1 The Jacobian matrix

In the last recitation, we solved a few problems with gradients, which are defined for functions that have vector input and scalar output. Now consider a function $f : \mathbb{R}^N \to \mathbb{R}^M$ which takes a vector $\mathbf{x} \in \mathbb{R}^N$ and outputs a vector $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^M$. The gradient doesn't make much sense here since there are multiple outputs. In fact, since each output f_i could be a function of each input x_j, there is a partial derivative for each combination of f_i and x_j. This is the intuition for the Jacobian matrix $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$, whose (i,j)th entry is defined to be $\frac{\partial f_i}{\partial x_j}$. Since i refers to the output vector and j refers to the input vector, $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ has shape MxN.

1. To make this idea more concrete, let's find the Jacobian for a few functions. Let $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{MxN}$ and $\mathbf{x} \in \mathbb{R}^N$.

   (a) What is the shape of $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$?

   The shape is MxN because the output has dimension M and the input has dimension N.

   (b) Express f_i in terms of $\mathbf{A}_{i,:}$ (the ith row of $\mathbf{A}$) and $\mathbf{x}$. Write this in summation form as well.

   $f_i = \mathbf{A}_{i,:}\mathbf{x}$. This is equivalent to $\sum_{j=1}^{N} A_{i,j}x_j$.

   (c) What is $\frac{\partial f_i}{\partial x_j}$?

   Since $f_i = \sum_{j=1}^{N} A_{i,j}x_j$, then $\frac{\partial f_i}{\partial x_j} = A_{i,j}$.

   (d) What is $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$? Does this coincide with your intuition based on scalar calculus?

   Since $(\frac{\partial \mathbf{f}}{\partial \mathbf{x}})_{i,j} = \frac{\partial f_i}{\partial x_j} = A_{i,j}$, then $\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \mathbf{A}$.

2. Let $\mathbf{f}(\mathbf{x}) = -\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^N$.

   (a) What is the shape of $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$?

   Since the input and output are both N-dimensional, the Jacobian has shape NxN.

   (b) What is $\frac{\partial f_i}{\partial x_i}$?

   $\frac{\partial f_i}{\partial x_i} = \frac{\partial}{\partial x_i} -x_i = -1$.

   (c) What is $\frac{\partial f_i}{\partial x_j}$ where $i \neq j$?

   $\frac{\partial f_i}{\partial x_j} = \frac{\partial}{\partial x_j} -x_i = 0$.

   (d) What is $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$? Does this coincide with your intuition based on scalar calculus?

   Since the diagonal is all -1's, and everything else is 0, then $\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = -\mathbf{I}$.

3. Let $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^N$.

   (a) What is the shape of $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$?

   Since the output is 1-dimensional and the output is N-dimensional, the Jacobian has shape 1xN. This is a row vector of length N.

   (b) Express f in summation form.

   $f(\mathbf{x}) = \sum_{i=1}^{N} x_i^2$.

   (c) What is $\frac{\partial f}{\partial x_i}$?

   $\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i} x_i^2 = 2x_i$.

   (d) What is $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$?

   Since $(\frac{\partial f}{\partial \mathbf{x}})_{1,i} = \frac{\partial f}{\partial x_i} = 2x_i$, then $\frac{\partial f}{\partial \mathbf{x}} = 2\mathbf{x}^T$.

   (e) What is the gradient of f with respect to $\mathbf{x}$? What does this suggest about the relationship between the gradient and the Jacobian?

   The gradient of f is $2\mathbf{x}$, and the Jacobian is $2\mathbf{x}^T$. This suggests that the gradient is the transpose of the Jacobian.

2 Closed-form solution to linear regression

Armed with our understanding of Jacobian matrices, we will now find the closed-form matrix solution to linear regression using the chain rule. Recall the quantity we are trying to minimize is $J(\mathbf{w}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2$, where $\mathbf{X} \in \mathbb{R}^{N \times M}$. This can be modeled as a composition of three functions:

$$\mathbf{f}_1(\mathbf{w}) = \mathbf{X}\mathbf{w}$$

$$\mathbf{f}_2(\mathbf{f}_1) = \mathbf{y} - \mathbf{f}_1$$

$$f_3(\mathbf{f}_2) = ||\mathbf{f}_2||_2^2$$

$$J(\mathbf{w}) = f_3(\mathbf{f}_2(\mathbf{f}_1(\mathbf{w})))$$

1. Using the chain rule, what is $\frac{\partial J}{\partial \mathbf{w}}$ in terms of the derivatives of the three functions?

   Working from outer to inner, the derivative is $\frac{1}{2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_1}{\partial \mathbf{w}}$.

2. What is $\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_2}$? What is its shape?

   From question 1.3, the answer is $2\mathbf{f}_2^T = 2(\mathbf{y} - \mathbf{X}\mathbf{w})^T$. The shape is 1xN because the output has dimension 1 and the input has dimension N.

3. What is $\frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}$? What is its shape?

   Since $\mathbf{y}$ is a constant, the derivative is 0, and as for $-\mathbf{f}_1$, the derivative is $-\mathbf{I}$ (see question 1.2). The shape is NxN because the output and input both have dimension N.

4. What is $\frac{\partial \mathbf{f}_1}{\partial \mathbf{w}}$? What is its shape?

   From question 1.1, the answer is $\mathbf{X}$. The shape is NxM because the output has dimension N and the input has dimension M.

5. What is $\frac{\partial J}{\partial \mathbf{w}}$? What is its shape?

   $$\frac{1}{2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_1}{\partial \mathbf{w}} = \frac{1}{2}(2(\mathbf{y} - \mathbf{X}\mathbf{w})^T)(-\mathbf{I})(\mathbf{X}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(-1)(\mathbf{I}\mathbf{X}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T\mathbf{X}$$

   The three shapes are (1XN),(NxN),(NxM), so multiplying all these together the shape is 1xM.

6. What is the optimal $\mathbf{w}$?

   We will set $\frac{\partial J}{\partial \mathbf{w}}$ equal to 0:

   $$\begin{aligned}(\mathbf{X}\mathbf{w} - \mathbf{y})^T\mathbf{X} &= 0\\ (\mathbf{w}^T\mathbf{X}^T - \mathbf{y}^T)\mathbf{X} &= 0\\ \mathbf{w}^T\mathbf{X}^T\mathbf{X} - \mathbf{y}^T\mathbf{X} &= 0\\ \mathbf{w}^T\mathbf{X}^T\mathbf{X} &= \mathbf{y}^T\mathbf{X}\\ \mathbf{X}^T\mathbf{X}\mathbf{w} &= \mathbf{X}^T\mathbf{y}\\ \mathbf{w} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

3 Gaussian Distribution

Review: 1-D Gaussian Distribution

The probability density function of $\mathcal{N}(\mu, \sigma^2)$ is given by:

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

Multivariate Gaussian Distribution

The multivariate Gaussian distribution in M dimensions is parameterized by a **mean vector** $\boldsymbol{\mu} \in \mathbb{R}^M$ and a **covariance matrix** $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times M}$, where $\boldsymbol{\Sigma}$ is a symmetric and positive-definite. This distribution is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and its probability density function is given by:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

Let $\boldsymbol{X} = [X_1, X_2, ..., X_m]^T$ be a vector-valued random variable where $\boldsymbol{X} = [X_1, X_2, ..., X_m]^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, we have:

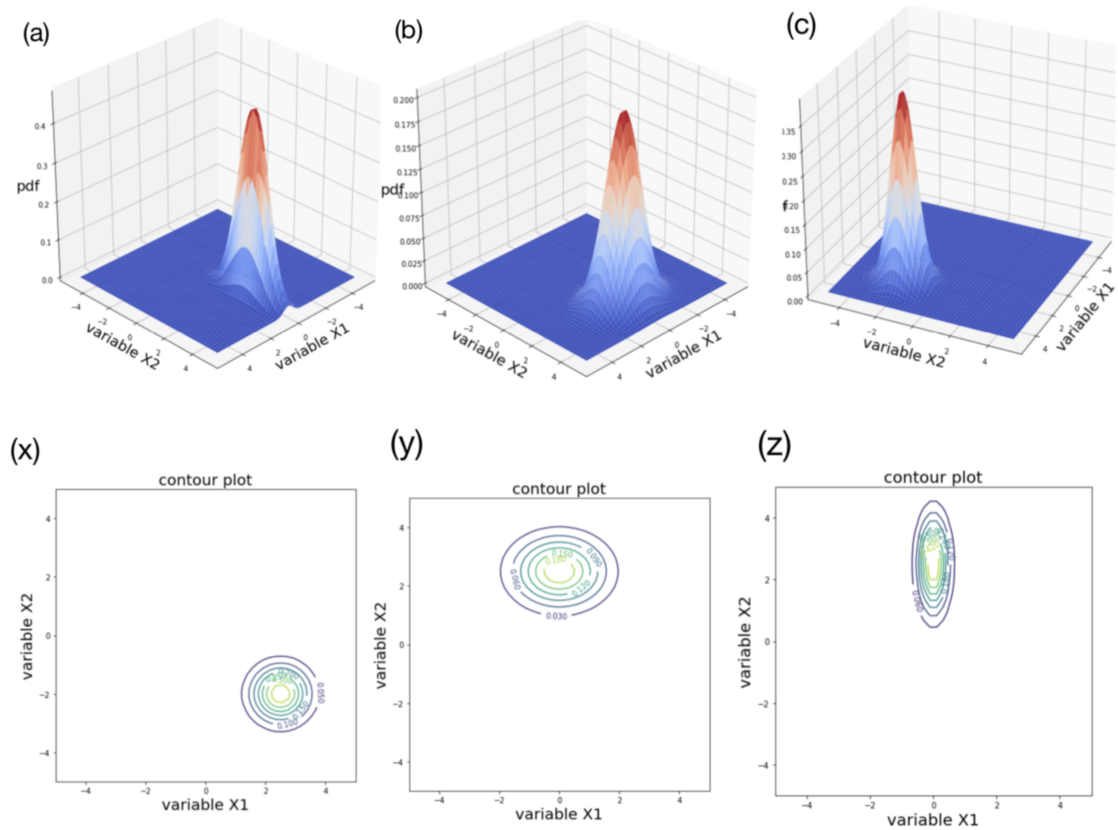
$$\boldsymbol{\Sigma} = Cov[\boldsymbol{X}] = \begin{bmatrix} Cov[X_1, X_1] = Var[X_1] & Cov[X_1, X_2] & \dots & Cov[X_1, X_M] \\ Cov[X_2, X_1] & Cov[X_2, X_2] = Var[X_2] & \dots & Cov[X_2, X_M] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[X_M, X_1] & Cov[X_M, X_2] & \dots & Cov[X_M, X_M] = Var[X_M] \end{bmatrix}$$

Note: Any arbitrary covariance matrix is positive semi-definite. However, since the pdf of a multivariate Gaussian requires Σ to have a strictly positive determinant, Σ has to be positive definite.

In order to get get an intuition for what a multivariate Gaussian is, consider the simple case where $M = 2$. Then, we have:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & Cov[X_1, X_2] \\ Cov[X_1, X_2] & \sigma_2^2 \end{bmatrix}$$

1. For each surface plot, (1) find the corresponding contour plot (2) use the plotting tool provided to find the parameter($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) of the distribution.



(a) → (z)

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 2.5 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}$$

(b) → (y)

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 2.5 \end{bmatrix}$$

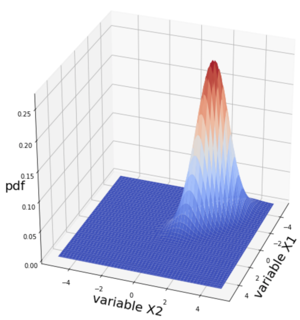
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$$

(c) → (x)

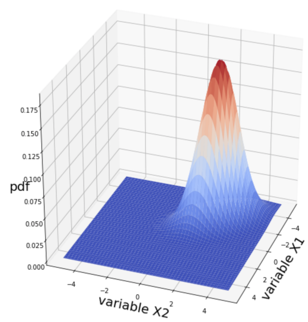
$$\boldsymbol{\mu} = \begin{bmatrix} 2.5 \\ -2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}$$

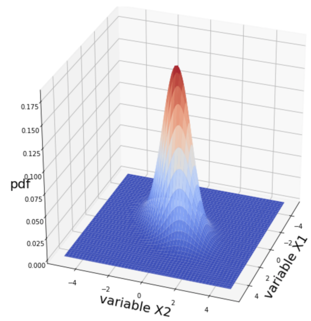
2. For each surface plot, find the corresponding contour plot and the corresponding parameters.



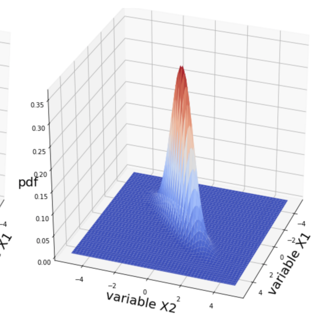
(a)



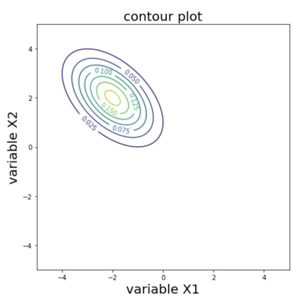
(b)



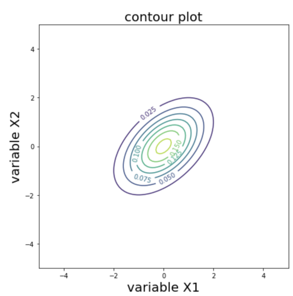
(c)



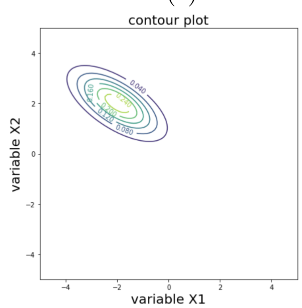
(d)



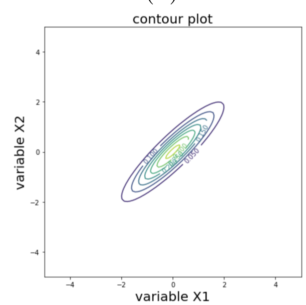
(h)



(i)



(j)



(k)

(x)

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

(y)

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

(z)

$$\boldsymbol{\mu} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

(w)

$$\boldsymbol{\mu} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 0.6 \end{bmatrix}$$

(a)→(j)→(w) (b)→(h)→(z) (c)→(i)→(x) (d)→(k)→(y)