Convex Optimization for Machine Learning and Computer Vision

Lecture: T. Wu
Exercises: E. Laude, T. Möllenhoff
Summer Semester 2017

Computer Vision Group
Institut für Informatik
Technische Universität München

Weekly Exercises 3

Room: 02.09.023
Monday, 22.05.2017, 12:15-14:00
Submission deadline: Wednesday, 17.05.2017, Room 02.09.023

Gradient and Subdifferential (12 Points + 4 Bonus)

Exercise 1 (4 Points)**.** Let $X \subset \mathbb{R}^n$ open and convex and let $f : X \to \mathbb{R}$ be twice continuously differentiable. Prove the equivalence of the following statements:

- f is convex.
- For all $x \in X$ the Hessian $\nabla^2 f(x)$ is positive semidefinite ($\forall\, v \in \mathbb{R}^n : v^\top \nabla^2 f(x) v \geq 0$).

Hints: You can use that for $x, y \in X$ it holds that f is convex iff

$$(y-x)^\top \nabla f(x) \leq f(y) - f(x).$$

Further recall that there are two variants of the Taylor expansion:

$$f(x+tv) = f(x) + tv^\top \nabla f(x) + \frac{t^2}{2} v^\top \nabla^2 f(x) v + o(t^2)$$

with $\lim_{t \to 0} \frac{o(t^2)}{t^2} = 0$ and

$$f(x+v) = f(x) + v^\top \nabla f(x) + \frac{1}{2} v^\top \nabla^2 f(x+tv) v$$

for appropriate $t \in (0,1)$.

Solution. Let f be convex, $x \in X$ and $v \in \mathbb{R}^n$. Since X is open there exists $\tau > 0$ s.t. for all $t \in (0, \tau]$ we have that $x + tv \in X$. Using the Taylor expansion given in the hint we obtain

$$0 \overset{\text{Hint}}{\leq} f(x+tv) - f(x) - tv^\top \nabla f(x) = \frac{t^2}{2} v^\top \nabla^2 f(x) v + o(t^2)$$

Multiplying both sides with $\frac{2}{t^2}$ yields

$$0 \leq v^\top \nabla^2 f(x) v + 2 \underbrace{\frac{o(t^2)}{t^2}}_{\to 0}.$$

Let conversely $\nabla^2 f(z)$ be positive semidefinite for all $z \in X$ and let $x, y \in X$. Using the Taylor expansion we have

$$f(y) = f(x+(y-x)) = f(x)+(y-x)^\top \nabla f(x)+\frac{1}{2}\underbrace{(y-x)^\top \nabla^2 f(x+t(y-x))(y-x)}_{\geq 0 \text{ by assumption.}}$$

and therefore

$$f(y) - f(x) \geq (y-x)^\top \nabla f(x),$$

which means that f is convex.

Exercise 2 (2 Points)**.** Let $X \subset \mathbb{R}^n$ open and convex, $A \in \mathbb{R}^{n \times n}$ positive semidefinite, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$. Show that that the quadratic form $f : X \to \mathbb{R}$ defined as

$$f(x) := \frac{1}{2} x^\top A x + b^\top x + c,$$

is convex.

Solution. To show that f is convex it suffices to show that the Hessian $\nabla^2 f(x)$ is positive semidefinite, since f is twice continuously differentiable. We start rewriting $f(x)$ in terms of finite sums:

$$\begin{aligned} f(x) &= \frac{1}{2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} a_{ij} x_j + \sum_{i=1}^{n} x_i b_i + c \\ &= \frac{1}{2}\sum_{i=1}^{n} x_i \sum_{\substack{j=1, \\ j \neq i}}^{n} a_{ij} x_j + \frac{1}{2}\sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} x_i b_i + c \end{aligned}$$

We now proceed computing the first and second order partial derivatives:

$$\begin{aligned} \frac{\partial f(x)}{\partial x_k} &= \frac{1}{2}\sum_{\substack{j=1, \\ j \neq k}} a_{kj} x_j + \frac{1}{2}\sum_{\substack{i=1, \\ i \neq k}} a_{ik} x_i + a_{kk} x_k + b_k \\ &= \frac{1}{2}\sum_{j=1} a_{kj} x_j + \frac{1}{2}\sum_{i=1} a_{ik} x_i + b_k \end{aligned}$$

Then we have for the gradient of f:

$$\nabla f(x) = \frac{1}{2}(A + A^\top)x + b.$$

The second order derivatives are given as:

$$\frac{\partial^2 f(x)}{\partial x_k^2} = \frac{1}{2} a_{kk} + \frac{1}{2} a_{kk} = a_{kk},$$

and

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{1}{2} a_{kl} + \frac{1}{2} a_{lk}.$$

The Hessian is then given as

$$\nabla^2 f(x) = \frac{1}{2}(A + A^\top).$$

Since A is positive semidefinite also the Hessian $\nabla^2 f(x)$ is positive semidefinite:

$$v^\top \frac{1}{2}(A + A^\top)v = v^\top A v \geq 0.$$

Exercise 3 (2 Points)**.** Let $\mathbb{E}$ be an Euclidean space, with norm $\|\cdot\|$. Show that the subdifferential at zero is given by

$$\partial \|\cdot\| (0) = \{y \in \mathbb{E} : \|y\|_* \leq 1\},$$

where $\|\cdot\|_*$ denotes the dual norm given by

$$\|y\|_* = \sup_{\|x\| \leq 1} \langle y, x \rangle.$$

Solution.

$$\begin{aligned} p \in \partial \|\cdot\| (0) &\Leftrightarrow \langle p, y \rangle \leq \|y\|, \forall y \in \mathbb{E} \\ &\Leftrightarrow \frac{\langle p, y \rangle}{\|y\|} \leq 1, \forall y \neq 0 \\ &\Leftrightarrow \sup_{y \neq 0} \frac{\langle p, y \rangle}{\|y\|} \leq 1. \\ &\Leftrightarrow \sup_{\|y\|=1} \langle p, y \rangle \leq 1 \Leftrightarrow \|p\|_* \leq 1. \end{aligned}$$

Exercise 4 (4 Points)**.** Compute the subdifferential of the following functions:

- $f : \mathbb{R}^n \to \mathbb{R}, f(x) = \|x\|_1$.
- $f : \mathbb{R}^n \to \mathbb{R}, f(x) = \|x\|_\infty$.
- $f : \mathbb{R}^{n \times n} \to \mathbb{R}, f(X) = \sum_{i=1}^n \left(\sum_{j=1}^n (X_{i,j})^2\right)^{1/2}$.
- $f : \mathbb{E} \to \overline{\mathbb{R}}, f(x) = \delta_C(x)$ for a closed convex set $C \subset \mathbb{E}$.

Solution. First, we show that in general it holds that

$$\begin{aligned} \partial \|\cdot\| (x) &= \{p \in \mathbb{E} : \langle p, x \rangle = \|x\|, \|p\|_* \leq 1\} \\ &\overset{\text{w.t.s.}}{=} \{p \in \mathbb{E} : \|x\| + \langle p, y - x \rangle \leq \|y\|, \forall y \in \mathbb{E}\}. \end{aligned} \tag{1}$$

for a norm $\|\cdot\|$ on an Euclidean space $\mathbb{E}$. Note that if $x = 0$ we recover the result from the previous exercise. For that, we need a generalized Cauchy-Schwarz inequality:

$$\langle x, y \rangle = \|x\| \langle \frac{x}{\|x\|}, y \rangle \leq \|x\| \cdot \sup_{\|z\| \leq 1} \langle z, y \rangle = \|x\| \|y\|_*, \ \forall x, y \in \mathbb{E}. \tag{2}$$

Now take $p \in \mathbb{E}$ with $\langle p, x \rangle = \|x\|$, $\|p\|_* \leq 1$. Then we have

$$\langle p, y - x \rangle + \|x\| = \langle p, y \rangle - \langle p, x \rangle + \|x\| = \langle p, y \rangle \leq \|y\| \, \|p\|_* \leq \|y\| \, , \forall y \in \mathbb{E}.$$

Hence $p \in \partial \|\cdot\| (x)$. Conversely take $p \in \partial \|\cdot\| (x)$. Then we have

$$\begin{aligned} & \langle p, y - x \rangle + \|x\| \leq \|y\| \, , \forall y \in \mathbb{E} \\ \Leftrightarrow \quad & \|x\| - \langle p, x \rangle + \sup_y \ \langle p, y \rangle - \|y\| \leq 0 \end{aligned} \tag{3}$$

The supremum evaluates as

$$\sup_y \ \langle p, y \rangle - \|y\| = \begin{cases} 0, & \|p\|_* \leq 1 \\ \infty, & \text{otherwise.} \end{cases} .$$

We show this as the following. Assume $\|p\|_* > 1$. Hence there is some vector $z \in \mathbb{E}$, $\|z\| \leq 1$ and $\langle p, z \rangle > 1$. It can be seen that the above supremum is unbounded, i.e. take some $y = tz$, $t(\langle p, z \rangle - \|z\|) \to \infty$ for $t \to \infty$. Now take $\|p\|_* \leq 1$, then we have $\langle p, y \rangle - \|y\| \leq \|y\| \, (\|p\|_* - 1) \leq 0$, where equality holds for $y = 0$.

Furthermore, we have

$$0 \geq -\langle p, x \rangle + \|x\| \geq - \|x\| \, \|p\|_* + \|x\| = \|x\| \, (1 - \|p\|_*) \geq 0$$

Hence $-\langle p, x \rangle + \|x\| = 0$ which implies $\|x\| = \langle p, x \rangle$.

- The dual norm of $\|\cdot\|_1$ is clearly $\|\cdot\|_\infty$ and vice versa. Hence,

$$\begin{aligned} \partial \|\cdot\|_1 (x) &= \{p \in \mathbb{R}^n : \|p\|_\infty \leq 1, \langle p, x \rangle = \|x\|_1\}, \\ &= \left\{ p \in \mathbb{R}^n : \begin{cases} p_i \in [-1, 1], & \text{if } \ x_i = 0 \\ p_i = \operatorname{sign}(x_i), & \text{otherwise.} \end{cases} \right\}. \end{aligned} \tag{4}$$

$$\partial \|\cdot\|_\infty (x) = \{p \in \mathbb{R}^n : \|p\|_1 \leq 1, \langle p, x \rangle = \|x\|_\infty\}. \tag{5}$$

- It can be easily verified that $f(X) = \sum_{i=1}^n \left(\sum_{j=1}^n (X_{i,j})^2 \right)^{1/2} =: \|X\|_{2,1}$ is a norm on $\mathbb{R}^{n \times n}$. The dual norm of $\|X\|_{2,1}$ is $\|X\|_{2,\infty} = \max_{1 \leq i \leq n} \|X_i\|_2$, where $X_i \in \mathbb{R}^n$ denotes the i-th row of X. Hence we have

$$\partial \|\cdot\|_{2,1} (X) = \{P \in \mathbb{R}^{n \times n} : \|P\|_{2,\infty} \leq 1, \langle P, X \rangle = \|X\|_{2,1}\}, \tag{6}$$

- Take a point $x \in \operatorname{dom} f$. Then the subgradients $g \in \partial f(x)$ fulfill

$$\langle g, y - x \rangle \leq 0, \forall y \in C \ \Leftrightarrow \ g \in N_C(x).$$

Hence $\partial f(x) = N_c(x)$.

Exercise 5 (4 Points)**.** Consider the nuclear norm $\|\cdot\|_{\text{nuc}} : \mathbb{R}^{n\times n} \to \mathbb{R}$ given by

$$\|X\|_{\text{nuc}} = \sum_{i=1}^{n} |\sigma_i(X)| = \|\sigma(X)\|_1 ,$$

where $\sigma_i(X) \in \mathbb{R}$ is the i-th singular value of $X \in \mathbb{R}^{n\times n}$. Show that the subdifferential at point $X \in \mathbb{R}^{n\times n}$ with $s \geq 0$ zero singular values is given as

$$\partial \|\cdot\|_{\text{nuc}} (X) = \left\{U_1 V_1^\top + U_2 M V_2^\top : M \in \mathbb{R}^{s\times s}, \|M\|_{\text{spec}} \leq 1\right\}, \tag{7}$$

where $U = \begin{bmatrix}U_1 & U_2\end{bmatrix}$ and $V = \begin{bmatrix}V_1 & V_2\end{bmatrix}$ are given by the singular value decomposition of $X = U\Sigma V^\top$, with U_1 and V_1 having $n - s$ columns. Furthermore $\|\cdot\|_{\text{spec}}$ denotes the spectral norm, i.e., the largest singular value.

Solution. Denote by $\langle X, Y\rangle = \text{tr}(X^T Y)$. First we show that the dual norm of the nuclear norm is the spectral norm, i.e.,

$$\sup_{\sum_i \sigma_i(Y)\leq 1} \langle X, Y\rangle = \sigma_1(X).$$

Clearly, $\sup_{\sum_i \sigma_i(Y)\leq 1} \langle X, Y\rangle \geq \sigma_1(X)$ since the supremum is bigger than the function at the feasible candidate $Y = u_1 v_1^T$ (for $X = U\Sigma V^T$) for which the supremum evaluates to $\langle u_1 v_1^T, U\Sigma V^T\rangle = \sigma_1(X)$. The other inequality (again with $X = U\Sigma V^T$) follows from von Neumann's trace inequality $\text{tr}(AB) \leq \sum_{i=1}^n \sigma_i(A)\sigma_i(B)$.

$$\sup_{\sum_i \sigma_i(Y)\leq 1} \langle Y, X\rangle = \sup_{\sum_i \sigma_i(Y)\leq 1} \text{tr}(Y^T X) \leq \sup_{\sum_i \sigma_i(Y)\leq 1} \sum_{i=1}^{n} \sigma_i(X)\sigma_i(Y) = \sigma_1(X). \tag{8}$$

Hence, from the previous solution, it then follows that

$$\partial \|X\|_{\text{nuc}} = \{Y \in \mathbb{R}^{n\times n} \ : \ \langle X, Y\rangle = \|X\|_{\text{nuc}} , \|Y\|_{\text{spec}} \leq 1\}. \tag{9}$$

We finish the proof by showing that (7) and (9) are the same. Denote by $X = U_1\Sigma V_1^T$ denote the compact SVD of X.

First we take some Y that satisfies (9), i.e., $\langle X, Y\rangle = \|X\|_{\text{nuc}}$ and $\|Y\|_{\text{spec}} \leq 1$ and show it is in (7). For that, consider the subspace $S = \{U_1 W V_1^T : W \in \mathbb{R}^{r\times r}\}$ where $r = n - s$ and its orthogonal complement $S^\perp = \{U_2 M V_2^T : M \in \mathbb{R}^{s\times s}\}$. Then we can write $Y = \Pi_S(Y) + \Pi_{S^\perp}(Y) = U_1 W V_1^T + U_2 M V_2^T$ for some W and M.

Since we have

$$\begin{aligned}\langle Y, X\rangle &= \langle U_1 W V_1^T + U_2 M V_2^T, U_1\Sigma V_1^T\rangle = \langle U_1 W V_1^T, U_1 \Sigma V_1^T\rangle \\ &= \text{tr}(V_1^T W^T U_1^T U \Sigma V_1) = \text{tr}(W^T\Sigma) \overset{\text{assumption}}{=} \text{tr}(\Sigma)\end{aligned} \tag{10}$$

we can conclude that $W = I$ and hence $Y = U_1 V_1^T + U_2 M V_2^T$. Since projections always have Lipschitz constant less or equal one we have that

$$\|M\|_{\text{spec}} = \left\|U_2 M V_2^T\right\|_{\text{spec}} = \|\Pi_{S^\perp}(Y)\|_{\text{spec}} \leq \|Y\|_{\text{spec}} \overset{\text{assumption}}{\leq} 1,$$

where we used the unitary invariance of the spectral norm in the first equality.

Conversely take some $U_1V_1^T + U_2MV_2^T$ from (7) with $\|M\|_{\text{spec}} \le 1$ and $X = U_1\Sigma V_1^T$. We show that it satisfies (9):

$$\langle U_1V_1^T + U_2MV_2^T, U_1\Sigma V_1^T\rangle = \mathrm{tr}(V_1U_1^TU\Sigma V_1^T) = \mathrm{tr}(\Sigma) = \|X\|_{\text{nuc}}.$$

For the spectral norm we use the fact that if $\|Ax\|^2 \le \|x\|^2$, then $\|A\|_{\text{spec}} \le 1$.

$$\begin{aligned}
\left\|(U_1V_1^T + U_2MV_2^T)x\right\|^2 &= \langle U_1V_1^Tx + U_2MV_2^Tx, U_1V_1^Tx + U_2MV_2^Tx\rangle \\
&= \langle x, (U_1V_1^T + U_2MV_2^T)^T(U_1V_1^T + U_2MV_2^T)x\rangle \\
&= \langle x, (V_1U_1^TU_1V_1^Tx\rangle + \langle x, V_2M^TU_2^TU_2MV_2^Tx\rangle \\
&\underbrace{+\langle x, V_1U_1^TU_2MV_2^Tx\rangle + \langle x, V_2M^TU_2^TU_1V_1^Tx\rangle}_{=0} \\
&= \langle V_1^Tx, (V_1^Tx\rangle + \langle MV_2^Tx, MV_2^Tx\rangle \\
&= \left\|V_1^Tx_1\right\|^2 + \left\|MV_2^Tx_2\right\| \\
&\overset{\text{assumption}}{\le} \|x_1\|^2 + \|x_2\|^2 = \|x\|^2,
\end{aligned} \tag{11}$$

where we decomposed $x = x_1 + x_2$ onto the subspace spanned by V_2^T and its orthogonal complement in the second to last step.

1 Image Cartooning

Finish the programming exercise from the second exercise sheet.