convex$_o pt$

ryanyuan42

March 2020

1 Subgradients and Proximal Operators

(i, 2 pts) Show that $\partial f(x)$ is a convex and closed set.

> Convex is easy to show.
> Closeness can be shown by arguing the complement of this set is open

(ii, 2 pts) Show that $\partial f(x) \subseteq N_{\{y:f(y)\leq f(x)\}}(x)$, where recall $N_C(x)$ denotes the normal cone to a set C at a point x. Give an example to show that this containment can be strict.

> if $g \in \partial f(x)$, then $f(y) \geq f(x) + g^T(y-x)$
> $N_{\{y:f(y)\leq f(x)\}}(x) = \{g : g^T x \geq g^T y, for\ y\ f(y) \leq f(x)\}$
>
> If $f(y) \leq f(x)$, $f(x) \geq f(y) \geq f(x) + g^T(y-x)$, $g^T x \geq g^T y$, i.e., $g \in N_{\{y:f(y)\leq f(x)\}}(x)$

(iii, 2 pts) Let $p, q > 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Consider the function $f(x) = ||x||_p = (\sum_{i=1}^n x_i^p)^{1/p}$. Show that $\forall x, y$:

$$x^T y \leq ||x||_p ||y||_q.$$

The above inequality is known as Hölder's inequality. Hint: you may use the dual representation of the ℓ_p norm, namely, $||x||_p = \max_{||z||_q\leq 1} z^T x$.

> Proof:
> $f(x) = ||x||_p = \max_{||z||_q\leq 1} z^T x \geq (\frac{y}{||y||_q})^T x$ therefore, $||x||_p ||y||_q \geq y^T x$

(iv, 3 pts) Use Hölder's inequality to show that for $f(x) = ||x||_p$, its subdifferential is $\partial f(x) =_{||z||_q \leq 1} z^T x$. (You are not allowed to use the rule for the subdifferential of a max of functions for this problem.)

Proof: we know the $\max_{||z||_q \leq 1} z^T x = ||x||_p$ So we want to show that $\partial ||x||_p = \{z \mid \|z\|_q \leq 1 \text{ and } z^\top x = \|x\|_p\}$,

To show $A = B$, we first show,

1. if $x \in A$, then $x \in B$

If $z \in \partial ||x||_p$, $\forall y$, we have $||y||_p \geq ||x||_p + z^T(y - x)$
Let y = 0 and y = 2x, we have $z^T x = ||x||_p$, plug it into the inequality we have $||y||_p \geq z^T y$
$\frac{z^T y}{||y||_p} \leq 1$, let $u = \frac{y}{||y||_p}$, $u^T z \leq 1$
we know that $\max_{||u||_p \leq 1} u^T z = ||z||_q$, so $||z||_q \leq 1$, which means $z \in \{z \mid \|z\|_q \leq 1 \text{ and } z^\top x = \|x\|_p\}$

2. if $x \in B$, then $x \in A$

If $z \in \{z \mid \|z\|_q \leq 1 \text{ and } z^\top x = \|x\|_p\}$ then by Hölder's inequality
$z^T y \leq ||y||_p ||z||_q \leq ||y||_p$
$||y||_p \geq ||x||_p + z^T - ||x||_p = ||x||_p + z^T(y - x)$
therefore, $z \in \partial ||x||_p$
$\partial ||x||_p = \{z \mid \|z\|_q \leq 1 \text{ and } z^\top x = \|x\|_p\} = argmax_{||z||_q \leq 1} z^T x$

2 Properties of Proximal Mappings and Subgradients

(a, 4pts) Prove one direction of the finite pointwise maximum rule for subdifferentials: The subdifferential of $f(x) = \max_{i=1,\ldots,n} f_i(x)$, for convex f_i, $i = 1, \ldots, m$, satisfies

$$\partial f(x) \supseteq \operatorname{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right). \tag{1}$$

Easy Proof, let set $S = \{i : f_i(x) = f(x)\}$ and check the convex combination of those functions and the subgradients of those functions (b, 4pts) Recall the definition of the proximal mapping: For a function h, the proximal mapping $prox_t$ is defined as

$$prox_t(x) = argmin_u \frac{1}{2t}\|x-u\|_2^2 + h(u). \tag{2}$$

Show that $prox_t(x) = u \Leftrightarrow h(y) \geq h(u) + \frac{1}{t}(x-u)^\top(y-u) \quad \forall y$.

Proof:
$prox_t(x) = argmin_u \frac{1}{2t}||x-u||_2^2 + h(u)$
1. If $h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u)$, and we know $\frac{1}{t}(x-u)^T(y-u) \geq \frac{1}{2t}(2x^Ty - 2x^Tu + u^Tu - y^Ty)$ (easy to find out by simple algebra, $(u-y)^T(u-y) \geq 0$)

then we have $\forall y$, $\frac{1}{2t}||x-y||_2^2 + h(y) \geq \frac{1}{2t}||x-u||_2^2 + h(u)$, i.e. $u = argmin_u \frac{1}{2t}||x-u||_2^2 + h(u) = prox_t(x)$

2. If $prox_t(x) = u$, we have $0 \in \frac{1}{t}(u-x) + \partial h(u)$, $\frac{1}{t}(x-u) \in \partial h(u)$, by the definition of subgradient,
$\forall y$, $h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u)$

Therefore, $prox_t(x) = u \Leftrightarrow h(y) \geq h(u) + \frac{1}{t}(x-u)^\top(y-u) \quad \forall y$

(c, 5 pts) Show how we can compose an affine mapping with the proximal operator. That is, assuming $f(x) = g(Ax+b)$, where $x \in R^n$, $A \in R^{m^*n}$, and $b \in R^m$, and also assuming $AA^T = aI_m$, for some scalar $a > 0$, then

$$prox_f(x) = x + \frac{1}{a}A^T\left(prox_{ag}(Ax+b) - Ax - b\right) \tag{3}$$

Hint: you may find it helpful to reparameterize $g(Ax+b)$ as $g(z)$ with the constraint that $z = Ax+b$, and then apply this constraint as a Lagrange multipler.

Proof:
$prox_f(x) = argmin_u \frac{1}{2}||u - x||_2^2 + f(u)$
$= argmin_u \frac{1}{2}||u - x||_2^2 + g(Au + b)$
$= argmin_{u,z} \frac{1}{2}||u - x||_2^2 + g(z), s.t, z = Au + b$

Lagranian multplier: $L(u, z, v) = \frac{1}{2}||u - x||_2^2 + g(z) + v^T(Au + b - z)$

We know the it has strong duality because the constraint satisfy the slater' condition, so we can apply KKT condtion that tells us

$u = x - A^T v$
$z = Au + b$
$0 \in \partial g(z) - v$
by first and second condition, we have $v = \frac{1}{a}(Ax + b - z)$
and by third condition, we have $0 \in \partial g(z) + \frac{1}{a}(z - Ax - b)$ so z minimize $g(z) + \frac{1}{2a}||z - Ax - b||_2^2$, so $z = prox_{ag}(Ax + b)$
$u = x - A^T v = x + \frac{1}{a}A^T(prox_{ag}(Ax + b) - Ax - b) = prox_f(x)$

(d, 5 pts) Show that if $\forall y \in \text{dom}(g)$, $\partial g(prox_f(y)) \supseteq \partial g(y)$, then

$$prox_{f+g}(x) = prox_f(prox_g(x)) \tag{4}$$

Hints:

1. Consider $prox_{f+g}(x)$, $prox_g(x)$, and $prox_f(prox_g(x))$.
2. The solution of the proximal can be characterized as:

$$u = prox_h(x) :=_u \frac{1}{2}\|u - x\|_2^2 + h(u) \iff 0 \in u - x + \partial h(u)$$

3. $\partial(f + g) = \partial f + \partial g$

Proof:
$prox_{f+g} = argmin \frac{1}{2}||z-x||_2^2 + f(z) + g(z)$
$prox_f = argmin \frac{1}{2}||z-x||_2^2 + f(z)$
$prox_g = argmin \frac{1}{2}||z-x||_2^2 + g(z)$

$0 \in prox_{f+g}(x) - x + \partial f(prox_{f+g}(x)) + \partial g(prox_{f+g}(x)))$
$0 \in prox_g(x) - x + \partial g(prox_g(x))$
$0 \in prox_f(prox_g(x)) - prox_g(x) + \partial f(prox_f(prox_g(x)))$

Sum up the last 2 statement, we know

$$0 \in prox_f(prox_g(x)) - x + \partial g(prox_g(x)) + \partial f(prox_f(prox_g(x)))$$

Let $y = prox_g(x)$, $0 \in prox_f(y) - x + \partial g(y) + \partial f(prox_f(y))$ because $\partial g(prox_f(y)) \supseteq \partial g(y)$, therefore we have

$$0 \in prox_f(y) - x + \partial g(prox_f(y)) + \partial f(prox_f(y))$$

Apparently this says, $prox_f(y)$ satisfy, $0 \in prox_{f+g}(x) - x + \partial f(prox_{f+g}(x)) + \partial g(prox_{f+g}(x)))$,
which means

$$prox_f(prox_g(x)) = prox_{f+g}(x)$$

3 Convergence Rate for Proximal Gradient Descent (20 pts) [Po-Wei]

In this problem, you will show the sublinear convergence for gradient descent and proximal gradient descent, which was presented in class.

To be clear, we assume that the objective $f(x)$ can be written as $f(x) = g(x) + h(x)$, where

(A1) g is convex, differentiable, and $dom(g) = R^n$.

(A2) ∇g is Lipschitz, with constant $L > 0$.

(A3) h is convex, not necessarily differentiable, and we take $dom(h) = R^n$ for simplicity.

(a) We begin with the simple case $f(x) = g(x)$; that is, $h(x) = 0$ and can be ignored. We will prove that the gradient descent converges sublinearly in this case. As a reminder, the iterates of gradient descent is computed by

$$x^+ = x - t\nabla g(x), \tag{5}$$

where x^+ is the iterate succeeding x. Henceforth, we will set $t = 1/L$ for simplicity.

(i, 3pt) Show that

$$g(x^+) - g(x) \le -\frac{1}{2L}\|\nabla g(x)\|^2.$$

That is, the objective value is monotonically decreasing in each update. This is why gradient descent is called a "descent method."

> By Lipschtiz continuity of the gradient, we have $g(y) \le g(x) + \nabla g(x)^T(y-x) + \frac{L}{2}||y-x||_2^2$
> Let $y = x^+ - t\nabla g(x)$, we have
> $g(x^+) \le g(x) - (t - \frac{t^2L}{2})||\nabla g(x)||_2^2$
> therefrore, $g(x^+) - g(x) \le -\frac{1}{2L}\|\nabla g(x)\|^2$.

(ii, 3pt) Using convexity of g, show the following helpful inequality:

$$g(x^+) - g(z) \le \nabla g(x)^T(x-z) - \frac{1}{2L}\|\nabla g(x)\|^2, \quad \forall z \in \mathbb{R}^n.$$

> By convexity, $\forall z$, $g(z) \ge g(x) + \nabla g(x)^T(z-x)$
> $g(x) - g(z) \le \nabla g(x)^T(x-z)$, and by (i),
> $g(x^+) - g(z) \le \nabla g(x)^T(x-z) - \frac{1}{2L}\|\nabla g(x)\|^2$

(iii, 2pt) Show that

$$g(x^+) - g(x^\star) \le \frac{L}{2}\left(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2\right),$$

> By (ii), we have $g(x^+) - g(x^*) \le \nabla g(x)(x - x^*) - \frac{1}{2L}\|\nabla g(x)\|^2 = \frac{L}{2}(\frac{2}{L}\nabla g(x)^T(x-x^*) - \frac{1}{L^2}\|\nabla g(x)\|^2) = -\frac{L}{2}(\|\frac{1}{L}\nabla g(x) - (x-x^*)\|^2 - \|x-x^*\|^2) = \frac{L}{2}(\|x-x^*\|^2 - \|x^+ - x^*\|^2)$

where $x^\star$ is the minimizer of g, assuming $g(x^\star)$ is finite.

(iv, 2pt) Now, aggregating the last inequality over all steps $i = 0, \ldots, k$, show that the accuracy of gradient descent at iteration k is $O(1/k)$, i.e.,

$$g(x^{(k)}) - g(x^\star) \le \frac{L}{2k}\|x^{(0)} - x^\star\|^2.$$

Put differently, for an ϵ-level accuracy, you need to run at most $O(1/\epsilon)$ iterations.

By summing over iteration, we have $\Sigma_{i=1}^{k} g(x^{(i)}) - g(x^*) \leq \frac{L}{2}(\|x^(0) - x^*\|_2^2)$
And we know the iteration is always dreasing in value, so $\Sigma_{i=1}^{k} g(x^{(i)}) - g(x^*) \geq kg(x^{(k)}) - kg(x^*)$ therefore,

$g(x^{(k)}) - g(x^*) \leq \frac{L}{2k}(\|x^{(0)} - x^*\|_2^2)$

(b) Now consider the general h in assumption (A3). We will prove that the proximal gradient descent converges sublinearly under such assumptions. Specifically, the iterates of proximal gradient descent is computed by

$$x^+ = \text{prox}_{th}\left(x - t\nabla g(x)\right), \tag{6}$$

where again we will set $t = 1/L$ for simplicity. Further, we define the useful notation

$$G(x) = \frac{1}{t}\left(x - x^+\right).$$

We will see (in the following proofs) that $G(x)$ behaves like $\nabla g(x)$ in gradient descent.

(i, 3pt) Show that

$$g(x^+) - g(x) \leq -\frac{1}{L}\nabla g(x)^T G(x) + \frac{1}{2L}\|G(x)\|^2.$$

we know $g(x^+) - g(x) \leq -\frac{1}{2L}\|\nabla g(x)\|^2$
and $\frac{1}{2L}(\|G(x)\|^2 - 2\nabla g(x)^T G(x) + \|\nabla g(x)\|^2) \geq 0$ therefore

$g(x^+) - g(x) \leq -\frac{1}{L}\nabla g(x)^T G(x) + \frac{1}{2L}\|G(x)\|^2$

(ii, 3pt) Show that

$$f(x^+) - f(z) \leq G(x)^T(x - z) - \frac{1}{2L}\|G(x)\|^2, \quad \forall z \in \mathbb{R}^n.$$

Note that setting $z := x$ verifies the proximal gradient descent is a "descent method." (Hint: Look back at what you did in Q2 part (b) and add the missing h to (i).)

By looking at Q2 part (b), we know $x^+ = \text{prox}_{th}(x - t\nabla g(x))$ means that
$h(x^+) - h(z) \le -\frac{1}{t}(x - t\nabla g(x) - x^+)^T(z - x^+) = G(x)^T(x^+ - x) + \nabla g(x)^T(z - x^+)$

$g(x^+) - g(z) \le \nabla g(x)^T(x - z) - \frac{1}{L}\nabla g(x)^T G(x) + \frac{1}{2L}\|G(x)\|^2$

Sum up the two inequlaities,
$f(x^+) - f(z) \le \nabla g(x)^T(x - x^+) - \frac{1}{L}\nabla g(x)^T G(x) + \frac{1}{2L}\|G(x)\|^2 + G(x)^T(x^+ - z)$

By some simple algebra, we can prove that $\nabla g(x)^T(x - x^+) - \frac{1}{L}\nabla g(x)^T G(x) + \frac{1}{2L}\|G(x)\|^2 + G(x)^T(x^+ - z) = G(x)^T(x - z) - \frac{1}{2L}||G(x)||^2$

Therefore, $f(x^+) - f(z) \le G(x)^T(x - z) - \frac{1}{2L}\|G(x)\|^2, \quad \forall z \in \mathbb{R}^n$.

(iii, 4pt) Show that

$$f(x^+) - f(x^\star) \le \frac{L}{2}\left(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2\right),$$

where $x^\star$ is the minimizer of f. Then show that

$$f(x^{(k)}) - f(x^\star) \le \frac{L}{2k}\|x^{(0)} - x^\star\|^2.$$

That is, the proximal descent method achieves $O(1/k)$ accuracy at the k-th iteration.

Proof:
It's very easy, let $z = x^*$ in b(ii), the rest follows the exact same logic as in a(iii)

Bonus. If we further assume g being strongly convex with constant m, show that the proximal gradient descent converges linearly, that is,

$$f(x^+) - f(x^\star) \le \left(1 - \frac{m}{L}\right)(f(x) - f(x^\star)).$$

You can use the following lemma. [Proximal Polyak-Łojasiewicz Inequality] Let $\lambda > 0$ be a scalar. Define

$$\phi(x;\lambda) = -2\lambda \min_y \left(\nabla g(x)^T(y - x) + \frac{\lambda}{2}\|y - x\|^2 + h(y) - h(x)\right),$$

then

$$\phi(x;\lambda_1) \le \phi(x;\lambda_2) \quad \text{if} \quad \lambda_1 \le \lambda_2.$$

Note that $\phi(x;\lambda)$ is related the minimum objective value in the proximal operators.
Hint: Bound $f(x) - f(x^\star)$ and $f(x) - f(x^+)$ using ϕ.

A very useful conclusion, remember, a lipschitz continuous gradient is suggesting a bound on the hessian matrix, why?

Proof:
$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
$\|\nabla f(x + hv) - \nabla f(x)\| \leq L\|hv\|$

By the definition of hessain, $\nabla^2 f(x)v = \lim_{h\to 0} \frac{\nabla f(x+hv) - \nabla f(x)}{h}$
and the definition of matrix norm(operator norm)
$\|A\| = sup\{\|Ax\|, \|x\| = 1\}$

from $\|\nabla f(x + hv) - \nabla f(x)\| \leq L\|hv\|$, we have
$\lim_{h\to 0} \frac{\|\nabla f(x+hv) - \nabla f(x)\|}{h} \leq L\|v\|$, i.e, $\|\nabla^2 f(x)v\| \leq L\|v\|$,
take supreme on both side, $\||\nabla^2 f(x)\| \leq sup_{\{\|v\|=1\}} L\|v\| = L$
since we have operator norm less or equal to L,

and from $\|\nabla^2 f(x)v\| \leq L\|v\|, \forall v$ we know that, $\forall v$, $\frac{\|\nabla^2 f(x)v\|}{\|v\|} \leq L$
Say v is one of the eigenvector of the hessian, we have, $\lambda \leq L$, therefore, the largest eigenvalue is less than L, that is saying, $\nabla^2 f(x) \preceq LI$

Strong convexity is giving lower bounds of the hessian, i.e.,$\nabla^2 f(x) \succeq mI$, why?
strong convexity says:
$g(x) = f(x) - \frac{m}{2}\|x\|_2^2$ is convex, i.e., $\nabla^2 g(x) \succeq 0$, i.e., $\nabla^2 f(x) \succeq mI$

This lemma is saying that, the condition number of the hessian is critical to the convergence speed, if the condition number is very large, i.e., ill-conditioned, $\frac{m}{L}$ is very small, close to 0, which makes each iteration doesn't get closer to optimal value