Lecture 2b: MAP Estimation

Priors, Posteriors, and the Regularization Connection

Where We Are

What if we have prior beliefs?

MLE
maximize $P(\text{data} \mid \theta)$

Last lecture

MAP
maximize $P(\theta \mid \text{data})$

This lecture

Bayes' Theorem for Parameters

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

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

The Three Ingredients

Prior $P(\theta)$

What you believed *before* seeing data

"Coins are usually close to fair"

$\times$

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

What the data says (same as MLE)

"7 heads in 10 flips"

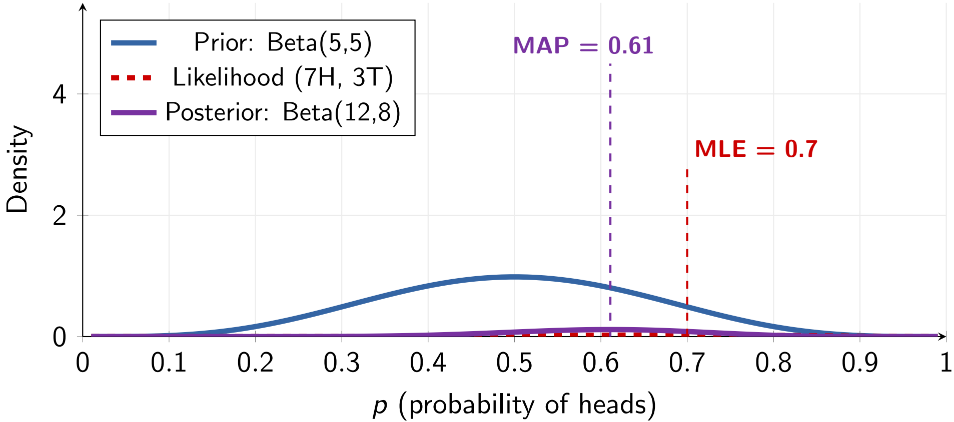
$=$

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

Updated belief *after* seeing data

"Coin is probably slightly biased"

Visualizing the Update: Coin Bias



Prior pulls the estimate from 0.7 toward 0.5. The posterior is a **compromise**.

MAP = Mode of the Posterior

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

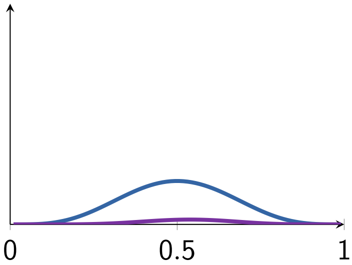
Maximize: log-likelihood + log-prior

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.

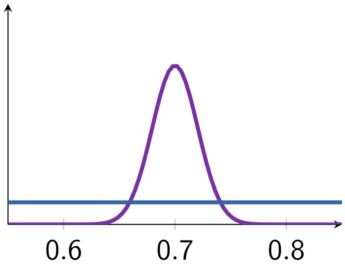
When Does the Prior Matter?

Small n (e.g., $n = 5$)



Prior has **large** influence

Large n (e.g., $n = 500$)



Prior has **negligible** influence

With enough data, the likelihood dominates $\Rightarrow$ MAP $\approx$ MLE.
The prior is "washed out" by the data.

The Key Connection: Regularization $=$ MAP

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

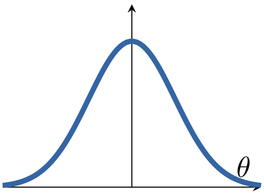
is the same as

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

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

Gaussian Prior $\Leftrightarrow$ Ridge (L2) Regression

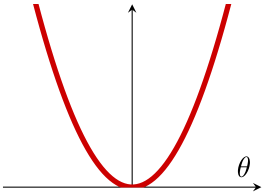
Gaussian prior



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

$\Rightarrow$

L2 penalty



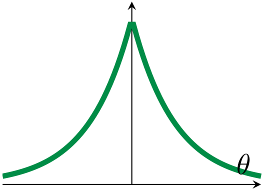
$$-\log P(\theta) \propto \|\theta\|_2^2$$

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

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

Laplace Prior $\Leftrightarrow$ Lasso (L1) Regression

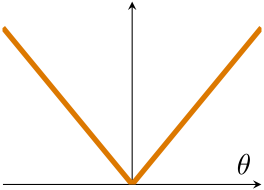
Laplace prior



$P(\theta) \propto e^{-|\theta|/b}$

$\Rightarrow$

L1 penalty



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

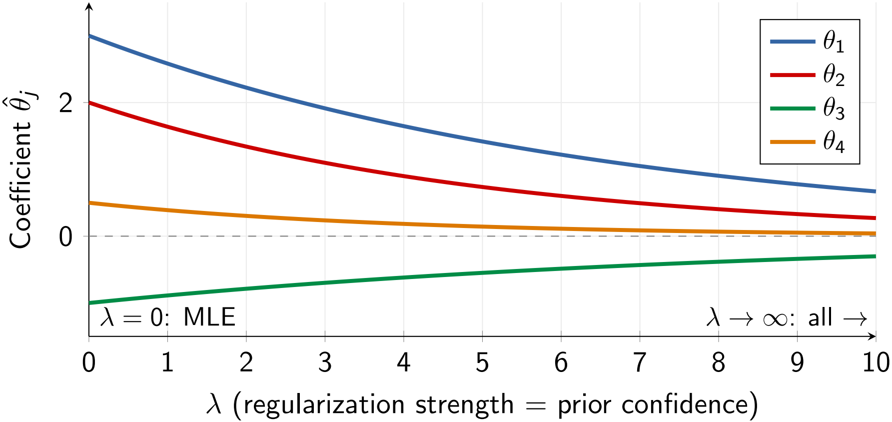
$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left[\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \theta)^2 + \lambda \|\theta\|_1\right]$$

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

The Regularization Map

Prior (Bayesian)		**Penalty (Frequentist)**	
Gaussian $\mathcal{N}(0, \tau^2)$	$\rightarrow$	Ridge: $\lambda\|\theta\|_2^2$	Shrinks all
Laplace$(0, b)$	$\rightarrow$	Lasso: $\lambda\|\theta\|_1$	Sets some
Uniform (flat)	$\rightarrow$	No penalty (MLE)	No shrinka

Visualizing Ridge Shrinkage



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

Three Philosophies

MLE

Mode of likelihood

No prior info used

Risk: overfitting with small n

$\hat{\theta} = \arg\max \ell(\theta)$

$\xrightarrow{\text{add prior}}$

MAP

Mode of posterior

Uses prior beliefs

Still a point estimate

$\hat{\theta} = \arg\max[\ell + \log P]$

$\xrightarrow{\text{keep full dist.}}$

Full Bayesian

Entire posterior

Full uncertainty

Computationally expensive (MCMC)

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

When to Use What

MLE when:
– Large n (prior doesn't matter)
– No reliable prior info
– Simplicity is valued

MAP when:
– Small n (need regularization)
– Have domain knowledge
– Want a point estimate fast

Full Bayesian when:
– Uncertainty quantification is critical (medical, safety)
– Model comparison needed
– Computational cost is acceptable

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

Questions?

Next lecture: Estimator Quality — Bias, Variance, and the Tradeoff