# PREDICTIVE DISTRIBUTION ESTIMATION FOR BAYESIAN MACHINE LEARNING USING A DIRICHLET PROCESS PRIOR


U.S. NAVAL RESEARCH LABORATORY

THE GEORGE WASHINGTON UNIVERSITY, WASHINGTON, DC

Paul Rademacher[1] and Miloš Doroslovački[2]

[1]Radar Division, U.S. Naval Research Laboratory ; [2]Department of Electrical and Computer Engineering, The George Washington University


## 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 $\theta$:

$$\mathrm{P}_{y,x,D\,|\,\theta}(y,x,D|\theta) = \mathrm{P}_{y,x\,|\,\theta}(y,x|\theta) \prod_{n=1}^{N} \mathrm{P}_{D_n\,|\,\theta}(D_n|\theta) = \theta(y,x) \prod_{y'\in\mathcal{Y}} \prod_{x'\in\mathcal{X}} \theta(y',x')^{\bar{N}(y',x';D)}$$

- $\mathrm{P}_{D\,|\,\theta}$ depends on $D$ only through the transform $\bar{N}(y,x;D) \equiv \sum_{n=1}^{N} \delta\left[(y,x), D_n\right]$

  ⇒ 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:

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

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

⇓ ⇓ ***Select Prior*** $\mathrm{p}_\theta$ ⇓ ⇓

Bayes Risk:

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

⇒ Decisions formulated using *Bayes predictive distribution*, $\mathrm{P}_{y\,|\,x,\bar{n}} = \mathrm{E}_{\theta\,|\,x,\bar{n}}\left[\mathrm{P}_{y\,|\,x,\theta}\right] = \mu_{\tilde{\theta}(x)\,|\,x,\bar{n}}$

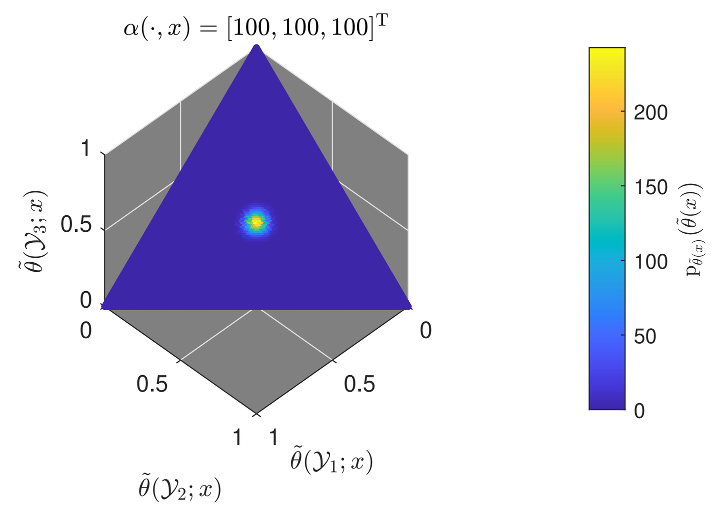
## Bayesian Prediction

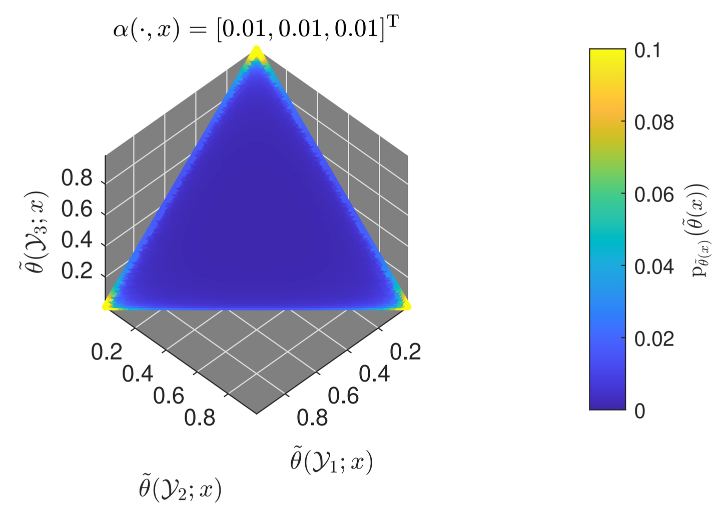
Dirichlet Priors:

$$\mathrm{p}_\theta(\theta) = \mathrm{Dir}(\theta;\alpha) = \beta(\alpha)^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha(y,x)-1}$$

$$\Downarrow$$

$$\mathrm{p}_{\tilde{\theta}}\left(\tilde{\theta}\right) = \prod_{x\in\mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x); \alpha(\cdot,x)\right)$$

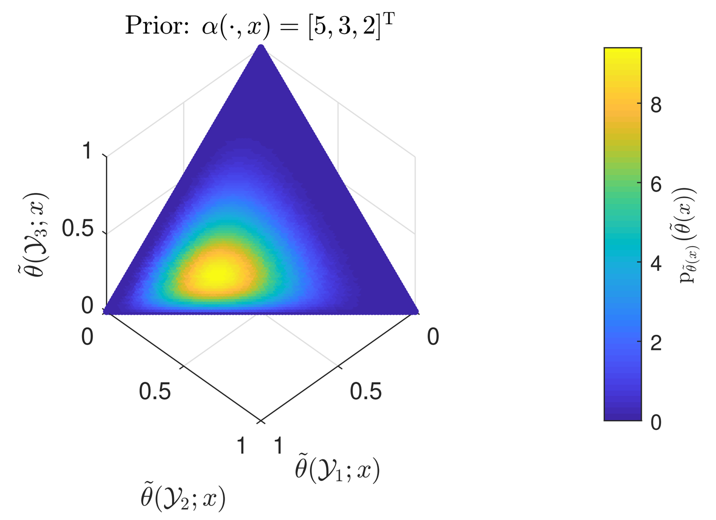


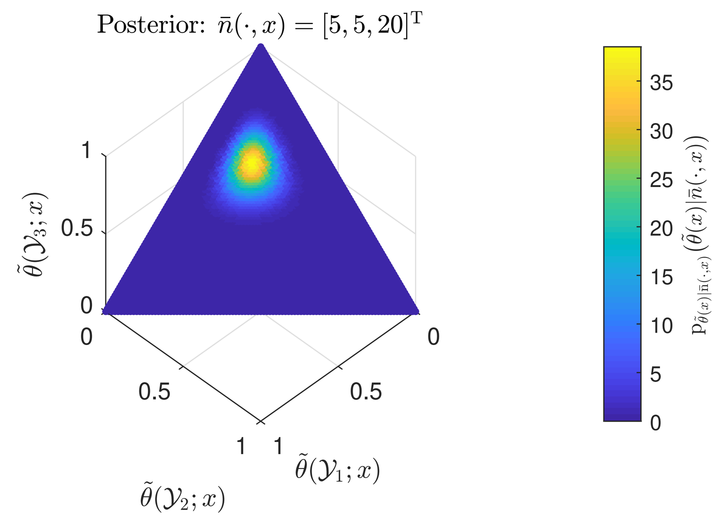




- Concentration parameters $\alpha'(x) \equiv \sum_{y\in\mathcal{Y}} \alpha(y,x)$ enable both *subjective* and *non-informative* priors
- Conjugate Prior for I.I.D. observations ⇒ Tractable Posterior

Dirichlet Posteriors:

$$\mathrm{p}_{\tilde{\theta}\,|\,x,\bar{n}}(\tilde{\theta}|x,\bar{n}) = \mathrm{p}_{\tilde{\theta}\,|\,\bar{n}}(\tilde{\theta}|\bar{n}) = \prod_{x'\in\mathcal{X}} \mathrm{p}_{\tilde{\theta}(x')\,|\,\bar{n}(\cdot,x')}\left(\tilde{\theta}(x')|\bar{n}(\cdot,x')\right) = \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x'); \alpha(\cdot,x') + \bar{n}(\cdot,x')\right)$$




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

**Bayes Predictive Distribution**:

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

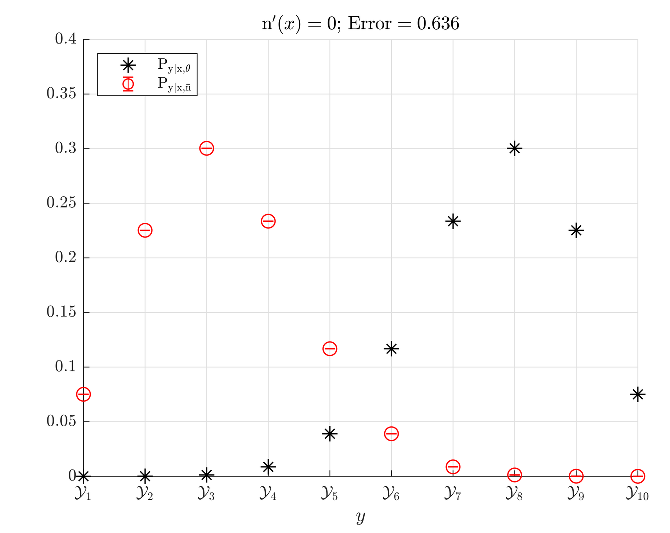
⇒ Convex combination of data-independent PMF and conditional empirical PMF

⇒ $\mathrm{E}_{\bar{n}\,|\,n',\theta}\left[\mathrm{P}_{y\,|\,x,\bar{n}}\right] = \left(\frac{\alpha'(x)}{\alpha'(x)+n'(x)}\right) \frac{\alpha(\cdot,x)}{\alpha'(x)} + \left(\frac{n'(x)}{\alpha'(x)+n'(x)}\right) \tilde{\theta}(x)$

## Density Estimation

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

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




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

$$\mathrm{Cov}(y,y';x,n',\theta) = \mathrm{C}_{\bar{n}\,|\,n',\theta}\left[\mathrm{P}_{y\,|\,x,\bar{n}}(\cdot\,|\,x,\bar{n})\right](y,y') = \frac{n'(x)}{\left(\alpha'(x)+n'(x)\right)^2}\left(\tilde{\theta}(y;x)\delta[y,y'] - \tilde{\theta}(y;x)\tilde{\theta}(y';x)\right)$$

$$\Downarrow \quad \Downarrow \quad \Downarrow$$

$$\mathcal{E}(y,y';x,n',\theta) = \mathrm{E}_{\bar{n}\,|\,n',\theta}\left[\Delta(y;x,\bar{n},\theta)\Delta(y';x,\bar{n},\theta)\right] = \mathbf{Bias}(y;x,n',\theta)\mathbf{Bias}(y';x,n',\theta) + \mathbf{Cov}(y,y';x,n',\theta)$$

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

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

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