Conditions

- Recall the main model: (P) $\quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$
- instance of **DC optimization** [review - Horst, Thoai '99]

General DC problem:

$$\max_{\mathbf{x}} s(\mathbf{x}) - t(\mathbf{x})$$

s, t - extended real-valued convex functions, $\mathrm{dom}(t) \subseteq \mathrm{dom}(s)$

Most fundamental necessary optimality condition: **CRITICALITY**

$$\bar{\mathbf{x}} \text{ opt. } \Rightarrow \underbrace{\partial s(\bar{\mathbf{x}}) \cap \partial t(\bar{\mathbf{x}}) \neq \emptyset}_{\text{criticality}}$$

Optimality Conditions

- Recall the main model: (P) $\max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$
- instance of **DC optimization** [review - Horst, Thoai '99]

General DC problem:

$$\max_{\mathbf{x}} s(\mathbf{x}) - t(\mathbf{x})$$

s, t - extended real-valued convex functions, $\mathrm{dom}(t) \subseteq \mathrm{dom}(s)$

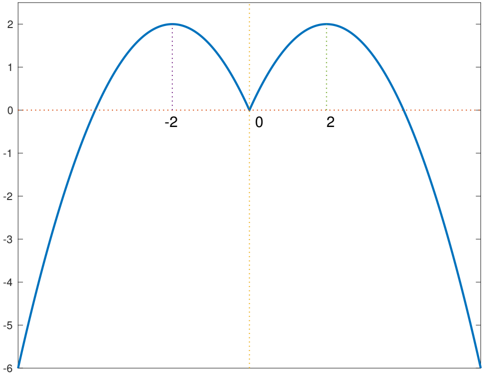
Most fundamental necessary optimality condition: **CRITICALITY**

$$\bar{\mathbf{x}} \text{ opt. } \Rightarrow \underbrace{\partial s(\bar{\mathbf{x}}) \cap \partial t(\bar{\mathbf{x}}) \neq \emptyset}_{\text{criticality}}$$

- can be replaced by $\partial s(\bar{\mathbf{x}}) \subseteq \partial t(\bar{\mathbf{x}})$
- another condition is stationarity = lack of feasible ascent directions.
- In general, criticality is **weaker** than **stationarity**. More results [Pang et. al. '17]

Stationarity vs. Criticality

The function $2|y| - \frac{y^2}{2}$ has three **critical** points $y = -2, 0, 2$.
Among them $y = -2, 2$ are **stationary** points



Back to the Main Model

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\},$$

- **Criticality.** $\mathbf{A}^T\partial f(\mathbf{A}\bar{\mathbf{x}}) \cap \partial g(\bar{\mathbf{x}}) \neq \emptyset$
- In all three PCA models, f is continuously differentiable over $\mathrm{dom}(g)$ and criticality $\iff$ stationarity.

Back to the Main Model

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\},$$

- **Criticality.** $\mathbf{A}^T\partial f(\mathbf{A}\bar{\mathbf{x}}) \cap \partial g(\bar{\mathbf{x}}) \neq \emptyset$
- In all three PCA models, f is continuously differentiable over $\mathrm{dom}(g)$ and criticality $\Longleftrightarrow$ stationarity.
- **Can we do better?**

Back to the Main Model

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\},$$

- **Criticality.** $\mathbf{A}^T\partial f(\mathbf{A}\bar{\mathbf{x}}) \cap \partial g(\bar{\mathbf{x}}) \neq \emptyset$
- In all three PCA models, f is continuously differentiable over $\mathrm{dom}(g)$ and criticality $\Longleftrightarrow$ stationarity.
- **Can we do better?** YES - through duality.

Toland Duality

main model:

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$$

Toland Duality

main model:

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$$

- use the fact that $f(\mathbf{A}\mathbf{x}) = f^{**}(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^n}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y})\}$

Toland Duality

main model:

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$$

- use the fact that $f(\mathbf{A}\mathbf{x}) = f^{**}(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^n}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y})\}$
- $\max_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^n}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) - g(\mathbf{x})\}$.

Toland Duality

main model:

$$(\mathrm{P}) \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$$

- use the fact that $f(\mathbf{A}\mathbf{x}) = f^{**}(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^n}\left\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y})\right\}$
- $\max_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^n}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) - g(\mathbf{x})\}$.
- $\max_{\mathbf{y}\in\mathbb{R}^n}\max_{\mathbf{x}\in\mathbb{R}^d}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) - g(\mathbf{x})\}$.

Toland Duality

main model:

$$(\mathrm{P}) \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}$$

- use the fact that $f(\mathbf{A}\mathbf{x}) = f^{**}(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^n}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y})\}$
- $\max_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^n}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) - g(\mathbf{x})\}$.
- $\max_{\mathbf{y}\in\mathbb{R}^n}\max_{\mathbf{x}\in\mathbb{R}^d}\{\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) - g(\mathbf{x})\}$.
- Obtain the Toland dual problem [Toland, '78,'79]:

$$(\mathrm{D}) \quad \max_{\mathbf{y}\in\mathbb{R}^n}\{q(\mathbf{y}) \equiv g^*(\mathbf{A}^T\mathbf{y}) - f^*(\mathbf{y})\}.$$

- DC problem (nonconvex)
- f^* - also $C^{1,1}$, g^* - real-valued

Duality Examples

- **PCA**

$$\text{(P)} \quad \max_{\|\mathbf{x}\|_2 \leq 1} 0.5\|\mathbf{A}\mathbf{x}\|_2^2 \qquad \text{(D-PCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - \frac{1}{2}\|\mathbf{y}\|_2^2 \right\}.$$

Duality Examples

- **PCA**

$$\text{(P)} \quad \max_{\|\mathbf{x}\|_2 \le 1} 0.5\|\mathbf{A}\mathbf{x}\|_2^2 \qquad \text{(D-PCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - \frac{1}{2}\|\mathbf{y}\|_2^2 \right\}.$$

- **sparse PCA**

$$\text{(SPCA)} \max\{0.5\|\mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \text{conv}(B_2[\mathbf{0},1] \cap C_s)\}$$

$$\text{(D-SPCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|T_s(\mathbf{A}^T\mathbf{y})\|_2 - \frac{1}{2}\|\mathbf{y}\|_2^2 \right\}.(T_s - \text{hard thresholding})$$

Duality Examples

- **PCA**

$$\text{(P)} \quad \max_{\|\mathbf{x}\|_2 \leq 1} 0.5\|\mathbf{A}\mathbf{x}\|_2^2 \qquad \text{(D-PCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - \frac{1}{2}\|\mathbf{y}\|_2^2 \right\}.$$

- **sparse PCA**

$$\text{(SPCA)} \max\{0.5\|\mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \operatorname{conv}(B_2[\mathbf{0},1] \cap C_s)\}$$

$$\text{(D-SPCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|T_s(\mathbf{A}^T\mathbf{y})\|_2 - \frac{1}{2}\|\mathbf{y}\|_2^2 \right\} .(T_s - \text{hard thresholding})$$

- **square-root PCA**

$$\text{(SRPCA)} \quad \max_{\|\mathbf{x}\|_2 \leq 1} -\sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x}\rangle^2 + \varepsilon^2}.$$

$$\text{(D-SRPCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - \sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2 + \varepsilon^2}\sqrt{y_i^2 + 1} \right\}.$$

Primal Dual Relations

global optimality:

- $\bar{\mathbf{y}}$ opt. for (D) $\Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ opt. for (P).
- $\bar{\mathbf{x}}$ opt. for (P) $\Rightarrow \bar{\mathbf{y}} \in \partial f(\mathbf{A}\bar{\mathbf{x}})$ opt. for (D).

Primal Dual Relations

global optimality:

- $\bar{\mathbf{y}}$ opt. for (D) $\Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ opt. for (P).
- $\bar{\mathbf{x}}$ opt. for (P) $\Rightarrow \bar{\mathbf{y}} \in \partial f(\mathbf{A}\bar{\mathbf{x}})$ opt. for (D).

optimality conditions:

- $\bar{\mathbf{y}}$ critical pt. of (D) $\Rightarrow$ any $\bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ s.t. $\nabla f^*(\bar{\mathbf{y}}) = \mathbf{A}\bar{\mathbf{x}}$ is critical for (P).

Primal Dual Relations

global optimality:

- $\bar{\mathbf{y}}$ opt. for (D) $\Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ opt. for (P).
- $\bar{\mathbf{x}}$ opt. for (P) $\Rightarrow \bar{\mathbf{y}} \in \partial f(\mathbf{A}\bar{\mathbf{x}})$ opt. for (D).

optimality conditions:

- $\bar{\mathbf{y}}$ critical pt. of (D) $\Rightarrow$ any $\bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ s.t. $\nabla f^*(\bar{\mathbf{y}}) = \mathbf{A}\bar{\mathbf{x}}$ is critical for (P).
- $\boxed{\bar{\mathbf{y}} \text{ stationary pt. of (D)} \Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}}) \text{ is critical for (P).}}$

Primal Dual Relations

global optimality:

- $\bar{\mathbf{y}}$ opt. for (D) $\Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ opt. for (P).
- $\bar{\mathbf{x}}$ opt. for (P) $\Rightarrow \bar{\mathbf{y}} \in \partial f(\mathbf{A}\bar{\mathbf{x}})$ opt. for (D).

optimality conditions:

- $\bar{\mathbf{y}}$ critical pt. of (D) $\Rightarrow$ any $\bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ s.t. $\nabla f^*(\bar{\mathbf{y}}) = \mathbf{A}\bar{\mathbf{x}}$ is critical for (P).
- $\boxed{\bar{\mathbf{y}} \text{ stationary pt. of (D)} \Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}}) \text{ is critical for (P).}}$

Definition: $\bar{\mathbf{x}} \in \mathrm{dom}(g)$ is a dual-stationary point of (P) if $\bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ for some stationary point $\bar{\mathbf{y}} \in \mathbb{R}^m$ of (D).

Primal Dual Relations

global optimality:

- $\bar{\mathbf{y}}$ opt. for (D) $\Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ opt. for (P).
- $\bar{\mathbf{x}}$ opt. for (P) $\Rightarrow \bar{\mathbf{y}} \in \partial f(\mathbf{A}\bar{\mathbf{x}})$ opt. for (D).

optimality conditions:

- $\bar{\mathbf{y}}$ critical pt. of (D) $\Rightarrow$ any $\bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ s.t. $\nabla f^*(\bar{\mathbf{y}}) = \mathbf{A}\bar{\mathbf{x}}$ is critical for (P).
- $\boxed{\bar{\mathbf{y}} \text{ stationary pt. of (D)} \Rightarrow \bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}}) \text{ is critical for (P).}}$

Definition: $\bar{\mathbf{x}} \in \text{dom}(g)$ is a dual-stationary point of (P) if $\bar{\mathbf{x}} \in \partial g^*(\mathbf{A}^T\bar{\mathbf{y}})$ for some stationary point $\bar{\mathbf{y}} \in \mathbb{R}^m$ of (D).

Result:

$$\boxed{\text{OPTIMALITY} \Rightarrow \text{DUAL STATIONARITY} \Rightarrow \text{CRITICALITY}}$$

Example

$$(P_1) \quad \max_{x_1,x_2} \left\{ \frac{1}{2}(x_1+x_2)^2 : |x_1| \le 1, |x_2| \le 1 \right\}.$$

- critical (=stationary points in this case) are

$$\{(x_1,x_2)^T : x_1+x_2=0, |x_1| \le 1, |x_2| \le 1\} \cup \{(-1,-1)^T, (1,1)^T\}.$$

- dual stationary pts. are $(-1,-1)^T, (1,1)^T$, which are the global optimal solutions.

So Far...

- Improved duality-based conditions.

So Far...

- Improved duality-based conditions.

Next

Devise duality-based methods that

(a) converge in some sense to dual-stationary points;

(b) able to tackle large-scale instances

So Far...

- Improved duality-based conditions.

Next

Devise duality-based methods that

(a) converge in some sense to dual-stationary points;

(b) able to tackle large-scale instances

Example of a scalable method for PCA:

Oja's method (variant of stochastic projected gradient):

$$\mathbf{x}^{k+1} = \frac{\tilde{\mathbf{x}}^{k+1}}{\|\tilde{\mathbf{x}}^{k+1}\|_2}, \text{ where } \tilde{\mathbf{x}}^{k+1} = \mathbf{x}^k + t_k \mathbf{a}_{i_k},$$

variants and more results [Shamir '16]

Objective: define a simple/cheap method for the general (P) that converges to dual-stationary pts.

Back to the Dual Problem

$$\text{(D)} \quad \max_{\mathbf{y}\in\mathbb{R}^n}\{q(\mathbf{y}) \equiv \underbrace{g^*(\mathbf{A}^T\mathbf{y})}_{\text{real}-\text{valued}} - \underbrace{f^*(\mathbf{y})}_{C^{1,1}\text{ function}}\}.$$

Equivalent to (D') $\quad \min_{\mathbf{y}\in\mathbb{R}^n}\{\underbrace{f^*(\mathbf{y})}_{C^{1,1}} - \underbrace{g^*(\mathbf{A}^T\mathbf{y})}_{\text{real}-\text{valued}}\}.$

Back to the Dual Problem

$$\text{(D)} \quad \max_{\mathbf{y}\in\mathbb{R}^n}\{q(\mathbf{y}) \equiv \underbrace{g^*(\mathbf{A}^T\mathbf{y})}_{\text{real}-\text{valued}} - \underbrace{f^*(\mathbf{y})}_{C^{1,1}\text{ function}}\}.$$

Equivalent to (D') $\quad \min_{\mathbf{y}\in\mathbb{R}^n}\{\underbrace{f^*(\mathbf{y})}_{C^{1,1}} - \underbrace{g^*(\mathbf{A}^T\mathbf{y})}_{\text{real}-\text{valued}}\}.$

Idea: employ a randomized coordinate descent (RCD) method on (D')

<u>**The RCD Method**</u>

Input. $(F, \mathbf{t}^0, r)$ where $F : \mathbb{R}^n \to \mathbb{R}, \mathbf{t}^0 \in \mathbb{R}^m, r \in (0, \infty]$

General Step. For any $k = 0, 1, ...$

(a) pick $i_k \in [n]$ at random (assume uniform for simplicity)

(b) compute $\alpha \in \operatorname{argmin}_{t\in[-r,r]} F(t_1^k, t_2^k, \ldots, t_{i_k-1}^k, t, t_{i_k+1}^k, \ldots, t_n^k)$;

(c) set $t_{i_k}^{k+1} = \alpha$ and $t_j^{k+1} = t_j^k$ for $j \neq i_k$.

Convergence of RCD in the Dual Space

Theorem [Beck, Hallak 2020] Let $F = f_1 - f_2$

- $f_1 : \mathbb{R}^n \to \mathbb{R}$ differentiable convex
- $f_2 : \mathbb{R}^n \to \mathbb{R}$ convex.

Let $\{\mathbf{y}^k\}_{k\geq 0}$ be generated by RCD. Then almost surely, all accumulation points of $\{\mathbf{y}^k\}_{k\geq 0}$ are stationary points of the problem $\min_{\mathbf{y}} F(\mathbf{y})$.

Dual RCD for Solving $\boxed{\text{(P) } \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}}$

Primal Sequence: $\mathbf{x}^k \in \partial g^*(\mathbf{A}^T\mathbf{y}^k)$ ($\mathbf{y}^k$ - dual sequence, $\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

Dual RCD for Solving $\boxed{\text{(P) } \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}}$

Primal Sequence: $\mathbf{x}^k \in \partial g^*(\mathbf{A}^T\mathbf{y}^k)$ ($\mathbf{y}^k$ - dual sequence, $\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

Dual RCD (Input: $(f, g, \mathbf{A})$, $r \in (0, \infty]$)
Initialization. $\mathbf{y}^0 = \mathbf{0} \in \mathbb{R}^n, \mathbf{z}^0 = \mathbf{0} \in \mathbb{R}^d$.
General Step. For any $k = 0, 1, .., K$,

(a) pick $i_k \in [n]$ at random

(b) compute

$$t_k \in \underset{t\in[-r,r]}{\operatorname{argmin}} \left\{f^*(\mathbf{y}^k + (t - y^k_{i_k})\mathbf{e}_{i_k}) - g^*(\mathbf{z}^k + (t - y_{i_k})\mathbf{a}_{i_k})\right\};$$

(c) update $\mathbf{y}^{k+1} = \mathbf{y}^k + (t_k - y^k_{i_k})\mathbf{e}_{i_k}$ and $\mathbf{z}^{k+1} = \mathbf{z}^k + (t_k - y_{i_k})\mathbf{a}_{i_k}$.

Output: $\mathbf{x}_{\text{out}} \in \partial g^*(\mathbf{z}^{K+1})$.

Form: $\boxed{\mathbf{z}^{k+1} = \mathbf{z}^k + s_k\mathbf{a}_{i_k}}$; $g = \delta_{B_2[\mathbf{0},1]} \Rightarrow$ normalization of output: $\mathbf{x}_{\text{out}} = \mathbf{z}^{K+1}/\|\mathbf{z}^{K+1}\|_2$.

Dual RCD for Solving $\boxed{\text{(P) } \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\}}$

Primal Sequence: $\mathbf{x}^k \in \partial g^*(\mathbf{A}^T\mathbf{y}^k)$ ($\mathbf{y}^k$ - dual sequence, $\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

<u>Dual RCD</u> (Input: $(f, g, \mathbf{A})$**,** $r \in (0, \infty]$)
Initialization. $\mathbf{y}^0 = \mathbf{0} \in \mathbb{R}^n, \mathbf{z}^0 = \mathbf{0} \in \mathbb{R}^d$.
General Step. For any $k = 0, 1, .., K$,

(a) pick $i_k \in [n]$ at random

(b) compute
$$t_k \in \operatorname*{argmin}_{t\in[-r,r]} \left\{ f^*(\mathbf{y}^k + (t - y_{i_k}^k)\mathbf{e}_{i_k}) - g^*(\mathbf{z}^k + (t - y_{i_k})\mathbf{a}_{i_k}) \right\};$$

(c) update $\mathbf{y}^{k+1} = \mathbf{y}^k + (t_k - y_{i_k}^k)\mathbf{e}_{i_k}$ and $\mathbf{z}^{k+1} = \mathbf{z}^k + (t_k - y_{i_k})\mathbf{a}_{i_k}$.

Output: $\mathbf{x}_{\rm out} \in \partial g^*(\mathbf{z}^{K+1})$.

Form: $\boxed{\mathbf{z}^{k+1} = \mathbf{z}^k + s_k\mathbf{a}_{i_k}}$; $g = \delta_{B_2[\mathbf{0},1]} \Rightarrow$ normalization of output: $\mathbf{x}_{\rm out} = \mathbf{z}^{K+1}/\|\mathbf{z}^{K+1}\|_2$.
Different then Oja's method for PCA (repeated normalization):
$$\mathbf{x}^{k+1} = \tilde{\mathbf{x}}^{k+1}/\|\tilde{\mathbf{x}}^{k+1}\|_2, \text{ where } \tilde{\mathbf{x}}^{k+1} = \mathbf{x}^k + t_k\mathbf{a}_{i_k},$$

Primal Convergence of Dual RCD

Theorem

- let $\{\mathbf{y}^k\}_{k\geq 0}$ be generated by RCD employed on
$$-q(\mathbf{y}) = f^*(\mathbf{y}) - g^*(\mathbf{A}^T\mathbf{y})$$
- $\boxed{\text{assume that } -q \textbf{ has bounded level sets}}$.
- let $\mathbf{x}^k \in \partial g^*(\mathbf{A}^T\mathbf{y}^k)$.

$\Rightarrow$ a.s. all accumulation pts. of $\{\mathbf{x}^k\}_{k\geq 0}$ are **dual stationary pts.** of (P).

- assumption also required to make the method well-defined.
- not always easy to verify

Replacing the Bounded Level Sets Assumption

- **asymptotic function of a proper h:** $h_\infty(\mathbf{d}) \equiv \liminf_{\mathbf{d}' \to \mathbf{d}, t \to \infty} \frac{h(t\mathbf{d}')}{t}$.
- **known result:** if $h_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0} \Rightarrow h$ has bounded level sets.

Replacing the Bounded Level Sets Assumption

- **asymptotic function of a proper h:** $h_\infty(\mathbf{d}) \equiv \liminf_{\mathbf{d}' \to \mathbf{d}, t \to \infty} \frac{h(t\mathbf{d}')}{t}$.
- **known result:** if $h_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0} \Rightarrow h$ has bounded level sets.
- **conclusion:** it is enough to assume

$$[C] \quad (-q_\infty)(\mathbf{d}) = (f^*(\cdot) - g^*(\mathbf{A}^T \cdot))_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0}.$$

However, condition [C] is not explicit. **Need a calculus rule for asymptotic functions!**

Replacing the Bounded Level Sets Assumption

- **asymptotic function of a proper** h**:** $h_\infty(\mathbf{d}) \equiv \liminf_{\mathbf{d}' \to \mathbf{d}, t \to \infty} \frac{h(t\mathbf{d}')}{t}$.
- **known result:** if $h_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0} \Rightarrow h$ has bounded level sets.
- **conclusion:** it is enough to assume

$$[C] \quad (-q_\infty)(\mathbf{d}) = (f^*(\cdot) - g^*(\mathbf{A}^T \cdot))_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0}.$$

However, condition [C] is not explicit. **Need a calculus rule for asymptotic functions!**

Lemma. Suppose u, v real-valued such that v is Lipschitz continuous and convex. Then

$$(u - v)_\infty = u_\infty - v_\infty.$$

Replacing the Bounded Level Sets Assumption

- **asymptotic function of a proper** h**:** $h_\infty(\mathbf{d}) \equiv \liminf_{\mathbf{d}' \to \mathbf{d}, t \to \infty} \frac{h(t\mathbf{d}')}{t}$.
- **known result:** if $h_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0} \Rightarrow h$ has bounded level sets.
- **conclusion:** it is enough to assume

$$[C] \quad (-q_\infty)(\mathbf{d}) = (f^*(\cdot) - g^*(\mathbf{A}^T \cdot))_\infty(\mathbf{d}) > 0 \ \forall \mathbf{d} \neq \mathbf{0}.$$

However, condition [C] is not explicit. **Need a calculus rule for asymptotic functions!**

Lemma. Suppose u, v real-valued such that v is Lipschitz continuous and convex. Then

$$(u - v)_\infty = u_\infty - v_\infty.$$

Result. Bounded level set assumption can be replaced by

$$[C] \quad (f^*)_\infty(\mathbf{d}) - (g^*)_\infty(\mathbf{A}^T\mathbf{d}) > 0$$

Validity of [C] for the 3 PCA Models

name	$f(\mathbf{x})$	$g(\mathbf{x})$
PCA	$\frac{1}{2}\\|\mathbf{x}\\|_2^2$	$\delta_{B_2[\mathbf{0},1]}(\mathbf{x})$
SPCA	$\frac{1}{2}\\|\mathbf{x}\\|_2^2$	$\delta_{\mathrm{conv}(B_2[\mathbf{0},1]\cap C_s)}(\mathbf{x})$
SRPCA	$-\sum_{i=1}^n \sqrt{\\|\mathbf{a}_i\\|_2^2+\varepsilon^2-z_i^2}$ $(\lvert z_i\rvert \le \sqrt{\\|\mathbf{a}_i\\|_2^2+\varepsilon^2})$	$\delta_{B_2[\mathbf{0},1]}(\mathbf{x})$

PCA,SPCA: $f^*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2 \Rightarrow (f^*)_\infty(\mathbf{d}) = \infty \ \ \forall \mathbf{d} \neq \mathbf{0}$

Validity of [C] for the 3 PCA Models

name	$f(\mathbf{x})$	$g(\mathbf{x})$
PCA	$\frac{1}{2}\|\mathbf{x}\|_2^2$	$\delta_{B_2[\mathbf{0},1]}(\mathbf{x})$
SPCA	$\frac{1}{2}\|\mathbf{x}\|_2^2$	$\delta_{\mathrm{conv}(B_2[\mathbf{0},1]\cap C_s)}(\mathbf{x})$
SRPCA	$-\sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2-z_i^2}$ $(\lvert z_i\rvert \leq \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2})$	$\delta_{B_2[\mathbf{0},1]}(\mathbf{x})$

PCA,SPCA: $f^*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2 \Rightarrow (f^*)_\infty(\mathbf{d}) = \infty \ \ \forall \mathbf{d} \neq \mathbf{0}$

SRPCA: $f^*(\mathbf{y}) = \sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2}\sqrt{y_i^2+1} \Rightarrow (f^*)_\infty(\mathbf{d}) = \sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2}|d_i|$. [C] follows by the fact that

$$\sum_{i=1}^n \sqrt{\|\mathbf{a}_i\|_2^2+\varepsilon^2}|d_i| - \|\mathbf{A}^T\mathbf{d}\|_2 > 0 \text{ for all } \mathbf{d} \neq \mathbf{0}.$$

(would not work without smoothing)

Convergence Rate when $g = \delta_{B_2[\mathbf{0},1]}$

problem (P) amounts to

$$\text{(P)} \quad \max_{\mathbf{x}} \{ f(\mathbf{A}\mathbf{x}) : \|\mathbf{x}\|_2 \leq 1 \}$$

Dual:
$$q_{\rm opt} = \max_{\mathbf{y}} \left\{ q(\mathbf{y}) \equiv \|\mathbf{A}^T\mathbf{y}\|_2 - f^*(\mathbf{y}) \right\}.$$

Assumption: $\operatorname{argmin}_{\mathbf{x}} f(\mathbf{x}) = \{\mathbf{0}\}$.

Convergence Rate when $g = \delta_{B_2[\mathbf{0},1]}$

problem (P) amounts to

$$\text{(P)} \quad \max_{\mathbf{x}} \{ f(\mathbf{A}\mathbf{x}) : \|\mathbf{x}\|_2 \leq 1 \}$$

Dual:
$$q_{\rm opt} = \max_{\mathbf{y}} \left\{ q(\mathbf{y}) \equiv \|\mathbf{A}^T\mathbf{y}\|_2 - f^*(\mathbf{y}) \right\}.$$

Assumption: $\operatorname{argmin}_{\mathbf{x}} f(\mathbf{x}) = \{\mathbf{0}\}$.

Lemma. If $\mathbf{y}^0 = \mathbf{0}$, then the dual objective function q is differentiable at $\mathbf{y}^k$ for all $k \geq 1$ (as well as all accumulation pts)

Convergence Rate when $g = \delta_{B_2[\mathbf{0},1]}$

problem (P) amounts to

$$(\mathrm{P}) \quad \max_{\mathbf{x}} \{ f(\mathbf{A}\mathbf{x}) : \|\mathbf{x}\|_2 \leq 1 \}$$

Dual:
$$q_{\rm opt} = \max_{\mathbf{y}} \left\{ q(\mathbf{y}) \equiv \|\mathbf{A}^T \mathbf{y}\|_2 - f^*(\mathbf{y}) \right\}.$$

Assumption: $\operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}) = \{\mathbf{0}\}$.

Lemma. If $\mathbf{y}^0 = \mathbf{0}$, then the dual objective function q is differentiable at $\mathbf{y}^k$ for all $k \geq 1$ (as well as all accumulation pts)

Theorem. Let $\{\mathbf{y}^k\}_{k\geq 0}$ be generated by the RCD method employed on $-q$ with initialization $\mathbf{y}^0 = \mathbf{0}$. Then

$$\min_{k=1,\ldots,N} \mathbb{E}(\|\nabla q(\mathbf{y}^k)\|_2^2) \leq \frac{2nL_{\max}}{N}(q_{\rm opt} - q(\mathbf{0})),$$

where $L_{\max}$ - maximum of coordinate Lipschitz constants of f^*.

Dual RCD Methods for PCA

$$\text{(PCA)} \quad \max_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{A}\mathbf{x}\|_2^2 \qquad \text{(D-PCA)} \quad \max_{\mathbf{y} \in \mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - 0.5\|\mathbf{y}\|_2^2 \right\}.$$

1-D minimization problem solved at each iteration ($\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

$$(1D) \quad \min_t \left\{ h(t) \equiv 0.5\|\mathbf{y}^k + (t - y_{i_k}^k)\mathbf{e}_{i_k}\|_2^2 - \|\mathbf{z}^k + (t - y_{i_k}^k)\mathbf{a}_{i_k}\|_2 \right\},$$

reduces to the finding roots of a quartic polynomial

Dual RCD Methods for PCA

$$\text{(PCA)} \quad \max_{\|\mathbf{x}\|_2 \le 1} \|\mathbf{A}\mathbf{x}\|_2^2 \qquad \text{(D-PCA)} \quad \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - 0.5\|\mathbf{y}\|_2^2 \right\}.$$

1-D minimization problem solved at each iteration ($\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

$$(1D) \quad \min_t \left\{ h(t) \equiv 0.5\|\mathbf{y}^k + (t - y_{i_k}^k)\mathbf{e}_{i_k}\|_2^2 - \|\mathbf{z}^k + (t - y_{i_k}^k)\mathbf{a}_{i_k}\|_2 \right\},$$

reduces to the finding roots of a quartic polynomial

Dual RCD Method for PCA

Initialization. $\mathbf{y}^0 = \mathbf{0}, \mathbf{z}^0 = \mathbf{0}$.

General Step. For any $k = 0, 1, \ldots K$,

(a) pick $i_k \in [n]$ at random.

(b) find a solution t_k of problem (1D)

(c) $\mathbf{y}^{k+1} = \mathbf{y}^k + (t_k - y_{i_k}^k)\mathbf{e}_{i_k}, \mathbf{z}^{k+1} = \mathbf{z}^k + (t_k - y_{i_k})\mathbf{a}_{i_k}$

Output: $\mathbf{x}_{\text{out}} = \frac{\mathbf{z}^{K+1}}{\|\mathbf{z}^{K+1}\|_2}$.

Dual RCD Methods for SRPCA

$$(P) \max_{\|\mathbf{x}\|_2 \le 1} -\sum_{i=1}^{n} \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x} \rangle^2 + \varepsilon^2}, (\mathrm{D}) \max_{\mathbf{y} \in \mathbb{R}^n} \left\{ \|\mathbf{A}^T \mathbf{y}\|_2 - \sum_{i=1}^{n} \sqrt{\|\mathbf{a}_i\|_2^2 + \varepsilon^2} \sqrt{y_i^2 + 1} \right\}.$$

1-D minimization solved at each iteration ($\mathbf{z}^k = \mathbf{A}^T \mathbf{y}^k, \tilde{\mathbf{z}}^k = \mathbf{z}^k - y_{i_k} \mathbf{a}_{i_k}$)

$$(1D) \min_t \left\{ h_3(t) \equiv \sqrt{\|\mathbf{a}_{i_k}\|_2^2 + \varepsilon^2} \sqrt{t^2 + 1} - \sqrt{\|\tilde{\mathbf{z}}^k\|_2^2 + 2t \mathbf{a}_{i_k}^T \tilde{\mathbf{z}}^k + t^2 \|\mathbf{a}_{i_k}\|_2^2} \right\},$$

reduces to the finding roots of a quartic polynomial

Dual RCD Methods for SRPCA

$$(P) \max_{\|\mathbf{x}\|_2 \le 1} -\sum_{i=1}^{n} \sqrt{\|\mathbf{a}_i\|_2^2 - \langle \mathbf{a}_i, \mathbf{x}\rangle^2 + \varepsilon^2}, \text{(D)} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \|\mathbf{A}^T\mathbf{y}\|_2 - \sum_{i=1}^{n} \sqrt{\|\mathbf{a}_i\|_2^2 + \varepsilon^2}\sqrt{y_i^2+1} \right\}.$$

1-D minimization solved at each iteration $(\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k, \tilde{\mathbf{z}}^k = \mathbf{z}^k - y_{i_k}\mathbf{a}_{i_k})$

$$(1D) \min_t \left\{ h_3(t) \equiv \sqrt{\|\mathbf{a}_{i_k}\|_2^2 + \varepsilon^2}\sqrt{t^2+1} - \sqrt{\|\tilde{\mathbf{z}}^k\|_2^2 + 2t\mathbf{a}_{i_k}^T\tilde{\mathbf{z}}^k + t^2\|\mathbf{a}_{i_k}\|_2^2} \right\},$$

reduces to the finding roots of a quartic polynomial

Dual RCD Method for SRPCA

Initialization. $\mathbf{y}^0 = \mathbf{0}, \mathbf{z}^0 = \mathbf{0}$.

General Step. For any $k = 0, 1, \dots K$,

(a) pick $i_k \in [n]$ at random.

(b) find a solution t_k of problem (1D)

(c) $\mathbf{y}^{k+1} = \mathbf{y}^k + (t_k - y_{i_k}^k)\mathbf{e}_{i_k}, \mathbf{z}^{k+1} = \mathbf{z}^k + (t_k - y_{i_k})\mathbf{a}_{i_k}$

Output: $\mathbf{x}_{\text{out}} = \frac{\mathbf{z}^{K+1}}{\|\mathbf{z}^{K+1}\|_2}$.

Dual RCD Methods for SPCA

$$\text{(P)} \max\{\|\mathbf{A}\mathbf{x}\|_2^2 : \|\mathbf{x}\|_2 \le 1, \|\mathbf{x}\|_0 \le s\} \quad \text{(D)} \max_{\mathbf{y}} \left\{\|T_s(\mathbf{A}^T\mathbf{y})\|_2 - 0.5\|\mathbf{y}\|_2^2\right\}.$$

1-D minimization problem solved at each iteration ($\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

$$(1D) \quad \min_t \left\{0.5\|\mathbf{y}^k + (t - y_{i_k}^k)\mathbf{e}_{i_k}\|_2^2 - \|T_s(\mathbf{z}^k + (t - y_{i_k}^k)\mathbf{a}_{i_k})\|_2\right\}$$

Dual RCD Methods for SPCA

$$\text{(P)}\max\{\|\mathbf{A}\mathbf{x}\|_2^2 : \|\mathbf{x}\|_2 \leq 1, \|\mathbf{x}\|_0 \leq s\} \quad \text{(D)}\max_{\mathbf{y}}\left\{\|T_s(\mathbf{A}^T\mathbf{y})\|_2 - 0.5\|\mathbf{y}\|_2^2\right\}.$$

1-D minimization problem solved at each iteration ($\mathbf{z}^k = \mathbf{A}^T\mathbf{y}^k$)

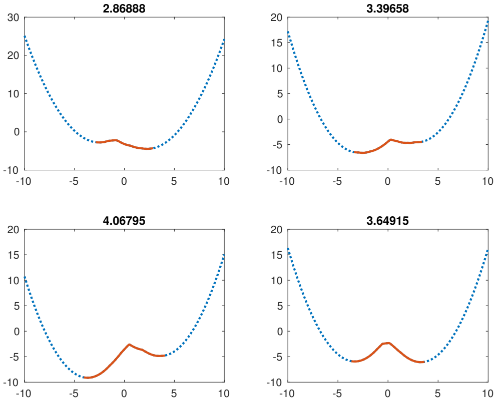
$$(1D) \quad \min_t \left\{0.5\|\mathbf{y}^k + (t - y_{i_k}^k)\mathbf{e}_{i_k}\|_2^2 - \|T_s(\mathbf{z}^k + (t - y_{i_k}^k)\mathbf{a}_{i_k})\|_2\right\}$$

- Same type of update formula $\mathbf{z}^{k+1} = \mathbf{z}^k + (t_k - y_{i_k})\mathbf{a}_{i_k}$
- No explicit formula for the solution of (1D) problem

Solving the 1D Problem

$$\min_t \left\{ R_{\mathbf{v},\mathbf{w}}(t) \equiv 0.5t^2 - \| T_s(\mathbf{v} + t\mathbf{w})\|_2 \right\}$$

- **Result 1:** all optimal solutions are in $[-\|\mathbf{w}\|_2, \|\mathbf{w}\|_2]$
- **Result 2:** $R_{\mathbf{v},\mathbf{w}}$ is $2\|\mathbf{w}\|_2$-Lipschitz continuous



Summary - Main Points

$$\text{(P)} \quad \max_{\mathbf{x}\in\mathbb{R}^d}\{f(\mathbf{A}\mathbf{x}) - g(\mathbf{x})\},$$

$$\text{(D)} \quad \max_{\mathbf{y}\in\mathbb{R}^n}\{q(\mathbf{y}) \equiv g^*(\mathbf{A}^T\mathbf{y}) - f^*(\mathbf{y})\}.$$

- duality-stationarity is a stronger condition than criticality.
- dual randomized coordinate descent algorithms converge a.s. to dual-stationary points
- form of dual RCD: $\mathbf{z}^{k+1} = \mathbf{z}^k + s_k\mathbf{a}_{i_k}$. s_k is a solution of a 1D problem.

THANK YOU FOR YOUR ATTENTION!!