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

U.S. NAVAL RESEARCH LABORATORY

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

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

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 $\mathrm{y} \in \mathcal{Y}$ given observed $\mathrm{x} \in \mathcal{X}$ and a training set $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$

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

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

- $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ 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 $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$ is a sufficient statistic for θ; decisions can depend on $\bar{\mathrm{n}}$ 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:

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

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

⇓ ⇓ ***Select Prior*** p_θ ⇓ ⇓

Bayes Risk:

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

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

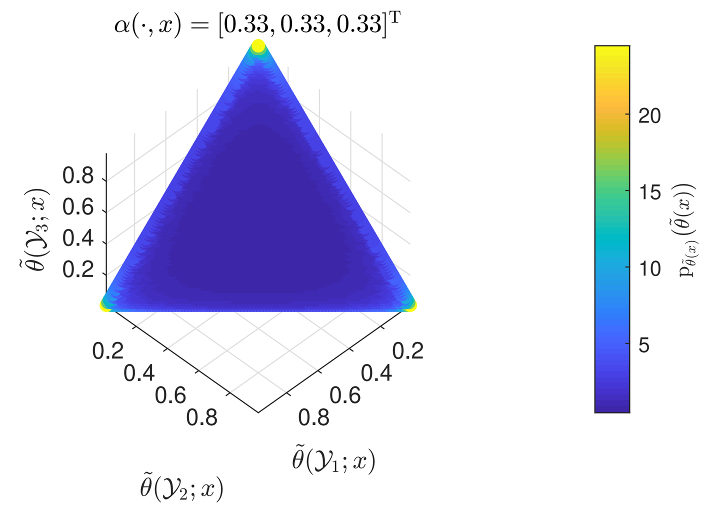
Bayesian Prediction

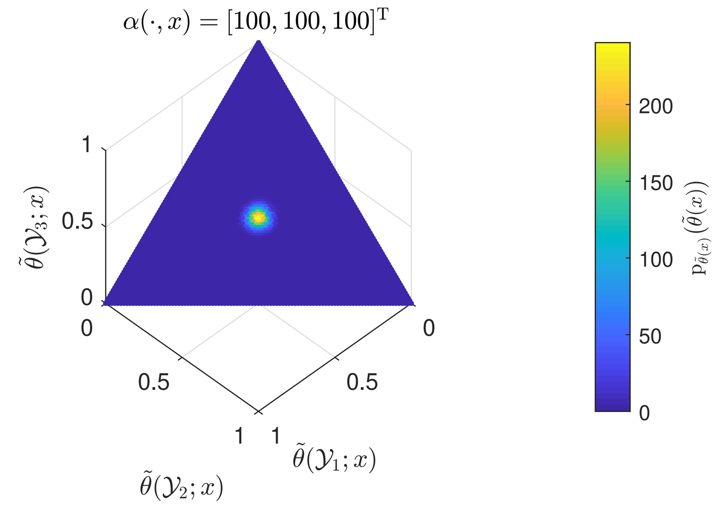
Dirichlet Priors:

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

$$\Downarrow$$

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

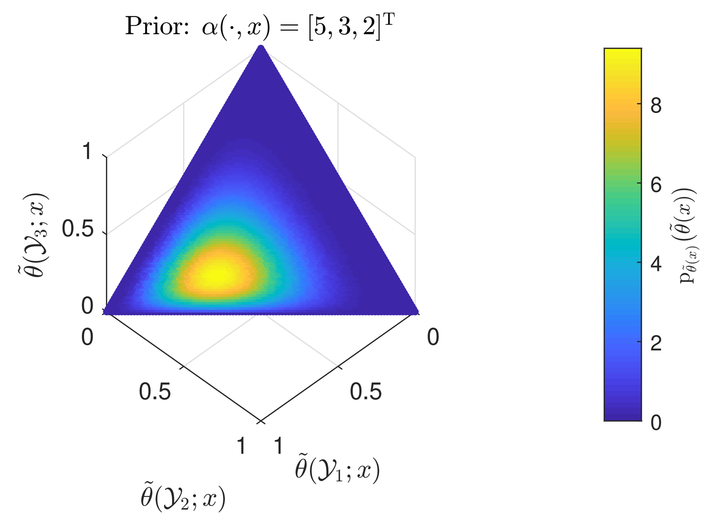


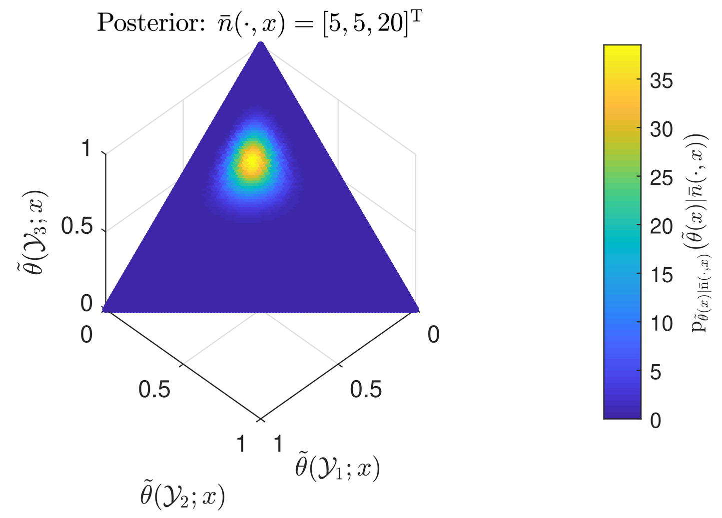


- Concentration parameters $\alpha'(x) \equiv \sum_{y\in\mathcal{Y}}\alpha(y,x)$ enable both *subjective* and *non-informative* priors
- Conjugate Prior for I.I.D. observations ⇒ Tractable Posterior

Dirichlet Posteriors:

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





- **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:

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

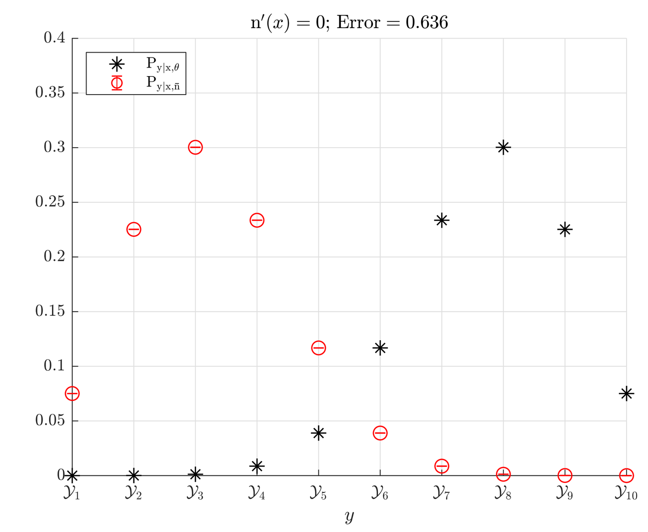
⇒ Convex combination of data-independent PMF and conditional empirical PMF

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

Density Estimation

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

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



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

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

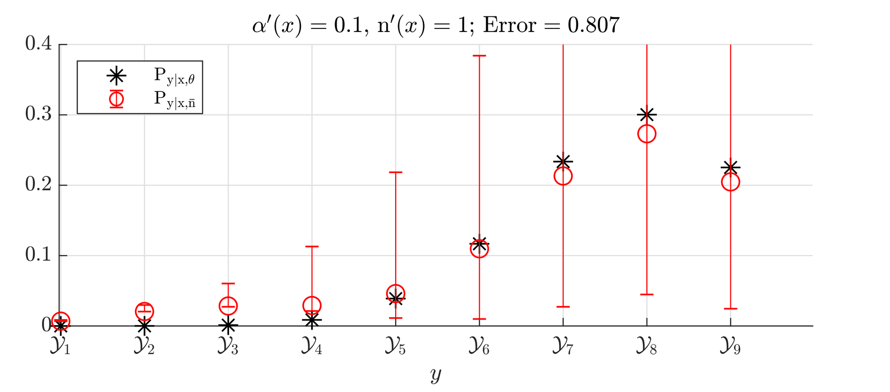
⇓ ⇓ ⇓

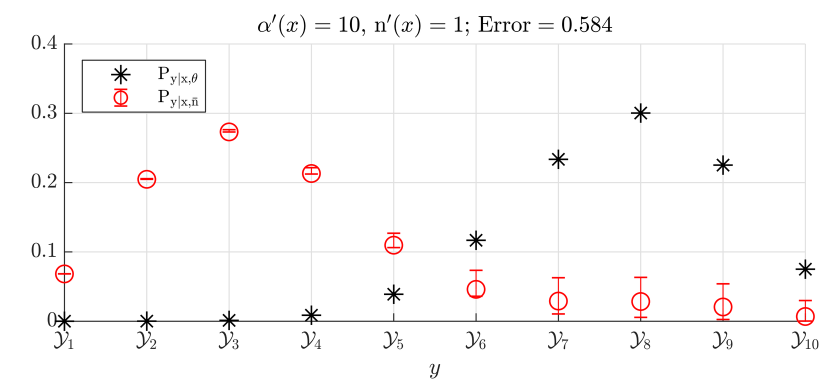
$$\mathcal{E}(y,y';\mathrm{x},\mathrm{n}',\theta) = \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(y;\mathrm{x},\bar{\mathrm{n}},\theta)\Delta(y';\mathrm{x},\bar{\mathrm{n}},\theta)\right] = \mathbf{Bias}(y;\mathrm{x},\mathrm{n}',\theta)\mathbf{Bias}(y';\mathrm{x},\mathrm{n}',\theta)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

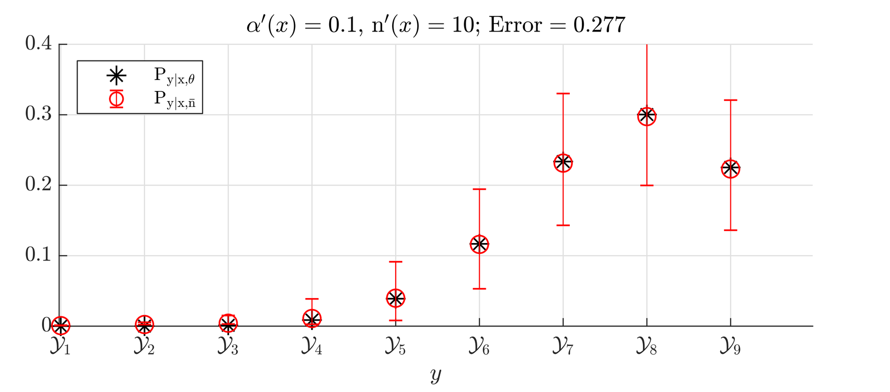
- **Concentration parameter controls a *Bias-Variance trade-off***

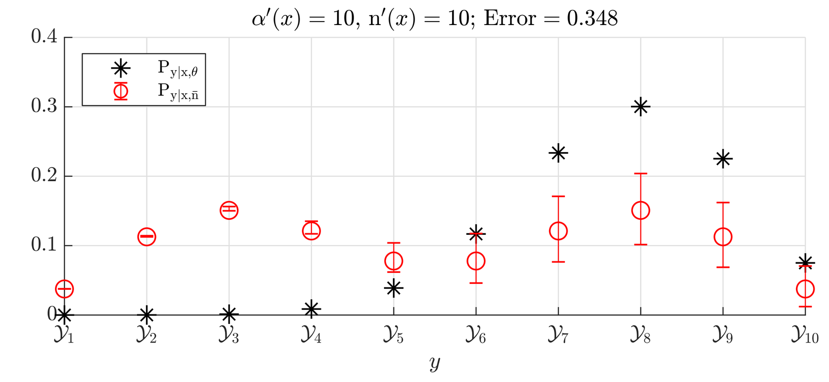
$$\mathrm{Error} = \sqrt{\sum_{y\in\mathcal{Y}}\mathcal{E}(y,y;\mathrm{x},\mathrm{n}',\theta)}$$

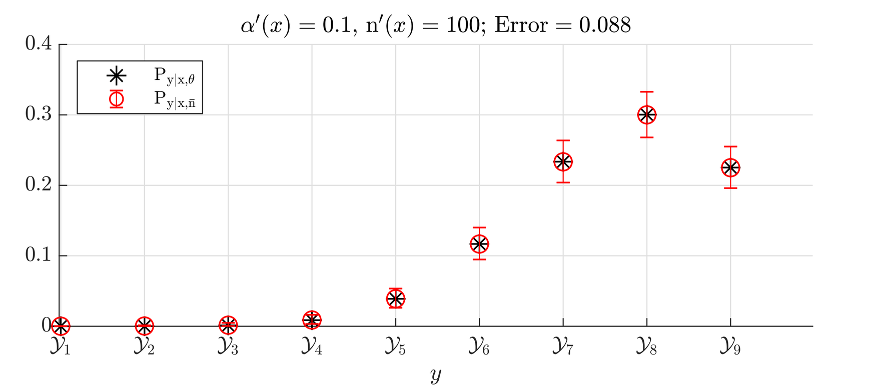
$\alpha'(\mathrm{x})$	Bias	Covariance
$\to\infty$	$\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} - \tilde{\theta}(\mathrm{x})$	0
$\to 0$	0	$\frac{\Sigma_{\bar{\mathrm{n}}(\cdot,\mathrm{x})\,|\,\mathrm{n}'(\mathrm{x}),\tilde{\theta}(\mathrm{x})}(y,y')}{\mathrm{n}'(\mathrm{x})^2}$

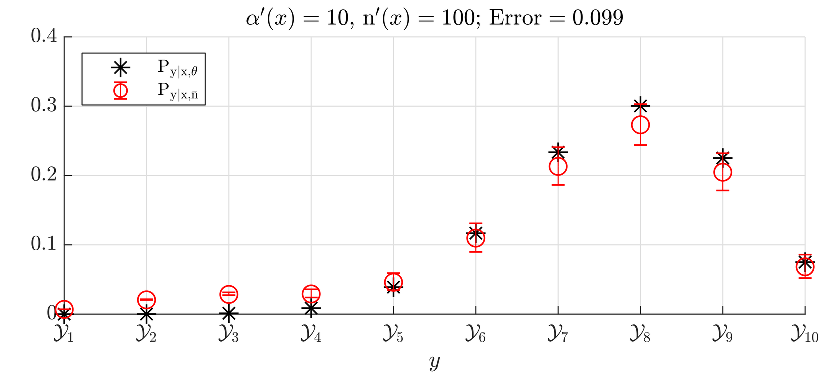












DISTRIBUTION A. Approved for public release: distribution unlimited