EP3260:Fundamentals of Machine Learning over Networks Homework 3(a) and 3(b)

Group 3

Notation

Upper-case letters with a double underline denotes matrices, e.g., $\boldsymbol{A}$. Lower-case letters with a single underline denotes vectors, e.g., $\boldsymbol{a}$. The i-th element of the vector $\boldsymbol{a}$ is denoted by either $a\,[i]$ or a_i, and element in the i-th row and j-th column of the matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}\,[i,j]$. An i-th column vector of a matrix $\boldsymbol{A}$ is denoted as either $\boldsymbol{A}_i$ or $\boldsymbol{A}[:,i]$.

HW3(a): Show that for convex and closed f: $\boldsymbol{Aw} - \boldsymbol{b} \in \partial g(\boldsymbol{\lambda})$, where $\partial g(\boldsymbol{\lambda})$ is the set of subgradients

The optimization problem is given by

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad f(\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{Aw} = \boldsymbol{b} \,. \tag{1}$$

The Lagrange dual function, utilizing the Lagrangian, can be written as

$$g(\boldsymbol{\lambda}) = \inf_{\boldsymbol{w}} \left\{ L(\boldsymbol{w}, \boldsymbol{\lambda}) \coloneqq f(\boldsymbol{w}) + \boldsymbol{\lambda}^{\mathrm{T}} (\boldsymbol{Aw} - \boldsymbol{b}) \right\} \tag{2}$$

The Lagrange dual problem, utilizing the conjugate function, can be expressed as

$$\underset{\boldsymbol{\lambda}}{\text{maximize}} \left\{ g(\boldsymbol{\lambda}) \coloneqq -f^* \left(-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}\right) - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} \right\} \tag{3}$$

There can at least be two ways to show that $\boldsymbol{Aw} - \boldsymbol{b} \in \partial g(\boldsymbol{\lambda})$:

- Utilizing the KKT conditions, the primal feasibility condition, in particular (2), yields that $\boldsymbol{Aw} - \boldsymbol{b} \in \partial g(\boldsymbol{\lambda})$ since f is closed and convex function. In other words, taking the derivative of the Lagrangian

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) \coloneqq f(\boldsymbol{w}) + \boldsymbol{\lambda}^{\mathrm{T}} (\boldsymbol{Aw} - \boldsymbol{b}) \tag{4}$$

with respect to the dual variable $\boldsymbol{\lambda}$ yields that

$$\boldsymbol{Aw} - \boldsymbol{b} \in \partial g(\boldsymbol{\lambda}) \,.$$

- Given $f : \mathbb{R}^n \to \mathbb{R}$, the conjugate function reads

$$f^*(\boldsymbol{y}) = \max_{\boldsymbol{x}} \ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} - f(\boldsymbol{x}) = -\left\{ \min_{\boldsymbol{x}} f(\boldsymbol{x}) - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} \right\} \,. \tag{5}$$

For a closed and convex function f, according to Fenchel duality theorem, any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\boldsymbol{x} \in \partial f^*(\boldsymbol{y}) \Longleftrightarrow \boldsymbol{y} \in \partial f(\boldsymbol{x}) \Longleftrightarrow \boldsymbol{x} \in \arg\min_{\boldsymbol{z}} f(\boldsymbol{z}) - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{z} \,. \tag{6}$$

To this end, if we take the derivative of (3) with respect to $\boldsymbol{\lambda}$ and utilizing the chain rule, i.e.,

$$\partial g(\boldsymbol{\lambda}) = \boldsymbol{A}\, \partial f^* \left(-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}\right) - \boldsymbol{b} \,. \tag{7}$$

Since $\boldsymbol{w} \in \partial f^* \left(\underbrace{-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}}_{=\,\boldsymbol{y}} \right)$, then $\boldsymbol{Aw} - \boldsymbol{b} \in \partial g(\boldsymbol{\lambda})$, where

$$\boldsymbol{w} \in \arg\min_{\boldsymbol{z}} f(\boldsymbol{z}) - \left(-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}\right)^{\mathrm{T}} \boldsymbol{z} = \arg\min_{\boldsymbol{z}} f(\boldsymbol{z}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{Az} \,.$$

HW3(b): Analyze the convergence of dual ascent for L-smooth and μ-strongly convex f. Is the solution primal feasible?

The dual ascent scheme for the problem (1) can be summarized as

$$\boldsymbol{w}^{(k+1)} = \arg\min_{\boldsymbol{w}} L\left(\boldsymbol{w}, \boldsymbol{\lambda}^{(k)}\right) \in \arg\min_{\boldsymbol{w}} f(\boldsymbol{w}) + \left(\boldsymbol{\lambda}^{(k)}\right)^{\mathrm{T}} \boldsymbol{A}\boldsymbol{w} \tag{8}$$

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + \alpha^{(k)} \left(\boldsymbol{A}\boldsymbol{w}^{(k+1)} - \boldsymbol{b}\right) , \tag{9}$$

where $L\left(\boldsymbol{w}, \boldsymbol{\lambda}^{(k)}\right)$ is defined in (4).

For brevity, here onwards, we redefine $\boldsymbol{x} \coloneqq \boldsymbol{w}$ and $\boldsymbol{u} \coloneqq \boldsymbol{\lambda}$.

Some preliminaries are presented in order to support the convergence of dual ascent scheme.

Since f is (strongly) convex, then according to Fenchel duality theorem, any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\boldsymbol{x} \in \partial f^*(\boldsymbol{y}) \iff \boldsymbol{y} \in \partial f(\boldsymbol{x}) \iff \boldsymbol{x} \in \arg\min_{\boldsymbol{z}} f(\boldsymbol{z}) - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{z} . \tag{10}$$

In addition, f is strongly convex, i.e., $\partial f^*(\boldsymbol{y}) = \{\nabla f^*(\boldsymbol{y})\}$, then

$$\boldsymbol{x} = \nabla f^*(\boldsymbol{y}) = \arg\min_{\boldsymbol{z}} \left\{ f(\boldsymbol{z}) - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{z} \right\} . \tag{11}$$

On the other hand, if a function, say $g(\boldsymbol{x})$, is also μ-strongly convex (where $\mu > 0$),

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 . \tag{12}$$

If $\boldsymbol{x}$ is a minimizer, then $\nabla g(\boldsymbol{x}) = 0$ yields

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 . \tag{13}$$

To this end, we can define $g_{\boldsymbol{u}}(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}$ such that the minimizer $\boldsymbol{x}_u = \nabla f^*(\boldsymbol{u})$, see (11), then for any $\boldsymbol{y}$

$$\underbrace{g(\boldsymbol{y})}_{=f(\boldsymbol{y}) - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{y}} \geq \underbrace{g(\boldsymbol{x})}_{=f(\boldsymbol{x}_u) - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_u} + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{14}$$

$$f(\boldsymbol{y}) - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{y} \geq f(\boldsymbol{x}_u) - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_u + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 . \tag{15}$$

Similarly, we define $g_{\boldsymbol{v}}(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}$ such that the minimizer $\boldsymbol{x}_v = \nabla f^*(\boldsymbol{v})$, then

$$f(\boldsymbol{y}) - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{y} \geq f(\boldsymbol{x}_v) - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}_v + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 . \tag{16}$$

Now, plugin $\boldsymbol{y} = \boldsymbol{x}_v$ in (15) and $\boldsymbol{y} = \boldsymbol{x}_u$ in (16), and thereby add them such that,

$$(\boldsymbol{u} - \boldsymbol{v})^{\mathrm{T}} (\boldsymbol{x}_u - \boldsymbol{x}_v) \geq \mu \|\boldsymbol{x}_u - \boldsymbol{x}_v\|_2^2 . \tag{17}$$

Now, we invoke Cauchy-Schwartz inequality on the left hand side and rearrange such that

$$\Bigg\| \underbrace{\boldsymbol{x}_u}_{=\nabla f^*(\boldsymbol{u})} - \underbrace{\boldsymbol{x}_v}_{=\nabla f^*(\boldsymbol{v})} \Bigg\|_2 \geq \frac{1}{\mu} \|(\boldsymbol{u} - \boldsymbol{v})\|_2 \tag{18}$$

$$\|\nabla f^*(\boldsymbol{u}) - \nabla f^*(\boldsymbol{v})\|_2 \geq \frac{1}{\mu} \|(\boldsymbol{u} - \boldsymbol{v})\|_2 . \tag{19}$$

Thus, if f is μ-strongly convex, then ∇f^* is Lipschitz with parameter $\frac{1}{\mu}$.

To this end, applying what we know about the convergence rate of (primal) gradient descent.

- If f is μ-strongly convex, then the dual ascent with constant step size $\alpha^{(k)} \leq \mu$ converges at sublinear rate $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ same as the primal gradient descent.

- If f is μ-strongly convex and also L-smooth, i.e., ∇f is Lipschitz, then the primal gradient descent converges at linear rate $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$.

- Since we have shown above that if f is μ-strongly convex, then ∇f^* is Lipschitz with parameter $\frac{1}{\mu}$, where the converse is also true, i.e., if ∇f^* is Lipschitz with parameter $\frac{1}{\mu}$ then f is μ-strongly convex. For closed and convex f, the dual of the dual is primal, i.e., $f^{**} = f$. So, if f has Lipschitz gradient and it is strongly convex, then it is also true for f^*. Thus, dual ascent also converges at the linear rate of $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$.

References

[1] Bubeck, Sbastien, ”Convex optimization: Algorithms and complexity,” Foundations and Trends in Machine Learning, vol. 8, no.3–4, pp. 231–357, 2015.

[2] L. Bottou, F. Curtis, J. Norcedal, ”Optimization Methods for Large-Scale Machine Learning,” SIAM Rev., 60, no. 2, pp. 223-311, 2018.

[3] T. Hastie, R. Tibshirani, and J. Friedman, ”The Elements of Statistical Learning: Data Mining, Inference and Prediction,” Second edition, Springer, 2009.

[4] C. D. Meyer, ”Matrix Analysis and Applied Linear Algebra,” SIAM, 2000.

[5] J. R. Magnus and H. Neudecker, ”Matrix differential calculus with applications in statistics and econometrics,” 2nd Edition, Wiley, UK, 1999.

[6] K. P. Murphy, ”Machine learning - a probabilistic perspective,” MIT Press, 2012.

[7] M. Schmidt,N. Le Roux, F. Bach, ”Minimizing finite sums with the stochastic average gradient,” Mathematical Programming, 2017.

[8] R. Johnson and T. Zhang, ”Accelerating stochastic gradient descent using predictive variance reduction,” NIPS, 2013.

HW 3(c)

Group 3

February 20, 2019

Problem Description Extend the dual decomposition of the following problem

$$\min \frac{1}{N}\Sigma_{i\in N} f_i(w_i) \tag{1}$$

$$\text{such that } w_i = w_j \qquad\qquad \forall i \in \mathcal{N}_i \tag{2}$$

To extend the dual decomposition of the above we first properly convert the equality constraint. Note that here $A = \texttt{lap}(G)$ such that $Aw = 0$ is a correct constraint, where the laplacian of a graph is defined as the following

$$\texttt{lap}(G) = D - Adj(G)$$

where D is a diagonal matrix with each entry $D_{ii} = deg(i)$. Where $deg(i)$ is the degree of node of node i. And $Adj(G)$ is the adjacency matrix of the network G. For example for a triangle graph between three nodes $1, 2, 3$, $\mathcal{L}(G)$ looks like the following

$$\texttt{lap}(\Delta) = \begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix}$$

Hence the problem becomes

$$\min \frac{1}{N}\Sigma_{i\in N} f_i(w_i) \tag{3}$$

$$\text{such that } L_i w_i = 0 \qquad\qquad \forall i \in \mathcal{N}_i \tag{4}$$

Where L_i is the i^{th} column vector of $\texttt{lap}(G)$

Dual Decomposition Here the dual of the problem is

$$\mathcal{L}(w, \lambda) = \sum_{i \in N} f_i(w_i) + \lambda^T A_i w - \frac{1}{N} \lambda^T b$$

Using the parameters from our problem this becomes

$$\mathcal{L}(w, \lambda) = \sum_{i \in N} f_i(w_i) + \lambda^T L_i w$$

Since the langrangian as described above is separable thus we can proceed with the usual Lagrangian separation on different nodes. Firstly we give the primal step.

Primal update for optmization In the primal step, each node minimizes $L_i(w_i, \lambda_k)$, that is for every node we need to update w_i^k which is given by the following

$$w_i^{k+1} \in \texttt{argmin}_{w_i} L_i(w_i, \lambda_k)$$

for $i \in \{1, \ldots, N\}$.

Dual update for optmization The dual update need to gather all the w_i within a connected component and update the dual as follows

$$\lambda_C^{k+1} = \lambda_C^k + \alpha_C^k \sum_{i \in [C]} A_i w_i^k$$

where C is a connected component. Recall that a connected component in an undirected graph is a set of nodes which has a path to each other. Note that here, we for each component we can optimize the problem independently.

1 Convergence cost in Random Geometric Graph

A random geometric graph $G(n, r)$ is a graph of n nodes where two nodes are connected to each other if the distance between them is less than r.

Recall that the diameter $diam(G)$ of a graph is the maximum eccentricity of any vertex in the graph. That is, $diam(G)$ is the greatest distance between any pair of vertices or, alternatively, $d = \max_{v \in V} \epsilon(v) d = \max_{v \in V} \epsilon(v) =$

$\max_{v\in V}\max_{u\in V} d(v,u)$, where $d(u,v)$ is the distance between any two nodes u and v.

When a random graph is connected, then we need all the nodes in the graph to agree on the same w_i. This is similar to the *average consensus building problem*. But the interesting case is when $G(n,r)$ is disconnected. Here each update of the dual variable cost the $diam(C)$, where C is component. The idea is as follows, every node updates its w_i and sends to a leader node of a C along a tree T edges. This tree T could be a BFS tree. And the cost is equal to the $Depth(T)$ This causes the leader to receive the sum in $Depth(T) = O(Diam(G))$ rounds.

References