# Lecture 4

## Generative Models for Discrete Data - Part 2

Luigi Freda

ALCOR Lab
DIAG
University of Rome "La Sapienza"

January 27, 2018

# Outline

# Outline

problem definition

- consider a series of $N$ **coin tosses**
- we would like to **infer the probability** $\theta \in [0, 1]$ that a coin shows up heads, given a series of observed coin tosses
- in this case we consider the **continuous random variable** $\theta$

N.B.: in the previous lesson we inferred a distribution over a discrete RV $h \in \mathcal{H}$ drawn from a finite space

# Outline

# The Beta-Binomial Model

## Likelihood

- for each $i$-th coin toss we have a discrete RV $X_i \sim \text{Ber}(\theta)$, where $X_i = 1$ represents "heads" and $X_i = 0$ represents "tails"
- the RV $\theta \in [0, 1]$ represents the probability of heads, i.e.

$$\theta = P_X(X = 1|\theta)$$

- since we assume to observe a set of **iid**[1] trials $\mathcal{D} = \{x_1, ..., x_N\}$ the **likelihood function** is

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \text{Ber}(x_i|\theta) = \prod_{i=1}^{N} \theta^{\mathbb{I}(x_i=1)}(1-\theta)^{\mathbb{I}(x_i=0)} =$$

$$= \theta^{N_1}(1-\theta)^{N_0}$$

  where $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ is the number of observed heads, and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ is the number of observed tails
- $N_1$ and $N_0$ are called the **counts**, one has $N = N_1 + N_0$

[1] Independent and Identically Distributed

- given that the likelihood function is

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

  all we need to specify it are the counts $N_1$ and $N_0$
- in this case $s(\mathcal{D}) = (N_1, N_0)$ are called the **sufficient statistics** of the data: all we need to know about $\mathcal{D}$ to infer $\theta$
- more formally $s(\mathcal{D})$ is a sufficient statistics for the data $\mathcal{D}$ if

$$p(\theta|\mathcal{D}) = p(\theta|s(\mathcal{D}))$$

- in this example, another sufficient statistics is $s(\mathcal{D}) = (N, N_1)$ (since $N_0 = N - N_1$)

- if we consider $N_1$ (the number of observed heads) as a RV

$$N_1 \sim \text{Bin}(N, \theta)$$

with the binomial distribution

$$\text{Bin}(N_1|N,\theta) = \binom{N_1}{N} \theta^{N_1}(1-\theta)^{N_0}$$

- hence if we consider the data $\mathcal{D}' = (N_1, N_0)$, we have

$$p(\mathcal{D}|\theta) \propto p(\mathcal{D}'|\theta) \propto \theta^{N_1}(1-\theta)^{N_0} \propto \text{Bin}(N_1|N,\theta)$$

since $\binom{N_1}{N}$ can be considered as a constant which does not depend on $\theta$

- here is the reason for the "binomial" part of the name **beta-binomial** model

# Outline

# The Beta-Binomial Model
## Prior

- we need a probability prior for $\theta$ which has support over $[0, 1]$
- given that

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

  if we had a prior of the same form, i.e.

$$p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_0}$$

  we could easily evaluate the posterior as

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) \propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1}(1-\theta)^{\gamma_0} = \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_0}$$

- when the prior and the posterior have the same form, we say that the prior is a **conjugate prior** for the corresponding likelihood
- in the case of the **Bernoulli**, the conjugate prior is the **beta distribution**

$$\text{Beta}(\theta|a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

- here is the reason for the "beta" part of the name **beta-binomial** model

- hence we select the conjugate prior

$$p(\theta) = \text{Beta}(\theta|a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

- in general the parameters $\boldsymbol{\pi}$ of the prior are called **hyper-parameters**, we can set them in order to encode our prior beliefs
- in this case $\boldsymbol{\pi} = (a, b)$
- for instance, given the beta distribution has mean $m$ and standard deviation $\sigma$

$$m = \frac{a}{a+b} \qquad \sigma = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$$

  if we want to represent our prior belief that $\theta$ has mean $m = 0.7$ and $\sigma = 0.2$, we can use these equations and compute $a = 2.975$ and $b = 1.275$
- if we know "nothing", we can use a **uniform prior** by setting $a = b = 1$ in order to have $p(\theta) = \text{Unif}(0, 1)$

**homework**: ex 3.15 and ex 3.16

# Outline

1. The Beta-Binomial Model
   - Problem Definition
   - Likelihood
   - Prior
   - Posterior
   - Posterior Predictive
   - Overfitting and the Black Swan Paradox

2. The Dirichlet-multinomial
   - Problem Definition
   - Likelihood
   - Prior
   - Posterior
   - Posterior Predictive

# The Beta-Binomial Model
## Posterior

- the posterior is obtained as a **beta-binomial model**

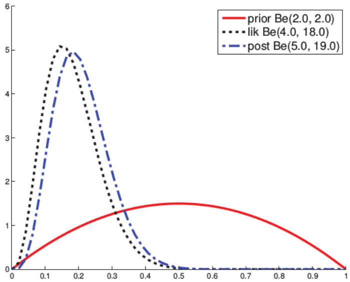
$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) \propto \text{Bin}(N_1|\theta, N_0 + N_1)\text{Beta}(\theta|a, b) \propto$$

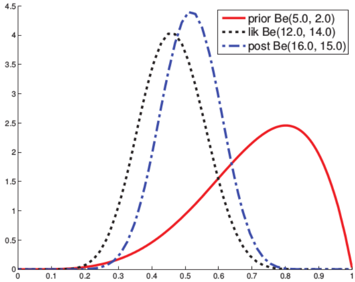
$$\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{a-1}(1-\theta)^{b-1} = \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

  hence we have

$$p(\theta|\mathcal{D}) \propto \text{Beta}(\theta|N_1 + a, N_0 + b)$$

- $N_1$ and $N_0$ are called the **empirical counts**
- the hyper-parameters $a$ and $b$ are called the **pseudo-counts**
- the pseudo-counts $a$ and $b$ play in the prior the same role that the empirical counts $N_1$ and $N_0$ play in the likelihood
- the strength of the prior, is given by the **equivalent sample size** $\alpha_0 = a + b$ which is the sum of the pseudo-counts
- $\alpha_0$ plays a role analogous to $N = N_1 + N_0$




$\alpha_0 = 4, \quad N = 20$

$\alpha_0 = 7, \quad N = 24$

strong prior due to $a = 5 > b = 2$

# The Beta-Binomial Model

## Sequential Posterior - Online Learning

let's see if updating the posterior sequentially is equivalent to updating in single batch

- **first sequence**: $\mathcal{D}'$ with sufficient statistics $N'_1, N'_0$ $\quad (N' = N'_1 + N'_0)$
- **second sequence**: $\mathcal{D}''$ with sufficient statistics $N''_1, N''_0$ $\quad (N'' = N''_1 + N''_0)$
- overall: $\mathcal{D} \triangleq \mathcal{D}' \cup \mathcal{D}''$, $N_1 \triangleq N'_1 + N''_1$ and $N_0 \triangleq N'_0 + N''_0$

**batch mode**

$$p(\theta|\mathcal{D}) = p(\theta|\mathcal{D}', \mathcal{D}'') \propto \text{Bin}(N_1|\theta, N_0 + N_1)\text{Beta}(\theta|a, b) \propto \text{Beta}(\theta|N_1 + a, N_0 + b)$$

**sequential mode**

1. first sequence posterior: $p(\theta|\mathcal{D}') \propto \text{Beta}(\theta|N'_1 + a, N'_0 + b)$
2. second sequence posterior: $p(\theta|\mathcal{D}', \mathcal{D}'') \propto \underbrace{p(\mathcal{D}''|\theta)}_{\text{lihelihood for} \mathcal{D}''} \times \underbrace{p(\theta|\mathcal{D}')}_{\text{prior for } \mathcal{D}'' \text{ based on } \mathcal{D}'} \propto$

$$\propto \text{Bin}(N''_1|\theta, N''_0 + N''_1)\text{Beta}(\theta|N'_1 + a, N'_0 + b) \propto$$

$$\propto \text{Beta}(\theta|N'_1 + N''_1 + a, N'_0 + N''_0 + b) \propto \text{Beta}(\theta|N_1 + a, N_0 + b)$$

- we have written the following equation by using intuition

$$p(\theta|\mathcal{D}',\mathcal{D}'') \propto \underbrace{p(\mathcal{D}''|\theta)}_{\text{lihelihood for}\mathcal{D}''} \times \underbrace{p(\theta|\mathcal{D}')}_{\text{prior for } \mathcal{D}'' \text{ based on } \mathcal{D}'}$$

but this can be shown as follows

$$p(\theta|\mathcal{D}',\mathcal{D}'') = \frac{p(\theta,\mathcal{D}''|\mathcal{D}')}{p(\mathcal{D}''|\mathcal{D}')} = \frac{p(\mathcal{D}''|\theta,\mathcal{D}')p(\theta|\mathcal{D}')}{p(\mathcal{D}''|\mathcal{D}')}$$

- note that $p(\mathcal{D}''|\theta,\mathcal{D}') = p(\mathcal{D}''|\theta)$ since $\mathcal{D}''$ and $\mathcal{D}'$ are independent
- hence we obtain the first equation above

N.B.: the above equation shows that Bayesian inference is well-suited for **online learning**

# Outline

let's revise the beta distribution

- $X$ is a continuous RV with values $x \in [0, 1]$
- $X \sim \text{Beta}(a, b)$, i.e. $X$ has a **beta distribution**

$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}$$

- requirements: $a > 0$ and $b > 0$
- the **beta function** is

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

- mean $\mathbb{E}[X] = \frac{a}{a+b}$
- mode $\frac{a-1}{a+b-2}$
- variance $\text{var}[X] = \frac{ab}{(a+b)^2(a+b+1)}$

**N.B.** the above equations will be used in the following slide

# The Beta-Binomial Model

## Posterior Mean and Mode

- $\theta_{MLE} = \arg\max_\theta p(\mathcal{D}|\theta) = \arg\max_\theta \left[\theta^{N_1}(1-\theta)^{N_0}\right] = \frac{N_1}{N}$ (**homework**: ex 3.1)
- **posterior mode**:

$$\theta_{MAP} = \arg\max_\theta p(\theta|\mathcal{D}) = \arg\max_\theta \text{Beta}(\theta|N_1 + a, N_0 + b) = \frac{a + N_1 - 1}{a + b + N - 2}$$

- **posterior mean**:

$$\mathbb{E}[\theta|\mathcal{D}] = \int_0^1 \theta p(\theta|\mathcal{D}) d\theta = \frac{a + N_1}{a + b + N} = \frac{a + N_1}{\alpha_0 + N}$$

- **prior mean**:

$$\mathbb{E}[\theta] = \int_0^1 \theta p(\theta) d\theta = \int_0^1 \theta \text{Beta}(\theta|a, b) d\theta = \frac{a}{\alpha_0}$$

where $a$ and $\alpha_0$ respectively play the role of $N_1$ and $N$

# The Beta-Binomial Model

## Posterior Mean and Mode

- $\theta_{MLE} = \frac{N_1}{N}$
- $\theta_{MAP} = \frac{a+N_1-1}{a+b+N-2}$
- $\mathbb{E}[\theta|\mathcal{D}] = \frac{a+N_1}{\alpha_0+N}$
- **prior mean**: $\mathbb{E}[\theta] = \int \theta p(\theta) d\theta = m_1 = \dfrac{a}{\alpha_0}$
- the posterior mean can decomposed as

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{m_1\alpha_0 + N_1}{\alpha_0 + N} = m_1 \frac{\alpha_0}{\alpha_0 + N} + \frac{N}{\alpha_0 + N}\frac{N_1}{N} = \lambda m_1 + (1-\lambda)\theta_{MLE}$$

were $\lambda \triangleq \dfrac{\alpha_0}{\alpha_0 + N}$

- the weaker the prior, the smaller $\lambda$, the closer $\mathbb{E}[\theta|\mathcal{D}]$ to $\theta_{MLE}$, hence

$$\lim_{N\to\infty} \mathbb{E}[\theta|\mathcal{D}] = \theta_{MLE}$$

# The Beta-Binomial Model

## Posterior Predictive

- now let's focus on prediction of future data
- the **posterior predictive** is

$$p(\tilde{x}=1|\mathcal{D}) = \int_0^1 p(\tilde{x}=1,\theta|\mathcal{D})d\theta = \int_0^1 p(\tilde{x}=1|\theta,\mathcal{D})p(\theta|\mathcal{D})d\theta =$$

$$\text{(data iid, } \tilde{x} \text{ independent from } \mathcal{D}) = \int_0^1 p(\tilde{x}=1|\theta)p(\theta|\mathcal{D})d\theta =$$

$$= \int_0^1 \theta \text{Beta}(\theta|N_1+a, N_0+b)d\theta = \mathbb{E}[\theta|\mathcal{D}]$$

- here we have used the Bayesian procedure of **integrating out** the unknown parameter
- if we reconsider the above equation

$$p(\tilde{x}|\mathcal{D}) = \int_0^1 p(\tilde{x}|\theta)p(\theta|\mathcal{D})d\theta = \int_0^1 \text{Ber}(\tilde{x}|\theta)p(\theta|\mathcal{D})d\theta$$

and we **plug-in**[2] $\hat{\theta} = \mathbb{E}[\theta|\mathcal{D}]$ we obtain $p(\tilde{x}|\mathcal{D}) = \text{Ber}(\tilde{x}|\mathbb{E}[\theta|\mathcal{D}])$

[2] recall the plug-in approximation $p(\theta|\mathcal{D}) \approx \delta_{\hat{\theta}}(\theta)$

# Outline

1. The Beta-Binomial Model
   - Problem Definition
   - Likelihood
   - Prior
   - Posterior
   - Posterior Predictive
   - Overfitting and the Black Swan Paradox
2. The Dirichlet-multinomial
   - Problem Definition
   - Likelihood
   - Prior
   - Posterior
   - Posterior Predictive

- let's consider the plug-in approximation with $\theta_{MLE} = N_1/N$, we obtain

$$p(\tilde{x}|\mathcal{D}) \approx \mathrm{Ber}(\tilde{x}|\theta_{MLE})$$

- the MLE estimate performs very bad with small datasets
- for instance, suppose we observed $N_1 = 0$ and $N_0 = 3$, in this case $\theta_{MLE} = 0$ and we predict that heads is impossible
- this is called the **zero count problem** or **sparse data problem**
- this problem is analogous to the **black swan paradox**: Western conception that all swans were white; black swans were discovered in Australia in the 17th Century

- now let's see the same problem in a Bayesian perspective
- assume a beta prior $p(\theta) = \text{Beta}(a, b)$ with $a = b = 1$ (uniform prior)
- as already computed

$$p(\tilde{x} = 1|\mathcal{D}) = \mathbb{E}[\theta|\mathcal{D}] = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

- this justifies the common practice of adding 1 to the counts (**add-one smoothing**)
- in this case even if $N_1 = 0$ and $N_0 = 3$ we have $p(\tilde{x} = 1|\mathcal{D}) = 1/4 \neq 0$

# Outline

problem definition

- consider a series of $N$ **dice rolls**
- the dice has $K$ faces
- we would like to **infer the probability** $\theta_j \in [0, 1]$ that the $j$-th dice face shows up, given a series of observations
- in this case we have a **continuous random variable** $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ with $\theta_j \in [0, 1]$ and $\sum_{j=1}^{K} \theta_j = 1$

# Outline

# The Dirichlet-Multinomial

## Likelihood

- suppose we observe $N$ dice rolls
- for each $i$-th dice roll we have a discrete RV $X_i \sim \text{Cat}(\boldsymbol{\theta})$, where $X_i = j$ means $j$-the face have shown up
- the dataset is $\mathcal{D} = \{x_1, ..., x_N\}$ where $x_i \in \{1, ..., K\}$ for $i \in 1, ..., N$
- since data is assumed iid, the **likelihood function** is

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \text{Cat}(x_i|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_i=k)} = \prod_{j=1}^{K} \theta_j^{N_k}$$

  where $N_k = \sum_{i=1}^{N} \mathbb{I}(x_i = k)$ is the number of times face $k$ is observed
- this likelihood function is proportional to the multinomial distribution

$$\mathbf{Mu}(N_1, ..., N_K|N, \boldsymbol{\theta}) = \binom{N}{N_1 ... N_K} \prod_{j=1}^{K} \theta_j^{N_k}$$

  since the multinomial coefficient $\binom{N}{N_1 ... N_K}$ does not depend on $\theta$

# Outline

- the RV $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ lives in a $K$-dimensional **probability simplex** $S_K$

$$S_K = \{\boldsymbol{\theta} \in \mathbb{R}^K : \ \theta_j \in [0,1], \quad \sum_{j=1}^{K} \theta_j = 1\}$$

- we need a prior that *(i)* supports the probability simplex and *(ii)* ideally is conjugate for the likelihood (prior and posterior have the same form)
- the Dirichlet distribution satisfies both criteria

$$\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \mathbb{I}(\boldsymbol{\theta} \in S_K)$$

# Outline

# The Dirichlet-Multinomial

## Posterior

- we obtain the posterior as usual

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto \prod_{j=1}^{K} \theta_j^{N_k}\theta_j^{\alpha_k - 1} = \prod_{j=1}^{K} \theta_j^{N_k+\alpha_k-1} \propto \text{Dir}(\boldsymbol{\theta}|\alpha_1+N_1, ..., \alpha_K+N_K)$$

where the $\alpha_j$ are the **pseudo-counts** and the $N_j$ are the **empirical counts**

- $\alpha_0 \triangleq \sum_{j=1}^{K} \alpha_j$ is the **equivalent sample size** of the prior and determines its strength

- the mode of the posterior can be derived by using a Lagrange multiplier
- we want to maximize $f(\boldsymbol{\theta}) = log(p(\boldsymbol{\theta}|\mathcal{D}))$ subject to $g(\boldsymbol{\theta}) \triangleq 1 - \sum_{j=1}^{N} \theta_j = 0$
- let's define the **Lagrangian function**

$$l(\boldsymbol{\theta}, \lambda) \triangleq f(\boldsymbol{\theta}) + \lambda g(\boldsymbol{\theta})$$

  where $\lambda$ is the Lagrange multiplier
- in order to optimize $f(\boldsymbol{\theta})$ subject to the constraint $g(\boldsymbol{\theta}) = 0$ we have to impose

$$\frac{\partial l}{\partial \lambda} = 0$$
$$\frac{\partial l}{\partial \theta_j} = 0 \quad \text{for} \quad j \in \{1, 2, ..., K\}$$

# The Dirichlet-Multinomial

## Posterior Mean and Mode

- we want to maximize $f(\boldsymbol{\theta}) = log(p(\boldsymbol{\theta}|\mathcal{D}))$ subject to $g(\boldsymbol{\theta}) \triangleq 1 - \sum_{j=1}^{K} \theta_j = 0$
- the **Lagrangian function** is

$$l(\boldsymbol{\theta}, \lambda) \triangleq f(\boldsymbol{\theta}) + \lambda g(\boldsymbol{\theta}) = log(p(\boldsymbol{\theta}|\mathcal{D})) + \lambda g(\boldsymbol{\theta}) =$$

$$= \sum_j N_j \log \theta_j + \sum_j (\alpha_j - 1) \log \theta_j + \lambda \Big( 1 - \sum_j \theta_j \Big)$$

- in order to solve the constrained optimization we impose

$$\frac{\partial l}{\partial \lambda} = 1 - \sum_{j=1}^{K} \theta_j = 0$$

$$\frac{\partial l}{\partial \theta_j} = \frac{N'_j}{\theta_j} - \lambda = 0 \quad \Rightarrow \quad N'_j = \lambda \theta_j$$

where $N'_j \triangleq N_j + \alpha_j - 1$

# The Dirichlet-Multinomial

## Posterior Mean and Mode

- we can solve the following equations by plugging-in the second in the first

$$1 - \sum_{j=1}^{K} \theta_j = 0$$

$$N'_j = \lambda \theta_j$$

and get

$$\sum_j N'_j = \lambda \quad \Rightarrow \quad N + \alpha_0 - K = \lambda$$

where $\alpha_0 = \sum_{j=1}^{K} \alpha_j$

- the MAP estimate is obtained as

$$\theta_j^{MAP} = \frac{N_j + \alpha_j - 1}{N + \alpha_0 - K}$$

- the MLE estimate is obtained by using a uniform prior[3], i.e. $\alpha_j = 1$

$$\theta_j^{MLE} = \frac{N_j}{N}$$

[3] recall that with $p(\boldsymbol{\theta}) \propto 1$ one has $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})$

# Outline

- now let's focus on prediction of future data
- the **posterior predictive** is

$$p(\tilde{x} = j|\mathcal{D}) = \int p(\tilde{x} = j, \boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} = \int p(\tilde{x} = j|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} =$$

$$(\text{data iid, } \tilde{x} \text{ independent from } \mathcal{D}) = \int p(\tilde{x} = j|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} =$$

$$= \int p(\tilde{x} = j|\theta_j)\left[\int p(\boldsymbol{\theta}_{-j}, \theta_j|\mathcal{D})d\boldsymbol{\theta}_{-j}\right]d\theta_j =$$

$$= \int \theta_j p(\theta_j|\mathcal{D})d\theta_j = \mathbb{E}[\theta_j|\mathcal{D}] = \frac{\alpha_j + N_j}{\sum_j(\alpha_j + N_j)} = \frac{\alpha_j + N_j}{\alpha_0 + N}$$

- $\boldsymbol{\theta}_{-j}$ is the vector $\boldsymbol{\theta}$ without the $j$-th component
- for the last two passages check the mean value of a Dirichlet distribution
- again we have used the Bayesian procedure of **integrating out** the unknown parameter
- as with the beta-binomial model, the Bayesian approach solves the zero-count problem (when for some $j \in \{1, .., K\}$ we observe $N_j = 0$)

# Credits

- Kevin Murphy's book