Lecture 6: MAP Estimation

Priors · Posteriors · Regularization Connection

Previously, on Lecture 5...

MoM: Set population moments = sample moments. Simple but can give impossible values.

MLE: $\hat{\theta} = \arg\max \ell(\theta)$. Pick the θ that makes the data most probable.

Properties: Consistent, asymptotically normal, efficient, invariant.

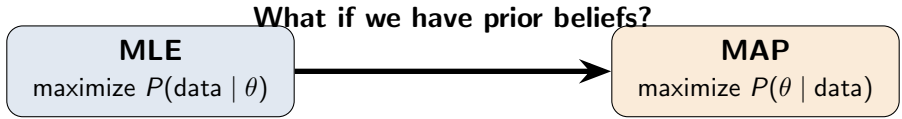
MLE = ML: Gaussian noise $\rightarrow$ MSE loss. Bernoulli $\rightarrow$ cross-entropy.

But: MLE can overfit with small n or flexible models.

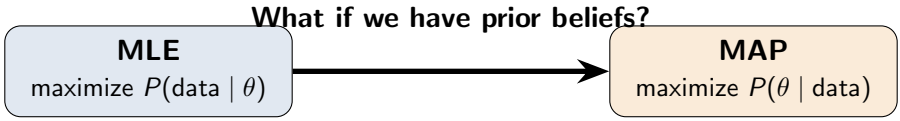
Today: What if we have **prior knowledge** about θ?

Can we do better than MLE by incorporating beliefs *before* seeing data?

Where We Are


What if we have prior beliefs?
MLE
maximize $P(\text{data} \mid \theta)$
MAP
maximize $P(\theta \mid \text{data})$
Lecture 5
This lecture

Where We Are



Why? A surgeon performs 3 operations, all successful. MLE:
$\hat{p} = 3/3 = 100\%$.
Would you bet your life on that? You *know* no one is perfect — that's **prior knowledge**.

Bayes' Theorem for Parameters

From likelihood to posterior

Before the Formula: An Example

A surgeon performs 3 heart surgeries — all 3 are successful.

What is their true success rate p?

Before the Formula: An Example

A surgeon performs 3 heart surgeries — all 3 are successful.

What is their true success rate p?

Prior belief

"Typical surgeons succeed about 85% of the time"

$p \approx 0.85$

By combining **prior knowledge** with data, we get a more sensible estimate. Bayes' theorem tells us *exactly* how to combine them. Let's see the formula.

Before the Formula: An Example

A surgeon performs 3 heart surgeries — all 3 are successful.

What is their true success rate p?

Prior belief

"Typical surgeons succeed about 85% of the time"

$p \approx 0.85$

Data says

3 out of 3 successful

MLE: $\hat{p} = 3/3 = 1.0$

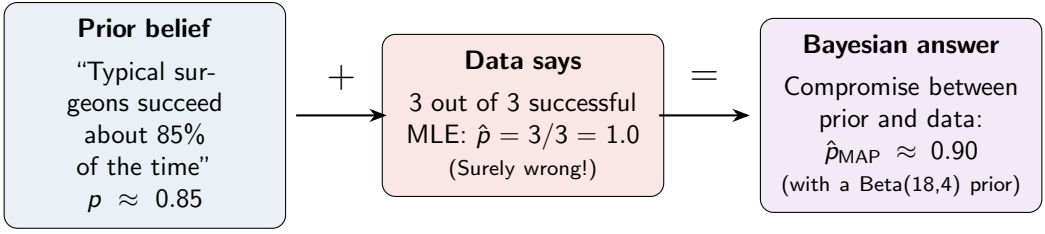
(Surely wrong!)

By combining **prior knowledge** with data, we get a more sensible estimate. Bayes' theorem tells us *exactly* how to combine them. Let's see the formula.

Before the Formula: An Example

A surgeon performs 3 heart surgeries — all 3 are successful.

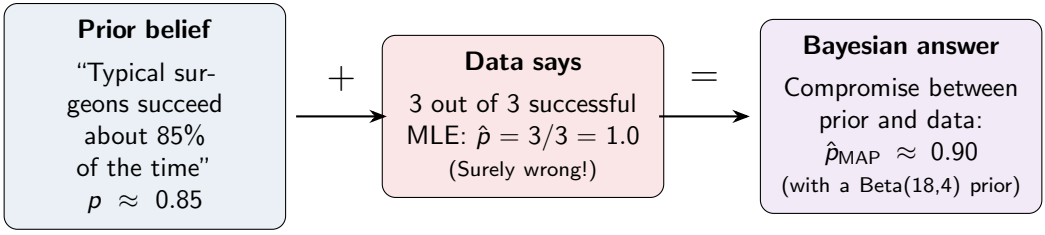
What is their true success rate p?



Before the Formula: An Example

A surgeon performs 3 heart surgeries — all 3 are successful.

What is their true success rate p?



By combining **prior knowledge** with data, we get a more sensible estimate. Bayes' theorem tells us *exactly* how to combine them. Let's see the formula.

Bayes' Theorem for Parameters

$$\underbrace{P(\theta \mid \text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data} \mid \theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\text{data})}_{\text{evidence}}}$$

Bayes' Theorem for Parameters

$$\underbrace{P(\theta \mid \text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data} \mid \theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\text{data})}_{\text{evidence}}}$$

Prior $P(\theta)$ — what you believed about θ *before* any data

Likelihood $P(\text{data} \mid \theta)$ — how probable is the data *if* θ were the true value

Posterior $P(\theta \mid \text{data})$ — your *updated* belief after combining both

Evidence $P(\text{data})$ — normalizing constant (same for all θ, so we can drop it)

Bayes' Theorem for Parameters

$$\underbrace{P(\theta \mid \text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data} \mid \theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\text{data})}_{\text{evidence}}}$$

Prior $P(\theta)$ — what you believed about θ *before* any data

Likelihood $P(\text{data} \mid \theta)$ — how probable is the data *if* θ were the true value

Posterior $P(\theta \mid \text{data})$ — your *updated* belief after combining both

Evidence $P(\text{data})$ — normalizing constant (same for all θ, so we can drop it)

Or simply: posterior $\propto$ likelihood $\times$ prior

($\propto$ means "proportional to" — equal up to a constant. We drop $P(\text{data})$ since it doesn't change $\arg\max_\theta$.)

The Three Ingredients

Prior $P(\theta)$

What you believed *before* seeing data

"Coins are usually close to fair"

The Three Ingredients

Prior $P(\theta)$

What you believed *before* seeing data

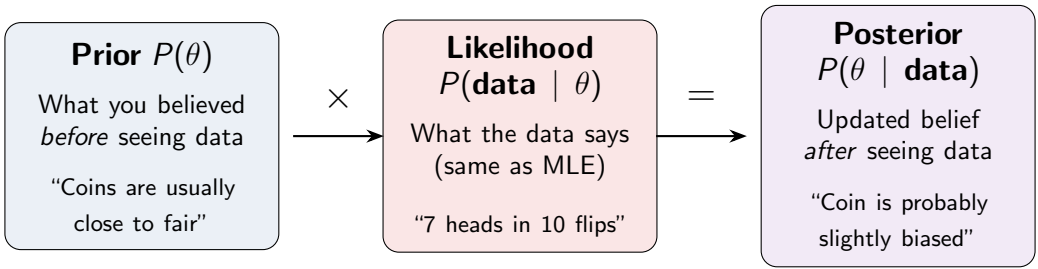
"Coins are usually close to fair"

Likelihood $P(\textbf{data} \mid \theta)$

What the data says (same as MLE)

"7 heads in 10 flips"

The Three Ingredients



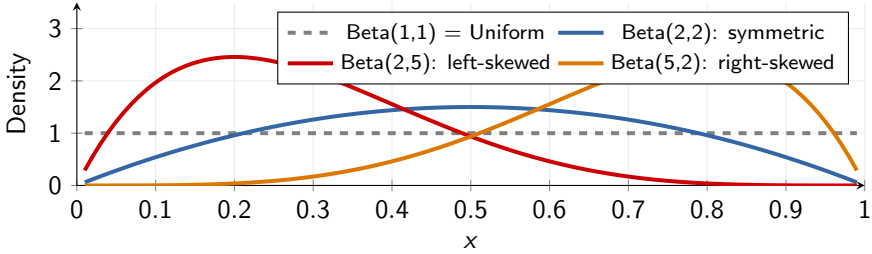
Meet the Beta Distribution

Before choosing a prior, let's meet a distribution that lives on $[0, 1]$ — perfect for modeling probabilities, proportions, or rates.

$$X \sim \text{Beta}(\alpha, \beta): \quad f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \le x \le 1$$

where $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the **Beta function** (just a normalizing constant)

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}, \quad \text{Var} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \quad \text{Mode} = \frac{\alpha-1}{\alpha+\beta-2}$$
$$(\alpha, \beta > 1)$$



The Beta Distribution: A Prior for Probabilities

We need a prior distribution for $p \in [0, 1]$. The **Beta distribution** is the natural choice:

$$\text{Beta}(\alpha, \beta): \quad f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \leq p \leq 1$$

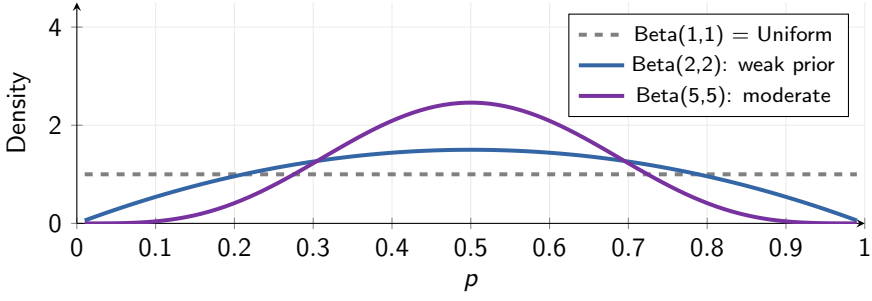
Mean $= \alpha/(\alpha+\beta)$. Higher $\alpha + \beta \Rightarrow$ more concentrated.

The Beta Distribution: A Prior for Probabilities

We need a prior distribution for $p \in [0, 1]$. The **Beta distribution** is the natural choice:

$$\text{Beta}(\alpha, \beta): \quad f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \le p \le 1$$

Mean $= \alpha/(\alpha+\beta)$. Higher $\alpha + \beta \Rightarrow$ more concentrated.

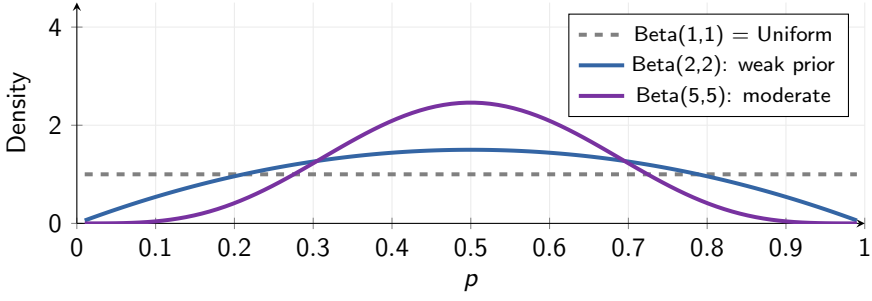


The Beta Distribution: A Prior for Probabilities

We need a prior distribution for $p \in [0, 1]$. The **Beta distribution** is the natural choice:

$$\text{Beta}(\alpha, \beta): \quad f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \le p \le 1$$

Mean $= \alpha/(\alpha+\beta)$. Higher $\alpha + \beta \Rightarrow$ more concentrated.



Why Beta? (1) Defined on $[0, 1]$, matching probabilities. (2) Very flexible shape. (3) **Conjugate** to the Binomial — posterior updates are trivial.

How Did We Compute the Posterior?

Prior: $p \sim \text{Beta}(5, 5)$: $\quad P(p) \propto p^{\alpha-1}(1-p)^{\beta-1} = p^4(1-p)^4$

How Did We Compute the Posterior?

Prior: $p \sim \text{Beta}(5, 5)$: $\quad P(p) \propto p^{\alpha-1}(1-p)^{\beta-1} = p^4(1-p)^4$

Likelihood: 7 heads in 10 flips: $\quad P(\text{data} \mid p) \propto p^7(1-p)^3$

How Did We Compute the Posterior?

Prior: $p \sim \text{Beta}(5, 5)$: $\quad P(p) \propto p^{\alpha-1}(1-p)^{\beta-1} = p^4(1-p)^4$

Likelihood: 7 heads in 10 flips: $\quad P(\text{data} \mid p) \propto p^7(1-p)^3$

Posterior = likelihood $\times$ prior:

$$P(p \mid \text{data}) \;\propto\; p^7(1-p)^3 \cdot p^4(1-p)^4 \;=\; p^{7+4}(1-p)^{3+4} \;=\; p^{11}(1-p)^7$$

How Did We Compute the Posterior?

Prior: $p \sim \text{Beta}(5,5)$: $\quad P(p) \propto p^{\alpha-1}(1-p)^{\beta-1} = p^4(1-p)^4$

Likelihood: 7 heads in 10 flips: $\quad P(\text{data} \mid p) \propto p^7(1-p)^3$

Posterior = likelihood $\times$ prior:

$$P(p \mid \text{data}) \;\propto\; p^7(1-p)^3 \cdot p^4(1-p)^4 \;=\; p^{7+4}(1-p)^{3+4} \;=\; p^{11}(1-p)^7$$

This is a Beta$(12, 8)$ distribution: just **add the exponents**!

$$\text{Beta}(\underbrace{5}_{\alpha} + \underbrace{7}_{k},\; \underbrace{5}_{\beta} + \underbrace{3}_{n-k}) = \text{Beta}(12, 8)$$

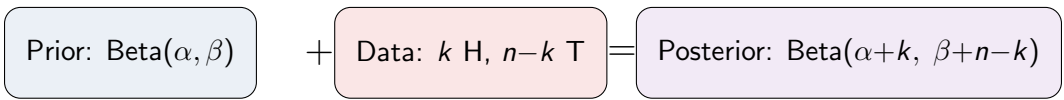
The prior parameters (α, β) simply absorb the data counts $(k, n-k)$.

This "addition rule" is why Beta–Binomial is called a **conjugate pair**.

Conjugate Priors as Pseudo-Observations

Conjugate prior = a prior whose posterior is in the *same distribution family*.

Beta is conjugate to Binomial: prior is Beta $\Rightarrow$ posterior is also Beta. No integrals needed!



Conjugate Priors as Pseudo-Observations

Conjugate prior = a prior whose posterior is in the *same distribution family*.

Beta is conjugate to Binomial: prior is Beta $\Rightarrow$ posterior is also Beta. No integrals needed!

Prior: Beta(α, β) + Data: k H, $n-k$ T = Posterior: Beta$(\alpha+k,\ \beta+n-k)$

The prior acts like "fake data" you've already seen:

Beta(α, β) = pretend you already observed $\alpha-1$ heads and $\beta-1$ tails.

Beta$(5, 5)$: "I've seen 4H and 4T" (8 pseudo-observations).

After 7H, 3T (10 real obs): posterior = Beta$(12, 8)$ = "11H, 7T out of 18 total."

As $n \to \infty$, the pseudo-observations become negligible $\Rightarrow$ posterior $\to$ likelihood $\Rightarrow$ MAP $\to$ MLE.

Common Conjugate Pairs

Likelihood	**Conjugate Prior**	**Posterior**	**Pseudo-data**
Binomial(n, p)	Beta(α, β)	Beta$(\alpha+k,\ \beta+n-k)$	$\alpha-1$ H, $\beta-1$ T
Poisson(λ)	Gamma(a, b)	Gamma$(a+\sum x_i,\ b+n)$	a events in b time
$N(\mu, \sigma_0^2)$	$N(m, \tau^2)$	N(precision-weighted mean, $\dots$)*	1 obs of value m
Exp(λ)	Gamma(a, b)	Gamma$(a+n,\ b+\sum x_i)$	a events in b time

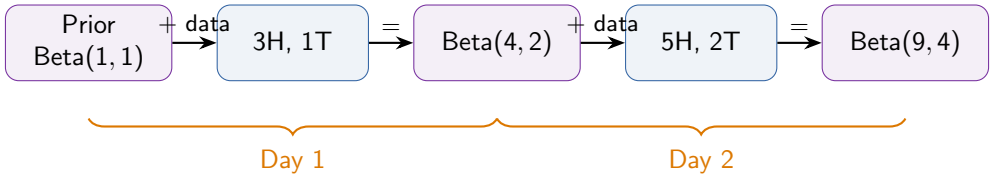
*Full formula derived later: $\hat{\mu} = w\bar{X} + (1-w)m$, a precision-weighted average (see "MAP for a Normal Mean" slide).

Pattern: if the likelihood is in the exponential family, there is always a conjugate prior.

The posterior update just adds the sufficient statistics.

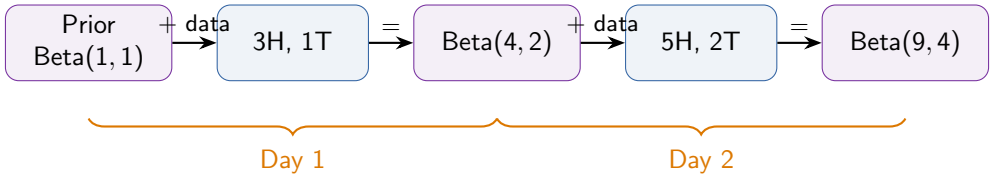
Bayesian Updating Is Sequential

A key Bayesian insight: **today's posterior becomes tomorrow's prior**.



Bayesian Updating Is Sequential

A key Bayesian insight: **today's posterior becomes tomorrow's prior**.



Order doesn't matter! Processing all 8H, 3T at once gives the same Beta(9, 4).

Bayesian updating is **consistent** — you can pause and resume at any time.

Beyond Conjugacy (Preview)

Conjugate priors make the math easy: posterior = same family, just update the parameters.

But what if the prior and likelihood aren't conjugate?

- The posterior $P(\theta \mid \text{data})$ may have **no closed form**
- We can still **compute MAP** (just an optimization problem — find the peak)
- For the **full posterior**, we need numerical methods:
  MCMC (Markov Chain Monte Carlo) — draw samples from the posterior
  Variational inference — approximate the posterior with a simpler distribution

Beyond Conjugacy (Preview)

Conjugate priors make the math easy: posterior = same family, just update the parameters.

But what if the prior and likelihood aren't conjugate?

- The posterior $P(\theta \mid \text{data})$ may have **no closed form**
- We can still **compute MAP** (just an optimization problem — find the peak)
- For the **full posterior**, we need numerical methods:
  MCMC (Markov Chain Monte Carlo) — draw samples from the posterior
  Variational inference — approximate the posterior with a simpler distribution

Why bother with the full posterior? It gives us more than a point estimate:

Posterior predictive: $P(x_{\text{new}} \mid \text{data}) = \int P(x_{\text{new}} \mid \theta)\, P(\theta \mid \text{data})\, d\theta$

This averages predictions over all plausible θ, automatically accounting for **uncertainty**.

Today we focus on **MAP** (the mode) — the simplest and most practical Bayesian point estimate.

MAP Estimation

Maximum **A** **P**osteriori

The mode of the posterior

MAP = Mode of the Posterior

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta \; P(\theta \mid \text{data}) = \arg\max_\theta \big[\ell(\theta) + \log P(\theta)\big]$$

Maximize: log-likelihood + log-prior

MLE: $\arg\max_\theta \; \ell(\theta)$ $\xrightarrow{+\text{ prior}}$ MAP: $\arg\max_\theta \; \ell(\theta) + \log P(\theta)$

MAP = Mode of the Posterior

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta \; P(\theta \mid \text{data}) = \arg\max_\theta \big[\ell(\theta) + \log P(\theta)\big]$$

Maximize: log-likelihood + log-prior

MLE: $\arg\max_\theta \; \ell(\theta)$ $\xrightarrow{+\text{ prior}}$ MAP: $\arg\max_\theta \; \ell(\theta) + \log P(\theta)$

MAP = MLE with an extra penalty/bonus term from the prior.

MAP Step by Step: Finding the Mode

We computed earlier that the posterior is $\text{Beta}(12, 8)$ for our coin example.

MAP = the **mode** (peak) of this posterior.

MAP Step by Step: Finding the Mode

We computed earlier that the posterior is $\text{Beta}(12, 8)$ for our coin example.

MAP = the **mode** (peak) of this posterior.

Step 1: Write the log-posterior.

$$\log P(p \mid \text{data}) = 11 \log p + 7 \log(1 - p) + \text{const}$$

MAP Step by Step: Finding the Mode

We computed earlier that the posterior is $\text{Beta}(12, 8)$ for our coin example.

MAP = the **mode** (peak) of this posterior.

Step 1: Write the log-posterior.

$$\log P(p \mid \text{data}) = 11 \log p + 7 \log(1 - p) + \text{const}$$

Step 2: Differentiate and set to zero.

$$\frac{d}{dp} \log P = \frac{11}{p} - \frac{7}{1-p} = 0 \quad \Rightarrow \quad 11(1-p) = 7p \quad \Rightarrow \quad p = \frac{11}{18}$$

MAP Step by Step: Finding the Mode

We computed earlier that the posterior is Beta$(12, 8)$ for our coin example.

MAP $=$ the **mode** (peak) of this posterior.

Step 1: Write the log-posterior.

$$\log P(p \mid \text{data}) = 11 \log p + 7 \log(1-p) + \text{const}$$

Step 2: Differentiate and set to zero.

$$\frac{d}{dp} \log P = \frac{11}{p} - \frac{7}{1-p} = 0 \quad \Rightarrow \quad 11(1-p) = 7p \quad \Rightarrow \quad p = \frac{11}{18}$$

General formula for Beta(α', β'):

$$\boxed{\hat{p}_{\text{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2}}$$

MAP Step by Step: Finding the Mode

We computed earlier that the posterior is Beta$(12, 8)$ for our coin example.

MAP $=$ the **mode** (peak) of this posterior.

Step 1: Write the log-posterior.

$$\log P(p \mid \text{data}) = 11 \log p + 7 \log(1-p) + \text{const}$$

Step 2: Differentiate and set to zero.

$$\frac{d}{dp} \log P = \frac{11}{p} - \frac{7}{1-p} = 0 \quad \Rightarrow \quad 11(1-p) = 7p \quad \Rightarrow \quad p = \frac{11}{18}$$

General formula for Beta(α', β'):

$$\boxed{\hat{p}_{\text{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2}}$$

For our example: $\hat{p}_{\text{MAP}} = \frac{12-1}{12+8-2} = \frac{11}{18} \approx 0.611$

Compare: $\hat{p}_{\text{MLE}} = \frac{7}{10} = 0.700$ (data alone, no prior)

MAP vs Posterior Mean vs Full Bayesian

MAP gives the **mode** of the posterior. But there are other ways to summarize it:

MAP

Mode of posterior

$\arg\max_\theta P(\theta \mid \text{data})$

Point estimate.
Connects to
regularization.

Our coin example (posterior = Beta(12,8)):

$$\text{MAP} = \frac{11}{18} \approx 0.611 \qquad \text{Posterior mean} = \frac{12}{20} = 0.600$$

Here the difference is small. For **skewed** posteriors (e.g., small n, asymmetric prior), mode and mean can differ substantially. Full Bayesian uses the whole distribution — including its spread for **uncertainty quantification**.

MAP vs Posterior Mean vs Full Bayesian

MAP gives the **mode** of the posterior. But there are other ways to summarize it:

MAP

Mode of posterior

$\arg\max_\theta P(\theta \mid \text{data})$

Point estimate.
Connects to regularization.

Posterior mean

$\mathbb{E}[\theta \mid \text{data}]$

Point estimate.
Minimizes MSE.
Often preferred.

Our coin example (posterior = Beta(12,8)):

$$\text{MAP} = \frac{11}{18} \approx 0.611 \qquad \text{Posterior mean} = \frac{12}{20} = 0.600$$

Here the difference is small. For **skewed** posteriors (e.g., small n, asymmetric prior), mode and mean can differ substantially. Full Bayesian uses the whole distribution — including its spread for **uncertainty quantification**.

MAP vs Posterior Mean vs Full Bayesian

MAP gives the **mode** of the posterior. But there are other ways to summarize it:

MAP

Mode of posterior
$\arg\max_\theta P(\theta \mid \text{data})$

Point estimate.
Connects to regularization.

Posterior mean

$\mathbb{E}[\theta \mid \text{data}]$

Point estimate.
Minimizes MSE.
Often preferred.

Full Bayesian

Use the *entire* posterior $P(\theta \mid \text{data})$

No information lost.
Gives uncertainty.

MAP vs Posterior Mean vs Full Bayesian

MAP gives the **mode** of the posterior. But there are other ways to summarize it:

MAP

Mode of posterior
$\arg\max_\theta P(\theta \mid \text{data})$

Point estimate.
Connects to regularization.

Posterior mean

$\mathbb{E}[\theta \mid \text{data}]$

Point estimate.
Minimizes MSE.
Often preferred.

Full Bayesian

Use the *entire* posterior $P(\theta \mid \text{data})$

No information lost.
Gives uncertainty.

Our coin example (posterior = Beta(12,8)):

$$\text{MAP} = \frac{11}{18} \approx 0.611 \qquad \text{Posterior mean} = \frac{12}{20} = 0.600$$

Here the difference is small. For **skewed** posteriors (e.g., small n, asymmetric prior), mode and mean can differ substantially. Full Bayesian uses the whole distribution — including its spread for **uncertainty quantification**.

Full Bayesian: How Does It Work?

Point estimates (MLE, MAP) pick *one* θ and predict with it.

Full Bayesian averages predictions over *all possible* θ, weighted by posterior probability.

Posterior predictive distribution:

$$P(x_{\text{new}} \mid \text{data}) = \int P(x_{\text{new}} \mid \theta) \cdot P(\theta \mid \text{data})\, d\theta$$

For each possible θ: make a prediction, then average them weighted by how plausible each θ is given the data.

Full Bayesian: How Does It Work?

Point estimates (MLE, MAP) pick *one* θ and predict with it.
Full Bayesian averages predictions over *all possible* θ, weighted by posterior probability.

Posterior predictive distribution:

$$P(x_{\text{new}} \mid \text{data}) = \int P(x_{\text{new}} \mid \theta) \cdot P(\theta \mid \text{data})\, d\theta$$

For each possible θ: make a prediction, then average them weighted by how plausible each θ is given the data.

Example: Coin with posterior $p \sim \text{Beta}(12, 8)$. Will the next flip be heads?

MAP / MLE approach:

Pick $\hat{p} = 0.611$, predict:
$P(H) = \hat{p} = 0.611$

Uses one value of p.
Ignores uncertainty in $\hat{p}$.

Full Bayesian:

Average over all p:
$P(H \mid \text{data}) = \int_0^1 p \cdot f(p \mid \text{data})\, dp$
$= E[p \mid \text{data}] = \frac{12}{20} = 0.600$

Uses entire posterior.

Full Bayesian: How Does It Work?

Point estimates (MLE, MAP) pick *one* θ and predict with it.
Full Bayesian averages predictions over *all possible* θ, weighted by posterior probability.

Posterior predictive distribution:

$$P(x_{\text{new}} \mid \text{data}) = \int P(x_{\text{new}} \mid \theta) \cdot P(\theta \mid \text{data})\, d\theta$$

For each possible θ: make a prediction, then average them weighted by how plausible each θ is given the data.

Example: Coin with posterior $p \sim \text{Beta}(12, 8)$. Will the next flip be heads?

MAP / MLE approach:

Pick $\hat{p} = 0.611$, predict:
$P(H) = \hat{p} = 0.611$

Uses one value of p.
Ignores uncertainty in $\hat{p}$.

Full Bayesian:

Average over all p:
$P(H \mid \text{data}) = \int_0^1 p \cdot f(p \mid \text{data})\, dp$
$= E[p \mid \text{data}] = \frac{12}{20} = 0.600$

Uses entire posterior.

MAP Under Different Priors

Same data (7 heads in 10 flips), but different priors:

Prior	**Posterior**	**MAP estimate**	**Pull toward 0.5**
$\text{Beta}(1,1) = \text{Uniform}$	$\text{Beta}(8,4)$	$7/10 = 0.700$	None ($=$ MLE)
$\text{Beta}(2,2)$	$\text{Beta}(9,5)$	$8/12 = 0.667$	Weak
$\text{Beta}(5,5)$	$\text{Beta}(12,8)$	$11/18 = 0.611$	Moderate
$\text{Beta}(50,50)$	$\text{Beta}(57,53)$	$56/108 = 0.519$	Strong

MAP Under Different Priors

Same data (7 heads in 10 flips), but different priors:

Prior	**Posterior**	**MAP estimate**	**Pull toward 0.5**
$\text{Beta}(1,1) = \text{Uniform}$	$\text{Beta}(8,4)$	$7/10 = 0.700$	None ($=$ MLE)
$\text{Beta}(2,2)$	$\text{Beta}(9,5)$	$8/12 = 0.667$	Weak
$\text{Beta}(5,5)$	$\text{Beta}(12,8)$	$11/18 = 0.611$	Moderate
$\text{Beta}(50,50)$	$\text{Beta}(57,53)$	$56/108 = 0.519$	Strong

Stronger prior (higher $\alpha + \beta$) $\Rightarrow$ more pull toward the prior mean.
More data ($n \uparrow$) $\Rightarrow$ data overwhelms the prior $\Rightarrow$ MAP $\to$ MLE.
A **flat prior** (Uniform) gives MAP $=$ MLE: no prior $=$ no regularization.

MAP for a Normal Mean

Model: $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ (known σ_0^2), prior $\mu \sim N(m, \tau^2)$.

MAP for a Normal Mean

Model: $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ (known σ_0^2), prior $\mu \sim N(m, \tau^2)$.

Log-posterior:

$$\log P(\mu \mid \text{data}) = \underbrace{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i - \mu)^2}_{\text{log-likelihood}} + \underbrace{\left(-\frac{(\mu - m)^2}{2\tau^2}\right)}_{\text{log-prior}} + \text{const}$$

MAP for a Normal Mean

Model: $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ (known σ_0^2), prior $\mu \sim N(m, \tau^2)$.

Log-posterior:

$$\log P(\mu \mid \text{data}) = \underbrace{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i - \mu)^2}_{\text{log-likelihood}} + \underbrace{\left(-\frac{(\mu - m)^2}{2\tau^2}\right)}_{\text{log-prior}} + \text{const}$$

Take derivative, set to 0:

$$\frac{n(\bar{X} - \mu)}{\sigma_0^2} - \frac{\mu - m}{\tau^2} = 0$$

MAP for a Normal Mean

Model: $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ (known σ_0^2), prior $\mu \sim N(m, \tau^2)$.

Log-posterior:

$$\log P(\mu \mid \text{data}) = \underbrace{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu)^2}_{\text{log-likelihood}} + \underbrace{\left(-\frac{(\mu-m)^2}{2\tau^2}\right)}_{\text{log-prior}} + \text{const}$$

Take derivative, set to 0:

$$\frac{n(\bar{X}-\mu)}{\sigma_0^2} - \frac{\mu-m}{\tau^2} = 0$$

Solve:

$$\hat{\mu}_{\text{MAP}} = \frac{\frac{n}{\sigma_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}} \cdot \bar{X} + \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}} \cdot m = w \cdot \bar{X} + (1-w) \cdot m$$

A **weighted average** of the sample mean $\bar{X}$ and the prior mean m.

More data ($n \uparrow$) $\Rightarrow w \to 1 \Rightarrow$ MAP $\to$ MLE. Stronger prior ($\tau^2 \downarrow$) $\Rightarrow w \to 0 \Rightarrow$ MAP $\to m$.

Choosing the Prior

Informative prior

Encodes real knowledge:
"Coins are close to fair"
$\rightarrow$ Beta$(50, 50)$

Stronger pull toward prior belief.
Fewer data needed.

Non-informative priors: Beta$(1, 1)$ = Uniform. Sounds "objective" but can be problematic — the result depends on how you parameterize!

Jeffreys prior $P(\theta) \propto \sqrt{I(\theta)}$ is invariant to reparameterization (uses Fisher info from Lecture 4).

Example: For Bernoulli, $I(p) = \frac{1}{p(1-p)}$ (Lecture 4), so Jeffreys prior $\propto p^{-1/2}(1-p)^{-1/2} =$ **Beta**$(\frac{1}{2}, \frac{1}{2})$ — a U-shape that concentrates near 0 and 1, expressing "I don't know, but extreme values aren't ruled out."

Choosing the Prior

Informative prior

Encodes real knowledge:
"Coins are close to fair"
$\rightarrow$ Beta$(50, 50)$

Stronger pull toward prior belief.
Fewer data needed.

Weakly informative prior

Gentle constraint, avoids extremes:
"p is probably not 0 or 1"
$\rightarrow$ Beta$(2, 2)$

Regularizes without strong bias.
Often the practical default.

Choosing the Prior

Informative prior

Encodes real knowledge:
"Coins are close to fair"
$\rightarrow$ Beta$(50, 50)$

Stronger pull toward prior belief. Fewer data needed.

Weakly informative prior

Gentle constraint, avoids extremes:
"p is probably not 0 or 1"
$\rightarrow$ Beta$(2, 2)$

Regularizes without strong bias. Often the practical default.

Non-informative priors: Beta$(1, 1)$ = Uniform. Sounds "objective" but can be problematic — the result depends on how you parameterize!

Jeffreys prior $P(\theta) \propto \sqrt{I(\theta)}$ is invariant to reparameterization (uses Fisher info from Lecture 4).

Example: For Bernoulli, $I(p) = \frac{1}{p(1-p)}$ (Lecture 4), so Jeffreys prior $\propto p^{-1/2}(1-p)^{-1/2} =$ **Beta**$(\frac{1}{2}, \frac{1}{2})$ — a U-shape that concentrates near 0 and 1, expressing "I don't know, but extreme values aren't ruled out."

Regularization as MAP

Priors are penalties in disguise

What Is Regularization?

In machine learning, models can **overfit**: they memorize noise in the training data.

What Is Regularization?

In machine learning, models can **overfit**: they memorize noise in the training data.

Regularization adds a **penalty** on the model parameters to keep them small:

$$\hat{\theta} = \arg\min_{\theta} \Big[\underbrace{\text{Loss}(\theta)}_{\text{data fit}} + \underbrace{\lambda \cdot \text{Penalty}(\theta)}_{\text{keep } \theta \text{ small}} \Big]$$

What Is Regularization?

In machine learning, models can **overfit**: they memorize noise in the training data.

Regularization adds a **penalty** on the model parameters to keep them small:

$$\hat{\theta} = \arg\min_{\theta} \Big[\underbrace{\text{Loss}(\theta)}_{\text{data fit}} + \underbrace{\lambda \cdot \text{Penalty}(\theta)}_{\text{keep } \theta \text{ small}} \Big]$$

Two common penalties:

L2 (Ridge)

Penalty $= \|\theta\|_2^2 = \sum \theta_j^2$

Shrinks all coefficients toward 0

L1 (Lasso)

Penalty $= \|\theta\|_1 = \sum |\theta_j|$

Drives some coefficients to **exactly 0**

What Is Regularization?

In machine learning, models can **overfit**: they memorize noise in the training data.

Regularization adds a **penalty** on the model parameters to keep them small:

$$\hat{\theta} = \arg\min_{\theta} \Big[\underbrace{\text{Loss}(\theta)}_{\text{data fit}} + \underbrace{\lambda \cdot \text{Penalty}(\theta)}_{\text{keep } \theta \text{ small}} \Big]$$

Two common penalties:

L2 (Ridge)

Penalty $= \|\theta\|_2^2 = \sum \theta_j^2$

Shrinks all coefficients toward 0

L1 (Lasso)

Penalty $= \|\theta\|_1 = \sum |\theta_j|$

Drives some coefficients to **exactly 0**

λ controls the trade-off: $\lambda = 0$ means no penalty ($=$ MLE); large λ means heavy penalty.

In practice, λ is chosen by **cross-validation** (we'll see this in a later lecture).

Key insight: these penalties are secretly **priors in disguise**. Let's see why.

The Key Connection: Regularization = MAP

$$\text{MAP:} \quad \hat{\theta} \;=\; \arg\max_\theta \big[\ell(\theta) \;+\; \log P(\theta)\big]$$

is the same as

$$\text{Regularization:} \quad \hat{\theta} \;=\; \arg\min_\theta \big[-\ell(\theta) \;+\; \lambda \cdot \text{penalty}(\theta)\big]$$

The log-prior acts as a **penalty on the parameters**.

From Prior to Penalty: The Algebra

Assume: $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \text{prior } \theta_j \sim N(0, \tau^2).$

From Prior to Penalty: The Algebra

Assume: $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \text{prior } \theta_j \sim N(0, \tau^2).$

MAP objective: $\max_\theta \big[\ell(\theta) + \log P(\theta)\big]$

From Prior to Penalty: The Algebra

Assume: $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad$ prior $\theta_j \sim N(0, \tau^2)$.

MAP objective: $\max_\theta \left[\ell(\theta) + \log P(\theta)\right]$

$$\ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2 + \text{const}$$

$$\log P(\theta) = -\frac{\|\theta\|_2^2}{2\tau^2} + \text{const}$$

Combine and flip sign (max $\to$ min, drop constants):

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left[\underbrace{\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2}_{\text{data fit (= MSE loss)}} + \underbrace{\frac{1}{2\tau^2} \|\theta\|_2^2}_{\text{from the prior}} \right]$$

Set $\lambda = \sigma^2/\tau^2$: **small** τ^2 (tight prior) $\Rightarrow$ **large** λ (strong penalty).
This is exactly **Ridge regression**.

From Prior to Penalty: The Algebra

Assume: $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad$ prior $\theta_j \sim N(0, \tau^2)$.

MAP objective: $\max_\theta \left[\ell(\theta) + \log P(\theta)\right]$

$$\ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \theta)^2 + \text{const}$$

$$\log P(\theta) = -\frac{\|\theta\|_2^2}{2\tau^2} + \text{const}$$

From Prior to Penalty: The Algebra

Assume: $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad$ prior $\theta_j \sim N(0, \tau^2)$.

MAP objective: $\max_\theta \left[\ell(\theta) + \log P(\theta)\right]$

$$\ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2 + \text{const}$$

$$\log P(\theta) = -\frac{\|\theta\|_2^2}{2\tau^2} + \text{const}$$

Combine and flip sign (max $\to$ min, drop constants):

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left[\underbrace{\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2}_{\text{data fit (= MSE loss)}} + \underbrace{\frac{1}{2\tau^2} \|\theta\|_2^2}_{\text{from the prior}} \right]$$

From Prior to Penalty: The Algebra

Assume: $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad$ prior $\theta_j \sim N(0, \tau^2)$.

MAP objective: $\max_\theta \big[\ell(\theta) + \log P(\theta)\big]$

$$\ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2 + \text{const}$$

$$\log P(\theta) = -\frac{\|\theta\|_2^2}{2\tau^2} + \text{const}$$

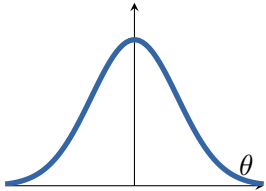
Combine and flip sign (max $\to$ min, drop constants):

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left[\underbrace{\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2}_{\text{data fit (= MSE loss)}} + \underbrace{\frac{1}{2\tau^2} \|\theta\|_2^2}_{\text{from the prior}} \right]$$

Set $\lambda = \sigma^2/\tau^2$: **small** τ^2 (tight prior) $\Rightarrow$ **large** λ (strong penalty).
This is exactly **Ridge regression**.

Gaussian Prior $\Leftrightarrow$ Ridge (L2) Regression

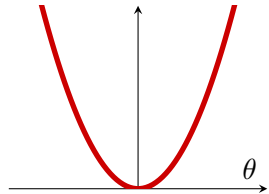
Gaussian prior



$P(\theta) = \mathcal{N}(0, \tau^2)$

$\Rightarrow$

L2 penalty



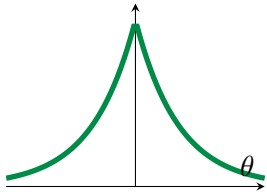
$-\log P(\theta) \propto \|\theta\|_2^2$

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left[\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2 + \lambda \|\theta\|_2^2\right]$$

This is exactly **Ridge regression**! $\lambda = \sigma^2/\tau^2$

Laplace Prior $\Leftrightarrow$ Lasso (L1) Regression

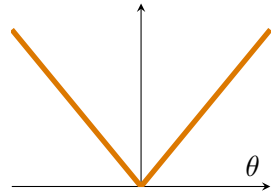
Laplace prior



$P(\theta) = \frac{1}{2b}\, e^{-|\theta|/b}$

$\Rightarrow$

L1 penalty



$-\log P(\theta) \propto \|\theta\|_1$

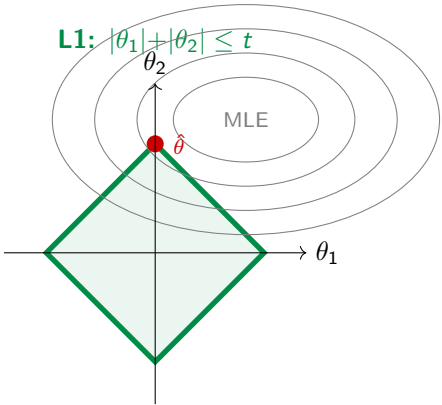
$$\hat{\theta}_{\text{MAP}} \;=\; \arg\min_\theta \left[\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2 + \lambda \|\theta\|_1\right]$$

This is exactly **Lasso regression**! Encourages **sparse** solutions ($\theta_j = 0$).

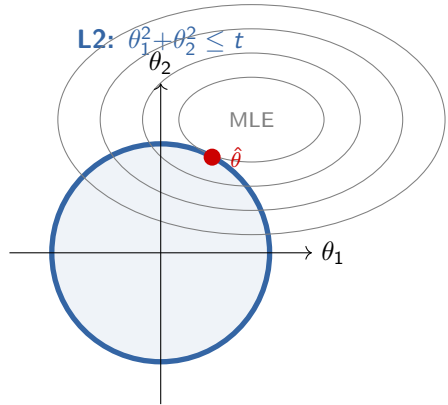
Why Does L1 Give Exact Zeros?

Setup: We minimize loss subject to $\|\theta\| \leq t$. The gray ellipses are **level curves** of the loss (centered at the unconstrained MLE). The colored shape is the **constraint region** (all θ allowed by the penalty).

The constrained solution is where the smallest ellipse **first touches** the constraint boundary.

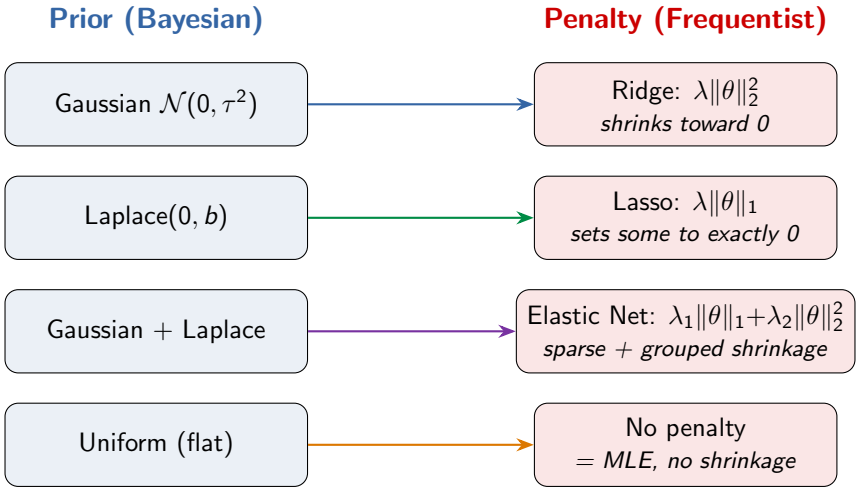


Ellipse hits a **corner**
$\Rightarrow \theta_1 = 0$ exactly



Ellipse hits **smooth edge**
$\Rightarrow$ both $\theta_j \neq 0$

The Regularization Map



Two Perspectives: Bayesian vs Frequentist

The Regularization Map reveals a deep duality between two schools of statistics:

Bayesian view

Parameters have *distributions*.

Start with a **prior** belief,
update with data via Bayes' rule.

"θ is probably near 0"
$\rightarrow$ Gaussian prior $\mathcal{N}(0, \tau^2)$

Result: posterior distribution.

Frequentist view

Parameters are *fixed* (unknown).

Estimate via optimization,
add **penalty** to avoid overfitting.

"Keep coefficients small"
$\rightarrow$ Ridge penalty $\lambda \|\theta\|_2^2$

Result: point estimate.

Two Perspectives: Bayesian vs Frequentist

The Regularization Map reveals a deep duality between two schools of statistics:

Bayesian view

Parameters have *distributions*.

Start with a **prior** belief,
update with data via Bayes' rule.

"θ is probably near 0"
$\rightarrow$ Gaussian prior $\mathcal{N}(0, \tau^2)$

Result: posterior distribution.

Frequentist view

Parameters are *fixed* (unknown).

Estimate via optimization,
add **penalty** to avoid overfitting.

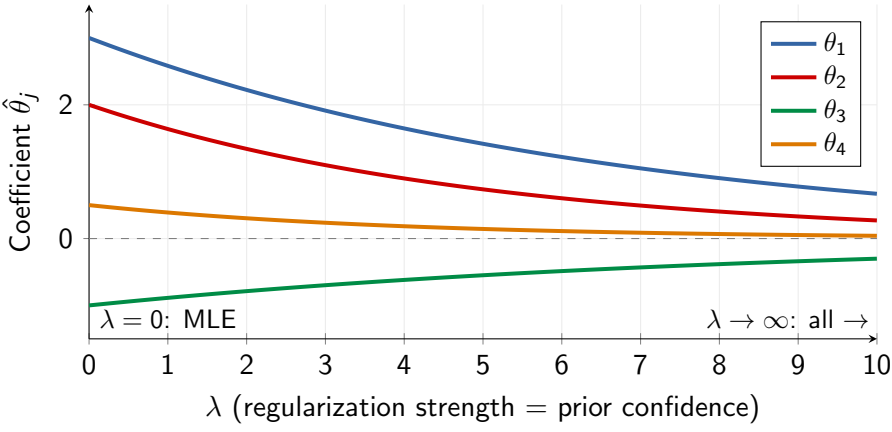
"Keep coefficients small"
$\rightarrow$ Ridge penalty $\lambda \|\theta\|_2^2$

Result: point estimate.

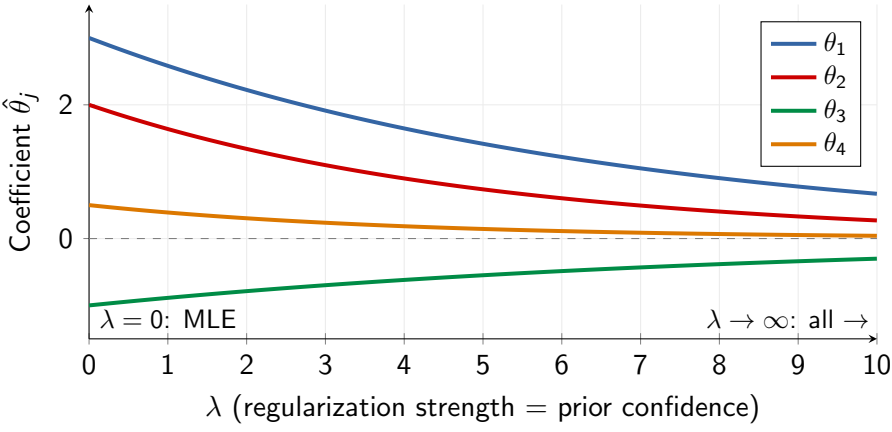
The punchline: they often give the *same answer*!

Gaussian prior $\Leftrightarrow$ Ridge. Laplace prior $\Leftrightarrow$ Lasso. MAP bridges both worlds.

Visualizing Ridge Shrinkage



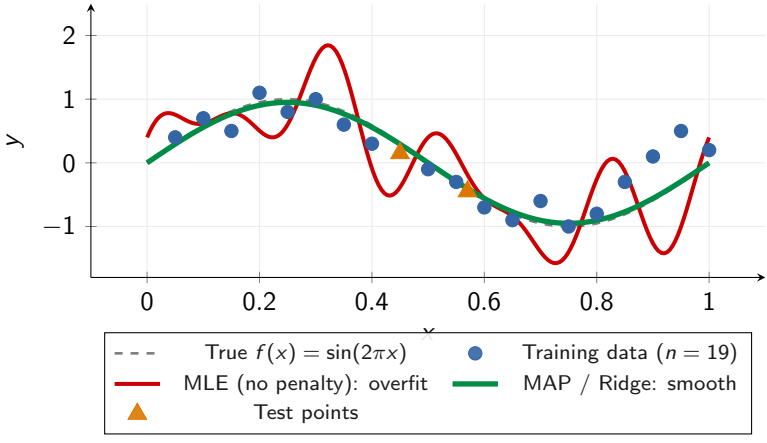
Visualizing Ridge Shrinkage



Increasing λ = stronger prior = more shrinkage = less overfitting (but more bias).

MLE vs MAP: The Overfitting Story

Fit a high-degree polynomial to noisy data (schematic). MLE uses all coefficients freely; MAP (Ridge) penalizes large ones.

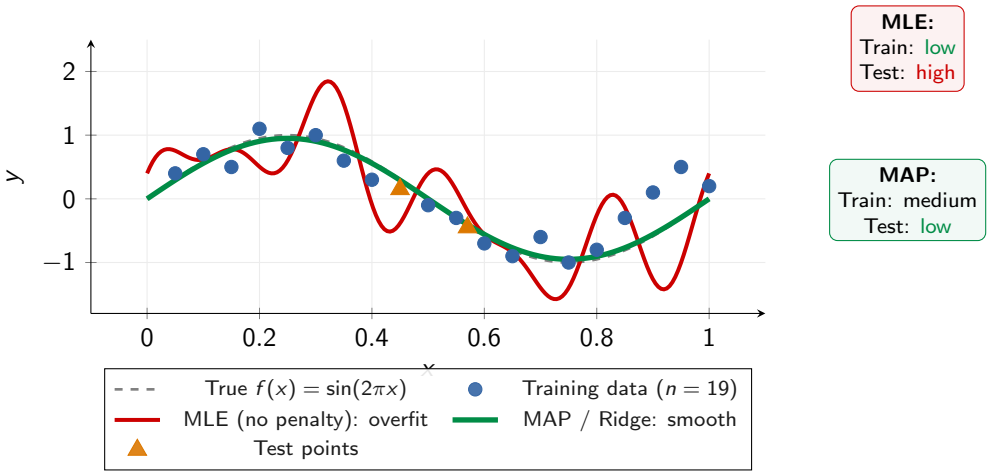


MLE:
Train: low
Test: high

MAP:
Train: medium
Test: low

MLE vs MAP: The Overfitting Story

Fit a high-degree polynomial to noisy data (schematic). MLE uses all coefficients freely; MAP (Ridge) penalizes large ones.



The prior says "coefficients should be small" $\Rightarrow$ smoother fit $\Rightarrow$ better **generalization**.

Summary: From MLE to MAP

Bayes' rule: posterior $\propto$ likelihood $\times$ prior. Update beliefs with data.

Conjugacy: Beta–Binomial, Normal–Normal. Prior acts like pseudo-observations.

MAP: $\hat{\theta} = \arg\max[\ell(\theta) + \log P(\theta)]$. MLE with a prior bonus.

Normal MAP: Weighted average $w\bar{X} + (1-w)m$. More data $\Rightarrow$ MAP $\to$ MLE.

Gaussian prior $\to$ Ridge: $-\log P(\theta) \propto \|\theta\|_2^2$. Shrinks toward 0.

Laplace prior $\to$ Lasso: $-\log P(\theta) \propto \|\theta\|_1$. Drives coefficients to exactly 0.

Flat prior $\to$ no penalty: MAP reduces to MLE. No prior = no regularization.

Bayesian = Frequentist: Gaussian prior $\Leftrightarrow$ Ridge, Laplace $\Leftrightarrow$ Lasso. Two views, same math.

Practical: Priors and Posteriors

1. **Coin bias estimation:**
   - Start with Beta(1,1), Beta(5,5), Beta(50,50) priors
   - Observe 7 heads in 10 flips
   - Plot prior, likelihood, and posterior for each
   - Compare the MAP estimates — how much does the prior pull?
2. **Ridge regression as MAP:**
   - Fit linear regression with $\lambda = 0, 0.1, 1, 10, 100$
   - Plot coefficients vs λ (shrinkage path)
   - Observe: larger λ = stronger prior = more shrinkage
3. **Visualize:** Plot the prior/likelihood/posterior for a simple 1D Normal with known σ^2, varying the prior variance τ^2

Homework

1. For $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ (known σ_0^2) with prior $\mu \sim N(m, \tau^2)$:
   derive the MAP estimator $\hat{\mu}_{\text{MAP}}$. Show it is a weighted average of $\bar{X}$ and m.
   What happens as $\tau^2 \to \infty$? As $n \to \infty$?

2. A coin is flipped 20 times with 14 heads. Compute the MAP estimate of p under:
   (a) Beta$(1,1)$, (b) Beta$(5,5)$, (c) Beta$(50,50)$ priors.
   Compare with the MLE. Which prior has the most influence?

3. Show that Ridge regression $\hat{\theta} = \arg\min \left[\|y - X\theta\|^2 + \lambda \|\theta\|^2 \right]$
   has the closed-form solution $\hat{\theta} = (X^\top X + \lambda I)^{-1} X^\top y$.
   Why does this always have a unique solution, even when $X^\top X$ is singular?

Recommended Visualizations & Resources (1/2)

Interactive: Bayesian Inference (R Psychologist)

`rpsychologist.com/d3/bayes` — interactive Bayesian two-sample t-test: adjust the prior, sample size, and effect size; watch the posterior, credible intervals, and Bayes factor update live.

Interactive: Seeing Theory — Bayesian Inference (Brown)

`seeing-theory.brown.edu/bayesian-inference` — drag priors to see posteriors update live. Beautiful animations for prior$\times$likelihood=posterior.

Video: 3Blue1Brown — Bayes' Theorem

`3blue1brown.com/lessons/bayes-theorem` — stunning animated walkthrough of how prior beliefs transform into posteriors. The geometric intuition directly maps to MAP.

Recommended Visualizations & Resources (2/2)

Interactive: Ridge & Lasso Visualization

`xavierbourretsicotte.github.io/ridge_lasso_visual.html` — contour plots of loss + penalty. See geometrically why Lasso gives sparse solutions.

Video: StatQuest — Bayes' Theorem, Clearly Explained

`youtu.be/9wCnvr7Xw4E` — step-by-step walkthrough of prior, likelihood, and posterior with clear examples. Great for building intuition.

Questions?

Next: Lecture 7 — Sampling distributions