10. Proximal point method

- proximal point method
- augmented Lagrangian method
- Moreau-Yosida smoothing
Proximal point method

a ‘conceptual’ algorithm for minimizing a closed convex function $f$:

$$x^{(k)} = \text{prox}_{t_k f}(x^{(k-1)})$$

$$= \arg\min_u \left( f(u) + \frac{1}{2t_k} \|u - x^{(k-1)}\|^2_2 \right)$$

- can be viewed as proximal gradient method (page 6-2) with $g(x) = 0$
- of interest if prox evaluations are much easier than minimizing $f$ directly
- a practical algorithm if inexact prox evaluations are used
- step size $t_k > 0$ affects number of iterations, cost of prox evaluations

basis of the augmented Lagrangian method
Convergence

assumptions

• $f$ is closed and convex (hence, $\text{prox}_{t_f}(x)$ is uniquely defined for all $x$)
• optimal value $f^*$ is finite and attained at $x^*$

result

$$f(x^{(k)}) - f^* \leq \left\| x^{(0)} - x^* \right\|^2_k \sum_{i=1}^{2} t_i$$

for $k \geq 1$

• implies convergence if $\sum_i t_i \to \infty$
• rate is $1/k$ if $t_i$ is fixed or variable but bounded away from zero
• $t_i$ is arbitrary; however cost of prox evaluations will depend on $t_i$
proof: apply analysis of proximal gradient method (lect. 6) with $g(x) = 0$

- since $g$ is zero, inequality (1) on page 6-12 holds for any $t > 0$

- from page 6-14, $f(x^{(i)})$ is nonincreasing and

$$t_i \left( f(x^{(i)}) - f^* \right) \leq \frac{1}{2} \left( \|x^{(i)} - x^*\|^2 - \|x^{(i-1)} - x^*\|^2 \right)$$

- combine inequalities for $i = 1$ to $i = k$ to get

$$\left( \sum_{i=1}^{k} t_i \right) \left( f(x^{(k)}) - f^* \right) \leq \sum_{i=1}^{k} t_i \left( f(x^{(i)}) - f^* \right) \leq \frac{1}{2} \|x^{(0)} - x^*\|^2$$
Accelerated proximal point algorithms

FISTA (take $g(x) = 0$ on p.7-8): choose $x^{(0)} = x^{(-1)}$ and for $k \geq 1$

$$x^{(k)} = \text{prox}_{t_k f} \left( x^{(k-1)} + \theta_k \frac{1 - \theta_{k-1}}{\theta_{k-1}} (x^{(k-1)} - x^{(k-2)}) \right)$$

Nesterov’s 2nd method (p. 7-23): choose $x^{(0)} = v^{(0)}$ and for $k \geq 1$

$$v^{(k)} = \text{prox}_{(t_k/\theta_k) f} (v^{(k-1)}), \quad x^{(k)} = (1 - \theta_k) x^{(k-1)} + \theta_k v^{(k)}$$

possible choices of parameters

- fixed steps: $t_k = t$ and $\theta_k = 2/(k + 1)$
- variable steps: choose any $t_k > 0$, $\theta_1 = 1$, and for $k > 1$, solve $\theta_k$ from

$$\frac{(1 - \theta_k) t_k}{\theta_k^2} = \frac{t_{k-1}}{\theta_{k-1}^2}$$
Convergence

assumptions

• $f$ is closed and convex (hence, $\text{prox}_t f(x)$ is uniquely defined for all $x$)
• optimal value $f^*$ is finite and attained at $x^*$

result

\[
f(x^{(k)}) - f^* \leq \frac{2 \|x^{(0)} - x^*\|^2}{\left(2\sqrt{t_1} + \sum_{i=2}^{k} \sqrt{t_i}\right)^2} \quad \text{for } k \geq 1
\]

• implies convergence if $\sum_i \sqrt{t_i} \to \infty$

• rate is $1/k^2$ if $t_i$ is fixed or variable but bounded away from zero
proof: follows from analysis in lecture 7 with \( g(x) = 0 \)

- since \( g \) is zero, first inequalities on p. 7-15 and p. 7-25 hold for any \( t > 0 \)
- therefore the conclusion on p. 7-16 and p. 7-26 holds:

\[
f(x^{(k)}) - f^* \leq \frac{\theta_k^2}{2t_k} \|x^{(0)} - x^*\|_2^2
\]

- for fixed step size \( t_k = t \), \( \theta_k = 2/(k + 1) \),

\[
\frac{\theta_k^2}{2t_k} = \frac{2}{(k + 1)^2 t}
\]

- for variable step size, we proved on page 7-19 that

\[
\frac{\theta_k^2}{2t_k} \leq \frac{2}{(2\sqrt{t_1} + \sum_{i=2}^{k} \sqrt{t_i})^2}
\]
Outline

• proximal point method

• **augmented Lagrangian method**

• Moreau-Yosida smoothing
Standard problem form

minimize \( f(x) + g(Ax) \)

- \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g : \mathbb{R}^m \to \mathbb{R} \) are closed convex functions; \( A \in \mathbb{R}^{m \times n} \)
- equivalent formulation with auxiliary variable \( y \):

\[
\begin{align*}
\text{minimize} & \quad f(x) + g(y) \\
\text{subject to} & \quad Ax = y
\end{align*}
\]

examples

- \( g \) is indicator function of \( \{b\} \): minimize \( f(x) \) subject to \( Ax = b \)
- \( g \) is indicator function of \( C \): minimize \( f(x) \) subject to \( Ax \in C \)
- \( g(y) = \|y - b\| : \text{minimize } f(x) + \|Ax - b\| \)
Dual problem

Lagrangian (of reformulated problem)

\[ L(x, y, z) = f(x) + g(y) + z^T(Ax - y) \]

dual problem

\[
\text{maximize } \inf_{x,y} L(x, y, z) = -f^*(-A^Tz) - g^*(z)
\]

optimality conditions: \(x, y, z\) are optimal if

- \(x, y\) are feasible: \(x \in \text{dom } f, y \in \text{dom } g\), and \(Ax = y\)
- \(x\) and \(y\) minimize \(L(x, y, z)\): \(-A^Tz \in \partial f(x)\) and \(z \in \partial g(y)\)

augmented Lagrangian method: proximal point method applied to dual
Proximal mapping of dual function

proximal mapping of \( h(z) = f^*(-A^Tz) + g^*(z) \) is defined as

\[
\text{prox}_{th}(z) = \arg\min_u \left( f^*(-A^Tu) + g^*(u) + \frac{1}{2t} \|u - z\|^2_2 \right)
\]

**dual expression:** \( \text{prox}_{th}(z) = z + t(A\hat{x} - \hat{y}) \) where

\[
(\hat{x}, \hat{y}) = \arg\min_{x,y} \left( f(x) + g(y) + z^T(Ax - y) + \frac{t}{2} \|Ax - y\|^2_2 \right)
\]

\( \hat{x}, \hat{y} \) minimize augmented Lagrangian (Lagrangian + quadratic penalty)
proof

• write augmented Lagrangian minimization as

\[
\begin{align*}
\text{minimize (over } x, y, w) & \quad f(x) + g(y) + \frac{t}{2}\|w\|^2 \\
\text{subject to} & \quad Ax - y + z/t = w
\end{align*}
\]

• optimality conditions ($u$ is multiplier for equality):

\[
\begin{align*}
Ax - y + \frac{1}{t}z &= w, \quad -A^T u \in \partial f(x), \quad u \in \partial g(y), \quad tw = u
\end{align*}
\]

• eliminating $x, y, w$ gives $u = z + t(Ax - y)$ and

\[
0 \in -A \partial f^*(-A^T u) + \partial g^*(u) + \frac{1}{t}(u - z)
\]

this is the optimality condition for problem in definition of $u = \text{prox}_{th}(z)$
Augmented Lagrangian method

choose initial $z^{(0)}$ and repeat:

1. minimize augmented Lagrangian

$$\hat{(x, y)} = \arg\min_{x, y} \left( f(x) + g(y) + \frac{t_k}{2} \left\| Ax - y + \frac{1}{t_k} z^{(k-1)} \right\|^2 \right)$$

2. dual update

$$z^{(k)} = z^{(k-1)} + t_k (A\hat{x} - \hat{y})$$

• also known as method of multipliers, Bregman iteration
• this is the proximal point method applied to the dual problem
• as variants, can apply the fast proximal point methods to the dual
• usually implemented with inexact minimization in step 1

Proximal point method 10-12
Examples

\[
\text{minimize } f(x) + g(Ax)
\]

equality constraints \((g \text{ is indicator of } \{b\})\):

\[
\hat{x} = \arg\min_x \left( f(x) + z^T Ax + \frac{t}{2} \|Ax - b\|^2_2 \right)
\]
\[
z := z + t(A\hat{x} - b)
\]

set constraint \((g \text{ indicator of convex set } C)\):

\[
\hat{x} = \arg\min \left( f(x) + \frac{t}{2} d(Ax + z/t)^2 \right)
\]
\[
z := z + t(A\hat{x} - P(A\hat{x} + z/t))
\]

\(P(u)\) is projection of \(u\) on \(C\), \(d(u) = \|u - P(u)\|_2\) is Euclidean distance
Outline

• proximal point method

• augmented Lagrangian method

• Moreau-Yosida smoothing
Moreau-Yosida smoothing

Moreau-Yosida regularization (Moreau envelope) of closed convex $f$ is

$$f(t)(x) = \inf_u \left( f(u) + \frac{1}{2t} \|u - x\|^2 \right) \quad (\text{with } t > 0)$$

$$= f(\text{prox}_{tf}(x)) + \frac{1}{2t} \|\text{prox}_{tf}(x) - x\|^2$$

**immediate properties**

- $f(t)$ is convex (infimum over $u$ of a convex function of $x$, $u$)
- domain of $f(t)$ is $\mathbb{R}^n$ (recall that $\text{prox}_{tf}(x)$ is defined for all $x$)
Examples

**indicator function:** smoothed $f$ is squared Euclidean distance

$$f(x) = I_C(x), \quad f_t(x) = \frac{1}{2t}d(x)^2$$

**1-norm:** smoothed function is Huber penalty

$$f(x) = \|x\|_1, \quad f_t(x) = \sum_{k=1}^{n} \phi_t(x_k)$$

$$\phi_t(z) = \begin{cases} 
\frac{z^2}{2t} & |z| \leq t \\
|z| - \frac{t}{2} & |z| > t
\end{cases}$$

[Graph of $\phi_t(z)$]
Conjugate of Moreau envelope

\[ f(t)(x) = \inf_u \left( f(u) + \frac{1}{2t} \| u - x \|_2^2 \right) \]

- \( f(t) \) is infimal convolution of \( f(u) \) and \( \| v \|_2^2 / (2t) \) (see page 8-14):

\[ f(t)(x) = \inf_{u+v=x} \left( f(u) + \frac{1}{2t} \| v \|_2^2 \right) \]

- from page 8-14, conjugate is sum of conjugates of \( f(u) \) and \( \| v \|_2^2 / (2t) \):

\[(f(t))^*(y) = f^*(y) + \frac{t}{2} \| y \|_2^2 \]

- hence, conjugate is strongly convex with parameter \( t \)
Gradient of Moreau envelope

\[ f_t(x) = \sup_y \left( x^T y - f^*(y) - \frac{t}{2} \|y\|_2^2 \right) \]

• maximizer in definition is unique and satisfies

\[ x - ty \in \partial f^*(y) \iff y \in \partial f(x - ty) \]

• maximizing \( y \) is the gradient of \( f_t \): from pages 6-7 and 9-4,

\[ \nabla f_t(x) = \frac{1}{t} \left( x - \text{prox}_{tf}(x) \right) = \text{prox}_{(1/t)f^*}(x/t) \]

• gradient \( \nabla f_t \) is Lipschitz continuous with constant \( 1/t \)

(follows from nonexpansiveness of prox; see page 6-9)
Interpretation of proximal point algorithm

apply gradient method to minimize Moreau envelope

\[
\minimize f_t(x) = \inf_u \left( f(u) + \frac{1}{2t} \|u - x\|^2_2 \right)
\]

this is an exact smooth reformulation of problem of minimizing \(f(x)\):

• solution \(x\) is minimizer of \(f\)
• \(f_t\) is differentiable with Lipschitz continuous gradient (\(L = 1/t\))

**gradient update:** with fixed \(t_k = 1/L = t\)

\[
x^{(k)} = x^{(k-1)} - t \nabla f_t(x^{(k-1)}) = \prox_{tf}(x^{(k-1)})
\]

... the proximal point update with constant step size \(t_k = t\)
Interpretation of augmented Lagrangian algorithm

\[
(\hat{x}, \hat{y}) = \arg\min_{x,y} \left( f(x) + g(y) + \frac{t}{2} \| Ax - y + (1/t) z \|_2^2 \right)
\]
\[
z := z + t (A\hat{x} - \hat{y})
\]

- with fixed \( t \), dual update is gradient step applied to smoothed dual
- if we eliminate \( y \), primal step can be interpreted as smoothing \( g \):

\[
\hat{x} = \arg\min_x \left( f(x) + g(1/t) (Ax + (1/t)z) \right)
\]

**example:** minimize \( f(x) + \| Ax - b \|_1 \)

\[
\hat{x} = \arg\min_x \left( f(x) + \phi_{1/t} (Ax - b + (1/t)z) \right)
\]

with \( \phi_{1/t} \) the Huber penalty applied componentwise (page 10-15)
proximal point algorithm and fast proximal point algorithm


augmented Lagrangian algorithm