Model
Have vector $y \in \mathbb{R}^n$ and covariates matrix $X \in \mathbb{R}^{n \times d+1}$ (+1 for bias). We model it as:

**Likelihood**: $y \sim N(Xw, \sigma^2 I)$

**Prior**: $w \sim N(0, \lambda^{-1} I)$

The unknown vector is $w \in \mathbb{R}^{d+1}$.

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

In Bayesian linear regression, we interpret $y_i \in \mathbb{R}$ as response and $x_i \in \mathbb{R}^{d+1}$ as its features. The $y_i$'s are independent, but we can combine using $\sigma^2 I$. 
**Maximum a posteriori inference**

MAP inference returns the maximum of the log joint likelihood.

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

**MAP solution**

\[
\hat{w}_{\text{MAP}} = \arg \max_w \ln p(y|w, X) + \ln p(w)
\]

\[
= \arg \max_w -\frac{1}{2\sigma^2} (y - Xw)^T (y - Xw) - \frac{\lambda}{2} w^T w + \text{const.}
\]

Interestingly, the solution for \(\hat{w}_{\text{MAP}}\) is the same as for ridge regression:

\[
\hat{w}_{\text{MAP}} = (\lambda \sigma^2 I + X^T X)^{-1} X^T y \iff \hat{w}_{\text{RR}}
\]

We also saw that this is the point that maximizes the posterior of \(w\).
Point estimates

\( \hat{w}_{\text{MAP}} \) and \( \hat{w}_{\text{ML}} \) 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 (via likelihood).
- **MAP**: Takes into account prior model (via joint likelihood).

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 $\mathbb{R}^{d+1}$, Bayes rule says that the posterior distribution of $w$ given $y, X$ is

$$p(w|y, X) = \frac{p(y|w, X)p(w)}{\int_{\mathbb{R}^{d+1}} p(y|w, X)p(w) \, dw}$$

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

prior $\rightarrow$ likelihood (data) $\rightarrow$ 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:

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

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

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

We can multiply and divide this by whatever we want to make it a probability distribution, *as long as it doesn’t involve* $w$. 
We need to normalize:

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

There are two key terms in the exponent:

\[ w^T (\lambda I + \sigma^{-2} X^T X)w - 2w^T X^T y/\sigma^2 \]

- quadratic in \( w \)
- linear in \( w \)

Since there is no \( w \) in front of \( \exp\{\} \), this tells us that \( p(w|y, X) \) is Gaussian.

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 \( w \).
The posterior distribution

Therefore, the posterior distribution of \( w \) is:

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

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

\[
\mu = (\lambda I + \sigma^{-2}X^TX)^{-1}X^Ty/\sigma^2,
\]

\[
= (\lambda \sigma^2 I + X^TX)^{-1}X^Ty \Leftarrow \hat{w}_{\text{MAP}}
\]

Things to notice:

- \( \mu = \hat{w}_{\text{MAP}} \) modulo the regularization parameter setting (\( \lambda' \leftarrow \lambda \sigma^2 \)).
- \( \Sigma \) captures uncertainty about \( w \) like Var[\( \hat{w}_{\text{LS}} \)] and Var[\( \hat{w}_{\text{RR}} \)] did before.
- But 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 $\hat{w}_{LS}$ and $\hat{w}_{RR}$. 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: The marginal distribution is $w_i | y, X \sim N(\mu_i, \Sigma_{ii})$.

Q: How do $w_i$ and $w_j$ relate?
A: Look at posterior correlation $\Sigma_{ij} / \sqrt{\Sigma_{ii} \Sigma_{jj}}$.

Predicting new data
The posterior $p(w|y, X)$ is perhaps most useful for 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 outputs: \(y_0 \approx x_0^T \hat{w}_{LS}\) or \(y_0 \approx x_0^T \hat{w}_{RR}\).

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

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

The extra \(X\) on the right of the conditioning bar doesn’t change this fact.
Predictive distribution (intuition)

According to our most recent distribution on $w$ — $p(w|y, X)$ — we have an infinite possible number of values for $w$.

Using calculus we can “consider all of them” according to their posterior probability. This is called the *predictive distribution*:

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

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

Model: \[ p(y_0|x_0, w) = N(y_0|x_0^T w, \sigma^2), \]
Bayes rule: \[ p(w|y, X) = N(\mu, \Sigma). \]

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

The predictive distribution can be calculated analytically with these distributions. The results is again 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, \( x_0^T \hat{w}_{\text{MAP}} \), but we now quantify our confidence in this prediction with the variance \( \sigma_0^2 \).
Active learning
Bayesian methods are naturally thought of as sequential processes. That is, the posterior after 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:

$$p(w|y_0, x_0, y, X) = N(w|\mu', \Sigma'),$$
$$\Sigma' = (\lambda I + \sigma^{-2}(x_0x_0^T + \sum_{i=1}^{n} x_ix_i^T))^{-1},$$
$$\mu' = (\lambda \sigma^2 I + (x_0x_0^T + \sum_{i=1}^{n} x_ix_i^T)^{-1}(x_0y_0 + \sum_{i=1}^{n} x_iy_i).$$
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 $D = \{x_1, \ldots, x_n\}$, can we come up with a good strategy?
**ACTIVE LEARNING**

An active learning strategy

Imagine we already have a measured set \((y, X)\) 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 \(x_0\), \(\sigma_0^2\) is how confident we are. This suggests the following:

1. Form the predictive distributions using the posterior \(p(w|y, X)\)
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
**ACTIVE LEARNING**

**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:

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

It’s a measure of the spread of the distribution. Larger values correspond to a more “uncertain” distribution.

The entropy of a multivariate Gaussian is

$$\mathcal{H}(N(w|\mu, \Sigma)) = \frac{d + 1}{2} \ln \left(2\pi e|\Sigma|\right).$$
ACTIVE LEARNING

The entropy of a Gaussian is a function of the covariance matrix only. With sequential Bayesian learning, the covariance transitions from

\[
\text{Prior : } (\lambda I + \sigma^{-2}X^TX)^{-1} \\
\downarrow \\
\text{Posterior : } (\lambda I + \sigma^{-2}(x_0x_0^T + X^TX))^{-1}
\]

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 want to minimize \( H \) 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.
Selecting \( \lambda \)

If we want to set \( \lambda \), we can do it by maximizing the evidence.

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

We can show that the distribution of \( y \) is 

\[
p(y|X, \lambda) = N(y|0, \sigma^2 I + \lambda^{-1} X^T X).
\]

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 between Type I & II is mostly in their perspectives. Here, Type-I doesn’t have a prior on \( w \), and Type-II has an implied one.