Introduction and learning objectives

Many widely used models, including diffusion models and variational autoencoders (VAEs), are built on properties of the multivariate Gaussian; for example, the formulas in Problem 2 give the distribution of the noisy data in a diffusion model. We cover VAEs in Week 11 and diffusion models in Week 12.

This tutorial works with the multivariate Gaussian from Lecture 2. We first find the distribution of some of its variables given the rest, then use the result for a Gaussian prior with a linear-Gaussian likelihood to derive the Bayesian linear regression posterior, which Lecture 2 stated without derivation. Both results are in this week's reading (PRML Sections 2.3.1 and 2.3.3).

By the end of the tutorial, you should be able to:

- Identify the mean and covariance of a Gaussian from the quadratic and linear terms in its exponent.
- Expand the exponent of a partitioned Gaussian in terms of the blocks of the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$.
- Derive the mean and covariance of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ by matching these terms.
- Match Bayesian linear regression to the linear-Gaussian model and derive the posterior mean $\mathbf{m}_N$ and covariance $\mathbf{S}_N$ for a general Gaussian prior.

1 Multivariate Gaussian Distribution

Assumptions. The densities and inverses in the questions require positive definite covariance matrices. In Problem 2, take $\mathbf{S}_0$ positive definite and $\beta > 0$, with independent observation noise given the weights and inputs. These assumptions ensure that the posterior covariance is well defined even when $\mathbf{X}$ does not have full column rank.

Problem 1. *Expanding the exponent.* Since $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda}$, the exponent is $z = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})$. Substitute the partitioned forms

$$\mathbf{x} - \boldsymbol{\mu} = \begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix},$$

and multiply out the blocks:

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) + (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ + (\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) + (\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b).\end{aligned}$$

The precision matrix is symmetric, because $\boldsymbol{\Sigma}$ is, so $\boldsymbol{\Lambda}_{ba} = \boldsymbol{\Lambda}_{ab}^\top$. The third term is a scalar, so it equals its own transpose:

$$(\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) = (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ba}^\top(\mathbf{x}_b - \boldsymbol{\mu}_b) = (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b).$$

The two cross terms are therefore equal, which gives the expression in the question:

$$z = -\frac{1}{2}\Big((\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) + 2(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) + (\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big).$$

Reading off the conditional. By the product rule, $p(\mathbf{x}_a|\mathbf{x}_b) = p(\mathbf{x}_a, \mathbf{x}_b)/p(\mathbf{x}_b)$. With $\mathbf{x}_b$ fixed, the denominator is a constant, so as a function of $\mathbf{x}_a$, $p(\mathbf{x}_a|\mathbf{x}_b) \propto \exp(z)$. We therefore collect the terms of z that depend on $\mathbf{x}_a$. They form a quadratic function of $\mathbf{x}_a$ whose quadratic part is $-\frac{1}{2}\mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\mathbf{x}_a$, and $\boldsymbol{\Lambda}_{aa}$ is positive definite (a diagonal block of the positive definite matrix $\boldsymbol{\Lambda}$), so it is invertible. The exponential of such a quadratic is proportional to a Gaussian density, so $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian, and we can find its parameters by matching coefficients. Expanding the first term, $(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) = \mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\mathbf{x}_a - 2\mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \text{const}$, so

$$\begin{aligned} \text{quadratic in } \mathbf{x}_a: &\quad -\tfrac{1}{2}\mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\mathbf{x}_a, \\ \text{linear in } \mathbf{x}_a: &\quad -\tfrac{1}{2}\Big(-2\mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + 2\mathbf{x}_a^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big) = \mathbf{x}_a^\top \Big(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big). \end{aligned}$$

A Gaussian $\mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ has exponent

$$-\tfrac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b})^\top \boldsymbol{\Sigma}_{a|b}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b}) = -\tfrac{1}{2}\mathbf{x}_a^\top \boldsymbol{\Sigma}_{a|b}^{-1}\mathbf{x}_a + \mathbf{x}_a^\top \boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b} + \text{const}.$$

Matching the quadratic terms gives $\boldsymbol{\Sigma}_{a|b}^{-1} = \boldsymbol{\Lambda}_{aa}$. Matching the linear terms gives $\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b} = \boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$, and multiplying on the left by $\boldsymbol{\Lambda}_{aa}^{-1}$:

$$\boxed{\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}, \qquad \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b).}$$

The conditional mean is an affine function of $\mathbf{x}_b$, and its covariance does not depend on $\mathbf{x}_b$. PRML Section 2.3.1 gives the same result and rewrites it in terms of the blocks of $\boldsymbol{\Sigma}$: $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$ and $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$.

Check with $D = 2$. Take $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ with $|\rho| < 1$. Then $\boldsymbol{\Lambda} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$, so $\Lambda_{aa} = \frac{1}{1-\rho^2}$ and $\Lambda_{ab} = \frac{-\rho}{1-\rho^2}$. The formulas give $\Sigma_{a|b} = 1 - \rho^2$ and $\mu_{a|b} = \rho\, x_b$: observing x_b shifts the mean of x_a by $\rho\, x_b$ and reduces its variance from 1 to $1 - \rho^2$.

Problem 2. *Matching.* The Bayesian linear regression model has the same form as the linear-Gaussian model in the question, with the weights $\mathbf{w}$ as the unknown variable. Note that in the question's result, $\mathbf{x}$ is the unknown variable, not an input. The prior does not depend on the inputs, $p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$, so we can apply the result with $\mathbf{X}$ held fixed. The matching is

Linear-Gaussian model	Bayesian linear regression	
$\mathbf{x}$ (unknown)	$\mathbf{w}$	
$\mathbf{y}$ (observed)	$\mathbf{t}$	
$\boldsymbol{\mu}$, $\boldsymbol{\Lambda}^{-1}$ (prior)	$\mathbf{m}_0$, $\mathbf{S}_0$	so $\boldsymbol{\Lambda} = \mathbf{S}_0^{-1}$
$\mathbf{A}$, $\mathbf{b}$	$\mathbf{X}$, $\mathbf{0}$	mean $\mathbf{Xw}$
$\mathbf{L}^{-1}$ (noise covariance)	$\beta^{-1}\mathbf{I}$	so $\mathbf{L} = \beta\mathbf{I}$

Posterior covariance. Substitute into $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$:

$$\mathbf{S}_N = \big(\mathbf{S}_0^{-1} + \mathbf{X}^\top(\beta\mathbf{I})\mathbf{X}\big)^{-1} \quad \Longrightarrow \quad \boxed{\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X}.}$$

Posterior mean. Substitute into $\boldsymbol{\Sigma}\{\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}$:

$$\mathbf{m}_N = \mathbf{S}_N\big\{\mathbf{X}^\top(\beta\mathbf{I})(\mathbf{t} - \mathbf{0}) + \mathbf{S}_0^{-1}\mathbf{m}_0\big\} \quad \Longrightarrow \quad \boxed{\mathbf{m}_N = \mathbf{S}_N\big(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{t}\big).}$$

Interpretation. The posterior precision is the prior precision $\mathbf{S}_0^{-1}$ plus the precision contributed by the data, $\beta\mathbf{X}^\top\mathbf{X}$. If $\mathbf{X}^\top\mathbf{X}$ is invertible, let $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$ be the least-squares solution; then $\beta\mathbf{X}^\top\mathbf{t} = \beta\mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}}$ and

$$\mathbf{m}_N = \big(\mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X}\big)^{-1}\big(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{X}\,\hat{\mathbf{w}}\big),$$

a matrix-weighted average of the prior mean and the least-squares solution: the weights $\mathbf{S}_N\mathbf{S}_0^{-1}$ and $\beta\mathbf{S}_N\mathbf{X}^\top\mathbf{X}$ come from the two precisions and sum to $\mathbf{I}$. If the prior becomes very broad ($\mathbf{S}_0^{-1} \to \mathbf{0}$), then $\mathbf{m}_N \to \hat{\mathbf{w}}$.

With Lecture 2's prior, $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$, the formulas become $\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{X}^\top\mathbf{X}$ and $\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{X}^\top\mathbf{t}$, the expressions in Lecture 2 (with $\boldsymbol{\Phi}$ in place of $\mathbf{X}$).