Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 15: Lagrangian Duality and Dual Decomposition Methods

Constrained Optimization

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

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

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

Lagrangian Duality

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:

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

Weak Duality

We have the following weak duality theorem.

Theorem

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* $[\mu, \lambda]$*.*

Strong Duality

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.

We are particularly interested in the situation that there exist

- primal feasible x_*
- dual feasible $[\mu_*, \lambda_*]$

such that the duality gap is zero:

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

This situation is called *strong duality*.

Strong Duality and KKT Conditions

Theorem

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.

Example

Example

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

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

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

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

The dual problem is:

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

Dual Decomposition

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

$$\phi(x, z) = f(x) + g(z) \qquad \text{subject to } Ax + Bz = c. \tag{3}$$

Its dual is

$$\phi_D(\alpha) = -\alpha^\top c - f^*(-A^\top \lambda) - g^*(-B^\top \lambda) \qquad \lambda \in C_D, . \tag{4}$$

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

Example

Example

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.

Example

$$\min_{x,z} \sum_{i=1}^m f_i(x_i) \qquad \text{subject to } x_1 - z = x_2 - z = \cdots x_m - z = 0.$$

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

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

Generalized Lasso

Another example of decomposition is generalized Lasso.

Example

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

We introduce 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

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

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

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

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

Return: x_T

Dual Proximal Gradient

We apply proximal gradient method to the dual problem:

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

and use an upper bound of $g^*(\cdot)$ for proximal iteration as below:

$$\begin{aligned}\alpha_t = & \arg\max_{\alpha} \Big[-c^\top \alpha - f^*(-A^\top \alpha) \\ & -g^*(-B^\top \alpha_{t-1}) - (-Bz_t)^\top (\alpha - \alpha_{t-1}) - \frac{1}{2\eta_t} \|(\alpha - \alpha_{t-1})\|_2^2 \Big],\end{aligned}$$

where

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

We can apply the Moreau's Identity to turn optimization in $f^*(\cdot)$ to optimization in $f(\cdot)$.

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

This leads to Algorithm 2.

Dual Proximal Gradient

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

If closed form solution for x_t can not be obtained, one may also use the following iterative method one for more times:

$$x_t = x_{t-1} - \tilde{\eta}_t A^\top [\tilde{\alpha}_t + \nabla f(x_{t-1})] \qquad \tilde{\alpha}_t = \alpha_{t-1} + \eta_t [Ax_{t-1} + Bz_t - c]. \quad (5)$$

Example

Example

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

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

- $x_t = \arg\min_x [\alpha_{t-1}^\top A x + f(x)]$
- $\alpha_t = \mathrm{proj}_{\{x : \|x\|_\infty \le \mu\}}(\alpha_{t-1} + \eta_t A x_t)$

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

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, Algorithm 1, to other algorithms, where we solve the optimization problem for x_t using multiple iterations of (5). 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 $\mu = 10^{-3}$)

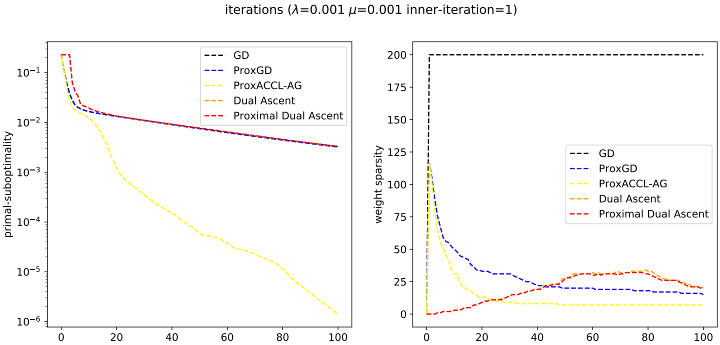

Figure: 1 inner-iteration of (5)

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

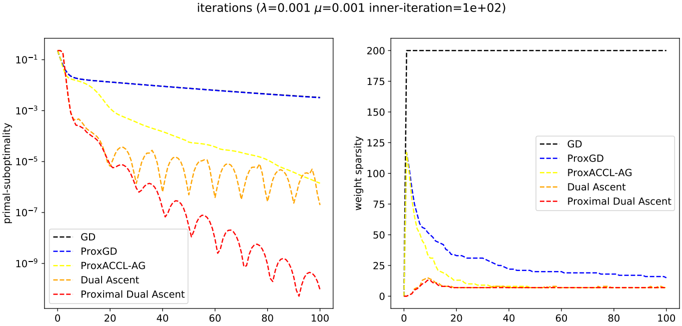

Figure: 100 inner-iterations of (5)

Summary

Lagrangian Duality

- Dual Formulation
- Weak and Strong Duality

Dual Decomposition

- Primal formulation: $\phi(x) = f(x) + g(z) \quad Ax + Bz = c$
- Dual formulation: $\phi_D(\alpha) = -f^*(-A^\top \alpha) - g^*(-B^\top \alpha) \quad \alpha \in C_D$

Dual Ascent Methods

- Dual Gradient Ascent
- Proximal Dual Graident Ascent