BAYESIAN LEARNING FOR REGRESSION USING DIRICHLET PRIOR DISTRIBUTIONS OF VARYING LOCALIZATION

Paul Rademacher

U.S. Naval Research Laboratory
Radar Division
Washington, DC 20375, USA
paul.rademacher@nrl.navy.mil

Miloš Doroslovački

The George Washington University
Department of Electrical and Computer Engineering
Washington, DC 20052, USA
doroslov@gwu.edu

ABSTRACT

When taking a Bayesian approach to machine learning applications, the performance of the learned function strongly depends on how well the prior distribution selected by the designer matches the true data-generating model. Dirichlet priors have a number of desirable properties - they result in closed-form posterior distributions given independent training data, have full support over the space of data probability distributions, and can be maximally informative or non-informative depending on their localization parameter. This paper assumes a Dirichlet prior and details the predictive distributions that characterize unobservable random quantities given observed data. The results are then applied to the most common loss function for regression, the squared-error loss. The optimal Bayes estimator and the resultant risk trends are presented for different prior localizations, demonstrating a bias/variance trade-off.

Index Terms— Bayesian learning, machine learning, regression, estimation, Dirichlet distribution, bias, variance, predictive distribution

1. INTRODUCTION

The success or failure of Bayesian learning methods hinge on how well the prior knowledge imparted by the designer matches reality. The chosen prior distribution over the set of data-generating probability distributions reflects the users confidence that different models are responsible for generating the observed/unobserved random elements. If a highly localized, "informative" prior is chosen that strongly weights the actual data model, low risk learning functions are possible even with limited training data; however, if the localized prior is poorly selected, a good solution may not be achieved. Conversely, a non-informative prior provides fast adaptation during training and can be suitable for a variety of problems; if data is limited, however, the learning function may not deliver satisfactory performance.

This work assumes that the observed and unobserved data elements are jointly characterized by a Dirichlet prior. The class of Dirichlet probability density functions (PDF) has the desirable properties of full support over the set of possible data-generating distributions and a closed-form posterior distribution for independently and identically distributed (i.i.d.) data [1]. Furthermore, control of the Dirichlet parameters enables both minimally and maximally informative priors, providing a wide range of learning solutions.

After motivating the Bayesian perspective and using the Dirichlet prior to generate the predictive distribution, the results will be applied to the squared error loss function. The Bayes optimal estimator and the achieved risk will be presented. Results will demonstrate a bias/variance risk trade-off dependent on prior parameterization, exemplifying the balance between prior knowledge and data utilization.

2. OBJECTIVE

Consider an observable scalar random variable $\mathrm{x} \in \mathcal{X} \subset \mathbb{R}$ and unobservable scalar random variable $\mathrm{y} \in \mathcal{Y} \subset \mathbb{R}$ which are jointly distributed according to an unknown probability mass function (PMF), $\theta \in \mathcal{P}(\mathcal{Y} \times \mathcal{X}) \equiv \left\{ \theta \in {\mathbb{R}_{\geq 0}}^{\mathcal{Y} \times \mathcal{X}} : \sum_{y,x} \theta(y,x) = 1 \right\}$, such that $\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\,\theta}(y,x|\theta) = \theta(y,x)$. The space $\mathcal{P}(\cdot)$ is the set of probability distributions over the argument set.

Also observed is a random sequence of N samples drawn from the model θ, denoted $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$. The N data pairs are identically distributed as $\mathrm{P}_{\mathrm{D}_n\,|\,\theta}(y,x|\theta) = \theta(y,x)$ for $n = 1, \ldots, N$ and are conditionally independent from one another and from the novel pair (y, x).

The regression objective is to design a learning function $f : \mathcal{D} \mapsto \mathbb{R}^{\mathcal{X}}$ which uses the training data to select an estimator from the set of functions $\mathcal{X} \mapsto \mathbb{R}$. The metric guiding the design is the squared-error loss function $\mathcal{L} : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$, defined as $\mathcal{L}(y', y) = (y' - y)^2$. The expected loss, or "risk",

is defined as

$$\mathcal{R}_\Theta(f;\theta) = \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y,x}\,|\,\theta}\left[\left(f(\mathrm{x};\mathrm{D}) - \mathrm{y}\right)^2\right]\right] \quad (1)$$
$$= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x,D}\,|\,\theta}\left[\left(f(\mathrm{x};\mathrm{D}) - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right],$$

where μ and Σ are the mean and covariance functions, respectively.

If the model θ were known, the "clairvoyant" [2] estimate would be $f_\Theta(\mathrm{x};\theta) = \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$, and the minimum risk $\mathcal{R}^*_\Theta(\theta) = \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$ would be achieved. Suboptimal learners induce excess risk $\mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta) \equiv \mathcal{R}_\Theta(f;\theta) - \mathcal{R}^*_\Theta(\theta)$ due to model uncertainty.

As the model θ is not observed, $\mathcal{R}_\Theta$ is not a feasible objective function for optimization. If the designer selects a PDF p_θ, the Bayes risk is formulated as

$$\mathcal{R}(f;\mathrm{p}_\theta) = \mathrm{E}_\theta\left[\mathcal{R}_\Theta(f;\theta)\right]$$
$$\equiv \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\left(f(\mathrm{x};\mathrm{D}) - \mathrm{y}\right)^2\right]\right] \quad (2)$$

and y, x, and D are treated as jointly distributed random variables. The optimal Bayesian estimate is expressed as

$$f^*(\mathrm{x};\mathrm{D}) = \underset{y'\in\mathbb{R}}{\arg\min}\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[(y'-\mathrm{y})^2\right] = \mu_{\mathrm{y}\,|\,\mathrm{x,D}}\,, \quad (3)$$

where the dependency of f^* on the prior p_θ (through $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$) is suppressed for brevity.

3. PROBABILITY DISTRIBUTIONS

3.1. Dirichlet Model Process

The Dirichlet prior PDF for the model random process $\theta \in \Theta$ is [3]

$$\mathrm{p}_\theta(\theta) = \beta(\alpha_0\alpha)^{-1}\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\theta(y,x)^{\alpha_0\alpha(y,x)-1}$$
$$\equiv \mathrm{Dir}(\theta;\alpha_0,\alpha)\,, \quad (4)$$

where β is the multivariate beta function, and the user-selected localization parameter $\alpha_0 \in \mathbb{R}^+$ and mean $\alpha \in \mathcal{P}(\mathcal{Y}\times\mathcal{X}) \cap \mathbb{R}^{+\mathcal{Y}\times\mathcal{X}}$ are introduced. Note that as $\alpha_0 \to \infty$, the PDF tends to a maximally informative prior $\mathrm{p}_\theta \to \delta(\cdot - \alpha)$, where δ is the Dirac delta function over $\mathcal{P}(\mathcal{Y}\times\mathcal{X})$.

3.1.1. Marginal and Conditional Models

An alternative representation is implemented via the bijection $\theta \Leftrightarrow (\theta_\mathrm{m}, \theta_\mathrm{c})$, where the marginal distribution $\theta_\mathrm{m} \equiv \sum_{y\in\mathcal{Y}}\theta(y,\cdot) \in \mathcal{P}(\mathcal{X})$ and the conditional distributions $\theta_\mathrm{c}(x) \equiv \theta(\cdot,x)/\,\theta_\mathrm{m}(x) \in \mathcal{P}(\mathcal{Y})$ for all $x \in \mathcal{X}$.

By the aggregation property [1], $\theta_\mathrm{m} \sim \mathrm{Dir}(\alpha_0,\alpha_\mathrm{m})$ is a Dirichlet random process parameterized by localization α_0 and mean $\alpha_\mathrm{m} \equiv \sum_{y\in\mathcal{Y}}\alpha(y,\cdot)$. Also, it can be shown that the predictive models $\theta_\mathrm{c}(x) \sim \mathrm{Dir}(\alpha_0\,\alpha_\mathrm{m}(x),\alpha_\mathrm{c})$ are independent Dirichlet processes, where $\alpha_\mathrm{c}(x) \equiv \alpha(\cdot,x)/\,\alpha_\mathrm{m}(x)$; they are independent of θ_m as well.

3.2. Training Data and the Empirical Statistic

Using the i.i.d. assumption, the distribution of the training data D conditioned on the model can be formulated as

$$\mathrm{P}_{\mathrm{D}\,|\,\theta}\left(D|\theta\right) = \left(\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\theta(y,x)^{\Psi(y,x;D)}\right)^N\,, \quad (5)$$

where the dependency on the training data D is expressed though a transform function $\Psi : \mathcal{D} \mapsto \Psi \subset \mathcal{P}(\mathcal{Y}\times\mathcal{X})$ defined as $\Psi(y,x;D) = N^{-1}\sum_{n=1}^N \delta\big[(y,x),D_n\big]$.

This transform produces the empirical distribution of the training data D. Since the distribution $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ depends on the training data only via Ψ, $\Psi(\mathrm{D})$ is a sufficient statistic [4] for the model θ. As such, the analysis is performed using a new random process $\psi \equiv \Psi(\mathrm{D}) \in \Psi$, noting that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(x,D) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi}\left(x,\Psi(D)\right)$.

Note that Ψ is a sampling of the distribution space $\mathcal{P}(\mathcal{Y}\times\mathcal{X})$ with cardinality $|\Psi| = \mathcal{M}\left((N,|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1)\right)$, where $\mathcal{M}$ is the multinomial coefficient; this can be shown using the stars-and-bars method [5]. Additionally, $|\Psi| \leq |\,\mathcal{D}\,| = \left(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\right)^N$ and thus the sufficient statistic efficiently represents the information in the training data.

The conditional distribution of the transformed process is

$$\mathrm{P}_{\psi|\theta}(\psi|\theta) = \mathcal{M}(N\psi)\left(\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\theta(y,x)^{\psi(y,x)}\right)^N$$
$$\equiv \mathrm{Emp}\left(\psi;N,\theta\right)\,,$$

an empirical PMF given θ. As $\psi|\theta$ is equivalent to a normalized multinomial random process [6], the mean and covariance functions are $\mu_{\psi|\theta} = \theta$ and

$$\Sigma_{\psi|\theta}(y,x,y',x'|\theta) \quad (6)$$
$$= \frac{1}{N}\theta(y,x)\big(\delta[y,y']\delta[x,x'] - \theta(y',x')\big)\,.$$

3.2.1. Marginal and Conditional Data Statistics

The empirical process can also be decomposed into marginal and conditional empirical processes via a bijection $\psi \Leftrightarrow (\psi_\mathrm{m},\psi_\mathrm{c})$, where $\psi_\mathrm{m} \equiv \sum_{y\in\mathcal{Y}}\psi(y,\cdot) \in \mathcal{P}(\mathcal{X})$ and $\psi_\mathrm{c}(x) \equiv \psi(\cdot,x)/\,\psi_\mathrm{m}(x) \in \mathcal{P}(\mathcal{Y})$. Like the multinomial random process [7], the empirical random process has an aggregation property, such that $\psi_\mathrm{m}\,|\,\theta_\mathrm{m} \sim \mathrm{Emp}(N,\theta_\mathrm{m})$. Also, the $|\,\mathcal{X}\,|$ functions $\psi_\mathrm{c}(x)|\,\psi_\mathrm{m}(x),\theta_\mathrm{c}(x) \sim \mathrm{Emp}\left(N\,\psi_\mathrm{m}(x),\theta_\mathrm{c}(x)\right)$ can be shown to be conditionally independent empirical processes.

3.3. Bayesian Predictive Distribution

As shown in Equation (3), the decision selected by the optimally designed function depends on the predictive distribution of y conditioned on all observable random variables. The

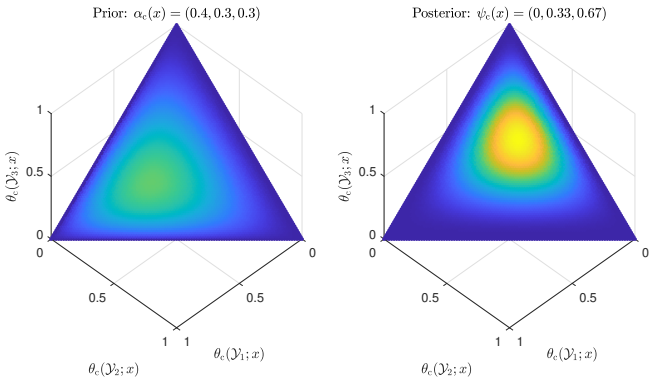

Fig. 1. Model prior $\mathrm{p}_{\theta_c(x)}$ and posterior $\mathrm{p}_{\theta_c(x)|\,\psi_m(x),\psi_c(x)}$

training data D is represented using the empirical sufficient statistic and $\mathrm{P}_{y\,|\,x,\psi}$ is used. Observe that $\mathrm{P}_{y\,|\,x,\theta} \equiv \theta_c(x)$ and thus that $\mathrm{P}_{y\,|\,x,\psi} = \mathrm{E}_{\theta|\,x,\psi}\left[\mathrm{P}_{y\,|\,x,\theta}\right] \equiv \mu_{\theta_c(x)|\,x,\psi}$; the Bayesian predictive PMF is the posterior mean [8] of the true predictive PMF.

Due to the independence of the conditional models $\theta_c(x)$ from one another and from the marginal model, $\mathrm{p}_{\theta_c(x)|\psi,x} \equiv \mathrm{p}_{\theta_c(x)|\,\psi_m(x),\psi_c(x)} \sim \mathrm{P}_{\psi_c(x)|\,\psi_m(x),\theta_c(x)}\,\mathrm{p}_{\theta_c(x)}$. Since the conditional PMF of the empirical statistic $\psi_c(x)$ has exponential form, the Dirichlet PDF $\mathrm{p}_{\theta_c(x)}$ is a conjugate prior [9]; thus, the true predictive models conditioned on the empirical statistic are independent Dirichlet processes,

$$\theta_c(x)|\,\psi_m(x),\psi_c(x) \sim \mathrm{Dir}\left(\alpha_0\,\alpha_m(x) + N\,\psi_m(x), \mu_{\theta_c(x)|\,\psi_m(x),\psi_c(x)}\right), \tag{7}$$

with mean functions

$$\begin{aligned}\mu_{\theta_c(x)|\,\psi_m(x),\psi_c(x)} &= \gamma(x;\psi_m)\,\alpha_c(x) \\ &\quad + \big(1-\gamma(x;\psi_m)\big)\,\psi_c(x) \\ &\equiv \mathrm{P}_{y\,|\,x,\psi}\,,\end{aligned} \tag{8}$$

where $\gamma(x;\psi_m) = \left(1 + \frac{N\,\psi_m(x)}{\alpha_0\,\alpha_m(x)}\right)^{-1} \in (0,1]$. Note that as the prior localization increases relative to the training data volume, $\gamma(x;\psi_m) \to 1$ and the posterior mean tends to $\alpha_c(x)$, reflecting confidence in the prior. Conversely, for large training sets, $\gamma(x;\psi_m) \to 0$ and the mean tends to $\psi_c(x)$. Also, as $N\,\psi_m(x) \to \infty$, the PDF $\mathrm{p}_{\theta_c(x)|\,\psi_m(x),\psi_c(x)} \to \delta\big(\cdot - \psi_c(x)\big)$ and the model is positively identified. This is a consequence of the full support of the Dirichlet distribution and does not hold in general. The adaptation of the posterior is demonstrated in Figure 1 for $\alpha_0\,\alpha_m(x) = 6$ and $N\,\psi_m(x) = 3$.

4. REGRESSION AND THE SQUARED-ERROR LOSS

4.1. Optimal Estimate: Bayesian Posterior Mean

To find the optimal Bayesian estimator, the Bayesian predictive distribution (8) is substituted into (3) and the posterior mean in terms of the empirical statistic is

$$\begin{aligned}f^*(x;\psi) \equiv \gamma(x;\psi_m)\sum_{y\in\mathcal{Y}}\alpha_c(y;x)\,y \\ +\big(1-\gamma(x;\psi_m)\big)\sum_{y\in\mathcal{Y}}\psi_c(y;x)\,y\,.\end{aligned} \tag{9}$$

The optimal estimate is a convex combination of two mean values - the first moment of the data-independent PMF $\mathrm{P}_{y\,|\,x} = \mu_{\theta_c(x)} = \alpha_c(x)$, and the conditional empirical mean.

The weighting factors are inherited from $\mathrm{P}_{y\,|\,x,\psi}$; thus, stronger prior information (larger $\alpha_0\,\alpha_m(x)$) provides more emphasis on the data-independent estimate $\mu_{y\,|\,x}$ and higher data volume $N\,\psi_m(x)$ puts emphasis on the empirical mean.

4.2. Squared-Error Risk

Substituting the Bayesian estimator (9) into (1), transforming to the empirical statistic, and exploiting the properties from Section 3.2.1, the excess squared-error risk is

$$\begin{aligned}\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{x,\psi|\theta}\left[\big(\mu_{y\,|\,x,\psi} - \mu_{y\,|\,x,\theta}\big)^2\right] \\ &\equiv \mathrm{E}_{x\,|\,\theta_m}\left[\lambda_{\mathrm{Bias}}(x)\,\big(\mu_{y\,|\,x} - \mu_{y\,|\,x,\theta_c}\big)^2 + \lambda_{\mathrm{Var}}(x)\Sigma_{y\,|\,x,\theta_c}\right],\end{aligned} \tag{10}$$

a conditional expectation of two weighted functions, where

$$\lambda_{\mathrm{Bias}}(x) = \mathrm{E}_{\psi_m(x)|\,\theta_m(x)}\left[\gamma(x;\psi_m)^2\right] \tag{11}$$

and

$$\lambda_{\mathrm{Var}}(x) = \mathrm{E}_{\psi_m(x)|\,\theta_m(x)}\left[\frac{\big(1-\gamma(x;\psi_m)\big)^2}{N\,\psi_m(x)}\right]. \tag{12}$$

The first function is dependent on the squared bias between the clairvoyant estimate $\mu_{y\,|\,x,\theta_c}$ and the data-independent estimate $\mu_{y\,|\,x}$. The second function depends on $\Sigma_{y\,|\,x,\theta_c}$ and measures the estimator variance in excess of the clairvoyant squared-error $\mathcal{R}^*_\Theta(\theta)$.

Both functions are scaled by factors dependent on the conditional prior localizations $\alpha_0\,\alpha_m(x)$ and on $\theta_m(x)$ and N via expectations with respect to $\mathrm{P}_{\psi_m(x)|\,\theta_m(x)}$. By the aggregation property of empirical distributions, $N\,\psi_m(x)|\,\theta_m(x) \sim \mathrm{Bi}\big(N,\theta_m(x)\big)$ is binomially distributed. Closed-forms have not been found for the empirical process expectations in (11) and (12).

4.2.1. Trends

Consider the trends of the excess squared-error risk (10) with training data volume N and with Dirichlet prior parameterization. As N tends to infinity, the distributions $\mathrm{P}_{\psi_m(x)|\,\theta_m(x)}$ concentrate such that $\psi_m(x) \to \theta_m(x)$; thus for $\theta_m(x) > 0$, the weights $\lambda_{\mathrm{Var}}(x)$ and $\lambda_{\mathrm{Bias}}(x)$ both tend to zero and $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to 0$. This desirable result is a consequence

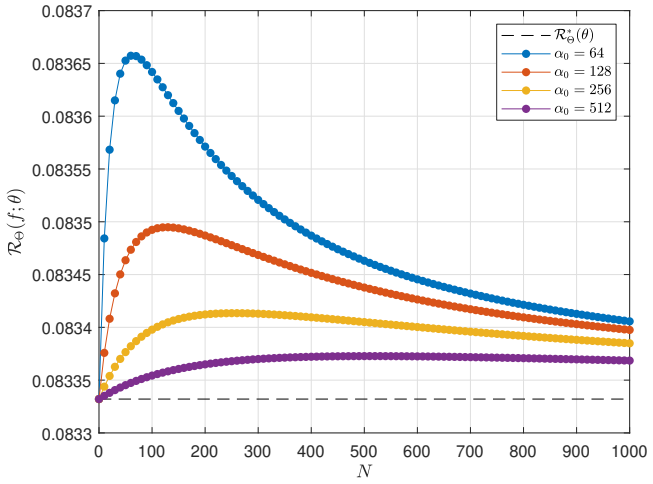

Fig. 2. Squared-Error Risk vs. N for unbiased estimators

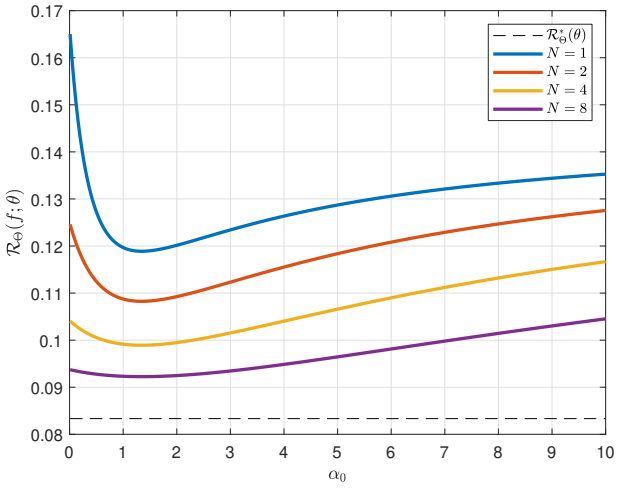

Fig. 3. Squared-Error Risk vs. α_0 for a biased estimator

of the full support of the Dirichlet prior, which ensures the convergence of $\theta_c(x)|\,\psi_m(x), \psi_c(x) \to \psi_c(x)$ as $N \to \infty$.

An interesting point regarding the dependency of the excess risk on N is that there may be a local maximum. To demonstrate, consider the case of $|\,\mathcal{X}\,| = 1$; treating N as a real number, there would be a maximum at

$$N = \alpha_0 \left(1 - 2\alpha_0 \frac{\left(\mu_{y\,|\,x} - \mu_{y\,|\,x,\theta_c}\right)^2}{\Sigma_{y\,|\,x,\theta_c}}\right) . \qquad (13)$$

If the prior estimate $\mu_{y\,|\,x}$ has low bias and yet the prior confidence α_0 is small, a local maximum may occur and additional training data may (temporarily) compromise the estimator performance.

Figure 2 exemplifies the excess conditional squared-error as a function of N for such estimators with varying localization α_0. The variable y is a discretization of the unit interval, such that $\mathcal{Y} = \{i/256 : i \in 0, \ldots, 255\}$, with $\theta_c(y;x) = 1/256$ and $\alpha_c(y;x) = \mathrm{Bi}(256y; 255, 0.5)$.

Next consider the effects of the conditional prior localizations $\alpha'(x) \equiv \alpha_0\,\alpha_m(x)$, which control a bias/variance risk trade-off. For maximal localization $\alpha'(x) \to \infty$, $\lambda_{\text{Bias}}(x) \to 1$ and $\lambda_{\text{Var}}(x) \to 0$. This is intuitive given that the estimator tends toward the data-independent solution. Conversely, for $\alpha'(x) \to 0$, $\lambda_{\text{Bias}}(x) \to \left(1 - \theta_m(x)\right)^N$ and $\lambda_{\text{Var}}(x) \to \mathrm{E}_{\psi_m(x)|\,\theta_m(x)}\left[\left(N\,\psi_m(x)\right)^{-1}\right]$. Note that the variance weight is equivalent to the first inverse moment of a positive binomial random variable [10].

Of primary interest are the values $\alpha'(x)$ that minimize the excess squared-error for given prior conditional distributions $\alpha_c(x)$. Calculating the first derivative, it can be shown that (for $N > 0$ and $\theta_m(x) > 0$) only one stationary point exists,

$$\alpha'(x) = \frac{\Sigma_{y\,|\,x,\theta_c}}{\left(\mu_{y\,|\,x} - \mu_{y\,|\,x,\theta_c}\right)^2} . \qquad (14)$$

Evaluation of the second derivative confirms the local minimum and comparison of the local risk value to the $\alpha'(x) \to 0$ and $\alpha'(x) \to \infty$ values confirms the global minimum.

Note that the minimizing localization values $\alpha'(x)$ are inversely proportional to the squared-bias of the prior conditional mean. This is sensible - the better the match between the true and prior predictive distributions, the more confidence should be expressed. Also, low localizations are preferable when the conditional models have low variance; these models can be accurately identified by learners that emphasize the empirical mean, even with limited training data. These trends are demonstrated in Figure 3 for a biased prior distribution with $\alpha_c(y;x) = \mathrm{Bi}(256y; 255, 0.75)$.

5. CONCLUSIONS

This paper has assumed a Dirichlet prior for Bayesian learning and applied the resultant predictive distribution to squared-error regression. Closed-forms have been provided for the optimal estimator and the achieved risk. Analysis and graphical examples highlight risk trends as a function of training data volume and the Dirichlet prior localization, demonstrating a bias/variance risk trade-off.

Future work will generalize these concepts for data drawn from continuous spaces using the continuous Dirichlet process [11]. Additionally, these estimators will be compared to classical methods such as generalized linear regression. The inherent adaptability due to the full prior support will highlight the utility of these Bayesian estimators for a wide range of regression problems; however, for data-limited problems, the "prior knowledge" implied by classical methods can yield superior performance when the data model is well understood.

6. REFERENCES

[1] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, ser. Signal Processing Series. Upper Saddle River, New Jersey: Prentice-Hall, 1998, vol. 2.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006.

[4] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2000.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., ser. Probability and Mathematical Statistics. New York, New York: John Wiley & Sons, 1971, vol. 2.

[6] T. P. Minka, "Bayesian inference, entropy, and the multinomial distribution," Microsoft Research, Tech. Rep., 2003.

[7] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*, ser. Probability and Statistics. John Wiley & Sons, 1997.

[8] K. P. Murphy, "Binomial and multinomial distributions," University of British Columbia, Tech. Rep., 2006.

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

[10] F. F. Stephan, "The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate," *The Annals of Mathematical Statistics*, vol. 16, no. 1, pp. 50–61, 1945.

[11] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, 2012.