

IIT KHARAGPUR

IIT GANDHINAGAR


NPTEL ONLINE
CERTIFICATION COURSES

Scalable Data Science

Lecture 22: Alternating Direction Method of Multipliers

Sourangshu Bhattacharya
Computer Science and Engineering
IIT KHARAGPUR

In this Lecture:

- Distributed Gradient Descent (recall)
- Distributed Optimization as Equality constrained problem
- Precursors to ADMM
- ADMM

Distributed Optimization

Distributed gradient descent

- Define $loss(\mathbf{x}) = \sum_{j=1}^{m} \sum_{i \in C_j} l_i(\mathbf{x}) + \lambda \Omega(\mathbf{x})$, where $l_i(\mathbf{x}) = l(\mathbf{x}, \mathbf{u}_i, v_i)$
- The gradient (in case of differentiable loss):

$$\nabla loss(\mathbf{x}) = \sum_{j=1}^{m} \nabla \left(\sum_{i \in C_j} l_i(\mathbf{x}) \right) + \lambda \Omega(\mathbf{x})$$

- Compute $\nabla l_j(\mathbf{x}) = \sum_{i \in C_j} \nabla l_i(\mathbf{x})$ on the j^{th} computer. Communicate to central computer.

Distributed gradient descent

- Compute $\nabla loss(\mathbf{x}) = \sum_{j=1}^{m} \nabla l_j(\mathbf{x}) + \Omega(\mathbf{x})$ at the central computer.
- The gradient descent update: $x^{k+1} = x^k - a\nabla loss(\mathbf{x})$.
- a chosen by a line search algorithm (distributed).
- For non-differentiable loss functions, we can use distributed sub-gradient descent algorithm.
  - Slow for most practical problems.

Dual Ascent

- Convex equality constrained problem:

$$\min_x f(x)$$
$$\text{subject to: } Ax = b$$

- Lagrangian: $L(x, y) = f(x) + y^T(Ax - b)$
- Dual function: $g(y) = \inf_x L(x, y)$
- Dual problem: $\max_y g(y)$
- Final solution: $x^* = \operatorname{argmin}_x L(x, y)$

Dual Ascent

- Gradient descent for dual problem:
  $y^{k+1} = y^k + a^k \nabla_{y^k} g(y^k)$
- $\nabla_{y^k} g(y^k) = A\tilde{x} - b$, where $\tilde{x} = \text{argmin}_x L(x, y^k)$
- Dual ascent algorithm:

$$x^{k+1} = \text{argmin}_x L(x, y^k)$$
$$y^{k+1} = y^k + a^k (Ax^{k+1} - b)$$

- Assumptions:
  - $L(x, y^k)$ is strictly convex. Else, the first step can have multiple solutions.
  - $L(x, y^k)$ is bounded below.

Dual Decomposition

- Suppose f is separable:

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

- L is separable in x :
  $L(x, y) = L_1(x_1, y) + \cdots + L_N(x_N, y) - y^T b$, where
  $L_i(x_i, y) = f_i(x_i) + y^T A_i x_i$
- x minimization splits into N separate problems:

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

Dual Decomposition

- Dual decomposition:

$$x_i^{k+1} = \text{argmin}_{x_i} L_i(x_i, y^k), i = 1, \ldots, N$$

$$y^{k+1} = y^k + a^k\left(\sum_{i=1}^{N} A_i x_i - b\right)$$

- Distributed solution:
  - Scatter y^k to individual nodes
  - Compute x_i in the i^{th} node (distributed step)
  - Gather $A_i x_i$ from the i^{th} node
- All drawbacks of dual ascent exist

Method of Multipliers

- Make dual ascent work under more general conditions
- Use **augmented Lagrangian**:

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

- Method of multipliers:

$$x^{k+1} = \text{argmin}_x L_\rho(x, y^k)$$
$$y^{k+1} = y^k + \rho(Ax^{k+1} - b)$$

Methods of Multipliers

- Optimality conditions (for differentiable f):
  - Primal feasibility: $Ax^* - b = 0$
  - Dual feasibility: $\nabla f(x^*) + A^T y^* = 0$

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

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

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

Alternating direction method of multipliers

- Problem with applying standard method of multipliers for distributed optimization:
  - there is no problem decomposition even if f is separable.
  - due to square term $\frac{\rho}{2}||Ax - b||_2^2$

Alternating direction method of multipliers

- ADMM problem:

$$\min_{x,z} f(x) + g(z)$$

subject to: $Ax + Bz = c$

- Lagrangian:
  $L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}||Ax + Bz - c||_2^2$
- ADMM:

$$x^{k+1} = \text{argmin}_x L_\rho(x, z^k, y^k)$$
$$z^{k+1} = \text{argmin}_z L_\rho(x^{k+1}, z, y^k)$$
$$y^{k+1} = y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

Alternating direction method of multipliers

- Problem with applying standard method of multipliers for distributed optimization:
  - there is no problem decomposition even if f is separable.
  - due to square term $\frac{\rho}{2}||Ax - b||_2^2$

- The above technique reduces to method of multipliers if we do joint minimization of x and z
- Since we split the joint x , z minimization step, the problem can be decomposed.

ADMM Optimality conditions

- Optimality conditions (differentiable case):
  - Primal feasibility: $Ax + Bz - c = 0$
  - Dual feasibility: $\nabla f(x) + A^T y = 0$ and $\nabla g(z) + B^T y = 0$
- Since z^{k+1} minimizes $L_\rho(x^{k+1}, z, y^k)$:

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

- So, the dual variable update satisfies the second dual feasibility constraint.
- Primal feasibility and first dual feasibility are satisfied asymptotically.

ADMM Optimality conditions

- Primal residual: $r^k = Ax^k + Bz^k - c$
- Since x^{k+1} minimizes $L_\rho(x, z^k, y^k)$:

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

or,

$$\rho A^T B(z^k - z^{k+1}) = \nabla f(x^{k+1}) + A^T y^{k+1}$$

- Hence, $s^{k+1} = \rho A^T B(z^k - z^{k+1})$ can be thought as dual residual.

ADMM with scaled dual variables

- Combine the linear and quadratic terms
  - Primal feasibility: $Ax + Bz - c = 0$
  - Dual feasibility: $\nabla f(x) + A^T y = 0$ and $\nabla g(z) + B^T y = 0$
- Since z^{k+1} minimizes $L_\rho(x^{k+1}, z, y^k)$:

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

- So, the dual variable update satisfies the second dual feasibility constraint.
- Primal feasibility and first dual feasibility are satisfied asymptotically.

Convergence of ADMM

- Assumption 1: Functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}$ are closed, proper and convex.
  - Same as assuming epi $f = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} | f(x) \leq t\}$ is closed and convex.
- Assumption 2: The unaugmented Lagrangian $L_0(x, y, z)$ has a saddle point (x^*, z^*, y^*):

$$L_0(x^*, z^*, y) \leq L_0(x^*, z^*, y^*) \leq L_0(x, z, y^*)$$

Convergence of ADMM

- Primal residual: $r = Ax + Bz - c$
- Optimal objective: $p^* = \inf_{x,z}\{f(x) + g(z) | Ax + Bz = c\}$
- Convergence results:
  - Primal residual convergence: $r^k \rightarrow 0$ as $k \rightarrow \infty$
  - Dual residual convergence: $s^k \rightarrow 0$ as $k \rightarrow \infty$
  - Objective convergence: $f(x) + g(z) \rightarrow p^*$ as $k \rightarrow \infty$
  - Dual variable convergence: $y^k \rightarrow y^*$ as $k \rightarrow \infty$

Stopping criteria

- Stop when primal and dual residuals small:

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

Hence, $||r^k||_2 \rightarrow 0$ and $||s^k||_2 \rightarrow 0$ as $k \rightarrow \infty$

Observations

- x - update requires solving an optimization problem

$$\min_x f(x) + \frac{\rho}{2}||Ax - v||_2^2$$

with, $v = Bz^k - c + u^k$

- Similarly for z-update.
- Sometimes has a closed form.
- ADMM is a meta optimization algorithm.

Decomposition

- If f is separable:

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

- A is conformably block separable; i.e. $A^T A$ is block diagonal.
- Then, x -update splits into N parallel updates of x_i

Consensus Optimization

- Problem:

$$\min_x f(x) = \sum_{i=1}^{N} f_i(x)$$

- ADMM form:

$$\min_{x_i, z} \sum_{i=1}^{N} f_i(x_i)$$

$$\text{s.t. } x_i - z = 0, i = 1, \ldots, N$$

- Augmented lagrangian:

$$L_\rho(x_1, \ldots, x_N, z, y) = \sum_{i=1}^{N} \left(f_i(x_i) + y_i^T(x_i - z) + \frac{\rho}{2}||x_i - z||_2^2\right)$$

Consensus Optimization

- ADMM algorithm:

$$x_i^{k+1} = \text{argmin}_{x_i}\left(f_i(x_i) + y_i^{kT}(x_i - z^k) + \frac{\rho}{2}||x_i - z^k||_2^2\right)$$

$$z^{k+1} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{k+1} + \frac{1}{\rho}y_i^k\right)$$

$$y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1})$$

- Final solution is z^k .

Consensus Optimization

z-update can be written as:

$$z^{k+1} = \bar{x}^{k+1} + \frac{1}{\rho}\bar{y}^{k+1}$$

Averaging the y-updates:

$$\bar{y}^{k+1} = \bar{y}^{k} + \rho(\bar{x}^{k+1} - z^{k+1})$$

Substituting first into second: $\bar{y}^{k+1} = 0$. Hence $z^k = \bar{x}^k$.

Revised algorithm:

$$x_i^{k+1} = \text{argmin}_{x_i}(f_i(x_i) + y_i^{kT}(x_i - \bar{x}^k) + \frac{\rho}{2}\|x_i - \bar{x}^k\|_2^2)$$
$$y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - \bar{x}^{k+1})$$

Final solution is z^k.

Loss minimization

- Problem:

$$\min_x l(Ax - b) + r(x)$$

- Partition A and b by rows:

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_N \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix},$$

where, $A_i \in \mathbb{R}^{m_i \times m}$ and $b_i \in \mathbb{R}^{m_i}$

- ADMM formulation:

$$\min_{x_i, z} \sum_{i=1}^{N} l_i(A_i x_i - b_i) + r(z)$$

$$\text{s.t.: } x_i - z = 0, \; i = 1, \ldots, N$$

Loss minimization

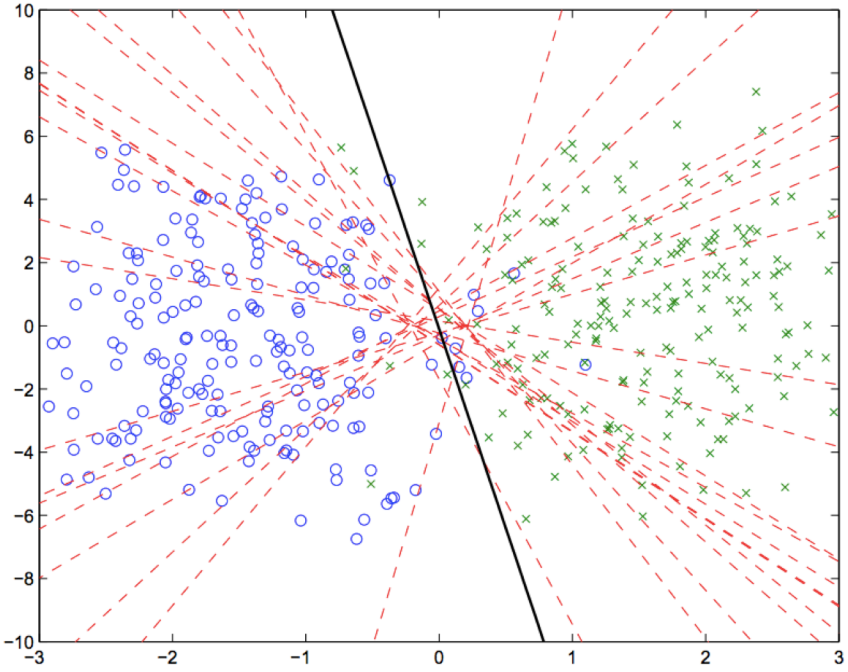
- ADMM solution:

$$x_i^{k+1} = \text{argmin}_{x_i}(l_i(A_i x_i - b_i) + \frac{\rho}{2}\|x_i - z^k + u_i^k\|_2^2)$$
$$u_i^{k+1} = u_i^k + x_i^{k+1} - z^{k+1})$$

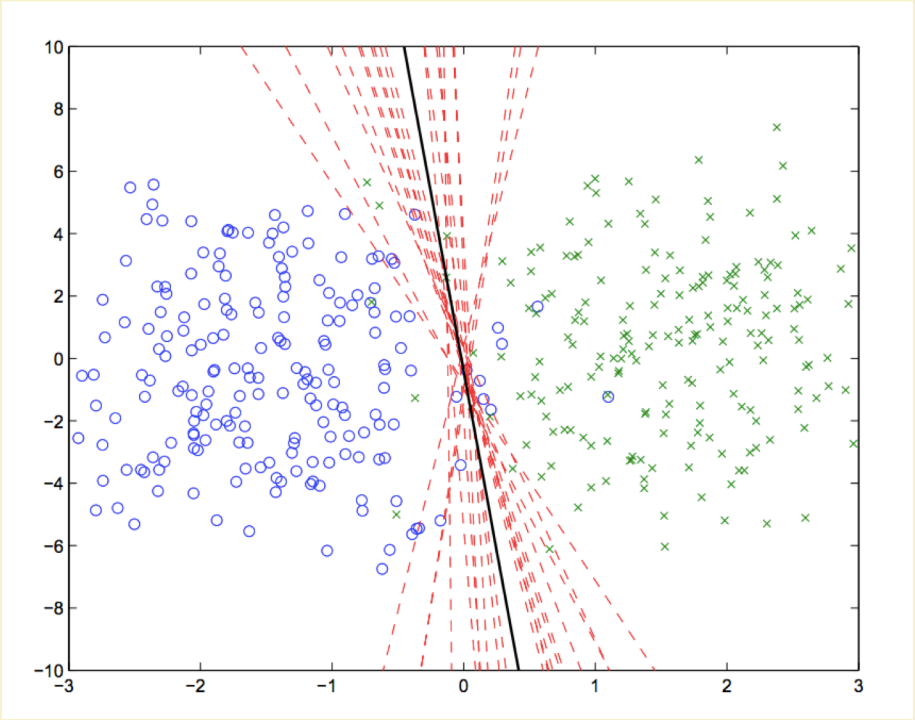
Distributed SVM

- hinge loss $l(u) = (1-u)_+$ with ℓ_2 regularization
- baby problem with $n = 2$, $N = 400$ to illustrate
- examples split into 20 groups, in worst possible way: each group contains only positive or negative examples

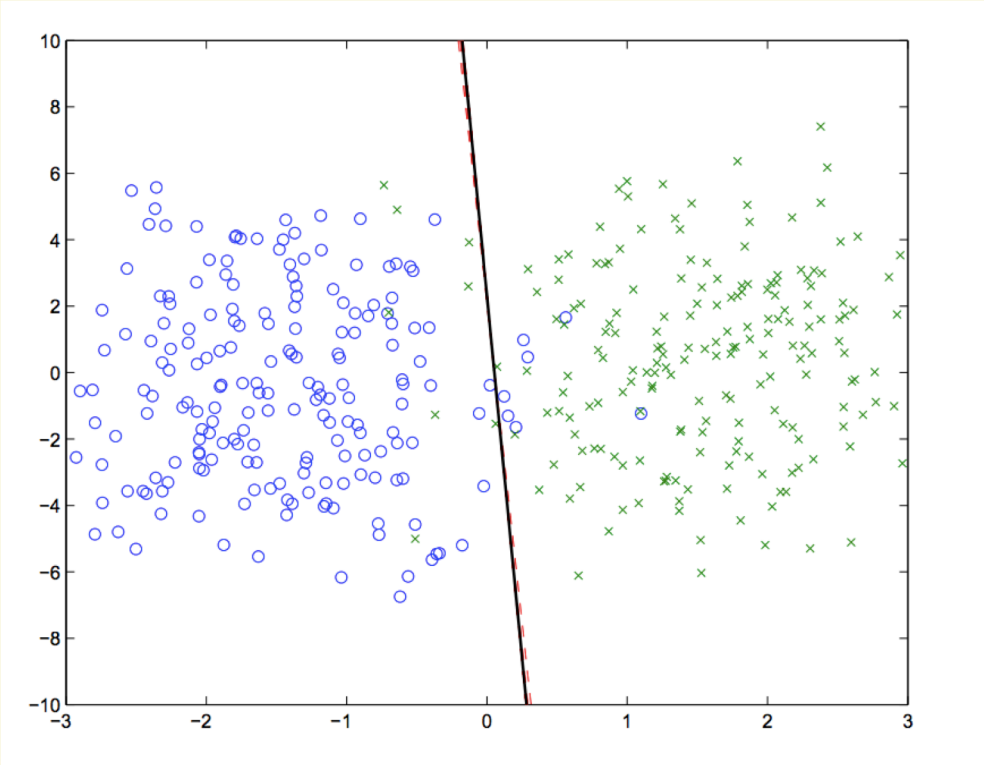
Iteration 1



Iteration 5



Iteration 40



References:

- S. Boyd, N. Pari**Distributed optimization and statistical learning via the alternating direction method of multipliers.**kh, E. Chu, B. Peleato, and J. Eckstein, 2011. http://web.stanford.edu/~boyd/admm.html

Thank You!!