Convex Optimization for Machine Learning and Computer Vision

Lecture: Dr. Tao Wu
Exercises: Yuesong Shen, Zhenzhang Ye
Winter Semester 2018/19

Computer Vision Group
Institut für Informatik
Technische Universität München

Weekly Exercises 8

Room: 01.09.014
Wednesday, 19.12.2018, 12:15-14:00
Submission deadline: Monday, 17.12.2018, 16:15, Room 01.09.014

Prox and Gradient descent (8+4 Points)

Exercise 1 (6 Points). Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and bounded from below. Consider the scaled gradient descent iteration:

$$x^{k+1} = x^k - \tau^k (H^k)^{-1} \nabla f(x^k). \tag{1}$$

For each k, assume that $\tau^k > 0$, $\nabla f(x^k) \neq 0$, and $H^k \in \mathbb{R}^{n \times n}$ is symmetric positive definite.

1. Prove that for given x^k and H^k, there exists some $\bar{\tau}^k > 0$ such that any $\tau^k \in (0, \bar{\tau}^k]$ will fulfill the following Armijo condition:

$$f(x^{k+1}(\tau^k)) \leq f(x^k) + c \left\langle \nabla f(x^k), x^{k+1}(\tau^k) - x^k \right\rangle, \tag{2}$$

   for some constant $0 < c < 1$.

2. Assume that for each k the condition (2) is satisfied with some chosen $\tau^k > 0$. In addition, assume that $\liminf_{k\to\infty} \tau^k = C_1 > 0$ and $\limsup_{k\to\infty} \lambda_{\max}(H^k) = C_2 < \infty$. Prove $\lim_{k\to\infty} \nabla f(x^k) = 0$.

Solution.

1. Consider both sides of (2) as functions of τ^k. Then we have $\mathrm{LHS}(0) = \mathrm{RHS}(0)$ and $\mathrm{LHS}'(0) - \mathrm{RHS}'(0) = (c-1)\left\langle \nabla f(x^k), (H^k)^{-1} \nabla f(x^k) \right\rangle < 0$. Hence $\mathrm{LHS}(\tau^k) < \mathrm{RHS}(\tau^k)$ as $\tau^k \to 0^+$. On the other hand, since $\mathrm{LHS}(\cdot)$ is bounded from below and $\mathrm{RHS}(\cdot)$ is strictly decreasing on $[0, \infty)$, they must intersect at some $\tau^k \in (0, \infty)$. Let $\bar{\tau}^k > 0$ be the first of such points, then $\mathrm{LHS}(\bar{\tau}^k) \leq \mathrm{RHS}(\bar{\tau}^k)$ for all $\tau^k \in (0, \bar{\tau}^k]$.

2. Note that $\{f(x^k)\}$ is a non-increasing sequence that is bounded from below. For sufficiently large k, we have $\tau^k \geq C_1/2$ and $\lambda_{\max}(H^k) \leq 2C_2$, and therefore $c\frac{C_1}{2}\frac{1}{2C_2}\|\nabla f(x^k)\|_2^2 \leq c\tau^k \left\langle \nabla f(x^k), (H^k)^{-1}\nabla f(x^k) \right\rangle \leq f(x^k) - f(x^{k+1}) \to 0$. Hence, $\nabla f(x^k) \to 0$.

Exercise 2 (6 points)**.** We want to show that the proximal operator of the nuclear norm is the proximal operator of the ℓ_1-norm applied to the singular values of the input argument. Formally, let $Y \in \mathbb{R}^{n\times n}$ and let $Y = U\Sigma V^\top$ be the singular value decomposition of Y. Our goal is to prove that

$$\mathrm{prox}_{\tau\|\cdot\|_{\mathrm{nuc}}}(Y) = U\mathrm{diag}(\{(\sigma_i - \tau)_+\})V^\top,$$

where $\mathrm{diag}(\{\sigma_i - \tau\}_+) := \mathrm{diag}(\{\max\{0, \sigma_i - \tau\}\}) = \mathrm{prox}_{\tau\|\cdot\|_1}(\{\sigma_i\})$ is the shrinkage (or soft thresholding) operator applied to the singular values σ_i of Y.

For this, we will argue in 2 steps:

1. In general, the proximal operator is well-defined and returns a unique minimizer, why? Give your argument. In our case, denote $\hat{X} = \mathrm{prox}_{\tau\|\cdot\|_{\mathrm{nuc}}}(Y)$, what do we have for the optimality condition?
2. Show that $\hat{X} = U\mathrm{diag}(\{(\sigma_i - \tau)_+\})V^\top$ verifies the optimality condition, and argue that this concludes our proof.

Hint: for step 2, recall from sheet 3 that the subdifferential at point $X \in \mathbb{R}^{n\times n}$ with $s \geq 0$ zero singular values is given as

$$\partial\left\|\cdot\right\|_{\mathrm{nuc}}(X) = \left\{U_1V_1^\top + U_2MV_2^\top : M \in \mathbb{R}^{s\times s}, \left\|M\right\|_{\mathrm{spec}} \leq 1\right\}, \tag{3}$$

where $\left\|\cdot\right\|_{\mathrm{spec}}$ denotes the spectral norm, i.e., the largest singular value.

Rewriting the expressions of X and Y with an appropriately defined decomposition $V = [V_1\ V_2]$, $U = [U_1\ U_2]$ can be helpful.

Solution.

1. Let $Y \in \mathbb{R}^{n\times n}$. We are interested in the solution of

   $$\mathrm{argmin}_X \frac{1}{2}\|X - Y\|_F^2 + \tau\|X\|_{\mathrm{nuc}}.$$

   whose solution is unique since the above problem is strictly convex. The optimality condition of the problem is given as

   $$0 \in \hat{X} - Y + \tau\partial\|\cdot\|_{\mathrm{nuc}}(\hat{X}). \tag{4}$$

   where $\partial\|\cdot\|_{\mathrm{nuc}}(X)$ is the subdifferential of the nuclear norm at X characterized on exercise sheet 3.

2. Our aim is to show that $\hat{X} := U\mathrm{diag}(\{(\sigma_i - \tau)_+\})V^\top$ meets the optimality condition. To this end we decompose $V = [V_1\ V_2]$, $U = [U_1\ U_2]$ and $\Sigma = \begin{bmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{bmatrix}$ so that

   $$Y = U_1\Sigma_1V_1^\top + U_2\Sigma_2V_2^\top,$$

where Σ_1 contains all singular values $\sigma_i > \tau$ and Σ_2 all singular values $\sigma_i \leq \tau$. We may then write $\hat{X}$ as

$$\hat{X} = U\mathrm{diag}(\{(\sigma_i - \tau)_+\})V^\top = U_1 \underbrace{(\Sigma_1 - \tau I)}_{\sigma_i > 0} V_1^\top + U_2 \underbrace{\mathrm{diag}(\{0\})}_{\sigma_i = 0} V_2^\top.$$

We will now show that $\hat{X}$ meets (4): $Y - \hat{X}$ is given as

$$Y - \hat{X} = \tau(U_1 V_1^\top + U_2 \tfrac{1}{\tau}\Sigma_2 V_2^\top).$$

By construction $\|\frac{1}{\tau}\Sigma_2\|_{\mathrm{spec}} \leq 1$. And therefore and due to sheet 3

$$Y - \hat{X} \in \tau \partial \| \cdot \|_{\mathrm{nuc}}(\hat{X})$$