Assignment 3: Duality and Backpropagation

AIGS/CSED515 Machine Learning
Instructor: Jungseul Ok
jungseul@postech.ac.kr

Due: 20:00pm Nov 5 , 2020

Remarks

- Group study and open discussion via LMS board are encouraged, however, assignment that your hand-in must be **of your own work**, and **hand-written** unless you're asked a coding task.
- Submit a scanned copy of your answer on LMS online in **a single PDF file**, to which you also print and add your code if apply, i.e., no zip file, just a single PDF containing everything.
- Delayed submission may get some penalty in score: 5% off for delay of $0 \sim 4$ hours; 20% off for delay of $4 \sim 24$ hours; and delay longer than 24 hours will not be accepted.

1. [8pt; Practice KKT] We learned how to make Lagrange dual problem from a constrained optimization problem, in which Karush-Kuhn-Tucker (KKT) conditions provide a set of necessary conditions on the optimization solution, c.f.,

https://www.cs.cmu.edu/ ggordon/10725-F12/slides/16-kkt.pdf
http://www.stat.cmu.edu/ ryantibs/convexopt-F16/scribes/kkt-scribed.pdf

Practice the use of KKT conditions (stationary, feasibility, and complementary slackness) and duality with the following optimization problems:

(a) [7pt] Write Lagrange function for the following minimization problem with dual variable μ. Write the stationary condition, and extract the property of (w_1, w_2) to guarantee the existence of μ. Combining the above observation and the feasibility condition, find the properties on (w_1, w_2). Check if $\mu > 0$ with (w_1, w_2) verifying all the above properties. State the complementary slackness condition, and specify optimal solution (w_1^*, w_2^*) from the KKT conditions.

$$\begin{aligned} \underset{w_1,w_2}{\text{minimize}} \quad & \sqrt{w_1^2 + w_2^2} \\ \text{subject to} \quad & w_1 + 2w_2 \geq 5 \end{aligned}$$

sol) We first write the corresponding Lagrange function:

$$\mathcal{L}(w_1, w_2, \mu) = \sqrt{w_1^2 + w_2^2} + \mu\,(5 - w_1 - 2w_2) \ .$$

The stationary condition, i.e., $\nabla\mathcal{L}(w_1, w_2, \mu) = 0$, includes

$$\begin{aligned} \frac{\partial}{\partial w_1}\mathcal{L}(w_1, w_2, \mu) &= \frac{w_1}{\sqrt{w_1^2 + w_2^2}} - \mu = 0 \ , \\ \frac{\partial}{\partial w_2}\mathcal{L}(w_1, w_2, \mu) &= \frac{w_2}{\sqrt{w_1^2 + w_2^2}} - 2\mu = 0 \ . \end{aligned}$$

For the existence of μ, we must have $2w_1 = w_2$. In addition to this, the feasibility condition provides that $5w_1 = \frac{5}{2}w_2 \geq 5$, i.e., $w_1 \geq 1$ and $w_2 \geq 2$. This implies $\mu > 0$. The complementary slackness is stated as follows:

$$\mu\,(5 - w_1 - 2w_2) = 0 \ ,$$

which implies $5 = w_1 + 2w_2 = 5w_1 = \frac{5}{2}w_2$. Hence, the KKT conditions specify $(w_1^*, w_2^*) = (1, 2)$ with maximum value $\sqrt{5}$.

(b) [1pt] Find the optimal solution of the following optimization:

$$\begin{aligned} \underset{w_1,w_2}{\text{maximize}} \quad & \frac{1}{\sqrt{w_1^2 + w_2^2}} \\ \text{subject to} \quad & w_1 + 2w_2 \geq 5 \end{aligned}$$

sol) This maximization is identical to Problem 1a, i.e., $(w_1^*, w_2^*) = (1, 2)$.

2. [15pt; Backpropagation] We want to train a simple deep neural network $f_{\boldsymbol{w}}(x)$ with $\boldsymbol{w} = (w_1, w_2, w_3)^\top \in \mathbb{R}^3$ and $x \in \mathbb{R}$, defined as:

$$f_{\boldsymbol{w}}(x) := w_3 \sigma_2(w_2 \sigma_1(w_1 x))$$

where $\sigma_1(u) = \sigma_2(u) = \frac{1}{1+\exp(-u)}$, i.e., sigmoid activation. You may denote $x_1 := w_1 x$ and $x_2 := w_2 \sigma_1(x_1)$ for notational convenience.

(a) [1pt] Illustrate a directed acyclic graph corresponding to the computation of $f_{\boldsymbol{w}}(x)$.

sol) f connected to w_2 and σ_2; σ_2 connected to w_2 and σ_1; σ_1 connected to w_1 and x

(b) [2pt] Compute $\frac{\partial \sigma_1}{\partial u}$ and provide the answer in two different forms: (i) using only u and the exponential functions; and (ii) using only $\sigma_1(u)$.

sol)

$$(i) \quad \frac{\partial \sigma_1}{\partial u} = \frac{\exp(-u)}{(1+\exp(-u))^2} \; ; \quad \text{and} \quad (ii) \quad \frac{\partial \sigma_1}{\partial u} = \sigma_1(u)(1-\sigma_1(u))$$

(c) [2pt] Describe briefly what is meant by a *forward pass* and a *backward pass*?

sol) Forward pass: given x and current parameters $\boldsymbol{w}$, we compute $f_{\boldsymbol{w}}$ from data to final result.

Backward pass: given the final and intermediate results we back-propagate the error to obtain the derivatives.

(d) [2pt] Compute $\frac{\partial f_{\boldsymbol{w}}}{\partial w_3}$. Which result should we retain from the forward pass in order for efficiently computing this derivative?

sol) $\frac{\partial f_{\boldsymbol{w}}}{\partial w_3} = \sigma_2(w_2 \sigma_1(w_1 x))$; retain $\sigma_2(w_2 \sigma_1(w_1 x))$ from forward pass to avoid having to recompute it.

(e) [3pt] Compute $\frac{\partial f_{\boldsymbol{w}}}{\partial w_2}$ using the second option in Problem 2b. Which results should we retain from the forward pass in order for efficiently computing this derivative?

sol)

$$\frac{\partial f_{\boldsymbol{w}}}{\partial w_2} = \frac{\partial f_{\boldsymbol{w}}}{\partial \sigma_2} \cdot \frac{\partial \sigma_2}{\partial x_2} \cdot \frac{\partial x_2}{\partial w_2} = w_3 \cdot \sigma_2(x_2)(1-\sigma_2(x_2)) \cdot \sigma_1(w_1 x) \; .$$

Retain x_2 and x_1 (or σ_2 and σ_1).

(f) [5pt] Compute $\frac{\partial f_{\boldsymbol{w}}}{\partial w_1}$ using the second option in Problem 2b. Which results should we retain from the forward pass in order for efficiently computing this derivative? In what order should we compute the derivatives $\frac{\partial f_{\boldsymbol{w}}}{\partial w_1}$, $\frac{\partial f_{\boldsymbol{w}}}{\partial w_2}$ and $\frac{\partial f_{\boldsymbol{w}}}{\partial w_3}$ in order for

maximizing computational efficiency? How is this order related to the forward pass?

sol)

$$\frac{\partial f_{\boldsymbol{w}}}{\partial w_1} = \frac{\partial f_{\boldsymbol{w}}}{\partial \sigma_2} \cdot \frac{\partial \sigma_2}{\partial x_2} \cdot \frac{\partial x_2}{\partial \sigma_1} \cdot \frac{\partial \sigma_1}{\partial x_1} \cdot \frac{\partial x_1}{\partial w_1} = w_3 \cdot \sigma_2(x_2)(1 - \sigma_2(x_2)) \cdot w_2 \cdot \sigma_1(x_1)(1 - \sigma_1(x_1)) \cdot x \ .$$

Retain results from Problem 2e and also keep σ_1 or x_1. Reverse order of forward pass is computationally most efficient, i.e., $\frac{\partial f_{\boldsymbol{w}}}{\partial w_3}$, $\frac{\partial f_{\boldsymbol{w}}}{\partial w_2}$ and $\frac{\partial f_{\boldsymbol{w}}}{\partial w_1}$.