Bayesian approach

Reminders/Comments

- Let's take a poll for concepts/sections that are particularly confusing to you

  - gradient descent

  - regularization, and its purpose

  - ML and MAP? e.g., examples with other distributions

  - formalizing prediction problems? weighted losses? other losses?

  - predicting multiple outputs?

  - basic optimization strategies?

  - generalized linear models?

  - neural networks

- Additional topics: boosting, bagging, decision trees

Bayesian learning

- Goal is to keep distribution over parameters

  - $p(w \mid D)$ rather than w^*

- Frequentist approach: find the most likely ("best") parameters

  - this is what we have been doing so far with ML and MAP

Bias of a coin

$$v^n = \begin{cases} 1 & \text{if on toss } n \text{ the coin comes up heads} \\ 0 & \text{if on toss } n \text{ the coin comes up tails} \end{cases}$$

Our aim is to estimate the probability θ that the coin will be a head, $p(v^n = 1|\theta) = \theta$ – called the 'bias' of the coin.
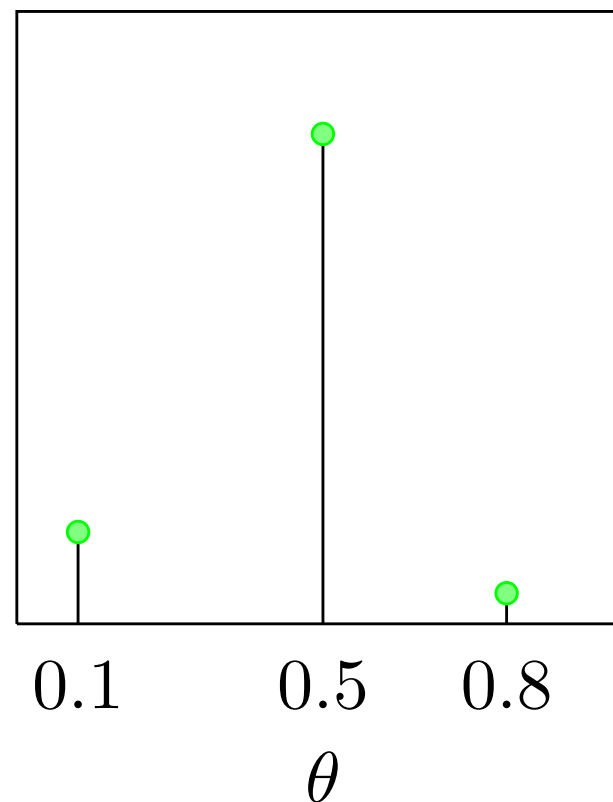
$$p(v^1, \ldots, v^N, \theta) = p(\theta) \prod_{n=1}^{N} p(v^n|\theta)$$

A prior for discrete parameters

We still need to fully specify the prior $p(\theta)$. To avoid complexities resulting from continuous variables, we'll consider a discrete θ with only three possible states, $\theta \in \{0.1, 0.5, 0.8\}$. Specifically, we assume

$$p(\theta = 0.1) = 0.15, \; p(\theta = 0.5) = 0.8, \; p(\theta = 0.8) = 0.05$$

The posterior for discrete parameters

$$p(\theta|v^1,\ldots,v^N) \propto p(\theta) \prod_{n=1}^{N} p(v^n|\theta)$$

$$= p(\theta) \prod_{n=1}^{N} \theta^{\mathbb{I}[v^n=1]} (1-\theta)^{\mathbb{I}[v^n=0]}$$

$$\propto p(\theta)\theta^{\sum_{n=1}^{N} \mathbb{I}[v^n=1]} (1-\theta)^{\sum_{n=1}^{N} \mathbb{I}[v^n=0]}$$
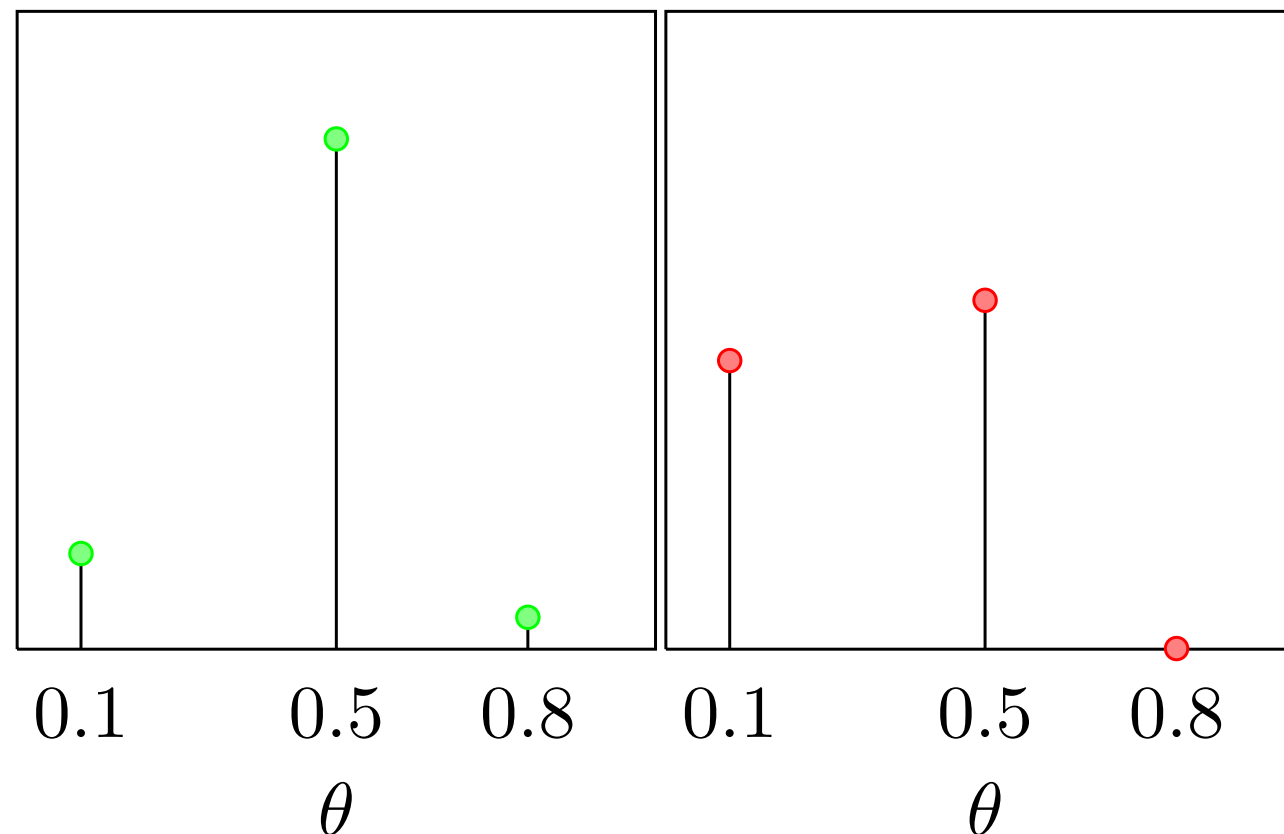
Hence

$$p(\theta|v^1,\ldots,v^N) \propto p(\theta)\theta^{N_H} (1-\theta)^{N_T}$$

$N_H = \sum_{n=1}^{N} \mathbb{I}[v^n = 1]$ is the number of occurrences of heads.
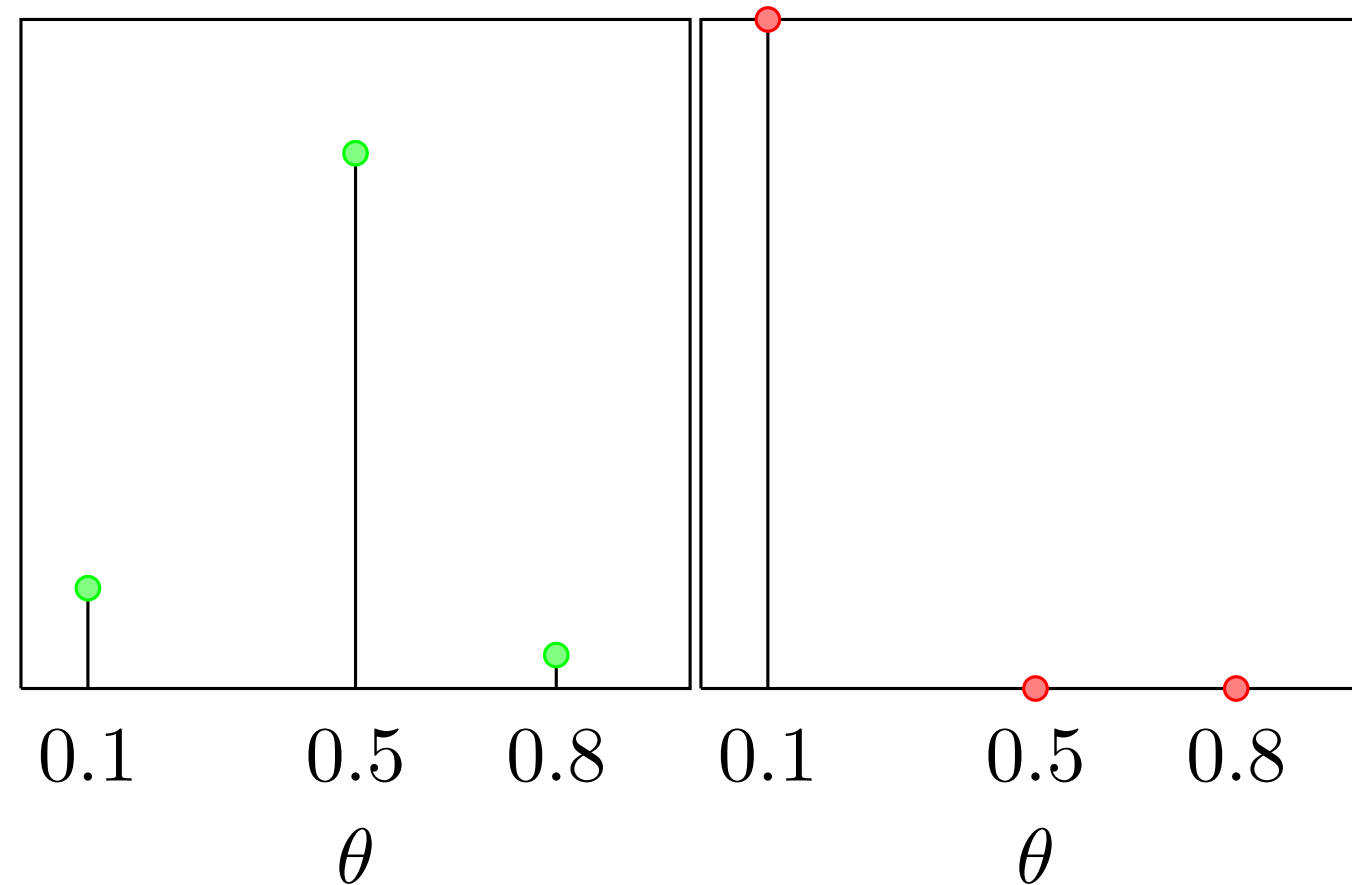$N_T = \sum_{n=1}^{N} \mathbb{I}[v^n = 0]$ is the number of tails.

For an experiment with $N_H = 2$, $N_T = 8$, the posterior distribution is

Posterior after 100 flips

Repeating the above with $N_H = 20$, $N_T = 80$, the posterior changes to

Continuous parameters

- Can ask the same question for continuous parameters

- Prior is then a density, rather than a set of probabilities

- Can do the same procedure but now the normalization is not as simple (have to integrate, or find closed form for integral)

  - for discrete parameter, we found p(theta | D) prop-to p(D | theta) p(theta), and then normalized the three values afterwards

Pros/Cons

✓ A Bayesian would like say that Bayesian approach are the "right" way to think about inference and estimation

✓ A good experts approach: Can more strongly influence learning with choice of prior

✓ Have a distribution over parameters, giving some measure of certainty

- Specifying a prior can be difficult (must carefully choose, limited often to a restricted set if computation matters)

- Can often involve numerical integration, which is computationally intensive

Whiteboard

- Bayesian approach for Poisson distribution

- Bayesian approach for linear regression