02471 Machine Learning for Signal Processing

Solution

Exercise 8: Dictionary learning and source separation

8.1 ICA and Gaussian signals

Exercise 8.1.1

We get (since A is orthogonal)

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det(J(\mathbf{x}, \mathbf{s}))|} \\ \mathbf{x} &= A\mathbf{s} \\ \mathbf{s} &= A^{-1}\mathbf{x} = A^T\mathbf{x} \end{aligned}$$

Next we need to compute the Jacobian. This is easiest done if we operate on each component of $\mathbf{x}$. From our knowledge of matrix multiplication, we know that the i'th component of x is $\mathrm{x}_i = A_{i,:}\mathbf{s}$, where $A_{i,:}$ denotes the i'th row of A, i.e

$$\mathbf{x} = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_l \end{bmatrix} = \begin{bmatrix} A_{1,1}\mathrm{s}_1 + A_{1,2}\mathrm{s}_2 + \cdots + A_{1,l}\mathrm{s}_l \\ A_{2,1}\mathrm{s}_1 + A_{2,2}\mathrm{s}_2 + \cdots + A_{2,l}\mathrm{s}_l \\ \vdots \\ A_{l,1}\mathrm{s}_1 + A_{l,2}\mathrm{s}_2 + \cdots + A_{l,l}\mathrm{s}_l \end{bmatrix}$$

So, we have for example for x_1:

$$\mathrm{x}_1 = A_{1,1}\mathrm{s}_1 + A_{1,2}\mathrm{s}_2 + \cdots + A_{1,l}\mathrm{s}_l$$

Taking the derivative with respect to s_1 yields

$$\begin{aligned} \frac{\partial}{\partial \mathrm{s}_1}\mathrm{x}_1 &= \frac{\partial}{\partial \mathrm{s}_1}(A_{1,1}\mathrm{s}_1 + A_{1,2}\mathrm{s}_2 + \cdots + A_{1,l}\mathrm{s}_l) \\ &= \frac{\partial}{\partial \mathrm{s}_1}A_{1,1}\mathrm{s}_1 + \frac{\partial}{\partial \mathrm{s}_1}A_{1,2}\mathrm{s}_2 + \cdots + \frac{\partial}{\partial \mathrm{s}_1}A_{1,l}\mathrm{s}_l \\ &= A_{1,1}\frac{\partial}{\partial \mathrm{s}_1}\mathrm{s}_1 + A_{1,2}\frac{\partial}{\partial \mathrm{s}_1}\mathrm{s}_2 + \cdots + A_{1,l}\frac{\partial}{\partial \mathrm{s}_1}\mathrm{s}_l \\ &= A_{1,1} \cdot 1 + A_{1,2} \cdot 0 + \cdots + A_{1,l} \cdot 0 \\ &= A_{1,1} \end{aligned}$$

Similarly, taking the derivative with respect to x_2 yields $A_{1,2}$, and so on.

$$\begin{aligned} \frac{\partial \mathrm{x}_1}{\partial \mathrm{s}_1} &= A_{1,1} \\ \frac{\partial \mathrm{x}_1}{\partial \mathrm{s}_2} &= A_{1,2} \\ &\vdots \\ \frac{\partial \mathrm{x}_1}{\partial \mathrm{s}_l} &= A_{1,l} \end{aligned}$$

Since the Jacobian matrix is defined as

$$J(\mathbf{x}, \mathbf{s}) = \begin{bmatrix} \frac{\partial x_1}{\partial s_1} & \frac{\partial x_1}{\partial s_2} & \cdots & \frac{\partial x_1}{\partial s_l} \\ \frac{\partial x_2}{\partial s_1} & \frac{\partial x_2}{\partial s_2} & \cdots & \frac{\partial x_2}{\partial s_l} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_l}{\partial s_1} & \frac{\partial x_l}{\partial s_2} & \cdots & \frac{\partial x_l}{\partial s_l} \end{bmatrix}$$

The first row of the Jacobian will become (by substituting all the partial derivatives)

$$J(\mathrm{x}_1, \mathbf{s}) = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,l} \end{bmatrix}$$

and so on. Hence, we get

$$\begin{aligned} J(\mathbf{x}, \mathbf{s}) &= \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,l} \\ \vdots & \vdots & \ddots & \vdots \\ A_{l,1} & A_{l,2} & \cdots & A_{l,l} \end{bmatrix} \\ &= A \end{aligned}$$

Exercise 8.1.2

We know that: $\det(A^{-1}) = \frac{1}{\det(A)}$, and $p_{\mathbf{s}}(\mathbf{s}) = \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right)$, and $\mathbf{s} = A^T\mathbf{x}$ hence by substitution we get

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det(J(\mathbf{x}, \mathbf{s}))|} \\ &= \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det(A)|} \\ &= p_{\mathbf{s}}(\mathbf{s})|\det(A^{-1})| \\ &= p_{\mathbf{s}}(\mathbf{s})|\det(A^T)| \\ &= \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right)|\det(A^T)| \\ &= \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|A^T\mathbf{x}\|^2}{2}\right)|\det(A^T)| \end{aligned}$$

Since A is orthogonal we have $\det(A^T) = \pm 1 \Rightarrow |\det(A^T)| = 1$. Additionally, we have

$$\|A^T\mathbf{x}\|^2 = (A^T\mathbf{x})^T A^T\mathbf{x} = \mathbf{x}^T A A^T \mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2$$

Using these two results, we get

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right)$$

8.2 Derivation of ICA based on mutual information

Exercise 8.2.1

From section 2.5 (equation 2.158):

$$\begin{aligned} I(\mathrm{x}, \mathrm{y}) &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y)(\ln p(x, y) - \ln p(x) - \ln p(y)) dxdy \end{aligned}$$

If we handle the terms individually, we get

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x)dxdy &= \int_{-\infty}^{\infty} \ln p(x)\left(\int_{-\infty}^{\infty} p(x,y)dy\right)dx \\
&= \int_{-\infty}^{\infty} \ln p(x)p(x)dx \\
&= -H(\mathrm{x}) \\
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(y)dxdy &= -H(\mathrm{y}) \\
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x,y)dxdy &= -H(\mathrm{x},\mathrm{y})
\end{aligned}$$

Combining yields

$$\begin{aligned}
I(\mathrm{x},\mathrm{y}) := &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)(\ln p(x,y) - \ln p(x) - \ln p(y))dxdy \\
= &-H(\mathrm{x},\mathrm{y}) + H(\mathrm{x}) + H(\mathrm{y})
\end{aligned}$$

This can

Which, in the general case (l–dimensional) leads to

$$I(\mathbf{z}) = -H(\mathbf{z}) + \sum_{i=1}^{l} H(\mathrm{z}_i)$$

E.g. if we set $l = 2$, we have $\mathbf{z} = [\mathrm{x}\ \mathrm{y}]^T$

Exercise 8.2.2

Using the result from the change of variables in 8.1 we have $p_{\mathbf{z}}(\mathbf{z}) = p_{\mathbf{x}}(\mathbf{x})/|\det(W)|$.

$$\begin{aligned}
p_{\mathbf{z}}(\mathbf{z}) &= p_{\mathbf{x}}(\mathbf{x})/|\det(W)| \Rightarrow \\
\ln p_{\mathbf{z}}(\mathbf{z}) &= \ln(p_{\mathbf{x}}(\mathbf{x})/|\det(W)|) \\
&= \ln p_{\mathbf{x}}(\mathbf{x}) - \ln|\det(W)| \Rightarrow \\
\mathbb{E}[\ln p_{\mathbf{z}}(\mathbf{z})] &= \mathbb{E}[\ln p_{\mathbf{x}}(\mathbf{x})] - \mathbb{E}[\ln|\det(W)|] \\
\mathbb{E}[\ln p_{\mathbf{z}}(\mathbf{z})] &= \mathbb{E}[\ln p_{\mathbf{x}}(\mathbf{x})] - \ln|\det(W)|
\end{aligned}$$

The last rewrite is made since W is deterministic.

The entropy can be written as $H(\mathbf{z}) = -\mathbb{E}[\ln p(\mathbf{z})]$, so we obtain

$$\begin{aligned}
\mathbb{E}[\ln p_{\mathbf{z}}(\mathbf{z})] &= \mathbb{E}[\ln p_{\mathbf{x}}(\mathbf{x})] - \ln|\det(W)| \Rightarrow \\
-H(\mathbf{z}) &= -H(\mathbf{x}) - \ln|\det(W)|
\end{aligned}$$

Using derivations from previous exercise, we get

$$\begin{aligned}
I(\mathbf{z}) &= -H(\mathbf{z}) + \sum_{i=1}^{l} H(\mathrm{z}_i) \\
&= -H(\mathbf{x}) - \ln|\det(W)| + \sum_{i=1}^{l} H(\mathrm{z}_i)
\end{aligned}$$

Exercise 8.2.3

From the definition we get

$$\begin{aligned}\hat{W} &= \arg\min_W \; I(\mathbf{z}) \\ &= \arg\min_W \; -H(\mathbf{x}) - \ln|\det(W)| + \sum_{i=1}^{l} H(\mathrm{z}_i) \\ &= \arg\min_W \; -H(\mathbf{x}) - \ln|\det(W)| - \sum_{i=1}^{l} \mathbb{E}[\ln p_i(\mathrm{z}_i)])\end{aligned}$$

where we in the last line used $H(\mathbf{z}) = -\mathbb{E}[\ln p(\mathbf{z})]$.

Since $H(\mathbf{x})$ is not a function of W, we can discard that in our optimization problem, and then change the minimization problem to a maximization problem by changing signs

$$\arg\min_W \; I(\mathbf{z}) = \arg\max_W \; \ln|\det(W)| + \mathbb{E}\left[\sum_{i=1}^{l} \ln p_i(\mathrm{z}_i)\right]$$

Exercise 8.2.4

We know that (the rule is given in the exercise text):

$$\frac{d}{dW}\det(W) = W^{-T}\det(W), \quad W^{-T} := (W^{-1})^T$$

and also that:

$$\frac{d}{dx}\ln x = \frac{1}{x}, x > 0$$

Using the chain rule, and assuming that $\det(W) > 0$, we get, since $|\det(W)| = \det(W)$:

$$\begin{aligned}\frac{d}{dW}\ln|\det(W)| &= \frac{\partial \ln\det(W)}{\partial \det(W)} \cdot \frac{d\det(W)}{dW} \\ &= \frac{1}{\det(W)} \cdot W^{-T}\det(W) \\ &= W^{-T}\end{aligned}$$

For $\det(W) < 0$, that is, $|\det(W)| = -\det(W)$, we get:

$$\begin{aligned}\frac{d}{dW}\ln|\det(W)| &= \frac{\partial \ln(-\det(W))}{\partial(-\det(W))} \cdot \frac{d(-\det(W))}{dW} \\ &= \frac{1}{-\det(W)} \cdot (-1) \cdot W^{-T}\det(W) \\ &= W^{-T}\end{aligned}$$

Hence, $\frac{d}{dW}\ln|\det(W)| = W^{-T}$ for $\det(W) \neq 0$.

Exercise 8.2.5

Since we are taking the logarithm to a distribution, we know for sure it will be non-negative. We can also assume that the probability will be greater than zero (albeit infinitely small) since we are searching for signals that we have a probability to observe, hence we can assume $p_i(\mathrm{z}_i) > 0$ and then $\ln p_i(\mathrm{z}_i)$ is integrable. Since $\log p_z(\mathbf{z}_i)$ is integrable we carry out the interchange of expectation and derivative.

$$\begin{aligned}
\frac{d}{dW}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \sum_{i=1}^{l}\frac{\partial \ln p_i(\mathrm{z}_i)}{\partial p_i(\mathrm{z}_i)}\frac{dp_i(\mathrm{z}_i)}{dW} \\
&= \sum_{i=1}^{l}\frac{\partial \ln p_i(\mathrm{z}_i)}{\partial p_i(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}\frac{d\mathrm{z}_i}{dW} \\
&= \sum_{i=1}^{l}\frac{1}{p_i(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}\frac{d\mathrm{z}_i}{dW}
\end{aligned}$$

Exercise 8.2.6

We first observe from the definition of $\boldsymbol{\phi}(\mathbf{z})$, that we have a vector with the i'th element

$$\boldsymbol{\phi}(\mathbf{z})_i = \frac{1}{p(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}$$

Thus we can rewrite

$$\begin{aligned}
\frac{d}{dW}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \sum_{i=1}^{l}\frac{1}{p_i(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}\frac{d\mathrm{z}_i}{dW} \\
&= \sum_{i=1}^{l}\boldsymbol{\phi}(\mathbf{z})_i\frac{d\mathrm{z}_i}{dW}
\end{aligned}$$

At this point it is easiest to consider the component-wise derivative wrt W. A scalar-by-matrix (of size $k \times m$) derivative is defined as

$$\frac{dx}{dA} = \begin{bmatrix}
\frac{dx}{dA_{1,1}} & \frac{dx}{dA_{1,2}} & \cdots & \frac{dx}{dA_{1,m}} \\
\frac{dx}{dA_{2,1}} & \frac{dx}{dA_{2,2}} & \cdots & \frac{dx}{dA_{2,m}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{dx}{dA_{k,1}} & \frac{dx}{dA_{k,2}} & \cdots & \frac{dx}{dA_{k,m}}
\end{bmatrix}$$

Hence if we have an expression for $\frac{d\mathrm{z}_i}{dW_{k,m}}$ we also have $\frac{d\mathrm{z}_i}{dW}$. To get an expression for this derivative, we consider the i'th z_i, where we have (from the ICA model):

$$\begin{aligned}
\mathrm{z}_i = W_{i,:}^T\mathbf{x} &= \sum_{j=1}^{l} W_{i,j}\mathrm{x}_j \\
&= W_{i,1}\mathrm{x}_1 + W_{i,2}\mathrm{x}_2 + \cdots + W_{i,l}\mathrm{x}_l
\end{aligned}$$

We immediately get

$$\begin{aligned}
\frac{d\mathrm{z}_i}{dW_{k,m}} &= \frac{d}{dW_{k,m}}(W_{i,1}\mathrm{x}_1 + W_{i,2}\mathrm{x}_2 + \cdots + W_{i,l}\mathrm{x}_l) \\
&= \mathrm{x}_m \quad \text{only if i = k, otherwise the derivative vanish}
\end{aligned}$$

Which we can use to create the component-wise derivative

$$\begin{aligned}\frac{d}{dW_{k,m}}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \sum_{i=1}^{l}\boldsymbol{\phi}(\mathbf{z})_i\frac{d\mathrm{z}_i}{dW_{k,m}} \\ &= \boldsymbol{\phi}(\mathbf{z})_k\frac{d\mathrm{z}_k}{dW_{k,m}} \\ &= \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_m\end{aligned}$$

Writing the full matrix we get

$$\begin{aligned}\frac{d}{dW}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \begin{bmatrix}\boldsymbol{\phi}(\mathbf{z})_1\mathrm{x}_1 & \boldsymbol{\phi}(\mathbf{z})_1\mathrm{x}_2 & \cdots & \boldsymbol{\phi}(\mathbf{z})_1\mathrm{x}_m \\ \boldsymbol{\phi}(\mathbf{z})_2\mathrm{x}_1 & \boldsymbol{\phi}(\mathbf{z})_2\mathrm{x}_2 & \cdots & \boldsymbol{\phi}(\mathbf{z})_2\mathrm{x}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_1 & \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_2 & \cdots & \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_m\end{bmatrix} \\ &= \boldsymbol{\phi}(\mathbf{z})\mathbf{x}^T\end{aligned}$$

Exercise 8.2.7

Gradient descent (equation 5.3 from the book) is written as:

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} - \mu_i\nabla J(\boldsymbol{\theta}^{(i-1)})$$

In our case, the parameter vector $\boldsymbol{\theta}$ we are optimizing is W, so plugging in we get

$$W^{(i)} = W^{(i-1)} - \mu_i\Big(\Big(\big(W^{(i-1)}\big)^{-1}\Big)^T + \mathbb{E}\big[\boldsymbol{\phi}(\mathbf{z})\mathbf{x}^T\big]\Big)$$

Denote $(W^{-1})^T := W^{-T}$. Our model for $\mathbf{x}$ can now be rewritten as

$$\mathbf{x} = W^{-1}\mathbf{z} \Rightarrow \mathbf{x}^T = \mathbf{z}^T W^{-T}$$

Substituting $\mathbf{x}^T$ then yields the final result:

$$\begin{aligned}W^{(i)} &= W^{(i-1)} - \mu_i\big(W^{(i-1)}\big)^{-T} + \mathbb{E}\Big[\boldsymbol{\phi}(\mathbf{z})\mathbf{z}^T\big(W^{(i-1)}\big)^{-T}\Big] \\ &= W^{(i-1)} - \mu_i\big(W^{(i-1)}\big)^{-T} + \mathbb{E}\big[\boldsymbol{\phi}(\mathbf{z})\mathbf{z}^T\big]\big(W^{(i-1)}\big)^{-T} \\ &= W^{(i-1)} - \mu_i\big(I + \mathbb{E}\big[\boldsymbol{\phi}(\mathbf{z})\mathbf{z}^T\big]\big)\big(W^{(i-1)}\big)^{-T}\end{aligned}$$