Singular value thresholding operator solves proximal operator of nuclear norm

Andersen Ang

Department of Combinatorics and Optimization,
University of Waterloo, Waterloo, Canada

ms$\mathbf{x}$ang@uwaterloo.ca, where $\mathbf{x} = \lfloor \pi \rfloor$
Homepage: angms.science

First draft: December 30, 2018 Last update: July 19, 2021

The minimization problem

Given a matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$, find a matrix $\mathbf{X}$ by solving the following problem:

$$\mathbf{X} = \underset{\mathbf{X}}{\operatorname{argmin}} f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau \|\mathbf{X}\|_*, \qquad (\mathcal{P})$$

where

- $\mathbf{X} \in \mathbb{R}^{m \times n}$ is the optimization variable
- $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau \|\mathbf{X}\|_*$ is the objective function
  - $\tau \geq 0$ is a regularization parameter
  - $\|\cdot\|_F : \mathbb{R}^{m \times n} \to \mathbb{R}$ is the Frobenius norm
  - $\|\cdot\|_* : \mathbb{R}^{m \times n} \to \mathbb{R}$ is the Ky Fan norm (the Nuclear norm), defined as the sum of singular values:

    $$\|\mathbf{A}\|_* = \sum_i |\sigma_i(\mathbf{A})| = \sum_i \sigma_i(\mathbf{A}).$$

    (Note that singular values are nonnegative so we can drop the absolute sign)
  - $\|\cdot\|_F$ is smooth (differentiable) but $\|\cdot\|_*$ is not

Singular value thresholding operator

- SVD: for all $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{A} \overset{\text{SVD}}{=} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\text{Diag}\big(\sigma_i(\mathbf{A})\big)\mathbf{V}^\top$.
- Thresholding operator $\mathbf{D}_a$ of $\mathbf{\Sigma}$: $\mathbf{D}_a(\mathbf{\Sigma}) = \text{Diag}\big([\sigma_i - a]_+\big) = \text{Diag}\big(\max\{\sigma_i(\mathbf{A}) - a, 0\}\big)$.
- Singular value thresholding (SVT) of a matrix $\mathbf{A}$ is

$$\text{SVT}_a(\mathbf{A}) = \mathbf{U}\mathbf{D}_a(\mathbf{\Sigma})\mathbf{V}^\top.$$

- What SVT does: subtract all σ_i by a, then replace any negative value by zero. In other words SVT shrinks the singular values of $\mathbf{A}$ towards zero by a amount.
- If all σ_i are larger than a, then we can drop the max operator and write it as

$$\text{SVT}_a(\mathbf{A}) = \mathbf{U}(\mathbf{\Sigma} - a\mathbf{I})\mathbf{V}^\top.$$

- Important fact: if all σ_i are smaller than a, then we get a zero matrix after SVT.

Theorem relating the SVT operator and Problem $(\mathcal{P})$

- **Theorem**[1] $\mathsf{SVT}_\tau(\mathbf{Y}) = \mathbf{U}\mathbf{D}_\tau(\Sigma)\mathbf{V}^\top$ is the solution to Problem $(\mathcal{P})$. That is,

$$\mathsf{SVT}_\tau(\mathbf{Y}) = \underset{\mathbf{X}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau \|\mathbf{X}\|_*.$$

- This can be proved by using von Neumann trace inequality, see here.

- The proof using von Neumann trace inequality is actually very elegant and short. But in this document we will prove this using subgradient and 1st-order optimality (FOC) condition of nuclear norm. This is a nice exercise to familiarize the notion of subgradient.

[1] Cai, Jian-Feng, Emmanuel Candés, and Zuowei Shen. "A singular value thresholding algorithm for matrix completion". *SIAM Journal on Optimization*, 2010

The prerequisites to understand the proof: subgradient

- Subgradient is used here to provide a characterization of the 1st-order optimality condition (FOC) for a convex (and possibly non-smooth) function

- Sub-gradient. For convex (but not necessarily smooth function $f : \mathbb{R}^{m \times n} \to \mathbb{R}$, $\mathbf{Z}$ is a sub-gradient of f at a point $\mathbf{X} = \mathbf{X}_0$, denoted as $\mathbf{Z} \in \partial f(\mathbf{X}_0)$, if

$$f(\mathbf{X}) \geq f(\mathbf{X}_0) + \langle \mathbf{Z}, \mathbf{X} - \mathbf{X}_0 \rangle$$

  is true for all $\mathbf{X}$.

- First order optimality condition. A point $\mathbf{X}_0$ minimizes $f(\mathbf{X})$ if and only if $\mathbf{0}$ is a sub-gradient of f at $\mathbf{X}_0$. That is,

$$\mathbf{0} \in \mathbf{X}_0 - \mathbf{Y} + \tau \partial \|\mathbf{X}_0\|_*, \tag{1}$$

The prerequisites to understand the proof: about the nuclear norm

- Ky Fan norm / nuclear norm is a norm, therefore it is a convex function. See the proof here.

- The sub-differential of $\|\cdot\|*$ at a point $\mathbf{A} \overset{\text{SVD}}{=} \mathbf{U\Sigma V}^\top$ is the set

$$\partial\|\mathbf{A}\|_* = \left\{\mathbf{UV}^\top + \mathbf{W} \,\middle|\, \mathbf{W} \in \mathbb{R}^{m\times n}, \mathbf{U}^\top\mathbf{W} = \mathbf{0}, \mathbf{WV} = \mathbf{0}, \|\mathbf{W}\|_2 \leq 1\right\}.$$

- For the proof, see G.A. Watson, "Characterization of the subdifferential of some matrix norms", Linear Algebra and its Applications, vol.170, pp 33-45, 1992

The idea behind the proof ... 1/2

$$\mathbf{X} = \underset{\mathbf{X}}{\operatorname{argmin}}\, f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau\|\mathbf{X}\|_* \qquad (\mathcal{P})$$

- Note that $(\mathcal{P})$ has a unique minimizer as the objective function f is strictly convex:
  - $\|\mathbf{X} - \mathbf{Y}\|_F^2$ is strictly convex in $\mathbf{X}$
  - $\|\mathbf{X}\|_*$ is convex (see previous slide)
  - A strictly convex function + a convex function = a strictly convex function
- What we need to do in the proof: show the global minimizer $\mathbf{X}^*$ of $(\mathcal{P})$ equals to $\text{SVT}_\tau(\mathbf{Y})$.
  - To do that, we use the FOC for f.
  - Since f has a non-differentiable part, so we use the sub-gradient version of the FOC.
- Apply the sub-gradient FOC: a point $\mathbf{X}_0$ minimizes f if and only if $\mathbf{0} \in \partial f(\mathbf{X}_0) = \mathbf{X}_0 - \mathbf{Y} + \tau\partial\|\mathbf{X}_0\|_*$. i.e., $\mathbf{Y} - \mathbf{X}_0 \in \tau\partial\|\mathbf{X}_0\|_*$.
- i.e. to prove the theorem, we have to show that, for $\mathbf{X}_0 = \text{SVT}_\tau(\mathbf{Y})$, it minimizes $f(\mathbf{X})$ if and only if you subtract $\mathbf{X}_0$ from $\mathbf{Y}$, the result belongs to the set of sub-gradient of the Ky Fan norm multiplied by the constant τ.

The idea behind the proof ... 2/2

- As we need to show such inclusion, we need to know the sub-differential (set of sub-gradients) $\partial \|\mathbf{X}_0\|_*$, which is

$$\partial \|\mathbf{X}\|_* = \left\{ \mathbf{U}\mathbf{V}^\top + \mathbf{W} \,\middle|\, \mathbf{W} \in \mathbb{R}^{m \times n}, \mathbf{U}^\top \mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\|_2 \leq 1 \right\},$$

  where $\mathbf{U}, \mathbf{V}$ are obtained from the SVD of the matrix $\mathbf{X}$.
- Therefore, what we have to do, is to show, for $\mathbf{X}_0 = \text{SVT}_\tau(\mathbf{Y})$, we have

$$\mathbf{Y} - \mathbf{X}_0 \in \tau \left\{ \mathbf{U}\mathbf{V}^\top + \mathbf{W} \,\middle|\, \mathbf{W} \in \mathbb{R}^{m \times n}, \mathbf{U}^\top \mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\|_2 \leq 1 \right\}$$

- The essence of the proof is to make use of this inclusion relation :
  - obtain the expression $\mathbf{Y} - \mathbf{X}_0 = \mathbf{Y} - \text{SVT}_\tau(\mathbf{Y})$
  - show that $\mathbf{Y} - \text{SVT}_\tau(\mathbf{Y})$ is inside the set $\partial \|\mathbf{X}_0\|_*$

The proof ... 1/2

1. Obtain the expression $\mathbf{Y}-$ SVT$_\tau(\mathbf{Y})$

Split the SVD $\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ into two groups based on the sign of $\sigma_i - \tau$:

$$\mathbf{Y} = \mathbf{U}_{>\tau}\boldsymbol{\Sigma}_{>\tau}\mathbf{V}_{>\tau}^\top + \mathbf{U}_{\leq\tau}\boldsymbol{\Sigma}_{\leq\tau}\mathbf{V}_{\leq\tau}^\top. \tag{*}$$

Here the SVT operator with parameter τ kills the second term in (*):

$$\begin{aligned}
\mathsf{SVT}_\tau(\mathbf{Y}) &= \mathsf{SVT}_\tau\Big(\mathbf{U}_{>\tau}\boldsymbol{\Sigma}_{>\tau}\mathbf{V}_{>\tau}^\top\Big) + \mathsf{SVT}_\tau\Big(\mathbf{U}_{\leq\tau}\boldsymbol{\Sigma}_{\leq\tau}\mathbf{V}_{\leq\tau}^\top\Big) \\
&= \mathbf{U}_{>\tau}\mathbf{D}_\tau(\boldsymbol{\Sigma}_{>\tau})\mathbf{V}_{>\tau}^\top + \mathbf{U}_{\leq\tau}\underbrace{\mathbf{D}_\tau\big(\boldsymbol{\Sigma}_{\leq\tau}\big)}_{\mathbf{0}}\mathbf{V}_{\leq\tau}^\top \\
&= \mathbf{U}_{>\tau}\underbrace{\mathsf{Diag}(\max\{\sigma_i-\tau,0\})}_{\text{all positive}}\mathbf{V}_{>\tau}^\top \\
&= \mathbf{U}_{>\tau}\big(\boldsymbol{\Sigma}_{>\tau} - \mathrm{Diag}(\tau\mathbf{1})\big)\mathbf{V}_{>\tau}^\top \\
&= \mathbf{U}_{>\tau}\boldsymbol{\Sigma}_{>\tau}\mathbf{V}_{>\tau}^\top - \tau\mathbf{U}_{>\tau}\mathbf{V}_{>\tau}^\top \\
\mathbf{Y} - \mathsf{SVT}_\tau(\mathbf{Y}) &\overset{(*)}{=} \mathbf{U}_{>\tau}\boldsymbol{\Sigma}_{>\tau}\mathbf{V}_{>\tau}^\top + \mathbf{U}_{\leq\tau}\boldsymbol{\Sigma}_{\leq\tau}\mathbf{V}_{\leq\tau}^\top - \mathbf{U}_{>\tau}\boldsymbol{\Sigma}_{>\tau}\mathbf{V}_{>\tau}^\top + \tau\mathbf{U}_{>\tau}\mathbf{V}_{>\tau}^\top \\
&= \mathbf{U}_{\leq\tau}\boldsymbol{\Sigma}_{\leq\tau}\mathbf{V}_{\leq\tau}^\top + \tau\mathbf{U}_{>\tau}\mathbf{V}_{>\tau}^\top \\
&= \tau(\mathbf{W} + \mathbf{U}_{>\tau}\mathbf{V}_{>\tau}^\top)
\end{aligned}$$

where $\mathbf{W} = \tau^{-1}\mathbf{U}_{\leq\tau}\boldsymbol{\Sigma}_{\leq\tau}\mathbf{V}_{\leq\tau}^\top$

The proof . . . 2/2

2. Show $\mathbf{Y}-$ SVT$_\tau(\mathbf{Y}) \in \tau\partial\|\mathbf{X}\|_*$

We have to show $\mathbf{W} + \mathbf{U}_{>\tau}\mathbf{V}_{>\tau}^\top$ is inside the set $\left\{\mathbf{U}\mathbf{V}^\top + \mathbf{W} \,\middle|\, \mathbf{W} \in \mathbb{R}^{m\times n}, \mathbf{U}^\top\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\|_2 \le 1\right\}$. by checking the three inclusion conditions:

- Does $\|\mathbf{W}\|_2 \le 1$? YES : $\|\mathbf{W}\|_2 = \tau^{-1}\|\mathbf{\Sigma}_{\le\tau}\|_2 \le 1$.
- Does $\mathbf{U}^\top\mathbf{W} = \mathbf{0}$? YES : as columns of $\mathbf{U}$ are orthonormal to each other, so
$$\mathbf{U}_{>\tau}^\top\mathbf{W} = \tau^{-1}\underbrace{\mathbf{U}_{>\tau}^\top\mathbf{U}_{\le\tau}}_{\mathbf{0}}\mathbf{\Sigma}_{\le\tau}\mathbf{V}_{\le\tau}^\top = \mathbf{0}$$
- Does $\mathbf{W}\mathbf{V}_{>\tau} = \mathbf{0}$?
YES : as columns of $\mathbf{V}$ are orthonormal to each other, so
$$\mathbf{W}\mathbf{V} = \tau^{-1}\mathbf{U}_{>\tau}^\top\mathbf{\Sigma}_{\le\tau}\underbrace{\mathbf{V}_{\le\tau}^\top\mathbf{V}_{>\tau}}_{\mathbf{0}} = \mathbf{0}.$$

End of document