Proximal gradient method

http://bicmr.pku.edu.cn/~wenzw/opt-2018-fall.html

Acknowledgement: this slides is based on Prof. Lieven Vandenberghes lecture notes

Outline

Proximal mapping

the proximal mapping (prox-operator) of a convex function h is defined as

$$\mathrm{prox}_h(x) = \underset{u}{\mathrm{argmin}} \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

examples

- $h(x) = 0 : \mathrm{prox}_h(x) = x$
- $h(x) = I_C(x)$ (indicator function of C): prox_h is projection on C

$$\mathrm{prox}_h(x) = \underset{u \in C}{\mathrm{argmin}} \, \|u - x\|_2^2 = P_C(x)$$

- $h(x) = \|x\|_1$: prox_h is the 'soft-threshold' (shrinkage) operation

$$\mathrm{prox}_h(x)_i = \begin{cases} x_i - 1, & x_i \geq 1 \\ 0, & |x_i| \leq 1 \\ x_i + 1, & x_i \leq -1 \end{cases} = \mathrm{sgn}(x_i) \max(|x_i| - 1, 0)$$

Proximal gradient method

unconstrained optimization with objective split in two components

$$\min \quad f(x) = g(x) + h(x)$$

- g convex, differentiable, $\mathbf{dom}\ g = \mathbf{R}^n$
- h convex with inexpensive prox-operator (many examples in the lecture on "proximal mapping")

proximal gradient algorithm

$$x^{(k)} = \mathrm{prox}_{t_k h}\left(x^{(k-1)} - t_k \nabla g(x^{(k-1)})\right)$$

$t_k > 0$ is step size, constant or determined by line search

Interpretation

$$x^+ = \mathrm{prox}_{th}(x - t\nabla g(x))$$

from definition of proximal mapping:

$$\begin{aligned} x^+ &= \operatorname*{argmin}_u \left(h(u) + \frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2 \right) \\ &= \operatorname*{argmin}_u \left(h(u) + g(x) + \nabla g(x)^\top (u - x) + \frac{1}{2t}\|u - x\|_2^2 \right) \end{aligned}$$

x^+ minimizes $h(u)$ plus a simple quadratic local model of $g(u)$ around x

Examples

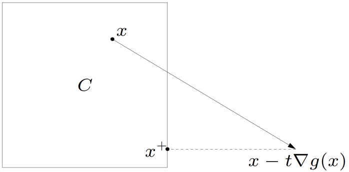
$$\min \quad g(x) + h(x)$$

gradient method: special case with $h(x) = 0$

$$x^+ = x - t\nabla g(x)$$

gradient projection method: special case with $h(x) = I_C(x)$

$$x^+ = P_C(x - t\nabla g(x))$$



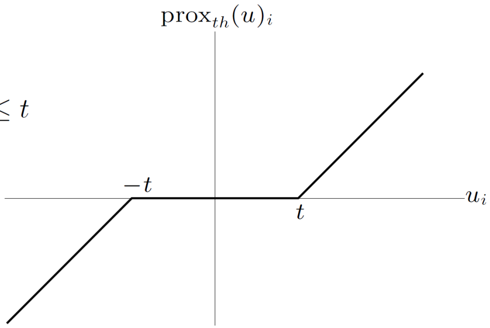
soft-thresholding: special case with $h(x) = \|x\|_1$

$$x^+ = \text{prox}_{th}(x - t\nabla g(x))$$

where $\text{prox}_{th}(u) = \text{sgn}(u)\max(|u| - t, 0)$

where

$$\text{prox}_{th}(u)_i = \begin{cases} u_i - t & u_i \geq t \\ 0 & -t \leq u_i \leq t \\ u_i + t & u_i \leq -t \end{cases}$$



Outline

Proximal mapping

if h is convex and closed (has a closed epigraph), then

$$prox_h(x) = \operatorname*{argmin}_u \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

exists and is unique for all x

- from optimality conditions of minimization in the definition:

$$\begin{aligned} u = \mathrm{prox}_h(x) \quad &\Leftrightarrow \quad x - u \in \partial h(u) \\ &\Leftrightarrow \quad h(z) \geq h(u) + (x - u)^\top (z - u) \quad \forall z \end{aligned}$$

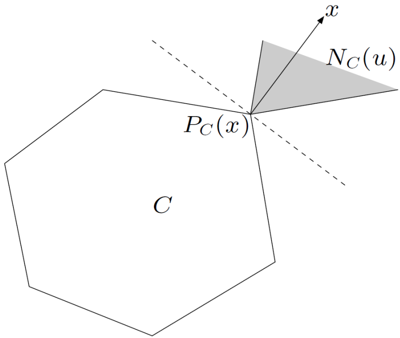
Projection on closed convex set

proximal mapping of indicator function I_C is Euclidean projection on C

$$\text{prox}_{I_C}(x) = \underset{u \in C}{\text{argmin}} \|u - x\|_2^2 = P_C(x)$$

subgradient characterization

$$u = P_C(x)$$
$$\Updownarrow$$
$$(x-u)^\top (z-u) \le 0 \quad \forall z \in C$$



we will see that proximal mappings have many properties of projections

Nonexpansiveness

if $u = \text{prox}_h(x), v = \text{prox}_h(y)$, then

$$(u-v)^\top (x-y) \geq \|u-v\|_2^2$$

prox_h is *firmly nonexpansive*, or *co-coercive* with constant 1

- follows from characterization of page 9 and monotonicity

$$x-u \in \partial h(u), y-v \in \partial h(v) \quad \Rightarrow \quad (x-u-y+v)^\top (u-v) \geq 0$$

- implies (from Cauchy-Schwarz inequality)

$$\|\text{prox}_h(x) - prox_h(y)\|_2 \leq \|x-y\|_2$$

proxh is *nonexpansive*, or *Lipschitz continuous* with constant 1

Outline

Convergence of proximal gradient method

to minimize $g + h$, choose $x^{(0)}$ and repeat

$$x^{(k)} = \mathrm{prox}_{t_k h}\left(x^{(k-1)} - t\nabla g(x^{(k-1)})\right), \quad k \geq 1$$

assumptions

- g convex with $\mathbf{dom}\, g = \mathbf{R}^n$; ∇g Lipschitz continuous with constant L:

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

- h is closed and convex (so that prox_{th} is well defined)
- optimal value f^* is finite and attained at x^* (not necessarily unique)

convergence result: $1/k$ rate convergence with fixed step size $t_k = 1/L$

Gradient map

$$G_t(x) = \frac{1}{t}(x - \mathrm{prox}_{th}(x - t\nabla g(x)))$$

$G_t(x)$ is the negative 'step' in the proximal gradient update

$$\begin{aligned} x^+ &= \mathrm{prox}_{th}(x - t\nabla g(x)) \\ &= x - tG_t(x) \end{aligned}$$

- $G_t(x)$ is not a gradient or subgradient of $f = g + h$
- from subgradient definition of prox-operator (page 9),

$$G_t(x) \in \partial g(x) + \partial h(x - tG_t(x))$$

- $G_t(x) = 0$ if and only if x minimizes $f(x) = g(x) + h(x)$

Consequences of Lipschitz assumption

recall upper bound (lecture on "gradient method") for convex g with Lipschitz continuous gradient

$$g(y) \leq g(x) + \nabla g(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y$$

- substitute $y = x - tG_t(x)$:

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^\top G_t(x) + \frac{t^2 L}{2}\|G_t(x)\|_2^2$$

- if $0 < t \leq 1/L$, then

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^\top G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \tag{1}$$

A global inequality

if the inequality (1) holds, then for all z,

$$f(x - tG_t(x)) \leq f(x) - G_t(x)^\top (x - z) + \frac{t}{2}\|G_t(x)\|_2^2 \qquad (2)$$

proof: (define $v = G_t(x) - \nabla g(x)$)

$$\begin{aligned}
f(x - tG_t(x)) &\leq g(x) - t\nabla g(x)^\top G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)) \\
&\leq g(z) - \nabla g(x)^\top (x - z) - t\nabla g(x)^\top G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + h(z) + v^\top (x - z - tG_t(x)) \\
&= g(z) + h(z) + G_t(x)^\top (x - z) - \frac{t}{2}\|G_t(x)\|_2^2
\end{aligned}$$

line 2 follows from convexity of g and h, and $v \in \partial h(x - tG_t(x))$

Progress in one iteration

$$x^+ = x - tG_t(x)$$

- inequality (2) with $z = x$ shows the algorithm is a descent method:

$$f(x^+) \leq f(x) - \frac{t}{2}\|G_t(x)\|_2^2$$

- inequality (2) with $z = x^*$

$$\begin{aligned} f(x^+) - f^* &\leq G_t(x)^\top (x - x^*) - \frac{t}{2}\|G_t(x)\|_2^2 \\ &= \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right) \end{aligned} \tag{3}$$

(hence, $\|x^+ - x^*\|_2^2 \leq \|x - x^*\|_2^2$, *i.e.*, distance to optimal set decreases)

Analysis for fixed step size

add inequalities (3) for $x = x^{(i-1)}, x^+ = x^{(i)}, t = t_i = 1/L$

$$\begin{aligned}
\sum_{i=1}^{k}(f(x^{(i)}) - f^*) &\le \frac{1}{2t}\sum_{i=1}^{k}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right) \\
&= \frac{1}{2t}\left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2\right) \\
&\le \frac{1}{2t}\|x^{(0)} - x^*\|_2^2
\end{aligned}$$

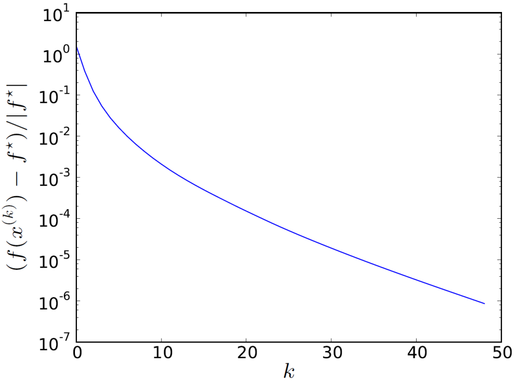
since $f(x^{(i)})$ is nonincreasing,

$$f(x^{(k)}) - f^* \le \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) \le \frac{1}{2kt}\|x^{(0)} - x^*\|_2^2$$

conclusion: reaches $f(x^{(k)}) - f^* \le \epsilon$ after $O(1/\epsilon)$ iterations

Quadratic program with box constraints

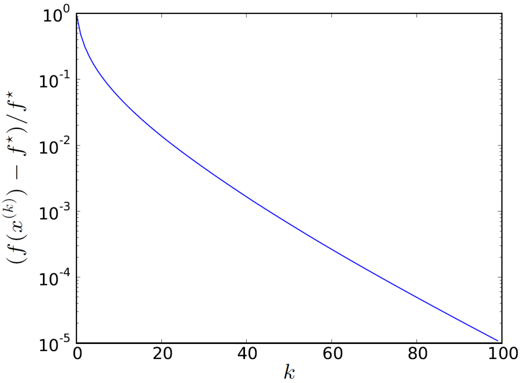
$$\begin{array}{ll} \min & (1/2)x^\top Ax + b^\top x \\ \text{s.t.} & 0 \preceq x \preceq 1 \end{array}$$



$n = 3000$; fixed step size $t = 1/\lambda_{\max}(A)$

1-norm regularized least-squares

$$\min \quad \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_1$$



randomly generated $A \in \mathbf{R}^{2000\times 1000}$; step $t_k = 1/L$ with $L = \lambda_{\max}(A^\top A)$

Outline

Line search

- the analysis for fixed step size starts with the inequality (1)

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^\top G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2$$

  this inequality is known to hold for $0 < t \leq 1/L$

- if L is not known, we can satisfy (1) by a backtracking line search: start at some $t := \hat{t} > 0$ and backtrack ($t := \beta t$) until (1) holds

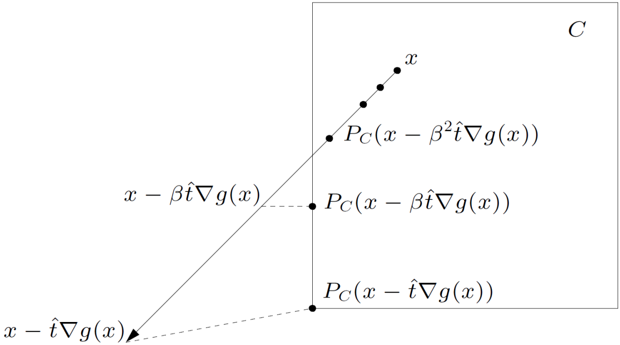
- step size t selected by the line search satisfies $t \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

- requires one evaluation of g and proxth per line search iteration

several other types of line search work

example: line search for projected gradient method

$$x^+ = P_C(x - t\nabla g(x)) = x - tG_t(x)$$



backtrack until $x - tG_t(x)$ satisfies 'sufficient decrease' inequality (1)

Analysis with line search

from page 17, if (1) holds in iteration i, then $f(x^{(i)}) < f(x^{(i-1)})$ and

$$\begin{aligned} f(x^{(i)}) - f^* &\leq \frac{1}{2t_i}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right) \\ &\leq \frac{1}{2t_{\min}}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right) \end{aligned}$$

- adding inequalities for $i = 1$ to $i = k$ gives

$$\sum_{i=1}^{k} f(x^{(i)}) - f^* \leq \frac{1}{2t_{\min}}\left(\|x^{(0)} - x^*\|_2^2\right)$$

- since $f(x^{(i)})$ is nonincreasing, obtain similar $1/k$ bound as for fixed t_i:

$$f(x^{(k)}) - f^* \leq \frac{1}{2kt_{\min}}\left(\|x^{(0)} - x^*\|_2^2\right)$$

Another Backtracking Line Search Scheme

Let $h(x) = \mu\|x\|_1$, consider

$$\min \quad f(x) = g(x) + h(x)$$

- Compute $\bar{x}^k = \mathrm{prox}_{\tau^k h}(x - \tau^k \nabla g(x^k))$, $d^k = \bar{x}^k - x^k$, $\Delta^k = \nabla g(x^k)^\top d^k + (h(\bar{x}^k) - h(x^k))$
- Choose $\alpha_\rho > 0$ and $\sigma, \rho \in (0,1)$. Choose $\alpha_k = \alpha_\rho \rho^h$ such that h is the smallest integer that satisfies

$$f(x^k + \alpha_\rho \rho^h d^k) \leq f(x^k) + \sigma \alpha_\rho \rho^h \Delta^k$$

- Set $x^{k+1} = x^k + \alpha^k d^k$.

Improving Backtracking Line Search Scheme

Choosing τ^k: Barzilai-Borwein method

- $s^{k-1} = x^k - x^{k-1}$ and $y^{k-1} = \nabla g(x^k) - \nabla g(x^{k-1})$
- $\tau^{k,BB1} = \frac{(s^{k-1})^\top s^{k-1}}{(s^{k-1})^\top y^{k-1}}$ or $\tau^{k,BB2} = \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}}$.

Choosing α^k: Nonmontone Armijo-like Line search

$$f(x^k + \alpha^k d^k) \leq C^k + \sigma \alpha^k \Delta^k$$

- $C^k = (\eta Q^{k-1} C^{k-1} + f(x^k))/Q^k$, $Q^k = \eta Q^{k-1} + 1$, $C^0 = f(x^0)$ and $Q^0 = 1$ (Zhang and Hager)

The continuation strategy

- Choose $\mu^0 > \mu^1 > \ldots > \mu^l = \mu$
- Solve $z(\mu^i) = \arg\min_x \; g(x) + \mu^i \|x\|_1$ starting from $z(\mu^{i-1})$

Outline

Inertial Proximal Algorithm

Consider the problem

$$\min \quad f(x) = g(x) + h(x),$$

where $\nabla g(x)$ is L-Lipschitz

- Choose x^0, set $x^{-1} = x^0$, choose $\beta \in [0, 1]$, set $\alpha < 2(1-\beta)/L$, the inertial proximal algorithm computes:

$$x^{k+1} = \text{prox}_{\alpha h}(x^k - \alpha \nabla g(x^k) + \beta(x^k - x^{k-1}))$$

- The term $\beta(x^k - x^{k-1})$: inertial term
- For $h(x) = 0$, the scheme is referred as the Heavy-ball method
- Ref: Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock, iPiano: Inertial Proximal Algorithm for Nonconvex Optimization, SIAM J. IMAGING SCIENCES, Vol. 7, No. 2

Outline

Conditional Gradient (CndG) Method: motivation

Let X be a compact set. Consider

$$\min_{x \in X} \quad f(x).$$

- Proximal gradient method:

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|y - x_k\|^2 \right\}.$$

  It is equivalent to the projected gradient method

$$x_{k+1} = \mathcal{P}_X(x_k - \alpha_k \nabla f(x_k)).$$

- Difficulty: the computational cost of the projection $\mathcal{P}_X(\cdot)$ may be expensive

Conditional Gradient (CndG) or Frank-Wolfe Method

- Given $y_0 = x_0$ and $\alpha_k \in (0, 1]$, the CndG method takes

$$\begin{aligned} x_k &= \operatorname*{argmin}_{x \in X} \langle \nabla f(y_{k-1}), x \rangle, \\ y_k &= (1 - \alpha_k) y_{k-1} + \alpha_k x_k \end{aligned}$$

- diminishing step sizes:

$$\alpha_k = \frac{2}{k+1}$$

or by exact line search

$$\alpha_k = \arg \min_{\alpha \in [0,1]} f((1 - \alpha) y_{k-1} + \alpha x_k)$$

Examples

考虑带某一范数$\|\cdot\|$ 约束的凸优化问题，

$$\min_x f(x) \quad \text{s.t.} \quad \|x\| \le t.$$

用条件梯度法求解该问题时，需要计算子问题，

$$\begin{aligned} x_k &\in \operatorname*{argmin}_{\|x\|\le t} \langle \nabla f(y_{k-1}), x\rangle \\ &= -t\cdot\left(\arg\max_{\|x\|\le 1} \langle \nabla f(y_{k-1}), x\rangle\right) \\ &= -t\cdot \partial\|\nabla f(y_{k-1})\|_*. \end{aligned} \tag{4}$$

其中$\|z\|_* = \sup\{z^T x, \|x\| \le 1\}$ 是$\|\cdot\|$ 的对偶范数。注意到(4) 条件梯度法的子问题相当于计算一个对偶范数的次梯度。如果计算$\|\cdot\|$ 范数的次梯度比计算在约束集合$X = \{x \in \mathbb{R}^n : \|x\| \le t\}$ 上的投影要简单，条件梯度法比投影梯度法效率更高。

Examples: ℓ_1 范数约束问题

由于ℓ_1 范数的对偶范数是ℓ_∞ 范数，因此用条件梯度法求解该问题时子问题为，

$$x_k \in -t \cdot \partial \|\nabla f(y_{k-1})\|_\infty.$$

考虑到ℓ_∞ 范数的次梯度为$\partial\|x\|_\infty = \{v : \langle v, x\rangle = \|x\|_\infty, \|v\|_1 \le 1\}$，子问题等价于，

$$\begin{aligned} i_k &\in \underset{i=1,\dots,n}{\operatorname{argmax}} |\nabla_i f(y_{k-1})| \\ x_k &= -t \cdot \mathrm{sgn}\left[\nabla_{i_k} f(y_{k-1})\right] \cdot e_{i_k}. \end{aligned}$$

其中$\nabla_i f(y_{k-1})$ 表示向量$\nabla f(y_{k-1})$ 的第i 个元素，e_i 表示第i 个元素为1 的单位向量。可以看到计算$\|\cdot\|_\infty$ 的次梯度和计算集合$X := \{x \in \mathbb{R}^n : \|x\|_1 \le t\}$ 上的投影都需要$\mathcal{O}(n)$ 的计算复杂度，但是条件梯度法子问题计算明显要更简单直接。

Examples: ℓ_p 范数约束问题, $1 \le p \le \infty$

由于ℓ_p 范数的对偶范数是ℓ_q 范数，其中$1/p + 1/q = 1$，因此用条件梯度法求解该问题时子问题为，

$$x_k \in -t \cdot \partial \|\nabla f(y_{k-1})\|_q.$$

注意到ℓ_q 范数的次梯度为$\partial \|x\|_q = \{v : \langle v, x\rangle = \|x\|_q, \|v\|_p \le 1\}$，子问题等价于，

$$x_k^{(i)} = -\beta \cdot \mathrm{sgn}\left[\nabla_i f(y_{k-1})\right] \cdot |\nabla_i f(y_{k-1})|^{p/q}.$$

其中β 是使得$\|x_k\|_q = t$ 的归一化常数。可以看到，除过$p = 1, 2, \infty$ 这些特殊情形，条件梯度法的子问题计算复杂度比直接计算点在集合$X = \{x \in \mathbb{R}^n : \|x\|_p \le t\}$ 上的投影要简单，后者投影计算需要单独解一个优化问题。

Example: 矩阵核范数约束优化问题

矩阵核范数$\|\cdot\|_*$的对偶范数是其谱范数$\|\cdot\|_2$:

$$\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X), \qquad \|X\|_2 = \max_{i=1,\dots,\min\{m,n\}} \sigma_i(X).$$

因此条件梯度法的子问题为$X_k \in -t \cdot \partial\|\nabla f(Y_{k-1})\|_2$. 对矩阵范数的次梯度: $\partial\|X\| = \{Y : \langle Y, X\rangle = \|X\|, \|Y\|_* \le 1\}$，设$u, v$分别是矩阵$\nabla f(Y_{k-1})$最大奇异值对应的左、右奇异向量，注意到，

$$\langle uv^T, \nabla f(Y_{k-1})\rangle = u^T \nabla f(Y_{k-1}) v = \sigma_{\max}(\nabla f(Y_{k-1})) = \|\nabla f(Y_{k-1})\|_2.$$

且$\|uv^T\|_* = 1$，因此矩阵$uv^T \in \partial\|\nabla f(Y_{k-1})\|_2$。则条件梯度法子问题等价于，

$$X_k \in -t \cdot uv^T. \tag{5}$$

可以看到，条件梯度法计算子问题时只需要计算矩阵最大的奇异值对应的左、右奇异向量。如果采用投影梯度法，其子问题是计算X到集合$\{X \in \mathbb{R}^{m\times n} : \|X\|_* \le t\}$的投影，需要对矩阵做全奇异值分解，计算量比条件梯度法复杂很多。

Convergence: Lemma

令$\gamma_t \in (0,1]$，$t = 1, 2, \ldots$，构造序列

$$\Gamma_t = \begin{cases} 1 & t = 1 \\ (1-\gamma_t)\Gamma_{t-1} & t \geq 2 \end{cases}.$$

如果序列$\{\Delta_t\}_{t\geq 0}$ 满足

$$\Delta_t \leq (1-\gamma_t)\Delta_{t-1} + B_t \quad t = 1, 2, \ldots.$$

则对任意的k 我们对Δ_k 有估计

$$\Delta_k \leq \Gamma_k(1-\gamma_1)\Delta_0 + \Gamma_k \sum_{t=1}^{k} \frac{B_t}{\Gamma_t}.$$

Convergence

Let $f(x)$ is convex, $\nabla f(x)$ is L-Lipschitz, $D_X = \sup_{x,y\in X} \|x-y\|$. Then

$$f(y_k) - f(x^*) \;\le\; \frac{2L}{k(k+1)} \sum_{i=1}^{k} \|x_i - y_{i-1}\|^2 \;\le\; \frac{2L}{k+1} D_X^2.$$

Proof: 令$\gamma_k = \frac{2}{k+1}$，记$\bar{y}_k = (1-\gamma_k)y_{k-1} + \gamma_k x_k$，则不管

$$\alpha_k = \frac{2}{k+1} \quad 或 \quad \alpha_k = \underset{\alpha\in[0,1]}{\operatorname{argmin}} f((1-\alpha)y_{k-1} + \alpha x_k).$$

对$y_k = (1-\alpha_k)y_{k-1} + \alpha_k x_k$，我们都有$f(y_k) \le f(\bar{y}_k)$。注意到$\bar{y}_k - y_{k-1} = \gamma_k(x_k - y_{k-1})$，由$f(x) \in C_L^{1,1}(X)$有

$$f(y_k) \le f(\bar{y}_k) \;\le\; f(y_{k-1}) + \langle \nabla f(y_{k-1}), \bar{y}_k - y_{k-1}\rangle + \frac{L}{2}\|\bar{y}_k - y_{k-1}\|^2 \tag{6}$$

$$\le\; (1-\gamma_k)[f(y_{k-1}) + \gamma_k[f(y_{k-1}) + \langle \nabla f(y_{k-1}), x - y_{k-1}\rangle] + \frac{L\gamma_k^2}{2}\|x_k - y_{k-1}\|^2 \tag{7}$$

$$\le\; (1-\gamma_k)f(y_{k-1}) + \gamma_k f(x) + \frac{L\gamma_k^2}{2}\|x_k - y_{k-1}\|^2, \qquad 对任意\ x\in X. \tag{8}$$

其中不等式(7) 是因为$x_k \in \min_{x\in X}\langle \nabla f(y_{k-1}), x\rangle$，由最优性条件我们可以得到对任意$x \in X$ 有$\langle x - x_k, \nabla f(y_{k-1})\rangle \geq 0$。将不等式(8) 稍做变换，对任意$x \in X$，

$$f(y_k) - f(x) \leq (1-\gamma_k)[f(y_{k-1}) - f(x)] + \frac{L}{2}\gamma_k^2\|x_k - y_{k-1}\|^2. \tag{9}$$

由引理可知，

$$f(y_k) - f(x) \leq \Gamma_k(1-\gamma_1)[f(y_0) - f(x)] + \frac{\Gamma_k L}{2}\sum_{i=1}^{k}\frac{\gamma_i^2}{\Gamma_i}\|x_i - y_{i-1}\|^2.$$

由$\gamma_k = \frac{2}{k+1}$，$\gamma_1 = 1$ 得到$\Gamma_k = \frac{2}{k(k+1)}$，我们可以得到收敛性不等式，

$$f(y_k) - f^* \ \leq \ \frac{2L}{k(k+1)}\sum_{i=1}^{k}\|x_i - y_{i-1}\|^2 \ \leq \ \frac{2L}{k+1}D_X^2.$$

令$\frac{2L}{k+1}D_X^2 \leq \epsilon$，可以得到分析复杂度结论。

References

convergence analysis of proximal gradient method

A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences (2009)

A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery*, in: Y. Eldar and D. Palomar (Eds.), *Convex Optimization in Signal Processing and Communications* (2009)