Alternating Direction Method of Multipliers

1 Introduction

In this lecture, we still consider the following dual decomposition problem:

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

Its dual formulation 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 $\phi_D(\cdot)$.

There are many applications such as concensus optimization, and generalized Lasso. We have investigated dual ascent and proximal dual ascent methods in the previous lecture. In this lecture, we will consider the ADMM method (Alternating Direction Method of Multipliers), which has been widely used in practice [1].

2 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:

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

The augmented Lagrangian formulation is ρ-strongly convex in $Ax+Bz$, which implies ρ^{-1} smoothness for the dual. We may take a stepsize of ρ in the dual formulation, which leads to Algorithm 1, referred to as Method of Multipliers [2]. It is also known as the *Augmented Lagrangian Method* in the literature, which can be applied to general constrained nonconvex optimization in addition to convex optimization.

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 $\alpha_t = \text{proj}_{C_D}(\alpha_{t-1} + \rho[Ax_t + Bz_t - c])$
Return: x_T

The Method of Multipliers is equivalent to the so-called *Proximal Point Method* for maximizing the dual objective function, as shown in the following proposition.

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

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

Proof From Algorithm 1, we have

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

The first order condition of x_t and z_t satisfy

$$\begin{aligned} 0 \in & \partial f(x_t) + A^\top \alpha_{t-1} + \rho A^\top [Ax_t + Bz_t - c], \\ 0 \in & \partial g(z_t) + B^\top \alpha_{t-1} + \rho B^\top [Ax_t + Bz_t - c]. \end{aligned}$$

Using the relationship of α_t and α_{t-1} in Algorithm 1, this can be rewritten as

$$-A^\top \alpha_t \in \partial f(x_t), -B^\top \alpha_t \in \partial g(z_t).$$

It implies that

$$x_t \in \partial f^*(-A^\top \alpha_t), z_t \in \partial g^*(-B^\top \alpha_t),$$

such that

$$Ax_t + Bz_t - c - \frac{1}{\rho}[\alpha_t - \alpha_{t-1}] = 0.$$

This is the optimality condition for α_t in (4), which proves the desired bound. ■

Algorithm 1 often converges better than the dual ascent algorithm, due to the added dual smoothness (follows from the extra augmented Lagrangian term), it is difficult to apply for many practical problems. For example, consider the generalized Lasso problem. If we want to apply Algorithm 1, then we have to solve the following problem

$$\min_{x,z} \left[\alpha^\top (Ax - z) + f(x) + \mu \|z\|_1 + \frac{\rho}{2} \|Ax - z\|_2^2 \right].$$

The optimization with respect to x and x' are coupled. Even if $f(x)$ is quadratic, we may not get a simple closed form solution for the above problem, so the optimization cannot be performed separately for each $f_i(x)$ on local nodes. Similarly, for the concensus optimization problem, we need to solve

$$\min_{x} \sum_{i=1}^{n} \left[\alpha_i^\top (x_i - z) + f_i(x_i) + \frac{\rho}{2} \|x_i - z\|_2^2 \right].$$

Again, all $f_i(\cdot)$ are coupled in this optimization problem. This coupling is undesirable, and it can be resolved using the Alternating Direction Method of Multipliers (ADMM), which we will introduce next.

3 ADMM

ADMM (Alternating Direction Method of Multipliers) tries to remedy the difficulty of dealing with a coupled problem in Algorithm 1. Instead of solving a joint problem as in the method of multipliers, we decouple x and z by solving the problem sequentially (alternating direction) as in Algorithm 2. It is also closely related to the proximal dual ascent algorithm presented in the previous lecture, where z_t is given by $z_t = \arg\min_z[\alpha_{t-1}^\top Bz + g(z)]$ without the augmented Lagrangian term.

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

Example 1 *For the concensus optimization problem:*

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

The ADMM method (we switch the order of sequential optimization for x_t and z_t) is (we take α_0 such that $\sum_i [\alpha_0]_i = 0$.

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

Note that after each update, we always have dual feasibility $\sum_i [\alpha_t]_i = 0$.

Example 2 *Similarly, 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.

If $f(x)$ is a quadratic function, then we can often solve the optimization problem for x_t in closed form.

4 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. For example, if $g(z) = \mu\|z\|_1$, then there is a closed form solution for z with $B = I$. However, there is no closed form solution for general B. To remedy this problem, we can introduce "preconditioners" so that such problems become easy to solve. This leads to Algorithm 3, where H and G are two symmetric positive semidefinite matrices.

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

In applications, we usually take $H = \eta_H^{-1} I - \rho A^\top A$ and $G = \eta_G^{-1} I - \rho B^\top B$. 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

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

In various applications, these proximal mapping problems are easier to solve than the original problems in Algorithm 2. Let $\|\cdot\|_2$ denote the spectral norm of a matrix, then we can choose η_H and η_G as

$$\eta_H \le \rho^{-1}\|A\|_2^{-2}, \qquad \eta_G \le \rho^{-1}\|B\|_2^{-2}.$$

We have the following convergence result for Preconditioned ADMM.

Theorem 1 *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

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

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

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

Since

$$\phi(\bar{x}_T, \bar{z}_T) + \alpha_*^\top(A\bar{x}_T + B\bar{z}_T - c) - \phi(x_*, z_*) \geq 0,$$

we have

$$\|A\bar{x}_T + B\bar{z}_T - c\|_2 \leq \frac{2\rho^{-1/2}R}{T},$$

which means that primal feasibility is satisfied as $T \to \infty$.

In the case that the smallest eigenvalue of $B^\top B$ is $\sigma_B^2 > 0$, and g is M_g Lipschitz, we let $\tilde{z}_T = B^{-1}(c - A\bar{x}_T)$. Then $[\bar{x}_T, \tilde{z}_T]$ satisfies the primal constraint, and we have

$$\begin{aligned}
&\phi(\bar{x}_T, \tilde{z}_T) - \phi(x_*, z_*) \\
\leq&\phi(\bar{x}_T, \bar{z}_T) + \alpha_*^\top(A\bar{x}_T + B\bar{z}_T - c) - \phi(x_*, z_*) + (\|\alpha_*\|_2 + \sigma_B^{-1}M_g)\|B(\bar{z}_T - \tilde{z}_T)\|_2 \\
\leq&\frac{2R^2}{T} + (\|\alpha_*\|_2 + \sigma_B^{-1}M_g)\|A\bar{x}_T + B\bar{z}_T - c\|_2 \\
\leq&\frac{2R}{T}(R + \rho^{-0.5}\|\alpha_*\|_2 + \rho^{-0.5}\sigma_B^{-1}M_g).
\end{aligned}$$

It is wroth mentioning that ADMM achieves the accelerated convergence rate, similar to Nesterov's method. Theorem 1 applies to the non-smooth non-strongly convex optimization, and the rate matches what can be obtained for Nesterov's method with smoothing. For smooth and strongly convex case, ADMM can convergence with a linear rate of $\tilde{O}(\sqrt{\kappa})$, which also matches that of the Nesterov's method. We do not discuss the general situation here.

5 Accelerated Linearized ADMM

In machine learning applications, we usually assume that minimization with respect to the regularizer $g(\cdot)$ is easier to do, but it is tricky to work directly with optimization of $f(\cdot)$. In this case, if $f(\cdot)$ is smooth, then we can use the linearized ADMM, which works with a quadratic upper bound of $f(\cdot)$ as follows:

$$f_H(\tilde{x}, x) = f(\tilde{x}) + \nabla f(\tilde{x})^\top(x - \tilde{X}) + \frac{1}{2}\|x - \tilde{x}\|_H^2.$$

For this formulation, we may apply Nesterov's acceleration to the primal variables, and obtain the accelerated linearized ADMM method. The resulting convergence rate can match that of the Nesterov's method.

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

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

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

6 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 Algorithm 2 and Algorithm 4 to other algorithms, where we solve the optimization problem for x_t using multiple 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.

7 Proof of Theorem 1

From the solution of z_t, we have

$$B^\top \alpha_{t-1} + \nabla g(z_t) + \rho A^\top (Ax_{t-1} + Bz_t - c) + G(z_t - z_{t-1}) = 0.$$

Therefore

$$B^\top \alpha_t + \nabla g(z_t) + \rho B^\top A(x_{t-1} - x_t) + G(z_t - z_{t-1}) = 0.$$

This means that

$$\begin{aligned}
&g(z_t) - g(z) + \alpha_t^\top B(z_t - z) \qquad (5)\\
\leq &\nabla g(z_t)^\top (z_t - z) + \alpha_t^\top B(z_t - z)\\
= &(z_t - z_{t-1})^\top G(z - z_t) - \rho(x_t - x_{t-1})^\top A^\top B(z - z_t)\\
= &\frac{1}{2}[\|z - z_{t-1}\|_G^2 - \|z - z_t\|_G^2 - \|z_t - z_{t-1}\|_G^2]\\
&+ \frac{\rho}{2}[\|Ax_{t-1} + Bz - c\|_2^2 + \|Ax_t + Bz_t - c\|_2^2 - \|Ax_t + Bz - c\|_2^2 - \|Ax_{t-1} + Bz_t - c\|_2^2]\\
\leq &\frac{1}{2}[\|z - z_{t-1}\|_G^2 - \|z - z_t\|_G^2] + \frac{\rho}{2}[\|Ax_{t-1} + Bz - c\|_2^2 - \|Ax_t + Bz - c\|_2^2 + \|\rho^{-1}(\alpha_t - \alpha_{t-1})\|_2^2].
\end{aligned}$$

From the solution of x_t, we obtain

$$A^\top \alpha_{t-1} + \nabla f(x_t) + \rho A^\top (Ax_t + Bz_t - c) + H(x_t - x_{t-1}) = 0.$$

Therefore

$$A^\top \alpha_t + \nabla f(x_t) + H(x_t - x_{t-1}) = 0.$$

We have

$$\begin{aligned}
&f(x_t) - f(x) + \frac{\lambda_f}{2}\|x_t - x\|_2^2 + \alpha_t^\top A(x_t - x) \qquad (6)\\
\leq &\nabla f(x_t)^\top (x_t - x) + \alpha_t^\top A(x_t - x)\\
= &(x_t - x_{t-1})^\top H(x - x_t) = \frac{1}{2}[\|x - x_{t-1}\|_H^2 - \|x - x_t\|_H^2 - \|x_t - x_{t-1}\|_H^2].
\end{aligned}$$

Finally we have

$$\begin{aligned}
&-(\alpha_t - \alpha)^\top (Ax_t + Bz_t - c) \qquad (7)\\
= &-\frac{1}{\rho}(\alpha_t - \alpha)^\top (\alpha_t - \alpha_{t-1}) = \frac{1}{2\rho}\left[\|\alpha - \alpha_{t-1}\|_2^2 - \|\alpha - \alpha_t\|_2^2 - \|\alpha_t - \alpha_{t-1}\|_2^2\right].
\end{aligned}$$

By adding (5), (6), and (7), we obtain

$$\begin{aligned}
&\phi(x_t, z_t) - \phi(x, z) + \alpha^\top (Ax_t + Bz_t - c) - \alpha_t^\top (Ax + Bz - c)\\
\leq &\frac{1}{2}[\|z - z_{t-1}\|_G^2 - \|z - z_t\|_G^2] + \frac{1}{2}[\|x - x_{t-1}\|_H^2 - \|x - x_t\|_H^2]\\
&+ \frac{\rho}{2}[\|Ax_{t-1} + Bz - c\|_2^2 - \|Ax_t + Bz - c\|_2^2] + \frac{1}{2\rho}\left[\|\alpha - \alpha_{t-1}\|_2^2 - \|\alpha - \alpha_t\|_2^2\right].
\end{aligned}$$

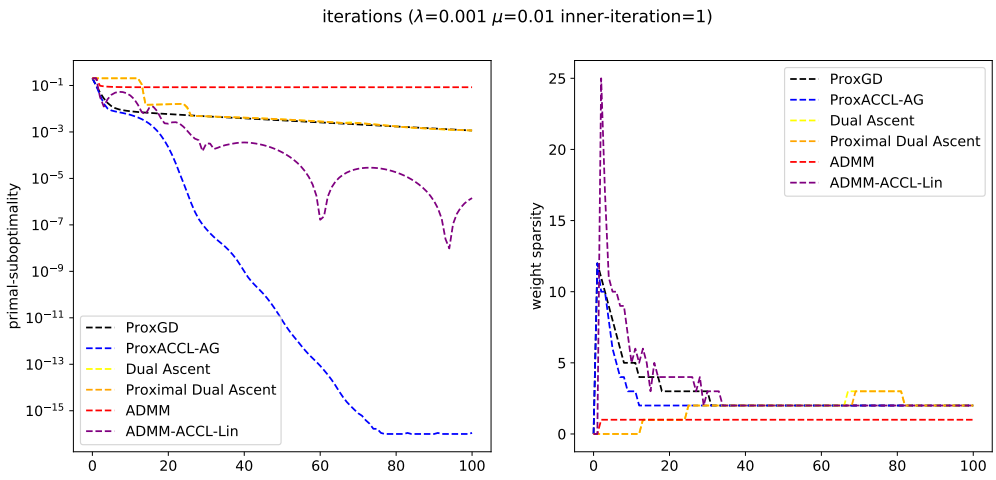
Summing from $t = 1$ to T, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \Big[\phi(x_t, z_t) - \phi(x, z) + \alpha^\top (Ax_t + Bz_t - c) - \alpha_t^\top (Ax + Bz - c)\Big] \\
\leq &\frac{1}{2}\|z - z_0\|_G^2 + \frac{1}{2}\|x - x_0\|_H^2 + \frac{\rho}{2}\|Ax_0 + Bz - c\|_2^2 + \frac{1}{2\rho}\|\alpha - \alpha_0\|_2^2.
\end{aligned}$$

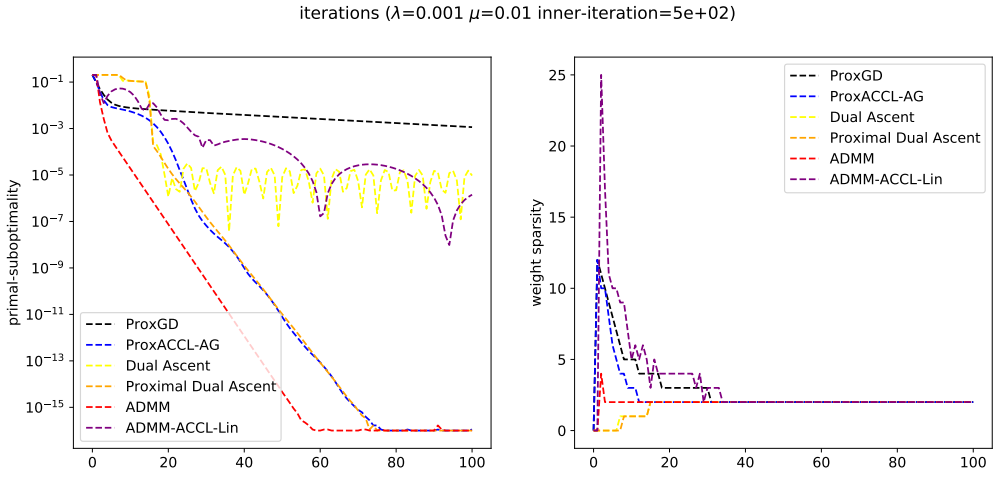
References

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

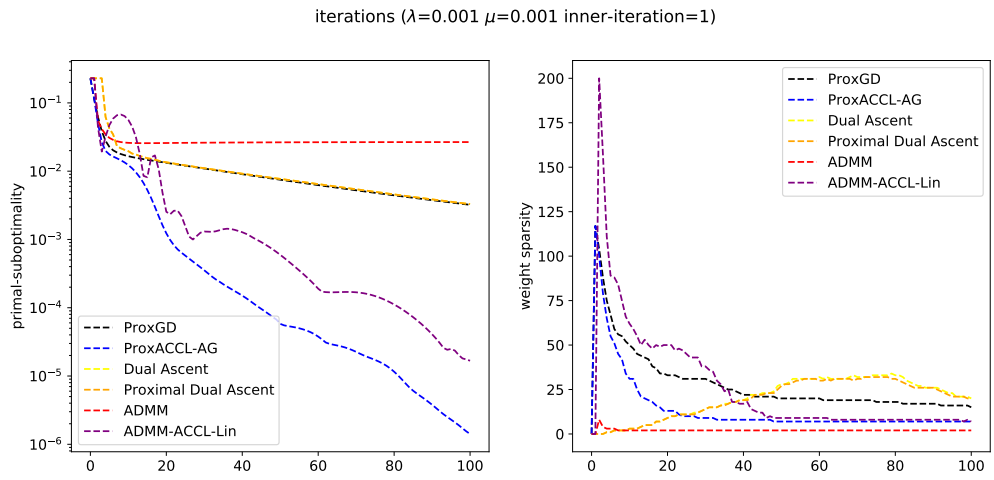


(a) 1 inner-iteration for solving x_t

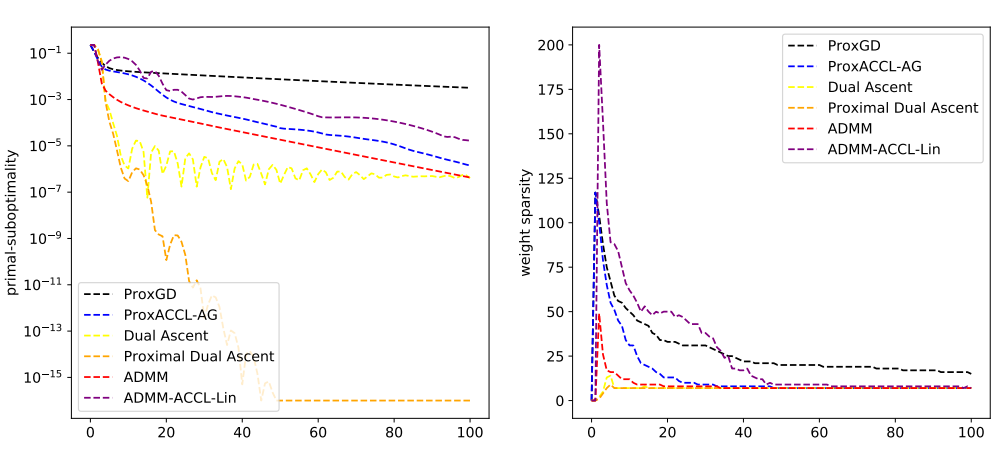


(b) 500 inner-iterations for solving x_t (except accelerated linearized ADMM)

Figure 1: $\lambda = 10^{-3}$ and $\mu = 10^{-2}$



(a) 1 inner-iteration for solving x_t



(b) 500 inner-iterations for solving x_t (except accelerated linearized ADMM)

Figure 2: $\lambda = 10^{-3}$ and $\mu = 10^{-3}$