Statistical Machine Learning: Coursework 3

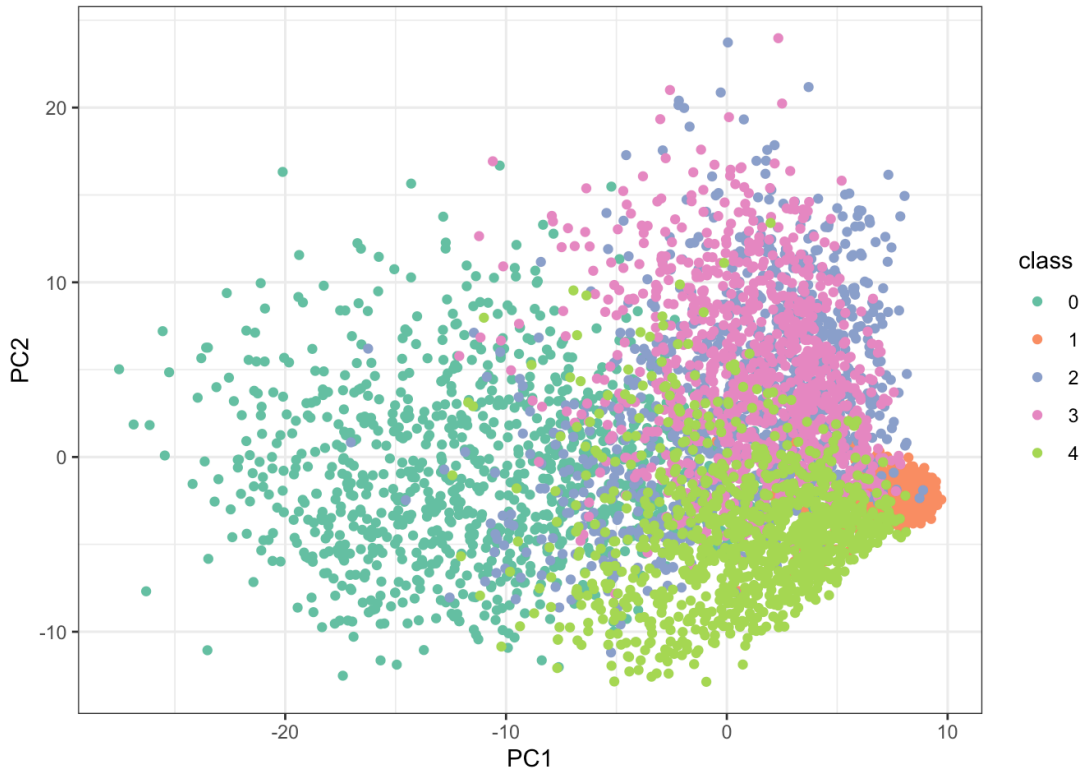
Student Name: Harleen Gulati , Student Number: 2101550

2024-02-28

Question 1

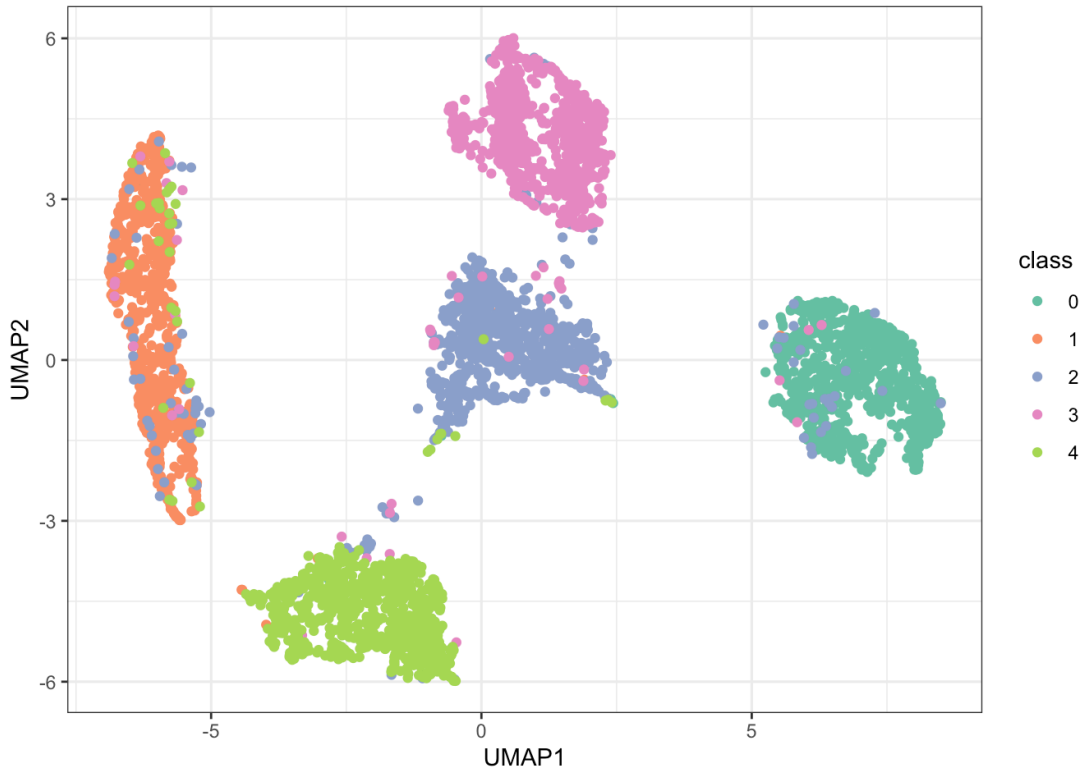
Question 1a

```
mnist_pca <- dim_red_fn(step_pca(num_comp=2),  data=mnist_df,label=class) # performing principal componen
t analysis with 2 components

mnist_pca%>% # visualising the data
  ggplot(aes(x=PC1,y=PC2,color=class))+
  geom_point()+
  scale_color_brewer(palette="Set2")+
  theme_bw()
```



Question 1b

```
mnist_kpca<- dim_red_fn(step_umap(num_comp=2),  data=mnist_df,label=class) # replaces step_pca with step_
umap so now performing non-linear dimensionality reduction

mnist_kpca%>% # visualizing the data
  ggplot(aes(x=UMAP1,y=UMAP2,color=class))+
  geom_point()+
  scale_color_brewer(palette="Set2")+
  theme_bw()
```



Question 1c

The linear method of principal components analysis has a lower computational complexity compared to non-linear methods. Thus, linear methods of dimensionality reduction may be faster to compute than non-linear methods, especially for large datasets and hence linear methods of principal component analysis may be preffered over non-linear methods for scenarios where computational efficiency is key to consider (e.g., for larger datasets).

Question 1d

The non-linear method used in part b (uniform manifold and projection) is preferable to the linear methods as it allows for the capturing of non-linear relationships present in the data which can then be seen in the visualizations. This allows us to capture relationships and intricate details of the data which may not have been able to be captured with linear methods; thus we can understand the underlying structure of the data better (e.g., how features are related). This may be preffered in many settings where datasets exhibit complex, non-linear relationships between features and we want to capture how these features are related more accurately.

Question 2

Question 2 a

We can write $Z = U \sum V^T V_k$

where U is the n x r matrix of left singular vectors ,V is the p x r matrix of right singular vectors, $\sum$ is the r x r diagonal matrix of singular values and V_k is the p x k matrix of the first k right singular vectors.

Now we observe $Z^T = (U \sum V^T V_k)^T = V_k^T V \sum U^T$

We also observe

$$\begin{aligned}(Z^T Z) &= (V_k^T V \sum U^T U \sum V^T V_k) \\ &= (V_k^T V \sum \sum V^T V_k) \\ &= (V_k^T V A V^T V_k)\end{aligned}$$

where we use the orthonormality of U and where $A = diag(\sigma_1^2, \ldots, \sigma_r^2)$.

Now we observe as follows:

$V_k^T V$ is a k x r matrix with entries 1 in each row i and column j where i = j and i $\leq$ k (i.e. up to the first k diagonals) and 0 elsewhere (which we can deduce from the orthonormality of V).

$V^T V_k$ is a r x k matrix with entries 1 in each row i and column j where i = j and i $\leq$ k (i.e. up to the first k diagonals) and 0 elsewhere (which we also deduce from the orthonormality of V).

Thus, from the above we can deduce $V_k^T VA$ to be a k x r matrix (since $V_k^T V$ is a k x r matrix and A is a r x r matrix) with entries in each row i and column j where i = j being σ_i^2 and all other entries being 0.

Now from the above, we deduce $Z^T Z = (V_k^T VA)(V^T V_k)$ to be a k x k matrix (since $(V_k^T VA)$ is a k x r matrix and $(V^T V_k)$ is a r x k matrix) with each diagonal having entries σ_i^2.

Thus, $(Z^T Z)^{-1}$ is a k x k matrix with each diagonal having entries $1/\sigma_i^2$.

Now consider $Z^T = V_k^T V \sum U^T$. We mentioned $V_k^T V$ to be a k x r matrix with entries 1 in each row i and column j with i = j and entries 0 elsewhere (from the orthonormality of V). Thus we can deduce $V_k^T V \sum$ to be a k x r matrix (since $V_k^T V$ is k x r and $\sum$ is r x r) with entries σ_i for each row i and column j with i = j and entries 0 elsewhere.

Thus $Z^T = V_k^T V \sum U^T$ is a k x n matrix (since $V_k^T V \sum$ is a k x r matrix and U^T is a r x n matrix) where each row i has value $\sigma_i u_i^T$

Now we use the fact that $(Z^T Z)^{-1}$ is a k x k matrix with each diagonal having entries $1/\sigma_i^2$ and $Z^T = V_k^T V \sum U^T$ is a k x n matrix where each row i has value $\sigma_i u_i^T$ to deduce that $(Z^T Z)^{-1} Z^T$ is a k x n matrix where each row i has value u_i^T/σ_i.

Now, using the fact that $(Z^T Z)^{-1} Z^T$ is a k x n matrix where each row i has value u_i^T/σ_i we can deduce that $(Z^T Z)^{-1} Z^T y^\circ$ is a k x 1 matrix (since y° is a n x 1 matrix) where each row i has entry $\frac{u_i^T y^\circ}{\sigma_i}$.

Now consider

$$\begin{aligned}\varphi(x)^T &= [V_k^T(x - \bar{x})]^T \\ &= (x - \bar{x})^T V_k \\ &= [(x - \bar{x})^T v_1, \ldots, (x - \bar{x})^T v_k]\end{aligned}$$

which is a 1xk matrix.

Using the fact $\varphi(x)^T$ is a 1 x k matrix and $(Z^T Z)^{-1} Z^T y^\circ$ is a k x 1 matrix where each row i has entry $\frac{u_i^T y^\circ}{\sigma_i}$ we deduce that $\varphi(x)^T (Z^T Z)^{-1} Z^T y^\circ$ is a scalar value as follows:

$$\begin{aligned}&\sum_{j=1}^{k} (x - \bar{x})^T v_j \frac{u_j^T y^\circ}{\sigma_j} \\ &= \sum_{j=1}^{k} \frac{u_j^T y^\circ}{\sigma_j} (x - \bar{x})^T v_j \\ &= \sum_{j=1}^{k} \frac{u_j^T y^\circ}{\sigma_j} v_j^T (x - \bar{x}) \\ &= \sum_{j=1}^{k} v_j^T \frac{u_j^T y^\circ}{\sigma_j} (x - \bar{x}) \\ &= \sum_{j=1}^{k} (\frac{u_j^T y^\circ}{\sigma_j} v_j)^T (x - \bar{x}) \\ &= w_k^T (x - \bar{x})\end{aligned}$$

where the first equality comes from using the fact that $\frac{u_j^T y^\circ}{\sigma_j}$ and $(x - \bar{x})^T v_j$ are scalar quantities so we can swap them around.

The second equality then comes from the fact that $(x - \bar{x})^T v_j = v_j^T (x - \bar{x})$ since this is an inner product.

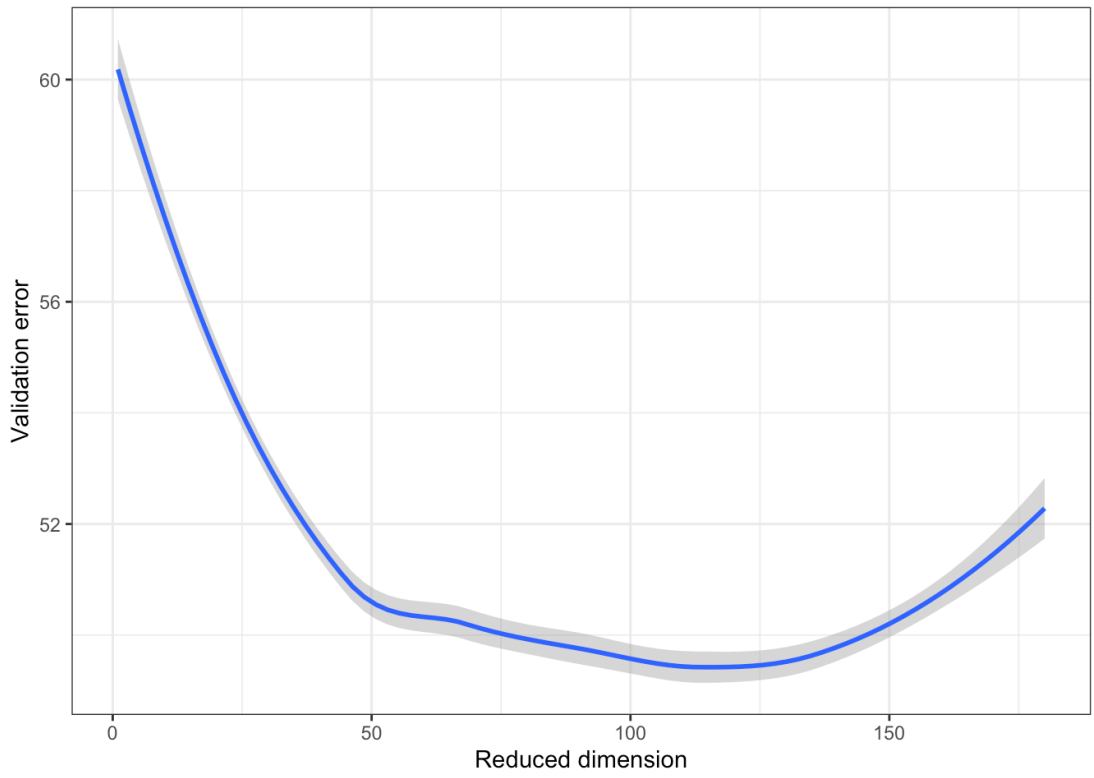
The third equality then follows through by the fact that $\frac{u_j^T y^\circ}{\sigma_j}$ is a scalar and so we can swap $\frac{u_j^T y^\circ}{\sigma_j}$ and v_j around and the fourth inequality also follows through on this principle (because the transpose of a scalar quantity is itself i.e. $((\frac{u_j^T y^\circ}{\sigma_j} v_j)^T = v_j^T (\frac{u_j^T y^\circ}{\sigma_j})^T = v_j^T \frac{u_j^T y^\circ}{\sigma_j}$).

Thus we have deduced that $\varphi(x)^T (Z^T Z)^{-1} Z^T = w_k^T (x - \bar{x})$ and hence $\phi_k(x) = w_k^T (x - \bar{x}) + \bar{y}$ as required.

Question 2 b

```
library(rsvd)
compute_pc_reg_weights<-function(x,y,k,rand_seed=0){

  # set random seed
  set.seed(rand_seed)

  # compute the truncated svd
  svd_x=rsvd(x, k = k)

  # estimate the weight vector
  y_bar = mean(y) # get mean of observations
  y_centered = y - y_bar # get centered y vector
  left_sing <- svd_x$u # left singular vectors of truncated SVD
  sing_vals <- svd_x$d # singular values of truncated SVD
  right_sing <- svd_x$v # right singular vectors of truncated SVD
  w = 0 # initial value of w
  for (x in 1:k) {
    w = w + (((t(left_sing[, x]) %*% y_centered) / sing_vals[x]) * right_sing[, x])
  } # sums from 1 to k using the value of w defined in the previous question

  return(w)

}
suppressWarnings(pc_reg_test_fn(compute_pc_reg_weights))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Question 3

Question 3a

We note that:

For $X \sim$ Unif[a,b], $\mathbb{E}[X] = \frac{a+b}{2}$ $(*)$, $Var(X) = \frac{(b-a)^2}{12}$ $(**)$ and for X a random variable $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

So for $Z_l(w)$ we can calculate $\mathbb{E}[Z_\ell(w)^2] = Var(Z_\ell(w)) + \mathbb{E}[Z_\ell(w)]^2$

Firstly, we find

$$\mathbb{E}[Z_\ell(w)] = \mathbb{E}[\sum_{j=1}^{p} a_{\ell j} w_j] = \sum_{j=1}^{p} \mathbb{E}[a_{\ell j} w_j] \text{ (by linearity of expectation)}$$

$$= \sum_{j=1}^{p} w_j \mathbb{E}[a_{\ell j}] \text{ (since w is fixed)}$$

$$= \sum_{j=1}^{p} w_j [\frac{-1+1}{2}] = \sum_{j=1}^{p} w_j * 0 = 0 \text{ (by property of a uniform distribution expectation as mentioned in (*))}$$

Thus we can deduce $\mathbb{E}[Z_\ell(w)]^2 = 0$

Now consider

$$Var(Z_\ell(w)) = Var(\sum_{j=1}^{p} a_{\ell j} w_j)$$

$$= \sum_{j=1}^{p} Var(a_{\ell j} w_j) \text{ (by mutual independence of } a_{\ell j} \text{ for all } \ell \in [1, .., k] \text{ and j} \in [1, .., p])$$

$$= \sum_{j=1}^{p} w_j^2 Var(a_{lj}) \text{ (since w is fixed)}$$

$$= \sum_{j=1}^{p} w_j^2 [\frac{(1-(-1))^2}{12}] = \sum_{j=1}^{p} w_j^2 * \frac{1}{3} \text{ (by property of variance of a uniform distribution as mentioned in (**))}$$

$$= \frac{1}{3} \sum_{j=1}^{p} w_j^2$$

Thus
$\mathbb{E}[Z_\ell(w)^2] = Var(Z_\ell(w)) + \mathbb{E}[Z_\ell(w)]^2 = \frac{1}{3} \sum_{j=1}^{p} w_j^2 + 0 = \frac{1}{3} \sum_{j=1}^{p} w_j^2 = \frac{1}{3} \sum_{j=1}^{p} 1 = \frac{1}{3} p$ (where we have used w is a unit vector)

Question 3b

Fix a unit vector $w \in \mathbb{R}^p$ and for each $\ell \in [1, .., k]$ set $Z_l(w) = \sum_{j=1}^{p} a_{lj} w_j$.

We first show that for $\lambda \leq \frac{k}{4}$ $\mathbb{E}[e^{\lambda[\frac{1}{k} \sum_{\ell=1}^{k} 3Z_l(w)^2 - 1]}] \leq e^{\frac{2\lambda^2}{k}}$ (*)

We have:

$$\mathbb{E}[e^{\lambda[\frac{1}{k} \sum_{l=1}^{k} 3Z_l(w)^2 - 1]}]$$

$$= \mathbb{E}[\prod_{l=1}^{k} e^{\frac{\lambda}{k}[3Z_l(w)^2 - 1]}]$$

$$= \prod_{l=1}^{k} \mathbb{E}[e^{\frac{\lambda}{k}[3Z_l(w)^2 - 1]}] \text{ (by mutual independence of } a_{lj} \text{ and by } Z_l(w) = \sum_{j=1}^{p} a_{lj} w_j)$$

$$\leq \prod_{l=1}^{k} e^{2(\frac{\lambda}{k})^2} \text{ (by Lemma M.G.F. given in the question)}$$

$$= e^{\frac{2\lambda^2 k}{k^2}} = e^{\frac{2\lambda^2}{k}}$$

thus we have shown (*)

We now will show that for $t \in [0, 1]$ $\mathbb{P}(|\frac{1}{k} \sum_{l=1}^{k} 3Z_l(w)^2 - 1| > t) \leq 2e^{\frac{-t^2 k}{8}}$ (**)

We take $> 0 $ and $\lambda \leq k/4$ and apply 1) Union Bound , 2) Markov Lemma and 3) (*)

This gives us:

$$\mathbb{P}[|\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1| > t]$$

$$= \sum_{\omega\in\{-1,+1\}} \mathbb{P}[\omega(\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1) > t]$$

$$= \sum_{\omega\in\{-1,+1\}} \mathbb{P}[e^{\lambda\omega(\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1)} > e^{\lambda t}]$$

$$\leq \sum_{\omega\in\{-1,+1\}} \mathbb{E}[e^{\lambda\omega(\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1)}]e^{-\lambda t}$$

$$\leq 2e^{\frac{2\lambda^2}{k} - \lambda t}$$

taking $\lambda = \frac{tk}{4}$ minimises the bound and thus we deduce

$$\mathbb{P}(|\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1| > t) \leq 2e^{\frac{-t^2k}{8}}$$

We will show now that for $\epsilon \in (0, 1)$

$$\mathbb{P}\{(1-\epsilon) \leq \frac{||\varphi(u) - \varphi(v)||^2}{||u - v||^2} \leq (1+\epsilon)\} \geq 1 - 2e^{\frac{-k\epsilon^2}{8}}$$

Now let $u, v \in \mathbb{R}^p$ be distinct points and let $w = \frac{u-v}{||u-v||}$

Consider

$$\frac{3}{k}\sum_{l=1}^{k} Z_l(w)^2$$

$$= \frac{3}{k}\sum_{l=1}^{k}(\sum_{j=1}^{p} a_{lj}w_j)^2$$

$$= \frac{3}{k}||(\sum_{j=1}^{p} a_{lj}w_j)_{l=1,..,k}||^2$$

$$= \frac{3}{k}||Aw||^2$$

$$= \frac{3||A(u-v)||^2}{k||u-v||^2}$$

$$= \frac{||\varphi(u) - \varphi(v)||^2}{||u-v||^2} (***)$$

Now (**) told us

$$\mathbb{P}(|\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1| > \epsilon) \leq 2e^{\frac{-t^2k}{8}}$$

$$\implies \mathbb{P}(|\frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1| \leq \epsilon) \geq 1 - 2e^{\frac{-t^2k}{8}}$$

$$\implies \mathbb{P}(-\epsilon \leq \frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 - 1 \leq \epsilon) \geq 1 - 2e^{\frac{-t^2k}{8}}$$

$$\implies \mathbb{P}(1-\epsilon \leq \frac{1}{k}\sum_{l=1}^{k} 3Z_l(w)^2 \leq 1+\epsilon) \geq 1 - 2e^{\frac{-t^2k}{8}}$$

$$\implies \mathbb{P}(1-\epsilon \leq \frac{||\varphi(u) - \varphi(v)||^2}{||u-v||^2} \leq 1+\epsilon) \geq 1 - 2e^{\frac{-t^2k}{8}} \text{ (using (***))}$$

Since $\mathbb{P}(1-\epsilon \leq \frac{||\varphi(u)-\varphi(v)||^2}{||u-v||^2} \leq 1+\epsilon) \geq 1-2e^{\frac{-\epsilon^2 k}{8}} \geq 1-\frac{2\delta}{|\mathbb{T}|(|\mathbb{T}|-1)}$ we can deduce $\mathbb{P}(\cap_{u,v\in\mathbb{T}}(1-\epsilon)||u-v||^2 \leq ||\varphi(u)-\varphi(v)||^2 \leq (1+\epsilon)||u-v||^2) \geq 1-\delta$ from the union bound since there are $\binom{|\mathbb{T}|}{2} = \frac{|\mathbb{T}|(|\mathbb{T}|-1)}{2}$ distinct pairs in $\mathbb{T}$

Question 3c

An attempt at the question is as follows:

Consider for $\lambda \leq \frac{1}{4}$

$$\begin{aligned}
&\mathbb{E}[e^{\lambda(3Z_l(w)^2-1)}] \\
&= \mathbb{E}[e^{3\lambda Z_l(w)^2-\lambda}] \\
&= \mathbb{E}[e^{-\lambda}e^{3\lambda Z_l(w)^2}] \\
&= e^{-\lambda}\mathbb{E}[e^{3\lambda Z_l(w)^2}]
\end{aligned}$$

Now consider $\mathbb{E}[e^{3\lambda Z_l(w)^2}]$

We know that e^x is convex and thus by Jensen's inequality

$$\begin{aligned}
\mathbb{E}[e^{3\lambda Z_l(w)^2}] &\geq e^{\mathbb{E}[3\lambda Z_{\ell'}(w)^2]} \\
&= e^{3\lambda\mathbb{E}[Z_{\ell'}(w)^2]} \\
&= e^{3\lambda\frac{1}{3}p} = e^{\lambda p}
\end{aligned}$$

Now suppose $\lambda < 0$ then by the hint given in the question we have

$$\begin{aligned}
e^{\lambda p} &\leq 1+\lambda p+\frac{\lambda^2 p^2}{2} \\
&\leq 1+\frac{\lambda^2 p^2}{2} \text{ (because } \lambda < 0) \\
&\leq 1+\frac{\lambda^2}{2} \text{ (since } p \in \mathbb{N}) \\
&\leq \frac{\lambda^2}{2}+\frac{\lambda^2}{2} (\text{since } \lambda \leq \frac{1}{4}) \\
&= \lambda^2 \leq 3\lambda^2
\end{aligned}$$

Suppose now that $0 < \lambda \leq \frac{1}{4}$ then by the hint in the question we have $e^{\lambda p} \geq 1+\lambda p$