Projection-Efficient First-Order Methods for Convex Low-Rank Matrix Optimization

Dan Garber
Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

One World Optimization Seminar

Low-Rank Matrix Recovery

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}:\ \mathrm{rank}(\mathbf{X})\leq r} f(\mathbf{X}) \qquad \text{or} \qquad \min_{\mathbf{X}\in\mathbb{S}^{n}:\ \mathbf{X}\succeq 0,\ \mathrm{rank}(\mathbf{X})\leq r} f(\mathbf{X})$$

$f(\mathbf{X})$ is (usually) convex, but not necessarily differentiable.

Applications:

- matrix completion
- robust principal component analysis
- sparse principal component analysis
- phase retrieval / synchronization
- matrix sensing

NP-Hard in general.

Rank-constrained problem is often not tackled directly.

Related work: nonconvex gradient methods

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}:\ \mathrm{rank}(\mathbf{X})\le r} f(\mathbf{X}) \quad \Longrightarrow \quad \min_{\mathbf{U}\in\mathbb{R}^{m\times r},\ \mathbf{V}\in\mathbb{R}^{m\times r}} f(\mathbf{U}\mathbf{V}^\top)$$

Unconstrained, but objective is nonconvex.

- Could be approached via simple gradient descent (when $f(\cdot)$ is diff.)
- Many exciting works in recent years on provable convergence of gradient methods to local/global minimum
- Strong results (efficient convergence to global minimum) often hold for very specific instances — specific choice of loss and strong assumptions on underlying data (e.g., follows some statistical model), and analysis is often quite involved.

Convex Relaxations

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}:\ \mathrm{rank}(\mathbf{X})\leq r} f(\mathbf{X}) \quad \Longrightarrow \quad \min_{\|\mathbf{X}\|_*\leq\tau} f(\mathbf{X})$$

$\|\mathbf{X}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{X})$ – sum of singular values (Trace/Nuclear norm).

$\|\cdot\|_*$ is convex surrogate for rank-sparsity in matrices in same way as $\|\cdot\|_1$ is convex surrogate for (entrywise) sparsity for vectors in $\mathbb{R}^n$.

- Have been studied extensively in past ~ 15 years
- In many cases well understood in terms of statistical theory and yield state-of-the-art recovery bounds (i.e., provable recovery of "ground truth" matrix under proper assumptions)
- Often yield strong empirical results in practice (recovery error, low-rank solutions)
- Easy to apply well-understood machinery for convex optimization and across various paradigms (stochastic, online, distributed, etc.)

Related work: conditional gradient for convex relaxation

$$\{\mathbf{X} \in \mathbb{R}^{m\times n} \mid \|\mathbf{X}\|_* \leq 1\} = \mathrm{conv}\{\mathbf{u}\mathbf{v}^\top \mid \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1\}$$

Admit the following simple iterations:

$$(\mathbf{u}_t, \mathbf{v}_t) \leftarrow \arg\min_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \mathbf{u}^\top \nabla f(\mathbf{X}_t)\mathbf{v}$$

$$\mathbf{X}_{t+1} \leftarrow (1-\eta_t)\mathbf{X}_t + \eta_t \tau \mathbf{u}_t \mathbf{v}_t^\top, \quad \eta_t \in [0,1]$$

- Efficient iterations - rank-one SVD (runtime $\propto \mathrm{nnz}(\nabla f)$)
- Suboptimal rates - $O(\beta/t)$, β — smoothness of $f(\cdot)$
- When number of iterations is moderately high require to store in memory high-rank matrices

Projected gradient methods for for convex relaxation

Methods such as Projected Grad, Accelerated Grad., FISTA, rely on the projected gradient mapping:

$$\mathbf{X}_t \to \Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}_t - \frac{1}{\beta}\nabla f(\mathbf{X}_t)]$$

- Enjoy optimal first-order convergence rates (Acc. Grad. / FISTA)
- Require in worst case to store high-rank matrices and compute SVDs of high-rank matrices (for the projection) — $O(\min\{m,n\}^2 \max\{m,n\})$ runtime per iteration

Convex relaxations — a wishful thinking

We solve convex relaxation to obtain low-rank solutions.

While it is plausible to assume that optimal solutions are indeed low-rank, it does not prevent standard first-order methods from going though high-rank matrices during the optimization process.

Question: can we expect, that under some plausible and quite general conditions, standard first-order methods will require to store and manipulate only low-rank matrices?

Euclidean projection onto nuclear norm ball

Lemma

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ *with SVD* $\mathbf{X} = \sum_{i=1}^{\min\{m,n\}} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, $\sigma_1 \geq \cdots \geq \sigma_n$. *If* $\|\mathbf{X}\|_* \geq 1$, *then the Euclidean projection of* $\mathbf{X}$ *onto unit nuclear norm ball is given by*

$$\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}] = \sum_{i=1}^{\min\{m,n\}} \max\{\sigma_i - \sigma, 0\} \mathbf{u}_i \mathbf{v}_i^\top,$$

where $\sigma \geq 0$ *is unique solution to the equation* $\sum_i \max\{\sigma_i - \sigma, 0\} = 1$.

- Projection thresholds to zero lower singular values
- If $\operatorname{rank}(\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}]) = r$, then only top r component in SVD of $\mathbf{X}$ are required.
  In particular, can be computed in time $\propto rmn << $ SVD time, when $r << \min\{m, n\}$

A simple approach (or wishful thinking)

Suppose convex relaxation indeed admits low-rank solutions.

We saw Euclidean projection cuts lower components of SVD

We know that $\mathrm{rank}(\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}^* - \frac{1}{\beta}\nabla f(\mathbf{X}^*)]) = \mathrm{rank}(\mathbf{X}^*)$

Question: *Focusing on projection-based methods, can we argue that, at least in neighborhoods of optimal solutions, the projected gradient mapping is guaranteed to be low-rank?*

What is the size of this neighborhood? How low is the rank?

Part 1: smooth minimization over unit nuclear norm ball

$$\min\{f(\mathbf{X}) \mid \|\mathbf{X}\|_* \leq 1\}$$

- $f(\cdot)$ is convex and β-smooth
- $\nabla f \neq 0$ over feasible set — optimal solutions do not perfectly fit data

Based on: *D. Garber, On the convergence of projected-gradient methods with low-rank projections for smooth convex minimization over trace-norm balls and related problems, SIAM Journal on Optimization 2021*.

Structure of optimal solutions

Lemma

Let $\mathbf{X}^*$ be an optimal solution, $rank(\mathbf{X}^*) = r^*$, and write its SVD $\mathbf{X}^* = \sum_{i=1}^{r^*} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Then, for all $i = 1, \ldots, r^*$ it holds that $\mathbf{u}_i^\top \nabla f(\mathbf{X}^*) \mathbf{v}_i = -\sigma_1(\nabla f(\mathbf{X}^*))$.

In particular: $\mathrm{rank}(\mathbf{X}^*) \leq \#\sigma_1(\nabla f(\mathbf{X}^*))$ — mult. of largest sing. value.

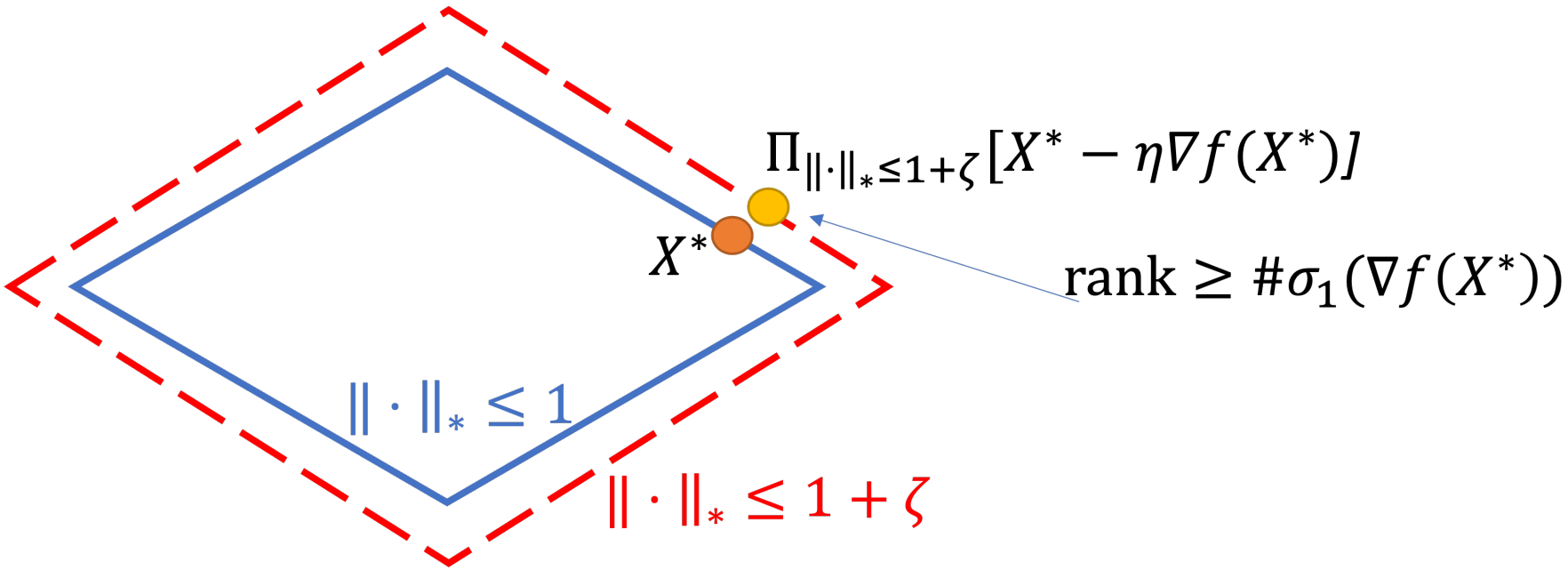
Equality need not hold in general:

$$\min_{\|\mathbf{X}\|_* \leq 1} \|\mathbf{X} - \mathrm{diag}(1+\sigma, \sigma, \ldots, \sigma)\|_F^2 \quad \Longrightarrow \quad \mathbf{X}^* = \mathbf{E}_{11}, \nabla f(\mathbf{X}^*) = \sigma \mathbf{I}$$

Note that in highly popular case: $f(\mathbf{X}) = g(\mathcal{A}\mathbf{X}) + \langle \mathbf{C}, \mathbf{X} \rangle$, $g(\cdot)$ str.convex, ∇f is constant over optimal set.

Can we expect that $\#\sigma_1(\nabla f(\mathbf{X}^*)) = \mathrm{rank}(\mathbf{X}^*)$?

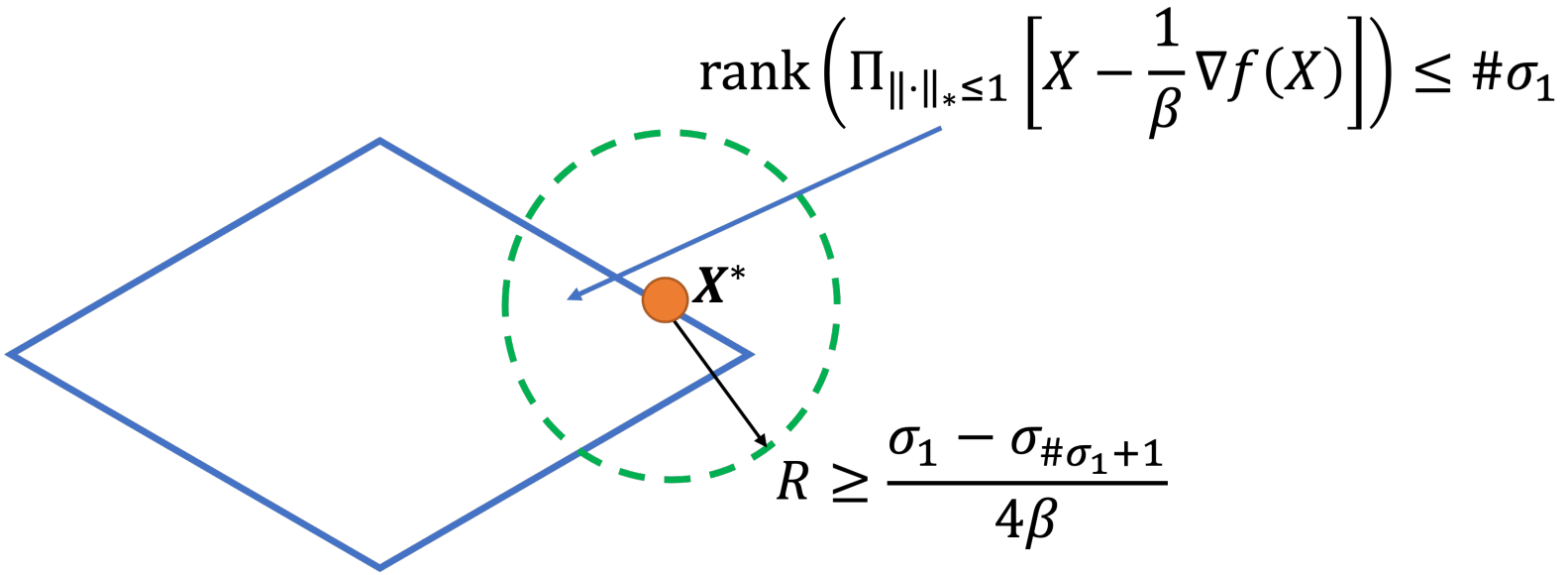
For any $f(\cdot)$ convex and differentiable with $\nabla f \neq 0$ over unit nuclear norm ball, for any $\eta > 0$, and $\zeta > 0$ small enough:



When $\mathrm{rank}(\mathbf{X}^*) < \#\sigma_1(\nabla f(\mathbf{X}^*))$ convex relaxation highly sensitive to arbitrary small misspecification of nuclear norm bound.
Suggests relaxation is ill-posed for low-rank matrix recovery.

Main result

Fix optimal solution $\mathbf{X}^*$ and denote $\sigma_i, i = 1, 2, \ldots$ the singular values of $\nabla f(\mathbf{X}^*)$. Let $\#\sigma_1$ denote number of sing. vals equal $\sigma_1(\nabla f(\mathbf{X}^*))$.

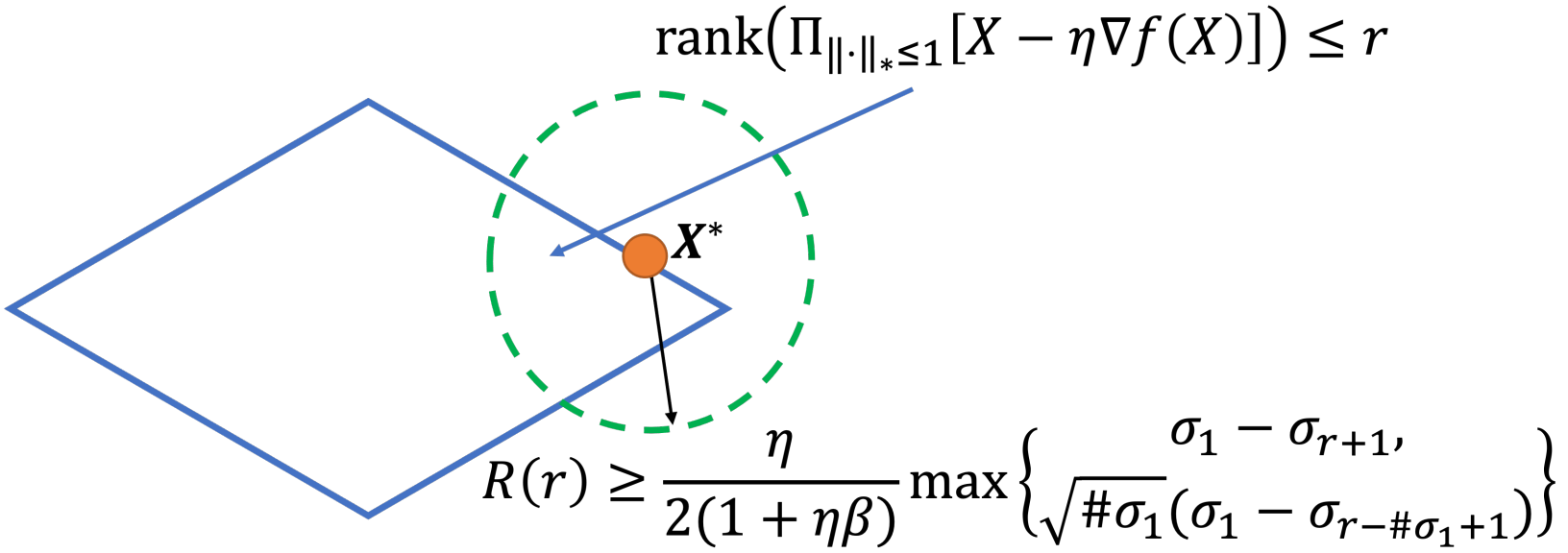


Inside ball it suffices to compute rank-r SVD, for $r \ge \#\sigma_1$ in order to compute projected gradient map.
If $\#\sigma_1 = \mathrm{rank}(\mathbf{X}^*)$ only $\mathrm{rank}(\mathbf{X}^*)$-SVD required.

Tightness: Lower bound (worst-case) tight up to factor $4\sqrt{2}$.

Main result - generalized

Fix optimal solution $\mathbf{X}^*$ and denote $\sigma_i, i = 1, 2, \ldots$ the singular values of $\nabla f(\mathbf{X}^*)$. Let $\#\sigma_1)$ denote number of sing. vals equal $\sigma_1(\nabla f(\mathbf{X}^*))$. For all $r \geq \#\sigma_1$ and $\eta > 0$:



Increasing rank of SVD computations can increase neighborhood in which projection is low-rank significantly

Proof sketch for $r = \#\sigma_1(\nabla f(\mathbf{X}^*))$

Lemma

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ *with SVD* $\mathbf{X} = \sum_{i=1}^{\min\{m,n\}} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, $\sigma_1 \geq \cdots \geq \sigma_n$. *If* $\|\mathbf{X}\|_* \geq 1$, *then*

$$\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}] = \sum_{i=1}^{\min\{m,n\}} \max\{\sigma_i - \sigma, 0\} \mathbf{u}_i \mathbf{v}_i^\top,$$

where $\sigma \geq 0$ *is unique solution to the equation* $\sum_i \max\{\sigma_i - \sigma, 0\} = 1$.

Lemma implies sufficient condition

$$\sum_{i=1}^{r} \sigma_i(\mathbf{X}) \geq 1 + r\sigma_{r+1}(\mathbf{X}) \quad \Longrightarrow \quad \operatorname{rank}(\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}]) \leq r$$

Proof sketch for $r = \#\sigma_1(\nabla f(\mathbf{X}^*))$

Let $\mathbf{X}^*$ be an optimal solution. Since singular vectors of $\mathbf{X}^*$ align with (minus) singular vectors of $\nabla f(\mathbf{X}^*)$ we have

$$\sigma_i(\mathbf{X}^* - \eta\nabla f(\mathbf{X}^*)) = \begin{cases} \sigma_i(\mathbf{X}^*) + \eta\sigma_1(\nabla f(\mathbf{X}^*)) & \text{if } i \le r; \\ \eta\sigma_i(\nabla f(\mathbf{X}^*)) & \text{if } i > r. \end{cases}$$

$$\sum_{i=1}^{r} \sigma_i(\mathbf{X}^* + \eta\nabla f(\mathbf{X}^*)) = \sum_{i=1}^{r} \sigma_i(\mathbf{X}^*) + \eta\sigma_1(\nabla f(\mathbf{X}^*)) = 1 + \eta r\sigma_1(\nabla f(\mathbf{X}^*))$$
$$= 1 + r\sigma_{r+1}(\mathbf{X}^* - \eta\nabla f(\mathbf{X}^*)) + r\eta(\sigma_1(\nabla f(\mathbf{X}^*)) - \sigma_{r+1}(\nabla f(\mathbf{X}^*)))$$

$\implies$ for $\mathbf{X}^*$ sufficient condition holds with positive slack.

Given some $\mathbf{X}$, apply smoothness and perturbation bounds for sing. vals (Ky Fan, Weyl) to replace $\mathbf{X}^*$ with $\mathbf{X}$, and obtain sufficient cond. holds for all $\mathbf{X}$ close enough to $\mathbf{X}^*$.

What if $\mathrm{rank}(\mathbf{X}^*) < \#\sigma_1(\nabla f(\mathbf{X}^*))$?

Should we still use SVD rank $\geq \#\sigma_1(\nabla f(\mathbf{X}^*))$?

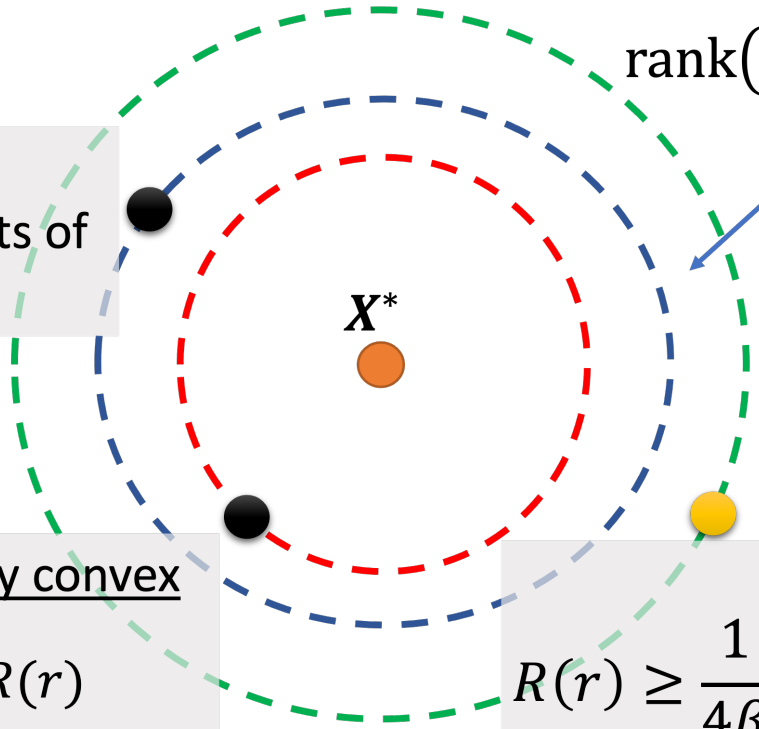
Theorem

Fix rank r, spectral gap $\sigma > 0$, and $\zeta > 0$ small enough. There exists a 1-smooth convex $f(\cdot)$ and feasible points $\mathbf{X}^*, \mathbf{X}$ such that $\mathbf{X}^*$ is a minimizer of $f(\cdot)$ over unit nuclear norm ball, $rank(\mathbf{X}^*) = rank(\mathbf{X}) = r$, and

1. $\#\sigma_1(\nabla f(\mathbf{X}^*)) = r + 1$, $\sigma_{r+1}(\nabla f(\mathbf{X}^*) - \sigma_{r+2}(\nabla f(\mathbf{X}^*) = \sigma$
2. $\|\mathbf{X} - \mathbf{X}^*\| \leq \zeta$
3. $\forall \eta \in (0, 1]$:
   $rank(\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X} - \eta \nabla f(\mathbf{X})]) = \#\sigma_1(\nabla f(\mathbf{X}^*)) > rank(\mathbf{X}^*)$

Some algorithmic implications

Methods converge with original convergence rates while requiring only rank-r SVD to compute projection.



AGD
depends on level sets of
$f(\cdot)$

AGD, α-strongly convex

$$\frac{\sqrt{\alpha}}{3\sqrt{\alpha+\beta}} R(r)$$

proj. grad.

$$R(r) \ge \frac{1}{4\beta} \max\left\{ \begin{matrix} \sigma_1 - \sigma_{r+1}, \\ \sqrt{\#\sigma_1}(\sigma_1 - \sigma_{r-\#\sigma_1+1}) \end{matrix} \right\}$$

Practice?

In practice, when computing the SVD, if we choose some SVD-rank parameter r, then as long as the projections have rank at most r, then we are guaranteed the projected gradient steps are accurate and method converges "correctly".

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\sigma_i = \sigma_i(\mathbf{X} - \eta \nabla f(\mathbf{X}))$, and fix $r \in \{1, \ldots, \min\{m, n\}\}$. From structure of Euclid. projection:

$$\operatorname{rank}(\Pi_{\|\cdot\|_* \leq 1}[\mathbf{X} - \eta \nabla f(\mathbf{X})]) \leq r \iff \sigma_{r+1} = 0 \text{ or } \sum_{i=1}^{r} \sigma_i \geq 1 + \sigma_{r+1}$$

Thus, suffices to compute rank-$(r+1)$ SVD to verify if low-rank projection is indeed correct.

Globally-convergent method via mix with Frank-Wolfe

Choose rank parameter r. On any iteration t:

1. compute (possibly in parallel):
   1. rank-$(r+1)$ SVD of $\mathbf{X}_t - \eta \nabla f(\mathbf{X}_t)$. Denote singular vals by $\sigma_1, \ldots, \sigma_{r+1}$
   2. $(\mathbf{u}_t, \mathbf{v}_t)$ - leading singular vectors of $\nabla f(\mathbf{X}_t)$
2. If $\sigma_{r+1} = 0$ or $\sum_{i=1}^{r} \geq 1 + \sigma_{r+1}$: take proj. grad step: $\mathbf{X}_{t+1} \leftarrow \Pi_{\|\cdot\|_* \leq 1}[\mathbf{X}_t - \eta \nabla f(\mathbf{X}_t)]$ (using above computed SVD).
3. Else: take Frank-Wolfe step: $\mathbf{X}_{t+1} \leftarrow (1 - \alpha_t)\mathbf{X}_t - \alpha_t \mathbf{u}_t \mathbf{v}_t^\top$ for some $\alpha_t \in [0, 1]$

Guarantees:

1. From any feasible initialization, method converges with rate $O(\beta/t)$.
2. Once $\|\mathbf{X}_t - \mathbf{X}^*\|_F \leq R(r)$ for some t, method only applies proj. grad. steps — maintains only a rank-r matrix in memory.

Empirical motivation from low-rank matrix completion

$$\min_{\|\mathbf{X}\|_* \leq \tau} \sum_{(i,j) \in S} (\mathbf{X}_{ij} - r_{ij})^2$$

S is the set of observed entries and r_{ij} denotes observed value.

dataset	trace	$\mathrm{rank}(\mathbf{X}^*)$	$\#\sigma_1(\nabla f^*)$	FISTA rank	PGD rank	MSE	gap
ML100k 943x1682	2500	3	3	3	3	1.3589	5.5844
	3000	10	10	10	10	0.9871	0.3234
	3500	41	41	42	41	0.7573	0.0456
	4000	70	70	71	70	0.5846	0.0227
	5000	117	117	118	117	0.3314	0.0148
ML1M 6040x3952	10000	3	3	3	3	1.2184	3.2861
	12000	12	12	12	12	0.9043	1.2056
	14000	74	74	75	74	0.7236	0.0698
	16000	155	155	157	155	0.5918	0.0119

Part 2: nonsmooth minimization

$$\min\{g(\mathbf{X}) \mid \mathbf{X} \succeq 0, \mathrm{Tr}(\mathbf{X}) = 1\}$$

Denote $\mathcal{S}_n = \{\mathbf{X} \in \mathbb{S}^n \mid \mathbf{X} \succeq 0, \mathrm{Tr}(\mathbf{X}) = 1\}$ – the spectrahedron

- $g(\cdot)$ is convex and nonsmooth
- $0 \notin \partial f$ over feasible set — optimal solutions do not perfectly fit data

Based on: *D. Garber and A.Kaplan, Low-rank extragradient method for nonsmooth and low-rank matrix optimization problems, NeurIPS 2021*.

Structure of optimal solutions

Definition

Let $\mathbf{X}^*$ be an optimal solution and let $\mathbf{G}^* \in \partial g(\mathbf{X}^*)$. If $\mathbf{G}^*$ satisfies: $\forall \mathbf{X} \in \mathcal{S}_n : \langle \mathbf{X} - \mathbf{X}^*, \mathbf{G}^* \rangle \geq 0$ (there always exists such $\mathbf{G}^*$), we call $\mathbf{G}^*$ a **special subgradient** at $\mathbf{X}^*$. .

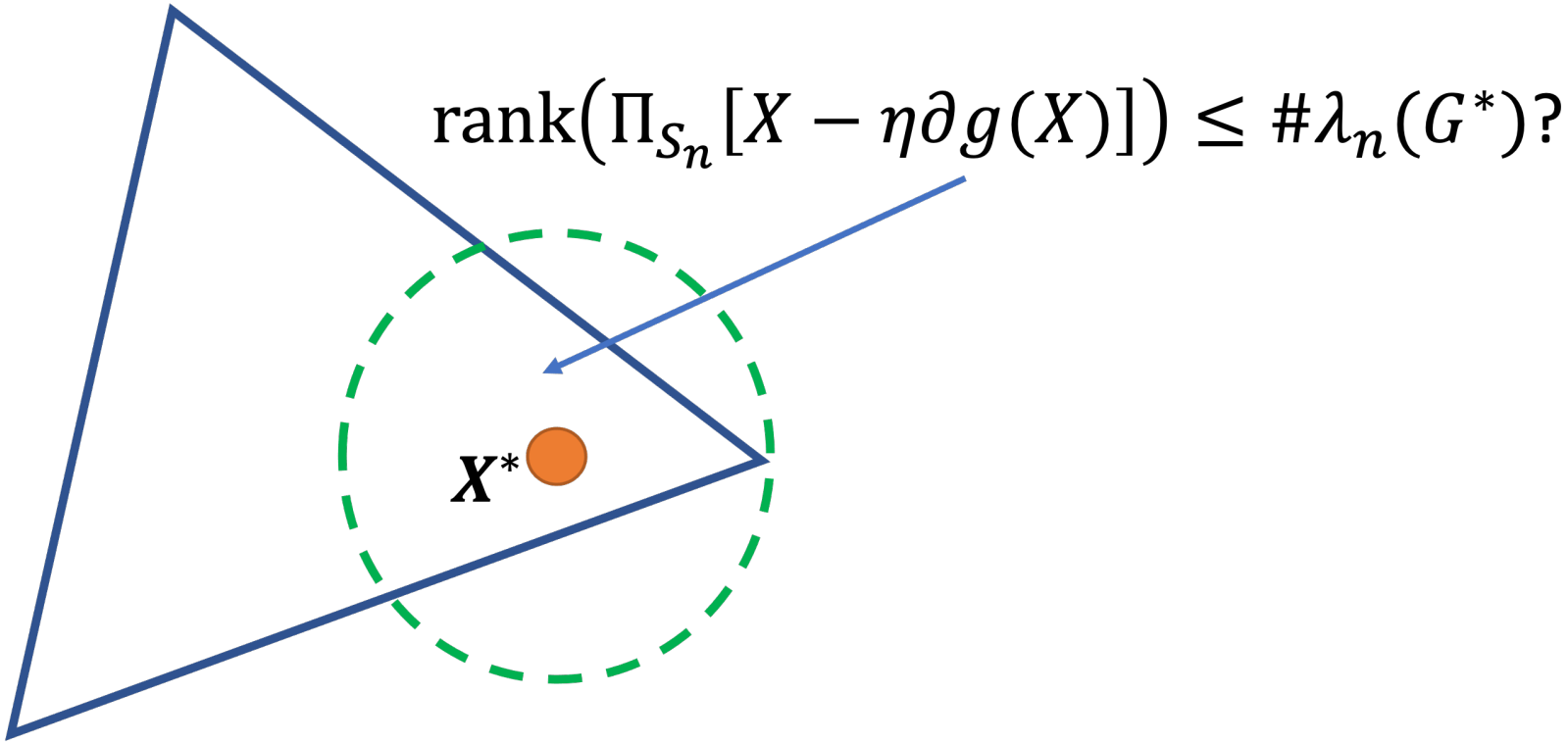
Lemma

Let $\mathbf{X}^*$ be an optimal solution with eigen-decomposition as $\mathbf{X}^* = \sum_{i=1}^{r^*} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$. Then, for any special subgradient $\mathbf{G}^* \in \partial g(\mathbf{X}^*)$, the set of vectors $\{\mathbf{v}_i\}_{i=1}^{r^*}$ is a set of eigenvectors of $\mathbf{G}^*$ corresponding to the eigenvalue $\lambda_n(\mathbf{G}^*)$.

In particular: $\operatorname{rank}(\mathbf{X}^*) \leq \#\lambda_n(\mathbf{G}^*)$.

Low-rank projected subgradient descent?

Projected Subgradient Descent is arguably the simplest and most general first-order method.



Failure of low-rank projected subgradient descent

Theorem

Consider the sparse-PCA problem

$$\min_{\mathbf{X}\in\mathcal{S}_n}\{g(\mathbf{X}) := -\left\langle \mathbf{z}\mathbf{z}^\top + \mathbf{z}_\perp\mathbf{z}_\perp^\top, \mathbf{X}\right\rangle + \frac{1}{2k}\|\mathbf{X}\|_1\},$$

$\mathbf{z} = (1/\sqrt{k},\ldots,1/\sqrt{k},0,\ldots,0)^\top$,
$\mathbf{z}_\perp = (0,\ldots,0,1/\sqrt{n-k},\ldots,1/\sqrt{n-k})^\top$, $k \le n/4$. *Then,* $\mathbf{z}\mathbf{z}^\top$ *is a (unique) rank-one optimal solution and* $\#\lambda_n(\mathbf{G}^*) = rank(\mathbf{z}\mathbf{z}^\top) = 1$. *However, for any* $\eta < \frac{2}{3}$ *and any* $\mathbf{v}\in\mathbb{R}^n$ *such that* $\|\mathbf{v}\| = 1$, $supp(\mathbf{v}) \subseteq supp(\mathbf{z})$, *and* $\|\mathbf{v}\mathbf{v}^\top - \mathbf{z}\mathbf{z}^\top\|_F \le \frac{1}{k}$, *it holds that*

$$rank\left(\Pi_{\mathcal{S}_n}[\mathbf{v}\mathbf{v}^\top - \eta\mathbf{G}_{\mathbf{v}\mathbf{v}^\top}]\right) > 1,$$

where $\mathbf{G}_{\mathbf{v}\mathbf{v}^\top} = -\mathbf{z}\mathbf{z}^\top - \mathbf{z}_\perp\mathbf{z}_\perp^\top + \frac{1}{2k}sign(\mathbf{v}\mathbf{v}^\top) \in \partial g(\mathbf{v}\mathbf{v}^\top)$.

Saddle-point approach

Consdier nonsmooth $g(\cdot)$ of the strucutre:

$$g(\mathbf{X}) = \max_{\mathbf{y} \in \mathcal{K}} f(\mathbf{X}, \mathbf{y}) = h(\mathbf{X}) + \max_{\mathbf{y} \in \mathcal{K}} \mathbf{y}^\top (\mathcal{A}\mathbf{X} - \mathbf{b})$$

$h(\cdot)$ is smooth and convex, $\mathcal{K}$ convex and compact

Lemma

$\mathbf{X}^*$ is minimizer of $g(\cdot)$ over $\mathcal{S}_n$ with special subgradient $\mathbf{G}^*$ if and only if there exists $\mathbf{y}^* \in \mathcal{K}$ such that $(\mathbf{X}^*, \mathbf{y}^*)$ is saddle point and $\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*) = \mathbf{G}^*$.

Solve via smooth convex-concave saddle-point methods.

Extragradient method [Korpelevich 76, Nemirovski 05]

$$\min_{\mathbf{X}\in\mathcal{S}_n}\ \max_{\mathbf{y}\in\mathcal{K}} f(\mathbf{X},\mathbf{y})$$

$$\mathbf{Z}_{t+1} \leftarrow \Pi_{\mathcal{S}_n}[\mathbf{X}_t - \eta\nabla_{\mathbf{X}} f(\mathbf{X}_t,\mathbf{y}_t)],\ \mathbf{w}_{t+1} \leftarrow \Pi_{\mathcal{K}}[\mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{X}_t,\mathbf{y}_t)]$$
$$\mathbf{X}_{t+1} \leftarrow \Pi_{\mathcal{S}_n}[\mathbf{X}_t - \eta\nabla_{\mathbf{X}} f(\mathbf{Z}_{t+1},\mathbf{w}_{t+1})],\ \mathbf{y}_{t+1} \leftarrow \Pi_{\mathcal{K}}[\mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{Z}_{t+1},\mathbf{w}_{t+1})]$$

Theorem

For $\eta \le \min\left\{\frac{1}{\beta_X+\beta_{Xy}}, \frac{1}{\beta_y+\beta_{yX}}, \frac{1}{\beta_X+\beta_{yX}}, \frac{1}{\beta_y+\beta_{Xy}}\right\}$ *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\max_{\mathbf{y}\in\mathcal{K}} f(\mathbf{Z}_{t+1},\mathbf{y}) - \frac{1}{T}\sum_{t=1}^{T}\min_{\mathbf{X}\in\mathcal{S}_n} f(\mathbf{X},\mathbf{w}_{t+1}) \le \frac{D^2}{2\eta T},$$

where $D := \sup_{(\mathbf{X},\mathbf{y}),(\tilde{\mathbf{X}},\tilde{\mathbf{y}})\in\mathcal{S}_n\times\mathcal{K}} \|(\mathbf{X},\mathbf{y}) - (\tilde{\mathbf{X}},\tilde{\mathbf{y}})\|_F$.

For $g(\mathbf{X}) = \max_{y\in\mathcal{K}} f(\mathbf{X},\mathbf{y})$ implies $\min_{t\in[T]} g(\mathbf{Z}_t) - g^* = O(1/T)$.

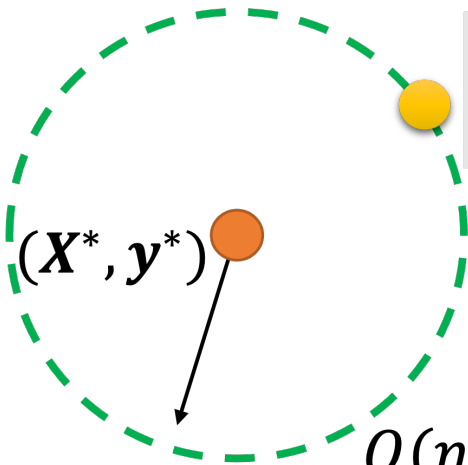
Main result for nonsmooth

Let $\mathbf{X}^*$ be optimal solution to

$$\min_{\mathbf{X}\in\mathcal{S}_n}\{h(\mathbf{X})+\max_{\mathbf{y}\in\mathcal{K}}\mathbf{y}^\top(\mathcal{A}\mathbf{X}-\mathbf{b})\}$$

Let $\mathbf{G}^*$ be special subgradient at $\mathbf{X}^*$ with eigenvalues $\lambda_n \le \lambda_{n-1}... \le \lambda_1$.

For all $r \ge \#\lambda_n(\mathbf{G}^*)$ we have:



Extragradient converges
With rank(projections) $\le r$

$$R(r) \ge \frac{O(\eta)}{1+\eta\max\{\beta_X,\beta_{Xy}\}}\max\left\{\begin{matrix}\lambda_{n-r}-\lambda_n \\ \sqrt{\#\lambda_n}(\lambda_{n-r+\#\lambda_n-1}-\lambda_n)\end{matrix}\right\}$$

Drawback: need "warm start" init. for $\mathbf{X}$ and $\mathbf{y}$

Empirical evidence - Sparse PCA

$$\min_{\substack{\mathrm{Tr}(\mathbf{X})=1,\\ \mathbf{X}\succeq 0}} \langle \mathbf{X}, -\mathbf{M}\rangle + \lambda\|\mathbf{X}\|_1 = \min_{\substack{\mathrm{Tr}(\mathbf{X})=1,\\ \mathbf{X}\succeq 0}} \max_{\|\mathbf{Y}\|_\infty \le 1} \{\langle \mathbf{X}, -\mathbf{M}\rangle + \lambda\langle \mathbf{X}, \mathbf{Y}\rangle\},$$

$$\mathbf{M} = \mathbf{z}\mathbf{z}^\top + \tfrac{c}{2}(\mathbf{N} + \mathbf{N}^\top)$$

dimension (n)	100	200	400	600
$\downarrow$ SNR $= 1 \downarrow$				
λ	0.006	0.003	0.0015	0.001
initialization error	0.1584	0.1464	0.1443	0.1411
recovery error	0.0059	0.0033	0.0019	0.0015
dual gap	8.6×10^{-4}	0.0031	0.0053	0.0060
$\lambda_{n-1}(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*)) - \lambda_n(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*))$	0.8406	0.8869	0.9178	0.9331
$\downarrow$ SNR $= 0.05 \downarrow$				
λ	0.04	0.02	0.01	0.005
initialization error	1.6701	1.6620	1.6542	1.6610
recovery error	0.0502	0.0234	0.0137	0.0109
dual gap	1.9×10^{-5}	0.0041	0.0534	0.0409
$\lambda_{n-1}(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*)) - \lambda_n(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*))$	0.2200	0.4076	0.5460	0.6788

All projections w.r.t. matrix variables are rank-one

Empirical evidence - Robust PCA

$$\min_{\substack{\mathrm{Tr}(\mathbf{X})=\tau,\\ \mathbf{X}\succeq 0}} \|\mathbf{X}-\mathbf{M}\|_1 = \min_{\substack{\mathrm{Tr}(\mathbf{X})=\tau,\\ \mathbf{X}\succeq 0}} \max_{\|\mathbf{Y}\|_\infty \le 1} \langle \mathbf{X}-\mathbf{M}, \mathbf{Y}\rangle,$$

$$\mathbf{M} = r\mathbf{Z}_0\mathbf{Z}_0^\top + \tfrac{1}{2}(\mathbf{N}+\mathbf{N}^\top)$$

dimension (n)	100	200	400	600
$\downarrow r = \mathrm{rank}(\mathbf{Z}_0\mathbf{Z}_0^\top) = 1 \downarrow$				
SNR	0.0021	7.2×10^{-4}	2.5×10^{-4}	1.3×10^{-4}
initialization error	1.3511	1.3430	1.2889	1.2606
recovery error	0.0084	0.0107	0.0109	0.0107
dual gap	0.0016	0.0029	0.0044	0.0069
$\lambda_{n-r}(\nabla_{\mathbf{X}} f(\mathbf{X}^*,\mathbf{y}^*)) - \lambda_n(\nabla_{\mathbf{X}} f(\mathbf{X}^*,\mathbf{y}^*))$	15.5944	41.2139	85.8117	140.5349
$\downarrow r = \mathrm{rank}(\mathbf{Z}_0\mathbf{Z}_0^\top) = 5 \downarrow$				
SNR	0.0110	0.0038	0.0013	6.9×10^{-4}
initialization error	1.5501	1.5527	1.5221	1.4833
recovery error	0.0092	0.0092	0.0087	0.0075
dual gap	0.0084	0.0390	0.1866	0.4721
$\lambda_{n-r}(\nabla_{\mathbf{X}} f(\mathbf{X}^*,\mathbf{y}^*)) - \lambda_n(\nabla_{\mathbf{X}} f(\mathbf{X}^*,\mathbf{y}^*))$	7.6734	26.2132	66.1113	108.7215

All projections w.r.t. matrix variables satisfy rank $=$ rank$(\mathbf{Z}_0\mathbf{Z}_0^\top)$

Empirical evidence - Low Rank & Sparse Recovery

$$\min_{\substack{\mathrm{Tr}(\mathbf{X})=1,\\ \mathbf{X}\succeq 0}} \frac{1}{2}\|\mathbf{X}-\mathbf{M}\|_F^2 + \lambda\|\mathbf{X}\|_1 = \min_{\substack{\mathrm{Tr}(\mathbf{X})=\tau,\\ \mathbf{X}\succeq 0}} \max_{\|\mathbf{Y}\|_\infty \le 1} \frac{1}{2}\|\mathbf{X}-\mathbf{M}\|_F^2 + \lambda\langle \mathbf{X}, \mathbf{Y}\rangle,$$

$$\mathbf{M} = \mathbf{Z}_0\mathbf{Z}_0^\top + \tfrac{c}{2}(\mathbf{N}+\mathbf{N}^\top)$$

dimension (n)	100	200	400	600
$\downarrow r = \mathrm{rank}(\mathbf{Z}_0\mathbf{Z}_0^\top) = 5$, SNR $= 2.4 \downarrow$				
λ	0.0012	0.0006	0.0003	0.0002
initialization error	0.2132	0.2103	0.1983	0.1907
recovery error	0.0641	0.0478	0.0349	0.0274
dual gap	9.0×10^{-4}	4.3×10^{-4}	1.4×10^{-4}	7.3×10^{-5}
$\lambda_{n-r}(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*)) - \lambda_n(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*))$	0.0148	0.0200	0.0257	0.0277
$\downarrow r = \mathrm{rank}(\mathbf{Z}_0\mathbf{Z}_0^\top) = 10$, SNR $= 4.8 \downarrow$				
λ	0.0007	0.0004	0.0002	0.0001
initialization error	0.1855	0.1661	0.1527	0.1473
recovery error	0.0702	0.0403	0.0268	0.0356
dual gap	4.9×10^{-4}	6.6×10^{-4}	4.2×10^{-4}	3.4×10^{-5}
$\lambda_{n-r}(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*)) - \lambda_n(\nabla_{\mathbf{X}} f(\mathbf{X}^*, \mathbf{y}^*))$	0.0072	0.0142	0.0187	0.0160

All projections w.r.t. matrix variables satisfy rank $=$ rank$(\mathbf{Z}_0\mathbf{Z}_0^\top)$

Additional results

- Lower bounds on neighborhoods scale with step-size η — inappropriate for SGD which needs $\eta \propto \epsilon$ — target accuracy. *D. Garber, On the convergence of stochastic gradient descent with low-rank projections for convex low-rank matrix problems, COLT 2020*, gives alternative analysis independent of η. However requires $\#\lambda_n(\nabla f(\mathbf{X}^*)) = \mathrm{rank}(\mathbf{X}^*)$ (strict complementarity).

- For PSD matrices (with unit trace), Euclidean norm often suboptimal. Working with Bregman distance induced by von Neuman entropy gives bounds which measure smoothness in spectral norm instead of Frobenius. *D. Garber and A.Kaplan, on the efficient implementation of the matrix exponentiated gradient algorithm for low-rank matrix optimization, In review*, gives a highly non-trivial extension. Also requires $\#\lambda_n(\nabla f(\mathbf{X}^*)) = \mathrm{rank}(\mathbf{X}^*)$.