10-725/36-725: Convex Optimization	**Fall 2019**

Lecture 25: November 20

Lecturer: Javier Peña | *Scribes: Addison Hu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Remark. If you find the content of this lecture interesting, consider 47-860, Convex Analysis, MW 3:30 - 5:20pm, Mini-3, 2020.

Remark. This lecture is based off of the paper: `https://arxiv.org/abs/1812.10198`.

25.1 Review: (Euclidean) proximal methods

Composite convex minimization. Consider the problem:

$$\min_{x \in \mathbb{R}^n} \{f(x) + \psi(x)\} \tag{25.1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is differentiable and convex, and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed and convex with $\mathrm{dom}(\psi) \subseteq \mathrm{dom}(f)$. Note that ψ tends to be a regularization term.

Let Prox_t be the following *proximal map*:

$$\mathrm{Prox}_t(x) := \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|z - x\|^2 + \psi(z) \right\} \tag{25.2}$$

There are a couple ways for us to approach (25.1).

Proximal gradient (PG).

$$\begin{aligned}
&\text{pick } t_k > 0 \\
&x_{k+1} = \mathrm{Prox}_{t_k}(x_k - t_k \nabla f(x_k))
\end{aligned}$$

Accelerated proximal gradient (APG).

$$\begin{aligned}
&\text{pick } \beta_k \geq 0, t_k > 0 \\
&y_k = x_k + \beta_k (x_k - x_{k-1}) \\
&x_{k+1} = \mathrm{Prox}_{t_k}(y_k - t_k \nabla f(y_k))
\end{aligned}$$

Stepsize. Choosing an appropriate stepsize for the generic update $z_+ = \text{Prox}_t(y - t\nabla f(y))$ motivates the Bregman distance. Observe that:

$$\text{Prox}_t(y - t\nabla f(y)) = \arg\min_{z\in\mathbb{R}^n} \left\{ f(y) + \langle \nabla f(y), z - y\rangle + \frac{1}{2t}\|z-y\|^2 + \psi(z) \right\}$$

Therefore, it makes sense to choose t such that z_+ satisfies:

$$f(z_+) + \psi(z_+) \leq f(y) + \langle(\nabla f(y), z_+ - y\rangle + \frac{1}{2t}\|z_+ - y\|^2 + \psi(z_+)$$

$$\underbrace{f(z_+) - f(y) - \langle \nabla f(y), z_+ - y\rangle}_{=D_f(z_+,y)} \leq \frac{1}{2t}\|z_+ - y\| \tag{25.3}$$

Definition 25.1 (Bregman distance.)

$$D_f(z,y) := f(z) - f(y) - \langle \nabla f(y), z - y\rangle$$

With this definition, the condition (25.3) may more succinctly be stated:

$$D_f(z_+, y) \leq \frac{1}{2t}\|z_+ - y\|^2$$

Definition 25.2 (L-smoothness.) *We say that a function f is L-smooth if for all $z, y \in \text{dom}(f)$,*

$$D_f(z,y) \leq \frac{L}{2}\|z-y\|^2$$

In this case condition (25.3) *holds for* $t = \frac{1}{L}$

Remark. f is L-smooth if ∇f is L-Lipschitz.

Convergence of Proximal Gradient. Suppose we solve (25.1) via $x_{k+1} = \text{Prox}_{t_k}(x_k - t_k\nabla f(x_k))$.

Theorem 25.3 *If the stepsizes t_k satisfy*

$$D_f(x_{k+1}, x_k) \leq \frac{1}{2t_k}\|x_{k+1} - x_k\|^2$$

then for all $\bar{x} \in \arg\min_x\{f(x) + \psi(x)\}$, the Proximal Gradient iterates satisfy

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{\|x_0 - \bar{x}\|^2}{2\sum_{i=0}^{k-1} t_i}$$

In particular, if each $t_k \geq \frac{1}{L} > 0$ then

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{L \cdot \|x_0 - \bar{x}\|^2}{2k} = O(1/k)$$

Convergence of Accelerated Proximal Gradient. Suppose we solve (25.1) using the updates:

$$y_k = x_k + \beta_k(x_k - x_{k-1})$$
$$x_{k+1} = \mathrm{Prox}_{t_k}(y_k - t_k \nabla f(y_k))$$

Theorem 25.4 (Beck & Teboulle 2009, Nesterov 2013) *Suppose $\beta_k = \frac{k-1}{k+2}$ and the stepsizes t_k satisfy $t_k \geq 1/L > 0$ and*

$$D_f(x_{k+1}, y_k) \leq \frac{1}{2t_k}\|x_{k+1} - y_k\|^2$$

Then for all $\bar{x} \in \arg\min_x\{f(x) + \psi(x)\}$ the Accelerated Proximal Gradient iterates satisfy

$$f(x_k) + \psi(x_k) - (f(\bar{x}) - \psi(\bar{x})) \leq \frac{2L \cdot \|x_0 - \bar{x}\|^2}{(k+1)^2} = O(1/k^2)$$

25.2 Bregman proximal methods

We will now generalize the Euclidean proximal map of Section 25.1 to the *Bregman proximal map*. In doing so, we will see that we may recover $O(1/k)$ and $O(1/k^2)$ convergence of proximal gradient methods when f is L-smooth.

Definition 25.5 (Bregman proximal map.)

$$g \mapsto \arg\min_{y \in \mathbb{R}^n} \left\{ \langle g, y \rangle + \frac{1}{t} D_h(y, x) + \psi(y) \right\}$$

The idea here is to replace $\frac{1}{2t}\|z - x\|^2$ with $\frac{1}{t}D_h(z, x)$. The Euclidean proximal map previously considered in Section 25.1 corresponds to the squared Euclidean norm reference function

$$h(x) = \frac{\|x\|^2}{2} \rightsquigarrow D_h(y, x) = \frac{\|y - x\|^2}{2}$$

25.2.1 Bregman proximal gradient

Consider problem (25.1) and suppose $h : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is a reference function. The **Bregman proximal gradient (BPG)** method does:

$$\begin{aligned}
&\text{pick } t_k > 0 \\
x_{k+1} &= \arg\min_{z \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), z \rangle + \frac{1}{t_k} D_h(z, x_k) + \psi(z) \right\} \\
&= \arg\min_{z \in \mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k), z - x_k \rangle + \frac{1}{t_k} D_h(z, x_k) + \psi(z) \right\}
\end{aligned}$$

Convergence. Bregman proximal gradient has $O(1/k)$ convergence when f is *smooth relative to h*, i.e., when

$$D_f(y, x) \leq L \cdot D_h(y, x) \tag{25.4}$$

for all $x, y \in \mathrm{dom}(f)$

25.2.2 Accelerated Bregman proximal gradient

For the same problem 25.1, the accelerated Bregman proximal gradient method (Gutman-Peña) generates sequences x_k, y_k, z_k for $k = 0, 1, \ldots$ as follows:

$$\begin{aligned}
&\text{pick } t_k > 0 \\
&z_{k+1} = \arg\min_{z\in\mathbb{R}^n} \left\{ \langle \nabla f(y_k), z\rangle + \frac{1}{t_k} D_h(z, z_k) + \psi(z) \right\} \\
&x_{k+1} = \frac{\sum_{i=0}^k t_i z_{i+1}}{\sum_{i=0}^k t_i} \\
&y_{k+1} = \frac{\sum_{i=0}^k t_i z_{i+1} + t_{k+1} z_{k+1}}{\sum_{i=0}^{k+1} t_i}
\end{aligned}$$

See related work by Hanzely-Richtarik-Xiao (2018).

Convergence. Accelerated Bregman proximal gradient has convergence $O(1/k^\gamma)$ if f is (L, γ)-smooth relative to h, as defined in the sequel.

25.2.3 Why Bregman proximal methods?

By generalizing the reference function beyond the Euclidean squared norm, we attain additional freedom which may aid the computation of the proximal mapping. For example, for

$$x \in \Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$$

the map

$$g \mapsto \arg\min_{y\in\Delta_{n-1}} \{\langle g, y\rangle + D_h(y, x)\}$$

is much simpler for $h(x) = \sum_{i=1}^n x_i \log(x_i)$ than for $h(x) = \|x\|^2/2$. We will generalize the L-smoothness assumption for convergence to *relative* L-smoothness.

The following two examples could be solved via Euclidean proximal methods, but they are more amenable to Bregman proximal methods with the *Burg entropy* reference function: $h(x) = -\sum_{i=1}^n \log(x_i)$:

- **D-optimal design problem (min-volume closing ellipsoid).**

  $$\min_{x\in\Delta_{n-1}} -\log(\det(HXH^\top))$$

  where $X = \mathrm{Diag}(x)$ and $H \in \mathbb{R}^{m\times n}$ with $m < n$.

- **Poisson linear inverse problem.**

  $$\min_{x\in\mathbb{R}^n_+} D_{KL}(b, Ax)$$

  where $b \in \mathbb{R}^n_{++}$ and $A \in \mathbb{R}^{m\times n}_+$ with $m > n$ and $D_{KL}(\cdot,\cdot)$ is the Kullback-Leibler divergence.

25.3 Convergence details for Bregman proximal methods

25.3.1 Fenchel duality

We first recall some details about duality.

Definition 25.6 (Convex conjugate) *For $\phi : \mathbb{R}^n \to \mathbb{R}$ let $\phi^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be defined via*

$$\phi^*(u) = \sup_{x \in \mathbb{R}^n} \{\langle u, x\rangle - \phi(x)\}$$

Consider the *primal* problem

$$\min_x \{f(x) + \psi(x)\}$$

the corresponding *Fenchel dual* problem is

$$\max_u \{-f^*(u) - \psi^*(-u)\}$$

Observe that if $f(\bar{x}) + \psi(\bar{x}) = -f^*(\bar{u}) - \psi^*(-\bar{u})$ then $\bar{x}, \bar{u}$ are optimal.

25.3.2 Warm-up towards convergence

Suppose an algorithm generates sequences x_k, v_k, w_k such that

$$f(x_k) + \psi(v_k) \le -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)$$

for some sequence of "distance" functions $d_k : \mathbb{R}^n \to \mathbb{R}$. Then for all $\bar{x} \in \arg\min_x\{f(x) + \psi(x)\}$ we have

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \le d_k(\bar{x}) \tag{25.5}$$

Observe that this gives us a suboptimality gap for free. For suitable t_k, Bregman proximal gradient and accelerated Bregman proximal gradient satisfy (25.5) for

$$d_k(z) = \frac{1}{\sum_{i=0}^k t_i} D_h(z, z_0)$$

We now state a key lemma for Bregman proximal methods. Suppose $y_k, z_k \in \mathrm{ri}(\mathrm{dom}(h)) \cap \mathrm{dom}(\psi)$, $g_k := \nabla f(y_k)$, and $t_k > 0$ satisfy

$$z_{k+1} = \arg\min_{z \in \mathbb{R}^n} \left\{ \langle g_k, z\rangle + \frac{1}{t_k} D_h(z, z_k) + \psi(z) \right\}$$

for $k = 0, 1, 2, \ldots$. We may rewrite this via optimality conditions as:

$$g_k + g_k^\psi + \frac{1}{t_k}(\nabla h(z_{k+1}) - \nabla h(z_k)) = 0 \tag{25.6}$$

for some $g_k^\psi \in \partial\psi(z_{k+1})$.

Let

$$v_k := \frac{\sum_{i=0}^k t_i g_i}{\sum_{i=0}^k t_i}, \quad w_k := \frac{\sum_{i=0}^k t_i g_i^\psi}{\sum_{i=0}^k t_i}$$

Lemma 25.7 *Suppose $y_k, z_k, g_k, g_k^\psi, t_k$ and v_k, w_k are as previously defined. Then*

$$\begin{aligned}\frac{\sum_{i=0}^k t_i(f(z_{i+1}) + \psi(z_{i+1}) - D_f(z_{i+1}, y_i)) + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k t_i} &= -\frac{\sum_{i=0}^k t_i(f^*(g_i) + \psi^*(g_i^\psi))}{\sum_{i=0}^k t_i} - d_k^*(-v_k - w_k)\\ &\leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)\end{aligned}$$

where $d_k(z) := \frac{1}{\sum_{i=0}^k t_i} D_h(z, z_0)$

25.3.3 Convergence of Bregman proximal gradient

Recall the Bregman proximal gradient algorithm from Section 25.2.1.

Theorem 25.8 (Gutman-Peña 2018) *Suppose each t_i is such that*

$$D_f(x_{i+1}, x_i) \leq \frac{1}{t_i} D_h(x_{i+1}, x_i) \tag{25.7}$$

Then for $\bar{x} \in \arg\min_{x \in \mathbb{R}^n}\{f(x) + \psi(x)\}$ the Bregman proximal gradient iterates satisfy

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{1}{\sum_{i=0}^k t_i} D_h(\bar{x}, x_0)$$

Proof: We apply Lemma 25.7 to $x_k = y_k = z_k$ and obtain:

$$\frac{\sum_{i=0}^k t_i(f(x_i + 1) + \psi(x_{i+1}) - D_f(x_{i+1}, x_i)) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)$$

Then, (25.7) implies

$$f(x_{k+1}) + \psi(x_{k+1}) \leq \frac{\sum_{i=0}^k t_i(f(x_{i+1}) + \psi(x_{i+1}))}{\sum_{i=0}^k t_i} \qquad \leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)$$

Thus for all $\bar{x} \in \arg\min_{x \in \mathbb{R}^n}\{f(x) + \psi(x)\}$

$$f(x_k) + \psi(x_k) \leq f(\bar{x}) + \psi(\bar{x}) + \frac{1}{\sum_{i=0}^k t_i} D_h(\bar{x}, x_0)$$

■

25.3.4 Relative smoothness

We will see that relative smoothness is a natural extension of smoothness beyond the Euclidean intuition. Suppose f, h are convex and differentiable on Q. We say that f is L-smooth relative to h on Q if for all $x, y \in Q$

$$D_f(y, x) \leq L \cdot D_h(y, x)$$

(Nguyen 2012, Bauschke et al. 2017, Lu et al. 2018).

If f is L-smooth relative to h on $\mathrm{dom}(\psi)$ then (25.7) holds for $t_i = 1/L, i = 0, 1, \ldots, k-1$ and the Bregman proximal gradient iterates satisfy

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{L \cdot D_h(\bar{x}, x_0)}{k}$$

This recovers results by Bauschke-Bolte-Teboulle (2017) and by Lu-Freund-Nesterov (2018). This also extends the $O(1/k)$ convergence rate of proximal gradient.

25.3.5 Convergence of the accelerated Bregman proximal gradient method

Recall the accelerated Bregman proximal gradient method from Section 25.2.2. By letting $\theta_k := \frac{t_k}{\sum_{i=0}^k t_i}$, the updates may be rewritten as

$$\begin{aligned} z_{k+1} &= \arg\min_{z\in\mathbb{R}^n} \left\{ \langle \nabla f(y_k), z\rangle + \frac{1}{t_k} D_h(z, z_k) + \psi(z) \right\} \\ x_{k+1} &= (1-\theta_k)x_k + \theta_k z_{k+1} \\ y_{k+1} &= (1-\theta_{k+1})x_{k+1} + \theta_{k+1} z_{k+1} \\ &= x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}(x_{k+1} - x_k) \end{aligned}$$

Notice that the x_{k+1} update is a convex combination of the past updates and the current update.

Theorem 25.9 (Gutman-Peña 2018) *Suppose each t_i and θ_i are such that*

$$D_f(x_{i+1}, y_i) - (1-\theta_i)D_f(x_i, y_i) \le \frac{\theta_i}{t_i} D_h(z_{i+1}, z_i) \tag{25.8}$$

Then for $\bar{x} \in \bar{X} := \arg\min_{x\in\mathbb{R}^n}\{f(x) + \psi(x)\}$ the accelerated Bregman proximal gradient iterates satisfy

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \le \frac{1}{\sum_{i=0}^k t_i} D_h(\bar{x}, x_0)$$

Proof: Similar to the argument for Bregman proximal gradient; use Lemma 25.7 and Fenchel duality. ■

25.3.6 Relative smoothness revisited

To accelerate as much as possible, choose $t_k > 0$, or equivalently, $\theta_k = \frac{t_k}{\sum_{i=0}^k t_i}$ as large as possible such that (25.8) holds.

Definition 25.10 ((L, γ) relative smoothness) *f is (L,γ)-smooth relative to h on Q if for all $x, y, z, \tilde{z} \in Q$ and $\theta \in [0,1]$*

$$D_f((1-\theta)x + \theta\tilde{z}, (1-\theta)x + \theta z) \le L\theta^\gamma D_h(\tilde{z}, z)$$

Remark. In the Euclidean case, the "anchor point" disappears as L-relative smoothness yields $(L, 2)$ relative smoothness.

So, how large may we push the stepsizes in acceleration?

Theorem 25.11 (Gutman-Peña 2018) *Suppose f is (L,γ) smooth relative to h on* $\mathrm{ri}(\mathrm{dom}(h)) \cap \mathrm{dom}(\psi)$ *for some $L > 0$ and $\gamma > 0$.*

Then the stepsizes t_k may be chosen such that the accelerated Bregman proximal gradient iterates satisfy

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \le \left(\frac{\gamma}{k+\gamma}\right) L \cdot D_h(\bar{X}, x_0)$$

This recovers the $O(1/k^2)$ rate when $h(x) = \frac{1}{2}\|x\|^2$ and f is L-smooth.

25.3.7 Implementation details

We wish to pick θ_k as large as possible such that (25.8) holds. To do this, we choose θ_k of the form

$$\theta_k = \frac{\gamma_k}{k + \gamma_k}$$

via backtracking on γ_k. If all $\gamma_k \geq \gamma > 0$ then we obtain

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \leq \left(\frac{\gamma}{k+\gamma}\right)^{\gamma} L \cdot D_h(\bar{X}, x_0)$$

Performing this with $\gamma = 2$ recovers the $O(1/k^2)$ rate, and this happens when $h(x) = \frac{1}{2}\|x\|^2$.

25.3.8 Conclusion

In this lecture, we analyzed Bregman proximal methods through Fenchel duality. The key observation is that this class of algorithms generate x_k, v_k, w_k such that:

$$f(x_{k+1}) + \psi(x_{k+1} \leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)$$

Related developments that were not discussed:

- Proximal subgradient method when f is non-differentiable
- Linear convergence via restarting
- Analogous results for conditional gradient

Current and future work:

- Saddle-point problems
- Stochastic first-order methods
- More computational experiments
- Role of γ in accelerated Bregman proximal methods

References

- Bauschke, Bolte, Teboulle (2017), "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications"
- Lu, Freund, Nesterov (2018), "Relatively smooth convex optimization by first-order methods, and applications"
- Hanzely, Richtarik, Xiao (2018), "Accelerated Bergman proximal gradient methods for relatively smooth convex optimization"
- Teboulle (2018), "A simplified view of first-order methods for optimization"
- Gutman and Peña (2018), "A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated schemes"