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:

$$\begin{aligned} x^{(k)} &= \mathrm{prox}_{t_k f}(x^{(k-1)}) \\ &= \operatorname*{argmin}_u \left(f(u) + \frac{1}{2t_k}\|u - x^{(k-1)}\|_2^2 \right) \end{aligned}$$

- can be viewed as proximal gradient method (page 6-3) 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, $\mathrm{prox}_{tf}(x)$ is uniquely defined for all x)
- optimal value $f^\star$ is finite and attained at $x^\star$

Result

$$f(x^{(k)}) - f^\star \le \frac{\left\|x^{(0)} - x^\star\right\|_2^2}{2\sum_{i=1}^{k} t_i} \qquad \text{for } k \ge 1$$

- 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 (lecture 6) with $g(x) = 0$

- since g is zero, inequality (3) on page 6-13 holds for any $t > 0$

- from page 6-15, $f(x^{(i)})$ is nonincreasing and

$$t_i \left(f(x^{(i)}) - f^\star \right) \le \frac{1}{2} \left(\|x^{(i)} - x^\star\|_2^2 - \|x^{(i-1)} - x^\star\|_2^2 \right)$$

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

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

Accelerated proximal point algorithms

- we take $g(x) = 0$ in FISTA on page 9-7:

$$
\begin{aligned}
x^{(1)} &= \operatorname{prox}_{t_1 f}(x^{(0)}) \\
x^{(k)} &= \operatorname{prox}_{t_k f}\left(x^{(k-1)} + \theta_k(\frac{1}{\theta_{k-1}} - 1)(x^{(k-1)} - x^{(k-2)})\right) \quad \text{for } k \geq 2
\end{aligned}
$$

- choose any $t_k > 0$, determine θ_k from equation

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

- convergence if $\sum_i \sqrt{t_i} \to \infty$ (lecture 9)

- rate is $1/k^2$ if t_i is fixed or variable but bounded away from zero

Outline

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

Standard problem format

Primal and dual problem (page 7-18)

$$\begin{array}{lll} \text{primal:} & \text{minimize} & f(x) + g(Ax) \\ \text{dual:} & \text{maximize} & -g^*(z) - f^*(-A^T z) \end{array}$$

Examples

- set constraints ($g(y) = \delta_C(y)$):

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax \in C \end{array}$$

- regularized norm approximation ($g(y) = \|y - b\|$):

$$\text{minimize} \quad f(x) + \|Ax - b\|$$

Augmented Lagrangian method: proximal point method applied to dual

Proximal mapping of dual function

Definition: proximal mapping of $h(z) = f^*(-A^T z) + g^*(z)$ is defined as

$$\mathrm{prox}_{th}(z) = \underset{u}{\mathrm{argmin}} \left(f^*(-A^T u) + g^*(u) + \frac{1}{2t}\|u - z\|_2^2 \right)$$

Dual expression: $\mathrm{prox}_{th}(z) = z + t(A\hat{x} - \hat{y})$ where

$$(\hat{x}, \hat{y}) = \underset{x,y}{\mathrm{argmin}} \left(f(x) + g(y) + z^T(Ax - y) + \frac{t}{2}\|Ax - y\|_2^2 \right)$$

$\hat{x}$, $\hat{y}$ minimize the *augmented Lagrangian* (Lagrangian + quadratic penalty)

Proof.

- write augmented Lagrangian minimization as

$$\begin{array}{ll} \text{minimize (over } x, y, w) & f(x) + g(y) + \frac{t}{2}\|w\|_2^2 \\ \text{subject to} & Ax - y + z/t = w \end{array}$$

- optimality conditions (u is multiplier for equality):

$$Ax - y + \frac{1}{t}z = w, \qquad -A^T u \in \partial f(x), \qquad u \in \partial g(y), \qquad tw = u$$

- 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 the definition of $u = \mathrm{prox}_{th}(z)$

Augmented Lagrangian method

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

1. minimize augmented Lagrangian

$$(\hat{x}, \hat{y}) = \underset{x,y}{\operatorname{argmin}} \left(f(x) + g(y) + \frac{t_k}{2} \left\| Ax - y + (1/t_k) z^{(k-1)} \right\|_2^2 \right)$$

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 accelerated proximal point methods to the dual
- usually implemented with inexact minimization in step 1

Examples

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

Equality constraints (g is indicator of $\{b\}$):

$$\begin{aligned} \hat{x} &= \operatorname*{argmin}_x \left(f(x) + z^T Ax + \frac{t}{2}\|Ax - b\|_2^2 \right) \\ z &:= z + t(A\hat{x} - b) \end{aligned}$$

Set constraint (g indicator of convex set C):

$$\begin{aligned} \hat{x} &= \operatorname*{argmin}_x \left(f(x) + \frac{t}{2} d\left(Ax + z/t\right)^2 \right) \\ z &:= z + t\left(A\hat{x} - P(A\hat{x} + z/t)\right) \end{aligned}$$

$P(u)$ 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

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

$$\begin{aligned} f_{(t)}(x) &= \inf_u \left(f(u) + \frac{1}{2t}\|u - x\|_2^2 \right) \qquad \text{(with } t > 0\text{)} \\ &= f\left(\operatorname{prox}_{tf}(x)\right) + \frac{1}{2t}\left\|\operatorname{prox}_{tf}(x) - x\right\|_2^2 \end{aligned}$$

Immediate properties

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

Examples

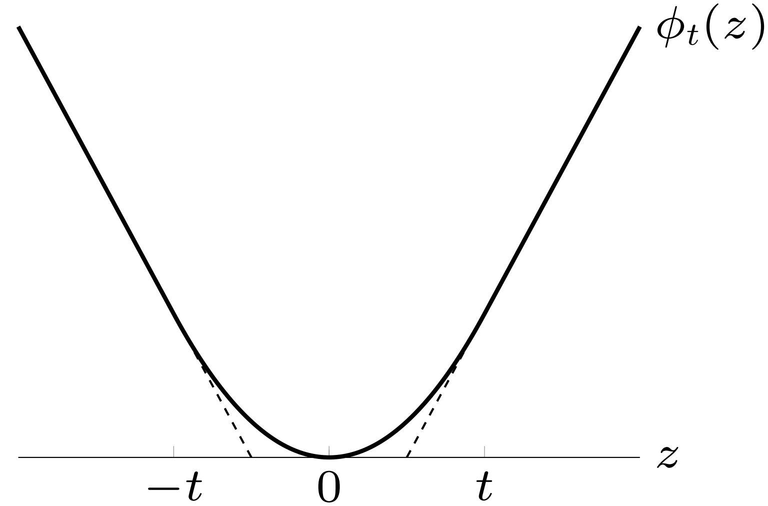
Indicator function: smoothed f is squared Euclidean distance

$$f(x) = \delta_C(x), \qquad f_{(t)}(x) = \frac{1}{2t} d(x)^2$$

1-norm: smoothed function is Huber penalty

$$f(x) = \|x\|_1, \qquad f_{(t)}(x) = \sum_{k=1}^{n} \phi_t(x_k)$$

$$\phi_t(z) = \begin{cases} z^2/(2t) & |z| \le t \\ |z| - t/2 & |z| \ge t \end{cases}$$



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 7-11):

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

- from page 7-11, 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

$$x - ty \in \partial f^*(y) \quad \Longleftrightarrow \quad y \in \partial f(x - ty)$$

- maximizing y is the gradient of $f_{(t)}$: from pages 6-7 and 8-4,

$$\nabla f_{(t)}(x) = \frac{1}{t} \left(x - \mathrm{prox}_{tf}(x) \right) = \mathrm{prox}_{(1/t)f^*}(x/t)$$

- gradient $\nabla f_{(t)}$ is Lipschitz continuous with constant $1/t$ (see p. 7-16 or p. 6-9)

Interpretation of proximal point algorithm

apply gradient method to minimize Moreau envelope

$$\text{minimize} \quad f_{(t)}(x) = \inf_u \left(f(u) + \frac{1}{2t}\|u - x\|_2^2 \right)$$

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 $t_k = 1/L = t$

$$x^{(k)} = x^{(k-1)} - t\nabla f_{(t)}(x^{(k-1)}) = \mathrm{prox}_{tf}(x^{(k-1)})$$

... the proximal point update with constant step size $t_k = t$

Interpretation of augmented Lagrangian algorithm

Augmented Lagrangian iteration

$$
\begin{aligned}
(\hat{x}, \hat{y}) &= \underset{x,y}{\operatorname{argmin}} \left(f(x) + g(y) + \frac{t}{2} \|Ax - y + (1/t)z\|_2^2 \right) \\
z &:= z + t(A\hat{x} - \hat{y})
\end{aligned}
$$

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

$$\hat{x} = \underset{x}{\operatorname{argmin}} \left(f(x) + g_{(1/t)}\left(Ax + (1/t)z\right) \right)$$

Example: minimize $f(x) + \|Ax - b\|_1$

$$\hat{x} = \underset{x}{\operatorname{argmin}} \left(f(x) + \phi_{1/t}(Ax - b + (1/t)z) \right)$$

with $\phi_{1/t}$ the Huber penalty applied componentwise (page 10-12)

References

Proximal point algorithm and fast proximal point algorithm

- O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control and Optimization (1991).
- O. Güler, *New proximal point algorithms for convex minimization*, SIOPT (1992).
- O. Güler, *Augmented Lagrangian algorithm for linear programming*, JOTA (1992).

Augmented Lagrangian algorithm

- D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (1982).