Lagrangian Duality and Dual Decomposition Methods

1 Lagrangian Duality

We consider the following optimization problem, referred to as the primal problem:

$$\begin{aligned} &\min_x \phi(x) \\ &\text{subject to } g(x) \leq 0 \\ &\text{and } h(x) = 0, \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^d$ is the primal parameter to be optimized. Here $g(x) = [g_1(x), \ldots, g_k(x)]$ and $h(x) = [h_1(x), \ldots, h_m(x)]$.

The Lagrangian function is

$$L(x, \mu, \lambda) = \phi(x) + \mu^\top g(x) + \lambda^\top h(x),$$

where $\mu \in \mathbb{R}^k_+$ and $\lambda \in \mathbb{R}^m$. The Lagrangian multiplier parameters $[\mu, \lambda]$ are called dual variables.

We may define the Lagrange dual function:

$$\phi_D(\mu, \lambda) = \inf_{x \in \mathbb{R}^d} L(x, \mu, \lambda),$$

which is in terms of the dual variables, and the dual optimization problem is:

$$\begin{aligned} &\max_{\mu,\lambda} \phi_D(\mu, \lambda) \\ &\text{subject to} \quad \mu \geq 0. \end{aligned} \tag{2}$$

We have the following weak duality theorem.

Theorem 1 *Given any primal feasible point $x \in C = \{x \in \mathbb{R}^d : g(x) \leq 0, h(x) = 0\}$, and any dual feasible point $[\mu, \lambda]$ with $\mu \geq 0$. We have*

$$\phi(x) \geq \phi_D(\mu, \lambda).$$

Moreover, $\phi_D(\cdot)$ is a concave function in $[\mu, \lambda]$.

Proof For x, μ, λ, we have

$$\phi(x) \geq L(x, \mu, \lambda) \geq \min_{x'} L(x', \mu, \lambda) = \phi_D(\mu, \lambda).$$

Since $L(x, \mu, \lambda)$ is a concave function of $[\mu, \lambda]$ parametrized by x, and taking the infimum over x, the resulting function is concave in $[\mu, \lambda]$. Therefore $\phi_D(\mu, \lambda)$ is concave in $[\mu, \lambda]$. ■

Given a primal x, and dual $[\mu, \lambda]$, the quantity

$$\phi(x) - \phi_D(\mu, \lambda)$$

is referred to as *duality gap*. It is always non-negative, and is an upper bound for primal suboptimality which one can often compute in practice to check convergence.

We are particularly interested in the situation that there exist primal feasible x_* and dual feasible $[\mu_*, \lambda_*]$ such that the duality gap is zero:

$$\phi(x_*) - \phi_D(\mu_*, \lambda_*) = 0.$$

This situation is called *strong duality*. If strong duality holds, then it is clear that x_* is the solution of the primal problem, and $[\mu_*, \lambda_*]$ is the solution of the dual problem. We can think $[\mu_*, \lambda_*]$ as a (dual) certificate that proves x_* is the optimal solution of the primal problem. For convex problem, we have the following result.

Theorem 2 *Assume that* (1) *is convex, and satisfies the Slater's condition: there exists* $x \in \mathbb{R}^d$ *such that*

$$g(x) < 0, \qquad h(x) = 0.$$

Then the strong duality holds: there exists primal feasible x_* *and dual feasible* $[\mu_*, \lambda_*]$ *such that*

$$\phi(x_*) = \phi_D(\mu_*, \lambda_*).$$

Moreover, such $[x_*, \mu_*, \lambda_*]$ *is the solution of the saddle point problem*

$$\min_x \max_{\mu,\lambda} L(x, \mu, \lambda) = \max_{\mu,\lambda} \min_x L(x, \mu, \lambda),$$

and satisfies the KKT conditions

- *Stationarity:* $\nabla_x L(x_*, \mu_*, \lambda_*) = 0$.
- *Primal Feasibility:* $g(x_*) \le 0, \qquad h(x_*) = 0$.
- *Dual Feasibility:* $\mu_* \ge 0$.
- *Complementary Slackness:* $\mu_* g(x_*) = 0$.

A classical example of dual formulation is linear programing. It can be stated as follows.

Example 1 *Primal problem (with* $\phi(x) = c^\top x$*):*

$$\min_x c^\top x \qquad \textit{subject to } Ax - b \le 0.$$

The Lagrangian function is

$$L(x, \lambda) = c^\top x + \lambda^\top (Ax - b).$$

The dual objective function is ($\lambda \ge 0$)*:*

$$\phi_D(\lambda) = \min_x [c^\top x + \lambda^\top (Ax - b)] = \begin{cases} -\lambda^\top b & \textit{if } A^\top \lambda + c = 0 \\ +\infty & \textit{otherwise} \end{cases}$$

The dual problem is:

$$\max_\lambda -b^\top \lambda \quad \textit{subject to } A^\top \lambda + c = 0, \qquad \lambda \ge 0.$$

2 Dual Decomposition

A major application of dual method in machine learning is to decompose (some times called splitting) a complex problem into multiple simpler problems.

The general formulation can be stated as optimization with linear equality constraint. Given a convex objective $\phi(\cdot)$, we consider the following primal optimization problem:

$$\min_x \phi(x) \qquad \text{subject to } Ax = b. \tag{3}$$

For this problem, the Lagrangian multiplier is:

$$L(x, \alpha) = \phi(x) + \alpha^\top (Ax - b).$$

The dual is

$$\phi_D(\alpha) = \min_x L(x, \alpha) = -\alpha^\top b - \sup_x [[-A^\top \alpha]^\top x - \phi(x)] = -\alpha^\top b - \phi^*(-A^\top \alpha) \quad \alpha \in C_D, \tag{4}$$

where C_D is the domain of $\phi^*(\cdot)$. The associated dual optimization problem is:

$$\max_\alpha \phi_D(\alpha), \qquad \text{subject to } \alpha \in C_D.$$

We have strong duality if the primal problem is feasible.

The linear constraint formulation in (3), with special choices of A, can be used to decompose a complex primal objective function into simpler subproblems that can be solved separately. Of specific interests, we consider a decomposition of the primal variable into two parts, which will be treated different in our optimization algorithms. We can rewrite the two parts as $[x, z]$ for clarity, and consider the following formulation.

$$\phi(x, z) = f(x) + g(z) \qquad \text{subject to } Ax + Bz = c. \tag{5}$$

This formulation is equivalent to (3), except for the notations. Its dual is:

$$\phi_D(\alpha) = -\alpha^\top c - f^*(-A^\top \lambda) - g^*(-B^\top \lambda) \qquad \lambda \in C_D, . \tag{6}$$

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

As an example of primal variable decomposition, we may consider the following composite optimization problem, which is a special case of (5)

Example 2 *Let $A = I$, $B = -I$, and $c = 0$. The constraint $Ax + Bz = c$ is $x = z$, and we have $\phi_D(x) = -f^*(-\alpha) - g^*(\alpha)$. This is consistent with the formulation for composite optimization in the last lecture.*

More generally, we have the following concensus optimization problem, which is also a special case of (5). It has applications in distributed computing, where we have m nodes, each node i contain a local function i to optimize.

Example 3 *Assume we want to solve*

$$\sum_{i=1}^m f_i(\bar{x}).$$

We may let $x = [x_1, \ldots, x_m]$ and $z = \bar{x}$, where each x_i is the local primal variable on node i. In this formulation we may let $f(x) = \sum_{i=1}^m f_i(x_i)$, $g(z) = 0$, and (5) becomes the following constrained optimization problem:

$$\min_{x,z} f(x) \qquad \text{subject to } x_1 - z = x_2 - z = \cdots x_m - z = 0.$$

We have the dual $\alpha = [\alpha_1, \ldots, \alpha_m]$, where each α_i is associated with the constraint $x_i - z = 0$. The dual problem is

$$\phi_D(\alpha) = \sum_{i=1}^m -f_i^*(-\alpha_i), \quad \text{subject to } \sum_{i=1}^m \alpha_i = 0.$$

Another example of decomposition is generalized Lasso, which we consider below.

Example 4 *We consider the following generalized Lasso problem (such as fused Lasso [3] or TV-norm regularization [2]), which optimizes*

$$\min_x [f(x) + \mu \|Ax\|_1].$$

It is difficult to apply proximal gradient for complex A. We may consider the constrained formulation that decouples the problem:

$$\phi([x, z]) = f(x) + g(z), \quad \text{subject to } Ax - z = 0,$$

where

$$g(z) = \mu \|z\|_1.$$

The dual problem is

$$\phi_D(\alpha) = -f^*(-A^\top \alpha) \qquad \alpha \in C_D = \{\alpha : \|\alpha\|_\infty \le \mu\}.$$

Another example is graphical Lasso, which tries to estimate a sparse inverse covariance matrix (also called precision matrix) given an observed sample covariance matrix $\hat{\Sigma}$ [1].

Example 5 *We want to solve the following problem, where X is a $d \times d$ symmetric positive definite matrix:*

$$\min_{X \succ 0} \left[-\ln \det(X) + \operatorname{trace}(\hat{\Sigma} X) + \lambda \|X\|_1 \right],$$

where $X \succ 0$ denotes that X is positive definite. It has the following decomposition of the (5):

$$\min_{X \succ 0} \left[-\ln \det(X) + \operatorname{trace}(\hat{\Sigma} X) + \lambda \|Z\|_1 \right], \qquad X - Z = 0.$$

We have the Lagrangian multiplier being a $d \times d$ symmetric matrix Λ, and

$$f(X) = -\ln \det(X) + \operatorname{trace}(\hat{\Sigma} X),$$

and

$$f^*(-\Lambda) = -\ln \det(\hat{\Sigma} + \Lambda) - d, \qquad \hat{\Sigma} + \Lambda \succ 0,$$

and

$$g^*(\Lambda) = 0 \qquad \|\Lambda\|_\infty \le \lambda.$$

Decomposition can also be used to deal with multiple regularizations, where it is possible to deal with a single regularization at each time, but not simultaneously.

Example 6 *Consider solving the following problem for matrix X, which means X is both low rank and sparse:*

$$\min_X \left[f(X) + \lambda \|X\|_* + \mu \|X\|_1 \right].$$

We may let $Z = [Z_1, Z_2]$ and rewrite it as

$$\min_X \left[f(X) + \lambda \|Z_1\|_* + \mu \|Z_2\|_1 \right] \quad \textit{subject to } X - Z_1 = 0,\ X - Z_2 = 0.$$

3 Dual Ascent Algorithms

We now consider solving (5) by performing gradient ascent in dual. In dual ascent algorithm, we have the update

$$\alpha_t = \alpha_{t-1} + \eta_t \nabla \phi_D(\alpha) = \alpha_{t-1} + \eta_t (A x_t + B z_t - c),$$

where $x_t \in \partial f^*(-A^\top \alpha_{t-1})$ and $z_t \in \partial g^*(-B^\top \alpha_{t-1})$. It means that $-A^\top \alpha_{t-1} \in \partial f(x_t)$, and $-B^\top \alpha_{t-1} \in \partial g(z_t)$, and thus

$$\begin{aligned} x_t &= \arg\min_x [\alpha_{t-1}^\top A x + f(x)] \\ z_t &= \arg\min_z [\alpha_{t-1}^\top B z + g(z)]. \end{aligned}$$

This method leads to Algorithm 1 for solving (5). It requires solving decomposed primal problems at each time. For some application, this may be easy to achieve.

Algorithm 1: Dual Ascent Method

Input: $\phi(\cdot)$, A, B, c, α_0, $\eta_1, \eta_2, \ldots$

Output: x_T

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

- Let $x_t = \arg\min_x [\alpha_{t-1}^\top A x + f(x)]$
- Let $z_t = \arg\min_z [\alpha_{t-1}^\top B z + g(z)]$
- Let $\alpha_t = \mathrm{proj}_{C_D}(\alpha_{t-1} + \eta_t [A x_t + B z_t - c])$

Return: x_T

The analysis of Algorithm 1 for the dual objective function is similar to that of the gradient descent method for the primal objective function.

We may also consider dual proximal gradient method, where we apply proximal gradient method to the dual problem:

$$\max_\alpha \left[-c^\top \alpha - f^*(-A^\top \alpha) - g^*(\alpha) \right].$$

In this case, we can assume that $g(\cdot)$ is strongly convex. It means that $g^*(\cdot)$ is smooth and thus, we can use an upper bound of $g^*(\cdot)$ for proximal iteration as below:

$$\alpha_t = \arg\max_\alpha \left[-c^\top \alpha - f^*(-A^\top \alpha) - g^*(-B^\top \alpha_{t-1}) - (-B z_t)^\top (\alpha - \alpha_{t-1}) - \frac{1}{2\eta_t} \|(\alpha - \alpha_{t-1})\|_2^2 \right],$$

where

$$z_t = \arg\min_z [\alpha_{t-1}^\top B z + g(z)] \quad \in \partial g^*(-B^\top \alpha_{t-1}).$$

We can apply the Moreau's Identity to turn optimization in $f^*(\cdot)$ to optimization in $f(\cdot)$. The smoothness of $g^*(\cdot)$ means that z_t is unique.

We can write first order condition for α_t: there exists $x_t \in \partial f^*(-A^\top \alpha_t)$ so that

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

Since $-A^\top \alpha_t \in \partial f(x_t)$, we obtain

$$0 \in (\partial f(x_t) + A^\top \alpha_{t-1}) + \eta_t A^\top (Ax_t + Bz_t - c),$$

which is the first order condition of the following optimization problem:

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

This leads to Algorithm 2.

Algorithm 2: Proximal Dual Ascent Method

Input: $\phi(\cdot)$, A, b, α_0, $\eta_1, \eta_2, \ldots$

Output: x_T

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

  Let $z_t = \arg\min_z[\alpha_{t-1}^\top Bz + g(z)]$

  Let $x_t = \arg\min_x[\alpha_{t-1}^\top Ax + 0.5\eta_t\|Ax + Bz_t - c\|_2^2 + f(x)]$

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

Return: x_T

Usually we assume that the closed form solution for z_t can be obtained easily. For some applications, solutions for x_t can also be obtained easily. For some other applications, we may use approximate solution, such as one or multiple steps of gradient descent as follows:

$$x_t = \arg\min_x \left[\alpha_{t-1}^\top Ax + f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + \frac{1}{2\tilde{\eta}_t}\|x - x_{t-1}\|_Q^2 + \frac{\eta_t}{2}\|Ax + Bz_t - c\|_2^2\right],$$

where we choose $Q = I - \tilde{\eta}_t \eta_t A^\top A$ such that the exact solution can be obtained as follows:

$$x_t = x_{t-1} - \tilde{\eta}_t A^\top[\tilde{\alpha}_t + \nabla f(x_{t-1})] \qquad \tilde{\alpha}_t = \alpha_{t-1} + \eta_t[Ax_{t-1} + Bz_t - c]. \tag{7}$$

Example 7 *If we apply Algorithm 1 to the concensus optimization problem, then we obtain (take α_0 such that $\sum_i [\alpha_0]_i = 0$)*

- $[x_t]_i = \arg\min_x[[\alpha_{t-1}]_i^\top x + f_i(x)]$ *($i = 1, \ldots, m$)*
- $z_t = m^{-1}\sum_{i=1}^m [x_t]_i$
- $[\alpha_t]_i = [\alpha_{t-1}]_i + \eta_t([x_t]_i - z_t)$ *($i = 1, \ldots, m$)*

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

Example 8 *If we apply Algorithm 1 to the generalized Lasso problem, we obtain (with $z_t = 0$)*

- $x_t = \arg\min_x[\alpha_{t-1}^\top Ax + f(x)]$

- $\alpha_t = \text{proj}_{\{x:\|x\|_\infty \le \mu\}}(\alpha_{t-1} + \eta_t A x_t)$

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-1} in closed form. One can also solve x_t approximately as in (7). The resulting method is similar to subgradient descent method, where α is the subgradient of $\mu\|\cdot\|_1$, which is aggregated instead of evaluated at x_{t-1} as in the standard subgradient descent.

In general, Algorithm 2 is preferred over Algorithm 1. It is proximal gradient applied to the dual problem. It is also possible to apply Nesterov's acceleration method. In practice, one often employs an improved version of Algorithm 2 called ADMM, which is closely related to the accelerated version of proximal dual gradient ascent. We will investigate ADMM in the next lecture.

4 Empirical Studies

While the main purpose of dual decomposition methods are to deal with complex problems (such as fused Lasso or TV-norm regularization) that cannot be handled directly by primal proximal gradient, in this study, we will investigate the simpler $L_1 - L_2$ regularization problem, so that its effectiveness can be compared to other methods.

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

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

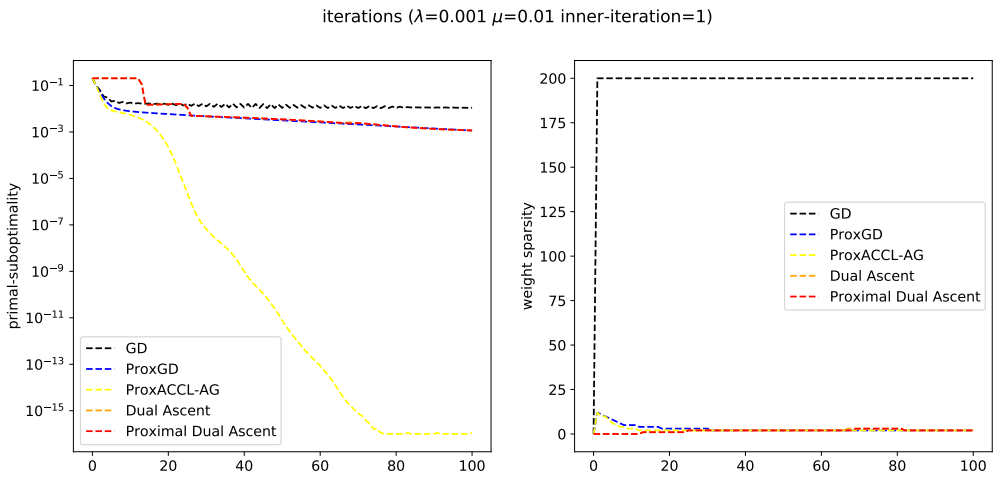
We compare Algorithm 2, Algorithm 1, to other algorithms, where we solve the optimization problem for x_t using multiple iterations of (7). Note that in our experiments, instead of using x_t as primal solution, we use z_t because z_t is sparser than x_t.

References

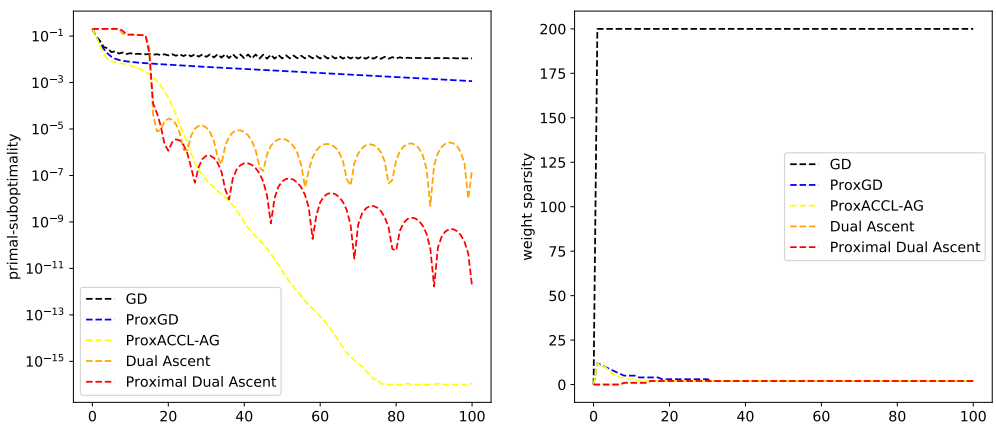
[1] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:485–516, 2008.

[2] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004.

[3] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused Lasso. *J. R. Statist. Soc. B*, 67:91–108, 2005.

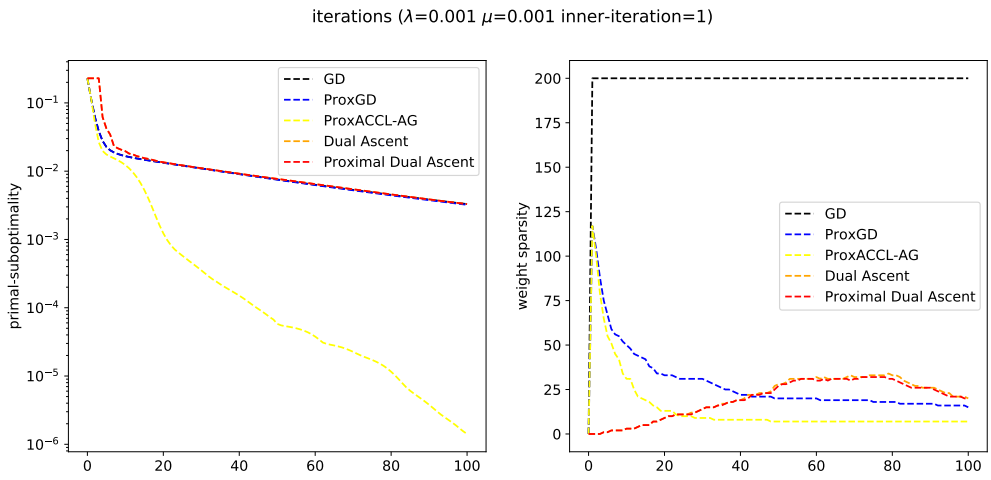


(a) 1 inner-iteration of (7)

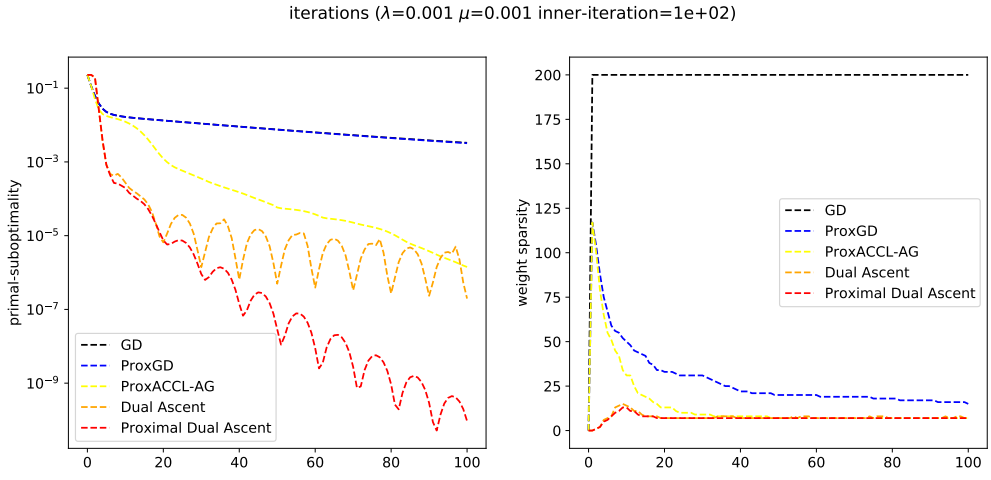


(b) 100 inner-iterations of (7)

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



(a) 1 inner-iteration of (7)



(b) 100 inner-iterations of (7)

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