Homework 03 Solutions

ECE 449/590, Fall 2022

1. *(30 points)* Let $\{x^{(1)}, \dots, x^{(m)}\}$ be sampled i.i.d. from a Gaussion distribution of mean μ anf variance σ^2. Let $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ be an estimator of Gaussian mean μ. Show that $\mathrm{Var}(\hat{\mu}_m) = \frac{\sigma^2}{m}$.

   Answer:

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}_m) &= \mathbb{E}\Big(\big(\hat{\mu}_m - \mathbb{E}(\hat{\mu}_m)\big)^2\Big) \\
&= \mathbb{E}\Big(\big(\frac{1}{m}\sum_{i=1}^{m} x^{(i)} - \mu\big)^2\Big) \\
&= \mathbb{E}\Big(\big(\frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu)\big)^2\Big) \\
&= \frac{1}{m^2}\Big(\sum_{i=1}^{m} \mathbb{E}\big((x^{(i)} - \mu)^2\big) + \sum_{i \neq j} \mathbb{E}\big((x^{(i)} - \mu)(x^{(j)} - \mu)\big)\Big) \\
&= \frac{1}{m^2}\Big(\sum_{i=1}^{m} \mathrm{Var}(x^{(i)}) + 0\Big) \\
&= \frac{\sigma^2}{m}
\end{aligned}
$$

2. *(70 points)* Considering the following neural network with inputs (x_1, x_2), outputs (z_1, z_2, z_3), and parameters $\boldsymbol{\theta} = (a, b, c, d, e, f, i, j, k, l, m, n, o, p, q)$,

$$
\begin{aligned}
\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}, \\
\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} &= \begin{pmatrix} \mathrm{ReLU}(g_1) \\ \mathrm{ReLU}(g_2) \end{pmatrix}, \\
\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} &= \begin{pmatrix} i & j \\ k & l \\ m & n \end{pmatrix}\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} o \\ p \\ q \end{pmatrix}.
\end{aligned}
$$

A. *(15 points)* For a minibatch containing a single training sample $(x_1, x_2, y = 2)$, apply softmax and write down the cross-entropy loss function $J(\boldsymbol{\theta})$ as a function of (z_1, z_2, z_3). Compute $\frac{\partial J}{\partial z_1}, \frac{\partial J}{\partial z_2}, \frac{\partial J}{\partial z_3}$ as functions of (z_1, z_2, z_3).

Answer: According to Lecture 19 Slide 13, note that $y = 2$:

$$\begin{aligned}
\frac{\partial J}{\partial z_1} &= \frac{exp(z_1)}{exp(z_1) + exp(z_2) + exp(z_3)}, \\
\frac{\partial J}{\partial z_2} &= \frac{exp(z_2)}{exp(z_1) + exp(z_2) + exp(z_3)} - 1, \\
\frac{\partial J}{\partial z_3} &= \frac{exp(z_3)}{exp(z_1) + exp(z_2) + exp(z_3)}.
\end{aligned}$$

B. *(20 points)* Base on A., apply backprop to compute $\frac{\partial J}{\partial i}, \frac{\partial J}{\partial j}, \frac{\partial J}{\partial k}, \frac{\partial J}{\partial l}, \frac{\partial J}{\partial m}, \frac{\partial J}{\partial n}, \frac{\partial J}{\partial o}, \frac{\partial J}{\partial p}, \frac{\partial J}{\partial q}, \frac{\partial J}{\partial h_1}, \frac{\partial J}{\partial h_2}$.

Answer: According to Lecture 19 Slide 16, derivatives to weights are:

$$\begin{aligned}
\frac{\partial J}{\partial i} &= h_1 \frac{\partial J}{\partial z_1}, & \frac{\partial J}{\partial j} &= h_2 \frac{\partial J}{\partial z_1}, \\
\frac{\partial J}{\partial k} &= h_1 \frac{\partial J}{\partial z_2}, & \frac{\partial J}{\partial l} &= h_2 \frac{\partial J}{\partial z_2}, \\
\frac{\partial J}{\partial m} &= h_1 \frac{\partial J}{\partial z_3}, & \frac{\partial J}{\partial n} &= h_2 \frac{\partial J}{\partial z_3}.
\end{aligned}$$

Derivatives to bias are:

$$\frac{\partial J}{\partial o} = \frac{\partial J}{\partial z_1}, \qquad \frac{\partial J}{\partial p} = \frac{\partial J}{\partial z_2}, \qquad \frac{\partial J}{\partial q} = \frac{\partial J}{\partial z_3}.$$

Derivatives to inputs are:

$$\frac{\partial J}{\partial h_1} = i \frac{\partial J}{\partial z_1} + k \frac{\partial J}{\partial z_2} + m \frac{\partial J}{\partial z_3}, \qquad \frac{\partial J}{\partial h_2} = j \frac{\partial J}{\partial z_1} + l \frac{\partial J}{\partial z_2} + n \frac{\partial J}{\partial z_3}.$$

C. *(20 points)* Base on B., apply backprop to compute $\frac{\partial J}{\partial a}, \frac{\partial J}{\partial b}, \frac{\partial J}{\partial c}, \frac{\partial J}{\partial d}, \frac{\partial J}{\partial e}, \frac{\partial J}{\partial f}$. Explain why you don't need to compute $\frac{\partial J}{\partial x_1}$ and $\frac{\partial J}{\partial x_2}$.
(Hint: use the step function $u(x)$ as the derivative of ReLU(x).)

Answer: First of all, we need to apply backprop to ReLU according to Lecture 19 Slide 11:

$$\frac{\partial J}{\partial g_1} = u(g_1) \frac{\partial J}{\partial h_1}, \qquad \frac{\partial J}{\partial g_2} = u(g_2) \frac{\partial J}{\partial h_2}.$$

Similar to B., we have,

$$\frac{\partial J}{\partial a} = x_1 \frac{\partial J}{\partial g_1}, \qquad \frac{\partial J}{\partial b} = x_2 \frac{\partial J}{\partial g_1},$$
$$\frac{\partial J}{\partial c} = x_1 \frac{\partial J}{\partial g_2}, \qquad \frac{\partial J}{\partial d} = x_2 \frac{\partial J}{\partial g_2},$$
$$\frac{\partial J}{\partial e} = \frac{\partial J}{\partial g_1}, \qquad \frac{\partial J}{\partial f} = \frac{\partial J}{\partial g_2}.$$

There is no need to compute $\frac{\partial J}{\partial x_1}$ and $\frac{\partial J}{\partial x_2}$ because we have reached the first layer.

D. *(15 points)* For the learning rate ϵ, show the equation to apply the simple SGD algorithm to update $\boldsymbol{\theta}$ for this minibatch.

Answer: We now have:

$$\begin{aligned}
\boldsymbol{\theta} &= (a, b, c, d, e, f, i, j, k, l, m, n, o, p, q)^\top, \\
\nabla_\theta J &= (\frac{\partial J}{\partial a}, \frac{\partial J}{\partial b}, \frac{\partial J}{\partial c}, \frac{\partial J}{\partial d}, \frac{\partial J}{\partial e}, \frac{\partial J}{\partial f}, \frac{\partial J}{\partial i}, \frac{\partial J}{\partial j}, \frac{\partial J}{\partial k}, \frac{\partial J}{\partial l}, \frac{\partial J}{\partial m}, \frac{\partial J}{\partial n}, \frac{\partial J}{\partial o}, \frac{\partial J}{\partial p}, \frac{\partial J}{\partial q})^\top.
\end{aligned}$$

The SGD update will be:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \nabla_\theta J.$$