Proximal Split Method

参考讲义：

1. Neal Parikh,Stephen Boyd.Proximal Algorithms.
2. Yuxin Chen.Proximal gradient methods
3. Patrick L.Proximal Splitting Methods in Signal Processing.

Strong Convex

A differentiable function f is strongly convex if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} ||y - x||^2$$

for some $\mu > 0$ and all x, y.

Proposition *The following conditions are all equivalent to the condition that a differentiable function f is strongly-convex with constant $\mu > 0$.*

$(i)\ f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} ||y - x||^2, \ \forall x, y.$

$(ii)\ g(x) = f(x) - \frac{\mu}{2} ||x||^2 \text{ is convex}, \ \forall x.$

$(iii)\ (\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu ||x - y||^2, \ \forall x, y.$

$(iv)\ f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y) - \frac{\alpha(1 - \alpha)\mu}{2} ||x - y||^2, \ \alpha \in [0, 1].$

http://xingyuzhou.org/blog/notes/strong-convexity

Proposition *For a continuously differentiable function f, the following conditions are all implied by strong convexity (SC) condition.*

$(a)\ \frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*),\ \forall x.$

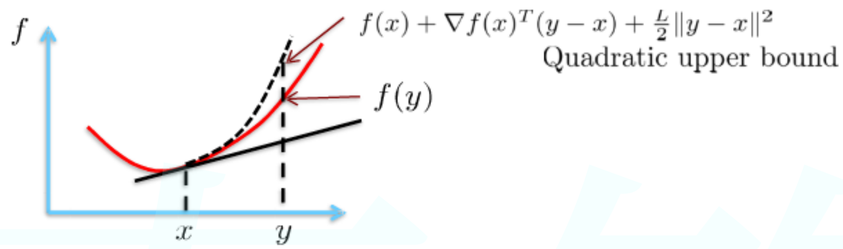
$(b)\ \|\nabla f(x) - \nabla f(y)\| \geq \mu\|x - y\|\ \forall x, y.$

$(c)\ f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2,\ \forall x, y.$

$(d)\ (\nabla f(x) - \nabla f(y)^T(x - y) \leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2,\ \forall x, y.$

f is convex and has Lipschitz continuous gradient iff one of the following holds:

$0 \leq f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2}\|y - x\|^2, \ \forall x, y \in \mathbb{R}^n$.



Proximal operator

The *proximal operator* $\mathbf{prox}_f : \mathbf{R}^n \to \mathbf{R}^n$ of f is defined by

$$\mathbf{prox}_f(v) = \operatorname*{argmin}_x \left(f(x) + (1/2)\|x - v\|_2^2 \right), \tag{1.1}$$

where $\|\cdot\|_2$ is the usual Euclidean norm. The function minimized on the righthand side is strongly convex and not everywhere infinite, so it has a unique minimizer for every $v \in \mathbf{R}^n$ (even when $\mathbf{dom}\, f \subsetneq \mathbf{R}^n$).

We will often encounter the proximal operator of the scaled function λf, where $\lambda > 0$, which can be expressed as

$$\mathbf{prox}_{\lambda f}(v) = \operatorname*{argmin}_x \left(f(x) + (1/2\lambda)\|x - v\|_2^2 \right). \tag{1.2}$$

This is also called the proximal operator of f with parameter λ. (To keep notation light, we write $(1/2\lambda)$ rather than $(1/(2\lambda))$.)

$\mathbf{prox}_f(v)$ is sometimes called a *proximal point* of v

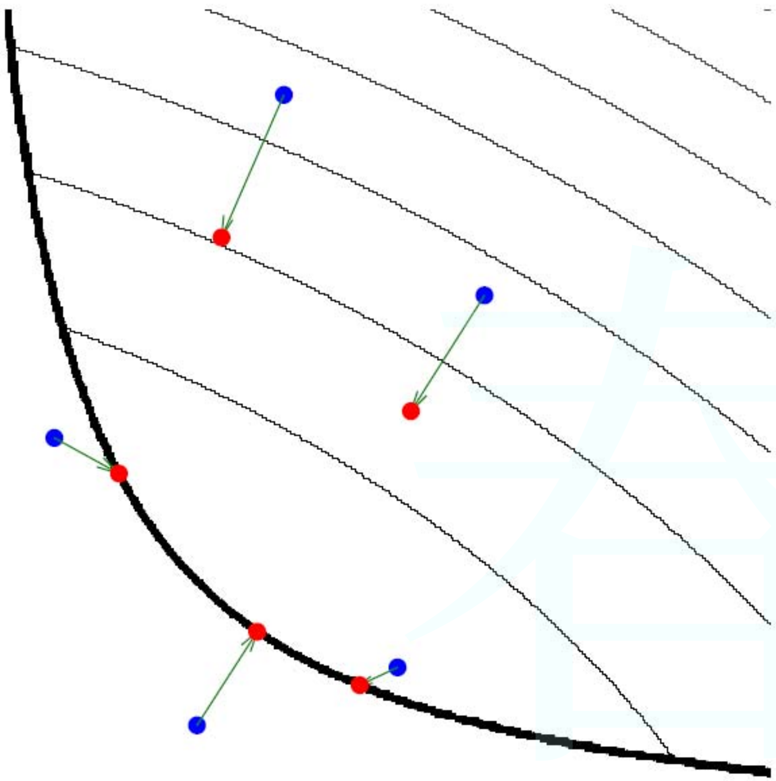
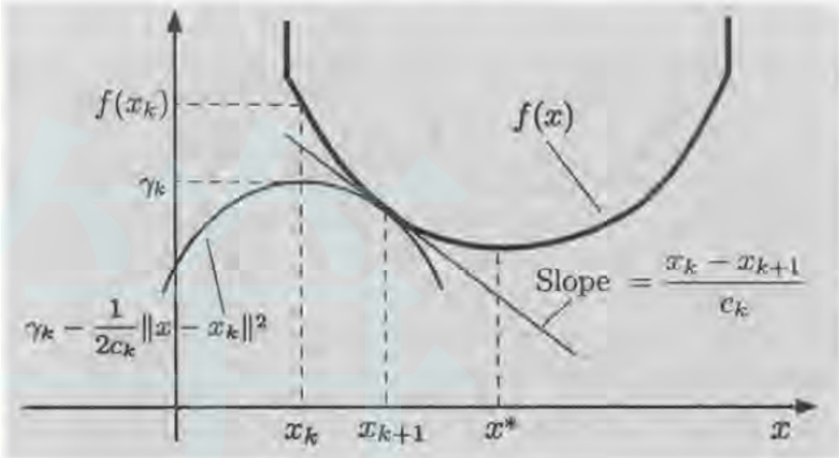


Figure 1.1: Evaluating a proximal operator at various points.



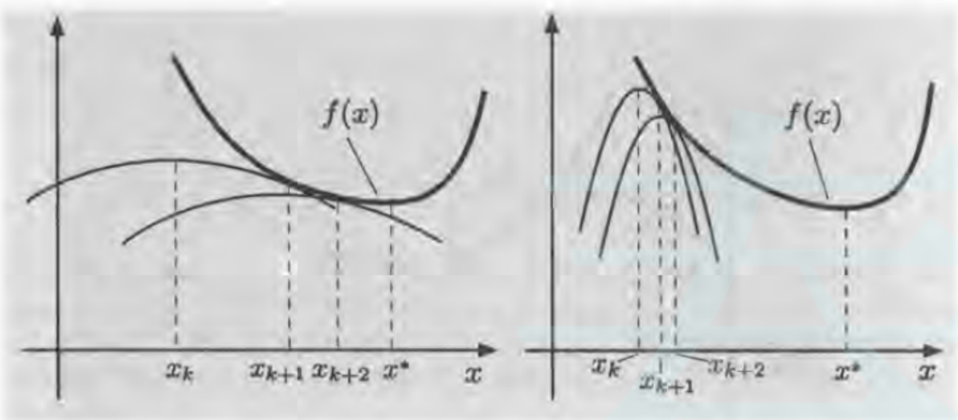


Figure 5.1.2. Illustration of the role of the parameter c_k in the convergence process of the proximal algorithm. In the figure on the left, c_k is large, the graph of the quadratic term is "blunt," and the method makes fast progress toward the optimal solution. In the figure on the right, c_k is small, the graph of the quadratic term is "pointed," and the method makes slow progress.

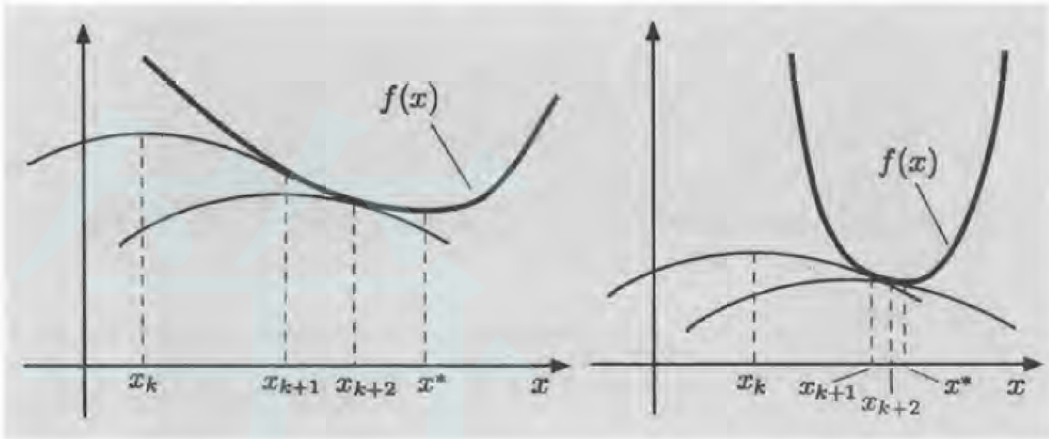


Figure 5.1.3. Illustration of the convergence rate of the proximal algorithm and the effect of the growth properties of f near the optimal solution set. In the figure on the left, f grows slowly and the convergence is slow. In the figure on the right, f grows fast and the convergence is fast.

In general, the problem we wish to solve is

$$\begin{array}{ll} \text{minimize} & f(x) + (1/2\lambda)\|x - v\|_2^2 \\ \text{subject to} & x \in \mathcal{C}, \end{array} \tag{6.1}$$

with variable $x \in \mathbf{R}^n$, where $\mathcal{C} = \mathbf{dom}\, f$ (which may be all of $\mathbf{R}^n$, in which case the problem is unconstrained).

1. Quadratic functions

   If $f(x) = (1/2)x^T Ax + b^T x + c$, with $A \in \mathbf{S}^n_+$, then

$$\mathbf{prox}_{\lambda f}(v) = (I + \lambda A)^{-1}(v - \lambda b).$$

2. affine

   if $f(x) = b^T x + c$, *i.e.*, if f is affine, then $\mathbf{prox}_{\lambda f}(v) = v - \lambda b$.

3. constant function

$$\mathbf{prox}_{\lambda f}(v) = v,$$

4. 2-norm

$$\mathbf{prox}_{\lambda f}(v) = \left(\frac{1}{1 + \lambda}\right) v,$$

Suppose H is the sum of a diagonal and a rank one matrix, *i.e.*,

$$H = D + zz^T,$$

where $D \in \mathbf{R}^{n\times n}$ is diagonal. By the matrix inversion lemma,

$$H^{-1} = D^{-1} - \frac{D^{-1}zz^T D^{-1}}{1 + z^T D^{-1} z},$$

5. log x

$$\mathbf{prox}_{\lambda f}(v) = \frac{v + \sqrt{v^2 + 4\lambda}}{2}.$$

6.nonsmooth $f(x) = |x|$:

$$\mathbf{prox}_{\lambda f}(v) = \begin{cases} v - \lambda & v \geq \lambda \\ 0 & |v| \leq \lambda \\ v + \lambda & v \leq -\lambda. \end{cases}$$

This operation is called *soft thresholding*

The projection problem

7.Polyhedra

$$\mathcal{C} = \{x \in \mathbf{R}^n \mid Ax = b,\ Cx \le d\},$$

where $A \in \mathbf{R}^{m\times n}$ and $C = \mathbf{R}^{p\times n}$. The projection problem is

$$\begin{array}{ll} \text{minimize} & (1/2)\|x - v\|_2^2 \\ \text{subject to} & Ax = b, \quad Cx \le d. \end{array}$$

The dual function of (6.4) is the concave quadratic

$$g(\nu, \eta) = -\frac{1}{2}\left\| \begin{bmatrix} A \\ C \end{bmatrix}^T \begin{bmatrix} \nu \\ \eta \end{bmatrix} \right\|_2^2 + \left(\begin{bmatrix} A \\ C \end{bmatrix} v - \begin{bmatrix} b \\ d \end{bmatrix} \right)^T \begin{bmatrix} \nu \\ \eta \end{bmatrix},$$

where $\nu \in \mathbf{R}^m$ and $\eta \in \mathbf{R}^p$ are dual variables. The dual problem is

$$\begin{array}{ll} \text{maximize} & g(\nu, \eta) \\ \text{subject to} & \eta \ge 0. \end{array}$$

$$x^\star = v - A^T\lambda^\star - C^T\nu^\star,$$

8.Affine set

$$\mathcal{C} = \{x \in \mathbf{R}^n \mid Ax = b\},$$

$$\Pi_{\mathcal{C}}(v) = v - A^\dagger(Av - b),$$

$$\Pi_{\mathcal{C}}(v) = v - A^T(AA^T)^{-1}(Av - b).$$

9. hyperplane $\mathcal{C} = \{x \mid a^T x = b\}$

$$\Pi_{\mathcal{C}}(v) = v + \left(\frac{b - a^T v}{\|a\|_2^2}\right) a.$$

10.Halfspace

If $\mathcal{C} = \{x \mid a^T x \le b\}$ is a halfspace, then

$$\Pi_{\mathcal{C}}(v) = v - \frac{(a^T v - b)_+}{\|a\|_2^2} a,$$

11. Box

Projection onto a *box* or *hyper-rectangle* $\mathcal{C} = \{x \mid l \leq x \leq u\}$ also takes a simple form:

$$(\Pi_{\mathcal{C}}(v))_k = \begin{cases} l_k & v_k \leq l_k \\ v_k & l_k \leq v_k \leq u_k \\ u_k & v_k \geq u_k, \end{cases}$$

.....太多了

Algorithms:

1. Proximal Point Method
2. Proximal Gradient Method
3. Accelerated Method

Proximal Point Method

The *proximal minimization algorithm*, also called *proximal iteration* or the *proximal point algorithm*, is

$$x^{k+1} := \mathbf{prox}_{\lambda f}(x^k), \qquad \longleftarrow \qquad \mathbf{prox}_{\lambda f}(v) = \operatorname*{argmin}_x \left(f(x) + (1/2\lambda)\|x - v\|_2^2 \right)$$

we often refer to gradient steps as *forward steps* and proximal steps as *backward steps*.

Proximal Gradient Descent

unconstrained optimization with objective split in two components

$$\text{minimize} \quad f(x) = g(x) + h(x)$$

- g convex, differentiable, $\mathbf{dom}\, g = \mathbf{R}^n$
- h convex with inexpensive prox-operator (many examples in lecture 8)

Proximal gradient algorithm

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

- $t_k > 0$ is step size, constant or determined by line search
- can start at infeasible $x^{(0)}$ (however $x^{(k)} \in \mathbf{dom}\, f = \mathbf{dom}\, h$ for $k \geq 1$)

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

from definition of proximal mapping:

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

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

Proximal gradient method

if L is not known (usually the case), can use the following line search:

---

given x^k, λ^{k-1}, and parameter $\beta \in (0,1)$.

Let $\lambda := \lambda^{k-1}$.

repeat

1. Let $z := \mathbf{prox}_{\lambda g}(x^k - \lambda \nabla f(x^k))$.
2. **break if** $f(z) \leq \hat{f}_\lambda(z, x^k)$.
3. Update $\lambda := \beta\lambda$.

return $\lambda^k := \lambda$, $x^{k+1} := z$.

For an upper bound of f, consider the function $\hat{f}_\lambda$ given by

$$\hat{f}_\lambda(x,y) = f(y) + \nabla f(y)^T(x-y) + (1/2\lambda)\|x-y\|_2^2,$$

Accelerated Method

$$
\begin{aligned}
y^{k+1} &:= x^k + \omega^k (x^k - x^{k-1}) \\
x^{k+1} &:= \mathbf{prox}_{\lambda^k g}(y^{k+1} - \lambda^k \nabla f(y^{k+1}))
\end{aligned}
$$

Where $\omega^k = \dfrac{k}{k+3}$.

given y^k, λ^{k-1}, and parameter $\beta \in (0,1)$.

Let $\lambda := \lambda^{k-1}$.

repeat

1. Let $z := \mathbf{prox}_{\lambda g}(y^k - \lambda \nabla f(y^k))$.
2. **break if** $f(z) \leq \hat{f}_\lambda(z, y^k)$.
3. Update $\lambda := \beta\lambda$.

return $\lambda^k := \lambda$, $x^{k+1} := z$.

Application

The lasso problem is

$$\text{minimize} \quad (1/2)\|Ax - b\|_2^2 + \gamma\|x\|_1$$

Consider the splitting

$$f(x) = (1/2)\|Ax - b\|_2^2, \qquad g(x) = \gamma\|x\|_1,$$

with gradient and proximal operator

$$\nabla f(x) = A^T(Ax - b), \qquad \mathbf{prox}_{\gamma g}(x) = S_\gamma(x),$$

where S_λ is the soft-thresholding operator