BAYESIAN LEARNING FOR REGRESSION USING A DIRICHLET PRIOR PROBABILITY DISTRIBUTION

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 distribution. Dirichlet priors have a number of desirable properties - they lead to closed-form posterior distributions given independent training data, have full support over the space of data probability distributions, and can be fully objective or subjective depending on their localization parameter. This paper assumes a Dirichlet prior and produces predictive distributions to 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 corresponding minimum risk are presented and interpreted for different values of the localization parameter; specific attention is given to the extremal values.

Index Terms— Bayesian learning, machine learning, regression, estimation, Dirichlet distribution, 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 distributions are responsible for generating the observed/unobserved random elements. If a highly localized prior is chosen that strongly weights the actual data probability distribution, low risk learning functions are possible even with limited training data; however, if the localized prior is poorly selected, a good solution is unlikely to be achieved. Conversely, a less localized prior that treats the different distributions without preference adapts more responsively during training; if data is limited, however, the learning function may not deliver the required performance.

This work assumes that the joint observed and unobserved data elements are drawn from finite sets. The probability mass function (PMF) generating the data is 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 data [1]. Furthermore, control of the Dirichlet parameters can enable both minimally and maximally localized priors; the objective uniform distribution and the subjective delta function distribution are special cases.

After introducing the problem and discussing the relevant probability distributions, the Bayesian framework will be applied to the most common loss function for regression: the squared error loss function. Algebraic forms of optimal estimators and their corresponding minimum risk will be presented. Specific attention will be given to performance for the asymptotic cases of Dirichlet prior localization.

2. OBJECTIVE

Consider an observable random element $\mathrm{x} \in \mathcal{X}$ and and unobservable random element $\mathrm{y} \in \mathcal{Y}$ which are jointly distributed according to an unknown probability distribution $\theta \in \Theta = \left\{\theta \in {\mathbb{R}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\theta(y,x) = 1\right\}$, such that $\mathrm{P}_{\mathrm{y,x}\,|\,\theta}(y,x|\theta) = \theta(y,x)$.

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

The objective is to design a decision function $f : \mathcal{D} \mapsto \mathcal{H}^{\mathcal{X}}$ which outputs a mapping from the space of the observed random element x to a decision space $\mathcal{H}$. The metric guiding the design is a loss function $\mathcal{L} : \mathcal{H}\times\mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ which penalizes the decision $h \in \mathcal{H}$ based on the value of y. The conditional expected loss, or conditional "risk", is defined as

$$\mathcal{R}_\Theta(f;\theta) = \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y,x}\,|\,\theta}\left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right]. \quad (1)$$

As the model θ is not observed, $\mathcal{R}_\Theta$ is not yet a feasible

objective function for optimization. If the designer selects a PDF p_θ, the Bayes risk is formulated as

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta\left[\mathcal{R}_\Theta(f;\theta)\right] \\ &= \mathrm{E}_{\mathrm{y,x,D}}\left[\mathcal{L}(f(\mathrm{x};\mathrm{D}),\mathrm{y})\right]\end{aligned} \tag{2}$$

and y, x, and D are treated as jointly distributed random elements. The optimal learning function is expressed as

$$f^*(\mathrm{x};\mathrm{D}) = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\mathcal{L}(h,\mathrm{y})\right] \tag{3}$$

and the corresponding minimum Bayes risk is

$$\mathcal{R}(f^*) = \mathrm{E}_{\mathrm{x,D}}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\mathcal{L}(h,\mathrm{y})\right]\right] . \tag{4}$$

3. PROBABILITY DISTRIBUTIONS

In this section, the joint PMF $\mathrm{P}_{\mathrm{y,x,D}}$ is determined using the Dirichlet prior p_θ. Other distributions of interest will be provided, including the training data PMF P_D and the predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$.

3.1. Model PDF, p_θ

The Dirichlet PDF for the model random process $\theta \in \Theta$ is [2]

$$\mathrm{p}_\theta(\theta) = \beta(\alpha)^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha(y,x)-1} , \tag{5}$$

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

The parameters α controls around which models θ the PDF concentrates and how strongly. For convenience, introduce the localization parameter $\alpha_0 \equiv \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \alpha(y,x)$. Note that $\mathrm{P}_{\mathrm{y,x}}(y,x) = \mu_\theta(y,x) = \alpha(y,x)/\alpha_0$, where μ denotes the mean function of a random variable.

Of specific interest is how p_θ changes as the localization parameter approaches its limiting values. For $\alpha_0 \to \infty$, the PDF concentrates at its mean, resulting in

$$\mathrm{p}_\theta(\theta) \to \delta\left(\theta - \frac{\alpha}{\alpha_0}\right) , \tag{6}$$

where $\delta(\cdot)$ represents the Dirac delta function over Θ. Conversely, for $\alpha_0 \to 0$, the PDF tends toward

$$\mathrm{p}_\theta(\theta) \to \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0}\delta\big(\theta(\cdot,\cdot) - \delta[\cdot,y]\delta[\cdot,x]\big) , \tag{7}$$

where $\delta[\cdot,\cdot]$ is the Kronecker delta function. Note that the PDF has support only for the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models with an ℓ_0 norm satisfying $\|\theta\|_0 = 1$.

These trends are demonstrated in Figure 1. The cardinalities $|\,\mathcal{Y}\,| = 3$ and $|\,\mathcal{X}\,| = 1$ are chosen to enable visualization, despite the implication that x is deterministic. Note that for the top sub-plot, $\alpha(y,x) < 1$ and the PDF tends to infinity at the domain boundaries; this cannot be captured by the plot color scale.

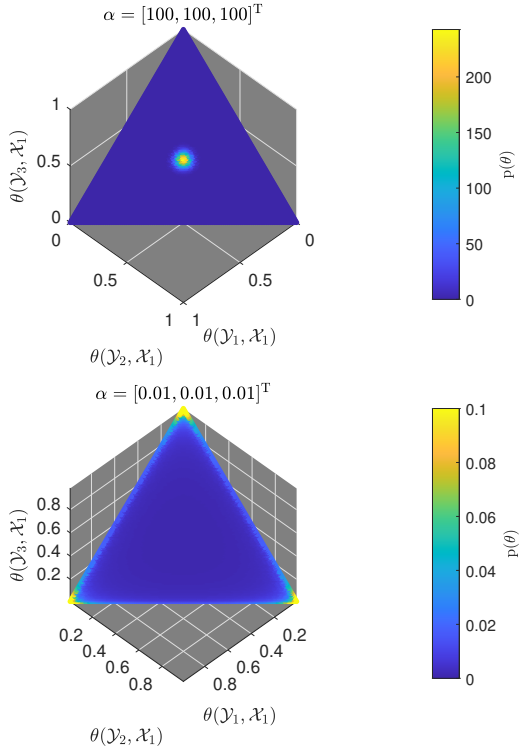

Fig. 1. Model prior PDF p_θ for different localizations α_0

3.2. Training Set PMF, P_D

The distribution of the training data D conditioned on the model can be formulated as

$$\mathrm{P}_{\mathrm{D}\,|\,\theta}\left(D|\theta\right) = \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{N}(y,x;D)} , \tag{8}$$

where the dependency on the training data D is expressed though a transform function $\bar{N} : \mathcal{D} \mapsto \bar{\mathcal{N}}$ defined as $\bar{N}(y,x;D) = \sum_{n=1}^{N} \delta\left[y,Y_n\right]\delta\left[x,X_n\right]$. This transform counts the occurences of each possible pair (y,x) in the training data; its range is a function space $\bar{\mathcal{N}} \subset \mathbb{Z}_{\geq 0}{}^{\mathcal{Y}\times\mathcal{X}}$.

As the conditional distribution $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ is of exponential form, it can be readily shown that the marginal distribution of the training data is [3]

$$\mathrm{P}_\mathrm{D}(D) = \beta(\alpha)^{-1}\beta\left(\alpha + \bar{N}(D)\right) . \tag{9}$$

Note that both $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ and P_D depend on the training data D only through the transform $\bar{N}$; consequently, $\bar{N}(\mathrm{D})$ is a sufficient statistic [4] for the model θ. As such, it is useful to define a new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$.

The cardinality of the random process' domain is $|\bar{\mathcal{N}}| = \mathcal{M}\left(\{N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\}\right)$, where $\mathcal{M}$ is the multinomial coefficient; this can be shown using the stars-and-bars method [5]. The cardinality of original set is $|\,\mathcal{D}\,| = \left(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\right)^N$; thus $|\bar{\mathcal{N}}| \leq |\,\mathcal{D}\,|$ and the sufficient statistic compactly represents the valuable information in the training data.

It is easily shown that the conditional PMF $\mathrm{P}_{\bar{\mathrm{n}}\,|\,\theta}$ is multinomial. As the Dirichlet distribution characterizes the parameters of this multinomial distribution, the marginal PMF of $\bar{\mathrm{n}}$ is a Dirichlet-Multinomial distribution [6] parameterized by α,

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) = \mathcal{M}(\bar{n})\beta(\alpha)^{-1}\beta(\alpha+\bar{n})\ . \tag{10}$$

The first and second joint moments of $\bar{\mathrm{n}}$ are

$$\mu_{\bar{\mathrm{n}}}(y,x) = N\frac{\alpha(y,x)}{\alpha_0} = N\mu_\theta(y,x) \tag{11}$$

and

$$\begin{aligned}&\mathrm{E}\left[\bar{\mathrm{n}}(y,x)\bar{\mathrm{n}}(y',x')\right] \\ &= \frac{N}{\alpha_0+1}\Big((\alpha_0+N)\mu_\theta(y,x)\delta[y,y']\delta[x,x'] \\ &\qquad + \alpha_0(N-1)\mu_\theta(y,x)\mu_\theta(y',x')\Big)\ .\end{aligned} \tag{12}$$

Again, the distributions for minimal and maximal α_0 are relevant. For $\alpha_0 \to \infty$, the model PDF p_θ concentrates at its mean and thus $\bar{\mathrm{n}}$ is characterized by a multinomial distribution,

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) = \mathcal{M}(\bar{n})\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\left(\frac{\alpha(y,x)}{\alpha_0}\right)^{\bar{n}(y,x)}\ . \tag{13}$$

Conversely, for $\alpha_0 \to 0$, the PMF tends toward

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\frac{\alpha(y,x)}{\alpha_0}\delta\big[\bar{n}(\cdot,\cdot), N\delta[\cdot,y]\delta[\cdot,x]\big] \tag{14}$$

and the training data are identical.

3.2.1. Aggregation Properties

Also of importance is the "marginalized" random process n', defined as $\mathrm{n}'(x) \equiv \sum_{y\in\mathcal{Y}}\bar{\mathrm{n}}(y,x)$ over the set $\mathcal{X}$. By the aggregation property of Dirichlet-Multinomial functions [6], the new function is distributed as $\mathrm{n}' \sim \mathrm{DM}(N,\alpha')$, where $\alpha'(x) = \sum_{y\in\mathcal{Y}}\alpha(y,x)$.

Also of interest is the distribution of $\bar{\mathrm{n}}$ conditioned on its aggregation n'. Using the Dirichlet-Multinomial properties presented in Appendix A,

$$\begin{aligned}&\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}(\bar{n}|n') \\ &\quad = \prod_{x\in\mathcal{X}}\Big[\mathcal{M}\left(\bar{n}(\cdot,x)\right)\beta\big(\alpha(\cdot,x)\big)^{-1}\beta\big(\alpha(\cdot,x)+\bar{n}(\cdot,x)\big)\Big]\ .\end{aligned} \tag{15}$$

Observe that conditioning on the aggregation renders the function segments $\bar{\mathrm{n}}(\cdot,x)$ independent of one another and that they are also Dirichlet-Multinomial.

3.3. Predictive PMF, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$

As shown in Equation (3), the decision selected by the optimally designed function depends on the predictive distribution of the unobserved y conditioned on all observable random elements. Note that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} = \mathrm{E}_{\theta|\,\mathrm{x,D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$ - the Bayesian predictive PMF is the expected value of the true predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ with respect to the model posterior distribution.

Since $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ is of exponential form, the Dirichlet prior p_θ is its conjugate prior [7]; thus, the model posterior PDF given the training data is

$$\begin{aligned}\mathrm{p}_{\theta|\,\mathrm{D}}(\theta|D) &= \frac{\mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta)\,\mathrm{p}_\theta(\theta)}{\mathrm{P}_{\mathrm{D}}(D)} \\ &= \beta\left(\alpha+\bar{N}(D)\right)^{-1}\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\theta(y,x)^{\alpha(y,x)+\bar{N}(y,x;D)-1}\ ,\end{aligned} \tag{16}$$

a Dirichlet distribution with parameter function $\alpha+\bar{N}(D)$.

Also, the localization parameter increases proportionately with the volume of training data; consequently, as $N\to\infty$, the posterior converges to $\mathrm{p}_{\theta|\,\mathrm{D}}(\theta|D) \to \delta\big(\theta-\bar{N}(D)/N\big)$ and the model is positively identified. This is a consequence of the full support of the Dirichlet distribution; general posterior distributions do not tend to a delta function at the empirical PMF. Conversely, as $\alpha_0\to\infty$, the prior model certainty is stronger and the posterior tends toward $\mathrm{p}_{\theta|\,\mathrm{D}}(\theta|D) \to \delta(\theta-\alpha/\alpha_0)$, independent of the training data.

The joint PMF of y and x conditioned on the training data is expressed as [8]

$$\begin{aligned}\mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{D}}(y,x|D) &= \mu_{\theta|\,\mathrm{D}}(y,x) \\ &= \frac{\alpha(y,x)+\bar{N}(y,x;D)}{\alpha_0+N}\end{aligned} \tag{17}$$

and the predictive distribution is generated via Bayes rule as

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|x,D) &= \frac{\alpha(y,x)+\bar{N}(y,x;D)}{\alpha'(x)+N'(x;D)} \\ &= \left(\frac{\alpha'(x)}{\alpha'(x)+N'(x;D)}\right)\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}(y|x) \\ &\quad + \left(\frac{N'(x;D)}{\alpha'(x)+N'(x;D)}\right)\frac{\bar{N}(y,x;D)}{N'(x;D)}\ ,\end{aligned} \tag{18}$$

where $N'(x;D) = \sum_{n=1}^{N}\delta[x,X_n]$. The last representation views the distribution as a convex combination of two conditional distributions. The first distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}(y|x) = \alpha(y,x)/\alpha'(x)$ is independent of the training data and based on the prior knowledge implied via the model PDF parameter; the second distribution is the conditional empirical PMF and depends solely on D. For both, only those values α and D corresponding to the observed value of x influence the distribution.

The weighting factors are dependent on these values as well. For $N'(x;D)=0$ or as $\alpha_0\to\infty$, the PMF tends toward the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$, which only depends

on the model parameter α. As the number of training examples increases or as $\alpha_0 \to 0$, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ tends towards the empirical conditional distribution.

4. REGRESSION AND THE SQUARED-ERROR LOSS

The squared-error (SE) loss function is arguably the most commonly used loss function for regression, or in fact for any estimation problem. This can be attributed to its quadratic form, which enables closed-form expression of the minimizing estimation function f^*.

It is assumed that the unobserved random element y is a scalar random variable; that is, $\mathcal{Y} \subset \mathbb{R}$. Additionally, the learning function's estimate is allowed to assume real numbers; thus, $\mathcal{H} = \mathbb{R} \supset \mathcal{Y}$.

The loss function is defined as

$$\mathcal{L}(h, y) = (h - y)^2 \,. \tag{19}$$

Substituting into (2), the Bayes risk for a general estimator is

$$\mathcal{R}(f) = \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\big(f(\mathrm{x};\mathrm{D}) - \mathrm{y}\big)^2\right]\right] . \tag{20}$$

4.1. Optimal Estimate: the Posterior Mean

To find the optimal estimator, the squared-error loss is substituted into (3); note that the objective function is quadratic over the argument h. It is easily shown that the function over h is positive-definite; as such, the minimizing decision h is the sole stationary point. Setting the first derivative of the function to zero, the optimal estimate is the expected value of y given the training data and the observed value x, such that

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \operatorname*{arg\,min}_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[(h-\mathrm{y})^2\right] = \mu_{\mathrm{y}\,|\,\mathrm{x,D}} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\mu_{\mathrm{y}\,|\,\mathrm{x}} \\ &\quad + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathrm{Y}_n}{N'(\mathrm{x};\mathrm{D})} \,. \end{aligned} \tag{21}$$

The optimal estimate is interpreted as a convex combination of two separate estimates - the expected value of y conditioned on the observed x and the average of the training values Y_n which have a value X_n matching the observed value of x. The weighting factors are the same as those of $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$; thus, stronger prior information (larger $\alpha'(x)$) provides more weight to the estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ and more voluminous training data puts emphasis on the empirical conditional mean.

Another interesting form for the optimal estimator is $f^*(\mathrm{x};\mathrm{D}) = \mathrm{E}_{\theta|\,\mathrm{x,D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$. If the model θ were known, then the clairvoyant estimate $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ would be optimal; instead, all such estimates are weighted and combined via the expectation over the model posterior PDF $\mathrm{p}_{\theta|\,\mathrm{x,D}}$.

4.2. Minimum Risk: the Expected Posterior Variance

Substituting the optimal estimator (21) into Equation (20), the minimum Bayes risk is the expected conditional variance

$$\begin{aligned} \mathcal{R}(f^*) &= \mathrm{E}_{\mathrm{x,D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \\ &= \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{C}_{\theta|\,\mathrm{x,D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\right] \,, \end{aligned} \tag{22}$$

where Σ is the variance of a random variable and the operator C is the variance of a function of a random variable. The second formula is of interest. The first term is the expected squared-error of the clairvoyant estimate $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$; the second term is its expected conditional variance with respect to the model posterior PDF $\mathrm{p}_{\theta|\,\mathrm{x,D}}$.

Using the sufficient statistic $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$, the minimum risk can also be represented as $\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right]$. Decompose the conditional variance as

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} \tag{23}$$

and assess the expected values of these terms separately. The first term is

$$\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2]\right] = \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2]\right] \,, \tag{24}$$

where the expectation over x is performed using $\mathrm{P}_{\mathrm{x}}(x) = \alpha'(x)/\alpha_0$. As proven in Appendix B, the second term is

$$\begin{aligned} &\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{N\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \big(\alpha_0\alpha'(\mathrm{x}) + N\alpha'(\mathrm{x}) + \alpha_0\big)\mu^2_{\mathrm{y}\,|\,\mathrm{x}}}{(\alpha_0 + N)\big(\alpha'(\mathrm{x}) + 1\big)}\right] . \end{aligned} \tag{25}$$

Combining, the mininum Bayes risk is

$$\begin{aligned} \mathcal{R}(f^*) &= \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + (\alpha_0 + N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + \alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] . \end{aligned} \tag{26}$$

The risk is a convex combination of scaled conditional variances for the different PMFs $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$. The convex coefficients are values from the prior marginal distribution P_{x}.

The scaling factor for each term $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ depends on the marginal PMF value $\mathrm{P}_{\mathrm{x}}(x)$, as well as on the prior localization α_0 and the number of training samples N. Observe that with no training data ($N = 0$), the scaling factor becomes unity and the risk is $\mathcal{R}(f^*) = \mathrm{E}_{\mathrm{x}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$. Conversely, as $N \to \infty$, the Bayes risk is $\mathcal{R}(f^*) = \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] = \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$. Also, as the model localization parameter $\alpha_0 \to 0$, the risk tends to zero (for $N > 0$); as $\alpha_0 \to \infty$, the risk tends toward $\mathrm{E}_{\mathrm{x}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$.

4.2.1. Examples

To illustrate these trends, explicitly define the sets $\mathcal{Y} = \{i/M_{\mathrm{y}} : i = 0, \ldots, M_{\mathrm{y}} - 1\}$ and $\mathcal{X} = \{i/M_{\mathrm{x}} : i =$

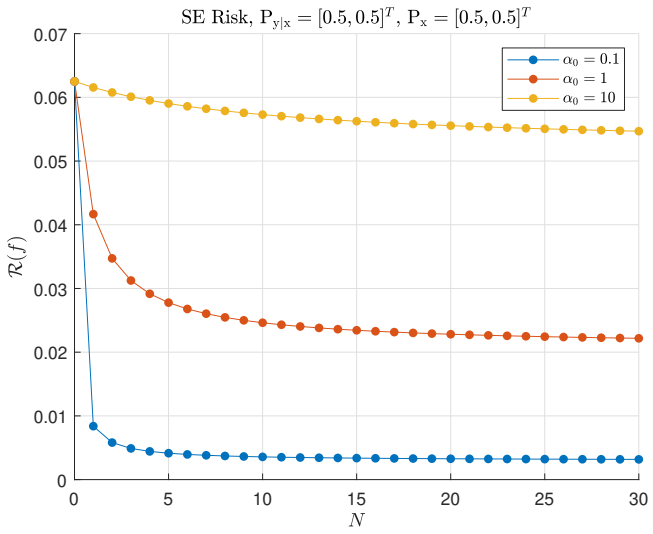

Fig. 2. Minimum SE Risk for different training set sizes N

$0, \ldots, M_{\mathrm{x}} - 1\}$. Let the conditional variances $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ be the same for each value $\mathrm{x} = x$; in this case, the squared-error becomes the conditional variance scaled by a factor dependent on the marginal distribution P_{x}, such that $\mathcal{R}(f^*) = \Sigma_{\mathrm{y}\,|\,\mathrm{x}}\, \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+(\alpha_0+N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}}\right]$. Figures 2 and 3 display how the risk changes with N and α_0 when $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$ and P_{x} are fixed.

It may not seem intutitve for the risk to decrease when α_0 is smaller – the variance of the model θ increases and the prior knowledge is less definitive. This is a result of the Dirichlet PDF weight shifting towards the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ different sparse models which have ℓ_0 norms satisfying $\|\theta\|_0 = 1$. Although these PMFs are maximally separated, they all have zero variance. The optimal learner (21) will simply use the empirical distribution supplied via the training data - this allows exact identification of θ with a single training pair.

It is also informative to visualize how the minimum squared-error changes with the marginal distribution P_{x} for fixed volume of training data N and prior localization α_0. Figure 4 demonstrates how the risk changes with this marginal PMF. Observe that the risk is maximal at the distributions satisfying $\|\,\mathrm{P}_{\mathrm{x}}\,\|_0 = 1$; the scaling factor for the conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ becomes $\frac{1+(\alpha_0+N)^{-1}}{1+\alpha_0^{-1}}$. Conversely, for $\mathrm{P}_{\mathrm{x}} = 1/|\,\mathcal{X}\,|$ the scaling factor becomes $\frac{|\,\mathcal{X}\,|^{-1}+(\alpha_0+N)^{-1}}{|\,\mathcal{X}\,|^{-1}+\alpha_0^{-1}}$ and the risk is minimal.

5. CONCLUSIONS

This paper has assumed a Dirichlet prior for Bayesian learning, established relevant distributions, and applied the framework to squared-error regression. Closed-forms have been provided for the optimal estimation function and the mini-

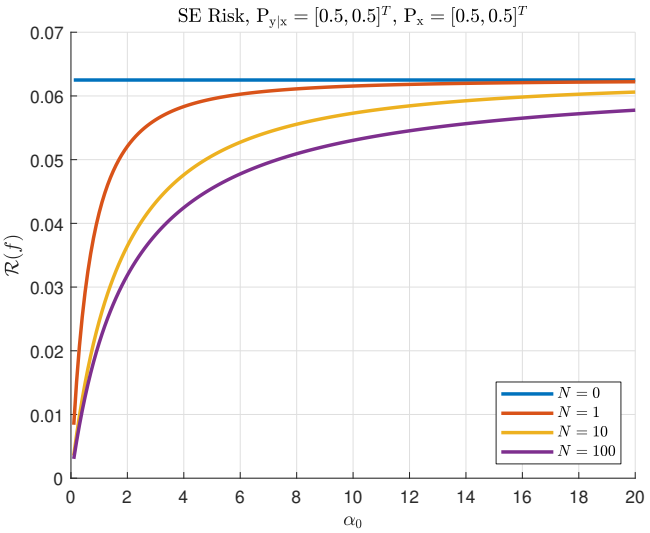

Fig. 3. Minimum SE Risk for different prior localization α_0

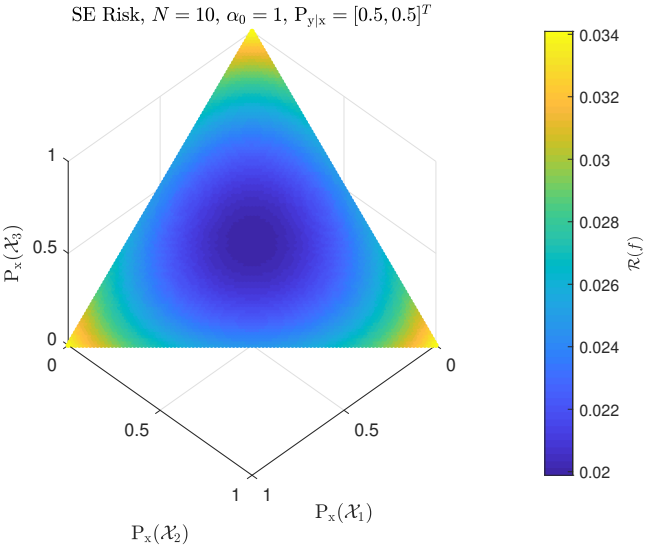

Fig. 4. Minimum SE Risk for different PMFs P_{x}

mum Bayes risk. Analysis and graphical examples highlight interesting trends for the estimation squared-error as a function of training data volume and the Dirichlet prior distribution parameters. Notably, Dirichlet priors with smaller localization parameters (and thus larger covariance) lead to optimal estimators with lower Bayes risk.

Future work will explore how the conditional risk (1) changes with different estimator parameterizations α for different true models θ. It will be shown that there is an optimal localization parameter dependent on the Dirichlet prior mean μ_θ and the actual model; whether the optimal localization parameter is large or small will depend on the quality of the match between μ_θ and θ. Additionally, a "minimax" function will be found that minimizes the worst-case conditional risk $\max_{\theta\in\Theta} \mathcal{R}_\Theta(f;\theta)$.

A. DIRICHLET-MULTINOMIAL RANDOM PROCESS CONDITIONED ON ITS AGGREGATION

A defining characteristic of a Dirichlet-Multinomial random process is that its aggregations are also Dirichlet-Multinomial [6]. Consider a DM random process $\bar{\mathrm{n}} \sim \mathrm{DM}(N,\alpha)$ over the set $\mathcal{Y}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots,\mathcal{S}(z),\ldots\}, z \in \mathcal{Z}$; the transformed random process $\mathrm{n}'(z) \equiv \sum_{y\in\mathcal{S}(z)} \bar{\mathrm{n}}(y)$ is neccessarily Dirichlet-Multinomial with parameterizing function $\alpha'(z) = \sum_{y\in\mathcal{S}(z)} \alpha(y)$.

Conditioned on the aggregation n', the segments $\big\{\bar{\mathrm{n}}(y) : y \in \mathcal{S}(z)\big\}$ of the original random process become independent Dirichlet-Multinomial random processes, such that

$$\begin{aligned}
\mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}'}(\bar{n}|n') &= \frac{\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n})}{\mathrm{P}_{\mathrm{n}'}(n')}\,\mathrm{P}_{\mathrm{n}'\,|\bar{\mathrm{n}}}(n'|\bar{n}) \qquad (27)\\
&= \frac{\mathcal{M}(\bar{n})\beta(\alpha)^{-1}\beta(\alpha+\bar{n})}{\mathcal{M}(n')\beta(\alpha')^{-1}\beta(\alpha'+n')}\\
&= \left(\prod_{z\in\mathcal{Z}} \frac{\Gamma\big(\alpha'(z)+n'(z)\big)}{n'(z)!\Gamma\big(\alpha'(z)\big)}\right)^{-1}\left(\prod_{y\in\mathcal{Y}} \frac{\Gamma\big(\alpha(y)+\bar{n}(y)\big)}{\bar{n}(y)!\Gamma\big(\alpha(y)\big)}\right)\\
&= \prod_{z\in\mathcal{Z}}\left[\frac{n'(z)!\Gamma\big(\alpha'(z)\big)}{\Gamma\big(\alpha'(z)+n'(z)\big)}\prod_{y\in\mathcal{S}(z)}\frac{\Gamma\big(\alpha(y)+\bar{n}(y)\big)}{\bar{n}(y)!\Gamma\big(\alpha(y)\big)}\right],
\end{aligned}$$

over domain $\bar{\mathrm{n}} \in \Big\{\mathbb{Z}_{\geq 0}{}^{\mathcal{Y}} : \sum_{y\in\mathcal{S}(z)} \bar{n}(y) = \mathrm{n}'(z), \quad \forall z \in \mathcal{Z}\Big\}$.

B. FORMULATION OF $\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right]$

The expected value of the conditional squared mean is compactly represented as

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \qquad (28)\\
&= \sum_{x,y,y'} yy'\,\mathrm{E}_{\bar{\mathrm{n}}}\Big[\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\,\bar{\mathrm{n}}}(y,x|\,\bar{\mathrm{n}})\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y'|x,\bar{\mathrm{n}})\Big]\\
&= \sum_{x,y,y'} yy'\,\mathrm{E}_{\mathrm{n}'}\left[\frac{1}{(\alpha_0+N)\big(\alpha'(x)+\mathrm{n}'(x)\big)}\right.\\
&\qquad \left.\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}\Big[\big(\alpha(y,x)+\bar{\mathrm{n}}(y,x)\big)\big(\alpha(y',x)+\bar{\mathrm{n}}(y',x)\big)\Big]\right]\\
&= \sum_{x\in\mathcal{X}} \frac{1}{(\alpha_0+N)\big(\alpha'(x)+1\big)}\,\mathrm{E}_{\mathrm{n}'}\Big[\mathrm{n}'(x)\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2](x)\\
&\qquad +\alpha'(x)\big(\alpha'(x)+\mathrm{n}'(x)+1\big)\mu^2_{\mathrm{y}\,|\,\mathrm{x}}(x)\Big]\\
&= \mathrm{E}_{\mathrm{x}}\left[\frac{N\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2]+\big(\alpha_0\alpha'(\mathrm{x})+N\alpha'(\mathrm{x})+\alpha_0\big)\mu^2_{\mathrm{y}\,|\,\mathrm{x}}}{(\alpha_0+N)\big(\alpha'(\mathrm{x})+1\big)}\right].
\end{aligned}$$

The formulation uses the statistical characterization of the aggregation, $\mathrm{n}' \sim \mathrm{DM}(N,\alpha')$; also used is the property that the Dirichlet-Multinomial random process $\bar{\mathrm{n}}$ conditioned on its aggregation n' yields independent conditional DM functions $\bar{\mathrm{n}}(\cdot,x)|\,\mathrm{n}'(x) \sim \mathrm{DM}\big(\mathrm{n}'(x),\alpha(\cdot,x)\big)$.

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

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

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

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

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