

## Overview

- In Bayesian treatments of machine learning, the success or failure of the estimator/classifier hinges on how well the prior distribution selected by the designer matches the actual data-generating model
- Highly localized Dirichlet priors can overcome the burden of a limited training set when the prior mean is well matched to the true distribution, but will degrade the approximation if the match is poor

## Objective

**Make inferences about unobservable  $y \in \mathcal{Y}$  given observed  $x \in \mathcal{X}$  and a training set  $D \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$**

- Joint elements  $(y, x)$  and  $D_n$  are distributed by an *unknown* PMF  $\theta$ :

$$\begin{aligned} P_{y,x,D|\theta}(y, x, D|\theta) &= P_{y,x|\theta}(y, x|\theta) \prod_{n=1}^N P_{D_n|\theta}(D_n|\theta) \\ &= \theta(y, x) \prod_{y' \in \mathcal{Y}} \prod_{x' \in \mathcal{X}} \theta(y', x')^{\bar{N}(y', x'; D)} \end{aligned}$$

- $P_{D|\theta}$  depends on  $D$  only through the transform  $\bar{N}(y, x; D) \equiv \sum_{n=1}^N \delta[(y, x), D_n]$   
 ⇒ Random process  $\bar{n} \equiv \bar{N}(D) \in \bar{\mathcal{N}}$  is a sufficient statistic for  $\theta$ ; decisions can depend on  $\bar{n}$  in place of  $D$

Design a decision function  $f: \bar{\mathcal{N}} \mapsto \mathcal{H}^{\mathcal{X}}$ , where  $\mathcal{H}$  is the decision space.

The metric is a loss function  $\mathcal{L}: \mathcal{H} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ .

Clairvoyant Risk:

$$\mathcal{R}_{\theta}(f) = E_{x, \bar{n}|\theta} \left[ E_{y|x, \theta} \left[ \mathcal{L}(f(x; \bar{n}), y) \right] \right]$$

⇒ Optimal decisions depend on the *true predictive distribution*,  $P_{y|x, \theta} = \theta(\cdot, x) / \sum_{y \in \mathcal{Y}} \theta(y, x) \equiv \tilde{\theta}(x)$

↓ ↓ **Select Prior  $p_{\theta}$**  ↓ ↓

Bayes Risk:

$$\mathcal{R}(f) = E_{\theta} \left[ \mathcal{R}_{\theta}(f(x; \bar{n})) \right] = E_{x, \bar{n}} \left[ E_{y|x, \bar{n}} \left[ \mathcal{L}(f(x; \bar{n}), y) \right] \right]$$

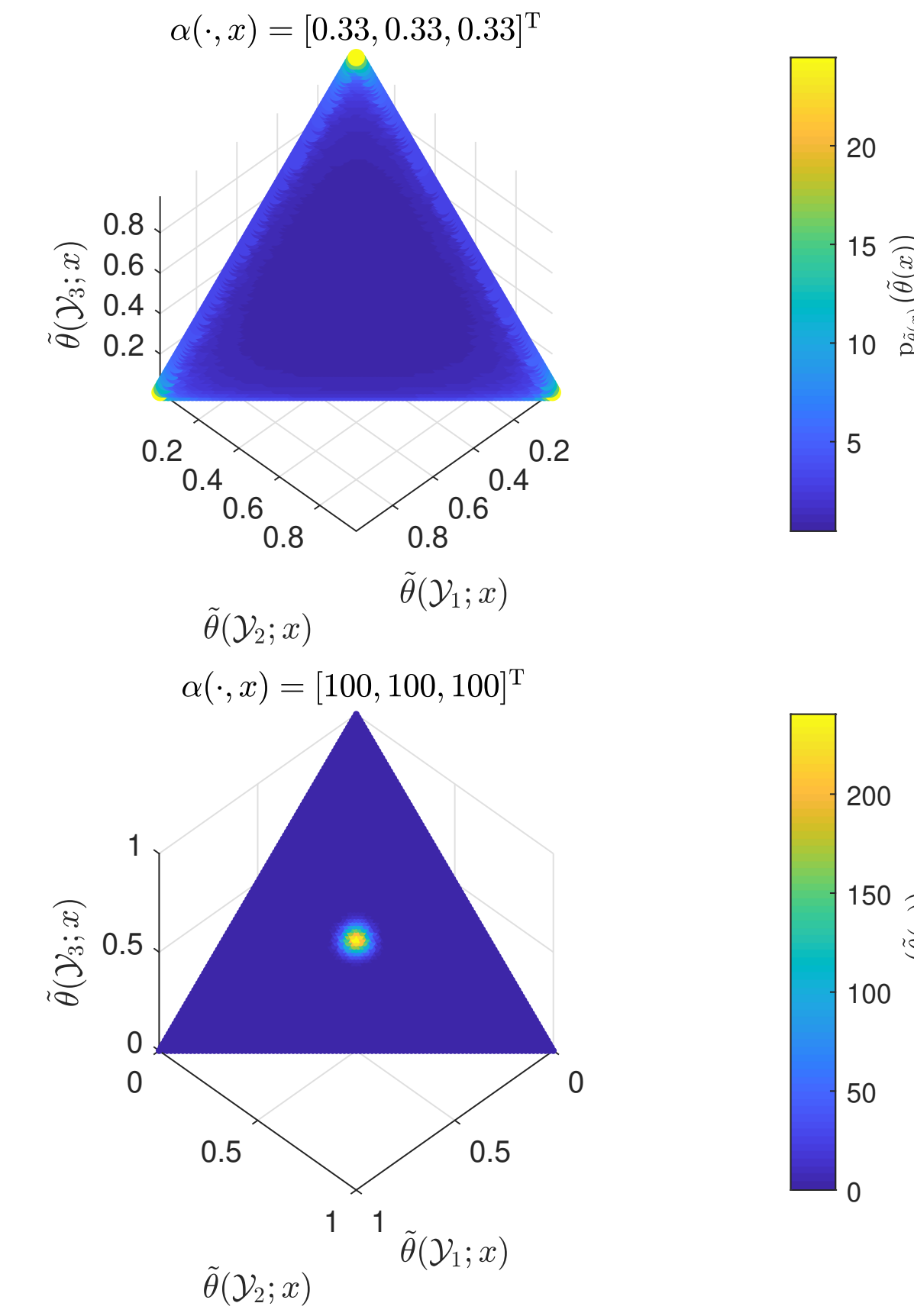
⇒ Decisions formulated using *Bayes predictive distribution*,  $P_{y|x, \bar{n}} = E_{\theta|x, \bar{n}} [P_{y|x, \theta}] = \mu_{\tilde{\theta}(x)|x, \bar{n}}$

## Bayesian Prediction

Dirichlet Priors:

$$\begin{aligned} p_{\theta}(\theta) &= \text{Dir}(\theta; \alpha) \\ &= \beta(\alpha)^{-1} \prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y, x)^{\alpha(y, x) - 1} \\ &\Downarrow \\ p_{\tilde{\theta}}(\tilde{\theta}) &= \prod_{x \in \mathcal{X}} \text{Dir}(\tilde{\theta}(x); \alpha(\cdot, x)) \end{aligned}$$

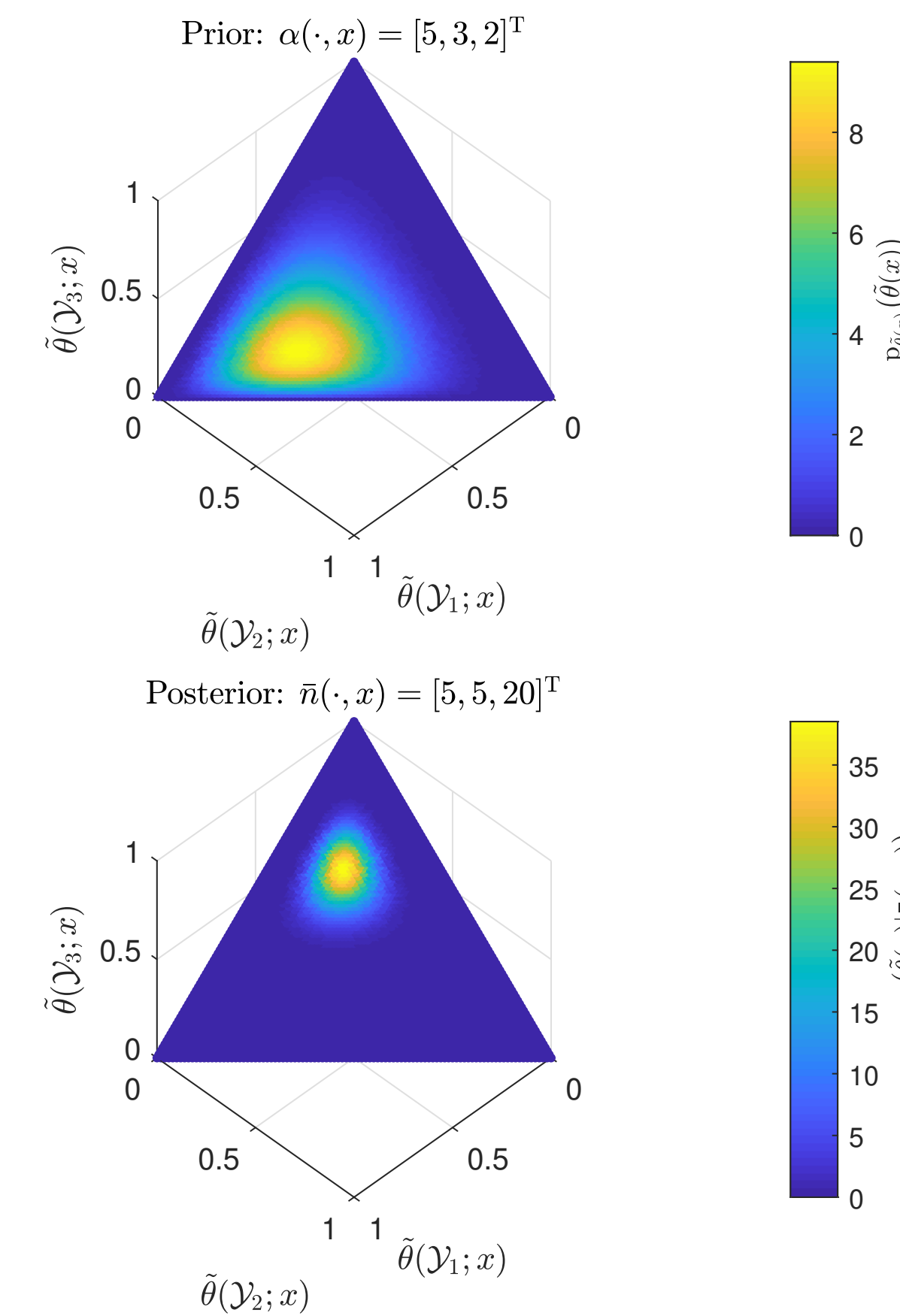
- Concentration parameters  $\alpha'(x) \equiv \sum_{y \in \mathcal{Y}} \alpha(y, x)$  enable both *subjective* and *non-informative* priors
- Conjugate Prior for I.I.D. observations ⇒ Tractable Posterior



Dirichlet Posteriors:

$$\begin{aligned} p_{\tilde{\theta}|x, \bar{n}}(\tilde{\theta}|x, \bar{n}) &= p_{\tilde{\theta}|\bar{n}}(\tilde{\theta}|\bar{n}) \\ &= \prod_{x' \in \mathcal{X}} p_{\tilde{\theta}(x')|\bar{n}(\cdot, x')}(\tilde{\theta}(x')|\bar{n}(\cdot, x')) \\ &= \prod_{x' \in \mathcal{X}} \text{Dir}(\tilde{\theta}(x'); \alpha(\cdot, x') + \bar{n}(\cdot, x')) \end{aligned}$$

- **Full support** over distribution space ensures identification of model  $\tilde{\theta}(x)$  as  $n'(x) \equiv \sum_y \bar{n}(y, x) \rightarrow \infty$



Bayes Predictive Distribution:

$$P_{y|x, \bar{n}} = \left( \frac{\alpha'(x)}{\alpha'(x) + \sum_y \bar{n}(y, x)} \right) \frac{\alpha(\cdot, x)}{\alpha'(x)} + \left( \frac{\sum_y \bar{n}(y, x)}{\alpha'(x) + \sum_y \bar{n}(y, x)} \right) \frac{\bar{n}(\cdot, x)}{\sum_y \bar{n}(y, x)}$$

⇒ Convex combination of data-independent PMF and conditional empirical PMF

$$\Rightarrow E_{\bar{n}|n', \theta} [P_{y|x, \bar{n}}] = \left( \frac{\alpha'(x)}{\alpha'(x) + n'(x)} \right) \frac{\alpha(\cdot, x)}{\alpha'(x)} + \left( \frac{n'(x)}{\alpha'(x) + n'(x)} \right) \tilde{\theta}(x)$$

## Density Estimation

- \* Density estimation accuracy assessed using the *Estimation Difference Function*:

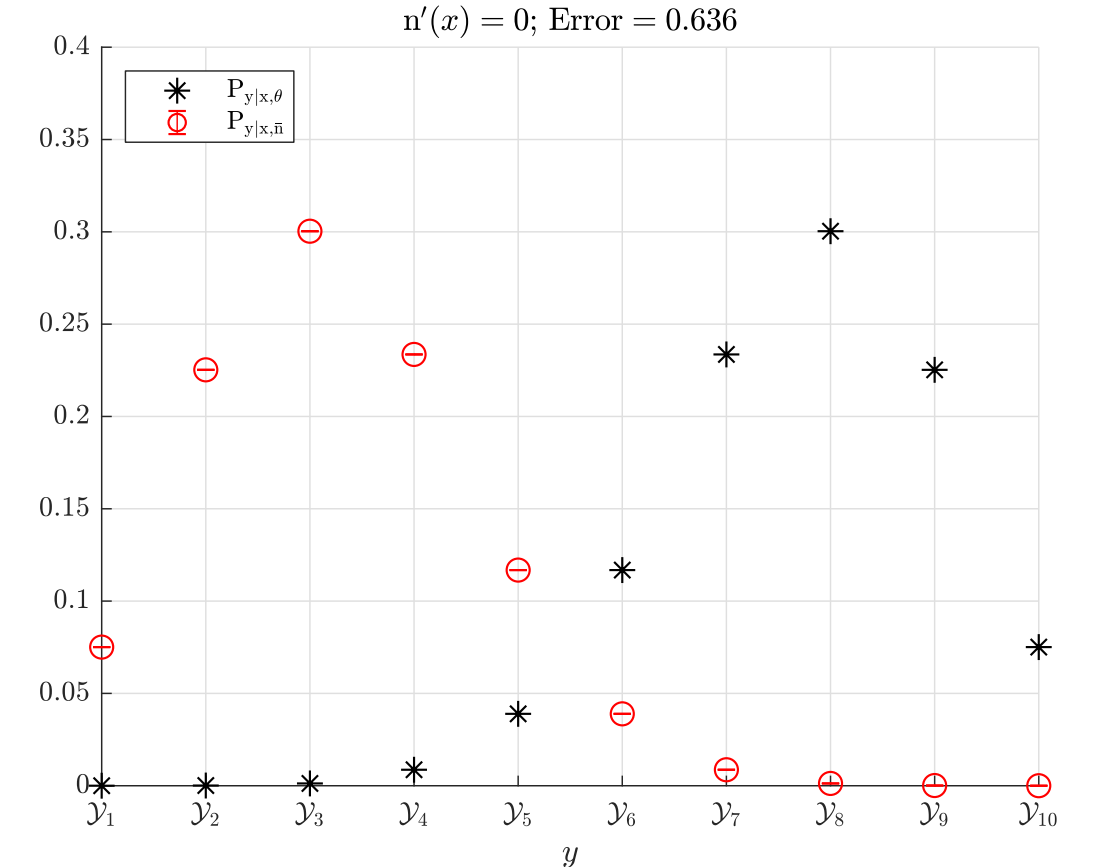
$$\Delta(x, \bar{n}, \theta) \equiv P_{y|x, \bar{n}} - P_{y|x, \theta} \in \mathbb{R}^{\mathcal{Y}}$$

$$\text{Bias}(x, n', \theta) = E_{\bar{n}|n', \theta} [\Delta(x, \bar{n}, \theta)] = \frac{\alpha'(x)}{\alpha'(x) + n'(x)} \left( \frac{\alpha(\cdot, x)}{\alpha'(x)} - \tilde{\theta}(x) \right)$$

$$\begin{aligned} \text{Cov}(y, y'; x, n', \theta) &= C_{\bar{n}|n', \theta} [P_{y|x, \bar{n}}(\cdot|x, \bar{n})](y, y') \\ &= \frac{n'(x)}{(\alpha'(x) + n'(x))^2} \left( \tilde{\theta}(y; x) \delta[y, y'] - \tilde{\theta}(y; x) \tilde{\theta}(y'; x) \right) \end{aligned}$$

↓ ↓ ↓

$$\begin{aligned} \mathcal{E}(y, y'; x, n', \theta) &= E_{\bar{n}|n', \theta} [\Delta(y; x, \bar{n}, \theta) \Delta(y'; x, \bar{n}, \theta)] \\ &= \text{Bias}(y; x, n', \theta) \text{Bias}(y'; x, n', \theta) + \text{Cov}(y, y'; x, n', \theta) \end{aligned}$$



- \* **Concentration parameter controls a Bias-Variance trade-off**

$$\text{Error} = \sqrt{\sum_{y \in \mathcal{Y}} \mathcal{E}(y, y; x, n', \theta)}$$

$\alpha'(x)$	Bias	Covariance
$\rightarrow \infty$	$\frac{\alpha(\cdot, x)}{\alpha'(x)} - \tilde{\theta}(x)$	0
$\rightarrow 0$	0	$\frac{\sum_{\bar{n}(\cdot, x)   n'(x), \tilde{\theta}(x)} (y, y')}{n'(x)^2}$

