ECE60131: Inference & Learning in Generative Models

Fall 2025

Homework 1

Due: 09/07

basic machinery

Some of these problems were adapted from those given by M.I. Jordan for U.C. Berkeley's CS281a.

1. **Conditional independence.**
   Show that

$$p(x, y|z) = p(x|z)p(y|z) \iff p(x|y, z) = p(x|z) \iff p(y|x, z) = p(y|z)$$

$$p(x,y|z) = p(x|z)p(y|z) \iff \frac{p(x,y,z)}{p(z)} = p(x|z)p(y|z)$$

$$\iff \frac{p(x,y,z)}{\underbrace{p(y|z)p(z)}_{= \frac{p(y,z)}{p(z)}p(z) = p(y,z)}} = p(x|z) \iff \frac{p(x,y,z)}{p(y,z)} = p(x|z) \iff$$

$$\iff p(x|y,z) = p(x|z) \quad \square \text{ (first equivalence)}$$

Similarly:

$$p(x,y|z) = p(x|z)p(y|z) \iff \frac{p(x,y,z)}{p(z)p(x|z)} = p(y|z) \iff$$

$$\iff \frac{p(x,y,z)}{p(x,z)} = p(y|z) \iff p(y|x,z) = p(y|z) \quad \square$$

And since it's equivalent to the first equation, it is equivalent to the second equation.

This shows the desired equivalence.

Minimal conditional independence. For a given variable X_i in a graphical model, what is the minimal set of nodes that renders X_i conditionally independent of all of the other variables? That is, what is the smallest set C such that $X_i \perp\!\!\!\perp X_{V\setminus\{i\}\cup C}$? (Note that V is the set of all nodes in the graph, so that $V\setminus\{i\}\cup C$ is the set of all nodes in the graph excluding i and C.)

2. **Undirected.**
   Solve the problem for an undirected graph.
3. **Directed.**
   Solve the problem for a directed graph.

In both cases, you may describe the set in words.

2. For an undirecteg graph, C is the set of all neighbors of X_i, i.e. all nodes directly connected to X_i. Mathematically: $C := \{V_j \mid (V_j, X_i) \in E\}$ where E is the set of all edges (undirected, so (V_j, X_i) is the same as (X_i, V_j)).

3. For a directed graph, first let $P(X_i)$ be the set of all parents of X_i (i.e. all nodes with outgoing edges to X_i), and $D(X_i)$ the set of all children of X_i (nodes with edges incoming from X_i). Finally, let $P'(X_i)$ be the set of all parents of children of X_i not equal to X_i, i.e. $P'(X_i) := \left(\bigcup_{X_j \in D(X_i)} P(X_j)\right) \setminus X_i$. Then, C is the union of these sets: $C := P(X_i) \cup D(X_i) \cup P'(X_i)$, i.e. the set of parents, children, and coparents of X_i's children.

Modeling factorizations. Consider a probability distribution that factors like this:

$$\hat{p}(\hat{x}_1, \hat{x}_2, \hat{x}_3) = f_a(\hat{x}_1, \hat{x}_2) f_b(\hat{x}_1, \hat{x}_3) f_c(\hat{x}_2, \hat{x}_3).$$

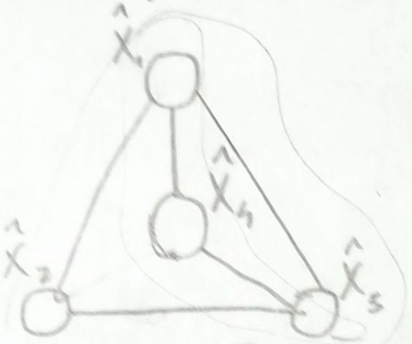
As discussed in class, no three-node graphical model, either directed or undirected, can enforce this factorization. However, it can be done with the help of auxiliary variables. Shows this for both types of graphical model.

4. **Undirected.**

5. **Directed.**

Assume $\hat{X}_i$ are discrete. (Hint: see the discussion on p. 16, Ch. 4 of IPGM.)

4. Introducing auxilary variable $\hat{X}_4$, such that the graph is:



Next, the new cliques are $(\hat{X}_1, \hat{X}_3, \hat{X}_4)$ and $(\hat{X}_1, \hat{X}_2, \hat{X}_3)$ and given potentials ψ_a, ψ_b, it follows that:

$$p(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4) = \psi_a(\hat{x}_1, \hat{x}_3, \hat{x}_4) \psi_b(\hat{x}_1, \hat{x}_2, \hat{x}_3)$$

Now let: $\psi_a(\hat{x}_1, \hat{x}_3, \hat{x}_4) = f_b(\hat{x}_1, \hat{x}_3) \delta(\hat{x}_4)$, $\psi_b(\hat{x}_1, \hat{x}_2, \hat{x}_3) = f_a(\hat{x}_1, \hat{x}_2) f_c(\hat{x}_2, \hat{x}_3)$

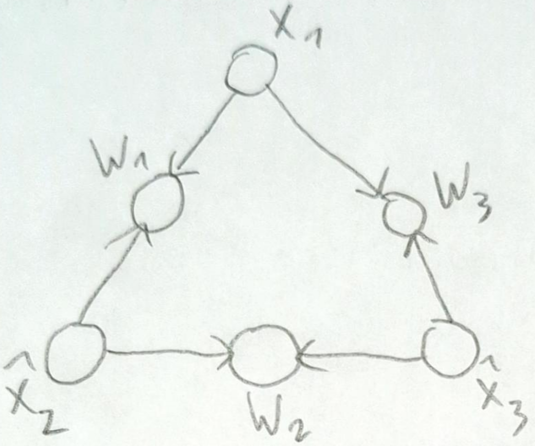
Then: $\sum_{\hat{x}_4} p(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4) = \sum_{\hat{x}_4} \psi_a(\hat{x}_1, \hat{x}_3, \hat{x}_4) \psi_b(\hat{x}_1, \hat{x}_2, \hat{x}_3) = \sum_{\hat{x}_4} f_b(\hat{x}_1, \hat{x}_3) \delta(\hat{x}_4) f_a(\hat{x}_1, \hat{x}_2) f_c(\hat{x}_2, \hat{x}_3)$

$= f_b(\hat{x}_1, \hat{x}_3) f_a(\hat{x}_1, \hat{x}_2) f_c(\hat{x}_2, \hat{x}_3) \underbrace{\sum_{\hat{x}_4} \delta(\hat{x}_4)} = f_a(\hat{x}_1, \hat{x}_2) f_b(\hat{x}_1, \hat{x}_3) f_c(\hat{x}_2, \hat{x}_3) =$

$= \hat{p}(\hat{x}_1, \hat{x}_2, \hat{x}_3)$, which demonstrates the factorization for an undirected graph

5. Using 3 new auxilary variables W_1, W_2, W_3 s.t.

$p(W_1 = 1 \mid \hat{x}_1, \hat{x}_2) = f_a(\hat{x}_1, \hat{x}_2)$ (and similarly rest, binary) and the graph is:



Then:

$$p(\hat{x}_1, \hat{x}_2, \hat{x}_3, W_1, W_2, W_3) = p(\hat{x}_1)\, p(\hat{x}_2)\, p(\hat{x}_3)\, p(W_1 \mid \hat{x}_1, \hat{x}_2)\, p(W_2 \mid \hat{x}_2, \hat{x}_3)\, p(W_3 \mid \hat{x}_1, \hat{x}_3)$$

Condition on $W_1 = 1, W_2 = 1, W_3 = 1$:

$$p(\hat{x}_1, \hat{x}_2, \hat{x}_3 \mid W_1 = 1, W_2 = 1, W_3 = 1) = p(\hat{x}_1)\, p(\hat{x}_2)\, p(\hat{x}_3) \underbrace{p(W_1 = 1 \mid \hat{x}_1, \hat{x}_2)\, p(W_2 = 1 \mid \hat{x}_2, \hat{x}_3)\, p(W_3 = 1 \mid \hat{x}_1, \hat{x}_3)}_{= f_a(\hat{x}_1, \hat{x}_2)\, f_b(\hat{x}_2, \hat{x}_3)\, f_c(\hat{x}_1, \hat{x}_3)}$$

$$\Downarrow$$

$$p(\hat{x}_1, \hat{x}_2, \hat{x}_3 \mid W_1 = 1, W_2 = 1, W_3 = 1) = p(\hat{x}_1)\, p(\hat{x}_2)\, p(\hat{x}_3)\, f_a(\hat{x}_1, \hat{x}_2)\, f_b(\hat{x}_2, \hat{x}_3)\, f_c(\hat{x}_1, \hat{x}_3)$$

So

$$p(\hat{x}_1, \hat{x}_2, \hat{x}_3 \mid W_1 = 1, W_2 = 1, W_3 = 1) \propto f_a(\hat{x}_1, \hat{x}_2)\, f_b(\hat{x}_2, \hat{x}_3)\, f_c(\hat{x}_1, \hat{x}_3)$$

And therefore:

$$\hat{p}(\hat{x}_1, \hat{x}_2, \hat{x}_3) = f_a(\hat{x}_1, \hat{x}_2)\, f_b(\hat{x}_2, \hat{x}_3)\, f_c(\hat{x}_1, \hat{x}_3)$$

Tree representations. Consider an undirected tree. We know that it is not possible in general to parameterize a distribution on a tree using marginal probabilities as the potentials. It is, however, possible to parameterize such a distribution using ratios of marginal probabilities. In particular, let:

$$\psi_i(x_i) = p(x_i)$$
$$\psi_{ij}(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$$

where i and j are neighbors in the tree, and the "marginal" probabilities $p(x_i)$ and $p(x_i, x_j)$ are all mutually consistent.

6. **Joint.**
Show that this setting of potentials yields a parameterization of a joint probability distribution on the tree under which $p(x_i)$ and $p(x_i, x_j)$ are marginals.

7. **Normalizer.**
What is the normalizer, Z, under this parameterization?

6. We know that for a tree graph, we can write the joint probability distribution as (as given by eq. 4.1 in Jordan's book):

$$p(X) = \frac{1}{Z}\left(\prod_{i\in V}\psi(x_i)\prod_{(i,j)\in\mathcal{E}}\psi(x_i, x_j)\right) \underset{\text{substitute}}{=}$$

$$= \frac{1}{Z}\prod_{i\in V}p(x_i)\prod_{(i,j)\in\mathcal{E}}\frac{p(x_i,x_j)}{p(x_i)p(x_j)} = \frac{1}{Z}\frac{\prod_{i\in V}p(x_i)\prod_{(i,j)\in\mathcal{E}}p(x_i,x_j)}{\prod_{k\in V}p(x_k)^{\deg(k)}} =$$

$$\prod_{(i,j)\in\mathcal{E}}p(x_i)p(x_j) = \prod_{k\in V}\underbrace{p(x_k)^{\deg(k)}}_{\text{every } p(x_k) \text{ appears } \deg(k) \text{ times}}$$

$$= \frac{1}{Z}\frac{\prod_{(i,j)\in\mathcal{E}}p(x_i,x_j)}{\prod_{i\in V}p(x_i)^{\deg(i)-1}}$$

Now to show they are marginals, I need to show: $\sum_{V\setminus\{i\}}p(x) = p(x_i)$.

This is equivalent to marginalizing out all other x_k's by summing over them. Suppose l is a leaf node, i.e. $\deg(l)=1$. To marginalize out l, compute $\sum_{x_l}p(x) = ?$

$$\sum_{x_l} p(x) = \frac{1}{Z}\sum_{x_l}\frac{\prod_{(i,j)\in\mathcal{E}} p(x_i,x_j)}{\prod_{i\in V} p(x_i)^{\deg(i)-1}} = \frac{1}{Z}\Bigg(\sum_{x_l}\frac{\overbrace{p(x_k,x_l)}^{\text{only 1 edge.}}}{\underbrace{p(x_l)^0}_{=1}}\Bigg)\Bigg(\frac{\prod_{(i,j)\in\mathcal{E}\setminus\{(k,l)\}} p(x_i,x_j)}{\prod_{i\in V\setminus\{l\}} p(x_i)^{\deg(i)-1}}\Bigg)$$

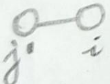
leaf has $\deg(l)-1=0$

$$= \frac{1}{Z}\underbrace{p(x_k)}_{\text{since } \sum_{x_l} p(x_k,x_l)=p(x_k)}\frac{\prod_{(i,j)\in\mathcal{E}'} p(x_i,x_j)}{\prod_{i\in V'} p(x_i)^{\deg(i)-1}} = \frac{1}{Z}\frac{\prod_{(i,j)\in\mathcal{E}'} p(x_i,x_j)}{\prod_{i\in V'} p(x_i)^{\deg'(i)-1}},$$

where $\mathcal{E}' := \mathcal{E}\setminus\{(k,l)\}$, $V' := V\setminus\{(l)\}$, $\deg'(i) := \begin{cases}\deg(i), & i\neq k\\ \deg(i)-1, & i=k\end{cases}$.

Note that this is just the original equation for a new tree without the node j and with k without j as the neighbor. Repeating this process until only i is left, by induction, yields:

$$\sum_{V\setminus\{i\}} p(x) \underset{\substack{\text{step before}\\ \text{last:}}}{=} \frac{\frac{1}{Z}\sum_{j'} p(j',i)}{p(j')^0 p(i)^0} = \frac{1}{Z}\underbrace{\sum_{j'} p(x_{j'},x_i)}_{}\cdot 1 = p(x_i),$$



$=1$, as shown below

so $p(x_i)$ is a marginal. Similar process follows for any $p(x_j, x_i)$.

7. Suppose in 6, $\frac{1}{Z} = ?$. Then, we know that:

$$\sum_V p(x) = 1 \underbrace{=}_{\substack{\text{from}\\ \text{last eq in 6.}}} \frac{1}{Z}\sum_i\sum_{j'} p(x_{j'},x_i) = \frac{1}{Z}\underbrace{\sum_{x_i} p(x_i)}_{\substack{\text{probability}\\ \text{must sum to 1}}} = \frac{1}{Z} \Rightarrow \frac{1}{Z} = 1, \text{ so } Z=1$$

8. **Multivariate normal distribution.**
The multivariate normal distribution can be expressed in exponential-family form,

$$p(x;\boldsymbol{\eta}) = h(x)\exp\{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{T}(x) - A(\boldsymbol{\eta})\}$$

with

$$\boldsymbol{T}(x) = \begin{pmatrix} x \\ \\ \mathrm{vec}\left[xx^{\mathrm{T}}\right] \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1}\mu \\ \\ -\frac{1}{2}\mathrm{vec}\left[\boldsymbol{\Sigma}^{-1}\right], \end{pmatrix}.$$

where vec[·] "vectorizes" its matrix argument by stacking its columns. Derive this. (Hint: you may want to use some properties of the matrix trace.)

8. By definition:

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$= \underbrace{\frac{1}{(2\pi)^{d/2}}}_{\triangleq h(x)}\exp\left(-\frac{1}{2}\left(x^T\Sigma^{-1}x - \underbrace{x^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x}_{\text{scalars, can combine}} + \mu^T\Sigma^{-1}\mu + \ln|\Sigma|\right)\right) =$$

$= h(x)\exp\left(-\frac{1}{2}\left(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu + \ln|\Sigma|\right)\right)$. Now, note that $x^T\Sigma^{-1}x$ is scalar, so: $x^T\Sigma^{-1}x = Tr(x^T\Sigma^{-1}x) = Tr(\Sigma^{-1}xx^T) =$

$= vec(\Sigma^{-1})^T vec(xx^T)$. Substitute back.

because $\Sigma^{-1} = (\Sigma^{-1})^T$

by properties of trace

$$p(x;\mu,\Sigma) = h(x)\exp\left(-\frac{1}{2}vec(\Sigma^{-1})^T vec(xx^T) + \overbrace{\mu^T(\Sigma^{-1})^T x}^{\text{same as } x^T\Sigma^{-1}\mu,} - \overbrace{\left(\frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\ln|\Sigma|\right)}^{A(\eta)}\right)$$

$$= h(x)\exp\left(\overbrace{\left[\Sigma^{-1}\mu \;\; -\frac{1}{2}vec(\Sigma^{-1})\right]^T}^{\eta^T}\overbrace{\begin{bmatrix} x \\ vec(xx^T)\end{bmatrix}}^{T(x)} - A(\eta)\right) =$$

$$= h(x)\exp\left(\eta^T T(x) - A(\eta)\right) \quad \square$$

9. **Cumulants of the normal distribution.**
Use the cumulant-generating property of the log-partition function to show that the third and fourth cumulants of the (univariate) normal distribution are 0.

First, for the normal distribution,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) =$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln\sigma\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln\sigma\right)}_{A(\eta)}\right).$$ Let.

$\eta_1 = \frac{\mu}{\sigma^2}$, $\eta_2 = -\frac{1}{2}\sigma^{-2}$, then:

$$A(\eta) = \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2).$$

Then:

$$K_3 = \frac{\partial^3 A}{\partial\eta_1^3} = \frac{\partial^2}{\partial\eta_1^2}\left(\frac{\partial A}{\partial\eta_1}\right) = \frac{\partial^2}{\partial\eta_1^2}\left(-\frac{\eta_1}{2\eta_2}\right) = \frac{\partial}{\partial\eta_1}\left(-\frac{1}{2\eta_2}\right) =$$

$= 0$, so the 3-rd cumulant is 0.

$$K_4 = \frac{\partial^4 A}{\partial\eta_1^4} = \frac{\partial}{\partial\eta_1}\left(\frac{\partial^3 A}{\partial\eta_1^3}\right) = \frac{\partial}{\partial\eta_1} 0 = 0$$

So this shows that the 3-rd and 4th cumulants of $N(\mu,\sigma^2)$ are 0