Exercises
Computational Intelligence Lab
SS 2017

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://cil.inf.ethz.ch/

Series 4 Solutions (Optimization)

Problem 1 (Lagrange Dual Problem for a Linear Program):

1. Recall $\mathbf{x} \in \mathbb{R}^m$, and $A \in \mathbb{R}^{p \times m}$. The *Lagrangian* is defined as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathbf{c}^T\mathbf{x} - \sum_{i=1}^{m} \lambda_i x_i + \boldsymbol{\nu}^T(A\mathbf{x} - \mathbf{b}) = -\mathbf{b}^T\boldsymbol{\nu} + (\mathbf{c} + A^T\boldsymbol{\nu} - \boldsymbol{\lambda})^T\mathbf{x}$$

   where $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\nu} \in \mathbb{R}^p$. Note that later on, we'll be interested in values of the Lagrangian where $\boldsymbol{\lambda} \geq 0$, however, $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is well defined for arbitrary values $\boldsymbol{\lambda}, \boldsymbol{\nu}$.

2. The *dual function* is defined as

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \;=\; \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \;=\; -\mathbf{b}^T\boldsymbol{\nu} + \inf_{\mathbf{x}} (A^T\boldsymbol{\nu} - \boldsymbol{\lambda} + \mathbf{c})^T\mathbf{x}$$

3. The analytical expression for the dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is
   $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = -\infty$ except when $A^T\boldsymbol{\nu} - \boldsymbol{\lambda} + \mathbf{c} = 0$, in which case it is $-\mathbf{b}^T\boldsymbol{\nu}$:

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\mathbf{b}^T\boldsymbol{\nu} & \text{if } A^T\boldsymbol{\nu} - \boldsymbol{\lambda} + \mathbf{c} = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

4. The Lagrange *dual problem* of the standard form LP is to maximize the dual function d subject to $\boldsymbol{\lambda} \geq 0$:

$$\begin{aligned} \underset{\boldsymbol{\lambda},\boldsymbol{\nu}}{\text{maximize}} \quad & d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\mathbf{b}^T\boldsymbol{\nu} & \text{if } A^T\boldsymbol{\nu} - \boldsymbol{\lambda} + \mathbf{c} = 0 \\ -\infty & \text{otherwise.} \end{cases} \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq 0. \end{aligned}$$

   Because d is finite only when $A^T\boldsymbol{\nu} - \boldsymbol{\lambda} + \mathbf{c} = 0$, we can form an equivalent problem by making these constraints explicit:

$$\begin{aligned} \underset{\boldsymbol{\lambda},\boldsymbol{\nu}}{\text{maximize}} \quad & -\mathbf{b}^T\boldsymbol{\nu} \\ \text{subject to} \quad & A^T\boldsymbol{\nu} - \boldsymbol{\lambda} + \mathbf{c} = 0 \\ & \boldsymbol{\lambda} \geq 0. \end{aligned}$$

   This is a linear program again. Finally, we can eliminate the $\boldsymbol{\lambda}$ variable by observing $\boldsymbol{\lambda} = A^T\boldsymbol{\nu} + \mathbf{c}$, which then leads to the dual linear program to the input problem:

$$\begin{aligned} \underset{\boldsymbol{\nu}}{\text{maximize}} \quad & -\mathbf{b}^T\boldsymbol{\nu} \\ \text{subject to} \quad & A^T\boldsymbol{\nu} + \mathbf{c} \geq 0. \end{aligned}$$

   This resulting dual linear program has m inequality constraints - as many as the original problem had variables.

Problem 2 (Dual Function is a Lower Bound on $f(\mathbf{x}^\star)$):

Recall that the general convex optimization problem was given in the following form:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \le 0, \quad i = 1, \dots, m \\ & h_j(\mathbf{x}) = 0, \quad j = 1, \dots, p \end{array}$$

We will here show a stronger statement (as given in the lecture slides). We will show that if $\boldsymbol{\lambda} \ge 0$ (and $\boldsymbol{\nu}$ arbitrary), then

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \le f(\tilde{\mathbf{x}}) \quad \text{for any feasible point } \tilde{\mathbf{x}} \tag{1}$$

(and 'feasible' meaning that $\tilde{\mathbf{x}}$ satisfies all constraints $g_i(.)$ and $h_j(.)$).

Note: Since an optimal $\mathbf{x}^\star$ must of course be feasible this statement will imply that $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \le f(\mathbf{x}^\star)$, as asked in the exercise.

Proof: Let $\tilde{\mathbf{x}}$ be a feasible point for the problem, i.e., $g_i(\tilde{\mathbf{x}}) \le 0$ and $h_j(\tilde{\mathbf{x}}) = 0$. Then, using that the Lagrange multipliers $\boldsymbol{\lambda}$ corresponding to the inequality constraints do satisfy $\boldsymbol{\lambda} \ge 0$, we have

$$\sum_{i=1}^{m} \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_{j=1}^{p} \nu_j h_j(\tilde{\mathbf{x}}) \le 0,$$

since each term in the first sum is nonpositive, and each term in the second sum is zero, and therefore

$$L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \;=\; f(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_{j=1}^{p} \nu_j h_j(\tilde{\mathbf{x}}) \;\le\; f(\tilde{\mathbf{x}}).$$

Hence

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \;=\; \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \;\le\; L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \;\le\; f(\tilde{\mathbf{x}}).$$

As this holds for any feasible $\tilde{\mathbf{x}}$, it will also hold for an optimal $\mathbf{x}^\star$, as we mentioned above.

Problem 3 (Convexity):

a) yes

b) no. (For example, take a point on the circle and its negative, and observe that their mid-point is not in the set)

c) no, it's $\le$

d) no. (For example, take a point $(\mathbf{u}, \mathbf{v}) = ([-1, 1], [1, -1])$ and consider its negative. Then observe that at their mid-point, we have a strictly higher function value, violating convexity of the function)

Problem 4 (SGD for Collaborative Filtering):

Consider the given objective function as a sum

$$f(\mathbf{U},\mathbf{Z}) = \frac{1}{|\Omega|}\sum_{(d,n)\in\Omega} \underbrace{\frac{1}{2}\big[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\big]^2}_{f_{d,n}}$$

and $\mathbf{U} \in Q_1 := \mathbb{R}^{D\times K}$, $\mathbf{Z} \in Q_2 := \mathbb{R}^{N\times K}$.

- **Stochastic Gradient:** For one fixed element (d,n) of the sum, we derive the gradient entry (d',k) of $\mathbf{U}$, that is $\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U},\mathbf{Z})$, and analogously for the $\mathbf{Z}$ part.

$$\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U},\mathbf{Z}) = \begin{cases} -\big[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\big] v_{n,k} & \text{if } d' = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial v_{n',k}} f_{d,n}(\mathbf{U},\mathbf{Z}) = \begin{cases} -\big[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\big] u_{d,k} & \text{if } n' = n \\ 0 & \text{otherwise} \end{cases}$$

- **Full Gradient:** We have access to all elements $(d,n)\in\Omega$, so we can calculate the partial derivatives of the full gradient for all $(d,n)\in\Omega$. For one specific $(d,n)\in\Omega$, the partial derivatives are the same as that in the stochastic gradient above.