10-801: Advanced Optimization and Randomized Methods

Homework 5: Proximal methods, monotone operators, incremental methods

(April 9, 2014)

Instructor: *Suvrit Sra* ***Due:*** *April 21, 2014*

1. Consider the convex optimization problem

$$\min \quad f(x) \quad x \in \mathcal{X},$$

where $\mathcal{X}$ is closed and convex, while $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous and convex. Suppose further that we have a function $D_\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y),\, x - y\rangle,$$

where ϕ is strongly convex with parameter μ and continuously differentiable on $\mathcal{X}$ (as a result $D_\phi(x, y) \geq \frac{\mu}{2}\|x - y\|_2^2$).

   (a) Show that $D_\phi(x, y)$ is strongly convex as a function of x by proving that

$$D_\phi(x, y) \geq D_\phi(z, y) + \langle \nabla_z D(z, y),\, x - z\rangle + \frac{\mu}{2}\|x - z\|_2^2.$$

   (b) Consider the general iterative algorithm

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathcal{X}}\{\langle x,\, g^k\rangle + \frac{1}{\alpha_k} D_\phi(x, x^k)\}, \quad g^k \in \partial f(x^k). \tag{5.1}$$

   - Write down the optimality conditions for (5.1)
   - Use these optimality conditions to write the above update explicitly in terms of $\nabla\phi$ and $\nabla\phi^*$.

   Hint: You'll need the fact: ϕ is strongly convex on $\mathcal{X}$, so ϕ^* is finite everywhere and differentiable, with $\nabla\phi^* \equiv (\nabla\phi + N_{\mathcal{X}})^{-1}$; then consider $u \in \nabla\phi(x) + N_{\mathcal{X}}(x)$, where $N_{\mathcal{X}}$ is the normal cone.

   (c) Show why the projected subgradient iteration

$$x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k g^k), \quad k = 0, 1, \ldots,$$

   is actually a special case of iteration (5.1).

2. The Douglas-Rachford iteration for minimizing $f(x) + g(x)$ is given by

$$\begin{aligned} x^k &= \operatorname{prox}_g(z^k) \\ v^k &= \operatorname{prox}_f(2x^k - z^k) \\ z^{k+1} &= z^k + \gamma_k(v^k - x^k) \end{aligned}$$

Show that for $\gamma_k = 1$, we can rewrite the above iteration using *averaged reflections* as

$$z^{k+1} = [\tfrac{1}{2}(R_f R_g + I)](z^k),$$

where the reflection operators are $R_f := 2\operatorname{prox}_f - I$, and $R_g := 2\operatorname{prox}_g - I$.

3. Consider the following separable convex optimization problem

$$\min_{x\in\mathbb{R}^n} \quad F(x) := \sum\nolimits_{i=1}^{m} f_i(x),$$

where each $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ (e.g., $f_i(x) = \delta_C(x)$ for some closed convex set C).

   (a) Derive a Douglas-Rachford (DR) iteration to optimize $F(x)$ using the "product space trick". Justify all your steps.

(b) Write $F(x) = f_1(x) + \sum_{i=2}^m f_i(x)$. Introduce variables $x_2, \ldots, x_m = x_1$. Now, obtain the (Lagrange) dual problem in terms of the conjugate functions f_i^*. Show how to solve this dual problem using DR.

(c) Compare (in words) the two formulations in (a) and (b) above. Are there situations where you would prefer one over the other?

4. Let $A_1, \ldots, A_T$ be matrices in $\mathbb{R}^{m\times n}$, and let $y_1, \ldots, y_T \in \mathbb{R}$. Consider the trace-norm regularized optimization problem

$$\min_{X\in\mathbb{R}^{m\times n}} \quad \sum\nolimits_{j=1}^T (y_j - \operatorname{tr}(X^T A_j))^2 + \lambda\|X\|_{\mathrm{tr}},$$

where the *trace norm* is $\|X\|_{\mathrm{tr}} := \sum_i \sigma_i(X)$ (sum of singular values).

(a) Derive a closed-form solution for the proximity operator of the trace-norm

$$\frac{1}{2}\|X - Y\|_{\mathrm{F}}^2 + \lambda\|X\|_{\mathrm{tr}},$$

Hint 1: If $r : \mathbb{R}^n \to \mathbb{R}$ is symmetric convex and *absolute* ($r(x) = r(|x_1|, |x_2|, \ldots, |x_n|)$), and $\sigma : \mathbb{R}^{m\times n} \to \mathbb{R}_+^n$ is the singular value map, then the conjugate of the composition $r \circ \sigma$, i.e., $(r\circ\sigma)^*$ is (no surprise) $r^* \circ \sigma$.

(b) Present pseudo-code for solving this problem via proximal-gradients. Comment on how to select the step-size parameter.

5. Let $\mathcal{X}$ be a closed and bounded convex set. Let f be strongly convex with parameter μ. Assume we run the stochastic gradient method

$$x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k g^k),$$

where g^k is a *stochastic subgradient*, i.e., $E[g^k \mid \xi_{[k-1]}] \in \partial f(x^k)$, that has *finite variance*, i.e., $E[\|g^k\|^2] \le \sigma^2$. In this exercise, we'll study a small modification to the simple convergence analysis from Lecture 19. In particular, we'll show that a weighted average of the iterates x^k demonstrates $O(1/k)$ convergence rate.

(a) Prove the following inequality (we essentially proved it in class already):

$$E[\|x^{k+1} - x^*\|^2] \le E\|x^k - x^*\|^2 + \alpha_k^2 E[\|g_k\|^2] - 2\alpha_k[f(x^k) - f(x^*) + \tfrac{\mu}{2}\|x^k - x^*\|^2].$$

(b) Show from this inequality it follows that

$$E[f(x^k)] - f(x^*) \le \frac{\alpha_k\sigma^2}{2} + \frac{\alpha_k^{-1} - \mu}{2}E[\|x^k - x^*\|^2] - \tfrac{1}{2\alpha_k}E[\|x^{k+1} - x^*\|^2]. \tag{5.2}$$

(c) Show that choosing stepsize $\alpha_k = \frac{2}{\mu(k+1)}$, implies that

$$Ef(\bar{x}^k) - f^* \le \frac{2\sigma^2}{\mu(k+1)},$$

where $\bar{x}^k := \frac{2}{k(k+1)}\sum_{t=1}^k t x^t$.

(d) Show how $\bar{x}^k$ can be efficiently updated from iteration $k \to k+1$.

6. Suppose f is a convex function on a set C. An alternative definition of strong convexity of f on C with coefficient $\mu > 0$ is

$$f(\alpha x + (1-\alpha)y) + \tfrac{\mu}{2}\alpha(1-\alpha)\|x - y\|_2^2 \le \alpha f(x) + (1-\alpha)f(y).$$

Suppose f is a continuously differentiable function on $\operatorname{int}(C)$. Show that the following two are equivalent:

(a) f is strongly convex with strong convexity coefficient μ

(b) $\langle\nabla f(x) - \nabla f(y),\, x - y\rangle \ge \mu\|x - y\|_2^2, \quad \forall x, y \in \operatorname{int}(C)$.

Hint: For part (b), it might help to define $z(\alpha) = \alpha x + (1-\alpha)y$ and invoke the integral representation

$$f(z(\alpha)) = f(x) + \int_0^1 \langle\nabla f(x + t(z(\alpha) - x)),\, z(\alpha) - x\rangle dt$$

7. If f is not convex, we can still define a prox-operator, which is now a set-valued map:

$$\mathrm{prox}_f^\lambda \equiv y \mapsto \operatorname*{Argmin}_{x \in \mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \lambda f(x), \qquad \lambda > 0.$$

Obtain the prox-maps for the following functions

(a) $f(x) = \|x\|_0$, i.e., the ℓ_0-"norm".
(b) $f(x) = \|x\|_{1/2} := (\sum_i |x_i|^{1/2})^2$
(c) Does there exist a nonconvex f for which the prox-map is a singleton (for $n > 1$)?

8. Consider the convex optimization problem

$$\min \quad f(x) + h(Ax), \tag{5.3}$$

where f and h are closed convex functions, and A has full column rank. Assume that $\partial(f + h \circ A) = \partial f + \partial(h \circ A)$ (assume a similar qualification on the dual if needed).

(a) Write the Fenchel dual of this problem
(b) Show why running Douglas-Rachford on the dual yields the Alternating Direction Method of Multipliers (ADMM) for solving (5.3). (*Hint:* You may need to use the full DR method, not just its averaged reflections incarnation).