1 Gaussian Fun Facts

We’ll add to these as we go along!

First, consider a Gaussian random vector with mean $\mu$ and covariance $C$. This is denoted $\vec{x} \sim \mathcal{N}(\vec{\mu}, C)$, and means that $\vec{x}$ has the density:

$$p(\vec{x}) = \frac{1}{\sqrt{|2\pi C|}} e^{-\frac{1}{2} (\vec{x} - \vec{\mu})^T C^{-1} (\vec{x} - \vec{\mu})},$$

where $\sqrt{|2\pi C|}$ denotes the square-root of the determinant of the matrix $2\pi C$. (This is a normalizing constant that ensures the density integrates to 1).

Then we have the following identities, where $\vec{x} \sim \mathcal{N}(\vec{\mu}, C)$ unless otherwise noted:

1. **Translation**: if $\vec{y} = \vec{x} + \vec{b}$, then $y \sim \mathcal{N}(\vec{\mu} + \vec{b}, C)$.

2. **Matrix multiplication**: if $\vec{y} = A\vec{x} + \vec{b}$, then $y \sim \mathcal{N}(A\vec{\mu}, ACA^T)$.

3. **Sums of independent Gaussian RVs**: if $\vec{x} \sim \mathcal{N}(\vec{\mu}_1, C_1)$ and $\vec{y} \sim \mathcal{N}(\vec{\mu}_2, C_2)$ then $\vec{x} + \vec{y} \sim \mathcal{N}(\vec{\mu}_1 + \vec{\mu}_2, C_1 + C_2)$.

4. **Products of Gaussian Densities**: The product of two Gaussian densities is proportional to a Gaussian density. Namely for two densities $\mathcal{N}(a, A)$ and $\mathcal{N}(b, B)$ we have:

$$\mathcal{N}(a, A) \cdot \mathcal{N}(b, B) \propto \mathcal{N}(c, C)$$

where $C = (A^{-1} + B^{-1})^{-1}$, $c = C(A^{-1}a + B^{-1}b)$

2 Brief of maximum likelihood for GLMs

So far we have talked mainly about maximum-likelihood (ML) estimators for GLMs, in which the parameter vector $\vec{w}$ was determined by maximizing the log-likelihood,

$$\hat{\vec{w}}_{\text{ml}} = \arg \max_{\vec{w}} \log p(Y | X, \vec{w}),$$

where $Y$ denotes the $n \times 1$ vector of responses and $X$ denotes the $n \times d$ design matrix of regressors.
Linear Gaussian model: In the case of the “linear Gaussian GLM”, we had a closed form solution for the maximum-likelihood estimator. Here, the likelihood was given by a multi-variate normal distribution:

\[ Y \mid X, \vec{w} \sim \mathcal{N}(X\vec{w}, \sigma^2 I), \]  

(4)

where \( \sigma^2 \) denotes the variance of the (additive Gaussian) noise in \( Y \). The maximum was given by

\[ \hat{w}_{\text{ml}} = (X^\top X)^{-1}X^\top Y. \]  

(5)

3 Bayesian approach

Today we will focus take a Bayesian approach and focus on the posterior distribution over \( \vec{w} \) given the data, which depends on a combination of the likelihood with a prior distribution. The posterior is given by Bayes rule:

\[ \frac{\text{posterior}}{p(\vec{w} \mid Y, X, \theta)} = \frac{\text{likelihood}}{P(Y \mid X, \vec{w})} \frac{\text{prior}}{P(\vec{w} \mid \theta)} \]  

"marginal likelihood" or "evidence"  

(6)

The prior \( p(\vec{w} \mid \theta) \) depends on a set of parameters \( \theta \) that we will typically refer as hyperparameters. These hyperparameters determine the shape (or strength) of the prior distribution.

The term in the denominator, known as the marginal likelihood or evidence is a normalizing constant that does not depend on \( \vec{w} \). Its purpose is simply to make sure that the posterior distribution integrates to 1, that is, \( \int p(\vec{w} \mid Y, X, \theta) d\vec{w} = 1 \). The denominator can therefore be seen as simply the integral of the stuff in the numerator:

\[ p(Y \mid X, \theta) = \int p(Y \mid X, \vec{w}) p(\vec{w} \mid \theta) d\vec{w} = \int p(Y, \vec{w} \mid X, \theta) d\vec{w}, \]  

(7)

where the rightmost expression shows that it is what we get by taking the joint distribution of \( Y \) and \( \vec{w} \) and marginalizing (integrating) over \( \vec{w} \) (hence the term “marginal likelihood”).

3.1 MAP inference / penalized maximum likelihood

The maximum a posteriori (MAP) estimator corresponds to the maximum of the posterior:

\[ \hat{w}_{\text{map}} = \arg \max_{\vec{w}} p(\vec{w} \mid Y, X, \theta). \]  

(8)

Note that this is equivalent to maximizing the product of likelihood and prior, since the evidence (denominator in Bayes’ rule) does not involve \( \vec{w} \) and does not change the location of the maximum:

\[ \hat{w}_{\text{map}} = \arg \max_{\vec{w}} p(Y \mid X, \vec{w}) P(\vec{w} \mid \theta) = \arg \max_{\vec{w}} \left[ \log P(Y \mid X, \vec{w}) + \log P(\vec{w} \mid \theta) \right]. \]  

(9)
The last expression shows that the MAP estimator can be thought of as maximizing the log-likelihood plus a “penalty” given by the log-prior. This view motivates the terminology that refers to MAP estimation as “penalized log-likelihood”.

3.2 Common priors / penalties

1) L2 penalty - also known as “L2 shrinkage” or “ridge penalty” or “ridge prior"

penalty = −\( \frac{1}{2} \lambda ||\vec{w}||^2 \) = −\( \frac{1}{2} \lambda \sum_{i=1}^{d} w_i^2 \) = −\( \frac{1}{2} \lambda \vec{w}^\top \vec{w} \) (10)

Corresponds to an independent Gaussian prior on the weights with variance \( \theta = \frac{1}{\lambda} \):

\( \vec{w} \sim \mathcal{N}(0, \theta I) \). (11)

The name “ridge prior” comes from the fact that the prior covariance matrix consists of a ridge of height \( \theta \) along the main diagonal.

To verify the equivalence:

\[ \log p(\vec{w}|\theta) = \frac{1}{2\theta} \vec{w}^\top \vec{w} + \text{const} \] (12)

2) L2 smoothing penalty - also known as “graph Laplacian”.

penalty = −\( \frac{1}{2} \lambda \sum_{i=2}^{d} (w_i - w_{i-1})^2 \) (13)

It turns out this operation (taking the sum squared pairwise difference between neighboring weights) can be embedded in the following matrix operation:

\[ \text{sum}_{i=2}^{d} (w_i - w_{i-1})^2 = \vec{w}^\top \Gamma \vec{w} \] (14)

where \( \Gamma \) is the graph Laplacian defined by:

\( \Gamma_{ii} = \text{degree}, \text{ or } \# \text{ of neighbors of weight } i \)

\( \Gamma_{ij} = -1 \text{ if } i \text{ and } j \text{ are neighbors} \) (15)

Thus for example, for a “1-dimensional” weight vector \( \vec{w} = [w_1, w_2, w_3, w_4] \), we can compute the squared pairwise differences via:

\[ ||D\vec{w}||^2 = \vec{w}^\top D^\top D \vec{w} = \vec{w}^\top \Gamma \vec{w} \] (16)

where \( D \) is the pairwise-difference computing matrix:

\[ D = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \] (17)
and $\Gamma$, the graph Laplacian is the tri-diagonal matrix:

$$D^\top D = \Gamma = \begin{bmatrix}
1 & -1 & & \\
-1 & 2 & -1 & \\
& -1 & 2 & -1 \\
& & -1 & 1
\end{bmatrix}. \quad (18)$$

Using this penalty is tantamount to a Gaussian prior given by:

$$\bar{w} \sim \mathcal{N}(0, \theta \Gamma^{-1}) \quad (19)$$

although $\Gamma$, which is of rank $d - 1$ is not invertible, meaning that this is a “degenerate” prior with infinite variance along the direction $\bar{w}_1 = \bar{w}_2 = \cdots = \bar{w}_d$.

3.3 MAP inference for linear-Gaussian model

For the linear-Gaussian model we can compute the MAP estimate in closed form. We have the model now specified by

$$\bar{w} \sim \mathcal{N}(0, C) \mid X, \bar{w} \sim \mathcal{N}(X \bar{w}, \sigma^2 I), \quad (20)$$

where $C$ is a prior covariance matrix (parametrized by $\theta$) and $\sigma^2$ is the variance of the additive noise in $Y$.

The log-posterior is given by:

$$\log p(Y \mid X, \bar{w}) + \log p(\bar{w} \mid \theta) + \text{const} = \frac{1}{2\sigma^2} (Y - X \bar{w})^\top (Y - X \bar{w}) - \frac{1}{2} \bar{w}^\top C^{-1} \bar{w} + \text{const}$$

$$= -\frac{1}{2} \bar{w}^\top \left( \frac{1}{\sigma^2} X^\top X \right) \bar{w} - \frac{1}{2} \bar{w}^\top C^{-1} \bar{w} - \bar{w} \left( \frac{1}{\sigma^2} X^\top Y \right) + \text{const}$$

$$= -\frac{1}{2} \bar{w}^\top \left( \frac{1}{\sigma^2} X^\top X + C^{-1} \right) \bar{w} - \bar{w} \left( \frac{1}{\sigma^2} X^\top Y \right) + \text{const}, \quad (21)$$

where $\text{const}$ is a catch-all that collects terms that do not contain $\bar{w}$.

Differentiating with respect to $\bar{w}$, setting to zero and solving gives:

$$\hat{w}_{\text{map}} = (X^\top X + C^{-1})^{-1} X^\top Y \quad (22)$$

In the case of ridge regression, where prior covariance $C = \theta I$, this corresponds to

$$\hat{w}_{\text{ridge}} = (X^\top X + \frac{\sigma^2}{\theta} I)^{-1} X^\top Y. \quad (23)$$

Thus the stimulus covariance $X^\top X$ is added to a diagonal matrix with diagonal value $\sigma^2/\theta$, the ratio of noise variance to prior variance. The larger this value, the more the coefficients of $\bar{w}_{\text{map}}$ are shrunk towards zero.