Homework 1: Mathematics of Deep Learning (EN 580.745)

Instructor: René Vidal, Biomedical Engineering, Johns Hopkins University

Due Date: 10/24/2019, 11:59PM Eastern Time

Instructions.

- You can discuss the problems with your peers at a high level, but you must write your own solutions.
- You can consult relevant background material, but don't seek out solutions to the problems themselves. Cite the outside material use.
- Submit a single PDF to Blackboard. Typed solutions are nice, but clearly handwritten solutions are also fine.

1. Properties of the ℓ_1 norm. Let $\boldsymbol{x} \in \mathbb{R}^n$ and recall the ℓ_1 norm $\|\boldsymbol{x}\|_1 \triangleq \sum_i |x_i|$.

(a) **(10 points)** The subdifferential $\partial f(\boldsymbol{x})$ of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}$ is defined to be

$$\partial f(\boldsymbol{x}) \triangleq \{\boldsymbol{z} \in \mathbb{R}^n \mid f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x} \rangle \text{ for all } \boldsymbol{y} \in \mathbb{R}^n\}. \tag{1}$$

Prove that

$$\partial \|\boldsymbol{x}\|_1 = \{\operatorname{sign}(\boldsymbol{x}) + \boldsymbol{w} \mid \operatorname{supp}(\boldsymbol{w}) \subseteq \operatorname{supp}(\boldsymbol{x})^c, \max_i |w_i| \leq 1\}, \tag{2}$$

where $\operatorname{sign}\colon \mathbb{R}^n \to \{-1, 0, 1\}$ denotes the sign function, and $\operatorname{supp}(\boldsymbol{x}) \subseteq [n]$ denotes the support of $\boldsymbol{x}$, that is the set of indices where $\boldsymbol{x}$ is non-zero.

(b) **(10 points)** Define the proximal operator $\operatorname{prox}_f\colon \mathbb{R}^n \to \mathbb{R}^n$ of a convex function f to be

$$\operatorname{prox}_f(\boldsymbol{x}) \triangleq \operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{x}\|_2^2 + f(\boldsymbol{a}). \tag{3}$$

Prove that $\operatorname{prox}_{\tau\|\cdot\|_1}(\boldsymbol{x}) = \mathcal{S}_\tau(\boldsymbol{x})$, where $\mathcal{S}_\tau$ is the elementwise soft-thresholding operator

$$\mathcal{S}_\tau(x_i) \triangleq \operatorname{sign}(x_i) \max(|x_i| - \tau, 0). \tag{4}$$

2. Properties of the nuclear norm. Let $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ be a matrix of rank r. Recall the nuclear norm $\|\boldsymbol{X}\|_* \triangleq \sum_{i=1}^r \sigma_i(\boldsymbol{X})$, where $\sigma_i(\boldsymbol{X})$ denotes the ith singular value of $\boldsymbol{X}$. Let $\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$ be the compact SVD, so that $\boldsymbol{U} \in \mathbb{R}^{D \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, and $\boldsymbol{V} \in \mathbb{R}^{N \times r}$. Recall also the spectral norm $\|\boldsymbol{X}\|_2 = \sigma_1(\boldsymbol{X})$.

(a) **(10 points)** Prove that [1]

$$\partial \|\boldsymbol{X}\|_* = \{\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W} \mid \boldsymbol{U}^\top \boldsymbol{W} = \boldsymbol{0}, \boldsymbol{W}\boldsymbol{V} = \boldsymbol{0}, \|\boldsymbol{W}\|_2 \leq 1\}. \tag{5}$$

(b) **(10 points)** Let $\tau > 0$. Prove that $\operatorname{prox}_{\tau\|\cdot\|_*}(\boldsymbol{X}) = \boldsymbol{U}\mathcal{S}_\tau(\Sigma)\boldsymbol{V}^\top$, where $\mathcal{S}_\tau$ is as in (4). Note that when generalizing (3) to matrices, the squared Frobenius norm $\frac{1}{2}\|\boldsymbol{A} - \boldsymbol{X}\|_F^2$ is used in place of the squared ℓ_2 norm ($\|\boldsymbol{X}\|_F \triangleq (\sum_{ij} X_{ij}^2)^{\frac{1}{2}}$).

3. Low rank matrix factorization. Let $\boldsymbol{Y} \in \mathbb{R}^{D \times N}$ be of rank r. Consider the matrix factorization problem

$$\operatorname*{minimize}_{\boldsymbol{U}, \boldsymbol{V}} \quad F(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2 + \frac{\tau}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2), \tag{6}$$

where $\boldsymbol{U} \in \mathbb{R}^{D \times r}$, $\boldsymbol{V} \in \mathbb{R}^{N \times r}$. This is in fact a regularized variant of the problem studied in (Baldi & Hornik, 1989). It is also closely related to the nuclear norm proximal operator from Problem 2(b), as we will see.

(a) **(10 points)** Let $\boldsymbol{Y} = \boldsymbol{P}\Sigma\boldsymbol{Q}^\top$ be the compact SVD. Define

$$\widehat{\boldsymbol{U}} = \boldsymbol{P}\mathcal{S}_\tau(\Sigma)^{\frac{1}{2}} \qquad \widehat{\boldsymbol{V}} = \boldsymbol{Q}\mathcal{S}_\tau(\Sigma)^{\frac{1}{2}}, \tag{7}$$

[1] *Hint:* the Von-Neumann trace inequality could be useful.

where the square root of a diagonal matrix is applied elementwise. Prove that $(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}})$ is a critical point of (6).

(b) **(bonus)** Prove that $(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}})$ is in fact a global minimizer.

Problem (6) and the nuclear norm proximal operator are therefore closely related, in the sense that the global minimizer $(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}})$ satisfies $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{V}}^\top = \mathrm{prox}_{\tau\|\cdot\|_*}(\boldsymbol{Y})$.