Homework 3 (Group 5)

HW3 (a)

$$\min f(w)$$
$$\text{s.t. } Aw=b$$

Define dual function $g(\lambda) = \inf_w L(w,\lambda) := f(w) + \lambda^T(Aw-b)$

For any given λ, denote $W(\lambda) = \arg\inf_w L(w,\lambda)$

Show that if $w \in W(\lambda)$, then $Aw-b \in \partial g(\lambda)$

proof: For any arbitrary $\bar{\lambda}$, we have that:

$$g(\bar{\lambda}) = \inf_{\bar{w}} \{f(\bar{w}) + \bar{\lambda}^T(A\bar{w}-b)\}$$
$$\le f(w) + \bar{\lambda}^T(Aw-b)$$
$$= f(w) + (\bar{\lambda}-\lambda)^T(Aw-b) + \lambda^T(Aw-b)$$
$$= \{g(\lambda) = f(w) + \lambda^T(Aw-b)\}$$
$$= g(\lambda) + (\bar{\lambda}-\lambda)^T(Aw-b)$$

Thus, we have shown that:

$$g(\bar{\lambda}) \le g(\lambda) + (\bar{\lambda}-\lambda)^T(Aw-b) \quad \forall \bar{\lambda}$$

which means $Aw-b \in \partial g(\lambda)$ ($g(\lambda)$ is concave)

HW3(b)

Analyze the convergence of dual ascent for L-smooth and μ-strongly convex f. Is the solution primal feasible?

Solution: For this problem since f is strongly convex, it holds that f^* is differentiable. Thus $g(\lambda)$ is differentiable and $Aw-b=\nabla g(\lambda)$.

Hence the following algorithm

$$\begin{cases} w_{k+1} \in \arg\min_w L(w,\lambda_k) \\ \lambda_{k+1} = \lambda_k + \alpha_k (Aw_k - b) \end{cases}$$

turns out to be dual gradient ascent.

In order to analyze the convergence of the above GD algorithm, firstly the property of $g(\lambda)$ should be investigated.

Claim: if f is strongly convex and x is its minimizer, then

$$f(y) \geq f(x) + \frac{\mu}{2}\|y-x\|^2 \quad \forall y \in \operatorname{dom} f \qquad (*)$$

Denote $F_u(x) = f(x) - u^T x$. Then we have the following properties:

① $f(x)$ is μ-strongly convex $\Rightarrow$ $F_u(x)$ is μ-strongly convex

② Denote $x_u \triangleq \nabla f^*(u)$. Since $\nabla f^*(u) = \arg\sup_w (u^T w - f(w))$, we know that x_u is the unique minimizer of $F_u(x)$.

Then by $(*)$ it holds that:

$$f(y) - u^T y \geq f(x_u) - u^T x_u + \frac{\mu}{2}\|y - x_u\|^2 \qquad \forall y$$

Denote $F_v(x) = f(x) - v^T x$ and $x_v \triangleq \nabla f^*(v)$, the same properties hold.

Thus we have:

$$\begin{cases} f(x_v) - u^T x_v \geq f(x_u) - u^T x_u + \frac{\mu}{2}\|x_v - x_u\|^2 \\ f(x_u) - v^T x_u \geq f(x_v) - v^T x_v + \frac{\mu}{2}\|x_v - x_u\|^2 \end{cases}$$

Combining the above equalities, it holds that:

$$\mu\|x_u - x_v\|^2 \leq (x_u - x_v)^T (u-v) \leq \|x_u - x_v\| \cdot \|u - v\|$$

$$\Rightarrow \|\nabla f^*(u) - \nabla f^*(v)\| \leq \frac{1}{\mu}\|u-v\|$$

Thus, f^* is $\frac{1}{\mu}$-smooth, which means that $-g(\lambda)$ is also $\frac{1}{\mu}$-smooth.

Now consider the problem

$$\max_{\lambda} g(\lambda)$$

where $-g(\lambda)$ is convex and $\frac{1}{\mu}$-smooth.

We know that the gradient ascent method with constant step-size $\alpha \le \mu$ converges at rate $O(\frac{1}{\varepsilon})$.

In addition, since the primal problem is convex with linear equality constraint, we know that strong duality holds.

Thus, the optimal solution of the dual problem also yields primal feasibility.

HW3 (c)

$$(P2) \quad \min \frac{1}{N}\sum_{i\in[N]} f_i(w_i) \quad \text{s.t. } w_i = w_j \text{ for all } j \in N_i$$

Solution: The constraint of local consensus is equivalent to

$$a_{ij}(w_i - w_j) = 0$$

for some doubly stochastic matrix $A = [a_{ij}]$ compatible with G.

Note that $a_{ij} = a_{ji}$ and $a_{ij} = 0$ if $(i,j) \notin \mathcal{E}$.

Then (P2) is equivalent to

$$(P2') \quad \min \frac{1}{N}\sum_{i\in[N]} f_i(w_i) \quad \text{s.t. } a_{ij}(w_i - w_j) = 0, \quad i,j = 1,\dots,N.$$

Then the Lagrangian for (P2') is

$$L(w,\lambda) = \frac{1}{N}\sum_{i=1}^{N} f_i(w_i) + \sum_{i=1}^{N}\sum_{j\neq i} \lambda_i^{(j)} a_{ij}(w_i - w_j)$$

$$= \sum_{i=1}^{N}\left[\frac{1}{N}f_i(w_i) + \sum_{j\neq i}\lambda_i^{(j)} a_{ij} w_i - \sum_{j\neq i}\lambda_j^{(i)} a_{ji} w_i\right]$$

$$= \sum_{i=1}^{N}\underbrace{\left[\frac{1}{N}f_i(w_i) + a_{ij}\sum_{j\neq i}(\lambda_i^{(j)} - \lambda_j^{(i)}) w_i\right]}_{\triangleq L_i(w_i,\lambda)}$$

(Note: the number of Lagrangian multipliers coincides with the number of constraints)

Thus, $L(w, \lambda)$ is seperable in w and dual decomposition can be applied.

$\boxed{\text{Dual Decomposition Algorithm}}$

- step 1 (primal update)

$$w_{i,k+1} \in \operatorname{argmin}_{w_i} L_i(w_i, \lambda_k) \qquad i=1,\cdots,N$$

- step 2 (dual update)

$$\lambda^{(j)}_{i,k+1} = \lambda^{(j)}_{i,k} + \alpha_k a_{ij}(w_{i,k} - w_{j,k}) \qquad i=1,\cdots,N, \quad j=1,\cdots,N$$

<u>Compare primal method and dual method</u>

1. communication cost

- primal: During each iteration, each note sends its value to the master and receives the values of its neighbors from the master.

$$\text{node } i: \quad [i] \underset{w_j\,(j\in N_i)}{\overset{w_i}{\rightleftarrows}} \text{Master}$$

- dual: During each iteration, each node sends its value to the master and receives the value of Lagrangian multipliers from the master.

$$\text{node } i: \quad [i] \underset{\lambda^{(j)}_i,\ \lambda^{(i)}_j\,(j\in N_i)}{\overset{w_i}{\rightleftarrows}} \text{Master}$$

Conclusion:

From the above analysis, we know that for (P2), usually the dual decomposition method requires more communication cost.

2. convergence

The convergence to a global minimum can be guaranteed for the primal method only when $f(w)$ is convex or initial guess w_0 is close to w^*. But for the dual method, since the dual problem is always convex regardless of f, convergence can be guaranteed under more general conditions.