Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 16: Alternating Direction Method of Multipliers

Dual Decomposition

We consider a decomposition of the primal variable into two parts $[x, z]$, which will be treated differently in our optimization algorithms.

$$\phi(x, z) = f(x) + g(z) \qquad \text{subject to } Ax + Bz = c. \tag{1}$$

Its dual is

$$\phi_D(\alpha) = -\alpha^\top c - f^*(-A^\top \alpha) - g^*(-B^\top \alpha) \qquad \alpha \in C_D, . \tag{2}$$

where C_D is the domain of $\phi_D(\cdot)$.

Augmented Lagrangian

In optimization, in order to improve convergence, one often employs the augmented Lagrangian function instead of the regular Lagrangian function, where the primal problem is reformulated as:

$$\min_x \phi_\rho(x,z) := \phi(x,z) + \frac{\rho}{2}\|Ax+Bz-c\|_2^2, \quad \text{subject to } Ax+Bz-c=0. \tag{3}$$

It is easy to see that this formulation is equivalent to (1). The corresponding Lagrangian function, often referred to as the augmented Lagrangian function, is:

$$L_\rho(x,\alpha) = \phi(x,z) + \frac{\rho}{2}\|Ax+Bz-c\|_2^2 + \alpha^\top(Ax+Bz-c).$$

Method of Multipliers

Algorithm 1: Method of Multipliers

Input: $\phi(\cdot)$, A, b, ρ, α_0

Output: x_T

for $t = 1, 2, \ldots, T$ **do**

Let $[x_t, z_t] = \arg\min_{x,z} \left[\alpha_{t-1}^\top [Ax + Bz] + \phi(x, z) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2\right]$

Let $\alpha_t = \mathrm{proj}_{C_D}(\alpha_{t-1} + \rho[Ax_t + Bz_t - c])$

Return: x_T

Equivalent Formulation of Method of Multiplier

Proposition (Dual Proximal Point)

Algorithm 1 is equivalent to the following update on the dual variable:

$$\alpha_t = \arg\max_{\alpha} \left[\phi_D(\alpha) - \frac{1}{2\rho} \|\alpha - \alpha_{t-1}\|_2^2 \right]. \tag{4}$$

Dual Equivalence of Method of Multiplier:

$$A \nabla f^*(-A^\top \alpha) + B \nabla g^*(B^\top \alpha_t) - c - \rho^{-1}[\alpha_t - \alpha_{t-1}] = 0.$$

Dual Equivalence of Dual proximal gradient ascent:

$$A \nabla f^*(-A^\top \alpha) + B \nabla g^*(B^\top \alpha_{t-1}) - c - \rho^{-1}[\alpha_t - \alpha_{t-1}] = 0.$$

ADMM

Algorithm 2: Alternating Direction Method of Multipliers (ADMM)

Input: $\phi(\cdot)$, A, B, c, ρ, α_0, x_0, z_0

Output: x_T

for $t = 1, 2, \ldots, T$ **do**

  Let $z_t = \arg\min_z \left[\alpha_{t-1}^\top Bz + g(z) + \frac{\rho}{2}\|Ax_{t-1} + Bz - c\|_2^2\right]$

  Let $x_t = \arg\min_x \left[\alpha_{t-1}^\top Ax + f(x) + \frac{\rho}{2}\|Ax + Bz_t - c\|_2^2\right]$

  Let $\alpha_t = \alpha_{t-1} + \rho[Ax_t + Bz_t - c]$

Return: x_T

The treatment of x_t and z_t are not symmetric due to the ordering. Compare to the Method of Multipliers:

$$z_t = \arg\min_z \left[\alpha_{t-1}^\top Bz + g(z) + \frac{\rho}{2}\|Ax_t + Bz - c\|_2^2\right]$$

Example

For the concensus optimization problem:

$$\min_{x,z} \sum_{i=1}^{m} f(x_i) \qquad \text{subject to } x_i = z \quad (i = 1, \ldots, m).$$

The ADMM method (we switch the order of sequential optimization for x_t and z_t) is

- $[x_t]_i = \arg\min_x [[\alpha_{t-1}]_i^\top x + f_i(x_i) + \frac{\rho}{2}\|x_i - z_{t-1}\|_2^2]$ $(i = 1, \ldots, m)$
- $z_t = m^{-1} \sum_{i=1}^m [x_t]_i$
- $[\alpha_t]_i = [\alpha_{t-1}]_i + \rho([x_t]_i - z_t)$ $(i = 1, \ldots, m)$

Example

We may apply ADMM to solve the generalized Lasso problem

$$\min_{x,z}[f(x) + \mu\|z\|_1], \quad Ax - z = 0,$$

and obtain

- $z_t = \arg\min_z[-\alpha_{t-1}^\top z + \mu\|z\|_1 + 0.5\rho\|z - Ax_{t-1}\|_2^2]$
- $x_t = \arg\min_x[\alpha_{t-1}^\top Ax + f(x) + 0.5\rho\|Ax - z_t\|_2^2]$
- $\alpha_t = \alpha_{t-1} + \rho(Ax_t - z_t)$

In this case, the generalized L_1 regularization, and the loss function $f(x)$ are decoupled.

Preconditioned ADMM

In the standard ADMM algorithm, the solutions for x_t and z_t involve the following optimization problems

$$\min_x \left[\frac{\rho}{2}\|Ax - \tilde{x}\|_2^2 + f(x)\right], \qquad \min_z \left[\frac{\rho}{2}\|Bz - \tilde{z}\|_2^2 + g(z)\right].$$

For general A and B, these problems may not be easy to solve.

Preconditioned ADMM: change the optimization problems into proximal mapping problems.

Preconditioned ADMM

Algorithm 3: Preconditioned ADMM

Input: $\phi(\cdot)$, A, B, c, H, G, ρ, α_0, x_0, z_0

Output: x_T, z_T, α_T

for $t = 1, 2, \ldots, T$ **do**

  Let $z_t = \arg\min_z \left[\alpha_{t-1}^\top Bz + g(z) + \frac{\rho}{2}\|Ax_{t-1} + Bz - c\|_2^2 + \frac{1}{2}\|z - z_{t-1}\|_G^2\right]$

  Let $x_t = \arg\min_x \left[\alpha_{t-1}^\top Ax + f(x) + \frac{\rho}{2}\|Ax + Bz_t - c\|_2^2 + \frac{1}{2}\|x - x_{t-1}\|_H^2\right]$

  Let $\alpha_t = \alpha_{t-1} + \rho[Ax_t + Bz_t - c]$

Return: x_T, z_T, α_T

Resulting Algorithm

In the resulting algorithm, we only need to solve the following proximal mapping problems for z_t and x_t:

$$
\begin{aligned}
\tilde{z}_t =& z_{t-1} - \eta_G B^\top [\alpha_{t-1} + \rho(Ax_{t-1} + Bz_{t-1} - c)] \\
z_t =& \arg\min_z \left[\frac{1}{2\eta_G} \|z - \tilde{z}_t\|_2^2 + g(z) \right],
\end{aligned}
$$

and

$$
\begin{aligned}
\tilde{x}_t =& x_{t-1} - \eta_H A^\top [\alpha_{t-1} + \rho(Ax_{t-1} + Bz_t - c)], \\
x_t =& \arg\min_x \left[\frac{1}{2\eta_H} \|x - \tilde{x}_t\|_2^2 + f(x) \right].
\end{aligned}
$$

Convergence Theroem

We have the following convergence result for Preconditioned ADMM.

Theorem

Consider Algorithm 3. Let

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \bar{z}_T = \frac{1}{T}\sum_{t=1}^{T} z_t, \quad \bar{\alpha}_T = \frac{1}{T}\sum_{t=1}^{T} \alpha_t,$$

we have

$$\phi(\bar{x}_T, \bar{z}_T) - \phi(x, z) + \alpha^\top (A\bar{x}_T + B\bar{z}_T - c) - \bar{\alpha}_T^\top (Ax + Bz - c)$$
$$\leq \frac{1}{2T}\left[\|z - z_0\|_G^2 + \|x - x_0\|_H^2 + \rho\|Ax_0 + Bz - c\|_2^2 + \rho^{-1}\|\alpha - \alpha_0\|_2^2\right].$$

Interpretation

In order to interpret the result, let $[x_*, z_*]$ be the optimal solution to the primal problem, and α_* is the optimal solution to the dual problem (optimal Lagrangian multiplier). We may let $[x, z] = [x_*, z_*]$ and $\alpha = \alpha_* + \sqrt{\rho} R\xi$, where

$$\xi = \|A\bar{x}_T + B\bar{z}_T - c\|_2^{-1}(A\bar{x}_T + B\bar{z}_T - c),$$

and

$$R^2 = \left[\|z_* - z_0\|_G^2 + \|x_* - x_0\|_H^2 + \rho\|Ax_0 + Bz_* - c\|_2^2 + 2\rho^{-1}\|\alpha_* - \alpha_0\|_2^2\right].$$

It follows that

$$\phi(\bar{x}_T, \bar{z}_T) + \alpha_*^\top(A\bar{x}_T + B\bar{z}_T - c) - \phi(x_*, z_*) + \sqrt{\rho}R\|A\bar{x}_T + B\bar{z}_T - c\|_2 \leq \frac{2R^2}{T}.$$

Accelerated Linearized ADMM

Algorithm 4: Accelerated Linearized ADMM

Input: $\phi(\cdot)$, A, B, c, H, G, η, β, α_0, x_0, z_0

Output: x_T, Z_T, α_T

Let $\bar{x}_0 = x_0$

Let $\bar{z}_0 = z_0$

for $t = 1, 2, \ldots, T$ **do**

- Let $z_t =$ $\arg\min_z \left[\alpha_{t-1}^\top Bz + g(z) + \frac{1}{2\eta}\|A\bar{x}_{t-1} + Bz - c\|_2^2 + \frac{1}{2}\|z - \bar{z}_{t-1}\|_G^2\right]$
- Let $x_t = \arg\min_x \left[\alpha_{t-1}^\top Ax + f_H(\bar{x}_{t-1}, x) + \frac{1}{2\eta}\|Ax + Bz_t - c\|_2^2\right]$
- Let $\alpha_t = \alpha_{t-1} + \eta^{-1}(1 - \beta)[Ax_t + Bz_t - c]$
- Let $\bar{z}_t = z_t + \beta(z_t - z_{t-1})$
- Let $\bar{x}_t = x_t + \beta(x_t - x_{t-1})$

Return: x_T, z_T, α_T

Implementation

In the implementation, we may take $H = \eta_H^{-1} I - \rho A^\top A$ and $G = \eta_G^{-1} I - \rho B^\top B$. Then in the resulting algorithm, we only need to solve the following proximal mapping problems for z_t:

$$\min_z \left[\frac{1}{2\eta_G} \|z - \tilde{z}_t\|_2^2 + g(z) \right],$$

where

$$\tilde{z}_t = \bar{z}_{t-1} - \eta_G B^\top [\alpha_{t-1} + \eta^{-1}(A\bar{x}_{t-1} + B\bar{z}_{t-1} - c)],$$

and

$$x_t = \bar{x}_{t-1} - \eta_H \nabla f(\bar{x}_{t-1}) - \eta_H A^\top [\alpha_{t-1} + \eta^{-1}(A\bar{x}_{t-1} + Bz_t - c)],$$

where η_H and η_G should satisfy

$$\eta_H \le \eta \|A\|_2^{-2}, \qquad \eta_G \le \eta \|B\|_2^{-2}.$$

Empirical Studies

We study the smoothed hinge loss function $\phi_\gamma(z)$ with $\gamma = 1$, and solves the following $L_1 - L_2$ regularization problem:

$$\min_w \left[\underbrace{\frac{1}{n}\sum_{i=1}^n \phi_\gamma(w^\top x_i y_i)}_{f(w)} + \underbrace{\frac{\lambda}{2}\|w\|_2^2 + \mu\|w\|_1}_{g(w)} \right].$$

We compare different algorithms, where we solve the optimization problem for x_t using one or multiple gradient descent iterations. Note that in our experiments, instead of using x_t as primal solution, we use z_t because z_t is sparser than x_t.

Results ($\lambda = 10^{-3}$ and $\mu = 10^{-3}$)

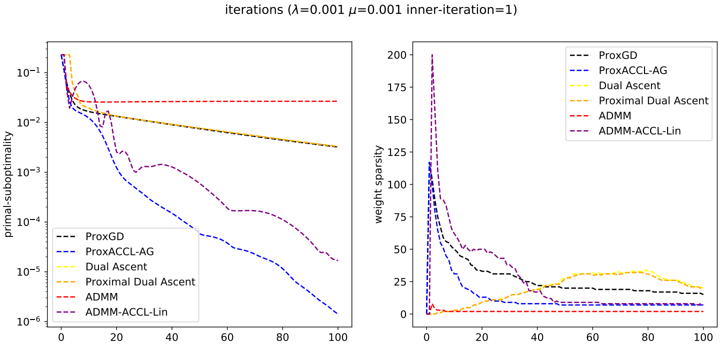

Figure: 1 inner-iteration for solving x_t

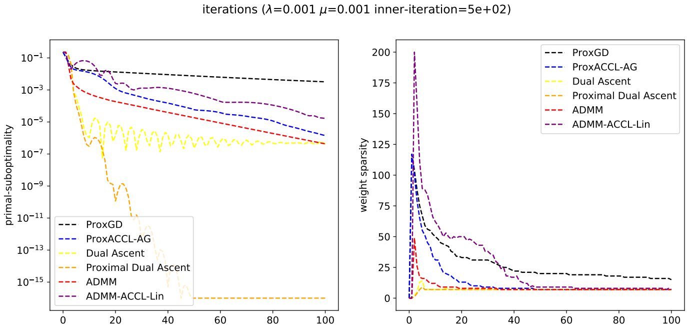

Figure: 500 inner-iterations for solving x_t

Summary

Augmented Lagrangian

- Method of Multipliers
- Dual Proximal Point Method

ADMM

- Standard and Preconditioned
- Solving proximal mapping problems with $f(\cdot)$ and $g(\cdot)$
- Rate is comparable to accelerated methods

Linearized ADMM (one-step gradient descent of $f(\cdot)$)

- Acceleration
- Comparable to Nesterov's method