Data-driven statistical modelling with optimisation VT21 HW5 (Revised)

Fanny Bergström

March 25, 2021

Exercise 5.8

In this exercise, we consider the composite gradient update (5.26 in [1]) given by

$$\text{prox}_{sh}(Z) = \underset{\Theta \in \mathbb{R}^{m\times n}}{\text{argmin}} \left\{ \frac{1}{2s} ||Z - \Theta||_F^2 + \lambda ||\Theta||_* \right\}, \tag{1}$$

where $||Z\Theta||_F^2 = \sum_{j=1}^{m} \sum_{k=}^{n} (Z_{jk}\Theta_{jk})^2$ and $||\Theta||_* = \sum_{j=1}^{m} \sigma_j(\Theta)$, with $\sigma_j(\Theta)$ being the j:th singular values of Θ. We will show that when h is given by the nuclear norm, the composite gradient update can be obtained by the following procedure:

a) Compute the singular value decomposition of the input matrix Z, that is $Z = UDV^T$ where $D =$diag$\{\sigma_j(Z)\}$ is a diagonal matrix of the singular values.

b) Apply the soft-thresholding operator (5.25) to compute the "shrunken" singular values

$$\gamma_j := S_{s\lambda}(\sigma_j(Z)), \text{ for } j = 1, ..., p.$$

c) Return the matrix $\hat{Z} = U\text{diag}\{\gamma_1, ..., \gamma_p\}V^T$.

Answer:

The solution to this exercise was derived with the help of theorem 3.1 (proof in the appendix) in the paper by Ji and Ye [2].

a) Because of convexity (quadratic term + norm), the optimal solution to Eq (1), is given by

$$0 \in -\frac{1}{s\lambda}(Z - \Theta) + \partial ||\Theta||_*, \tag{2}$$

where ∂ is the subgradient. If we let $\Theta = P_1 \Sigma P_2^T$ be the SVD of Θ, it is known [3] that

$$\partial ||\Theta||_* = \{P_1 P_2^T + S : S \in \mathbb{R}^{m \times n}, P^T S = 0, SP_2 = 0, ||S||_2 \leq 1\}, \qquad (3)$$

where $|| \cdot ||_2$ is the spectral norm.

b) Decomposing the SVD of Z into $Z = U_0 D_0 V_0^T + U_1 D_1 V_1^T$, where $U_0 D_0 V_0^T$ corresponds to the part with singular values greater than $s\lambda$. Since the singular values of Z only takes positive values, computing the "shrunken" singular values of Z with the soft threshold operator yields

$$\begin{aligned} \gamma_j := S_{s\lambda}(\sigma_j(Z)) &= \max(\sigma_j - s\lambda, 0) \\ &= \max(\sigma_{0j} - s\lambda, 0), \end{aligned}$$

where σ_{0j} is the jth diagonal element of D_0.

c) Next, we return the full matrix

$$\hat{Z} = U \text{diag}\{\gamma_1, ..., \gamma_p\} V^T = U_0 (D_0 - s\lambda I) V_0^T.$$

and we have that

$$\begin{aligned} \frac{1}{s\lambda}(Z - \hat{Z}) = &\frac{1}{s\lambda}(UDV^T - U_0(D_0 - s\lambda I)V_0^T) \\ &\frac{1}{s\lambda}(U_0 D_0 V_0^T + U_1 D_1 V_1^T - U_0 D_0 V_0^T - s\lambda U_0^T I V_0^T) \\ &\frac{1}{s\lambda} U_1 D_1 V_1^T + U_0^T V_0^T. \end{aligned}$$

If we let $S = \frac{1}{s\lambda} U_1 D_1 V_1^T)$, we see that this corresponds to the conditions in Eq (3) are true such that it is the subgraient and it follows that Eq (**??**) will hold if we let $\Theta = \hat{Z}$. We have then showed how the composite gradient update can be obtained with the steps a)-c).

References

[1] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* Chapman Hall/CRC, 2015.

[2] Shuiwang Ji and Jieping Ye. "An Accelerated Gradient Method for Trace Norm Minimization". In: *Proceedings of the 26th Annual International Conference on Machine Learning.* Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 457–464.

[3] G.A. Watson. "Characterization of the subdifferential of some matrix norms". In: *Linear Algebra and its Applications* 170 (1992), pp. 33–45.