# Predictive Distribution Estimation for Bayesian Machine Learning using a Dirichlet Process Prior

Paul Rademacher
U.S. Naval Research Laboratory
Radar Division
Washington, DC 20375, USA

Miloš Doroslovački
The George Washington University
Department of Electrical and Computer Engineering
Washington, DC 20052, USA

***Abstract*—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. This paper assumes that the data-generating distribution is a realization of a Dirichlet process and assesses the mismatch between the true predictive distribution and the predictive distribution approximated using the training data. It is shown that 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. A bias/variance tradeoff will be demonstrated with illustrative examples.**

## I. Extended Summary

This article investigates how a Bayesian perspective influences the predictive distributions used to make decisions in machine learning applications. The efficacy of Bayesian learning methods depends on how well the prior knowledge imparted by the designer matches the true data-generating probability mass function (PMF) $\mathrm{\theta}$. The chosen prior distribution $\mathrm{p}_\mathrm{\theta}$ over the set of data-generating PMFs reflects the users confidence that different PMFs $\mathrm{\theta}$ are responsible for generating the novel pair $(\mathrm{y},\mathrm{x}) \in \mathcal{Y} \times \mathcal{X}$ and the observed training data $\mathrm{D} = (\mathrm{Y},\mathrm{X}) \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$. The elements $(\mathrm{y},\mathrm{x})$ are characterized by $\mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{\theta}}(y,x|\theta) = \theta(y,x)$ and the training data by

$$\begin{aligned} \mathrm{P}_{\mathrm{D}\,|\,\mathrm{\theta}}\left(D|\theta\right) &= \prod_{n=1}^{N} \theta(Y_n, X_n) \qquad (1) \\ &= \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{N}(y,x;D)} \ . \end{aligned}$$

The dependency on $\mathrm{D}$ is expressed though the function

$$\bar{N}(y,x;D) = \sum_{n=1}^{N} \delta\left[y, Y_n\right] \delta\left[x, X_n\right] \ , \qquad (2)$$

where $\delta[\cdot,\cdot]$ is the Kronecker delta function. The random elements $\mathrm{x}$ and $\mathrm{D}$ are observable and provide predictive information about the unobserved value of $\mathrm{y}$.

If a highly localized prior distribution $\mathrm{p}_\mathrm{\theta}$ is chosen that strongly weights the actual data PMF, low risk learning functions are possible even with limited training data $\mathrm{D}$; however, if the true PMFs weighting is low, a good solution will never be achieved. Conversely, a high-variance prior probability density function (PDF) leads to decision functions that will always be able to adapt with enough training data; if data is limited, however, the function may not deliver the required performance.

This article assumes that the data-generating PMF is a realization of a Dirichlet process. The Dirichlet distribution for the PMF model $\mathrm{\theta} \in \Theta$ is [1]

$$\mathrm{p}_\mathrm{\theta}(\theta) = \beta(\alpha)^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha(y,x)-1} \ , \qquad (3)$$

where the user-selected PDF parameters $\alpha : \mathcal{Y} \times \mathcal{X} \mapsto \mathbb{R}^+$ are introduced and $\beta$ is the generalized beta function.

The class of Dirichlet PDFs have the desirable properties of full support over the set of possible PMFs and a tractable posterior distribution for independently and identically distributed data [2].

Additionally, control of the Dirichlet parameters can enable both maximally and minimally localized prior distributions. The parameters $\alpha$ control around which models $\mathrm{\theta}$ the PDF concentrates and how strongly. For convenience, introduce the concentration parameter $\alpha_0 \equiv \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \alpha(y,x)$.

Of specific interest is how $\mathrm{p}(\mathrm{\theta})$ changes as the concentration parameter approaches its limiting values. For $\alpha_0 \to \infty$, the PDF concentrates at its mean, resulting in

$$\mathrm{p}_\mathrm{\theta}(\theta) = \delta\left(\theta - \frac{\alpha}{\alpha_0}\right) \ , \qquad (4)$$

where $\delta(\cdot)$ is the Dirac delta function over the set $\Theta$. Conversely, for $\alpha_0 \to 0$, the PDF trends toward

$$\mathrm{p}_\mathrm{\theta}(\theta) = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0} \delta\big(\theta - \delta[\cdot,y]\delta[\cdot,x]\big) \ , \qquad (5)$$

which distributes its weight among the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models with an $\ell_0$ norm satisfying $\|\theta\|_0 = 1$. These trends are demonstrated in Figure 1.

Define the decision function $f : \mathcal{D} \mapsto \mathcal{H}^\mathcal{X}$, where $\mathcal{H}$ is the decision space. The metric guiding the design is the conditional expected loss, or conditional "risk",

$$\mathcal{R}_\Theta(f;\mathrm{\theta}) = \mathrm{E}_{\mathrm{D}\,|\,\mathrm{\theta}}\left[\mathrm{E}_{\mathrm{y,x}\,|\,\mathrm{\theta}}\left[\mathcal{L}\left(f(\mathrm{x},\mathrm{D}),\mathrm{y}\right)\right]\right] \ , \qquad (6)$$

where $\mathcal{L} : \mathcal{H} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ is the user-specified loss function.

If the model $\mathrm{\theta}$ were observable, a "clairvoyant" decision function $f_\Theta : \Theta \mapsto \mathcal{H}^\mathcal{X}$ could be designed that is dependent

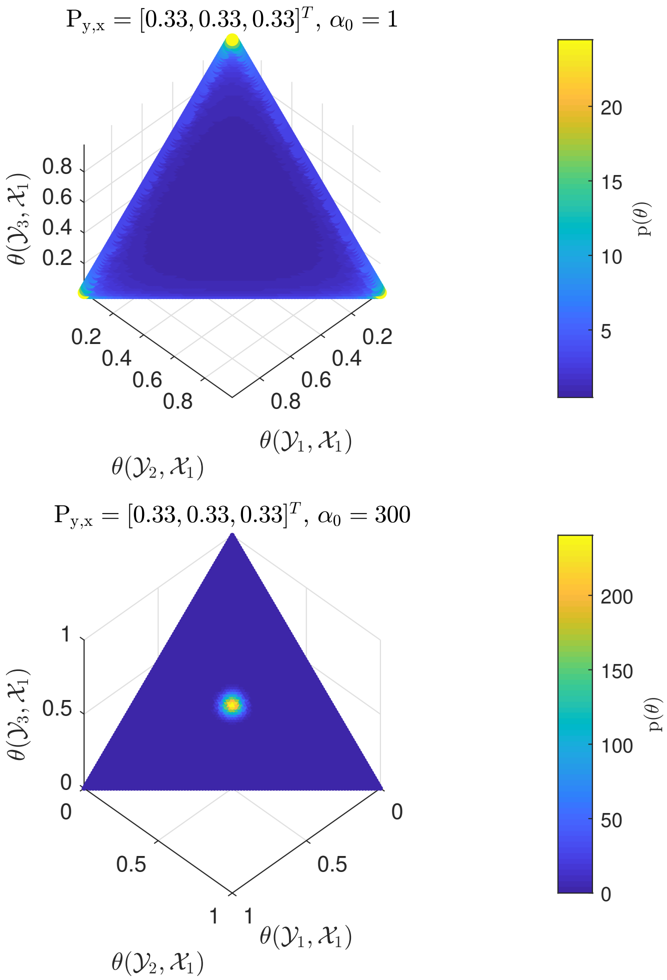

Fig. 1. Model prior PDF $\mathrm{p}(\theta)$ for different concentrations $\alpha_0$

only on the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$. However, as the model $\theta$ is unknown, the prior distribution $\mathrm{p}_\theta$ is assumed and the Bayes risk can be formulated as

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta\left[\mathcal{R}_\Theta\left(f(\mathrm{x},\mathrm{D});\theta\right)\right] \\ &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\mathcal{L}\left(f(\mathrm{x},\mathrm{D}),\mathrm{y}\right)\right]\right] .\end{aligned} \tag{7}$$

Now, the decision function is dependent on the predictive distribution

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|x,D) &= \frac{\alpha(y,x)+\bar{N}(y,x;D)}{\alpha'(x)+N'(x;D)} \\ &= \left(\frac{\alpha'(x)}{\alpha'(x)+N'(x;D)}\right)\frac{\alpha(y,x)}{\alpha'(x)} \\ &\quad + \left(\frac{N'(x;D)}{\alpha'(x)+N'(x;D)}\right)\frac{\bar{N}(y,x;D)}{N'(x;D)} ,\end{aligned} \tag{8}$$

where $\alpha'(x) \equiv \sum_{y\in\mathcal{Y}} \alpha(y,x)$ and $N'(x;D) = \sum_{n=1}^{N} \delta\left[x, X_n\right]$. This PMF can be interpreted as an estimate of the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$.

It is informative to compare the Bayesian predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ to the unobserved predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ and investigate the effects of the Dirichlet prior localization. As $\bar{N}(\mathrm{D})$ is a sufficient statistic for the PMF $\theta$, any distributions dependent on $\mathrm{D}$ will be replaced by their corresponding distributions of $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$, simplifying the analysis. Note that $\bar{\mathrm{n}}\,|\theta \sim \mathrm{Mult}(N,\theta)$ [4].

For a given observation $\mathrm{x} = x$ and corresponding number of training samples $\mathrm{n}'(x) \equiv \sum_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)$, the expected value of the predictive PMF estimate condtioned on the true model $\theta$ is

$$\begin{aligned}&\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|x,\bar{\mathrm{n}})\right] \\ &\quad = \left(\frac{\alpha'(x)}{\alpha'(x)+\mathrm{n}'(x)}\right)\frac{\alpha(y,x)}{\alpha'(x)} + \left(\frac{\mathrm{n}'(x)}{\alpha'(x)+\mathrm{n}'(x)}\right)\frac{\theta(y,x)}{\theta'(x)} .\end{aligned} \tag{9}$$

where the aggregation property for multinomial distributions has been used [3].

To aid further characterization of the Bayesian predictive distribution, define the difference between the estimated and true predictive PMFs as the random process $\Delta(\cdot;x,\bar{\mathrm{n}}) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot|x,\bar{\mathrm{n}}) - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(\cdot|x,\theta)$. Given $\mathrm{x} = x$ and a corresponding number of training samples $\mathrm{n}'(x)$, the bias of the conditional PMF estimate is

$$\begin{aligned}\mathrm{Bias}(y;x,\mathrm{n}') &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(y;x,\bar{\mathrm{n}})\right] \\ &= \frac{\alpha'(x)}{\alpha'(x)+\mathrm{n}'(x)}\left(\frac{\alpha(y,x)}{\alpha'(x)} - \frac{\theta(y,x)}{\theta'(x)}\right)\end{aligned} \tag{10}$$

and its covariance function is

$$\begin{aligned}&\mathrm{Cov}(y,y';x,\mathrm{n}') = \mathrm{C}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot|x,\bar{\mathrm{n}})\right](y,y') \\ &\quad = \frac{\mathrm{n}'(x)}{\left(\alpha'(x)+\mathrm{n}'(x)\right)^2}\left(\frac{\theta(y,x)}{\theta'(x)}\delta[y,y'] - \frac{\theta(y,x)}{\theta'(x)}\frac{\theta(y',x)}{\theta'(x)}\right) ,\end{aligned} \tag{11}$$

where $C$ is the covariance operator for functions of random elements.

Combining the estimator bias and variance, the conditional second moments of $\Delta(\cdot;x,\bar{\mathrm{n}})$ are

$$\begin{aligned}\mathcal{E}(y,y';x,\mathrm{n}') &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(y;x,\bar{\mathrm{n}})\Delta(y';x,\bar{\mathrm{n}})\right] \\ &= \mathrm{Bias}(y;x,\mathrm{n}')\mathrm{Bias}(y';x,\mathrm{n}') + \mathrm{Cov}(y,y';x,\mathrm{n}') .\end{aligned} \tag{12}$$

To exemplify how the model estimate $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ approximates $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$, consider a scenario with $|\mathcal{Y}| = 10$. The data-independent PMF $\alpha/\alpha_0$ and true model $\theta$ are shown in Figure 2 - note the significant mismatch.

Figures 3 and 4 show how the expected value and variance of the PMF estimate (represented by the blue markers and error bars) change for different values of $\mathrm{n}'(x)$ and $\alpha'(x)$. Note that the upper and lower error bars represent the expected squared deviation above and below the conditional mean $\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|x,\bar{\mathrm{n}})\right]$, respectively. Each individual plot heading provides the error $\sqrt{\sum_{y\in\mathcal{Y}} \mathcal{E}(y,y;x,\mathrm{n}')}$ to assess the quality of the PMF estimate.

Observe that for $\mathrm{n}'(x) = 1$, the high variance of the $\alpha'(x) = 0.1$ estimate (favoring the empirical PMF) renders it worse than the $\alpha_0 = 10$ estimate; in fact, the variance is so high that the error exceeds that of the data-independent estimate $\alpha(\cdot,x)/\alpha'(x)$ (Figure 2). Conversely, for $\mathrm{n}'(x) = 10$, the

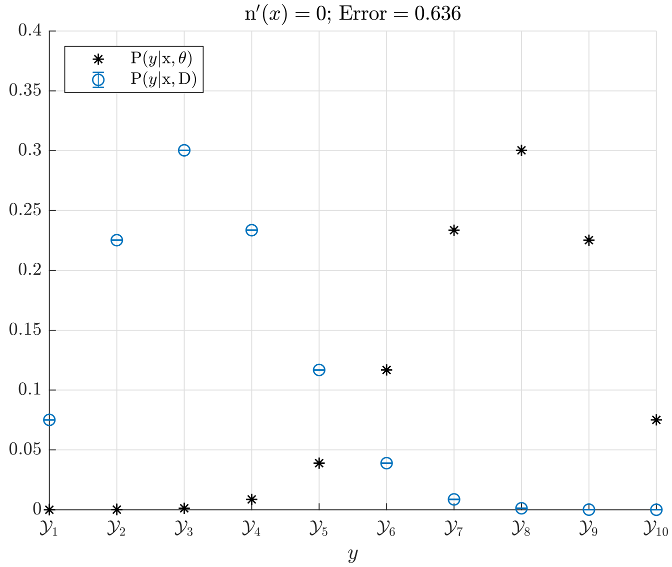

Fig. 2. Model θ estimate, no training data

confidence of the $\alpha'(x) = 10$ estimate leads to high bias and the $\alpha'(x) = 0.1$ estimate is superior. For $\mathrm{n}'(x) = 100$, both the $\alpha'(x) = 0.1$ and $\alpha'(x) = 10$ estimates begin converging to the true predictive distribution - this is guaranteed due to the full support of the Dirichlet prior.

The full paper will provide full detail in determining the predictive distributions and expand the discussion on how they are affected by the Dirichlet prior parameters. Additional results will investigate the bias/variance trade-off with other informative examples. The conditional second moments of the difference between the two predicitve PMFs have important applications for Bayesian regression, specifically for determining the expected squared-error loss. This will be a primary focus of future work.

## References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[2] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[3] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Probability and Statistics. John Wiley & Sons, 1997.

[4] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier, 2015.

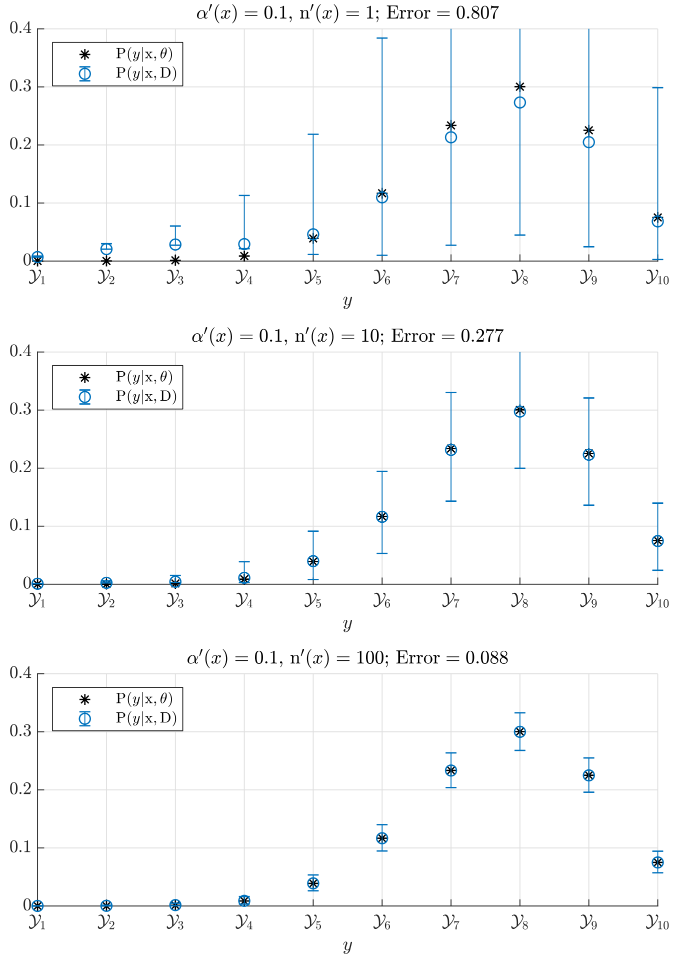

Fig. 3. Model θ estimates, $\alpha_0 = 0.1$

$\alpha'(x) = 10$, $\mathrm{n}'(x) = 1$; Error = 0.584

$\alpha'(x) = 10$, $\mathrm{n}'(x) = 10$; Error = 0.348

$\alpha'(x) = 10$, $\mathrm{n}'(x) = 100$; Error = 0.099

Fig. 4. Model θ estimates, $\alpha_0 = 10$