Proximal methods

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

December 30, 2023

Outline

1. Alternating direction method of multipliers (ADMM)
   - constrained convex optimization
   - General patterns of ADMM
2. Proximal methods
   - Proximal operator
   - ADMM in proximal form
   - Projection on some convex sets
   - Proximal operators in closed-form
3. ADMM in applications
   - Lasso
   - Global consensus
   - Allocation
4. ADMM convergence

Alternating direction method of multipliers (ADMM)

Problem structures

some structures that are amenable for applying the methods in this chapter

- global consensus: minimizing $\sum_{i=1}^N f_i(x)$ is equivalent to

  $$\text{minimize} \quad \sum_{i=1}^N f_i(x_i) \;\; \text{subject to } x_1 = x_2 = \cdots = x_N$$

  (minimizing local objective on a global x)
- exchange problem: minimizing social cost subject to market clearing

  $$\text{minimize} \quad \sum_{i=1}^N f_i(x_i) \quad \text{subject to} \quad \sum_{i=1}^N x_i = 0$$

- allocation problem

  $$\text{minimize} \sum_{i=1}^N f_i(x) \;\; \text{subject to} \;\; x \succeq 0, \quad \sum_{i=1}^N x_i = b_i$$

Problem format for ADMM

ADMM solves problems in the form

$$\begin{array}{ll} \text{minimize}_{x,z} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c \end{array} \tag{1}$$

- $f, g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ are closed proper convex (can be nonsmooth)
- the objective function is separable across splitting variable x and z
- the **augmented Lagrangian** associated with the problem is

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

  where $\rho > 0$ is a penalty parameter and $y \in \mathbf{R}^n$ is a dual variable
- L_ρ is the usual Lagrangian with an quadratic penalty on the equality constraint

ADMM algorithm

consider the problem (1), ADMM consists of the iterations

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} \; L_\rho(x, z^k, y^k) \\ z^{k+1} &= \underset{z}{\operatorname{argmin}} \; L_\rho(x^{k+1}, z, y^k) \\ y^{k+1} &= y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \end{aligned}$$

- in x- and z- update steps, L_ρ is minimized over the variable using the most recent value of the other primal variable and the dual variable
- the method of multipliers has the form

$$(x^{k+1}, z^{k+1}) = \underset{x,z}{\operatorname{argmin}} \;\; L_\rho(x, z, y^k), \quad y^{k+1} = y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

hence, the term *alternating direction* in ADMM accounts for the alternating update in x, z

Scaled form of ADMM

- $u = (1/\rho)y$: the **scaled** dual variable
- $r = Ax + Bz - c$: residual and complete the square

$$
\begin{aligned}
y^T r + (\rho/2)\|r\|^2 &= (\rho/2)\|r + y/\rho\|^2 - (1/2\rho)\|y\|^2 \\
&= (\rho/2)\|r + u\|^2 - (\rho/2)\|u\|^2
\end{aligned}
$$

using the scaled dual variable, we can express ADMM in scaled form as

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

(u^k is the running sum of the residuals)

Example: constrained convex optimization

the generic constrained convex optimization

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in C, \qquad f \text{ and set } C \text{ are convex}$$

can be rewritten in ADMM format using $g(x) = I_C(x)$ as

$$\underset{x,z}{\text{minimize}} \quad f(x) + g(z) \quad \text{subject to} \quad x - z = 0$$

the scaled form of ADMM is

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

ADMM is beneficial if the x-update and the projection on C are computationally simple

Example: quadratic cost and linear constraints

$$\underset{x}{\text{minimize}} \;\; (1/2)x^T P x + q^T x \quad \text{subject to} \quad Ax = b, \; x \succeq 0, \quad P \in \mathbf{S}_+^n$$

it can be expressed in ADMM format on page 8 with

$$f(x) = (1/2)x^T P x + q^T x, \quad \mathbf{dom}\, f = \{x \mid Ax = b\}, \quad g(x) = I_{\mathbf{R}_+^n}(x)$$

the x-update step becomes an equality-constrained quadratic minimization

$$x^{k+1} = \underset{Ax=b}{\text{argmin}} \;\; (1/2)x^T P x + q^T x + (\rho/2)\|x - z^k + u^k\|_2^2)$$

(KKT condition is a linear system — hence, can be solved easily)

the z-update step is simply a projection on the non-negative orthant

$$z^{k+1} = \Pi_{\mathbf{R}_+^n}(x^{k+1} + u^k) = \max(0, x^{k+1} + u^k) \triangleq (x^{k+1} + u^k)_+$$

General patterns of ADMM

general cases that will be encountered repeatedly

we illlustrate with the x-update which has the form

$$x^+ = \underset{x}{\text{argmin}} \; (f(x) + (\rho/2)\|Ax - v\|_2^2), \quad v = -Bz + c$$

- proximal operator: when $A = I$
- f is quadratic: $f(x) = (1/2)x^T Px + q^T x + r$
- decomposition: $f(x) = \sum_i f_i(x_i)$
- ℓ_1-norm: $f(x) = \lambda\|x\|_1$

Proximal methods

Proximal operator

let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a closed proper convex function

the *proximal operator* $\mathbf{prox}_{\lambda f} : \mathbf{R}^n \to \mathbf{R}^n$ of f with parameter $\lambda > 0$ is defined by

$$\mathbf{prox}_{\lambda f}(v) = \underset{x}{\operatorname{argmin}} \left(f(x) + \frac{1}{2\lambda}\|x - v\|_2^2 \right)$$

$\mathbf{prox}_{\lambda f}(v)$ is a point that compromises between minimizing f and being near v

- when f is the *indicator function*: $I_C(x) = 0$ if $x \in C$ and $I_C(x) = +\infty$ otherwise
$$\mathbf{prox}_f(v) = \Pi_C(v) = \operatorname{argmin}_{x \in C} \|x - v\|_2$$
- if $f(x, y) = f_1(x) + f_2(y)$ then $\mathbf{prox}_f(u, v) = (\mathbf{prox}_{f_1}(u), \mathbf{prox}_{f_2}(v))$
- if $f(x) = ag(x) + b$ with $a > 0$ then $\mathbf{prox}_{\lambda f}(v) = \mathbf{prox}_{a\lambda g}(v)$

ADMM in proximal form

the problem of minimizing $f(x) + g(x)$ has the ADMM format as

$$\underset{x,z}{\text{minimize}} \quad f(x) + g(z) \quad \text{subject to} \quad x - z = 0$$

the ADMM update in scaled form is

$$\begin{aligned} x^{k+1} &= \underset{x}{\text{argmin}}\ f(x) + (\rho/2)\|x - z^k + u^k\|_2^2 \\ z^{k+1} &= \underset{z}{\text{argmin}}\ g(z) + (\rho/2)\|x^{k+1} - z + u^k\|_2^2 \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1} \end{aligned}$$

- x-update step is to find $\mathbf{prox}_{f/\rho}(z^k - u^k)$
- z-update step is to find $\mathbf{prox}_{g/\rho}(x^{k+1} + u^k)$
- ADMM is a **proximal algorithm**; favorable when the proximal operators can be efficiently computed

Projection on some convex sets

proximal operator of $I_C(x)$ is the projection on C

set	C	$\Pi_C(v)$
nonnegative orthant	$\mathbf{R}^n_+$	$\max(0, v)$
affine set	$\{x \mid Ax = b\}$	$v - A^\dagger(Av - b)$
		$v - A^T(AA^T)^{-1}(Av - b)$, A is fat
hyperplane	$\{x \mid a^Tx = b\}$	$v + \left(\frac{b - a^Tv}{\\|a\\|_2^2}\right) a$
box	$\{x \mid l \preceq x \preceq u\}$	$(\Pi_C)_k = \begin{cases} l_k, & v_k \leq l_k \\ v_k, & l_k \leq v_k \leq u_k \\ u_k, & v_k \geq u_k \end{cases}$
probability simplex	$\{x \mid x \succeq 0, \mathbf{1}^Tx = 1\}$	$(v - \alpha\mathbf{1})_+$ with $\mathbf{1}^T(v - \alpha\mathbf{1})_+ = 1$
2-norm ball	$\{x \mid \\|x\\|_2 \leq 1\}$	$\Pi_C(v) = \begin{cases} v/\\|v\\|_2, & \\|v\\|_2 > 1 \\ v, & \\|v\\|_2 \leq 1 \end{cases}$
consensus	$\{x \in \mathbf{R}^N \mid x_1 = \cdots = x_N\}$	$(1/N)\sum_{i=1}^N v_i$

Projection on probability simplex

problem: minimize$_x (1/2)\|x - v\|_2$ subject to $x \succeq 0$ and $\mathbf{1}^T x = 1$

Lagrangian: $L(x, \lambda, \nu) = (1/2)\|x - v\|_2^2 - \lambda^T x + \nu(\mathbf{1}^T x - 1)$

zero gradient: $\nabla_x L = 0$ gives $x = \lambda + v - \nu\mathbf{1}$

dual function: $g(\lambda, \nu) = -(1/2)\|\lambda - (\nu\mathbf{1} - v)\|_2^2 - \nu + (1/2)\|v\|_2^2$

dual problem: maximize$_{\lambda,\nu}\, g(\lambda, \nu)$ subject to $\lambda \succeq 0$

- any vector can be split as $u = u_+ + u_- = \max(0, u) + \min(0, u)$
- minimize $\|\lambda - c\|_2^2$ subject to $\lambda \succeq 0$ gives $\lambda^\star = \max(0, c) = c_+$
- $\tilde{g}(\nu) = g(\lambda^\star, \nu) = (-1/2)\| - (\nu\mathbf{1} - v)_-\|_2^2 - \nu + (1/2)\|v\|_2$
- dual problem: minimize$_\nu$ $(1/2)\|(v - \nu\mathbf{1})_+\|_2^2 + \nu$
- optimal primal: $x = (\nu\mathbf{1} - v)_+ - (\nu\mathbf{1} - v) = -(\nu\mathbf{1} - v)_- = (v - \nu\mathbf{1})_+$

  with feasibility: $\mathbf{1}^T(v - \nu\mathbf{1})_+ = 1$ (we can use bisection to solve for ν)

Some proximal operators in closed-form

$f(x)$	$\mathbf{prox}_{\lambda f}(v)$
$(1/2)x^TPx + q^Tx + c,\ P \in \mathbf{S}^n_+$	$(I + \lambda P)^{-1}(v - \lambda q)$
$\\|x\\|_1$ (soft thresholding)	$(\mathbf{prox}_{\lambda f}(v))_i = \begin{cases} v_i - \lambda, & v_i \geq \lambda \\ 0, & \|v_i\| \leq \lambda \\ v_i + \lambda, & v_i \leq -\lambda \end{cases}$
	or $\mathbf{sign}(v)(\|v\| - \lambda)_+ \triangleq S_\lambda(v)$
$\\|x\\|_2$ (block soft thresholding)	$\begin{cases} (1 - \frac{\lambda}{\\|v\\|_2})v, & \\|v\\|_2 \geq \lambda \\ 0, & \\|v\\|_2 < \lambda \end{cases}$

ADMM in applications

Solving lasso with ADMM

problem: minimize $\quad (1/2)\|Ax - b\|_2^2 + \lambda\|x\|_1$

ADMM format: minimize $\quad f(x) + \lambda g(z)$ subject to $x - z = 0$ with

$$f(x) = (1/2)\|Ax - b\|_2^2 = (1/2)x^TA^TAx - (A^Tb)^Tx + b^Tb \text{ and } g(z) = \|z\|_1$$

ADMM updates are

$$\begin{aligned} x^{k+1} &= (A^TA + \rho I)^{-1}(A^Tb + \rho(z^k - u^k)) \quad \text{(main computation)} \\ z^{k+1} &= S_{\lambda/\rho}(x^{k+1} + u^k) \qquad \text{(soft thresholding)} \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1} \end{aligned}$$

which follows from

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} \; f(x) + (\rho/2)\|x - z^k + u^k\|_2^2) = \mathbf{prox}_{f/\rho}(z^k - u^k) \\ z^{k+1} &= \underset{z}{\operatorname{argmin}} \; \lambda g(z) + (\rho/2)\|x^{k+1} - z + u^k\|_2^2 = \mathbf{prox}_{\lambda g/\rho}(x^{k+1} + u^k) \end{aligned}$$

ADMM for global consensus problem

define the **consensus set**

$$C = \{\ (x_1, x_2, \ldots, x_N) \mid x_1 = x_2 = \cdots = x_N\ \}, \quad \text{each } x_i \text{ is a vector}$$

problem in canonical form: minimize $\sum_{i=1}^N f_i(x_i) + I_C(x_1, x_2, \ldots, x_N)$

problem in ADMM format: $f(x) = \sum_{i=1}^N f_i(x_i)$ and $g(z) = I_C(z_1, \ldots, z_N)$

- proximal of f can be separable:

$$\mathbf{prox}_{\lambda f}(u) = (\mathbf{prox}_{\lambda f_1}(u_1), \mathbf{prox}_{\lambda f_2}(u_2), \ldots, \mathbf{prox}_{\lambda f_N}(u_N))$$

- proximal of g is the projection on C

$$\mathbf{prox}_{\lambda g}(v) = \Pi_C(v) = (1/N) \sum_{i=1}^N v_i = \bar{v} \qquad \text{(the average)}$$

ADMM updates (after simplifying) are as follows for $i = 1, 2, \ldots, N$

$$\bar{x}^k = (1/N)\sum_{i=1}^{N} x_i^k, \quad x_i^{k+1} = \mathbf{prox}_{f_i/\rho}(\bar{x}^k - u_i^k), \quad u_i^{k+1} = u_i^k + x_i^{k+1} - \bar{x}^{k+1}$$

- the updates can be distributed in parallel to obtain u_i^{k+1} and x_i^{k+1}
- when $f_i(x_i)$ is a goodness of fit using the ith data set, the prox step on x can be interpreted as ℓ_2-regularized estimation
- the ADMM steps follows from page 13 and are simplified from

$$z_i^{k+1} = (1/N)\sum_{i=1}^{N}\left(x_i^{k+1} + u_i^k\right) \triangleq \bar{x}^{k+1} + \bar{u}^k, \quad u^{k+1} = u^k + x^{k+1} - z^{k+1}$$

plugging the 1st eq into the 2nd eq gives $\bar{u}^{k+1} = 0$ (dual variable has zero average)

ADMM for allocation problem

define the **allocation set**

$$C = \{ \, (x_1, \ldots, x_N) \mid x_i \geq 0, \;\; x_1 + x_2 + \cdots + x_N = b \, \}$$

problem: minimize $\sum_{i=1}^N f_i(x_i)$ subject to $x \succeq 0$ and $\sum_{i=1}^N x_i = b$

problem in ADMM format: $f(x) = \sum_{i=1}^N f_i(x_i)$ and $g(z) = I_C(z_1, z_2, \ldots, z_N)$

- proximal of f can be separable:

$$\mathbf{prox}_{\lambda f}(u) = (\mathbf{prox}_{\lambda f_1}(u_1), \mathbf{prox}_{\lambda f_2}(u_2), \ldots, \mathbf{prox}_{\lambda f_N}(u_N))$$

- proximal of g is the projection on C (similar to projection on probability simplex)

ADMM updates for $i = 1, 2, \ldots, N$

$$x_i^{k+1} = \mathbf{prox}_{f_i/\rho}(z^k - u^k), \quad z^{k+1} = \Pi_C(x^{k+1} + u^k), \quad u^{k+1} = u^k + x^{k+1} - z^{k+1}$$

- the x-update can be done in parallel
- the z-update is a projection on probability simplex that can be solved from the dual, using bisection

ADMM convergence

ADMM convergence

assumptions:

1. the extended functions f and g are closed, proper, and convex (implying that the x- and z-updates are solvable
2. the unaugmented Lagrangian L has a saddle point $(x^\star, z^\star, y^\star)$ (not unique)

$$L(x^\star, z^\star, y) \leq L(x^\star, z^\star, y^\star) \leq L(x, z, y^\star)$$

convergence results: as $k \to \infty$, ADMM iterations satisfy

1. residual convergence: $r^k \to 0$
2. objective convergence: $f(x^k) + g(z^k) \to p^\star$ (ADMM objective approaches the optimal value)
3. dual variable convergence: $y^k \to y^\star$ where $y^\star$ is a dual optimal point

Stopping criterion

define the **primal** and **dual residuals** at iteration $k+1$ as

$$s^{k+1} = \rho A^T B(z^{k+1} - z^k), \quad r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$$

in a convergence proof of ADMM, it can be shown that when $\|x^k - x^\star\|_2 \leq d$,

$$f(x^k) + g(z^k) - p^\star \leq -(y^k)^T r^k + d\|s^k\|_2 \leq \|y^k\|_2 \|r^k\|_2 + d\|s^k\|_2$$

this suggests a stopping rule that the primal and dual residuals must be small

$$\|r^k\|_2 \leq \epsilon^{\rm pri} \quad \text{and} \quad \|s^k\|_2 \leq \epsilon^{\rm dual}$$

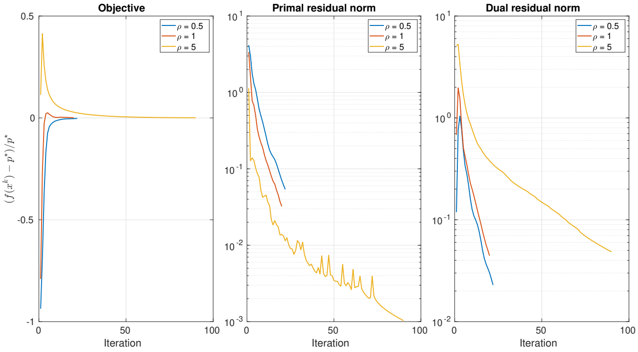
denote $\epsilon^{\rm abs}$ and $\epsilon^{\rm rel}$ the absolute and relative tolerance values, we can choose

$$\begin{aligned} \epsilon^{\rm pri} &= \sqrt{p}\epsilon^{\rm abs} + \epsilon^{\rm rel} \max\{\ \|Ax^k\|_2, \|Bz^k\|_2, \|c\|_2\ \},\ A \in \mathbf{R}^{p\times n} \\ \epsilon^{\rm dual} &= \sqrt{n}\epsilon^{\rm abs} + \epsilon^{\rm rel}\|A^T y^k\|_2 \end{aligned}$$

ADMM iterations: lasso

problem parameters: $(m, n) = (150, 500), \lambda = 0.1\lambda_{\max}$

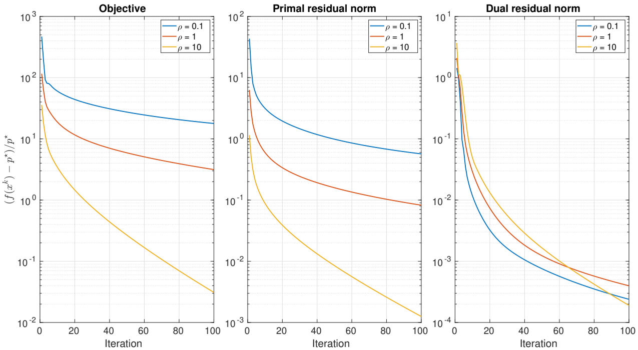
ADMM parameter: $\rho \in \{0.5, 1, 5\}$, tolerance: $\epsilon^{\text{abs}} = 10^{-4}, \epsilon^{\text{rel}} = 10^{-2}$



elapsed time is around 0.01 sec (and around 0.7 sec for $(m, n) = (1500, 5000)$)

ADMM iterations: consensus

local objective: $f_i(x) = (1/2)x^T P_i x + q_i^T x$ for $i = 1, 2, \ldots, N = 10$ and $x \in \mathbf{R}^{100}$

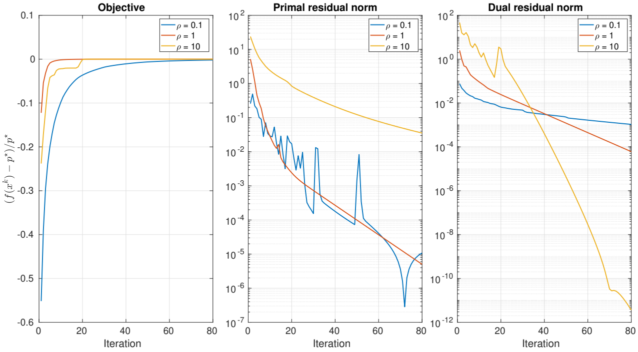


small ρ corresponds to slow convergence in primal residual

elapsed time: 0.1-0.2 sec (not parallel, CVX took 1.1 sec)

ADMM iterations: allocation

local objective: $f_i(x) = (1/2)a_i x^2 + b_i x$ for $i = 1, 2, \ldots, N = 100$ and $x \in \mathbf{R}$



elapsed time: 0.0007 sec (not parallel, CVX took 1 sec)

ADMM parameter (ρ) is chosen to obtain good convergence in both r and s

Summary

- for some problem structures, ADMM has a low computational cost, suitable for large-scale problems
- ADMM solutions can be returned with moderate accuracy (when high accuracy is not crucial)
- ADMM parameter (ρ) is typically tuned by users; it is often problem-dependent, where literature on adaptive penalty approach exists
- ADMM can be applied to non-convex problems where convergence is guaranteed in some problem types

References

1. S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, *Distributed Optimization and Statistical Learnign via the Alternating Direction Method of Multipliers*, Foundations and Trends in Machine Learning, 2011
2. N. Parikh and S. Boyd, *Proximal Algorithms*, Foundations and Trends in Optimization, 2013
3. J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient Projections onto the* ℓ_1*-ball for learning in high dimensions*, ICML, 2008, `https://stanford.edu/~jduchi/projects/DuchiShSiCh08.pdf`