Intro to Bayesian Machine Learning

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Material Covered

Models: Linear models, basis expansion, logistic regression, neural networks, Prob. densities, Bayesian density estimation

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity, Maximum likelihood, maximum a posteriori, LOTUS, change of variables, expectation identities, equating coefficients, epistemic/aleatoric uncertainty

Settings: Regression, Classification, Density Estimation

This lecture: Bayesian linear regression, joint Gaussian, posterior predictive, marginal likelihood

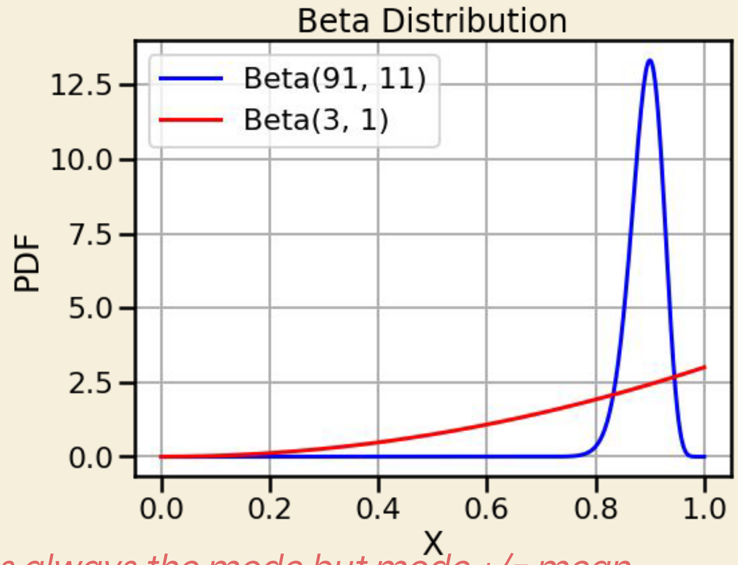
Errata from last lecture

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

Seller 1 mean: 0.892
(Variance: 0.0009)

Seller 2 mean: 0.75
(Variance: 0.05)



Takeaway: MAP is always the mode but mode +/= mean

Recap: Bayes theorem & uncertainty

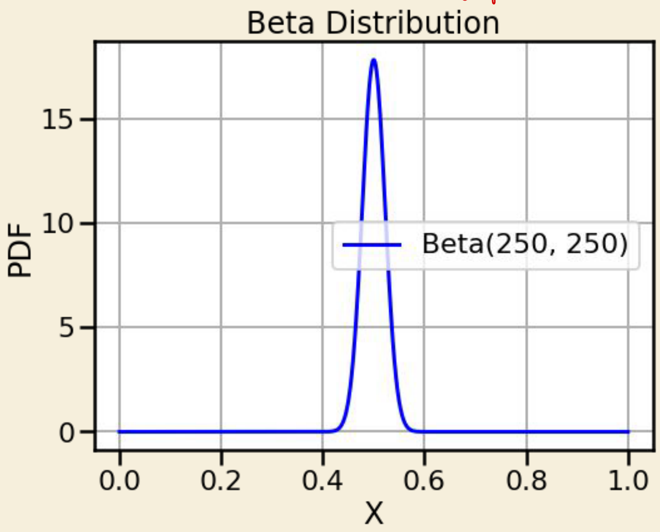
Recall that the fundamental modelling choice with Bayesian probability/statistics is that the parameter is a random variable.

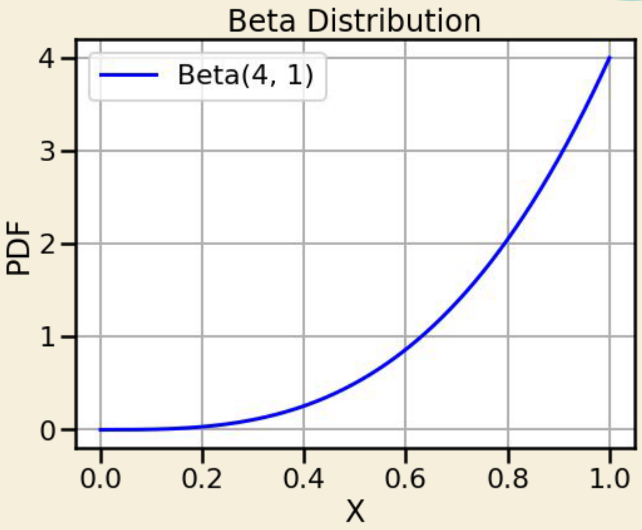
$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(D)}$$

Given data, we can compute the "posterior probability" according to Bayes theorem

Uncertainty in ML: Aleatoric

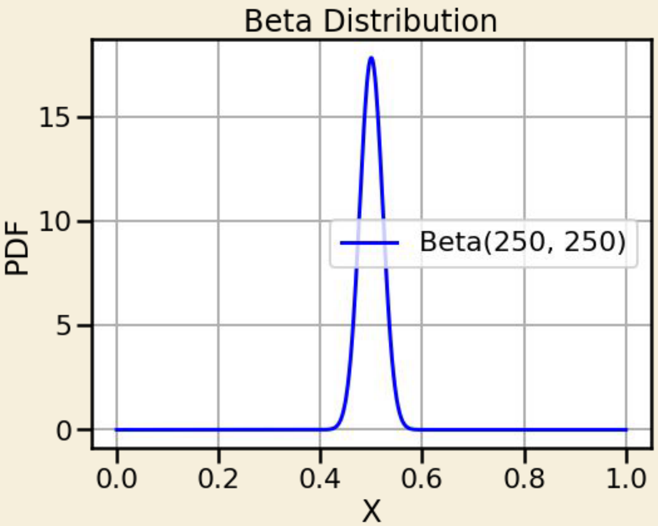
aleatoric - uncertainty from noise

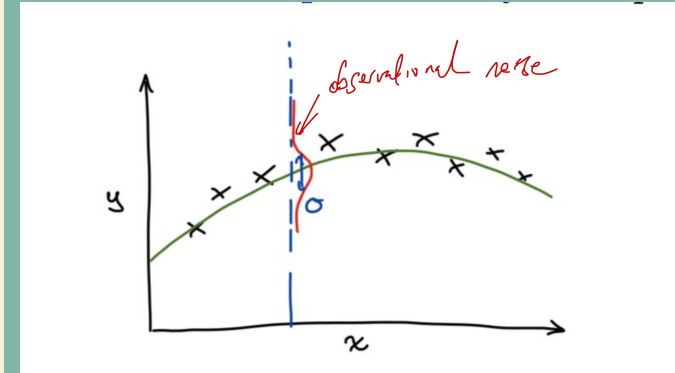




We discussed uncertainty last lecture. Can you recall which one of these models irreducible/aleatoric uncertainty?

Uncertainty in ML: Aleatoric

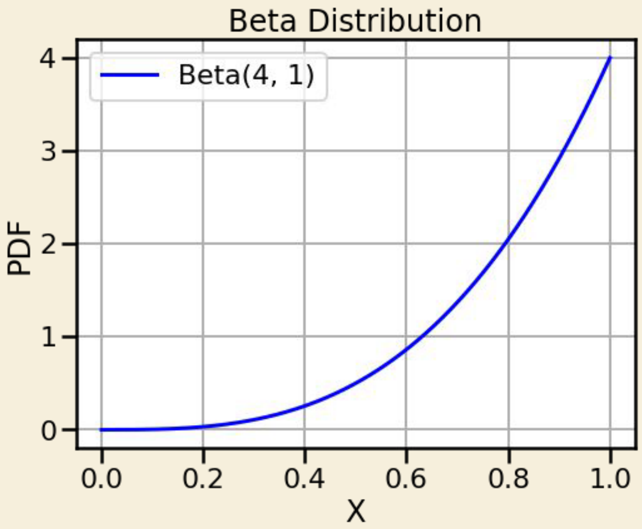




$$(\boldsymbol{\theta}^*, \sigma^{2*}) = \text{argmax}_{\boldsymbol{\theta},\sigma^2} \log p(\mathbf{y}|\boldsymbol{\theta}, \sigma^2, X)$$

We have seen aleatoric uncertainty before in our conditional density estimation/linear regression setting.

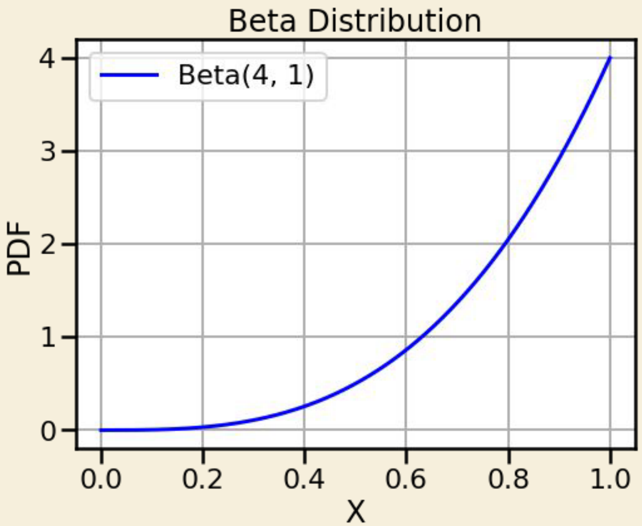
Uncertainty in ML: Epistemic

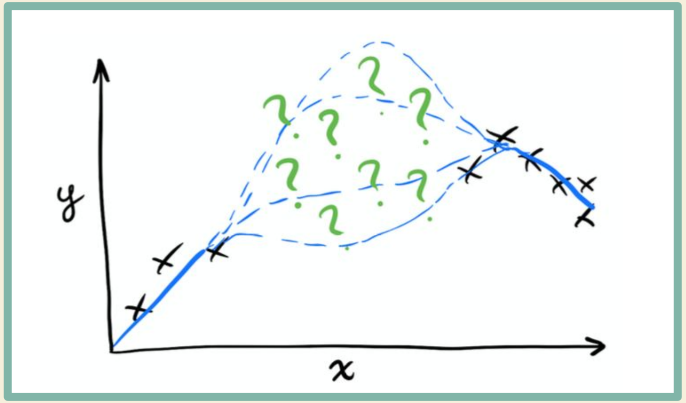


Epistemic uncertainty on the other hand might be a little less obvious in the linear regression setting. But recall: this is the uncertainty from lack of data.

Uncertainty in ML: Epistemic

randomness in fn as a
result of lack of data

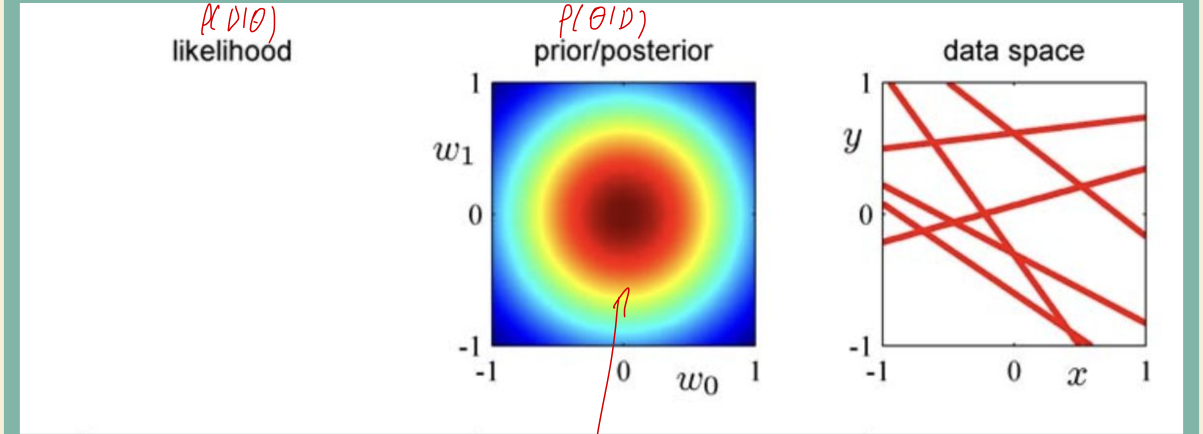




Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.

Uncertainty in ML: Epistemic

Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.

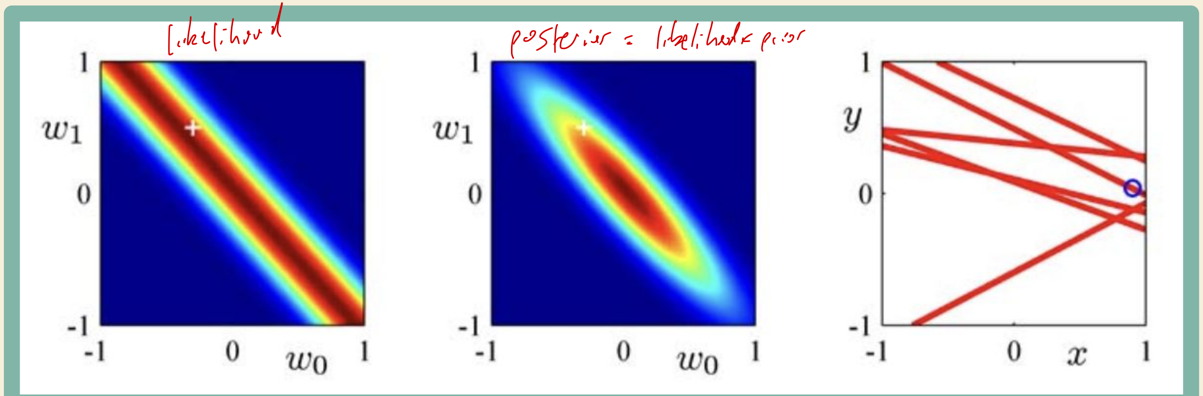


Joint gaussian

Uncertainty in ML: Epistemic

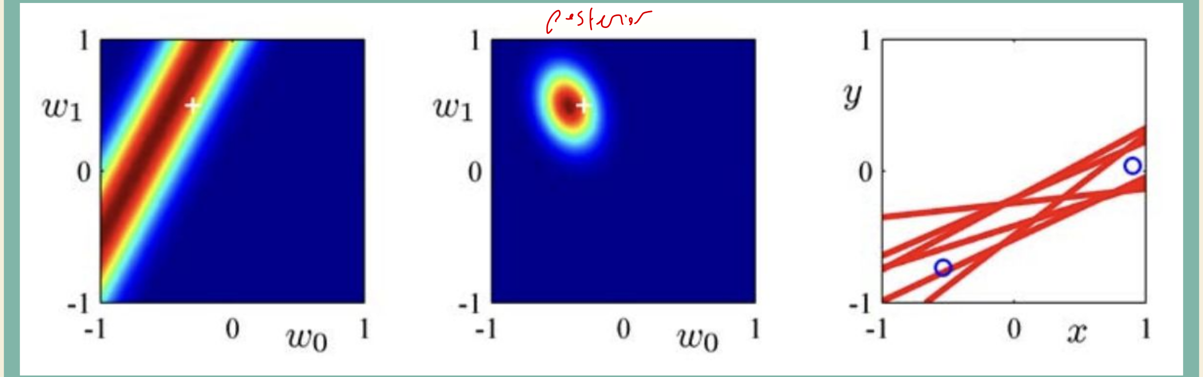
Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.

observed a data point:



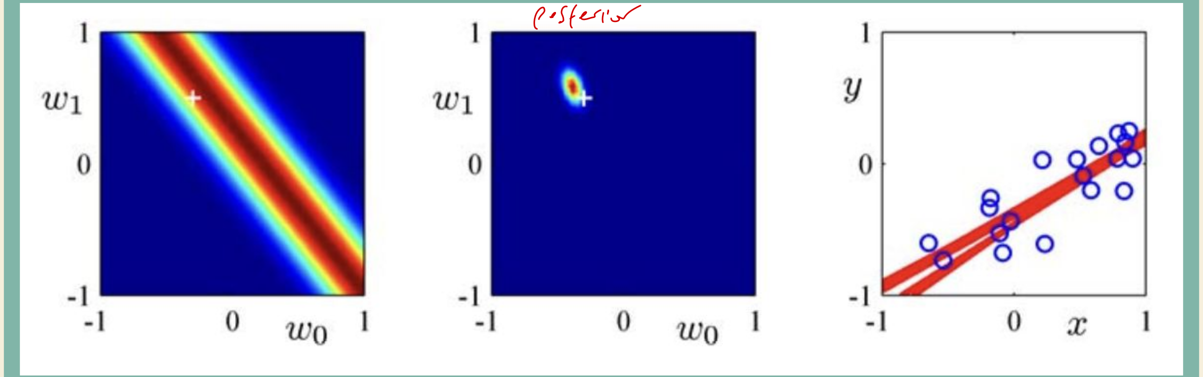
Uncertainty in ML: Epistemic

Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



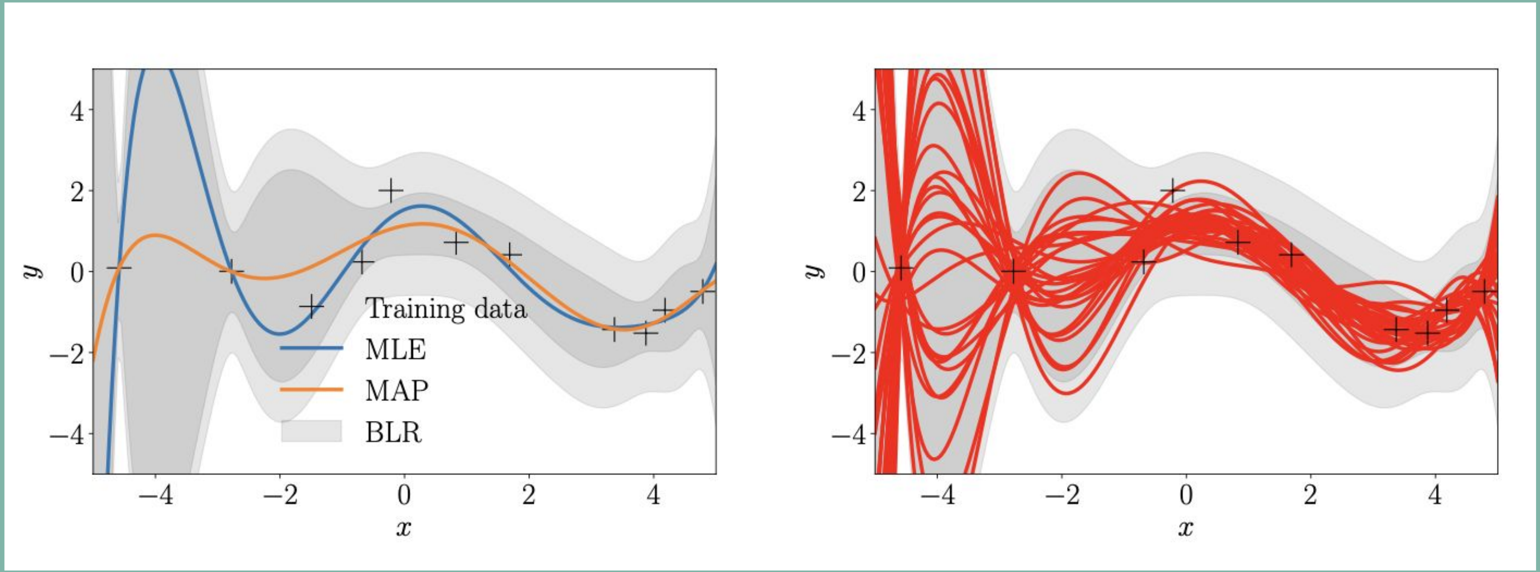
Uncertainty in ML: Epistemic

Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



Epistemic uncertainty: basis expansion

Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



Our favorite: linear regression

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{\top(1)}, \\ \mathbf{x}^{\top(2)}, \\ \vdots, \\ \mathbf{x}^{\top(N)}, \end{bmatrix} = \begin{bmatrix} x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}, \\ x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}, \\ \vdots, \ddots, \vdots \\ x_1^{(N)}, x_2^{(N)}, \ldots, x_n^{(N)} \end{bmatrix}$$

Design matrix / feature matrix

Our favorite: linear regression

$\phi(x_1, x_2, x_3 \dots x_n) = \phi(x_1), \phi(x_2), \dots$

eg. $\phi(x_1) = 1 + x_1 + x_1^2$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{\top(1)}, \\ \mathbf{x}^{\top(2)}, \\ \vdots, \\ \mathbf{x}^{\top(N)}, \end{bmatrix} \qquad \mathbf{\Phi} = \begin{bmatrix} \phi(\mathbf{x}^{(1)})^\top, \\ \phi(\mathbf{x}^{(2)})^\top, \\ \vdots, \\ \phi(\mathbf{x}^{(N)})^\top \end{bmatrix}$$

does basis expansion

Design matrix

Linear regression

$$\hat{\mathbf{y}} = \mathbf{X}\theta$$

Recall:

OLS

1. Pick your loss
2. Calculate the gradient
3. Set equal to zero
4. Solve!

Linear regression

$$\hat{\mathbf{y}} = \mathbf{X}\theta + \epsilon$$

Maximum likelihood

Recall:

1. Epsilon is our noise model
2. Write out likelihood (iid)
3. Calculate the gradient (neg log likelihood)
4. Solve!

Bayesian linear regression

data (x) is not
a RV

$$p(\theta|\mathbf{X}, \mathbf{y}) = p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)$$

We see the above which is how we may write out the posterior of Bayesian linear regression. However, it is important for us to remember the design matrix is not a random variable, so this is an abuse of notation, but one that is generally accepted.

Linear regression prior

posterior

likelihood

prior

$$p(\theta|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)$$

Isotropic Gaussian: $$p(\theta) = \mathcal{N}(0, \alpha\mathbf{I})$$

params centered around 0 — no initial bias

We discussed some philosophies behind prior selection, and we will see one more in our next lecture, but for now we leave this topic to those who want to take the probabilistic inference course

Likelihood model

$$p(\theta|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\theta)p(\theta)$$

Gaussian obser. noise:

$$\mathbf{y} = \mathbf{X}\theta + \epsilon$$

$$= \mathcal{N}(\mathbf{X}\theta, \sigma^2)$$

Why MLE for linear regression?

0 mean, σ^2 variance

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

$$\hat{y}^{(i)} \sim \mathcal{N}(\underbrace{\mathbf{x}^{(i)\top}\theta}_{\text{mean}}, \underbrace{\sigma^2\mathbf{I}}_{\text{var}})$$

Given an additive isotropic Gaussian noise model, we can simply recenter our Gaussian at the prediction and now we have a nice form for our probabilistic model

We have seen this in our MLE/MAP derivation

$$\underset{n \times m}{y} = \underset{n \times p}{X}\;\underset{p \times m}{\theta}$$

Combining the above

$$p(\theta|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)$$
$$\propto \mathcal{N}(\mathbf{X}\theta, \sigma^2)\mathcal{N}(0, \alpha\mathbf{I})$$

Now we see that we have a product of Gaussian distributions, so we know that we can apply the technique of equating coefficients, but this time we need to use linear algebra skills.

$N(\cdots) \times N(\cdots) = N(?)$

Equating coefficients in BLR

$$p(\mathbf{y}|\mathbf{X},\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left(\frac{(\mathbf{y} - \overbrace{\phi(\mathbf{X})\theta}^{\mu})^\top(\mathbf{y} - \phi(\mathbf{X})\theta)}{2\sigma^2}\right)$$

$$p(\theta) = \frac{1}{(2\pi\tau^2)^{k/2}} \cdot \exp\left(\frac{\theta^\top\theta}{2\tau^2}\right) \quad \leftarrow \text{mean} = 0$$

What techniques have we seen in MLE and and last lecture that will make this algebra simpler?

Equating coefficients in BLR

$$p(\mathbf{y}|\mathbf{X},\theta) = \cancel{\frac{1}{(2\pi\sigma^2)^{N/2}}} \cdot \exp\left(\frac{(\mathbf{y}-\phi(\mathbf{X})\theta)^\top(\mathbf{y}-\phi(\mathbf{X})\theta)}{2\sigma^2}\right)$$

$$p(\theta) = \cancel{\frac{1}{(2\pi\tau^2)^{k/2}}} \cdot \exp\left(\frac{\theta^\top\theta}{2\tau^2}\right)$$

What techniques have we seen in MLE and and last lecture that will make this algebra simpler?

Proportionality

Equating coefficients in BLR

$$p(\mathbf{y}|\mathbf{X},\theta) = \cancel{\frac{1}{(2\pi\sigma^2)^{N/2}}} \cdot \cancel{\exp}\left(\frac{(\mathbf{y}-\phi(\mathbf{X})\theta)^\top(\mathbf{y}-\phi(\mathbf{X})\theta)}{2\sigma^2}\right)$$

$$p(\theta) = \cancel{\frac{1}{(2\pi\tau^2)^{k/2}}} \cdot \cancel{\exp}\left(\frac{\theta^\top\theta}{2\tau^2}\right)$$

What techniques have we seen in MLE and and last lecture that will make this algebra simpler?

Proportionality

Working with log likelihood/log density

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X}, \mathbf{y})) \propto \frac{1}{2\sigma^2}((\mathbf{y} - \mathbf{X}\theta)^\top(\mathbf{y} - \mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta)$$

$$= \frac{1}{2\sigma^2}\left(\theta^\top\mathbf{X}^\top\mathbf{X}\theta - 2\mathbf{y}^\top\mathbf{X}\theta + \mathbf{y}^\top\mathbf{y}\right) + \frac{1}{\tau^2}\theta^\top\theta$$

What techniques have we seen in MLE and and last lecture that will make this algebra simpler?

Proportionality

Working with log likelihood/log density

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta)$$

$$= \frac{1}{2\sigma^2}\left(\theta^\top\mathbf{X}^\top\mathbf{X}\theta - 2\mathbf{y}^\top\mathbf{X}\theta + \mathbf{y}^\top\mathbf{y}\right) + \frac{1}{\tau^2}\theta^\top\theta$$

$$= \frac{1}{2\sigma^2}\theta^\top\mathbf{X}^\top\mathbf{X}\theta + \frac{2}{2\sigma^2}\mathbf{y}^\top\mathbf{X}\theta - \frac{1}{2\sigma^2}\mathbf{y}^\top\mathbf{y} + \frac{1}{2\tau^2}\theta^\top\theta$$

Crunching densities!

Equating coefficients in BLR

$$
\begin{aligned}
\log(p(\theta|\mathbf{X},\mathbf{y})) &\propto \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta) \\
&= \frac{1}{2\sigma^2}\left(\theta^\top\mathbf{X}^\top\mathbf{X}\theta - 2\mathbf{y}^\top\mathbf{X}\theta + \mathbf{y}^\top\mathbf{y}\right) + \frac{1}{\tau^2}\theta^\top\theta \\
&= \frac{1}{2\sigma^2}\theta^\top\mathbf{X}^\top\mathbf{X}\theta + \frac{2}{2\sigma^2}\mathbf{y}^\top\mathbf{X}\theta - \frac{1}{2\sigma^2}\mathbf{y}^\top\mathbf{y} + \frac{1}{2\tau^2}\theta^\top\theta \\
&\propto \sigma^2\theta^\top\mathbf{X}^\top\mathbf{X}\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y} + \tau^2\theta^\top\theta
\end{aligned}
$$

$x \sim N(\mu, \sigma^2) \to e^{-(x-\mu)^T \Sigma^{-1}(x-\mu)} \to$

$(x-\mu)^T\Sigma^{-1}(x-\mu) \to (x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu) \propto x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu + \underbrace{2\sigma^2\mu^T\Sigma^{-1}\mu}_{\text{const}}$

$\to x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu$ - get in this form

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta)$$

$$\propto \sigma^2\theta^\top\mathbf{X}^\top\mathbf{X}\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y} + \tau^2\theta^\top\theta$$

Before, I said that the expression at this point *looked* like a normal distribution to me. Do we see the same structure here?

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta)$$

$$\propto \sigma^2 \boxed{\theta^\top\mathbf{X}^\top\mathbf{X}\theta} + 2\sigma^2\mathbf{y}^\top\mathbf{X}\boxed{\theta} - 2\sigma^2\mathbf{y}^\top\mathbf{y} + \boxed{\tau^2\theta^\top\theta}$$

Before, I said that the expression at this point *looked* like a normal distribution to me. Do we see the same structure here?

Quadratic in theta! So we know how we need to proceed with equating coefficients.

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta)$$

$$\propto \sigma^2 \boxed{\theta^\top\mathbf{X}^\top\mathbf{X}\theta} + 2\sigma^2\mathbf{y}^\top\mathbf{X}\boxed{\theta} - 2\sigma^2\mathbf{y}^\top\mathbf{y} + \boxed{\tau^2\theta^\top\theta}$$

$$= \sigma^2\theta^\top\mathbf{X}^\top\mathbf{X}\theta + \tau^2\theta^\top\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \frac{1}{2\sigma^2}((\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta)$$

$$\propto \sigma^2\boxed{\theta^\top\mathbf{X}^\top\mathbf{X}\theta} + 2\sigma^2\mathbf{y}^\top\mathbf{X}\boxed{\theta} - 2\sigma^2\mathbf{y}^\top\mathbf{y} + \boxed{\tau^2\theta^\top\theta}$$

$$= \theta^\top(\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

Now this again looks a lot like a normal distribution to me!

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X}, \mathbf{y})) \propto \theta^\top(\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\log\left(\mathcal{N}(\theta; \mu, \Sigma)\right) \propto (\theta - \mu)^\top\Sigma^{-1}(\theta - \mu)$$

Now this again looks a lot like a normal distribution to me!

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \theta^\top(\sigma^2\mathbf{X}^\top\mathbf{X}+\tau^2\mathbf{I})\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\begin{aligned}\log\left(\mathcal{N}(\theta;\mu,\Sigma)\right) &\propto (\theta-\mu)^\top\Sigma^{-1}(\theta-\mu)\\ &= \theta^\top\Sigma^{-1}\theta - 2\theta^\top\Sigma^{-1}\mu + \mu^\top\Sigma^{-1}\mu\end{aligned}$$

Now this again looks a lot like a normal distribution to me!

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X}, \mathbf{y})) \propto \theta^\top(\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\begin{aligned}\log\big(\mathcal{N}(\theta; \mu, \Sigma)\big) &\propto (\theta - \mu)\Sigma^{-1}(\theta - \mu) \\ &= \theta^\top\Sigma^{-1}\theta - 2\theta^\top\Sigma^{-1}\mu + \cancel{\mu^\top\Sigma^{-1}\mu} \\ &= \theta^\top\Sigma^{-1}\theta - 2\theta^\top\Sigma^{-1}\mu + \text{const.}\end{aligned}$$

Now this again looks a lot like a normal distribution to me!

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \theta^\top \boxed{(\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})}\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\log(\mathcal{N}(\theta;\mu,\Sigma)) \propto \theta^\top \boxed{\Sigma^{-1}}\theta - 2\theta^\top\Sigma^{-1}\mu$$

invertible $\longrightarrow$ $$\Sigma^{-1} = (\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})$$

It is very natural to see what we must set our covariance matrix to.

Aside: the inverse of the covariance matrix is called the precision matrix

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \theta^\top \boxed{(\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})}\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\log(\mathcal{N}(\theta;\mu,\Sigma)) \propto \theta^\top \boxed{\Sigma^{-1}}\theta - 2\theta^\top \boxed{\Sigma^{-1}}\boxed{\mu}$$

$$\Sigma^{-1} = (\sigma^2\mathbf{X}^\top\mathbf{X} + \tau^2\mathbf{I})$$

Now we need to set our mean equal to something that makes two equations above work out

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \theta^\top(\sigma^2\mathbf{X}^\top\mathbf{X}+\tau^2\mathbf{I})\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\log(\mathcal{N}(\theta;\mu,\Sigma)) \propto \theta^\top\Sigma^{-1}\theta - 2\theta^\top\boxed{\Sigma^{-1}}\boxed{\mu}$$

$$2\theta^\top\Sigma^{-1}\mu = 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta$$

$$\mu = \sigma^2\Sigma\mathbf{X}^\top\mathbf{y}$$

$(2\theta^\top\Sigma^{-1})^{-1}\,2\sigma^2 y^\top X\theta$

$= \Sigma\theta^{-\top}$

$$\Sigma^{-1} = (\sigma^2\mathbf{X}^\top\mathbf{X}+\tau^2\mathbf{I})$$

Equating coefficients in BLR

$$\log(p(\theta|\mathbf{X},\mathbf{y})) \propto \theta^\top(\sigma^2\mathbf{X}^\top\mathbf{X}+\tau^2\mathbf{I})\theta + 2\sigma^2\mathbf{y}^\top\mathbf{X}\theta - 2\sigma^2\mathbf{y}^\top\mathbf{y}$$

$$\log\left(\mathcal{N}(\theta;\mu,\Sigma)\right) \propto \theta^\top\Sigma^{-1}\theta - 2\theta^\top\boxed{\Sigma^{-1}}\boxed{\mu}$$

$$\mu = \sigma^2\Sigma\mathbf{X}^\top\mathbf{y} \qquad \Sigma^{-1} = (\sigma^2\mathbf{X}^\top\mathbf{X}+\tau^2\mathbf{I})$$

Hurray! We have computed our closed form Bayesian posterior in the case of linear regression:

$$\mathcal{N}\left(\theta; \underbrace{\sigma^2\Sigma\phi(\mathbf{X})^\top\mathbf{y}}_{\mu}, \underbrace{(\sigma^2\phi(\mathbf{X})^\top\phi(\mathbf{X})+\tau^2\mathbf{I})^{-1}}_{\Sigma^{-1}}\right)$$

Diving into some steps: inversion

$$\mu = \sigma^2 \Sigma \mathbf{X}^\top \mathbf{y} \qquad \boxed{\Sigma^{-1} = (\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})}$$

Hurray! We have computed our closed form Bayesian posterior in the case of linear regression:

$$\mathcal{N}\left(\theta; \sigma^2 \boxed{\Sigma} \phi(\mathbf{X})^\top \mathbf{y}, (\sigma^2 \phi(\mathbf{X})^\top \phi(\mathbf{X}) + \tau^2 \mathbf{I})^{-1}\right)$$

In prior lectures we have skipped the step of showing that we *can* invert a matrix and I have hand-waved it. So here I will take a second to prove it to give you an example.

Diving into some steps: inversion

$$\mu = \sigma^2 \Sigma \mathbf{X}^\top \mathbf{y} \qquad \boxed{\Sigma^{-1} = (\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})}$$

First, let's abstract away some of the complexity by identifying what objects we are dealing with.

$$\mathbf{A} + \alpha \mathbf{I}$$

Diving into some steps: inversion

$$\mu = \sigma^2 \Sigma \mathbf{X}^\top \mathbf{y}$$

$$\boxed{\Sigma^{-1} = (\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})}$$

First, let's abstract away some of the complexity by identifying what objects we are dealing with.

$$\mathbf{A} + \alpha \mathbf{I}$$

as $X^T X$

Square, positive semi-definite matrix

$\lambda \geq 0$

Constant times the identity matrix

Diving into some steps: inversion

$$\Sigma^{-1} = (\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})$$

We know that if we have a positive definite matrix, then it is invertible. Positive definiteness is the condition when all our eigenvalues are positive.

$$\mathbf{A} + \alpha \mathbf{I}$$

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

'v' is an eigenvector of A, and we know that lambda is non-negative

$$c\mathbf{I}\mathbf{v} = c\mathbf{v}$$

every vector is an eigenvector of the identity, so we can write this down for 'v'

Diving into some steps: inversion

$$\Sigma^{-1} = (\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})$$

Using the eigenvector v, we can show it is an eigenvector of the resulting matrix A + cI and that the eigenvalue is strictly positive.

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v} \qquad c\mathbf{I}\mathbf{v} = c\mathbf{v}$$

$$(\mathbf{A} + c\mathbf{I})\mathbf{v} = \mathbf{A}\mathbf{v} + c\mathbf{I}\mathbf{v} = \lambda\mathbf{v} + c\mathbf{v} = (\lambda + \boxed{c})\mathbf{v}$$

Strictly positive!

Since it is positive definite, we can invert this matrix, so our posterior is well defined.

Reasoning about mean

$$\mu = \sigma^2 \Sigma \mathbf{X}^\top \mathbf{y} \qquad \Sigma^{-1} = (\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})$$

Closed form Bayesian posterior

$$\mathcal{N}\left(\theta; \sigma^2 \Sigma \mathbf{X}^\top \mathbf{y}, (\sigma^2 \mathbf{X}^\top \mathbf{X} - \tau^2 \mathbf{I})^{-1}\right)$$

$$\begin{aligned}
\mu &= \sigma^2 \Sigma \mathbf{X}^\top \mathbf{y} \\
&= \sigma^2 \left(\sigma^2 \mathbf{X}^\top \mathbf{X} + \tau^2 \mathbf{I})^{-1}\right) \mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{X}^\top \mathbf{X} + \frac{\tau^2}{\sigma^2} \mathbf{I})^{-1}) \mathbf{X}^\top \mathbf{y} \quad = MAP
\end{aligned}$$

Method 2: Joint Gaussian

$$p(\theta, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \end{bmatrix}, \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \theta], \mathbb{V}[\mathbf{y}] \end{bmatrix} \right)$$

Computing the posterior via joint Gaussian

Theorem 3.1. ***Marginalization*** *Given a Gaussian random variable:*

$$\mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}; \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{a,a}, \Sigma_{a,b} \\ \Sigma_{b,a}, \Sigma_{b,b} \end{bmatrix}\right)$$

The marginal distribution of a *is given by:*

$$p(a) = \mathcal{N}(a; \mu_a, \Sigma_{a,a})$$

Computing the posterior via joint Gaussian

Theorem 3.2. ***Conditioning*** *Given a Gaussian random variable:*

$$\mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}; \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{a,a}, \Sigma_{a,b} \\ \Sigma_{b,a}, \Sigma_{b,b} \end{bmatrix}\right)$$

formula in equation sheet

The conditional distribution of $p(a|b)$ *is given by:*

$$\begin{aligned} p(a|b) &= \mathcal{N}(a; \mu_{a|b}, \Sigma_{a|b}), \\ \mu_{a|b} &= \mu_a + \Sigma_{a,b}\Sigma_{b,b}^{-1}(b - \mu_b) \\ \Sigma_{a|b} &= \Sigma_{a,a} - \Sigma_{a,b}\Sigma_{b,b}^{-1}\Sigma_{b,a} \end{aligned}$$

Method 2: Joint Gaussian

$$p(\theta, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \mathbf{y} \end{bmatrix} ; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \end{bmatrix} , \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \theta], \mathbb{V}[\mathbf{y}] \end{bmatrix} \right)$$

$$\mathbb{E}[\theta] = \mathbf{0}$$

isotropic

$$\mathbb{E}[\mathbf{y}] = \mathbf{0}$$

linear transformation (preserves origin) so centered at 0

Method 2: Joint Gaussian

$$p(\theta, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \end{bmatrix}, \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \theta], \mathbb{V}[\mathbf{y}] \end{bmatrix} \right)$$

$$\mathbb{E}[\theta] = \mathbf{0} \qquad \mathbb{V}[\theta] = \mathbf{I}_n$$

$V(X\theta + \epsilon)$: $V(X\theta) + V(\epsilon)$ as X & ϵ indep

$$\mathbb{E}[\mathbf{y}] = \mathbf{0} \qquad \mathbb{V}[\mathbf{y}] = \mathbb{V}[\mathbf{X}\theta] + \mathbb{V}[\epsilon]$$

By independence

$$= \mathbf{X}^\top \mathbf{X} + \sigma^2$$

By linearity of expectation

$V(X\theta)$: $E\left[(X\theta - \underbrace{E[X\theta]}_{E[X\theta] = XE[\theta] = 0})(X\theta - E[X\theta])^T\right]$

$= E[(X\theta)(X\theta)^T] = E[X\theta\theta^T X^T]$

$Var(AX)$: $A\, Var(X)\, A^T$

Method 2: Joint Gaussian

$$p(\theta, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \mathbf{y} \end{bmatrix} ; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \end{bmatrix} , \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \theta], \mathbb{V}[\mathbf{y}] \end{bmatrix} \right)$$

$$\mathbb{E}[\theta] = \mathbf{0}$$
$$\mathbb{E}[\mathbf{y}] = \mathbf{0}$$

$$\mathbb{V}[\theta] = \mathbf{I}_n$$
$$\begin{aligned} \mathbb{V}[\mathbf{y}] &= \mathbb{V}[\mathbf{X}\theta] + \mathbb{V}[\epsilon] \\ &= \mathbf{X}^\top \mathbf{X} + \sigma^2 \end{aligned}$$

$E(XY) - E(X)E(Y)$

$$\begin{aligned} \mathbb{C}[\theta, \mathbf{y}] &= \mathbb{E}[\theta y^\top] - \overbrace{\mathbb{E}[\theta]}^{=0}\overbrace{\mathbb{E}[\mathbf{y}]}^{=0} \\ &= \mathbb{E}[\theta(\mathbf{X}\theta + \epsilon)^\top] \\ &= \mathbb{E}[\theta\theta^\top \mathbf{X}^\top] + \underbrace{\mathbb{E}[\theta]}_{=0}\mathbb{E}[\epsilon] \\ &= \mathbf{X}^\top \end{aligned}$$

$E(\theta y^T) = E(\theta(X\theta + \epsilon)^T) = E(\theta\theta^T X^T) + E(\theta\epsilon^T)$

$= E$

Method 2: Joint Gaussian

$$p(\theta, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \mathbf{y} \end{bmatrix} ; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \end{bmatrix} , \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \theta], \mathbb{V}[\mathbf{y}] \end{bmatrix} \right)$$

$$\mathbb{E}[\theta] = \mathbf{0}$$
$$\mathbb{E}[\mathbf{y}] = \mathbf{0}$$

$$\begin{aligned} \mathbb{V}[\theta] &= \mathbf{I}_n \\ \mathbb{V}[\mathbf{y}] &= \mathbb{V}[\mathbf{X}\theta] + \mathbb{V}[\epsilon] \\ &= \mathbf{X}^\top \mathbf{X} + \sigma^2 \end{aligned}$$

$$\begin{aligned} \mathbb{C}[\theta, \mathbf{y}] &= \mathbb{E}[\theta y^\top] - \mathbb{E}[\theta]\mathbb{E}[\mathbf{y}] \\ &= \mathbb{E}[\theta(\mathbf{X}\theta + \epsilon)^\top] \\ &= \mathbb{E}[\theta\theta^\top \mathbf{X}^\top] + \cancel{\mathbb{E}}[\theta]\cancel{\mathbb{E}}[\epsilon] \\ &= \mathbf{X}^\top \end{aligned}$$

Theta is normally distributed and so the expectation is the covariance which is the identity

Method 2: Joint Gaussian

$$p(\theta, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \mathbf{y} \end{bmatrix} ; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \end{bmatrix} , \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \theta], \mathbb{V}[\mathbf{y}] \end{bmatrix} \right)$$

$$\mu_{\theta|y} = \mu_\theta + \Sigma_{\theta,y}\Sigma_{y,y}^{-1}(y - \mu_y)$$

$$\Sigma_{\theta|y} = \Sigma_\theta - \Sigma_{\theta,y}\Sigma_y^{-1}\Sigma_{y,\theta}$$

Using our formulas from the conditioning identity we had a few slides ago, we can plug in the values we computed for means and covariances to get our posterior

Method 2: Joint Gaussian

$$\mu_{\theta|y} = \mu_\theta + \Sigma_{\theta,y}\Sigma_y^{-1}(y - \mu_y) \qquad \Sigma_{\theta|y} = \Sigma_\theta - \Sigma_{\theta,y}\Sigma_y^{-1}\Sigma_{y,\theta}$$

Plugging in the computed values for our joint Gaussian above we get:

$$p(\theta|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\theta; \underbrace{\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}}_{\mu_{\theta|y}}, \underbrace{\mathbf{I} - \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top - \sigma^2\mathbf{I}\right)\mathbf{X}}_{\Sigma_{\theta|y}}\right)$$

Method 2: Joint Gaussian

Plugging in the computed values for our joint Gaussian above we get:

$$p(\theta|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\theta; \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, \mathbf{I} - \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top - \sigma^2\mathbf{I}\right)\mathbf{X}\right)$$

When we crunched out densities we got:

$$p(\theta|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\theta; \left(\mathbf{X}^\top\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}, \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{I}\right)^{-1}\right)$$

Method 2: Joint Gaussian

Plugging in the computed values for our joint Gaussian above we get:

$$p(\theta|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\theta; \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, \mathbf{I} - \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top - \sigma^2\mathbf{I}\right)\mathbf{X}\right)$$

When we crunched out densities we got:

$$p(\theta|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\theta; \left(\mathbf{X}^\top\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}, \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{I}\right)^{-1}\right)$$

Comparing the two posteriors

$$AB, A \in \mathbb{R}^{M \times N}, B \in \mathbb{R}^{N \times M} \qquad O(NM^2)$$

$$A^{-1}, A \in \mathbb{R}^{N \times N} \qquad O(N^3)$$

Additional exercise:
Compute the computational complexity of the two posterior we computed. Joint is worse when feature dimension is much larger than number of datapoints

Posterior Predictive Distribution

We have now covered Bayesian inference in linear regression models, however, having a distribution over parameters also slightly complicates how we make predictions

Given a new input x*, how do we make a prediction with a distribution over our model parameters?

Posterior Predictive Distribution

We have now covered Bayesian inference in linear regression models, however, having a distribution over parameters also slightly complicates how we make predictions

p(y): p(y|x) p(x)

We marginalize over our parameters! This is also known as Bayesian model averaging.

$$p(y|\mathbf{x}^*) = p(y|\mathbf{x}^*, \theta)p(\theta|\mathbf{X}, \mathbf{y})$$

unseen

p(y|x

training data

Posterior Predictive Distribution

We marginalize over our parameters! This is also known as Bayesian model averaging.

integrate out θ

$$p(y|\mathbf{x}^*) = p(y|\mathbf{x}^*, \theta)p(\theta|\mathbf{X}, \mathbf{y})$$

$$p(y|\mathbf{x}^*) = \int_{\Theta} p(y|\mathbf{x}^*, \theta)p(\theta|\mathbf{X}, \mathbf{y})d\theta$$

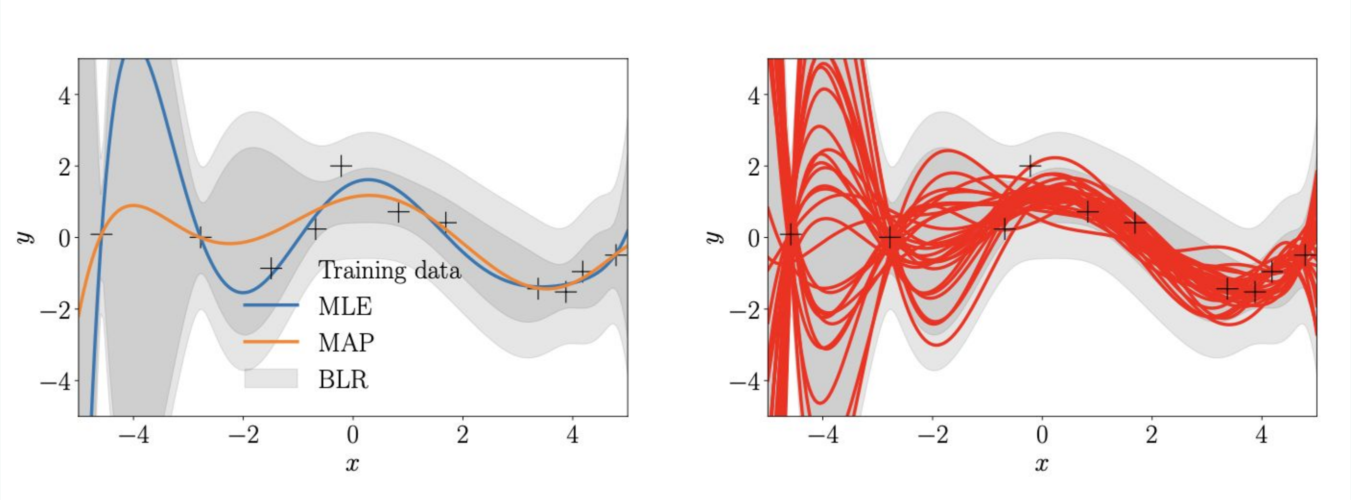
$$= \mathbb{E}_{p(\theta|\mathbf{X},\mathbf{y})}[p(y|\mathbf{x}^*, \theta)]$$

Posterior Predictive Distribution

$$p(y|\mathbf{x}^*) = \int_\Theta p(y|\mathbf{x}^*, \theta)p(\theta|\mathbf{X}, \mathbf{y})d\theta$$
$$= \mathbb{E}_{p(\theta|\mathbf{X},\mathbf{y})}[p(y|\mathbf{x}^*, \theta)]$$

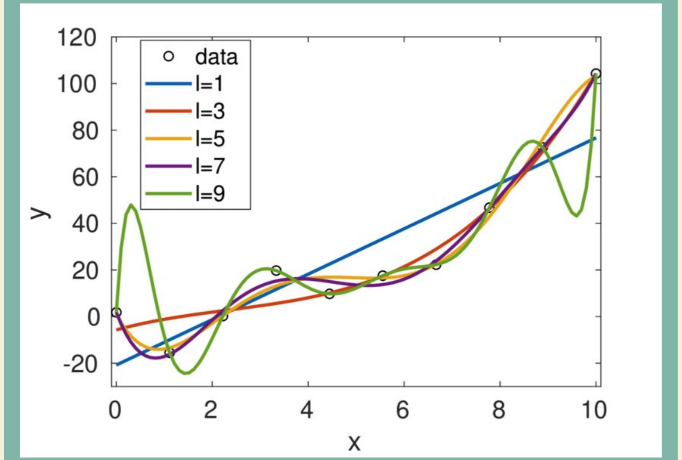
Linking back to our discussion of epistemic and aleatory uncertainty, we can see that the epistemic uncertainty is best captured as the variance about this expectation while the aleatoric uncertainty is the noise from our likelihood

Posterior Predictive Distribution



Linking back to our discussion of epistemic and aleatory uncertainty, we can see that the epistemic uncertainty is best captured as the variance about this expectation while the aleatoric uncertainty is the noise from our likelihood

Marginal likelihood

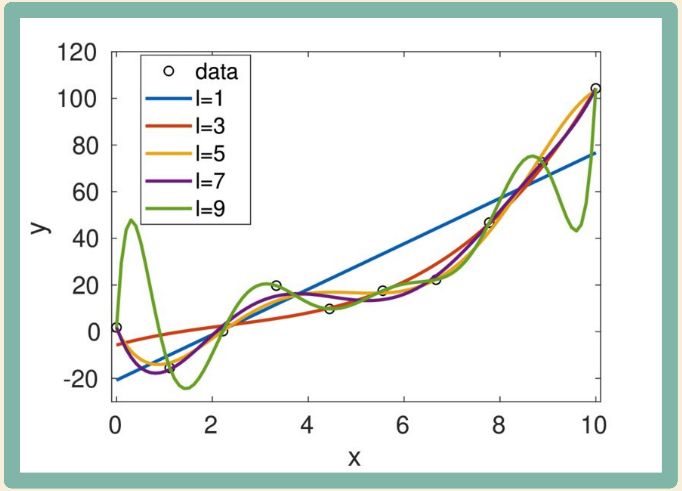


In this course, we have talked about several different learning paradigms: MLE, MAP, Bayesian inference. But we have not yet talked about how to choose between different models.

Marginal likelihood

$$p(\mathcal{D}) = \int_{\Theta} p(\mathcal{D}|\theta)p(\theta)d\theta$$

prob of data given model

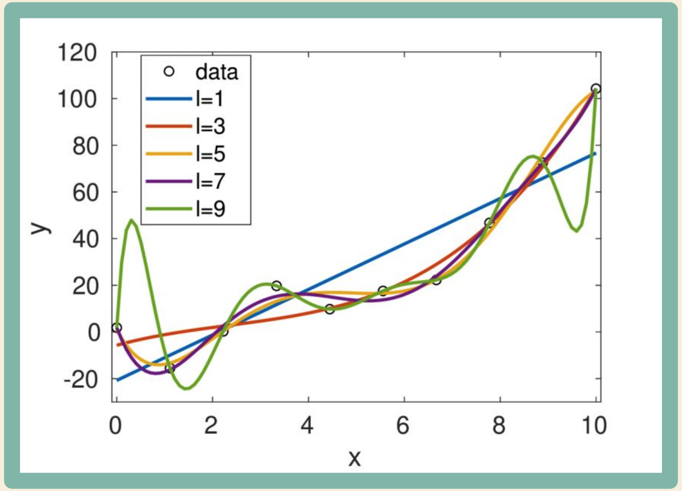


So, how can we use this marginal likelihood to compare models?

Marginal likelihood

$$p(\mathcal{D}) = \int_{\Theta} p(\mathcal{D}|\theta)p(\theta)d\theta$$

So, how can we use this marginal likelihood to compare models?



Marginal likelihood

Marginal likelihood of a linear model

$$\int_{\Theta} p(\mathcal{D}|\theta^{(1)})p(\theta^{(1)})d\theta^{(1)}$$

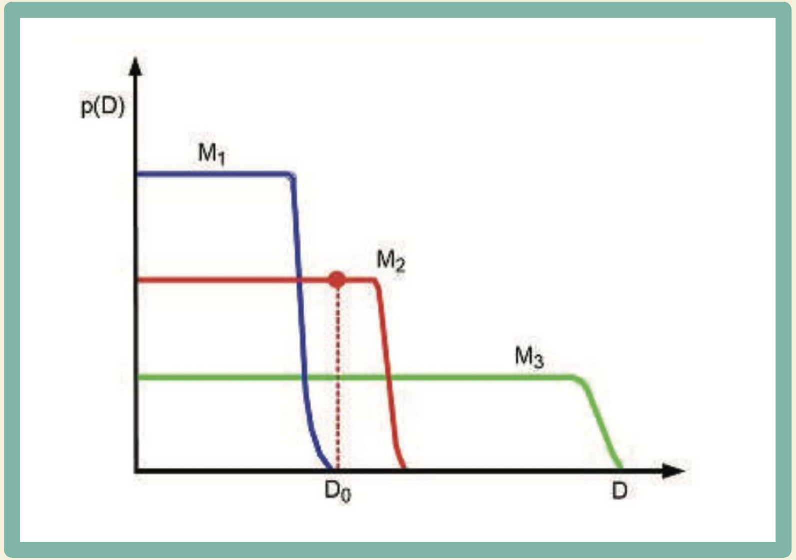
Marginal likelihood of a quadratic model

$$\int_{\Theta} p(\mathcal{D}|\theta^{(2)})p(\theta^{(2)})d\theta^{(2)}$$

Marginal likelihood of a cubic model

$$\int_{\Theta} p(\mathcal{D}|\theta^{(3)})p(\theta^{(3)})d\theta^{(3)}$$

Marginal likelihood



On the x-axis we order datasets by their complexity. Given that all probability distributions sum to 1, the complex models will have to spread their mass thinly over all of the complex datasets they can fit. So, by selecting a model with the highest marginal likelihood, we are in essence picking a model that is "just right" for the data complexity we observe.

Next lecture: Testing and Validation of ML