CS7643: Deep Learning
Fall 2018
HW0 Solutions

Instructor: Dhruv Batra (dbatra@gatech.edu)

September 23, 2018

1 Probability and Statistics

1. (1 Point) Assuming a fair die, there is a 1/6 chance of landing on any number,

$$p(1) = \frac{1}{6}; \qquad p(\text{not } 1) = \frac{5}{6} \tag{1}$$

The expected outcome for a turn is

$$\$1\left(\frac{1}{6}\right) - \$\frac{1}{4}\left(\frac{5}{6}\right) = -\$\frac{1}{24} \tag{2}$$

So, we will lose money. Thus, it is not a good deal.

2.

$$\begin{align} C(x) &= \int_0^x p(z)dz \tag{3} \\ &= \begin{cases} \int_0^x 4z & 0 \leq x \leq 1/2 \\ \int_0^{1/2} 4zdz + \int_{1/2}^x (-4z+4)dz & 1/2 \leq x \leq 1 \end{cases} \tag{4} \\ &= \begin{cases} 2x^2 & 0 \leq x \leq 1/2 \\ 1/2 + -2x^2 + 1/2 + (4x-2) & 1/2 \leq x \leq 1 \end{cases} \tag{5} \\ &= \begin{cases} 2x^2 & 0 \leq x \leq 1/2 \\ -2x^2 + 4x - 1 & 1/2 \leq x \leq 1 \end{cases} \tag{6} \end{align}$$

3. (1 Point) If a random variable X has the expected value (mean) $\mu = E[X]$, then the variance of x is given by:

$$\begin{align} Var[X] &= E[(X-\mu)^2] \tag{7a} \\ &= E[X^2 - 2\mu X + \mu^2] \tag{7b} \\ &= E[X^2] - 2\mu^2 + \mu^2 \tag{7c} \\ &= E[X^2] - \mu^2 \tag{7d} \\ &= E[X^2] - (E[X])^2 \tag{7e} \end{align}$$

4. (1 Point) For standard normal distribution, we have,

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{8a}$$

$$\int_{-\infty}^{\infty} p(x)xdx = E(X) = 0 \tag{8b}$$

$$\int_{-\infty}^{\infty} p(x)x^2dx = E(X^2) = VAR(X) + [E(X)]^2 = 1 + 0 = 1 \tag{8c}$$

Hence,

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx = a + c \tag{9}$$

2 Proving Stuff

1. (2 Points) Define function $g(x)$ where:

$$g(x) = \log_e x - x + 1 \leq 0 \tag{10}$$

$g(x)$ is a strictly concave function ($g''(x) = -x^{-2} < 0$), therefore it is enough to show that the maximum is non-positive. At the maximum of $g(x)$ we must have $g'(x) = 0$. Therefore: $g'(x) = \frac{1}{x} - 1 = 0$. Solving this for x shows that the maximum of $g(x)$ is reached at $x = 1$. As the function value, there is $g(x = 1) = \log(1) - 1 + 1 = 0$. We know that $g(x) \leq 0$ for all $x \geq 0$.

2. (3 Points)

(a) Let $x = \frac{q_i}{p_i}$, we have,

$$\log\left(\frac{q_i}{p_i}\right) \leq \frac{q_i}{p_i} - 1 \tag{11}$$

$$\begin{aligned} KL(p,q) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) &= -\sum_{i=1}^{k} p_i \log\left(\frac{q_i}{p_i}\right) && \text{(12a)} \\ &\geq -\sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) && \text{(12b)} \\ &= -\sum_{i=1}^{k} (q_i - p_i) && \text{(12c)} \\ &= -\sum_{i=1}^{k} q_i + \sum_{i=1}^{k} p_i = 0 && \text{(12d)} \end{aligned}$$

(b) $KL(p,q) = 0$ if and only if $p_i = q_i \forall i$.

(c) Let $p = [1/2, 1/2]$ and $q = [1/4, 3/4]$. Then

$$KL(p,q) = \frac{1}{2}\log(\frac{1/2}{1/4}) + \frac{1}{2}\log(\frac{1/2}{3/4}) \approx 0.144 \tag{13}$$
$$KL(q,p) = \frac{1}{4}\log(\frac{1/4}{1/2}) + \frac{3}{4}\log(\frac{3/4}{1/2}) \approx 0.131 \tag{14}$$
$$\tag{15}$$

So $KL(p,q) \neq KL(q,p)$.

3 Calculus

1. (3 Points) Let

$$z_1 = 5\max\{x_1, x_2\}\frac{x_3}{x_4} - 5(x_5 + x_6) \tag{16}$$
$$z_2 = \log(z_1) + \frac{1}{2} \tag{17}$$
$$z_3 = \sigma(z_2) \qquad (= f(x)) \tag{18}$$
$$\tag{19}$$

Then

$$\nabla_x f^T = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \frac{\partial f}{\partial x_4} \\ \frac{\partial f}{\partial x_5} \\ \frac{\partial f}{\partial x_6} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_3} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_4} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_5} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_6} \end{bmatrix} \tag{20}$$

Now compute the partials listed above:

$$\begin{bmatrix} \frac{\partial z_1}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_1}{\partial x_3} \\ \frac{\partial z_1}{\partial x_4} \\ \frac{\partial z_1}{\partial x_5} \\ \frac{\partial z_1}{\partial x_6} \end{bmatrix} = \begin{bmatrix} 5\frac{x_3}{x_4}[\![x_1 > x_2]\!] \\ 5\frac{x_3}{x_4}[\![x_2 > x_1]\!] \\ 5\frac{\max\{x_1,x_2\}}{x_4} \\ -5\frac{\max\{x_1,x_2\}x_3}{x_4^2} \\ -5 \\ -5 \end{bmatrix} \tag{21}$$

$$\frac{\partial z_2}{\partial z_1} = \frac{1}{z_1} \tag{22}$$

$$\frac{\partial z_3}{\partial z_2} = \frac{e^{-z_2}}{(1+e^{-z_2})^2} = \sigma(z_2)(1-\sigma(z_2)) = z_3(1-z_3) \tag{23}$$

All that's left is plugging in values. First compute $f(\hat{x})$:

$$\hat{z}_1 = 2.5 \tag{24}$$

$$\hat{z}_2 \approx 1.416 \tag{25}$$

$$f(\hat{x}) = \hat{z}_3 \approx 0.805 \tag{26}$$

And finally plug numbers into the gradient at $\hat{x}$. Start with the scalars

$$\frac{\partial z_2}{\partial z_1}|_{\hat{x}} = \frac{1}{2.5} = 0.4 \tag{27}$$

$$\frac{\partial z_3}{\partial z_2}|_{\hat{x}} = \hat{z}_3(1 - \hat{z}_3) \approx 0.157 \tag{28}$$

$$\nabla_x f(x)^T|_{\hat{x}} \approx \begin{bmatrix} 0.157 \cdot 0.4 \frac{\partial z_1}{\partial x_1} \\ 0.157 \cdot 0.4 \frac{\partial z_1}{\partial x_2} \\ 0.157 \cdot 0.4 \frac{\partial z_1}{\partial x_3} \\ 0.157 \cdot 0.4 \frac{\partial z_1}{\partial x_4} \\ 0.157 \cdot 0.4 \frac{\partial z_1}{\partial x_5} \\ 0.157 \cdot 0.4 \frac{\partial z_1}{\partial x_6} \end{bmatrix} \approx \begin{bmatrix} 0.157 \cdot 0.4 \cdot 2.5 \\ 0.157 \cdot 0.4 \cdot 0 \\ 0.157 \cdot 0.4 \cdot 2.08 \\ 0.157 \cdot 0.4 \cdot -1.04 \\ 0.157 \cdot 0.4 - 5 \\ 0.157 \cdot 0.4 - 5 \end{bmatrix} \approx \begin{bmatrix} 0.1571 \\ 0 \\ 0.1309 \\ -0.0655 \\ -0.3142 \\ -0.3142 \end{bmatrix} \tag{29}$$

4 Softmax

1. The implementation is available on Canvas in `hw0_sol.zip`.
2. Note that the loss function decomposes into a score function and a log-sum-exp function:

$$L(W) = -\log(p_y) \tag{30}$$

$$= -\log\left(\frac{e^{z_j}}{\sum_k e^{z_k}}\right) \tag{31}$$

$$= -(\log(e^{z_j}) - \log(\sum_k e^{z_k})) \tag{32}$$

$$= -(z_j - \log(\sum_k e^{z_k})) \tag{33}$$

Following convex function composition rules, the loss is convex as long as each term (score and log-sum-exp) is convex. Luckily, $-z_j$ is linear in W, so all that remains is to show that $g(z) = \log(\sum_k e^{z_k})$ is convex.

Let $s_k = e^{z_k}$ be the exponentiated score. The gradient of g is

$$\nabla_z g(z) = \frac{s}{1^T s}. \tag{34}$$

The Hessian is

$$\nabla_z^2 g(z) = \frac{1}{1^T s} \operatorname{diag}(s) - \frac{1}{(1^T s)^2} s s^T. \tag{35}$$

Consider an arbitrary vector of reals x with the same dimensionality as z. I want to show

$$x^T \nabla_z^2 g(z) x \geq 0 \tag{36}$$

Now consider the vectors $t_k = \sqrt{s_k}$ and $u_k = \sqrt{s_k} x_k$. Then

$$\begin{align} x^T \nabla_z^2 g(z) x &= x^T \left(\frac{1}{1^T s} \operatorname{diag}(s) - \frac{1}{(1^T s)^2} s s^T \right) x \tag{37} \\ &= \frac{1}{t^T t} u^T u - \frac{1}{(t^T t)^2} (t^T u)^2 \tag{38} \\ & \tag{39} \end{align}$$

By the Cauchy-Schwarz inequality,

$$\begin{align} ||t|| ||u|| &\geq (t^T u) \tag{40} \\ (t^T t)(u^T u) &\geq (t^T u) \tag{41} \\ (t^T t)(u^T u) - (t^T u) &\geq 0 \tag{42} \\ \frac{1}{(t^T t)} (u^T u) - \frac{1}{(t^T t)^2} (t^T u) &\geq 0 \tag{43} \end{align}$$

That means the Hessian is positive semi-definite, so the log-sum-exp function is convex and the entire loss function is convex.