EP3260: Fundamentals of Machine Learning over Networks
Homework Assignment 3, Due Date: March 13, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Problem 1. Consider the optimization problem on slide 11 of Lecture 6. Show that for convex and closed $f : A\boldsymbol{w} - \boldsymbol{b} \in \partial g(\boldsymbol{\lambda})$, where ∂ is the set of subgradients.

Proof.

Def. 1 (Subgradient). *We say a vector $\boldsymbol{c} \in \mathbb{R}^n$ is a subgradient of $g : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x} \in \boldsymbol{dom}\ g$ if for all $\boldsymbol{z} \in \boldsymbol{dom}\ g$,*

$$g(\boldsymbol{z}) \geq g(\boldsymbol{x}) + \boldsymbol{c}^T(\boldsymbol{z} - \boldsymbol{x}). \tag{1}$$

We know that $g(\boldsymbol{\lambda}) := \inf_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}) = f(\boldsymbol{w}) + \boldsymbol{\lambda}^T(A\boldsymbol{w} - \boldsymbol{b})$. Therefore,

$$g(\boldsymbol{\lambda}_1) = f(\boldsymbol{w}) + \boldsymbol{\lambda}_1^T(A\boldsymbol{w} - \boldsymbol{b}) \tag{2a}$$
$$g(\boldsymbol{\lambda}_2) = f(\boldsymbol{w}) + \boldsymbol{\lambda}_2^T(A\boldsymbol{w} - \boldsymbol{b}). \tag{2b}$$

Hence,

$$g(\boldsymbol{\lambda}_2) = g(\boldsymbol{\lambda}_1) + (A\boldsymbol{w} - \boldsymbol{b})^T(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1), \tag{3}$$

here the equality holds, which means that $(A\boldsymbol{w} - \boldsymbol{b}) \in \partial g(\boldsymbol{\lambda})$. □

Problem 2. Consider the dual ascent algorithm on slide 11 of Lecture 6. Analyze the convergence of dual ascent for L-smooth and μ-strongly convex f. Is the solution primal feasible?

Problem Description

Consider

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & z = f(\boldsymbol{w}) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b} \end{aligned} \tag{4}$$

Let the Lagrange dual function be

$$g(\boldsymbol{\lambda}) = \inf_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\lambda}) := f(\boldsymbol{w}) + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}). \tag{5}$$

A lower bound for problem (4) can then be found by the Dual Ascent Algorithm in which each iteration consists of two steps. Step one is to update the primal variable according to

$$\boldsymbol{w}_{k+1} \in \underset{\boldsymbol{w}}{\operatorname{argmin}}\, L(\boldsymbol{w}, \boldsymbol{\lambda}), \tag{6}$$

and step two is to update the dual variable according to

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha_k(\boldsymbol{A}\boldsymbol{w}_{k+1} - \boldsymbol{b}). \tag{7}$$

The problem is to analyze the convergence of dual ascent for L-smooth and μ-strongly convex f, and then to check if the solution is primal feasible.

Convergence Analysis - Primal Variable Update

The function $L(\boldsymbol{w}, \boldsymbol{\lambda})$ consists of the sum of two functions f and $\boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$. f is μ-strongly convex and L-smooth per the problem description, and $\boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$ is convex since it's an affine function in $\boldsymbol{w}$. As we showed in HW1 [1], the sum of a convex function and a μ-strongly convex function is still μ-strongly convex, so $L(\boldsymbol{w}, \boldsymbol{\lambda})$ is μ-strongly convex.

Similarly, we can prove that $L(\boldsymbol{w}, \boldsymbol{\lambda})$ is L-smooth. First, since $\boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$ is affine, it can be upper-bounded by a quadratic with no curvature, in other words it is Lipschitz-smooth with Lipschitz-constant zero. Let's define $g(\boldsymbol{x}) := \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$, and call it L_2-smooth. We can then use a condition for smoothness from HW1 [1] on both $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$ to get:

$$f(\boldsymbol{x}_2) \leq f(\boldsymbol{x}_1) + \nabla^T f(\boldsymbol{x}_1)(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{L}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \tag{8}$$

$$g(\boldsymbol{x}_2) \leq g(\boldsymbol{x}_1) + \nabla^T g(\boldsymbol{x}_1)(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{L_2}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \tag{9}$$

Taking the sum of (8) and (9) yields:

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = g(\boldsymbol{x}_2) + f(\boldsymbol{x}_2) \leq g(\boldsymbol{x}_1) + f(\boldsymbol{x}_1) + (\nabla g(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1))^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{L_2 + L}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \tag{10}$$

which is equivalent to saying that $L(\boldsymbol{w}, \boldsymbol{\lambda})$ is $L_2 + L$-smooth, but $L_2 = 0$ so it is L-smooth. In total, we now know that, just like f, $L(\boldsymbol{w}, \boldsymbol{\lambda})$ is a μ-strongly convex and L-smooth function. If we use gradient descent with constant step size to find the minimum of $L(\boldsymbol{w}, \boldsymbol{\lambda})$, it will converge in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ operations [2] (slide 19).

Convergence Analysis - Dual Variable Update

Since $\boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$ is in the set of subgradients for the lagrange dual function (as proven in the previous problem), the dual variable update is being performed with gradient ascent. So once again, if we can establish

strong convexity and Lipschitz-smoothness of the lagrange dual function $g(\boldsymbol{\lambda})$, the algorithm will converge in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ operations.

$g(\boldsymbol{\lambda})$ consists of the sum of two functions, f^* and $\boldsymbol{\lambda}^T\boldsymbol{b}$. As explained in the previous section, the sum of a strongly convex and Lipschitz-smooth function with an affine function is also strongly convex and Lipschitz-smooth. Since $\boldsymbol{\lambda}^T\boldsymbol{b}$ is affine in $\boldsymbol{\lambda}$ it is sufficient to show that f^* is strongly convex and Lipschitz-smooth.

For all μ-strongly convex functions f, the convex conjugate f^* is $\frac{1}{\mu}$-smooth [5] (slide 19). Also, due to the biconjugate property [5] (slide 13), the opposite holds as well. Using these two properties we can establish that f^* of the f in our problem is $\frac{1}{L}$-strongly convex and $\frac{1}{\mu}$-smooth. Thus, we have established that $g(\boldsymbol{\lambda})$ is strongly convex and Lipschitz smooth. Which means that the dual ascent algorithm will converge in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ operations.

Primal feasibility

Since the primal problem is a convex programming problem with no inequality constraints and only linear equality constraints, we know that Slater's condition holds and thus that we have strong duality [4] (theorem 1). Since we have strong duality, the solution is not only primal feasible but also primal optimal.

EP3260: Fundamentals of Machine Learning over Networks
Homework Assignment 3, Due Date: March 13, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Problem 3. Consider the optimization problem (P2) on slide 21 of Lecture 6. Extend the dual decomposition of Slide 6–12 to solve (P2). Compare it to the primal method (analytically or numerically) in terms of total communication cost and convergence rate on a random geometric communication graph.

Proof. Let us consider the problem

$$\begin{aligned} &\text{minimize } \frac{1}{N}\sum_{i\in[N]} f_i\left(w_i\right) \\ &\text{s.t. } w_i = w_j, \quad \text{for all } j \in \mathcal{N}_i \end{aligned} \tag{11}$$

The Lagrangian function in the dual formulation will be

$$\mathcal{L}(\lambda) = \frac{1}{N}\sum_{i\in[N]} f_i\left(w_i\right) - \sum_{i\in[N]}\sum_{j\in\mathcal{N}_i} \lambda_{i,j}^T(w_i - w_j) \tag{12}$$

where we denote the dual variable $\lambda_{i,j}\ \forall\ i,j \in \mathcal{V}$, and $\lambda := \left[\lambda_{11}^T,\ldots,\lambda_{1|\mathcal{N}_1|}^T,\lambda_{21}^T,\ldots,\lambda_{2|\mathcal{N}_2|}^T,\ldots,\lambda_{N|\mathcal{N}_N|}^T\right]^T$. The corresponding dual problem with respect to the primal variables w_i, $\forall i \in [N]$:

$$\mathcal{D}(\lambda) = \inf_{w_i:i\in[N]} L(\lambda) = \inf_{w_i:i\in[N]} \frac{1}{N}\sum_{i\in[N]} f_i\left(w_i\right) - \sum_{i\in[N]}\sum_{j\in\mathcal{N}_i} \lambda_{ij}^T(w_i - w_j). \tag{13}$$

We introduce the sub-problem:

$$\eta_i(\lambda) = \inf_{w_i} f_i\left(w_i\right) + \sum_{j\in\mathcal{N}_i}^{N} \lambda_{ij}^T w_i \tag{14}$$

and we note that each sub-problem $\eta_i(\lambda)$ can be solved locally at node i and independently of all other node $j \neq i$. The whole dual problem in Equation 13 can be written as the sum of these sub-problems

$$\mathcal{D}(\lambda) = \sum_{i\in[N]} \eta_i(\lambda) \tag{15}$$

We know that $\mathcal{D}(\lambda)$ is a lower bound of the solution in the primal space. We are interested in the tightest lower bound that is given by the solution of the the following of the dual problem

$$d^* = \sup_{\lambda} \mathcal{D}(\lambda) = \sup_{\lambda} \sum_{i=1}^{N} \eta_i(\lambda). \tag{16}$$

The dual problem can be solve with sub-gradient method, by computing:

$$\lambda_{ij}^{k+1} = \lambda_{ij}^{k} + \alpha_k g(\mathcal{D}(\lambda^k)) \tag{17}$$

where $g(\cdot)$ denotes the sub-gradient operator with respect to the dual variable λ. Since the subgradient can be computed and is equal to $g(\mathcal{D}(\lambda)) = (w_{i,k+1} - w_{j,k+1})$, where $w_{i,k+1}$ and is solution of the minimization problem $\eta_i(\lambda)$ evaluated at λ^k

$$\lambda_{ij}^{k+1} = \lambda_{ij}^{k} + \alpha_k\left(w_{i,k+1} - w_{j,k+1}\right) \tag{18}$$

Dual decomposition algorithm
for $k = 1, 2, \ldots,$ do
for $i \in [N]$ do
Master node broadcasts λ_{ij} for all $j \in \mathcal{N}_i$
Solve in parallel dual subproblems η_i finding $(w_{i,k+1}) \in \operatorname{arginf}_{w_i} \eta_i(\lambda)$
Master node collects $\boldsymbol{w}_{1,k+1}$ a and computes $\lambda_{ij}^{k+1} = \lambda_{ij}^{k} + \alpha_k\left(w_{i,k+1} - w_{j,k+1}\right)$

For what concerns the communication cost we have two contributions: the first concerns the distribution of the primal variables w_j such that $j \in \mathcal{N}_i$ and the second is about accumulating the dual variables λ_{ij}. The cost will be

$$\sum_{i=1}^{N}(|\mathcal{N}_{ij}| - 1) \cdot (|\lambda_{ij}| + |w_{i,k+1}|) \tag{19}$$

where we subtract 1 to the neighbour's cardinality because we exclude the cost of the node i that does not to be communicated. Compared to the transmission cost of the dual, where only the transmission of the $a_{i,j}$ weights for every node $i \in [n]$ is required, we can see that the dual solution requires approximately double of the information cost with respect to the primal method. □

References

[1] Machine Learning over Networks, homework 1. https://github.com/hshokrig/EP3260-MLoNs-2020/blob/master/Assignments/Homework1.pdf Visited 2020-03-10

[2] Machine Learning over Networks, lecture 2. https://github.com/hshokrig/EP3260-MLoNs-2020/blob/master/Lectures%20(2020)/Lecture2.pdf Visited 2020-03-10

[3] Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[4] Laurent El Ghaoui. Convex optimization, lecture 8. https://people.eecs.berkeley.edu/ elghaoui/Teaching/EE227A/lecture8.pdf Visited 2020-03-10

[5] Lieven Vandenberghe, Optimization Methods for Large-Scale Systems, lecture 5 http://www.seas.ucla.edu/ vandenbe/236C/lectures/conj.pdf Visited 2020-03-10