Advanced Topics in Control 2020: Large-Scale Convex Optimization

Exercise 5: Gradient methods

Goran Banjac, Mathias Hudoba de Badyn, Andrea Iannelli,
Angeliki Kamoutsi, Ilnura Usmanova

April 2, 2020

Date due: April 9, 2020 at 23:59.

Please submit your solutions via Moodle as a PDF with filename `Ex05_Surname.pdf`, replacing `Surname` with your surname.

All the functions f featured in this Homework are assumed to be convex.

1 Problem 1 - Gradient descent method

Consider the unconstrained minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

where f is continuously differentiable in $\mathbb{R}^n$, and a generic iterative algorithm to find stationary points of f:

$$x^{k+1} = x^k + t^k d^k, \tag{2}$$

where t^k is the stepsize and d^k is a descent direction, that is, its directional derivative $\nabla f(x^k)^\top d^k$ is strictly negative for every k.

(a) Assume the objective function is of the form $f(x) = \frac{1}{2} x^\top A x - b^\top x$, where $A \succ 0$, $b \in \mathbb{R}^n$. Find the expression for the stepsize $t^\star$ that minimizes f along the line $x^k + td^k$. This choice of stepsize (irrespective of the type of function to which is applied) is also known as exact line search.

(b) In the gradient (or steepest) descent method, the direction is chosen as the opposite of the gradient of f at the current point x^k. Show that $d^k = -\frac{\nabla f(x^k)}{||\nabla f(x^k)||_2}$ achieves the minimal directional derivative among all unit-norm descent directions d.

(c) Show that when using the steepest descent with exact line search for t^k, the sequence $\{x^k\}_{k \geq 0}$ generated by equation (2) is such that $x^{k+1} - x^k$ is orthogonal to $x^{k+2} - x^{k+1}$ for every k.

(d) Suppose that f is L-Lipschitz smooth. Show that at every iteration of the steepest descent method the following holds:

$$f(x^{k+1}) \leq f(x^k) - t^k \left(1 - \frac{Lt^k}{2}\right) ||\nabla f(x^k)||_2^2, \tag{3}$$

and determine the stepsize t^k which gives the largest function value decrease.

(e) Consider the case where f is given by:

$$f(x_1, x_2) = \begin{cases} \sqrt{x_1^2 + \gamma x_2^2}, & \text{if } |x_2| \leq x_1, \\ \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}}, & \text{else,} \end{cases}$$

where $\gamma > 1$. Show that, when the steepest descent algorithm with exact line search is applied starting from $\left(x_1^0, x_2^0\right) = (\gamma, 1)$, the iterates are:

$$x_1^k = \gamma \left(\frac{\gamma - 1}{\gamma + 1} \right)^k, \quad x_2^k = \left(-\frac{\gamma - 1}{\gamma + 1} \right)^k.$$

Where does the sequence generated by the gradient method converge to? Is this the minimizer?
Hint: Notice that f is not differentiable over its domain. However, if you can show that the iterates always stay in the interior of one of the regions, then the function is differentiable therein and its gradient is well defined. You can then prove the statement by induction.

2 Problem 2 - Subgradient method

Consider the optimization in Eq. (1) where f is convex but not necessarily continuously differentiable in $\mathbb{R}^n$ and a subgradient method is used to find its minimum. In its simplest form, this consists of the following iterative algorithm:

$$x^{k+1} = x^k - t^k g^k, \quad g^k \in \partial f(x^k). \tag{4}$$

Assume also that f is L-Lipschitz continuous.

(a) Show that $||g||_2 \leq \text{L}$ for any $g \in \partial f(x)$ and any x.
Hint: Use the Lipschitz property and the definition of subgradient.

(b) Show that after k iterations of algorithm (4), the following holds:

$$f_{\text{best}}^k \leq f(x) + \frac{L^2 \sum_{i=0}^k (t^i)^2 + ||x - x^0||_2^2}{2 \sum_{i=0}^k t^i}, \quad \forall x \in \mathbb{R}^n, \tag{5}$$

where $f_{\text{best}}^k = \min_{0 \leq i \leq k} f(x^i)$.
Hint: Start off by writing out $||x^{i+1} - x||_2^2$ for a generic i such that $0 \leq i \leq k$ and bound this distance with respect to $f(x^{i+1})$ and $f(x)$ using the definition of subdifferential and the property proved in part (a). Then, use the telescopic summation to find a bound on f_{best}^k.

(c) Consider the case where a constant step length is chosen, i.e. $t^k = \frac{\beta}{||g^k||_2}$, with $\beta > 0$. Show the convergence properties of the algorithm (4).

3 Problem 3 - Conjugate gradient method

Consider the same quadratic function f analyzed in part (a) of Problem 1, and recall that $x \in \mathbb{R}^n$. A generic conjugate *direction* method can be written as:

$$x^{k+1} = x^k + t^k p^k, \tag{6}$$

where $\{p^k\}_{k=0}^{n-1}$ are conjugate directions with respect to A, and t^k is the step size obtained with the exact line search. We will denote by r^k the current value of the gradient (or residual), i.e. $\nabla f(x^k) = Ax^k - b$.

(a) Show that, for any initial value x^0, the sequence $\{x^k\}_{k=1}^n$ generated by (6) is such that $x^n = x^\star$, where $x^\star$ is the minimizer of f.
Hint: Show that generating iterates with the conjugate direction method is equivalent to writing $x^\star - x^0$ as a linear combination of the conjugate directions (which indeed form a basis of $\mathbb{R}^n$).

(b) In the conjugate *gradient* method, the initial direction $p^0 = -r^0$, while for $k > 0$ it holds $p^k = -r^k + \beta^{(k)} p^{k-1}$, where β^k is chosen such that p^k are conjugate directions. Derive expressions for:

(1) t^k of the form $\frac{(z^k)^\top z^k}{(y^k)^\top y^k}$ where $z^k, y^k \in \mathbb{R}^n$;

(2) β_k of the form $\frac{(z^k)^\top z^k}{(z^{k-1})^\top z^{k-1}}$ (z^k is the same vector used for t^k).

Hint: z^k and y^k are fictitious symbols, but in the solution of the exercise they will have to be some function of the vectors featuring in the method, e.g. p^k. You will have to use the properties of this algorithm, for example that r^k is orthogonal to all previous search directions p^i $(i = 0, ..., k-1)$ and that p^k are conjugate directions. Note that you have already determined an exact line search t^k in part (a) of Problem 1 for the same quadratic function. Recall finally that the Cholesky factorization of a positive definite matrix B allows you to write it as $B = U^\top U$, where the matrix U is called Cholesky factor.

(c) Show that the conjugate gradient method starting with $x^0 \neq 0$ is equivalent to the same method applied to a function $f(y) = \frac{1}{2} y^\top A y - \hat{b}^\top y$ starting with $y^0 = 0$, where $\hat{b} = b - Ax^0$.
Hint: Write the algorithm for $f(y)$ and show that: the directions and the residuals of the two problems coincide for every k $\geq$0; as a consequence, you should also find that the iterates satisfy $y^k = x^k - x^0$.