Biostatistics 682: Applied Bayesian Inference
Lecture 5: More on multiparameter model

Jian Kang

Department of Biostatistics
University of Michigan, Ann Arbor

Motivating example: Clinical trial

- A new treatment has three possible outcomes for patients:
  - 1: Improved
  - 2: No change
  - 3: Worsened
- Data from $n = 20$ patients:

$$(y_1, y_2, y_3) = (12, 6, 2).$$

- Prior: $\theta = (\theta_1, \theta_2, \theta_3) \sim \text{Dirichlet}(2, 2, 2)$.

What questions might clinicians ask?

- What is the chance that more than half of patients improve?

$$P(\theta_1 > 0.5 \mid y).$$

- What is the chance fewer than 20% of patients worsen?

$$P(\theta_3 < 0.2 \mid y).$$

- What is the probability that the treatment helps more than it harms?

$$P(\theta_1 > \theta_3 \mid y).$$

- In 10 new patients, what is the probability that at least 7 improve? Suppose $(z_1, z_2, z_3) \mid \theta \sim \text{Multinomial}(\theta, m)$, where $m = 10$.

$$P(z_1 \geq 7 \mid y) =$$

Gamma construction of the Dirichlet

- A convenient representation of the Dirichlet distribution uses Gamma random variables.
- Let
$$X_i \overset{\text{ind}}{\sim} \Gamma(\alpha_i, 1), \quad i = 1, \ldots, k.$$
- Define
$$\theta_i = \frac{X_i}{\sum_{j=1}^{k} X_j}, \quad i = 1, \ldots, k.$$
- Then
$$\theta = (\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k).$$
- Key advantage: This representation makes it easy to prove important properties:
  - Collapsing categories
  - Re-normalization
  - Marginal Beta distributions

Dirichlet property: collapsing categories

- Gamma construction: $X_i \sim \Gamma(\alpha_i, 1)$ independent, $\theta_i = X_i / \sum_j X_j$.
- Combine categories i and j:

$$\theta_i + \theta_j = \frac{X_i + X_j}{\sum_{l=1}^{k} X_l}.$$

- Since $X_i + X_j \sim \Gamma(\alpha_i + \alpha_j, 1)$, the vector

$$(\theta_i + \theta_j,\, \theta_{\text{others}}) \sim \text{Dirichlet}(\alpha_i + \alpha_j,\, \alpha_{\text{others}}).$$

- Interpretation: categories can be grouped without leaving the Dirichlet family.

Dirichlet property: re-normalization

- Gamma construction: $X_i \sim \Gamma(\alpha_i, 1)$, independent.
- For subset $S \subseteq \{1, \ldots, k\}$:

$$\frac{\theta_i}{\sum_{j \in S} \theta_j} = \frac{X_i}{\sum_{j \in S} X_j}, \quad i \in S.$$

- But $\{X_i : i \in S\}$ are independent Gammas with shape parameters $(\alpha_i, i \in S)$.
- Hence

$$\left(\frac{\theta_i}{\sum_{j \in S} \theta_j}, \ i \in S \right) \sim \text{Dirichlet}(\alpha_i, \ i \in S),$$

  independent of $\sum_{j \in S} \theta_j \sim \Gamma(\sum_{j \in S} \alpha_j, 1)$.
- Example: $\text{Dirichlet}(1, 2, 3) \Rightarrow (\theta_1/(\theta_1 + \theta_3), \theta_3/(\theta_1 + \theta_3)) \sim \text{Dirichlet}(1, 3)$.

Dirichlet distribution: Marginals

- Gamma construction: $X_i \sim \Gamma(\alpha_i, 1)$, $\theta_i = X_i / \sum_j X_j$.
- Consider $(X_i,\ \sum_{j \neq i} X_j)$.
  - $X_i \sim \Gamma(\alpha_i, 1)$
  - $\sum_{j \neq i} X_j \sim \Gamma(\alpha_0 - \alpha_i, 1)$, independent.
- Then

$$\theta_i = \frac{X_i}{X_i + \sum_{j \neq i} X_j} \sim \text{Beta}(\alpha_i,\ \alpha_0 - \alpha_i).$$

- Moments:

$$\mathbb{E}[\theta_i] = \frac{\alpha_i}{\alpha_0}, \qquad \text{Var}(\theta_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}.$$

- Interpretation: each α_i is a pseudo-count.

Marginal distribution: Integration approach

- Let $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$ denote all coordinates except θ_i.
- Marginal density obtained by integrating out θ_{-i}:

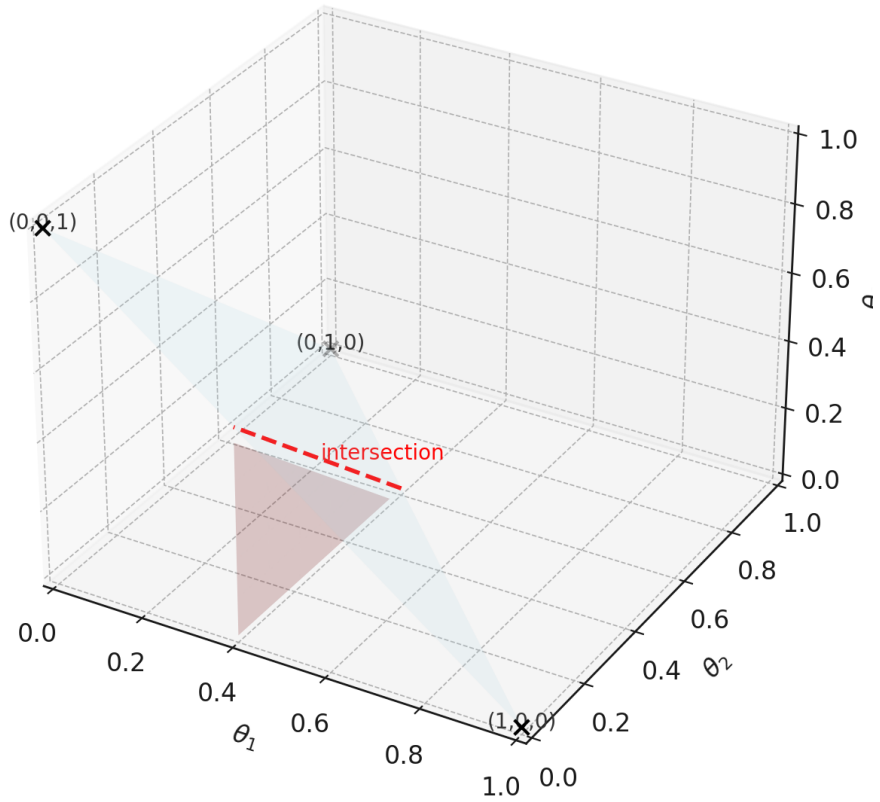
$$\pi(\theta_i) = \int_{\{\theta_{-i} \geq 0,\ \sum_{j \neq i} \theta_j = 1 - \theta_i\}} \pi(\theta)\, d\theta_{-i}.$$

  Consider variable transformation $\theta_j^* = \theta_j/(1-\theta_i)$. Then $\theta_{-i}^* = (\theta_1^*, \ldots, \theta_{i-1}^*, \theta_{i+1}^*, \ldots, \theta_k^*)$ follows a Dirichlet distribution with precision parameters $(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_k)$.
- Substitution gives:

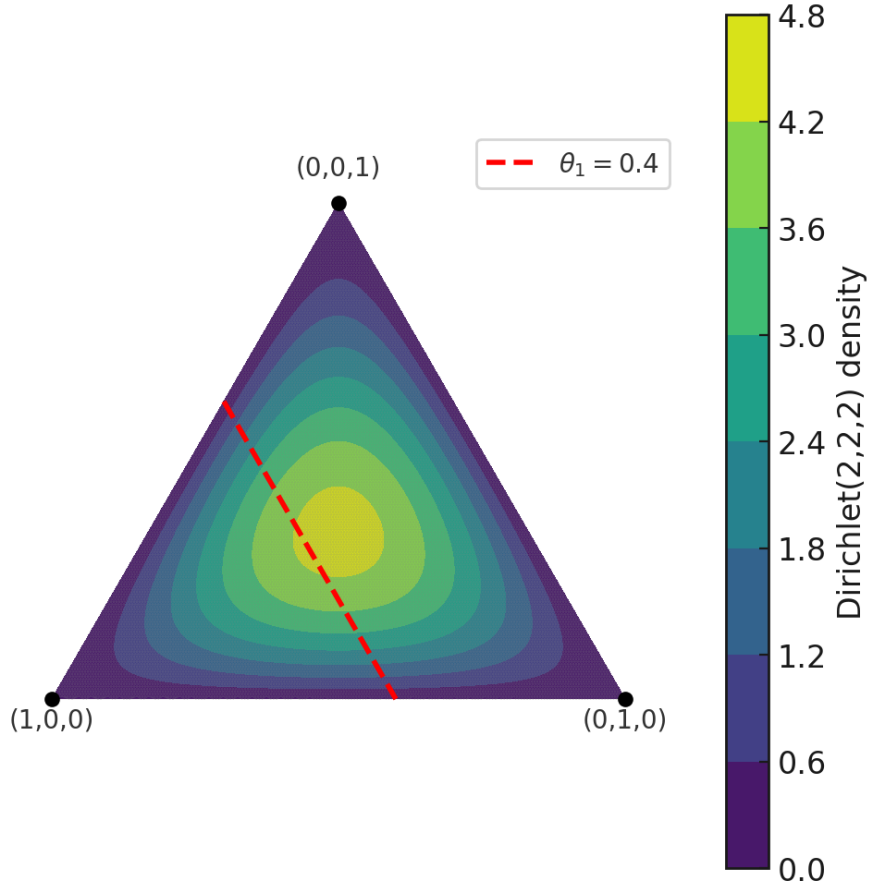
$$f(\theta_i) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_i)\,\Gamma(\alpha_0 - \alpha_i)}\ \theta_i^{\alpha_i - 1}\,(1-\theta_i)^{\alpha_0 - \alpha_i - 1}.$$

- Result: $\theta_i \sim \mathrm{Beta}(\alpha_i,\ \alpha_0 - \alpha_i)$.

Dirichlet(2,2,2): geometry and density



3D view: The simplex (blue triangle) lies in $\theta_1 + \theta_2 + \theta_3 = 1$. Fixing $\theta_1 = 0.4$ defines a vertical plane (red), whose intersection with the simplex is a line segment.



2D density: The Dirichlet(2,2,2) density is highest in the interior. The red dashed line shows the slice $\theta_1 = 0.4$, which integrates to a Beta marginal density for θ_1 at 0.4.

Posterior marginals

- Data: $y = (y_1, \ldots, y_k) \sim \text{Multinomial}(n, \theta)$ where $\sum_{j=1}^{k} y_j = n$.
- Posterior:
$$\theta \mid y \sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_k + y_k).$$
- Marginal for category i:
$$\theta_i \mid y \sim \text{Beta}(\alpha_i + y_i,\ \alpha_0 + n - (\alpha_i + y_i)).$$
where $\alpha_0 = \sum_{j=1}^{k} \alpha_j$.
- Posterior mean:
$$\mathbb{E}[\theta_i \mid y] = \frac{\alpha_i + y_i}{\alpha_0 + n}.$$
- Interpretation: Posterior = prior pseudo-counts + observed counts.

Posterior predictive distribution

- Posterior after observing counts $y = (y_1, \ldots, y_k)$:

$$\theta \mid y \sim \text{Dirichlet}(\alpha'_1, \ldots, \alpha'_k), \quad \alpha'_i = \alpha_i + y_i.$$

- Predictive for new counts $z = (z_1, \ldots, z_k)$ with m trials:

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

- Likelihood:

$$\pi(z \mid \theta) = \frac{m!}{\prod_{i=1}^k z_i!} \prod_{i=1}^k \theta_i^{z_i}.$$

- Integration $\Rightarrow$ Dirichlet–Multinomial distribution:

$$\pi(z \mid y) = \frac{m!}{\prod_{i=1}^k z_i!} \frac{\Gamma(\alpha'_0)}{\Gamma(\alpha'_0 + m)} \prod_{i=1}^k \frac{\Gamma(\alpha'_i + z_i)}{\Gamma(\alpha'_i)}, \quad \alpha'_0 = \sum_{i=1}^k \alpha'_i.$$

Collapsed case: Beta–Binomial predictive

- Suppose we collapse categories into two groups:

$$z_A = \sum_{i \in A} z_i, \qquad z_B = \sum_{i \in B} z_i, \qquad A \cup B = \{1, \ldots, k\}, \quad A \cap B = \emptyset;$$

- Then

$$(z_A, z_B) \sim \text{DirichletMultinomial}\big(m;\ \alpha'_A,\ \alpha'_B\big),$$

where $\alpha'_A = \sum_{i \in A} \alpha'_i$ and $\alpha'_B = \sum_{i \in B} \alpha'_i$.

- This reduces to the Beta–Binomial distribution:

$$z_A \sim \text{BetaBin}(m;\ \alpha'_A,\ \alpha'_B),$$

with pmf

$$P(z_A = z) = \binom{m}{z} \frac{B(z + \alpha'_A,\ m - z + \alpha'_B)}{B(\alpha'_A,\ \alpha'_B)}.$$

- Example (clinical trial): Improved vs Not improved

$$z_1 \sim \text{BetaBin}(10;\ \alpha'_1 = 14,\ \alpha'_2 + \alpha'_3 = 12).$$

Recap: Motivating example

- Clinical trial with $n = 20$ patients, three outcomes:

$$(y_1, y_2, y_3) = (12, 6, 2), \quad 1 = \text{Improved}, \ 2 = \text{No change}, \ 3 = \text{Worsened}.$$

- Prior: $\theta = (\theta_1, \theta_2, \theta_3) \sim \text{Dirichlet}(2, 2, 2)$.
- Posterior: $\theta \mid y \sim \text{Dirichlet}(14, 8, 4)$.
- Clinical questions to answer:
  1. $P(\theta_1 > 0.5 \mid y)$ (majority improve)
  2. $P(\theta_3 < 0.2 \mid y)$ (safety margin)
  3. $P(\theta_1 > \theta_3 \mid y)$ (benefit vs harm)
  4. $P(z_1 \geq 7 \mid y, m = 10)$ (predictive success in 10 new patients)
- Next: use properties of the Dirichlet distribution (marginals, collapsing) to solve each.

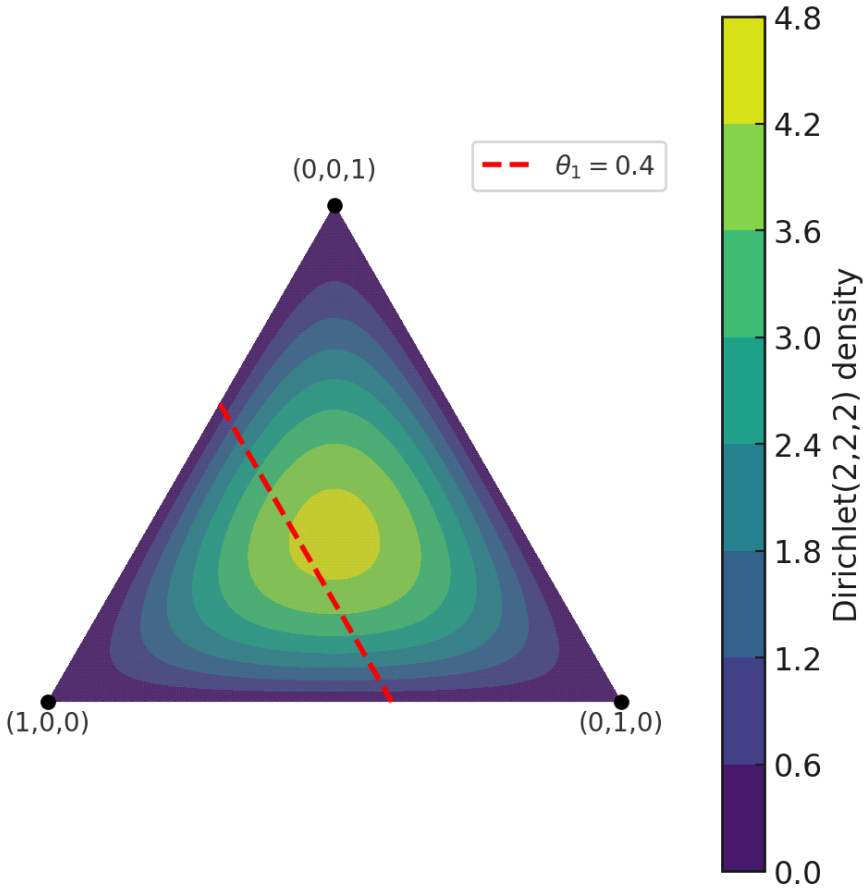
More than half improve

- Posterior: $\theta \mid y \sim \text{Dirichlet}(\alpha_1', \alpha_2', \alpha_3') = (14, 8, 4)$.
- Marginal: $\theta_1 \mid y \sim \text{Beta}\big(\alpha_1', \alpha_0' - \alpha_1'\big) = \text{Beta}(14, 12)$.
- Compute

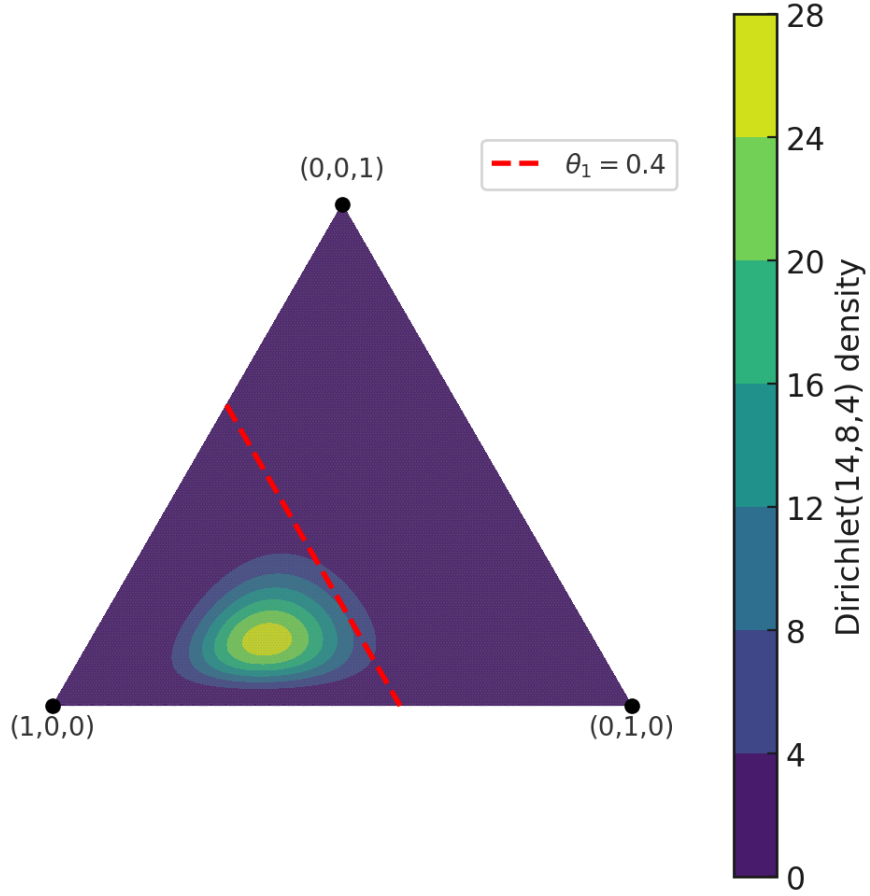
$$P(\theta_1 > 0.5 \mid y) = 1 - F_{\text{Beta}(14,12)}(0.5) \approx 0.655.$$

where F_{Beta} is the Beta CDF.

Prior vs Posterior Dirichlet densities



Prior: Dirichlet(2,2,2) Symmetric, most mass in the interior. The slice $\theta_1 = 0.4$ integrates to Beta(2,4).



Posterior: Dirichlet(14,8,4) Concentrated near "Improved" corner. The slice $\theta_1 = 0.4$ integrates to Beta(14,12).

Fewer than 20% worsen

- Posterior marginal: $\theta_3 \mid y \sim \mathrm{Beta}\big(\alpha_3', \alpha_0' - \alpha_3'\big) = \mathrm{Beta}(4, 22)$.
- Compute
$$P(\theta_3 < 0.2 \mid y) = F_{\mathrm{Beta}(4,22)}(0.2) \approx 0.766.$$
- Interpretation: posterior probability that the harm rate is below 20%.

Treatment helps more than it harms

- Posterior: $\theta \mid y \sim \text{Dirichlet}(14, 8, 4)$.
- Re-normalization property: For any subset of categories, the normalized proportions follow a Dirichlet with the corresponding parameters.
- Apply to categories 1 (Improved) and 3 (Worsened):

$$U = \frac{\theta_1}{\theta_1 + \theta_3} \sim \text{Beta}(14, 4).$$

- Event: $\theta_1 > \theta_3 \iff U > 0.5$.
- Numerical result:

$$P(\theta_1 > \theta_3 \mid y) \approx 0.994.$$

Predictive: at least 7 of 10 improve

- Posterior predictive:

$$z_1 \sim \text{BetaBin}(m = 10;\ \alpha_1' = 14,\ \alpha_2' + \alpha_3' = 12).$$

- Compute:

$$P(z_1 \geq 7 \mid y) = \sum_{z=7}^{10} \binom{10}{z} \frac{B(z + 14,\ 10 - z + 12)}{B(14, 12)}.$$

- Numerical result:

$$P(z_1 \geq 7 \mid y) \approx 0.279.$$

The Simplest Bayesian Language Model

- In statistics papers, the word `data` often appears in many contexts:
  - `data analysis`, `data are`, `data processing`, `data model`, ...
- Question: Given observed text, how can we model the distribution of words that appear immediately after `data`?
- Goal: Build a simple probabilistic model that allows us to
  1. Summarize the frequency of words following `data`.
  2. Estimate the probability of the next word (e.g., $P(\text{“analysis”} \mid \texttt{data})$).
  3. Make predictions about unseen or rare words.
- This is the simplest form of a Bayesian language model — a categorical distribution with a Dirichlet prior.

Bigram Dirichlet Language Model

- Context c = previous token (here $c =$ `'data'`).
- Next word $w \in \mathcal{V}$ with probabilities $\theta^{(c)} = (\theta_1^{(c)}, \ldots, \theta_K^{(c)})$.
- Prior: $\theta^{(c)} \sim \text{Dirichlet}(\alpha_1^{(c)}, \ldots, \alpha_K^{(c)})$.
- Counts from corpus: $y_j^{(c)} = \#\{w_j \text{ follows } c\}$.
- Posterior: $\theta^{(c)} \mid y^{(c)} \sim \text{Dirichlet}(\alpha_j^{(c)} + y_j^{(c)})$.
- Posterior predictive (next word):

$$P(w_{\text{next}} = w_j \mid c, y^{(c)}) = \frac{\alpha_j^{(c)} + y_j^{(c)}}{\sum_k (\alpha_k^{(c)} + y_k^{(c)})}.$$

Practical choices for real papers

- Vocabulary: lemmatize ("analyses" $\rightarrow$ "analysis"), lowercase, drop stopwords if desired; keep a `<other>` bucket.
- Prior: symmetric $\alpha = 0.5$ (mild smoothing), or informative prior (e.g., higher α on "analysis", "set").
- Unknowns: all terms outside top-K map to `<other>` with its own α.
- Quality checks: compare MLE $y_j^{(c)} / \sum y^{(c)}$ vs. posterior mean; compute top-k accuracy on a held-out set.
- Extensions: backoff/Interpolated models with previous two tokens (trigram), hierarchical Dirichlet priors across contexts, or Kneser–Ney style discounting.

Predicting the word after `data`

- Prior: symmetric $\alpha = 0.5$ over $\mathcal{V}$.
- Observed counts $y^{(\texttt{data})}$ from a stats corpus (examples): "analysis"=34, "set"=18, "is"=15, "are"=12, "model"=11, <other>=25, ...
- Posterior predictive (top terms):

$$P(\text{“analysis”} \mid \texttt{data}) \approx 0.186,$$
$$P(\texttt{<other>} \mid \texttt{data}) \approx 0.138,$$
$$P(\text{“set”} \mid \texttt{data}) \approx 0.100,$$
$$P(\text{“is”}) \approx 0.084,$$
$$P(\text{“are”}) \approx 0.067,$$

Why Dirichlet smoothing?

- MLE assigns zero probability to unseen words: $P_{\text{MLE}}(w|c) = 0$ when $y_w^{(c)} = 0$.
- Bayesian fix: $P(w|c, y) = \dfrac{\alpha_w + y_w}{\sum_k (\alpha_k + y_k)}$ keeps a nonzero mass for rare/unseen words.
- Interpretation: α_w are pseudo-counts encoding prior beliefs.
- Outcome: better generalization on held-out text; smoother top-k predictions.