1 Some Fun Facts

1.1 Useful Matrix Identities

1. “inverse flip” identity:
\[(I_n + AB^\top)^{-1}A = A(I_d + B^\top A),\text{ for any } n \times d \text{ matrices } A, B.\]
Proof is easy: start from the fact that \(A + AB^\top A = A(I_d + B^\top A) = (I_n + AB^\top)A.\)

2. Matrix Inversion Lemma:
\[(A + UBU^\top)^{-1} = A^{-1} - A^{-1}U(B^{-1} + U^\top A^{-1}U)^{-1}U^\top A^{-1}\]

Both of these allow us to flip between a matrix inverses of two different sizes.

1.2 Gaussian fun facts

1. Matrix multiplication:
if \(\vec{x} \sim \mathcal{N}(\vec{\mu}, C)\), then \(\vec{y} = A\vec{x}\), has marginal distribution \(y \sim \mathcal{N}(A\vec{\mu}, ACA^\top).\)

2. Sums:
if \(\vec{x} \sim \mathcal{N}(\vec{\mu}, C)\) and \(\vec{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)\) then \(\vec{y} = \vec{x} + \vec{\epsilon}\) has marginal \(\vec{y} \sim \mathcal{N}(\vec{\mu}, C + \sigma^2 I).\)

Note that the two facts above allow us to perform marginalizations that often come up in regression. Suppose for example we see the marginal:
\[
p(\vec{y}) = \int p(\vec{y}|\vec{x})p(\vec{x})d\vec{x} = \int \mathcal{N}(\vec{y}|A\vec{x}, \sigma^2 I)\mathcal{N}(\vec{x}|\vec{\mu}, C)d\vec{x}
\]

Rather than writing out the densities explicitly and trying to manipulate them, we can recognize this as a case where we can apply the two rules above, since this is equivalent to asking for the marginal distribution of \(\vec{y} = A\vec{x} + \vec{\epsilon}\), with \(\vec{x} \sim \text{Nrm}(\mu, C)\) and \(\vec{\epsilon} \sim (0, \sigma^2 I)\). Namely, we first apply fun fact #1 to see that \(A\vec{\mu} \sim \mathcal{N}(A\vec{\mu}, ACA^\top)\) and then apply #2 to obtain the distribution of the sum: \(A\vec{x} + \vec{\epsilon} = \vec{y} \sim \mathcal{N}(A\vec{\mu}, ACA^\top + \sigma^2 I).\)

2 Recap of Distributions arising in Bayesian regression

Let’s recap the basic distributions we’ve covered so far:
1. prior: \( p(\vec{w}) \).
   (sometimes denoted \( p(\vec{w}|\theta) \), to emphasize dependence on some hyperparameters \( \theta \))

2. likelihood: \( p(Y|X,\vec{w}) = \prod_{i=1}^{n} p(\vec{y}_i|\vec{x}_i, \vec{w}) \)
   (aka “observation model”, “encoding distribution” or “conditional” when considered over \( Y \)):

3. posterior \( p(\vec{w}|X,Y) \)
   (from Bayes’ rule)

4. marginal likelihood: \( p(Y|X) = \int p(Y|X,\vec{w})p(\vec{w})d\vec{w} \)
   (denominator in Bayes’ rule)

5. posterior predictive distribution: \( p(Y^*|X^*, Y, X) = \int p(Y^*|X^*, \vec{w})p(\vec{w}|Y, X)d\vec{w}. \)
   (distribution over new data \( Y^* \) given new stimuli \( X^* \) and observed data \( (X, Y) \) used for fitting; involves integrating over uncertainty in \( \vec{w} \)).

3 The linear Gaussian model

In the case of the linear-Gaussian model (with a zero-mean Gaussian prior over weights \( \vec{w} \) and linear observation model with additive Gaussian noise), these distributions are all Gaussians we can compute analytically:

1. prior: \( \vec{w} \sim \mathcal{N}(0, C). \)

2. likelihood: \( Y|X, \vec{w} \sim \mathcal{N}(X\vec{w}, \sigma^2 I_d). \)

3. posterior: \( \vec{w}|X,Y \sim \mathcal{N}\left((X^\top X + \sigma^2 C^{-1})^{-1}X^\top Y, (\frac{1}{\sigma^2}X^\top X + C^{-1})^{-1}\right) \)
   (can be computed by completing the square)

4. marginal likelihood: \( p(Y|X) = \mathcal{N}\left(0, XCX^\top + \sigma^2 I_n\right). \)
   (see notes on marginalization using Gaussian fun facts above).

5. posterior predictive:
   \( Y^*|X^*, Y, X \sim \mathcal{N}\left(X^*_s(X^\top X + \sigma^2 C^{-1})^{-1}X^\top Y, X^*_s(\frac{1}{\sigma^2}X^\top X + C^{-1})^{-1}X^*_s + \sigma^2 I_s\right) \)
   (where \( I_s \) is an identity matrix of size given by the number of rows (stimuli) in \( X^*_s \); can be derived with same marginalization tricks as the marginal likelihood).

4 Primal vs. Dual space

The formulas above for posterior and posterior predictive distribution are both known as primal space or weight space formulas because all the matrices that require inverting are size \( d \times d \), where \( d \) is the dimensionality of the weights \( \vec{w} \). An alternative is to work in the dual space or function space, which instead uses matrices of size \( n \times n \), where \( n \) is the number of stimuli. Obviously, if \( n > d \), it makes sense to work with primal space formulas. (This is the standard case
in linear-Gaussian regression). However, if \( n < d \) then it makes more sense to work with dual space formulas. We will see that dual space formulas allow us to work with with infinite-dimensional feature spaces, i.e., where the effective weight vector would be (in principle) infinite.

5 Dual Space formulas for linear-Gaussian model

We can use the two matrix identities given above to convert the primal space formulas to their dual space equivalents. (The first identity suffices for the means, while the latter applies to the covariances).

In the following we write \( I_n \) to emphasize that these are \( n \times n \) identity matrices, while in the primal space formulas involved \( I_d \), the \( d \times d \) identity matrix.

1. posterior:
\[
\bar{w}|X, Y \sim \mathcal{N}
\left(
X^\top (XCX^\top + \sigma^2 I_n)^{-1}Y, \ C - CX^\top (XCX^\top + \sigma^2 I_n)^{-1}XC
\right)
\]

2. posterior predictive:
\[
Y_*|X_*, Y, X \sim \mathcal{N}
\left(
X_*CX_*^\top (XCX^\top + \sigma^2 I_n)^{-1}Y, \ X_*CX_*^\top - X_*CX_*^\top (XCX^\top + \sigma^2 I_d)^{-1}XCX_*^\top + \sigma^2 I_*\right)
\]

6 Gram Matrix and Kernel Functions

Note that in the dual space formula for the posterior predictive distribution above, we never have to explicitly represent a matrix or vector of size \( d \); the non-diagonal matrices all take the form or \( XCX^\top \) or \( X_*CX_*^\top \) or \( X_*CX_*^\top \), which are matrices of size \( n \times n \), \( n_* \times n \) or \( n \times n_* \), respectively.

If we inspect the forms of these matrices, their elements all involve inner products of pairs of stimulus points. That is, the \( i, j \)'th element of \( XCX^\top \) is
\[
\left[ XCX^\top \right]_{ij} = \bar{x}_i^\top C \bar{x}_j
\] (2)

and similarly
\[
\left[ X_*CX_*^\top \right]_{ij} = \bar{x}_*i^\top C \bar{x}_j
\] (3)

Let \( K = XCX^\top \), which is known as the Gram matrix, consisting of the function known generally as the kernel function \( k(\cdot, \cdot) \) applied to all pairs of stimuli, where here \( k(\bar{x}_i, \bar{x}_j) = \bar{x}_i^\top C \bar{x}_j \).

Let \( K_* = X_*CX_*^\top \) denote the \( n_* \times n_* \) matrix formed by applying the kernel function to all test stimuli \( \times \) all training stimuli, and \( K_{**} = X_*CX_*^\top \) be the \( n_* \times n_* \) matrix formed by applying the kernel to all pairs of test stimuli. Then the posterior predictive distribution above can be written much more simply as:
\textbf{posterior predictive:}

\[ Y_*|X_*,Y,X \sim \mathcal{N}(K_*(K + \sigma^2 I)^{-1}Y, K_{**} - K_*(K + \sigma^2 I)^{-1}K_*^\top + \sigma^2 I_*) \]  \hfill (4)

As we will see in the next lecture, these formulas allow us to perform regression for models with more complex feature spaces, even infinite dimensional feature spaces. The canonical kernel function used in Gaussian Process regression is given by:

\[ k(\vec{x}_i, \vec{x}_j) = \rho \exp\left(-\frac{||\vec{x}_i - \vec{x}_j||^2}{2\delta^2}\right). \hfill (5) \]

Although we will not give the proof explicitly, this kernel function cannot be written as \( \phi(\vec{x}_i) \cdot \phi(\vec{x}_j) \), a dot product between two finite dimensional feature vectors \( \phi(\vec{x}_i) \) and \( \phi(\vec{x}_j) \).

Regression using this kernel function (often called the “Gaussian”, “radial basis function (RBF)” or “squared exponential (SE)” kernel) is therefore tantamount to nonlinear regression in an infinite dimensional weight space (making use of the dual space formulation essential!).

Of course, if we stick to the plain old linear kernel \( k(\vec{x}_i, \vec{x}_j) = \vec{x}_i^\top C \vec{x}_j \), then we are back in the (finite dimensional) world of linear regression.