Bayesian linear regression

Model
Have vector $y \in \mathbb{R}^n$ and covariates matrix $X \in \mathbb{R}^{n \times d}$ (or $d + 1$ if w/ bias). The $i$th row of $y$ and $X$ correspond to the $i$th observation $(y_i, x_i)$.

In a Bayesian setting, we model this data as:

\begin{align*}
\text{Likelihood} : \quad & y \sim N(Xw, \sigma^2 I) \\
\text{Prior} : \quad & w \sim N(0, \lambda^{-1} I)
\end{align*}

The unknown model variable is $w \in \mathbb{R}^d$.

- The “likelihood model” says how well the observed data agrees with $w$.
- The “prior model” is our prior belief (or constraints) on $w$.

This Bayesian linear regression because we put distributions on the data and use a prior distribution on the unknown parameter.
**MAP solution**

MAP inference returns the maximum of the log joint likelihood.

\[
\text{Joint Likelihood : } p(y, w|X) = p(y|w, X)p(w)
\]

Using Bayes rule that this point also maximizes the posterior of \( w \).

\[
w_{\text{MAP}} = \underset{w}{\text{arg max}} \ln p(w|y, X) \\
= \underset{w}{\text{arg max}} \ln p(y|w, X) + \ln p(w) \\
= \underset{w}{\text{arg max}} - \frac{1}{2\sigma^2} (y - Xw)^T (y - Xw) - \frac{\lambda}{2} w^T w + \text{const.}
\]

We saw that this solution for \( w_{\text{MAP}} \) is the same as for ridge regression:

\[
w_{\text{MAP}} = (\lambda \sigma^2 I + X^T X)^{-1} X^T y \iff w_{\text{RR}}
\]
Point estimates

\( w_{\text{MAP}} \) and \( w_{\text{ML}} \) are referred to as *point estimates* of the model parameters.

They find a specific value (point) of the vector \( w \) that maximizes an objective function (MAP or ML).

- **ML**: Only consider data model: \( p(y|w, X) \).
- **MAP**: Takes into account prior model: \( p(y, w|X) = p(y|w, X)p(w) \).

Bayesian inference

Bayesian inference goes one step further by characterizing uncertainty about the values in \( w \) using Bayes rule.
Bayes rule and linear regression

Posterior calculation
Since $w$ is a continuous-valued random variable in $\mathbb{R}^d$, Bayes rule says that the posterior distribution of $w$ given $y, X$ is

$$p(w|y, X) = \frac{p(y|w, X)p(w)}{\int_{\mathbb{R}^d} p(y|w, X)p(w) \, dw}$$

That is, we get an updated distribution on $w$ through the transition

prior $\rightarrow$ likelihood $\rightarrow$ posterior

Quote: “The posterior of __ is proportional to the likelihood times the prior.”
Bayesian linear regression

In this case, we can update the posterior distribution $p(w|y, X)$ analytically.

We work with the proportionality first:

$$p(w|y, X) \propto p(y|w, X)p(w)$$

$$\propto \left[e^{-\frac{1}{2\sigma^2}(y-Xw)^T(y-Xw)}\right]\left[e^{-\frac{1}{2}w^Tw}\right]$$

$$\propto e^{-\frac{1}{2}\{w^T(\lambda I+\sigma^{-2}X^TX)w-2\sigma^{-2}w^TX^Ty\}}$$

We can multiply and divide this by whatever we want to make it a probability distribution, *as long as it doesn’t involve $w$*. 
Bayesian inference for linear regression

We need to normalize:

\[ p(w|y, X) \propto e^{-\frac{1}{2} \left\{ w^T (\lambda I + \sigma^{-2} X^T X) w - 2\sigma^{-2} w^T X^T y \right\}} \]

There are two key terms in the exponent:

1. \( w^T (\lambda I + \sigma^{-2} X^T X) w \) is quadratic in \( w \).
2. \( 2\sigma^{-2} w^T X^T y / \sigma^2 \) is linear in \( w \).

We can conclude that \( p(w|y, X) \) is Gaussian. Why?

1. We can multiply and divide by anything not involving \( w \).
2. A Gaussian has \( (w - \mu)^T \Sigma^{-1} (w - \mu) \) in the exponent.
3. We can “complete the square” by adding terms not involving \( w \).
Bayesian inference for linear regression

**Compare:** In other words, a Gaussian looks like:

\[
p(w|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (w^T \Sigma^{-1} w - 2w^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu)}
\]

and we’ve shown for some setting of \( Z \) that

\[
p(w|y, X) = \frac{1}{Z} e^{-\frac{1}{2} (w^T (\lambda I + \sigma^{-2} X^T X) w - 2w^T X^T y / \sigma^2)}
\]

**Conclude:** What happens if in the original Gaussian we define:

\[
\Sigma^{-1} = (\lambda I + \sigma^{-2} X^T X), \quad \Sigma^{-1} \mu = X^T y / \sigma^2
\]

Using these specific values of \( \mu \) and \( \Sigma \) we only need to set

\[
Z = (2\pi)^{d/2} |\Sigma|^{1/2} e^{\frac{1}{2} \mu^T \Sigma^{-1} \mu}
\]
The posterior distribution

Therefore, the posterior distribution of $w$ is:

$$p(w|y, X) = N(w|\mu, \Sigma),$$

$$\Sigma = (\lambda I + \sigma^{-2}X^TX)^{-1},$$

$$\mu = (\lambda \sigma^2 I + X^TX)^{-1}X^Ty \iff w_{MAP}$$

Things to notice:

- $\mu = w_{MAP}$ modulo a redefinition of the regularization parameter $\lambda$.
- $\Sigma$ captures uncertainty about $w$ as $\text{Var}[w_{LS}]$ and $\text{Var}[w_{RR}]$ did before.
- However, now we have a full probability distribution on $w$. 
Uses of the Posterior Distribution

Understanding $w$

We saw how we could calculate the variance of $w_{LS}$ and $w_{RR}$. Now we have an entire distribution. Some questions we can ask are:

**Q:** Is $w_i > 0$ or $w_i < 0$? Can we confidently say $w_i \neq 0$?

**A:** Use the **marginal posterior distribution**: $w_i \sim N(\mu_i, \Sigma_{ii})$.

**Q:** How do $w_i$ and $w_j$ relate?

**A:** Use their joint marginal posterior distribution:

$$
\begin{bmatrix}
w_i \\
w_j
\end{bmatrix} \sim N
\begin{bmatrix}
\mu_i \\
\mu_j
\end{bmatrix},
\begin{bmatrix}
\Sigma_{ii} & \Sigma_{ij} \\
\Sigma_{ji} & \Sigma_{jj}
\end{bmatrix}
$$

Predicting new data

The posterior $p(w|y, X)$ is perhaps most useful for predicting new data.
PREDICTING NEW DATA
**PREDICTING NEW DATA**

**Recall:** For a new pair \((x_0, y_0)\) with \(x_0\) measured and \(y_0\) unknown, we can predict \(y_0\) using \(x_0\) and the LS or RR (i.e., ML or MAP) outputs:

\[
y_0 \approx x_0^T w_{\text{LS}} \quad \text{or} \quad y_0 \approx x_0^T w_{\text{RR}}
\]

With Bayes rule, we can make a *probabilistic* statement about \(y_0\):

\[
p(y_0|x_0, y, X) = \int_{\mathbb{R}^d} p(y_0|x_0, w) p(w|y, X) \, dw
\]

This is just a more complicated looking integral form of

\[
P(A) = \sum_b P(A, B = b) = \sum_b P(A|B = b)P(B = b).
\]

Having \((x_0, y, X)\) on the right of the conditioning bar doesn’t change this fact.
Predictive distribution (intuition)

We evaluate the likelihood of a $y_0$ for a given $w$ and $x_0 \rightarrow p(y_0|x_0, w)$, weighted by our current belief about its probability $\rightarrow p(w|y, X)$.

Using integration we can consider the infinite possible values for $w$.

This is called the *predictive distribution*:

$$p(y_0|x_0, y, X) = \int_{\mathbb{R}^d} p(y_0|x_0, w) p(w|y, X) \, dw$$

With this equation, we are integrating out all the uncertainty about $w$. 
We know from the model and Bayes rule that

Model: \[ p(y_0|x_0, w) = N(y_0|x_0^Tw, \sigma^2), \]

Bayes rule: \[ p(w|y, X) = N(\mu, \Sigma). \]

With \( \mu \) and \( \Sigma \) calculated on a previous slide.

The predictive distribution can be calculated exactly with these distributions. (We skip this derivation.) Again we get a Gaussian distribution:

\[ p(y_0|x_0, y, X) = N(y_0|\mu_0, \sigma_0^2), \]

\[ \mu_0 = x_0^T\mu, \]
\[ \sigma_0^2 = \sigma^2 + x_0^T\Sigma x_0. \]

Notice that the expected value is the MAP prediction, \( x_0^T w_{\text{MAP}} \), but we now quantify our confidence in this prediction with the variance \( \sigma_0^2 \).
ACTIVE LEARNING
Bayesian learning is naturally thought of as a sequential process. That is, the posterior after seeing some data becomes the prior for the next data.

Let $y$ and $X$ be “old data” and $y_0$ and $x_0$ be some “new data”. By Bayes rule

$$p(w|y_0, x_0, y, X) \propto p(y_0|w, x_0)p(w|y, X).$$

The posterior after $(y, X)$ has become the prior for $(y_0, x_0)$.

Simple modifications can be made sequentially:

$$p(w|y_0, x_0, y, X) = N(w|\mu', \Sigma'),$$

$$\Sigma' = (\lambda I + \sigma^{-2}(x_0x_0^T + \sum_{i=1}^n x_ix_i^T))^{-1},$$

$$\mu' = (\lambda \sigma^2 I + (x_0x_0^T + \sum_{i=1}^n x_ix_i^T)^{-1}(x_0y_0 + \sum_{i=1}^n x_iy_i)).$$
Of course, we could also have written

\[ p(w|y_0, x_0, y, X) \propto p(y_0, y|w, X, x_0)p(w) \]

but often we want to use the sequential aspect of inference to help us learn.

Learning \( w \) and making predictions for new \( y_0 \) is a two-step procedure:

- Form the predictive distribution \( p(y_0|x_0, y, X) \).
- Update the posterior distribution \( p(w|y, X, y_0, x_0) \).

**Question:** Can we learn \( p(w|y, X) \) intelligently?

That is, if we’re in the situation where we can pick which \( y_i \) to measure with the knowledge of \( \mathcal{D} = \{x_1, \ldots, x_n\} \), can we come up with a good strategy?
**ACTIVE LEARNING**

An “active learning” strategy

Imagine we already have a measured dataset \((y, X)\) and posterior \(p(w|y, X)\). We can construct the predictive distribution for every remaining \(x_0 \in D\).

\[
p(y_0|x_0, y, X) = N(y_0|\mu_0, \sigma_0^2),
\]

\[
\mu_0 = x_0^T \mu,
\]

\[
\sigma_0^2 = \sigma^2 + x_0^T \Sigma x_0.
\]

For each \(x_0\), \(\sigma_0^2\) tells how confident we are. This suggests the following:

1. Form the predictive distribution \(p(y_0|x_0, y, X)\) using \(p(w|y, X)\)
2. Pick the \(x_0 \in D\) for which \(\sigma_0^2\) is largest and measure \(y_0\)
3. Update the posterior \(p(w|y_0, x_0, y, X)\)
4. Return to \#1 using the updated posterior
Entropy (i.e., uncertainty) minimization

When devising a procedure such as this one, it’s useful to know what *objective function* is being optimized in the process.

We introduce the concept of the *entropy* of a distribution. Let $p(x)$ be a continuous distribution, then its (differential) entropy is:

$$\mathcal{H}(p) = -\int p(x) \ln p(x) dx.$$  

This is a measure of the spread of the distribution. Larger values correspond to a more “uncertain” distribution (more variance).

The entropy of a multivariate Gaussian is

$$\mathcal{H}(N(w|\mu, \Sigma)) = \frac{d}{2} \ln \left(2\pi e|\Sigma|\right).$$
The entropy of a Gaussian changes with its covariance matrix. With sequential Bayesian learning, the covariance transitions from

\[
\text{Prior} : \quad (\lambda I + \sigma^{-2}X^T X)^{-1}
\]

\[
\downarrow
\]

\[
\text{Posterior} : \quad (\lambda I + \sigma^{-2}(x_0 x_0^T + X^T X))^{-1}
\]

We can show that the following two “rules” are equivalent:

1. Selecting \(x_0\) with the largest variance for \(y_0\).
2. Selecting the \(x_0\) that reduces the posterior entropy the most.

We are minimizing \(\mathcal{H}\) myopically, so this is called a “greedy algorithm”.
Model selection
**Selecting $\lambda$**

We’ve discussed $\lambda$ as a “nuisance” parameter that can impact performance.

To set it, we considered:

- The “degrees of freedom” (Lecture 3, last slide & ESL pg. 68)
- Cross-validation (Lecture 4)

Bayes rule gives a principled way to do this via *evidence maximization*:

$$
p(w|y, X, \lambda) = \frac{p(y|w, X)p(w|\lambda)}{p(y|X, \lambda)}.
$$

The “evidence” gives the likelihood of the data with $w$ integrated out. It’s a measure of how good our model and parameter assumptions are.
Let's consider the selection of the hyperparameter $\lambda$. We can set $\lambda$ by maximizing the evidence.

$$\hat{\lambda} = \arg \max_\lambda \ln p(y|X, \lambda).$$

We can show that the distribution of $y$ is $p(y|X, \lambda) = N(y|0, \sigma^2 I + \lambda^{-1}X^TX)$. This requires an algorithm to maximize over $\lambda$.

We notice that this looks exactly like maximum likelihood, and it is:

**Type-I ML**: Maximize the likelihood over the “main parameter” ($w$).

**Type-II ML**: Integrate out “main parameter” ($w$) and maximize over the “hyperparameter” ($\lambda$). Also called *empirical Bayes*.

The difference is only in their perspective.

This approach requires that we can solve this integral, but often we can’t for more complex models. Cross-validation is the method that always works.