Subgradient of singular value function

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : June 2, 2020
Last update : June 4, 2020

A subgradient formula

- Given a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the SVD of $\mathbf{X}$.
  - Denote $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{X})$ the vector of singular value of $\mathbf{X}$.
  - Denote σ_i be the ith largest singular value of $\mathbf{X}$.

- Given a function $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ that maps $\mathbf{X}$ into a real number. Suppose this f is a function g of $\boldsymbol{\sigma}$:

$$f(\mathbf{X}) = g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big).$$

- Question : what is the gradient (subgradient) of f with respect to $\mathbf{X}$ at a point $\mathbf{X}_0$?

- Answer : there is a close form formula to comput that.

$$\partial f(\mathbf{X})\Big|_{\mathbf{X}=\mathbf{X}_0} = \mathbf{U}\mathrm{Diag}\Big\{\partial g\Big(\boldsymbol{\sigma}(\mathbf{X}_0)\Big)\Big\}\mathbf{V}^\top.$$

Defintions : Absolutely Symmetric Function (ASF)

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is an ASF if

$$f(x_1, x_2, \ldots, x_n) = f(|x_{\pi(1)}|, |x_{\pi(2)}|, \ldots, |x_{\pi(n)}|)$$

  holds for any permutation π on $[n] = \{1, 2, \ldots, n\}$.

- Other names of ASF : Totally symmetric function, Symetric Gauge.

- Examples of ASF.
  - L_p norm : $\left(\sum_i x^p\right)^{1/p}$
  - $\sum_i \log(|x_i| + c)$
  - $\prod_i x_i^{2k}$, $k \in \{0, 1, \ldots\}$

- Functions that are not ASF.
  - $\sum_i \log x_i$: because $\log x_i \neq \log(-x_i)$
  - $\sum_i x_i$: for $\mathbf{x} = [-1, 1]$, then $-1 + 1 \neq |-1| + |1|$.

  We see that for a function to be ASF, it has to be invariant under signed permutation.

Definition : Orthogonally Invariant Function (OIF)

- A function $f : \mathbb{R}^{m\times n} \to \{\mathbb{R} \cup \pm\infty\}$, with $n \leq m$, is an OIF if

$$f(\mathbf{U}^\top \mathbf{X}\mathbf{V}) = f(\mathbf{X})$$

  holds for all $\mathbf{X} \in \mathbb{R}^{m\times n}$ and all orthogonal matrices $\mathbf{U} \in \mathbb{R}^{m\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$.

- Note : the conditon $n \leq m$ can be assumed without loss of generality. If a matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ has $n > m$, we can just consider its transpose.

- Examples of OIF : trace function, Frobenius norm.

A theorem linking OIF and ASF

- A function $f : \mathbb{R}^{m \times n} \to \{\mathbb{R} \cup \pm\infty\}$, with $n \leq m$, is an OIF if and only if it is a function g of singular value of $\mathbf{X}$ in the form

$$f(\mathbf{X}) = g \circ \boldsymbol{\sigma}(\mathbf{X})$$

  where g is ASF. In other words, if f is an OIF, then it can be expressed as an ASF on the singular values of $\mathbf{X}$.

- A slightly more "intuitive" expression is

$$f(\mathbf{X}) = g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big).$$

  We will stick with this notation instead of $g \circ \boldsymbol{\sigma}(\mathbf{X})$.

- The proof of this theorem : J. von Neumann, "Some matrix inequalities and metrization of metric spaces", Tomsk Univ. Rev. 1:286-300 (1937).

Subdifferential of singular value function

- A useful formula concerning subdifferential of a singular value function f at the point $\mathbf{X} \in \mathbb{R}^{m \times n}$ in terms of subdifferential of the corresponding ASF g at the singular vector $\boldsymbol{\sigma}(\mathbf{X})$.

- Let $g : \mathbb{R}^n \to \{\mathbb{R} \cup \pm\infty\}$ be an ASF, then subdifferential of the corresponding singular value function $g\big(\boldsymbol{\sigma}(\mathbf{X})\big)$ at the point $\mathbf{X} \in \mathbb{R}^{m \times n}$ is

$$\partial f(\mathbf{X}) = \partial g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big) = \mathbf{U}\mathrm{Diag}\Big\{\partial g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big)\Big\}\mathbf{V}^\top,$$

  where $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{\Sigma} = \mathrm{Diag}\big\{\boldsymbol{\sigma}(\mathbf{X})\big\} \in \mathbb{R}^{n \times n}$.

- For the proof of the formula, see Theorem 7.1 in : A.S. Lewis and H.S. Sendov "Nonsmooth analysis of singular values, part I: theory", Set-Valued Analysis, 13 (3) (2005), pp. 213-241

More about the formula

- To use the formula on a function $f(\mathbf{X})$:
  - First we need to make sure that f is an OIF
  - Then, by the Theorem linking OIF ans ASF, $f(\mathbf{X}) = g\big(\boldsymbol{\sigma}(\mathbf{X})\big)$, we need to find the expression of g.
  - Use the formula to compute the subgradient

$$\partial f(\mathbf{X}) = \partial g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big) = \mathbf{U}\mathrm{Diag}\Big\{\partial g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big)\Big\}\mathbf{V}^\top.$$

- This formula is useful (in optimization) when
  - we need to compute the subgradient of a function of singular values.
  - we need a singular value characterization of the subgradient of f.

Example 1 : Frobenius norm squared ... (1/2))

- Let $f(\mathbf{X}) = \|\mathbf{X}\|_F^2$. We know that
  - $\|\mathbf{X}\|_F^2 = \sum_i \sigma_i^2$.
  - $\|\mathbf{X}\|_F^2 = \operatorname{Tr} \mathbf{X}^\top \mathbf{X}$.
  - $\nabla_{\mathbf{X}} f(\mathbf{X}) = 2\mathbf{X}$.

  Now we use the subgradient formula to derive the same $\nabla_{\mathbf{X}} f(\mathbf{X})$.

- First, we show f is an OIF.

$$\begin{aligned} f(\mathbf{U}^\top \mathbf{X} \mathbf{V}) &= \operatorname{Tr} (\mathbf{U}^\top \mathbf{X} \mathbf{V})^\top \mathbf{U}^\top \mathbf{X} \mathbf{V} \\ &= \operatorname{Tr} \mathbf{V}^\top \mathbf{X}^\top \underbrace{\mathbf{U}\mathbf{U}^\top}_{=\mathbf{I}_m} \mathbf{X} \mathbf{V} \\ &= \operatorname{Tr} \mathbf{X}^\top \mathbf{X} \underbrace{\mathbf{V}\mathbf{V}^\top}_{=\mathbf{I}_n} \qquad = \quad f(\mathbf{X}). \end{aligned}$$

  So, $f(\mathbf{X}) = g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big)$ for some g that is ASF.

- Next, we have $g(\mathbf{t}) = \sum t_i^2 = \|\mathbf{t}\|_2^2$, which is ASF (trivial).

Example 1 : Frobenius norm squared ... (2/2))

- Compute the gradient of g. In fact from $g(\mathbf{t}) = \|\mathbf{t}\|_2^2$, by vector calculus we know directly $\nabla_{\mathbf{t}} g(\mathbf{t}) = 2\mathbf{t}$. But here we still evaluate it from the definition :

$$\nabla_{\mathbf{t}} g(\mathbf{t}) = \begin{bmatrix} \frac{\partial}{\partial t_1} \sum t_i^2 \\ \vdots \\ \frac{\partial}{\partial t_r} \sum t_i^2 \end{bmatrix} = \begin{bmatrix} 2t_1 \\ \vdots \\ 2t_r \end{bmatrix} = 2\mathbf{t},$$

  so $\text{Diag}\Big(\partial g(\mathbf{t})\Big) = 2\text{Diag}(\mathbf{t})$.

- Finally $\partial f(\mathbf{X}) = \partial g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big)$ and

$$\partial f(\mathbf{X}) = \mathbf{U}\text{Diag}\Big\{\partial g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big)\Big\}\mathbf{V}^\top = 2\mathbf{U}\underbrace{\text{Diag}\Big\{\boldsymbol{\sigma}(\mathbf{X})\Big\}}_{=\boldsymbol{\Sigma}}\mathbf{V}^\top = 2\mathbf{X}.$$

Example 2 : a trace function ... (1/2)

- Let $f(\mathbf{X}) = \mathrm{Tr}\left((\mathbf{X}^\top\mathbf{X} + \delta\mathbf{I})^p\right)$, where $\mathbf{X} \in \mathbb{R}^{m\times n}$.
- This f is a special case of the smooth Schattern-p norm.
- The singular value expression :

$$
\begin{aligned}
f(\mathbf{X}) &= \mathrm{Tr}\left\{(\mathbf{X}^\top\mathbf{X} + \delta\mathbf{I})^p\right\} \\
&= \mathrm{Tr}\left\{\left((\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top) + \delta\mathbf{I}\right)^p\right\} \\
&= \mathrm{Tr}\left\{\left(\mathbf{V}\mathbf{\Sigma}^\top\underbrace{\mathbf{U}^\top\mathbf{U}}_{=\mathbf{I}_r}\mathbf{\Sigma}\mathbf{V}^\top + \delta\underbrace{\mathbf{I}}_{=\mathbf{V}\mathbf{V}^\top}\right)^p\right\} \\
&= \mathrm{Tr}\left\{\left(\left(\mathbf{V}(\mathbf{\Sigma}^2 + \delta\mathbf{I})\mathbf{V}^\top\right)^p\right\} \\
&= \mathrm{Tr}\left\{\mathbf{V}(\mathbf{\Sigma}^2 + \delta\mathbf{I})\mathbf{V}^\top\mathbf{V}(\mathbf{\Sigma}^2 + \delta\mathbf{I})\mathbf{V}^\top \ldots \mathbf{V}(\mathbf{\Sigma}^2 + \delta\mathbf{I})\mathbf{V}^\top\right\} \\
&= \mathrm{Tr}\left\{\mathbf{V}(\mathbf{\Sigma}^2 + \delta\mathbf{I})^p\mathbf{V}^\top\right\} \\
&= \mathrm{Tr}\left\{(\mathbf{\Sigma}^2 + \delta\mathbf{I})^p\underbrace{\mathbf{V}^\top\mathbf{V}}_{=\mathbf{I}_r}\right\} = \mathrm{Tr}\left\{(\mathbf{\Sigma}^2 + \delta\mathbf{I})^p\right\}
\end{aligned}
$$

As $\mathbf{\Sigma}^2 + \delta\mathbf{I}$ is a diagonal matrix, so the power to p times is a diagonal matrix. Thus the trace is $\sum_i(\sigma_i^2 + \delta)^p$.

Example 2 : a trace function ... (2/2)

- We have $g(\mathbf{t}) = \sum_i (t_i^2 + \delta)^p$.
- It is easy to check g is ASF.
- The gradient of g with respect to one element is

$$\frac{\partial}{\partial t_i} g(\mathbf{t}) = \frac{\partial}{\partial t_i} \sum_i (t_i^2 + \delta)^p = p(t_i^2 + \delta)^{p-1} 2t_i = 2p \frac{t_i}{(t_i^2 + \delta)^{1-p}}$$

- So the whole gradient is

$$\partial g = 2p \left[\frac{t_i}{(t_i^2 + \delta)^{1-p}} \right]_i .$$

- The subgradient of f is then

$$\begin{aligned} \partial f(\mathbf{X}) &= \mathbf{U}\mathrm{Diag}\{\partial g(\boldsymbol{\sigma}(\mathbf{X}))\}\mathbf{V}^\top \\ &= \mathbf{U}\mathrm{Diag}\left\{ 2p \frac{\sigma_i}{(\sigma_i^2 + \delta)^{1-p}} \right\} \mathbf{V}^\top \\ &= 2p\mathbf{U}\mathrm{Diag}\left\{ \frac{\sigma_i}{(\sigma_i^2 + \delta)^{1-p}} \right\} \mathbf{V}^\top . \end{aligned}$$

Last page - summary

- Absolutely Symmetric Function (ASF).
- Orthogonally Invariant Function (OIF.)
- If f is OIF, then $f(\mathbf{X}) = g\Big(\boldsymbol{\sigma}(\mathbf{X})\Big)$, where g is ASF.
- If f that is OIF, then the subgradient of f w.r.t. $\mathbf{X}$ at the point $\mathbf{X}_0$ is

$$\partial f(\mathbf{X})\Big|_{\mathbf{X}=\mathbf{X}_0} = \mathbf{U}\mathrm{Diag}\Big\{\partial g\Big(\boldsymbol{\sigma}(\mathbf{X}_0)\Big)\Big\}\mathbf{V}^\top.$$

End of document