Lecture: Proximal Point Method

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

Outline

1. Proximal point method

2. Augmented Lagrangian method

3. Moreau-Yosida smoothing

Proximal Point Method

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

$$\begin{aligned} x^{(k)} &= \text{prox}_{t_k f}(x^{(k-1)}) \\ &= \underset{u}{\text{argmin}}(f(u) + \frac{1}{2t_k}||u - x^{(k-1)}||_2^2) \end{aligned} \tag{1}$$

- 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*

LASSO

考虑LASSO问题：

$$\min_{x\in\mathbb{R}^n} \quad \psi(x) := \mu\|x\|_1 + \frac{1}{2}\|Ax-b\|_2^2. \tag{2}$$

引入变量$y = Ax - b$，问题(2)可以等价地转化为

$$\min_{x,y} \quad f(x,y) := \mu\|x\|_1 + \frac{1}{2}\|y\|_2^2, \qquad \text{s.t.} \quad Ax - y - b = 0. \tag{3}$$

对于问题(3)，我们采用近似点算法进行求解，其第k步迭代为

$$(x^{k+1}, y^{k+1}) \approx \operatorname*{argmin}_{(x,y)\in\mathcal{D}} \left\{ f(x,y) + \frac{1}{2t_k}\left(\|x - x^k\|_2^2 + \|y - y^k\|_2^2\right) \right\}, \tag{4}$$

其中$\mathcal{D} = \{(x,y) \mid Ax - y = b\}$为可行域，$t_k$ 为步长。由于问题(4)没有显式解，我们需要采用迭代算法来进行求解，比如罚函数法，增广拉格朗日方法等等。

除了直接求解问题(4)，一种比较实用的方式是通过对偶问题的解来构造(x^{k+1}, y^{k+1})。引入拉格朗日乘子z，问题(4)的对偶函数为：

$$
\begin{aligned}
\Phi_k(z) = & \inf_x \left\{ \mu \|x\|_1 + z^T Ax + \frac{1}{2t_k} \|x - x^k\|_2^2 \right\} \\
& + \inf_y \left\{ \frac{1}{2} \|y\|_2^2 - z^T y + \frac{1}{2t_k} \|y - y^k\|_2^2 \right\} - b^T z \\
= & \mu \Gamma_{\mu t_k}(x^k - t_k A^T z) - \frac{1}{2t_k} \left(\|x_k - t_k A^T z\|_2^2 - \|x_k\|_2^2 \right) \\
& - \frac{1}{2(t_k + 1)} (\|z\|_2^2 + 2(y^k)^T z - \|y^k\|_2^2) - b^T z.
\end{aligned}
$$

这里，

$$
\Gamma_{\mu t_k}(u) = \inf_x \left\{ \|x\|_1 + \frac{1}{2\mu t_k} \|x - u\|_2^2 \right\}.
$$

通过简单地计算，并记函数$q_{\mu t_k}:\mathbb{R}\rightarrow\mathbb{R}$ 为

$$q_{\mu t_k}(v)=\begin{cases}\frac{v^2}{2\mu t_k}, & |v|\le t,\\ |v|-\frac{\mu t_k}{2}, & |v|>t,\end{cases}$$

我们有$\Gamma_{\mu t_k}(u)=\sum_{i=1}^n q_{\mu t_k}(u_i)$, 其为极小点$x=\mathrm{prox}_{\mu t_k\|x\|_1}(u)$ 处的目标函数值。易知$\Gamma_{\mu t_k}(u)$ 是关于u 的连续可微函数且导数为：

$$\nabla_u\Gamma_{\mu t_k}(u)=u-\mathrm{prox}_{\mu t_k\|x\|_1}(u).$$

那么，问题(4)的对偶问题为

$$\min_z \quad \Phi_k(z).$$

设对偶问题的逼近最优解为z^{k+1}，那么根据问题(4)的最优性条件，我们有

$$\begin{cases}x^{k+1}=\mathrm{prox}_{\mu t_k\|x\|_1}\left(x^k-t_kA^Tz^{k+1}\right),\\ y^{k+1}=\frac{1}{t_k+1}(y^k+t_kz^{k+1}).\end{cases}$$

在第k步迭代，LASSO (2) 问题的近似点算法的迭代格式写为：

$$\begin{cases} z^{k+1} \approx \underset{z}{\operatorname{argmax}} \quad \Phi_k(z), \\ x^{k+1} = \operatorname{prox}_{\mu t_k \|x\|_1}\left(x^k - t_k A^T z^{k+1}\right), \\ y^{k+1} = \dfrac{1}{t_k+1}(y^k + t_k z^{k+1}). \end{cases} \tag{5}$$

根据$\Phi_k(z)$的连续可微性，我们可以调用梯度法进行求解。另外可以证明$\Phi_k(z)$是半光滑的，从而调用半光滑牛顿法来更有效地求解。为了保证算法(5)的收敛性，我们采用以下z^{k+1}满足以下不精确收敛准则：

$$\begin{aligned} &\|\nabla\Phi_k(z^{k+1})\|_2 \le \sqrt{\alpha/t_k}\epsilon_k,\ \epsilon_k \ge 0, \sum_k^{\infty} \epsilon_k < \infty, \\ &\|\nabla\Phi_k(z^{k+1})\|_2 \le \sqrt{\alpha/t_k}\delta_k\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|^2,\ \delta_k \ge 0, \sum_k^{\infty} \delta_k < \infty, \end{aligned} \tag{6}$$

其中ϵ_k, δ_k是人为设定的参数，α为Φ_k的强凹参数（即$-\Phi_k$的强凸参数）。

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

$$\begin{aligned}(\sum_{i=1}^{k} t_i)(f(x^{(k)}) - f^*)) &\leq \sum_{i=1}^{k} t_i(f(x^{(i)}) - f^*) \\ &\leq \frac{1}{2}||x^{(0} - x^*||_2^2\end{aligned} \tag{7}$$

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, $\mathrm{prox}_{tf}(x)$ 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 $g(x) = 0$

- therefore the conclusion 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 that

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

Outline

General augmented Lagrangian framework

Consider

$$\min_x \quad f(x), \quad \text{s.t.} \quad c_i(x) = 0, \quad i = 1, \dots, m,$$

where $f(x)$, $c_i(x)$ are differentiable functions.

- Define the Lagrangian function: $L(x, \lambda) = f(x) - \sum_{i=1}^m \lambda_i c_i(x)$
- The KKT condition is

$$\begin{aligned} \nabla_x L(x, \lambda) &= \nabla f(x) - \sum_{i=1}^m \lambda_i \nabla c_i(x) = 0, \\ c_i(x) &= 0. \end{aligned}$$

General augmented Lagrangian framework

Define the augmented Lagrangian function:

$$L_t(x, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{t}{2}\|c(x)\|_2^2.$$

At each iteration, the augmented Lagrangian method:

- for a given λ, solves the minimization problem:

$$x^+ = \operatorname*{argmin}_x L_t(x, \lambda),$$

  which implies that

$$\nabla f(x^+) - \sum_{i=1}^{m} (\lambda_i - tc_i(x^+)) \nabla c_i(x^+) = 0$$

- then it updates $\lambda^+ = \lambda_i - tc_i(x^+)$.

Hope $c_i(x^+) \to 0$?

Framework for problem with inequality constraints

Consider

$$\min_x f(x), \text{ s.t. } c_i(x) \geq 0, \ i = 1, \cdots, m.$$

An equivalent reformulation is

$$\min_{x,v} f(x), \quad \text{s.t. } c_i(x) - v_i = 0, \quad v_i \geq 0, \ i = 1, \cdots, m.$$

At each iteration, the augmented Lagrangian framework solves

$$\begin{aligned} (x^+, v^+) = \operatorname*{argmin}_{x,v} f(x) + \sum_i \left\{ -\lambda_i (c_i(x) - v_i) + \frac{t}{2} (c_i(x) - v_i)^2 \right\} \\ \text{s.t. } v_i \geq 0, \ i = 1, \cdots, m, \end{aligned} \tag{8}$$

then updates

$$\lambda_i^+ = \lambda_i - t(c_i(x^+) - v_i^+) \tag{9}$$

Framework for problem with inequality constraints

In (8), eliminating the variable v gives

$$v_i^+ = \max(c_i(x^+) - \lambda_i/t, 0).$$

Then (8) becomes:

$$x^+ = \operatorname*{argmin}_x L_t(x, \lambda) := f(x) + \sum_i \psi(c_i(x), \lambda_i, t), \tag{10}$$

where

$$\psi(c_i(x), \lambda_i, t) = \begin{cases} -\lambda_i c_i(x) + \frac{t}{2} c_i^2(x), & \text{if } c_i(x) - \lambda_i/t \le 0 \\ -\frac{\lambda_i^2}{2t}, & \text{otherwise .} \end{cases}$$

The update (9) becomes:

$$\lambda_i^+ = \max(\lambda_i - t c_i(x^+), 0).$$

Splitting + Augmented Lagrangian

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

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

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

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 *augmented Lagrangian* (Lagrangian + quadratic penalty) *proof*

- write augmented Lagrangian minimization as

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

- 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

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

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

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

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):

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

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

Outline

1. Proximal point method
2. Augmented Lagrangian method
3. Moreau-Yosida smoothing

Moreau-Yosida smoothing

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

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

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 $\text{prox}_{tf}(x)$ is defined for all x)

Examples

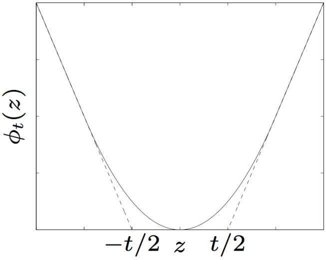
indicator function: smoothed f is squared Euclidean distance

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

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

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

- maximizer in definition is unique and satisfies

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

- Since $x \in \partial (f_{(t)})^*(y) \iff y \in \partial f_{(t)}(x)$, the maximizer 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

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

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

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

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

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

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)