Solution 1: 19 February 2019
CS-526 Learning Theory

Exercise 3.1

The hypothesis class $\mathcal{H}$ being PAC learnable with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$ means that there is a learning algorithm A such that when running A on $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ i.i.d. samples generated by $\mathcal{D}$ and labeled by f, with probability at least $1-\delta$, A returns a hypothesis $h \in \mathcal{H}$ with $L_{D,f}(h) \leq \epsilon$.

Given $0 < \epsilon_1 \leq \epsilon_2 < 1$, consider $m \geq m_{\mathcal{H}}(\epsilon_1,\delta)$, we have that with probability at least $1-\delta$, A returns a hypothesis $h \in \mathcal{H}$ with $L_{D,f}(h) \leq \epsilon_1 \leq \epsilon_2$. This implies that $m_{\mathcal{H}}(\epsilon_1,\delta)$ is a sufficient number of samples for accuracy ϵ_2. Therefore, $m_{\mathcal{H}}(\epsilon_1,\delta) \geq m_{\mathcal{H}}(\epsilon_2,\delta)$.

The proof of $m_{\mathcal{H}}(\epsilon,\delta_1) \geq m_{\mathcal{H}}(\epsilon,\delta_2)$ for $0 < \delta_1 \leq \delta_2 < 1$ follows analogously from the definition.

Exercise 3.3

The realizability assumption for $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$ implies that there is a circle such that any x inside it has label $y = 1$, and the learning task here is to distinguish this circle. Now consider an ERM algorithm which given a training sequence $S = \{(x_i, y_i)\}_{i=1}^m$, returns the hypothesis $\hat{h}$ corresponding to the tightest circle which contains all the positive instances in S where $y_i = 1$ and does not allow false negative predictions. With the realizability assumption let h^* be the circle with zero training error and r^* be the corresponding radius.

Let $\bar{r} \leq r^*$ be a scalar such that $\mathbb{P}_{x\sim\mathcal{D}}(x : \bar{r} \leq \|x\| \leq r^*) = \epsilon$ and $E = \{x \in \mathbb{R}^2 : \bar{r} \leq \|x\| \leq r^*\}$. We have

$$\begin{aligned}\mathbb{P}(L_{\mathcal{D}}(h_S) \geq \epsilon) &\leq \mathbb{P}(\text{no points in } S \text{ belongs to } E)\\ &= (1-\epsilon)^m\\ &\leq e^{-\epsilon m}\end{aligned}$$

The desired bound on the sample complexity follows from requiring $e^{-\epsilon m} \leq \delta$.

Exercise 3.7

Let g be any (potentially probabilistic) classifier from $\mathcal{X}$ to $\{0,1\}$. Note that for 0-1 loss

$$\begin{aligned}L_{\mathcal{D}}(g) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{g(x)\neq y}] = \mathbb{E}_{x\sim\mathcal{D}}\big[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}[\mathbb{1}_{g(x)\neq y}]\big] = \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{P}_{y\sim\mathcal{D}_{Y|x}}(g(X)\neq Y|X=x)],\\ L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{P}_{y\sim\mathcal{D}_{Y|x}}(f_{\mathcal{D}}(X)\neq Y|X=x)].\end{aligned}$$

We should compare the two conditional probabilities inside the expectation. Let $x \in \mathcal{X}$ and $a_x = \mathbb{P}(Y = 1|X = x)$. We have

$$\begin{aligned}
\mathbb{P}(g(X) \neq Y|X = x) &= \mathbb{P}(g(X) = 0|X = x) \cdot \mathbb{P}(Y = 1|X = x) \\
&+ \mathbb{P}(g(X) = 1|X = x) \cdot \mathbb{P}(Y = 0|X = x) \\
&= \mathbb{P}(g(X) = 0|X = x) \cdot a_x + \mathbb{P}(g(X) = 1|X = x) \cdot (1 - a_x) \\
&\geq \mathbb{P}(g(X) = 0|X = x) \cdot \min\{a_x, 1 - a_x\} \\
&+ \mathbb{P}(g(X) = 1|X = x) \cdot \min\{a_x, 1 - a_x\} \\
&= \min\{a_x, 1 - a_x\}.
\end{aligned}$$

When $g = f_{\mathcal{D}}$ we should replace $\mathbb{P}(g(X) = 0|X = x)$ by $\mathbb{1}_{a_x<1/2}$ and $\mathbb{P}(g(X) = 1|X = x)$ by $\mathbb{1}_{a_x\geq 1/2}$. Then the above inequality is tight:

$$\mathbb{P}(f_{\mathcal{D}}(X) \neq Y|X = x) = \mathbb{1}_{a_x<1/2} \cdot a_x + \mathbb{1}_{a_x\geq 1/2} \cdot (1 - a_x) = \min\{a_x, 1 - a_x\}.$$

Therefore, we have $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Exercise 3.8

1. Solved already in Exercise 3.7.
2. We have shown in Exercise 3.7 that the Bayes optimial predictor $f_{\mathcal{D}}$ is optimal w.r.t. $\mathcal{D}$; in other words, $f_{\mathcal{D}}$ is always better than any other learning algorithm w.r.t. $\mathcal{D}$.
3. Take $\mathcal{D}$ to be any probability distribution and $B = f_{\mathcal{D}}$.

Exercise 4.1

$\underline{1 \Rightarrow 2}$: Assume for every $\epsilon, \delta > 0$ there exists $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$

$$\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) < \delta. \tag{1}$$

Then using the definition of expectation

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] &\leq \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) \cdot 1 + \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) \leq \epsilon) \cdot \epsilon \\
&\leq \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) + \epsilon \\
&\leq \delta + \epsilon,
\end{aligned}$$

where the last inequality follows from the assumption (1). Now set $\delta = \epsilon$. We have for every $\epsilon > 0$ there exists $m(\epsilon, \epsilon)$ such that $\forall m \geq m(\epsilon, \epsilon)$

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq 2\epsilon. \tag{2}$$

So it is valid to pass both sides of (2) to the limit $\lim_{m\to\infty} \lim_{\epsilon\to 0}$, which gives

$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq 0.$$

Also by definition $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq 0$. Thus we conclude $\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$.

2 $\Rightarrow$ 1: Assume that $\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$. For every $\epsilon, \delta \in (0,1)$ there exists some $m_0 \in \mathbb{N}$ such that for every $m \geq m_0$, $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \epsilon\delta$. By Markov's inequality,

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) &\leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))]}{\epsilon} \\ &\leq \frac{\epsilon\delta}{\epsilon} \\ &= \delta.\end{aligned}$$

Exercise 4.2

Using Hoeffding's inequality on $L_{\mathcal{D}} \in [a,b]$ we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon) \leq 2\exp\Big(-\frac{2m\epsilon^2}{(b-a)^2}\Big).$$

Then we substitute this into the step where the union bound is used:

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}(\exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon) &\leq \sum_{h\in\mathcal{H}} \mathbb{P}_{S\sim\mathcal{D}^m}(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon) \\ &\leq 2|\mathcal{H}|\exp\Big(-\frac{2m\epsilon^2}{(b-a)^2}\Big)\end{aligned}$$

The desired bound on the sample complexity follows from requiring $2|\mathcal{H}|\exp\Big(-\frac{2m\epsilon^2}{(b-a)^2}\Big) \leq \delta$.

Solution 2: 26 February 2019
CS-526 Learning Theory

1. For every $\alpha \in [0,1]$, a convex function f satisfies
$$f(\alpha a + (1-\alpha)b) \le \alpha f(a) + (1-\alpha) f(b).$$
Substituting $f(X) = e^{\lambda X}$ and $\alpha = \frac{b-X}{b-a} \in [0,1]$ we get
$$e^{\lambda X} \le \frac{b-X}{b-a} e^{\lambda a} + \frac{X-a}{b-a} e^{\lambda b}.$$
Taking the expectation on both sides and using $\mathbb{E}[X] = 0$ we have
$$\mathbb{E}[e^{\lambda X}] \le \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}.$$

2. With $p = -a/(b-a)$ and $h = \lambda(b-a)$, we have
$$\begin{aligned}\log\left(\frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}\right) &= \log(e^{\lambda a}) + \log\left(\frac{b}{b-a} - \frac{a}{b-a} e^{\lambda(b-a)}\right)\\ &= \lambda a + \log\left(1 + \frac{a}{b-a} - \frac{a}{b-a} e^{\lambda(b-a)}\right)\\ &= -hp + \log(1 - p + pe^h).\end{aligned}$$

3. Let $\theta = \frac{pe^h}{1-p+pe^h}$. One can compute
$$L'(h) = -p + \theta, \qquad L''(h) = \theta(1-\theta) = -(\theta - \frac{1}{2})^2 + \frac{1}{4} \le \frac{1}{4}.$$
One can also verify $L(0) = L'(0) = 0$. Using these remarks on the equation $L(h) = L(0) + hL'(0) + (h^2/2)L''(\xi)$, we obtain $L(h) \le h^2/8$. Combining with the previous steps implies
$$\mathbb{E}[e^{\lambda X}] \le e^{L(\lambda(b-a))} \le e^{-\lambda^2 (a-b)^2/8}.$$

4. Let $X_i = Z_i - \mu$ and $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$. Using the monotonicity of the exponent function and Markov's inequality, we have
$$\mathbb{P}\big(\bar{X} \ge \epsilon\big) = \mathbb{P}(e^{\lambda \bar{X}} \ge e^{\lambda\epsilon}) \le e^{-\lambda\epsilon}\ \mathbb{E}[e^{\lambda \bar{X}}].$$
As X_i are independent, we have $\mathbb{E}[e^{\lambda \bar{X}}] = \prod_{i=1}^m \mathbb{E}[e^{\lambda X_i/m}]$. Also, the previous exercise provides $\mathbb{E}[e^{\lambda X_i/m}] \le e^{-\lambda^2(a-b)^2/(8m^2)}$. So we conclude
$$\mathbb{P}(\bar{X} \ge \epsilon) \le \exp\big(-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}\big).$$

5. The exponent $-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}$ is a quadratic (convex) function of λ. It is minimized when $\lambda = 4m\epsilon/(b-a)^2$. This optimization gives the desired bound.

Solution of Graded Homework 1: 19 March 2019
CS-526 Learning Theory

5.1 We simply apply lemma from the hint to obtain

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) \geq 1/8) &= \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) \geq 1 - 7/8)\\ &\geq \frac{\mathbb{E}[L_{\mathcal{D}}(A(S))] - (1 - 7/8)}{7/8}\\ &\geq \frac{1/8}{7/8} = 1/7.\end{aligned}$$

Alternatively, if you dislike Lemma B.1, you can also prove by contrapositive, i.e., showing that if $\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) \geq 1/8) < 1/7$ then $\mathbb{E}[L_{\mathcal{D}}(A(S)] < 1/4$. This is easily seen because

$$L_{\mathcal{D}}(A(S)) < 1 \cdot \mathbb{1}_{L_{\mathcal{D}}(A(S))\geq 1/8} + \frac{1}{8} \cdot \mathbb{1}_{L_{\mathcal{D}}(A(S))<1/8}$$

and under the hypothesis

$$\mathbb{E}[L_{\mathcal{D}}(A(S))] < 1 \cdot \frac{1}{7} + \frac{1}{8} \cdot \frac{6}{7} = 1/4.$$

6.2 (a) Consider a set of k+1 elements. All-one labeling cannot be obtained, so VCdim$(\mathcal{H}) \leq k$. Analogously, for a set of $|\mathcal{X}| - k + 1$ elements all-zero labeling cannot be obtained, so VCdim$(\mathcal{H}_{=k}) \leq \min(k, |\mathcal{X}| - k)$.

Take a set C of size $m = \min(k, |\mathcal{X}| - k)$ and a labeling $(y_1, \ldots, y_m)$ with s ones, $0 \leq s \leq m$. We can pick a hypothesis $h \in \mathcal{H}_{=k}$ such that $h(x_i) = y_i$ for all $x_i \in C$ and it has $k - s$ ones at the set $\mathcal{X} \setminus C$. Therefore, C is shattered and VCdim$(\mathcal{H}_{=k}) \geq \min(k, |\mathcal{X}| - k)$.

(b) Consider set of $2k + 2$ elements. It is clear that any labeling with $k + 1$ ones and $k + 1$ zeros cannot be obtained, so VCdim$(\mathcal{H}_{at-most-k}) \leq 2k + 1$. Note that it may happen that $2k + 1 > |\mathcal{X}|$, so the bound should be VCdim$(\mathcal{H}_{at-most-k}) \leq \min(2k + 1, |\mathcal{X}|)$.

Take a set of $\min(2k+1, |\mathcal{X}|)$ elements. Any labeling on this set has either $\leq k$ zeros or $\leq k$ ones, so it is shattered by $\mathcal{H}_{at-most-k}$. Therefore, VCdim$(\mathcal{H}_{at-most-k}) = \min(2k + 1, |\mathcal{X}|)$

6.5 We simply generalize the proof from the two-dimensional case. Let's first formally state the hypothesis class

$$\mathcal{H} = \{h_{(a_i,b_i)} | a_i \leq b_i, h_{(a_i,b_i)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} \mathbb{1}_{a_i\leq x_i\leq b_i}\}$$

Consider set $\{\mathbf{x}_1, \dots, \mathbf{x}_{2d}\}$, where $\mathbf{x}_i = \mathbf{e}_i$ for $1 \le i \le d$ and $\mathbf{x}_i = -\mathbf{e}_{i-d}$ for $d+1 \le i \le 2d$. For any labeling $(y_1, \dots, y_{2d})$, pick $a_i = -2$ if $y_{d+i} = 1$ and $a_i = -0.5$ otherwise. Similarly, pick $b_i = 2$ if $y_i = 1$ and $b_i = 0.5$ otherwise. Then $h_{(a_i,b_i)}(\mathbf{x}_i) = y_i$ and hence $\text{VCdim}(\mathcal{H}) \ge 2d$.

For a set C of size $2d+1$, by the pigeonhole principle there exists an element $\mathbf{x}$ s.t. $\forall j \in [d]$ there exist $\mathbf{x}', \mathbf{x}'' \in C : x'_j \le x_j \le x''_j$. This means that labeling with only $\mathbf{x}$ negative and all other elements positive cannot be obtained and therefore $\text{VCdim}(\mathcal{H}) \le 2d$.

6.8 Let's prove the lemma first.

$$\begin{aligned}\sin(2^m \pi x) &= \sin(2^m \pi \cdot (0.x_1 x_2 \dots)) = \sin(2\pi \cdot (x_1 x_2 \dots x_{m-1}.x_m x_{m+1} \dots)) \\ &= \sin(2\pi \cdot (0.x_m x_{m+1} \dots))\end{aligned}$$

For $x_m = 0$, we know that $\exists k \ge m$ s.t. $x_k = 1$, i.e. the number $0.0x_{m+1}\dots$ is nonzero. This means that $2\pi \cdot (0.0x_{m+1}\dots) \in (0, \pi)$, where $\sin(x)$ is positive, which gives the label 1. For $x_m = 1$, we get $2\pi \cdot (0.1x_{m+1}\dots) \in (\pi, 2\pi)$, where $\sin(x)$ is negative, which gives the label 0. Proof completed.

To prove that $\mathcal{H}$ has infinite VC-dimension, we need to show that for any n there is a set x of n points in $\mathbb{R}$ on which we can obtain all 2^n possible labelings. Consider $x_1, \dots, x_n \in [0, 1]$ so that first 2^n bits of their binary expansions give all possible labelings.

Example for $n = 3$:

$$\begin{array}{lccccccccc} x_1 & 0. & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & \dots \\ x_2 & 0. & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & \dots \\ x_3 & 0. & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \dots \end{array}$$

Using the lemma, invoking the function $\lceil sin(2^i \pi x) \rceil$ on the set $\{x_1, \dots, x_n\}$ for $1 \le i \le 2^n$ allows to obtain all possible labelings. Hence, $\mathcal{H}$ shatters the set $\{x_1, \dots, x_n\}$

6.9 $\text{VCdim}(\mathcal{H}) = 3$. In order to prove it, let's recall the unsigned intervals class $\mathcal{H}_+$, which was studied during the class. It can be seen that if labeling $(y_0, y_1, \dots)$ is obtained by $h_{a,b} \in \mathcal{H}_+$, then $h_{a,b,+} \in \mathcal{H}$ gives the same labeling and $h_{a,b,-} \in \mathcal{H}$ gives its inverse $(1-y_0, 1-y_1, \dots)$. Labeling $(0, 1, 0)$ can be obtained by an interval, so signed intervals can label $(1, 0, 1)$ and therefore $\text{VCdim}(\mathcal{H}) \ge 3$.

Consider the set of 4 points. Labels $(0, 1, 0, 1)$ and $(1, 0, 1, 0)$ cannot be obtained with any signed interval, so $\text{VCdim}(\mathcal{H}) \le 3$, which concludes the proof.

7.3 (a) For any $h \in \mathcal{H}$ and given $n(h), |\mathcal{H}_{n(h)}|$, we can set $w(h) = \frac{2^{-n(h)}}{|\mathcal{H}_{n(h)}|}$. This gives

$$\sum_{h \in \mathcal{H}} w(h) = \sum_{h \in \mathcal{H}} \frac{2^{-n(h)}}{|\mathcal{H}_{n(h)}|} = \sum_{n \in \mathbb{N}} \frac{2^{-n}}{|\mathcal{H}_n|} \sum_{\substack{h \in \mathcal{H}_n \\ h \notin \mathcal{H}_{n'}, n' < n}} \mathbf{1} \le \sum_{n \in \mathbb{N}} \frac{2^{-n}}{|\mathcal{H}_n|} \sum_{h \in \mathcal{H}_n} \mathbf{1} = \sum_{n \in \mathbb{N}} 2^{-n} = 1.$$

The equality is achieved when all $\mathcal{H}_n$ are disjoint

(b) Since $\mathcal{H}_n$ is countable, we can enumerate all $h \in \mathcal{H}_n$ as $h_{n,1}, h_{n,2}, \ldots$. Consider $w(h_{n,k}) = 2^{-n}2^{-k}$. Similarly to the previous exercise, we get

$$\sum_{h \in \mathcal{H}} w(h) \leq \sum_{n \in \mathbb{N}} 2^{-n} \sum_{k \in \mathbb{N}} 2^{-k} = 1.$$

It should be noted that for some $\mathcal{H}_n$ hypotheses $h_{n,k}$ may not exist for sufficiently big k (e.g. $\mathcal{H}_n$ is finite), but we are only interested in upper bound, so it does not change anything.

Solution 4: 19 March 2019
CS-526 Learning Theory

Exercise 1

1. $f(x) = \max_{1\le i\le m} f_i(\mathbf{x})$ where $f_i(\mathbf{x}) = \mathbf{a}_i^T\mathbf{x} + b_i$ is convex differentiable with gradient $\nabla f_i(\mathbf{x}) = \mathbf{a}_i$. By Claim 14.6, it follows that $\forall \mathbf{x} : \mathbf{a}_j \in \partial f(\mathbf{x})$ where $j \in \arg\max_i f_i(\mathbf{x})$.

2. $f(x) = \max_{1\le i\le m} f_i(\mathbf{x})$ where $f_i(\mathbf{x}) = |\mathbf{a}_i^T\mathbf{x} + b_i|$ is convex subdifferentiable. Fix $\mathbf{x}$, let $j \in \arg\max_i f_i(\mathbf{x})$ and choose $\mathbf{v} \in \partial f_j(\mathbf{x})$ as follows:

$$\mathbf{v} = \begin{cases} -\mathbf{a}_j & \text{if } \mathbf{a}_j^T\mathbf{x} + b_j < 0\,, \\ 0 & \text{if } \mathbf{a}_i^T\mathbf{x} + b_i = 0\,, \\ +\mathbf{a}_j & \text{if } \mathbf{a}_j^T\mathbf{x} + b_j > 0\,. \end{cases}$$

A straightforward generalization of Claim 14.6 shows that $\mathbf{v}$ is a subgradient of f at $\mathbf{x}$.

3. Note that the sup is really a maximum as $t \mapsto p(t,\mathbf{x})$ is a continuous function on a compact. Hence $f(\mathbf{x}) = \max_{t\in[0,1]} p(t,\mathbf{x})$ and $\forall t \in [0,1] : \nabla_{\mathbf{x}} p(t,\mathbf{x}) = [1, t, \dots, t^{n-1}]^T \in \mathbb{R}^n$. A straightforward generalization of Claim 14.6 shows that $[1, t(\mathbf{x}), \dots, t(\mathbf{x})^{n-1}]^T \in \partial f(\mathbf{x})$, where $t(\mathbf{x}) \in \arg\max_{t\in[0,1]} p(t,\mathbf{x})$.

Exercise 2

1. v is a subgradient of f at 0 if $\forall u > 0 : f(u) \ge f(0) + (u-0)v$, i.e.,

$$\forall u > 0 : 0 \ge 1 + uv\,. \tag{1}$$

Clearly v must be negative for the later to hold, and if v is negative then $0 \ge 1 + uv \Leftrightarrow u \ge 1/|v|$. Whatever v, (1) cannot hold on the whole interval $[0, +\infty)$. Hence f is not subdifferentiable at 0.
2. v is a subgradient of f at 0 if $\forall u > 0 : f(u) \ge f(0) + (u-0)v$, i.e.,

$$\forall u > 0 : -1 \ge \sqrt{u}v\,. \tag{2}$$

Clearly v must be negative for the later to hold, and if v is negative then $-1 \ge \sqrt{u}v \Leftrightarrow u \ge 1/v^2$. Whatever v, (2) cannot hold on the whole interval $[0, +\infty)$. Hence f is not subdifferentiable at 0.

Exercise 3

Fix $\mathbf{w}$, $\mathbf{u}$. The function f is λ-strongly convex, so for all $\alpha \in [0,1]$ we have:

$$\begin{aligned} & f((1-\alpha)\mathbf{w}+\alpha\mathbf{u}) \leq (1-\alpha)f(\mathbf{w})+\alpha f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 \\ \Leftrightarrow \quad & f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w})) - f(\mathbf{w}) \leq \alpha\Big(f(\mathbf{u}) - f(\mathbf{w}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\Big) \end{aligned} \tag{3}$$

Let $\mathbf{v} \in \partial f(\mathbf{w})$. Then, $\forall \alpha \in [0,1] : f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w})) \geq f(\mathbf{w}) + \langle \alpha(\mathbf{u}-\mathbf{w}), \mathbf{v}\rangle$. Combining this inequality and (3) gives:

$$\begin{aligned} & \langle \alpha(\mathbf{u}-\mathbf{w}), \mathbf{v}\rangle \leq \alpha\Big(f(\mathbf{u}) - f(\mathbf{w}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\Big) \\ \Leftrightarrow \quad & \langle \mathbf{u}-\mathbf{w}, \mathbf{v}\rangle \leq f(\mathbf{u}) - f(\mathbf{w}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 \\ \Leftrightarrow \quad & \langle \mathbf{w}-\mathbf{u}, \mathbf{v}\rangle \geq f(\mathbf{w}) - f(\mathbf{u}) + \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 \end{aligned}$$

Taking the limit $\alpha \to 0+$ ends the proof: $\langle \mathbf{w}-\mathbf{u}, \mathbf{v}\rangle \geq f(\mathbf{w}) - f(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2$.

Exercise 4

To prove that $\pi_C(\cdot)$ is Lipschiztian, we first show an important property of projection onto a closed convex set:

Lemma 1. *If C is a non-empty closed convex subset of a Hilbert space H then $\forall(\mathbf{x},\mathbf{y}) \in H \times C : \langle \mathbf{x} - \pi_C(\mathbf{x}), \mathbf{y} - \pi_C(\mathbf{x})\rangle \leq 0$.*

Proof. Let $\alpha \in (0,1)$. By definition of $\pi_C(\cdot)$, we have:

$$\begin{aligned} 0 &\leq \|\mathbf{x} - (1-\alpha)\pi_C(\mathbf{x}) - \alpha\mathbf{y}\|^2 - \|\mathbf{x} - \pi_C(\mathbf{x})\|^2 \\ &= \|\mathbf{x} - \pi_C(\mathbf{x}) - \alpha(\mathbf{y} - \pi_C(\mathbf{x}))\|^2 - \|\mathbf{x} - \pi_C(\mathbf{x})\|^2 \\ &= \alpha^2\|\mathbf{y} - \pi_C(\mathbf{x})\|^2 - 2\alpha\langle \mathbf{x} - \pi_C(\mathbf{x}), \mathbf{y} - \pi_C(\mathbf{x})\rangle \,. \end{aligned}$$

Dividing the final inequality by α and taking the limit $\alpha \to 0$ ends the proof. □

We can now prove that $\pi_C(\cdot)$ is 1-Lipschitz. $\forall \mathbf{x}_0, \mathbf{x}_1$:

$$\begin{aligned} \|\pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\|^2 &= \langle \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1), \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle \\ &= \underbrace{\langle \pi_C(\mathbf{x}_0) - \mathbf{x}_0, \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle}_{\leq 0} + \langle \mathbf{x}_0 - \pi_C(\mathbf{x}_1), \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle \\ &\leq \langle \mathbf{x}_0 - \pi_C(\mathbf{x}_1), \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle \\ &\leq \underbrace{\langle \mathbf{x}_1 - \pi_C(\mathbf{x}_1), \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle}_{\leq 0} + \langle \mathbf{x}_0 - \mathbf{x}_1, \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle \\ &\leq \langle \mathbf{x}_0 - \mathbf{x}_1, \pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\rangle \\ &\leq \|\mathbf{x}_0 - \mathbf{x}_1\|\|\pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\| \qquad \text{(Cauchy-Schwarz inequality)} \end{aligned}$$

It directly implies $\|\pi_C(\mathbf{x}_0) - \pi_C(\mathbf{x}_1)\| \leq \|\mathbf{x}_0 - \mathbf{x}_1\|$. Note that for $\mathbf{x}_0, \mathbf{x}_1 \in C$ this inequality is an equality, hence the it cannot be improved.

Solution 5 (2^{nd} graded homework): 26 March 2019
CS-526 Learning Theory

Exercise 1

a) Fix $A, B \in \mathcal{S}_n^+$ and $\alpha \in [0,1]$. Let $\mathbf{e} \in \mathbb{R}^n$ a unit-norm eigenvector of $\alpha A + (1-\alpha)B$ associated to the maximum eigenvalue, i.e., $(\alpha A + (1-\alpha)B)\mathbf{e} = \lambda_{\max}(\alpha A + (1-\alpha)B)\mathbf{e}$ and $\|\mathbf{e}\| = 1$. We have:

$$\begin{aligned} f(\alpha A + (1-\alpha)B) = \mathbf{e}^T(\alpha A + (1-\alpha)B)\mathbf{e} &= \alpha \mathbf{e}^T A \mathbf{e} + (1-\alpha)\mathbf{e}^T B \mathbf{e} \\ &\le \alpha \lambda_{\max}(A) + (1-\alpha)\lambda_{\max}(B) \\ &= \alpha f(A) + (1-\alpha) f(B)\,. \end{aligned}$$

This shows that f is convex.
b) Let $A \in \mathcal{S}_n^+$. A subgradient of f at A is a matrix $V \in \mathbb{R}^{n\times n}$ that satisfies:

$$\forall B \in \mathcal{S}_n^+ : f(B) \ge f(A) + \mathrm{Tr}\big((B-A)^T V\big)\,.$$

Consider any $\mathbf{e} \in \mathbb{R}^n$ which is a unit-norm eigenvector of A associated to the maximum eigenvalue, i.e., $A\mathbf{e} = \lambda_{\max}(A)\mathbf{e}$ and $\|\mathbf{e}\| = 1$. Then for all $B \in \mathcal{S}_n^+$:

$$\begin{aligned} f(A) = \lambda_{\max}(A) = \mathbf{e}^T A \mathbf{e} = \mathbf{e}^T B \mathbf{e} + \mathbf{e}^T (A-B)\mathbf{e} &\le \lambda_{\max}(B) + \mathbf{e}^T(A-B)\mathbf{e} \\ &= f(B) + \mathrm{Tr}(\mathbf{e}^T(A-B)\mathbf{e}) \\ &= f(B) + \mathrm{Tr}((A-B)^T \mathbf{e}\mathbf{e}^T)\,. \end{aligned}$$

In the last equality we used that $(A-B)^T = A - B$ and that the trace is preserved by cyclic permutations. We see that $\mathbf{e}\mathbf{e}^T$ satisfies the definition of a subgradient: $\mathbf{e}\mathbf{e}^T \in \partial f(A)$.

Exercise 2

a) $\min_{\|\mathbf{w}\|\le\|\mathbf{w}^*\|} f(\mathbf{w}) \le f(\mathbf{w}^*) \le 0$ because $\forall i \in [m] : y_i\langle \mathbf{w}^*, \mathbf{x}_i\rangle \ge 1$. Suppose there exists $\mathbf{w}$ satisfying both $\|\mathbf{w}\| \le \|\mathbf{w}^*\|$ and $f(\mathbf{w}) < 0$. Then $\mathbf{w}$ can be slightly modify to obtain a vector $\tilde{\mathbf{w}}$ such that $\|\tilde{\mathbf{w}}\| < \|\mathbf{w}^*\|$, while still having $f(\tilde{\mathbf{w}}) \le 0$. It contradicts $\mathbf{w}^*$'s definition, hence $\min_{\|\mathbf{w}\|\le\|\mathbf{w}^*\|} f(\mathbf{w}) \ge 0$. It proves $\min_{\|\mathbf{w}\|\le\|\mathbf{w}^*\|} f(\mathbf{w}) = 0$.

b) If $f(\mathbf{w}) < 1$ then $\forall i \in [m] : y_i\langle \mathbf{w}^*, \mathbf{x}_i\rangle > 0$, i.e., $\mathbf{w}$ separates the examples.

c) For all $i \in [m]$ the gradient of $f_i : \mathbf{w} \mapsto 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle$ is $-y_i\mathbf{x}_i$. Applying Claim 14.6, we get that a subgradient of f at $\mathbf{w}$ is given by $-y_{i^*}\mathbf{x}_{i^*}$ where $i^* \in \arg\max_{i\in[m]}\{1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\}$.

d) The algorithm is inialized with $\mathbf{w}^{(1)} = 0$. At each iteration, if $f(\mathbf{w}^{(t)}) \ge 1$ then it chooses $i^* \in \arg\min_{i\in[m]}\{y_i\langle \mathbf{w}^{(t)}, \mathbf{x}_i\rangle\}$ and updates $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_{i^*}\mathbf{x}_{i^*}$. Otherwise, if

$f(\mathbf{w}^{(t)}) < 1$, $\mathbf{w}^{(t)}$ separates all the examples and we stop. To analyze the speed of convergence of the subgradient algorithm, first notice that $\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \rangle - \langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle = \eta y_{i^*} \langle \mathbf{w}^*, \mathbf{x}_{i^*} \rangle \geq \eta$. Therefore, after performing T iterations, we have

$$\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle = \langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle - \langle \mathbf{w}^*, \mathbf{w}^{(1)} \rangle = \sum_{t=1}^{T} \langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \rangle - \langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle \geq \eta T \,. \tag{1}$$

Besides, $\|\mathbf{w}^{(t+1)}\|^2 = \|\mathbf{w}^{(t)}\|^2 + \eta^2 y_{i^*}^2 \|\mathbf{x}_i\|^2 + 2\eta y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_{i^*} \rangle \leq \|\mathbf{w}^{(t)}\|^2 + \eta^2 R^2$. The last inequality follows from $\|\mathbf{x}_i\| \leq R$ and $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_{i^*} \rangle \leq 0$ (we update only if $f(\mathbf{w}^{(t)}) \geq 1$). Then

$$\|\mathbf{w}^{(T+1)}\| \leq \eta R \sqrt{T} \,. \tag{2}$$

Combining Cauchy-Schwarz inequality, (1) and (2), we obtain

$$1 \geq \frac{\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle}{\|\mathbf{w}^{(T+1)}\| \|\mathbf{w}^*\|} \geq \frac{\sqrt{T}}{R \|\mathbf{w}^*\|} \,. \tag{3}$$

The subgradient algorithm must stop in less than $R^2 \|\mathbf{w}^*\|^2$ iterations. We see that η does not affect the speed of convergence. The algorithm is almost identical to the Batch Perceptron algorithm with two modifications. First, the Batch Perceptron updates with any example for which $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$, while the current algorithm chooses the example for which $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle$ is minimal. Second, the current algorithm employs the parameter η. However, the only difference with the case $\eta = 1$ is that it scales $\mathbf{w}^{(t)}$ by η.

Exercise 3

We prove the following Theorem:

Theorem 1. *Let $B, \rho > 0$. Let f be a convex function and let $\mathbf{w}^\star \in \arg\min_{\mathbf{w}:\|\mathbf{w}\| \leq B} f(\mathbf{w})$. Assume that SGD is run for T iterations with $\eta_t = \frac{B}{\rho\sqrt{t}}$. Assume also that for all t, $\mathbb{E}\|\mathbf{v}_t\|^2 \leq \rho^2$. Then*

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{3\rho B}{\sqrt{T}}$$

Proof. By Jensen's inequality, we have:

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \right] . \tag{4}$$

As $\forall t : \mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$, we can reproduce what is done in Theorem 14.8 to get the inequality:

$$\mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \right] \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \right] . \tag{5}$$

We now have to prove an upper bound on the right-hand side of (5). This is similar to what is done in Lemma 14.10, except that we have to take into account the time-dependence of the steps η_t. For all $t \in \{1\dots,T\}$:

$$\begin{aligned}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle = \frac{1}{\eta_t}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \eta_t\mathbf{v}_t\rangle &= \frac{1}{2\eta_t}\big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^\star - \eta_t\mathbf{v}_t\|^2 + \eta_t^2\|\mathbf{v}_t\|^2\big)\\ &= \frac{1}{2\eta_t}\big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1/2)} - \mathbf{w}^\star\|^2 + \eta_t^2\|\mathbf{v}_t\|^2\big)\\ &\le \frac{1}{2\eta_t}\big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2\big) + \frac{\eta_t}{2}\|\mathbf{v}_t\|^2. \quad (6)\end{aligned}$$

Let $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \le B\}$. The last inequality follows from $\mathbf{w}^{(t+1)} = \pi_{\mathcal{H}}(\mathbf{w}^{(t+1/2)})$ and the 1-Lipschitzianity of $\pi_{\mathcal{H}}$ (see Homework 4, Exercise 4):

$$\|\pi_{\mathcal{H}}(\mathbf{w}^{(t+1/2)}) - \mathbf{w}^\star\| = \|\pi_{\mathcal{H}}(\mathbf{w}^{(t+1/2)}) - \pi_{\mathcal{H}}(\mathbf{w}^\star)\| \le \|\mathbf{w}^{(t+1/2)} - \mathbf{w}^\star\|\,.$$

Summing the inequality (6) over t, we have:

$$\begin{aligned}\sum_{t=1}^T\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle &\le \sum_{t=1}^T \frac{1}{2\eta_t}\big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2\big) + \frac{\eta_t}{2}\|\mathbf{v}_t\|^2\\ &= \frac{1}{2\eta_1}\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \sum_{t=1}^{T-1}\frac{\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2}{2}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\\ &\qquad - \frac{1}{2\eta_T}\|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2 + \sum_{t=1}^T\frac{\eta_t}{2}\|\mathbf{v}_t\|^2\\ &\le \frac{1}{2\eta_1}\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \sum_{t=1}^{T-1}\frac{\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2}{2}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + \sum_{t=1}^T\frac{\eta_t}{2}\|\mathbf{v}_t\|^2\\ &\le 2B^2\left(\frac{1}{\eta_1} + \sum_{t=1}^{T-1}\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + \sum_{t=1}^T\frac{\eta_t}{2}\|\mathbf{v}_t\|^2\\ &= \frac{2B^2}{\eta_T} + \sum_{t=1}^T\frac{\eta_t}{2}\|\mathbf{v}_t\|^2\,. \qquad (7)\end{aligned}$$

Taking the expectation of inequality (7) and diving by T, we obtain:

$$\mathbb{E}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^T\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle\right] \le \frac{2B^2}{T\eta_T} + \sum_{t=1}^T\frac{\eta_t}{2T}\mathbb{E}\|\mathbf{v}_t\|^2 \le \frac{2\rho B}{\sqrt{T}} + \frac{\rho^2}{2T}\sum_{t=1}^T\eta_t\,. \qquad (8)$$

The last inequality follows from the assumption $\mathbb{E}\|\mathbf{v}_t\|^2 \le \rho^2$ and η_T's definition. Besides

$$\sum_{t=1}^T\eta_t = \frac{B}{\rho}\sum_{t=1}^T\frac{1}{\sqrt{t}} \le \frac{B}{\rho}\left(1 + \sum_{t=2}^T\int_{t-1}^t\frac{dx}{\sqrt{x}}\right) = \frac{B}{\rho}\left(1 + \int_1^T\frac{dx}{\sqrt{x}}\right) = \frac{B}{\rho}\big(2\sqrt{T} - 1\big)\,.$$

Combining this last inequality with (4), (5) and (8), we finally obtain:

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \le \frac{2\rho B}{\sqrt{T}} + \frac{\rho B}{2T}\big(2\sqrt{T} - 1\big) \le \frac{3\rho B}{\sqrt{T}}\,.$$

It concludes the proof. $\square$

Exercise 4

$\mathcal{H}_{n-parity}$ is a finite class, therefore (see paragraph 6.3.4):

$$\text{VCdim}(\mathcal{H}_{n-parity}) \leq \log_2 |\mathcal{H}_{n-parity}| = \log_2 2^n = n\,.$$

We now show that this upperbound on VCdim($\mathcal{H}_{n-parity}$) is tight, i.e., there exists n points in $\{0,1\}^n$ that are shattered by $\mathcal{H}_{n-parity}$. Let $\mathbf{e}^{(j)} \in \{0,1\}^n$ be such that $\mathbf{e}^{(j)}_j = 1$ and $\forall i \neq j : \mathbf{e}^{(j)}_i = 0$. The subset $C = \{\mathbf{e}^{(j)}\}_{j=1}^n$ of n points is shattered by $\mathcal{H}_{n-parity}$. Indeed, given $(y_1, \ldots, y_n) \in \{0,1\}^n$, we can define $J = \{j \in \{1,\ldots,n\} : y_j = 1\}$ and see that:

$$\forall j \in \{1,\ldots,n\} : h_J(\mathbf{e}^{(j)}) = \sum_{i \in J} \mathbf{e}^{(j)}_i \mod 2 = \sum_{i=1}^{n} \mathbf{e}^{(j)}_i y_i \mod 2 = y_j\,.$$

Hence VCdim($\mathcal{H}_{n-parity}$) $= n$.

Exercise 6: 9 April 2019
CS-526 Learning Theory

Problem 1

1) The joint distribution is (up to normalisation factors of Gaussians)

$$p(y_1, \ldots, y_m, x_1, \ldots, x_m, w_1, \ldots, w_p) \propto \prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{i=1}^{m} P_0(x_i) \prod_{a=1}^{p} e^{-\alpha w_a^2}$$

2) Here the x_i is a parent of y_i (for all $i = 1, \ldots, m$) and $w_1, \ldots, w_p$ are parents of each $y_i, i = 1, \ldots, m$.

3) The ML principle says that you maximize the log-likelihood $\log P(data \mid w_1, \ldots, w_p)$. Since

$$P(data \mid w_1, \ldots, w_p) \propto \prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{i=1}^{m} P_0(x_i)$$

this is equivalent to minimising

$$\mathcal{E}_{data}(f) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2$$

over functions in the class $\mathcal{H} \ni f(x) = \sum_{a=1}^{p} w_a x^a$.

4) The posterior distribution is

$$\begin{aligned} P(w_1, \ldots, w_p \mid data) &= \frac{p(y_1, \ldots, y_m, x_1, \ldots, x_m, w_1, \ldots, w_p)}{\int \prod_{a=1}^{p} dw_a p(y_1, \ldots, y_m, x_1, \ldots, x_m, w_1, \ldots, w_p)} \\ &= \frac{\prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{a=1}^{p} e^{-\alpha w_a^2}}{\int \prod_{a=1}^{p} dw_a \prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{a=1}^{p} e^{-\alpha w_a^2}} \end{aligned}$$

The MAP principle says you maximise the posterior which is equivalent to minimizing

$$\frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + 2\alpha\sigma^2 \sum_{a=1}^{p} w_a^2$$

over the functions in the class $\mathcal{H} \ni f(x) = \sum_{a=1}^{p} w_a x^a$.

5) The optimal regression function is $f_{regr}(x) = \mathbb{E}_{w|data}\mathbb{E}_{y|x,w}[y]$. From the model it is clear that

$$\mathbb{E}_{y|x,w}[y] = \sum_{a=1}^{p} w_a x^a$$

Further average over the posterior gives

$$f_{regr}(x) = \sum_{a=1}^{p} \mathbb{E}_{w|data}[w_a]x^a$$

Problem 2

1) $a \perp\!\!\!\perp b|c$ because $p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$. But a, b are not independent because $p(a,b) = \sum_c p(a|c)p(b|c) \neq p(a)p(b)$.

2) $a \perp\!\!\!\perp b|c$ because $p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(b|c)p(c|a)p(a)}{p(c)} = p(b|c)\frac{p(c|a)p(a)}{p(c)} = p(b|c)p(a|c)$. But a, b are not independent because $p(a,b) = \sum_c p(a)p(c|a)p(b|c) = p(a)p(b|a) \neq p(a)p(b)$.

3) $a \perp\!\!\!\perp b$ because

$$p(a,b) = \sum_{c,d} p(a,b,c,d) = \sum_{c,d} p(a)p(b)p(c|a,b)p(d|c) = p(a)p(b)\sum_{c,d} p(c|a,b)p(d|c) = p(a)p(b).$$

However, we don't have $a \perp\!\!\!\perp b|c$ because $p(a,b|c) = \frac{p(a)p(b)p(c|a,b))}{p(c)}$ cannot be decomposed.

Problem 3

The left hand side is

$$p(x_i|\mathbf{x}_{\sim i}) = \frac{p(\mathbf{x})}{\int dx_i\ p(\mathbf{x})} \tag{1}$$

where

$$p(\mathbf{x}) = p(x_i|\{x_v\}_{v\in \text{pa}(i)}) \prod_{k\in \text{child}(j)} p(x_j|\{x_v\}_{v\in \text{pa}(k)}) \prod_{\substack{l\neq i \\ l\neq \text{child}(i)}} p(x_l|\{x_v\}_{v\in \text{pa}(l)}).$$

The product $\prod_{\substack{l\neq i \\ l\neq \text{child}(i)}} p(x_l|\{x_v\}_{v\in \text{pa}(l)})$ is independent of x_i. It cancels with the same factor in the denominator of (1). So we have

$$p(x_i|\mathbf{x}_{\sim i}) = \frac{p(x_i|\{x_v\}_{v\in \text{pa}(i)}) \prod_{k\in \text{child}(j)} p(x_j|\{x_v\}_{v\in \text{pa}(k)})}{\int dx_i\ p(x_i|\{x_v\}_{v\in \text{pa}(i)}) \prod_{k\in \text{child}(j)} p(x_j|\{x_v\}_{v\in \text{pa}(k)})} \tag{2}$$

On the other hand, the right hand side is

$$p(x_i|\{x_v\}_{v\in \text{MB}(i)}) = \frac{p(x_i, \{x_v\}_{v\in \text{MB}(i)})}{\int dx_i\ p(x_i, \{x_v\}_{v\in \text{MB}(i)})} \tag{3}$$

where

$$
\begin{aligned}
&p(x_i, \{x_v\}_{v\in\mathrm{MB}(i)}) \\
&= \int d\mathbf{x}_{\sim i,\mathrm{MB}(i)}\; p(\mathbf{x}) \\
&= \int d\mathbf{x}_{\sim i,\mathrm{MB}(i)}\; p(x_i|\{x_v\}_{v\in\mathrm{pa}(i)}) \prod_{k\in\mathrm{child}(j)} p(x_j|\{x_v\}_{v\in\mathrm{pa}(k)}) \prod_{\substack{l\neq i\\ l\neq\mathrm{child}(i)}} p(x_l|\{x_v\}_{v\in\mathrm{pa}(l)}) \\
&=\; p(x_i|\{x_v\}_{v\in\mathrm{pa}(i)}) \prod_{k\in\mathrm{child}(j)} p(x_j|\{x_v\}_{v\in\mathrm{pa}(k)}) \Big[\int d\mathbf{x}_{\sim i,\mathrm{MB}(i)} \prod_{\substack{l\neq i\\ l\neq\mathrm{child}(i)}} p(x_l|\{x_v\}_{v\in\mathrm{pa}(l)}) \Big]
\end{aligned}
$$

We identify $\int d\mathbf{x}_{\sim i,\mathrm{MB}(i)} \prod_{\substack{l\neq i\\ l\neq\mathrm{child}(i)}} p(x_l|\{x_v\}_{v\in\mathrm{pa}(l)})$ independent of x_i. It cancels with the same factor in the denominator of (3). So (3) is reduced to the same expression as (2).

Problem 4 (Bishop, p.371 & 419, Exercise 8.7)

Using $\mathbb{E}[x_i] = \sum_{j\in\mathrm{pa}(i)} w_{ij}\mathbb{E}[x_j] + b_i$ gives

$$
\begin{aligned}
\mu_1 &= \sum_{j\in\emptyset} w_{1j}\mathbb{E}[x_j] + b_1 = b_1 \\
\mu_2 &= \sum_{j\in\{1\}} w_{2j}\mathbb{E}[x_j] = w_{21}b_1 + b_2 \\
\mu_3 &= \sum_{j\in\{2\}} w_{3j}\mathbb{E}[x_j] + b_3 = w_{32}(w_{21}b_1 + b_2) + b_3
\end{aligned}
$$

Using $\mathrm{cov}[x_i, x_j] = \sum_{k\in\mathrm{pa}(j)} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij}v_j$ for $i \leq j$ and $\mathrm{cov}[x_i, x_j] = \mathrm{cov}[x_j, x_i]$ gives

$$
\begin{aligned}
\mathrm{cov}[x_1, x_1] &= \sum_{k\in\emptyset} w_{1j}\mathrm{cov}[x_1, x_k] + v_1 = v_1 \\
\mathrm{cov}[x_1, x_2] &= \sum_{k\in\{1\}} w_{2j}\mathrm{cov}[x_1, x_k] = w_{21}v_1 \\
\mathrm{cov}[x_1, x_3] &= \sum_{k\in\{2\}} w_{3j}\mathrm{cov}[x_1, x_k] = w_{32}(w_{21}v_1) \\
\mathrm{cov}[x_2, x_2] &= \sum_{k\in\{1\}} w_{2j}\mathrm{cov}[x_2, x_k] + v_2 = w_{21}(w_{21}v_1) + v_2 \\
\mathrm{cov}[x_2, x_3] &= \sum_{k\in\{2\}} w_{3j}\mathrm{cov}[x_2, x_k] = w_{32}(w_{21}^2 v_1 + v_2) \\
\mathrm{cov}[x_3, x_3] &= \sum_{k\in\{2\}} w_{3j}\mathrm{cov}[x_3, x_k] + v_3 = w_{32}^2(w_{21}^2 v_1 + v_2) + v_3
\end{aligned}
$$

Problem 5 (Barber, p.75, Exercise 4.4)

1) First note that

$$p(\mathbf{h}|\mathbf{v}) \propto e^{(\mathbf{v}^\top \mathbf{W}+\mathbf{b}^\top)\mathbf{h}} = \prod_i e^{h_i(b_i+\sum_j W_{ji}v_j)}$$

So $p(\mathbf{h}|\mathbf{v}) = \prod_i p(h_i|\mathbf{v})$. Recall $h_i \in \{0,1\}$. Thus we have

$$p(h_i = 1|\mathbf{v}) = \frac{e^{b_i+\sum_j W_{ji}v_j}}{\sum_{h_i\in\{0,1\}} e^{h_i(b_i+\sum_j W_{ji}v_j)}} = \sigma\Big(b_i + \sum_j W_{ji}v_j\Big).$$

2)

$$p(\mathbf{v}|\mathbf{h}) = \prod_i p(v_i|\mathbf{h}), \qquad \text{with } p(v_i = 1|\mathbf{h}) = \sigma\Big(a_i + \sum_j W_{ij}h_j\Big)$$

3) No. Because the term $\mathbf{v}^\top\mathbf{W}\mathbf{h}$ in $p(\mathbf{v},\mathbf{h})$ introduces dependence between $\mathbf{v}$ and $\mathbf{h}$.

4) For a general $\mathbf{W}$ there is no known efficient way to compute Z efficiently. The dependence between $\mathbf{v}$ and $\mathbf{h}$ does not allow always decomposition of $p(\mathbf{v},\mathbf{h})$.

Problem 6 (Barber, p.77, Exercise 4.14)

We write

$$\begin{aligned}\phi_{ij}(x_i,x_j) &= e^{\ln\phi_{ij}(x_i,x_j)}\\ &= e^{\mathbb{I}(x_i=0,x_j=0)\ln\phi_{ij}(0,0)+\mathbb{I}(x_i=0,x_j=1)\ln\phi_{ij}(0,1)+\mathbb{I}(x_i=1,x_j=0)\ln\phi_{ij}(1,0)+\mathbb{I}(x_i=1,x_j=1)\ln\phi_{ij}(1,1)}\end{aligned}$$

With $x_i \in \{0,1\}$ we can replace $\mathbb{I}[\cdot]$ by

$$\begin{aligned}&\mathbb{I}(x_i=0,x_j=0) = (1-x_i)(1-x_j), && \mathbb{I}(x_i=0,x_j=1) = (1-x_i)x_j,\\ &\mathbb{I}(x_i=1,x_j=0) = x_i(1-x_j), && \mathbb{I}(x_i=1,x_j=1) = x_ix_j.\end{aligned}$$

So $\phi_{ij}(x_i,x_j)$ is in the form $e^{W_{ij}x_ix_j+b_ix_i+b_jx_j+\text{constant}}$ and $p(\mathbf{x}) = \frac{1}{Z'}e^{\sum_{ij\in\mathcal{E}}W_{ij}x_ix_j+\sum_i \deg(i)b_ix_i}$ is the Boltzmann machine.

Problem 7

Fix a subset $S \subseteq V$. We have:

$$p(\mathbf{x}_S,\mathbf{x}_{V\setminus S}) = p(\mathbf{x}) = \frac{1}{Z}\prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}}\psi_{C'}(\mathbf{x}_{C'}) \cdot \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}}\psi_C(\mathbf{x}_C)\,;$$

$$p(\mathbf{x}_{V\setminus S}) = \sum_{\mathbf{x}_S} p(\mathbf{x}) = \frac{1}{Z}\prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}}\psi_{C'}(\mathbf{x}_{C'}) \cdot \left(\sum_{\mathbf{x}_S}\prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}}\psi_C(\mathbf{x}_C)\right).$$

Therefore, the conditional distribution of $\mathbf{x}_S$ given $\mathbf{x}_{V\setminus S}$ reads:

$$p(\mathbf{x}_S|\mathbf{x}_{V\setminus S}) = \frac{p(\mathbf{x}_S,\mathbf{x}_{V\setminus S})}{p(\mathbf{x}_{V\setminus S})} = \frac{\prod_{C:S\cap C\neq\emptyset}\psi_C(\mathbf{x}_C)}{\sum_{\widetilde{\mathbf{x}}_S}\prod_{C:S\cap C\neq\emptyset}\psi_C(\widetilde{\mathbf{x}}_C)}\,. \tag{4}$$

To write the denominator in the last equality, we implitcitly introduced $\widetilde{\mathbf{x}} = (\widetilde{\mathbf{x}}_S, \mathbf{x}_{V\setminus S})$, while $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{V\setminus S})$.
Consider any maximal clique C such that $S \cap C \neq \emptyset$ and let $i \in S \cap C$. If $j \in C \setminus S$ then $j \in \partial S$ because $\{i, j\} \in E$ ($i \in C$ and C is a clique). Therefore $C \subseteq S \cup \partial S$. It follows:

$$p(\mathbf{x}_S, \mathbf{x}_{\partial S}) = \sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} p(\mathbf{x}) = \frac{1}{Z} \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}} \psi_C(\mathbf{x}_C) \cdot \left(\sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} \prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \right);$$

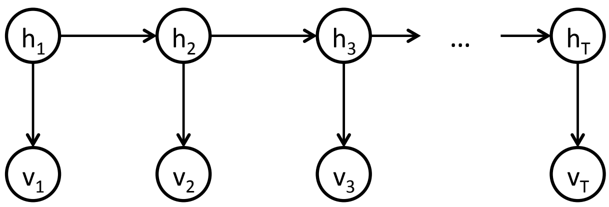
$$\begin{aligned} p(\mathbf{x}_{\partial S}) &= \sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} p(\mathbf{x}_{V\setminus S}) = \sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} \frac{1}{Z} \prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \cdot \left(\sum_{\mathbf{x}_S} \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}} \psi_C(\mathbf{x}_C) \right) \\ &= \frac{1}{Z} \left(\sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} \prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \right) \left(\sum_{\mathbf{x}_S} \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}} \psi_C(\mathbf{x}_C) \right). \end{aligned}$$

It comes

$$p(\mathbf{x}_S|\mathbf{x}_{\partial S}) = \frac{p(\mathbf{x}_S, \mathbf{x}_{\partial S})}{p(\mathbf{x}_{\partial S})} = \frac{\prod_{C:S\cap C\neq\emptyset} \psi_C(\mathbf{x}_C)}{\sum_{\widetilde{\mathbf{x}}_S} \prod_{C:S\cap C\neq\emptyset} \psi_C(\widetilde{\mathbf{x}}_C)}. \tag{5}$$

The final equalities in (4) and (5) are the same, thus proving that $p(\mathbf{x}_S|\mathbf{x}_{V\setminus S})$ and $p(\mathbf{x}_S|\mathbf{x}_{\partial S})$ are equal.

Problem 8 (Barber, p.99, Exercise 5.4)



1)

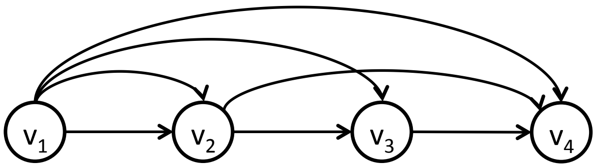
2) A simple linear chain for $p(\mathbf{h})$ can be easily seen from

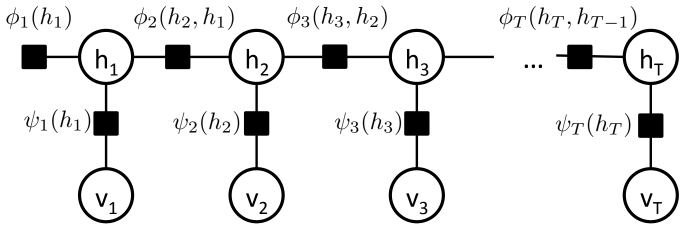
$$p(\mathbf{h}) = \sum_{\mathbf{v}} p(\mathbf{v}, \mathbf{h}) = p(h_1) \prod_{t=2}^{T} p(h_t|h_{t-1})$$

On the other hand, $p(\mathbf{v})$ is a fully connected cascade belief network because the marginal probability does not admit any decomposition. For example $T = 4$,

$$\begin{aligned} p(v_1, v_2, v_3, v_4) &= \sum_{h_1,h_2,h_3,h_4} p(v_1, v_2, v_3, v_4, h_1, h_2, h_3, h_4) \\ &= \sum_{h_4} p(v_4|h_4) \sum_{h_3} \Big(p(v_3, h_4|h_3) \sum_{h_2} \left(p(v_2, h_3|h_2) p(v_1, h_2) \right) \Big) \end{aligned}$$

We see that v_1, v_2, v_3, h_4 are all coupled.





3) The factors are $\psi_t(h_t) = p(v_t|h_t)$, $\phi_1(h_1) = p(h_1)$ and $\phi_t(h_t, h_{t-1}) = p(h_t|h_{t-1})$ for $t \geq 2$.

4) Suppose our observation is $\hat{\mathbf{v}} = (\hat{v}_1, \dots, \hat{v}_T)$. Since

$$p(\mathbf{h}|\mathbf{v} = \hat{\mathbf{v}}) \propto p(\mathbf{h}, \mathbf{v} = \hat{\mathbf{v}}),$$

we can use a sum-product algorithm to compute the marginal $p(h_t, \hat{\mathbf{v}})$ and then it is easy to obtain $p(h_t|\hat{\mathbf{v}}) = \frac{p(h_t, \hat{\mathbf{v}})}{\sum_{h_t} p(h_t, \hat{\mathbf{v}})}$. Recall that

$$\begin{aligned} p(\hat{\mathbf{v}}, h_t) = \sum_{\mathbf{h}_{\sim t}} p(\hat{\mathbf{v}}, \mathbf{h}) &= \sum_{\mathbf{h}_{\sim t}} p(h_1) p(\hat{v}_1|h_1) \prod_{i=2}^{T} p(\hat{v}_i|h_i) p(h_i|h_{i-1}) \\ &= \sum_{\mathbf{h}_{\sim t}} \phi_1(h_1) \psi_1(h_1) \prod_{i=2}^{T} \psi_i(h_i) \phi_i(h_i, h_{i-1}) \end{aligned}$$

To compute the sum efficiently we define messages propagating from the two ends of the factor graph. For the forward propagation we define the factor-to-variable message

$$\mu_{\psi_i \to h_i}(h_i) = \psi(h_i), \quad \mu_{\phi_i \to h_i}(h_i) = \sum_{h_{i-1}} \phi_i(h_i, h_{i-1}) \mu_{h_{i-1} \to \phi_i}(h_{i-1}) \text{ with } \phi_1(h_1, h_0) \triangleq \phi_1(h_1)$$

and variable-to-factor message

$$\mu_{h_i \to \phi_{i+1}}(h_i) = \mu_{\psi_i \to h_i}(h_i) \mu_{\phi_i \to h_i}(h_i)$$

We compute the messages in the order $(\mu_{\psi_1 \to h_1}, \mu_{\phi_1 \to h_1}) \to \mu_{h_1 \to \phi_2} \to (\mu_{\psi_2 \to h_2}, \mu_{\phi_2 \to h_2}) \to \mu_{h_2 \to \phi_3} \to \cdots \to (\mu_{\psi_t \to h_t}, \mu_{\phi_t \to h_t})$. So we have

$$\mu_{\phi_t \to h_t} = \sum_{h_1, \dots, h_{t-1}} \psi_1(h_1) \psi_1(h_1) \prod_{i=2}^{t} \psi_i(h_i) \phi_i(h_i, h_{i-1})$$

It does not harm to continue the forward propagation up to $(\mu_{\psi_T \to h_T}, \mu_{\phi_T \to h_T})$ but here it is unneccessary. Next, we start the backward propagation with factor-to-variable message

$$\mu_{\phi_i \to h_{i-1}}(h_{i-1}) = \sum_{h_i} \phi_i(h_i, h_{i-1}) \mu_{h_i \to \phi_i}(h_i)$$

and variable-to-factor message

$$\mu_{h_i\to\phi_i}(h_i) = \mu_{\psi_i\to h_i}(h_i)\mu_{\phi_{i+1}\to h_i}(h_i) \text{ with } \mu_{\phi_{T+1}\to h_T}(h_T) \triangleq 1$$

We proceed with $\mu_{\psi_T\to h_T} \to \mu_{h_T\to\phi_T} \to (\mu_{\psi_{T-1}\to h_{T-1}}, \mu_{\phi_T\to h_{T-1}}) \to \mu_{h_{T-1}\to\phi_{T-1}} \to \cdots \to \mu_{\phi_{t+1}\to h_t}$. So we have

$$\mu_{\phi_t\to h_t}(h_t) = \sum_{h_{t+1},\ldots,h_T} \prod_{i=t+1}^{T} \psi_i(h_i)\phi_i(h_i, h_{i-1})$$

and therefore

$$\begin{aligned} p(h_t, \hat{\mathbf{v}}) &= \mu_{\phi_t\to h_t}(h_t)\mu_{\psi_t\to h_t}(h_t)\mu_{\phi_{t+1}\to h_t}(h_t), \\ p(h_t|\hat{\mathbf{v}}) &= \frac{\mu_{\phi_t\to h_t}(h_t)\mu_{\psi_t\to h_t}(h_t)\mu_{\phi_{t+1}\to h_t}(h_t)}{\sum_{h_t} \mu_{\phi_t\to h_t}(h_t)\mu_{\psi_t\to h_t}(h_t)\mu_{\phi_{t+1}\to h_t}(h_t)}. \end{aligned}$$

5) Like the starting argument in the last question, we need to compute $\sum_{\mathbf{h}_{\sim t,t+1}} p(h_t, h_{t+1}, \hat{\mathbf{v}})$ where $\mathbf{h}_{\sim t,t+1}$ means h_t and h_{t+1} are excluded. So with the same message passing rules we obtain

$$p(h_t, h_{t+1}|\hat{\mathbf{v}}) \propto \mu_{\phi_t\to h_t}(h_t)\mu_{\psi_t\to h_t}(h_t)\phi_{t+1}(h_t, h_{t+1})\mu_{\phi_{t+2}\to h_{t+1}}(h_{t+1})\mu_{\psi_{t+1}\to h_{t+1}}(h_{t+1})$$

Problem 9 (Barber, p.98, Exercise 5.1)

The underlying undirected graph of a singly connected network with N nodes is a tree. We denote the tree with N nodes by $\mathcal{T}_N$. By definition it contains a leaf i which is connected to node j. The tree structure ensures the decomposition

$$Z = \sum_{\mathbf{x}_{\sim i}} \prod_{\substack{k\sim l \\ k\neq i \\ l\neq i}} \phi_{k,l}(x_k, x_l) \sum_{x_i} \phi_{i,j}(x_i, x_j).$$

where $\mathbf{x}_{\sim i}$ means x_i is excluded. So we can start the following recursion with $\mathcal{T}_N$.

1. Find a leaf i which is connected to node j.
2. Compute $\psi_{i,j}(x_j) = \sum_{x_i} \phi_{i,j}(x_i, x_j)$.
3. If node j has another neighbor node k,

3a. obtain $\mathcal{T}_{n-1}$ by removing node i and updating $\phi_{j,k}(x_j, x_k) \to \psi_{i,j}(x_j)\phi_{j,k}(x_j, x_k)$, and go to step 1 with $\mathcal{T}_{n-1}$;

3b. otherwise, there remain only node i and j, so we output $Z = \sum_{x_j} \psi_{i,j}(x_j)$.

The above algorithm ends with N iterations and therefore the time complexity is $O(N)$.

Problem 10 (Bishop, p.397 & 421, Exercise 8.16 & 8.17)

1) Given the observation $x_N = \hat{x}_N$, the initial message for β-recursion becomes

$$\mu_\beta(x_{N-1}) = \phi_{N-1,N}(x_{N-1}, \hat{x}_N).$$

Note that this initial message does not sum over x_N. The other message passing equations are unchanged. This message passing allows us to compute $p(x_n|x_N = \hat{x}_N)$.

2) Given the observation $x_3 = \hat{x}_3$, the algorithm suggests

$$p(x_2) = \frac{1}{Z}\mu_\alpha(x_2)\mu_\beta(x_2)$$

where

$$\begin{aligned}
&\mu_\beta(x_2) = \phi_{2,3}(x_2, \hat{x}_3)\mu_\beta(\hat{x}_3),\\
&Z = \sum_{x_2} \mu_\alpha(x_2)\mu_\beta(x_2) = \sum_{x_2} \mu_\alpha(x_2)\phi_{2,3}(x_2, \hat{x}_3)\mu_\beta(\hat{x}_3).
\end{aligned}$$

We can simplify the expression to

$$p(x_2) = \frac{\mu_\alpha(x_2)\phi_{2,3}(x_2, \hat{x}_3)}{\sum_{x_2} \mu_\alpha(x_2)\phi_{2,3}(x_2, \hat{x}_3)}.$$

Different x_5 will rescale $\mu_\beta(\hat{x}_3)$ but it changes nothing on $p(x_2)$. This aligns with the fact that $x_2 \perp\!\!\!\perp x_5 | x_3$.

Solution 8 (4^{th} graded homework): 21 Mai 2019
CS-526 Learning Theory

Problem 1

1) For every $i \in [K]$, $\underline{d}_i$ is the i^{th} canonical basis vector of $\mathbb{R}^K$ and we define the latent random vector $\underline{h} \in \{\underline{d}_i : i \in [K]\}$ whose distribution is $\forall i \in [K] : \mathbb{P}(\underline{h} = \underline{d}_i) = w_i$. Finally, let $\underline{x} = \sum_{i=1}^K h_i \underline{a}_i + \underline{z}$ where $\underline{z} \sim \mathcal{N}(0, \sigma^2 I_{D\times D})$ is independent of $\underline{h}$. The random vector $\underline{x}$ has a probability density function $p(\cdot)$. We have:

$$\mathbb{E}[\underline{x}] = \sum_{i=1}^K \mathbb{E}[h_i]\underline{a}_i + \mathbb{E}[\underline{z}] = \sum_{i=1}^K w_i\, \underline{a}_i \quad ;$$

$$\begin{aligned}\mathbb{E}[\underline{x}\underline{x}^T] &= \mathbb{E}[\underline{z}\underline{z}^T] + \sum_{i=1}^K \mathbb{E}[h_i] \underbrace{\mathbb{E}[\underline{z}]}_{=0} \underline{a}_i^T + \mathbb{E}[h_i]\underline{a}_i \mathbb{E}[\underline{z}]^T + \sum_{i,j=1}^K \underbrace{\mathbb{E}[h_i h_j]}_{=w_i\delta_{ij}} \underline{a}_i \underline{a}_j^T \\ &= \sigma^2 I_{D\times D} + \sum_{i=1}^K w_i\, \underline{a}_i \underline{a}_i^T \ .\end{aligned}$$

Finally, to compute the third moment tensor, note that $\mathbb{E}[\underline{z} \otimes \underline{z} \otimes \underline{z}] = 0$ and that for every $(i,j) \in [K]^2$: $\mathbb{E}[\underline{a}_i \otimes \underline{a}_j \otimes \underline{z}] = \mathbb{E}[\underline{a}_i \otimes \underline{z} \otimes \underline{a}_j] = \mathbb{E}[\underline{z} \otimes \underline{a}_i \otimes \underline{a}_j] = 0$. Hence:

$$\begin{aligned}\mathbb{E}[\underline{x} \otimes \underline{x} \otimes \underline{x}] = &\sum_{i,j,k=1}^K \underbrace{\mathbb{E}[h_i h_j h_k]}_{=w_i\delta_{ij}\delta_{ik}} \underline{a}_i \otimes \underline{a}_j \otimes \underline{a}_k \\ &+ \sum_{i=1}^K \mathbb{E}[h_i]\mathbb{E}[\underline{a}_i \otimes \underline{z} \otimes \underline{z}] + \mathbb{E}[h_i]\mathbb{E}[\underline{z} \otimes \underline{a}_i \otimes \underline{z}] + \mathbb{E}[h_i]\mathbb{E}[\underline{z} \otimes \underline{z} \otimes \underline{a}_i] \\ = &\sum_{i=1}^K w_i\, \underline{a}_i \otimes \underline{a}_i \otimes \underline{a}_i + \sigma^2 \sum_{j=1}^D \sum_{i=1}^K w_i (\underline{a}_i \otimes \underline{e}_j \otimes \underline{e}_j + \underline{e}_j \otimes \underline{e}_j \otimes \underline{a}_i + \underline{e}_j \otimes \underline{a}_i \otimes \underline{e}_j) \,.\end{aligned}$$

2) Let $A = [\underline{a}_1, \underline{a}_2, \dots, \underline{a}_K] \in \mathbb{R}^{D\times K}$ and $A' = [\underline{a}'_1, \underline{a}'_2, \dots, \underline{a}'_K] \in \mathbb{R}^{D\times K}$. By definition, $\widetilde{R} = \Sigma^{-1} R \Sigma$ where Σ is the diagonal matrix such that $\Sigma_{ii} = \sqrt{w_i}$ and $A' = A\widetilde{R}^T$. We can directly apply the formula of question 1) to compute the second moment matrix of the new mixture of Gaussians:

$$\begin{aligned}\mathbb{E}[\underline{x}\underline{x}^T] &= \sigma^2 I_{D\times D} + A'\Sigma^2 A'^T = \sigma^2 I_{D\times D} + A\widetilde{R}^T \Sigma^2 \widetilde{R} A^T \\ &= \sigma^2 I_{D\times D} + A\Sigma R^T R \Sigma A^T = \sigma^2 I_{D\times D} + A\Sigma^2 A^T \ .\end{aligned}$$

Problem 2: Examples of tensors and their rank

1) The matrices corresponding to B, P, E are:

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \ ; \ P = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \ ; \ E = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

The frontal slices of G and W are:

$$G_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} , G_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} ; W_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} , W_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} .$$

The matricizations of G and W are:

$$G_{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} ; G_{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} ; G_{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} ;$$
$$W_{(1)} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} ; W_{(2)} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} ; W_{(3)} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} .$$

2) B and E are clearly rank-2 matrices, while $P = (e_0 + e_1) \otimes (e_0 + e_1)$ is a rank-1 matrix.
By its definition, G is at most rank 2. Assume it is rank 1: $G = a \otimes b \otimes c$ with $a, b, c \in \mathbb{R}^2$. We have $a_1 b_1 c_1 = G_{111} = 1$ and $a_2 b_1 c_1 = G_{211} = 0$ so we must have $a_2 = 0$. Besides, $a_2 b_2 c_2 = G_{222} = 1$ and $a_1 b_2 c_2 = G_{122} = 0$ so $a_1 = 0$. Hence $a^T = (0, 0)$ and G is the all-zero tensor. This is a contradiction and we conclude that G is rank 2.
By its definition, W is at most rank 3. To prove the rank cannot be smaller than 3, we will proceed by contradiction:

- Assume W is rank 1: $W = a \otimes b \otimes c$ with $a, b, c \in \mathbb{R}^2$. We have $a_1 b_1 c_1 = W_{111} = 0$ and $a_2 b_1 c_1 = W_{211} = 1$ so $a_1 = 0$. Besides, $a_1 b_1 c_2 = W_{112} = 1$ and $a_2 b_1 c_2 = W_{212} = 0$ so $a_2 = 0$. Then $a = (0, 0)^T$ and W is the all-zero tensor, which is a contradiction.

- Assume W is rank 2: $W = a \otimes b \otimes c + d \otimes e \otimes f$. We claim that *$a$ and d must be linearly independent*. Indeed, suppose they are parallel and take a vector x perpendicular to both a and d. Then
$$W(x, I, I) = (x^T a) b \otimes c + (x^T d) e \otimes f = 0$$
but also
$$W(x, I, I) = (x^T e_0) e_0 \otimes e_1 + (x^T e_0) e_1 \otimes e_0 + (x^T e_1) e_0 \otimes e_0 = \begin{bmatrix} x^T e_1 & x^T e_0 \\ x^T e_0 & 0 \end{bmatrix}$$
which cannot be zero since x cannot be perpendicular to both e_0 and e_1. Now, we take x perpendicular to d. We have
$$W(x, I, I) = (x^T a) b \otimes c$$
which is rank one. Therefore, we must have $x^T e_0 = 0$ which implies that x is parallel to e_1 and thus $\underline{d \text{ parallel to } e_0}$. Now, if we take x perpendicular to a, the matrix
$$W(x, I, I) = (x^T d) e \otimes f$$
is rank one and, once again, we must have $x^T e_0 = 0$, which implies x parallel to e_1 and thus $\underline{a \text{ parallel to } e_0}$. Hence, we have shown that a and d are linearly independent but also that both are parallel to e_0. This is a contradiction.

3) Writing everything in terms of matrix product, it comes:

$$(Oe_0) \otimes (Oe_0) + (Oe_1) \otimes (Oe_1) = Oe_0 e_0^T O^T + Oe_1 e_1^T O^T = OO^T = B .$$

so B does not have a unique decomposition.

For G we have $G = \underline{a}_1 \otimes \underline{b}_1 \otimes \underline{c}_1 + \underline{a}_2 \otimes \underline{b}_2 \otimes \underline{c}_2$ with

$$A = [\underline{a}_1, \underline{a}_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ; B = [\underline{b}_1, \underline{b}_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ; C = [\underline{c}_1, \underline{c}_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} .$$

A, B, C are full column rank and G has rank 2: by Jennrich's algorithm, the decomposition is unique (up to trivial rank permutation and feature scaling).

For W we have $W = \underline{a}_1 \otimes \underline{b}_1 \otimes \underline{c}_1 + \underline{a}_2 \otimes \underline{b}_2 \otimes \underline{c}_2 + \underline{a}_3 \otimes \underline{b}_3 \otimes \underline{c}_3$ with

$$A = [\underline{a}_1, \underline{a}_2, \underline{a}_3] = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \; ; \; B = [\underline{b}_1, \underline{b}_2, \underline{b}_3] = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \; ; \; C = [\underline{c}_1, \underline{c}_2, \underline{c}_3] = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} .$$

A, B, C are not full column rank: Jennrich's theorem does not allow to conclude that the decomposition of W is unique.

4) We expand the tensor products in the definition of D_ϵ:

$$\begin{aligned} D_\epsilon &= \frac{1}{\epsilon}\Big[(e_0 + \epsilon e_1) \otimes (e_0 + \epsilon e_1) \otimes (e_0 + \epsilon e_1) - e_0 \otimes e_0 \otimes e_0\Big] \\ &= \frac{1}{\epsilon}\Big[e_0 \otimes e_0 \otimes e_0 + \epsilon\, e_0 \otimes e_0 \otimes e_1 + \epsilon\, e_0 \otimes e_1 \otimes e_0 + \epsilon\, e_1 \otimes e_0 \otimes e_0 \\ &\qquad + \epsilon^2\, e_1 \otimes e_1 \otimes e_0 + \epsilon^2\, e_1 \otimes e_0 \otimes e_1 + \epsilon^2\, e_0 \otimes e_1 \otimes e_1 + \epsilon^3\, e_1 \otimes e_1 \otimes e_1 - e_0 \otimes e_0 \otimes e_0\Big] \\ &= e_0 \otimes e_0 \otimes e_1 + e_0 \otimes e_1 \otimes e_0 + e_1 \otimes e_0 \otimes e_0 \\ &\qquad + \epsilon(e_1 \otimes e_1 \otimes e_0 + e_1 \otimes e_0 \otimes e_1 + e_0 \otimes e_1 \otimes e_1) + \epsilon^2\, e_1 \otimes e_1 \otimes e_1 \\ &= W + \epsilon(e_1 \otimes e_1 \otimes e_0 + e_1 \otimes e_0 \otimes e_1 + e_0 \otimes e_1 \otimes e_1) + \epsilon^2\, e_1 \otimes e_1 \otimes e_1 \; . \end{aligned}$$

Hence $\lim_{\epsilon \to 0} D_\epsilon = 0$.

Problem 3

1) There cannot be an analogous general result for tensors. Indeed, the order-3 tensor W of Problem 2 is rank 3 and we show in 4) that $\lim_{\epsilon \to 0} \|W - D_\epsilon\|_F = 0$. So there is no minimum attained in the space of rank 2 tensors. In this sense, there is simply no *best* rank-two approximation of W.

2) Let M a matrix of rank $R + 1$ with singular values $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_R \geq \sigma_{R+1} > 0$. By the Eckart-Young-Mirsky theorem, the minimum of $\|M - \widehat{M}\|_F$ over rank R matrices $\widehat{M}$ is equal to $\sigma_{R+1} > 0$. Therefore, there cannot be a sequence of matrices M_n given by a sum of R rank-one matrices such that $\lim_{n \to +\infty} \|M - M_n\|_F = 0$.

4) In the real-valued case, we have:

$$|T(R_1, R_2, R_3)^{\alpha\beta\gamma}|^2 = \sum_{\delta,\epsilon,\zeta,\delta',\epsilon',\delta'} R_1^{\delta\alpha} R_1^{\delta'\alpha} R_2^{\epsilon\beta} R_2^{\epsilon'\beta} R_3^{\zeta\gamma} R_3^{\zeta'\gamma} T^{\delta\epsilon\zeta} T^{\delta'\epsilon'\zeta'} \; .$$

Summing over α, β, γ and using the orthogonality of rotation matrices, we find:

$$\sum_\alpha R_1^{\delta\alpha} R_1^{\delta'\alpha} = \delta_{\delta\delta'}, \quad \sum_\beta R_2^{\epsilon\beta} R_2^{\epsilon'\beta} = \delta_{\beta\beta'}, \quad \sum_\gamma R_3^{\zeta\gamma} R_3^{\zeta'\gamma} = \delta_{\zeta\zeta'} \; .$$

The result directly follows:

$$\|T(R_1, R_2, R_3)\|_F^2 = \sum_{\delta\epsilon\zeta} |T(R_1, R_2, R_3)^{\alpha\beta\gamma}|^2 = \sum_{\delta\epsilon\zeta} |T^{\delta\epsilon\zeta}|^2 = \|T\|_F^2 \; .$$

Problem 4

1) To show that $A \odot_{KhR} B$ is full column rank, we have to prove that the kernel of the linear application $\underline{x} \mapsto (A \odot_{KhR} B)\underline{x}$ is $\{0\}$. Let $\underline{x} \in \mathbb{R}^R$ with components $(x^1, x^2, \cdots, x^R)$ be such that $(A \odot_{KhR} B)\underline{x} = 0$. Then, $\forall \alpha \in [I_1]$:

$$\sum_{r=1}^{R} a_r^\alpha x^r \underline{b}_r = 0 \,.$$

Because B is full column rank, $\sum_{r=1}^{R} a_r^\alpha x^r \underline{b}_r = 0$ implies that $\forall r \in [R] : a_r^\alpha x^r = 0$. Note that:

$$\forall \alpha \in [I_1], \forall r \in [R] : a_k^\alpha x^r = 0 \Leftrightarrow A\underline{x} = 0 \,.$$

A is full column rank and $A\underline{x} = 0$, hence $\underline{x} = 0$. $A \odot_{KhR} B$ is full column rank.

2) Suppose we are given a tensor (the weights λ_r that usually appear in the sum are absorbed in the vectors $\underline{a}_r$)

$$\mathcal{X} = \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r \,, \tag{1}$$

where $A = [\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_R] \in \mathbb{R}^{I_1 \times R}$, $B = [\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_R] \in \mathbb{R}^{I_2 \times R}$ and $C = [\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_R] \in \mathbb{R}^{I_3 \times R}$ are full column rank. By Jennrich's algorithm, the decomposition (1) is unique up to trivial rank permutation and feature scaling and Jennrich's algorithm is a way to recover this decomposition. At the end of the step (5) of the algorithm, we have computed A, B and it remains to recover C. We now show how the result in question **1)** allows to recover C uniquely. For each $\gamma \in [I_3]$, define the slice $\mathcal{X}_\gamma$ as the $I_1 \times I_2$ matrix with entries $(\mathcal{X}_\gamma)^{\alpha\beta} = \mathcal{X}^{\alpha\beta\gamma}$ and denote $F(\mathcal{X}_\gamma)$ the $I_1 I_2$ column vector with entries $F(\mathcal{X}_\gamma)^{\beta + I_2(\alpha-1)} = \mathcal{X}^{\alpha\beta\gamma}$. We have:

$$\forall (\alpha, \beta) \in [I_1] \times [I_2] : F(\mathcal{X}_\gamma)^{\beta + I_2(\alpha-1)} = \sum_{r=1}^{R} a_r^\alpha b_r^\beta c_r^\gamma = \sum_{r=1}^{R} (A \odot_{KhR} B)^{\beta + I_2(\alpha-1), r} c_r^\gamma \,.$$

Therefore, the $I_1 I_2 \times I_3$ matrix $F(\mathcal{X}) = [F(\mathcal{X}_1), F(\mathcal{X}_2), \ldots, F(\mathcal{X}_{I_3})]$ satisfies:

$$F(\mathcal{X}) = (A \odot_{KhR} B) C^T \,.$$

Because $A \odot_{KhR} B$ is full column rank, we can invert the system with the Moore-Penrose pseudoinverse: $C^T = (A \odot_{KhR} B)^\dagger F(\mathcal{X})$.

Problem 5

1) To apply Jennrich's algorithm we need to prove that the matrix $E = [\underline{c}_1 \otimes_{Kro} \underline{d}_1, \ldots, \underline{c}_R \otimes_{Kro} \underline{d}_R]$ is full column rank (A, B are full column rank by assumption). Note that teh same proof as the one in Problem 4 question 1 applies. Nevertheless we repeat the argument here.
Let $\underline{v} \in \mathbb{R}^R$ a column vector in the kernel of E, i.e., $E\underline{v} = 0$. Then:

$$\forall \gamma \in [I_3] : \sum_{r=1}^{R} (c_r^\gamma v^r) \underline{d}_r = 0 \;\Rightarrow\; \forall \gamma \in [I_3], \forall r \in [R] : c_r^\gamma v^r = 0 \;\Rightarrow\; C\underline{v} = 0 \;\Rightarrow\; \underline{v} = 0 \,.$$

The first implication follows from D being full column rank and the third one from C being full column rank. We conclude that the kernel of E is $\{0\}$: E is full column rank.
We can therefore apply Jennrich's algorithm.

2) We recover the rank R as well as A, B and E by applying Jennsen's algorithm to $\widetilde{T}$. From E we can then determine C and D. Fix $r \in [R]$. Since C is full column rank, there exists $\alpha \in [I_3]$ such that $c_r^\alpha \neq 0$. As $c_r^\alpha \neq 0$, we can use the I_4-dimensional column vector $c_r^\alpha \underline{d}_r$ contained in the r^{th} column of E to recover $\underline{d}_r$. Doing this for every $r \in [R]$ we recover the matrix D. Finally, for every $r \in R$, pick $\beta \in I_4$ such that $d_r^\beta \neq 0$ (such β exists because D is full column rank) and use the entries $c_r^\alpha d_r^\beta$, $\alpha \in [I_3]$, to recover $\underline{c}_r$. C has then been recovered.

Problem 6

1) Define $\Sigma^\dagger$ as the $N \times M$ diagonal matrix with diagonal entries:

$$\forall i \in \{1, 2, \ldots, \min\{M, N\}\} : (\Sigma^\dagger)_{ii} \begin{cases} 1/\Sigma_{ii} & \text{if} \quad \Sigma_{ii} \neq 0 \;; \\ 0 & \text{otherwise.} \end{cases}$$

Then both $\Sigma^\dagger\Sigma \in \mathbb{C}^{N\times N}$ and $\Sigma\Sigma^\dagger \in \mathbb{C}^{M\times M}$ are diagonal square matrices with diagonal entries:

$$\forall i \in [N] : (\Sigma^\dagger\Sigma)_{ii} = \begin{cases} 1 & \text{if} \quad i \leq \min\{M, N\} \text{ and } \Sigma_{ii} \neq 0 \;; \\ 0 & \text{otherwise.} \end{cases}$$

$$\forall i \in [M] : (\Sigma\Sigma^\dagger)_{ii} = \begin{cases} 1 & \text{if} \quad i \leq \min\{M, N\} \text{ and } \Sigma_{ii} \neq 0 \;; \\ 0 & \text{otherwise.} \end{cases}$$

It is then easy to check that $\Sigma^\dagger$ satisfies the first two conditions of the Moore-Penrose pseudoinverse: $\Sigma\Sigma^\dagger\Sigma = \Sigma$ and $\Sigma^\dagger\Sigma\Sigma^\dagger = \Sigma^\dagger$. Besides, $\Sigma^\dagger\Sigma$ and $\Sigma\Sigma^\dagger$ being real diagonal matrices, the last two conditions are clearly satisfied too.

2) We can check that the matrix $V\Sigma^\dagger U^*$ satisfies the four conditions of the Moore-Penrose pseudoinverse, i.e., $A^\dagger = V\Sigma^\dagger U^*$:

$$\begin{aligned} A[V\Sigma^\dagger U^*]A &= U\Sigma(V^*V)\Sigma^\dagger(U^*U)\Sigma V^* = U\Sigma\Sigma^\dagger\Sigma V^* = U\Sigma V^* = A \;; \\ [V\Sigma^\dagger U^*]A[V\Sigma^\dagger U^*] &= V\Sigma^\dagger(U^*U)\Sigma(V^*V)\Sigma^\dagger U^* = V\Sigma^\dagger\Sigma\Sigma^\dagger U^* = V\Sigma^\dagger U^* \;; \\ (AV\Sigma^\dagger U^*)^* &= (U\Sigma\Sigma^\dagger U^*)* = U(\Sigma\Sigma^\dagger)^*U^* = U\Sigma\Sigma^\dagger U^* = AV\Sigma^\dagger U^* \;; \\ (V\Sigma^\dagger U^*A)^* &= (V\Sigma^\dagger\Sigma V^*)* = V(\Sigma^\dagger\Sigma)^*V^* = V\Sigma^\dagger\Sigma V^* = V\Sigma^\dagger U^*A \,. \end{aligned}$$

3) A is full column rank, therefore A^*A is a full rank $N \times N$ matrix and has a unique inverse $(A^*A)^{-1}$. The matrix $(A^*A)^{-1}A^*$ satisfies the four conditions:

$$A[(A^*A)^{-1}A^*]A = A \;;\; [(A^*A)^{-1}A^*]A[(A^*A)^{-1}A^*] = (A^*A)^{-1}A^* \;;$$
$$(A[(A^*A)^{-1}A^*])^* = A[(A^*A)^{-1}A^*] \;;\; ([(A^*A)^{-1}A^*]A)^* = A^*A(A^*A)^{-1} = I_{N\times N} = ([(A^*A)^{-1}A^*]A \,.$$

Hence $A^\dagger = (A^*A)^{-1}A^*$.

4) A is full row rank, therefore AA^* is a full rank $M \times M$ matrix and has a unique inverse $(AA^*)^{-1}$. The matrix $A^*(AA^*)^{-1}$ satisfies the four conditions:

$$A[A^*(AA^*)^{-1}]A = A \;;\; [A^*(AA^*)^{-1}]A[A^*(AA^*)^{-1}] = A^*(AA^*)^{-1} \;;$$
$$(A[A^*(AA^*)^{-1}])^* = (AA^*)^{-1}AA^* = I_{M\times M} = AA^\dagger \;;\; ([A^*(AA^*)^{-1}]A)^* = A^*(AA^*)^{-1}A \,.$$

Hence $A^\dagger = A^*(AA^*)^{-1}$.

5) We have $AA^{-1}A = A$, $A^{-1}AA^{-1} = A^{-1}$, $(AA^{-1})^* = I_{M\times M} = AA^{-1}$, $(A^{-1}A)^* = I_{N\times N} = A^{-1}A$. Hence $A^\dagger = A^{-1}$.

6) A is full column rank so $A^\dagger A = I_{M\times M}$ and B is full column rank so $BB^\dagger = I_{N\times N}$. Therefore:

$$\begin{aligned}
(AB)(B^\dagger A^\dagger)(AB) &= A(BB^\dagger)(A^\dagger A)B = AI_{M\times M}I_{N\times N}B = AB\ ;\\
(B^\dagger A^\dagger)(AB)(B^\dagger A^\dagger) &= B^\dagger(A^\dagger A)(BB^\dagger)A^\dagger = B^\dagger I_{N\times N}I_{M\times M}A^\dagger = B^\dagger A^\dagger\ ;\\
(ABB^\dagger A^\dagger)^* &= (AI_{N\times N}A^\dagger)^* = (AA^\dagger)^* = AA^\dagger = (AB)(B^\dagger A^\dagger)\ ;\\
(B^\dagger A^\dagger AB)^* &= (B^\dagger I_{M\times M}B)^* = (B^\dagger B)^* = B^\dagger B = (B^\dagger A^\dagger)(AB)\ .
\end{aligned}$$

Hence $(AB)^\dagger = B^\dagger A^\dagger$.