10-702 Statistical Machine Learning: Assignment 3 Solution

1. (a)

$$\log p(x) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \sum_{j<k}^{d} \beta_{jk} X_j X_k + \cdots + \sum_{j<k<\ell}^{d} \beta_{jk\ell} X_j X_k X_\ell + \cdots + \tag{1}$$

Since $\beta_A = 0$ whenever $\{1, 2\} \subset A$, all the terms in the above linear right hand side can be partitioned into 3 parts :

- containing $X_3, ...X_d$ and no terms containing X_1 and X_2
- containing $X_1, X_3, ...X_d$ and no terms containing X_2
- containing $X_2, X_3, ...X_d$ and no terms containing X_1

That is, there will not be any terms containing both X_1 and X_2, since their corresponding β_A will be zero.

$$\log p(x) = f_1(X_3, ...X_d) + f_2(X_1, X_3, ...X_d) + f_3(X_3, X_3, ...X_d)$$
$$p(x) = e^{f_1(X_3,...X_d)}\ e^{f_2(X_1,X_3,...X_d)}\ e^{f_3(X_2,X_3,...X_d)}$$

Using appropriate probability normalization, we can express this as

$$p(x) = P(X_3, ...X_d)P(X_1|X_3, ...X_d)P(X_2|X_3, ...X_d) \tag{2}$$

However, from factorization, we know that

$$p(x) = P(X_3, ...X_d)P(X_1, X_2|X_3, ...X_d) \tag{3}$$

From the above two equations, we see that

$$P(X_1|X_3, ...X_d)P(X_2|X_3, ...X_d) = P(X_1, X_2|X_3, ...X_d) \tag{4}$$

which implies that

$$X_1 \amalg X_2 \,\Big|\, X_3, \ldots, X_d.$$

Thus proved.

(b) For any i,

$$\begin{aligned}
\max_{x_j, j\neq i} p\left(x_1, \ldots, x_i^*, \ldots, x_d\right) &= m_i\left(x_i^*\right) \\
&= \max_{x_i} m_i\left(x_i\right) \quad \text{(with uniqueness)} \\
&= \max_{x_i} \max_{x_j, j\neq i} p\left(x_1, \ldots, x_d\right)
\end{aligned}$$

which implies

$$x_i^* = \arg\max_{x_i} \left[\max_{x_j, j\neq i} p\left(x_1, \ldots, x_d\right)\right] \quad \text{(with uniqueness)} \tag{5}$$

We may then conclude that

$$x^* = \left(x_1^*, \ldots, x_d^*\right) = \arg\max_{x} p\left(x_1, \ldots, x_d\right) \quad \text{(with uniqueness)}$$

i.e. x^* is the unique mode of p. Proof: suppose x^* is not the unique mode of p. Then there exists $x' = \arg\max_x p\left(x_1, \ldots, x_d\right)$ such that $x' \neq x^*$. This implies

$$x_i' = \arg\max_{x_i} \left[\max_{x_j, j\neq i} p\left(x_1, \ldots, x_d\right)\right]$$

for all i, which contradicts equation (5) for any i such that $x_i' \neq x_i^*$.

(c) One set of integers is $m_i = 1 - D_i$, where D_i is the degree of vertex x_i. Proof: G is a tree, so we can number the vertices such that x_1 is the root, and x_j is a descendant of x_i for $j > i$. Let all edges $(i,j) \in E$ be such that $i < j$. Then

$$\begin{aligned}
f_m(x_1, \ldots, x_d) &= \prod_{i=1}^{d} p_i(x_i)^{1-D_i} \prod_{(i,j)\in E} p_{ij}(x_i, x_j) \\
&= \frac{\prod_{(i,j)\in E} p_{ij}(x_i, x_j)}{\prod_{i=1}^{d} p_i(x_i)^{D_i - 1}} \\
&= \frac{\prod_{(i,j)\in E} p_{j|i}(x_j \mid x_i)\, p_i(x_i)}{\prod_{i=1}^{d} p_i(x_i)^{D_i - 1}} \\
&= \frac{\left[\prod_{(i,j)\in E} p_{j|i}(x_j \mid x_i)\right] \left[p_1(x_1)^{D_i}\right] \left[\prod_{i=2}^{d} p_i(x_i)^{D_i - 1}\right]}{\prod_{i=1}^{d} p_i(x_i)^{D_i - 1}} \\
&\quad \text{(every vertex has one parent, except for } x_1\text{)} \\
&= p_1(x_1) \prod_{(i,j)\in E} p_{j|i}(x_j \mid x_i)
\end{aligned}$$

Observe that for any j, $\int p_{j|i}(x_j \mid x_i)\, dx_j = 1$ for any value of x_i. Also note that each vertex x_j for $j \geq 2$ appears exactly once in $\prod_{(i,j)\in E} p_{j|i}(x_j \mid x_i)$ (not counting the x_i being conditioned upon), while x_1 does not appear at all. Hence we can integrate out one term at a time to get $\int \cdots \int f_m(x_1, \ldots, x_d)\, dx_1 \ldots dx_d = 1$. Finally, f_m is nonnegative since $p_i(x_i)$ and $p_{ij}(x_i, x_j)$ are nonnegative.

2

(a)

(i)

The distribution of X is not in the exponential family.

Assume for a contradiction that X is in the exponential family. Then for some $\phi(x)$, the pdf of x takes the form

$$f_\theta(x) = a(x) \exp\left(\theta^\top \phi(x) - \Psi_{a,\phi}(\theta)\right)$$

where θ is a vector-valued function of p, and $a(x)$ does not depend on θ (and hence p). We know that $f_\theta(x) = 0$ for $x < 0$ or $x > p$. Since the exponential function is never zero, it must be the case that $a(x) = 0$ for $x < 0$ or $x > p$, implying that $a(x)$ depends on p. Contradiction, hence the distribution of X is not in the exponential family.

(ii)

The distribution of Y is in the exponential family.

We need to show that

$$f_{Y,\theta}(y) = a(y) \exp\left(\theta^\top \phi(y) - \Psi_{a,\phi}(\theta)\right)$$

for some $a(y), \theta, \phi(y), \Psi_{a,\phi}(\theta)$. Observe that $y(x) = \exp(x)$ is a monotone, 1-to-1 transformation. Hence

$x(y) = \log y$ and

$$\begin{aligned}
f_Y(y) &= f_X(x(y))\left|\frac{dx(y)}{dy}\right| \\
&= f_X(\log y)\frac{1}{y} \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(\log y)^2\right)\frac{1}{y} \\
&= \exp\left(-\frac{1}{2\sigma^2}(\log y)^2 - \log y - \log\sqrt{2\pi\sigma^2}\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}(\log y)^2 - \log y - \frac{1}{2}\log 2\pi\sigma^2\right)
\end{aligned}$$

Let

$$\begin{aligned}
a(y) &= 1 \\
\theta &= \begin{bmatrix} -\frac{1}{2\sigma^2} \\ -1 \end{bmatrix} \\
\phi(y) &= \begin{bmatrix} (\log y)^2 \\ \log y \end{bmatrix} \\
\Psi_{a,\phi}(\theta) &= \frac{1}{2}\log 2\pi\sigma^2 = \frac{1}{2}\log\frac{-\pi}{\theta}
\end{aligned}$$

and confirm that

$$\begin{aligned}
\log\int_0^\infty \exp\left(\theta^\top\phi(y)\right)\,dy &= \log\int_0^\infty \exp\left(-\frac{1}{2\sigma^2}(\log y)^2 - \log y\right)\,dy \\
&= \log\int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}x^2 - x\right)\exp(x)\,dx \\
&= \log\int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}x^2\right)\,dx \\
&= \log\left(\sqrt{2\pi}\sigma\right) \\
&= \Psi_{a,\phi}(\theta)
\end{aligned}$$

Hence

$$f_{Y,\theta}(y) = a(y)\exp\left(\theta^\top\phi(y) - \Psi_{a,\phi}(\theta)\right)$$

which was to be shown.

(iii)

The distribution of X is in the exponential family.
We need to show that

$$f_\theta(x) = a(x)\exp\left(\theta^\top\phi(x) - \Psi_{a,\phi}(\theta)\right)$$

for some $a(x), \theta, \phi(x), \Psi_{a,\phi}(\theta)$. We have that

$$f(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

Note that $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{1}{B(a,b)}$, where $B(a,b)$ is the beta function defined by

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

Hence

$$\begin{aligned} f(x;a,b) &= \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\ &= \exp\left((a-1)\log x + (b-1)\log(1-x) - \log B(a,b)\right) \end{aligned}$$

Let

$$\begin{aligned} a(x) &= 1 \\ \theta &= \begin{bmatrix} a-1 \\ b-1 \end{bmatrix} \\ \phi(x) &= \begin{bmatrix} \log x \\ \log(1-x) \end{bmatrix} \\ \Psi_{a,\phi}(\theta) &= \log B(a,b) = \log B(\theta_1+1, \theta_2+1) \end{aligned}$$

and confirm that

$$\begin{aligned} \log \int_0^1 \exp\left(\theta^\top \phi(x)\right) dx &= \log \int_0^1 \exp\left((a-1)\log x + (b-1)\log(1-x)\right) dx \\ &= \log \int_0^1 x^{a-1}(1-x)^{b-1}\,dx \\ &= \log B(a,b) \\ &= \Psi_{a,\phi}(\theta) \end{aligned}$$

Hence

$$f_\theta(x) = a(x)\exp\left(\theta^\top \phi(x) - \Psi_{a,\phi}(\theta)\right)$$

which was to be shown.

(b)

Rewrite the optimization problem as

$$\begin{aligned} \min_{p_1,\ldots,p_m} \sum_{j=1}^m p_j \log p_j & \\ \text{s.t.} & \\ -p_j &\le 0 \quad j \in \{1,\ldots,m\} \\ \left(\sum_{j=1}^m p_j\right) - 1 &= 0 \\ \left(\sum_{j=1}^m p_j \phi_k(j)\right) - \mu_k &= 0 \quad k \in \{1,\ldots,d\} \end{aligned}$$

The Lagrangian is

$$\mathcal{L}(p,\lambda,\alpha,\beta) = \left[\sum_{j=1}^m p_j \log p_j\right] + \left[\sum_{j=1}^m \lambda_j(-p_j)\right] + \left[\alpha\left(-1 + \sum_{j=1}^m p_j\right)\right] + \left[\sum_{k=1}^d \beta_k \left(-\mu_k + \sum_{j=1}^m p_j \phi_k(j)\right)\right]$$

and the dual function is

$$\ell(\lambda,\alpha,\beta) = \inf_p \mathcal{L}(p,\lambda,\alpha,\beta)$$

Solving for the infimum with respect to p,

$$\begin{aligned}
\frac{d\mathcal{L}}{dp_j} &= 0 \\
(\log p_j + 1) - \lambda_j + \alpha + \sum_{k=1}^{d} \beta_k \phi_k(j) &= 0 \\
\log p_j &= \lambda_j - 1 - \alpha - \beta^\top \phi(j) \\
p_j^* = p_j &= \exp\left(\lambda_j - \alpha - \beta^\top \phi(j) - 1\right)
\end{aligned}$$

Hence

$$\begin{aligned}
\ell(\lambda,\alpha,\beta) &= \left[\sum_{j=1}^{m} p_j^* \log p_j^*\right] - \left[\sum_{j=1}^{m} p_j^* \lambda_j\right] + \left[\alpha\left(-1 + \sum_{j=1}^{m} p_j^*\right)\right] + \left[\sum_{k=1}^{d} \beta_k \left(-\mu_k + \sum_{j=1}^{m} p_j^* \phi_k(j)\right)\right] \\
&= \left[\sum_{j=1}^{m} p_j^* \log p_j^*\right] - \left[\sum_{j=1}^{m} p_j^* \lambda_j\right] - \alpha + \left[\sum_{j=1}^{m} p_j^* \alpha\right] + \left[\sum_{j=1}^{m} p_j^* \beta^\top \phi(j)\right] - \beta^\top \mu \\
&= \left[\sum_{j=1}^{m} p_j^* \left(\log p_j^* - \lambda_j + \alpha + \beta^\top \phi(j)\right)\right] - \alpha - \beta^\top \mu \\
&= \left[-\sum_{j=1}^{m} p_j^*\right] - \alpha - \beta^\top \mu \\
&= -\beta^\top \mu - \alpha - e^{-\alpha-1} \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)
\end{aligned}$$

Note that $p_j^* = \exp\left(\lambda_j - \beta^\top \phi(j)\right) \exp(-\alpha - 1)$ satisfies $\sum_{j=1}^{m} p_j^* = 1$, and therefore $\exp(-\alpha-1)$ must be a normalizing factor:

$$\begin{aligned}
\exp(-\alpha - 1) &= \frac{1}{\sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)} \\
\exp(\alpha + 1) &= \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right) \\
\alpha &= \left[\log \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)\right] - 1
\end{aligned}$$

Thus we can eliminate α:

$$\begin{aligned}
\ell(\lambda,\alpha,\beta) &= -\beta^\top \mu - \alpha - e^{-\alpha-1} \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right) \\
\ell(\lambda,\beta) &= -\beta^\top \mu - \left[\log \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)\right] + 1 - \frac{\sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)}{\sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)} \\
\ell(\lambda,\beta) &= -\beta^\top \mu - \log \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^\top \phi(j)\right)
\end{aligned}$$

Finally, observe that

$$
\begin{aligned}
\exp \ell(\lambda, \beta) &= \exp\left(-\beta^{\top}\mu\right) \exp\left(-\log \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^{\top}\phi(j)\right)\right) \\
&= \frac{\exp\left(-\beta^{\top}\mu\right)}{\sum_{j=1}^{m} \exp\left(\lambda_j - \beta^{\top}\phi(j)\right)}
\end{aligned}
$$

which is the likelihood of an exponential family, provided that β, λ satisfy $m \exp\left(-\beta^{\top}\mu\right) = \sum_{j=1}^{m} \exp\left(\lambda_j - \beta^{\top}\phi(j)\right)$.

3

(a)

Theorem 26.18: Fix any $\delta > 0$. Then

$$
\sup_{p \in \Sigma(\beta, L)} \mathbb{P}\left(\left|\widehat{p}(x) - p(x)\right| > \sqrt{\frac{C \log(2/\delta)}{nh^d}} + ch^{\beta}\right) < \delta
$$

We now repeat the proof with Hoeffding's inequality. By definition, $\widehat{p}(x) = n^{-1} \sum_{i=1}^{n} Z_i$ where

$$
Z_i = \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right)
$$

Let $p_h(x) = \mathbb{E}(\widehat{p}(x))$. Observe that

$$
\mathbb{E}\left(\widehat{p}(x) - p_h(x)\right) = 0
$$

and

$$
|Z_i| \leq \frac{c_1}{h^d}
$$

where $c_1 = K(0)$ (the kernel is maximized at $K(0)$), which in turn implies

$$
|Z_i - p_h(x)| \leq \frac{c_1}{h^d}
$$

We then apply Hoeffding's inequality:

$$
\begin{aligned}
\mathbb{P}\left(\left|\widehat{p}(x) - p_h(x)\right| > \epsilon\right) &< 2 \exp\left\{\frac{-2n\epsilon^2}{4c_1^2/h^{2d}}\right\} \\
&= 2 \exp\left\{\frac{-nh^{2d}\epsilon^2}{2c_1^2}\right\}
\end{aligned}
$$

Choosing $\epsilon = \sqrt{C \log(2/\delta) / nh^{2d}}$ where $C = 2c_1^2$ gives

$$
\mathbb{P}\left(\left|\widehat{p}(x) - p_h(x)\right| > \sqrt{\frac{C \log(2/\delta)}{nh^{2d}}}\right) < \delta \tag{6}
$$

Observe the h^{2d} factor where Bernstein's inequality would have given h^d. By the triangle inequality, for any p we have that

$$
\left|\widehat{p}(x) - p(x)\right| \leq \left|\widehat{p}(x) - p_h(x)\right| + \left|p_h(x) - p(x)\right|
$$

From Lemma 26.11, $|p_h(x) - p(x)| \le ch^\beta$ for some c, and therefore

$$|\widehat{p}(x) - p(x)| \le |\widehat{p}(x) - p_h(x)| + ch^\beta$$

for any p. Comparing this with (6) gives the result

$$\sup_{p\in\Sigma(\beta,L)} \mathbb{P}\left(|\widehat{p}(x) - p(x)| > \sqrt{\frac{C\log(2/\delta)}{nh^{2d}}} + ch^\beta\right) \le \mathbb{P}\left(|\widehat{p}(x) - p_h(x)| + ch^\beta > \sqrt{\frac{C\log(2/\delta)}{nh^{2d}}} + ch^\beta\right) < \delta$$

The $\sqrt{h^{-2d}}$ factor (as opposed to $\sqrt{h^{-d}}$ from Bernstein's inequality) makes the corresponding term in the probability statement larger, hence the bound is weaker. Compare Bernstein's inequality

$$\mathbb{P}\left(\left|\bar{Z} - \mu\right| > \epsilon\right) < 2\exp\left\{-\frac{n\epsilon^2}{2\sigma_Z^2 + 2M_Z\epsilon/3}\right\}$$

with Hoeffding's inequality

$$\mathbb{P}\left(\left|\bar{Z} - \mu\right| > \epsilon\right) < 2\exp\left\{-\frac{2n\epsilon^2}{\left(b_{Z-\mu} - a_{Z-\mu}\right)^2}\right\}$$

Observe that the denominator in Bernstein's inequality is $\mathrm{O}\left(\sigma_Z^2 + M_Z\right)$, while the denominator in Hoeffding's inequality is $\mathrm{O}\left(\left(b_{Z-\mu} - a_{Z-\mu}\right)^2\right)$. Because $|Z_i| \le M_Z = \frac{c_1}{h^d}$ and $\sigma_Z^2 \le \frac{c_2}{h^d}$ (Lemma 26.13), the denominator in Bernstein's inequality is $\mathrm{O}\left(h^{-d}\right)$. But $b_{Z-\mu} - a_{Z-\mu} \le \frac{2c_1}{h^d}$, so the denominator in Hoeffding's inequality is $\mathrm{O}\left(h^{-2d}\right)$. In short, the reason why Bernstein's inequality yields the better rate since it utilizes the information of variance.

(b)

The LOOCV estimator of risk, for a particular bandwidth h, is

$$\widehat{R}(h) = \int \left(\widehat{p}(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^n \widehat{p}_{(-i)}(X_i)$$

where

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right)$$

Suppose $x_a = x_b$ for some $a \ne b$, and assume this is the only tie in the data. Consider the LOOCV estimator with x_a held-out, evaluated at x_a:

$$\begin{aligned}
\widehat{p}_{(-a)}(x_a) &= \frac{1}{n-1}\frac{1}{h^d}K\left(\frac{\|x_a - x_b\|}{h}\right) + \frac{1}{n-1}\sum_{i\notin\{a,b\}} \frac{1}{h^d}K\left(\frac{\|x_a - x_i\|}{h}\right) \\
&= \frac{1}{n-1}\frac{1}{h^d}K\left(\frac{0}{h}\right) + \frac{1}{n-1}\sum_{i\notin\{a,b\}} \frac{1}{h^d}K\left(\frac{\|x_a - x_i\|}{h}\right)
\end{aligned}$$

As $h \to 0$, the distribution of the kernel approaches a point mass at 0. Hence the first term approaches ∞ and the second term approaches 0. Thus $\lim_{h\to 0}\widehat{p}_{(-a)}(x_a) = \infty$, and

$$\begin{aligned}
\lim_{h\to 0}\widehat{R}(h) &= \lim_{h\to 0}\left[\int \left(\widehat{p}(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^n \widehat{p}_{(-i)}(X_i)\right] \\
&= \lim_{h\to 0}\left[\int \left(\widehat{p}(x)\right)^2 dx - \frac{2}{n}\left(\widehat{p}_{(-a)}(X_a) + \sum_{i\ne a}^n \widehat{p}_{(-i)}(X_i)\right)\right] \\
&= -\infty
\end{aligned}$$

Therefore cross-validation will choose $\widehat{h} = 0$ because it yields the smallest estimated risk.

To fix this problem, we can remove all but one of the K tied data points, and "reweigh" the remaining point by K. Let us refer to the earlier example — in this case, we remove x_b from the data, and double the weight of the kernel x_a to get the following kernel density estimator:

$$\widehat{p}^*(x) = \frac{1}{n}\left[\frac{2}{h^d}K\left(\frac{\|x-x_a\|}{h}\right) + \sum_{i\neq a}\frac{1}{h^d}K\left(\frac{\|x-x_i\|}{h}\right)\right]$$

and the following LOOCV estimator:

$$\widehat{p}^*_{(-j)}(x) = \begin{cases} \frac{1}{n-1}\left[\frac{2}{h^d}K\left(\frac{\|x-x_a\|}{h}\right) + \sum_{i\notin\{a,j\}}\frac{1}{h^d}K\left(\frac{\|x-x_i\|}{h}\right)\right] & j \neq a \\ \frac{1}{n-2}\sum_{i\neq j}\frac{1}{h^d}K\left(\frac{\|x-x_i\|}{h}\right) & j = a \end{cases}$$

We also double the weight of x_a in the LOOCV risk estimator:

$$\widehat{R}^*(h) = \int\left(\widehat{p}(x)\right)^2 dx - \frac{2}{n}\left[2\widehat{p}^*_{(-a)}(X_a) + \sum_{i\neq a}\widehat{p}^*_{(-i)}(X_i)\right]$$

Observe the following:

(a) $\widehat{p}^*(x) = \widehat{p}(x)$, i.e. the new kernel density estimator is identical to the previous one.

(b) $\widehat{p}^*_{(-j)}(x_j) = \widehat{p}_{(-j)}(x_j)$ when $j \neq a$, i.e. the LOOCV estimator is identical when the held-out data point is not x_a.

(c) The only difference occurs when x_a is held out, that is to say $\widehat{p}^*_{(-a)}(x_a) \neq \widehat{p}_{(-a)}(x_a)$. However, $\lim_{h\to 0}\widehat{p}^*_{(-a)}(x_a) = 0$ because there are no ties ($x_a \neq x_i$ for any $i \neq a$ since x_b was removed). Hence $\lim_{h\to 0}\widehat{R}^*(h) \neq -\infty$, so the problem has been fixed.

(c)

$$\begin{aligned}
\widehat{L}(D) &= \int_{[0,1]} \widehat{f}_{X,D}^{2}(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{X,D}^{(i)}(X_i) \\
&= \int_{[0,1]}\left(\frac{D}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i\in B(x)\}\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\left(\frac{D}{n-1}\sum_{j\neq i}^{n}\mathbb{I}\{X_j\in B(X_i)\}\right) \\
&= \frac{D^2}{n^2}\int_{[0,1]}\left(\sum_{i=1}^{n}\mathbb{I}\{X_i\in B(x)\}\right)^2 dx - \frac{2D}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbb{I}\{X_j\in B(X_i)\} \\
&= \frac{D^2}{n^2}\sum_{k=1}^{D}\frac{1}{D}\left(\sum_{i=1}^{n}\mathbb{I}\{X_i\in \mathrm{Bin}(k)\}\right)^2 - \frac{2D}{n(n-1)}\sum_{i=1}^{n}(|B(X_i)|-1) \\
&= \frac{D}{n^2}\sum_{k=1}^{D}|\mathrm{Bin}(k)|^2 - \frac{2D}{n(n-1)}\sum_{k=1}^{D}|\mathrm{Bin}(k)|(|\mathrm{Bin}(k)|-1) \\
&= \frac{D}{n^2}\sum_{k=1}^{D}|\mathrm{Bin}(k)|^2 - \frac{2D}{n(n-1)}\sum_{k=1}^{D}|\mathrm{Bin}(k)|^2 + \frac{2D}{n(n-1)}\sum_{k=1}^{D}|\mathrm{Bin}(k)| \\
&= \frac{D}{n-1}\left[\left(\frac{n-1}{n^2}-\frac{2}{n}\right)\sum_{k=1}^{D}|\mathrm{Bin}(k)|^2\right] + \frac{2D}{n-1} \\
&= \frac{2D}{n-1} + \frac{D}{n-1}\left[\left(\frac{-n-1}{n^2}\right)\sum_{k=1}^{D}|\mathrm{Bin}(k)|^2\right] \\
&= \frac{2D}{n-1} - \frac{D(n+1)}{n-1}\sum_{j=1}^{D}\left(\frac{|\mathrm{Bin}(j)|}{n}\right)^2
\end{aligned}$$

which was to be shown.

4

Summary of results:

a	0.1	0.5	0.95
glasso best ℓ	95310	94974	94924
glasso λ from best ℓ	1×10^{-5}	1×10^{-5}	3.3×10^{-6}
glasso $\left\|\widehat{\Sigma}-\Sigma\right\|_F$ from best ℓ	0.101	0.169	3.91
thresholding best ℓ	95328	93368	97852
thresholding M from best ℓ	3.3×10^{-4}	3.3×10^{-4}	1×10^{-3}
thresholding $\left\|\widehat{\Sigma}-\Sigma\right\|_F$ from best ℓ	0.101	0.170	3.91

Values of λ and M were selected from $\{1\times10^{-7}, 3.3\times10^{-7}, 1\times10^{-6}, \ldots, 3.3\times10^{-1}, 1\}$.

- At their optimal tuning parameters, both glasso and thresholding perform equally well in terms of log-likelihood and $\left\|\widehat{\Sigma}-\Sigma\right\|_F$. According to the analytical expression for $\mathrm{cov}(t_1,t_2)$, there is a continuum between the non-sparse entries on the diagonal and the sparse entries at the upper-right and lower-left corners — that is to say, the distinction between sparse and non-sparse entries is unclear. In principle, glasso should perform better — it minimizes the negative log-likelihood subject to an ℓ_1 penalty, while the thresholding procedure uses a cutoff that merely depends on n and T; glasso considers statistical properties of the data that the thresholding procedure ignores. However, the aforementioned continuum suggests that an appropriately chosen cutoff is adequate for the problem. Hence glasso and thresholding perform equally well under their optimal tuning parameters.

- According to the analytical expression for $\text{cov}(t_1, t_2)$, the true covariance matrix has its largest elements $\sigma^2 \frac{1-a^{2t_1}}{1-a^2}$ on the diagonal, while the off-diagonal elements decrease exponentially at the rate of $a^{|t_2-t_1|}$. Hence the the proportion of sparse entries decreases as $a \to 1$. Both glasso and thresholding favor sparse estimates of the covariance, consequently $\left\|\widehat{\Sigma} - \Sigma\right\|_F$ increases for both methods as we increase a (and hence decrease sparsity).

We now give the analytical expression for $\text{cov}(t_1, t_2)$. We first assume that $t_1 \leq t_2$:

$$\begin{aligned}
\text{cov}(t_1, t_2) &= \text{cov}(X_{t_1}, X_{t_2}) \\
&= \mathbb{E}\left[\left(X_{t_1} - \bar{X}_{t_1}\right)\left(X_{t_2} - \bar{X}_{t_2}\right)\right] \\
&= \mathbb{E}[X_{t_1} X_{t_2}] \quad \text{(all } X_t\text{s have mean 0)} \\
&= \mathbb{E}[X_{t_1}(aX_{t_2-1} + \epsilon_{t_2-1})] \\
&= \mathbb{E}[X_{t_1}(a(aX_{t_2-2} + \epsilon_{t_2}) + \epsilon_{t_2-1})] \\
&\vdots \\
&= \mathbb{E}\left[X_{t_1}\left(a^{t_2-t_1} X_{t_1} + a^{t_2-t_1-1}\epsilon_{t_1} + a^{t_2-t_1-2}\epsilon_{t_1+1} + \cdots + a\epsilon_{t_2-2} + \epsilon_{t_2-1}\right)\right] \\
&= a^{t_2-t_1}\mathbb{E}\left[X_{t_1}^2\right] + a^{t_2-t_1-1}\mathbb{E}[X_{t_1}\epsilon_{t_1}] + \cdots + \mathbb{E}[X_{t_1}\epsilon_{t_2-1}]
\end{aligned}$$

Observe that $\mathbb{E}[X_{t_1}\epsilon_{t_i}] = \mathbb{E}\left[\left(X_{t_1} - \bar{X}_{t_1}\right)(\epsilon_{t_i} - \bar{\epsilon}_{t_i})\right] = \text{cov}(X_{t_1}, \epsilon_{t_i}) = 0$ for all $t_i > t_1$. Hence

$$\begin{aligned}
\text{cov}(t_1, t_2) &= a^{t_2-t_1}\mathbb{E}\left[X_{t_1}^2\right] \\
&= a^{t_2-t_1}\mathbb{E}\left[\left(X_{t_1} - \bar{X}_{t_1}\right)^2\right] \quad (X_{t_1} \text{ has mean 0}) \\
&= a^{t_2-t_1}\mathbb{V}[X_{t_1}] \\
&= a^{t_2-t_1}\mathbb{V}\left[a^{t_1}X_0 + a^{t_1-1}\epsilon_0 + a^{t_1-2}\epsilon_1 + \cdots + \epsilon_{t_1-1}\right] \\
&= a^{t_2-t_1}\left(0 + a^{2(t_1-1)}\sigma^2 + a^{2(t_1-2)}\sigma^2 \cdots + \sigma^2\right) \quad (\epsilon_t\text{s are uncorrelated}) \\
&= a^{t_2-t_1}\sigma^2 \sum_{i=0}^{t_1-1} a^{2i} \\
&= a^{t_2-t_1}\sigma^2 \frac{1-a^{2t_1}}{1-a^2} \\
&= \sigma^2 \frac{a^{t_2-t_1} - a^{t_2+t_1}}{1-a^2}
\end{aligned}$$

Since $\text{cov}(t_1, t_2) = \text{cov}(t_2, t_1)$, we have that

$$\text{cov}(t_1, t_2) = \sigma^2 \frac{a^{|t_2-t_1|} - a^{t_2+t_1}}{1-a^2}$$