How to train your program

A probabilistic programming pattern for Bayesian learning from data

[arXiv:2105.03650](https://arxiv.org/abs/2105.03650)

PUB+ Ben-Gurion University of the Negev KAIST

David Tolpin
Hongseok Yang

Hierarchical models

Definition

- Each group y_i is conditioned on θ_i.
- All θ_i are conditioned on τ.

$$\begin{aligned}\tau &\sim H\\ \theta_i &\sim D(\tau)\\ y_i &\sim F(\theta_i)\end{aligned}$$

Example

- Multiple boxes are randomly filled by K marbles from a bag.
- y_{ij} is the jth draw from the ith box.
- Infer the the number of blue marbles θ_i in each box.

$$\begin{aligned}\tau &\sim \mathrm{Beta}(1, 1)\\ \theta_i &\sim \mathrm{Binomial}(K, \tau)\\ y_{ij} &\sim \mathrm{Bernoulli}\left(\frac{\theta_i}{K}\right)\end{aligned}$$

A hierarchical model cannot compress training data

- Learns from data — each box's θ_i is influenced by draws from all boxes.
- Learning from *training data* is computationally inefficient — full data set must be held.

Stump and fungus

Observation

- In Bayesian modelling, information about data is con–veyed through conditioning on the data.
- In a hierarchical model, influence of the ith group on hy–perparameter τ passes through group parameters θ_i.

Pattern

- Training is accomplished through inference on a hierar–chical model, in the usual way.
- Training outcomes are summarized as a collection of samples $\tilde\theta$, representing the mixture distribution of θ_i of all groups.
- For inference on new data item y, a stump–and–fungus model is employed:

$$\tilde\theta \sim \mathrm{Hierarchical}(Y)$$

– – – – – – – – – – – – – –

$$\begin{aligned}\tau &\sim H\\ \tilde\theta, \theta|\tau &\sim D(\tau)\\ y|\theta &\sim F(\theta)\end{aligned}$$

Problem: learning from data

- Population $\mathcal{Y}$ is a set of sets $y_i \in Y$ of observations $y_{ij} \in y_i$.
- Members of each y_i are drawn from distribution F with unobserved parameter θ_i, $y_{ij} \sim F(\theta_i)$.
- θ_i are drawn from a common distribution H.

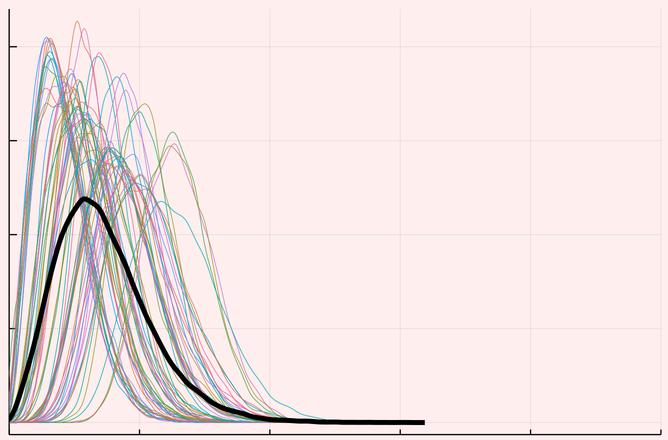
- **Goal:** devise a scheme that,
  - given a *training set* $Y \subset \mathcal{Y}$,
  - infers the posterior distribution of $\theta_k|Y, y_k$ for any $y_k \in \mathcal{Y}$
  - **in a shorter amortized time** than running inference on a hierarchical model $Y \cup \{y_k\}$.

Tumor incidence in rats

Hierarchical

$$\begin{aligned}\alpha, \beta &\sim \mathrm{Uniform}(0, \infty)\\ p_i|\alpha, \beta &\sim \mathrm{Beta}(\alpha, \beta)\\ y_i|p_i &\sim \mathit{Binomial}(n_i, p_i)\end{aligned}$$

Posterior



Stump and fungus

$$p_{-k} \sim \mathrm{Hierarchical}(n_{-k}, y_{-k})$$

– –

$$\begin{aligned}\alpha, \beta &\sim \mathrm{Uniform}(0, \infty)\\ p|\alpha, \beta &\sim \mathrm{Beta}(\alpha, \beta)\\ y_k|p_k &\sim \mathrm{Binomial}(n_k, p_k)\end{aligned}$$

Posterior