Lecture: Proximal Point Method

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

- Proximal point method
- Augmented Lagrangian method
- Moreau-Yosida smoothing

Proximal Point Method

A 'conceptual' algorithm for minimizing a closed convex function f:

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

- can be viewed as proximal gradient method 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^* is finite and attained at x^*

result

$$f(x^{(k)}) - f^* \leq \frac{||x^{(0)} - x^*||_2^2}{2\sum_{i=1}^{k} t_i} \qquad \text{for } k \geq 1$$

- 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

Convergence

proof: apply analysis of proximal gradient method with $g(x) = 0$

- since g is zero, inquality (1) in "lect-proxg.pdf" on holds for any $t > 0$
- from "lect-proxg.pdf", $f(x^{(i)})$ is nonincreasing and

$$t_i(f(x^{(i)}) - f^*) \leq \frac{1}{2}(||x^{(i)} - x^*||_2^2 - ||x^{(i-1)} - x^*||_2^2)$$

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

$$\begin{aligned} (\sum_{i=1}^{k} t_i)(f(x^{(k)}) - f^*)) &\leq \sum_{i=1}^{k} t_i(f(x^{(i)}) - f^*) \\ &\leq \frac{1}{2}||x^{(0} - x^*||_2^2 \end{aligned} \quad (2)$$

Accelerated proximal point algorithms

FISTA (take $g(x) = 0$): choose $x^{(0)} = x^{(-1)}$ and for $k > 1$

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

Nesterov's 2nd method : choose $x^{(0)} = v^{(0)}$ and for $k \geq 1$

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

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 θ_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, $\text{prox}_{tf}(x)$ id uniquely defined for all x)
- optimal value f^* is finite and attained at x^*

result

$$f(x^{(k)} - f^*) \leq \frac{2||x^{(0)} - x^*||_2^2}{(2\sqrt{t_1} + \sum_{i=2}^{k} \sqrt{t_i})^2} \quad k \geq 1$$

- 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

Convergence

proof: follows from analysis in the "lecture on fast proximal point method" with $g(x) = 0$

- since g is zero, first inequalities on p. 15 and p.25 hold for any $t > 0$
- therefore the conclusion on p. 16 and p. 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 19 that

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

Standard problem form

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

- $f : \mathbb{R}^n \to \mathbb{R}$ 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:

$$\begin{array}{ll} \text{minimize} & f(x) + g(y) \\ \text{subject to} & Ax = y \end{array}$$

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 $Ax \in C$
- $g(y) = ||y - b||$: 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

$$\text{maximize} \quad \inf_{x,y} L(x, y, z) = -f^*(-A^T z) - g^*(z)$$

optimality conditions: x, y, z are optimal if

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

augmented Lagrangian method: proximal point method applied to dual

Proximal mapping of dual function

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

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

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

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

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

- write augmented Lagrangian minimization as

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

- optimality comditions (u is multiplier for equality):

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

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

Augmented Lagrangian method

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

1. minimize augmented Lagrangian

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

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 fast proximal point menthods to the dual
- usually implemented with inexact minimization in step 1

Examples

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

equality constraints (g is indicator of {b})

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

set constraint (g indicator of convex set C):

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

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

Moreau-Yosida smoothing

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

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

immediate properties

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

Examples

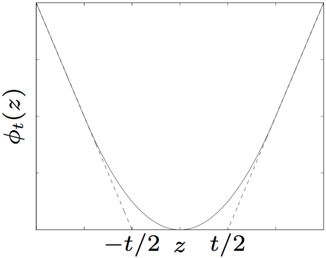
indicator function: smoothed f is squared Euclidean distance

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

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

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

- $f^{(t)}$ infimal convolution of $f(u)$ and $||v||_2^2/(2t)$:

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

- 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

$$f_{(t)}(x) = \sup_y (x^T y - f^*(y) - \frac{t}{2}||y||_2^2)$$

- maximizer in definition is unique and satisfies

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

- maximizing y is the gradient of $f_{(t)}$:

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

- gradient $\nabla f_{(t)}$ is Lipschitz continuous with constant $1/t$ (follows from nonexpansiveness of prox;)

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 $t_k = 1/L = t$

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

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

Interpretation of augmented Lagrangian algorithm

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

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

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

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

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)