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 7

Room: 02.09.023
Monday, 26.06.2017, 12:15-14:00
Submission deadline: Wednesday, 21.06.2017, Room 02.09.023

Majorization minimization and Convex Analysis Revisited (6 + 4 Points)

Exercise 1 (2 Points)**.** Consider the smooth approximation of the absolute value function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \sqrt{x^2 + \varepsilon}$ for some $\varepsilon > 0$. Show that

$$\widehat{f}(x; x_k) = f(x_k) + \frac{1}{2f(x_k)} \left[x^2 - x_k^2 \right],$$

is a *majorizing surrogate* at $x_k \in \mathbb{R}$, i.e., prove that

- $\widehat{f}(x_k; x_k) = f(x_k)$,
- $\widehat{f}(x; x_k) \geq f(x)$, $\forall x \in \mathbb{R}$.

Solution. The first part is trivial, since for $x = x_k$ we immediately have $\widehat{f}(x; x_k) = f(x_k)$ by definition.

For the second part, we linearize the concave function $\sqrt{1+t}$ at some point $t_0 \in \mathbb{R}_{\geq 0}$ to get the estimate:

$$\sqrt{t + \varepsilon} \leq \sqrt{t_0 + \varepsilon} + \frac{1}{2\sqrt{t_0 + \varepsilon}}(t - t_0).$$

Using this estimate with $t = x^2$ and $t_0 = x_k^2$ yields the desired inequality

$$f(x) = \sqrt{x^2 + \varepsilon} \leq \sqrt{x_k^2 + \varepsilon} + \frac{1}{2\sqrt{x_k^2 + \varepsilon}}(x^2 - x_k^2) = \widehat{f}(x; x_k).$$

Exercise 2 (2 Points)**.** Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$ and let $\|\cdot\|_*$ denote its dual norm. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and twice continuously differentiable. Let $x \in \mathbb{R}^n$. Let

$$\Delta x_n := \text{argmin}_{v : \|v\| = 1} \nabla f(x)^\top v = \text{argmin}_{v : \|v\| \leq 1} \nabla f(x)^\top v,$$

and

$$\Delta x = \|\nabla f(x)\|_* \Delta x_n,$$

be the normalized and unnormalized steepest descent directions at x. Prove the following identities.

- $\nabla f(x)^\top \Delta x_n = -\|\nabla f(x)\|_*$
- $\nabla f(x)^\top \Delta x = -\|\nabla f(x)\|_*^2$
- $\Delta x = \operatorname{argmin}_v \nabla f(x)^\top v + \frac{1}{2}\|v\|^2$

Solution. For the first part we observe, that Δx_n is determined via a convex conjugate of a unit ball of a norm:

$$\min_{v:\|v\|\leq 1} \nabla f(x)^\top v = -\max_v -\nabla f(x)^\top v - \delta\{\|v\| \leq 1\} = -\|\cdot\|_*(-\nabla f(x)).$$

Since Δx_n minimizes the above equation we obtain the desired result.

For the second part we may use the above result and observe that,

$$\nabla f(x)^\top \Delta x = \nabla f(x)^\top(\|\nabla f(x)\|_* \Delta x_n) = \nabla f(x)^\top \Delta x_n \|\nabla f(x)\|_* = -\|\nabla f(x)\|_*^2.$$

For the last part we first observe that the convex conjugate of the squared norm is the squared dual norm $(\frac{1}{2}\|\cdot\|^2)^* = (\frac{1}{2}\|\cdot\|_*^2)$: To this end let $x \in \mathbb{R}^n$ and observe that

$$y^\top x - \frac{1}{2}\|x\|^2 \leq \|x\|\|y\|_* - \frac{1}{2}\|x\|^2 \leq \frac{1}{2}\|y\|_*^2,$$

since the above is concave and quadratic in $\|x\|$ and it is maximized for $\|x\| = \|y\|_*$. On the other hand, observe that for x chosen such that $\|x\| = \|y\|_*$ and $y^\top x = \|y\|_*\|x\|$ (note that such an x exists) we have

$$y^\top x - \frac{1}{2}\|x\|^2 \geq \frac{1}{2}\|y\|_*^2.$$

This yields the desired result. Again, we may understand the minimization problem in terms of a convex conjugate:

$$\begin{aligned}\min_v \nabla f(x)^\top v + \frac{1}{2}\|v\|^2 &= -\max_v -\nabla f(x)^\top v - \frac{1}{2}\|v\|^2 \\ &= -\frac{1}{2}\|\cdot\|_*^2(-\nabla f(x)) \\ &= \frac{1}{2}\nabla f(x)^\top \Delta x.\end{aligned}$$

Since

$$\nabla f(x)^\top \Delta x + \frac{1}{2}\|\nabla f(x)\|_*^2 \underbrace{\|\Delta x_n\|^2}_{=1} = \frac{1}{2}\nabla f(x)^\top \Delta x,$$

Δx minimizes the above expression.

Exercise 3 (2 Points)**.** Steepest descent method in ℓ_∞-norm. Explain how to find a steepest descent direction in the ℓ_∞-norm, and give a simple interpretation.

Solution. Since $\|x\|_\infty = \max_i |x_i|$, clearly, the minimum of the objective

$$\Delta x_n := \operatorname{argmin}_{v:\|v\|_\infty \le 1} \nabla f(x)^\top v,$$

is attained at a vertex of the ℓ_∞-norm unit ball (we minimize a linear cost function over a polytope), i.e. Δx_n has the form,

$$(\Delta x_n)_i = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

for some j. Here, j is picked as the index corresponding to the entry of $(\nabla f(x))_j$ which has minimal value. Since the descent direction is zero for all but one component, this means, that our descent method decreases the objective function coordinate-wise: All components of the current iterate x^k are kept fixed except for one that is decreased.

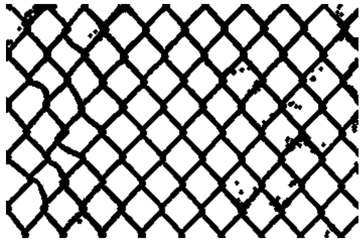
Exercise 4 (4 Points)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Let conv f be the largest convex function majorized by f, meaning that $(\text{conv } f)(x) \le f(x)$ for all $x \in \mathbb{R}^n$. Show the following identity

$$(\text{conv } f)(x) = \inf \left\{ \sum_{i=1}^{n+1} \lambda_i f(x_i) : \sum_{i=1}^{n+1} \lambda_i x_i = x, \lambda_i \ge 0, \sum_{i=1}^{n+1} \lambda_i = 1 \right\}.$$

Hint: Apply Caratheodory's Theorem to the epigraph of f.

Solution. We apply Caratheodory's Theorem to epi $f \subset \mathbb{R}^{n+1}$: every point of conv epi f is a convex combination of at most $n+2$ points of epi f. Actually, at most $n+1$ points x_i at a time are needed in determining the values of conv f, since a point $(\bar{x}, \bar{\alpha})$ of conv epi f not representable by fewer than $n+2$ points of epi f would lie in the interior of some $n+1$-simplex S in conv epi f. The vertical line through $(\bar{x}, \bar{\alpha})$ would meet the boundary of S in a point $(\bar{x}, \tilde{\alpha})$ with $\bar{\alpha} > \tilde{\alpha}$, and such a boundary point is representable by $n+1$ of the points in question. Thus, $(\text{conv } f)(x)$ is the infimum of all numbers α such that there exist $n+1$ points (x_i, α_i) and scalars $\lambda_i \ge 0$, with $\sum_{i=1}^{n+1} \lambda_i (x_i, \alpha_i)$, $\sum_{i=1}^{n+1} \lambda_i = 1$. This description translates to the formula claimed.

Input image



Inpainting mask

Minimizer of (1)

Programming: Image inpainting (8 Points)

Exercise 5. In this exercise the goal is to fill regions specified by a mask in an image by through minimizing the energy

$$E(u) = \frac{\lambda}{2} \left\| M(u - f) \right\|^2 + \left\| Du \right\|_\varepsilon . \tag{1}$$

Here, $f \in \mathbb{R}^{n_x \cdot n_y \cdot n_c}$ denotes the input image, $M : \mathbb{R}^{n_x \cdot n_y \cdot n_c} \to \mathbb{R}^{n_x \cdot n_y \cdot n_c}$ denotes a diagonal matrix consisting of zero/one values specifying the inpainting mask and $D : \mathbb{R}^{n_x \cdot n_y \cdot n_c} \to \mathbb{R}^{2 \cdot n_x \cdot n_y \cdot n_c}$ is the usual discrete gradient operator from the first exercise sheet. $\lambda > 0$ is a data fidelity parameter determining the smoothness of the solution. As in the second programming exercise, $\left\| x \right\|_\varepsilon = \sum_i \sqrt{x^2 + \varepsilon}$ denotes the smoothed ℓ_1 norm. Your tasks are the following:

1. Find a minimizer of (1) using gradient descent.
2. Minimize (1) using a majorization minimization approach. For this, use the result from exercise 1 to majorize the term $\left\| \cdot \right\|_\varepsilon$ at the current solution u^k with a quadratic function. Since the upper bound is quadratic, it can be efficiently minimized by solving a linear system. For that, use the backslash operator in MATLAB.
3. Compare the gradient descent approach to the MM scheme. Which one converges faster?

The deadline for handing in the programming solution is **June 28th, 2017, 23:59pm**.