Homework 3 — Group 2

Ali Bemani **Ali.Bemani@hig.se**
Oscar Bautista Gonzalez **Oscar.Bautista.Gonzalez@hig.se**
Martin Hellkvist **Martin.Hellkvist@angstrom.uu.se**

Problem 3.1

Consider the optimization problem on slide 11 of Lecture 6. Show that for convex and closed $f : Aw - b \in \partial g(\lambda)$ *where* ∂ *is the set of subgradients*

Solution: We have the Lagrange dual function of the minimization problem as:

$$g(\lambda) = \inf_{w \in \mathbf{dom} f} f(w) + \lambda^{\mathrm{T}}(Aw - b). \tag{1}$$

Using that the domain of f is convex and closed we can equivalently use the minimum, rather than the infimum:

$$g(\lambda) = \min_{w \in \mathbf{dom} f} f(w) + \lambda^{\mathrm{T}}(Aw - b). \tag{2}$$

We let w_λ denote the minimizer of $f(w) + \lambda^{\mathrm{T}}(Aw - b)$ for a given λ, so that

$$g(\lambda) = f(w_\lambda) + \lambda^{\mathrm{T}}(Aw_\lambda - b). \tag{3}$$

By the rules of subgradients, we can express $\partial g(\lambda)$ as

$$\begin{align} \partial g(\lambda) &= \partial f(w_\lambda) + \partial(\lambda^{\mathrm{T}}(Aw_\lambda - b)) \tag{4}\\ &= \partial(\lambda^{\mathrm{T}}(Aw_\lambda - b)) \tag{5}\\ &= Aw_\lambda - b, \tag{6} \end{align}$$

where in the last step we use that $\lambda^{\mathrm{T}}(Aw_\lambda - b)$ is differentiable, so the subgradient is equal to the gradient.

We have showed that for a given λ, $Aw_\lambda - b$ is a subgradient of $g(\lambda)$, where w_λ is a minimizer of $f(w) + \lambda^{\mathrm{T}}(Aw - b)$.

Problem 3.2

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

Solution: We will show that for a μ-strongly convex and L-smooth differentiable function f, the conjugate function f^* is $\frac{1}{\mu}$-smooth and $\frac{1}{L}$-strongly convex. Then we analyze the convergence of gradient ascent on the dual function $g(\lambda)$, using convergence properties of gradient descent.

Convexity and smoothness of the conjugate function

For a differentiable function f, and its conjugate function f^*, we have for an arbitrary z that

$$f^*(y) = z^{\mathrm{T}} \nabla f(z) - f(z), \tag{7}$$

if $y = \nabla f(z)$.

Define the following variables:

$$x_y = \nabla f^*(y), \quad y = \nabla f(x_y), \tag{8}$$
$$x_z = \nabla f^*(z), \quad z = \nabla f(x_z). \tag{9}$$

Using that f is μ-strongly convex, we have:

$$f(x_y) \geq f(x_z) + \nabla f(x_z)^{\mathrm{T}}(x_y - x_z) + \frac{\mu}{2}\|x_z - x_y\|_2^2, \tag{10}$$
$$f(x_z) \geq f(x_y) - \nabla f(x_y)^{\mathrm{T}}(x_y - x_z) + \frac{\mu}{2}\|x_z - x_y\|_2^2. \tag{11}$$

By addition of the two inequalities and simplifying, we obtain

$$\mu\|x_z - x_y\|_2^2 \leq \nabla f(x_z)^{\mathrm{T}}(x_z - x_y) - \nabla f(x_y)(x_z - x_y) \tag{12}$$
$$= (\nabla f(x_z) - \nabla f(x_y))^{\mathrm{T}}(x_z - x_y) \tag{13}$$
$$\leq \|\nabla f(x_z) - \nabla f(x_y)\|_2 \|x_z - x_y\|_2, \tag{14}$$

where in the last step we used Cauchy-Schwarz inequality. Dividing by $\|x_z - x_y\|_2$, we obtain:

$$\|x_z - x_y\|_2 \leq \frac{1}{\mu}\|\nabla f(x_z) - \nabla f(x_y)\|_2. \tag{15}$$

Plugging in the definitions in (8)-(9), we conclude that f^* is $\frac{1}{\mu}$-smooth, by the definition in (2) of Homework 1:

$$\|\nabla f^*(z) - \nabla f^*(y)\|_2 \leq \frac{1}{\mu}\|z - y\|_2. \tag{16}$$

This can be directly seen by using μ-strong convexity and the implied inequality from Homework 1, question 1.b), but this proof shows quite clearly step by step what is happening.

We now prove the converse relation: if f^* is L-smooth, then f is $\frac{1}{L}$-strongly convex. Assume the f^* is L-smooth. From the property of smoothness in Homework 1, question 2.c) we have:

$$\frac{1}{L}\|\nabla f^*(y) - \nabla f^*(z)\|_2^2 \leq (\nabla f^*(y) - \nabla f^*(z))^{\mathrm{T}}(y - z). \tag{17}$$

Plugging in the definitions in (8) - (9), we obtain

$$\frac{1}{L}\|x_y - x_z\|_2^2 \leq (x_y - x_z)^{\mathrm{T}}(\nabla f(x_y) - \nabla f(x_z)), \tag{18}$$

which is equivalent to f being $1/L$-strongly convex, by the second equivalence of smoothness in Homework 1, question 1.

We have now proven that for a differentiable function f we have that "f is μ-strongly convex" $\Leftrightarrow$ "f^* is $\frac{1}{\mu}$-smooth". Because $f = f^{**}$, we have that "f^* is $\frac{1}{L}$-strongly convex" $\Leftrightarrow$ "f is L-smooth".

So for a differentiable μ-strongly convex and L-smooth function f, its conjugate f^* is $\frac{1}{L}$-strongly convex and $\frac{1}{\mu}$-smooth.

Convergence Discussion

Because the conjugate function is always convex, we have that the Lagrange dual function $g(\lambda)$ is concave.

Performing gradient ascent on the concave function g is equivalent of doing gradient descent on $-g$. We have that $-g(\lambda) = f^*(-A^{\mathrm{T}}\lambda) + \lambda^{\mathrm{T}}b$, which is is $\frac{1}{L}$-strongly convex and $\frac{1}{\mu}$-smooth. The convergence rate of gradient ascent on g is therefore linear when the step-size α_k is chosen appropriately as $\alpha_k = 2/(\frac{1}{L} + \frac{1}{\mu})$.

The solution is primal feasible because when the algorithm is converging the gradient is vanishing, i.e., $Aw_k - b \to 0$, meaning that the primal constraint is fulfilled.

Problem 3.3

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.

Solution: We first clarify some notation to use:

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} \in \mathbb{R}^{(Nn)\times 1},\ w_i \in \mathbb{R}^{n\times 1}, \tag{19}$$

where w is the vector of all the N parameter vectors w_i in the network. All the local parameter vectors are of dimension n for this problem. The set of neighbors to node i is denoted by $\mathcal{N}_i$, and the number of neighbors to node i is denoted by p_i.

We introduce a new structure of parameter vectors to solve the problem, $\bar{w}_i$, constructed by the local parameters as well as the neighbors' parameters:

$$\bar{w}_i = \begin{bmatrix} w_i \\ w_i^{(1)} \\ \vdots \\ w_i^{(p_i)} \end{bmatrix}, \tag{20}$$

where $w_i^{(j)}$ is the parameter of the j^{th} neigbor of node i. We now reformulate the local constraints $w_i = w_j, \forall j \in \mathcal{N}_i$ in matrix form:

$$A_i \bar{w}_i = 0, \tag{21}$$

where the matrix $A_i \in \mathbb{R}^{n\times((p_i+1)n)}$ is defined as

$$A_i = \begin{bmatrix} I_n & -I_n & 0 & 0 & \cdots & 0 \\ I_n & 0 & -I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ I_n & 0 & 0 & \cdots & -I_n & 0 \\ I_n & 0 & 0 & \cdots & 0 & -I_n \end{bmatrix} \tag{22}$$

$$= \begin{bmatrix} I_n & \\ \vdots & I_{p_i n} \\ I_n & \end{bmatrix} \in \mathbb{R}^{(np_i)\times(n(p_i+1))} \tag{23}$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix. We will study an iterative problem, so we introduce a communication delay between neighbors as:

$$\bar{w}_{i,k+1} = \begin{bmatrix} w_{i,k+1} \\ w_{i,k}^{(1)} \\ \vdots \\ w_{i,k}^{(p_i)} \end{bmatrix}. \tag{24}$$

We get the local Lagrange function

$$L_i(\bar{w}_i, \lambda) = f_i(\bar{w}_i) + \lambda^{\mathrm{T}} A_i \bar{w}_i, \tag{25}$$

The presented formulation of A_i makes it possible for different number of rows in A_i over different nodes. This can be solved by for example zero padding the smaller matrices and corresponding $\bar{w}_i$ so to obtain the same dimensions of A_i $\forall\, i$.

The algorithm becomes:

Step 1: $\bar{w}_{i,k} \in \arg\min_{\bar{w}_i} L_i(\bar{w}_i, \lambda_k)$,

Step 2: $\lambda_{k+1} = [\lambda_k + \alpha_k(\sum_{i=1}^{N} A_i \bar{w}_{i,k})]_+$.