למידה עמוקה 046211
גליון יבש 2

גל גרנות	**315681593**
ניר טבת	**208731422**

שאלה 1

הגדרנו את ה-Bayes Risk ואת הסיכון של w_μ בתור:

$$\bar{R}(w) = E_{\epsilon, w_{true}}(R), \qquad R(w_\mu) = \left\|w_\mu - w_{true}\right\|^2 = \left\|\left(H_\mu^{-1}H - I\right)w_{true} + H_\mu^{-1}X^T\epsilon\right\|^2$$

נוכיח כי:

$$\bar{R}(w_\mu) = \sum_{i=1}^{d} \frac{(\sigma_w^2/d)\mu^2 + \sigma_\epsilon^2\lambda_i}{(\lambda_i + \mu)^2}$$

עבור:

1. $\epsilon \sim N(0, \sigma_\epsilon^2 I)$
2. $w_{true} \sim N\left(0, \frac{\sigma_w^2}{d} I\right)$
3. $H_\mu = \mu I + X^T X$
4. $H = X^T X$
5. λ_i העו"ע של המטריצה H
6. $w_\mu = H_\mu^{-1} X^T y = (\mu I + X^T X)^{-1} X^T y$

הוכחה:

$$\left\|\left(H_\mu^{-1}H-I\right)w_{true}+H_\mu^{-1}X^T\epsilon\right\|^2$$

$$=\left(\left(H_\mu^{-1}H-I\right)w_{true}+H_\mu^{-1}X^T\epsilon\right)^T\left(\left(H_\mu^{-1}H-I\right)w_{true}+H_\mu^{-1}X^T\epsilon\right)$$

$$=\left(\left(\left(H_\mu^{-1}H-I\right)w_{true}\right)^T+\left(H_\mu^{-1}X^T\epsilon\right)^T\right)\left(\left(H_\mu^{-1}H-I\right)w_{true}+H_\mu^{-1}X^T\epsilon\right)$$

$$=\left\|\left(H_\mu^{-1}H-I\right)w_{true}\right\|^2+\left\|H_\mu^{-1}X^T\epsilon\right\|^2+\text{cross terms}=*$$

מכיוון ש-ϵ ו-w_{true} חסרי קורלציה ($E(\epsilon w_{true})=0$) איברי ה-cross יתאפסו תחת התוחלת ולכן נתעלם מהם. נחשב בנפרד את התוחלת עבור הנורמות של שני האיברים:

$$=\left\|\left(H_\mu^{-1}H-I\right)w_{true}\right\|^2=\left(\left(H_\mu^{-1}H-I\right)w_{true}\right)^T\left(H_\mu^{-1}H-I\right)w_{true}$$

$$=w_{true}^T\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)w_{true}$$

הביטוי המתקבל הוא סקלר, או מטריצה מסדר 1×1 ולכן שווה לעקבה של עצמו. נוכיח זהות קצרה בשימוש בזהות הציקלית של העקבה:

$$a^Tb=Tr(ba^T)\Longrightarrow a^TMa=Tr(Maa^T)$$

נשתמש בזהות עם $a=w_{true}$ ו-$M=\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)$:

$$w_{true}^T\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)w_{true}=Tr\left(\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)w_{true}w_{true}^T\right)$$

נפעיל את התוחלת ונשתמש בלינאריות העקבה והתוחלת, כאשר $E(w_{true}w_{true}^T)$ היא מטריצת הקוואריאנס של הוקטור האקראי w_{true} שלפי הנתון היא $\frac{\sigma_w}{d}I$:

$$E\left(Tr\left(\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)w_{true}w_{true}^T\right)\right)=Tr(E\left(\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)w_{true}w_{true}^T\right)$$

$$Tr\left(\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)E(w_{true}w_{true}^T)\right)=Tr\left(\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)\frac{\sigma_w^2}{d}I\right)$$

$$\frac{\sigma_w^2}{d}Tr\left(\left(H_\mu^{-1}H-I\right)^T\left(H_\mu^{-1}H-I\right)\right)$$

נחזור לאיבר השני ב-*, ונשתמש בעובדה ש-H_μ^{-1} סימטרית ולכן שווה לשחלוף שלה:

$$\left\|H_\mu^{-1}X^T\epsilon\right\|^2=\left(H_\mu^{-1}X^T\epsilon\right)^TH_\mu^{-1}X^T\epsilon=\epsilon^TXH_\mu^{-1}H_\mu^{-1}X^T\epsilon$$

$$E\left(\left\|H_\mu^{-1}X^T\epsilon\right\|^2\right)=E\left(\epsilon^TXH_\mu^{-1}H_\mu^{-1}X^T\epsilon\right)=\sum_{i=1}^N\sum_{j=1}^N E\left(\epsilon_i\epsilon_j\right)\left(XH_\mu^{-1}\right)_i\left(H_\mu^{-1}X^T\right)_j$$

$$=\sum_{i=1}^N\sum_{j=1}^N\delta_{ij}\sigma_\epsilon^2\left(XH_\mu^{-1}\right)_i\left(H_\mu^{-1}X^T\right)_j=\sigma_\epsilon^2\sum_{i=1}^N\left(XH_\mu^{-1}\right)_i\left(H_\mu^{-1}X^T\right)_i=\sigma_\epsilon^2Tr\left(XH_\mu^{-2}X^T\right)$$

נאחד את התוצאות שקיבלנו:

$$\bar{R}(w_\mu) = \frac{\sigma_w^2}{d} Tr\left(\left(H_\mu^{-1}H - I\right)^T\left(H_\mu^{-1}H - I\right)\right) + \sigma_\epsilon^2 Tr\left(XH_\mu^{-2}X^T\right)$$

H היא מטריצה סימטרית ולכן קיים לה לכסון אורתוגונלי $U\Lambda U^T$ עבור U אורתוגונלית ו-Λ אלכסונית. נקבל:

$$= \frac{\sigma_w^2}{d} Tr\left(\left(H_\mu^{-1}H - I\right)^2\right) + \sigma_\epsilon^2 Tr\left(XX^T H_\mu^{-2}\right)$$

$$= \frac{\sigma_w^2}{d} Tr(((H + \mu I)^{-1}H - I)^2) + \sigma_\epsilon^2 Tr(H(H + \mu I)^{-2})$$

$$= \frac{\sigma_w^2}{d} Tr((U\Lambda U^T + \mu I)^{-1}U\Lambda U^T - I)^2) + \sigma_\epsilon^2 Tr(U\Lambda U^T(U\Lambda U^T + \mu I)^{-2})$$

נשתמש בעובדה כי $UU^T = I$ ובחילוף לכפל של מטריצת היחידה μI, נכפול משמאל וימין בהתאמה:

$$= \frac{\sigma_w^2}{d} Tr((U\Lambda U^T + U\mu I U^T)^{-1}U\Lambda U^T - I)^2) + \sigma_\epsilon^2 Tr(U\Lambda U^T(U\Lambda U^T + U\mu I U^T)^{-2})$$

$$= \frac{\sigma_w^2}{d} Tr((U(\Lambda + \mu I)U^T)^{-1}U\Lambda U^T - I)^2) + \sigma_\epsilon^2 Tr(U\Lambda U^T(U\Lambda U^T + U\mu I U^T)^{-2})$$

$$= \frac{\sigma_w^2}{d} Tr\left(U\left((\Lambda + \mu I)\right)^{-1}U^T U\Lambda U^T - I\right)^2) + \sigma_\epsilon^2 Tr(U\Lambda U^T U(\Lambda + \mu I)^{-2}U^T)$$

$$= \frac{\sigma_w^2}{d} Tr(U(\Lambda + \mu I)^{-1}U^T - I)^2) + \sigma_\epsilon^2 Tr(U\Lambda(\Lambda + \mu I)^{-2}U^T)$$

$$= \frac{\sigma_w^2}{d} Tr(U(\Lambda + \mu I)^{-1}\Lambda U^T - I)^2 + \sigma_\epsilon^2 Tr(U\Lambda(\Lambda + \mu I)^{-2}U^T)$$

המטריצה $\Lambda + \mu I$ אלכסונית ולכן ההופכית שלה היא אלכסונית עם כל האלמנטים ההפוכים:

$$\Lambda + \mu I = diag(\lambda_i) + diag(\mu) = diag(\lambda_i + \mu)$$

$$= \frac{\sigma_w^2}{d} Tr\left(U \cdot diag\left(\frac{1}{\lambda_i + \mu}\right) \cdot diag(\lambda_i)U^T - I\right)^2 + \sigma_\epsilon^2 Tr\left(U \cdot diag(\frac{\lambda_i}{(\lambda_i + \mu)^2}U^T\right)$$

$$= \frac{\sigma_w^2}{d} Tr\left(diag\left(\frac{\lambda_i}{\lambda_i + \mu} - 1\right)\right)^2 + \sigma_\epsilon^2 Tr\left(diag\left(\frac{\lambda_i}{(\lambda_i + \mu)^2}\right)\right)$$

$$= \frac{\sigma_w^2}{d}\sum_{i=1}^{d}\left(\frac{\lambda_i - \lambda_i - \mu}{\lambda_i + \mu}\right)^2 + \sigma_\epsilon^2\sum_{i=1}^{d}\frac{\lambda_i}{(\lambda_i + \mu)^2} = \sum_{i=1}^{d}\left(\frac{\frac{\sigma_w^2}{d}\mu^2}{(\lambda_i + \mu)^2} + \frac{\sigma_\epsilon^2}{(\lambda_i + \mu)^2}\right)$$

$$\sum_{i=1}^{d}\frac{\frac{\sigma_w^2}{d}\mu^2 + \sigma_\epsilon^2\lambda_i}{(\lambda_i + \mu)^2}$$

כנדרש.

שאלה 2

1. נתון ש-w נדגם בצורה אחידה מתוך Q אשר מכיל q ערכים. וכן נתון שכל רכיב בוקטור w נדגם באופן i.i.d לשאר הרכיבים.

אם w היה וקטור באורך 1, קל היה לראות שמתקבל: $P_{w\sim P_w}(w = w_\star) = \frac{1}{q}$

לכן כיוון ש-w באורך k ומכיל k איברים i.i.d ניתן לומר כי:

$$P_{w\sim P_w}(w = w_\star) = \left(\frac{1}{q}\right)^k$$

2. כדי שהביטוי יתקיים נצטרך שכל המשקולות יתאימו בין הרשתות. נשים לב שיש במטריצה W_1, $d_0 * d_1$ משקולות, וכן במטריצה W_2 יש d_1 משקולות. כמו שראינו בסעיף 1, לכל משקולת יש הסתברות של $\frac{1}{q}$ להתאים בין שתי הרשתות. כמו כן אנחנו יודעים שיש רק $d_{1\star} < d_1$ משקולות שאינן אפס בשכבה 1. סך הכל נקבל:

$$p_\star \geq \left(\frac{1}{q}\right)^{d_0 d_{\star 1} + d_1} = q^{-d_1 - d_0 d_{1\star}}$$

3. T הוא זמן העצירה של הרשת. ניתן להסתכל על T כעל רצף של ניסויי ברנולי עם פרמטר הצלחה $p_\star$ שהוגדר בסעיף 2. כלומר T מתפלג גאומטרית.
לכן ניתן לומר:

$$P(t > T) = (1 - p_\star)^T$$

נפעיל לוג על שני האגפים:

$$\log\big(P(t > T)\big) = T \cdot log(1 - p_\star)$$

$$T = \frac{\log\big(P(t > T)\big)}{log(1 - p_\star)} \rightarrow \lfloor T \rfloor \leq \frac{\log\big(P(t > T)\big)}{log(1 - p_\star)}$$

4. משילוב של שני הקירובים שנתונים ברמז והצבה שלהם בתוצאה של סעיף 3 נקבל:

$$T = \frac{\log\big(P(t > T)\big)}{-p_\star}$$

נציב את התוצאה מסעיף 1 ונקבל:

$$T \leq -(\log\big(P(t > T)\big)\, q^{d_1 + d_0 d_{1\star}}$$

נגדיר $\eta = \log\big(P(t > T)\big)$

כמו כן במקרה שלנו:

$$\log(|\mathcal{F}|) = \log T \leq \log\left(\frac{log\eta}{-p_\star}\right) = loglog\eta + logp_\star \leq loglog\left(\frac{1}{\eta}\right) + (d_0 d_{1\star} + d_1)logq$$

כעת נציב הכל במשפט 2 ונקבל:

$$\epsilon < \frac{loglog\left(\frac{1}{\eta}\right) + (d_0 d_{1\star} + d_1)logq + \log\left(\frac{1}{\delta}\right)}{N}$$

כנדרש.

5. ביטוי 4 עבור משקולות שיכולות לקבל q ערכים שונים נקבל:

$$\log|\mathcal{F}| = \mathrm{k}\log\mathrm{q} = (d_0 d_1 + d_1)\log\mathrm{q}$$

נקבל ביטוי שגדול יותר מהביטוי בסעיף 3 ולכן ההכללה של ביטוי 4 תהיה פחות טובה.

שאלה 3

1. בכל הגרפים הנקודה הקריטית היא הנקודה שבה ה- test lossמפסיק לעלות ומתחיל לרדת שוב.
2.

a. הגרף הכחול והירוק הם שני מודלים שונים, הכחול הוא מודל שמתרגם מגרמנית לאנגלית והירוק הוא מודל שמתרגם מאנגלית לצרפתית . אנו יכולים לראות כי ה-test loss לאחר הגדלת גודל המודל מתחיל לעלות בשלב מסוים ולאחר גודל מסוים מתחיל לרדת, בנוסף ניתן לראות כי ה – train loss מונוטוני יורד עם הגדלת גודל המודל ולכן זהו model-wise double decent מכיוון שהאזור הקריטי מופיע ב-test loss.

b. אנו יכולים לראות כי זהmodel-wise double decent ככל שכמות הפרמטרים במודל גדלה כך השגיאה קטנה, עבור מודלים גדולים, בניגוד לשגיאה הקלאסית. הנקודה הקריטית מתקבלת בנק' מקסימום בה הגרף מתחיל לרדת לאחר העלייה (בערך ברוחב 10 פרמטרים)

c. אנו יכולים לראות כי epoch-wise double decentכיוון שעבור המודל הגדול (האדום) אנחנו רואים שעלייה בכמות ה EPOCHSגורמת לירידה בשגיאה.

שאלה 4

1. ההתפלגות של W סימטרית -> ההתפלגות של U סימטרית:

$$E[\varphi^2(u_{l-1})] = E[\max(0, u_{l-1})^2] = \frac{1}{2}E\left[u_{l-1}^2\right] = \frac{1}{2}\sigma_{u_{l-1}}^2$$

ניתן לראות כי הביטוי המחושב זהה לחישוב השונות של u אך החלק השלילי מאופס. מכיוון שההתפלגות של u סימטרית, החלק השלילי תרומה שווה לתרומה של החלק החיובי ונקבל חצי מהשונות של u.

$$\sigma_l = \frac{1}{\sqrt{\sum_j E[\varphi^2(u_{l-1}[j])]}} = \frac{1}{\sqrt{\sum_j \frac{1}{2}\sigma_{u_{l-1}[j]}^2}} = \sqrt{\frac{2}{d_{l-1}}}$$

תחת ההנחה שהשונות של u_{l-1} היא 1.

2. לפי תאוריית הגבול המרכזי:

נדיר משתנה Z שמתפלג נורמלי עם תוחלת 0 ושונות 1.

$$\varphi(u_{l-1}) = \max(0, u_{l-1}) = \sigma_{u_{l-1}} \max(0, z)$$

$$E[\varphi^2(u_{l-1})] = E\left[\left(\sigma_{u_{l-1}} \max(0, z)\right)^2\right] = \sigma_{u_{l-1}}^2 E[\max(0, z)^2] = \frac{1}{2}\sigma_{u_{l-1}}^2 \sigma_z = \frac{1}{2}\sigma_{u_{l-1}}^2$$

וההמשך זהה לסעיף 1.

שאלה 5

<u>צד ראשון:</u> אם מתקיים $\forall \tau \in H, W[i,j] = W[\tau(i), \tau(j)]$ אז f_W היא *equivariant*:

$$f_W(\tau x) = \phi\big(W(\tau x)\big) = \phi\big(\tau(Wx)\big)$$

כאשר השוויון האחרון נובע מהנתון ומכך שϕ פועלת על כל אלמנט בנפרד.

מ-*equivariance* של *f* ל-τ נקבל:

$$\phi\big(\tau(Wx)\big) = \tau\phi(Wx) = \tau \cdot f_W(x)$$

ובסך הכול קיבלנו *equivariance* של *f* ל-*H*.

<u>צד שני:</u> אם f_W היא *equivariant* אז מתקיים $\forall \tau \in H, W[i,j] = W[\tau(i), \tau(j)]$

$$f_W(\tau x) = \tau f_W(x) \to \phi(W\tau x) = \tau \cdot \phi(Wx)$$

מחד-חד ערכיות של ϕ ניתן לומר שחייב להתקיים שוויון בארגומנטים:

$$W\tau x = \tau(Wx) \to \tau W = W\tau$$

ומכאן ניתן להסיק שלכל איבר ב*W* מתקיים:

$$W[i,j] = W[\tau(i), \tau(j)]$$

שאלה 6

Layer	Output_dim	Number of parametes
INPUT	224X224X3	0
CONV3-64	224X224X64	(3X3X3+1)X64=1792
ReLU	224X224X64	0
POOL2	112X112X64	0
CONV3-128	112X112X128	(3X3X64+1)X128=73856
ReLU	112X112X128	0
POOL2	56X56X128	0
CONV3-256	56X56X256	(3X3X128+1)X256=295168
ReLU	56X56X256	0
CONV3-256	56X56X256	(3X3X256+1)X256=590080
ReLU	56X56X256	0
POOL2	28X28X256	0
CONV3-512	28X28X512	(3X3X256+1)X512 = 2359808
ReLU	28X28X512	0
CONV3-512	28X28X512	(3X3X512+1)X512 = 4718592
ReLU	28X28X512	0
POOL2	14X14X512	0
CONV3-512	14X14X512	(3X3X512+1)X512 = 4718592
ReLU	14X14X512	0
CONV3-512	14X14X512	(3X3X512+1)X512 = 4718592
ReLU	14X14X512	0
POOL2	7X7X512	0
FC-4096	4096X1	(7X7X512+1)X4096 = 102764544
FC-4096	4096X1	(4096+1)X4096 = 16781312
FC-1000	1000X1	(4096+1)X1000 = 4097000
SOFTMAX	1000X1	0

1. מספר הפרמטרים הכולל: 114,119,336
2. החלק היחסי של הפרמטרים של שכבות ה-*FC*:

$$\frac{123,642,856}{141,119,336} \cdot 100\% = \ 87.62\%$$