PURDUE

BME 64600 – 001 and ECE 60146 – 001

Midterm #2, Spring 2023

NAME _____

PUID _____

Exam instructions:

- You have 75 minutes to work the exam.
- This is a closed-book and closed-note exam. You may not use or have access to your book, notes, any supplementary reference, a calculator, or any communication device including a cell-phone or computer.
- You may not communicate with any person other than the official proctor during the exam.

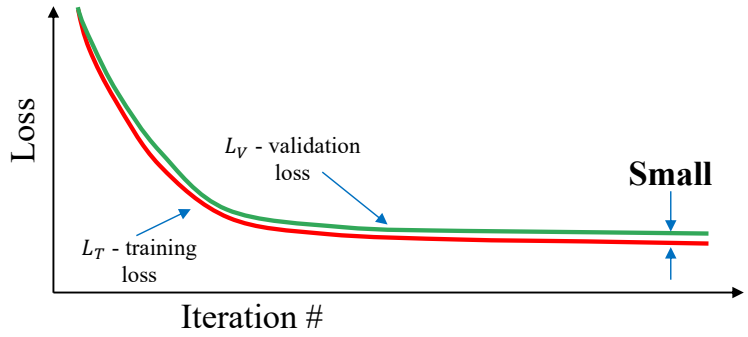
To ensure Gradescope can read your exam:

- Write your full name and PUID above and on the top of every page.
- Answer all questions in the area designated for each problem.
- Write only on the front of the exam pages.
- DO NOT run over to the next question.

Name/PUID: Key

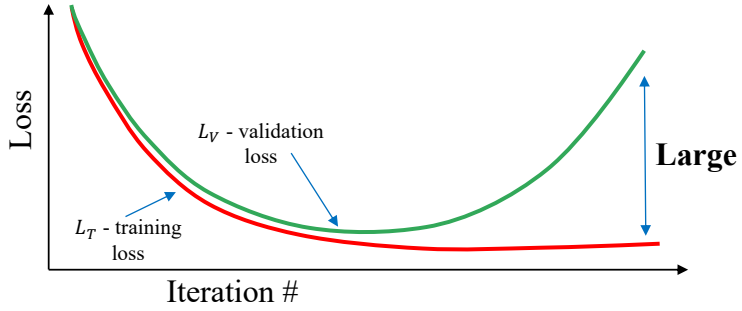
Problem 1. (18pt) Training Convergence

For each of the plots below, answer the associated questions.



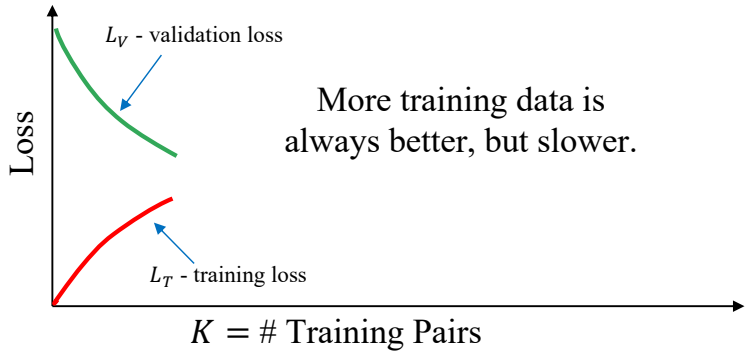
a) What is the data telling you? Based on this, what action(s) might you take?

Solution: The capacity of the network may be too small. You might want to increase the capacity by adding more layers, structures or parameters.



b) What is the data telling you? Based on this, what action(s) might you take?

Solution: The capacity of the network may be too large. You might want to decrease the capacity by a) removing layers or parameters, b) increasing the amount of training data, or c) introducing some regularization of the network parameters.



c) What is the data telling you? Based on this, what action(s) might you take?

Solution: You probably do not have enough training data. You might want to increase the amount of training data, or reduce the number of parameters in the network, or introduce some regularization of the network parameters.

Name/PUID: _____

Problem 2. (48pt) Stochastic Gradient Descent

Consider a problem in which you are doing supervised training of an inference network, $f_\theta : \Re^{N_y} \rightarrow \Re^{N_x}$, for $\theta \in \Re^p$. Let the loss function for each training pair, (x_k, y_k), be given by

$$L_k(\theta) = \frac{1}{2}\|x_k - f_\theta(y_k)\|^2 \ ,$$

and let the total loss function be given by

$$L(\theta) = \sum_{k=0}^{K-1} L_k(\theta) \ .$$

Also assume the gradient of the inference network is given by

$$A_k = \nabla_\theta f_\theta(y_k) \ ,$$

where $A_k \in \Re^{N_x \times p}$, and define $g_k \in \Re^p$ given by

$$g_k = \nabla_\theta L_k(\theta) \ ,$$

to be a column vector equal to the gradient of the loss function for the k^{th} training pair.

Let G be a p-dimensional random column vector produced by selecting the gradient of the loss for a single training pair at random with a uniform distribution. Then the probability density of G is given by

$$p(g) = \frac{1}{K}\sum_{k=0}^{K-1} \delta(g - g_k) \ ,$$

where $\delta(x)$ is a multi-dimensional delta function.

Furthermore, let $G_0, \cdots, G_{K_b-1}$ be K_b independently sampled (with replacement) gradients, and let

$$\hat{G} = \frac{1}{K_b}\sum_{k=0}^{K_b-1} G_k \ ,$$

be the average of their gradients.

a) Calculate an expression for $g_k = \nabla_\theta L_k(\theta)$.

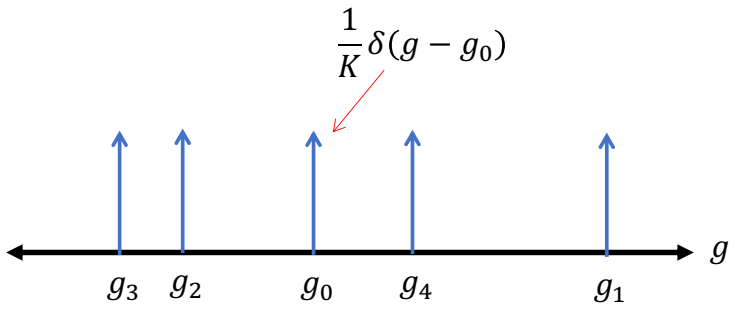
Solution:

$$g_k = [A_k]^t \left(x_k - f_\theta(y_k)\right)$$

---

b) Sketch the probability density $p(g)$ for a typical case. Label your sketch to explain the meaning of its structure. For your sketch, you can assume $p = 1$.

Solution:



---

c) Calculate μ, the mean of G.

Solution:

$$\begin{aligned}
\mu &= E[G] \\
&= \int_{\Re^p} g \frac{1}{K} \sum_{k=0}^{K-1} \delta(g - g_k) dg \\
&= \frac{1}{K} \sum_{k=0}^{K-1} \int_{\Re^p} g\, \delta(g - g_k) dg \\
&= \frac{1}{K} \sum_{k=0}^{K-1} g_k
\end{aligned}$$

---

d) Calculate R, the covariance of G.

Solution:

$$\begin{aligned} R &= E\left[(G-\mu)(G-\mu)^t\right] \\ &= \int_{\Re^p} (g-\mu)(g-\mu)^t \ \frac{1}{K}\sum_{k=0}^{K-1} \delta(g-g_k)dg \\ &= \frac{1}{K}\sum_{k=0}^{K-1} \int_{\Re^p} (g-\mu)(g-\mu)^t \ \delta(g-g_k)dg \\ &= \frac{1}{K}\sum_{k=0}^{K-1} (g_k-\mu)(g_k-\mu)^t \end{aligned}$$

---

e) Calculate $\hat{\mu}$, the mean $\hat{G}$.

Solution:

$$\begin{aligned} \hat{\mu} &= E\left[\hat{G}\right] \\ &= E\left[\frac{1}{K_b}\sum_{k=0}^{K_b-1} G_k\right] \\ &= \frac{1}{K_b}\sum_{k=0}^{K_b-1} E\left[G_k\right] \\ &= \frac{1}{K_b}\sum_{k=0}^{K_b-1} \mu \\ &= \mu \end{aligned}$$

---

f) Calculate an expression for $\hat{R}$, the covariance of $\hat{G}$.
(Hint: This calculation is a bit tricky, so you might just want to guess at the answer for partial credit.)

Solution:

$$
\begin{aligned}
\hat{R} &= E\left[\left(\hat{G}-\mu\right)\left(\hat{G}-\mu\right)^t\right] \\
&= E\left[\left(\frac{1}{K_b}\sum_{k=0}^{K_b-1} G_k-\mu\right)\left(\frac{1}{K_b}\sum_{k=0}^{K_b-1} G_k-\mu\right)^t\right] \\
&= E\left[\left(\frac{1}{K_b}\sum_{k=0}^{K_b-1} (G_k-\mu)\right)\left(\frac{1}{K_b}\sum_{k=0}^{K_b-1} (G_k-\mu)\right)^t\right] \\
&= E\left[\left(\frac{1}{K_b}\sum_{i=0}^{K_b-1} (G_i-\mu)\right)\left(\frac{1}{K_b}\sum_{j=0}^{K_b-1} (G_j-\mu)\right)^t\right] \\
&= E\left[\frac{1}{K_b^2}\sum_{i=0}^{K_b-1}\sum_{j=0}^{K_b-1} (G_i-\mu)(G_j-\mu)^t\right] \\
&= \frac{1}{K_b^2}\sum_{i=0}^{K_b-1}\sum_{j=0}^{K_b-1} E\left[(G_i-\mu)(G_j-\mu)^t\right] \\
&= \frac{1}{K_b^2}\sum_{i=0}^{K_b-1}\sum_{j=0}^{K_b-1} E\left[(G_i-\mu)(G_i-\mu)^t\right]\delta(i-j) \\
&= \frac{1}{K_b^2}\sum_{i=0}^{K_b-1} E\left[(G_i-\mu)(G_i-\mu)^t\right] \\
&= \frac{1}{K_b^2}\sum_{i=0}^{K_b-1} R \\
&= \frac{1}{K_b}R
\end{aligned}
$$

---

g) What is/are advantage(s) of choosing K_b to be larger?

Solution: By choosing K_b large, you can reduce the variance in your estimate of the gradient. This is better for exactly getting at the precise minimum of the loss function.

---

h) What is/are advantage(s) of choosing K_b to be smaller?

Solution: By choosing K_b smaller, you increase the variance in your estimate of the gradient. This can be advantageous in certain situations since it can help to jump out of local minima of the loss function. It is also faster to do each update since the number of training pairs in the batch is smaller.

Name/PUID: _____

Problem 3. (48pt) Optimization

Consider a problem where $\tilde{Y} \sim \tilde{p}(y)$ and $Y \sim p(y)$ are two random vectors in $\Re^N$, and define the likelihood ratio as

$$R(y) = \frac{\tilde{p}(y)}{p(y)} \ .$$

Also, assume the technical condition that $\tilde{p}(y)$ is absolutely continuous with respect to $p(y)$ i.e., that $\forall y \in \Re^N$, $R(y) < \infty$.[1]

Also define the function

$$C(R) = E\left[(1 + R(Y)) \log(1 + R(Y))\right] \ ,$$

and let R_a and R_b be any two valid likelihood ratios.

a) Is $R(y) = 2$ a valid likelihood ratio? Why or why not?

Solution: No, $R(y) = 2$ is not a valid likelihood ratio. To see this, notice that then $\tilde{p}(y) = 2p(y)$, but this implies that $\int_{\Re^N} \tilde{p}(y)dy = \int_{\Re^N} 2p(y)dy = 2$. However, this is not possible, so the likelihood ratio must not be valid.

---

b) Show that for every valid likelihood ratio, $E\left[R(Y)\right] = 1$.

Solution:

$$\begin{aligned} E\left[R(Y)\right] &= \int_{\Re^N} R(y)\, p(y)dy \\ &= \int_{\Re^N} \frac{\tilde{p}(y)}{p(y)}\, p(y)dy \\ &= \int_{\Re^N} \tilde{p}(y)dy \\ &= 1 \end{aligned}$$

---

[1] This is not exactly the correct definition of absolutely continuous, but it's good enough.

c) Show that for any valid likelihood ratio, R, and integrable function, $f : \Re^N \to \Re$, then

$$E\left[f(\tilde{Y})\right] = E\left[f(Y)R(Y)\right] \ .$$

Solution:

$$\begin{aligned} E\left[f(\tilde{Y})\right] &= \int_{\Re^N} f(y)\,\tilde{p}(y)dy \\ &= \int_{\Re^N} f(y)\frac{\tilde{p}(y)}{p(y)}\,p(y)dy \\ &= \int_{\Re^N} f(y)R(y)\,p(y)dy \\ &= E\left[f(Y)R(Y)\right] \end{aligned}$$

---

d) Show that the function $h(x) = (1+x)\log(1+x)$ is strictly convex for all $x \geq 0$.

Solution:

$$\begin{aligned} \frac{dh(x)}{dx} &= \log(1+x) + \frac{1+x}{1+x} \\ &= \log(1+x) + 1 \end{aligned}$$

So then

$$\frac{d^2h(x)}{dx^2} = \frac{1}{1+x} \ .$$

So then $\forall x \geq 0$, we have that $\frac{d^2h(x)}{dx^2} > 0$, which implies that $h(x)$ is strictly convex.

---

e) Show that $\forall \lambda \in (0,1)$, then $R = \lambda R_a + (1-\lambda) R_b$ is a valid likelihood ratio.

Solution:

$$\begin{aligned} E\left[R(Y)\right] &= E\left[\lambda R_a(Y) + (1-\lambda) R_b(Y)\right] \\ &= \lambda E\left[R_a(Y)\right] + (1-\lambda) E\left[R_b(Y)\right] \\ &= \lambda + (1-\lambda) = 1 \end{aligned}$$

Furthermore, if R_a and R_b are valid, then $R(y) < \infty$.

---

f) Show that C is a strictly convex function by showing that $\forall \lambda \in (0,1)$, then

$$C(R) < \lambda C\left(R_a\right) + (1-\lambda) C\left(R_b\right) \ .$$

Solution:

$$\begin{aligned} C(R) &= E\left[(1+R(Y)) \log(1+R(Y))\right] \\ &= E\left[h(R(Y))\right] \\ &= E\left[h(\lambda R_a + (1-\lambda) R_b)\right] \\ &\leq E\left[\lambda h(R_a) + (1-\lambda) h(R_b)\right] \\ &= \lambda E\left[h(R_a)\right] + (1-\lambda) E\left[h(R_b)\right] \\ &= \lambda C(R_a) + (1-\lambda) C(R_b) \end{aligned}$$

---

g) Use these results to argue that if $C(R)$ has a local minimum, then the local minimum must be the unique global minimum.

Solution: Since it is a strictly convex function, any local minimum must be the unique global minimum.

---

h) Use these results to argue that the unique global minimum of $C(R)$ must occur for the likelihood ratio given by $R(y) = 1$.

Solution: First notice that for $x = 1$, $\frac{dh(x)}{dx} = 1 + \log 2$.

To do this, we need to your the method of variations. However, we get the correct answer if we differentiate with respect to R as if it was a finite dimensional vector. We also need to enforce the constraint that $E[R(Y)] = 1$, so we will minimize the following Lagrangian

$$L(R) = C(R) + \alpha E[R(Y) \ ,$$

by differentiating with respect to R.

$$\begin{aligned}
\nabla_R L(R) &= \nabla_R \left[C(R) + \alpha E[R(Y)] \right] \\
&= \nabla_R \left[E\left[h(R(Y)) \right] + \alpha E\left[R(Y) \right] \right] \\
&= \nabla_R E\left[h(R(Y)) + \alpha R(Y) \right] \\
&= \nabla_R \int_{\Re^N} \left[h(R(y)) + \alpha R(y) \right] p(y) dy \\
&= \int_{\Re^N} \nabla_R \left[h(R(y)) + \alpha R(y) \right] p(y) dy \\
&= \int_{\Re^N} \left[1 + \log 2 + \alpha 1 \right] p(y) dy \\
&= 1 + \log 2 + \alpha 1
\end{aligned}$$

If we choose $\alpha = \frac{-1}{1+\log 2}$, then we find that

$$\nabla_R L(R) = 0 \ ,$$

which means that $R = 1$ is a minimum to the constrained optimization problem.
