MathDNN HW 9

2018-13260 차재현

1 Problem 3

For a set $X \in \{1, \cdots .n\}$, define $X(i)$ be the i-th smallest element of X.

Let σ be a permutation represented by

$$[\Omega, \Omega^{\complement}] = [\Omega(1), \cdots, \Omega(|\Omega|), \Omega^{\complement}(1), \cdots, \Omega^{\complement}(|\Omega^{\complement}|)],$$

i.e., $\sigma(i) = \Omega(i)$ if $i \leq |\Omega|$, $\Omega^{\complement}(i - |\Omega|)$ else. Such σ defines a linear map $\mathbb{R}^n \to \mathbb{R}^n$, denoted as P_σ by $P_\sigma(\mathbf{e}_i) = \mathbf{e}_{\sigma(i)}$.

Let $\overline{x}, \overline{z}$ be preimages of x, z respectably under P_σ. By definition,

$$\begin{aligned}\overline{z}_{\{1,\cdots,|\Omega|\}} &= \overline{x}_{\{1,\cdots,|\Omega|\}}, \\ \overline{z}_{\{|\Omega|+1,\cdots,n\}} &= e^{s_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}})} \odot \overline{x}_{\{|\Omega|+1,\cdots,n\}} + t_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}}).\end{aligned}$$

We have derivative

$$\frac{\partial \overline{z}}{\partial \overline{x}} = \begin{pmatrix} I_{|\Omega|} & 0 \\ * & \operatorname{diag}(e^{s_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}})}) \end{pmatrix} \tag{1}$$

and log-determinant

$$\log \left| \frac{\partial \overline{z}}{\partial \overline{x}} \right| = \log \prod_{i \leq n-|\Omega|} e^{s_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}})} = \mathbf{1}_{n-|\Omega|}^\top s_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}}). \tag{2}$$

Now, since P_σ is linear, its derivative is just P_σ. Together with (1) and the relation

$$y_\Omega = \overline{y}_{\{1,\cdots,|\Omega\}}, \qquad y_{\Omega^{\complement}} = \overline{y}_{\{|\Omega+1|,\cdots,n\}}$$

for $y = x, z$ we have

$$\frac{\partial z}{\partial x} = P_\sigma \begin{pmatrix} I_{|\Omega|} & 0 \\ * & \operatorname{diag}(e^{s_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}})}) \end{pmatrix} P_\sigma^{-1}. \tag{3}$$

Since $\det(P_\sigma P_\sigma^{-1}) = 1$, (1) and (3) has same determinant. Therefore by (2),

$$\log \left| \frac{\partial z}{\partial x} \right| = \mathbf{1}_{n-|\Omega|}^\top s_\theta(\overline{x}_{\{1,\cdots,|\Omega|\}}).$$

2 Problem 4

(a) Rewrite $D_{\mathrm{KL}}(X||Y)$ as

$$D_{\mathrm{KL}}(X||Y) = \int_{\mathbb{R}^d} f(x) \left(-\log \frac{g(x)}{f(x)}\right) dx.$$

Since $-\log$ is convex, Jensen's inequality gives

$$D_{\mathrm{KL}}(X||Y) \geq -\log \int_{\mathbb{R}^d} f(x) \frac{g(x)}{f(x)} dx = -\log \int_{\mathbb{R}^d} g(x) dx = -\log(1) = 0.$$

(b) Denote f_i be a pdf of each X_i, and p_X, p_{X_i} be probabilities of sample space of X, X_i respectably. That $X = (X_1, \cdots, X_d)$ and $X_1, \cdots, X_d$ are independent is equivalent to the statement

$$f = p_X \cdot X^{-1} = \prod_{1 \leq i \leq d} p_{X_i} \cdot X_i^{-1} = \prod_i f_i.$$

The same is analogous for Y. Since pdf is integrable, we can use Fubini's theorem. We have

$$\begin{aligned}
\int_{\mathbb{R}^n} f(x) \log \frac{f_i(x_i)}{g_i(x_i)} dx &= \int_{\mathbb{R}^n} \prod_j f_j(x_j) \log \frac{f_i(x_i)}{g_i(x_i)} dx \\
&= \int_{\mathbb{R}} f_i(x_i) \log \frac{f_i(x_i)}{g_i(x_i)} \left(\prod_{j \neq i} \int_{\mathbb{R}} f_j(x_j) dx_j \right) dx_i \\
&= \int_{\mathbb{R}} f_i(x_i) \log \frac{f_i(x_i)}{g_i(x_i)} dx_i = D_{\mathrm{KL}}(X_i||Y_i).
\end{aligned}$$

Therefore

$$\begin{aligned}
D_{\mathrm{KL}}(X||Y) &= \int_{\mathbb{R}^n} f(x) \log \frac{f_(x)}{g_(x)} dx \\
&= \int_{\mathbb{R}^n} \prod_j f(x_j) \sum_i \log \frac{f_i(x)}{g_i(x)} dx \\
&= \sum_i \int_{\mathbb{R}^n} \prod_j f(x_j) \log \frac{f_i(x)}{g_i(x)} dx \\
&= \sum_i D_{\mathrm{KL}}(X_i||Y_i).
\end{aligned}$$

3 Problem 5

Make use of pdf of multivariate Gaussian distribution

$$\mathcal{N}(\mu, \Sigma) \sim \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}},$$

we can write $D_{\mathrm{KL}}(\mathcal{N}(\mu_0, \Sigma_0) || \mathcal{N}(\mu_1, \Sigma_1))$ as

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}}$$
$$\cdot \frac{1}{2}\left(\log \frac{|\Sigma_1|}{|\Sigma_0|} - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0) + (x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1)\right) dx$$

Now we divide into parts.

1. The log-term becomes

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} \log \frac{|\Sigma_1|}{|\Sigma_0|} dx$$
$$= \log \frac{|\Sigma_1|}{|\Sigma_0|} \int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} dx = \log \frac{|\Sigma_1|}{|\Sigma_0|}.$$

   because it is just a multiplication of constant and pdf. Since both covariance matrices is assumed to be positive-definite, $|\Sigma| = \det(\Sigma)$ for $\Sigma = \Sigma_0, \Sigma_1$.

2. We assume Σ_0 is positive-definite. If v is an eigenvector with corresponding eigenvalue λ, we have

$$v^\top \Sigma_0 v = \langle v, \Sigma_0 v \rangle = \overline{\lambda} \langle v, v \rangle > 0,$$

   so $\lambda > 0$. This means, spectra $\sigma(\Sigma_0)$ of Σ_0 consists of finite positive numbers.

   Regard Σ_0 as a linear operator on Banach space $\mathbb{R}^d$. Let Δ be its maximal ideal space, and $\hat{\Sigma_0}$ be Galfand transformation of Σ_0. Since Σ_0 is convariance matrix, it is self-adjoint. Applying spectral theorem on normal operator (note that self-adjoint implies normal), there is a measure E, μ defined on $\Delta, \sigma(\Sigma_0)$ respectably, satisfying

$$\Sigma_0 = \int_\Delta \hat{\Sigma_0} dE = \int_{\sigma(\Sigma_0)} \lambda d\mu(\lambda).$$

Define

$$\Sigma_0' = \int_{\sigma(\Sigma_0)} \sqrt{\lambda} d\mu(\lambda).$$

Then Σ_0' is self-adjoint and $\Sigma_0'^2 = \Sigma_0$. In this case, taking change of variable $t = \Sigma_0'^{-1}(x-\mu_0)/\sqrt{2}$, we get

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)dx \tag{4}$$

$$= \int \frac{t^\top t}{\sqrt{\pi^d}} e^{-t^\top t} dt. \tag{5}$$

Denote $t = (t_1 \cdots t_d)^\top$. Then expression becomes

$$\int \frac{\sum_i t_i^2}{\sqrt{\pi^d}} e^{-t^\top t} dt = \sum_i \int \frac{t_i^2}{\sqrt{\pi^d}} e^{-\sum_j t_j^2} dt. \tag{6}$$

Fix i. Since integrand is absolutely integrable, we apply Fubini theorem to get

$$\begin{aligned}\int \frac{t_i^2}{\sqrt{\pi^d}} e^{-\sum_j t_j^2} dt &= \int_{t_1} \cdots \int_{t_d} \int_{t_i} \frac{t_i^2}{\sqrt{\pi}} e^{-t_i^2} dt_i \frac{e^{-t_d^2}}{\sqrt{\pi}} dt_d \cdots \frac{e^{-t_1^2}}{\sqrt{\pi}} dt_1 \\ &= \prod_{j \neq i} \int_{t_j} \frac{e^{-t_j^2}}{\sqrt{\pi}} dt_j \int_{t_i} \frac{t_i^2}{\sqrt{\pi}} e^{-t_i^2}.\end{aligned}$$

Appeal to undergraduate calculus (specifically, using polar coordinate), we have $\int_{y \in \mathbb{R}} e^{-y^2} dy = \sqrt{\pi}$. To treat the last case, use the change of variable $u = \sqrt{2}t$, we get

$$\int_{t_i} \frac{t_i^2}{\sqrt{\pi}} e^{-t_i^2} = \int_u \frac{u^2}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du.$$

Since this is just a variance of standard normal distribution, it is equal to 1. Therefore every summands of (6) is 1, and (6), equivalently (4), is d.

3. To fit mean, rewrite remained part as

$$\begin{aligned}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) &= (x-\mu_0)^\top \Sigma_1^{-1}(x-\mu_0) + (\mu_0-\mu_1)^\top \Sigma_1^{-1}(\mu_0-\mu_1) \\ &= (\mu_0-\mu_1)^\top \Sigma_1^{-1}(x-\mu_0) + (x-\mu_0)^\top \Sigma_1^{-1}(\mu_0-\mu_1).\end{aligned}$$

First,

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (\mu_0-\mu_1)^\top \Sigma_1^{-1}(x-\mu_0)dx$$

$$= (\mu_0 - \mu_1)^\top \Sigma_1^{-1} \int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (x - \mu_0) dx$$

$$= (\mu_0 - \mu_1)^\top \Sigma_1^{-1} \mathbb{E}(x - \mu_0) = 0,$$

which can also be applied for the term $(x - \mu_0)^\top \Sigma_1^{-1} (\mu_0 - \mu_1)$. Next,

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (\mu_0 - \mu_1)^\top \Sigma_1^{-1} (\mu_0 - \mu_1) dx$$

$$= (\mu_0 - \mu_1)^\top \Sigma_1^{-1} (\mu_0 - \mu_1) \int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} dx$$

$$= (\mu_0 - \mu_1)^\top \Sigma_1^{-1} (\mu_0 - \mu_1) \cdot 1.$$

Thus the only term remained is $(x - \mu_0)^\top \Sigma_1^{-1} (x - \mu_0)$. Make use of $t = \Sigma_0'^{-1}(x - \mu_0)/\sqrt{2}$ again, we have

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (x - \mu_0)^\top \Sigma_1^{-1} (x - \mu_0) dx$$

$$= \int \frac{t^\top \Sigma_0' \Sigma_1^{-1} \Sigma_0' t}{\sqrt{\pi^d}} e^{t^\top t} dt.$$

Let $\Sigma = \Sigma_0' \Sigma_1^{-1} \Sigma_0'$. Then we have

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (x - \mu_0)^\top \Sigma_1^{-1} (x - \mu_0) dx$$

$$= \sum_{i,j} \int \frac{\Sigma_{ij} t_i t_j}{\sqrt{\pi^d}} e^{t^\top t} dt.$$

When $i \neq j$, integrand is odd w.r.t both t_i, t_j, so it becomes zero after integrating over $\mathbb{R}$. Thus

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (x - \mu_0)^\top \Sigma_1^{-1} (x - \mu_0) dx$$

$$= \sum_{i,j} \int \frac{\Sigma_{ij} t_i t_j}{\sqrt{\pi^d}} e^{t^\top t} dt = \sum_i \int \frac{\Sigma_{ii} t_i^2}{\sqrt{\pi^d}} e^{t^\top t} dt.$$

From the result of second part, we conclude

$$\int \frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} e^{-\frac{(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}{2}} (x - \mu_0)^\top \Sigma_1^{-1} (x - \mu_0) dx$$

$$= \sum_i \int \frac{\Sigma_{ii} t_i^2}{\sqrt{\pi^d}} e^{t^\top t} dt = \sum_i \Sigma_{ii}.$$

Thus, our proof finishes when we show $\sum_i \Sigma_{ii} = \mathrm{tr}(\Sigma_1^{-1}\Sigma_0)$. This can be done by element-wise calculation. Take a look on $\sum_i \Sigma_{ii}$. We have

$$\sum_i \Sigma_{ii} = \sum_i \sum_k \sum_l \Sigma'_{0il} {\Sigma_1^{-1}}_{lk} \Sigma'_{0ki},$$

where the summation of indices are over all possible range; $1 \sim d$. Next, together with $\Sigma_0 = \Sigma_0' \Sigma_0'$ we get

$$\begin{aligned} \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) &= \sum_l \sum_k {\Sigma_1^{-1}}_{lk} \Sigma_{0kl} \\ &= \sum_l \sum_k {\Sigma_1^{-1}}_{lk} \sum_i \Sigma'_{0ki} \Sigma'_{0il}. \end{aligned}$$

After reordering summation (possible because this is finite sum), our proof finishes.

4 Problem 6

Assume θ fixed. Since the sum $g + h$ is independent of ϕ, g obtains its maximum at ϕ minimizing $h(\theta, \phi)$ – that $h(\theta, \phi) = 0$. Such ϕ always exist by assumption. Choose $\phi = \phi(\theta)$ making h zero. Of course such ϕ is not unique, but it is enough to choose one of them. Then we have

$$f(\theta) = g(\theta, \phi).$$

Now θ maximizes f if and only if it maximizes g. Therefore

$$\mathrm{argmax}\ f = \{\theta \mid (\theta, \phi(\theta)) \in \mathrm{argmax}\ g\}.$$