CSC/DSCC 574 Continuous Algorithms for Optimization and Sampling Lecture 4

Proximal Gradient Method

Lecturer: Jiaming Liang *February 1, 2024*

1 Proximal operator

Definition 1. *Given a function f, the proximal mapping of f is given by*

$$\text{prox}_f(x) = \text{argmin}_{u\in\mathbb{R}^n}\left\{f(u) + \frac{1}{2}\|u-x\|^2\right\}, \quad \forall x \in \mathbb{R}^n.$$

Note that if f is closed and convex then $\text{prox}_f(x)$ is a singleton for any $x \in \mathbb{R}^n$.

Example: soft-thresholding.

For some $\alpha > 0$, the proximal mapping for the one-dimensional function $\alpha|\cdot|$ is

$$\text{prox}_{\alpha|\cdot|}(y) = \mathcal{T}_\alpha(y) = [|y|-\alpha]_+ \,\text{sgn}(y) = \begin{cases} y-\alpha, & y \geq \alpha, \\ 0, & |y| < \alpha, \\ y+\alpha, & y \leq -\alpha. \end{cases}$$

Hence, the proximal mapping for $\alpha\|x\|_1$ is

$$\text{prox}_{\alpha\|\cdot\|_1}(x) = \mathcal{T}_\alpha(x) \equiv (\mathcal{T}_\alpha(x_j))_{j=1}^n = [|x| - \alpha\mathbf{1}]_+ \odot \text{sgn}(x), \tag{1}$$

where $\odot$ denotes componentwise multiplication. Now, consider a symmetric matrix $X \in \mathbb{S}^n$ with eigenvalue decomposition

$$X = U\,\text{diag}(\lambda(X))U^\top,$$

where $\lambda(X)$ denotes the eigenvalues of X in a vector form. Recall the nuclear norm $\|\cdot\|_*$ for symmetric matrices is defined as

$$\|X\|_* = \sum_{i=1}^n |\lambda_i(X)| = \|\lambda(X)\|_1.$$

Then, the proximal mapping for $\alpha\|X\|_*$ is

$$\text{prox}_{\alpha\|\cdot\|_*}(X) = U\,\text{diag}(\mathcal{T}_\alpha(\lambda(X)))U^\top,$$

where $\mathcal{T}_\alpha(\cdot)$ is as in (1).

Theorem 1. *Let $Q \subseteq \mathbb{R}^n$ be nonempty. Then, $\text{prox}_{I_Q}(x) = \text{proj}_Q(x)$ for any $x \in \mathbb{R}^n$. Let $Q \subseteq \mathbb{R}^n$ be a nonempty closed convex set. Then, $\text{prox}_{I_Q}(x) = \text{proj}_Q(x)$ is a singleton for any $x \in \mathbb{R}^n$.*

Theorem 2. *Let f be a closed and convex function. Then for any $x, y \in \mathbb{R}^n$, we have*

(i) $\|\operatorname{prox}_f(x) - \operatorname{prox}_f(y)\|^2 \leq \langle \operatorname{prox}_f(x) - \operatorname{prox}_f(y), x - y\rangle$;

(ii) $\|\operatorname{prox}_f(x) - \operatorname{prox}_f(y)\| \leq \|x - y\|$.

Proof. (a) Let $u = \operatorname{prox}_f(x)$ and $v = \operatorname{prox}_f(y)$. It follows from the defintion of proximal mapping that

$$u = \operatorname{argmin}_{w \in \mathbb{R}^n} \left\{ f(w) + \frac{1}{2}\|w - x\|^2 \right\}$$

and

$$x - u \in \partial f(u).$$

The inclusion is equivalent to

$$f(w) \geq f(u) + \langle x - u, w - u\rangle \quad \forall w \in \mathbb{R}^n.$$

Taking $w = v$, we have

$$f(v) \geq f(u) + \langle x - u, v - u\rangle.$$

Following the same argument for $v = \operatorname{prox}_f(y)$, we have

$$f(u) \geq f(v) + \langle y - v, u - v\rangle$$

Adding the above two inequalities, we obtain

$$0 \geq \langle y - x + u - v, u - v\rangle,$$

i.e.,

$$\langle x - y, u - v\rangle \geq \|u - v\|^2.$$

Plugging $u = \operatorname{prox}_f(x)$ and $v = \operatorname{prox}_f(y)$ into the above inequality, we prove (a).

(b) This statement simply follows from (a) using the Cauchy-Schwarz inequality. □

2 Proximal gradient method

2.1 Composite optimization

$$\min\{\phi(x) := f(x) + h(x)\}$$

- h is closed and convex;
- f is closed and convex, $\operatorname{dom} f$ is convex, $\operatorname{dom} h \subseteq \operatorname{int}(\operatorname{dom} f)$, and f is L-smooth over $\operatorname{int}(\operatorname{dom} f)$;
- the optimal set X_* is nonempty.

2.2 Proximal gradient

Algorithm 1 Proximal gradient method

Input: Initial point $x_0 \in \operatorname{dom} h$, stepsize $\lambda > 0$

for $k \geq 0$ **do**

Compute $x_{k+1} = \operatorname{prox}_{\lambda h}(x_k - \lambda \nabla f(x_k))$.

end for

Theorem 3. *Functions f and h are as assumed in Subsection 2.1. Choose $\lambda \in (0, 1/L]$. Then, the proximal gradient method generates a sequence of points $\{x_k\}$ satisfying*

$$\phi(x_k) - \phi_* \leq \frac{\|x_0 - x_*\|^2}{2\lambda k}, \quad \forall k \geq 1.$$

Proof. It is easy to verify that one iteration of the proximal gradient method can be written as

$$x_{k+1} = \min_{x \in \mathbb{R}^n} \left\{ \ell_f(x; x_k) + h(x) + \frac{1}{2\lambda}\|x - x_k\|^2 \right\}.$$

Using Theorem 2 of Lecture 2 and the fact that the above objective function is $(1/\lambda)$-strongly convex, we have for every $x \in \operatorname{dom} h$,

$$\begin{aligned}
\ell_f(x; x_k) + h(x) + \frac{1}{2\lambda}\|x - x_k\|^2 &\geq \ell_f(x_{k+1}; x_k) + h(x_{k+1}) + \frac{1}{2\lambda}\|x_{k+1} - x_k\|^2 + \frac{1}{2\lambda}\|x - x_{k+1}\|^2 \\
&\geq \ell_f(x_{k+1}; x_k) + h(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\lambda}\|x - x_{k+1}\|^2 \\
&\geq f(x_{k+1}) + h(x_{k+1}) + \frac{1}{2\lambda}\|x - x_{k+1}\|^2,
\end{aligned}$$

where the second inequality is due to $\lambda \leq 1/L$ and the last inequality is due to smoothness of f. It then follows from the convexity of f that

$$f(x) + h(x) + \frac{1}{2\lambda}\|x - x_k\|^2 \geq f(x_{k+1}) + h(x_{k+1}) + \frac{1}{2\lambda}\|x - x_{k+1}\|^2.$$

Taking $x = x_k$, we have

$$f(x_k) + h(x_k) \geq f(x_{k+1}) + h(x_{k+1}) + \frac{1}{2\lambda}\|x_{k+1} - x_*\|^2 \geq f(x_{k+1}) + h(x_{k+1}),$$

which shows that the function value of the iterates is a nonincreasing sequence. Taking $x = x_*$, we have

$$f(x_*) + h(x_*) + \frac{1}{2\lambda}\|x_k - x_*\|^2 \geq f(x_{k+1}) + h(x_{k+1}) + \frac{1}{2\lambda}\|x_{k+1} - x_*\|^2,$$

i.e.,

$$(f + h)(x_{k+1}) - (f + h)(x_*) \leq \frac{1}{2\lambda}\|x_k - x_*\|^2 - \frac{1}{2\lambda}\|x_{k+1} - x_*\|^2.$$

Summing the above inequality and using the monotinicity of $\{(f+h)(x_k)\}$, we obtain

$$k\left[(f+h)(x_k)-(f+h)(x_*)\right] \le \sum_{i=0}^{k-1}(f+h)(x_{i+1})-(f+h)(x_*) \le \frac{1}{2\lambda}\|x_0-x_*\|^2-\frac{1}{2\lambda}\|x_k-x_*\|^2.$$

Thus, the claim of the theorem follows. □

3 Dual proximal method

Consider the problem

$$\min\{f(x)+h(Ax) : x\in\mathbb{R}^n\}$$

where $A\in\mathbb{R}^{m\times n}$ and

- h is closed and convex;
- f is closed and μ-strongly convex;
- there exist $\hat{x}\in \mathrm{ri}\,(\mathrm{dom}\, f)$ and $\hat{z}\in \mathrm{ri}\,(\mathrm{dom}\, h)$ such that $A\hat{x}=\hat{z}$.

Strong duality holds in this case.

3.1 Dual problem

Consider an equivalent problem

$$\begin{aligned}&\min_{x,z\in\mathbb{R}^n} f(x)+h(z)\\ &\text{s.t. } Ax-z=0.\end{aligned}$$

We define the Lagrangian as

$$L(x,z;y)=f(x)+h(z)-y^\top(Ax-z), \tag{2}$$

and the dual function is

$$\begin{aligned}d(y)&=\inf_{x,z} L(x,z;y)\\ &=\inf_x\left\{f(x)-y^\top Ax\right\}+\inf_z\left\{h(z)+y^\top z\right\}\\ &=-\sup_x\left\{(A^\top y)^\top x-f(x)\right\}-\sup_z\left\{(-y)^\top z-h(z)\right\}\\ &=-f^*(A^\top y)-h^*(-y),\end{aligned}$$

where f^* and h^* denote the conjugates of f and h, respectively. Thus, the dual problem is

$$\max_{y\in\mathbb{R}^n} d(y).$$

We consider the dual problem in its minimization form

$$\min_{y\in\mathbb{R}^m} F(y) + H(y) \tag{3}$$

where

$$F(y) = f^*(A^\top y), \quad H(y) = h^*(-y).$$

Lemma 1. *We have F is convex and L_F-smooth where $L_F = \|A\|^2/\mu$ and H is closed and convex.*

Proof. Since f is μ-strongly convex, by conjugacy, we know f^* is $(1/\mu)$-smooth. Thus, for any $y_1, y_2 \in \mathbb{R}^m$, we have

$$\begin{aligned}
\|\nabla F(y_1) - \nabla F(y_2)\| &= \|A\nabla f^*(A^\top y_1) - A\nabla f^*(A^\top y_2)\| \\
&\le \|A\|\|\nabla f^*(A^\top y_1) - \nabla f^*(A^\top y_2)\| \\
&\le \frac{\|A\|}{\mu}\|A^\top(y_1 - y_2)\| \\
&\le \frac{\|A\|^2}{\mu}\|y_1 - y_2\|.
\end{aligned}$$

By conjugacy and the fact that convexity preserves under composition of a convex function and a linear mapping, we know both F and H are convex. □

3.2 Dual proximal method

Since the dual problem (3) is the sum of a convex smooth function $F(y)$ and a convex composite function $H(y)$, which is exactly the setting for proximal gradient method, we apply Algorithm 1 to (3).

Algorithm 2 Dual proximal method

Input: Initial point $y_0 \in \mathbb{R}^m$
for $k \ge 0$ **do**
  Compute $y_{k+1} = \text{prox}_{\lambda H}(y_k - \lambda\nabla F(y_k))$.
end for

Since F is convex and L_F-smooth and H is closed and convex, invoking Theorem 3, we obtain the convergence rate of the dual sequence.

Theorem 4. *Choose $\lambda \in (0, 1/L_F]$. Then, Algorithm 2 generates a sequence of points $\{y_k\}$ satisfying*

$$d^* - d(y_k) \leq \frac{\|y_0 - y^*\|^2}{2\lambda k}, \quad \forall k \geq 1.$$

Lemma 2. *The dual iteration $y_{k+1} = \text{prox}_{\lambda H}(y_k - \lambda \nabla F(y_k))$ can be equivalently rewritten as*

$$x_{k+1} = \text{argmax}_{x \in \mathbb{R}^n} \{\langle x, A^\top y_k \rangle - f(x)\}, \tag{4}$$

$$y_{k+1} = y_k - \lambda A x_{k+1} + \lambda \, \text{prox}_{\frac{1}{\lambda} h}\left(A x_{k+1} - \frac{1}{\lambda} y_k\right). \tag{5}$$

Proof. Note that the dual proximal update can be written as

$$y_{k+1} = \min_{y \in \mathbb{R}^m} \left\{ \ell_F(y; y_k) + H(y) + \frac{1}{2\lambda} \|y - y_k\|^2 \right\}.$$

Its optimality condition is

$$0 \in \nabla F(y_k) + \partial H(y_{k+1}) + \frac{y_{k+1} - y_k}{\lambda}. \tag{6}$$

It follows from Proposition 1 of Lecture 3 and (4) that

$$\nabla F(y_k) = A \nabla f^*(A^\top y_k) = A \, \text{argmax}_x \{\langle A^\top y_k, x \rangle - f(x)\} \overset{(4)}{=} A x_{k+1}.$$

Define

$$z_{k+1} = \frac{y_{k+1} - y_k}{\lambda} + \nabla F(y_k) = \frac{y_{k+1} - y_k}{\lambda} + A x_{k+1}. \tag{7}$$

Then, it follows from the optimality condition (6) that

$$-z_{k+1} \in \partial H(y_{k+1}) = -\partial h^*(-y_{k+1}).$$

Using Theorem 3 of Lecture 3, we have

$$\partial h(z_{k+1}) \ni -y_{k+1}.$$

Hence,

$$0 \in y_{k+1} + \partial h(z_{k+1}).$$

Equivalently, by (7), we have

$$0 \in \partial h(z_{k+1}) + y_k + \lambda z_{k+1} - \lambda A x_{k+1}.$$

It is interesting to see that the above inclusion is also the optimality condition of

$$z_{k+1} = \text{argmin}_{z \in \mathbb{R}^m} \left\{ h(z) + \langle z, y_k \rangle + \frac{\lambda}{2} \|z - A x_{k+1}\|^2 \right\}.$$

Using Definition 1, we have

$$z_{k+1} = \text{prox}_{\frac{1}{\lambda} h}\left(A x_{k+1} - \frac{1}{\lambda} y_k\right).$$

Finally, it follows from (7) and the above formula for z_{k+1} that (5) holds. □

Note: The proof of Lemma 2 can be simplified using Moreau decomposition formula, see Amir Beck's book Lemma 12.5.

Using Lemma 2, we can rewrite Algorithm 2 in its primal form.

Algorithm 3 Dual proximal method (primal form)

Input: Initial point $y_0 \in \operatorname{dom} g$

for $k \geq 0$ **do**

Compute $x_{k+1} = \operatorname{argmax}_{x\in\mathbb{R}^n}\{\langle x, A^\top y_k\rangle - f(x)\}$.

Compute $y_{k+1} = y_k - \lambda A x_{k+1} + \lambda \operatorname{prox}_{\frac{1}{\lambda}h}\left(Ax_{k+1} - \frac{1}{\lambda}y_k\right)$.

end for

It is clear from the proof of Lemma 2 that the dual proximal method has another presentation in the alternating minimization form using the z sequence.

Algorithm 4 Dual proximal method (alternating minimization form)

Input: Initial point $y_0 \in \operatorname{dom} g$

for $k \geq 0$ **do**

Compute $x_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n}\{f(x) - \langle x, A^\top y_k\rangle\}$.

Compute $z_{k+1} = \operatorname{argmin}_{z\in\mathbb{R}^m}\left\{h(z) + \langle z, y_k\rangle + \frac{\lambda}{2}\|z - Ax_{k+1}\|^2\right\}$.

Compute $y_{k+1} = y_k - \lambda A x_{k+1} + \lambda z_{k+1}$.

end for

In fact, Algorithm 4 can be understood from the augmented Lagrangian perspective. Recall the Lagrange function $L(x, z; y)$ is defined in (2). We define the augmented Lagrange function as follows

$$L_\lambda(x, z; y) := L(x, z; y) + \frac{\lambda}{2}\|Ax - z\|^2 = f(x) + h(z) - y^\top(Ax - z) + \frac{\lambda}{2}\|Ax - z\|^2.$$

Then, we rewrite Algorithm 4 as

$$\begin{aligned}
x_{k+1} &= \operatorname{argmin}_{x\in\mathbb{R}^n} L(x, z_k; y_k),\\
z_{k+1} &= \operatorname{argmin}_{z\in\mathbb{R}^m} L_\lambda(x_{k+1}, z; y_k),\\
y_{k+1} &= y_k - \lambda(Ax_{k+1} - z_{k+1}) = y_k + \lambda\nabla_y L(x_{k+1}, z_{k+1}; y).
\end{aligned}$$

With the understanding that the dual proximal method is the proximal gradient method applied to the dual problem in mind, we also note that the dual ascent method is the subgradient method applied to the dual problem, and the augmented Lagrangian method is the proximal point method applied to the dual problem.

Theorem 5. *Choose $\lambda \in (0, 1/L_F]$. Then, Algorithm 3 generates a sequence of pairs $\{(x_k, y_k)\}$ satisfying*

$$\|x_k - x_*\|^2 \leq \frac{2}{\mu}[d^* - d(y_k)]. \tag{8}$$

Thus, we have

$$\|x_k - x_*\|^2 \leq \frac{\|y_0 - y^*\|^2}{\lambda \mu k}, \quad \forall k \geq 1. \tag{9}$$

Proof. Relation (8) holds in view of Lemma 12.7 of Amir Beck's book. We omit the proof and suggest the interested readers to read the book for the proof of Lemma 12.7. (We should be also able to prove (8) by first verifying the dual proximal framewotk being an inexact proximal point method and then using the generic convergence results of the framework.) It follows from (8) and Theorem 4 that (9) holds. □

Note that similar to (8), using the μ-strong convexity of ϕ from the primal perspective, we also have

$$\phi(x_k) - \phi_* \geq \frac{\mu}{2}\|x_k - x_*\|^2.$$