10. Multiplier methods

- proximal point algorithm
- Moreau envelope
- augmented Lagrangian method
- alternating direction method of multipliers (ADMM)

Recall: proximal gradient method

unconstrained problem with composite cost (slightly different notation from lecture 7)

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

- g convex, differentiable, with $\mathbf{dom}\, g = \mathbf{R}^n$
- h convex, possibly nondifferentiable, with inexpensive prox-operator

proximal gradient algorithm

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

$t_k > 0$ is step size, constant or determined by line search

Proximal point algorithm

a (conceptual) algorithm for minimizing a closed convex function f

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

- special case of the proximal gradient method with $g(x) = 0$
- step size $t_k > 0$ affects #iterations, cost of $\mathbf{prox}$ evaluations
- a practical algorithm if inexact $\mathbf{prox}$ evaluations are used
- of interest if prox evaluations are much easier than original problem

basis of the *method of multipliers* or *augmented Lagrangian method*

Convergence

assumptions

- f is closed and convex (hence, $\mathbf{prox}_{tf}(x)$ uniquely defined for all x)
- optimal value $f^\star$ is finite and attained at $x^\star$
- exact evaluations of prox-operator

result

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

- implies convergence if $\sum_i t_i \to \infty$
- rate is $1/k$ if t_i is constant
- t_i is arbitrary; however cost of prox evaluations will depend on t_i (when no closed form, and we choose to do inexactly)

proof: follows from analysis of prox grad method (lecture 7), setting $g = 0$.

- can also apply *accelerated proximal method* (with $g = 0$)
- different variants from lecture 7 can be used

Moreau envelope

Moreau envelope (Moreau-Yosida regularization, Moreau-Yosida smoothing) of closed convex f is defined as

$$f_{(\mu)}(x) = \inf_u \left(f(u) + \frac{1}{2\mu}\|u - x\|_2^2 \right) \qquad \text{(with } \mu > 0\text{)}$$

minimizer in the definition is $u = \mathbf{prox}_{\mu f}(x)$

immediate properties

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

Examples

indicator function (of closed convex set C)

$$f(x) = I_C(x), \qquad f_{(\mu)}(x) = \frac{1}{2\mu}\,\mathbf{dist}(x)^2$$

$\mathbf{dist}(x)$ is the Euclidean distance to C

1-norm

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

ϕ_μ is the Huber penalty

Conjugate of Moreau envelope

$$(f_{(\mu)})^*(y) = f^*(y) + \frac{\mu}{2}\|y\|_2^2$$

proof:

$$\begin{aligned}
(f_{(\mu)})^*(y) &= \sup_x \left(y^T x - f_{(\mu)}(x)\right) \\
&= \sup_{x,u} \left(y^T x - f(u) - \frac{1}{2\mu}\|u - x\|_2^2\right) \\
&= \sup_u \left(y^T(u + \mu y) - f(u) - \frac{\mu}{2}\|y\|_2^2\right) \\
&= f^*(y) + \frac{\mu}{2}\|y\|_2^2
\end{aligned}$$

- note: $(f_{(\mu)})^*$ is strongly convex with parameter μ

Gradient of Moreau envelope

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

- $f_{(\mu)}$ is differentiable; gradient is Lipschitz continuous with constant $1/\mu$
- maximizer in definition satisfies

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

- the maximizing y is the gradient of $f_{(\mu)}$: from p. 6-15 and p. 6-27,

$$\begin{aligned} \nabla f_{(\mu)}(x) &= \frac{1}{\mu} \left(x - \mathbf{prox}_{\mu f}(x) \right) \\ &= \mathbf{prox}_{f^*/\mu}(x/\mu) \end{aligned}$$

Interpretation of proximal point algorithm

apply gradient method to minimize Moreau envelope:

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

this is an exact smooth reformulation of original problem:

- solution x is minimizer of f
- $f_{(\mu)}$ is differentiable with Lipschitz continuous gradient $(L = 1/\mu)$

gradient update: with fixed $t_k = 1/L = \mu$

$$\begin{aligned} x^{(k)} &= x^{(k-1)} - \mu \nabla f_{(\mu)}(x^{(k-1)}) \\ &= \mathbf{prox}_{\mu f}(x^{(k-1)}) \end{aligned}$$

this is the proximal point algorithm with constant step size $t_k = \mu$

Outline

- proximal point algorithm
- Moreau envelope
- **augmented Lagrangian method**
- alternating direction method of multipliers (ADMM)

Augmented Lagrangian method

convex problem and dual (linear constraints for simplicity)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array} \qquad\qquad \begin{array}{ll} \text{maximize} & -F(\lambda, \nu) \end{array}$$

where

$$F(\lambda, \nu) = \begin{cases} h^T\lambda + b^T\nu + f^*(-G^T\lambda - A^T\nu) & \lambda \succeq 0 \\ +\infty & \text{otherwise} \end{cases}$$

augmented Lagrangian method:

proximal point algorithm applied to the dual

Prox-operator of negative dual function

from p. 9-35

$$\mathbf{prox}_{tF}(\lambda, \nu) = \begin{bmatrix} \lambda + t(G\hat{x} + \hat{s} - h) \\ \nu + t(A\hat{x} - b) \end{bmatrix}$$

where $(\hat{x}, \hat{s})$ is the solution of

$$\begin{array}{ll} \text{minimize} & \mathcal{L}(x, s, \lambda, \nu) \\ \text{subject to} & s \succeq 0 \end{array}$$

cost function is augmented Lagrangian

$$\mathcal{L}(x, s, \lambda, \nu) =$$

$$f(x) + \lambda^T(Gx + s - h) + \nu^T(Ax - b) + \frac{t}{2}\left(\|Gx + s - h\|_2^2 + \|Ax - b\|_2^2\right)$$

Algorithm

choose λ, ν, $t > 0$

1. minimize the augmented Lagrangian

$$(\hat{x}, \hat{s}) := \operatorname*{argmin}_{x, s \succeq 0} \mathcal{L}(x, s, \lambda, \nu)$$

2. dual update

$$\lambda := \lambda + t(G\hat{x} + \hat{s} - h), \qquad \nu := \nu + t(A\hat{x} - b)$$

- this is the proximal point algorithm applied to dual problem
- equivalently, gradient method applied to Moreau-Yosida regularized dual
- as a variant, can apply fast proximal point algorithm to the dual

Applications

augmented Lagrangian method is useful when subproblems

$$
\begin{array}{ll}
\text{minimize} & f(x) + \frac{t}{2}\left(\|Gx - h + s + \frac{1}{t}\lambda\|_2^2 + \|Ax - b + \frac{1}{t}\nu\|_2^2\right) \\
\text{subject to} & s \succeq 0
\end{array}
$$

are substantially easier than original problem

(note: apply 'completion of squares' to aug. Lagrangian on page 10-12)

example

$$
\begin{array}{ll}
\text{minimize} & \|x\|_1 \\
\text{subject to} & Ax = b
\end{array}
$$

- solve sequence of ℓ_1-regularized least-squares problems

Outline

- proximal point algorithm
- Moreau-Yosida regularization
- augmented Lagrangian method
- **alternating direction method of multipliers (ADMM)**

Goals

robust methods for

- arbitrary-scale optimization
  - machine learning/statistics with huge data-sets
  - dynamic optimization on large-scale network
- decentralized optimization
  - devices/processors/agents coordinate to solve large problem, by passing relatively small messages
- ideas go back to the 60's; recent surge of interest

([Gabay,Mercier '76], [Glowinski,Marrocco '75], . . .)

Dual decomposition

convex problem with separable objective

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

augmented Lagrangian

$$\mathcal{L}(x, y, \nu) = f(x) + h(y) + \nu^T(Ax + By - b) + \frac{t}{2}\|Ax + By - b\|_2^2$$

- difficulty: quadratic penalty destroys separability of Lagrangian
- solution: replace joint minimization over (x, y) by alternating minimization

Alternating direction method of multipliers

apply one cycle of alternating minimization steps (also known as Gauss-Siedel, block-coordinate descent, etc.) to augmented Lagrangian

1. minimize augmented Lagrangian over x:

$$x^{(k)} = \operatorname*{argmin}_x \mathcal{L}(x, y^{(k-1)}, \nu^{(k-1)})$$

2. minimize augmented Lagrangian over y:

$$y^{(k)} = \operatorname*{argmin}_y \mathcal{L}(x^{(k)}, y, \nu^{(k-1)})$$

3. dual update:

$$\nu^{(k)} := \nu^{(k-1)} + t\left(Ax^{(k)} + By^{(k)} - b\right)$$

can be shown to converge under weak assumptions

Example

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

f convex (not necessarily strongly)

reformulated problem

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

augmented Lagrangian

$$\begin{aligned} \mathcal{L}(x, y, z) &= f(x) + \|y\| + z^T(y - Ax + b) + \frac{t}{2}\|y - Ax + b\|_2^2 \\ &= f(x) + \|y\| + \frac{t}{2}\|y - Ax + b + \frac{1}{t}z\|_2^2 - \frac{1}{2t}\|z\|_2^2 \end{aligned}$$

alternating minimization

1. minimization over x

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

2. minimization over y involves projection on dual norm ball

$$\begin{aligned} \underset{y}{\operatorname{argmin}}\, \mathcal{L}(x, y, z) &= \mathbf{prox}_{\|\cdot\|/t}\,(Ax - b - (1/t)z) \\ &= \frac{1}{t}\,(P_C\,(z - t(Ax - b)) - (z - t(Ax - b))) \end{aligned}$$

where $C = \{u \mid \|u\|_* \leq 1\}$

3. dual update

$$z := z + t(y - Ax - b) = P_C(z - t(Ax - b))$$

comparison with dual proximal gradient algorithm (lecture 9)

- ADMM does not require strong convexity of f, can use larger values of t
- dual updates are identical
- ADMM step 1 may be more expensive, *e.g.*, for $f(x) = (1/2)\|x - a\|_2^2$:

$$x := (I + tA^T A)^{-1}(a + A^T(z + t(y - b)))$$

  as opposed to $x := a + A^T z$ in the dual proximal gradient method

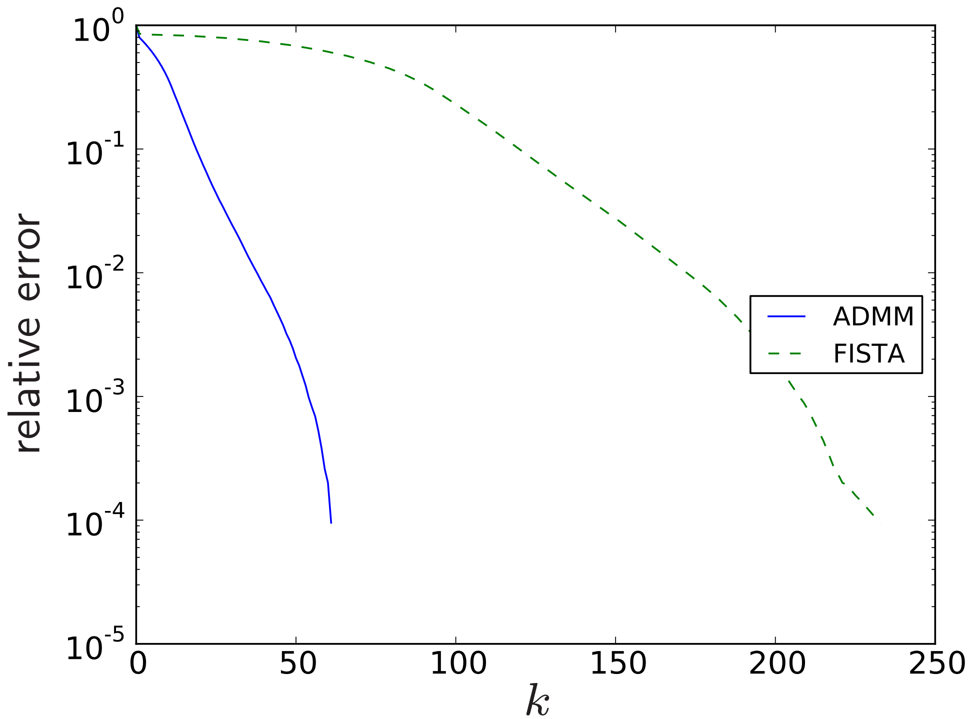
related algorithms (see references)

- split Bregman method with linear constraints
- fast alternating minimization algorithms

example: nuclear norm approximation (problem instance of p. 9-18)

$$\text{minimize} \quad \frac{1}{2}\|x-a\|_2^2 + \|A(x)-B\|_*$$

$\|\cdot\|_*$ is nuclear norm; $A : \mathbf{R}^n \times \mathbf{R}^{p\times q}$ with $A(x) = \sum^n x_i A_i$



FISTA step size is $1/L = 1/\|A\|_2^2$; ADMM step size is $t = 100/\|A\|_2^2$

(recall FISTA is a variant of Nesterov's 1st method covered in lecture 8)

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)

alternating direction method of multipliers and related algorithms

- S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers* (2010)
- D. Goldfarb, S. Ma, K. Scheinberg, *Fast alternating linearization methods for minimizing the sum of two convex functions*, (2010)
- T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imag. Sciences (2009)