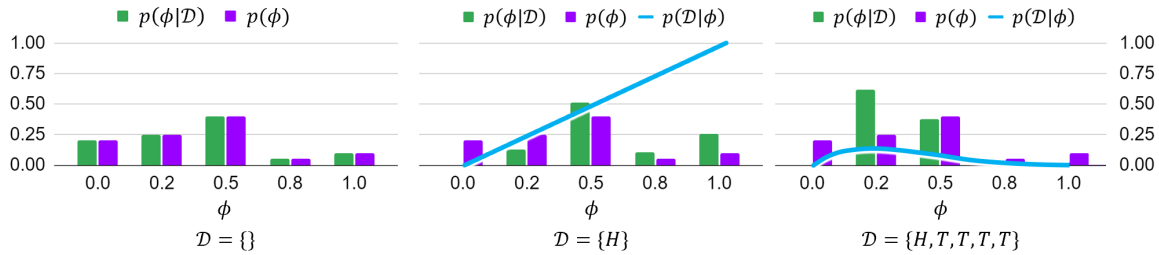
10-315 Notes
Maximum *a posteriori* Estimation

Carnegie Mellon University
Machine Learning Department

Contents

1 MLE and MAP Summary

Here's a quick preview of our trick coin example, where we start with a prior belief about our parameter, $p(\phi)$, and see how the posterior, $p(\phi \mid \mathcal{D})$ changes as we start to collect data and incorporate the likelihood, $p(\mathcal{D} \mid \phi)$:



To get a better understanding of how the prior, posterior, and data interact, experiment with different possible combinations of coin flip data with this handy visualization.

Overview highlighting the differences and similarities between MLE and MAP estimation in one table:

	Maximum likelihood estimation	Maximum *a posteriori* estimation
Formulation	$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ p(\mathcal{D} \mid \theta)$ $= \underset{\theta}{\text{argmax}} \prod_{i=1}^{N} p(y^{(i)} \mid \theta)$	$\theta_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta \mid \mathcal{D})$ $= \underset{\theta}{\text{argmax}} \frac{p(\theta)\, p(\mathcal{D} \mid \theta)}{p(\mathcal{D})}$ $= \underset{\theta}{\text{argmax}} \frac{p(\theta) \prod_{i=1}^{N} p(y^{(i)} \mid \theta)}{p(\mathcal{D})}$ $= \underset{\theta}{\text{argmax}}\ p(\theta) \prod_{i=1}^{N} p(y^{(i)} \mid \theta)$
Recipe for optimization	1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$ 2. Set objective $J(\theta)$ equal to negative log of the likelihood, $\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$ 3. Compute derivative of objective, $\partial J / \partial \theta$ 4. Find $\hat{\theta}$, either a. Set derivative equal to zero and solve for θ b. Use (stochastic) gradient descent to step towards better θ	1. Formulate the likelihood times the prior, $p(\mathcal{D} \mid \theta)\, p(\theta)$ 2. Set objective $J(\theta)$ equal to negative log of the likelihood times the prior, $\mathrm{J}(\theta) = -\log (p(\mathcal{D} \mid \theta) p(\theta))$ 3. Compute derivative of objective, $\partial J / \partial \theta$ 4. Find $\hat{\theta}$, either a. Set derivative equal to zero and solve for θ b. Use (stochastic) gradient descent to step towards better θ
Data and model assumptions	▪ With no prior model assumptions, we have no information about our parameters before collecting data ▪ With a small amount of data, we are at greater risk of overfitting ▪ As the number of data points increases, MLE is sufficient, as we can rely solely on the data to estimate our parameters	▪ Even with no data, we can have an initial estimate for our parameters, however, we must use our knowledge to formulate a prior (additional modeling assumptions) ▪ If our prior model assumptions are reasonably accurate, we only need a small amount of data to compute a good estimate of our parameters ▪ As the number of data points increases, the MAP estimate and the MLE estimate converge to the same value

2 Bayes Rule

2.1 Bayes rule: General version

Bayes rule is a super powerful theorem that allows us to convert the conditional probability $p(b \mid a)$ into $p(a \mid b)$ as long as we also know $p(b)$:

$$p(a \mid b) = \frac{p(b \mid a)\ p(a)}{p(b)}$$

It can be useful to write Bayes rule without the denominator, $p(b)$ using the "proportional to" notation, $\propto$:

$$p(a \mid b) \propto p(b \mid a)\ p(a)$$

The stems from the following set of probability rules:

- Product rule: $p(b \mid a)\ p(a) = p(a, b)$
- Marginalization: If we can model the joint distribution $p(a, b)$, then we can compute $p(b) = \sum_a p(a, b)$ (if a is discrete) or $\int_a p(a, b)$ (if a is continuous)
- Normalization: If we can model the numerator of Bayes rule, $p(b \mid a)p(a)$, we can compute the left-hand side, by normalizing the values of the numerator over all possible values of a. Specifically, combining the product rule and marginalization and assuming a is discrete, we can compute $Z = p(b) = \sum_a p(b \mid a)\ p(a)$, and then write Bayes rule as $p(a \mid b) = \frac{1}{Z} p(b \mid a)p(a)$
- If we know that b is a fixed value, then we can treat $Z = p(b)$ as a constant and write Bayes rule as $p(a \mid b) \propto p(b \mid a)\ p(a)$.

2.2 Bayes rule: Data and parameters

With the dataset $\mathcal{D} = \{y^{(i)}\}_{i=1}^N$ and generic model parameter(s) θ, we write Bayes rule as follows:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\ p(\theta)}{p(\mathcal{D})}$$

where:

- $p(\mathcal{D} \mid \theta)$ is the **likelihood**,
- $p(\theta)$ is the **prior** on the parameters, and
- $p(\theta \mid \mathcal{D})$ is the **posterior**

Normalization: When trying to find the optimal value for our parameters, we treat $\mathcal{D}$ and $p(\mathcal{D})$ as constant and write Bayes rule as:

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\ p(\theta)$$

3 Bernoulli MAP example: Trick coin

Suppose we wander into a joke shop where there is a container full of coins with a sign that says, "NEW: Randomly-weighted Trick Coins!" Statistical curiosity drives us to purchase one coin.

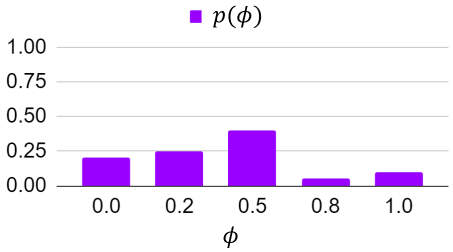
Before flipping the coin, we don't know what to expect. Will the coin always come up heads? Always tails? Did we just pay money for a fair coin? Or is weighted in an interesting way such that comes up heads with some other probability.

Just as we start to formulate our MLE experiment, out of the corner of our eye, we spot a piece of paper sticking out of the trash bin next to the joke shop. The visible portion of the paper says, "Invoice: Weighted Coins." Pulling out the paper, we see the full contents of the receipt listing the five different types of coins and the quantities purchased! We excitedly start scribbling some notes on the invoice about the probability of selecting each type of coin from the store:

Invoice: Weighted Coins Customer: Torch Tricks, Inc.

Item	Quantity	Coin type, ϕ	$p(\phi)$
0% Heads Coin	40/200	0.0	0.20
20% Heads Coin	50/200	0.2	0.25
50% Heads Coin	80/200	0.5	0.40
80% Heads Coin	10/200	0.8	0.05
100% Heads Coin	20/200	1.0	0.10
Total: 200			✓ 1.00

Now, before we starting flipping the coin at all, we already have some information about the Bernoulli coin flip parameter, $\phi = P(Y = 1 \mid \phi)$, the probability of the coin coming up heads. If we were forced to guess what type of coin we have prior to flipping it, we'd probably choose, $\phi = 0.5$ because that has the highest probability based on our prior information from the invoice, i.e. $p(\phi = 0.5) = 80/200 = 0.4$, which is greater than all other possible values of ϕ (other coin types).



We know that flipping coins helps us collect information about the likelihood, $p(\mathcal{D} \mid \phi) = \prod_{i=1}^{N} p(y^{(i)} \mid \phi)$. Bayes rule allows us to combine the likelihood with the prior on the coin flip parameter:

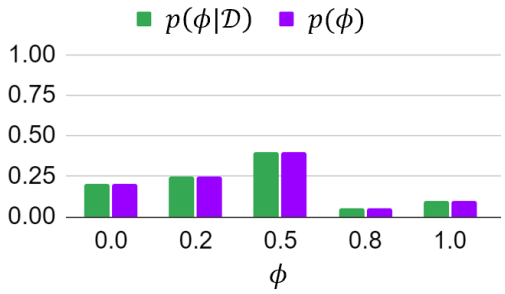
$$p(\phi \mid \mathcal{D}) \propto p(\phi) \prod_{i=1}^{N} p(y^{(i)} \mid \phi)$$

As we flip more and more coins, the likelihood term contains more and more factors

$\mathcal{D}$	$p(\phi) \prod_{i=1}^{N} p(y^{(i)} \mid \phi)$
$\{\}$	$p(\phi)$
$\{H\}$	$p(\phi)\ \phi$
$\{H, T\}$	$p(\phi)\ \phi(1-\phi)$
$\{H, T, T\}$	$p(\phi)\ \phi(1-\phi)(1-\phi)$

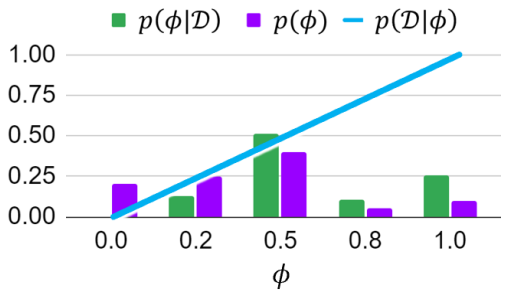
The detailed example below shows how the posterior distribution of $p(\phi \mid \mathcal{D})$ is initially equal to the prior distribution and then changes as we collect more data and add more factors to the likelihood term.

$N = 0: \quad \mathcal{D} = \{\}$



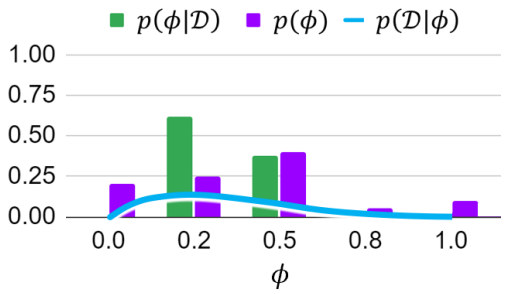
ϕ	$p(\phi \mid \mathcal{D})$	$p(\phi)$	$p(\mathcal{D} \mid \phi)$
0.0	0.200000	0.20	
0.2	0.250000	0.25	
0.5	0.400000	0.40	
0.8	0.050000	0.05	
1.0	0.100000	0.10	
		$\mathcal{D}$:	

$N = 1: \quad \mathcal{D} = \{H\}$



ϕ	$p(\phi \mid \mathcal{D})$	$p(\phi)$	$p(\mathcal{D} \mid \phi)$
0.0	0.000000	0.20	0.0
0.2	0.128205	0.25	0.2
0.5	0.512821	0.40	0.5
0.8	0.102564	0.05	0.8
1.0	0.256410	0.10	1.0
		$\mathcal{D}$:	H

$N = 5: \quad \mathcal{D} = \{H, T, T, T, T\}$



ϕ	$p(\phi \mid \mathcal{D})$	$p(\phi)$	$p(\mathcal{D} \mid \phi)$				
0.0	0.000000	0.20	0.0	1.0	1.0	1.0	1.0
0.2	0.619780	0.25	0.2	0.8	0.8	0.8	0.8
0.5	0.378284	0.40	0.5	0.5	0.5	0.5	0.5
0.8	0.001937	0.05	0.8	0.2	0.2	0.2	0.2
1.0	0.000000	0.10	1.0	0.0	0.0	0.0	0.0
		$\mathcal{D}$:	H	T	T	T	T

Note that the tables above are missing the details on how to compute the posterior value for each parameter value. While the prior value is simply multiplied times the product of all of the likelihood factors, there is still a normalization step over the probabilities for all possible values of ϕ that happens to get specific values for the posterior. Here is the derivation of Z for the specific $N = 5$ $\mathcal{D} = \{H, T, T, T, T\}$ example above:

$$p(\phi \mid \mathcal{D}) = \frac{1}{Z} p(\phi) \prod_{i=1}^{N} p(y^{(i)} \mid \phi)$$

$$Z = \sum_{\phi} p(\phi) \prod_{i=1}^{N} p(y^{(i)} \mid \phi) = \sum_{\phi} p(\phi)\ \phi(1-\phi)(1-\phi)(1-\phi)(1-\phi) = \sum_{\phi} p(\phi)\ \phi(1-\phi)^4$$

$$= 0.20 \cdot 0.0 \cdot 1.0^4 + 0.25 \cdot 0.2 \cdot 0.8^4 + 0.40 \cdot 0.5 \cdot 0.5^4 + 0.05 \cdot 0.8 \cdot 0.2^4 + 0.10 \cdot 1.0 \cdot 0.0^4$$

$$= 0.033044$$

$$p(\phi = 0.2 \mid \mathcal{D}) = \frac{1}{0.033044} 0.25 \cdot 0.2 \cdot 0.8^4$$

$$= 0.619780$$

4 Recipe for Maximum *a posteriori* (MAP) estimation

Because there are only five possible values of the parameter ϕ in our joke shop coin example, we can compute the posterior value for all five parameter values and then save the parameter value $\hat{\phi}_{MAP}$ that led to the highest posterior. This value is the **maximum *a posteriori*** estimate.

More generally, we would need to try an infinite set of possible parameter values, so instead we turn to optimization techniques.

The general recipe for finding MAP estimate is extremely similar to the MLE recipe:

1. Formulate the likelihood times the prior, $p(\mathcal{D} \mid \theta)\ p(\theta)$
2. Set objective $J(\theta)$ equal to the negative log of the likelihood times the prior:

$$J(\theta) = -\log\left(p(\mathcal{D} \mid \theta)\ p(\theta)\right)$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$ by either:
   (a) Setting derivative equal to zero and solve for θ
   (b) Using (stochastic) gradient descent to step towards better θ

5 Gaussian MAP example: Course hours

We started taking an advanced systems class at CMU. We're part way through the semester and are starting to wonder what the mean number of hours per week students are spending on this class. Asking a few friends in the course, we collect a dataset of their hours per week:

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^{4} = \{18, 20, 14, 10\}$$

Not wanting to overfit our small dataset, we do a little digging in past course evaluation data and find statistics saying that over the past several years the reported hours per week have a mean of $\nu = 23.9$ and a standard deviation of $\tau = 1.56$. We use these statistics to create a Gaussian prior on our mean parameter, $\mu \sim \mathcal{N}(\nu = 23.9, \tau = 1.56)$. (To avoid confusion, we chose different symbols for this second set of mean and standard deviation.)

Note that we now have two different Gaussian distributions, one for our hours data, $x \sim \mathcal{N}(\mu, \sigma)$, (with unknown parameters, μ and σ) and one for our mean parameter, $\mu \sim \mathcal{N}(\nu = 23.9, \tau = 1.56)$.

In this example problem, we are going focus on using MAP to estimate the mean for our data distribution, μ. To do this, we'll hold the standard deviation, σ constant. (The mean and standard deviation parameters prior distribution, ν and τ, are already constants.)

5.1 Formulate the likelihood times the prior

$$p(\mu \mid \mathcal{D}) \propto p(\mu) \prod_{i=1}^{4} p(x^{(i)} \mid \mu) \tag{1}$$

$$= \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mu-\nu)^2} \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2} \tag{2}$$

5.2 Set the objective function to the negative log likelihood times the prior

$$J(\mu) = -\log\left(p(\mu)\prod_{i=1}^{4} p(x^{(i)} \mid \mu)\right) \tag{3}$$

$$= -\log\left(\frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mu-\nu)^2} \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2}\right) \tag{4}$$

$$= -\log\left(\frac{1}{\sqrt{2\pi\tau^2}}\right) + \frac{1}{2\tau^2}(\mu-\nu)^2 - N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{4}\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2$$

(where $N = 4$, our number of datapoints)

5.3 Compute derivative of objective, $\partial J/\partial\theta$

Thankfully, the remaining log terms don't contain μ, so they are dropped in the derivative.

$$\frac{\partial J}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-\log\left(\frac{1}{\sqrt{2\pi\tau^2}}\right) + \frac{1}{2\tau^2}(\mu-\nu)^2 - \log\left(\frac{N}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{4}\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2\right) \tag{5}$$

$$= \frac{1}{\tau^2}(\mu-\nu) - \sum_{i=1}^{4}\frac{1}{\sigma^2}(x^{(i)}-\mu) \tag{6}$$

5.4 Find $\hat{\theta}$ by setting derivative equal to zero and solve for θ

$$\frac{\partial J}{\partial \mu} = 0 \tag{7}$$

$$0 = \frac{1}{\tau^2}(\mu-\nu) - \sum_{i=1}^{4}\frac{1}{\sigma^2}(x^{(i)}-\mu) \tag{8}$$

$$0 = \frac{1}{\tau^2}\mu - \frac{\nu}{\tau^2} + \frac{N}{\sigma^2}\mu - \frac{1}{\sigma^2}\sum_{i=1}^{4} x^{(i)} \tag{9}$$

$$\frac{1}{\tau^2}\mu + \frac{N}{\sigma^2}\mu = \frac{\nu}{\tau^2} + \frac{1}{\sigma^2}\sum_{i=1}^{4} x^{(i)} \qquad \text{// Next, multiply both side by } \tau^2\sigma^2 \tag{10}$$

$$\sigma^2\mu + N\tau^2\mu = \sigma^2\nu + \tau^2\sum_{i=1}^{4} x^{(i)} \tag{11}$$

$$\left(\sigma^2 + N\tau^2\right)\mu = \sigma^2\nu + \tau^2\sum_{i=1}^{4} x^{(i)} \tag{12}$$

$$\hat{\mu}_{MAP} = \frac{1}{\sigma^2 + N\tau^2}\left(\sigma^2\nu + \tau^2\sum_{i=1}^{4} x^{(i)}\right) \tag{13}$$

Ok, admittedly, this is a bit dissatisfying because 1) it looks complicated and 2) the MAP estimate of μ is written in terms of the standard deviation, σ. But, this is cool, watch what happens if we (rather naively) assume that the standard deviation for this semester's data is the same as the prior standard deviation, $\sigma = \tau$:

$$\hat{\mu} = \frac{1}{1+N}\left(\nu + \sum_{i=1}^{4} x^{(i)}\right)$$

This estimate is the same as if we treated the prior mean as just one other data point and included it into our MLE mean!

$$\hat{\mu} = \frac{1}{5}\left(23.9 + (18 + 20 + 14 + 10)\right) = 17.8 \text{ hours per week}$$