Lecture 13: Transformations of Random Variables

In this lecture we will discuss transformations of random vectors and review some additional properties of random vectors. These concepts are essential for understanding the state estimation approaches (e.g., Kalman filtering) to be presented later.

Linear Transformation of a Gaussian Random Variable

Consider a Gaussian random variable $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and the linear scalar-valued transformation

$$y = ax + b \,. \tag{1}$$

The r.v. is transformed through the above function to yield a new r.v. $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. To find the mean μ_y we take the expected operator of both sides:

$$\begin{aligned} \mu_y = E[y] &= E[ax + b] && (2)\\ &= aE[x] + E[b] && (3)\\ \implies \mu_y &= a\mu_x + b && (4) \end{aligned}$$

and the variance σ_y^2 is similarily obtained by applying the definition of variance to the new r.v. y with (1) and (4):

$$\begin{aligned} \sigma_y^2 &= E[(y - \mu_y)^2] \\ &= E[(ax + b - (a\mu_x + b))^2] \\ &= E[(ax - a\mu_x)^2] \\ &= E[a^2(x - \mu_x)^2] \\ &= a^2 E[(x - \mu_x)^2] \\ &= a^2\sigma_x^2 \\ \implies \sigma_y &= a\sigma_x \end{aligned}$$

Thus, the trasformed r.v. y has a mean that is obtained by propagating $x = \mu_y$ through the function $y = ax + b$ and the variance scales by a factor of a^2.

Aside: In MATLAB the function `randn()` generates a zero-mean and unit variance random number (i.e., from $\mathcal{N}(0, 1)$). We can use the property described above to draw a number from $\mathcal{N}(\mu_y, \sigma_y^2)$ by transforming `randn()` appropriately, for example:

```
mu_y = 2; %  desired mean
var_y = 3; % desired variance
r = randn([100 1]); % a hundred unit variance random numbers
y = sqrt(var_y)*r + mu_y % random numbers with mean mu_y and variance var_y
```

Linear Transformation of a Gaussian Random Vector

Now we revisit the previous example for the case of a random vector instead of a random variable. Consider a Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{P}_x)$ with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\boldsymbol{P}_x \in \mathbb{R}^{n \times n}$ that is transformed through the linear matrix equation:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\ , \tag{5}$$

where $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{P}_y)$ is a new Gaussian random vector and $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{b} \in \mathbb{R}^m$. To find the mean μ_y we take the expected operator of both sides of (5):

$$\begin{aligned} \boldsymbol{\mu_y} = E[\boldsymbol{y}] &= E[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] && (6)\\ &= \boldsymbol{A}E[\boldsymbol{x}] + E[\boldsymbol{b}] && (7)\\ \implies \boldsymbol{\mu_y} &= \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b} && (8) \end{aligned}$$

Similarily, we can obtain the covariance by using the definition of covariance of a random vector and subsituting (5) and (8)

$$\begin{aligned} \boldsymbol{P_y} &= E[\boldsymbol{y}\boldsymbol{y}^T] - \boldsymbol{\mu_y}\boldsymbol{\mu_y}^T && (9)\\ &= E[(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})^T] - (\boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b})(\boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b})^T && (10)\\ &= E[(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})(\boldsymbol{x}^T\boldsymbol{A}^T + \boldsymbol{b}^T)] - (\boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b})(\boldsymbol{\mu}_x^T\boldsymbol{A}^T + \boldsymbol{b}^T) && (11)\\ &= E[(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{x}\boldsymbol{b}^T + \boldsymbol{b}\boldsymbol{x}^T\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T)] - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{b}^T - \boldsymbol{b}\boldsymbol{\mu}_x^T\boldsymbol{A}^T - \boldsymbol{b}\boldsymbol{b}^T && (12)\\ &= \boldsymbol{A}E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{A}^T + \boldsymbol{A}E[\boldsymbol{x}]\boldsymbol{b}^T + \boldsymbol{b}E[\boldsymbol{x}^T]\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{b}^T - \boldsymbol{b}\boldsymbol{\mu}_x^T\boldsymbol{A}^T - \boldsymbol{b}\boldsymbol{b}^T && (13)\\ &= \boldsymbol{A}E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{b}^T + \boldsymbol{b}\boldsymbol{\mu}_x^T\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{b}^T - \boldsymbol{b}\boldsymbol{\mu}_x^T\boldsymbol{A}^T && (14)\\ &= \boldsymbol{A}E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\boldsymbol{A}^T && (15)\\ &= \boldsymbol{A}\left[E[\boldsymbol{x}\boldsymbol{x}^T] - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\right]\boldsymbol{A}^T && (16)\\ \implies \boldsymbol{P_y} &= \boldsymbol{A}\boldsymbol{P_x}\boldsymbol{A}^T && (17) \end{aligned}$$

Aside: We can use the above property to generate a random vector with a desired mean and covariance by transforming a zero-mean and unit-diagonal covariance random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{P}_x)$ by the above linear equation where $\boldsymbol{P}_x = \boldsymbol{1}_{n \times n}$ is the identity matrix. If the target distribution has a covariance $\boldsymbol{P}_y$ then we can obtain the desired transformation by finding a matrix $\boldsymbol{A}$ such that:

$$\begin{aligned} \boldsymbol{P_y} &= \boldsymbol{A}\boldsymbol{P_x}\boldsymbol{A}^T\\ &= \boldsymbol{A}\boldsymbol{1}\boldsymbol{A}^T\\ &= \boldsymbol{A}\boldsymbol{A}^T \end{aligned}$$

The $\boldsymbol{A}$ matrix can be found using Cholsky decomposition (i.e., `chol` in MATLAB).

Transformation of a Random Vector through a Nonlinear Function [1, Sec. 14.1]

Consider a random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}_x$ and a symmetric p.d.f. around the mean (e.g., a Gaussian p.d.f.). Now, suppose this random vector is transformed by a nonlinear function

$$\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) \tag{18}$$

to give a new random vector $\boldsymbol{y}$; we wish to determine its mean and covariance. Begin by expanding (18) as a Taylor series around a nominal value $\boldsymbol{\mu}_x$:

$$\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{\mu}_x) + D_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + \frac{1}{2!}D^2_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + \frac{1}{3!}D^3_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + \cdots \tag{19}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{\mu}_x$ is a new r.v. describing the deviation from the nominal value and the operator $D^k_{\tilde{\boldsymbol{x}}}\ \boldsymbol{f}$:

$$\begin{aligned} D^k_{\tilde{\boldsymbol{x}}}\ \boldsymbol{f} &= \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2} + \cdots + \tilde{x}_n\frac{\partial}{\partial x_n}\right)^k \boldsymbol{f}(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \\ &= \left(\sum_{i=1}^{n}\tilde{x}_i\frac{\partial}{\partial x_i}\right)^k \boldsymbol{f}(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \end{aligned}$$

where the partial derivative applies element-wise to the column vector of functions $\boldsymbol{f}(\boldsymbol{x})$.

Note: In our earlier lecture on linearization of ODEs we saw this operator written in a slightly different form using the Jacobian, however both forms are equivalent:

$$(\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{h})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\tilde{\boldsymbol{x}} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial x_1} & \cdots & \frac{\partial h_n}{\partial x_n} \end{bmatrix}_{\boldsymbol{x}=\boldsymbol{\mu}_x} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix} \tag{20}$$

$$= \begin{bmatrix} \frac{\partial h_1}{\partial x_1}\tilde{x}_1 + \cdots + \frac{\partial h_1}{\partial x_n}\tilde{x}_n \\ \vdots \\ \frac{\partial h_n}{\partial x_1}\tilde{x}_1 + \cdots + \frac{\partial h_n}{\partial x_n}\tilde{x}_n \end{bmatrix}_{\boldsymbol{x}=\boldsymbol{\mu}_x} \tag{21}$$

$$= \tilde{x}_1 \begin{bmatrix} \frac{\partial h_1}{\partial x_1} \\ \vdots \\ \frac{\partial h_n}{\partial x_1} \end{bmatrix}_{\boldsymbol{x}=\boldsymbol{\mu}_x} + \cdots + \tilde{x}_n \begin{bmatrix} \frac{\partial h_1}{\partial x_n} \\ \vdots \\ \frac{\partial h_n}{\partial x_n} \end{bmatrix}_{\boldsymbol{x}=\boldsymbol{\mu}_x} \tag{22}$$

$$= \tilde{x}_1 \left(\frac{\partial}{\partial x_1}\right)\boldsymbol{h}(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} + \cdots + \tilde{x}_n \left(\frac{\partial}{\partial x_n}\right)\boldsymbol{h}(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \tag{23}$$

$$= \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \cdots + \tilde{x}_n\frac{\partial}{\partial x_n}\right)\boldsymbol{h}(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \tag{24}$$

$$= D_{\tilde{\boldsymbol{x}}} \tag{25}$$

By construction, this new r.v. is zero-mean, $E[\tilde{\boldsymbol{x}}] = 0$. To find the mean of the transformed r.v. $\boldsymbol{y}$ take the expected value of both sides:

$$\boldsymbol{\mu}_y \triangleq E[\boldsymbol{y}] = E[\boldsymbol{h}(\boldsymbol{\mu}_x) + D_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + \frac{1}{2!}D^2_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + \frac{1}{3!}D^3_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + \cdots] \tag{26}$$

$$= E[\boldsymbol{h}(\boldsymbol{\mu}_x)] + E[D_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] + E[\frac{1}{2!}D^2_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] + E[\frac{1}{3!}D^3_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] + E[\cdots] \tag{27}$$

Since $h(\boldsymbol{\mu}_x)$ is not a random variable then $E[h(\boldsymbol{\mu}_x)] = h(\boldsymbol{\mu}_x)$ and the first term in (27) can be simplified. To simplify the second term in (27), use the definition of $D_{\tilde{\boldsymbol{x}}}\boldsymbol{h}$ and take its expected value:

$$E[D_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] = E\left[\sum_{i=1}^{n}\tilde{x}_i\frac{\partial}{\partial x_i}\boldsymbol{h}(\boldsymbol{x})\,\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\right] \tag{28}$$

$$= \sum_{i=1}^{n}E[\tilde{x}_i]\frac{\partial}{\partial x_i}\boldsymbol{h}(\boldsymbol{x})\,\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \tag{29}$$

$$= 0 \tag{30}$$

In the above expression we distributed the expected value to only the first term since $\frac{\partial}{\partial x_i}\boldsymbol{h}(\boldsymbol{x})\,\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}$ is a constant when evaluated at the known mean value and we used the fact that $E[\tilde{\boldsymbol{x}}] = 0$. Now, consider the fourth (cubic) term in (27). Again, use the definition of $D_{\tilde{\boldsymbol{x}}}\boldsymbol{h}$ and take the expected value:

$$E[D^3_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] = E\left[\sum_{i=1}^{n}\left[\tilde{x}_i\frac{\partial}{\partial x_i}\right]^3\boldsymbol{h}(\boldsymbol{x})\,\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\right] \tag{31}$$

$$= \sum_{i=1}^{n}E[\tilde{x}_i^3]\left[\frac{\partial}{\partial x_i}\right]^3\boldsymbol{h}(\boldsymbol{x})\,\Big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \tag{32}$$

$$= 0 \tag{33}$$

This follows from the fact that all odd moments are zero (e.g., $E[\tilde{x}_i^3] = 0$ for all $i = 1, \ldots, n$) for a pdf that is zero-mean and symmetrically distributed (i.e., $f_x(-x) = f_x(x)$). For a proof of this fact see the Appendix and [1, p.58-59]. Thus, the Taylor series (27) only contains even terms:

$$\boldsymbol{\mu}_y \triangleq E[\boldsymbol{y}] = E[\boldsymbol{h}(\boldsymbol{\mu})] + E[\frac{1}{2!}D^2_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] + E[\frac{1}{4!}D^4_{\tilde{\boldsymbol{x}}}\boldsymbol{h}] + E[\cdots] \tag{34}$$

We have shown that if we simply propagate the mean of $\boldsymbol{x}$ (i.e., $\boldsymbol{\mu}_x$) through the nonlinear function to give an approximate mean $\boldsymbol{\mu}_y = \boldsymbol{h}(\boldsymbol{\mu}_x)$ then we will be accurate only to first order. To improve accuracy we can evaluate the second order term in the Taylor series, and so on.

First-Order Nonlinear Transformation of a Gaussian Random Vector

We have just shown above that the Taylor series expansion of the expected value of a random vector contains only even terms (34). Here we go back to the nonlinear transformation we started with (19) but only consider the first two terms (i.e., a linear approximation):

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{\mu}_x) + D_{\tilde{\boldsymbol{x}}}\boldsymbol{h} \\ &= \boldsymbol{h}(\boldsymbol{\mu}_x) + (\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{h})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\tilde{\boldsymbol{x}}\end{aligned}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{\mu}_x$. This expression is in the form (5) with $\boldsymbol{A} = (\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{h})\big|_{\boldsymbol{x}}$ and $\boldsymbol{b} = \boldsymbol{h}(\boldsymbol{\mu}_x)$. Thus, for the first-order linear approximation we can use our previous results for transformations of linear variables (8) and (17). That is, to first-order, the mean and covariance of the transformed random vector are:

$$\boldsymbol{\mu}_y = \boldsymbol{h}(\boldsymbol{\mu}_x) \tag{35}$$

$$\boldsymbol{P}_y = \left[(\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{h})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\right] \boldsymbol{P}_x \left[(\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{h})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\right]^{\mathrm{T}} \tag{36}$$

Example: First-order Error Propagation (1D case). Suppose we have a random variable $X \sim \mathcal{N}(\mu_x = 9, \sigma_x^2 = 0.2)$ that is transformed through the nonlinear function $z = f(x) = \sqrt{x}$ into a new random variable $Z \sim \mathcal{N}(\mu_z, \sigma_z^2)$. We wish to determine μ_z and σ_z^2.

Solution. The mean μ_z is found by propagating the mean μ_x through the transformation.

$$\mu_z = f(\mu_x) = \sqrt{9} = 3$$

The first-order error propagation law (36) reduces in the 1D case to:

$$\sigma_z^2 = \left[J_f\big|_{x=\mu_x}\right]^2 \sigma_x^2$$

where J_f is the derivative of $f(x)$ with respect to x. Thus

$$J = \frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$$

Evaluating with the mean μ_x

$$J_f\big|_{x=\mu_x} = \frac{1}{2\sqrt{\mu_x}} = \frac{1}{2\sqrt{9}} = \frac{1}{6}$$

Finally, applying the 1D error-propagation law:

$$\sigma_z^2 = \left[J_f\big|_{x=\mu_x}\right]^2 \sigma_x^2 = \left[\frac{1}{6}\right]^2 (0.2) = 0.00556$$

Thus, the random variable Y is characterized as $Z \sim \mathcal{N}(3, 0.00556)$.

The following example illustrates the same principle for a two-dimensional Gaussian random vector.

Example: First-order Error Propagation (2D case).
Given the random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{P}_x)$ where

$$\boldsymbol{\mu}_x = \begin{bmatrix} 16 \\ 30 \end{bmatrix} \qquad \boldsymbol{P}_x = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

find the mean $\boldsymbol{\mu}_z$ and covariance $\boldsymbol{P}_z$ which characterizes the random variable $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{P}_z)$ (to first-order) that results from the transformation

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{x})$$
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} x_1^{1/4} + x_2 \\ x_1 x_2 / 10 \end{bmatrix}$$

Solution. The mean $\boldsymbol{\mu}_z$ is obtained by propagating the mean $\boldsymbol{\mu}_x$ through the system:

$$\boldsymbol{\mu}_z = \boldsymbol{f}(\boldsymbol{\mu}_x) = \begin{bmatrix} 16^{1/4} + 30 \\ 30(16)/10 \end{bmatrix} = \begin{bmatrix} 32 \\ 48 \end{bmatrix}$$

Next, compute the Jacobian

$$\mathbf{J}_{\boldsymbol{f}} = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 \end{bmatrix} = \begin{bmatrix} \left[\frac{1}{4}\right] x_1^{-3/4} & 1 \\ x_2/10 & x_1/10 \end{bmatrix}$$

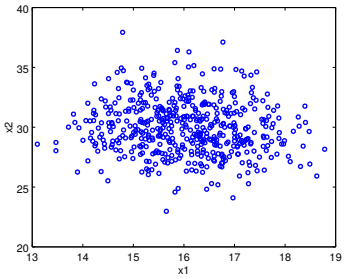
and evaluate the Jacobian at the mean $\boldsymbol{\mu}_x$

$$(\boldsymbol{J}_x \boldsymbol{f})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x} = \begin{bmatrix} \left[\frac{1}{4}\right] (16)^{-3/4} & 1 \\ 30/10 & 16/10 \end{bmatrix} = \begin{bmatrix} 0.03125 & 1 \\ 3 & 1.6 \end{bmatrix}$$

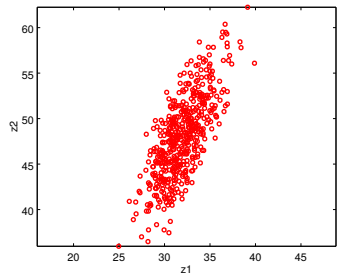
We have all the necessary elements to compute the covariance $\boldsymbol{P}_z$ using the error-propagation law:

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{z}} &= \left[(\boldsymbol{J}_x \boldsymbol{f})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\right] \boldsymbol{P}_x \left[(\boldsymbol{J}_x \boldsymbol{f})\big|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\right]^{\mathrm{T}} \\ &= \begin{bmatrix} 0.03125 & 1 \\ 3 & 1.6 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 0.03125 & 1 \\ 3 & 1.6 \end{bmatrix}^{\mathrm{T}} \\ &= \begin{bmatrix} 5.001 & 8.094 \\ 8.094 & 21.800 \end{bmatrix} \end{aligned}$$

The solution can be visualized numerically by generating samples from $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{P}_x)$



and propagating them through the system



Joint Gaussian Vector Distributions

Suppose that $\boldsymbol{x} \sim (\boldsymbol{\mu}_x, \boldsymbol{Q})$ is a Gaussian random vector taking values in $\mathbb{R}^n$ and $\boldsymbol{y} \sim (\boldsymbol{\mu}_y, \boldsymbol{R})$ is a Gaussian random vector taking values in $\mathbb{R}^m$. Then the stacked random vector $\boldsymbol{z} = [\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}$ jointly describes both random vectors (i.e., $p(\boldsymbol{x}, \boldsymbol{y})$). The mean is, of course,

$$E[\boldsymbol{z}] = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \tag{37}$$

and the covariance is

$$\begin{aligned}
\boldsymbol{P}_z &= E\left[\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}\right)\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}\right)^{\mathrm{T}}\right] && (38)\\
&= E\left[\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}\right)\left(\begin{bmatrix} \boldsymbol{x}^{\mathrm{T}} & \boldsymbol{y}^{\mathrm{T}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x^{\mathrm{T}} & \boldsymbol{\mu}_y^{\mathrm{T}} \end{bmatrix}\right)\right] && (39)\\
&= E\left[\left(\begin{bmatrix} \boldsymbol{x} - \boldsymbol{\mu}_x \\ \boldsymbol{y} - \boldsymbol{\mu}_y \end{bmatrix}\right)\left(\begin{bmatrix} (\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_x^{\mathrm{T}}) & (\boldsymbol{y}^{\mathrm{T}} - \boldsymbol{\mu}_y^{\mathrm{T}}) \end{bmatrix}\right)\right] && (40)\\
&= E\left[\left(\begin{bmatrix} (\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^{\mathrm{T}} & (\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^{\mathrm{T}} \\ (\boldsymbol{y} - \boldsymbol{\mu}_y)(\boldsymbol{y} - \boldsymbol{\mu}_y)^{\mathrm{T}} & (\boldsymbol{y} - \boldsymbol{\mu}_y)(\boldsymbol{y} - \boldsymbol{\mu}_y)^{\mathrm{T}} \end{bmatrix}\right)\right] && (41)\\
&= \begin{bmatrix} \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{x}) & \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) \\ \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x}) & \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{y}) \end{bmatrix} && (42)\\
&= \begin{bmatrix} \boldsymbol{Q} & \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) \\ \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x}) & \boldsymbol{R} \end{bmatrix} && (43)
\end{aligned}$$

The terms $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ and $\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x})$ describe the covariance of one random vector with the other. If they are entirely independent then these are zero matrices.

Conditional Gaussian Vector Distributions

Having described the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ above we now consider the conditional distributions (e.g., $p(\boldsymbol{x}|\boldsymbol{y})$ which is the p.d.f. of $\boldsymbol{x}$ given a particular value of $\boldsymbol{y}$). Consider again the case above of a stacked random vector $\boldsymbol{z} = [\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}$ with the mean partitioned as

$$E[\boldsymbol{z}] = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}. \tag{44}$$

Denote the partitioned covariance matrix (45) as

$$\boldsymbol{P}_z = \begin{bmatrix} \boldsymbol{P}_x & \boldsymbol{P}_{xy} \\ \boldsymbol{P}_{yx} & \boldsymbol{P}_y \end{bmatrix} \tag{45}$$

The symmetry of $\boldsymbol{P}_z$ implies that $\boldsymbol{P}_{xy} = \boldsymbol{P}_{yx}^{\mathrm{T}}$. In many situations, it is convenient to work with the inverse of the covariance matrix, called the *precision matrix*, which we denote as:

$$\boldsymbol{\Lambda}_z := \boldsymbol{P}_z^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy} \\ \boldsymbol{\Lambda}_{yx} & \boldsymbol{\Lambda}_{yy} \end{bmatrix} \tag{46}$$

Note however, that $\boldsymbol{\Lambda}_{xx}$ is a subblock of the precision matrix but $\boldsymbol{\Lambda}_{xx} \neq \boldsymbol{P}_{xx}^{-1}$. To obtain the mean and covariance as a function of the original covariance matrix (45) we can use the matrix identity associated with Schur's complement.

Aside (Schur's complement): Consider a matrix $\boldsymbol{P}$ that can be partitioned into four sub-blocks as

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}. \tag{47}$$

The inverse of this matrix is given by the following identity:

$$\boldsymbol{P}^{-1} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{B} & (\boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1}) \end{bmatrix} \tag{48}$$

where the new matrix is introduced

$$\boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \tag{49}$$

Applying the above inverse identity to (45) and (46) it follows that

$$\boldsymbol{\Lambda}_{xx} = (\boldsymbol{P}_{xx} - \boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}\boldsymbol{P}_{yx})^{-1} \tag{50}$$

$$\boldsymbol{\Lambda}_{xy} = -(\boldsymbol{P}_{xx} - \boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}\boldsymbol{P}_{yx})^{-1}\boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1} \tag{51}$$

The quantity $\boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy}$ that appears in the mean then simplifies as

$$\boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy} = -(\boldsymbol{P}_{xx} - \boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}\boldsymbol{P}_{yx})[(\boldsymbol{P}_{xx} - \boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}\boldsymbol{P}_{yx})^{-1}\boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}] \tag{52}$$

$$= -\boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1} \tag{53}$$

so an equivalent expression for the mean and covariance of the condition probability is

$$\implies \boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{54}$$

$$\implies \boldsymbol{P}_{x|y} = \boldsymbol{P}_{xx} - \boldsymbol{P}_{xy}\boldsymbol{P}_{yy}^{-1}\boldsymbol{P}_{yx} \tag{55}$$

Marginal Gaussian Vector Distributions

For the partitioned random vector $\boldsymbol{z} = [\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}$ we have thusfar defined the the joint distribution $\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y})$ as well as the condition distributions $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{y})$ and $\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$. The marginal distributions for each component are defined by the integrals

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \tag{56}$$

$$p(\boldsymbol{y}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} \tag{57}$$

where we "integrate out" the undesired variables. It can be shown (see [2, Sec 2.3.2]) that the marginal distribution is also Gaussian, and is intuitively given by

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{P}_{xx}) \tag{58}$$

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{P}_{yy}) \tag{59}$$

Product of Two Gaussians

Consider two indpendent distributions $\mathcal{N}(\boldsymbol{a}, \boldsymbol{A})$ and $\mathcal{N}(\boldsymbol{b}, \boldsymbol{B})$ defined for random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, of the same size. For a given $\boldsymbol{x}$ or $\boldsymbol{y}$ each of the p.d.f.s evaluate to a scalar value and the product of these two density values yields a new density that is also Gaussian. This new distribution is $Z^{-1}\mathcal{N}(\boldsymbol{c}, \boldsymbol{C})$ where

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \tag{60}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}) \tag{61}$$

and the normalizing constant is

$$Z^{-1} = (2\pi)^{-n/2}|\boldsymbol{A} + \boldsymbol{B}|^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^{\mathrm{T}}(\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})) \tag{62}$$

Appendix: Linearization Review

Consider the vector-valued nonlinear function $\boldsymbol{f}(\cdot)$ with a scalar argument x. We can approximate the function $\boldsymbol{f}(x)$ around some nominal value $x = \bar{x}$ by defining a Taylor series. Defining the deviation from the nominal point as $\tilde{x} = x - \bar{x}$, the Taylor series expansion of $f(x)$ is:

$$\begin{aligned}\boldsymbol{f}(x) &= \boldsymbol{f}(\bar{x}) + \frac{d\boldsymbol{f}}{dx}\bigg|_{\bar{x}} \tilde{x} + \frac{1}{2!}\frac{d^2\boldsymbol{f}}{dx^2}\bigg|_{\bar{x}} \tilde{x}^2 + \frac{1}{3!}\frac{d^3\boldsymbol{f}}{dx^3}\bigg|_{\bar{x}} \tilde{x}^3 + \cdots \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\frac{d^{(k)}\boldsymbol{f}}{dx^{(k)}}(\tilde{x})^k\bigg|_{x=\bar{x}}\end{aligned}$$

Now suppose the function argument $\boldsymbol{x}$ is a vector. For example, if $\boldsymbol{x}$ is a 2x1 vector, then the Taylor series expansion is:

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\bar{\boldsymbol{x}}) &+ \left(\frac{\partial \boldsymbol{f}}{\partial x_1}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_1 + \frac{\partial \boldsymbol{f}}{\partial x_2}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_2\right) + \frac{1}{2!}\left(\frac{\partial^2 \boldsymbol{f}}{\partial x_1^2}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_1^2 + \frac{\partial^2 \boldsymbol{f}}{\partial x_2^2}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_2^2 + 2\frac{\partial^2 \boldsymbol{f}}{\partial x_1 x_2}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_1\tilde{x}_2\right) \\ &+ \frac{1}{3!}\left(\frac{\partial^3 \boldsymbol{f}}{\partial x_1^3}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_1^3 + \frac{\partial^3 \boldsymbol{f}}{\partial x_2^3}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_2^3 + 3\frac{\partial^3 \boldsymbol{f}}{\partial x_1^2 x_2}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_1^2\tilde{x}_2 + 3\frac{\partial^3 \boldsymbol{f}}{\partial x_1 x_2^2}\bigg|_{\bar{\boldsymbol{x}}} \tilde{x}_1\tilde{x}_2^2\right) + \cdots \\ = \boldsymbol{f}(\bar{\boldsymbol{x}}) &+ \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right)\boldsymbol{f}\bigg|_{\bar{\boldsymbol{x}}} + \frac{1}{2!}\left(\tilde{x}_1^2\frac{\partial^2}{\partial x_1^2} + \tilde{x}_2^2\frac{\partial^2}{\partial x_2^2} + \tilde{x}_1\tilde{x}_2\frac{\partial^2}{\partial x_1 x_2}\right)\boldsymbol{f}\bigg|_{\bar{\boldsymbol{x}}} \\ &+ \frac{1}{3!}\left(\tilde{x}_1^3\frac{\partial^3}{\partial x_1^3} + \tilde{x}_3^2\frac{\partial^3}{\partial x_3^2} + \tilde{x}_1^2\tilde{x}_2\frac{\partial^3}{\partial x_1^2 x_2} + \tilde{x}_1\tilde{x}_2^2\frac{\partial^3}{\partial x_1 x_2^2}\right)\boldsymbol{f}\bigg|_{\bar{\boldsymbol{x}}} + \cdots\end{aligned}$$

Define the operator $D^k_{\tilde{\boldsymbol{x}}}\ \boldsymbol{f}$

$$\begin{aligned}D^k_{\tilde{\boldsymbol{x}}}\ \boldsymbol{f} &= \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2} + \cdots + \tilde{x}_n\frac{\partial}{\partial x_n}\right)^k \boldsymbol{f}(\boldsymbol{x})\bigg|_{\boldsymbol{x}=\boldsymbol{\mu}_x} \\ &= \left(\sum_{i=1}^{n} \tilde{x}_i\frac{\partial}{\partial x_i}\right)^k \boldsymbol{f}(\boldsymbol{x})\bigg|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\end{aligned}$$

Then using this definition we may write the Taylor expansion for any sized vector argument $\boldsymbol{x}$ as

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}) &= \boldsymbol{f}(\bar{\boldsymbol{x}}) + D_{\tilde{\boldsymbol{x}}}\boldsymbol{f} + \frac{1}{2!}D^2_{\tilde{\boldsymbol{x}}}\boldsymbol{f} + \frac{1}{3!}D^3_{\tilde{\boldsymbol{x}}}\boldsymbol{f} + \cdots \\ &= \sum_{k=0}^{\infty}\frac{1}{k!}D^k_{\tilde{\boldsymbol{x}}}\ \boldsymbol{f}\end{aligned} \tag{63}$$

Appendix: Odd moments of a symmetric probability distribution are zero

Suppose there is a random variable X with mean of zero $E[X] = 0$, and with a symmetric probability density function (pdf): $p(x) = p(-x)$. The i-th moment of X can be written

$$\begin{aligned}m_i &= E[X^i] \\ &= \int_{-\infty}^{\infty} x^i p(x)dx \\ &= \int_{-\infty}^{0} x^i p(x)dx + \int_{0}^{\infty} x^i p(x)dx\end{aligned}$$

Changing the limits of the first term:

$$\int_{-\infty}^{0} x^i p(x)dx = -\int_{0}^{-\infty} x^i p(x)dx$$

Then substituting $u = -x$, the right hand side becomes:

$$\begin{aligned} -\int_{0}^{-\infty} x^i p(x)dx &= \int_{0}^{\infty} (-u)^i p(-u)du \\ &= (-1)^i \int_{0}^{\infty} u^i p(u)du \end{aligned}$$

Thus

$$m_i = (-1)^i \underbrace{\int_{0}^{\infty} u^i p(u)du}_{A} + \underbrace{\int_{0}^{\infty} x^i p(x)dx}_{A}$$

and it is clear that the moment is zero for odd i.

Appendix: Conditional Gaussian Distribution

Recall that the multivariate p.d.f. is given by

$$f_z(\boldsymbol{z}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_z)^{\mathrm{T}}\boldsymbol{P}_z^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right)}{(2\pi)^{n/2}\sqrt{|\boldsymbol{P}_z|}} \tag{64}$$

First we establish the fact that, for a general Gaussian random vector $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{P}_z)$ the argument of the exponent in (64) has components that are quadratic in $\boldsymbol{z}$, linear in $\boldsymbol{z}$, and independent of $\boldsymbol{z}$:

$$-\frac{1}{2}\Delta = -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_z)^{\mathrm{T}}\boldsymbol{P}_z^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z) \tag{65}$$

$$= -\frac{1}{2}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{z} - \boldsymbol{z}^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{\mu}_z - \boldsymbol{\mu}_z^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{z} + \boldsymbol{\mu}_z^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{\mu}_z \tag{66}$$

$$= \underbrace{-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{z}}_{\text{quadratic in } \boldsymbol{z}} + \underbrace{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{\mu}_z}_{\text{linear in } \boldsymbol{z}} - \frac{1}{2}\underbrace{\boldsymbol{\mu}_z^{\mathrm{T}}\boldsymbol{P}_z^{-1}\boldsymbol{\mu}_z}_{\text{independent of } \boldsymbol{z}} \tag{67}$$

Recall that Δ is the Mahalanobis distance between $\boldsymbol{z}$ and $\boldsymbol{\mu}$. Our strategy to find the conditional probability is to expand this expression with the components of the stacked vector $\boldsymbol{z}$ while treating $\boldsymbol{y}$ as a fixed quantity and equating terms quadratic, linear, and independent of $\boldsymbol{x}$ to (67).

Expand the Mahalanobis distance first:

$$-\frac{1}{2}\Delta = (\boldsymbol{z}-\boldsymbol{\mu}_z)^{\mathrm{T}}\boldsymbol{P}_z^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z) \tag{68}$$

$$= -\frac{1}{2}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}\right)^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy}\\ \boldsymbol{\Lambda}_{yx} & \boldsymbol{\Lambda}_{yy}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}\right) \tag{69}$$

$$= -\frac{1}{2}\begin{bmatrix}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}} & (\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy}\\ \boldsymbol{\Lambda}_{yx} & \boldsymbol{\Lambda}_{yy}\end{bmatrix}\begin{bmatrix}(\boldsymbol{x}-\boldsymbol{\mu}_x)\\ (\boldsymbol{y}-\boldsymbol{\mu}_y)\end{bmatrix} \tag{70}$$

$$= -\frac{1}{2}\begin{bmatrix}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}} & (\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{xx}(\boldsymbol{x}-\boldsymbol{\mu}_x)+\boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y)\\ \boldsymbol{\Lambda}_{yx}(\boldsymbol{x}-\boldsymbol{\mu}_x)+\boldsymbol{\Lambda}_{yy}(\boldsymbol{y}-\boldsymbol{\mu}_y)\end{bmatrix} \tag{71}$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}(\boldsymbol{\Lambda}_{xx}(\boldsymbol{x}-\boldsymbol{\mu}_x)+\boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y)) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}(\boldsymbol{\Lambda}_{yx}(\boldsymbol{x}-\boldsymbol{\mu}_x)+\boldsymbol{\Lambda}_{yy}(\boldsymbol{y}-\boldsymbol{\mu}_y)) \tag{72}$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}(\boldsymbol{x}-\boldsymbol{\mu}_x) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\boldsymbol{\Lambda}_{yx}(\boldsymbol{x}-\boldsymbol{\mu}_x) \tag{73}$$

$$- \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\boldsymbol{\Lambda}_{yy}(\boldsymbol{y}-\boldsymbol{\mu}_y) \tag{74}$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}(\boldsymbol{x}-\boldsymbol{\mu}_x) - (\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\boldsymbol{\Lambda}_{yy}(\boldsymbol{y}-\boldsymbol{\mu}_y) \tag{75}$$

$$= -\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x - \frac{1}{2}\boldsymbol{\mu}_x^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{y} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{\mu}_y + \boldsymbol{\mu}_x^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{y} - \boldsymbol{\mu}_x^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{\mu}_y \tag{76}$$

$$- \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\boldsymbol{\Lambda}_{yy}(\boldsymbol{y}-\boldsymbol{\mu}_y) \tag{77}$$

$$= \underbrace{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{x}}_{\text{quadratic in } \boldsymbol{x}} \underbrace{+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{y} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{\mu}_y}_{\text{linear in } \boldsymbol{x}} \tag{78}$$

$$\underbrace{-\frac{1}{2}\boldsymbol{\mu}_x^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x + \boldsymbol{\mu}_x^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{y} - \boldsymbol{\mu}_x^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{\mu}_y - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_y)^{\mathrm{T}}\boldsymbol{\Lambda}_{yy}(\boldsymbol{y}-\boldsymbol{\mu}_y)}_{\text{independent of } \boldsymbol{x}} \tag{79}$$

Comparing the quadratic term of (67) with (79) it is evident that the covariance of the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$ is given by

$$\implies \boldsymbol{P}_{x|y} = \boldsymbol{\Lambda}_{xx}^{-1}\,. \tag{80}$$

Also, the linear term in (79) can be written as

$$\begin{aligned}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{y} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{\mu}_y &= \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x - \boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y))\\ &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{xx}(\boldsymbol{\mu}_x - \boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y))\end{aligned}$$

and comparing with (79) it follows that

$$\implies \boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x - \boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy}(\boldsymbol{y}-\boldsymbol{\mu}_y)\,. \tag{81}$$

Thus, we can conclude that the condition probability distribution is characterized by

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{P}_{x|y}) \tag{82}$$

References

[1] Dan Simon. *Optimal State Estimation: Kalman, H infinity, and Nonlinear Approaches*. John Wiley & Sons, 2006.

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*. Springer, 2006.