Homework 5: Optimization

CS 205A: Mathematical Methods for Robotics, Vision, and Graphics (Fall 2013), Stanford University

Due Monday, November 11, at midnight

Problem 1 (20 points). *A few warm-up problems to get you thinking about conjugate gradients:*

(a) *Suppose A is symmetric and positive definite. Show that the minimizer $\vec{x}^*$ of $f(\vec{x}) \equiv \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top \vec{x} + c$ satisfies $\vec{x}^* = A^{-1}\vec{b}$. Be sure to argue that $\vec{x}^*$ is a minimizer instead of a critical point.*

(b) *Suppose $\vec{v}_1, \ldots, \vec{v}_k$ are linearly independent in $\mathbb{R}^n$. We will say two vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ are A-conjugate when $\vec{x}^\top A\vec{y} = 0$. Give a variant of the Gram-Schmidt algorithm for finding a set of A-conjugate vectors $\vec{w}_1, \ldots, \vec{w}_n$ with span $\{\vec{w}_1, \ldots, \vec{w}_n\}$ = span $\{\vec{v}_1, \ldots \vec{v}_n\}$.*

Problem 2 (20 points). *In this problem we will study the* linear programming *problem.*

(a) *A linear program in "standard form" is given by:*

$$\begin{array}{rl} \textit{minimize}_{\vec{x}} & \vec{c}^\top \vec{x} \\ \textit{such that} & A\vec{x} = \vec{b} \\ & \vec{x} \geq \vec{0} \end{array}$$

Here, the optimization is over $\vec{x} \in \mathbb{R}^n$; the remaining variables are constants $A \in \mathbb{R}^{m \times n}$, $\vec{b} \in \mathbb{R}^m$, and $\vec{c} \in \mathbb{R}^n$. Find the KKT conditions of this system.

(b) *Suppose we add a constraint of the form $\vec{v}^\top \vec{x} \leq d$ for some fixed $\vec{v} \in \mathbb{R}^n$ and $d \in \mathbb{R}$. Explain how such a constraint can be added while keeping a linear program in standard form.*

(c) *The "dual" of this linear program is another optimization:*

$$\begin{array}{rl} \textit{maximize}_{\vec{y}} & \vec{b}^\top \vec{y} \\ \textit{such that} & A^\top \vec{y} \leq \vec{c} \end{array}$$

Assuming that the primal and dual have exactly one stationary point, show that the optimal value of the primal and dual objectives coincide.
Hint: *Show that the KKT multipliers of one problem can be used to solve the other.*
Note: *This property is called "strict duality." The famous simplex algorithm for solving linear programs maintains estimates of $\vec{x}$ and $\vec{y}$, terminating when $\vec{c}^\top \vec{x}^* - \vec{b}^\top \vec{y}^* = 0$.*

Problem 3 (20 points). *Here we examine some changes to the gradient descent algorithm described in class for unconstrained optimization on a function f.*

(a) *In machine learning, the* stochastic gradient descent *algorithm can be used to optimize many common objective functions:*

(i) *Give an example of a practical optimization problem (from lecture or of your own design) with an objective taking the form $f(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} g(\vec{x}_i - \vec{x})$ for some function $g : \mathbb{R}^n \to \mathbb{R}$.*

(ii) *Propose a randomized approximation of ∇f summing no more than k terms (for some $k \ll N$) assuming the $\vec{x}_i$'s are similar to one another. Discuss advantages and drawbacks of using such an approximation.*

(b) *The "line search" part of gradient descent must be considered carefully:*

(i) *Suppose an iterative optimization routine gives a sequence of estimates $\vec{x}_1, \vec{x}_2, \ldots$ of the position $\vec{x}^*$ of the minimum of f. Is it enough to assume $f(\vec{x}_k) < f(\vec{x}_{k-1})$ to guarantee that the $\vec{x}_k$'s converge to a local minimum? Why?*

(ii) *Suppose we run gradient descent. Using the notation from §8.4.1 of the lecture notes, if we suppose $f(\vec{x}) \geq 0$ for all $\vec{x}$ and that we are able to find t^* exactly in each iteration, show that $f(\vec{x}_k)$ converges as $k \to \infty$.*

EC. *Explain how the optimization in (ii) for t^* can be overkill. In particular, explain how the Wolfe conditions (you will have to look these up!) relax the assumption that we can find t^*.*

Problem 4 (20 points)**.** *Sometimes we are greedy and wish to optimize multiple objectives simultaneously. For example, we might want to fire a rocket to reach an optimal point in time* and *space. It may not be possible to do this simultaneously, but some theories attempt to reconcile multiple optimization objectives.*

Suppose we are given functions $f_1(\vec{x}), f_2(\vec{x}), \ldots, f_k(\vec{x})$. A point $\vec{x}$ is said to Pareto dominate *another point $\vec{y}$ if $f_i(\vec{x}) \leq f_i(\vec{y})$ for all i and $f_j(\vec{x}) < f_j(\vec{y})$ for some $j \in \{1, \ldots, k\}$. A point $\vec{x}^*$ is* Pareto optimal *if it is not dominated by any point $\vec{y}$. Assume $f_1, \ldots, f_k$ are convex and in particular have unique minimizers.*

(a) *Show that the set of Pareto optimal points is nonempty in this case.*

(b) *Suppose $\sum_i \gamma_i = 1$ and $\gamma_i > 0$ for all i. Show that the minimizer $\vec{x}^*$ of $g(\vec{x}) \equiv \sum_i \gamma_i f_i(\vec{x})$ is Pareto optimal.*
Note: *One strategy for multi-objective optimization is to promote $\vec{\gamma}$ to a variable with constraints $\vec{\gamma} \geq \vec{0}$ and $\sum_i \gamma_i = 1$.*

(c) *Suppose $\vec{x}_i^*$ minimizes $f_i(\vec{x})$ over all possible $\vec{x}$. Write vector $\vec{z} \in \mathbb{R}^k$ with components $z_i = f_i(\vec{x}_i^*)$. Show that the minimizer $\vec{x}^*$ of $h(\vec{x}) \equiv \sum_i (f_i(\vec{x}) - z_i)^2$ is Pareto optimal.*
Note: *This part and the previous part represent two possible* scalarizations *of the multi-objective optimization problem by reducing to a single scalar objective.*

Problem 5 (20 points; Heath 6.8)**.** *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$f(\vec{x}) \equiv \frac{1}{2}(x_1^2 - x_2)^2 + \frac{1}{2}(1 - x_1)^2.$$

(a) *At what point does f attain a minimum?*

(b) *Perform one iteration of Newton's method for minimizing f using the starting point $\vec{x}_0 = (2, 2)$.*

(c) *In what sense is this a good step?*

(d) *In what sense is this a bad step?*