Alternating Direction Method of Multipliers

Ernest K. Ryu

MATH 273A: Optimization Theory
University of California, Los Angeles
Department of Mathematics

source:

Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers (Boyd, Parikh, Chu, Peleato, Eckstein)

Goals

robust methods for

- arbitrary-scale optimization
  - machine learning/statistics with huge data-sets
  - dynamic optimization on large-scale network
  - computer vision
- decentralized optimization
  - devices/processors/agents coordinate to solve large problem, by passing relatively small messages

Outline

Outline

Dual problem

- convex equality constrained optimization problem

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

- Lagrangian: $L(x, y) = f(x) + y^T(Ax - b)$
- dual function: $g(y) = \inf_x L(x, y)$
- dual problem: maximize $g(y)$
- recover $x^\star = \operatorname{argmin}_x L(x, y^\star)$

Dual ascent

- gradient method for dual problem: $y^{k+1} = y^k + \alpha^k \nabla g(y^k)$
- $\nabla g(y^k) = A\tilde{x} - b$, where $\tilde{x} = \operatorname{argmin}_x L(x, y^k)$
- dual ascent method is

$$\begin{array}{lcll} x^{k+1} & := & \operatorname{argmin}_x L(x, y^k) & \textit{// } x\textit{-minimization} \\ y^{k+1} & := & y^k + \alpha^k (Ax^{k+1} - b) & \textit{// dual update} \end{array}$$

- works, with lots of strong assumptions

Dual decomposition

- suppose f is separable:

$$f(x) = f_1(x_1) + \cdots + f_N(x_N), \quad x = (x_1, \ldots, x_N)$$

- then L is separable in x:
  $L(x, y) = L_1(x_1, y) + \cdots + L_N(x_N, y) - y^T b$,

$$L_i(x_i, y) = f_i(x_i) + y^T A_i x_i$$

- x-minimization in dual ascent splits into N separate minimizations

$$x_i^{k+1} \quad := \quad \operatorname*{argmin}_{x_i} L_i(x_i, y^k)$$

  which can be carried out in parallel

Dual decomposition

- dual decomposition (Everett, Dantzig, Wolfe, Benders 1960–65)

$$\begin{aligned} x_i^{k+1} &:= \operatorname{argmin}_{x_i} L_i(x_i, y^k), \quad i = 1, \ldots, N \\ y^{k+1} &:= y^k + \alpha^k (\textstyle\sum_{i=1}^N A_i x_i^{k+1} - b) \end{aligned}$$

- scatter y^k; update x_i in parallel; gather $A_i x_i^{k+1}$

- solve a large problem
  - by iteratively solving subproblems (in parallel)
  - dual variable update provides coordination

- works, with lots of assumptions; often slow

Outline

Method of multipliers

- a method to robustify dual ascent
- use **augmented Lagrangian** (Hestenes, Powell 1969), $\rho > 0$

$$L_\rho(x, y) = f(x) + y^T(Ax - b) + (\rho/2)\|Ax - b\|_2^2$$

- method of multipliers (Hestenes, Powell; analysis in Bertsekas 1982)

$$\begin{aligned} x^{k+1} &:= \operatorname*{argmin}_x L_\rho(x, y^k) \\ y^{k+1} &:= y^k + \rho(Ax^{k+1} - b) \end{aligned}$$

(note specific dual update step length ρ)

Method of multipliers dual update step

- optimality conditions (for differentiable f):

$$Ax^\star - b = 0, \qquad \nabla f(x^\star) + A^T y^\star = 0$$

  (primal and dual feasibility)

- since x^{k+1} minimizes $L_\rho(x, y^k)$

$$\begin{array}{rcl} 0 & = & \nabla_x L_\rho(x^{k+1}, y^k) \\ & = & \nabla_x f(x^{k+1}) + A^T \left(y^k + \rho(Ax^{k+1} - b)\right) \\ & = & \nabla_x f(x^{k+1}) + A^T y^{k+1} \end{array}$$

- dual update $y^{k+1} = y^k + \rho(x^{k+1} - b)$ makes (x^{k+1}, y^{k+1}) *dual feasible*
- *primal feasibility* achieved in limit: $Ax^{k+1} - b \to 0$

Method of multipliers

(compared to dual decomposition)

- *good news*: converges under much more relaxed conditions (f can be nondifferentiable, take on value $+\infty$, ...)

- *bad news*: quadratic penalty destroys splitting of the x-update, so can't do decomposition

Outline

Alternating direction method of multipliers

- a method
  - with good robustness of method of multipliers
  - which can support decomposition
- "robust dual decomposition" or "decomposable method of multipliers"
- proposed by Gabay, Mercier, Glowinski, Marrocco in 1976

Alternating direction method of multipliers

- ADMM problem form (with f, g convex)

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

  - two sets of variables, with separable objective

- $L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)\|Ax + Bz - c\|_2^2$

- ADMM:

$$\begin{array}{lll} x^{k+1} & := & \operatorname{argmin}_x L_\rho(x, z^k, y^k) \qquad // \ x\text{-minimization} \\ z^{k+1} & := & \operatorname{argmin}_z L_\rho(x^{k+1}, z, y^k) \qquad // \ z\text{-minimization} \\ y^{k+1} & := & y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \qquad // \ \text{dual update} \end{array}$$

Alternating direction method of multipliers

- if we minimized over x and z jointly, reduces to method of multipliers
- instead, we do one pass of a Gauss-Seidel method
- we get splitting since we minimize over x with z fixed, and vice versa

ADMM and optimality conditions

- optimality conditions (for differentiable case):
  - primal feasibility: $Ax + Bz - c = 0$
  - dual feasibility: $\nabla f(x) + A^T y = 0, \quad \nabla g(z) + B^T y = 0$

- since z^{k+1} minimizes $L_\rho(x^{k+1}, z, y^k)$ we have

$$\begin{aligned} 0 &= \nabla g(z^{k+1}) + B^T y^k + \rho B^T (Ax^{k+1} + Bz^{k+1} - c) \\ &= \nabla g(z^{k+1}) + B^T y^{k+1} \end{aligned}$$

- so with ADMM dual variable update, $(x^{k+1}, z^{k+1}, y^{k+1})$ satisfies second dual feasibility condition

- primal and first dual feasibility are achieved as $k \to \infty$

ADMM with scaled dual variables

- combine linear and quadratic terms in augmented Lagrangian

$$\begin{aligned} L_\rho(x,z,y) &= f(x) + g(z) + y^T(Ax+Bz-c) + (\rho/2)\|Ax+Bz-c\|_2^2 \\ &= f(x) + g(z) + (\rho/2)\|Ax+Bz-c+u\|_2^2 + \text{const.} \end{aligned}$$

with $u^k = (1/\rho)y^k$

- ADMM (scaled dual form):

$$\begin{aligned} x^{k+1} &:= \operatorname*{argmin}_x \left(f(x) + (\rho/2)\|Ax+Bz^k-c+u^k\|_2^2 \right) \\ z^{k+1} &:= \operatorname*{argmin}_z \left(g(z) + (\rho/2)\|Ax^{k+1}+Bz-c+u^k\|_2^2 \right) \\ u^{k+1} &:= u^k + (Ax^{k+1}+Bz^{k+1}-c) \end{aligned}$$

Convergence

- assume (very little!)
  - f, g convex, closed, proper
  - L_0 has a saddle point
- then ADMM converges:
  - iterates approach feasibility: $Ax^k + Bz^k - c \to 0$
  - objective approaches optimal value: $f(x^k) + g(z^k) \to p^\star$

Related algorithms

- operator splitting methods
  (Douglas, Peaceman, Rachford, Lions, Mercier, . . . 1950s, 1979)
- Dykstra's alternating projections algorithm (1983)
- Spingarn's method of partial inverses (1985)
- Rockafellar-Wets progressive hedging (1991)
- proximal methods (Rockafellar, many others, 1976–)
- saddle-point proximal methods (Chambolle, Pock 2005–)
- Bregman iterative methods (2008–)
- most of these are special cases of the proximal point algorithm
  (Rockafellar 1976)

Outline

Common patterns

- x-update step requires minimizing $f(x) + (\rho/2)\|Ax - v\|_2^2$ (with $v = Bz^k - c + u^k$, which is constant during x-update)
- similar for z-update
- several special cases come up often
- can simplify update by exploiting structure in these cases

Decomposition

- suppose f is block-separable,

$$f(x) = f_1(x_1) + \cdots + f_N(x_N), \qquad x = (x_1, \ldots, x_N)$$

- A is conformably block separable: $A^T A$ is block diagonal
- then x-update splits into N parallel updates of x_i

Proximal operator

- consider x-update when $A = I$

$$x^+ = \operatorname*{argmin}_x \left(f(x) + (\rho/2)\|x - v\|_2^2\right) = \mathbf{prox}_{f,\rho}(v)$$

- some special cases:

$f = I_C$ (indicator fct. of set C)	$x^+ := \Pi_C(v)$ (projection onto C)
$f = \lambda\|\cdot\|_1$ (ℓ_1 norm)	$x_i^+ := S_{\lambda/\rho}(v_i)$ (soft thresholding)

($S_a(v) = (v - a)_+ - (-v - a)_+$)

Quadratic objective

- $f(x) = (1/2)x^T Px + q^T x + r$
- $x^+ := (P + \rho A^T A)^{-1}(\rho A^T v - q)$
- use matrix inversion lemma when computationally advantageous

$$(P + \rho A^T A)^{-1} = P^{-1} - \rho P^{-1} A^T (I + \rho A P^{-1} A^T)^{-1} A P^{-1}$$

- (direct method) cache factorization of $P + \rho A^T A$ (or $I + \rho A P^{-1} A^T$)
- (iterative method) warm start, early stopping, reducing tolerances

Smooth objective

- f smooth
- can use standard methods for smooth minimization
  - gradient, Newton, or quasi-Newton
  - preconditionned CG, limited-memory BFGS (scale to very large problems)
- can exploit
  - warm start
  - early stopping, with tolerances decreasing as ADMM proceeds

Outline

Constrained convex optimization

- consider ADMM for generic problem

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

- ADMM form: take g to be indicator of $\mathcal{C}$

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & x - z = 0 \end{array}$$

- algorithm:

$$\begin{aligned} x^{k+1} &:= \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2)\|x - z^k + u^k\|_2^2 \right) \\ z^{k+1} &:= \Pi_{\mathcal{C}}(x^{k+1} + u^k) \\ u^{k+1} &:= u^k + x^{k+1} - z^{k+1} \end{aligned}$$

Lasso

- lasso problem:

$$\text{minimize} \quad (1/2)\|Ax - b\|_2^2 + \lambda\|x\|_1$$

- ADMM form:

$$\begin{array}{ll} \text{minimize} & (1/2)\|Ax - b\|_2^2 + \lambda\|z\|_1 \\ \text{subject to} & x - z = 0 \end{array}$$

- ADMM:

$$\begin{array}{rcl} x^{k+1} & := & (A^T A + \rho I)^{-1}(A^T b + \rho z^k - y^k) \\ z^{k+1} & := & S_{\lambda/\rho}(x^{k+1} + y^k/\rho) \\ y^{k+1} & := & y^k + \rho(x^{k+1} - z^{k+1}) \end{array}$$

Lasso example

- example with dense $A \in \mathbb{R}^{1500 \times 5000}$ (1500 measurements; 5000 regressors)
- computation times

factorization (same as ridge regression)	1.3s
subsequent ADMM iterations	0.03s
lasso solve (about 50 ADMM iterations)	2.9s
full regularization path (30 λ's)	4.4s

- not bad for a *very short* Matlab script

Outline

Dual decomposition

Method of multipliers

Alternating direction method of multipliers

Common patterns

Examples

Conclusions

Summary and conclusions

ADMM

- is the same as, or closely related to, many methods with other names
- has been around since the 1970s
- gives simple single-processor algorithms that can be competitive with state-of-the-art
- can be used to coordinate many processors, each solving a substantial problem, to solve a very large problem