Mathematical Exercises

Exercise 1: Variance of the Sum of Two Random Variables

We want to show that for two random variables X and Y we have:

$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\cdot\mathrm{Cov}(X, Y)$

Using the definitions $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ and $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$, and linearity of expectation:

$$\begin{aligned}\mathrm{Var}(X + Y) &= E[(X + Y)^2] - (E[X] + E[Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] - (E[X]^2 + 2E[X]E[Y] + E[Y]^2) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)\end{aligned}$$

Exercise 3: Bernoulli–Gaussian Prior and ℓ_0 Regularization

The Bernoulli–Gaussian prior assumes each weight w_i is zero with probability $(1 - \pi)$ and Gaussian with probability π. This leads to a sparse-inducing penalty.

$P(w_i) = (1 - \pi)\cdot\delta(w_i = 0) + \pi\cdot N(w_i ; 0, \sigma^2)$

Taking the negative log of the joint prior yields a cost proportional to the number of non-zero weights:

$-\log P(w) = C + \lambda\cdot||w||_0$

Thus, the Bernoulli–Gaussian prior naturally induces ℓ_0 regularization.

Exercise 4: LASSO, Sparsity, and Ridge

ℓ_1 regularization (LASSO) encourages sparse solutions because the ℓ_1-ball has sharp corners aligned with the coordinate axes, making exact zeros likely.

ℓ_2 regularization (ridge) shrinks coefficients smoothly but seldom makes them exactly zero.

From a probabilistic viewpoint: ℓ_1 corresponds to a Laplace prior (sharp peak, heavy tails), promoting sparsity; ℓ_2 corresponds to a Gaussian prior (smooth), not forcing zeros.

LASSO is used when feature selection is important. Ridge is used when all features matter and multicollinearity must be handled.