Advanced Topics in Control 2020: Large-Scale Convex Optimization

Solution to Exercise 5

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

April 11, 2020

---

1 Problem 1 - Gradient methods

(a) This problem can be recast as the following optimization problem:

$$\min_{t\in\mathbb{R}} h(t) = f(x+td), \tag{1}$$

where $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ and we have dropped for simplicity of notation the superscript k. We can write out h in this way:

$$\begin{aligned} h(t) =& \frac{1}{2}(x+td)^\top A(x+td) - b^\top(x+td) \\ =& \frac{1}{2}(d^\top Ad)t^2 + (d^\top Ax - d^\top b)t + \frac{1}{2}x^\top Ax - b^\top x \\ =& \frac{1}{2}(d^\top Ad)t^2 + (d^\top Ax - d^\top b)t + f(x) \end{aligned}$$

To find the optimal t we can compute the derivative of h with respect to t:

$$h'(t) = (d^\top Ad)t + d^\top(Ax-b) = (d^\top Ad)t + d^\top\nabla f(x), \tag{2}$$

and observe that $h''(t) > 0$. Therefore, the minimizer is attained by setting (2) to zero.

$$t^\star = -\frac{d^\top\nabla f(x)}{(d^\top Ad)}.$$

The denominator of $t^\star$ is clearly positive, while for the numerator we can observe that, since d is a descent direction, this is negative. Thus $t^\star > 0$. The complete expression with the index k is:

$$t^k = -\frac{{d^k}^\top\nabla f(x^k)}{\left({d^k}^\top Ad^k\right)}.$$

(b) This problem can also be recast as an optimization problem. Let us define the directional derivative as the function $z(x,d)=\nabla f(x)^\top d$. Then we need to show that, for any non-stationary point x (i.e. $||\nabla f(x)||_2 \neq 0$), $d^\star = -\frac{\nabla f(x)}{||\nabla f(x)||_2}$ is the minimizer of:

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & z(x,d), \\ \text{s.t.} \quad & ||d||_2 = 1. \end{aligned} \tag{3}$$

By the Cauchy–Schwarz inequality we have

$$z(x,d) \geq -||\nabla f(x)||_2||d||_2 = -||\nabla f(x)||_2$$

Thus, $-||\nabla f(x)||_2$ is a lower bound on the optimal value of (3). We need to prove now that this value is attained for $d = \frac{\nabla f(x)}{||\nabla f(x)||_2}$. If we plug-in this value of d in $z(x,d)$, we obtain:

$$z(x, -\frac{\nabla f(x)}{||\nabla f(x)||_2}) = -\nabla f(x)^\top \left(\frac{\nabla f(x)}{||\nabla f(x)||_2}\right) = -||\nabla f(x)||_2,$$

which shows that indeed the lower bound on z is attained for $d = -\frac{\nabla f(x)}{||\nabla f(x)||_2} = d^\star$.

(c) The orthogonality property to be proven in the exercise can be concisely written as:

$$\left(x^{k+1} - x^k\right)^\top \left(x^{k+2} - x^{k+1}\right) = 0. \tag{4}$$

We first note that the iterates of the gradient methods are such that the following holds:

$$x^{k+1} - x^k = -t^k \nabla f(x^k),$$
$$x^{k+2} - x^{k+1} = -t^{k+1} \nabla f(x^{k+1}),$$

Therefore, Eq. (4) holds if and only is:

$$\nabla f(x^k)^\top \nabla f(x^{k+1}) = 0. \tag{5}$$

To show this, we use the fact that t^k is an exact line search. As analysed in part (a), this means that, for any k, t^k is a solution of the following problem (now specialized to the gradient descent method):

$$\min_{t \in \mathbb{R}} f(x^k - t\nabla f(x^k)),$$

By setting the derivative of the objective function to zero, we finally obtain:

$$-\nabla f(x^k)^\top \nabla f(x^k - t^k \nabla f(x^k)) = \nabla f(x^k)^\top \nabla f(x^{k+1}),$$

which proves (5) and thus the statement.

(d) To show the statement, we start from the following inequality which holds for L-Lipschitz smooth functions f:

$$f(x^{k+1}) - f(x^k) \leq \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}||x^{k+1} - x^k||_2^2.$$

By using this result and the gradient method iteration, we get:

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}||x^{k+1} - x^k||_2^2, \\ &= -t^k \nabla f(x^k)^\top \nabla f(x^k) + \frac{L(t^k)^2}{2}||\nabla f(x^k)||_2^2, \\ &= -t^k \left(1 - \frac{Lt^k}{2}\right) ||\nabla f(x^k)||_2^2, \end{aligned}$$

The function value decrease is upper bounded by:

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

Thus, we need to minimize $h(t^k) = -t^k \left(1 - \frac{Lt^k}{2}\right)$. Since h is differentiable and convex in t^k, the minimum is found at zero gradient: $\nabla h = -1 + Lt^k = 0$. Thus, the optimal value is $t^\star = \frac{1}{L}$.

(e) It is easy to see (since $\gamma > 1$ and $\gamma - 1 < \gamma + 1$) that $|x_2| \leq x_1$, thus the iterates are always in the interior of the region where $f(x_1, x_2) = \sqrt{x_1^2 + \gamma x_2^2}$. In this region, the function is differentiable and it holds:

$$\nabla f(x_1, x_2) = \frac{1}{\sqrt{x_1^2 + \gamma x_2^2}} \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix}$$

thus the direction of the gradient is $(x_1, \gamma x_2)$. The coefficient $\frac{1}{\sqrt{x_1^2+\gamma x_2^2}}$ does not matters in the remainder since it will be included in the stepsize (which is chosen to do an exact line search). We now verify that the iterates of the gradient descent with exact line search are given by:

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

The value of these iterates matches with the starting point (at $k = 0$), i.e. $\left(x_1^0, x_2^0\right) = (\gamma, 1)$. As for $k > 0$, we proceed by induction and note first (using the previously found direction of the gradient) that the exact line search minimizes f along the line:

$$\begin{bmatrix} x_1^k - t^k x_1^k \\ x_2^k - \gamma x_2^k \end{bmatrix} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k \begin{bmatrix} (1 - t^k)\gamma \\ (1 - \gamma t^k)(-1)^k \end{bmatrix} \tag{6}$$

If we find the value of t^k corresponding to the exact line search, we can substitute it back in (6) and solve for the iterates. To find t^k, we compute f along the line:

$$f\left(x_1^k(1 - t^k), x_2^k(1 - \gamma t^k)\right) = \sqrt{\gamma^2(1 - t^k)^2 + \gamma(1 - \gamma t^k)^2} \left(\frac{\gamma - 1}{\gamma + 1}\right)^k,$$

which is minimized by $t^k = t = \frac{2}{1+\gamma}$. We finally obtain:

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k \begin{bmatrix} (1 - t)\gamma \\ (1 - \gamma t)(-1)^k \end{bmatrix} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^{k+1} \begin{bmatrix} \gamma \\ (-1)^{k+1} \end{bmatrix} \tag{7}$$

which is the sought expression.

2 Problem 2 - Subgradient methods

(a) This relationship follows directly from the definition of Lipschitz continuity of f:

$$|f(y) - f(x)| \leq L||y - x||_2, \quad \forall x, y \in \mathbb{R}^n.$$

and by noting that, by definition of subgradient at x:

$$f(y) - f(x) \geq g^\top (y - x), \quad \forall y. \tag{8}$$

If we take $y = x + g$, we can write from (8):

$$||g||_2^2 = g^\top g \leq f(x + g) - f(x) \leq |f(x + g) - f(x)| \leq ||g||_2, \quad \forall y.$$

From which it can be concluded that $||g||_2 \leq$ L for any $g \in \partial f(x)$.

(b) As suggested, we start off by writing explicitly $||x^{i+1} - x||_2$ for a generic k:

$$\begin{aligned} ||x^{i+1} - x||_2^2 &= ||x^i - t^i g^i - x||_2^2 = ||x^i - x||_2^2 + t^{i^2}||g^i||_2^2 + 2t^i g^{i^\top}(x - x^i) \\ &\leq ||x^i - x||_2^2 + t^{i^2}||g^i||_2^2 + 2t^i(f(x) - f(x^i)), \quad \text{def. subradient} \\ &\leq ||x^i - x||_2^2 + t^{i^2}L^2 + 2t^i(f(x) - f(x^i)), \quad \text{because } ||g||_2 \leq L \end{aligned}$$

This can be rearranged as:

$$2t^i f(x^i) \leq 2t^i f(x) + {t^i}^2 L^2 + ||x^i - x||_2^2 - ||x^{i+1} - x||_2^2. \tag{9}$$

We denote $f_{\text{best}}^k = \min_{0 \leq i \leq k} f(x^i)$. We can the sum up (9) over i to obtain:

$$2\sum_{i=0}^{k} t^i f_{\text{best}}^k \leq 2\sum_{i=0}^{k} t^i f(x^i) \leq 2\sum_{i=0}^{k} t^i f(x) + \sum_{i=0}^{k} {t^i}^2 L^2 + ||x_{(0)} - x||_2^2 - ||x_{(k+1)} - x||_2^2.$$

The last term is non-negative, thus can be eliminated. This allows the final expression to be obtained:

$$f_{\text{best}}^k \leq f(x) + \frac{L^2 \sum_{i=0}^{k} {t^i}^2 + ||x_{(0)} - x||_2^2}{2\sum_{i=0}^{k} t^i}.$$

(c) We assume there is a minimizer of f and denote it by $x^\star$. For simplicity, we also assume that there exists $R > 0$ such that $R \geq ||x_{(0)} - x||_2^2$ (this is not necessary but it simplifies the notation). Then the result obtained in part (b) can be specialized to $x = x^\star$:

$$f_{\text{best}}^k - f(x^\star) \leq \frac{L^2 \sum_{i=0}^{k} {t^i}^2 + R^2}{2\sum_{i=0}^{k} t^i}. \tag{10}$$

In the constant step length case, i.e. $t_{(k)} = \frac{\beta}{||g^k||_2}$, Eq. (10) becomes:

$$f_{\text{best}}^k - f(x^\star) \leq L\frac{\beta^2 k + R^2}{2\beta k}.$$

using the fact that $t^i > \frac{\beta}{L}$. The limit for $k \to \infty$ of this expression tends to $\frac{L\beta}{2}$.

3 Problem 3 - Conjugate gradient method

Consider the problem of minimizing the function $f(x) = \frac{1}{2}x^\top Ax - b^\top x$, where $A \succ 0$, $b \in \mathbb{R}^n$, $\{p^k\}_{k=0}^{n-1}$ are conjugate directions with respect to A, and t^k is an exact line search. Define $r^k = Ax^k - b$.

(a) The property of achieving the optimal solution in at most n steps is a well known property of conjugate direction methods. Recall that the iterative algorithm can be written as:

$$x^{k+1} = x^k + t^k p^k, \tag{11}$$

and therefore, starting from an initial value $x_{(0)}$, the k-th value of x generated by the algorithm is:

$$x_k = x_{(0)} + \sum_{i=0}^{k-1} t^i p^i \tag{12}$$

The stepsize t^k is chosen such that it is an exact line search. Notice that the value of the gradient is $\nabla f(x)^k = Ax^k - b = r^k$. Therefore, the solution of part (a) in Problem 1 can be used to write the stepsize as:

$$t^k = -\frac{{r^k}^\top p^k}{{p^k}^\top A p^k}, \tag{13}$$

Recall also that by definition of conjugate directions the following holds:

$${p^i}^\top A p^j = 0, \quad \forall i \neq j. \tag{14}$$

Moreover, the conjugate directions $\{p^k\}_{k=0}^{n-1}$ form a basis of $\mathbb{R}^n$ and thus span the whole space. Therefore, we can write the difference between the solution $x^\star$ and the initial value $x_{(0)}$ as:

$$x^\star - x^0 = \sum_{k=0}^{n-1} \sigma_k p^k, \tag{15}$$

for some coefficients σ_k. To find them, we can premultiply (15) by ${p^k}^\top A$ (for $k = 0, ...n-1$) and exploit the property (14), in this way showing that:

$$\sigma_k = \frac{{p^k}^\top A(x^\star - x^0)}{{p^k}^\top A p^k}. \tag{16}$$

Eq. (15) shows that if the coefficients σ_k just derived are used as stepsize, then starting from $x_{(0)}$ the sequence of $\{x^k\}$ will indeed converge to $x^\star$ after n steps. Therefore, what is left to show is that (16) coincides with (13). Since the denominators are the same, we need to show that the numerators coincide.

To see this, premultiply (12) by ${p^k}^\top A$ (for $k = 0, ...n-1$) and use again the conjugacy property to obtain:

$${p^k}^\top A \left(x^k - x^0\right) = 0.$$

Thus, the numerator of (16) can be written as:

$$\begin{aligned} {p^k}^\top A(x^\star - x^0) &= {p^k}^\top A(x^\star - x^k + (x^k - x^0)) = {p^k}^\top A(x^\star - x^k) \\ &= {p^k}^\top (b - Ax^k) = -{p^k}^\top r^k, \end{aligned}$$

which coincides with the numerator of (13).

(b) In the conjugate gradient method, the coefficients β^k are chosen in such a way that the directions computed with the following formula are conjugate:

$$p^k = -r^k + \beta^k p^{k-1} \tag{17}$$

In the first iteration $p^0 = -\nabla f(x^0)$. Consider then a generic iteration k and premultiply (17) by ${p^{k-1}}^\top A$. By imposing the condition that ${p^{k-1}}^\top A p^k = 0$, one obtains:

$$\beta^k = \frac{{r^k}^\top A p^{k-1}}{{p^{k-1}}^\top A p^{k-1}}.$$

We now derive the sought expressions for t^k and β^k.

Note first that from the definition of r and Eq. (11) the following expression holds for the update of r:

$$r^{k+1} = r^k + t^k A p^k \tag{18}$$

From this, and the fact that all the directions are conjugate, it follows:

$${r^k}^\top p^j = 0, \quad \forall i = 0, ..., k-1. \tag{19}$$

Eq. (19), together with Eq. (17), shows that

$$-{r^k}^\top p^k = {r^k}^\top r^k \tag{20}$$

and thus the stepsize can be written as:

$$t^k = \frac{{r^k}^\top r^k}{{p^k}^\top A p^k} = \frac{{z^k}^\top z^k}{{y^k}^\top y^k},$$

where $z^k = r^k$ and $y^k = Up^k$ where $A = U^\top U$ is the Cholesky factorization of A.

As for β^k, we will use (18) to simplify its expression. By using previously derived formula, the denominator can be written as:

$$p^{k-1^\top} A p^{k-1} = \frac{1}{t^k} p^{k-1^\top}(r^k - r^{k-1}) = -\frac{1}{t^k} p^{k-1^\top} r_{k-1} = \frac{1}{t^k} r^{k-1^\top} r^{k-1}.$$

while the denominator

$$r^{k^\top} A p^{k-1} = \frac{1}{t^k} r^{k^\top}(r^k - r^{k-1}) = \frac{1}{t^k} r^{k^\top} r^k = \frac{1}{t^k} r^{k-1^\top} r^{k-1}.$$

Thus their ratio is

$$\beta_k = \frac{r^{k^\top} r^k}{r^{k-1^\top} r^{k-1}}.$$

as desired.

(c) The conjugate gradient method for the function $f(y) = \frac{1}{2} y^\top A y - \hat{b}^\top y$ is:

$$y^{k+1} = y^k + \gamma^k q^k,$$

where γ and q are used to denote the stepsize and conjugate directions in the new problem. The stepsize γ, direction q, gradient s, and parameter δ (analogous of β in the new problem) updates are given by:

$$\begin{aligned}
\gamma^k &= \frac{s^{k^\top} s^k}{q^{k^\top} A q^k}, \\
s^{k+1} &= s^k + \gamma^k A q^k, \\
q^{k+1} &= -s^{k+1} + \delta^{k+1} q^k, \\
\delta^{k+1} &= \frac{s^{k+1^\top} s^{k+1}}{s^{k^\top} s^k}.
\end{aligned}$$

Note that $y^0 = 0$, therefore $s^0 = Ay^0 - \hat{b} = -\hat{b} = Ax^0 - b = r^0$, and by definition the direction at the first timestep is $q^0 = -s^0 = -r^0 = p^0$. Therefore, directions and gradients are the same at $k = 0$. Suppose that they also coincide for some $k > 0$, i.e. assume $s^k = r^k$ and $q^k = p^k$. Then we can write:

$$\begin{aligned}
s^{k+1} &= s^k + \frac{s^{k^\top} s^k}{q^{k^\top} A q^k} A q^k = r^k + \frac{r^{k^\top} r^k}{p^{k^\top} A p^k} A p^k = r^{k+1}, \\
q^{k+1} &= -s^{k+1} + \frac{s^{k+1^\top} s^{k+1}}{s^{k^\top} s^k} q^k = -r^{k+1} + \frac{r^{k+1^\top} r^{k+1}}{r^{k^\top} r^k} p_k = p_{k+1},
\end{aligned}$$

It follows by induction that $s_k = r_k$ and $q_k = p_k \ \forall k \geq 0$. Moreover, we can show that:

$$y^{k+1} - y^k = \frac{s^{k^\top} s^k}{q^{k^\top} A q^k} q^k = \frac{r^{k^\top} r^k}{p^{k^\top} A p^k} p^k = x^{k+1} - x^k, \quad \text{for any } \ k \geq 0$$

Thus, one can see that $y^k = x^k - x^0$.