ESE 546: PRINCIPLES OF DEEP LEARNING

FALL 2020

HOMEWORK 0

INSTRUCTOR SOLUTIONS

Solution 1. (a) We know $\forall(x', y') \in A \times B$, we have :

$$f(x', y') \leq \max_{x \in A} f(x, y')$$

Therefore $\forall x' \in A$

$$\min_{y \in B} f(x', y) \leq \min_{y \in B} \max_{x \in A} f(x, y)$$

We know that $\min\limits_{y \in B} \max\limits_{x \in A} f(x, y)$ is a fixed quantity since A and B are finite and the above inequality holds for all x in A. Therefore $\min\limits_{y \in B} \max\limits_{x \in A} f(x, y)$ must be greater than or equal to the left hand side for any choice of x', even the value of x' that maximizes $\min_{y \in B} f(x', y)$.

$$\max_{x \in A} \min_{y \in B} f(x, y) \leq \min_{y \in B} \max_{x \in A} f(x, y) \qquad \square$$

(b) Let $g(x')$ represent $\inf\limits_{y \in B} f(x', y)$. We have:

$$g(x') = \inf_{y \in B} f(x', y) \qquad , \forall x' \in A$$

$\because$ g is infimum of f, we have

$$g(x') \leq f(x', y) \qquad , \forall (x', y) \in A \times B$$

$$g(x') \leq \sup_{x \in A} f(x, y) \qquad , \forall (x', y) \in A \times B$$

$$\sup_{x \in A} g(x) \leq \sup_{x \in A} f(x, y) \qquad , \forall y \in B$$

Since $\sup_{x\in A} g(x)$ is less than $\sup_{x\in A} f(x,y)$, it should be less even for the y for which $\sup_{x\in A} f(x,y)$ is minimum, therefore

$$\sup_{x\in A} g(x) \leq \inf_{y\in B} \sup_{x\in A} f(x,y)$$

$$\sup_{x\in A} \inf_{y\in B} f(x,y) \leq \inf_{y\in B} \sup_{x\in A} f(x,y) \qquad \square$$

Solution 2. For the function $f(x) = 2x_1^2 - 1.05x_1^4 + \frac{1}{6}x_1^6 - x_1x_2 + x_2^2$

$$f_{x_1} = \frac{\partial f(x)}{\partial x_1} = 4x_1 - 4.2x_1^3 + x_1^5 - x_2$$
$$f_{x_1x_1} = \frac{\partial^2 f(x)}{\partial x_1^2} = 4 - 12.6x_1^2 + 5x_1^4$$
$$f_{x_2} = \frac{\partial f(x)}{\partial x_2} = -x_1 + 2x_2$$
$$f_{x_2x_2} = \frac{\partial^2 f(x)}{\partial x_2^2} = 2$$
$$f_{x_1x_2} = \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} = -1$$

Setting $f_{x_1} = f_{x_2} = 0$ we get

$$x_2 = \frac{1}{2}x_1$$
$$2x_1^5 - 8.4x_1^3 + 7x_1 = 0$$

Solving the equations we get the stationary points (x_1, x_2) = (0,0), (1.07, 0.53), (1.74, 0.87), (-1.07, -0.53), (-1.74, -0.87) in the region $-3 \leq x_1 \leq 3$ and $-3 \leq x_2 \leq 3$. We have to check the second derivative for its sign

$$f_{x_1x_1}f_{x_2x_2} - f_{x_1x_2}^2 > 0$$
$$f_{x_1x_1} > 0$$
$$f_{x_2x_2} > 0$$

The points (0,0), (1.74, 0.87), (-1.74, -0.87) satisfy these conditions which shows that they are minimizers. Further $f(0,0)$ is the least of all the minimizers, which shows that it is a global minimum.

Similarly, $f_{x_1x_1}f_{x_2x_2} - f_{x_1x_2}^2 < 0$ for (1.07, 0.53) and (-1.07, -0.53) which shows that they are saddle points.

Solution 3. We are tasked to solve for

$$\min f(x,y) = x^2 + y^2 - 6xy - 4x - 5y$$
$$\text{such that } y \le -(x-2)^2 + 4 \Rightarrow (x-2)^2 + y - 4 \le 0$$
$$y \ge -x + 1 \Rightarrow -x - y + 1 \le 0$$

Define the Lagrangian

$$L(x,y,\lambda_1,\lambda_2) = x^2 + y^2 - 6xy - 4x - 5y + \lambda_1((x-2)^2 + y - 4) + \lambda_2(-x - y + 1),$$

take the partial derivatives and set them to 0.

$$\frac{\partial L}{\partial x} = 2x - 6y - 4 + 2\lambda_1(x-2) - \lambda_2 = 0 \tag{1}$$
$$\frac{\partial L}{\partial y} = 2y - 6x - 5 + \lambda_1 - \lambda_2 = 0 \tag{2}$$
$$\frac{\partial L}{\partial \lambda_1} = (x-2)^2 + y - 4 = 0 \tag{3}$$
$$\frac{\partial L}{\partial \lambda_2} = -x - y + 1 = 0 \tag{4}$$

1. Case 1: no constraints are active, so $\lambda_1, \lambda_2 = 0$. This gives $(x,y) = (-1.1875, -1.0625)$ after solving (1) and (2). This solution does not satisfy the first constraint, so at least one constraint is binding.
2. Case 2: assume both constraints are active. We solve for equations (1–4) by hand, or by using Mathematica/symbolic MATLAB/sympy, to see that both $\lambda_1 < 0$ and $\lambda_2 < 0$. This cannot happen.
3. Case 3: assume only the first constraint is active. We set $\lambda_2 = 0$ and solve for (1–3) to get $\lambda_1 = 14.14$ and $(x,y) = (2.69, 3.51)$. This solution satisfies both constraints. For this solution we have
$$f(x,y) = -65.6.$$
4. Case 4: only the second constraint is active. Set $\lambda_1 = 0$ and solve for (1–4) which is a linear system of equations. The solution is $(x,y) = (0.438, 0.563)$ and $\lambda_2 = -6.5$. Since $\lambda_2 < 0$, the solution does not satisfy the KKT conditions and it is invalid.

The second part of the problem involves changing the constraint

$$(x-2)^2 + y - 4 \le 0 \quad \text{to} \quad (x-2)^2 + y - 4.1 \le 0,$$

i.e., a change of -0.1. We know that the Lagrange multiplier is the rate at which the function value changes with respect to the constriant. We have

$$\Delta f^* = \Delta c_1 \, \lambda_1 = -1.415.$$

This matches the output of the code in problem3.py.

Solution 4. (a) It is clear that $X^2 \equiv 1$, $Y^2 \equiv 1$. Then for $z \in \mathcal{X}$,

$$\begin{aligned}\mathbb{P}(Z = z) = \mathbb{P}(XY = z) = \mathbb{P}(X^2Y = zX) = \mathbb{P}(Y = zX)& \\ &= \sum_{x\in\mathcal{X}} \mathbb{P}(Y = zx, X = x) \\ &= \sum_{x\in\mathcal{X}} \mathbb{P}(Y = zx)\mathbb{P}(X = x) \\ &= \frac{1}{2}\sum_{x\in X} P(X = x) = \frac{1}{2}\end{aligned}$$

Now we are able to compute the conditional distribution:

$$\begin{aligned}\mathbb{P}(Y|Z = z) = \frac{\mathbb{P}(Y, XY = z)}{\mathbb{P}(Z = z)} &= \frac{\mathbb{P}(Y, X = zY)}{\mathbb{P}(Z = z)} \\ &= \frac{\mathbb{P}(Y)\mathbb{P}(X = zY)}{\mathbb{P}(Z = z)} \\ &= \mathbb{P}(X = zY) = \frac{1 + zY}{2}q + \frac{1 - zY}{2}(1 - q). \end{aligned} \tag{5}$$

(b) Solution:

$$\begin{aligned}\mathbb{E}[Y|Z = z] &= \sum_{y\in\mathcal{X}} y\mathbb{P}(y|z) \\ &= \sum_{y\in\mathcal{X}} \frac{y + z}{2}q + \frac{y - z}{2}(1 - q) = 2qz - z.\end{aligned} \tag{6}$$

(c) Solution: By computation in (b) and definition:

$$\mu_{Y|Z} = \mathbb{E}[Y|Z] = \sum_{y\in\mathcal{X}} y\mathbb{P}(y|Z) = (2q - 1)Z \tag{7}$$

If $2q - 1 = 0$, then $\mu_{Y|Z} \equiv 0$. If not, by (a) with probability $\frac{1}{2}$, $\mu_{Y|Z} = 2q - 1$, with probability $\frac{1}{2}$, $\mu_{Y|Z} = 1 - 2q$.

Solution 5. (a) By definition

$$\begin{aligned}
f(x|y) &= \frac{f(x,y)}{f(y)} \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_x} \exp\left\{\frac{-1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{\sigma_x\sigma_y}\right)+\frac{y^2}{2\sigma_y^2}\right\} \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_x} \exp\left\{-\frac{\left(x-\frac{\rho\sigma_x y}{\sigma_y}\right)^2}{2(1-\rho^2)\sigma_x^2}\right\}.
\end{aligned}$$

This means that

$$X|Y \sim N\left(\frac{\rho\sigma_x y}{\sigma_y}, (1-\rho^2)\sigma_x^2\right).$$

(b) We now use the fact that

$$H(X|Y=y) = \frac{1}{2}\log\left(2\pi e(1-\rho^2)\sigma_x^2\right)$$

for all y. Therefore, again using the data given in the problem we have

$$H(X|Y) = \int f(y)\, H(X|Y=y)\, \mathrm{d}y = \frac{1}{2}\log\left(2\pi e(1-\rho^2)\sigma_x^2\right)$$

because $\int f(y)\, \mathrm{d}y$ integrates to 1.

Solution 6. Let the three sides be X, Y, Z. Suppose we fix Z to be longest side. The probability that X, Y, Z form a triangle is the area of the triangle bounded by

$$\{(X, Y) : 0 \leq X \leq Z,\ 0 \leq Y \leq Z, X + Y \geq Z\}.$$

This region on the XY plane is bounded by the X-axis, Y-axis and the line $x + y = z$. It is a right angled triangle with both height and width equal to Z and its area is $\frac{z^2}{2}$. Since Z is uniformly distributed on $[0, 1]$ we integrate this area from 0 to 1 to get $\frac{1}{6}$. We now multiply this by 3 because we fixed one of the sides to be the longest one to get the final answer as $\frac{1}{2}$.

Solution 7. We would like to find w^* and b^* such that $w^*, b^* = \text{argmin } \ell(w, b)$.

In order to simplify the analytical derivation, we incorporate the bias term into w by adding a column of ones to the matrix X and by increasing the dimension of w by 1. The additional parameter in w is thus equivalent to the bias term b.

Let $\bar{X} = \begin{bmatrix} X & \mathbb{1} \end{bmatrix}$ and let $\bar{w} = \begin{bmatrix} w & b \end{bmatrix}$. The residual loss is hence alternately expressed as:

$$\ell(w, b) = \ell(\bar{w}) = \frac{1}{2n}||Y - \bar{X}\bar{w}||_2^2$$

Differentiating the above expression with respect to $\bar{w}$ we have:

$$\begin{aligned}
\frac{\partial \ell}{\partial \bar{w}} &= \frac{\partial}{\partial \bar{w}} \left(\frac{1}{2n}||Y - \bar{X}\bar{w}||_2^2 \right) \\
&= \frac{\partial}{\partial \bar{w}} \left(\frac{1}{2n}(Y - \bar{X}\bar{w})^T (Y - \bar{X}\bar{w}) \right) \\
&= \frac{\partial}{\partial \bar{w}} \left(\frac{1}{2n} \left(Y^T Y + \bar{w}^T \bar{X}^T Y - Y^T \bar{X}\bar{w} - \bar{w}^T \bar{X}^T \bar{X}\bar{w} \right) \right) \\
&= \frac{1}{2n} \left(-2\bar{X}^T y + 2\bar{X}^T \bar{X}\bar{w} \right)
\end{aligned}$$

By equating the derivative to 0 and solving for $\bar{w}$, we obtain:

$$\begin{aligned}
& -2\bar{X}^T y + 2\bar{X}^T \bar{X}\bar{w}^* = 0 \\
\Rightarrow \quad & \bar{X}^T \bar{X}\bar{w}^* = \bar{X}^T y \\
\Rightarrow \quad & \bar{w}^* = (\bar{X}^T \bar{X})^{-1} \bar{X}^T y
\end{aligned}$$

Note that $\bar{X}^T \bar{X}$ is invertible *iff* the matrix X is full-rank. The quantity $(\bar{X}^T \bar{X})^{-1} X^T$ is known as the Moore-Penrose pseudo inverse and is a generalization of the matrix inverse to rectangular matrices.

See the Jupyter notebook on Canvas for the code.