CS 148a Problem Set 2

Question 1

Prove that

$$\mathbb{E}[\mathcal{L}(W)] = ||y - f(p)Xw||^2 + g(p)||\Gamma w||^2$$

where $f(p)$ and $g(p)$ are functions of p, and Γ is a diagonal matrix with the standard deviations of features in the data matrix X

Proof. Let D be a diagonal matrix where $D_{ii} \sim \text{Bernoulli}(p)$ represents the dropout mask applied to the features. The modified matrix will be XD, with some features destroyed by the mask in every sample. The loss function is:

$$\begin{aligned}\mathbb{E}[||y - XDw||^2] &= \mathbb{E}[(y - XDw)^T(y - XDw)] \\ &= \mathbb{E}[y^Ty - 2y^TXDw + (XDw)^T(XDw)] \\ &= y^Ty - 2y^TX\mathbb{E}[D]w + \mathbb{E}[w^TD^TX^TXDw]\end{aligned}$$

First, handle the linear term. Since $\mathbb{E}[D] = pI$:

$$-2y^TX\mathbb{E}[D]w = -2py^TXw$$

Now we handle the quadratic term. Let $A = X^TX$. We need to evaluate $\mathbb{E}[w^TDADw]$. Expanding the matrix multiplication into a summation:

$$\begin{aligned}\mathbb{E}[w^TDADw] &= \mathbb{E}\left[\sum_i\sum_j(Dw)_iA_{ij}(Dw)_j\right] && \text{(sum of a quadratic form)} \\ &= \sum_i\sum_j(D_{ii}w_i)A_{ij}(D_{jj}w_j) && \text{(since D is diagonal, so multiply w by D just scales w_n by D_{nn})} \\ &= \sum_i\sum_j w_iw_jA_{ij}\mathbb{E}[D_{ii}D_{jj}] \\ & && \text{(move deterministic values and use linearity of expectation over random variablers)}\end{aligned}$$

- If $i \neq j$: $\mathbb{E}[D_{ii}D_{jj}] = \mathbb{E}[D_{ii}]\mathbb{E}[D_{jj}] = p \cdot p = p^2$, since the dropout probabilities are independent
- If $i = j$: $\mathbb{E}[D_{ii}D_{jj}] = \mathbb{E}[D_{ii}^2] = p$ (since $x^2 = x$ for $x \in \{0, 1\}$ and the probabilities are perfectly correlated)

We split the sum:

$$\begin{aligned}\mathbb{E}[w^TDADw] &= \sum_i\sum_{j\neq i} w_iw_jA_{ij}p^2 + \sum_i w_i^2A_{ii}p \\ &= \sum_i\sum_{j\neq i} w_iw_jA_{ij}p^2 + \sum_i w_i^2A_{ii}p^2 - \sum_i w_i^2A_{ii}p^2 + \sum_i w_i^2A_{ii}p \\ &= \sum_i\sum_j w_iw_jA_{ij}p^2 + \sum_i w_i^2A_{ii}[p - p^2] \\ &= p^2\underbrace{\sum_{i,j} w_iw_jA_{ij}}_{w^TAw} + p(1-p)\underbrace{\sum_i w_i^2A_{ii}}_{w^T\text{diag}(A)w}\end{aligned}$$

By definition[1], $\Gamma^2 = \text{diag}(X^TX) = \text{diag}(A)$. So we can write the sum as

$$\begin{aligned}\mathbb{E}[w^TDX^TXDw] &= p^2w^TX^TXw + p(1-p)w^T\Gamma^2w \\ &= ||pXw||^2 + p(1-p)||\Gamma w||^2\end{aligned}$$

Finally, substitute the linear and quadratic terms back into the original expression:

$$\begin{aligned}\mathbb{E}[||y - XDw||^2] &= ||y||^2 - 2py^TXw + ||pXw||^2 + p(1-p)||\Gamma w||^2 \\ &= ||y - pXw||^2 + p(1-p)||\Gamma w||^2\end{aligned}$$

□

[1] from Piazza

Question 2

(a) 1.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta_j} &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \cdot \frac{\partial Y_{ij}}{\partial \beta_j} \\
&= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} (1) \\
&= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \\
\implies \frac{\partial \mathcal{L}}{\partial \beta} &= \frac{\partial \mathcal{L}}{\partial Y}
\end{aligned}$$

2.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \gamma_j} &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \cdot \frac{\partial Y_{ij}}{\partial \gamma_j} \\
&= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \cdot \hat{X}_{ij} \\
\implies \frac{\partial \mathcal{L}}{\partial \gamma} &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{i}} \circ \hat{X}_{i}
\end{aligned}$$

3.

4. By the chain rule, we can decompose the gradient into

$$\frac{\partial \mathcal{L}}{\partial X_{ij}} = \frac{\partial \mathcal{L}}{\partial \hat{X}_{ij}} \frac{\partial \hat{X}_{ij}}{\partial X_{ij}} + \frac{\partial \mathcal{L}}{\partial \sigma_j^2} \frac{\partial \sigma_j^2}{\partial X_{ij}} + \frac{\partial \mathcal{L}}{\partial \mu_j} \frac{\partial \mu_j}{\partial X_{ij}}$$

We now tackle each component

–

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \hat{X}_{ij}} &= \frac{\partial \mathcal{L}}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial \hat{X}_{ij}} \\
&= \frac{\partial \mathcal{L}}{\partial Y_{ij}} \gamma_j
\end{aligned}$$

$$\frac{\partial \hat{X}_{ij}}{\partial X_{ij}} = \frac{1}{\sqrt{\sigma_j^2 + \epsilon}}$$

–

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma_j^2} &= \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} \frac{\partial \hat{X}_{kj}}{\partial \sigma_j^2} \\
&= \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} \cdot (X_{kj} - \mu_j) \cdot (-\frac{1}{2}) (\sigma_j^2 + \epsilon)^{-3/2}
\end{aligned}$$

$$\frac{\partial \sigma_j^2}{\partial X_{ij}} = \frac{2(X_{ij} - \mu_j)}{N}$$

–

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu_j} &= \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} \frac{\partial \hat{X}_{kj}}{\partial \mu_j} \\
&= \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} \left(\frac{-1}{\sqrt{\sigma_j^2 + \epsilon}} \right)
\end{aligned}$$

$$\frac{\partial \mu_j}{\partial X_{ij}} = \frac{1}{N}$$

So the final gradient is

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial X_{ij}} =& \frac{\partial \mathcal{L}}{\partial Y_{ij}} \gamma_j \frac{1}{\sqrt{\sigma_j^2 + \epsilon}} \\
& + \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} (X_{kj} - \mu_j) \cdot (-\frac{1}{2})(\sigma_j^2 + \epsilon)^{-3/2} \cdot \frac{2(X_{ij} - \mu_j)}{N} \\
& + \frac{1}{N} \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} \left(\frac{-1}{\sqrt{\sigma_j^2 + \epsilon}} \right) \\
=& \frac{\gamma_j}{\sqrt{\sigma_j^2 + \epsilon}} \left[\frac{\partial \mathcal{L}}{\partial Y_{ij}} - \frac{1}{N} \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} - \frac{(X_{ij} - \mu_j)}{\sigma_j^2 + \epsilon} \cdot \frac{1}{N} \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{X}_{kj}} (X_{kj} - \mu_j) \right] \\
& = \frac{\gamma_j}{\sqrt{\sigma_j^2 + \epsilon}} \left[\frac{\partial \mathcal{L}}{\partial Y_{ij}} - \frac{1}{N} \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{kj}} \gamma_j - \hat{X}_{ij} \cdot \frac{1}{N} \sum_{k=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{kj}} \gamma_j \hat{X}_{kj} \right]
\end{aligned}
$$

Question 3

(a) 1.

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}}$$

2.

$$\frac{\partial \mathcal{L}}{\partial \gamma_j} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial \gamma_j} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial Y_{ij}} \hat{X}_{ij}$$

3.

$$\frac{\partial \mathcal{L}}{\partial X_{ij}} = \frac{\partial \mathcal{L}}{\partial \hat{X}_{ij}} \frac{\partial \hat{X}_{ij}}{\partial X_{ij}} + \frac{\partial \mathcal{L}}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial X_{ij}} + \frac{\partial \mathcal{L}}{\partial \mu_i} \frac{\partial \mu_i}{\partial X_{ij}}$$

1.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{X}_{ij}} &= \frac{\partial \mathcal{L}}{\partial Y_{ij}} \gamma_j \\ \frac{\partial \hat{X}_{ij}}{\partial X_{ij}} &= \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \end{aligned}$$

2.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \sigma_i^2} &= \sum_{k=1}^{D} \frac{\partial \mathcal{L}}{\partial \hat{X}_{ik}} \frac{\partial \hat{X}_{ik}}{\partial \sigma_i^2} = \sum_{k=1}^{D} \frac{\partial \mathcal{L}}{\partial \hat{X}_{ik}} \cdot (X_{ik} - \mu_i) \cdot \left(-\frac{1}{2}\right) (\sigma_i^2 + \epsilon)^{-3/2} \\ \frac{\partial \sigma_i^2}{\partial X_{ij}} &= \frac{2(X_{ij} - \mu_i)}{D} \end{aligned}$$

3.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_i} &= \sum_{k=1}^{D} \frac{\partial \mathcal{L}}{\partial \hat{X}_{ik}} \frac{\partial \hat{X}_{ik}}{\partial \mu_i} = \sum_{k=1}^{D} \frac{\partial \mathcal{L}}{\partial \hat{X}_{ik}} \left(\frac{-1}{\sqrt{\sigma_i^2 + \epsilon}}\right) \\ \frac{\partial \mu_i}{\partial X_{ij}} &= \frac{1}{D} \end{aligned}$$

So the final gradient, using a similar series of simplifications as in the previous part, will be

$$\frac{\partial \mathcal{L}}{\partial X_{ij}} = \frac{1}{D\sqrt{\sigma_i^2 + \epsilon}} \left[D\gamma_j \frac{\partial \mathcal{L}}{\partial Y_{ij}} - \sum_{k=1}^{D} \gamma_k \frac{\partial \mathcal{L}}{\partial Y_{ik}} - \hat{X} ij \sum_{k=1}^{D} \gamma_k \frac{\partial \mathcal{L}}{\partial Y_{ik}} \hat{X}_{ik} \right]$$

(c) In the case where β and γ are set to 0 and 1 (vectors), we have the following expressions for elements in the public matrix Y and private matrix Z:

$$\begin{aligned} Y_{ij} &= 1 \cdot \hat{X}_{ij} + 0 \\ &= \frac{X_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}} \\ Z_{ij} &= 1 \cdot \overline{X}_{ij} + 0 \\ &= \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \\ Y_{ji} &= \hat{X}_{ji} = \frac{X_{ji} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} && \text{(looking at transposed indices)} \\ &= \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} && \text{(since } X \text{ is symmetric)} \\ \implies Y_{ji} &= Z_{ij} \end{aligned}$$

Therefore, $Y = Z^T$. Give knowledge of Y, we know every element in the 16×16 matrix Z, such that we know all 256 elements in Z.

In the case where β and γ may take arbitrary values,

$$\begin{aligned}
Y_{ij} &= \gamma_j \hat{X}_{ij} + \beta_j \\
&= \gamma_j \frac{X_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}} + \beta_j \\
\implies Y_{ji} &= \gamma_i \frac{X_{ji} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i \\
Z_{ij} &= \gamma_j \overline{X}_{ij} + \beta_j \\
&= \gamma_j \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_j \\
\implies \frac{Z_{ij}}{Y_{ji}} &= \frac{\gamma_j \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_j}{\gamma_i \frac{X_{ji} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i} \\
&= \frac{\gamma_j \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_j}{\gamma_i \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i} && \text{(just changing the indices in the denominato since } X \text{ is symmetric)} \\
\implies \frac{Z_{ij} - \beta_j}{Y_{ji} - \beta_i} &= \frac{\gamma_j \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}}{\gamma_i \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}} && \text{(editing the original ratio)} \\
\implies \frac{Z_{ij} - \beta_j}{Y_{ji} - \beta_i} &= \frac{\gamma_j}{\gamma_i} \\
\implies Z_{ij} &= \frac{\gamma_j}{\gamma_i} \left(Y_{ji} - \beta_i \right) + \beta_j
\end{aligned}$$

Along the diagonal elements, when $i = j$, this simplifies to

$$Z_{ii} = Y_i i$$

However, this is not true for $i \neq j$ in general. Therefore, for arbitrary β and γ, the only known elements are the 16 diagonal elements.

Question 4

Question (a)(3) and (b)(1) are the same here, so I only submit the below proof once

(b) We assume that the ℓ_2 regularized update expression also includes $\nabla_w \mathcal{L}$, such that we have to show

$$w_{t+1} \leftarrow w_t - \eta(\nabla_w \mathcal{L} + \lambda_{wd} w_t) \tag{1}$$

$$\leftrightarrow w_{t+1} \leftarrow w_t - \eta(\nabla_w \mathcal{L} + \nabla_w \lambda_{l2} \sum_i w_i^2) \tag{2}$$

It is therefore sufficient to show that $\nabla_w \lambda_{l2} \sum_i w_i^2$ is equivalent to $\lambda_{wd} w_t$ up to a constant scale factor. The j-th component of the gradient vector will be $\frac{\partial}{\partial w_j} \sum_i w_i^2 = \frac{\partial}{\partial w_j}[w_1^2 + w^2 + \dots w_j^2 + \dots w_n^2] = 2w_j$. So

$$\lambda_{l2} \nabla_w \sum_i w_i^2 = \lambda_{l2}(2w)$$

Note that this w is the same as w_t, with the subscript having been ommitted for notational simplicity.

This is equigavlent to $\lambda_{wd} w_t$ if we set $\lambda_{wd} = 2\lambda_{l2}$. So with SGD, ℓ_2 regularized updates are equivalent to weight decay updates.

(c) 1. In Adam, we normalize the gradient by the square root of its variance to ensure constant scale for all parameter updates, such that

$$w_{t+1,i} = w_{t,i} - \eta \frac{\bar{s}_{t+1,i}}{\sqrt{\bar{r}_{t+1,i} + \delta}}$$

where

$$g_{t,i} = \frac{\partial \mathcal{L}(w_{t,i})}{\partial w_i}$$

$$s_{t+1,i} = \beta_1 s_{t,i} + (1-\beta_1) g_t \qquad \bar{s}_{t+1,i} = \frac{s_{t+1,i}}{1-\beta_1^{t+1}} \tag{1}$$

$$r_{t+1,i} = \beta_2 r_{t,i} + (1-\beta_2) g_{t,i}^2 \qquad \bar{r}_{t+1,i} = \frac{r_{t+1,i}}{1-\beta_2^{t+1}} \tag{2}$$

We require that, in expectation, the factor $s_{t+1,i}$ is equal to the mean of the gradient for that parameter $\mu = \mathbb{E}[g_i]$.

$$\begin{aligned}
s_{t+1,i} &= (1-\beta_1)\sum_{k=1}^{t} \beta_1^{t-k} g_{k,i} && \text{(unrolling the recurrence)}\\
\implies \mathbb{E}[s_{t+1,i}] &= (1-\beta_1)\sum_{k=1}^{t} \beta_1^{t-k} \mathbb{E}[g_k] \\
&= (1-\beta_1)\mu \sum_{k=1}^{t} \beta_1^{t-k} && \text{(letting } \mu = \mathbb{E}[g_k]\text{)}\\
&= \mu(1-\beta_1)(1)\frac{1-\beta_1^t}{1-\beta_1} && \text{(sum of geometric series)}\\
&= \mu(1-\beta_1^t)
\end{aligned}$$

So while we require $\mathbb{E}[g_i] = \mu$, it is instead equal to $\mu(1-\beta_1^t)$. We therefore divide by this extra factor at each step, using

$$\bar{s}_{t+1,i} = \frac{s_{t+1,i}}{1-\beta_1^t}$$

.

2. In **SGD with momentum**, updates are made as

$$v_{t+1} = \beta v_t + (1-\beta)\nabla_w \mathcal{L}(w_t) \tag{1}$$

$$w_{t+1} = w_t - \eta v_{t+1} \tag{2}$$

With ℓ_2 regularization, the momentum term used in (2) would become

$$v_{t+1} = \beta v_t + (1-\beta)\nabla_w \mathcal{L}(w_t) + 2\lambda w_t$$

Such that the regularization term is accumulated in the momentum and couples regularization with the optimization dynamics in a complicated manner. This would cause regularization to build up over time through the moving average v_t with weight β, and cause the effective regularization strength to depend on β in addition to λ.

Using **RMSProp**, updates are made as

$$\begin{aligned} v_{t+1} &= \beta v_t + (1-\beta)(\nabla_w \mathcal{L}(w_t))^2 \\ w_{t+1} &= w_t - \frac{\eta}{\sqrt{v_{t+1}} + \epsilon} \nabla_w \mathcal{L}(w_t) \end{aligned}$$

With ℓ_2 regularization, the update rules would both change:

$$\begin{aligned} v_{t+1} &= \beta v_t + (1-\beta)(\nabla_w \mathcal{L}(w_t) + 2\lambda w_t)^2 \\ &\qquad \text{(the factor of 2 appears since we differentiate } \lambda||w||^2 \text{ with respect to } w) \\ w_{t+1} &= w_t - \frac{\eta}{\sqrt{v_{t+1}} + \epsilon}(\nabla_w \mathcal{L}(w_t) + 2\lambda w_t) \end{aligned}$$

This is problematic because the regularization term $2\lambda w_t$ gets squared and added to the adaptive learning rate denominator v_{t+1}. For parameters with large weights, this inflates v_{t+1}, which reduces the effective learning rate via the denominator term in the weight update $\frac{1}{\sqrt{v_{t+1}}}$. This means parameters that need the most regularization (those with large values) receive smaller gradient updates, thereby weakening the regularization effect. The regularization strength becomes coupled with the adaptive learning rate mechanism in an undesirable way.

Using **Adam**, updates are made as

$$w_{t+1} = w_t - \eta \frac{\bar{s}_{t+1}}{\sqrt{\bar{r}_{t+1} + \delta}}$$

where

$$\begin{aligned} g_t &= \nabla_w \mathcal{L}(w_t) \\ s_{t+1} &= \beta_1 s_t + (1-\beta_1) g_t \qquad \bar{s}_{t+1} = \frac{s_{t+1}}{1-\beta_1^{t+1}} \qquad (1) \\ r_{t+1} &= \beta_2 r_t + (1-\beta_2) g_t^2 \qquad \bar{r}_{t+1} = \frac{r_{t+1}}{1-\beta_2^{t+1}} \qquad (2) \end{aligned}$$

With ℓ_2 regularization, the regularization term is accumulated into both the moving average and variance terms:

$$\begin{aligned} g_t &= \nabla_w \mathcal{L}(w_t) + 2\lambda w_t \\ \implies s_{t+1} &= \beta_1 s_t + (1-\beta_1)(\nabla_w \mathcal{L}(w_t) + 2\lambda w_t) \\ \text{and } r_{t+1} &= \beta_2 r_t + (1-\beta_2)(\nabla_w \mathcal{L}(w_t) + 2\lambda w_t)^2 \end{aligned}$$

This combines the issue of SGD and RMSProp when ℓ_2 regularization is applied. The regularization term $2\lambda w_t$ accumulates in the momentum s_t, coupling regularization strength with β_1. Additionally, the squared regularization term inflates v_t, reducing the effective step size for large weights. Critically, parameters with large weights receive smaller updates, which weakens regularization when it should be strongest.

3. The AdamW update rule applies the weight penalty after the adaptive moment-based update, applying weight decay directly to the weights:

$$w_{t+1} = w_t - \eta \frac{\bar{s}_{t+1}}{\sqrt{\bar{r}_{t+1} + \delta}} - \eta \lambda w_t$$

g_t, and therefore s_t and r_t, are maintained as in the original Adam formulation:

$$\begin{aligned} g_t &= \nabla_w \mathcal{L}(w_t) \\ s_{t+1} &= \beta_1 s_t + (1-\beta_1) g_t \\ r_{t+1} &= \beta_2 r_t + (1-\beta_2) g_t^2 \end{aligned}$$

This ensures that all parameters receive regularization proportional to their magnitude regardless of their gradient history, that the regularization strength is decoupled from the optimizer's momentum and adaptive learning rate mechanisms, and that large weights are penalized consistently.