# Probabilistic Machine Learning (4): Parameter Estimation: MAP and Bayesian Inference

CS771: Introduction to Machine Learning

Piyush Rai

# MLE and Its Shortcomings..

- MLE finds parameters that make the observed data most probable

Log-likelihood

Neg. log-likelihood (NLL)

$$\theta_{MLE} = \underset{\theta}{\text{argmax}} \sum_{n=1}^{N} \log p(y_n|\theta) = \underset{\theta}{\text{argmin}} \sum_{n=1}^{N} -\log p(y_n|\theta)$$

- No provision to control overfitting (MLE is just like minimizing training loss)

- How do we regularize probabilistic models in a principled way?

- Also, MLE gives only a single "best" answer ("point estimate")
  - .. and it may not be very reliable, especially when we have very little data
  - Desirable: Report a probability distribution over the learned params instead of point est

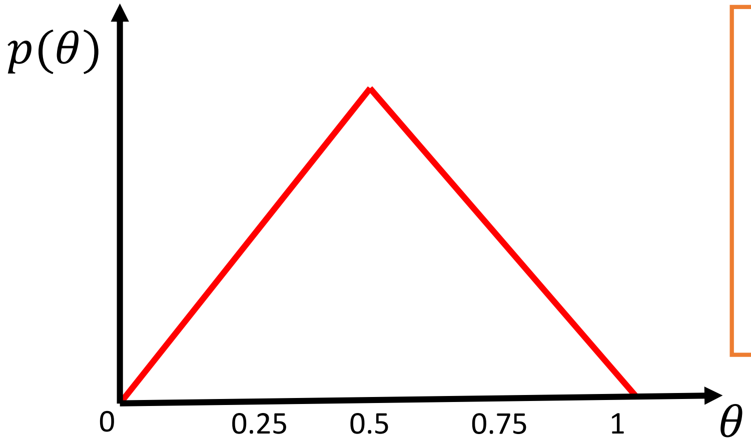
This distribution can give us a sense about the uncertainty in the parameter estimate

- Prior distributions provide a nice way to accomplish such things!

# Priors

- Can specify our prior belief about likely param values via a prob. dist., e.g.,

This is a rather simplistic/contrived prior. ☺ Just to illustrate the basic idea. We will see more concrete examples of priors shortly. Also, the prior usually depends (assumed conditioned on) on some fixed/learnable hyperparameters (say some $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ , and written as $p(\theta|\alpha,\beta)$



A possible prior for the coin bias estimation problem. The unknown $\boldsymbol{\theta}$ is being treated as a random variable, not simply a fixed unknown as we treated it as in MLE

- Once we observe the data $\boldsymbol{y}$, apply Bayes rule to update prior into posterior

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})}$$

Prior

Likelihood

Posterior

Marginal likelihood

Note: Marginal lik. is hard to compute in general as it requires a summation or integral which may not be easy (will briefly look at this in CS771, although will stay away going too deep in this course – CS775 does that in more detail)

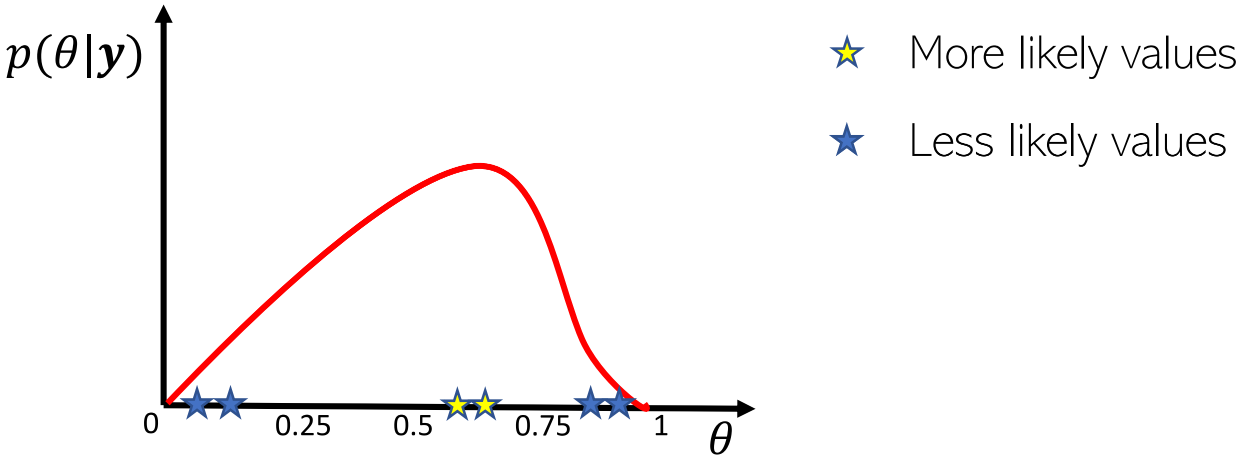
- Two way now to report the answer now:
  - Report the maxima (mode) of the posterior: $\arg\max_{\theta} p(\theta|\boldsymbol{y})$
  - Report the full posterior (and its properties, e.g., mean, mode, variance, quantiles, etc)

Maximum-a-posteriori (MAP) estimation

Fully Bayesian inference

# Posterior

- Posterior distribution tells us how probable different parameter values are after we have observed some data
- Height of posterior at each value gives the posterior probability of that value



- Can think of the posterior as a "hybrid" obtained by combining information from the likelihood and the prior

# Maximum-a-Posteriori (MAP) Estimation

- The MAP estimation approach reports the maxima/mode of the posterior

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|y) = \arg\max_{\theta} \log p(\theta|y) = \arg\max_{\theta} \log \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})}$$

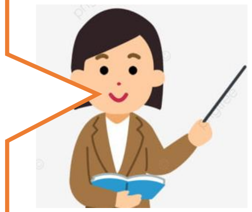
- Since $p(y)$ is constant w.r.t. $\theta$, the above simplifies to

$$\theta_{MAP} = \arg\max_{\theta} [\log p(y|\theta) + \log p(\theta)]$$

$$= \arg\min_{\theta} [-\log p(y|\theta) - \log p(\theta)]$$

$$\boxed{\theta_{MAP} = \arg\min_{\theta} [NLL(\theta) - \log p(\theta)]}$$

The NLL term acts like the training loss and the (negative) log-prior acts as regularizer. Keep in mind this analogy. ☺



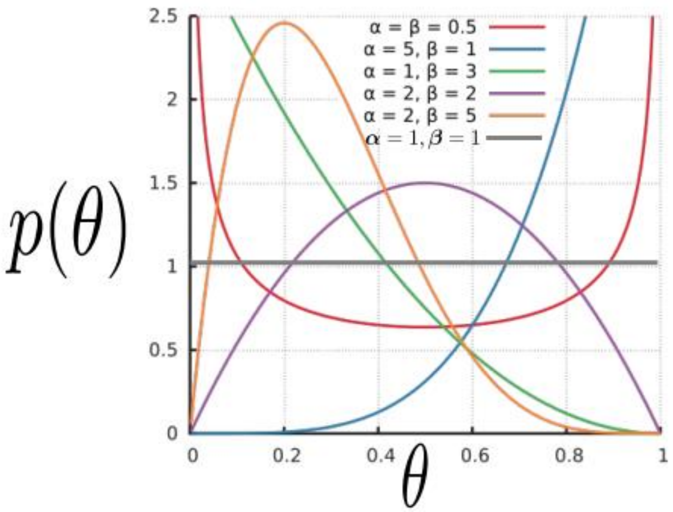
- Same as MLE with an extra log-prior-distribution term (acts as a regularizer) ☺
- If the prior is absent or <u>uniform</u> (all values equally likely a prior) then MAP=MLE

# MAP Estimation: An Example

- Let's again consider the coin-toss problem (estimating the bias of the coin)
- Each likelihood term is Bernoulli

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Also need a prior since we want to do MAP estimation
- Since $\theta \in (0,1)$, a reasonable choice of prior for $\theta$ would be Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The gamma function

$\alpha$ and $\beta$ (both non-negative reals) are the two hyperparameters of this Beta prior

Using $\alpha = 1$ and $\beta = 1$ will make the Beta prior a uniform prior

Can set these based on intuition, cross-validation, or even learn them

# MAP Estimation: An Example (Contd)

- The log posterior for the coin-toss model is log-lik + log-prior

$$LP(\theta) = \sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta|\alpha,\beta)$$

- Plugging in the expressions for Bernoulli and Beta and ignoring any terms that don't depend on $\boldsymbol{\theta}$, the log posterior simplifies to

$$LP(\theta) = \sum_{n=1}^{N} [y_n \log \theta + (1-y_n)\log(1-\theta)] + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Maximizing the above log post. (or min. of its negative) w.r.t. $\boldsymbol{\theta}$ gives

$$\theta_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N + \alpha + \beta - 2}$$

Using $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$ gives us the same solution as MLE

Recall that $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$ for Beta distribution is in fact equivalent to a uniform prior (hence making MAP equivalent to MLE)

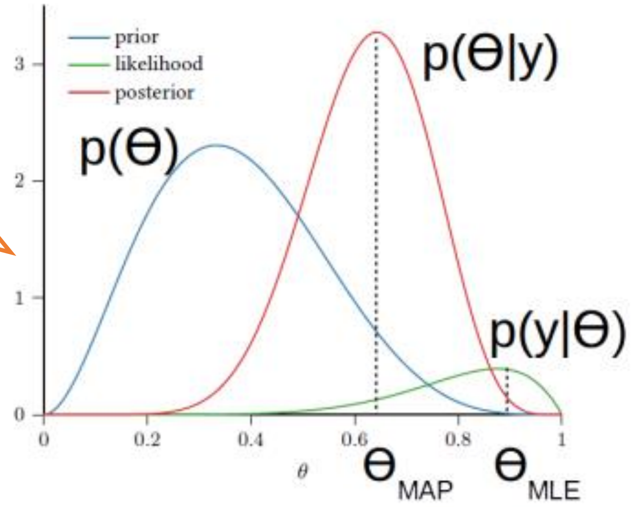
Prior's hyperparameters have an interesting interpretation. Can think of $\alpha - 1$ and $\beta - 1$ as the number of heads and tails, respectively, before starting the coin-toss experiment (akin to "pseudo-observations")

Such interpretations of prior's hyperparameters as being "pseudo-observations" exist for various other prior distributions as well (in particular, distributions belonging to "exponential family" of distributions

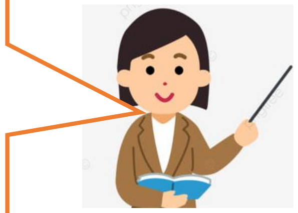
# Fully Bayesian Inference

- MLE/MAP only give us a point estimate of $\boldsymbol{\theta}$

MAP estimate is more robust than MLE (due to the regularization effect) but the estimate of uncertainty is missing in both approaches – both just return a single “optimal” solution by solving an optimization problem



Interesting fact to keep in mind: Note that the use of the prior is making the MLE solution move towards the prior (MAP solution is kind of a “compromise between MLE solution of the mode of the prior) ☺



Fully Bayesian inference

- If we want more than just a point estimate, we can compute the full posterior

Computable analytically only when the prior likelihood are “friends” with each other (i.e., they form a conjugate pair of distributions (distributions from exponential family have conjugate priors

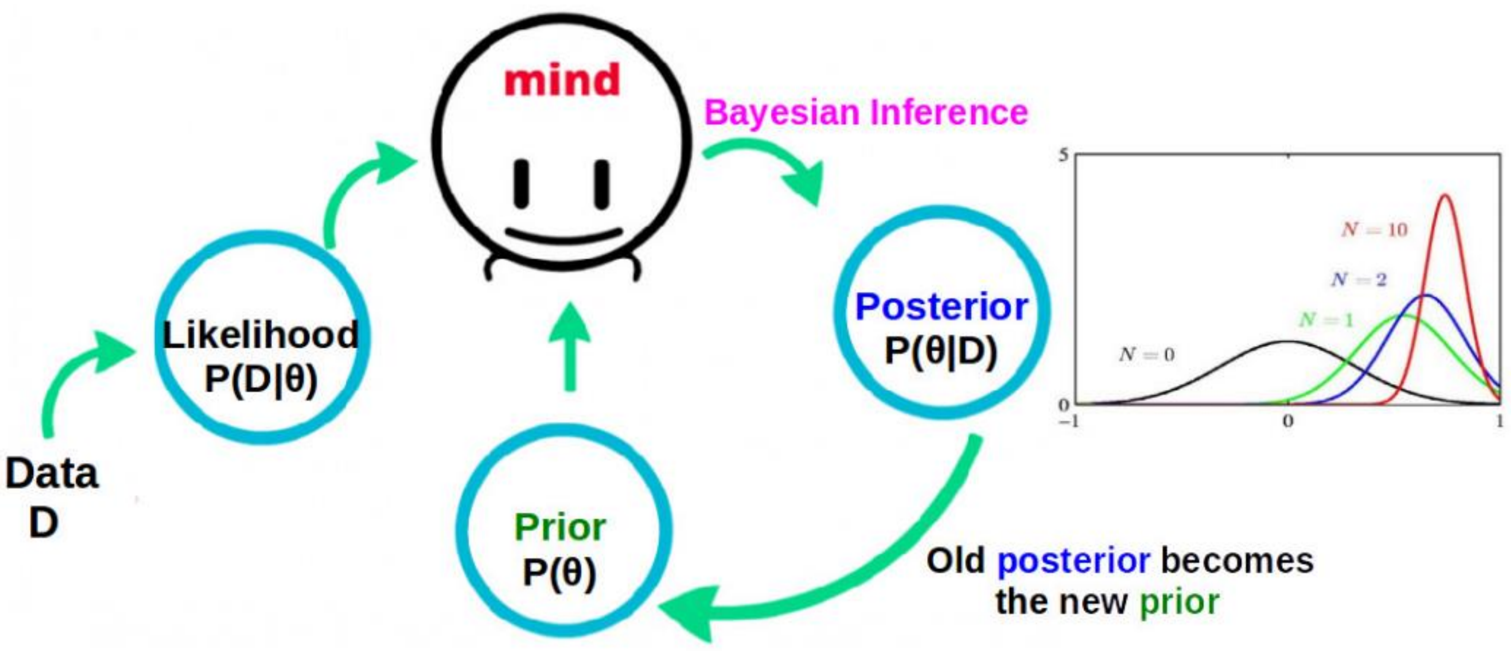
$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})}$$

In other cases, the posterior needs to be approximated (will see 1-2 such cases in this course; more detailed treatment in the advanced course on probabilistic modeling and inference)

An example: Bernoulli and Beta are conjugate. Will see some more such pairs

# “Online” Nature of Bayesian Inference

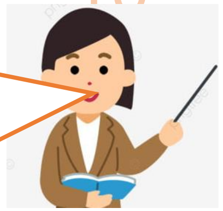
- Fully Bayesian inference fits naturally into an “online” learning setting



Also, the posterior becomes more and more “concentrated” as the number of observations increases. For very large N, you may expect it to be peak around the MLE solution

- Our belief about $\theta$ keeps getting updated as we see more and more data

# Fully Bayesian Inference: An Example



Posterior is the same distribution as the prior (both Beta), just with updated hyperparameters (property when likelihood and prior are conjugate to each other)

Also, if you get more observations, you can treat the current posterior as the new prior and obtain a new posterior using these extra observations

- Let's again consider the coin-toss problem
- Bernoulli likelihood: $p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$
- Beta prior: $p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior can be computed as

Number of heads $(N_1)$

Number of tails $(N_0)$

$$\theta^{\sum_{n=1}^N y_n}(1-\theta)^{N-\sum_{n=1}^N y_n}$$

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})} = \frac{p(\theta)\prod_{n=1}^N p(y_n|\theta)}{p(\boldsymbol{y})} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^N\theta^{y_n}(1-\theta)^{1-y_n}}{\int\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^N\theta^{y_n}(1-\theta)^{1-y_n}d\theta}$$

This is the numerator integrated/marginalized over $\theta: p(\mathbf{y}) = \int p(\theta,\mathbf{y})d\theta = \int p(\theta)p(\mathbf{y}|\theta)d\theta$

In general, hard but with conjugate pairs of prior and likelihood, we don't need to compute this, as we will see in this example ☺

Parts coming from the numerator, which consist of $\theta$ terms. We have ignored other constants in the numerator, and the whole denominator which is also constant w.r.t. $\theta$

$$\propto \theta^{\alpha+N_1-1}(1-\theta)^{\beta+N_0-1}$$

Aha! This is nothing but $\text{Beta}(\theta|\alpha+N_1,\beta+N_0)$

This, of course, is not always possible but only in simple cases like this

Found the posterior just by simple inspection without having to calculate the constant of proportionality ☺

# Conjugacy

- Many pairs of distributions are conjugate to each other
  - Bernoulli (likelihood) + Beta (prior) ⇒ Beta posterior
  - Binomial (likelihood) + Beta (prior) ⇒ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) ⇒ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) ⇒ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) ⇒ Gaussian posterior
  - and many other such pairs ..

Not true in general, but in some cases (e.g., when mean of the Gaussian prior is fixed)

- Tip: If two distr are conjugate to each other, their functional forms are similar
  - Example: Bernoulli and Beta have the forms

$$\text{Bernoulli}(y|\theta) = \theta^{y}\,(1-\theta)^{1-y}$$

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

This is why, when we multiply them while computing the posterior, the exponents get added and we get the same form for the posterior as the prior but with just updated hyperparameter. Also, we can identify the posterior and its hyperparameters simply by inspection

# Probabilistic Models: Making Predictions

- Having estimated $\boldsymbol{\theta}$, we can now use it to make predictions
- Prediction entails computing the predictive distribution of a new observation, say $\boldsymbol{y}_*$

For example, PMF of the label of a new test input in classification

$$p(y_*|\boldsymbol{y}) = \int p(y_*, \theta|\boldsymbol{y})d\theta$$

Marginalizing over the unknown $\boldsymbol{\theta}$

$$= \int p(y_*|\theta, \boldsymbol{y})p(\theta|\boldsymbol{y})d\theta$$

Decomposing the joint using chain rule

$$= \int p(y_*|\theta)p(\theta|\boldsymbol{y})d\theta$$

Assuming i.i.d. data, given $\boldsymbol{\theta}$, $\boldsymbol{y}_*$ does not depend on $\boldsymbol{y}$

Conditional distribution of the new observation, given past observations

- When doing MLE/MAP, we approximate the posterior $\boldsymbol{p(\theta|y)}$ by a single point $\boldsymbol{\theta_{opt}}$

$$p(y_*|\boldsymbol{y}) = \int p(y_*|\theta)p(\theta|\boldsymbol{y})d\theta \approx p(y_*|\theta_{opt})$$

A "plug-in prediction" (simply plugged in the singe estimate we had)

- When doing fully Bayesian est, getting the predictive dist. Will require computing

$$p(y_*|\boldsymbol{y}) = \int p(y_*|\theta)p(\theta|\boldsymbol{y})d\theta$$

$$\mathbb{E}_{p(\theta|\boldsymbol{y})}[p(y_*|\theta)]$$

This computes the predictive distribution by averaging over the full posterior – basically calculate $p(y_*|\theta)$ for each possible $\boldsymbol{\theta}$, weighs it by how likely this $\boldsymbol{\theta}$ is under the posterior $\boldsymbol{p(\theta|y)}$, and sum all such posterior weighted predictions. Note that not each value of theta is given equal importance here in the averaging

# Probabilistic Models: Making Predictions (Example)

- For coin-toss example, let's compute probability of the $(N+1)^{th}$ toss showing head
- This can be done using the MLE/MAP estimate, or using the full posterior

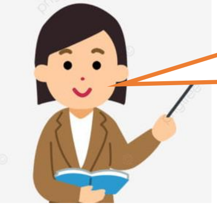
$$\theta_{MLE} = \frac{N_1}{N} \qquad \theta_{MAP} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2} \qquad p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

- Thus for this example (where observations are assumed to come from a Bernoulli)

MLE prediction: $p(y_{N+1} = 1|\mathbf{y}) = \int p(y_{N+1} = 1|\theta)p(\theta|\mathbf{y})d\theta \approx p(y_{N+1} = 1|\theta_{MLE}) = \theta_{MLE} = \frac{N_1}{N}$

MAP prediction: $p(y_{N+1} = 1|\mathbf{y}) = \int p(y_{N+1} = 1|\theta)p(\theta|\mathbf{y})d\theta \approx p(y_{N+1} = 1|\theta_{MAP}) = \theta_{MAP} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}$

Fully Bayesian: $p(y_{N+1} = 1|\mathbf{y}) = \int p(y_{N+1} = 1|\theta)p(\theta|\mathbf{y})d\theta = \int \theta p(\theta|\mathbf{y})d\theta = \int \theta \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta = \frac{N_1 + \alpha}{N + \alpha + \beta}$



Again, keep in mind that the posterior weighted averaged prediction used in the fully Bayesian case would usually not be as simple to compute as it was in this case. We will look at some hard cases later

Expectation of $\boldsymbol{\theta}$ under the Beta posterior that we computed using fully Bayesian inference

# Probabilistic Modeling: A Summary

- Likelihood corresponds to a loss function; prior corresponds to a regularizer
- Can choose likelihoods and priors based on the nature/property of data/parameters
- MLE estimation = unregularized loss function minimization
- MAP estimation = regularized loss function minimization
- Allows us to do fully Bayesian learning (learning the full distribution of the parameters)
- Makes robust predictions by posterior averaging (rather than using point estimate)
- Many other benefits, such as
    - Estimate of confidence in the model's prediction (useful for doing Active Learning)
    - Can do automatic model selection, hyperparameter estimation, handle missing data, etc.
    - Formulate latent variable models
    - .. and many other benefits (a proper treatment deserves a separate course, but we will see some of these in this course, too)

# Coming up next

- Probabilistic modeling for regression and classification problems