# 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 $\theta$ that makes the data most probable.

**Properties:** Consistent, asymptotically normal, efficient, invariant.

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

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

**Today:** What if we have **prior knowledge** about $\theta$?

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

# Where We Are

**MLE**

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

Lecture 5

**What if we have prior beliefs?**

**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"

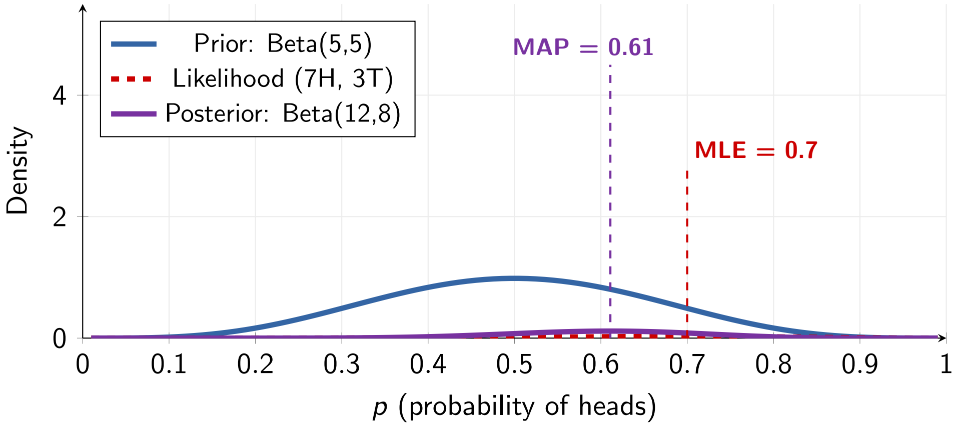
$=$

**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**.

# Conjugate Priors as Pseudo-Observations

Why did the math work out so neatly? Because Beta is **conjugate** to Binomial:

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

# Conjugate Priors as Pseudo-Observations

Why did the math work out so neatly? Because Beta is **conjugate** to Binomial:

Prior: $\text{Beta}(\alpha, \beta)$ + Data: $k$ H, $n-k$ T = Posterior: $\text{Beta}(\alpha+k,\ \beta+n-k)$

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

$\text{Beta}(\alpha, \beta)$ = pretend you already observed $\alpha-1$ heads and $\beta-1$ tails.

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

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

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

# MAP = Mode of the Posterior

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

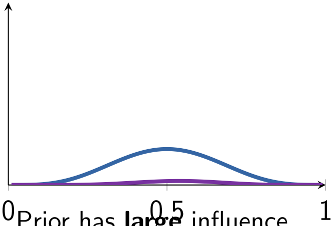
Maximize: log-likelihood + log-prior

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

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

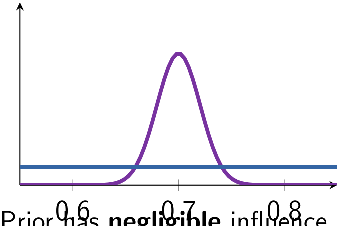
# When Does the Prior Matter?

**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

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

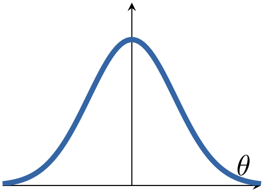
is the same as

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

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

# Gaussian Prior $\Leftrightarrow$ Ridge (L2) Regression

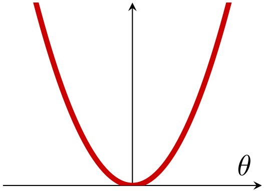
**Gaussian prior**




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

$\Rightarrow$

**L2 penalty**



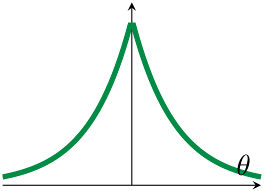

$-\log P(\theta) \propto \|\theta\|_2^2$

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

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

# Laplace Prior $\Leftrightarrow$ Lasso (L1) Regression

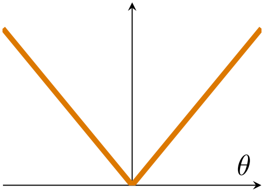
**Laplace prior**




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

$\Rightarrow$

**L1 penalty**




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

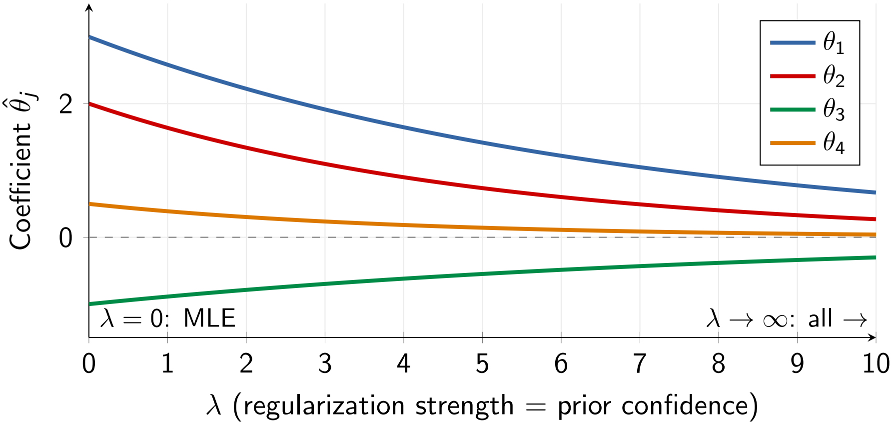
$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left[\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\theta)^2 + \lambda\|\theta\|_1\right]$$

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

# The Regularization Map

| **Prior (Bayesian)** | **Penalty (Frequentist)** |
|---|---|
| Gaussian $\mathcal{N}(0, \tau^2)$ | Ridge: $\lambda\|\theta\|_2^2$ <br> *shrinks toward 0* |
| Laplace$(0, b)$ | Lasso: $\lambda\|\theta\|_1$ <br> *sets some to exactly 0* |
| Uniform (flat) | No penalty <br> $=$ *MLE, no shrinkage* |

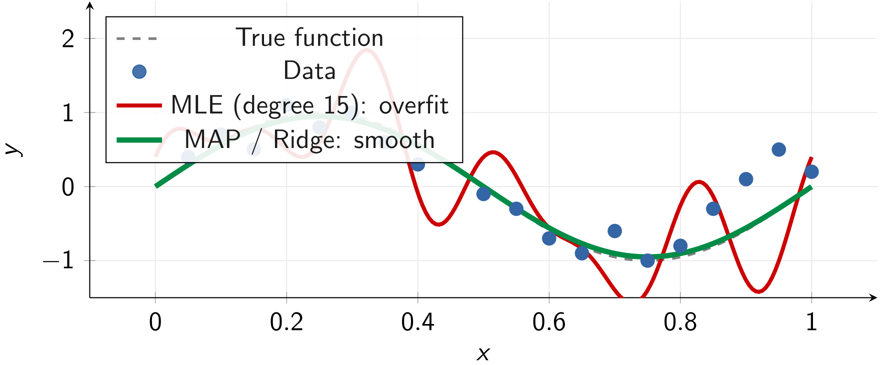
# Visualizing Ridge Shrinkage




Increasing $\lambda$ = stronger prior = more shrinkage = less overfitting (but more bias).

# MLE vs MAP: The Overfitting Story

Fit a polynomial to noisy data. MLE uses all parameters freely; MAP penalizes large coefficients.




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

# 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 matters (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 $\lambda$ (shrinkage path)
   - Observe: larger $\lambda$ = stronger prior = more shrinkage
3. **Visualize:** Plot the prior/likelihood/posterior for a simple 1D Normal with known $\sigma^2$, varying the prior variance $\tau^2$

# Homework

1. For $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ (known $\sigma_0^2$) 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) $\text{Beta}(1, 1)$, (b) $\text{Beta}(5, 5)$, (c) $\text{Beta}(50, 50)$ 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?

# Questions?

Next: Sampling distributions and confidence intervals