Bayesian Learning using a Dirichlet Prior for Regression and Classification

Paul Rademacher

June 9, 2021

Contents

Todo list

Operator/functional terminology?

likelihood function terminology

empirical risk terms/discussion?

ALL figure notation: theta font + Ycal indexing. use R,f opt?

Generalize to semi-supervised joint decisions??? training/test!

Investigate N lim for psi given theta, risks. Model support, bounded prior?

priors = sparse conditionals; w/ sufficient statistics

NFLT investigation? try sim examples

generalize y,x,h from scalars to functions!!!

jeffrey prior, fisher info?

HALDANE PRIOR

bibliography

Theo: (mult moments), DP agg, Dir moments

Bishop: (dir eq), dir posterior, moments, mode

Ferguson: (agg Dir), agg DP - via theo, DP posterior, moments

Gershman: agg DP - ref ferg, discrete draws

Johnson GET PDF: mult moments, (mult agg, DM agg, DM moments)

Add Theo-PR???

Chapter 1

Introduction

1.1 Background

PGR: complete rework!!

This report details a Bayesian perspective on statistical learning theory for when both the observations and unobserved quantities are jointly distributed according to an unknown probability distribution function. While the validity of Bayesian methods for statistical signal processing and machine learning has long been contended, the author believes it to be a justified approach that does not necessarily imply that the distribution model is 'random'; rather, it simply reflects the desire of the user to formulate risk as a weighted sum of learner performance across the space of distributions.

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 informative prior [4] is chosen that is concentrated around the actual data probability distribution, low risk learning functions are possible even with limited training data; however, if the informative prior is poorly designed, a

good solution may not be achieved. Conversely, a non-informative prior that weights the different distributions without preference provides a more robust solution for all models, but may under-perform relative to learners based on well-selected informative priors.

This work assumes that the prior distribution is Dirichlet. The class of Dirichlet probability density functions (PDF) and processes have the desirable properties of full support over the set of possible data-generating distributions and an analytic posterior distribution for independently and identically distributed data [7]. Furthermore, control of the Dirichlet parameters can enable both non-informative and informative prior knowledge. Special cases including the uniform prior will be given specific attention.

After introducing the problem and discussing the relevant data probability distributions, the Bayesian framework will be applied to two of the most common loss functions in machine learning: the squared error loss function (common for regression), and the 0-1 loss function [1] (common for classification). Optimal estimators/classifiers and their corresponding minimum risk will be presented for different Dirichlet prior distributions. Specific attention will be given to various asymptotic cases to show the differing performance for non-informative and informative Dirichlet priors.

1.2 Notation

Discuss arithmetic ops on functions? Upcasting?!

This section details the mathematical notation and typesetting conventions used throughout. Note that many variable scalars and functions including x, y, g, etc. are repeatedly redefined and reused to avoid introducing an excessive volume of symbols; unless explicitly stated, none of these variable definitions will hold in subsequent sections.

Sets and Function Arguments

Sets will typically be typeset with a calligraphic font, such as $\mathcal{X}$. Exceptions include common number sets such as the real numbers, which are typeset using blackboard bold $\mathbb{R}$. Function spaces such as the set of functions $\mathcal{X} \mapsto \mathcal{Y}$ are compactly represented as $\mathcal{Y}^{\mathcal{X}}$.

non-calligraphic for risk, loss, etc?

Various mappings will be defined for which the domain and/or the range [17] are function spaces. The set of functions $\mathcal{X} \mapsto \mathcal{Y}$ is denoted $\mathcal{Y}^{\mathcal{X}}$. For a mapping $g : \mathcal{Z} \mapsto \mathcal{Y}^{\mathcal{X}}$, the argument notation $g(z) \in \mathcal{Y}^{\mathcal{X}}$ denotes a function, while $g(x; z) \in \mathcal{Y}$ is a specific value of that function. Semicolons are used to distinguish between the arguments referring to the domain and arguments that access the resulting function. The mapping $\{1, \ldots, N\} \mapsto \mathcal{Y}$ will be represented as $\mathcal{Y}^N$ for brevity. Spaces of indexed tuples will be notated as $g \in \mathcal{Y}^N$ and items of a tuple are accessed with subscripts rather than parentheses, such that $g_i \in \mathcal{Y}$.

The Cartesian product of sets will be frequently used, such that for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For a general product of sets $\mathcal{S}_i$, the notation $\prod_i \mathcal{S}_i = \mathcal{S}_1 \times \mathcal{S}_2 \ldots$ is used.

The convention adopted for natural numbers is $\mathbb{N} = \{1, 2, \ldots\}$; the set of non-negative integers is denoted $\mathbb{Z}_{\geq 0} = \mathbb{N} \cup \{0\}$. The set of positive real numbers $\mathbb{R}^+$ excludes zero, while non-negative real numbers are represented as $\mathbb{R}_{\geq 0} = \mathbb{R}^+ \cup \{0\}$. The cardinality of countably infinite sets, including the set of natural numbers, is denoted $\aleph_0 = |\,\mathbb{N}\,|$; the cardinality of uncountable sets such as $\mathbb{R}$ is at least $\aleph_1$.

Numerous probability distribution functions will be defined over different domains. As such, for a given set $\mathcal{X}$, define a set function $\mathcal{P}$ such that $\mathcal{P}(\mathcal{X})$ is the set of distributions over $\mathcal{X}$. If $\mathcal{X}$ is countable, the set is defined as $\mathcal{P}(\mathcal{X}) = \left\{p \in {\mathbb{R}_{\geq 0}}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} p(x) = 1\right\}$; if $\mathcal{X}$ is a Euclidean space, the set is defined as $\mathcal{P}(\mathcal{X}) = \left\{p \in {\mathbb{R}_{\geq 0}}^{\mathcal{X}} : \int_{\mathcal{X}} p(x)\mathrm{d}x = 1\right\}$.

aleph reference?

Special Operators and Functions

Various operators commonly used in linear algebra will be generalized for functions. Specifically, the outer product operator $\otimes$ is used on two real-valued functions $f \in \mathbb{R}^{\mathcal{X}}$ and $g \in \mathbb{R}^{\mathcal{Y}}$, such that $\big(f \otimes g\big) \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ with $\big(f \otimes g\big)(x,y) = f(x)g(y)$. A general outer product of functions is denoted $\bigotimes_i f_i = f_1 \otimes f_2 \ldots$ for $i = 1, \ldots$, where $\Big(\bigotimes_i f_i\Big)(x_1, x_2, \ldots) = f_1(x_1)f_x(x_2)\ldots$, is used as well. Also, the diagonal operator operates on a single real-valued function, such that $\mathrm{diag}(f) \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$. For countable sets X, the operator values are $\mathrm{diag}(f)(x,x') = f(x)\delta[x,x']$; for Euclidean sets, the operator values are $\mathrm{diag}(f)(x,x') = f(x)\delta(x-x')$.

tensor product?

A variety of special functions will be used throughout. Both the Dirac and Kronecker delta functions will frequently required. The Dirac delta function over a Euclidean domain $\mathcal{X}$ is represented as $\delta(\cdot)$; it has support only at the point $x = 0$ and satisfies

$$\int_{\mathcal{X}} \delta(x)\mathrm{d}x = 1\ . \tag{1.1}$$

Consequently, it also satisfies

$$\int_{\mathcal{X}} g(x)\delta(x)\mathrm{d}x = g(0)\ . \tag{1.2}$$

Consider a set $\mathcal{X}$; the Kronecker delta function has domain $\mathcal{X}\times\mathcal{X}$ and is defined as

$$\delta[x,x'] = \begin{cases} 1 & \text{if } x = x', \\ 0 & \text{if } x \neq x'. \end{cases} \tag{1.3}$$

As both functions are denoted by the symbol δ, they are distinguished by the use of parentheses or square brackets. The relation $\delta(\cdot - x) = \delta(0)\delta[\cdot, x]$ may be used to relate the two functions.

Add Dirac subscript to explictly define domain?

reference Dirac/Kronecker

The multinomial coefficient and multivariate beta function, which typically operate on sequences, are defined more generally for function inputs. The multinomial operator $\mathcal{M}$ is used for functions $g : \mathcal{X} \mapsto \mathbb{Z}_{\geq 0}$ that map to non-negative integers from an arbitrary countable domain $\mathcal{X}$. The output of the operator is

$$\mathcal{M}(g) = \frac{\left(\sum_{x\in\mathcal{X}} g(x)\right)!}{\prod_{x\in\mathcal{X}} g(x)!} \, . \tag{1.4}$$

Similarly, the beta function β operates on functions $g : \mathcal{X} \mapsto \mathbb{R}^+$ that map to positive real numbers from an arbitrary countable domain $\mathcal{X}$, such that

$$\beta(g) = \frac{\prod_{x\in\mathcal{X}} \Gamma\big(g(x)\big)}{\Gamma\left(\sum_{x\in\mathcal{X}} g(x)\right)} \, . \tag{1.5}$$

Note that the countable domains of the input functions may have an infinite number of elements.

For a given subset $\mathcal{S} \subset \mathcal{X}$, the indicator function $\chi(\mathcal{S}) : \mathcal{X} \mapsto \{0, 1\}$, defined as

$$\chi(x; \mathcal{S}) = \begin{cases} 1 & \text{if } x \in \mathcal{S}, \\ 0 & \text{if } x \notin \mathcal{S} \, , \end{cases} \tag{1.6}$$

will be used repeatedly.

Random elements, variables, and processes

Introduce convergence in probability, relate to delta PDF convergence?

Random elements are denoted with roman font (e.g., x), while specific values are denoted with italics (e.g., x). Random elements that assume numerical scalars/functions are referred to as random variables/processes, respectively.

Consider a random element $\mathrm{x} \in \mathcal{X}$. If $\mathcal{X}$ is countable, either finite with $|\,\mathcal{X}\,| \in \mathbb{N}$ or countably infinite with $|\,\mathcal{X}\,| = \aleph_0$, then x is a discrete random element and is characterized by a probability mass function (PMF) [14], denoted $\mathrm{P_x} \in \mathcal{P}(\mathcal{X})$. If $\mathcal{X}$ is a Euclidean space and is thus uncountable with $|\,\mathcal{X}\,| \geq \aleph_1$, then x is a continuous random variable/process characterized by a probability density function (PDF), denoted $\mathrm{p_x} \in \mathcal{P}(\mathcal{X})$.

explicit PMF/PDF formula with P of events?

Consider x conditioned on another random element $\mathrm{z} \in \mathcal{Z}$. The conditional distribution is represented as $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}} : \mathcal{Z} \mapsto \mathcal{P}(\mathcal{X})$, such that $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}(z)$ is a PMF over $\mathcal{X}$ and $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}(x|z)$ is a specific value of that PMF. Often, the dependency on the conditional variable z will not be expressed in terms of a specific value z, but will be left in terms of the random element itself; in this case, the more compact notation $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}$ is used to imply $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}(\mathrm{z})$, a function of z.

Ever need the full functional?

Many distributions will be repeatedly used and thus special functions will be defined for the PDF's and PMF's of interest. For example, consider a random process $\mathrm{x} \in \mathcal{X}$ characterized by a Multinomial distribution with parameters $N \in \mathbb{Z}_{\geq 0}$ and $\theta \in \Theta$; the PDF will be notated as $\mathrm{Multi} : \mathbb{Z}_{\geq 0} \times \Theta \mapsto \mathcal{P}(\mathcal{X})$, where the range is the set of valid PDF's. More compactly, the notation $\mathrm{x} \sim \mathrm{Multi}(N, \theta)$ implies that $\mathrm{P}_{\mathrm{x}} = \mathrm{Multi}(N, \theta)$. Other distribution functions repeatedly used include Dir, DE, DP, and DEP, representing the Dirichlet distribution, the Dirichlet-Empirical distribution, the Dirichlet process, and the Dirichlet-Empirical process.

Add PDF citations. Introduce Empirical process?

Expectation Operators

For a discrete random element x, the expectation operator E_{x} is defined as

$$\mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right] = \sum_{x} \mathrm{P}_{\mathrm{x}}(x) g(x) \;, \tag{1.7}$$

where the argument g is an arbitrary scalar function of x with range $\mathbb{R}$. Additionally, define the variance operator C_{x} as

$$\mathrm{C}_{\mathrm{x}}\left[g(\mathrm{x})\right] = \mathrm{E}_{\mathrm{x}}\left[\Big(g(\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right]\Big)^2\right] . \tag{1.8}$$

When x is a random variable and the function g is the identity operator, such that $g(\mathrm{x}) = \mathrm{x}$, the mean and variance are compactly represented as μ_{x} and Σ_{x}, respectively.

These operations can be performed with respect to a conditional distribution as well. In this case, the expectation operator is a function of the observed value of z, such that

$$\mathrm{E}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](z) = \sum_{x} \mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}(x|z)g(x)\ . \tag{1.9}$$

Similarly, the conditional variance is notated $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](z)$. When g is the identity operator, the conditional mean and variance as represented by $\mu_{\mathrm{x}\,|\,\mathrm{z}}(z)$ and $\Sigma_{\mathrm{x}\,|\,\mathrm{z}}(z)$, respectively.

As with conditional distributions, it is common that an explicit value z of the conditional random element will not be used, but rather the expectation will be left as a function of the random element z. In these cases, the argument is suppressed and the notation $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right]$ implies the dependency on z. This convention also holds for the conditional variance operator $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{z}}$, as well as for the $\mu_{\mathrm{x}\,|\,\mathrm{z}}$ and $\Sigma_{\mathrm{x}\,|\,\mathrm{z}}$ functions.

If the range of g is a Hilbert space, such that $g(\mathrm{x})$ is itself a function with a domain $\mathcal{Y}$, then the notation for these operators is expanded. The output of the expectation operator is a function over $\mathcal{Y}$ represented by

$$\mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right](y) = \sum_{x} \mathrm{P}_{\mathrm{x}}(x)g(y;x)\ . \tag{1.10}$$

Similarly, the covariance function notation is modified and the output is a function over $\mathcal{Y}\times\mathcal{Y}$,

$$\mathrm{C}_{\mathrm{x}}\left[g(\mathrm{x})\right](y,y') = \mathrm{E}_{\mathrm{x}}\left[\Big(g(y;\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(y;\mathrm{x})\right]\Big)\Big(g(y';\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(y';\mathrm{x})\right]\Big)\right]\ . \tag{1.11}$$

$$\mathrm{C}_{\mathrm{x}}\left[g(\mathrm{x})\right] = \mathrm{E}_{\mathrm{x}}\left[\Big(g(\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right]\Big) \otimes \Big(g(\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right]\Big)\right]\ . \tag{1.12}$$

As before, the notation is simplified when the function g is the identity operator. If x is a random process over a domain $\mathcal{Y}$, then the mean and covariance functions are defined over domains $\mathcal{Y}$ and $\mathcal{Y}\times\mathcal{Y}$ with values notated such as $\mu_{\mathrm{x}}(y)$ and $\Sigma_{\mathrm{x}}(y,y')$.

If the expectations are evaluated with respect to a conditional distribution $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}$, the additional argument for the observed random element is added and the notation

for the above operators is extended to $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](y|z)$ and $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](y, y'|z)$ for non-scalar outputs. When g is the identity operator, the notation $\mu_{\mathrm{x}\,|\,\mathrm{z}}(y|z)$ and $\Sigma_{\mathrm{x}\,|\,\mathrm{z}}(y, y'|z)$ is used. As for probability distributions, it is common for the conditional random element z to be left as a random quantity instead of being explicitly defined; in these cases, the dependency on z is implied.

BELOW NOTATION CREATES AMBIGUITY!!!!!!! Check for residual uses...

In such cases, the italic z is dropped from the arguments and the formulas $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](y)$, $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](y, y')$, $\mu_{\mathrm{x}\,|\,\mathrm{z}}(y)$, and $\Sigma_{\mathrm{x}\,|\,\mathrm{z}}(y, y')$ imply dependence on z.

Chapter 2

Problem Statement

Generalize for limited dimensionality models?

2.1 Data Model

italic theta font before Bayes?

Consider an observable random element $\mathrm{x} \in \mathcal{X}$ and an unobservable random element $\mathrm{y} \in \mathcal{Y}$ which are jointly distributed according to an unknown probability distribution $\theta \in \Theta \equiv \mathcal{P}(\mathcal{Y} \times \mathcal{X})$, such that $\mathrm{P}_{\mathrm{y,x}\,|\theta} = \theta$. Note that the uppercase PMF notation used throughout this section implies that the random elements are discrete; PDF's are used when x and/or y are continuous random variables/processes.

Also observed is a random sequence of N samples from θ, denoted $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$; an alternative representation that can be used is $\mathrm{D} \Leftrightarrow \big((\mathrm{Y}_1, \mathrm{X}_1), \ldots, (\mathrm{Y}_N, \mathrm{X}_N)\big)$, where $\mathrm{Y} \in \mathcal{Y}^N$ and $\mathrm{X} \in \mathcal{X}^N$. The N data pairs are conditionally independent from one another and are identically distributed as $\mathrm{P}_{\mathrm{D}_n\,|\theta} = \mathrm{P}_{\mathrm{y,x}\,|\theta}$. The samples are also conditionally independent from (y, x). Thus $\mathrm{P}_{\mathrm{y,x,D}\,|\theta} = \mathrm{P}_{\mathrm{y,x}\,|\theta} \otimes \left(\bigotimes_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\theta}\right) = \theta \otimes \left(\bigotimes_{n=1}^{N} \theta\right)$, or explicitly,

$$\mathrm{P}_{\mathrm{y,x,D}\,|\theta}(y, x, D|\theta) = \mathrm{P}_{\mathrm{y,x}\,|\theta}(y, x|\theta) \prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\theta}\left(Y_n, X_n|\theta\right) . \tag{2.1}$$

2.1.1 Marginal and Conditional Model Distributions

As only y is unobservable, it will be useful to alternatively represent the model distribution via the bijection $\theta \Leftrightarrow (\theta_{\mathrm{m}}, \theta_{\mathrm{c}})$. First, introduce the marginal distribution $\theta_{\mathrm{m}} \equiv \sum_{y \in \mathcal{Y}} \theta(y, \cdot) \in \mathcal{P}(\mathcal{X})$; note that the summation is replaced by an integral when y is a continuous random variable. Next, introduce the conditional distributions $\theta_{\mathrm{c}} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$ defined as $\theta_{\mathrm{c}}(x) \equiv \theta(\cdot, x) / \theta_{\mathrm{m}}(x)$. Observe that $\mathrm{P}_{\mathrm{x}\,|\theta} \equiv \mathrm{P}_{\mathrm{x}\,|\,\theta_{\mathrm{m}}} = \theta_{\mathrm{m}}$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}} = \theta_{\mathrm{c}}(\mathrm{x})$.

2.2 Sufficient Statistic: the Empirical Distribution

continuous? empirical process?

For countable sets $\mathcal{Y}$ and $\mathcal{X}$, the distribution of D conditioned on the model can be formulated as

$$\begin{aligned} \mathrm{P}_{\mathrm{D}\,|\theta}\left(D|\theta\right) &= \prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\theta}\left(D_n|\theta\right) = \prod_{n=1}^{N} \theta(D_n) \\ &= \prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y,x)^{N\Psi(y,x;D)} \\ &= \left(\prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y,x)^{\Psi(y,x;D)} \right)^{N}, \end{aligned} \tag{2.2}$$

where the dependency on the training data D is expressed though a transform function $\Psi : \mathcal{D} \mapsto \Psi \subset \Theta$, defined as

$$\begin{aligned} \Psi(D) &= \frac{1}{N} \sum_{n=1}^{N} \delta\big[\cdot, D_n\big] \\ &\equiv \frac{1}{N} \sum_{n=1}^{N} \delta\left[\cdot, Y_n\right] \delta\left[\cdot, X_n\right] \end{aligned} \tag{2.3}$$

with range

$$\begin{aligned} \Psi &= \big\{ \Psi(D) : D \in \mathcal{D} \big\} \\ &= \left\{ \frac{n}{N} : n \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y} \times \mathcal{X}}, \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} n(y,x) = N \right\}. \end{aligned} \tag{2.4}$$

This function determines the empirical probability of the pair (y, x). Note that the set Ψ is a finite subset of Θ and thus that the empirical model $\Psi(\mathrm{D})$ is a valid probability distribution.

The distribution $\mathrm{P}_{\mathrm{D}\,|\theta}$ depends on the training data D only through the transform Ψ; as such, it is useful to define a new random process $\psi \equiv \Psi(\mathrm{D}) \in \Psi$. It can be shown using Neyman-Pearson factorization [11] that the data D is conditionally independent of the model given ψ – as such, $\Psi(\mathrm{D})$ is a sufficient statistic for the model θ.

D AND x jointly sufficient!?

The cardinality of the random process' domain is $|\Psi| = \mathcal{M}\left((N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1)\right)$; this can be shown using the stars-and-bars method [6]. The cardinality of original set is $|\,\mathcal{D}\,| = \left(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\right)^N$; thus $|\Psi| \leq |\,\mathcal{D}\,|$ and the sufficient statistic compactly represents the valuable information in the training data.

sufficient statistic savings in memory bits?

Conditioned on the model θ, the PMF of ψ is

$$\begin{aligned}
\mathrm{P}_{\psi|\theta}(\psi|\theta) &= \sum_{D\in\{\Psi(D)=\psi\}} \mathrm{P}_{\mathrm{D}\,|\theta}(D|\theta) \\
&= \left|\{D : \Psi(D) = \psi\}\right| \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{N\psi(y,x)} \\
&= \mathcal{M}(N\psi) \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{N\psi(y,x)} \\
&= \mathrm{Multi}\left(N\psi; N, \theta\right) \\
&= \mathcal{M}(N\psi) \left(\prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\psi(y,x)}\right)^N \\
&= \mathrm{Emp}\left(\psi; N, \theta\right) .
\end{aligned} \tag{2.5}$$

Observe that the Empirical process is equivalent to a Multinomial process within a scale factor.

The first and second joint moments of this Empirical distribution (derived from Multinomial moments [19]) are

$$\mu_{\psi|\theta} = \theta \tag{2.6}$$

and

$$\mathrm{E}_{\psi|\theta}\left[\psi \otimes \psi\right] = \frac{1}{N}\,\mathrm{diag}(\theta) + \left(1 - \frac{1}{N}\right)\theta \otimes \theta \tag{2.7}$$

and the covariance function is

$$\Sigma_{\psi|\theta} = \frac{1}{N}\big(\,\mathrm{diag}(\theta) - \theta \otimes \theta\big)\ . \tag{2.8}$$

The first and second moments of the Empirical distribution are proportionate to those of the Multinomial distribution.

Observe that for larger training data volumes N, the set Ψ becomes a denser grid of samples from the set Θ and the covariance tends to zero, concentrating the Empirical PMF around the model θ. Thus, as $N \to \infty$,

$$\mathrm{P}_{\psi|\theta}(\psi|\theta) \to \delta[\psi, \theta]\ . \tag{2.9}$$

This trend underscores the identifiability of the model θ.

CHECK!???? SCALING FACTOR???

FIGS!

Cite Glivenko–Cantelli theorem?

Also, using the maximum likelihood estimate of a Multinomial distribution [15], the maximum likelihood estimate of θ given the training data empirical model is simply

$$\theta_{\mathrm{ML}}\big(\psi\big) = \arg\max_{\theta \in \Theta} \mathrm{P}_{\psi|\theta}(\psi|\theta) = \psi\ . \tag{2.10}$$

2.2.1 Marginal and Conditional Data Distributions

Also of interest are the marginal and conditional distributions of the joint training data sequences Y and X. The marginal distribution given θ for the observations X

alone is

$$\begin{aligned} \mathrm{P}_{\mathrm{X}\,|\theta}\left(X|\theta\right) &= \prod_{n=1}^{N} \mathrm{P}_{\mathrm{X}_n\,|\theta}\left(X_n|\theta\right) \equiv \prod_{n=1}^{N} \theta_{\mathrm{m}}(X_n) \\ &\equiv \prod_{x\in\mathcal{X}} \theta_{\mathrm{m}}(x)^{N\,\Psi_{\mathrm{m}}(x;X)} \\ &= \left(\prod_{x\in\mathcal{X}} \theta_{\mathrm{m}}(x)^{\Psi_{\mathrm{m}}(x;X)}\right)^{N}, \end{aligned} \tag{2.11}$$

where the dependency on θ is only through the marginal model θ_{m}. Additionally, note that the dependency on the training observations X is expressed though a "marginal" distribution function $\Psi_{\mathrm{m}} : \mathcal{X}^N \mapsto \Psi_{\mathrm{m}} \subset \mathcal{P}(\mathcal{X})$ defined as

$$\Psi_{\mathrm{m}}(X) = \frac{1}{N}\sum_{n=1}^{N} \delta\big[\cdot, X_n\big] \equiv \sum_{y\in\mathcal{Y}} \Psi(y, \cdot; D) \tag{2.12}$$

with range

$$\begin{aligned} \Psi_{\mathrm{m}} &= \left\{\, \Psi_{\mathrm{m}}(X) : X \in \mathcal{X}^N \,\right\} \\ &= \left\{ \frac{n}{N} : n \in {\mathbb{Z}_{\geq 0}}^{\mathcal{X}},\ \sum_{x\in\mathcal{X}} n(x) = N \right\}. \end{aligned} \tag{2.13}$$

The conditional distribution of the values Y given the corresponding X and the model θ can be found using Bayes theorem as

$$\begin{aligned} \mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta}\left(Y|X,\theta\right) &= \prod_{n=1}^{N} \frac{\mathrm{P}_{\mathrm{Y}_n,\mathrm{X}_n\,|\theta}\left(Y_n, X_n|\theta\right)}{\mathrm{P}_{\mathrm{X}_n\,|\theta}\left(X_n|\theta\right)} \equiv \prod_{n=1}^{N} \theta_{\mathrm{c}}(Y_n; X_n) \\ &\equiv \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} \theta_{\mathrm{c}}(y; x)^{N\Psi(y,x;Y,X)} \\ &= \prod_{x\in\mathcal{X}} \left(\prod_{y\in\mathcal{Y}} \theta_{\mathrm{c}}(y; x)^{\Psi_{\mathrm{c}}(y;x;Y,X)}\right)^{N\,\Psi_{\mathrm{m}}(x;X)}, \end{aligned} \tag{2.14}$$

where the function $\Psi_{\mathrm{c}} : \{\mathcal{Y}\times\mathcal{X}\}^N \mapsto \Psi_{\mathrm{c}} \subset \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$ is defined as

$$\Psi_{\mathrm{c}}(x; Y, X) = \frac{\Psi(\cdot, x; Y, X)}{\Psi_{\mathrm{m}}(x; X)} = \frac{\sum_{n=1}^{N} \delta\big[\cdot, Y_n\big]\delta\big[x, X_n\big]}{\sum_{n=1}^{N} \delta\big[x, X_n\big]}. \tag{2.15}$$

Note that the conditional distribution is defined only for values $x \in \mathcal{X}_{\mathrm{s}}\left(\Psi_{\mathrm{m}}(X)\right)$, where $\mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}}) = \{x \in \mathcal{X} : \psi_{\mathrm{m}}(x) > 0\}$. Observe that the range of the transform is

$$\Psi_{\mathrm{c}} = \bigcup_{\psi_{\mathrm{m}}\in\Psi_{\mathrm{m}}} \prod_{x\in\mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})} \left\{ \frac{n}{N\,\psi_{\mathrm{m}}(x)} : n \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}},\ \sum_{y\in\mathcal{Y}} n(y) = N\,\psi_{\mathrm{m}}(x) \right\}. \tag{2.16}$$

and that the dependency of the conditional distribution on the model θ is expressed only through the conditional models $\theta_{\mathrm{c}}(x)$.

Analogous to the decomposition of the model θ into its marginal and conditional models, the empirical process can be decomposed into marginal and conditional empirical processes via a bijection $\psi \Leftrightarrow (\psi_{\mathrm{m}}, \psi_{\mathrm{c}})$. Introduce the "marginalized" random process ψ_{m} over the set $\mathcal{X}$, defined as $\psi_{\mathrm{m}} \equiv \sum_{y \in \mathcal{Y}} \psi(y, \cdot) \equiv \Psi_{\mathrm{m}}(\mathrm{X}) \in \Psi_{\mathrm{m}}$. Similar to the Multinomial random processes [9], the Empirical random process has an aggregation property (Appendix A.1). Using this principle, it can be shown that conditioned on the model θ, the marginal model is distributed as $\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}} \sim \mathrm{Emp}(N, \theta_{\mathrm{m}})$.

Also of interest is the conditional distribution of $\psi_{\mathrm{c}} \in \Psi_{\mathrm{c}}$, where $\psi_{\mathrm{c}}(x) \equiv \psi(\cdot, x) / \psi_{\mathrm{m}}(x)$. Using the Empirical process properties proven in Appendix A.1, it can be shown that when conditioned on ψ and on the model θ, the PMF of ψ_{c} is

$$\begin{aligned} \mathrm{P}_{\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}, \theta}(\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}, \theta) &\equiv \mathrm{P}_{\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}, \theta_{\mathrm{c}}}(\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}, \theta_{\mathrm{c}}) \\ &= \prod_{x \in \mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})} \left[\mathcal{M}\left(N\, \psi_{\mathrm{m}}(x)\, \psi_{\mathrm{c}}(x)\right) \left(\prod_{y \in \mathcal{Y}} \theta_{\mathrm{c}}(y; x)^{\psi_{\mathrm{c}}(y;x)} \right)^{N\, \psi_{\mathrm{m}}(x)} \right] \\ &= \prod_{x \in \mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})} \mathrm{Emp}\left(\psi_{\mathrm{c}}(x); N\, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x) \right), \end{aligned} \tag{2.17}$$

$$\begin{aligned} \mathrm{P}_{\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}, \theta} &\equiv \bigotimes_{x \in \mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})} \mathrm{P}_{\psi_{\mathrm{c}}(x) | \, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x)} \\ &= \bigotimes_{x \in \mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})} \mathrm{Emp}\left(N\, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x) \right), \end{aligned} \tag{2.18}$$

over the domain $\prod_{x \in \mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})} \left\{ \frac{n}{N\, \psi_{\mathrm{m}}(x)} : n \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}},\ \sum_{y \in \mathcal{Y}} n(y) = N\, \psi_{\mathrm{m}}(x) \right\}$. Observe that conditioning on the marginal empirical process renders the conditional processes $\psi_{\mathrm{c}}(x)$ independent of one another and that they are also Empirically distributed, such that $\psi_{\mathrm{c}}(x) | \, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x) \sim \mathrm{Emp}\left(N\, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x)\right)$ for every $x \in \mathcal{X}_{\mathrm{s}}(\psi_{\mathrm{m}})$.

otimes for all? ordered set? use sim notation?

2.3 Learning Objective

use marginal/conditional model? D or psi?

Continuous? Comment on PMF notation

The task in supervised machine learning is to design a learning function $f : \mathcal{D} \mapsto \mathcal{H}^{\mathcal{X}}$ which produces a mapping from the space of the observed random elements to a decision space $\mathcal{H}$. Define the function space $\mathcal{F} = \left\{\mathcal{H}^{\mathcal{X}}\right\}^{\mathcal{D}}$, such that $f \in \mathcal{F}$. The decision functions $f(D)$ are non-parametric and there are no restrictions on the set of achievable functions $\mathcal{H}^{\mathcal{X}}$.

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 objective is to minimize the expected loss, or "risk",

$$\begin{aligned} \mathcal{R}_{\Theta}(f;\theta) &= \mathrm{E}_{\mathrm{y,x,D}\,|\theta}\left[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right] \\ &= \mathrm{E}_{\mathrm{D}\,|\theta}\left[\,\mathrm{E}_{\mathrm{y,x}\,|\theta}\left[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right] \\ &= \mathrm{E}_{\mathrm{D}\,|\theta}\left[\,\mathrm{E}_{\mathrm{x}\,|\theta}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right]\right], \end{aligned} \tag{2.19}$$

where the conditional independence of random element y from the training data D given the model θ is used. As the model θ is not observed, $\mathcal{R}_{\Theta} : \Theta \mapsto {\mathbb{R}_{\geq 0}}^{\mathcal{F}}$ is not a feasible objective function for optimization. This is the fundamental challenge of supervised learning – the true risk objective cannot be evaluated and the designer can never be precisely sure how well any learning function performs.

2.3.1 Clairvoyant Decision

subscript Theta? Use theta sub and remove argument like a cond dist?

It is instructive to formulate the optimal decision function assuming the model θ was in fact observed; it will be referred to as the "clairvoyant" function, following terminology used in [10]. This clairvoyant decision function $f_{\Theta} : \Theta \mapsto \mathcal{F}$ is represented

by

$$f_\Theta(\theta) = \arg\min_{f \in \mathcal{F}} \mathcal{R}_\Theta(f; \theta) \ . \tag{2.20}$$

For a given training set D, the function $f_\Theta(\theta)$ selects the decision function $g \in \mathcal{H}^\mathcal{X}$ that minimizes $\mathrm{E}_{\mathrm{y},\mathrm{x}\,|\theta}\Big[\,\mathcal{L}\left(g(\mathrm{x}), \mathrm{y}\right)\Big]$. Note the conditional independence of (y, x) from D in (2.19) – the knowledge of θ renders the training data D valueless. As such, the range of the clairvoyant function is recast as $f_\Theta : \Theta \mapsto \mathcal{H}^\mathcal{X}$, and

$$f_\Theta(\theta) = \arg\min_{g \in \mathcal{H}^\mathcal{X}} \mathrm{E}_{\mathrm{y},\mathrm{x}\,|\theta}\Big[\,\mathcal{L}\left(g(\mathrm{x}), \mathrm{y}\right)\Big] \ . \tag{2.21}$$

For a given observation, the clairvoyant decision is

$$f_\Theta(\mathrm{x}; \theta) = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h, \mathrm{y})\right] \ . \tag{2.22}$$

The corresponding "irreducible" risk for a given model θ is

$$\begin{aligned} \mathcal{R}^*_\Theta(\theta) &\equiv \mathcal{R}_\Theta\left(f_\Theta(\theta); \theta\right) \\ &= \min_{g \in \mathcal{H}^\mathcal{X}} \mathrm{E}_{\mathrm{y},\mathrm{x}\,|\theta}\Big[\,\mathcal{L}\left(g(\mathrm{x}), \mathrm{y}\right)\Big] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h, \mathrm{y})\right]\right] \ . \end{aligned} \tag{2.23}$$

Additionally, define the excess risk $\mathcal{R}_{\Theta,\mathrm{ex}}(f; \theta) \equiv \mathcal{R}_\Theta(f; \theta) - \mathcal{R}^*_\Theta(\theta)$; minimization of the learning objective (2.19) is equivalent to minimization of this function.

Note that using the marginal/conditional model representations introduced previously, the clairvoyant decision will depend only on the conditional model θ_c.

2.3.2 Bayes Decision

To design an optimal learning function $f \in \mathcal{F}$, an operator must be chosen to remove the dependency of the risk $\mathcal{R}_\Theta$ on θ and form an objective function $\mathcal{F} \mapsto \mathbb{R}_{\geq 0}$. One choice is to integrate over Θ; to ensure a non-negative objective value, the weighting function should be non-negative. Also, as scaling the objective function will not change its minimizing argument, the weighting function can be constrained to integrate to

one. These are the requirements for a valid probability density function (PDF); as such, the model θ is treated as a random process and a Bayesian approach can be adopted.

Define the PDF $\mathrm{p}_\theta \in \mathcal{P}(\Theta)$. Now the Bayes risk can be formulated as

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta \left[\mathcal{R}_\Theta(f;\theta) \right] \\ &= \mathrm{E}_{\mathrm{y,x,D}} \left[\mathcal{L}(f(\mathrm{x};\mathrm{D}),\mathrm{y}) \right] \\ &= \mathrm{E}_\mathrm{D} \left[\mathrm{E}_{\mathrm{y,x}\,|\,\mathrm{D}} \left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right) \right] \right] \\ &= \mathrm{E}_\mathrm{D} \left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}} \left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}} \left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right) \right] \right] \right]\end{aligned} \tag{2.24}$$

and y, x, and D are treated as jointly distributed random elements.

Finally, express the optimal learning function

$$f^* = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f) \ , \tag{2.25}$$

such that

$$f^*(\mathrm{D}) = \arg\min_{g \in \mathcal{H}^\mathcal{X}} \mathrm{E}_{\mathrm{y,x}\,|\,\mathrm{D}} \left[\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right) \right] . \tag{2.26}$$

The decision expressed by the non-parametric learning function $f^*(\mathrm{D})$ for a given novel observation x is

$$f^*(\mathrm{x};\mathrm{D}) = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}} \left[\mathcal{L}(h,\mathrm{y}) \right] \tag{2.27}$$

and the minimum Bayes risk is

$$\begin{aligned}\mathcal{R}^* &\equiv \mathcal{R}(f^*) \\ &= \min_{f \in \mathcal{F}} \mathcal{R}(f) \\ &= \mathrm{E}_\mathrm{D} \left[\min_{g \in \mathcal{H}^\mathcal{X}} \mathrm{E}_{\mathrm{y,x}\,|\,\mathrm{D}} \left[\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right) \right] \right] \\ &= \mathrm{E}_\mathrm{D} \left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}} \left[\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}} \left[\mathcal{L}(h,\mathrm{y}) \right] \right] \right] .\end{aligned} \tag{2.28}$$

2.3.2.1 Model Posteriors

Continuous? sums, PMFs...

Observe that the marginal and conditional Bayesian distributions can be represented as $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{D}}\left[\mathrm{P}_{\mathrm{x}\,|\,\theta}\right] \equiv \mu_{\theta_{\mathrm{m}}\,|\,\mathrm{D}}$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} = \mathrm{E}_{\theta|\,\mathrm{x,D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \equiv \mu_{\theta_{\mathrm{c}}(\mathrm{x})|\,\mathrm{x,D}}$, the expected values of the corresponding clairvoyant distributions with respect to the model posteriors $\mathrm{p}_{\theta_{\mathrm{m}}\,|\,\mathrm{D}}$ and $\mathrm{p}_{\theta_{\mathrm{c}}\,|\,\mathrm{x,D}}$, respectively. The predictive distribution can also be represented as $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} \equiv \mu_{\theta_{\mathrm{c}}\,|\,\mathrm{x,D}}(\mathrm{x};\mathrm{x},\mathrm{D})$ or $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(x,\mathrm{D}) = \mu_{\theta_{\mathrm{c}}(x)|\,\mathrm{x,D}}(x,\mathrm{D})$. Also, $\mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{D}}\left[\mathrm{P}_{\mathrm{y,x}\,|\theta}\right] \equiv \mu_{\theta|\,\mathrm{D}}$. Thus, the Bayesian approach to prediction uses the model posterior given the observable random elements to integrate out the model dependency of the risk $\mathcal{R}_{\Theta}(f;\theta)$.

The relevant posteriors can be represented as

$$\mathrm{p}_{\theta_{\mathrm{m}}\,|\,\mathrm{Y,X}}(\theta_{\mathrm{m}}\,|Y,X) = \frac{\mathrm{E}_{\theta_{\mathrm{c}}\,|\,\theta_{\mathrm{m}}}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta_{\mathrm{c}}}(Y|X,\theta_{\mathrm{c}})\right](\theta_{\mathrm{m}})}{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X)} \frac{\mathrm{P}_{\mathrm{X}\,|\,\theta_{\mathrm{m}}}(X|\,\theta_{\mathrm{m}})}{\mathrm{P}_{\mathrm{X}}(X)}\,\mathrm{p}_{\theta_{\mathrm{m}}}(\theta_{\mathrm{m}}) \tag{2.29}$$

and

$$\mathrm{p}_{\theta_{\mathrm{c}}\,|\,\mathrm{Y,X,x}}(\theta_{\mathrm{c}}\,|Y,X,x) = \frac{\mathrm{E}_{\theta_{\mathrm{m}}\,|\,\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{X,x}\,|\,\theta_{\mathrm{m}}}(X,x|\,\theta_{\mathrm{m}})\right](\theta_{\mathrm{c}})}{\mathrm{P}_{\mathrm{X,x}}(X,x)} \frac{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta_{\mathrm{c}}}(Y|X,\theta_{\mathrm{c}})}{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X,x}}(Y|X,x)}\,\mathrm{p}_{\theta_{\mathrm{c}}}(\theta_{\mathrm{c}}) \tag{2.30}$$

where $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}} = \mathrm{E}_{\theta_{\mathrm{m}}\,|\,\mathrm{X}}\left[\mathrm{E}_{\theta_{\mathrm{c}}\,|\,\theta_{\mathrm{m}}}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta_{\mathrm{c}}}\right]\right]$ and $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X,x}} = \mathrm{E}_{\theta_{\mathrm{m}}\,|\,\mathrm{X,x}}\left[\mathrm{E}_{\theta_{\mathrm{c}}\,|\,\theta_{\mathrm{m}}}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta_{\mathrm{c}}}\right]\right]$.

marginal/conditional independence discuss? x independence!

Additionally, since $\Psi(\mathrm{D})$ is a sufficient statistic for the model θ, the Bayesian distributions of interest P_{D}, $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ will also depend on D only through $\Psi(\mathrm{D})$; as such, the training data can be transformed into the empirical process ψ for Bayesian prediction without incurring any additional risk.

For this approach, the distributions P_{ψ}, $\mathrm{P}_{\mathrm{x}\,|\psi}$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi}$ are required. Note that $\mathrm{P}_{\mathrm{D}\,|\theta}(D|\theta) = \mathcal{M}\left(N\Psi(D)\right)^{-1}\mathrm{P}_{\psi|\theta}\left(\Psi(D)|\theta\right)$. Also, observe that the relevant posterior distributions satisfy $\mathrm{p}_{\theta_{\mathrm{m}}\,|\,\mathrm{D}} = \mathrm{p}_{\theta_{\mathrm{m}}\,|\psi}\left(\Psi(\mathrm{D})\right)$ and $\mathrm{p}_{\theta_{\mathrm{c}}(x)|\,\mathrm{x,D}} = \mathrm{p}_{\theta_{\mathrm{c}}(x)|\,\mathrm{x},\psi}\left(\mathrm{x},\Psi(\mathrm{D})\right)$, due to the sufficiency of the statistic $\Psi(\mathrm{D})$ [2]; consequently, $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} = \mathrm{P}_{\mathrm{x}\,|\psi}\left(\Psi(\mathrm{D})\right)$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi}\left(\mathrm{x},\Psi(\mathrm{D})\right)$. Also, $\mathrm{p}_{\theta|\,\mathrm{D}} = \mathrm{p}_{\theta|\psi}\left(\Psi(\mathrm{D})\right)$ and thus $\mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{D}} = \mathrm{P}_{\mathrm{y,x}\,|\psi}\left(\Psi(\mathrm{D})\right)$.

The posteriors can be represented as

$$\mathrm{p}_{\theta_\mathrm{m}\,|\,\psi_\mathrm{c},\psi_\mathrm{m}}(\theta_\mathrm{m}\,|\,\psi_\mathrm{c},\psi_\mathrm{m}) = \frac{\mathrm{E}_{\theta_\mathrm{c}\,|\,\theta_\mathrm{m}}\left[\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\theta_\mathrm{c}}\left(\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\theta_\mathrm{c}\right)\right](\theta_\mathrm{m})}{\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m}}\left(\psi_\mathrm{c}\,|\,\psi_\mathrm{m}\right)}\frac{\mathrm{P}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left(\psi_\mathrm{m}\,|\,\theta_\mathrm{m}\right)}{\mathrm{P}_{\psi_\mathrm{m}}\left(\psi_\mathrm{m}\right)}\,\mathrm{p}_{\theta_\mathrm{m}}(\theta_\mathrm{m}) \tag{2.31}$$

and

$$\mathrm{p}_{\theta_\mathrm{c}\,|\,\psi_\mathrm{c},\psi_\mathrm{m},\mathrm{x}}(\theta_\mathrm{c}\,|\,\psi_\mathrm{c},\psi_\mathrm{m},x) = \frac{\mathrm{p}_{\theta_\mathrm{m}\,|\,\theta_\mathrm{c}}\left[\mathrm{P}_{\psi_\mathrm{m},\mathrm{x}\,|\,\theta_\mathrm{m}}\left(\psi_\mathrm{m},x|\,\theta_\mathrm{m}\right)\right](\theta_\mathrm{c})}{\mathrm{P}_{\psi_\mathrm{m},\mathrm{x}}\left(\psi_\mathrm{m},x\right)}\frac{\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\theta_\mathrm{c}}\left(\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\theta_\mathrm{c}\right)}{\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\mathrm{x}}\left(\psi_\mathrm{c}\,|\,\psi_\mathrm{m},x\right)}\,\mathrm{p}_{\theta_\mathrm{c}}(\theta_\mathrm{c}) \tag{2.32}$$

where $\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m}} = \mathrm{E}_{\theta_\mathrm{m}\,|\,\psi_\mathrm{m}}\left[\mathrm{E}_{\theta_\mathrm{c}\,|\,\theta_\mathrm{m}}\left[\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\theta_\mathrm{c}}\right]\right]$ and $\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\mathrm{x}} = \mathrm{E}_{\theta_\mathrm{m}\,|\,\psi_\mathrm{m},\mathrm{x}}\left[\mathrm{E}_{\theta_\mathrm{c}\,|\,\theta_\mathrm{m}}\left[\mathrm{P}_{\psi_\mathrm{c}\,|\,\psi_\mathrm{m},\theta_\mathrm{c}}\right]\right]$.

conditional EP independence? product notation?

2.3.3 Risk trends

Location???

Prior support? Must be bounded??

"Universal consistency"

FIXXXXXX lim

The trends in risk as $N \to \infty$ are of specific interest. Recall that $\mathrm{P}_{\psi|\theta}(\psi|\theta) \to \delta[\psi,\theta]$. For priors with full support, observe that as the number of training samples increases, the statistic PMF tends toward $\mathrm{P}_\psi(\psi) \approx N^{1-|\mathcal{Y}||\mathcal{X}|}\,\mathrm{p}_\theta(\psi)$; this can be proven using Gautschi's inequality [20]. Thus,

$$\begin{aligned}\mathrm{p}_{\theta|\psi}(\theta|\psi) &= \frac{\mathrm{P}_{\psi|\theta}(\psi|\theta)}{\mathrm{P}_\psi(\psi)}\,\mathrm{p}_\theta(\theta)\\ &\to \delta(\theta-\psi)\end{aligned} \tag{2.33}$$

and $\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\psi} = \mu_{\theta|\psi} \to \psi$.

Representing $\psi|\theta \xrightarrow{p} \theta$ as $\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\psi}\,|\theta \xrightarrow{p} \mathrm{P}_{\mathrm{y},\mathrm{x}\,|\theta}$, it is clear that $f^*(\mathrm{D})|\theta \xrightarrow{p} f_\Theta(\theta)$, such

that the clairvoyant learner is precisely identified, achieving

$$\begin{aligned}\mathcal{R}_\Theta(f;\theta) &= \mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathcal{L}\left(f^*(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right] \\ &= \mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\theta}\left[\mathcal{L}\left(f_\Theta(\mathrm{x};\theta),\mathrm{y}\right)\right] \\ &= \mathcal{R}^*_\Theta(\theta)\ . \end{aligned} \tag{2.34}$$

This demonstrates the consistency of the full-support Bayesian learner.

The irreducible risk (2.23) for a given model satisfies $\mathcal{R}^*_\Theta(\theta) \leq \mathcal{R}_\Theta(f;\theta) \quad \forall f \in \mathcal{F},\ \theta \in \Theta$. Consequently, the Bayes risk satisfies $\mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] \leq \mathcal{R}(f) \quad \forall f \in \mathcal{F}$. Note that this inequality holds for any number of training samples N and that the lower bound does not depend on N. Thus, even with unlimited training data, no learning function can provide a Bayes risk lower than this value.

2.4 Predictive Model Estimation

bias-variance CITE??

It is instructive to treat the Bayesian predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ as an estimator of the clairvoyant predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \theta_\mathrm{c}(\mathrm{x})$ and investigate the effects of prior knowledge. Of specific interest are the mean $\mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right]$ and covariance $\mathrm{C}_{\mathrm{D}\,|\,\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right]$ with respect to the true model θ.

just in terms of normal theta?

To aid characterization of the estimator, define the random process $\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c}) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}} \in \mathbb{R}^{\mathcal{Y}}$. The expected bias of the estimator is thus

$$\begin{aligned}\mathrm{Bias}(\mathrm{x};\theta_\mathrm{m},\theta_\mathrm{c}) &= \mathrm{E}_{\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c})\right] \\ &= \mathrm{E}_{\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\ , \end{aligned} \tag{2.35}$$

measuring the difference between the clairvoyant distribution and the expected Bayesian distribution.

Defining the covariance of the Bayesian predictive distribution as

$$\mathrm{Cov}(\mathrm{x};\theta_\mathrm{m},\theta_\mathrm{c}) = \mathrm{C}_{\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] \in \mathbb{R}^{+\mathcal{Y}\times\mathcal{Y}}\ , \tag{2.36}$$

the conditional second moments of $\Delta(\mathrm{x}; \mathrm{D}, \theta_\mathrm{c})$ can be readily shown to be

$$\begin{aligned} &\mathrm{E}_{\mathrm{D} \mid \theta_\mathrm{m}, \theta_\mathrm{c}} \left[\Delta(\mathrm{x}; \mathrm{D}, \theta_\mathrm{c}) \otimes \Delta(\mathrm{x}; \mathrm{D}, \theta_\mathrm{c}) \right] \\ &\quad = \mathrm{Bias}(\mathrm{x}; \theta_\mathrm{m}, \theta_\mathrm{c}) \otimes \mathrm{Bias}(\mathrm{x}; \theta_\mathrm{m}, \theta_\mathrm{c}) + \mathrm{Cov}(\mathrm{x}; \theta_\mathrm{m}, \theta_\mathrm{c}) \end{aligned} \tag{2.37}$$

over the domain $\mathcal{Y} \times \mathcal{Y}$.

2.5 Applications to Common Loss Functions

marginal/conditional??? D or psi? Continuous?

In this section, loss functions typical for classification and regression applications, specifically the 0-1 loss function and the squared error loss function, are adopted. The risk (2.19) is assessed, clairvoyant decision functions (2.22) are found, and the irreducible risk (2.23) is expressed.

2.5.1 Regression: 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 a closed-form expression of the minimizing estimation function.

It is assumed that the unobserved random element y is a scalar random variable; that is, $\mathcal{Y} \subseteq \mathbb{R}$. Additionally, the estimator's output is allowed to assume real numbers; thus, $\mathcal{H} = \mathbb{R} \supseteq \mathcal{Y}$.

Restrict estimate to discrete values? Rounding? Discuss, at least.

The loss function is defined as

$$\mathcal{L}(h, y) = (h - y)^2 \, . \tag{2.38}$$

Substituting the squared error loss into (2.19), the squared error risk is

$$\begin{aligned}\mathcal{R}_\Theta(f;\theta) &= \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y,x}\,|\,\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\big)^2\right]\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\big)^2\right]\right]\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(\mathrm{y}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})^2\right]\right] + \mathrm{E}_{\mathrm{x,D}\,|\,\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x,D}\,|\,\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\right],\end{aligned} \tag{2.39}$$

a sum of two terms. The first term is the expected conditional variance of the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$. The second term is the expected squared bias between the estimate and the true conditional mean $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$.

Note that the risk can also be represented as $\mathcal{R}_\Theta(f;\theta) = \mathcal{R}^*_\Theta(\theta) + \mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta)$, where the first term is the irreducible squared error (as demonstrated in the next sub-section) and the second term is the excess squared error,

$$\begin{aligned}\mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta) &= \mathrm{E}_{\mathrm{x,D}\,|\,\theta}\left[\big(f(\mathrm{x};\mathrm{D})-f_\Theta(\mathrm{x};\theta)\big)^2\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\Big(\mathrm{E}_{\mathrm{D}\,|\,\theta}\left[f(\mathrm{x};\mathrm{D})\right]-f_\Theta(\mathrm{x};\theta)\Big)^2 + \mathrm{C}_{\mathrm{D}\,|\,\theta}\left[f(\mathrm{x};\mathrm{D})\right]\right],\end{aligned} \tag{2.40}$$

where $f_\Theta(\mathrm{x};\theta)$ is the clairvoyant estimator. Observe that the excess risk can be further decomposed as a sum of the estimator's expected bias and variance, respectively.

additional bias-variance trade-off discussion?

2.5.1.1 Clairvoyant Estimation

plots?

To find the clairvoyant estimator, the squared error loss is substituted into (2.22); note that the objective function is quadratic over the argument $h \in \mathcal{H} = \mathbb{R}$. 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 clairvoyant estimate is the expected value of y given the model θ and the

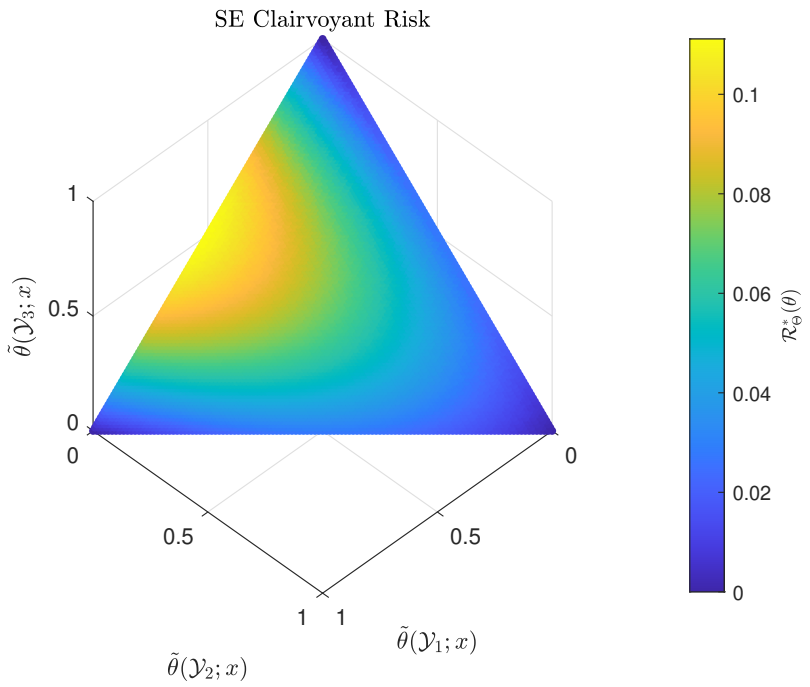

Figure 2.1: Irreducible Squared-Error, constant $\theta_c(x)$

observed value x, such that

$$f_\Theta(\mathrm{x};\theta) = \arg\min_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(h-\mathrm{y})^2\right] \qquad (2.41)$$
$$= \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\ .$$

Substituting the loss and clairvoyant function into (2.23), the irreducible squared error is

$$\mathcal{R}^*_\Theta(\theta) = \mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(\mathrm{y}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})^2\right]\right] \qquad (2.42)$$
$$= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\ .$$

Observe that the general risk (2.39) can be represented as $\mathcal{R}_\Theta(f;\theta) = \mathcal{R}^*_\Theta(\theta) + \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(f(\mathrm{x};\mathrm{D}) - f_\Theta(\mathrm{x};\theta)\right)^2\right]$. The first summand is equal to the irreducible squared error; the second term is dependent on the difference between the general estimate and the clairvoyant estimate.

Fig. 2.1 displays the irreducible risk for predictive models $\theta_c(x)$ independent of x.

conditional location?

2.5.1.2 Bayesian Estimation

Optimal Estimate: the Posterior Mean To find the optimal estimator, the squared error loss is substituted into (2.27). Again, 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

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[(h-\mathrm{y})^2\right] \\ &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \ . \end{aligned} \tag{2.43}$$

An interesting form for the optimal estimator is $f^*(\mathrm{x};\mathrm{D}) = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}}\left[f_\Theta(\mathrm{x};\theta)\right]$. Substituting the squared error loss into the second line of (2.27), the optimal Bayes estimator is the conditional expected value of the clairvoyant estimate with respect to the model posterior distribution.

Minimum Bayes Risk: the Expected Posterior Variance The Bayes squared error risk for a general learning function is

$$\begin{aligned} \mathcal{R}(f) &= \mathrm{E}_\theta\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{y},\mathrm{x}\,|\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\big)^2\right]\right]\right] \\ &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\big)^2\right]\right] \\ &= \mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] + \mathrm{E}_{\mathrm{x},\mathrm{D},\theta}\left[\big(f(\mathrm{x};\mathrm{D})-f_\Theta(\mathrm{x};\theta)\big)^2\right] \\ &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\big)^2\right] . \end{aligned} \tag{2.44}$$

Substituting the optimal estimator (2.43) into Equation (2.44), the minimum Bayes risk is the expected conditional variance

$$\begin{aligned} \mathcal{R}^* &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] \\ &= \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{C}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\right] \\ &= \mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{C}_{\theta|\,\mathrm{x},\mathrm{D}}\left[f_\Theta(\mathrm{x};\theta)\right]\right] . \end{aligned} \tag{2.45}$$

The first term is the expected irreducible risk. The second term is the expected variance of the clairvoyant estimate $f_\Theta(\mathrm{x};\theta) = \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ with respect to the model posterior PDF $\mathrm{p}_{\theta|\,\mathrm{x},\mathrm{D}}$.

2.5.1.3 Squared-Error

Rename section

Substituting in the clairvoyant and Bayesian estimators, the excess squared error (2.40) can be represented as

$$\begin{aligned}\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right] \\ &\equiv \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y'\,\mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\Delta(y;\mathrm{x};\mathrm{D},\theta_\mathrm{c})\Delta(y';\mathrm{x};\mathrm{D},\theta_\mathrm{c})\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\left(\mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2 + \mathrm{C}_{\mathrm{D}\,|\,\theta}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right]\right] \\ &\equiv \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\left(\sum_{y\in\mathcal{Y}} y\ \mathrm{Bias}(y;\mathrm{x};\theta_\mathrm{m},\theta_\mathrm{c})\right)^2 + \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y'\ \mathrm{Cov}(y,y';\mathrm{x};\theta_\mathrm{m},\theta_\mathrm{c})\right]\ ,\end{aligned} \tag{2.46}$$

where the formulae for the predictive distribution bias and variance from Section 2.4 have been used.

2.5.2 Classification: the 0-1 Loss

In this section, the developed framework is applied to a common machine learning task: classification. In classification problems, the set $\mathcal{Y}$ is countable and typically finite. Furthermore, the hypothesis space is usually identical to the unobserved variable space; that is, $\mathcal{H} = \mathcal{Y}$. The 0-1 loss function is the most widely used for these problems; it is represented as

$$\mathcal{L}(h,y) = 1 - \delta[h,y]\ . \tag{2.47}$$

Applying the 0-1 loss, the risk (2.19) for a general classifier is

$$\mathcal{R}_\Theta(f;\theta) = 1 - \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left(f(\mathrm{x};\mathrm{D})|\,\mathrm{x},\theta\right)\right]\right]\ . \tag{2.48}$$

2.5.2.1 Clairvoyant Hypothesis

decision region figures??

To find the clairvoyant classifier, the 0-1 loss is substituted into (2.22); given an observation x, the optimum hypothesis is simply the value y that maximizes the conditional model $\theta_c(x)$,

$$\begin{aligned} f_\Theta(\mathrm{x};\theta) &= \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[1-\delta[h,y]\right] \\ &= \arg\max_{h\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(h|\,\mathrm{x},\theta) \\ &= \arg\max_{y\in\mathcal{Y}} \theta(y,\mathrm{x})\ . \end{aligned} \tag{2.49}$$

Substituting the 0-1 loss and clairvoyant hypothesis into (2.23), the resulting irreducible probability of error is

$$\mathcal{R}^*_\Theta(\theta) = 1 - \mathrm{E}_{\mathrm{x}\,|\theta}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(y|\,\mathrm{x},\theta)\right]\ . \tag{2.50}$$

Fig. 2.2 displays the irreducible risk for predictive models $\theta_c(x)$ independent of x. Intuitively, the models that are more concentrated lead to lower probability of error.

conditional location?

2.5.2.2 Bayesian Classification

Optimal Hypothesis: Conditional Maximum *a posteriori* To determine the optimal learning function, the 0-1 loss from Equation (2.47) is substituted into Equation (2.27) to find

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[1-\delta[h,\mathrm{y}]\right] \\ &= \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\ . \end{aligned} \tag{2.51}$$

The optimal classifier chooses the value $y \in \mathcal{Y}$ that maximizes the conditional PMF for the observed values of x and D.

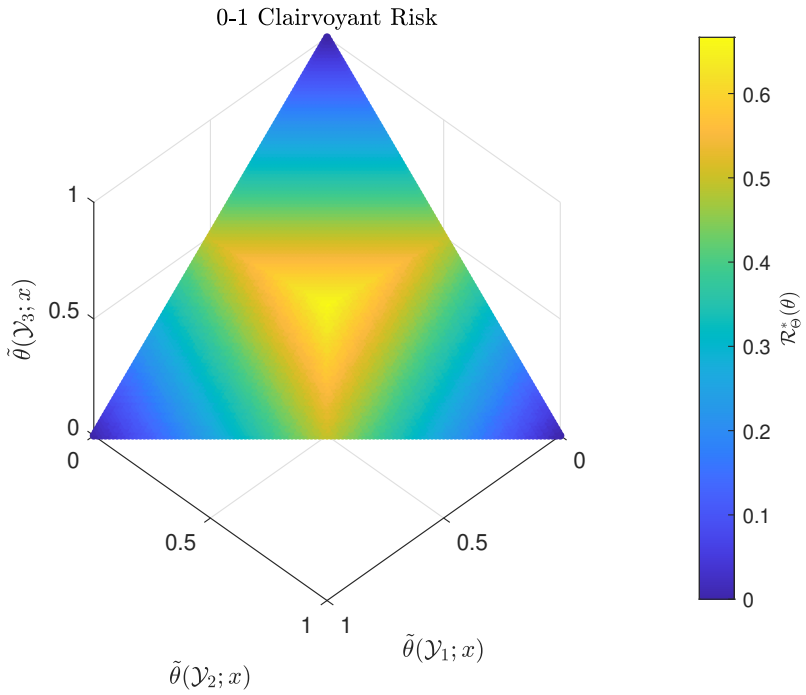

Figure 2.2: Irreducible probability of error, constant $\theta_{\mathrm{c}}(x)$

Minimum Bayes Risk: Probability of Error Using the 0-1 loss, the Bayes probability of error (2.28) is

$$\mathcal{R}(f) = 1 - \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}\left(f(\mathrm{x};\mathrm{D})|\,\mathrm{x},\mathrm{D}\right)\right] \ . \tag{2.52}$$

Substituting the optimal learning function (2.51) into the general risk (2.52), the minimum probability of error is

$$\mathcal{R}^* = 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \ . \tag{2.53}$$

Chapter 3

Discrete-Domain Dirichlet Model

Generalize discussion/math for infinite-countable domains?

Figs BROKEN from alpha changes

Uncoupled m/c alphas = non-Dir Ptheta??

This chapter determines the optimal learning functions when the sets $\mathcal{Y}$ and $\mathcal{X}$ have a finite number of elements and the model θ is characterized by a Dirichlet distribution.

3.1 Probability Distributions

To determine the optimal learning function, the joint PMF $P_{y,x,D}$ is required. Having already defined the distribution conditioned on the model θ, all that remains is to select a PDF p_θ reflecting the user's prior knowledge. In this section, the Dirichlet distribution is used. The Dirichlet distribution possesses the desirable property of being the conjugate prior for the multinomial conditional distribution characterizing the data; as such, it will provide analytic forms for the model posterior distribution and lead to closed form expressions for the data conditional distribution used to design the learning function.

Other distributions of interest will be provided, such as the training data PMF

$\mathrm{P_D}$ and the conditional distribution $\mathrm{P_{y\,|\,x,D}}$ used to form a decision given specific observations.

3.1.1 Model PDF, $\mathrm{p_\theta}$

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

$$\begin{aligned} \mathrm{p_\theta}(\theta) &= \beta(\alpha_0\alpha)^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha_0\alpha(y,x)-1} \\ &= \mathrm{Dir}\left(\theta;\alpha_0,\alpha\right) , \end{aligned} \tag{3.1}$$

where the user-selected parameterizing distribution $\alpha \in \left\{\mathbb{R}^{+\mathcal{Y}\times\mathcal{X}} : \sum_{y,x}\alpha(y,x) = 1\right\} \subset \Theta$ and concentration parameter $\alpha_0 \in \mathbb{R}^+$ are introduced. Note that β is the generalized beta function.

Use of alpha is REDUNDANT?! Just use mu??????????????

The first and second joint moments of the model are

$$\mu_\theta = \alpha \tag{3.2}$$

and

$$\mathrm{E_\theta}\left[\theta\otimes\theta\right] = \frac{1}{\alpha_0+1}\,\mathrm{diag}(\alpha) + \frac{\alpha_0}{\alpha_0+1}\alpha\otimes\alpha\ . \tag{3.3}$$

Observe that $\mathrm{P_{y,x}} = \mu_\theta = \alpha$. The covariance is

$$\begin{aligned} \Sigma_\theta &= \mathrm{E_\theta}\left[\left(\theta-\mu_\theta\right)\otimes\left(\theta-\mu_\theta\right)\right] \\ &= \frac{\mathrm{diag}(\alpha)-\alpha\otimes\alpha}{\alpha_0+1}\ . \end{aligned} \tag{3.4}$$

Also, for PDF's satisfying $\alpha(y,x) > \alpha_0^{-1}$, the maximizing value of the distribution is

$$\theta_{\max} = \arg\max_{\theta\in\Theta} \mathrm{p_\theta}(\theta) = \frac{\alpha-\alpha_0^{-1}}{1-\alpha_0^{-1}|\,\mathcal{Y}\,||\,\mathcal{X}\,|}\ . \tag{3.5}$$

This can be easily shown by maximizing the logarithm of the distribution using the method of Lagrange multipliers.

Reference? Mode proof for all values of alpha?

HALDANE PRIOR

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

$$\mathrm{p}_\theta(\theta) \to \delta\left(\theta - \alpha\right) \ . \tag{3.6}$$

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

$$\mathrm{p}_\theta(\theta) \to \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \alpha(y,x) \delta\big(\theta - \delta[\cdot, y]\delta[\cdot, x]\big) \ , \tag{3.7}$$

which distributes its weight among the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models with an ℓ_0 norm satisfying $\|\theta\|_0 = 1$. Note that the Dirac delta for these formulas is defined on the set Θ, such that $\int_\Theta \delta(\theta) \mathrm{d}\theta = 1$.

formal proof for limiting PDFs??? stirling/gautschi?

These trends are demonstrated with Fig. 3.1. The cardinalities $|\,\mathcal{Y}\,| = 3$ and $|\,\mathcal{X}\,| = 1$ are chosen to enable visualization, despite the implication that x is deterministic; these cardinalities will be used for many subsequent figures as well. Note that for $\alpha_0 = 2.99 < |\,\mathcal{Y}\,||\,\mathcal{X}\,|$, the PDF values at the boundaries of the domain tend to infinity; this is not captured by the plot color scale.

Uniform Prior When the Dirichlet parameters are $\alpha(y,x) = \big(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\big)^{-1}$ and $\alpha_0 = |\,\mathcal{Y}\,||\,\mathcal{X}\,|$, the distribution becomes a uniform PDF and is represented as

$$\mathrm{p}_\theta = \big(|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\big)! \ . \tag{3.8}$$

3.1.1.1 Marginal and Conditional Distributions

PGR: move/add Dir figs here?

The marginal distribution θ_m and the conditional distribution θ_c are also of interest. For brevity, introduce the bijection $\alpha \Leftrightarrow (\alpha_\mathrm{m}, \alpha_\mathrm{c})$, where $\alpha_\mathrm{m} \equiv \sum_{y \in \mathcal{Y}} \alpha(y, \cdot)$, and $\alpha_\mathrm{c}(x) \equiv \alpha(\cdot, x) / \,\alpha_\mathrm{m}(x)$ for each $x \in \mathcal{X}$. Observe that $\alpha_\mathrm{m} \in \mathcal{P}(\mathcal{X})$ and $\alpha_\mathrm{c} \in \mathcal{P}(\mathcal{Y})^\mathcal{X}$.

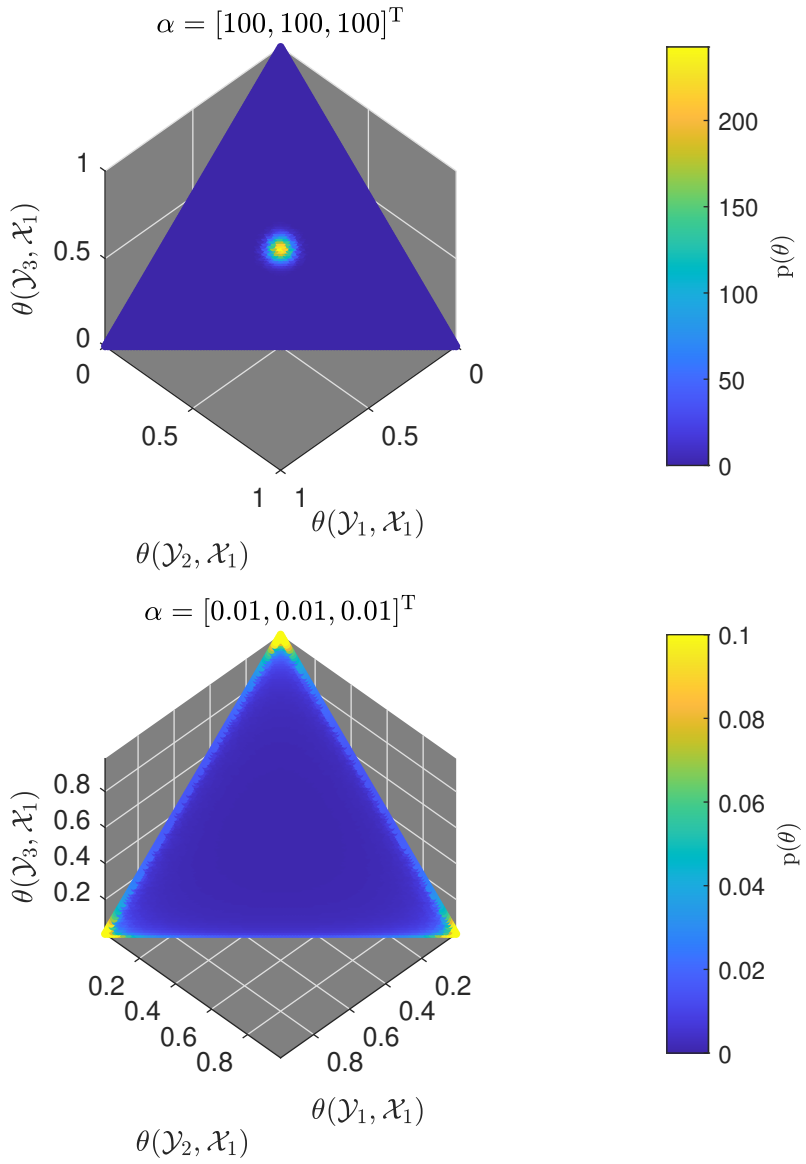

Figure 3.1: Model prior PDF for different concentrations α_0

By the aggregation property [7], $\Theta_{\mathrm{m}} \sim \mathrm{Dir}(\alpha_0, \alpha_{\mathrm{m}})$ is a Dirichlet random process parameterized by concentration α_0 and distribution α_{m}; observe that $\mathrm{P}_{\mathrm{x}} = \mu_{\Theta_{\mathrm{m}}} = \alpha_{\mathrm{m}}$. Also of interest is the distribution of the predictive model Θ_{c} conditioned on the marginal Θ_{m}. As demonstrated in Appendix A.2, these random processes are jointly distributed as

$$\begin{aligned} \mathrm{p}_{\Theta_{\mathrm{c}} \,|\, \Theta_{\mathrm{m}}}(\theta_{\mathrm{c}} \,|\, \theta_{\mathrm{m}}) &= \mathrm{p}_{\Theta_{\mathrm{c}}}(\theta_{\mathrm{c}}) \\ &= \prod_{x \in \mathcal{X}} \left[\beta\big(\alpha_0\, \alpha_{\mathrm{m}}(x)\, \alpha_{\mathrm{c}}(x)\big)^{-1} \prod_{y \in \mathcal{Y}} \theta_{\mathrm{c}}(y;x)^{\alpha_0\, \alpha_{\mathrm{m}}(x)\, \alpha_{\mathrm{c}}(y;x) - 1} \right] \\ &= \prod_{x \in \mathcal{X}} \mathrm{Dir}\left(\theta_{\mathrm{c}}(x); \alpha_0\, \alpha_{\mathrm{m}}(x), \alpha_{\mathrm{c}}(x) \right) , \end{aligned} \tag{3.9}$$

$$\Theta_{\mathrm{c}} \,|\, \Theta_{\mathrm{m}} \sim \Theta_{\mathrm{c}} \sim \bigotimes_{x \in \mathcal{X}} \mathrm{Dir}\left(\alpha_0\, \alpha_{\mathrm{m}}(x), \alpha_{\mathrm{c}}(x) \right) , \tag{3.10}$$

a product of Dirichlet distributions defined on $\theta_{\mathrm{c}} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$. As shown, the processes $\Theta_{\mathrm{c}}(x)$ are Dirichlet with parameterizing distributions $\alpha_{\mathrm{c}}(x)$ and concentrations $\alpha_0\, \alpha_{\mathrm{m}}(x)$, independent of one another, and independent of the marginal distribution Θ_{m}. Note that $\mathrm{P}_{\mathrm{y} \,|\, \mathrm{x}} = \mu_{\Theta_{\mathrm{c}}}(\mathrm{x}) = \alpha_{\mathrm{c}}(\mathrm{x})$.

3.1.2 Training Set PMF, P_{D}

EVIDENCE TERMINOLOGY??

Next, the conditional distribution $\mathrm{P}_{\mathrm{D} \,|\, \Theta}$ will be used to determine the marginal PMF, P_{D} and properties will be discussed.

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 [12]

$$\begin{aligned} \mathrm{P}_{\mathrm{D}}(D) &= \mathrm{E}_{\Theta} \left[\prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n \,|\, \Theta}\left(D_n | \Theta\right) \right] \\ &= \mathrm{E}_{\Theta} \left[\left(\prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y,x)^{\Psi(y,x;D)} \right)^{N} \right] \\ &= \frac{\beta\big(\alpha_0 \alpha + N \Psi(D)\big)}{\beta(\alpha_0 \alpha)} \, . \end{aligned} \tag{3.11}$$

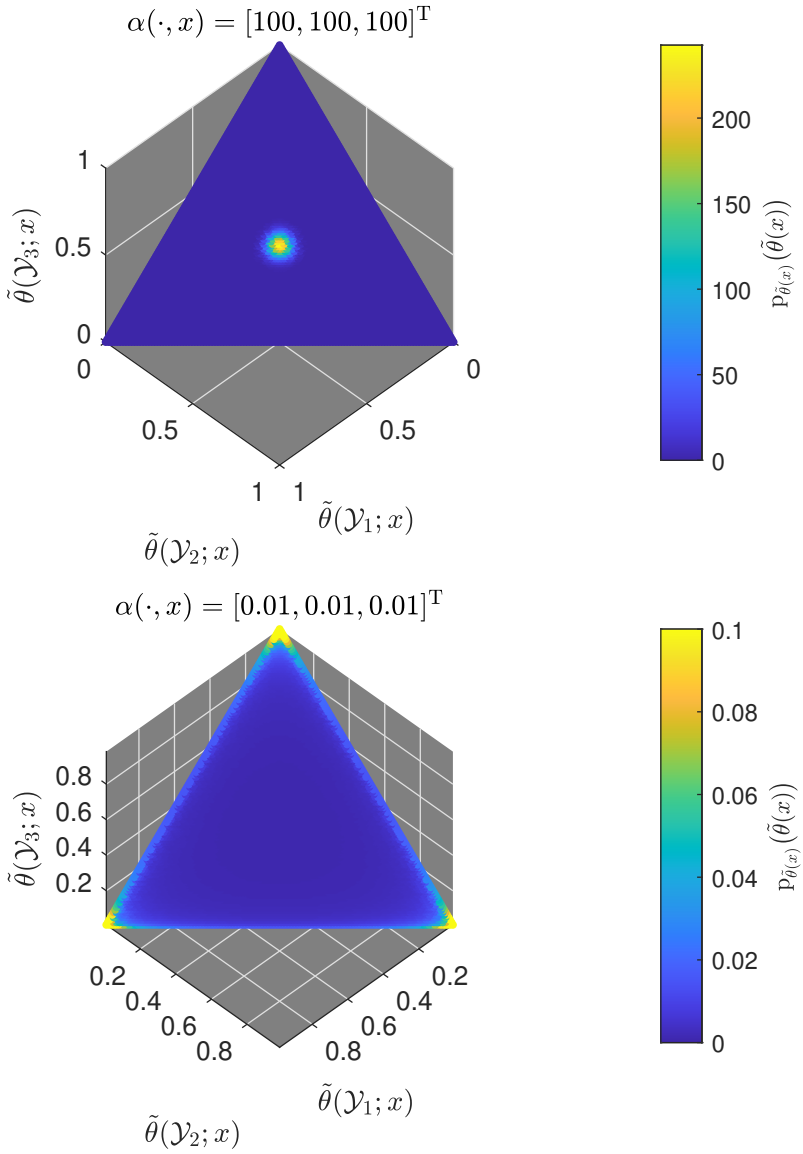

Figure 3.2: Model prior PDF for different concentrations $\alpha_0\,\alpha_{\mathrm{m}}(x)$

Note that values of the PMF $\mathrm{P_D}$ are equivalent to joint moments of the model θ.

It is instructive to consider the limiting forms of this distribution for the extreme values of the model concentration parameter α_0. As $\alpha_0 \to \infty$, the model concentrates at its mean and the training data D distribution is

$$\begin{aligned} \mathrm{P_D}(D) &\to \mathrm{E}_\theta \left[\prod_{n=1}^{N} \theta(Y_n, X_n) \right] \\ &= \prod_{n=1}^{N} \alpha\big(Y_n, X_n\big) \ . \end{aligned} \tag{3.12}$$

Conversely, as $\alpha_0 \to 0$, the distribution becomes

$$\mathrm{P_D}(D) \to \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \alpha(y,x) \prod_{n=1}^{N} \delta\big[D_n, (y,x)\big] \tag{3.13}$$

and the training data are identical.

Next, the distribution of the sufficient statistic ψ will be represented. As a Dirichlet distribution characterizes the parameters of the Empirical distribution $\mathrm{P}_{\psi_\mathrm{m}|\theta}$, the PMF of ψ is a Dirichlet-Empirical distribution (related to the Dirichlet-Multinomial distribution [9]) for N samples, concentration α_0, and parameter distribution α, such that

$$\begin{aligned} \mathrm{P}_\psi(\psi) &= \mathcal{M}(N\psi) \frac{\beta(\alpha_0 \alpha + N\psi)}{\beta(\alpha_0 \alpha)} \\ &= \mathrm{DE}\left(\psi; N, \alpha_0, \alpha\right) \ . \end{aligned} \tag{3.14}$$

The first and second joint moments of the empirical model ψ are

$$\mu_\psi = \alpha = \mu_\theta \tag{3.15}$$

and

$$\mathrm{E}_\psi\left[\psi \otimes \psi\right] = \frac{\alpha_0^{-1} + N^{-1}}{1 + \alpha_0^{-1}} \, \mathrm{diag}(\alpha) + \frac{1 - N^{-1}}{1 + \alpha_0^{-1}} \alpha \otimes \alpha \ .$$

The covariance function is

$$\begin{aligned} \Sigma_\psi &= \frac{\alpha_0^{-1} + N^{-1}}{1 + \alpha_0^{-1}} \big(\mathrm{diag}(\alpha) - \alpha \otimes \alpha \big) \\ &= \left(1 + \frac{\alpha_0}{N}\right) \Sigma_\theta \ . \end{aligned} \tag{3.16}$$

Observe that as $N \to \infty$, the variance $\Sigma_\psi \to \Sigma_\theta$.

Observe that as the number of training samples increases, the statistic PMF tends toward $\mathrm{P}_\psi(\psi) \approx N^{1-|\mathcal{Y}||\mathcal{X}|}\,\mathrm{p}_\theta(\psi)$; this can be proven using Gautschi's inequality [20]. Fig. 3.3 shows how a specific model prior influences the data PMF differently for different N.

Again, the data PMF's for minimal and maximal concentration α_0 are relevant. For $\alpha_0 \to \infty$, the model PDF p_θ concentrates at its mean, α, and thus ψ is characterized by an Empirical distribution,

$$\mathrm{P}_\psi(\psi) \to \mathcal{M}(N\psi) \left(\prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \alpha(y,x)^{\psi(y,x)} \right)^N \tag{3.17}$$

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

$$\mathrm{P}_\psi(\psi) \to \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \alpha(y,x) \delta\big[\psi, \delta[\cdot, y]\delta[\cdot, x]\big] \ . \tag{3.18}$$

formal proofs for limiting PMFs? stirling/gautschi?

Fig. 3.4 displays example distributions of ψ for $N = 10$ and different model concentrations α_0. Observe that for large α_0, the distribution approaches an Empirical distribution $\psi \sim \mathrm{Emp}(N, \alpha)$.

Uniform Prior For the uniform prior distribution, $\alpha(y,x) = \big(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\big)^{-1}$ and $\alpha_0 = |\,\mathcal{Y}\,||\,\mathcal{X}\,|$,

$$\mathrm{P}_\mathrm{D}(D) = \mathcal{M}\left((N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1)\right)^{-1} \mathcal{M}\left(N\Psi(D)\right)^{-1} \tag{3.19}$$

and

$$\mathrm{P}_\psi = |\Psi|^{-1} = \mathcal{M}\left((N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1)\right)^{-1} . \tag{3.20}$$

The distribution of ψ is uniform over the set Ψ. The PMF for D depends on the training data only through the multinomial coefficient; consequently, more "concentrated" training sets are more probable.

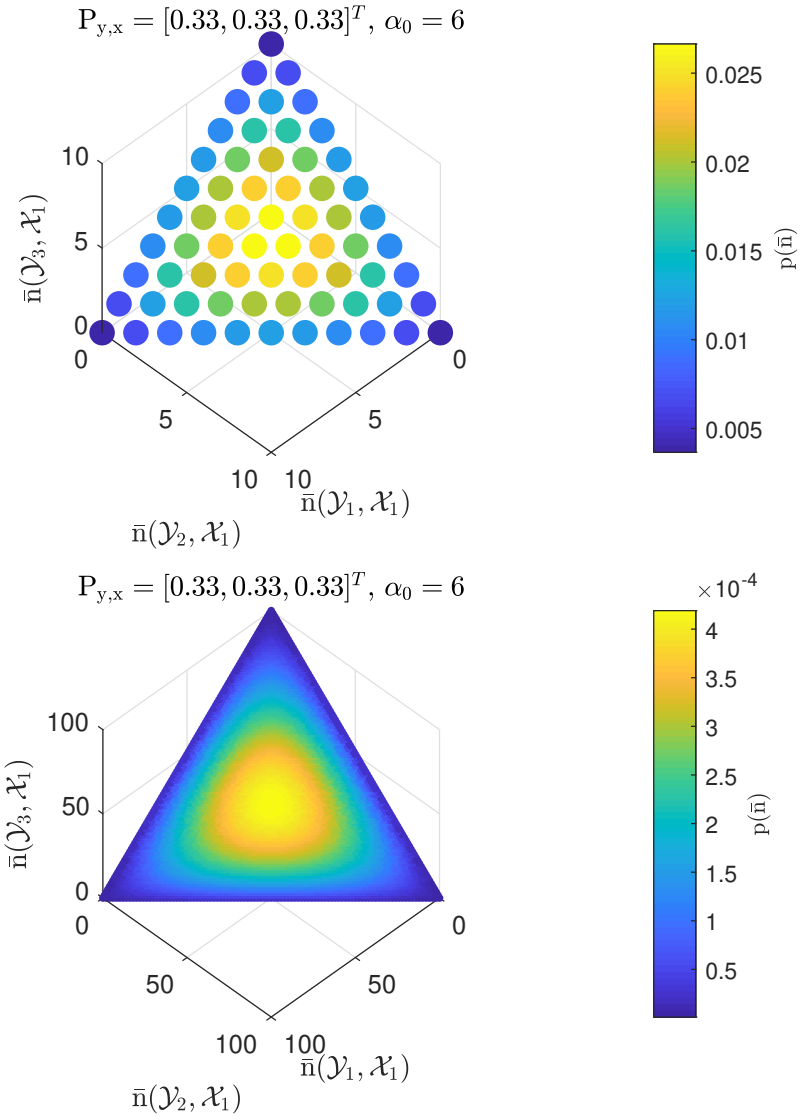

Figure 3.3: P_{ψ} for different training set sizes N

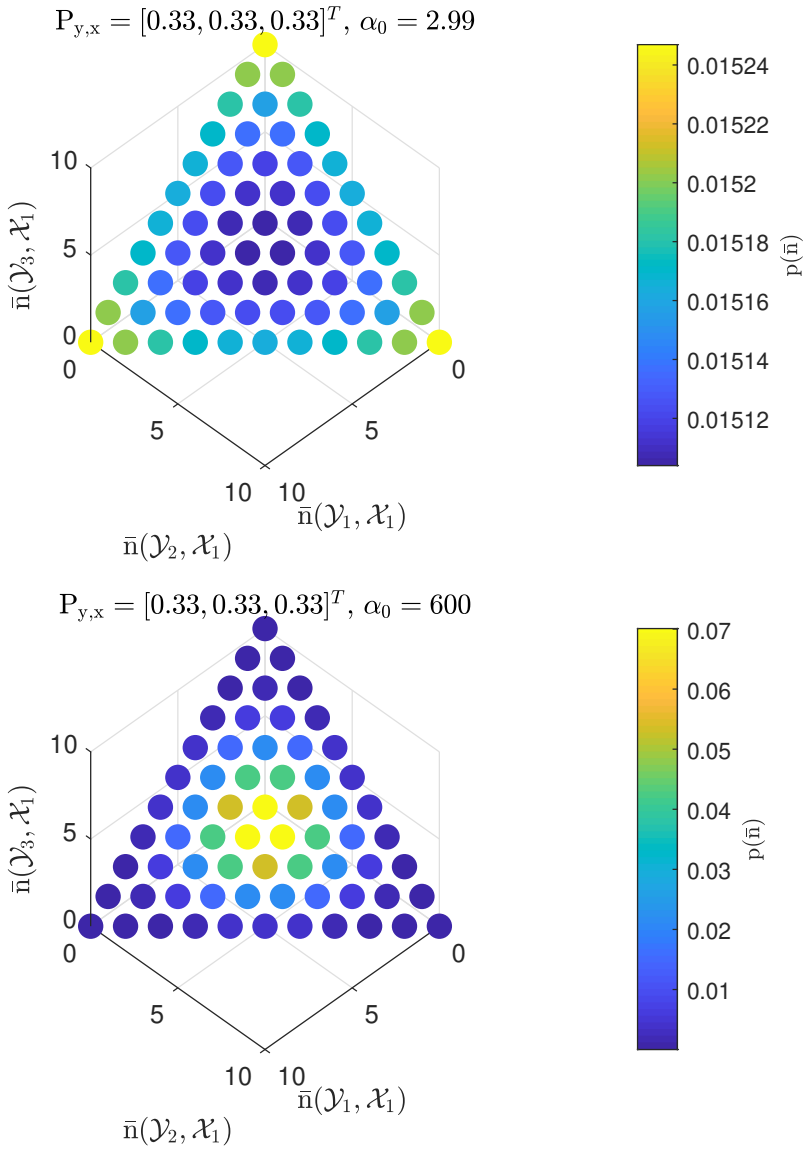

Figure 3.4: P_ψ for different prior concentrations α_0

3.1.2.1 Marginal and Conditional Distributions

It is also useful to express the marginal and conditional distributions for the training data given the Dirichlet prior. As $\mathrm{P}_{\mathrm{X}\,|\,\theta}$ is of exponential form with respect to the marginal model θ_{m}, the marginal distribution of X can be expressed as

$$\begin{aligned}\mathrm{P}_{\mathrm{X}}(X) &= \mathrm{E}_{\theta_{\mathrm{m}}}\left[\,\mathrm{P}_{\mathrm{X}\,|\,\theta_{\mathrm{m}}}\,\right](X) \\ &= \mathrm{E}_{\theta}\left[\prod_{n=1}^{N}\mathrm{P}_{\mathrm{X}_n\,|\,\theta}\left(X_n|\theta\right)\right] \\ &= \mathrm{E}_{\theta_{\mathrm{m}}}\left[\left(\prod_{x\in\mathcal{X}}\theta_{\mathrm{m}}(x)^{\Psi_{\mathrm{m}}(x;X)}\right)^{N}\right] \\ &= \frac{\beta\big(\alpha_0\,\alpha_{\mathrm{m}} + N\,\Psi_{\mathrm{m}}(X)\big)}{\beta(\alpha_0\,\alpha_{\mathrm{m}})}\,.\end{aligned} \tag{3.21}$$

As the model marginal θ_{m} and conditional θ_{c} are independent, the distribution $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}$ can be represented as

$$\begin{aligned}\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X) &= \mathrm{E}_{\theta_{\mathrm{c}}}\left[\,\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta_{\mathrm{c}}}\,\right](Y;X) \\ &= \prod_{x\in\mathcal{X}}\mathrm{E}_{\theta_{\mathrm{c}}(x)}\left[\prod_{y\in\mathcal{Y}}\theta_{\mathrm{c}}(y;x)^{N\,\Psi_{\mathrm{m}}(x;X)\,\Psi_{\mathrm{c}}(y;x;Y,X)}\right] \\ &= \prod_{x\in\mathcal{X}}\frac{\beta\big(\alpha_0\,\alpha_{\mathrm{m}}(x)\,\alpha_{\mathrm{c}}(x) + N\,\Psi_{\mathrm{m}}(x;X)\,\Psi_{\mathrm{c}}(x;Y,X)\big)}{\beta\big(\alpha_0\,\alpha_{\mathrm{m}}(x)\,\alpha_{\mathrm{c}}(x)\big)}\,.\end{aligned} \tag{3.22}$$

The corresponding distributions for the sufficient statistics will be expressed as well. Recall that $\psi_{\mathrm{m}}\,|\theta \sim \mathrm{Emp}(N,\theta_{\mathrm{m}})$; by the aggregation property of Dirichlet-Empirical functions (inherited from the Dirichlet-Multinomial properties [9]), the random process is distributed as $\psi_{\mathrm{m}} \sim \mathrm{DE}(N,\alpha_0,\alpha_{\mathrm{m}})$.

Also of interest is the distribution of ψ_{c} conditioned on its aggregation ψ_{m}. Using the Dirichlet-Empirical properties presented in Appendix A.3, it can be shown that

$$\begin{aligned}\mathrm{P}_{\psi_{\mathrm{c}}\,|\,\psi_{\mathrm{m}}}(\psi_{\mathrm{c}}\,|\,\psi_{\mathrm{m}}) &= \prod_{x\in\mathcal{X}}\left[\mathcal{M}\left(N\,\psi_{\mathrm{m}}(x)\,\psi_{\mathrm{c}}(x)\right)\frac{\beta\big(\alpha_0\,\alpha_{\mathrm{m}}(x)\,\alpha_{\mathrm{c}}(x) + N\,\psi_{\mathrm{m}}(x)\,\psi_{\mathrm{c}}(x)\big)}{\beta\big(\alpha_0\,\alpha_{\mathrm{m}}(x)\,\alpha_{\mathrm{c}}(x)\big)}\right] \\ &= \prod_{x\in\mathcal{X}}\mathrm{DE}\Big(\,\psi_{\mathrm{c}}(x);N\,\psi_{\mathrm{m}}(x),\alpha_0\,\alpha_{\mathrm{m}}(x),\alpha_{\mathrm{c}}(x)\Big)\end{aligned} \tag{3.23}$$

over the domain $\prod_{x\in\mathcal{X}}\left\{\frac{n}{N\,\psi_{\mathrm{m}}(x)} : n\in{\mathbb{Z}_{\geq 0}}^{\mathcal{Y}} : \sum_{y\in\mathcal{Y}}n(y) = N\,\psi_{\mathrm{m}}(x)\right\}$. Observe that conditioning on the marginal empirical process renders the conditional processes $\psi_{\mathrm{c}}(x)$

independent of one another and that they are also Dirichlet-Empirical, such that $\psi_{\mathrm{c}}(x)|\,\psi_{\mathrm{m}}(x) \sim \mathrm{DE}\left(N\,\psi_{\mathrm{m}}(x), \alpha_0\,\alpha_{\mathrm{m}}(x), \alpha_{\mathrm{c}}(x)\right)$.

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

As shown in Equation (2.27), the decision selected by the optimally designed function depends on $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$, the distribution of the unobserved y conditioned on all observable random elements. This PMF will be expressed next.

First observe that since $\mathrm{P}_{\mathrm{D}\,|\theta}$ is of exponential form, the Dirichlet prior p_θ is its conjugate prior [19]; thus, the model posterior PDF given the training data is

$$\mathrm{p}_{\theta|\,\mathrm{D}}(\theta|D) = \beta\,(\alpha_0\alpha + N\Psi(D))^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha_0\alpha(y,x)+N\Psi(y,x;D)-1}\,, \qquad (3.24)$$

a Dirichlet distribution with concentration $\alpha_0 + N$ and parameter distribution

$$\begin{aligned} \mu_{\theta|\,\mathrm{D}} &= \frac{\alpha_0\alpha + N\Psi(\mathrm{D})}{\alpha_0 + N} \\ &= \gamma\alpha + (1-\gamma)\Psi(\mathrm{D})\,, \end{aligned} \qquad (3.25)$$

where the weight

$$\gamma = \left(1 + \frac{N}{\alpha_0}\right)^{-1} \in (0,1] \qquad (3.26)$$

is introduced. Note that when $N = 0$, the empirical transform Ψ is undefined; however, since $1 - \gamma = 0$, it is understood that $\mu_{\theta|\,\mathrm{D}} = \alpha$. This convention is used throughout for brevity wherever the empirical transform is needed.

This posterior distribution is of specific interest in the machine learning literature. While Bayesian techniques are used here, often point estimates of the model θ are formed; perhaps the most common approach is to form the Maximum a posteriori estimate,

$$\begin{aligned} \theta_{\mathrm{MAP}}(D) &= \arg\max_{\theta\in\Theta} \mathrm{P}_{\theta|\,\mathrm{D}}(\theta|D) = \frac{\alpha_0\alpha + N\Psi(D) - 1}{\alpha_0 + N - |\,\mathcal{Y}\,||\,\mathcal{X}\,|} \\ &= \frac{\alpha_0}{\alpha_0 - |\,\mathcal{Y}\,||\,\mathcal{X}\,| + N}(\alpha - \alpha_0^{-1}) + \frac{N}{\alpha_0 - |\,\mathcal{Y}\,||\,\mathcal{X}\,| + N}\Psi(\mathrm{D})\,. \end{aligned} \qquad (3.27)$$

This maximizing value is only valid if $\mu_{\theta|\,\mathrm{D}} > (\alpha_0 + N)^{-1}$. For the uniform model prior, the maximizing value of the posterior is the empirical model $\Psi(D)$.

MAP discussion out of place?

Observe that the concentration parameter increases proportionately with both the training data volume and the prior concentration. Consequently, as $N \to \infty$, $\gamma \to 0$ and the posterior converges to $\mathrm{p}_{\theta|\,\mathrm{D}} \to \delta\big(\,\cdot - \Psi(\mathrm{D})\big)$; as more data is collected, the model can be more positively identified and used to formulate minimum risk decisions. Conversely, as $\alpha_0 \to \infty$, $\gamma \to 1$ reflection confidence in the prior and the posterior tends toward $\mathrm{p}_{\theta|\,\mathrm{D}} \to \delta(\cdot - \alpha)$, independent of the training data.

Fig. 3.5 shows the influence of the training data on the model distribution; after conditioning on the training data (via Ψ), the PDF concentration shifts away from the models favored by the prior knowledge and towards other models that better account for the observations.

Recall that the joint PMF of y and x conditioned on the training data is equivalent to the posterior mean $\mu_{\theta|\,\mathrm{D}}$, such that [13]

$$\mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{D}} = \gamma\alpha + (1-\gamma)\Psi(\mathrm{D})\;. \tag{3.28}$$

This is a mixture distribution of the prior mean $\mu_\theta = \alpha$ and the empirical distribution $\Psi(\mathrm{D})$. The more informative the model prior (i.e., larger α_0), the more the prior mean is favored; the more data, the more the empirical model is favored. The marginal distribution for x given D is

$$\begin{aligned}\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} \equiv \mathrm{P}_{\mathrm{x}\,|\,\mathrm{X}} &= \frac{\alpha_0\,\alpha_\mathrm{m} + N\,\Psi_\mathrm{m}(\mathrm{X})}{\alpha_0 + N} \\ &= \gamma\,\alpha_\mathrm{m} + (1-\gamma)\,\Psi_\mathrm{m}(\mathrm{X})\;.\end{aligned} \tag{3.29}$$

Finally, the predictive distribution of interest is generated via Bayes rule as

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} &= \frac{\alpha_0\alpha(\cdot,\mathrm{x}) + N\Psi(\cdot,\mathrm{x};\mathrm{D})}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x}) + N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})} \\ &= \left(\frac{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x})}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x}) + N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}\right)\alpha_\mathrm{c}(\mathrm{x}) + \left(\frac{N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x}) + N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}\right)\Psi_\mathrm{c}(\mathrm{x};\mathrm{D}) \\ &\equiv \gamma_\mathrm{m}(\mathrm{x};\mathrm{X})\,\alpha_\mathrm{c}(\mathrm{x}) + \big(1-\gamma_\mathrm{m}(\mathrm{x};\mathrm{X})\big)\,\Psi_\mathrm{c}(\mathrm{x};\mathrm{D})\;,\end{aligned} \tag{3.30}$$

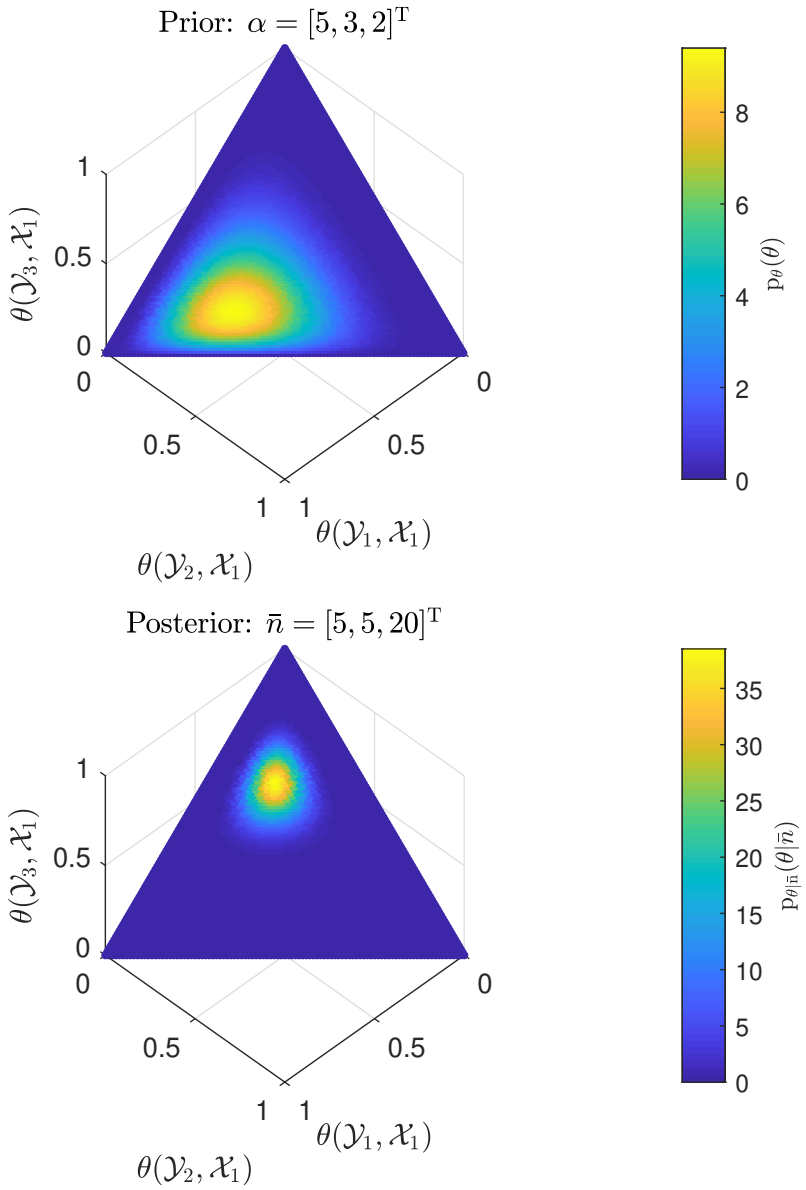

Figure 3.5: Model θ PDF, prior and posterior

where the "marginal" weighting function

$$\gamma_{\mathrm{m}}(X) = \left(1 + \frac{N\,\Psi_{\mathrm{m}}(X)}{\alpha_0\,\alpha_{\mathrm{m}}}\right)^{-1} \in (0,1]^{\mathcal{X}} \tag{3.31}$$

is introduced. The last representation views the distribution as a convex combination of two conditional distributions. The first distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha_{\mathrm{c}}(\mathrm{x})$ 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 model and depends on D, not on α.

Recall that the weighting factors $\alpha_0\,\alpha_{\mathrm{m}}(x)$ and $N\,\Psi_{\mathrm{m}}(x;\mathrm{X})$ are the concentration of the conditional prior $\theta_{\mathrm{c}}(x)$ and the number of training samples characterizing the conditional empirical model $\psi_{\mathrm{c}}(x)$ (samples satisfying $X_n = \mathrm{x}$), respectively. As the former increases relative to the latter, the weight value $\gamma_{\mathrm{m}}(x;\mathrm{X}) \to 0$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ tends away from the prior function $\alpha_{\mathrm{c}}(x)$ and towards the empirical conditional distribution $\Psi_{\mathrm{c}}(x;\mathrm{D})$.

Uniform Prior For the uniform model prior PDF, the conditional distribution is

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} &= \frac{N\Psi(\cdot,\mathrm{x};\mathrm{D}) + 1}{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X}) + |\,\mathcal{Y}\,|} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{|\,\mathcal{Y}\,| + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\frac{1}{|\,\mathcal{Y}\,|} + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{|\,\mathcal{Y}\,| + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\Psi_{\mathrm{c}}(\mathrm{x};\mathrm{D})\ .\end{aligned} \tag{3.32}$$

Now the prior PMF contribution $\alpha_{\mathrm{c}}(x)$ is a uniform distribution over the $|\,\mathcal{Y}\,|$ possible outputs. The weighting factors are dependent on conditional prior concentration $\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) = |\,\mathcal{Y}\,|$; the more possible outcomes $|\,\mathcal{Y}\,|$ there are for a given training set size, the more the Bayesian predictive distribution tends toward the uniform PMF.

3.1.3.1 Via the Conditional Model Distribution

PGR: reference posterior equations!

PGR: DIR FIGS? for PDF asymptotics?

The Bayesian distributions $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ can also be found from the posterior distributions $\mathrm{p}_{\theta_{\mathrm{m}}\,|\,\mathrm{D}}$ and $\mathrm{p}_{\theta_{\mathrm{c}}\,|\,\mathrm{x,D}}$, respectively. As the Dirichlet assumption renders θ_{m}

and θ_c independent, it can be shown that $P_{Y|X} = E_{\theta_c}\left[P_{Y|X,\theta_c}\right]$ and thus that θ_m is conditionally independent of Y given X. Furthermore, the Dirichlet distribution p_{θ_m} is a conjugate prior for the likelihood $P_{X|\theta_m}$. As a result, $\theta_m \,|\, D \sim \mathrm{Dir}\left(\alpha_0 + N, \mu_{\theta_m|X}\right)$ and

$$\begin{aligned} P_{x|D} &= \mu_{\theta_m|D} \\ &\equiv \mu_{\theta_m|X} = \gamma\,\alpha_m + (1-\gamma)\,\Psi_m(X)\ . \end{aligned} \tag{3.33}$$

Similarly, the distribution can be expressed in terms of the empirical model sufficient statistic as

$$\begin{aligned} P_{x|\psi} &= \mu_{\theta_m|\psi} \\ &\equiv \mu_{\theta_m|\psi_m} = \gamma\,\alpha_m + (1-\gamma)\,\psi_m\ , \end{aligned} \tag{3.34}$$

where the dependency on ψ is expressed only through the marginal random process ψ_m.

independence of conditionals too

The posterior $p_{\theta_c|x,D}$ can be simplified by noting that the independence of θ_m and θ_c implies $P_{Y|X,x} = E_{\theta_c}\left[P_{Y|X,\theta_c}\right] = P_{Y|X}$. Consequently, θ_c is conditionally independent of x given D. Thus, as p_{θ_c} is a conjugate prior for $P_{Y|X,\theta_c}$ the posterior distribution is

$$\begin{aligned} p_{\theta_c|x,D}(\theta_c|x,D) = p_{\theta_c|D}(\theta_c|D) &= \prod_{x'\in\mathcal{X}} p_{\theta_c(x')|D}\left(\theta_c(x')|D\right) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\theta_c(x'); \alpha_0\,\alpha_m(x') + N\,\Psi_m(x';D), \mu_{\theta_c|D}(x';D)\right)\ , \end{aligned} \tag{3.35}$$

$$\begin{aligned} p_{\theta_c|x,D} = p_{\theta_c|D} &= \prod_{x'\in\mathcal{X}} p_{\theta_c(x')|D} \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\alpha_0\,\alpha_m(x') + N\,\Psi_m(x';D), \mu_{\theta_c(x')|D}\right)\ , \end{aligned} \tag{3.36}$$

where

$$\begin{aligned} \mu_{\theta_c|x,D}(x';x,D) &= \mu_{\theta_c|D}(x';D) \\ &\equiv \gamma_m(x';X)\,\alpha_c(x') + \left(1-\gamma_m(x';X)\right)\Psi_c(x';D) \end{aligned} \tag{3.37}$$

$$\begin{aligned}\mu_{\theta_{\mathrm{c}}\,|\,\mathrm{x,D}} = \mu_{\theta_{\mathrm{c}}\,|\,\mathrm{D}} &= \bigotimes_{x'\in\mathcal{X}} \mu_{\theta_{\mathrm{c}}(x')|\,\mathrm{D}} \\ &\equiv \bigotimes_{x'\in\mathcal{X}} \big(\gamma_{\mathrm{m}}(x';X)\,\alpha_{\mathrm{c}}(x') + \big(1-\gamma_{\mathrm{m}}(x';\mathrm{X})\big)\,\Psi_{\mathrm{c}}(x';\mathrm{D})\big)\end{aligned} \tag{3.38}$$

and the distinct model conditional PMF's are independent from one another. The Bayes predictive PMF can thus be expressed as

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} &= \mu_{\theta_{\mathrm{c}}(x)|\,\mathrm{x,D}} = \mu_{\theta_{\mathrm{c}}(\mathrm{x})|\,\mathrm{D}} \\ &\equiv \gamma_{\mathrm{m}}(\mathrm{x};X)\,\alpha_{\mathrm{c}}(\mathrm{x}) + \big(1-\gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\big)\,\Psi_{\mathrm{c}}(\mathrm{x};\mathrm{D})\;.\end{aligned} \tag{3.39}$$

A similar treatment demonstrates that

$$\begin{aligned}\mathrm{p}_{\theta_{\mathrm{c}}\,|\,\mathrm{x},\psi}(\theta_{\mathrm{c}}\,|x,\psi) = \mathrm{p}_{\theta_{\mathrm{c}}\,|\psi}(\theta_{\mathrm{c}}\,|\psi) &\equiv \mathrm{p}_{\theta_{\mathrm{c}}\,|\,\psi_{\mathrm{m}},\psi_{\mathrm{c}}}(\theta_{\mathrm{c}}\,|\,\psi_{\mathrm{m}},\psi_{\mathrm{c}}) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{p}_{\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')}\big(\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')\big) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\Big(\theta_{\mathrm{c}}(x');\alpha_0\,\alpha_{\mathrm{m}}(x') + N\,\psi_{\mathrm{m}}(x'),\mu_{\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')}\big(\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')\big)\Big)\;,\end{aligned} \tag{3.40}$$

$$\begin{aligned}\mathrm{p}_{\theta_{\mathrm{c}}\,|\,\mathrm{x},\psi} = \mathrm{p}_{\theta_{\mathrm{c}}\,|\psi} &\equiv \mathrm{p}_{\theta_{\mathrm{c}}\,|\,\psi_{\mathrm{m}},\psi_{\mathrm{c}}} \\ &= \bigotimes_{x'\in\mathcal{X}} \mathrm{p}_{\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')} \\ &= \bigotimes_{x'\in\mathcal{X}} \mathrm{Dir}\Big(\alpha_0\,\alpha_{\mathrm{m}}(x') + N\,\psi_{\mathrm{m}}(x'),\mu_{\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')}\Big)\;,\end{aligned} \tag{3.41}$$

where

$$\mu_{\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')} = \gamma_{\mathrm{m}}(x';\psi_{\mathrm{m}})\,\alpha_{\mathrm{c}}(x') + \big(1-\gamma_{\mathrm{m}}(x';\psi_{\mathrm{m}})\big)\,\psi_{\mathrm{c}}(x')\;.$$

$$\begin{aligned}\mu_{\theta_{\mathrm{c}}\,|\,\psi_{\mathrm{m}},\psi_{\mathrm{c}}} &= \bigotimes_{x'\in\mathcal{X}} \mu_{\theta_{\mathrm{c}}(x')|\,\psi_{\mathrm{m}}(x'),\psi_{\mathrm{c}}(x')} \\ &= \bigotimes_{x'\in\mathcal{X}} \Big(\gamma_{\mathrm{m}}(x';\psi_{\mathrm{m}})\,\alpha_{\mathrm{c}}(x') + \big(1-\gamma_{\mathrm{m}}(x';\psi_{\mathrm{m}})\big)\,\psi_{\mathrm{c}}(x')\Big)\;.\end{aligned} \tag{3.42}$$

and the modified weighting function

$$\gamma_{\mathrm{m}}(\psi_{\mathrm{m}}) = \left(1 + \frac{N\,\psi_{\mathrm{m}}}{\alpha_0\,\alpha_{\mathrm{m}}}\right)^{-1} \in (0,1]^{\mathcal{X}} \tag{3.43}$$

is introduced, operating on the empirical data distribution.

Reconcile different weighting funcs for D and psi...

Observe that when the conditioning is performed using the sufficient statistic, the independent conditional models $\theta_{\mathrm{c}}(x)$ are only dependent on the marginal empirical model value $\psi_{\mathrm{m}}(x)$ and on the corresponding conditional empirical model $\psi_{\mathrm{c}}(x)$.

The Bayes predictive PMF can thus be expressed as

$$\begin{aligned} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi} = \mu_{\theta_{\mathrm{c}}(\mathrm{x})|\,\mathrm{x},\psi} &\equiv \mu_{\theta_{\mathrm{c}}(\mathrm{x})|\,\psi_{\mathrm{m}}(\mathrm{x}),\psi_{\mathrm{c}}(\mathrm{x})} \\ &= \gamma_{\mathrm{m}}(\mathrm{x};\psi_{\mathrm{m}})\,\alpha_{\mathrm{c}}(\mathrm{x}) + \big(1 - \gamma_{\mathrm{m}}(\mathrm{x};\psi_{\mathrm{m}})\big)\,\psi_{\mathrm{c}}(\mathrm{x})\ . \end{aligned} \tag{3.44}$$

3.2 Predictive Model Estimation

Use lambda weight notation from conf. papers? And table!

This section analyzes the bias and variance of the Dirichlet-based $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ when used to estimate the clairvoyant predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \theta_{\mathrm{c}}(\mathrm{x})$. To simplify the analysis, the training data D will be represented using the marginal and conditional sufficient statistics $(\psi_{\mathrm{m}}, \psi_{\mathrm{c}})$, such that $\mathrm{E}_{\mathrm{D}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] = \mathrm{E}_{\psi_{\mathrm{m}},\psi_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi_{\mathrm{m}},\psi_{\mathrm{c}}}\right]$ and $\mathrm{C}_{\mathrm{D}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] = \mathrm{C}_{\psi_{\mathrm{m}},\psi_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi_{\mathrm{m}},\psi_{\mathrm{c}}}\right]$.

For a given x, the expected value of the estimate conditioned on the true model is

$$\begin{aligned} \mathrm{E}_{\psi_{\mathrm{m}},\psi_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi_{\mathrm{m}},\psi_{\mathrm{c}}}\right] &= \mathrm{E}_{\psi_{\mathrm{m}}\,|\,\theta_{\mathrm{m}}}\left[\gamma_{\mathrm{m}}(\mathrm{x};\psi_{\mathrm{m}})\right]\alpha_{\mathrm{c}}(\mathrm{x}) \\ &\quad + \mathrm{E}_{\psi_{\mathrm{m}}\,|\,\theta_{\mathrm{m}}}\left[1 - \gamma_{\mathrm{m}}(\mathrm{x};\psi_{\mathrm{m}})\right]\theta_{\mathrm{c}}(\mathrm{x})\ , \end{aligned} \tag{3.45}$$

where the properties of an Empirical distribution (Appendix A.1) conditioned on its aggregation have been used. The result is a convex combination of the conditional data-independent distribution $\alpha_{\mathrm{c}}(\mathrm{x})$ and the true conditional distribution $\theta_{\mathrm{c}}(\mathrm{x})$.

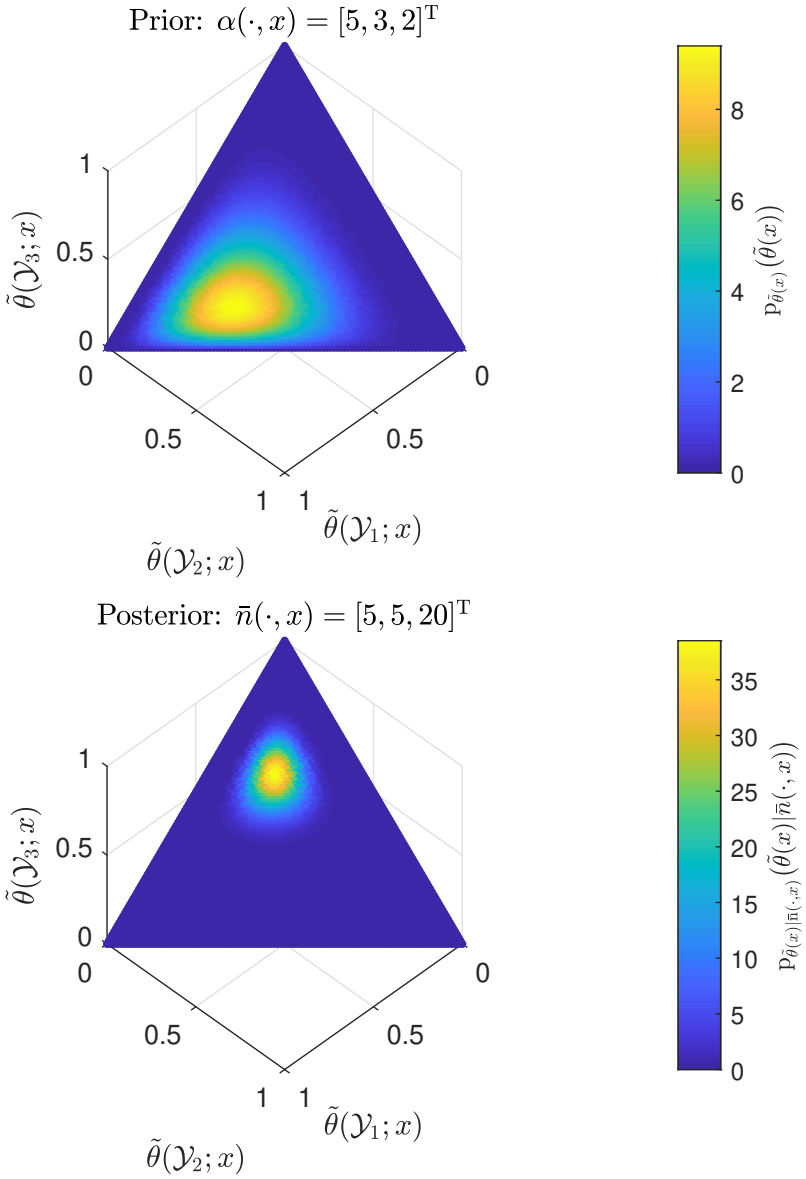

Figure 3.6: Model PDF, prior and posterior

Substituting into (2.35), the expected bias is

$$\mathrm{Bias}(\mathrm{x};\theta_\mathrm{m},\theta_\mathrm{c}) = \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\,\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right]\left(\,\alpha_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) . \tag{3.46}$$

and noting that

$$\begin{aligned}
&\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi_\mathrm{m},\psi_\mathrm{c}} - \mathrm{E}_{\psi_\mathrm{m},\psi_\mathrm{c}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi_\mathrm{m},\psi_\mathrm{c}}\,\right] \\
&\quad = \Big(\,\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m}) - \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\,\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right]\Big)\left(\,\alpha_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) \\
&\qquad + \left(1 - \gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right)\left(\,\psi_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) ,
\end{aligned} \tag{3.47}$$

the covariance (2.36) of the estimate can be represented as

$$\begin{aligned}
&\mathrm{Cov}(\mathrm{x};\theta_\mathrm{m},\theta_\mathrm{c}) = \mathrm{C}_{\psi_\mathrm{m},\psi_\mathrm{c}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi_\mathrm{m},\psi_\mathrm{c}}\,\right] \\
&\quad = \mathrm{C}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\,\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right]\left(\,\alpha_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) \otimes \left(\,\alpha_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) \\
&\qquad + \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\frac{\left(1 - \gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right)^2}{N\,\psi_\mathrm{m}(\mathrm{x})}\right]\Big(\,\mathrm{diag}\left(\,\theta_\mathrm{c}(\mathrm{x})\right) - \theta_\mathrm{c}(\mathrm{x}) \otimes \theta_\mathrm{c}(\mathrm{x})\Big) .
\end{aligned} \tag{3.48}$$

Substituting the estimator bias and variance into (2.37), the conditional second moments of $\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c})$ are

$$\begin{aligned}
&\mathrm{E}_{\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c}) \otimes \Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c})\right] \\
&\quad \equiv \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})^2\right]\left(\,\alpha_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) \otimes \left(\,\alpha_\mathrm{c}(\mathrm{x}) - \theta_\mathrm{c}(\mathrm{x})\right) \\
&\qquad + \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\frac{\left(1 - \gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right)^2}{N\,\psi_\mathrm{m}(\mathrm{x})}\right]\Big(\,\mathrm{diag}\left(\,\theta_\mathrm{c}(\mathrm{x})\right) - \theta_\mathrm{c}(\mathrm{x}) \otimes \theta_\mathrm{c}(\mathrm{x})\Big) .
\end{aligned} \tag{3.49}$$

Note that the bias is proportionate to the difference between the true conditional model and the data-independent estimate, while the covariance also depends on $\Sigma_{\psi_\mathrm{c}(x)\,|\,\psi_\mathrm{m}(x),\theta_\mathrm{c}(x)}$.

3.2.1 Trends

The trends of the bias and variance with the Dirichlet parameters and with the data volume N are of interest. These values effect a bias-variance trade-off via expectations of functions of $\gamma_\mathrm{m}(\psi_\mathrm{m})$ given the marginal model θ_m (and implicitly the marginal prior

mean α_{m}). Note that as a result of its aggregation property, the empirical process value $\psi_{\mathrm{m}}(x)$ conditioned on the model $\theta_{\mathrm{m}}(x)$ is distributed as

$$\begin{aligned} \mathrm{P}_{\psi_{\mathrm{m}}(x)|\,\theta_{\mathrm{m}}(x)}\left(\,\psi_{\mathrm{m}}(x)|\,\theta_{\mathrm{m}}(x)\right) &= \mathrm{Emp}\left(\left(\,\psi_{\mathrm{m}}(x), 1-\psi_{\mathrm{m}}(x)\right); N, \left(\,\theta_{\mathrm{m}}(x), 1-\theta_{\mathrm{m}}(x)\right)\right) \\ &= \mathrm{Bi}\left(N\,\psi_{\mathrm{m}}(x); N, \theta_{\mathrm{m}}(x)\right)\,, \end{aligned} \tag{3.50}$$

$$N\,\psi_{\mathrm{m}}(x)|\,\theta_{\mathrm{m}}(x) \sim \mathrm{Bi}\left(N, \theta_{\mathrm{m}}(x)\right)\,, \tag{3.51}$$

where Bi is the binomial PMF. Closed-forms have not been found for the expectations of the relevant functions.

binomial inverse moment review citations?

First consider the effects of the training data volume. Clearly, if $N = 0$, $\gamma_{\mathrm{m}}(\psi_{\mathrm{m}})$ is equal to one. Consequently, the prior mean $\alpha_{\mathrm{c}}(\mathrm{x})$ is used and the bias weight is maximal at unity; as this estimator is data-independent, the covariance is zero. As $N \to \infty$, the data PMF tends to $\mathrm{P}_{\psi_{\mathrm{m}}\,|\,\theta_{\mathrm{m}}} \to \delta[\cdot, \theta_{\mathrm{m}}]$ and $\gamma_{\mathrm{m}}(\psi_{\mathrm{m}}) \to \left(1 + \frac{N\,\theta_{\mathrm{m}}}{\alpha_0\,\alpha_{\mathrm{m}}}\right)^{-1}$. As a result, the bias scaling factor tends to zero for all values $x \in \mathcal{X}$ satisfying $\theta_{\mathrm{m}}(x) > 0$ and to one otherwise. Thus for observations falling in the support of the marginal model θ_{m}, the empirical data distribution $\psi_{\mathrm{c}}(x)$ is used and there is no bias. Similarly, the expectations affecting the covariance tend to zero. This demonstrates that the Dirichlet-based Bayesian estimate $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ converges to the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}$ in the limit of training data volume; this is guaranteed due to the full support of the prior.

Next consider the effects of the Dirichlet parameterization. As $\alpha_0 \to \infty$, $\gamma_{\mathrm{m}}(\psi_{\mathrm{m}})$ is again equal to one, the bias is maximized and the covariance is zero. Conversely, as $\alpha_0 \to 0$, the empirical data distribution is emphasized and the bias scaling factor tends to its minimal value $\left(1 - \theta_{\mathrm{m}}(x)\right)^N$; this value is equivalent to $\mathrm{P}_{\psi_{\mathrm{m}}(x)|\,\theta_{\mathrm{m}}}\left(0|\,\theta_{\mathrm{m}}\right)$, the probability that no training samples satisfy $\mathrm{X}_n = x$ and thus that $\psi_{\mathrm{c}}(x)$ is undefined. Additionally, the expectations affecting the covariance tend to

$$\mathrm{C}_{\psi_{\mathrm{m}}\,|\,\theta_{\mathrm{m}}}\left[\,\gamma_{\mathrm{m}}(\mathrm{x};\psi_{\mathrm{m}})\right] \to \left(1 - \theta_{\mathrm{m}}(\mathrm{x})\right)^N \left(1 - \left(1 - \theta_{\mathrm{m}}(\mathrm{x})\right)^N\right)\,, \tag{3.52}$$

proportionate to the variance of the number of matching samples $N\,\psi_{\mathrm{m}}(x)$, and

$$\mathrm{E}_{\psi_{\mathrm{m}}\,|\,\theta_{\mathrm{m}}}\left[\frac{\big(1-\gamma_{\mathrm{m}}(\mathrm{x};\psi_{\mathrm{m}})\big)^2}{N\,\psi_{\mathrm{m}}(\mathrm{x})}\right] \to \sum_{n=1}^{N}\binom{N}{n}\,\theta_{\mathrm{m}}(\mathrm{x})^n\big(1-\theta_{\mathrm{m}}(\mathrm{x})\big)^{N-n}\frac{1}{n}\,. \quad (3.53)$$

Note that the latter is equivalent to the first inverse moment of a positive binomial random variable [18].

Intuitive explanation for empirical predictor variance?

This demonstrates the critical bias-variance trade-off that is controlled by selection of the Dirichlet concentration parameter. For common applications, both distribution estimate bias and variance contribute to the overall risk and an optimal value of α_0 can be found to optimally balance these competing sources of error.

3.2.2 Example

To exemplify how the model estimate $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi}$ approximates $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$, consider a scenario with $|\,\mathcal{Y}\,| = 60$; the visualization will be provided for an arbitrary observation value x. The true model $\theta_{\mathrm{c}}(x)$ and prior mean $\alpha_{\mathrm{c}}(x)$ are distributed as $\alpha_{\mathrm{c}}(y;x) = \mathrm{Bi}(y/60;60,0.7)$ and $\theta_{\mathrm{c}}(y;x) = \mathrm{Bi}(y/60;60,0.3)$, respectively, and are shown in Fig. 3.7 – note the significant mismatch. The model and prior marginal values used are $\theta_{\mathrm{m}}(x) = 0.5$ and $\alpha_{\mathrm{m}}(x) = 0.5$.

Figs. 3.8 and 3.9 show how the bias and variance of the estimate changes with different values of N and α_0. The plot lines represent the mean of the estimator, $\mathrm{E}_{\psi|\theta}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi}\,\right]$; the shaded regions represent the square-root of the expected variance $\mathrm{C}_{\psi|\theta}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi}\,\right]$ above and below the conditional mean.

Observe that for $N = 5$, the $\alpha_0 = 0.1$ estimate (favoring the empirical PMF) has negligible bias but massive variance where θ is highest; conversely, the $\alpha_0 = 10$ estimate has low variance but high bias, favoring the erroneous prior mean α. The $N = 500$ figure shows that, given sufficient data, the $\alpha_0 = 0.1$ and $\alpha_0 = 10$ estimators will eliminate their high variance and high bias, respectively, leading to perfect estimation of the true model θ. The consistency of this estimate is guaranteed due to the full

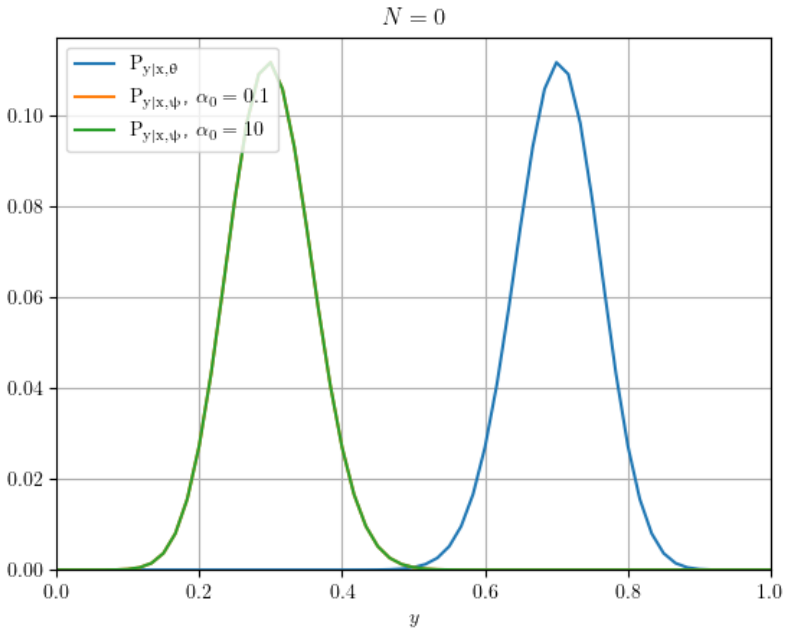

Figure 3.7: Model θ estimate, $N = 0$

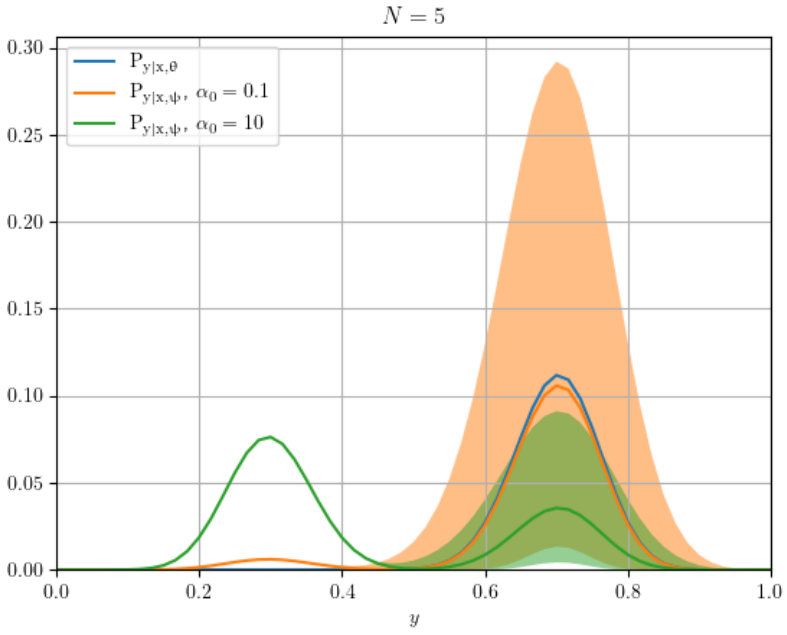

Figure 3.8: Model θ estimate, $N = 5$

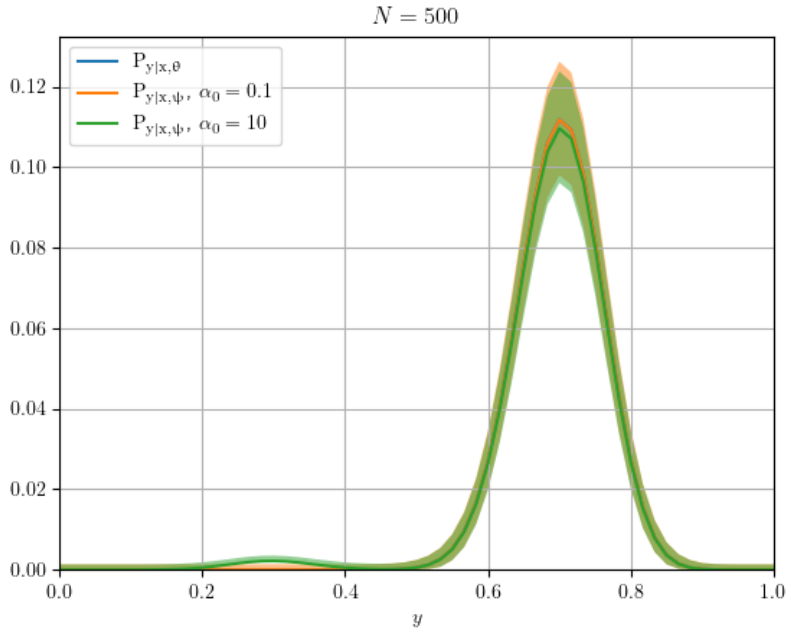

Figure 3.9: Model θ estimate, $N = 500$

support of the Dirichlet prior.

3.3 Applications to Common Loss Functions

REPLACE bayes figs with fixed alpha0, varying learner parameterization? Should be apples-to-apples comparison!!

REMOVE GENERAL, RELOCATED MATERIAL

In this section, the Dirichlet prior is applied to the regression and classification applications. Optimal learners f^* are found, the corresponding minimum Bayes risk $\mathcal{R}^*$ is assessed, and the risk $\mathcal{R}_\Theta(f^*;\theta)$ is analyzed.

It is informative to substitute the Bayes distribution using the Dirichlet prior (3.28) into (2.26), expressing the decision function for a given training set D as

$$\begin{aligned} f^*(\mathrm{D}) &= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\mathrm{D}}\Big[\,\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right)\Big] \qquad (3.54)\\ &= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \gamma \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \alpha(y,x)\,\mathcal{L}\left(g(x),y\right) + (1-\gamma)\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \Psi(y,x;\mathrm{D})\,\mathcal{L}\left(g(x),y\right)\\ &= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \gamma\,\mathrm{E}_{\mathrm{y},\mathrm{x}}\Big[\,\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right)\Big] + (1-\gamma)\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(g(\mathrm{X}_n),\mathrm{Y}_n\right)\\ &= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \alpha_0\,\mathrm{E}_{\mathrm{y},\mathrm{x}}\Big[\,\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right)\Big] + \sum_{n=1}^{N}\mathcal{L}\left(g(\mathrm{X}_n),\mathrm{Y}_n\right). \end{aligned}$$

Observe that the objective function can be represented as a convex combination of two expected losses, one with respect to the Dirichlet prior mean α, and one with respect to the empirical distribution $\Psi(D)$. As N increases, γ decreases and more emphasis is placed on the empirical loss. The final representation casts the learning optimization task as regularized empirical risk minimization. Note that the regularizing term is scaled using the Dirichlet localization α_0 – the stronger the prior knowledge, the more extreme the regularization. Importantly, the selection of the probability distribution α allows the designer to effect very different types of regularization from those typically used in the literature, such as the ℓ^2 norm for parametric learners.

Regularized loss REFERENCE? See TheoML p.72

It is useful to substitute the Bayes predictive distribution using the Dirichlet prior

(3.30) into (2.27), expressing the decision for a given input x and training set D as

$$
\begin{aligned}
f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] \qquad (3.55)\\
&= \arg\min_{h\in\mathcal{H}} \frac{\alpha_0 \sum_{y\in\mathcal{Y}} \alpha(y,\mathrm{x})\,\mathcal{L}(h,y) + N\sum_{y\in\mathcal{Y}} \Psi(y,\mathrm{x};\mathrm{D})\,\mathcal{L}(h,y)}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x}) + N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}\\
&= \arg\min_{h\in\mathcal{H}} \left(\frac{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x})}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x}) + N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}\right)\sum_{y\in\mathcal{Y}} \alpha_\mathrm{c}(y;\mathrm{x})\,\mathcal{L}(h,y)\\
&\quad + \left(\frac{N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{x}) + N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})}\right)\sum_{y\in\mathcal{Y}} \Psi_\mathrm{c}(y;\mathrm{x};\mathrm{D})\,\mathcal{L}(h,y)\\
&= \arg\min_{h\in\mathcal{H}} \gamma_\mathrm{m}(\mathrm{x};\mathrm{X})\sum_{y\in\mathcal{Y}} \alpha_\mathrm{c}(y;\mathrm{x})\,\mathcal{L}(h,y) + \big(1-\gamma_\mathrm{m}(\mathrm{x};\mathrm{X})\big)\sum_{y\in\mathcal{Y}} \Psi_\mathrm{c}(y;\mathrm{x};\mathrm{D})\,\mathcal{L}(h,y)\\
&= \arg\min_{h\in\mathcal{H}} \gamma_\mathrm{m}(\mathrm{x};\mathrm{X})\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathcal{L}(h,\mathrm{y})\right] + \big(1-\gamma_\mathrm{m}(\mathrm{x};\mathrm{X})\big)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathcal{L}\left(h,\mathrm{Y}_n\right)}{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]}\\
&= \arg\min_{h\in\mathcal{H}} \alpha_0\sum_{y\in\mathcal{Y}} \alpha(y,\mathrm{x})\,\mathcal{L}(h,y) + \sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathcal{L}\left(h,\mathrm{Y}_n\right).
\end{aligned}
$$

The metric to be minimized can be represented as a convex combination of two expected losses. The first expected loss is evaluated with respect to the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha_\mathrm{c}(\mathrm{x})$, which reflects the prior knowledge of the model parameter. The second term is the conditional empirical risk, or the average loss among samples Y_n whose corresponding values X_n match the observed value x. The convex weights are inherited from the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$; thus, for a given observation x, the model prior concentration $\alpha_0\,\alpha_\mathrm{m}(\mathrm{x})$ and the number of matching training samples $N\,\Psi_\mathrm{m}(\mathrm{x};\mathrm{X})$ dictate which of the two expectations are emphasized.

3.3.1 Regression: the Squared-Error Loss

PGR: Use finite hypothesis space instead, wait for continuous DP???

The elements of the finite cardinality set $\mathcal{Y}$ are real numbers, such that $\mathcal{Y} \subset \mathbb{R}$. Again, $\mathcal{H} = \mathbb{R} \supset \mathcal{Y}$.

3.3.1.1 Bayesian Estimation

Optimal Estimate: the Posterior Mean Substituting in the Bayes predictive distribution for a Dirichlet prior (3.30) into (2.43), the optimal Bayesian estimate is

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \\ &= \left(\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\sum_{y\in\mathcal{Y}} y\,\alpha_{\mathrm{c}}(y;\mathrm{x}) \\ &\quad + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\sum_{y\in\mathcal{Y}} y\,\Psi_{\mathrm{c}}(y;\mathrm{x};\mathrm{D}) \\ &\equiv \gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \big(1-\gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\big)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathrm{Y}_n}{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]}\,. \end{aligned} \tag{3.56}$$

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 mean of the training values Y_n which have a value X_n matching the observed value x. The weighting factors are the same as those of $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$; thus, stronger prior information (larger $\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})$) provides more weight to the estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ and more voluminous training data puts emphasis on the empirical conditional mean.

Uniform Prior The optimal estimator for a uniform prior is

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \left(\frac{|\,\mathcal{Y}\,|}{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X}) + |\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X}) + |\,\mathcal{Y}\,|}\right)\sum_{y\in\mathcal{Y}} y\,\Psi_{\mathrm{c}}(y;\mathrm{x};\mathrm{D}) \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X}) + |\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X}) + |\,\mathcal{Y}\,|}\right)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathrm{Y}_n}{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]}\,. \end{aligned} \tag{3.57}$$

Now, the model prior contribution to the weighting factors depends on the cardinality $|\,\mathcal{Y}\,|$ and the prior expectation is simply the average of the elements of $\mathcal{Y}$.

Minimum Bayes Risk: the Expected Posterior Variance The minimum Bayes squared error is $\mathcal{R}^* = \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right]$. Using the sufficient statistic $\psi \equiv \Psi(\mathrm{D})$, it can also be represented as $\mathrm{E}_{\mathrm{x},\psi}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\psi}\right]$; as such, the expectations are performed over ψ. Decompose the conditional variance as

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x},\psi} = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\psi}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\psi} \tag{3.58}$$

and assess the expected values of these terms separately using distributions derived from the Dirichlet prior. The first term is simply

$$\begin{aligned}\mathrm{E}_{\mathrm{x},\psi}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\psi}[\mathrm{y}^2]\right] &= \mathrm{E}_{\mathrm{y}}[\mathrm{y}^2] = \sum_{y\in\mathcal{Y}} y^2 \left(\sum_{x\in\mathcal{X}} \alpha(y,x)\right) \\ &= \mathrm{E}_{\mathrm{x}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2]\right] = \sum_{x\in\mathcal{X}} \alpha_{\mathrm{m}}(x) \sum_{y\in\mathcal{Y}} y^2\, \alpha_{\mathrm{c}}(y;x)\ ,\end{aligned}$$

where the different functions of α are represented by the PMF's of y and x. Next, find,

$$\begin{aligned}\mathrm{E}_{\mathrm{x},\psi}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\psi}\right] &\equiv \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\psi_{\mathrm{c}},\psi_{\mathrm{m}}\,|\,\mathrm{x}}\left[\frac{\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + N\,\psi_{\mathrm{m}}(\mathrm{x})\sum_{y\in\mathcal{Y}} y\,\psi_{\mathrm{c}}(y;\mathrm{x})\big)^2}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\psi_{\mathrm{m}}(\mathrm{x})\big)^2}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\psi_{\mathrm{c}},\psi_{\mathrm{m}}}\left[\frac{\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + N\,\psi_{\mathrm{m}}(\mathrm{x})\sum_{y\in\mathcal{Y}} y\,\psi_{\mathrm{c}}(y;\mathrm{x})\big)^2}{\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\psi_{\mathrm{m}}(\mathrm{x})\big)(\alpha_0+N)}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\psi_{\mathrm{m}}}\left[\frac{\mathrm{E}_{\psi_{\mathrm{c}}\,|\,\psi_{\mathrm{m}}}\left[\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + N\,\psi_{\mathrm{m}}(\mathrm{x})\sum_{y\in\mathcal{Y}} y\,\psi_{\mathrm{c}}(y;\mathrm{x})\big)^2\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\psi_{\mathrm{m}}(\mathrm{x})\big)(\alpha_0+N)}\right]\right] \\ &= \ldots \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{E}_{\psi_{\mathrm{m}}}\left[N\,\psi_{\mathrm{m}}(\mathrm{x})\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\psi_{\mathrm{m}}(\mathrm{x}) + 1\big)\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + 1\big)(\alpha_0+N)}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{N\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\alpha_{\mathrm{m}}(\mathrm{x}) + 1\big)\mu^2_{\mathrm{y}\,|\,\mathrm{x}}}{\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + 1\big)(\alpha_0+N)}\right]\ .\end{aligned} \tag{3.59}$$

PGR: provide additional steps? Check psi work?!

The above formulation exploits the statistical characterization of the aggregation, $\psi_{\mathrm{m}} \sim \mathrm{DE}(N, \alpha_0, \alpha_{\mathrm{m}})$; also used is the property that the Dirichlet-Empirical random process ψ_{c} conditioned on its aggregation ψ_{m} yields independent conditional DE functions $\psi_{\mathrm{c}}(x)|\,\psi_{\mathrm{m}}(x) \sim \mathrm{DE}\left(N\,\psi_{\mathrm{m}}(x), \alpha_0\,\alpha_{\mathrm{m}}(x), \alpha_{\mathrm{c}}(x)\right)$.

PGR: move to appendix???

Finally, combine the two formulas to represent the mininum Bayes risk,

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_{\mathrm{x},\psi}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\psi}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\psi}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\alpha_{\mathrm{m}}(\mathrm{x}) + 1\big)}{\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + 1\big)(\alpha_0+N)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_{\mathrm{m}}(\mathrm{x}) + (\alpha_0+N)^{-1}}{\alpha_{\mathrm{m}}(\mathrm{x}) + \alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]\ .\end{aligned} \tag{3.60}$$

The minimum Bayes squared error is the expected value of the scaled conditional variance with respect to $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha_{\mathrm{c}}(\mathrm{x})$. The expectation is taken with respect to the prior marginal distribution $\mathrm{P}_{\mathrm{x}} = \alpha_{\mathrm{m}}$.

The scaling factor for each term $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ depends on the marginal P_{x}, as well as on the prior concentration 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 $\mathcal{R}^* = \mathrm{E}_{\mathrm{x}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$. Conversely, as $N \to \infty$, the Bayes risk is $\mathcal{R}^* \to \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_{\mathrm{m}}(\mathrm{x})}{\alpha_{\mathrm{m}}(\mathrm{x})+\alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$; note that this is equivalent to the expected irreducible risk $\mathrm{E}_{\theta}\left[\mathcal{R}^*_{\Theta}(\theta)\right] = \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$. Also, as the model concentration parameter $\alpha_0 \to 0$, the risk tends to zero (for $N > 0$); as $\alpha_0 \to \infty$, the risk tends toward $\mathrm{E}_{\mathrm{x}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$.

PGR: first/second derivatives of alpha0??

To illustrate these trends, explicitly define the sets $\mathcal{Y} = \{i/M_{\mathrm{y}} : i = 0, \ldots, M_{\mathrm{y}} - 1\}$ and $\mathcal{X} = \{i/M_{\mathrm{x}} : i = 0, \ldots, M_{\mathrm{x}} - 1\}$. Assume that the conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ is independent of x; in this case, the squared error becomes the conditional variance scaled by a factor dependent on the marginal distribution P_{x}, such that $\mathcal{R}^* = \Sigma_{\mathrm{y}\,|\,\mathrm{x}}\,\mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_{\mathrm{m}}(\mathrm{x})+(\alpha_0+N)^{-1}}{\alpha_{\mathrm{m}}(\mathrm{x})+\alpha_0^{-1}}\right]$. Figs. 3.10 and 3.11 display how the risk changes with N and α_0 when $\alpha_{\mathrm{c}}(\mathrm{x})$ and α_{m} are fixed.

It may not seem intuitive 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 $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models which have ℓ_0 norms satisfying $\|\theta\|_0 = 1$. Although these PMF's are maximally separated (and uncorrelated), they all have zero variance. The optimal learner (3.56) 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 instructional to visualize how the minimum squared error changes for fixed volume of training data N and a fixed prior concentration α_0. First, consider how the risk changes with the conditional PMF $\alpha_{\mathrm{c}}(\mathrm{x})$. Fig. 3.12 demonstrates how the squared error tends towards zero for PMFs that have ℓ_0-norm equal to one.

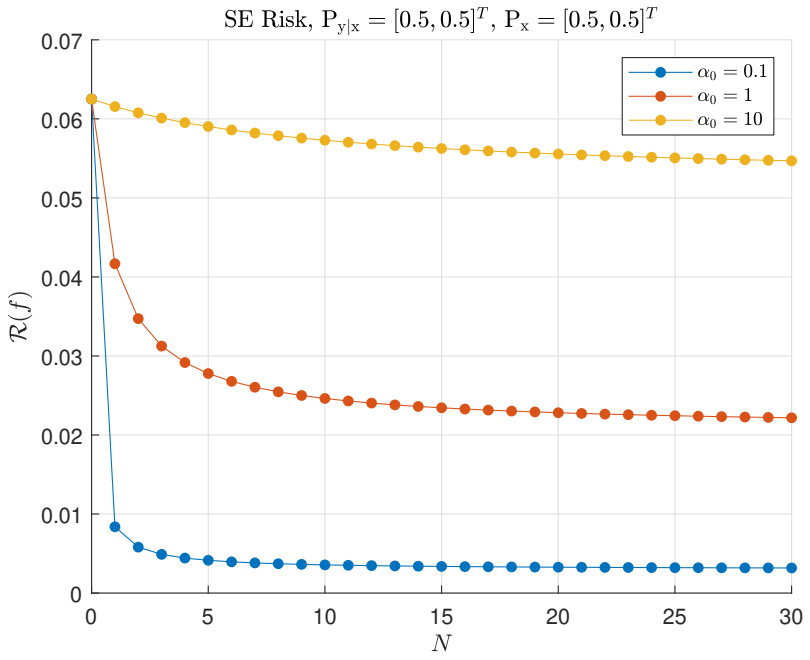

Figure 3.10: Minimum SE Risk for different training set sizes N

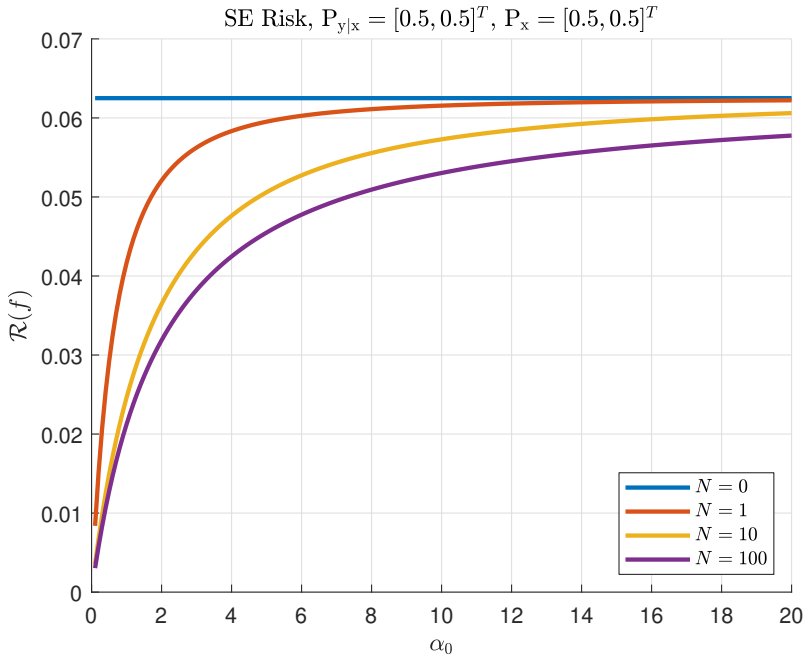

Figure 3.11: Minimum SE Risk for different prior concentrations α_0

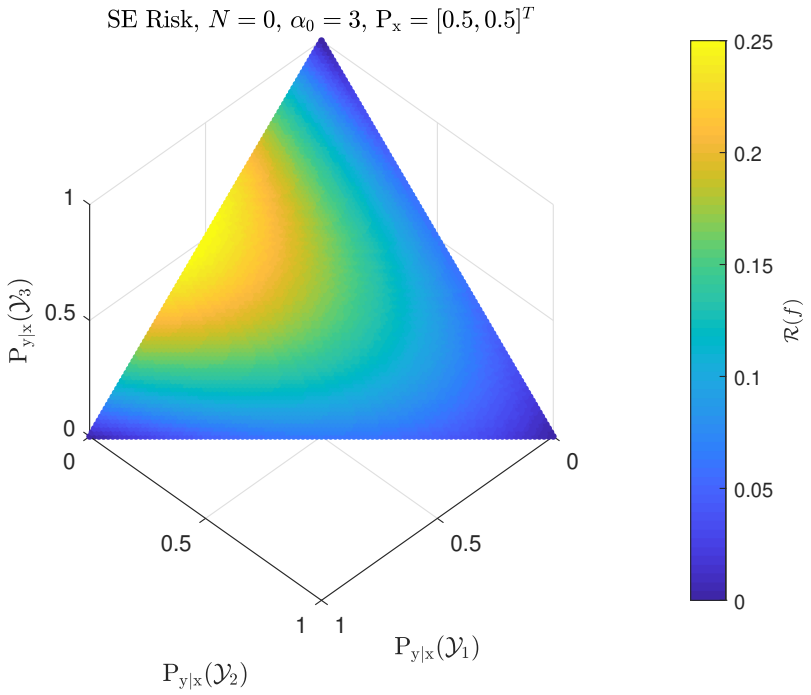

Figure 3.12: Minimum SE Risk for different prior means $\mathrm{P_{y\,|\,x}}$

Next, consider the effect of the marginal distribution α_m. Fig. 3.13 demonstrates how the risk changes with this marginal PMF. Observe that the risk is maximal at the distributions satisfying $\|\,\alpha_\mathrm{m}\,\|_0 = 1$; 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 $\alpha_\mathrm{m} = |\,\mathcal{X}\,|^{-1}$ the scaling factor becomes $\frac{|\,\mathcal{X}\,|^{-1}+(\alpha_0+N)^{-1}}{|\,\mathcal{X}\,|^{-1}+\alpha_0^{-1}}$ and the risk is minimal. Figs. 3.14 and 3.15 show how different marginals α_m affect the risk as a function of N and α_0, respectively.

Uniform Prior For the uniform model prior, the risk reduces to

$$\begin{aligned}\mathcal{R}^* &= \frac{|\,\mathcal{Y}\,|\big(N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,| + 1\big)}{\big(|\,\mathcal{Y}\,| + 1\big)\big(N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|\big)}\left[\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y^2\right) - \left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y\right)^2\right] \\ &= \frac{1 + \big(N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|\big)^{-1}}{1 + |\,\mathcal{Y}\,|^{-1}}\left[\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y^2\right) - \left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y\right)^2\right].\end{aligned} \tag{3.61}$$

Since all possible values of x are equally probable and the conditional probability $\alpha_\mathrm{c}(\mathrm{x})$ is uniform and independent of x, the risk simply becomes the variance of the set $\mathcal{Y}$ scaled by a factor dependent on $|\,\mathcal{Y}\,|$ and on $N/|\,\mathcal{X}\,|$. Without training data

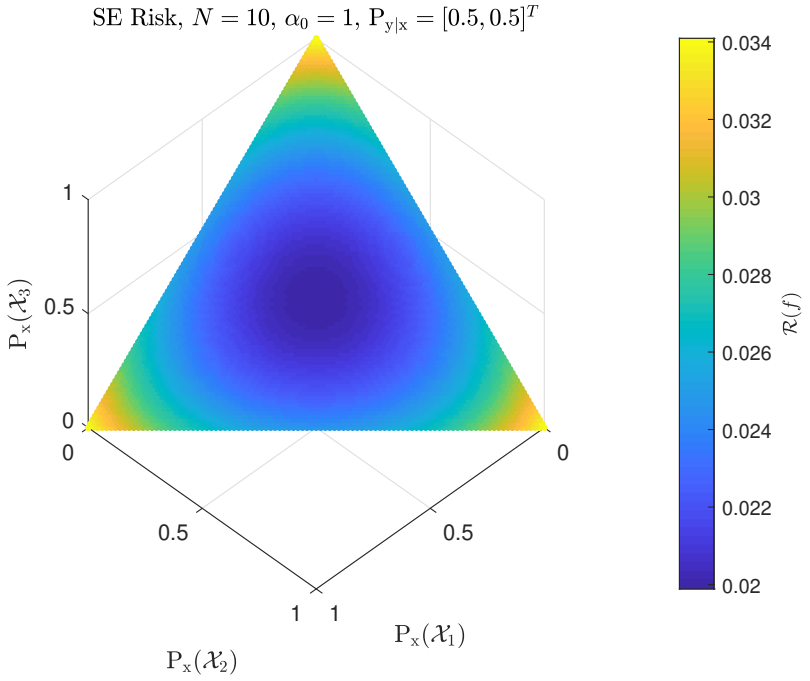

Figure 3.13: Minimum SE Risk for different prior means P_{x}

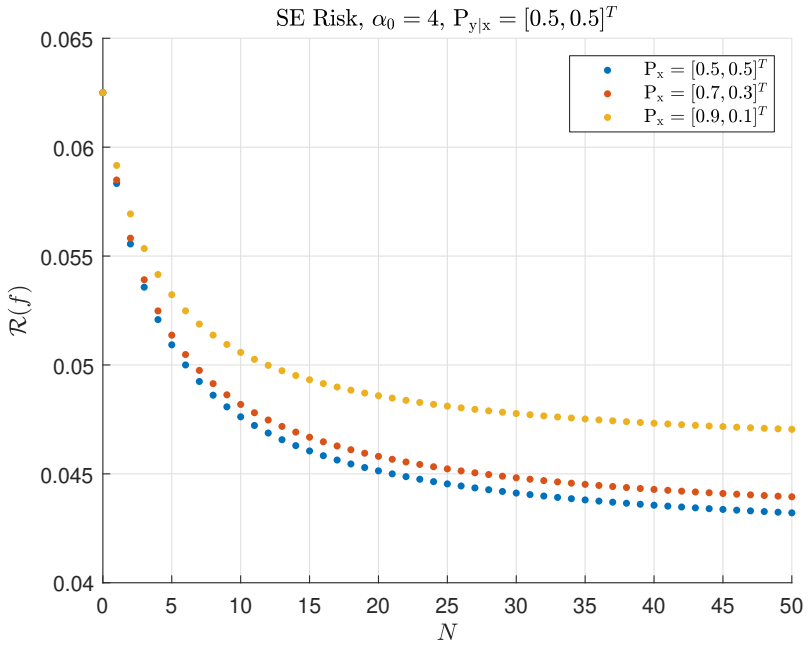

Figure 3.14: Minimum SE Risk for different training set volumes N

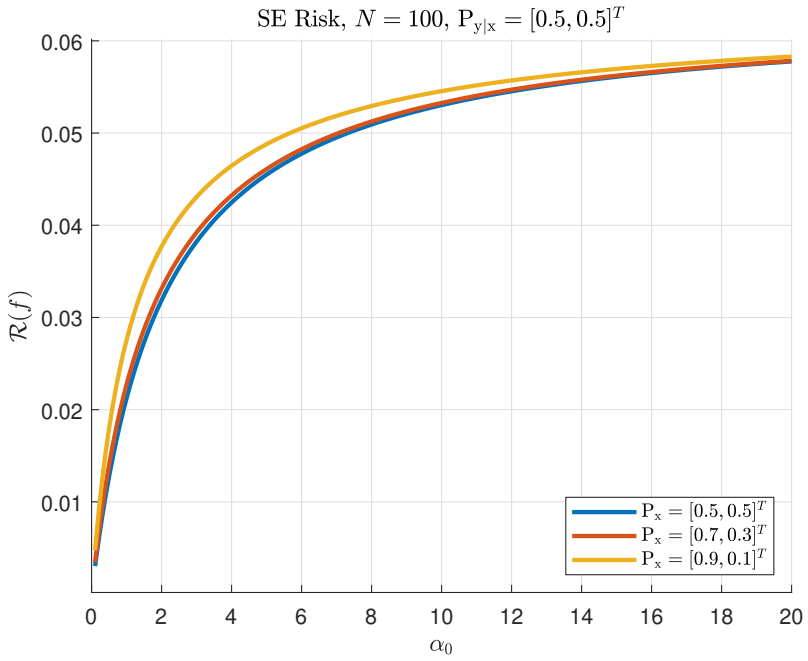

Figure 3.15: Minimum SE Risk for different prior concentrations α_0

$(N = 0)$, the scaling is unity; as $N/|\,\mathcal{X}\,| \to \infty$, the scaling factor is $\left(1 + |\,\mathcal{Y}\,|^{-1}\right)^{-1}$.

To visualize the performance, use the explicit sets $\mathcal{Y}$ and $\mathcal{X}$ defined earlier. The conditional variance becomes

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x}} = \frac{|\,\mathcal{Y}\,|^2 - 1}{12|\,\mathcal{Y}\,|^2} = \frac{1 - |\,\mathcal{Y}\,|^{-2}}{12} \tag{3.62}$$

and the minimum Bayes risk is expressed as

$$\begin{aligned} \mathcal{R}^* &= \frac{\left(1 - |\,\mathcal{Y}\,|^{-1}\right)\left(1 + \left(N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|\right)^{-1}\right)}{12} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|}\right)\frac{1 - |\,\mathcal{Y}\,|^{-2}}{12} + \left(\frac{N/|\,\mathcal{X}\,|}{N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|}\right)\frac{1 - |\,\mathcal{Y}\,|^{-1}}{12}\,. \end{aligned} \tag{3.63}$$

Interestingly, the minimum squared error for the uniform prior can be represented as a convex combination of two separate risk values with weighting factors dependent on $|\,\mathcal{Y}\,|$ and $N/|\,\mathcal{X}\,|$. Thus for a uniform prior, the risk depends on the number of elements in $\mathcal{Y}$ and the number of training samples "per element of $\mathcal{X}$". Note the relationship of these weighting factors to those of the conditional PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$, which depend on $\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})$ and on $N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})$. For the uniform prior, $\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) = |\,\mathcal{Y}\,|$ and $N\,\mathrm{E}_{\mathrm{X}}\left[\,\Psi_{\mathrm{m}}(\mathrm{X})\right] = N/|\,\mathcal{X}\,|$.

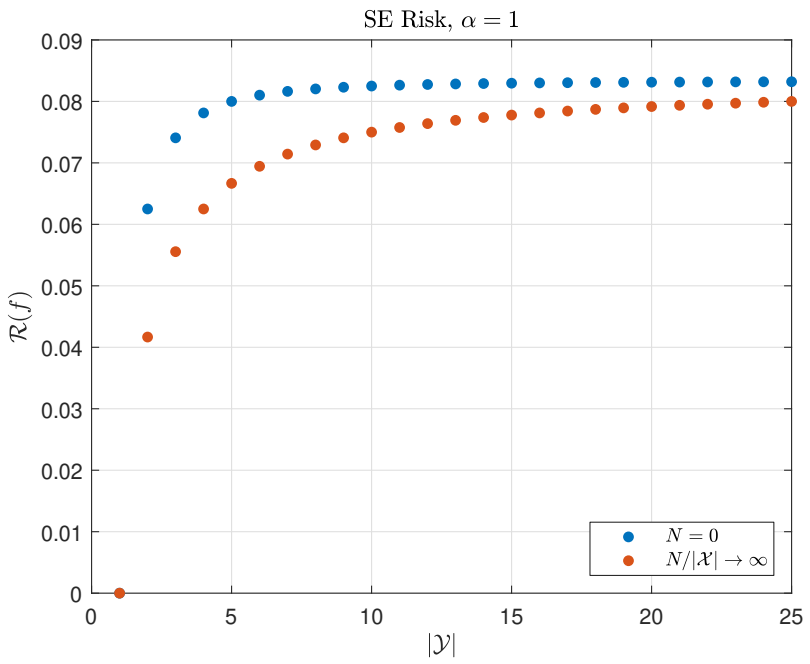

Figure 3.16: Minimum SE Risk, Uniform Prior, zero and infinite training data

The first risk is the conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ – this is intuitively satisfying as the corresponding weight becomes unity when $N = 0$. The second risk is the squared error with infinite training data. Note that the reduction of the risk between these two extreme cases is modest, and that the attenuating factor increases towards unity for applications with more possible outcomes. Fig. 3.16 illustrates the difference between these cases.

PGR: additional figures for uniform case?

3.3.1.2 Squared-Error Trends

Having derived the optimal estimator based on a Dirichlet model prior, it is important to consider the risk $\mathcal{R}_\Theta(f^*;\theta)$ and analyze how different prior parameterizations α influence the squared error for different models θ.

Substituting the second moments of $\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c})$ (3.49) into (2.46), the excess

squared error can be represented as

$$\begin{aligned}
\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y' \, \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}} \Big[\Delta(y;\mathrm{x};\mathrm{D},\theta_\mathrm{c})\Delta(y';\mathrm{x};\mathrm{D},\theta_\mathrm{c})\Big] \\
&= \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}} \left[\big(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\big)^2 \, \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}} \left[\left(\frac{\alpha_0 \, \alpha_\mathrm{m}(\mathrm{x})}{\alpha_0 \, \alpha_\mathrm{m}(\mathrm{x}) + N \, \psi_\mathrm{m}(\mathrm{x})} \right)^2 \right] \right] \\
&\quad + \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}} \left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}} \, \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}} \left[\frac{N \, \psi_\mathrm{m}(\mathrm{x})}{\big(\alpha_0 \, \alpha_\mathrm{m}(\mathrm{x}) + N \, \psi_\mathrm{m}(\mathrm{x})\big)^2} \right] \right] \\
&= \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}} \Big[\mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}} \big[\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})^2 \big] \big(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\big)^2 \Big] \\
&\quad + \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}} \left[\mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}} \left[\frac{\big(1 - \gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m}\big)^2}{N \, \psi_\mathrm{m}(\mathrm{x})} \right] \Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}} \right] . \qquad (3.64)
\end{aligned}$$

The excess risk can thus be represented as the conditional expectation (with respect to $\mathrm{P}_{\mathrm{x}\,|\theta} = \theta_\mathrm{m}$) of a sum of two functions of x. The first function is dependent on the squared bias between the clairvoyant estimate $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}$ and the data-independent estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$. This term alone is influenced by the data-independent Bayes predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha_\mathrm{c}(\mathrm{x})$. The second function measures the additional variance beyond that of the clairvoyant estimator (i.e., the irreducible squared error); like the irreducible squared error, it depends on $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}$, the conditional variance of the clairvoyant estimate for a given observation of x. These two second-order terms (of y) are scaled by factors dependent on the conditional prior localizations $\alpha_0 \, \alpha_\mathrm{m}(x)$ and on $\theta_\mathrm{m}(x)$ and N via conditional expectations with respect to $\psi_\mathrm{m}(x)$.

It is instructional to consider the trends of the squared error risk (3.64) with training data volume N and with Dirichlet prior parameterization. As these weights depend on the bias and variance scaling factors introduced in Section 3.2, the analysis performed there is directly applicable.

First consider how the excess risk changes with the training volume N. For $N = 0$, it is evident that the excess risk is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta} \left[\big(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2 \right]$, the expected squared bias between the clairvoyant and data-independent estimators. Recall that as $N \to \infty$, the bias and variance both vanish; as a result, $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to 0$. This desirable estimator property is a consequence of the full support of the Dirichlet prior,

ensuring that the model posterior concentrates at the empirical PMF.

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

$$N = \alpha_0 \left(1 - 2\alpha_0 \frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}} \right) . \tag{3.65}$$

CHECK, consider x dependency + below

Note that as the squared bias of the prior mean increases relative to the clairvoyant estimator variance, the maximizing value decreases (even below zero). Thus, the worse the prior estimate, the more likely the excess squared error will decrease monotonically with N. Conversely, if the prior estimate is accurate, a local maximum may occur and additional training data may (temporarily) compromise the estimator performance. Also consider the effect of prior concentration; informative priors with sufficiently high α_0 will not have the local maxima.

The excess risk at this potentially non-integral value would be

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \frac{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}{4\alpha_0 \left(1 - \alpha_0 \frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}} \right)} . \tag{3.66}$$

PGR: better form above???

Figs. 3.17 and 3.18 exemplify the excess squared error as a function of N for estimators based on Dirichlet priors of varying concentration α_0. The former shows local maxima for an unbiased estimator; note that higher concentration results in superior performance. The latter uses biased estimators and as such, learners based on low concentration achieve lower risk.

Next consider the effects of the Dirichlet prior parameters. The analysis will interpret the Dirichlet parameters as the conditional prior means $\alpha_\mathrm{c}(x)$ and the corresponding concentrations $\bar{\alpha}_0(x) \equiv \alpha_0\, \alpha_\mathrm{m}(x)$; the latter affects the risk through the value $\gamma_\mathrm{m}(\psi)$.

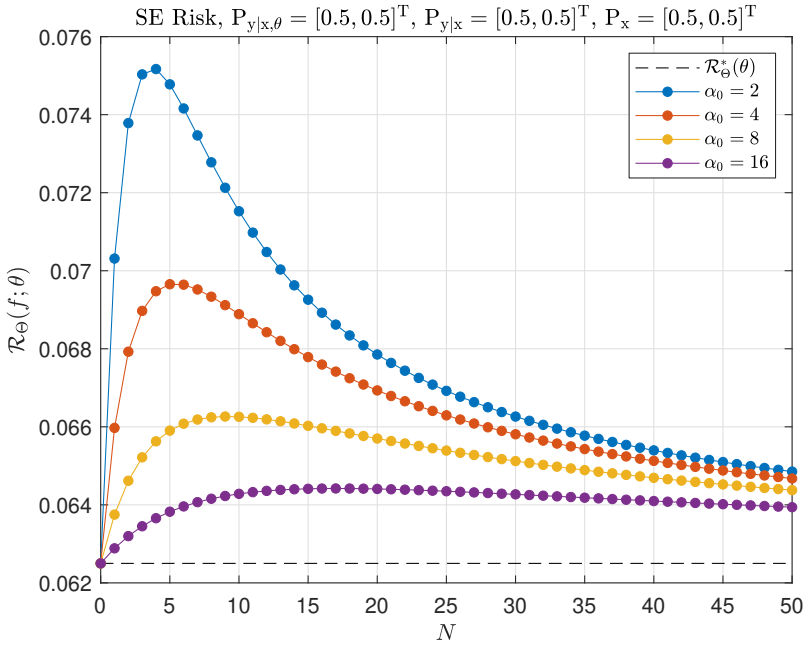

Figure 3.17: Conditional SE Risk versus N, unbiased Dirichlet estimators of varying concentration

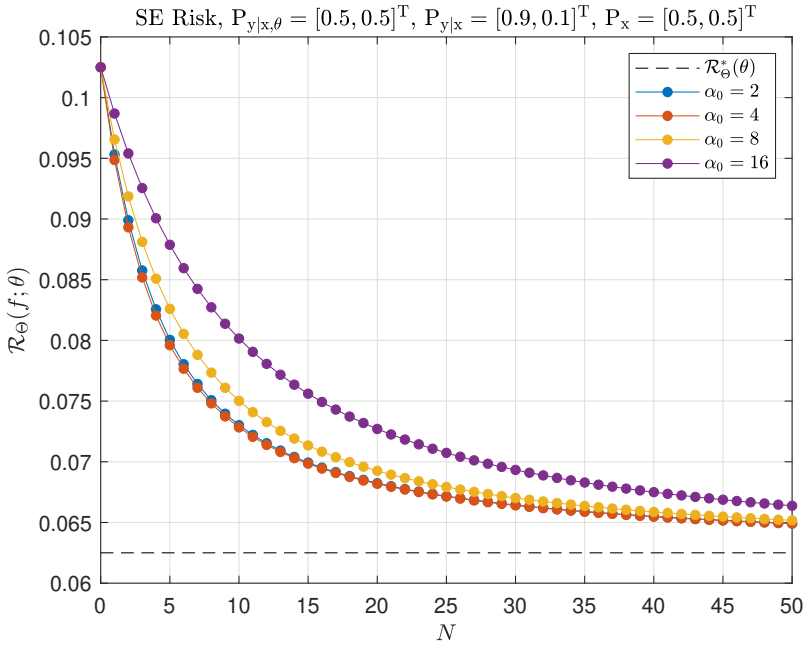

Figure 3.18: Conditional SE Risk versus N, biased Dirichlet estimators of varying concentration

First consider the conditional prior PMF's $\alpha_{\mathrm{c}}(x)$; as shown, they manifest themselves in the risk through the squared estimator bias. It is clear that regardless of how the values α_0 and $\alpha_{\mathrm{m}}(x)$ are chosen, the best selections for these conditional priors must have first moments matching those of the corresponding true predictive distributions $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ for each $x \in \mathcal{X}$. The resultant estimators $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ are unbiased and the excess risk is equivalent to the first term in (3.64), measuring additional variance due to model uncertainty.

The concentrations $\bar{\alpha}_0(x)$ of the conditional distributions $\theta_{\mathrm{c}}(x)$ control important bias-variance trade-offs via the two scaling factors in (3.64). First, consider the asymptotic trends, again referencing the distribution estimation analysis in Section 3.2.

Consider how the excess risk tends as the priors become maximally concentrated. Recall that as $\bar{\alpha}_0(x) \to \infty$, the estimate of θ_{c} is maximally biased and has no variance; thus, the excess risk tends to $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, the expected squared error between the means of the Bayesian predictive PMF and the true predictive PMF. This is intuitive given that the estimator tends toward a data-independent solution; the estimator may be biased, but will have no variance due to the training data statistics.

Conversely, if concentrations $\bar{\alpha}_0(x) \to 0$ are chosen, the Bayesian estimate tends to the empirical mean, independent of α_{c}. Using the limits displayed in Section 3.2, it can be shown that the excess risk tends to

$$\begin{aligned}\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to\ & \mathrm{E}_{\mathrm{x}\,|\,\theta_{\mathrm{m}}}\left[\left(1-\theta_{\mathrm{m}}(\mathrm{x})\right)^N\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}\right)^2\right] \\ & + \mathrm{E}_{\mathrm{x}\,|\,\theta_{\mathrm{m}}}\left[\left(\sum_{n=1}^{N}\binom{N}{n}\theta_{\mathrm{m}}(\mathrm{x})^n\left(1-\theta_{\mathrm{m}}(\mathrm{x})\right)^{N-n}\frac{1}{n}\right)\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}\right] .\end{aligned}$$

Observe that the squared bias contributes to the sum for a given value x, proportionate to the probability that no training samples are observed matching this value.

Of further interest are the values $\bar{\alpha}_0(x)$ that minimize the excess squared error for given prior conditional distributions $\alpha_{\mathrm{c}}(x)$. With the asymptotic values of the excess risk known, all that remains is to determine any local minima. Since the $|\,\mathcal{X}\,|$

summands of the excess risk depend only on the corresponding concentrations $\bar{\alpha}_0(x)$, each of these values can be optimized separately.

add the derivative details below???

Calculating the first derivative with respect to $\bar{\alpha}_0(x)$, it can be shown that for $N > 0$ and $\theta_{\mathrm{m}}(x) > 0$, only one stationary point exists, at

$$\bar{\alpha}_0(\mathrm{x}) = \frac{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}}{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}\right)^2} \,. \tag{3.67}$$

Calculation of the second derivative confirms that this value is a local minimum. Furthermore, the excess risk evaluated at these values is

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x}\,|\,\theta_{\mathrm{m}}}\left[\mathrm{E}_{\psi_{\mathrm{m}}\,|\,\theta_{\mathrm{m}}}\left[\left(N\,\psi_{\mathrm{m}}(\mathrm{x})\Sigma^{-1}_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}} + \left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}\right)^{-2}\right)^{-1}\right]\right] \,, \tag{3.68}$$

which can be easily shown to be less than both the asymptotic values for $\bar{\alpha}_0(x) \to 0$ and $\bar{\alpha}_0(x) \to \infty$. Thus the concentration values (3.67) can be used to find the optimal values of α_0 and $\alpha_{\mathrm{m}}(x)$, yielding the minimum excess risk for the given prior conditional distributions $\alpha_{\mathrm{c}}(x)$.

Note that the minimizing concentration values $\bar{\alpha}_0(x)$ 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 concentrations are preferable when the conditional model has low variance. Such models can be accurately identified by learners that prioritize the empirical mean over the prior estimate, even with limited training data volume N; quick adaptation to the data is effective, with minimal risk due to overfitting. Additionally, note that these values $\bar{\alpha}_0(x)$ do not depend on the training volume N.

Figs. 3.19 and 3.20 show how the excess squared error trends as a function of the Dirichlet learner concentration. Note that the latter is based on a biased prior estimate and thus the optimal Dirichlet concentration value is lower.

PGR: plot captions, alpha zero or x???

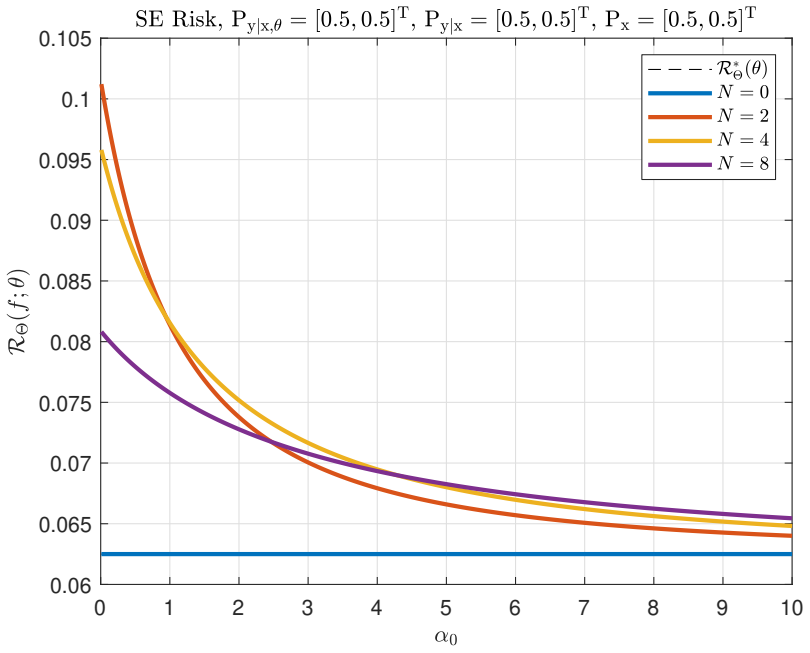

Figure 3.19: Conditional SE Risk versus $\alpha_0\,\alpha_{\mathrm{m}}(x)$, unbiased Dirichlet estimator using varying training set volumes

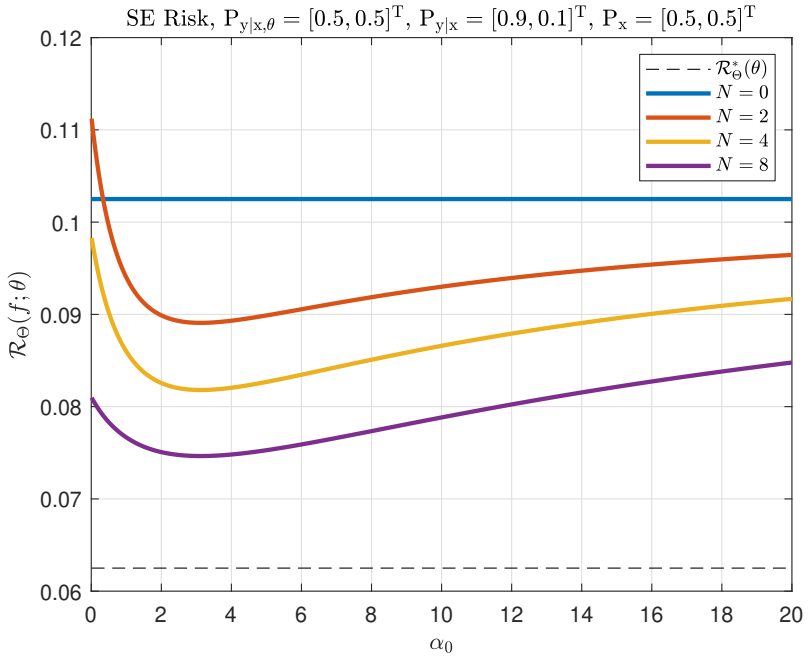

Figure 3.20: Conditional SE Risk versus α_0, biased Dirichlet estimator using varying training set volumes

3.3.1.3 Example

Use harder non-linearity that fails even for higher-order poly??

use markers for predict plots?!?

To demonstrate the efficacy of Dirichlet-based regressors, consider a data model where $\mathcal{X}$ and $\mathcal{Y}$ are 128-point discretizations of the unit interval $[0, 1]$, inclusive, such that $\mathcal{X} = \mathcal{Y} = \{i/127 : i = 0, \ldots, 127\}$. The Dirichlet estimator is parameterized by $\alpha_{\mathrm{m}} = 1/128$ and $\alpha_{\mathrm{c}}(y; x) = \mathrm{DE}\big((y, 1-y); 127, 4.164, (0.5, 0.5)\big)$, such that the prior estimator $\mu_{\mathrm{y}\,|\,\mathrm{x}} = 0.5$ is constant; various localizations α_0 will be used. Results were generated by averaging 50,000 iterations of a novel Python learning simulation.

Simulation details, reference? Github release??

Figs. 3.21 and 3.22 display the Bayesian squared error realized by the Dirichlet-based regressor. The model θ is randomly selected from a Dirichlet distribution with the same α_{m} and α_{c} used for the regressor design and with the localization fixed at $\alpha_0 = 400$. Observe that with increasing training data volume N, all the regressors tend towards the expected irreducible risk $\mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] = \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \approx 0.038$. Additionally, the regressor performs best when its localization matches $\alpha_0 = 400$, regardless of the value N, achieving the Bayesian squared error (3.60).

Next, consider the risk trends when the true model is fixed. The PMF $\theta_{\mathrm{m}} = 1/128$ is uniform and $\theta_{\mathrm{c}}(y; \mathrm{x}) = \mathrm{DE}\big((y, 1-y); 127, 4.164, (\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}, 1-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})\big)$, where the clairvoyant regressor is $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta} = 1/\big(2 + \sin(2\pi\,\mathrm{x})\big)$. Note that the Dirichlet prior estimator $\mu_{\mathrm{y}\,|\,\mathrm{x}} = 0.5$ is significantly biased. The true predictive variance is $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta} = 0.2\,\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}(1-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})$ and thus the irreducible squared error is $\mathcal{R}^*_\Theta(\theta) \approx 0.039$. For comparison, a Bayesian linear regressor (Appendix C) using $\Sigma_{\mathrm{y}} = 0.1$, basis functions $\phi(x) = (1, x)$, and a Normal prior $\mathcal{N}\big((0.5, 0), \Sigma_\theta\big)$ is evaluated, as well; different prior covariance functions Σ_θ will be used. Note that the linear regressor will also predict $\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = 0.5$ when $N = 0$.

Figs. 3.23 and 3.24 provide visualization of the statistics of the achieved regression functions. The lines represent $\mathrm{E}_{\mathrm{D}\,|\theta}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right]$, the expectation of the Bayesian regressors

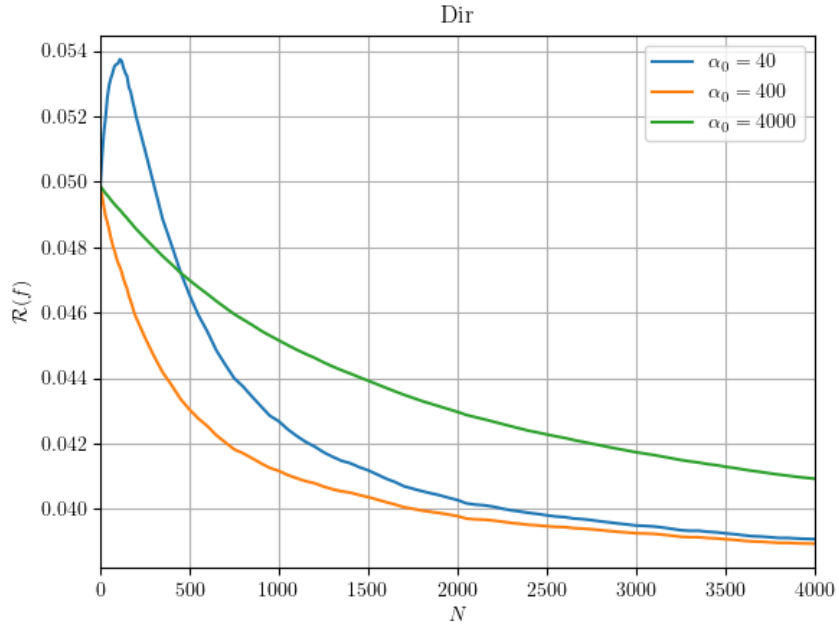

Figure 3.21: Bayes Squared-Error vs. N

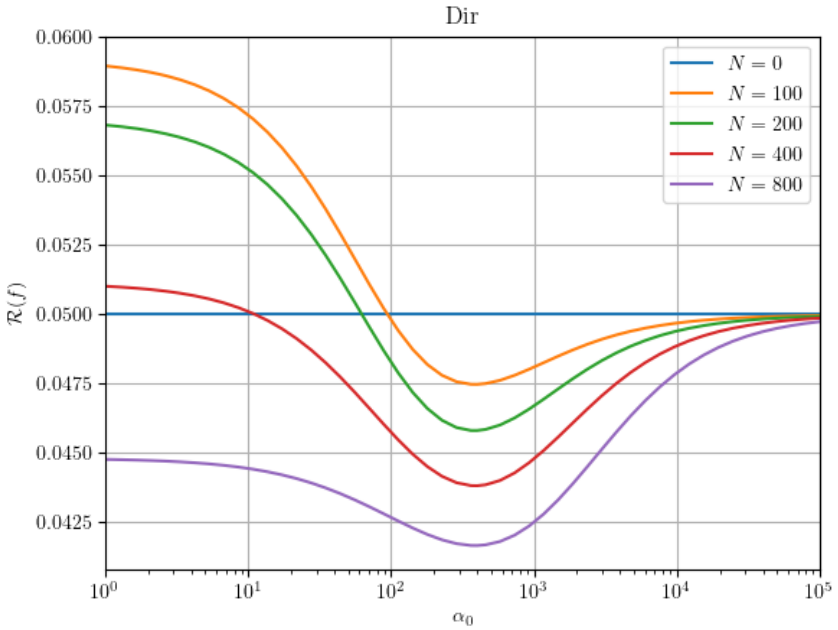

Figure 3.22: Bayes Squared-Error vs. prior localization α_0

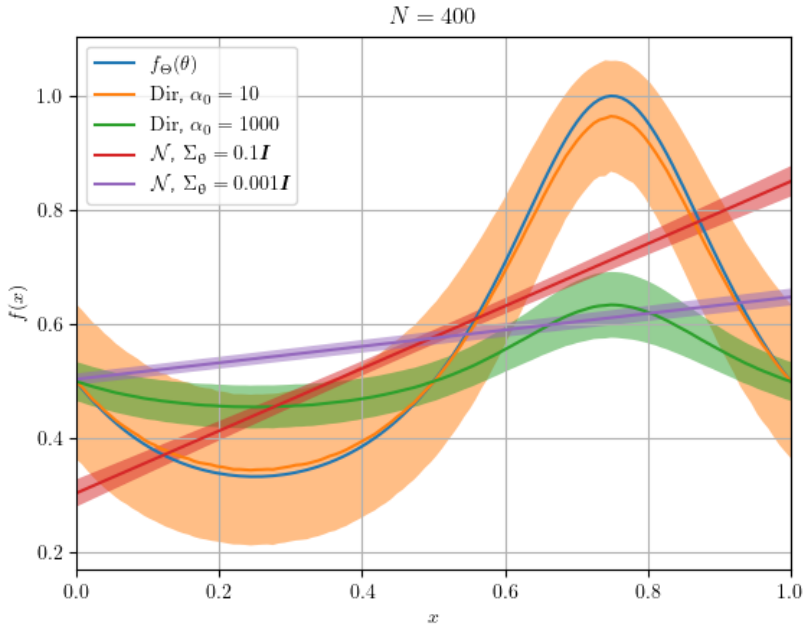

Figure 3.23: Predictor mean/variance, comparative

with respect to the training data; the filled regions represent the square-root of $\mathrm{C}_{\mathrm{D}\,|\theta}\left[\mu_{\mathrm{y}\,|\,\mathrm{x,D}}\right]$, the regressor variance. Observe that the Dirichlet-based estimator with $\alpha_0 = 10$ has lower prediction bias, but will incur additional error due to high variance. Conversely, the estimator using $\alpha_0 = 1000$ is hindered by its confidence in the biased estimator $\mu_{\mathrm{y}\,|\,\mathrm{x}}$, but is less sensitive to variations in the observed training set. Also note that both the bias and variance of the estimator tend to zero in the limit $N \to \infty$, even when using the more concentrated $\alpha_0 = 1000$ prior.

The Bayesian linear regressor with the Normal prior is highly biased, as its set of achievable estimator functions is critically limited; it also has lower variance than the Dirichlet-based estimators. This is a consequence of the lower-dimensionality, limited-support prior – there are fewer models $\theta \in \mathcal{P}(\mathcal{Y} \times \mathcal{X})$ that are considered and thus fewer functions $f_\Theta(\mathrm{x};\theta) = \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ that can be realized. Contrasting, the Dirichlet estimator uses a full-support prior, which is necessary to ensure that any complex clairvoyant estimator can be learned.

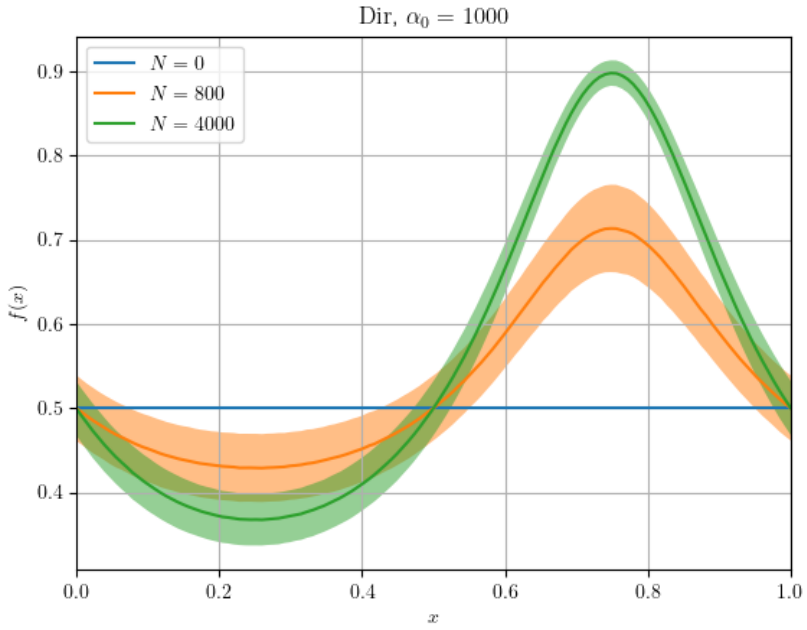

Figure 3.24: Dirichlet-based predictor mean/variance, varying N

Fig. 3.25 shows the resultant expected squared error as a function of N. Observe that the Dirichlet-based estimator trends toward the irreducible risk $\mathcal{R}^*_\Theta(\theta)$, regardless of how much confidence in the data-independent regressor $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ is indicated through the prior localization α_0. Furthermore this trend holds no matter how severe the bias of α_c might be. This is a consequence of the Bayesian predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ being a consistent estimator of the true predictive model θ_c, which is guaranteed by the Dirichlet prior's full support. In contrast, the Bayesian linear regressor will generally result in non-zero excess squared error, no matter how much training data is available.

Note that due to their initial bias, both the Dirichlet-based estimator and Bayesian linear regressor perform better when projecting relatively low confidence in their data-independent prediction via high prior variance (low α_0, high Σ_θ). Fig. 3.26 demonstrates the error trends of the Dirichlet-based estimator for different prior localizations α_0; observe that the value of α_0 that optimizes the bias-variance trade-off

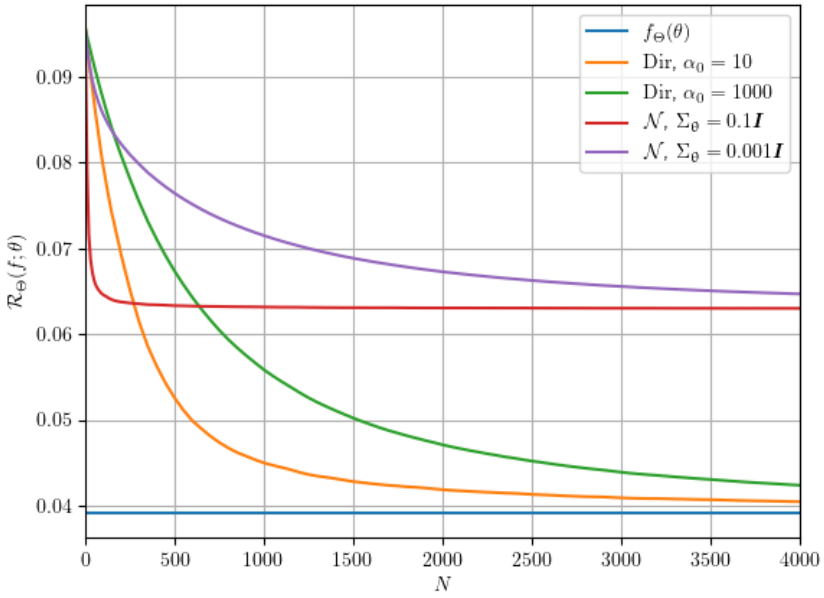

Figure 3.25: Squared-Error vs. training data volume N

is independent of the training volume N.

3.3.2 Classification: the 0-1 Loss

This section derives 0-1 loss classifiers based on the Dirichlet prior distribution and assesses their performance.

3.3.2.1 Bayesian Classification

Optimal Hypothesis: Conditional Maximum *a posteriori* PGR: decision region figures??

PGR: weighted conditional majority decision

To determine the optimal learning function, the 0-1 loss from Equation (2.47) is

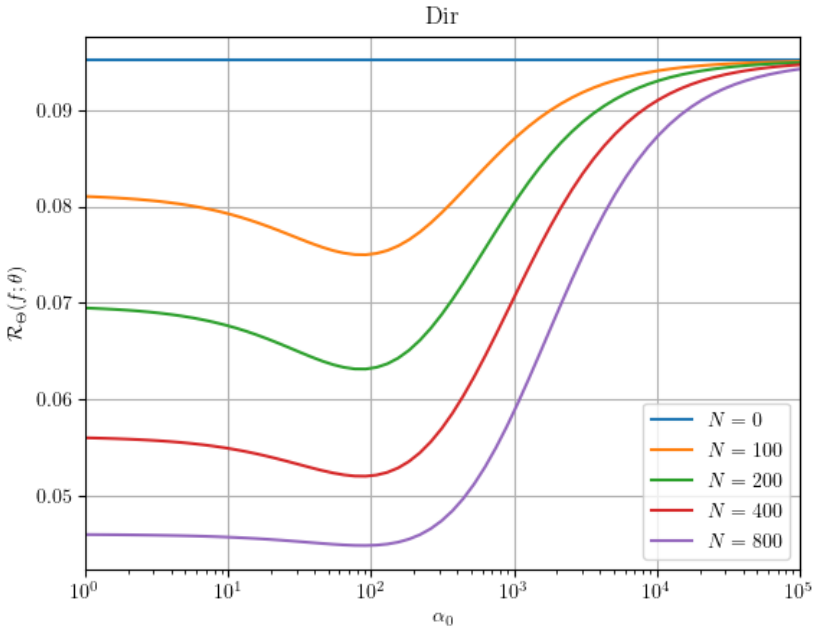

Figure 3.26: Squared-Error vs. prior localization α_0

substituted into Equation (3.55) and Equation (2.27) to find

$$\begin{aligned} f^*(x;D) &= \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|x,D) \\ &= \arg\max_{y\in\mathcal{Y}} \frac{\alpha_0\,\alpha_{\mathrm{m}}(x)\,\alpha_{\mathrm{c}}(y;x) + N\,\Psi_{\mathrm{m}}(x;D)\,\Psi_{\mathrm{c}}(y;x;D)}{\alpha_0\,\alpha_{\mathrm{m}}(x) + N\,\Psi_{\mathrm{m}}(x;D)} \\ &= \arg\max_{y\in\mathcal{Y}} \big(\alpha_0\alpha(y,x) + N\Psi(y,x;D)\big) \\ &= \arg\max_{y\in\mathcal{Y}} \big(\alpha_0\,\alpha_{\mathrm{m}}(x)\,\alpha_{\mathrm{c}}(y;x) + N\,\Psi_{\mathrm{m}}(x;D)\,\Psi_{\mathrm{c}}(y;x;D)\big)\ . \end{aligned} \tag{3.69}$$

Using the Dirichlet prior, different classes are "scored" by counting the number of training samples with a value of X_n matching that of x and combining with the prior parameters α_0 and $\alpha(\cdot,\mathrm{x})$.

Uniform Prior When the uniform prior is used, the Bayes classifier simplifies to

$$f^*(x;D) = \arg\max_{y\in\mathcal{Y}} \Psi_{\mathrm{c}}(y;x;D)\ , \tag{3.70}$$

the maximizing argument of the conditional empirical model. This effects a conditional majority decision which chooses the class from $\mathcal{Y}$ most often represented among training set samples D with a matching input value x. This is intuitive, as the model PDF parameter α imparts no confidence as to which classes may be most likely.

Minimum Bayes Risk: Probability of Error

no closed-forms found??? Find closed-form BOUNDS???

Evaluating the minimum Bayes risk (2.53) using the distributions derived from the Dirichlet prior, the Bayes minimum probability of error is

$$\begin{aligned}\mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \\ &= 1 - \mathrm{E}_{\mathrm{x},\psi_\mathrm{m},\psi_\mathrm{c}}\left[\frac{\max_{y\in\mathcal{Y}}\left(\alpha_0\,\alpha_\mathrm{m}(x)\,\alpha_\mathrm{c}(y;x) + N\,\psi_\mathrm{m}(x)\psi(y;x)\right)}{\alpha_0\,\alpha_\mathrm{m}(x) + N\,\psi_\mathrm{m}(x)}\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \frac{\mathrm{E}_\psi\left[\max_{y\in\mathcal{Y}}\left(\alpha_0\alpha(y,x) + N\psi(y,x)\right)\right]}{\alpha_0 + N}\,.\end{aligned} \tag{3.71}$$

Figs. 3.27 and 3.28 plot the minimum Bayes probability of error against training data volume N and prior concentration α_0, respectively. Note that for $N = 0$, the Bayes risk is $\mathcal{R}^* = 1 - \sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \alpha(y,x)$. Additionally, consider the risk for maximal/minimal values of the Dirichlet concentration. For $\alpha_0 \to 0$ (and $N > 1$), the risk is $\mathcal{R}^* = 0$; conversely, for $\alpha_0 \to \infty$, the risk tends to $\mathcal{R}^* \to 1 - \sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \alpha(y,x)$. These trends can be visualized in Figs. 3.29 and 3.30.

PGR: risk for $N \to \infty$?

PGR: missing info for Dir gen graphics? fixed y given x conditional alpha???

PGR: comment on simulation!

Uniform Prior PGR: COMPUTATIONAL COMPLEXITY savings for risk formula?

PGR: Can uniform minimal risk be approximated as a function of My and Mx/N, as is for SE loss???

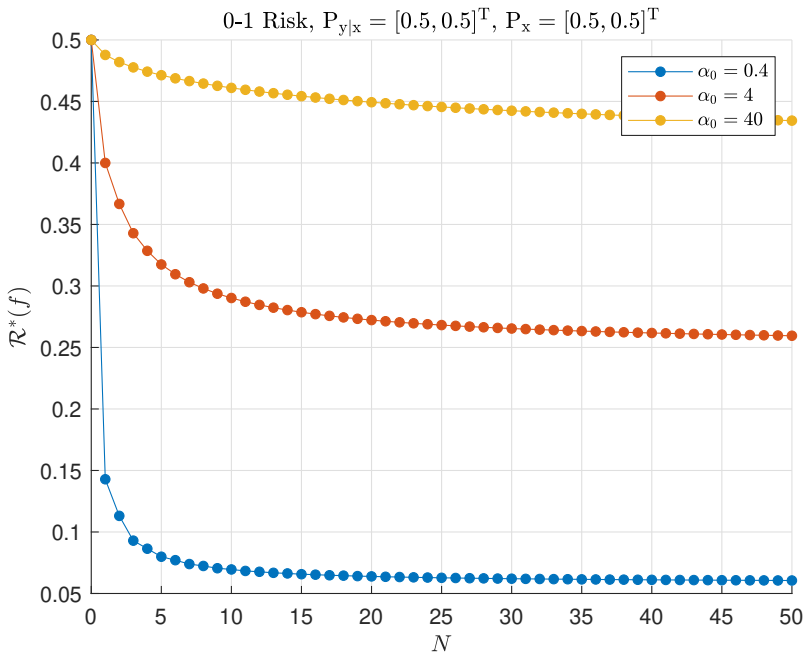

Figure 3.27: Minimum 0-1 Risk for different training data volumes N

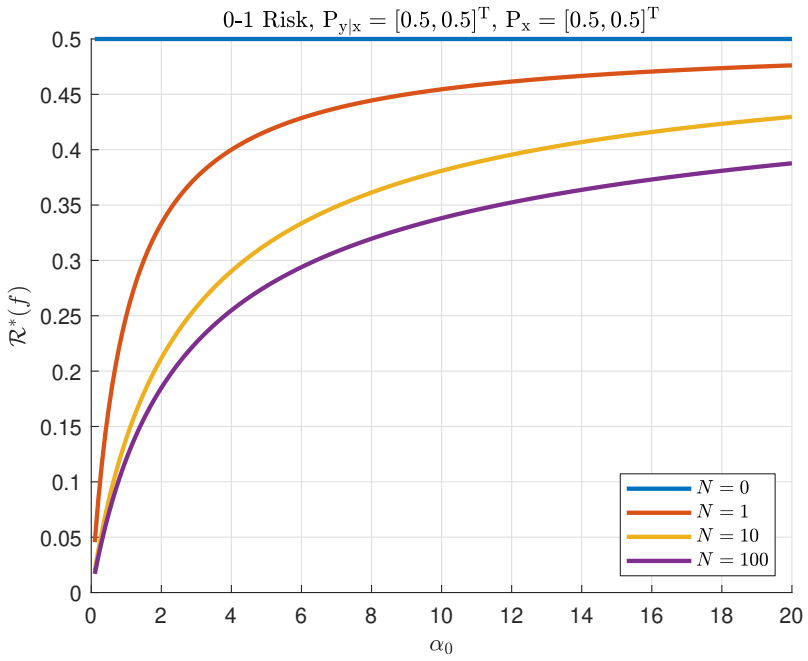

Figure 3.28: Minimum 0-1 Risk for different prior concentrations α_0

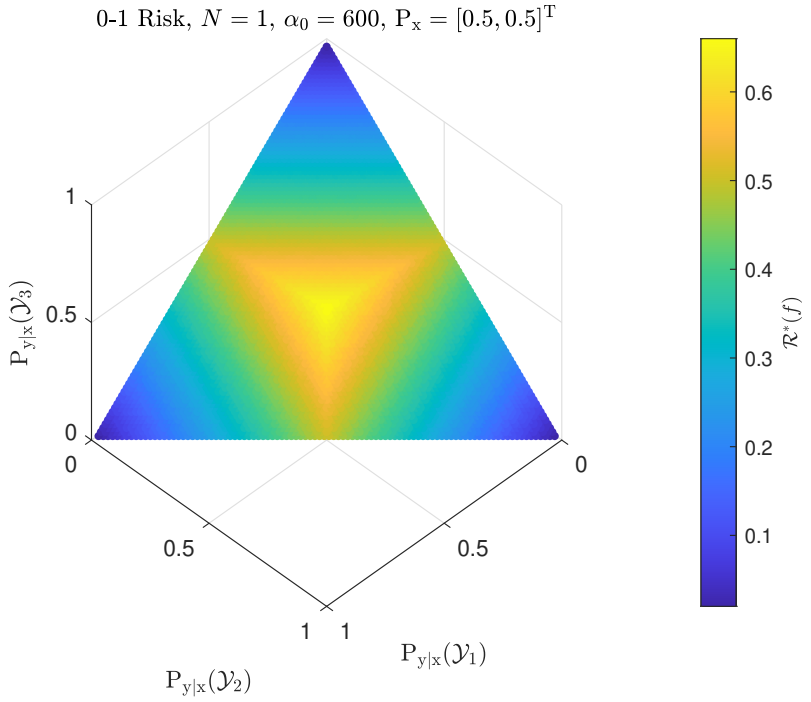

Figure 3.29: Minimum 0-1 Risk for different prior means $\mathrm{P_{y\,|\,x}}$

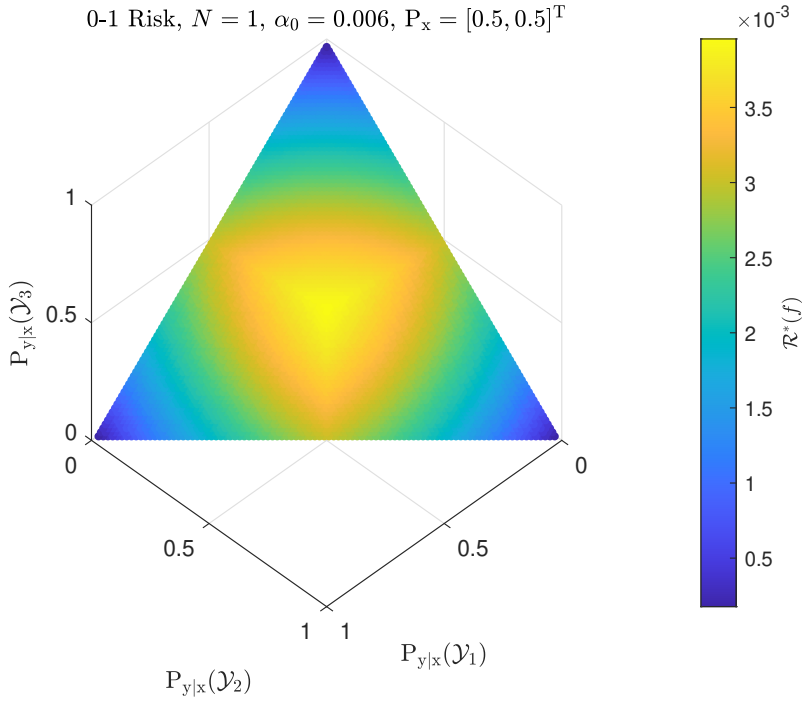

Figure 3.30: Minimum 0-1 Risk for different prior means $\mathrm{P_{y\,|\,x}}$

PGR: use Mcal not binom!

PGR: add nmax CDF fig!

Using the uniform prior, the minimum Bayes 0-1 risk is

$$\begin{aligned}\mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \frac{1 + N\,\mathrm{E}_\psi\left[\max_{y\in\mathcal{Y}}\psi(y,x)\right]}{|\,\mathcal{Y}\,||\,\mathcal{X}\,| + N} \\ &= 1 - \frac{1 + N|\,\mathcal{X}\,|^{-1}\sum_{x\in\mathcal{X}}\mathrm{E}_\psi\left[\max_{y\in\mathcal{Y}}\psi(y,x)\right]}{|\,\mathcal{Y}\,| + N|\,\mathcal{X}\,|^{-1}} \\ &= 1 - \frac{1 + N|\,\mathcal{X}\,|^{-1}\sum_{x\in\mathcal{X}}\mathrm{E}_{\psi_\mathrm{m}(x)}\left[\psi_\mathrm{m}(x)\,\mathrm{E}_{\psi_\mathrm{c}(x)|\,\psi_\mathrm{m}(x)}\left[\max_{y\in\mathcal{Y}}\psi_\mathrm{c}(y;x)\right]\right]}{|\,\mathcal{Y}\,| + N|\,\mathcal{X}\,|^{-1}}\,.\end{aligned} \tag{3.72}$$

The expectation operates on the maximum value from a subset of a uniform Dirichlet-Empirical random process. Via the Dirichlet-Empirical aggregation property (related to the Dirichlet-Multinomial property [9]), a consequence of the the uniform PMF P_ψ is that the individual segments $\psi(\cdot,x)$ are identically distributed; thus, the expectation will be same for every value x.

To evaluate this expectation, new random variables $\psi_{\max}(x) \equiv \max_{y\in\mathcal{Y}}\psi(y,x) \in \{n/N : n \in 0,\ldots,N\}$ are introduced and characterized by their identical PMF. To this end, the probability of the event $\mathrm{P}\left(\psi_{\max}(x) \geq n/N\right) = \mathrm{P}\left(\cup_{y\in\mathcal{Y}}\{\psi(y,x) \geq n/N\}\right)$ will be determined. As the distribution of ψ is uniform, the event probability is proportionate to the cardinality of the set $\cup_{y\in\mathcal{Y}}\{\psi : \psi(y,x) \geq n/N\}$. Using the inclusion-exclusion principle [5], the cardinality is represented as

$$\begin{aligned}&\left|\,\cup_{y\in\mathcal{Y}}\{\psi : \psi(y,x) \geq n/N\}\right| \\ &\quad= \begin{cases} \binom{N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} & \text{if } n < 0, \\ \sum_{m=1}^{|\,\mathcal{Y}\,|}\binom{|\,\mathcal{Y}\,|}{m}(-1)^{m-1}\binom{N-mn+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}H\left(\left\lfloor\frac{N}{m}\right\rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N, \end{cases}\end{aligned} \tag{3.73}$$

where $H : \mathbb{Z} \mapsto \{0,1\}$ is the discrete Heaviside step function. For $n < 0$, the cardinality is equivalent to $|\Psi|$.

For $0 \leq n < N$, the cardinality is an alternating binomial summation where the m^{th} term accounts for the different intersections of m of the $|\,\mathcal{Y}\,|$ individual sets $\{\psi : \psi(y,x) \geq n/N\}$. Observe that the cardinality of the intersections is only dependent on the number of contributing sets m and not on which sets intersect. Furthermore, note the dependency of the intersection cardinalities on the argument n. The step function contributes such that if $n > \lfloor \frac{N}{m} \rfloor$, only up to $m-1$ individual sets will intersect. The binomial coefficient $\mathcal{M}\left((N - mn, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1)\right)$ provides the intersection cardinality for a given m; note the similarity to the cardinality $|\Psi|$ – the only difference is the number of points that grid the $|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1$ dimensional region.

The probability of interest can thus be expressed as

$$\mathrm{P}\left(\psi_{\max}(x) \geq n/N\right) = \binom{N + |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1}^{-1} \left|\, \cup_{y \in \mathcal{Y}} \{\psi : \psi(y,x) \geq n/N\} \right| \quad (3.74)$$

$$= \begin{cases} 1 & \text{if } n < 0, \\ \sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right) H\left(\lfloor \frac{N}{m} \rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N. \end{cases}$$

PGR: use Mcal op?

PGR: Heaviside reference?

As the PMF of $\psi_{\max}(x)$ has support on $\{n/N : n \in 0, \ldots, N\}$, the expectation over ψ is evaluated as

$$\begin{aligned} \mathrm{E}_{\psi}\left[\psi_{\max}(x)\right] &= \sum_{n=0}^{N} \frac{n}{N} \Big(\mathrm{P}\left(\psi_{\max}(x) \geq n/N\right) - \mathrm{P}\left(\psi_{\max}(x) \geq (n+1)/N\right) \Big) \quad (3.75) \\ &= -\frac{1}{N} + \frac{1}{N} \sum_{n=0}^{N} \mathrm{P}\left(\psi_{\max}(x) \geq n/N\right) \\ &= -\frac{1}{N} + \frac{1}{N} \sum_{m=1}^{|\mathcal{Y}|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} \sum_{n=0}^{\lfloor \frac{N}{m} \rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right) \end{aligned}$$

and the minimum 0-1 risk is

$$\mathcal{R}^* = 1 - \frac{\sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \sum_{n=0}^{\lfloor \frac{N}{m} \rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right)}{|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|}\,. \quad (3.76)$$

It is instructional to express the risk for minimal and maximal volumes of training data. Using the binomial summation identity [8]

$$\sum_{m=0}^{M} \binom{M}{m} (-1)^m g(m) = 0 \,, \tag{3.77}$$

where g is a polynomial function of degree less than M, it can be shown that for $N = 0$, the minimum Bayes risk is $\mathcal{R}^* = 1 - |\,\mathcal{Y}\,|^{-1}$. This is sensible, as the classes are equiprobable with $\mathrm{P_y} = |\,\mathcal{Y}\,|^{-1}$.

PGR: use ruiz citation for identity?

PGR: find min Bayes risk explicitly from theta?

To find the risk for $N \to \infty$, note that

$$\begin{aligned} &\lim_{N\to\infty} \left(|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|\right)^{-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \left(1 - \frac{mn}{N+l}\right) \\ &\quad = \lim_{N/m\to\infty} \frac{|\,\mathcal{X}\,|}{m} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \left(1 - \frac{mn}{N}\right)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \frac{m}{N} \\ &\quad = \frac{|\,\mathcal{X}\,|}{m} \int_0^1 (1-t)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \mathrm{d}t \\ &\quad = \frac{1}{m|\,\mathcal{Y}\,|} \,. \end{aligned} \tag{3.78}$$

The minimum Bayes probability of error for the uniform prior tends toward

$$\begin{aligned} \mathcal{R}^* &\to 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\,\mathcal{Y}\,|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} m^{-1} \\ &= 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\,\mathcal{Y}\,|} m^{-1} \,, \end{aligned} \tag{3.79}$$

providing a lower bound for the achievable 0-1 Bayes risk. The above formulation has made use of the alternating summation identity from [16] to display the risk with a form including the $|\,\mathcal{Y}\,|^{\text{th}}$ harmonic number $H_{|\,\mathcal{Y}\,|} \equiv \sum_{m=1}^{|\,\mathcal{Y}\,|} m^{-1}$. Observe that the minimum Bayes risk does not depend on the cardinality $|\,\mathcal{X}\,|$.

harmonic reference

Fig. 3.31 demonstrates how the minimum 0-1 risk decreases with training volume N; observe that the risk is more severe for sequences corresponding to higher $|\,\mathcal{Y}\,|$. It

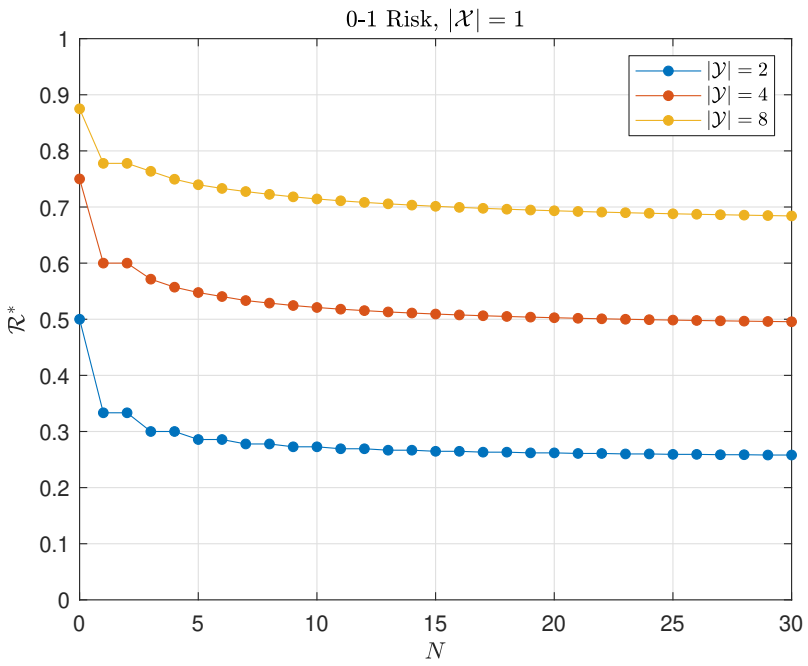

Figure 3.31: Minimum 0-1 Risk vs training set volume N

is sensible that the probability of error should increase when more classes have to be considered. Fig. 3.32 illustrates the Bayes risk with multiple sequences for different cardinalities $|\mathcal{X}|$. Note that risk increases with $|\mathcal{X}|$. Considering $N \mathrm{E}_{\mathrm{X}}\left[\Psi_{\mathrm{m}}(\mathrm{X})\right] = N\mu_{\Psi_{\mathrm{m}}} = N/|\mathcal{X}|$, this should be intuitive – each conditional empirical distribution $\Psi_{\mathrm{c}}(x; D)$ is forced to approximate $\theta_{\mathrm{c}}(x)$ with fewer data.

Further insight into how $|\mathcal{X}|$ affects the risk can be acquired by plotting the risk as a function of $N/|\mathcal{X}|$. In Fig. 3.33, it is shown that the minimal risk can be approximated by a function dependent only on $N/|\mathcal{X}|$; of the series plotted, only the series for $|\mathcal{X}| = 1$ shows notable non-negligible from the others.

It is also useful to graph the $N = 0$ and $N \to \infty$ Bayes risk as a function of $|\mathcal{Y}|$; both formulas are independent of $|\mathcal{X}|$. Fig. 3.34 displays these bounds; note the margin in the probability of error between the optimal $N = 0$ and $N \to \infty$ classifiers. For $|\mathcal{Y}| = 2$ binary classification, both sequences are at their minimum and infinite training data provides a reduction in expected probability of error from 0.5 to 0.25. As $|\mathcal{Y}|$ increases, the classification risk for both the $N = 0$ and $N \to \infty$ cases tend to

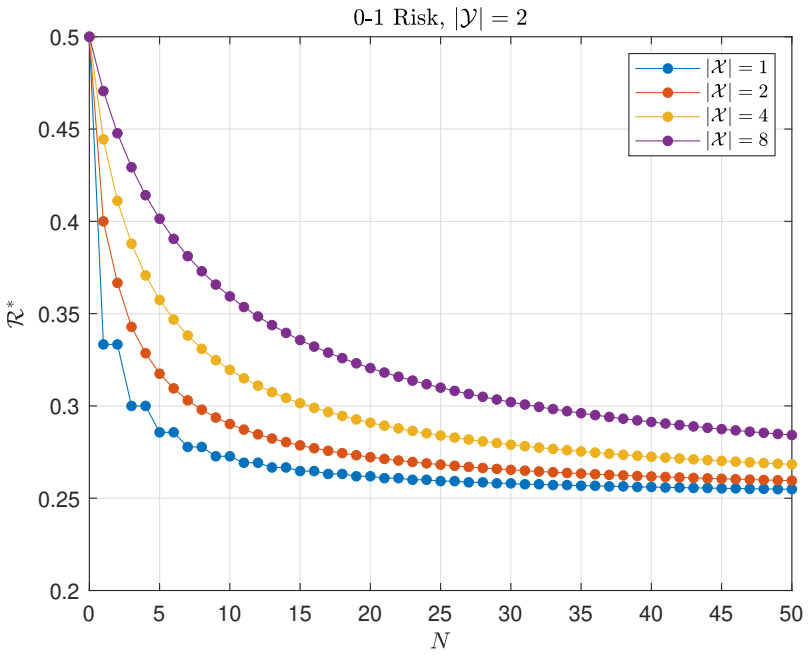

Figure 3.32: Minimum 0-1 Risk vs training set volume N

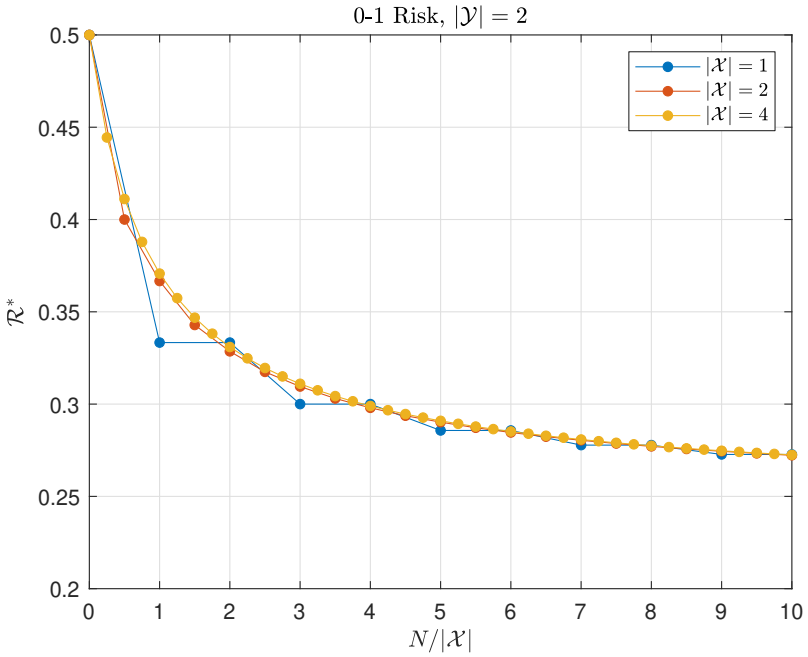

Figure 3.33: Minimum 0-1 Risk vs $N/|\,\mathcal{X}\,|$

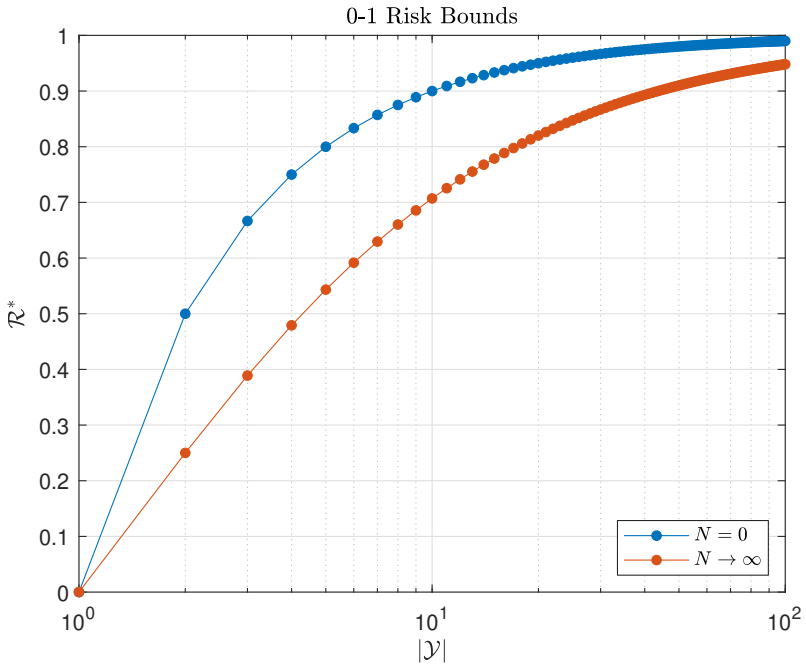

Figure 3.34: Minimum 0-1 Risk vs $|\mathcal{Y}|$

unity and the error reduction for $N \to \infty$ decreases.

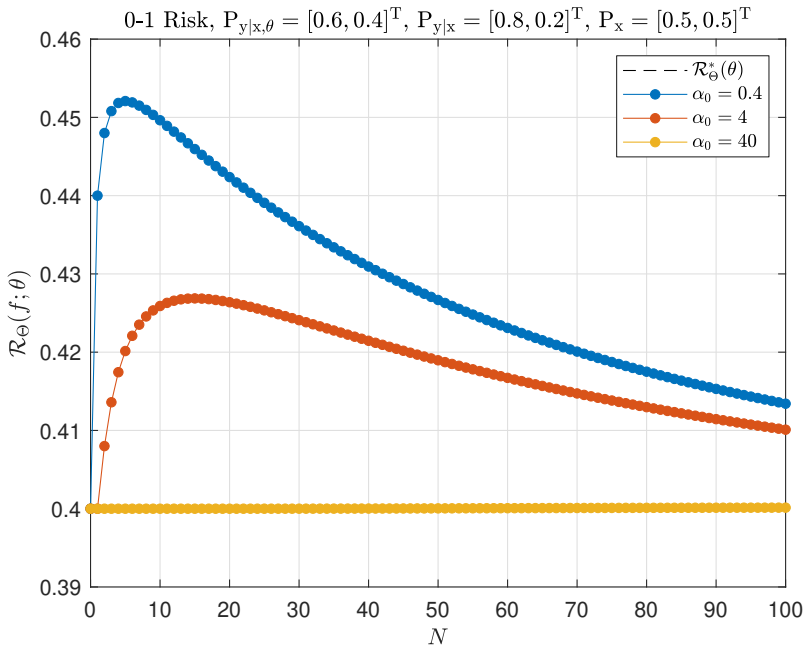

Figure 3.35: Excess probability of error, well-matched informative Dirichlet-based classifier

PGR: newpage

3.3.2.2 Probability of Error Trends

PGR: INCOMPLETE

PGR: comment on alpha0 versus alphax simplification

Substituting the optimal Dirichlet-based classifier into the formula for the probability of error (2.48), the risk is

$$\mathcal{R}_\Theta(f;\theta) = 1 - \sum_{x\in\mathcal{X}} \theta_m(x)\, E_{\psi|\,\theta_m,\theta_c}\left[\theta_c\left(\arg\max_{y\in\mathcal{Y}}\left(N\psi(y,x) + \alpha_0\alpha(y,x)\right);x\right)\right] \quad (3.80)$$

Figs. 3.35 and 3.36 show how the risk trends for classifiers based on well-matched and poorly-matched informative Dirichlet priors, respectively. Note that the well-matched prior does better with higher prior concentrations α_0; this is reflective of the fact that the maximizing arguments $y \in \mathcal{Y}$ of both the true model $\theta_c(x)$ and the prior mean $\alpha_c(x)$ are the same.

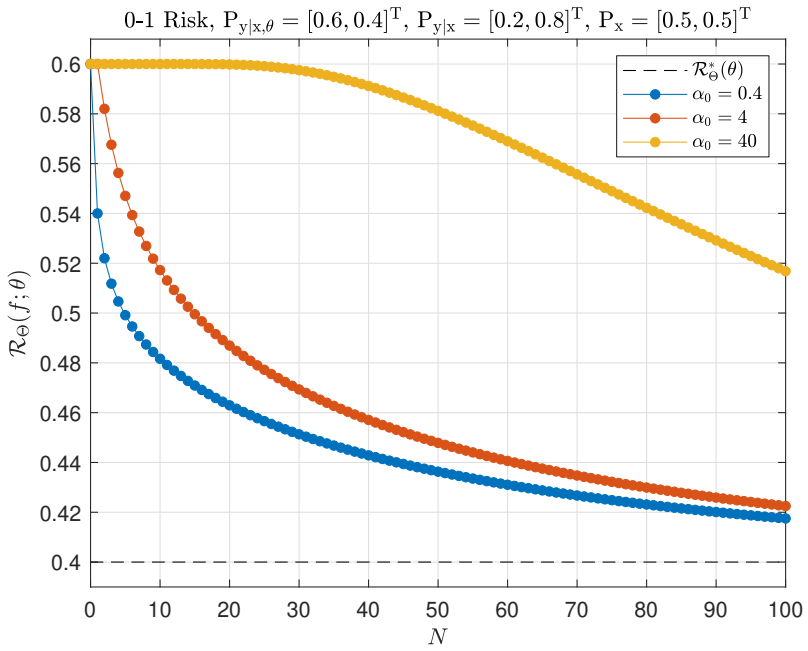

Figure 3.36: Excess probability of error, poorly-matched informative Dirichlet-based classifier

Also, it is important to consider how a given classifier performs for varying models $\theta_{\mathrm{c}}(x)$. Figs. 3.37 and 3.38 demonstrate the excess probability of error achieved by the conditional majority decision (based on a non-informative Dirichlet prior) and by a classifier derived from an informative Dirichlet prior, respectively. Note that while the former has fewer models for which the error is critically high, the latter has more models for which the irreducible risk $\mathcal{R}^*_\Theta(\theta)$ is achieved. This a fundamental trade-off between Bayesian learners based on non-informative versus informative priors.

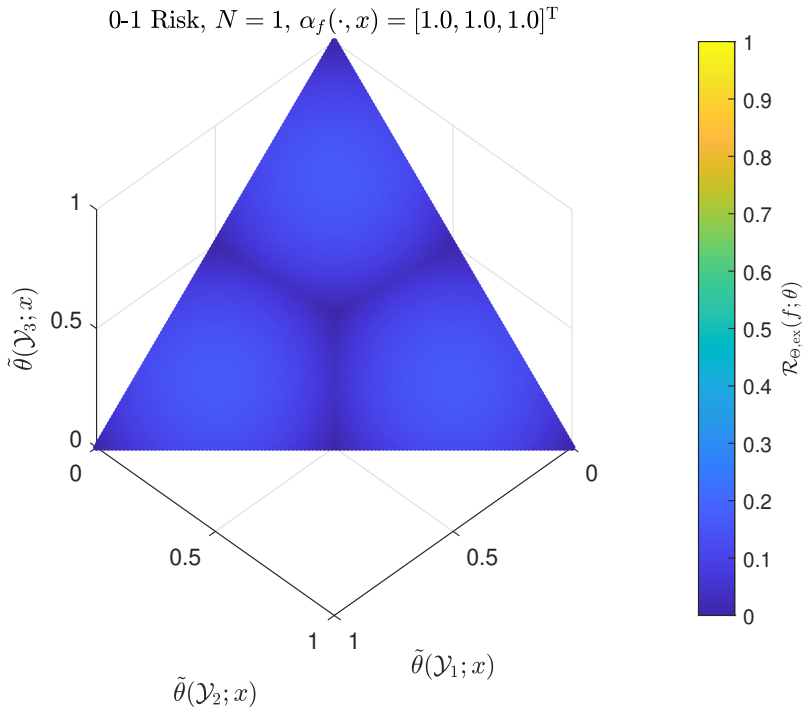

Figure 3.37: Excess probability of error, conditional majority decision

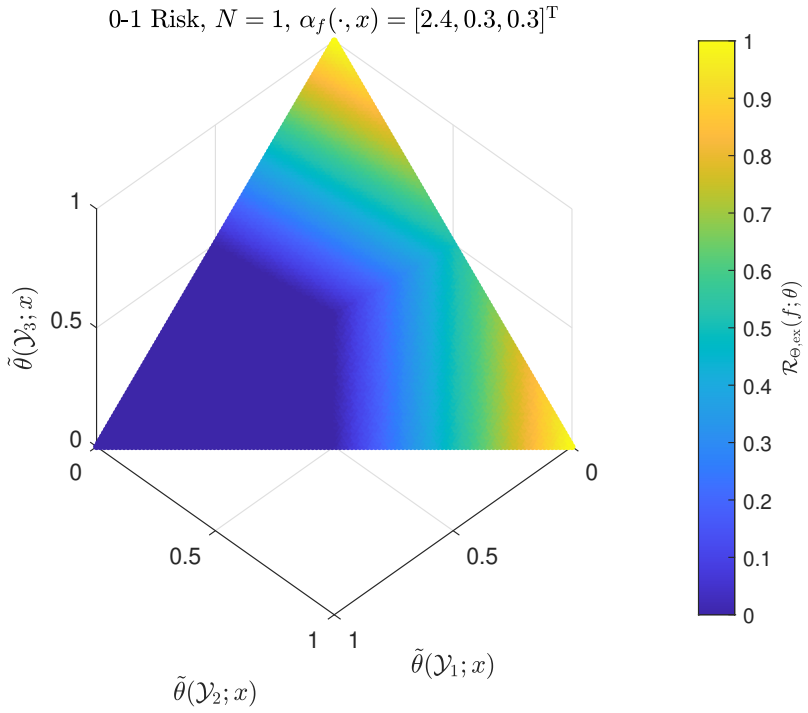

Figure 3.38: Excess probability of error, informative Dirichlet-based classifier

Chapter 4

Continuous-Domain Dirichlet Model

PGR: SPECIFY EUCLIDEAN/HILBERT??

This chapter extends further to the case where $\mathcal{Y}$ and $\mathcal{X}$ are continuous spaces and the model θ is a continuous-domain random process. Note that $|\mathcal{Y}| \geq \aleph_1$ and $|\mathcal{X}| \geq \aleph_1$.

4.1 Problem PGR MOD?

LOCATION?? Up front?

4.1.1 Model

The model θ is a PDF and has a continuous domain; the space $\Theta \equiv \mathcal{P}(\mathcal{Y} \times \mathcal{X})$ is an infinite-dimensional function space. Thus, the marginal model is defined as $\theta_{\mathrm{m}} \equiv \int_{\mathcal{Y}} \theta(y, \cdot) \mathrm{d}y \in \mathcal{P}(\mathcal{X})$.

4.1.2 Empirical Sufficient Statistic

The training data is distributed as

$$\begin{aligned} \mathrm{p}_{\mathrm{D}\,|\theta}\left(D|\theta\right) &= \prod_{n=1}^{N} \mathrm{p}_{\mathrm{D}_n\,|\theta}\left(D_n|\theta\right) = \prod_{n=1}^{N} \theta(D_n) \\ &= \exp\left(\sum_{n=1}^{N} \ln\left(\theta(D_n)\right)\right) \\ &= \exp\left(\iint_{\mathcal{Y}\times\mathcal{X}} N\Psi(y,x;D)\ln\left(\theta(y,x)\right)\mathrm{d}y\mathrm{d}x\right) \\ &\equiv \prod_{\mathcal{Y}\times\mathcal{X}} \left(\theta(y,x)^{N\Psi(y,x;D)}\right)^{\mathrm{d}y\mathrm{d}x} , \end{aligned} \tag{4.1}$$

where the operator $\prod$ is the geometric integral, the continuous analog of the discrete product operator.

CITE for geometric integral!! Just remove if unused?

Note that the dependency on the training data is expressed via the empirical transform $\Psi : \mathcal{D} \mapsto \Psi \subset \Theta$, redefined for continuous data as

$$\begin{aligned} \Psi(D) &= \frac{1}{N}\sum_{n=1}^{N} \delta\big(\,\cdot - D_n\big) \\ &\equiv \frac{1}{N}\sum_{n=1}^{N} \delta(\cdot - Y_n)\delta(\cdot - X_n)\ . \end{aligned} \tag{4.2}$$

Define the new random process $\psi \equiv \Psi(\mathrm{D}) \in \Psi$. Since the likelihood function only depends on the data through $\Psi(\mathrm{D})$, the data is conditionally independent of the model θ given ψ; consequently, the empirical model ψ is a sufficient statistic.

PDF for psi using geometric integral for Mcal?

As shown in Appendix B.1, given θ, the empirical model is a continuous-domain empirical process $\psi|\theta \sim \mathrm{EP}(N,\theta)$. The mean and covariance functions take the same form as those of the discrete Empirical process (using the continuous set form of the diag operator.)

4.1.2.1 Marginal and Conditional Data Distributions

The marginal and conditional distributions of the data using the representation $\mathrm{D} \Leftrightarrow (\mathrm{Y}, \mathrm{X})$ are of use.

The dependency of

$$\begin{aligned} \mathrm{p}_{\mathrm{X}\,|\,\theta}\left(X|\theta\right) &\equiv \prod_{n=1}^{N} \mathrm{p}_{\mathrm{X}_n\,|\,\theta_\mathrm{m}}\left(X_n|\,\theta_\mathrm{m}\right) = \prod_{n=1}^{N} \theta_\mathrm{m}(X_n) \\ &= \prod_{\mathcal{X}} \left(\theta_\mathrm{m}(x)^{N\,\Psi_\mathrm{m}(x;X)}\right)^{\mathrm{d}x} \ , \end{aligned} \tag{4.3}$$

on X is expressed via the marginal empirical statistic $\Psi_\mathrm{m} : \mathcal{X}^N \mapsto \Psi_\mathrm{m} \subset \mathcal{P}(\mathcal{X})$, defined as

$$\Psi_\mathrm{m}(X) = \frac{1}{N}\sum_{n=1}^{N} \delta\big(\,\cdot - X_n\big) = \int_{\mathcal{Y}} \Psi(y, \cdot; D)\mathrm{d}y \ . \tag{4.4}$$

Note that the dependency on θ is only through the marginal model θ_m.

The conditional distribution of the values Y given the corresponding X and the model θ,

$$\begin{aligned} \mathrm{p}_{\mathrm{Y}\,|\,\mathrm{X},\theta}\left(Y|X,\theta\right) &= \prod_{n=1}^{N} \frac{\mathrm{p}_{\mathrm{Y}_n,\mathrm{X}_n\,|\theta}\left(Y_n, X_n|\theta\right)}{\mathrm{p}_{\mathrm{X}_n\,|\theta}\left(X_n|\theta\right)} = \prod_{n=1}^{N} \theta_\mathrm{c}(Y_n; X_n) \\ &= \prod_{\mathcal{X}} \left(\prod_{\mathcal{Y}} \theta_\mathrm{c}(y; x)^{\Psi_\mathrm{c}(y;x;Y,X)\mathrm{d}y} \right)^{N\,\Psi_\mathrm{m}(x;X)\mathrm{d}x} \end{aligned} \tag{4.5}$$

depends only on the conditional models $\theta_\mathrm{c}(x)$. The dependency on the data D can be expressed using Ψ_m and $\Psi_\mathrm{c} : \{\mathcal{Y} \times \mathcal{X}\}^N \mapsto \Psi_\mathrm{c} \subset \mathcal{P}(\mathcal{Y})^\mathcal{X}$, defined as

$$\begin{aligned} \Psi_\mathrm{c}(x; Y, X) &= \frac{\Psi(\cdot, x; Y, X)}{\Psi_\mathrm{m}(x; X)} \\ &= \frac{\sum_{n=1}^{N} \delta(\cdot - Y_n)\delta(x - X_n)}{\sum_{n=1}^{N} \delta(x - X_n)} = \frac{\sum_{n=1}^{N} \delta(\cdot - Y_n)\delta[x, X_n]}{\sum_{n=1}^{N} \delta[x, X_n]} \ . \end{aligned} \tag{4.6}$$

As in the discrete-domain case, the conditional empirical distribution is defined only for observations $x \in \mathcal{X}_\mathrm{s}\left(\,\Psi_\mathrm{m}(X)\right) = \bigcup_{n=1}^{N}\{X_n\}$.

The same bijection can be used to decompose the empirical process into marginal and conditional empirical processes. The "marginalized" random process $\psi_\mathrm{m} \in \Psi_\mathrm{m} \subset$

$\mathcal{P}(\mathcal{X})$ is now defined as $\psi_{\mathrm{m}} \equiv \int_{\mathcal{Y}} \psi(y, \cdot)\mathrm{d}y$. Using the aggregation property of empirical processes detailed in Appendix B.1, it can be shown that conditioned on the model θ, the marginal model is also an Empirical process, $\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}} \sim \mathrm{EP}(N, \theta_{\mathrm{m}})$. Note the conditional independence from θ_{c}.

Additionally, given ψ_{m} and the model θ, the conditional process $\psi_{\mathrm{c}} \in \Psi_{\mathrm{c}}$ is comprised of independent Empirical processes $\psi_{\mathrm{c}}(x) |\, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x) \sim \mathrm{EP}\left(\delta(0)^{-1} N\, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x)\right)$. They are independent of the marginal model θ_{m}. Note that the conditional empirical process $\psi_{\mathrm{c}}(x)$ is characterized by the number of matching samples $\delta(0)^{-1} N\, \psi_{\mathrm{m}}(x) \equiv \sum_{n=1}^{N} \delta\big[x, X_n\big]$.

delta domain? dx?

4.2 Probability Distributions

4.2.1 Model θ Characterization

The model is characterized by a Dirichlet process $\theta \sim \mathrm{DP}(\alpha_0, \alpha)$ with concentration α_0 and mean function $\alpha \in \mathcal{P}(\mathcal{Y} \times \mathcal{X})$. In Appendix B.2, it is shown that the mean and covariance functions of the continuous process have the same form as those of the discrete process (using the continuous variant of the diag operator).

4.2.1.1 Marginal and Conditional Distributions

The marginal distribution θ_{m} and the conditional distribution θ_{c} are also of interest. Define the bijection $\alpha \Leftrightarrow (\alpha_{\mathrm{m}}, \alpha_{\mathrm{c}})$, where $\alpha_{\mathrm{m}} \equiv \int_{\mathcal{Y}} \alpha(y, \cdot)\mathrm{d}y$ (and $\alpha_{\mathrm{c}}(x) \equiv \alpha(\cdot, x) / \alpha_{\mathrm{m}}(x)$) for each $x \in \mathcal{X}$. Again, $\alpha_{\mathrm{m}} \in \mathcal{P}(\mathcal{X})$ and $\alpha_{\mathrm{c}} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$.

By the continuous-domain Dirichlet process properties detailed in Appendix B.2, $\theta_{\mathrm{m}} \sim \mathrm{DP}(\alpha_0, \alpha_{\mathrm{m}})$ is a Dirichlet random process parameterized by concentration α_0 and distribution α_{m}; observe that the PDF $\mathrm{p}_{\mathrm{x}} = \mu_{\theta_{\mathrm{m}}} = \alpha_{\mathrm{m}}$. Also, the functions $\theta_{\mathrm{c}}(x) \sim \mathrm{DP}\left(\delta(0)^{-1}\alpha_0\, \alpha_{\mathrm{m}}(x), \alpha_{\mathrm{c}}(x)\right)$ are independent Dirichlet processes and are independent of θ_{m} as well. Note that $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x}} = \mu_{\theta_{\mathrm{c}}}(\mathrm{x}) = \alpha_{\mathrm{c}}(\mathrm{x})$.

4.2.2 Predictive PDF, $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}$

By the properties of the Dirichlet process proven in Appendix B.4.3, the model conditioned on the training data D is Dirichlet with concentration $\alpha_0 + N$ and mean function

$$\mu_{\theta|\,\mathrm{D}} = \gamma\alpha + (1-\gamma)\Psi(\mathrm{D})\ . \tag{4.7}$$

Again, by Bayes rule the marginal PDF is

$$\mathrm{p}_{\mathrm{x}\,|\,\mathrm{D}} \equiv \mathrm{p}_{\mathrm{x}\,|\,\mathrm{X}} = \gamma\,\alpha_{\mathrm{m}} + (1-\gamma)\,\Psi_{\mathrm{m}}(\mathrm{X})\ , \tag{4.8}$$

and the conditional PDF of interest is

$$\begin{aligned}
\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}} &= \frac{\alpha_0\alpha(\cdot,\mathrm{x}) + N\Psi(\cdot,\mathrm{x};\mathrm{D})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})} \\
&= \left(\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\alpha_{\mathrm{c}}(\mathrm{x}) + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\Psi_{\mathrm{c}}(\mathrm{x};\mathrm{D}) \\
&= \gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\,\alpha_{\mathrm{c}}(\mathrm{x}) + \big(1-\gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\big)\,\Psi_{\mathrm{c}}(\mathrm{x};\mathrm{D})\ .
\end{aligned} \tag{4.9}$$

The conditional distribution when $\mathcal{X}$ is a continuous space has notable differences from its form for a countable set $\mathcal{X}$. Specifically, since $\Psi_{\mathrm{m}}(D)$ is a Dirac delta function mixture, its values are either zero or tend towards infinity; thus, if the prior mean α_{m} is upper-bounded, the weight $\gamma_{\mathrm{m}}(D)$ is either zero or one. Consequently, the predictive distribution will equal the conditional empirical distribution $\Psi_{\mathrm{c}}(x;D)$ at all values $x \in \mathcal{X}$ that have been observed in the training data. The empirical distribution is used for prediction whenever it is available, similar to the discrete-domain case when $\alpha_0 \to 0$. In this case, the Dirichlet localization α_0 has no effect.

4.2.2.1 Via the Conditional Model Process

As the Dirichlet prior implies that θ_{m} is independent from θ_{c}, the likelihood $\mathrm{p}_{\mathrm{D}\,|\,\theta_{\mathrm{m}}}$ is proportionate to $\mathrm{p}_{\mathrm{X}\,|\,\theta_{\mathrm{m}}} = \bigotimes_{n=1}^{N}\theta_{\mathrm{m}}$. Thus, using the properties detailed in Appendix B.4.3, it can be shown that the marginal process satisfies $\theta_{\mathrm{m}}\,|\,\mathrm{D} \sim \theta_{\mathrm{m}}\,|\,\mathrm{X} \sim$

$\mathrm{DP}(\alpha_0 + N, \mu_{\theta_m \,|\, X})$, where

$$\mu_{\theta_m \,|\, X} = \gamma\, \alpha_m + (1-\gamma)\, \Psi_m(X)\ . \tag{4.10}$$

Similarly, using the empirical statistic representation, the model likelihood of $\psi_c, \psi_m \,|\, \theta_m$ is proportionate to $\psi_m \,|\, \theta_m \sim \mathrm{EP}(N, \theta_m)$. As a result, observe that $\theta_m \,|\, \psi_m, \psi_c \sim \theta_m \,|\, \psi_m \sim \mathrm{DP}(\alpha_0 + N, \mu_{\theta_m \,|\, \psi_m})$, where

$$\mu_{\theta_m \,|\, \psi_m} = \gamma\, \alpha_m + (1-\gamma)\, \psi_m\ . \tag{4.11}$$

Recall that $p_{x \,|\, \psi} \equiv \mu_{\theta_m \,|\, \psi_m}$.

The independence of the marginal and conditional models also implies that the likelihood $p_{x,D \,|\, \theta_c}$ is proportionate to $p_{Y \,|\, X, \theta_c} = \bigotimes_{n=1}^{N} \theta_c(X_n)$, which can be factored into separate functions of $\theta_c(x)$. Thus, it can be shown that

$$\theta_c(x) |\, x, D \sim \theta_c(x) |\, D \sim \mathrm{DP}\left(\frac{\alpha_0\, \alpha_m(x) + N\, \Psi_m(x; X)}{\delta(0)}, \mu_{\theta_c(x) |\, D} \right)\ , \tag{4.12}$$

where

$$\mu_{\theta_c(x) |\, D} = \gamma_m(x; X)\, \alpha_c(x) + \big(1 - \gamma_m(x; X)\big)\, \Psi_c(x; D)\ .$$

Similarly, using the empirical statistic representation, the likelihood of $x, \psi_c, \psi_m \,|\, \theta_c$ is proportionate to $\psi_c \,|\, \psi_m, \theta_m \sim \bigotimes_{x \in \mathcal{X}} \mathrm{EP}(N\, \psi_m(x), \theta_c(x))$. Consequently, observe that

$$\begin{aligned} \theta_c(x) |\, x, \psi_m, \psi_c &\sim \theta_c(x) |\, \psi_m(x), \psi_c(x) \\ &\sim \mathrm{DP}\left(\frac{\alpha_0\, \alpha_m(x) + N\, \psi_m(x)}{\delta(0)}, \mu_{\theta_c(x) |\, \psi_m(x), \psi_c(x)} \right)\ , \end{aligned} \tag{4.13}$$

where

$$\mu_{\theta_c(x) |\, \psi_m(x), \psi_c(x)} = \gamma_m(x; \psi_m)\, \alpha_c(x) + \big(1 - \gamma_m(x; \psi_m)\big)\, \psi_c(x)\ .$$

Recall that $p_{y \,|\, x, \psi} \equiv \mu_{\theta_c(x) |\, \psi_m(x), \psi_c(x)}$.

sim, otimes notation??

4.2.3 Training Data PDF, $\mathrm{p_D}$

move before predictive

geometric integral beta function??

Using the Dirichlet process properties shown in Appendix B.2,

$$\begin{aligned} \mathrm{p}_{\mathrm{D}_{n+1}\,|\,\mathrm{D}_n,\ldots,\mathrm{D}_1} &= \mu_{\theta|\,\mathrm{D}_n,\ldots,\mathrm{D}_1} \\ &\equiv \frac{\alpha_0\alpha + \sum_{i=1}^{n}\delta\big(\,\cdot - \mathrm{Y}_i\,\big)\delta\big(\,\cdot - \mathrm{X}_i\,\big)}{\alpha_0 + n} \end{aligned} \tag{4.14}$$

and thus the training data PDF is

$$\begin{aligned} \mathrm{p_D}(D) &= \mathrm{E}_\theta\left[\prod_{n=1}^{N}\theta\big(D_n\big)\right] \\ &= \mathrm{p}_{\mathrm{D}_1}\left(D_1\right)\prod_{n=2}^{N}\mathrm{p}_{\mathrm{D}_n\,|\,\mathrm{D}_{n-1},\ldots,\mathrm{D}_1}\left(D_n|D_{n-1},\ldots,D_1\right) \\ &\equiv \alpha\big(Y_1,X_1\big)\prod_{n=2}^{N}\frac{\alpha_0\alpha\big(Y_n,X_n\big) + \sum_{i=1}^{n-1}\delta\big(Y_n - Y_i\big)\delta\big(X_n - X_i\big)}{\alpha_0 + n - 1}\,. \end{aligned} \tag{4.15}$$

PGR: BELOW, just integrate??

It is instructional to find the PDF's for the training output values Y given the input values X, as well as the marginal PDF for the input values alone. Observe that since the independent observations $\mathrm{X}_n\,|\,\theta_\mathrm{m}$ are characterized by the Dirichlet process $\theta_\mathrm{m} \sim \mathrm{DP}(\alpha_0, \alpha_\mathrm{m})$, the PDF for X can be represented as

$$\begin{aligned} \mathrm{p_X}(X) &= \mathrm{E}_\theta\left[\,\mathrm{p}_{\mathrm{X}\,|\theta}(X|\theta)\right] \equiv \mathrm{E}_{\theta_\mathrm{m}}\left[\prod_{n=1}^{N}\theta_\mathrm{m}\left(X_n\right)\right] \\ &= \alpha_\mathrm{m}\left(X_1\right)\prod_{n=2}^{N}\frac{\alpha_0\,\alpha_\mathrm{m}\left(X_n\right) + \sum_{i=1}^{n-1}\delta\big(X_n - X_i\big)}{\alpha_0 + n - 1}\,. \end{aligned} \tag{4.16}$$

Note that the marginal PDF's are $\mathrm{p}_{\mathrm{X}_n} = \mathrm{p_x} = \mu_{\theta_\mathrm{m}} = \alpha_\mathrm{m}$.

Using Bayes theorem,

$$\mathrm{p}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X) = \mathrm{E}_{\theta_\mathrm{c}}\left[\prod_{n=1}^{N}\theta_\mathrm{c}\left(Y_n;X_n\right)\right] \tag{4.17}$$

$$= \alpha_\mathrm{c}\left(Y_1;X_1\right)\prod_{n=2}^{N}\frac{\alpha_0\,\alpha_\mathrm{m}(X_n)\,\alpha_\mathrm{c}\left(Y_n;X_n\right) + \sum_{i=1}^{n-1}\delta\big(Y_n - Y_i\big)\delta\big(X_n - X_i\big)}{\alpha_0\,\alpha_\mathrm{m}\left(X_n\right) + \sum_{i=1}^{n-1}\delta\big(X_n - X_i\big)} \tag{4.18}$$

express conditional using Dir aggregation conditional independence properties?

Note that the independent observations $\mathrm{Y}_n \,|\, \mathrm{X}_n, \theta_\mathrm{c}$ are conditionally independent of X_i, $i \neq n$, and that they are characterized by the independent processes $\theta_\mathrm{c}(x) \sim \mathrm{DP}\big(\delta(0)^{-1}\alpha_0\,\alpha_\mathrm{m}(x), \alpha_\mathrm{c}(x)\big)$. Consequently, the joint distribution of any samples from Y given X will only depend on the matching training samples.

The first and second order conditional distributions are of specific interest. The first order conditional distributions are

$$\mathrm{p}_{\mathrm{Y}_n \,|\, \mathrm{X}}(X) = \mathrm{p}_{\mathrm{Y}_n \,|\, \mathrm{X}_n}(X_n) = \mathrm{p}_{\mathrm{y} \,|\, \mathrm{x}}(X_n) = \mu_{\theta_\mathrm{c}(X_n)} = \alpha_\mathrm{c}(X_n)\,. \tag{4.19}$$

To determine the second order conditional distributions, first note that for $\mathrm{X}_n \neq \mathrm{X}_{n'}$, $\mathrm{p}_{\mathrm{Y}_n,\mathrm{Y}_{n'} \,|\, \mathrm{X}_n,\mathrm{X}_{n'}} = \mu_{\theta_\mathrm{c}}(\mathrm{X}_n) \otimes \mu_{\theta_\mathrm{c}}(\mathrm{X}_{n'}) = \alpha_\mathrm{c}(\mathrm{X}_n) \otimes \alpha_\mathrm{c}(\mathrm{X}_{n'})$. Conversely, if $\mathrm{X}_n = \mathrm{X}_{n'}$, then

$$\begin{aligned}
\mathrm{p}_{\mathrm{Y}_n,\mathrm{Y}_{n'} \,|\, \mathrm{X}_n,\mathrm{X}_{n'}} &= \mathrm{E}_{\theta_\mathrm{c}}\left[\,\theta_\mathrm{c}(\mathrm{X}_n) \otimes \theta_\mathrm{c}(\mathrm{X}_n)\right] \\
&= \frac{\mathrm{diag}\left(\,\alpha_\mathrm{c}(\mathrm{X}_n)\right) + \delta(0)^{-1}\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n)\,\alpha_\mathrm{c}(\mathrm{X}_n) \otimes \alpha(\mathrm{X}_n)}{\delta(0)^{-1}\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n) + 1} \\
&= \frac{\delta(0)}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n) + \delta(0)}\,\mathrm{diag}\left(\,\alpha_\mathrm{c}(\mathrm{X}_n)\right) + \frac{\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n)}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n) + \delta(0)}\,\alpha_\mathrm{c}(\mathrm{X}_n) \otimes \alpha(\mathrm{X}_n)\,.
\end{aligned} \tag{4.20}$$

Combining, the second order distribution formula is

$$\begin{aligned}
\mathrm{p}_{\mathrm{Y}_n,\mathrm{Y}_{n'} \,|\, \mathrm{X}_n,\mathrm{X}_{n'}} = \frac{\delta(\mathrm{X}_n - \mathrm{X}_{n'})}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n) + \delta(\mathrm{X}_n - \mathrm{X}_{n'})}\,\mathrm{diag}\left(\,\alpha_\mathrm{c}(\mathrm{X}_n)\right) \\
+\frac{\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n)}{\alpha_0\,\alpha_\mathrm{m}(\mathrm{X}_n) + \delta(\mathrm{X}_n - \mathrm{X}_{n'})}\,\alpha_\mathrm{c}(\mathrm{X}_n) \otimes \alpha(\mathrm{X}_{n'})
\end{aligned} \tag{4.21}$$

PGR: Dirichlet-Empirical Process perspective

The marginalized data can also be represented using the empirical transform. As demonstrated in Appendix B.4, the transformed process is Dirichlet-Empirical $\psi \equiv \Psi(\mathrm{D}) \sim \mathrm{DEP}(N, \alpha_0, \alpha)$. The mean and correlation functions have the same form as those of a discrete DEP (with the continuous space definitions of the diag operator).

Observe that by the aggregation principle, $\psi_\mathrm{m} \sim \mathrm{DEP}(N, \alpha_0, \alpha_\mathrm{m})$ is a DEP over the set $\mathcal{X}$. Additionally, the 1-dimensional subsets conditioned on the marginalized

DEP are characterized as

$$\psi_{\mathrm{c}}(x) \big| \, \psi_{\mathrm{m}}(x) \sim \mathrm{DEP}\left(\frac{N \, \psi_{\mathrm{m}}(x)}{\delta(0)}, \frac{\alpha_0 \, \alpha_{\mathrm{m}}(x)}{\delta(0)}, \alpha_{\mathrm{c}}(x) \right) . \tag{4.22}$$

4.3 Predictive Model Estimation

For continuous data spaces, the analysis of the Bayesian predictive distribution as an estimate of the true predictive distribution has some unique differences from the analysis put forth in Section 3.2 for discrete sets. Redefine the difference function $\Delta(\mathrm{x}; \mathrm{D}, \theta_{\mathrm{c}}) \equiv \mathrm{p}_{\mathrm{y} \,|\, \mathrm{x}, \mathrm{D}} - \mathrm{p}_{\mathrm{y} \,|\, \mathrm{x}, \theta_{\mathrm{c}}} \in \mathbb{R}^{\mathcal{Y}}$. Using the properties of the continuous-domain empirical process conditioned on its aggregation, the covariance function takes on the new form

$$\begin{aligned} \mathrm{Cov}(\mathrm{x}; \theta_{\mathrm{m}}, \theta_{\mathrm{c}}) &= \mathrm{C}_{\psi_{\mathrm{m}}, \psi_{\mathrm{c}} \,|\, \theta_{\mathrm{m}}, \theta_{\mathrm{c}}} \left[\, \mathrm{p}_{\mathrm{y} \,|\, \mathrm{x}, \psi_{\mathrm{m}}, \psi_{\mathrm{c}}} \, \right] \\ &= \mathrm{C}_{\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}}} \left[\, \gamma_{\mathrm{m}}(\mathrm{x}; \psi_{\mathrm{m}}) \right] \big(\, \alpha_{\mathrm{c}}(\mathrm{x}) - \theta_{\mathrm{c}}(\mathrm{x}) \big) \otimes \big(\, \alpha_{\mathrm{c}}(\mathrm{x}) - \theta_{\mathrm{c}}(\mathrm{x}) \big) \\ &\quad + \mathrm{E}_{\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}}} \left[\frac{\big(1 - \gamma_{\mathrm{m}}(\mathrm{x}; \psi_{\mathrm{m}}) \big)^2}{\delta(0)^{-1} N \, \psi_{\mathrm{m}}(\mathrm{x})} \right] \big(\, \mathrm{diag} \, \big(\, \theta_{\mathrm{c}}(\mathrm{x}) \big) - \theta_{\mathrm{c}}(\mathrm{x}) \otimes \theta_{\mathrm{c}}(\mathrm{x}) \big) \end{aligned} \tag{4.23}$$

and the expectation of the second moments of the difference function is thus

$$\begin{aligned} &\mathrm{E}_{\mathrm{D} \,|\, \theta_{\mathrm{m}}, \theta_{\mathrm{c}}} \Big[\Delta(\mathrm{x}; \mathrm{D}, \theta_{\mathrm{c}}) \otimes \Delta(\mathrm{x}; \mathrm{D}, \theta_{\mathrm{c}}) \Big] \\ &\quad \equiv \mathrm{E}_{\mathrm{x} \,|\, \theta_{\mathrm{m}}} \left[\mathrm{E}_{\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}}} \left[\, \gamma_{\mathrm{m}}(\mathrm{x}; \psi_{\mathrm{m}})^2 \right] \big(\, \alpha_{\mathrm{c}}(\mathrm{x}) - \theta_{\mathrm{c}}(\mathrm{x}) \big) \otimes \big(\, \alpha_{\mathrm{c}}(\mathrm{x}) - \theta_{\mathrm{c}}(\mathrm{x}) \big) \right] \\ &\qquad + \mathrm{E}_{\mathrm{x} \,|\, \theta_{\mathrm{m}}} \left[\mathrm{E}_{\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}}} \left[\frac{\big(1 - \gamma_{\mathrm{m}}(\mathrm{x}; \psi_{\mathrm{m}}) \big)^2}{\delta(0)^{-1} N \, \psi_{\mathrm{m}}(\mathrm{x})} \right] \Big(\, \mathrm{diag} \, \big(\, \theta_{\mathrm{c}}(\mathrm{x}) \big) - \theta_{\mathrm{c}}(\mathrm{x}) \otimes \theta_{\mathrm{c}}(\mathrm{x}) \Big) \right] . \end{aligned} \tag{4.24}$$

Note that the dependency on the marginal empirical model can be expressed through the conditional number of samples $\delta(0)^{-1} N \, \psi_{\mathrm{m}}(\mathrm{x})$.

4.3.1 Trends

The bias-variance trade-off for continuous-domain data has notable differences from the discrete-domain results. Note that by the aggregation property of Empirical distributions, the empirical process value $\psi_{\mathrm{m}}(x)$ conditioned on the model $\theta_{\mathrm{m}}(x)$ is

now distributed as

$$\begin{aligned} \mathrm{p}_{\psi_{\mathrm{m}}(x)|\,\theta_{\mathrm{m}}(x)}\left(\psi_{\mathrm{m}}(x)|\,\theta_{\mathrm{m}}(x)\right) &= \mathrm{Emp}\left(\left(\frac{\psi_{\mathrm{m}}(x)}{\delta(0)}, 1-\frac{\psi_{\mathrm{m}}(x)}{\delta(0)}\right); N, \left(\frac{\theta_{\mathrm{m}}(x)}{\delta(0)}, 1-\frac{\theta_{\mathrm{m}}(x)}{\delta(0)}\right)\right) \\ &= \mathrm{Bi}\left(\frac{N\,\psi_{\mathrm{m}}(x)}{\delta(0)}; N, \frac{\theta_{\mathrm{m}}(x)}{\delta(0)}\right) , \end{aligned} \tag{4.25}$$

$$\frac{N\,\uppsi_{\mathrm{m}}(x)}{\delta(0)}\Big|\,\theta_{\mathrm{m}}(x) \sim \mathrm{Bi}\left(N, \frac{\theta_{\mathrm{m}}(x)}{\delta(0)}\right) , \tag{4.26}$$

where Bi is the binomial PMF.

Thus for bounded $\theta_{\mathrm{m}}(x)$, the expected number of samples $\delta(0)^{-1} N\,\uppsi_{\mathrm{m}}(x) \to 0$, the value $\gamma_{\mathrm{m}}(x;\uppsi_{\mathrm{m}})$ tends to unity, and the expected value of the predictive distribution tends to the conditional prior mean $\alpha_{\mathrm{c}}(x)$ regardless of the value $\alpha_0\,\alpha_{\mathrm{m}}(x)$. As a result, the bias is always maximal and the covariance is always zero, no matter how large N is – the quality of the distribution estimate at x is entirely dependent on the bias.

If instead $\uptheta_{\mathrm{m}}$ includes a Dirac delta function at $x \in \mathcal{X}$, the conditional prior concentration $\alpha_0\,\alpha_{\mathrm{m}}(x)$ can affect the bias and variance. If the marginal prior mean value $\alpha_{\mathrm{m}}(x)$ is bounded, the empirical distribution is used for prediction whenever available and the weight $\mathrm{E}_{\uppsi_{\mathrm{m}}\,|\,\uptheta_{\mathrm{m}}}\left[\,\gamma_{\mathrm{m}}(\mathrm{x};\uppsi_{\mathrm{m}})\right]$ assumes its minimal value $\left(1-\delta(0)^{-1}\,\uptheta_{\mathrm{m}}(x)\right)^N$, which is not necessarily equal to one. The estimate has minimal bias and maximal variance; the weighting values in (4.24) are analogous to those provided in Section 3.2. Conversely, if α_{m} also has a Dirac delta function at x, the bias weight behaves as for the discrete-domain case, depending on the relative values of $\delta(0)^{-1}\,\uptheta_{\mathrm{m}}(x)$ and $\delta(0)^{-1}\,\alpha_{\mathrm{m}}(x)$.

Plot realizations (instead of stats) for meaningful viz?

4.4 Applications to Common Loss Functions

Discuss overfitting, like discrete with zero concentration

PGR: COPIED, incomplete

$$
\begin{aligned}
f^*(\mathrm{D}) &= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\mathrm{D}}\Big[\,\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right)\Big] \qquad (4.27)\\
&= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \gamma \iint_{\mathcal{Y}\times\mathcal{X}} \alpha(y,x)\,\mathcal{L}\left(g(x),y\right)\mathrm{d}y\mathrm{d}x + (1-\gamma)\iint_{\mathcal{Y}\times\mathcal{X}} \Psi(y,x;\mathrm{D})\,\mathcal{L}\left(g(x),y\right)\mathrm{d}y\mathrm{d}x\\
&= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \gamma\,\mathrm{E}_{\mathrm{y},\mathrm{x}}\Big[\,\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right)\Big] + (1-\gamma)\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(g(\mathrm{X}_n),\mathrm{Y}_n\right)\\
&= \arg\min_{g\in\mathcal{H}^{\mathcal{X}}} \alpha_0\,\mathrm{E}_{\mathrm{y},\mathrm{x}}\Big[\,\mathcal{L}\left(g(\mathrm{x}),\mathrm{y}\right)\Big] + \sum_{n=1}^{N}\mathcal{L}\left(g(\mathrm{X}_n),\mathrm{Y}_n\right).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\big[\,\mathcal{L}(h,\mathrm{y})\big] &= \int_{\mathcal{Y}} \mathcal{L}(h,y)\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\mathrm{d}y \qquad (4.28)\\
&= \left(\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\int_{\mathcal{Y}} \alpha_{\mathrm{c}}(y;\mathrm{x})\,\mathcal{L}(h,y)\mathrm{d}y\\
&\quad + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\int_{\mathcal{Y}} \Psi_{\mathrm{c}}(y;\mathrm{x};\mathrm{D})\,\mathcal{L}(h,y)\mathrm{d}y\\
&= \gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\big[\,\mathcal{L}(h,\mathrm{y})\big] + \big(1-\gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\big)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathcal{L}\left(h,\mathrm{Y}_n\right)}{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]}.
\end{aligned}
$$

4.4.1 Regression: the Squared-Error Loss

Now we choose for the regression function to map to $\mathcal{H} = \mathcal{Y} = \mathbb{R}$.

4.4.1.1 Bayesian Estimation

Optimal Estimator The optimal function is the expected value of the output conditional PDF,

$$
\begin{aligned}
f^*(\mathrm{x};\mathrm{D}) &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \qquad (4.29)\\
&= \left(\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\int_{\mathcal{Y}} y\,\alpha_{\mathrm{c}}(y;\mathrm{x})\mathrm{d}y\\
&\quad + \left(\frac{N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\Psi_{\mathrm{m}}(\mathrm{x};\mathrm{X})}\right)\int_{\mathcal{Y}} y\,\Psi_{\mathrm{c}}(y;\mathrm{x};\mathrm{D})\mathrm{d}y\\
&= \gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \big(1-\gamma_{\mathrm{m}}(\mathrm{x};\mathrm{X})\big)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathrm{Y}_n}{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]}.
\end{aligned}
$$

As discussed in Section 4.2, the weighting function $\gamma_{\mathrm{m}}(D)$ will tend to zero if α_{m} is upper-bounded and matching samples $\mathrm{X}_n = \mathrm{x}$ are observed; otherwise, the weight is one and the prior estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ is used. In this case, the localization α_0 has no effect.

Minimum Bayes Risk To determine the minimum Bayes squared error, $\mathcal{R}^* = \mathrm{E}_{\mathrm{x},\psi}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\psi}\right]$, a new evaluation of $\mu^2_{\mathrm{y}\,|\,\mathrm{x},\psi}$ must be performed.

PGR: D PERSPECTIVE

To evaluate the expectation directly using training samples Y and X, first note that $\mu_{\mathrm{Y}_n\,|\,\mathrm{X}} = \mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\mathrm{X}_n\big)$ and $\mathrm{E}_{\mathrm{Y}_n\,|\,\mathrm{X}}\left[\mathrm{Y}_n^2\right] = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\big(\mathrm{X}_n\big)$, and that

$$\begin{aligned}&\mathrm{E}_{\mathrm{Y}_n,\mathrm{Y}_{n'}\,|\,\mathrm{X}}\left[\mathrm{Y}_n\,\mathrm{Y}_{n'}\right] \\ &\quad= \frac{\alpha_0\,\alpha_{\mathrm{m}}\left(\mathrm{X}_n\right)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\mathrm{X}_n\big)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\mathrm{X}_{n'}\big) + \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\big(\mathrm{X}_n\big)\delta\big(\mathrm{X}_n - \mathrm{X}_{n'}\big)}{\alpha_0\,\alpha_{\mathrm{m}}\left(\mathrm{X}_n\right) + \delta\big(\mathrm{X}_n - \mathrm{X}_{n'}\big)}\,.\end{aligned} \tag{4.30}$$

Solving,

$$\begin{aligned}\mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\left(\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)}\right)^2\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{D}}\left[\frac{\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)^2}{\alpha_{\mathrm{m}}(\mathrm{x})\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)(\alpha_0+N)}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{X}}\left[\frac{\mathrm{E}_{\mathrm{Y}\,|\,\mathrm{X}}\left[\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)^2\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)(\alpha_0+N)}\right]\right]\,.\end{aligned}$$

Evaluating the expectation over Y given X, we have

$$\mathrm{E}_{\mathrm{Y}\,|\,\mathrm{X}}\left[\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}+\sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\right)^2\right] \tag{4.31}$$

$$\begin{aligned}
&=\alpha_0^2\,\alpha_{\mathrm{m}}(\mathrm{x})^2\mu_{\mathrm{y}\,|\,\mathrm{x}}^2+2\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}\sum_{n=1}^{N}\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\,\mathrm{X}_n\,\big)\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\\
&\quad+\sum_{n=1}^{N}\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\big(\,\mathrm{X}_n\,\big)\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)^2\\
&\quad+\sum_{n\neq n'}\frac{\alpha_0\,\alpha_{\mathrm{m}}\left(\,\mathrm{X}_{n'}\,\right)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\,\mathrm{X}_n\,\big)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\,\mathrm{X}_{n'}\,\big)+\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\big(\,\mathrm{X}_n\,\big)\delta\big(\,\mathrm{X}_n-\mathrm{X}_{n'}\,\big)}{\alpha_0\,\alpha_{\mathrm{m}}\left(\,\mathrm{X}_{n'}\,\right)+\delta\big(\,\mathrm{X}_n-\mathrm{X}_{n'}\,\big)}\\
&\qquad\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\delta\big(\,\mathrm{x}-\mathrm{X}_{n'}\,\big)\\
&=\dots\\
&=\alpha_0^2\,\alpha_{\mathrm{m}}(\mathrm{x})^2\mu_{\mathrm{y}\,|\,\mathrm{x}}^2+2\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)+\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)^2\\
&\quad+\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2+\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)}\sum_{n\neq n'}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\delta\big(\,\mathrm{x}-\mathrm{X}_{n'}\,\big)\\
&\dots\\
&=\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)}\\
&\quad\left(\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)+\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)+\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\right)\right)\,.
\end{aligned}$$

Plugging,

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x,D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x,D}}^2\right]\\
&\quad=\mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{X}}\left[\frac{\mathrm{E}_{\mathrm{Y}\,|\,\mathrm{X}}\left[\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}+\sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\right)^2\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\right)(\alpha_0+N)}\right]\right]\\
&\quad=\mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{E}_{\mathrm{X}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)+\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)+\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\right)\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)(\alpha_0+N)}\right]
\end{aligned}$$

Evaluating the expectation over X,

$$\begin{aligned}
&\mathrm{E}_{\mathrm{X}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)+\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)+\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\right)\right]\\
&\quad=\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)N\,\alpha_{\mathrm{m}}(\mathrm{x})+\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)+N\,\alpha_{\mathrm{m}}(\mathrm{x})\big)\\
&\quad=\alpha_{\mathrm{m}}(\mathrm{x})\Big(\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)N+\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)+N\,\alpha_{\mathrm{m}}(\mathrm{x})\big)\Big)\,.
\end{aligned}$$

Plugging,

$$\begin{aligned} &\mathrm{E}_{\mathrm{x,D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \\ &\quad = \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)N + \mu^2_{\mathrm{y}\,|\,\mathrm{x}}\alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \delta(0) + N\,\alpha_{\mathrm{m}}(\mathrm{x})\big)}{\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \delta(0)\big)(\alpha_0 + N)}\right] . \end{aligned} \tag{4.33}$$

Combining with the second moment produces the risk,

$$\begin{aligned} \mathcal{R}^* &= \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\,\mathrm{y}^2\,\right] - \mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \delta(0) + N\,\alpha_{\mathrm{m}}(\mathrm{x})\big)}{(\alpha_0 + N)\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + \delta(0)\big)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x}) + (\alpha_0 + N)^{-1}}{\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x}) + \alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] . \end{aligned} \tag{4.34}$$

PGR: DEP PERSPECTIVE???

To perform the expectation over the Dirichlet-Empirical process $\psi \sim \mathrm{DEP}(N, \alpha_0, \alpha)$, split the expectation into an expectation over the marginal DEP $\psi_{\mathrm{m}} \sim \mathrm{DMP}(N, \alpha_0, \alpha_{\mathrm{m}})$ and a conditional expectation over ψ_{c} given ψ_{m}. The characterization of the conditional DEP is found in Appendix B.3.

$$\begin{aligned} \mathrm{E}_{\mathrm{x},\psi}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\psi}\right] &= \mathrm{E}_{\mathrm{x},\psi}\left[\left(\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + N\int_{\mathcal{Y}} y\psi(y,\mathrm{x})\mathrm{d}y}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\int_{\mathcal{Y}}\psi(y,\mathrm{x})\mathrm{d}y}\right)^2\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\psi_{\mathrm{m}}}\left[\frac{\mathrm{E}_{\psi_{\mathrm{c}}\,|\,\psi_{\mathrm{m}}}\left[\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + N\,\psi_{\mathrm{m}}(\mathrm{x})\int_{\mathcal{Y}} y\,\psi_{\mathrm{c}}(y;\mathrm{x})\mathrm{d}y\big)^2\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x}) + N\,\psi_{\mathrm{m}}(\mathrm{x})\big)(\alpha_0 + N)}\right]\right] . \end{aligned}$$

Evaluating the conditional expectation,

$$
\begin{aligned}
&\mathrm{E}_{\psi_{\mathrm{c}}\,|\,\psi_{\mathrm{m}}}\left[\left(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}+N\,\psi_{\mathrm{m}}(\mathrm{x})\int_{\mathcal{Y}} y\,\psi_{\mathrm{c}}(y;\mathrm{x})\mathrm{d}y\right)^2\right] \qquad (4.35)\\
&=\alpha_0^2\,\alpha_{\mathrm{m}}(\mathrm{x})^2\mu_{\mathrm{y}\,|\,\mathrm{x}}^2+2\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})N\,\psi_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}\int_{\mathcal{Y}} y\,\alpha_{\mathrm{c}}(y;\mathrm{x})\mathrm{d}y\\
&\quad+\frac{N^2\,\psi_{\mathrm{m}}(\mathrm{x})^2}{1+\frac{\delta(0)}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})}}\int_{\mathcal{Y}}\int_{\mathcal{Y}} yy'\left[\left(1-\frac{\delta(0)}{N\,\psi_{\mathrm{m}}(\mathrm{x})}\right)\alpha_{\mathrm{c}}(y;\mathrm{x})\,\alpha_{\mathrm{c}}(y';\mathrm{x})\right.\\
&\qquad\left.+\left(\frac{\delta(0)}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})}+\frac{\delta(0)}{N\,\psi_{\mathrm{m}}(\mathrm{x})}\right)\alpha_{\mathrm{c}}(y;\mathrm{x})\delta(y-y')\right]\mathrm{d}y\mathrm{d}y'\\
&=\alpha_0^2\,\alpha_{\mathrm{m}}(\mathrm{x})^2\mu_{\mathrm{y}\,|\,\mathrm{x}}^2+2\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})N\,\psi_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\\
&\quad+\frac{N\,\psi_{\mathrm{m}}(\mathrm{x})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)}\Big[\big(N\,\psi_{\mathrm{m}}(\mathrm{x})-\delta(0)\big)\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}^2+\delta(0)\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+N\,\psi_{\mathrm{m}}(\mathrm{x})\big)\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\Big]\\
&=\frac{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+N\,\psi_{\mathrm{m}}(\mathrm{x})}{\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)}\Big[\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+N\,\psi_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)+\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)N\,\psi_{\mathrm{m}}(\mathrm{x})\Big]\,.
\end{aligned}
$$

Plugging,

$$
\begin{aligned}
&\mathrm{E}_{\mathrm{x},\psi}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\psi}^2\right] \qquad (4.36)\\
&=\mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{E}_{\psi_{\mathrm{m}}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+N\,\psi_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)+\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)N\,\psi_{\mathrm{m}}(\mathrm{x})\right]}{\alpha_{\mathrm{m}}(\mathrm{x})\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)(\alpha_0+N)}\right]\\
&=\mathrm{E}_{\mathrm{x}}\left[\frac{\mu_{\mathrm{y}\,|\,\mathrm{x}}^2\alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+N\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)+\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\delta(0)N}{(\alpha_0+N)\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)}\right]\,.
\end{aligned}
$$

Combining produces the risk,

$$
\begin{aligned}
\mathcal{R}^* &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)+N\,\alpha_{\mathrm{m}}(\mathrm{x})\big)}{(\alpha_0+N)\big(\alpha_0\,\alpha_{\mathrm{m}}(\mathrm{x})+\delta(0)\big)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] \qquad (4.37)\\
&= \mathrm{E}_{\mathrm{x}}\left[\frac{\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x})+(\alpha_0+N)^{-1}}{\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x})+\alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]\,.
\end{aligned}
$$

Change form to have conditional prior concentration?

The minimum Bayesian squared error equation is similar to the discrete-domain formula (3.60), with one difference – the scaling factor for $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ depends on the marginal prior mean through $\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x})$.

Note that as $N\to\infty$, the Bayesian risk tends to $\mathcal{R}^*\to\mathrm{E}_{\mathrm{x}}\left[\frac{\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x})}{\delta(0)^{-1}\,\alpha_{\mathrm{m}}(\mathrm{x})+\alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$, which can readily be shown to be the expected irreducible risk, $\mathrm{E}_{\theta}\left[\,\mathcal{R}_{\Theta}^*(\theta)\right]=$

$\mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$. This result is dependent on the fact that models θ drawn from the prior p_θ will be Dirac delta mixtures with countable support, enabling effective learning of the clairvoyant regressor $f_\Theta(\theta)$; as shown in the following section, if the true model θ is bounded, then the excess risk will never vanish and the irreducible squared error will not be achieved.

Other notable difference from the discrete set results occur if α_m is bounded; specifically, $\mathcal{R}^* \approx (1 + N/\alpha_0)^{-1}\,\mathrm{E}_\mathrm{x}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$. This is a consequence of the estimator preference for using the empirical data distribution when available; the scaling factor $(1+N/\alpha_0)^{-1}$ is equal to the probability that the novel data pair (y,x) is not represented in the training set and thus that the prior estimator is used. Additionally, as $N \to \infty$, the Bayes squared error tends to zero. With continuous data, for any localization α_0, the probability that two of the Dirac delta functions comprising the mixture p_θ are located at the same value x is zero. As such, every predictive distribution $\theta_\mathrm{c}(x)$ will be supported at a single value y, the conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}$ is zero, and the irreducible squared error is zero. With sufficient data, the probability that the novel data pair (y,x) is represented in the training set tends to one, and thus the empirical predictive distribution $\psi_\mathrm{c}(\mathrm{x})$ will precisely identify $\theta(\mathrm{x})$. This is analogous to the $\alpha_0 \to 0$ case for discrete data, where the model θ will be concentrated at a single value (y, x).

improve above discussion?? NEED DP CITATIONS!!!

4.4.1.2 Squared-Error Trends

Using the second moments of the random process $\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c})$ formulated in (4.24), the excess squared error is redefined as

$$\begin{aligned}\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right] \\ &\equiv \int_{\mathcal{Y}} y \int_{\mathcal{Y}} y'\, \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\Delta(y;\mathrm{x};\mathrm{D},\theta_\mathrm{c})\Delta(y';\mathrm{x};\mathrm{D},\theta_\mathrm{c})\right]\mathrm{d}y\mathrm{d}y' \\ &= \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\,\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})^2\right]\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right)^2\right] \\ &\quad + \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\frac{\left(1-\gamma_\mathrm{m}(\mathrm{x};\psi_\mathrm{m})\right)^2}{\delta(0)^{-1}N\,\psi_\mathrm{m}(\mathrm{x})}\right]\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right]\,. \end{aligned} \tag{4.38}$$

It is instructional to consider the trends of the excess squared error (4.38) with training data volume N and with Dirichlet prior parameterization. Referencing the predictive distribution estimation discussion in Section 4.3, the bias-variance trends are used. Importantly, if the marginal model θ_m is upper-bounded, the biased, fixed prior estimator $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ will always be used (on average) and the excess risk is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \approx \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, regardless of how large the training data volume N may be.

If instead θ_m includes Dirac delta functions, there are values $x \in \mathcal{X}$ with non-zero probabilities of being observed. At these values, if $\alpha_\mathrm{m}(x)$ is bounded, the weights tend to

$$\mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\,\gamma_\mathrm{m}(x;\psi_\mathrm{m})^2\right] \to \left(1 - \frac{\theta_\mathrm{m}(x)}{\delta(0)}\right)^N \tag{4.39}$$

and

$$\mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\frac{\left(1-\gamma_\mathrm{m}(x;\psi_\mathrm{m})\right)^2}{\delta(0)^{-1}N\,\psi_\mathrm{m}(x)}\right] \to \sum_{n=1}^{N}\binom{N}{n}\left(\frac{\theta_\mathrm{m}(x)}{\delta(0)}\right)^n\left(1-\frac{\theta_\mathrm{m}(x)}{\delta(0)}\right)^{N-n}\frac{1}{n}\,, \tag{4.40}$$

representing the minimum bias and maximum variance incurred by the empirical estimate. If α_m also has a Dirac delta function at x, the value $\gamma_\mathrm{m}(x;\psi_\mathrm{m})$ may assume values between zero and one and the estimate may convexly combine the prior estimate with the empirical estimate; in this case, the bias-variance weights may take on the same full range of values as for regression using discrete-domain data.

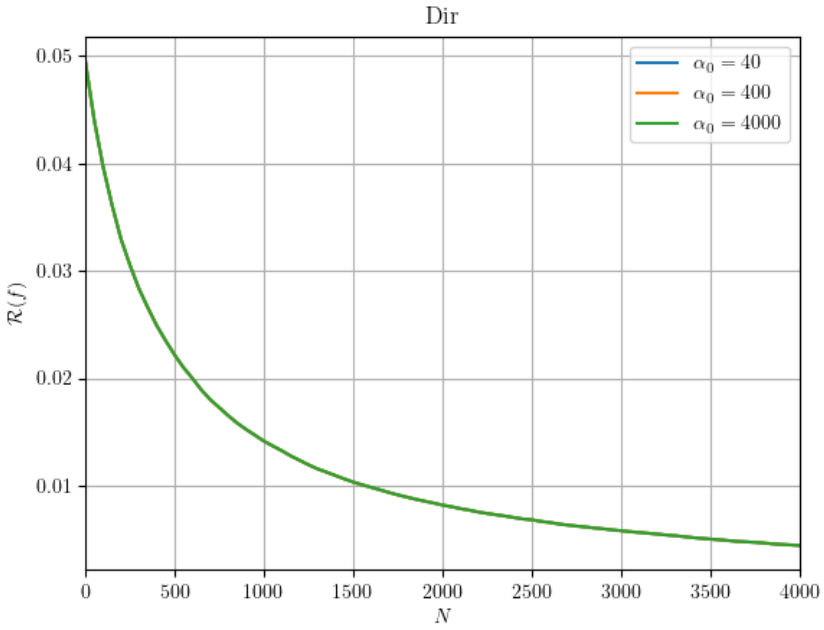

Figure 4.1: Bayes Squared-Error vs. N

comment on total risk with N - equal to prob thetam is continuous times sq bias?!

4.4.1.3 Example

Consider a data model where $\mathcal{X} = \mathcal{Y} = [0, 1]$, the closed unit interval. The Dirichlet estimator is parameterized by $\alpha_{\mathrm{m}} = 1$ and $\alpha_{\mathrm{c}}(x) = \mathrm{Beta}(2, 2)$, such that the prior estimator $\mu_{\mathrm{y}\,|\,\mathrm{x}} = 0.5$ is constant; various localizations α_0 will be used.

Fig. 4.1 displays the Bayesian squared error realized by the Dirichlet-based regressor. The model θ is randomly selected from a Dirichlet process with the same α_{m} and α_{c} used for the regressor design and with the localization fixed at $\alpha_0 = 400$. Since α_{m} is bounded, the Bayes risk is $\mathcal{R}^* \approx (1 + N/\alpha_0)^{-1}\, \mathrm{E}_{\mathrm{x}} \left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$. Observe that the risk is independent of the value α_0 chosen for the estimator, since $\delta(0)^{-1}\alpha_0\, \alpha_{\mathrm{m}}(x) \to 0$ and thus the empirical estimate will always be used when available. Additionally, note that as $N \to \infty$, the regressors always tend to $\mathcal{R}^* \to 0$, which is equivalent to the expected irreducible squared error.

Next, consider the risk trends when the true model is fixed. The marginal distribution $\theta_{\mathrm{m}} = 1$ is uniform and $\theta_{\mathrm{c}}(\mathrm{x}) = \mathrm{Beta}\big(4\,\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}, 4\,(1 - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta})\big)$, where the clairvoyant regressor is $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta} = 1/\big(2 + \sin(2\pi\,\mathrm{x})\big)$. Comparing with the discrete-space models used for demonstration in Section 3.3.1.3, observe that the optimal regressor $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ and true predictive variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta} = 0.2\,\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}(1 - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta})$ are the same; the irreducible squared error $\mathcal{R}^*_\Theta(\theta) \approx 0.038$ is approximately equal as well. However, since θ_{m} is bounded, no effective learning is possible, the Bayesian estimator tends to $\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \to \mu_{\mathrm{y}\,|\,\mathrm{x}}$, and the excess risk tends to the maximum bias value $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \approx \mathrm{E}_{\mathrm{x}\,|\theta}\Big[\big(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\Big] \approx 0.058$ regardless of the data volume N.

Chapter 5

Discretized Dirichlet Model

Reconsider full thesis structure. Orig stats sec after problem statement!?

Separate general feature theory from discretization. If Tcal not subset of Xcal!?

Develop from low-dim prior! More SUBJECTIVE priors!!

CITE for discretization work

5.1 From the continuous DP

As seen in Chapter 4, the flexibility of using a full-support prior, while effective for countable data spaces, has limited practical value when operating on observations drawn from continuous spaces. This section develops a new Bayesian predictive distribution using discretization that may be used to realize the benefits of Dirichlet process priors while avoiding the aforementioned limitation.

Define a discretizing data transformation $T : \mathcal{X} \mapsto \mathcal{T}$, where the range is a countable subset $\mathcal{T} \subset \mathcal{X}$, such that $|\mathcal{T}| \leq \aleph_0$. This transform is applied to each of the training values X_n, as well as to the novel observation x. Note that T reduces the cardinality of the observation set; this is conceptually similar to the dimensionality reduction induced by feature transforms commonly used in machine learning and yields similar benefits. Use T to define a partitioning $\{\ldots, \mathcal{X}'(t), \ldots\}$ of the observation

space $\mathcal{X}$, where $\mathcal{X}'(t) = \{x \in \mathcal{X} : T(x) = t\}$. It is assumed that these subsets are all connected spaces; furthermore, assume that $t \in \mathcal{X}'(t),\ \forall\ t \in \mathcal{T}$, such that the transform maps $T(t) = t$.

Having effectively changed the domain of the observations, the Dirichlet prior marginal mean is re-defined as $\alpha_{\mathrm{m}} = \sum_{t\in\mathcal{T}} \alpha'_{\mathrm{m}}(t)\delta(\cdot - t)$, a mixture of delta functions at the discretized values, where $\alpha'_{\mathrm{m}} \in \mathcal{P}(\mathcal{T})$ is the discrete marginal function.

discretized conditional alpha?? generic marginal alpha to start?

Starting from the Dirichlet-based Bayesian predictive distribution (4.9), substitute for the discretized observations, such that

$$\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}} = \gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\right)\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) + \Big(1 - \gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\right)\Big)\,\Psi'_{\mathrm{c}}\left(T(\mathrm{x});\mathrm{D}\right), \tag{5.1}$$

where the discretized weighting function $\gamma'_{\mathrm{m}} : \mathcal{X}^N \mapsto (0,1]^{\mathcal{T}}$ is defined as

$$\begin{aligned}\gamma'_{\mathrm{m}}(t;X) &= \left(1 + \frac{\sum_{n=1}^{N}\delta\big(t - T(X_n)\big)}{\alpha_0 \sum_{t'\in\mathcal{T}} \alpha'_{\mathrm{m}}(t')\delta(t-t')}\right)^{-1} \\ &= \left(1 + \frac{\sum_{n=1}^{N}\delta\big[t, T(X_n)\big]}{\alpha_0 \sum_{t'\in\mathcal{T}} \alpha'_{\mathrm{m}}(t')\delta[t,t']}\right)^{-1} \\ &= \left(1 + \frac{N\,\Psi'_{\mathrm{m}}(t;X)}{\alpha_0\,\alpha_{\mathrm{m}}{}'(t)}\right)^{-1}. \end{aligned} \tag{5.2}$$

Note that the Dirac delta functions over $\mathcal{X}$ are recast as Kronecker delta functions over $\mathcal{T}$. The discretized marginal empirical function is

$$\Psi'_{\mathrm{m}}(X) = \frac{1}{N}\sum_{n=1}^{N}\delta\big[\cdot, T(X_n)\big] \in \mathcal{P}(\mathcal{T}) \tag{5.3}$$

and the discretized conditional empirical transform $\Psi'_{\mathrm{c}}(D) \in \mathcal{P}(\mathcal{Y})^{\mathcal{T}}$ is defined as

$$\Psi'_{\mathrm{c}}(t;D) = \frac{\sum_{n=1}^{N}\delta\big(\,\cdot - Y_n\big)\delta\big[t, T(X_n)\big]}{\sum_{n=1}^{N}\delta\big[t, T(X_n)\big]}. \tag{5.4}$$

Observe that for a novel observation x, the convex weight will now depend on $\sum_{n=1}^{N}\delta\big[T(\mathrm{x}), T(X_n)\big]$, the number of training samples that discretize to the same value. This contrasts with the pure continuous-domain Dirichlet predictive distribution, which only counts training values that match precisely. Similarly, the conditional empirical distribution being mixed in (5.1) is formed using the same, larger group of training observations.

5.2 Sufficient Statistic: Discretized Empirical

Inspecting the new predictive distribution $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}$, it is clear that the dependency on the joint observations (x, D) can be represented using the sufficient statistics $\big(T(\mathrm{x}), \Psi'_{\mathrm{m}}(\mathrm{X}), \Psi'_{\mathrm{c}}(\mathrm{D})\big)$. Again, it is useful to define and characterize new random processes.

Defining the novel feature $\mathrm{t} \equiv T(\mathrm{x})$, it is evident that $\mathrm{P}_{\mathrm{t}\,|\,\theta} = \Pr\big(T(\mathrm{x}) = \mathrm{t}\,|\theta\big) = \Pr\big(\,\mathrm{x} \in \mathcal{X}'(\mathrm{t})|\theta\big)$. As a result, $\mathrm{P}_{\mathrm{t}\,|\,\theta} \equiv \theta'_{\mathrm{m}} \in \mathcal{P}(\mathcal{T})$, where θ'_{m} is the discretized marginal model defined as $\theta'_{\mathrm{m}}(t) = \int_{\mathcal{X}'(t)} \theta_{\mathrm{m}}(x)\mathrm{d}x$. Also, note that given the discretization, the observation PDF is

$$\mathrm{p}_{\mathrm{x}\,|\,\mathrm{t},\theta} \equiv \frac{\theta_{\mathrm{m}}}{\theta'_{\mathrm{m}}(\mathrm{t})}\chi\big(\,\mathcal{X}'(\mathrm{t})\big)\ . \tag{5.5}$$

Next, characterize the new empirical processes. Define the discretized empirical process $\psi' \equiv \Psi'(D) \in \mathcal{P}(\mathcal{Y} \times \mathcal{T})$, where

$$\begin{aligned}
\Psi'(y,t;D) &\equiv \frac{1}{N}\sum_{n=1}^{N} \delta(y - Y_n)\delta\big[t, T(X_n)\big] \\
&= \frac{1}{N}\sum_{n=1}^{N} \delta(y - Y_n)\chi\big(X_n; \mathcal{X}'(t)\big) \\
&= \frac{1}{N}\sum_{n=1}^{N} \delta(y - Y_n)\int_{\mathcal{X}'(t)} \delta(x - X_n)\mathrm{d}x \\
&= \int_{\mathcal{X}'(t)} \Psi(y,x;D)\mathrm{d}x\ .
\end{aligned} \tag{5.6}$$

Observe that ψ' is a transform of the original empirical process ψ. By the aggregation property of Empirical processes (Appendix B.1), it is shown that $\psi'\,|\theta \sim \mathrm{EP}(N, \theta')$, where the dependency on the model is expressed through the discretized $\theta' \in \mathcal{P}(\mathcal{Y} \times \mathcal{T})$, defined as $\theta'(y,t) = \int_{\mathcal{X}'(t)} \theta(y,x)\mathrm{d}x$.

Consequently, when conditioned on the model, the marginal process $\psi'_{\mathrm{m}} \equiv \Psi'_{\mathrm{m}}(\mathrm{X}) \equiv \int_{\mathcal{Y}} \psi'(y,\cdot)\mathrm{d}y$ is also Empirical with N samples and mean θ'_{m}. Additionally, the conditional processes $\psi'_{\mathrm{c}}(t) \equiv \Psi'_{\mathrm{c}}(t;\mathrm{D})$ are conditionally independent and

distributed as $\psi'_{\mathrm{c}}(t)|\,\psi'_{\mathrm{m}}(t),\theta'_{\mathrm{c}}(t) \sim \mathrm{EP}\left(N\,\psi'_{\mathrm{m}}(t),\theta'_{\mathrm{c}}(t)\right)$, where

$$\theta'_{\mathrm{c}}(t) \equiv \frac{\theta'(\cdot,t)}{\theta'_{\mathrm{m}}(t)} = \int_{\mathcal{X}'(t)} \theta_{\mathrm{c}}(x)\frac{\theta_{\mathrm{m}}(x)}{\theta'_{\mathrm{m}}(t)}\mathrm{d}x\ , \tag{5.7}$$

convexly combining the true predictive distributions $\theta_{\mathrm{c}}(x)$ for observations $x \in \mathcal{X}'(t)$. The discretized conditional model can be simply represented as $\theta'_{\mathrm{c}}(\mathrm{t}) = \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_{\mathrm{m}}}\left[\,\theta_{\mathrm{c}}(\mathrm{x})\right]$.

5.2.1 PGR predictive dist w psi

The Bayesian predictive distribution (5.1) can be represented in terms of the discretized sufficient statistics as $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}} = \mathrm{p}_{\mathrm{y}\,|\,\mathrm{t},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\left(T(\mathrm{x}),\Psi'_{\mathrm{m}}(\mathrm{X}),\Psi'_{\mathrm{c}}(\mathrm{D})\right)$, where

$$\mathrm{p}_{\mathrm{y}\,|\,\mathrm{t},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}} = \gamma'_{\mathrm{m}}(\mathrm{t};\psi'_{\mathrm{m}})\,\alpha_{\mathrm{c}}(\mathrm{t}) + \left(1-\gamma'_{\mathrm{m}}(\mathrm{t};\psi'_{\mathrm{m}})\right)\psi'_{\mathrm{c}}(\mathrm{t})\ . \tag{5.8}$$

Note that the weighting function is redefined to operate on the discretized empirical distribution rather than the raw observations, such that $\gamma'_{\mathrm{m}} : \mathcal{P}(\mathcal{T}) \mapsto (0,1]^{\mathcal{T}}$ and

$$\gamma'_{\mathrm{m}}(\psi_{\mathrm{m}}{}') = \left(1+\frac{N\,\psi'_{\mathrm{m}}}{\alpha_0\,\alpha'_{\mathrm{m}}}\right)^{-1}\ . \tag{5.9}$$

5.3 Predictive Model Estimation

To evaluate the bias and covariance functions of the Bayesian predictive distribution $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}$, the training data will be represented by the discretized sufficient statistics $(\psi'_{\mathrm{m}},\psi'_{\mathrm{c}})$. Note that $\mathrm{E}_{\mathrm{D}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}\right] = \mathrm{E}_{\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\right]$ and $\mathrm{C}_{\mathrm{D}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}\right] = \mathrm{C}_{\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\right]$.

In terms of the transformed observation t, the expected value of the estimate conditioned on the true model is

$$\begin{aligned}\mathrm{E}_{\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{t},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\right] &= \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\left[\,\gamma'_{\mathrm{m}}(\mathrm{t};\psi'_{\mathrm{m}})\right]\alpha_{\mathrm{c}}(\mathrm{t}) \\ &\quad + \left(1-\mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\left[\,\gamma'_{\mathrm{m}}(\mathrm{t};\psi'_{\mathrm{m}})\right]\right)\theta'_{\mathrm{c}}(\mathrm{t})\ .\end{aligned} \tag{5.10}$$

Comparing to (3.45), observe that the mixture distribution no longer includes the true predictive distribution θ_{c}, but rather the discretized θ'_{c}.

Using the equivalence $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}} = \mathrm{p}_{\mathrm{y}\,|\,\mathrm{t},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\left(T(\mathrm{x}),\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}\right)$ and substituting into (2.35), the expected bias is

$$\begin{aligned}\mathrm{Bias}(\mathrm{x};\theta_{\mathrm{m}},\theta_{\mathrm{c}}) = \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\Big[\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\psi'_{\mathrm{m}}\right)\Big]\Big(\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\Big) \\ +\Big(\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})\Big)\,. \end{aligned} \tag{5.11}$$

While similar in form to (3.46), observe that the bias includes an additional term that quantifies a "discretization bias". Since the discretized conditional empirical function $\psi_{\mathrm{c}}(t)$ integrates all training samples falling in the subset $\mathcal{X}'(t) \subset \mathcal{X}$, its expected value integrates the true predictive distributions $\theta_{\mathrm{c}}(x)$ in the same region.

Following a procedure similar to that employed in Section 3.2, the covariance of the estimate in terms of the transformed value t is

$$\begin{aligned}&\mathrm{C}_{\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{t},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\right] \\ &\quad= \mathrm{C}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\left[\gamma'_{\mathrm{m}}(\mathrm{t};\psi'_{\mathrm{m}})\right]\left(\alpha_{\mathrm{c}}(\mathrm{t}) - \theta'_{\mathrm{c}}(\mathrm{t})\right)\otimes\left(\alpha_{\mathrm{c}}(\mathrm{t}) - \theta'_{\mathrm{c}}(\mathrm{t})\right) \\ &\qquad+ \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\left[\frac{\left(1-\gamma'_{\mathrm{m}}(\mathrm{t};\psi'_{\mathrm{m}})\right)^2}{N\,\psi'_{\mathrm{m}}(\mathrm{t})}\right]\Big(\mathrm{diag}\left(\theta'_{\mathrm{c}}(\mathrm{t})\right) - \theta'_{\mathrm{c}}(\mathrm{t})\otimes\theta'_{\mathrm{c}}(\mathrm{t})\Big)\end{aligned} \tag{5.12}$$

and thus the covariance (2.36) can be represented as

$$\mathrm{Cov}(\mathrm{x};\theta_{\mathrm{m}},\theta_{\mathrm{c}}) = \mathrm{C}_{\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{t},\psi'_{\mathrm{m}},\psi'_{\mathrm{c}}}\right]\left(T(\mathrm{x});\theta_{\mathrm{m}},\theta_{\mathrm{c}}\right)\,. \tag{5.13}$$

Substituting the estimator bias and variance into (2.37), the conditional second moments of $\Delta(\mathrm{x};\mathrm{D},\theta_{\mathrm{c}})$ are

$$\begin{aligned}&\mathrm{E}_{\mathrm{D}\,|\,\theta_{\mathrm{m}},\theta_{\mathrm{c}}}\left[\Delta(\mathrm{x};\mathrm{D},\theta_{\mathrm{c}})\otimes\Delta(\mathrm{x};\mathrm{D},\theta_{\mathrm{c}})\right] \\ &\quad\equiv \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\Big[\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\psi'_{\mathrm{m}}\right)^2\Big]\Big(\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\Big)\otimes\Big(\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\Big) \\ &\qquad+ \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\left[\frac{\left(1-\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\psi'_{\mathrm{m}}\right)\right)^2}{N\,\psi'_{\mathrm{m}}\left(T(\mathrm{x})\right)}\right]\Big(\mathrm{diag}\left(\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\right) - \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\otimes\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\Big) \\ &\qquad+ \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\Big[\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\psi'_{\mathrm{m}}\right)\Big]\Big(\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\Big)\otimes\Big(\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})\Big) \\ &\qquad+ \mathrm{E}_{\psi'_{\mathrm{m}}\,|\,\theta'_{\mathrm{m}}}\Big[\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\psi'_{\mathrm{m}}\right)\Big]\Big(\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})\Big)\otimes\Big(\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)\Big) \\ &\qquad+ \Big(\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})\Big)\otimes\Big(\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})\Big)\,.\end{aligned} \tag{5.14}$$

Note that the first two terms are analogous to the two terms in (3.49), with $T(\mathrm{x})$ in place of x and the appropriate substitutions for the discretized functions; the remaining terms quantify additional deviation due to the discretization bias.

5.3.1 Trends

The bias-variance trends for the discretized Dirichlet distribution estimate (5.1) have similarities with the trends of the discrete-domain Dirichlet distribution as detailed in Section 3.2. The bias and covariance formulae contain the same expectation forms as before, with θ_{m}, θ_{c}, ψ_{m}, and γ_{m} being replaced by their new discretization variants; the earlier analysis is directly applicable with the appropriate substitutions. Note that $N\,\psi'_{\mathrm{m}}(t)|\,\theta'_{\mathrm{m}}(t) \sim \mathrm{Bi}\left(N, \theta'_{\mathrm{m}}(t)\right)$; thus, the relevant expectations will implicitly depend on the selection of the discretization function T.

There is one critical difference from the previous results: since the distribution estimate (5.1) mixes the discretized θ'_{c} instead of the true model θ_{c}, the new bias formula (5.11) contains an additional discretization term. This term depends neither on the training data volume N nor on the Dirichlet parameterization $(\alpha_0, \alpha_{\mathrm{m}}, \alpha_{\mathrm{c}})$. Consider the effect of the training volume N on the bias. For $N = 0$, the data-independent distribution $\alpha_{\mathrm{c}}\left(T(\mathrm{x})\right)$ is again maximally biased and has zero variance; the bias $\mathrm{Bias}(\mathrm{x}; \theta_{\mathrm{m}}, \theta_{\mathrm{c}}) = \alpha_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})$ now depends on the conditional prior mean only at values $t \in \mathcal{T} \subset \mathcal{X}$. As $N \to \infty$, for values x satisfying $\theta_{\mathrm{m}}(x) > 0$, the expected value of the Bayesian predictive distribution tends to the discretized predictive model $\theta'_{\mathrm{c}}\left(T(\mathrm{x})\right)$. As a result, only the first term in the bias formula (5.11) vanishes; the residual $\mathrm{Bias}(\mathrm{x}; \theta_{\mathrm{m}}, \theta_{\mathrm{c}}) = \theta'_{\mathrm{c}}\left(T(\mathrm{x})\right) - \theta_{\mathrm{c}}(\mathrm{x})$ can not be reduced any further. It is clear that, unlike its discrete-domain relative (3.30), the distribution estimate (5.1) will *not* necessarily converge to the true predictive distribution θ_{c} in the limit of training data volume.

Next, consider the effects of the user-selected parameters. The dependency of the bias-variance trade-off on the Dirichlet prior parameterization is conceptually

identical to that of the discrete-domain scenario in Section 3.2. However, the design of the discretization transform T also affects the bias and variance of the predictive distribution estimation. The extreme cases will be analyzed.

First, consider the finest discretization, such that $|\,\mathcal{T}\,| \to \infty$; assume that $\mathcal{X}$ is bounded and the function T is selected such that the volume of the sets $\mathcal{X}'(t)$ tend to $\int_{\mathcal{X}'(t)} \mathrm{d}x \to 0$. Consequently, $\mathrm{p}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}} \to \delta(\cdot - \mathrm{t})$ and thus $\theta'_\mathrm{c}(t) \to \theta_\mathrm{c}(t)$. Based on our assumptions regarding the transform, $T(x) \approx x$ and any smooth function g over the domain $\mathcal{X}$ will satisfy $g(x) \approx g(t),\ \forall x \in \mathcal{X}'(t)$. Together, these properties dictate that the discretization bias $\left\| \theta'_\mathrm{c}\left(T(\mathrm{x})\right) - \theta_\mathrm{c}(\mathrm{x}) \right\| \to 0$. However, the overall bias may still be high. If the marginal model θ_m is bounded, the discretized model tends to $\theta'_\mathrm{m} \to 0$ and thus $\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m} \to 0$. As a result, $\gamma'_\mathrm{m}(\psi'_\mathrm{m}) \to 1$ and the first term in (5.11) is maximal. Also, observe that the covariance function (5.13) will tend to zero. Note that these trends are the same as those exhibited by the continuous-domain Dirichlet predictive distribution, as detailed in Section 4.3; the higher the cardinality of the transformed observation set, the larger the training volume N required to effectively learn the true model.

Next, consider $\mathcal{T}$ to be a singleton, such that $|\,\mathcal{T}\,| = 1$ and thus $\mathcal{X}'(t) = \mathcal{X}$. The discretization bias is at its highest, as the discretized conditional model is $\theta'_\mathrm{c}(t) = \int_\mathcal{X} \theta_\mathrm{c}(x)\,\theta_\mathrm{m}(x)\mathrm{d}x$ and the same predictive model $\theta'_\mathrm{c}\left(T(x)\right)$ is used for all $x \in \mathcal{X}$. The benefit to such coarse discretization is that, on average, more data is available for each transformed observation. In this case, $\theta'_\mathrm{m} = 1$, such that $\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m} \to 1$ and the weighting function tends to $\gamma'_\mathrm{m}(\psi'_\mathrm{m}) \to \left(1 + \frac{N}{\alpha_0\,\alpha'_\mathrm{m}(t)}\right)^{-1}$. As a result, the first term in the bias formula (5.11) is minimal (assuming the Dirichlet parameters are held constant). Additionally, if the training data volume N is sufficiently high, this discretization result in minimal predictor variance; this is a consequence of the conditional variance of ψ'_c being inversely proportional to ψ'_m.

Evidently, the choice of how heavy a discretization to use effects its own trade-off. Finer discretization is required to avoid data-independent discretization bias, yet it may

incur higher bias overall if the prior mean α_{c} is poorly selected. Coarser discretization provides more data for the estimation of each distribution $\theta'_{\mathrm{c}}(t)$, reducing the variance, but the discretization bias may severe if the true predictive distribution θ_{c} changes rapidly with x.

5.4 Applications to Common Loss Functions

New regularized risk form?? Empirical loss now in feature space?!??

$$\begin{aligned}
\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] &= \int_{\mathcal{Y}} \mathcal{L}(h,y)\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\mathrm{d}y \\
&= \gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\,\right)\int_{\mathcal{Y}} \alpha_{\mathrm{c}}\left(y;T(\mathrm{x})\right)\mathcal{L}(h,y)\mathrm{d}y \\
&\quad +\Big(1-\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\,\right)\Big)\int_{\mathcal{Y}} \Psi'_{\mathrm{c}}\left(y;T(\mathrm{x});\mathrm{D}\,\right)\mathcal{L}(h,y)\mathrm{d}y \\
&= \gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\,\right)\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathcal{L}(h,\mathrm{y})\right] \\
&\quad +\Big(1-\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\,\right)\Big)\frac{\sum_{n=1}^{N}\delta\big[T(\mathrm{x}),T(\mathrm{X}_n)\big]\,\mathcal{L}\left(h,\mathrm{Y}_n\right)}{\sum_{n=1}^{N}\delta\big[T(\mathrm{x}),T(\mathrm{X}_n)\big]}\,.
\end{aligned} \tag{5.15}$$

Above requires low-dim prior definition

5.4.1 Regression: the Squared-Error Loss

5.4.1.1 Bayesian Estimation

Optimal Estimator The optimal function is the expected value of the output conditional PDF,

$$\begin{aligned}
f^*(\mathrm{x};\mathrm{D}) &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \\
&= \gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\,\right)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(T(\mathrm{x})\big) + \Big(1-\gamma'_{\mathrm{m}}\left(T(\mathrm{x});\mathrm{X}\,\right)\Big)\frac{\sum_{n=1}^{N}\delta\big[T(\mathrm{x}),T(\mathrm{X}_n)\big]\,\mathrm{Y}_n}{\sum_{n=1}^{N}\delta\big[T(\mathrm{x}),T(\mathrm{X}_n)\big]}\,.
\end{aligned} \tag{5.16}$$

Inheriting the properties of the discretized predictive distribution (5.1), the empirical mean for a given observation x averages all values Y_n whose corresponding observation satisfy $\mathrm{X}_n \in \mathcal{X}'\left(T(\mathrm{x})\right)$.

Minimum Bayes Risk PGR

5.4.1.2 Squared-Error Trends

The effects of discretization on the squared error will be analyzed next. Using the second moments of the random process $\Delta(\mathrm{x};\mathrm{D},\theta_\mathrm{c})$ formulated in (5.14), the excess squared error is formulated as

$$\begin{aligned}
\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right] \\
&\equiv \int_{\mathcal{Y}} y \int_{\mathcal{Y}} y'\, \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}\left[\Delta(y;\mathrm{x};\mathrm{D},\theta_\mathrm{c})\Delta(y';\mathrm{x};\mathrm{D},\theta_\mathrm{c})\right]\mathrm{d}y\mathrm{d}y' \\
&\equiv \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\mathrm{E}_{\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m}}\left[\,\gamma'_\mathrm{m}(\mathrm{t};\psi'_\mathrm{m})^2\right]\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}(\mathrm{t}) - \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right)^2\right] \\
&\quad + \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\mathrm{E}_{\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m}}\left[\frac{\left(1-\gamma'_\mathrm{m}(\mathrm{t};\psi'_\mathrm{m})\right)^2}{N\,\psi'_\mathrm{m}(\mathrm{t})}\right]\left(\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}] + \mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right)\right] \\
&\quad + \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\,\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right] \\
&\equiv \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\mathrm{E}_{\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m}}\left[\,\gamma'_\mathrm{m}(\mathrm{t};\psi'_\mathrm{m})^2\right]\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}(\mathrm{t}) - \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right)^2\right] \\
&\quad + \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\mathrm{E}_{\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m}}\left[\frac{\left(1-\gamma'_\mathrm{m}(\mathrm{t};\psi'_\mathrm{m})\right)^2}{N\,\psi'_\mathrm{m}(\mathrm{t})}\right]\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right] \\
&\quad + \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\left(1 + \mathrm{E}_{\psi'_\mathrm{m}\,|\,\theta'_\mathrm{m}}\left[\frac{\left(1-\gamma'_\mathrm{m}(\mathrm{t};\psi'_\mathrm{m})\right)^2}{N\,\psi'_\mathrm{m}(\mathrm{t})}\right]\right)\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right], \qquad (5.17)
\end{aligned}$$

a weighted summation of three terms, each of which can be viewed as second-order in terms of y. Note that the expectations over the observations space are performed using the operator equivalence $\mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}} \equiv \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\,\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}$, allowing select terms in (5.14) to be evaluated strictly in terms of the transformed observation t. The representation $\theta'_\mathrm{c}(\mathrm{t}) = \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}\left[\,\theta_\mathrm{c}(\mathrm{x})\right] = \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$ is used throughout.

Comparing with the discrete-domain excess squared error (3.64), there are both similar and new terms. The first two terms are directly comparable; instead of integrating over the observation space $\mathcal{X}$, however, the expectation is evaluated using the transformed observation t. Additionally, note that the clairvoyant regressor (2.41) and the its conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}$ are replaced by their expectations with respect to the conditional distribution $\mathrm{p}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}$. Note that $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$ represents the

discretized clairvoyant regressor; the expected value of the Bayesian regressor tends to this function in the limit of training data volume. The additional term quantifies the discretization error via $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$, measuring the variation of the clairvoyant regressor within each discretization subset $\mathcal{X}'(t)$.

Next, consider the trends of the excess squared error (5.17); the trends will have similarities with those of the discrete-domain observation problem discussed in Section 3.3.1, with the appropriate deviations detailed in Section 5.3.

First, consider the trends with training data volume N. For $N = 0$, the three weighting factors equate to one, zero, and one, respectively; as such, the excess squared error measures the deviation between the data-independent regressor $\mu_{\mathrm{y}\,|\,\mathrm{x}}(\mathrm{t})$ and the discretized clairvoyant regressor $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$, as well as the variance of the clairvoyant estimator within the discretization subsets. Combined the formula can also be represented as $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(T(\mathrm{x})\big) - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right)^2\right]$, demonstrating error due to the maximal bias. Conversely, as $N \to \infty$, only the new discretization error remains, such that $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\,\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right]$, the total average variance of the clairvoyant estimate within the subsets $\mathcal{X}'(t)$. Because of this unavoidable bias, the discretized regressor is unable to achieve the irreducible risk $\mathcal{R}^*_\Theta(\theta)$, no matter how much data is available for training – this is the fundamental drawback of using discretization.

Next, consider the trends with the user-defined parameters. Again, for $\alpha_0 \to \infty$, the excess risk is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(T(\mathrm{x})\big) - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right)^2\right]$ due to the maximally biased estimation. For $\alpha_0 \to 0$, the expectations with respect to ψ'_m will be analogous to those provided in Section 3.3.1. Also, the optimal conditional prior concentrations $\bar{\alpha}'_0(t) \equiv \alpha_0\,\alpha'_\mathrm{m}(t)$ can be shown to be

$$\bar{\alpha}'_0(\mathrm{t}) = \frac{\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}] + \mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]}{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}(\mathrm{t}) - \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right)^2} \tag{5.18}$$

for a given model θ and conditional prior mean α_c. As with the discrete-domain results, the optimal conditional concentrations are directly proportionate to the conditional model variance and inversely proportional to the squared bias between the prior

regressor and the optimal regressor. Note that the dependency is measured through the expectation of these values with respect to $\mathrm{p}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}$. Additionally, observe that the optimal concentrations $\bar{\alpha}'_0(\mathrm{t})$ are directly proportional to the clairvoyant regressor variations $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$; discretization adds this additional source of uncertainty in the target value y, increasing the risk of overfitting and motivating decreased weight on the empirical mean.

Lastly, consider the excess risk trends with the selection of the discretization transform T. As discussed in Section 5.3, with $|\,\mathcal{T}\,| \to \infty$ and $\int_{\mathcal{X}'(t)} \mathrm{d}x \to 0$, the conditional distribution $\mathrm{p}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}$ concentrates and the discretized conditional model θ'_c tends to the true model θ_c; as a result, $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}] \to 0$ and the discretization error is eliminated. However, with a decreasing expected volume of data $N\,\psi'_\mathrm{m}(t)$ matching each transformed observation, the expected value of the regressor tends to $f^*(\mathrm{x};\mathrm{D}) \to \mu_{\mathrm{y}\,|\,\mathrm{x}}\big(T(\mathrm{x})\big) \approx \mu_{\mathrm{y}\,|\,\mathrm{x}}$ and the excess risk tends to the high bias, zero variance value, $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{t}\,|\,\theta_\mathrm{m}}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}(\mathrm{t}) - \mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\right)^2\right] \approx \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right)^2\right]$. Note that this is the same as the continuous-domain Dirichlet regressor error for bounded θ_m, as detailed in Section 4.4.1.

Contrasting, if the transform set is singleton, the expectations of ψ'_m used for weighting will be at their lowest values (given sufficient training data); see the discussion in Section 5.3. However, the extremity of the averaging in $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$ and in the discretization variance term $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]$ will generally cause the total excess error to be prohibitively high. In practice, the degree of discretization must balance these competing sources of squared error risk.

Optimal discretizer for fixed Dirichlet params? Aggregate continuous alpha...

5.4.1.3 Example

To demonstrate the efficacy of the discretized Dirichlet regressor, the scenario detailed in Section 4.4.1.3 will be again used; $\mathcal{X}$ and $\mathcal{Y}$ are the closed unit interval and the true model dictates a non-linear clairvoyant regressor. The Bayesian linear regressor

detailed in Section 3.3.1.3 will be used for comparison; note that it can operate on the continuous data without any modification.

The set of discretized observations is defined as

$$\mathcal{T} \equiv \begin{cases} \left\{ \frac{i}{M-1} : i = 0, \ldots, M-1 \right\} & \text{if } M > 1, \\ \{0\} & \text{if } M = 1\ , \end{cases} \tag{5.19}$$

where the cardinality $|\,\mathcal{T}\,| \equiv M$ is selected by the designer. Note that when $M = 128$, this set is equivalent to the set used for the discrete-domain Dirichlet example in Section 3.3.1.3. The discretization transform used is $T(x) = \arg\min_{t \in \mathcal{T}} \|x - t\|$, rounding each observation to the nearest discretized value.

Recall that the continuous-domain Dirichlet learner previously used was parameterized by $\alpha_{\mathrm{m}} = 1$ and $\alpha_{\mathrm{c}}(x) = \mathrm{Beta}(2, 2)$. For the discretized regressor, the discrete set marginal function is defined as $\alpha'_{\mathrm{m}}(t) = \int_{\mathcal{X}'(t)} \alpha_{\mathrm{m}}(x)\mathrm{d}x$, an aggregation of the continuous-domain parameterizing function; using the discretization transform, α'_{m} is approximately uniform. As before, the prior estimator $\mu_{\mathrm{y}\,|\,\mathrm{x}} = 0.5$ is constant.

Fig. 5.1 provides visualization of the Dirichlet-based predictor statistics when different degrees of discretization $|\,\mathcal{T}\,|$ are used. The clairvoyant regressor and Bayesian linear regressor are included, as well. Observe that the $|\,\mathcal{T}\,| = 4$ regressor implements the coarsest discretization. As a result, it has low variance and also has low bias relative to the discretized clairvoyant regressor $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_{\mathrm{m}}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}]\big(T(x)\big)$. However, the large size of the discretization subsets $\mathcal{X}'(t)$ result in high discretization bias proportionate to $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{t},\theta_{\mathrm{m}}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}]$. The fine $|\,\mathcal{T}\,| = 4096$ discretization results in a predictor with negligible discretization bias, but with high variance and severe bias relative to the discretized clairvoyant regressor due to fewer average training observations per transformed value $t \in \mathcal{T}$. It is visually evident that the $|\,\mathcal{T}\,| = 128$ predictor has the lowest total bias, but also suffers from variance due to limited data volume.

Note that the Bayesian linear regressor has prediction statistics similar to those of the coarsest discretization regressors – it has high bias and low variance. This is intuitive, as it compresses the training data into a 2-dimensional parameter space, and

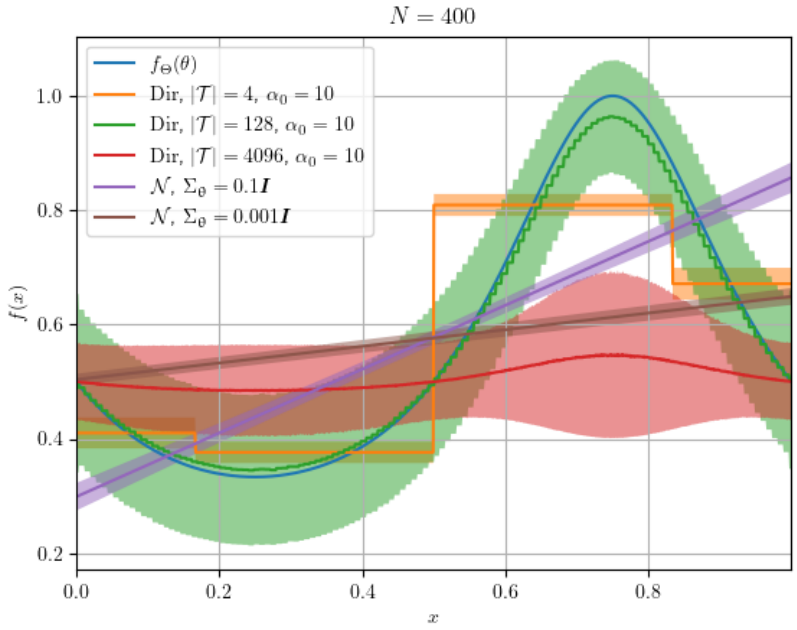

Figure 5.1: Predictor mean/variance, comparative

thus has fewer degrees-of-freedom than even the $|\mathcal{T}| = 4$ Dirichlet-based regressor.

Expand parameter space dimensionality discussion!?

This bias-variance trade-off is directly comparable to the trade-off effected by the selection of the prior concentration α_0. Using lower $|\mathcal{T}|$ imposes a more serious restriction on the regressors that can be realized by the learner. Similarly, using high α_0 limits the sensitivity of the learning to the training data. Both parameterizations thus lead to higher prediction bias and lower variance. Conversely, finer discretization with high $|\mathcal{T}|$ provides a similar effect to low α_0; that is, the empirical distribution is emphasized for prediction. Refer to Section 3.3.1.3 for demonstration of the predictor statistics with different prior concentration.

To underscore the limitation of predictors with coarse discretization, consider Fig. 5.2. Unlike the predictors visualized in Section 3.3.1.3, the discretized predictors converge to $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{t},\theta_\mathrm{m}}[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}]\big(T(x)\big)$ in the limit $N \to \infty$, not to the clairvoyant regressor $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}$; this effects the independence of the discretization bias from the data volume

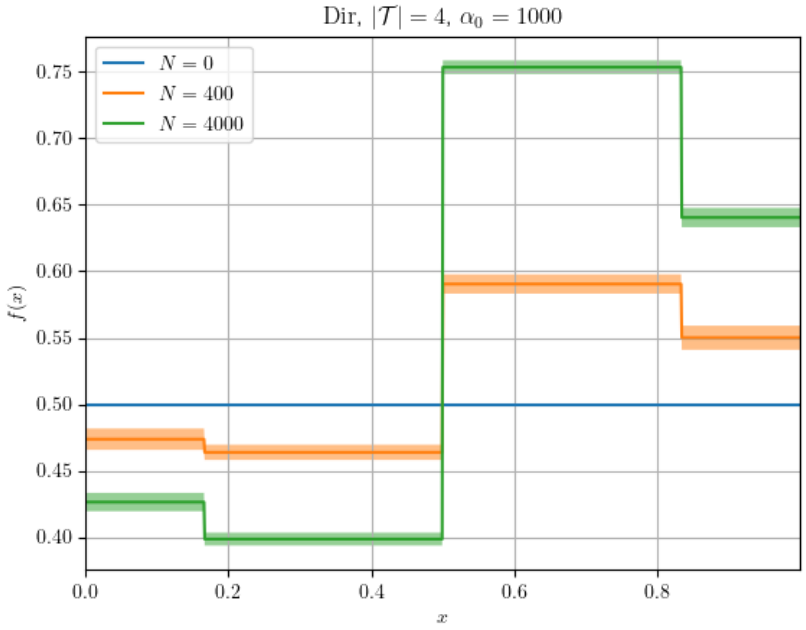

Figure 5.2: Dirichlet-based predictor mean/variance, varying N

N. As such, the quality of the match is dependent on the variation of the clairvoyant regressor and on the size of the subsets $\mathcal{X}'(t)$. Note that the persistent discretization bias is inherently related to the performance of the discretized Bayesian predictive distribution (5.1) as an estimator of the true model θ_c; due to the discretization, consistent estimation is no longer guaranteed.

Fig. 5.3 displays the squared error achieved by the Dirichlet-based regressors for varying data volumes N and for different discretization set sizes $|\mathcal{T}|$. Unlike the regressors derived from full-support priors (see the discrete domain results in Section 3.3.1.3), these regressors do not achieve the irreducible squared error $\mathcal{R}^*_\Theta(\theta)$ in the limit of training data volume N. The $|\mathcal{T}| = 4$ regressor is extremely sensitive to N and consequently outperforms the learners using finer discretization if N is relatively small; however, it produces the highest discretization error, making the discretization excessive when the training set is more voluminous. The $|\mathcal{T}| = 4096$ regressor uses such fine discretization that it it under-utilizes the data and it barely improves over

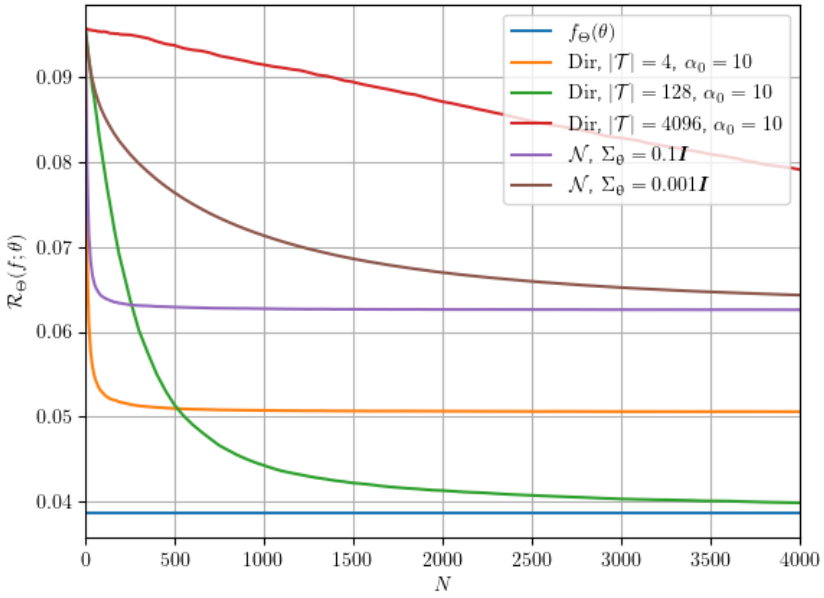

Figure 5.3: Squared-Error vs. training data volume N

the range of values N shown. Nonetheless, it has the smallest discretization error; if enough data is collected, it will eventually outperform all the other regressors and tend to a loss even lower than that realized by the $|\mathcal{T}| = 128$ regressor. Clearly, for more middling training data volumes, the $|\mathcal{T}| = 128$ regressor strikes the best balance between the two sources of risk. Also note that the Bayesian linear regressors suffer from the worst excess squared error due to their low-dimensionality priors and poorly-matched set of achievable prediction functions.

Fig. 5.4 demonstrates the error trends of different discretization learners for varying prior localizations α_0. To aid visualization, the risk is plotted against $\alpha_0/|\mathcal{T}|$; this normalized concentration is equivalent to the average of the conditional prior concentrations (5.18). The optimal values of α_0 optimize the bias-variance trade-off depending on the match between the true predictive model and the data-independent regressor $\mu_{\mathrm{y}|\mathrm{x}}$. Again, the optimal concentrations $\bar{\alpha}_0'(t)$ are independent of the data volume N; however, they are not independent of the selected discretization transform T.

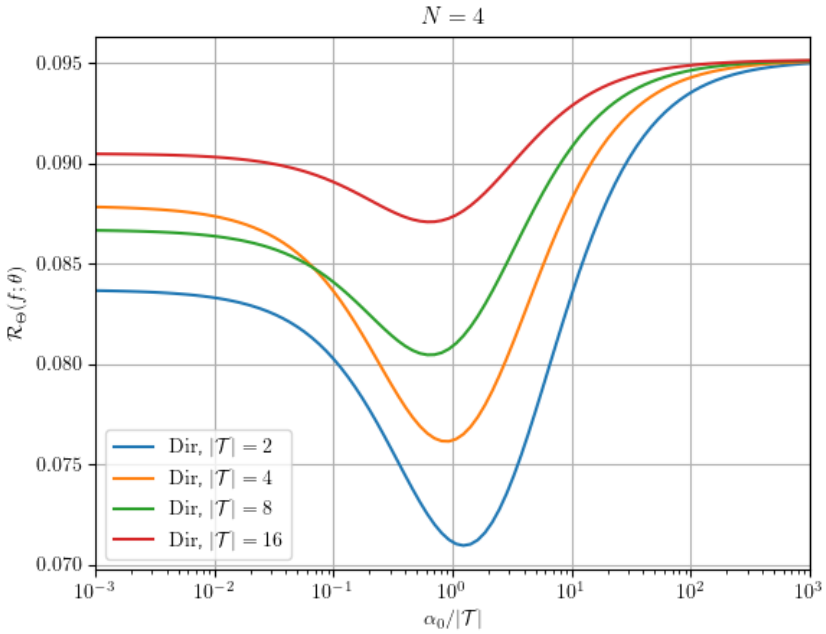

Figure 5.4: Squared-Error vs. prior localization α_0

Observe that the optimal conditional concentrations are higher when the discretization is coarser – the discretization causes the learner to perceive higher variation in the values Y_n satisfying $T(\mathrm{X}_n) = t$, such that $\bar{\alpha}_0'(t)$ is increased to de-emphasize the empirical mean.

To further demonstrate the importance of the selection of the discretization function, Figs. 5.5 and 5.6 demonstrate the error trends as a function of $|\mathcal{T}|$ for different data volumes and prior concentrations, respectively. In the former, note that the transform set cardinality that minimizes the squared error is directly proportionate to the training data volume N. Conforming with the previous results, more data allows finer discretizers to be used, reducing the discretization error without the severe consequences of lower data sensitivity.

Remove zoom? Remove alpha fig or do hifi, add discussion! Equal argmins???

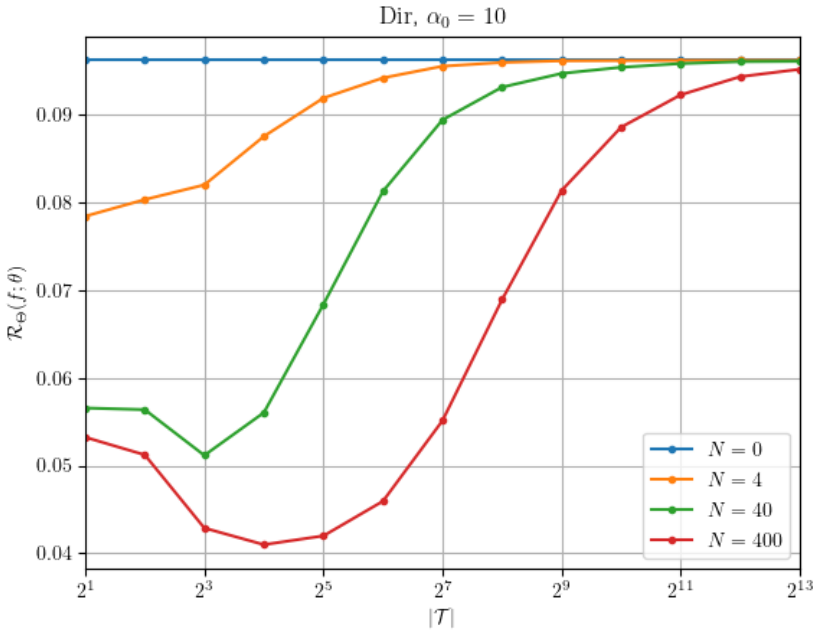

Figure 5.5: Squared-Error vs. discretization $|\mathcal{T}|$, various N

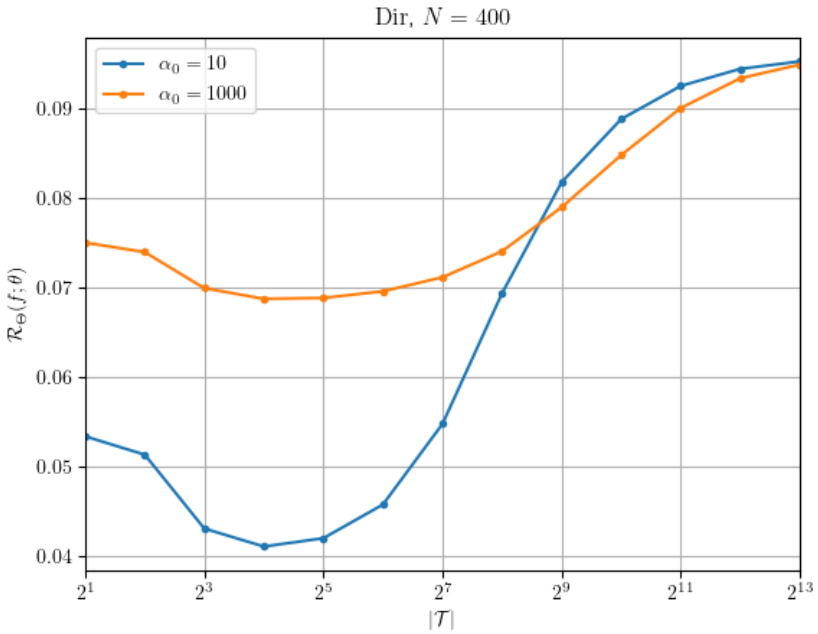

Figure 5.6: Squared-Error vs. discretization $|\mathcal{T}|$, various α_0

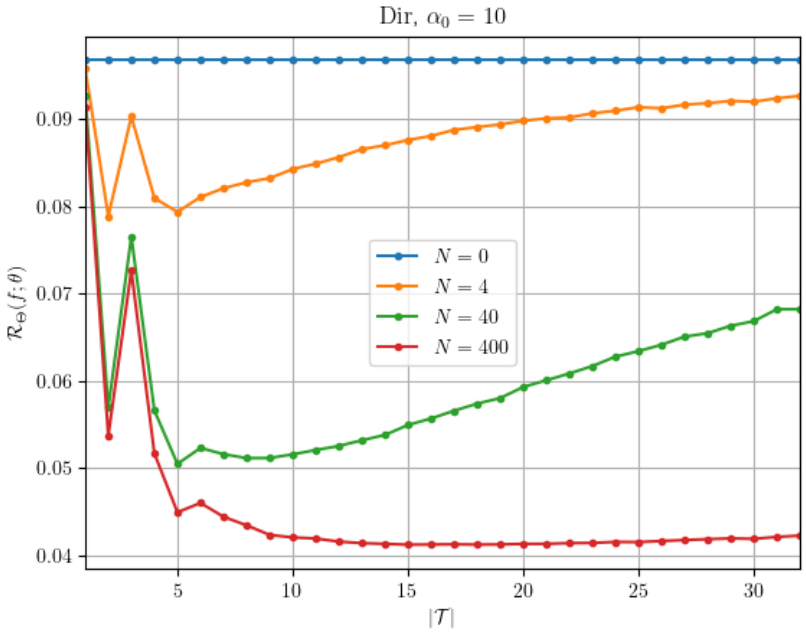

Figure 5.7: Squared-Error vs. discretization $|\mathcal{T}|$, various N

Appendix A

Discrete-Domain Random Processes

LOTS of redundancy...

This chapter details the properties of various discrete-domain random processes. The domain $\mathcal{Y}$ is assumed countable.

A.1 Empirical Distribution Properties

A.1.1 Aggregation

integer z? remove? use i?

concatenation notation?

A characteristic of an Empirical random process is that its aggregations are also Empirical processes. Consider a random process $\psi \sim \text{Emp}(N, \theta)$ drawn from $\Psi \subset \mathcal{P}(\mathcal{Y})$ for N samples and mean function θ. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$ and the corresponding function partitions $\psi_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\psi = (\ldots, \psi_z, \ldots)$, and $\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\theta = (\ldots, \theta_z, \ldots)$. The transformed random process $\psi_{\text{m}} \in \Psi_{\text{m}} \subset \mathcal{P}(\mathcal{Z})$, defined as $\psi_{\text{m}}(z) \equiv \sum_{y \in \mathcal{S}_z} \psi_z(y)$, is distributed as $\psi_{\text{m}} \sim \text{Emp}(N, \theta_{\text{m}})$ with a parameterizing distribution θ_{m} defined as $\theta_{\text{m}}(z) = \sum_{y \in \mathcal{S}_z} \theta_z(y)$.

To prove this principle, define the subset

$$\begin{aligned}\Psi'(\psi_{\mathrm{m}}) &= \prod_{z\in\mathcal{Z}} \Psi'_z\big(\,\psi_{\mathrm{m}}(z)\big) \\ &= \prod_{z\in\mathcal{Z}} \left\{ n_z/N : n_z \in {\mathbb{Z}_{\geq 0}}^{\mathcal{S}_z},\ \sum_{y\in\mathcal{S}_z} n_z(y) = N\,\psi_{\mathrm{m}}(z) \right\} \subset \Psi \end{aligned} \tag{A.1}$$

and observe that

$$\begin{aligned}\mathrm{P}_{\psi_{\mathrm{m}}\,|\theta}(\psi_{\mathrm{m}}\,|\theta) &= \sum_{\psi\in\Psi'(\psi_{\mathrm{m}})} \mathrm{P}_{\psi|\theta}(\psi|\theta) = \sum_{\psi\in\Psi'(\psi_{\mathrm{m}})} \mathcal{M}(N\psi) \left(\prod_{y\in\mathcal{Y}} \theta(y)^{\psi(y)} \right)^N \\ &= \mathcal{M}(N\,\psi_{\mathrm{m}}) \prod_{z\in\mathcal{Z}} \sum_{\psi_z\in\Psi'_z\left(\,\psi_{\mathrm{m}}(z)\right)} \mathcal{M}(N\psi_z) \left(\prod_{y\in\mathcal{S}_z} \theta_z(y)^{\psi_z(y)} \right)^N \\ &\equiv \mathcal{M}(N\,\psi_{\mathrm{m}}) \left(\prod_{z\in\mathcal{Z}} \theta_{\mathrm{m}}(z)^{\psi_{\mathrm{m}}(z)} \right)^N = \mathrm{Emp}(\psi_{\mathrm{m}}; N, \theta_{\mathrm{m}})\ , \end{aligned} \tag{A.2}$$

where the multinomial theorem [8] has been used.

A.1.2 Conditioned on its Aggregation

If the Empirical random process ψ is conditioned on its aggregation ψ_{m} over the partition $\{\ldots,\mathcal{S}_z,\ldots\}$, $z\in\mathcal{Z}$, the distinct segments ψ_z become independent random processes, such that for $\psi\in\Psi'(\psi_{\mathrm{m}})$,

$$\begin{aligned}\mathrm{P}_{\psi|\,\psi_{\mathrm{m}},\theta}(\psi|\,\psi_{\mathrm{m}},\theta) &= \frac{\mathrm{P}_{\psi,\psi_{\mathrm{m}}\,|\theta}(\psi,\psi_{\mathrm{m}}\,|\theta)}{\mathrm{P}_{\psi_{\mathrm{m}}\,|\theta}(\psi_{\mathrm{m}}\,|\theta)} \equiv \frac{\mathcal{M}(N\psi)}{\mathcal{M}(N\,\psi_{\mathrm{m}})} \left(\frac{\prod_{y\in\mathcal{Y}} \theta(y)^{\psi(y)}}{\prod_{z\in\mathcal{Z}} \theta_{\mathrm{m}}(z)^{\psi_{\mathrm{m}}(z)}} \right)^N \\ &= \prod_{z\in\mathcal{Z}} \left[\mathcal{M}(N\psi_z) \left(\frac{\prod_{y\in\mathcal{S}_z} \theta_z(y)^{\psi_z(y)}}{\theta_{\mathrm{m}}(z)^{\psi_{\mathrm{m}}(z)}} \right)^N \right] \\ &= \prod_{z\in\mathcal{Z}} \left[\mathcal{M}(N\psi_z) \left(\prod_{y\in\mathcal{S}_z} \left(\frac{\theta_z(y)}{\theta_{\mathrm{m}}(z)} \right)^{\psi_z(y)} \right)^N \right]. \end{aligned} \tag{A.3}$$

While these function segments are independent, they are not Empirical processes. Introducing the conditional distributions $\theta_{\mathrm{c}}(z) \equiv \theta_z/\,\theta_{\mathrm{m}}(z)$ and the normalized seg-

ments $\psi_c(z) \equiv \psi_z / \psi_m(z) \in \mathcal{P}(\mathcal{S}_z)$, it can be shown that

$$\begin{aligned} P_{\psi_c \,|\, \psi_m, \theta}(\psi_c \,|\, \psi_m, \theta) &\equiv \prod_{z \in \mathcal{Z}} \left[\mathcal{M}\left(N\, \psi_m(z)\, \psi_c(z)\right) \left(\prod_{y \in \mathcal{S}_z} \theta_c(y; z)^{\psi_c(y;z)} \right)^{N\, \psi_m(z)} \right] \\ &= \prod_{z \in \mathcal{Z}} \mathrm{Emp}\left(\psi_c(z); N\, \psi_m(z), \theta_c(z) \right) . \end{aligned} \tag{A.4}$$

Thus, when also conditioned on the aggregation ψ_m, the individual random processes $\psi_c(z)$ are independent Empirical processes of $N\, \psi_m(z)$ samples, parameterized by the distributions $\theta_c(z) \in \mathcal{P}(\mathcal{S}_z)$.

A.2 Dirichlet Distribution Properties

A.2.1 Aggregation

It is known that Dirichlet aggregations are also Dirichlet [7]. Let the random process $\theta \sim \mathrm{Dir}(\alpha_0, \alpha)$ drawn from $\Theta \equiv \mathcal{P}(\mathcal{Y})$ be Dirichlet with concentration $\alpha_0 \in \mathbb{R}^+$ and mean function $\alpha \in \left\{ \mathbb{R}^{+\mathcal{Y}} : \sum_{y \in \mathcal{Y}} \alpha(y) = 1 \right\}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$ and the corresponding function partitions $\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\theta = (\ldots, \theta_z, \ldots)$ and $\alpha_z \in \mathbb{R}^{+\mathcal{S}_z}$, such that $\alpha = (\ldots, \alpha_z, \ldots)$. The transformed random process $\theta_m \in \mathcal{P}(\mathcal{Z})$, defined as $\theta_m(z) \equiv \sum_{y \in \mathcal{S}_z} \theta_z(y)$, is distributed as $\theta_m \sim \mathrm{Dir}(\alpha_0, \alpha_m)$ with a parameterizing distribution α_m defined as $\alpha_m(z) = \sum_{y \in \mathcal{S}_z} \alpha_z(y)$.

To prove this principle, define the subset

$$\begin{aligned} \Theta'(\theta_m) &= \prod_{z \in \mathcal{Z}} \Theta'_z\big(\theta_m(z)\big) \\ &= \prod_{z \in \mathcal{Z}} \left\{ \theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z} : \sum_{y \in \mathcal{S}_z} \theta_z(y) = \theta_m(z) \right\} \subset \Theta \end{aligned} \tag{A.5}$$

and note that

$$\begin{aligned}
\mathrm{p}_{\theta_\mathrm{m}}(\theta_\mathrm{m}) &= \int_{\Theta'(\theta_\mathrm{m})} \mathrm{p}_\theta(\theta)\mathrm{d}\theta = \int_{\Theta'(\theta_\mathrm{m})} \beta(\alpha_0\alpha)^{-1} \prod_{y\in\mathcal{Y}} \theta(y)^{\alpha_0\alpha(y)-1}\mathrm{d}\theta \\
&= \beta(\alpha_0\alpha)^{-1} \prod_{z\in\mathcal{Z}} \int_{\Theta'_z\left(\theta_\mathrm{m}(z)\right)} \prod_{y\in\mathcal{S}_z} \theta_z(y)^{\alpha_0\alpha_z(y)-1}\mathrm{d}\theta_z \\
&= \beta(\alpha_0\alpha)^{-1} \prod_{z\in\mathcal{Z}} \theta_\mathrm{m}(z)^{\alpha_0\,\alpha_\mathrm{m}(z)-1} \int_{{\Theta''}_z} \prod_{y\in\mathcal{S}_z} \theta_\mathrm{c}(y;z)^{\alpha_0\alpha_z(y)-1}\mathrm{d}\,\theta_\mathrm{c}(z) \\
&= \beta(\alpha_0\alpha)^{-1} \prod_{z\in\mathcal{Z}} \theta_\mathrm{m}(z)^{\alpha_0\,\alpha_\mathrm{m}(z)-1} \beta(\alpha_0\alpha_z) \\
&= \beta(\alpha_0\,\alpha_\mathrm{m})^{-1} \prod_{z\in\mathcal{Z}} \theta_\mathrm{m}(z)^{\alpha_0\,\alpha_\mathrm{m}(z)-1} = \mathrm{Dir}(\theta_\mathrm{m};\alpha_0,\alpha_\mathrm{m})
\end{aligned} \tag{A.6}$$

where the transform $\theta_\mathrm{c}(z) \equiv \theta_z/\,\theta_\mathrm{m}(z) \in {\Theta''}_z = \left\{\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z} : \sum_{y\in\mathcal{S}_z}\theta_z(y) = 1\right\}$ has been used. Note that the determinant of the transform dictates $\mathrm{d}\,\theta_\mathrm{c}(z) = \theta_\mathrm{m}(z)^{1-|\mathcal{S}_z|}\mathrm{d}\theta_z$.

cite Jacobian?

A.2.2 Conditioned on its Aggregation

This section details another important property of Dirichlet distributed random processes – when conditioned on its own aggregation θ_m, the partitioned segments θ_z of the process become independent. Furthermore, the normalized functions $\theta_\mathrm{c}(z)$ are also Dirichlet processes.

The PDF of the original random process θ conditioned on its aggregation θ_m can be formulated as

$$\begin{aligned}
\mathrm{p}_{\theta|\,\theta_\mathrm{m}}(\theta|\,\theta_\mathrm{m}) &= \frac{\beta(\alpha_0\,\alpha_\mathrm{m})\prod_{y\in\mathcal{Y}}\theta(y)^{\alpha_0\alpha(y)-1}}{\beta(\alpha_0\alpha)\prod_{z\in\mathcal{Z}}\theta_\mathrm{m}(z)^{\alpha_0\,\alpha_\mathrm{m}(z)-1}} \\
&\equiv \prod_{z\in\mathcal{Z}}\left[\beta(\alpha_0\alpha_z)^{-1}\frac{\prod_{y\in\mathcal{S}_z}\theta_z(y)^{\alpha_0\alpha_z(y)-1}}{\theta_\mathrm{m}(z)^{\alpha_0\,\alpha_\mathrm{m}(z)-1}}\right] \\
&= \prod_{z\in\mathcal{Z}}\left[\frac{\theta_\mathrm{m}(z)^{1-|\mathcal{S}_z|}}{\beta(\alpha_0\alpha_z)}\prod_{y\in\mathcal{S}_z}\left(\frac{\theta_z(y)}{\theta_\mathrm{m}(z)}\right)^{\alpha_0\alpha_z(y)-1}\right],
\end{aligned} \tag{A.7}$$

which is defined for $\prod_{z\in\mathcal{Z}}\left\{\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z} : \sum_{y\in\mathcal{S}_z}\theta_z(y) = \theta_\mathrm{m}(z)\right\}$.

Observe that the partitioned segements θ_z are conditionally independent. Introducing the normalized functions $\alpha_c(z) \equiv \alpha_z / \alpha_m(z)$ and the normalized random processes $\theta_c(z) \equiv \theta_z / \theta_m(z) \in \mathcal{P}(\mathcal{S}_z)$, it can be shown that

$$\begin{aligned} p_{\theta_c \,|\, \theta_m}(\theta_c \,|\, \theta_m) &= \prod_{z \in \mathcal{Z}} \left[\beta\big(\alpha_0 \, \alpha_m(z) \, \alpha_c(z)\big)^{-1} \prod_{y \in \mathcal{S}_z} \theta_c(y;z)^{\alpha_0 \, \alpha_m(z) \, \alpha_c(y;z) - 1} \right] \\ &= \prod_{z \in \mathcal{Z}} \mathrm{Dir}\left(\theta_c(z); \alpha_0 \, \alpha_m(z), \alpha_c(z)\right) . \end{aligned} \tag{A.8}$$

Thus after conditioning, the normalized processes $\theta_c(z)$ are Dirichlet distributed, independent of one another, and independent of the aggregation θ_m.

PGR: discuss transform Jacobian and dimensionality?

A.3 Dirichlet-Empirical Distribution Properties

A.3.1 Aggregation

The Dirichlet-Empirical distribution is so named since it is the expectation of a Empirical distribution $\mathrm{Emp}(N, \theta)$ with respect to its mean function, a Dirichlet process $\theta \sim \mathrm{Dir}(\alpha_0, \alpha)$. Naturally, these random processes share many properties with the Empirical distribution.

A characteristic of a Dirichlet-Empirical random process is that its aggregations are also Dirichlet-Empirical – this is inherited from the related Dirichlet-Multinomial distribution [9]. Consider a DE random process $\psi \sim \mathrm{DE}(N, \alpha_0, \alpha)$ drawn from $\Psi \subset \mathcal{P}(\mathcal{Y})$ of N samples, concentration α_0 and mean function α. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$ and the corresponding function partitions $\psi_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\psi = (\ldots, \psi_z, \ldots)$, and $\alpha_z \in {\mathbb{R}^+}^{\mathcal{S}_z}$, such that $\alpha = (\ldots, \alpha_z, \ldots)$. The transformed random process $\psi_m \in \Psi_m \subset \mathcal{P}(\mathcal{Z})$, defined as $\psi_m(z) \equiv \sum_{y \in \mathcal{S}_z} \psi_z(y)$, is distributed as $\psi_m \sim \mathrm{DE}(N, \alpha_0, \alpha_m)$ with mean function α_m defined as $\alpha_m(z) = \sum_{y \in \mathcal{S}_z} \alpha_z(y)$.

To prove this principle, define the subset

$$\begin{aligned}\Psi'(\psi_{\mathrm{m}}) &= \prod_{z\in\mathcal{Z}} \Psi'_z\big(\,\psi_{\mathrm{m}}(z)\big) \\ &= \prod_{z\in\mathcal{Z}} \left\{ n_z/N : n_z \in {\mathbb{Z}_{\geq 0}}^{\mathcal{S}_z},\ \sum_{y\in\mathcal{S}_z} n_z(y) = N\,\psi_{\mathrm{m}}(z) \right\} \subset \Psi\ . \end{aligned} \tag{A.9}$$

Next, observe that

$$\begin{aligned}\mathrm{P}_{\psi_{\mathrm{m}}}(\psi_{\mathrm{m}}) &= \sum_{\psi\in\Psi'(\psi_{\mathrm{m}})} \mathrm{P}_{\psi}(\psi) = \sum_{\psi\in\Psi'(\psi_{\mathrm{m}})} \mathcal{M}(N\psi)\frac{\beta(\alpha_0\alpha + N\psi)}{\beta(\alpha_0\alpha)} \\ &= \mathcal{M}(N\,\psi_{\mathrm{m}})\frac{\beta(\alpha_0\,\alpha_{\mathrm{m}} + N\,\psi_{\mathrm{m}})}{\beta(\alpha_0\alpha)} \prod_{z\in\mathcal{Z}} \sum_{\psi_z\in\Psi'_z\left(\,\psi_{\mathrm{m}}(z)\right)} \mathcal{M}(N\psi_z)\frac{\beta(\alpha_0\alpha_z + N\psi_z)}{\beta(\alpha_0\alpha_z)} \\ &= \mathcal{M}(N\,\psi_{\mathrm{m}})\frac{\beta(\alpha_0\,\alpha_{\mathrm{m}} + N\,\psi_{\mathrm{m}})}{\beta(\alpha_0\alpha)} = \mathrm{DE}(\psi_{\mathrm{m}}; N, \alpha_0, \alpha_{\mathrm{m}})\ , \end{aligned} \tag{A.10}$$

where the identity

$$\sum_{\substack{n\in{\mathbb{Z}_{\geq 0}}^{\mathcal{Y}}:\\ \sum_y n(y)=N}} \mathcal{M}(n)\beta(a+n) = \beta\,(a) \tag{A.11}$$

has been used.

more proof steps?

cite identity

A.3.2 Conditioned on its Aggregation

If the Dirichlet-Empirical random process $\boldsymbol{\psi}$ is conditioned on its aggregation $\boldsymbol{\psi}_{\mathrm{m}}$ over the partition $\{\ldots,\mathcal{S}_z,\ldots\}$, $z\in\mathcal{Z}$, the distinct segements $\boldsymbol{\psi}_z$ become independent random processes, such that for $\psi\in\Psi'(\psi_{\mathrm{m}})$,

$$\begin{aligned}\mathrm{P}_{\boldsymbol{\psi}|\,\boldsymbol{\psi}_{\mathrm{m}}}(\psi|\,\psi_{\mathrm{m}}) &= \frac{\mathcal{M}(N\psi)\beta(\alpha_0\alpha)^{-1}\beta(\alpha_0\alpha + N\psi)}{\mathcal{M}(N\,\psi_{\mathrm{m}})\beta(\alpha_0\,\alpha_{\mathrm{m}})^{-1}\beta(\alpha_0\,\alpha_{\mathrm{m}} + N\,\psi_{\mathrm{m}})} \\ &= \left(\prod_{z\in\mathcal{Z}} \frac{\Gamma\big(\alpha_0\,\alpha_{\mathrm{m}}(z) + N\,\psi_{\mathrm{m}}(z)\big)}{\big(N\,\psi_{\mathrm{m}}(z)\big)!\,\Gamma\big(\,\alpha_{\mathrm{m}}(z)\big)}\right)^{-1} \left(\prod_{y\in\mathcal{Y}} \frac{\Gamma\big(\alpha_0\alpha(y) + N\psi(y)\big)}{\big(N\psi(y)\big)!\,\Gamma\big(\alpha_0\alpha(y)\big)}\right) \\ &= \prod_{z\in\mathcal{Z}} \left[\frac{\big(N\,\psi_{\mathrm{m}}(z)\big)!\,\Gamma\big(\,\alpha_{\mathrm{m}}(z)\big)}{\Gamma\big(\alpha_0\,\alpha_{\mathrm{m}}(z) + N\,\psi_{\mathrm{m}}(z)\big)} \prod_{y\in\mathcal{S}_z} \frac{\Gamma\big(\alpha_0\alpha_z(y) + N\psi_z(y)\big)}{\big(N\psi_z(y)\big)!\,\Gamma\big(\alpha_0\alpha_z(y)\big)}\right] \\ &= \prod_{z\in\mathcal{Z}} \mathcal{M}(N\psi_z)\frac{\beta(\alpha_0\alpha_z + N\psi_z)}{\beta(\alpha_0\alpha_z)}\ . \end{aligned} \tag{A.12}$$

While these individual function segments are independent, they are not Dirichlet-Empirical processes. Defining the functions $\alpha_{\mathrm{c}}(z) \equiv \alpha_z / \alpha_{\mathrm{m}}(z) \in \mathcal{P}(\mathcal{S}_z)$ and the normalized segments $\psi_{\mathrm{c}}(z) \equiv \psi_z / \psi_{\mathrm{m}}(z) \in \mathcal{P}(\mathcal{S}_z)$, it can be shown that

$$\begin{aligned} \mathrm{P}_{\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}}(\psi_{\mathrm{c}} \,|\, \psi_{\mathrm{m}}) &= \prod_{z \in \mathcal{Z}} \mathcal{M}\left(N\, \psi_{\mathrm{m}}(z)\, \psi_{\mathrm{c}}(z)\right) \frac{\beta\big(\alpha_0\, \alpha_{\mathrm{m}}(z)\, \alpha_{\mathrm{c}}(z) + N\, \psi_{\mathrm{m}}(z)\, \psi_{\mathrm{c}}(z)\big)}{\beta\big(\alpha_0\, \alpha_{\mathrm{m}}(z)\, \alpha_{\mathrm{c}}(z)\big)} \\ &= \prod_{z \in \mathcal{Z}} \mathrm{DE}\left(\psi_{\mathrm{c}}(z); N\, \psi_{\mathrm{m}}(z), \alpha_0\, \alpha_{\mathrm{m}}(z), \alpha_{\mathrm{c}}(z)\right), \end{aligned} \tag{A.13}$$

Thus, when conditioned on the aggregation ψ_{m}, the individual functions $\psi_{\mathrm{c}}(z) \in \mathcal{P}(\mathcal{S}_z)$ are independent Dirichlet-Empirical processes of $N\, \psi_{\mathrm{m}}(z)$ samples and concentration $\alpha_0\, \alpha_{\mathrm{m}}(z)$, with mean functions $\alpha_{\mathrm{c}}(z)$.

Double check.

Appendix B

Continuous-Domain Random Processes

This chapter details the properties of various continuous-domain random processes. The domain $\mathcal{Y}$ is assumed to be a continuous set.

B.1 Empirical Process Properties

B.1.1 Definition

This section introduces a new continuous-domain random process, referred to as the Empirical process (EP). It is the generalization of the Empirical distribution for i.i.d. samples drawn from a continuous set. The Empirical process $\psi \sim \mathrm{EP}(N, \theta)$ is parameterized by N samples and mean function $\theta \in \mathcal{P}(\mathcal{Y})$; it assumes functions from the set $\Psi = \left\{ N^{-1} \sum_{n=1}^{N} \delta(\cdot - D_n) : D \in \mathcal{Y}^N \right\} \subset \mathcal{P}(\mathcal{Y})$.

The continuous-domain Empirical process is characterized by the same aggregation property as its discrete-domain variant. Define a countable partition of $\mathcal{Y}$, $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$, and the corresponding function partitions $\psi_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\psi = (\ldots, \psi_z, \ldots)$, and $\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\theta = (\ldots, \theta_z, \ldots)$. Thus, for an Empirical process $\psi \in \Psi$, the aggregation $\psi_{\mathrm{m}} \in \Psi_{\mathrm{m}} \subset \mathcal{P}(\mathcal{X})$ satisfying $\psi_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z} \psi_z(y) \mathrm{d}y$

is an Empirical process with N samples and mean function θ_{m} satisfying $\theta_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z} \theta_z(y)\mathrm{d}y$.

Additionally, when further conditioned on the aggregation Ψ_{m}, the normalized random processes $\Psi_{\mathrm{c}}(z) \equiv \Psi_z / \Psi_{\mathrm{m}}(z)$ are independent continuous-domain Empirical processes, $\Psi_{\mathrm{c}}(z) | \Psi_{\mathrm{m}}(z), \Theta_{\mathrm{c}}(z) \sim \mathrm{EP}\left(N \Psi_{\mathrm{m}}(z), \Theta_{\mathrm{c}}(z)\right)$, where $\theta_{\mathrm{c}}(z) \equiv \theta_z / \theta_{\mathrm{m}}(z) \in \mathcal{P}(\mathcal{S}_z)$.

B.1.2 Mean and Correlation Functions

In this section, it is shown that the expected value of an Empirical process $\Psi \sim \mathrm{EP}(N, \theta)$ is

$$\mu_\Psi = \theta \ . \tag{B.1}$$

A defining characteristic of Empirical processes is that their aggregations are also Empirical. Define the partition of $\mathcal{Y} = \mathbb{R}$, $\left\{ \mathcal{S}(y), S^{\mathrm{c}}(y) \right\}$ where $\mathcal{S}(y) = (-\infty, y]$. The transform random process $\left(\Psi_{\mathrm{m}}, {\Psi_{\mathrm{m}}}^{\mathrm{c}} \right)$, where $\Psi_{\mathrm{m}} \equiv \int_\infty^y \Psi(t)\mathrm{d}t$, is thus a discrete-domain Empirical random process for N samples and mean $\left(\theta_{\mathrm{m}}, {\theta_{\mathrm{m}}}^{\mathrm{c}} \right)$, where $\theta_{\mathrm{m}} = \int_{-\infty}^y \theta(t)\mathrm{d}t$ and ${\theta_{\mathrm{m}}}^{\mathrm{c}} = \int_y^\infty \theta(t)\mathrm{d}t$. Dependency on y is suppressed for brevity. Observe that $\mu_{\Psi_{\mathrm{m}}}(y) = \theta_{\mathrm{m}}$ and thus that

$$\mu_{\Psi_{\mathrm{m}}} = \int_{-\infty}^y \theta(t)\mathrm{d}t = \int_{-\infty}^y \mu_\Psi(t)\mathrm{d}t \ .$$

Differentiating with respect to y, we have the expected value of the EP.

Next, the correlation function is shown to be

$$\mathrm{E}_\Psi \left[\Psi(y_1)\Psi(y_2) \right] = \frac{1}{N}\Theta(y_1)\delta(y_1 - y_2) + \left(1 - \frac{1}{N}\right) \Theta(y_1)\Theta(y_2) \ . \tag{B.2}$$

First, assume $y_2 \geq y_1$ and define a new partition of $\mathcal{Y}$, $\left\{(-\infty, y_1], (y_1, y_2], (y_2, \infty)\right\}$. By the aggregation property, the random triplet $\left(\int_{-\infty}^{y_1} \Psi(t)\mathrm{d}t, \int_{y_1}^{y_2} \Psi(t)\mathrm{d}t, \int_{y_2}^{\infty} \Psi(t)\mathrm{d}t \right)$ is Empirical with N samples and mean $\left(\int_{-\infty}^{y_1} \theta(t)\mathrm{d}t, \int_{y_1}^{y_2} \theta(t)\mathrm{d}t, \int_{y_2}^{\infty} \theta(t)\mathrm{d}t \right)$.

Define the function

$$
\begin{aligned}
g(t_1,t_2) &= \mathrm{E}_\psi\left[\int_{-\infty}^{y_1}\psi(t_1)\mathrm{d}t_1\int_{-\infty}^{y_2}\psi(t_2)\mathrm{d}t_2\right] \qquad \text{(B.3)}\\
&= \mathrm{E}_\psi\left[\left(\int_{-\infty}^{y_1}\psi(t_1)\mathrm{d}t_1\right)^2+\left(\int_{-\infty}^{y_1}\psi(t_1)\mathrm{d}t_1\right)\left(\int_{y_1}^{y_2}\psi(t_2)\mathrm{d}t_2\right)\right]\\
&= \frac{1}{N}\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)\left(1+(N-1)\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)+\left(1-\frac{1}{N}\right)\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)\left(\int_{y_1}^{y_2}\theta(t_2)\mathrm{d}t_2\right)\\
&= \frac{1}{N}\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)+\left(1-\frac{1}{N}\right)\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)\left(\int_{-\infty}^{y_2}\theta(t_2)\mathrm{d}t_2\right) \quad \forall y_2\geq y_1\ .
\end{aligned}
$$

Following the same steps provides the values of g for $t_2 \leq t_1$; the combined formula can be given as

$$
g(t_1,t_2) = \frac{1}{N}\left(\int_{-\infty}^{\min(y_1,y_2)}\theta(t_1)\mathrm{d}t_1\right)+\left(1-\frac{1}{N}\right)\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)\left(\int_{-\infty}^{y_2}\theta(t_2)\mathrm{d}t_2\right) \qquad \text{(B.4)}
$$

and, finally,

$$
\begin{aligned}
\mathrm{E}_\psi\left[\psi(y_1)\psi(y_2)\right] &= \frac{\mathrm{d}^2}{\mathrm{d}t_1\mathrm{d}t_2}g(t_1,t_2) \qquad \text{(B.5)}\\
&= \frac{\mathrm{d}}{\mathrm{d}t_2}\left[\frac{1}{N}u(t_2-t_1)\theta\big(\min(t_1,t_2)\big)+\left(1-\frac{1}{N}\right)\theta(y_1)\left(\int_{-\infty}^{y_2}\theta(t_2)\mathrm{d}t_2\right)\right]\\
&= \frac{1}{N}\theta(y_1)\delta(y_1-y_2)+\left(1-\frac{1}{N}\right)\theta(y_1)\theta(y_2)\ .
\end{aligned}
$$

B.1.3 Continuous aggregation

The aggregation property has been stated for countable partitions of the process domain – it also holds for continuous aggregations. Define an Empirical process $\psi \in \Psi$ parameterized by N samples and a mean function $\theta \in \mathcal{P}(\mathcal{Y}\times\mathcal{X})$. The process assumes functions from the set $\Psi = \left\{N^{-1}\sum_{n=1}^{N}\delta(\cdot-Y_n)\delta(\cdot-X_n) : Y\in\mathcal{Y}^N, X\in\mathcal{X}^N\right\}$. The aggregation $\psi_\mathrm{m} \equiv \int_\mathcal{Y}\psi(y,\cdot)\mathrm{d}y$ is an Empirical process with N samples and mean function $\theta_\mathrm{m} \equiv \int_\mathcal{Y}\theta(y,\cdot)\mathrm{d}y$; this characterization is inherited from the original Empirical process, as aggregations of ψ_m are equivalent to aggregations of ψ. Note that $\psi_\mathrm{m} \in \Psi_\mathrm{m} = \left\{N^{-1}\sum_{n=1}^{N}\delta(\cdot-X_n) : X\in\mathcal{X}^N\right\}$.

Additionally, when also conditioned on the aggregation ψ_m, the normalized random processes $\psi_\mathrm{c}(x) \equiv \psi(\cdot,x)/\,\psi_\mathrm{m}(x)$ are independent continuous-domain Em-

pirical processes, $\psi_c(x)|\,\psi_m(x), \theta_c(x) \sim \text{EP}\left(\delta(0)^{-1} N\,\psi_m(x), \theta_c(x)\right)$, where $\theta_c(x) \equiv \theta(\cdot, x)/\,\theta_m(x) \in \mathcal{P}(\mathcal{Y})$.

To demonstrate this, use $\mathcal{Y} = \mathcal{X} = \mathbb{R}$ for simplicity. Define the countable partition of $\mathcal{Y} \times \mathcal{X}$, $\{\ldots, \mathcal{S}_i, \ldots\}$, $i \in \mathbb{Z}$, such that $\mathcal{S}_i = \mathbb{R} \times \big[i\Delta, (i+1)\Delta\big)$, where Δ is an arbitrarily small interval in $\mathcal{X}$. The corresponding function partitions are $\psi_i \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_i}$, such that $\psi = (\ldots, \psi_i, \ldots)$, and $\theta_i \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_i}$, such that $\theta = (\ldots, \theta_i, \ldots)$.

Introduce the aggregation process ${\psi_m}'(i) = \int_{\mathcal{S}_i} \psi_i(y, x)\mathrm{d}y\mathrm{d}x = \int_{i\Delta}^{(i+1)\Delta} \psi_m(x)\mathrm{d}x$, which is Empirical with N samples and mean function ${\theta_m}'(i) = \int_{\mathcal{S}_i} \theta_i(y, x)\mathrm{d}y\mathrm{d}x = \int_{i\Delta}^{(i+1)\Delta} \theta_m(x)\mathrm{d}x$. By the properties of Empirical processes, the normalized functions ${\psi_c}'(i) \equiv \psi_i/\,{\psi_m}'(i)$ conditioned on ${\psi_m}'$ are independent Empirical processes of $N\,{\psi_m}'(i)$ samples and mean functions ${\theta_c}'(i) \equiv \theta_i/\,{\theta_m}'(i)$. Next, use the conditional aggregation to define ${\psi_c}''(i) = \int_{i\Delta}^{(i+1)\Delta} {\psi_c}'(\cdot, x; i)\mathrm{d}x \sim \text{EP}\left(N\,{\psi_m}'(i), {\theta_c}''(i)\right)$, where ${\theta_c}''(i) = \int_{i\Delta}^{(i+1)\Delta} {\theta_c}'(\cdot, x; i)\mathrm{d}x$.

As $\Delta \to 0$, the conditional processes tend to ${\psi_c}''(i) \to \psi_c(i\Delta) \sim \text{EP}\left(N\,\psi_m(i\Delta)\Delta, \psi_c(i\Delta)\right)$. Setting $x \equiv i\Delta$ and using $\delta(0) \equiv \Delta^{-1}$, the conditional model characterization given the marginal model is proven. Also, observe that $\left(\delta(0)^{-1}\,\psi_m(x), 1 - \delta(0)^{-1}\,\psi_m(x)\right) \sim \text{Emp}\left(N, \left(\delta(0)^{-1}\,\theta_m(x), 1 - \delta(0)^{-1}\,\theta_m(x)\right)\right)$ and thus that $\delta(0)^{-1} N\,\psi_m(x) \sim \text{Bi}\left(N, \delta(0)^{-1}\,\theta_m(x)\right)$.

B.2 Dirichlet Process Properties

B.2.1 Definition

The Dirichlet process $\theta \sim \text{Dir}(\alpha_0, \alpha)$ assumes distributions from from $\Theta \equiv \mathcal{P}(\mathcal{Y})$ and is parameterized by concentration $\alpha_0 \in \mathbb{R}^+$ and mean function $\alpha \in \left\{\mathbb{R}^{+\mathcal{Y}} : \int_{\mathcal{Y}} \alpha(y)\mathrm{d}y = 1\right\}$. The continuous-domain Dirichlet process is characterized by the same aggregation property as its discrete-domain variant. Define a countable partition of $\mathcal{Y}$, $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$, and the corresponding function partitions $\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\theta = (\ldots, \theta_z, \ldots)$ and $\alpha_z \in {\mathbb{R}^+}^{\mathcal{S}_z}$, such that $\alpha = (\ldots, \alpha_z, \ldots)$. Thus, the transformed random process $\theta_m \in \mathcal{P}(\mathcal{Z})$, defined as $\theta_m(z) \equiv \int_{\mathcal{S}_z} \theta_z(y)\mathrm{d}y$, is Dirichlet with

concentration α_0 and mean function α_{m}, defined as $\alpha_{\mathrm{m}}(z) = \int_{\mathcal{S}_z} \alpha_z(y)\mathrm{d}y$.

Additionally, the normalized random processes $\theta_{\mathrm{c}}(z) \equiv \theta_z / \theta_{\mathrm{m}}(z) \in \mathcal{P}(\mathcal{S}_z)$ are independent continuous-domain Dirichlet processes $\theta_{\mathrm{c}}(z) \sim \mathrm{Dir}\big(\alpha_0\, \alpha_{\mathrm{m}}(z), \alpha_{\mathrm{c}}(z)\big)$, where $\alpha_{\mathrm{c}}(z) \equiv \alpha_z / \alpha_{\mathrm{m}}(z)$, and are independent of the aggregation process θ_{m}.

B.2.2 Mean and Correlation Functions

In this section, it is shown that the expected value of a Dirichlet process $\theta \sim \mathrm{DP}(\alpha_0, \alpha)$ is

$$\mu_\theta = \alpha \; . \tag{B.6}$$

A defining characteristic of Dirichlet processes is that their aggregations are also Dirichlet. Define the partition of $\mathcal{Y} = \mathbb{R}$, $\big\{\, \mathcal{S}(y), S^{\mathrm{c}}(y)\big\}$ where $\mathcal{S}(y) = (-\infty, y]$. The transform random variable $\theta_{\mathrm{m}} \equiv \int_{-\infty}^{y} \theta(t)\mathrm{d}t$ is thus a Beta random variable with parameters $\lambda = \alpha_0 \int_{-\infty}^{y} \alpha(t)\mathrm{d}t$ and $\lambda^{\mathrm{c}} = \alpha_0 \int_{y}^{\infty} \alpha(t)\mathrm{d}t$. Dependency on y is suppressed for brevity. Observe that $\mu_{\theta_{\mathrm{m}}} \equiv \int_{-\infty}^{y} \mu_\theta(t)\mathrm{d}t$ and that using the formula for the expected value of a beta random variable [14],

$$\begin{aligned} \mu_{\theta_{\mathrm{m}}} &= \frac{\lambda}{\lambda + \lambda^{\mathrm{c}}} \\ &= \int_{-\infty}^{y} \alpha(t)\mathrm{d}t = \int_{-\infty}^{y} \mu_\theta(t)\mathrm{d}t \; . \end{aligned} \tag{B.7}$$

Differentiating with respect to y, we have the expected value of the DP.

Next, the correlation function is shown to be

$$\mathrm{E}_\theta\left[\theta(y_1)\theta(y_2)\right] = \frac{\alpha(y_1)\delta(y_1 - y_2) + \alpha_0 \alpha(y_1)\alpha(y_2)}{\alpha_0 + 1} \; . \tag{B.8}$$

First, assume $y_2 \geq y_1$ and define a new partition of $\mathcal{Y}$, $\big\{(-\infty, y_1], (y_1, y_2], (y_2, \infty)\big\}$. By the aggregation property, the random triplet $\Big(\int_{-\infty}^{y_1} \theta(t)\mathrm{d}t, \int_{y_1}^{y_2} \theta(t)\mathrm{d}t, \int_{y_2}^{\infty} \theta(t)\mathrm{d}t\Big)$ is Dirichlet with concentration α_0 and mean $\Big(\int_{-\infty}^{y_1} \alpha(t)\mathrm{d}t, \int_{y_1}^{y_2} \alpha(t)\mathrm{d}t, \int_{y_2}^{\infty} \alpha(t)\mathrm{d}t\Big)$.

Define the function

$$
\begin{aligned}
g(t_1,t_2) &= \mathrm{E}_\theta\left[\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\int_{-\infty}^{y_2}\theta(t_2)\mathrm{d}t_2\right] \qquad \text{(B.9)}\\
&= \mathrm{E}_\theta\left[\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)^2+\left(\int_{-\infty}^{y_1}\theta(t_1)\mathrm{d}t_1\right)\left(\int_{y_1}^{y_2}\theta(t_2)\mathrm{d}t_2\right)\right]\\
&= \frac{\left(\int_{-\infty}^{y_1}\alpha(t_1)\mathrm{d}t_1\right)\left(1+\alpha_0\int_{-\infty}^{y_1}\alpha(t_1)\mathrm{d}t_1\right)+\alpha_0\left(\int_{-\infty}^{y_1}\alpha(t_1)\mathrm{d}t_1\right)\left(\int_{y_1}^{y_2}\alpha(t_2)\mathrm{d}t_2\right)}{\alpha_0+1}\\
&= \frac{\left(\int_{-\infty}^{y_1}\alpha(t_1)\mathrm{d}t_1\right)+\alpha_0\left(\int_{-\infty}^{y_1}\alpha(t_1)\mathrm{d}t_1\right)\left(\int_{-\infty}^{y_2}\alpha(t_2)\mathrm{d}t_2\right)}{\alpha_0+1} \quad \forall\ y_2\geq y_1\ .
\end{aligned}
$$

Following the same steps provides the values of g for $t_2 \leq t_1$; the combined formula can be given as

$$
g(t_1,t_2) = \frac{\left(\int_{-\infty}^{\min(y_1,y_2)}\alpha(t_1)\mathrm{d}t_1\right)+\alpha_0\left(\int_{-\infty}^{y_1}\alpha(t_1)\mathrm{d}t_1\right)\left(\int_{-\infty}^{y_2}\alpha(t_2)\mathrm{d}t_2\right)}{\alpha_0+1}\ . \qquad \text{(B.10)}
$$

Finally,

$$
\begin{aligned}
\mathrm{E}_\theta\left[\theta(y_1)\theta(y_2)\right] &= \frac{\mathrm{d}^2}{\mathrm{d}t_1\mathrm{d}t_2}g(t_1,t_2) \qquad \text{(B.11)}\\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}t_2}\left[u(t_2-t_1)\alpha\big(\min(t_1,t_2)\big)+\alpha_0\alpha(y_1)\left(\int_{-\infty}^{y_2}\alpha(t_2)\mathrm{d}t_2\right)\right]}{\alpha_0+1}\\
&= \frac{\alpha(y_1)\delta(y_1-y_2)+\alpha_0\alpha(y_1)\alpha(y_2)}{\alpha_0+1}\ .
\end{aligned}
$$

B.2.3 Continuous Aggregation

provide full proof here?

The aggregation property has been stated for countable partitions of the process domain – it also holds for continuous aggregations. Define an Dirichlet process $\theta \in \mathcal{P}(\mathcal{Y}\times\mathcal{X})$ parameterized by concentration α_0 and mean function α. Using a procedure similar to that used in Appendix B.1, the aggregation $\theta_{\mathrm{m}} \equiv \int_{\mathcal{Y}}\theta(y,\cdot)\mathrm{d}y \in \mathcal{P}(\mathcal{X})$ is shown to be a Dirichlet process with concentration α_0 and parameterizing function $\alpha_{\mathrm{m}} \equiv \int_{\mathcal{Y}}\alpha(y,\cdot)\mathrm{d}y$.

Additionally, the normalized functions $\theta_{\mathrm{c}}(x) \equiv \theta(\cdot,x)/\,\theta_{\mathrm{m}}(x) \in \mathcal{P}(\mathcal{Y})$ are independent continuous-domain Dirichlet processes, $\theta_{\mathrm{c}}(x) \sim \mathrm{DP}\left(\delta(0)^{-1}\alpha_0\,\alpha_{\mathrm{m}}(x), \alpha_{\mathrm{c}}(x)\right)$,

where $\alpha_{\mathrm{c}}(x) \equiv \alpha(\cdot, x) / \alpha_{\mathrm{m}}(x) \in \mathcal{P}(\mathcal{Y})$, and are independent of the aggregation process θ_{m}.

B.3 Dirichlet-Empirical Process Properties

B.3.1 Definition

This section introduces a new random process, referred to as the Dirichlet-Empirical process (DEP). It is the generalization of the Dirichlet-Empirical distribution for i.i.d. samples drawn from a continuous set $\mathcal{Y}$; that is, it is the expectation of an Empirical process $\psi|\theta \sim \mathrm{EP}(N, \theta)$ with respect to its mean function $\theta \sim \mathrm{DP}(\alpha_0, \alpha)$, a Dirichlet process prior with concentration α_0 and mean function α. The Dirichlet-Empirical process $\psi \sim \mathrm{DEP}(N, \alpha_0, \alpha)$ is parameterized by N samples, concentration α_0, and mean function α; it assumes functions from the set $\Psi = \left\{ N^{-1} \sum_{n=1}^{N} \delta(\cdot - D_n) : D \in \mathcal{Y}^N \right\}$.

Analogous to the Dirichlet and Dirichlet-Empirical distributions for countable spaces, the Dirichlet-Empirical process inherits the aggregation property from the Dirichlet process prior. Consider a Dirichlet-Empirical process $\psi \in \Psi$ and a countable partition of $\mathcal{Y}$, $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$, and the corresponding function partitions $\psi_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\psi = (\ldots, \psi_z, \ldots)$, and $\alpha_z \in {\mathbb{R}^{+}}^{\mathcal{S}_z}$, such that $\alpha = (\ldots, \alpha_z, \ldots)$. The transformed random process ψ_{m}, defined as $\psi_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z} \psi_z(y) \mathrm{d}y$, is necessarily Dirichlet-Empirical for N samples, concentration α_0, and mean function α_{m}, defined as $\alpha_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z} \alpha_z(y) \mathrm{d}y$.

Also, when conditioned on the aggregation ψ_{m}, the normalized functions $\psi_{\mathrm{c}}(z) \equiv \psi_z / \psi_{\mathrm{m}}(z)$ are independent continuous-domain Dirichlet-Empirical processes, $\psi_{\mathrm{c}}(z) | \psi_{\mathrm{m}}(z) \sim \mathrm{DEP}\left(N \psi_{\mathrm{m}}(z), \alpha_0 \alpha_{\mathrm{m}}(z), \alpha_{\mathrm{c}}(z)\right)$, where $\alpha_{\mathrm{c}}(z) \equiv \alpha_z / \alpha_{\mathrm{m}}(z)$.

B.3.2 Mean and Correlation Functions

In this section, it is shown that the expected value of a Dirichlet-Empirical process $\psi \sim \mathrm{DEP}(N, \alpha_0, \alpha)$ is

$$\mu_\psi = \alpha \; . \tag{B.12}$$

A defining characteristic of Dirichlet-Empirical processes is that their aggregations are also Dirichlet-Empirical. Define the partition of $\mathcal{Y} = \mathbb{R}$, $\left\{ \mathcal{S}(y), S^{\mathrm{c}}(y) \right\}$ where $\mathcal{S}(y) = (-\infty, y]$. The transform random process $\left(\psi_{\mathrm{m}}, {\psi_{\mathrm{m}}}^{\mathrm{c}} \right)$, where $\psi_{\mathrm{m}} \equiv \int_\infty^y \psi(t)\mathrm{d}t$, is thus a discrete-domain Dirichlet-Empirical random process for N samples, concentration α_0, and mean $\left(\alpha_{\mathrm{m}}, {\alpha_{\mathrm{m}}}^{\mathrm{c}} \right)$, where $\alpha_{\mathrm{m}} = \int_{-\infty}^y \alpha(t)\mathrm{d}t$ and ${\alpha_{\mathrm{m}}}^{\mathrm{c}} = \int_y^\infty \alpha(t)\mathrm{d}t$. Dependency on y is suppressed for brevity. Observe that $\mu_{\psi_{\mathrm{m}}}(y) = \alpha_{\mathrm{m}}$ and thus that

$$\mu_{\psi_{\mathrm{m}}} = \int_{-\infty}^y \alpha(t)\mathrm{d}t = \int_{-\infty}^y \mu_\psi(t)\mathrm{d}t \; .$$

Differentiating with respect to y, we have the expected value of the DEP.

Next, the correlation function is shown to be

$$\mathrm{E}_\psi \left[\psi(y_1)\psi(y_2) \right] = \frac{(\alpha_0^{-1} + N^{-1})\alpha(y_1)\delta(y_1 - y_2) + (1 - N^{-1})\alpha(y_1)\alpha(y_2)}{1 + \alpha_0^{-1}} \; . \tag{B.13}$$

First, assume $y_2 \geq y_1$ and define a new partition of $\mathcal{Y}$, $\left\{ (-\infty, y_1], (y_1, y_2], (y_2, \infty) \right\}$. By the aggregation property, the random triplet $\left(\int_{-\infty}^{y_1} \psi(t)\mathrm{d}t, \int_{y_1}^{y_2} \psi(t)\mathrm{d}t, \int_{y_2}^{\infty} \psi(t)\mathrm{d}t \right)$ is Dirichlet-Empirical with N samples, concentration α_0, and mean $\left(\int_{-\infty}^{y_1} \alpha(t)\mathrm{d}t, \int_{y_1}^{y_2} \alpha(t)\mathrm{d}t, \int_{y_2}^{\infty} \alpha(t)\mathrm{d}t \right)$.

Define the function

$$\begin{aligned}
g(t_1, t_2) &= \mathrm{E}_\psi \left[\int_{-\infty}^{y_1} \psi(t_1)\mathrm{d}t_1 \int_{-\infty}^{y_2} \psi(t_2)\mathrm{d}t_2 \right] \\
&= \mathrm{E}_\psi \left[\left(\int_{-\infty}^{y_1} \psi(t_1)\mathrm{d}t_1 \right)^2 + \left(\int_{-\infty}^{y_1} \psi(t_1)\mathrm{d}t_1 \right) \left(\int_{y_1}^{y_2} \psi(t_2)\mathrm{d}t_2 \right) \right] \\
&= \frac{\left(\int_{-\infty}^{y_1} \alpha(t_1)\mathrm{d}t_1 \right) \left((\alpha_0^{-1} + N^{-1}) + (1 - N^{-1}) \int_{-\infty}^{y_1} \alpha(t_1)\mathrm{d}t_1 \right) + (1 - N^{-1}) \left(\int_{-\infty}^{y_1} \alpha(t_1)\mathrm{d}t_1 \right) \left(\int_{y_1}^{y_2} \alpha(t_2)\mathrm{d} \right.}{1 + \alpha_0^{-1}} \\
&= \frac{(\alpha_0^{-1} + N^{-1}) \left(\int_{-\infty}^{y_1} \alpha(t_1)\mathrm{d}t_1 \right) + (1 - N^{-1}) \left(\int_{-\infty}^{y_1} \alpha(t_1)\mathrm{d}t_1 \right) \left(\int_{-\infty}^{y_2} \alpha(t_2)\mathrm{d}t_2 \right)}{1 + \alpha_0^{-1}} \quad \forall y_2 \geq y_1 \; .
\end{aligned} \tag{B}$$

Following the same steps provides the values of g for $t_2 \leq t_1$; the combined formula can be given as

$$g(t_1, t_2) \quad = \frac{(\alpha_0^{-1} + N^{-1}) \left(\int_{-\infty}^{\min(y_1, y_2)} \alpha(t_1) \mathrm{d}t_1 \right) + (1 - N^{-1}) \left(\int_{-\infty}^{y_1} \alpha(t_1) \mathrm{d}t_1 \right) \left(\int_{-\infty}^{y_2} \alpha(t_2) \mathrm{d}t_2 \right)}{1 + \alpha_0^{-1}} \quad \forall\ y_2 \geq y$$

Finally,

$$\begin{aligned} \mathrm{E}_\psi \left[\psi(y_1) \psi(y_2) \right] &= \frac{\mathrm{d}^2}{\mathrm{d}t_1 \mathrm{d}t_2} g(t_1, t_2) \qquad \text{(B.15)} \\ &= \frac{\frac{\mathrm{d}}{\mathrm{d}t_2} \left[(\alpha_0^{-1} + N^{-1}) \Big(u(t_2 - t_1) \alpha\big(\min(t_1, t_2) \big) \Big) + (1 - N^{-1})\, \alpha(y_1) \left(\int_{-\infty}^{y_2} \alpha(t_2) \mathrm{d}t_2 \right) \right]}{1 + \alpha_0^{-1}} \\ &= \frac{(\alpha_0^{-1} + N^{-1}) \alpha(y_1) \delta(y_1 - y_2) + (1 - N^{-1}) \alpha(y_1) \alpha(y_2)}{1 + \alpha_0^{-1}} \,. \end{aligned}$$

B.3.3 Continuous aggregation

provide full proof here?

The aggregation property has been stated for countable partitions of the process domain – it also holds for continuous aggregations. Define an Dirichlet-Empirical process $\psi \in \mathcal{P}(\mathcal{Y} \times \mathcal{X})$ parameterized by N samples, concentration α_0, and mean function α. Using a procedure similar to that used in Appendix B.1, the aggregation $\psi_\mathrm{m} \equiv \int_\mathcal{Y} \theta(y, \cdot) \mathrm{d}y \in \mathcal{P}(\mathcal{X})$ is shown to be a Dirichlet-Empirical process with N samples, concentration α_0, and parameterizing function $\alpha_\mathrm{m} \equiv \int_\mathcal{Y} \alpha(y, \cdot) \mathrm{d}y$.

Additionally, when conditioned on the aggregation ψ_m, the normalized functions $\psi_\mathrm{c}(x) \equiv \psi(\cdot, x) / \psi_\mathrm{m}(x) \in \mathcal{P}(\mathcal{Y})$ are independent continuous-domain Dirichlet-Empirical processes, $\psi_\mathrm{c}(x) | \psi_\mathrm{m}(x) \sim \mathrm{DEP}\left(\delta(0)^{-1} N \psi_\mathrm{m}(x), \delta(0)^{-1} \alpha_0 \alpha_\mathrm{m}(x), \alpha_\mathrm{c}(x) \right)$, where $\alpha_\mathrm{c}(x) \equiv \alpha(\cdot, x) / \alpha_\mathrm{m}(x) \in \mathcal{P}(\mathcal{Y})$.

B.4 Training Data representations and distributions

B.4.1 Proof: $\psi \equiv N^{-1}\sum_{n=1}^{N}\delta(\cdot - \mathrm{D}_n)$ given θ is an Empirical Process

It is demonstrated that conditioned on θ, the random process $\psi \equiv \Psi(\mathrm{D}) = N^{-1}\sum_{n=1}^{N}\delta(\cdot - \mathrm{D}_n)$ is an EP, given that $\mathrm{p}_{\mathrm{D}\,|\theta}(D|\theta) = \prod_{n=1}^{N}\theta(D_n)$.

Define the aggregation ψ_{m}, where $\psi_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z}\psi(y)\mathrm{d}y \equiv N^{-1}\sum_{n=1}^{N}\chi(\mathrm{D}_n;\mathcal{S}_z)$ and note that $\mathrm{P}\left(\chi(\mathrm{D}_n;\mathcal{S}_z)=1\middle|\theta\right) \equiv \theta_{\mathrm{m}}(z)$, where $\theta_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z}\theta(y)\mathrm{d}y$. As the events $\mathrm{D}_n \in \mathcal{S}_z$ are independent given θ, ψ_{m} conditioned on the model θ is characterized by an Empirical distribution

$$\mathrm{P}_{\psi_{\mathrm{m}}\,|\theta}(\psi_{\mathrm{m}}\,|\theta) \equiv \mathcal{M}(N\,\psi_{\mathrm{m}})\prod_{z\in\mathcal{Z}}\left(\theta_{\mathrm{m}}(z)^{\psi_{\mathrm{m}}(z)}\right)^{N} = \mathrm{Emp}\left(\,\psi_{\mathrm{m}};N,\theta_{\mathrm{m}}\,\right) \tag{B.16}$$

of N samples with parameters θ_{m}. Since the aggregation property is satisfied, ψ is a Empirical process.

B.4.2 Proof: $\psi \equiv N^{-1}\sum_{n=1}^{N}\delta\big(y - \mathrm{D}_n\,\big)$ is a DEP

Next, it is demonstrated that the random process $\psi \equiv \Psi(\mathrm{D}) = N^{-1}\sum_{n=1}^{N}\delta\big(\,\cdot - \mathrm{D}_n\,\big)$ is a DEP, given that $\mathrm{p}_{\mathrm{D}\,|\theta}(D|\theta) = \prod_{n=1}^{N}\theta(D_n)$ and $\theta \sim \mathrm{DP}(\alpha_0,\alpha)$.

Define the aggregation ψ_{m}, where $\psi_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z}\psi(y)\mathrm{d}y \equiv N^{-1}\sum_{n=1}^{N}\chi(\mathrm{D}_n;\mathcal{S}_z)$, and θ_{m}, where $\theta_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z}\theta(y)\mathrm{d}y$. By the aggregation properties of Empirical and Dirichlet processes, $\psi_{\mathrm{m}}\,|\theta$ and θ are Empirical and Dirichlet, respectively. As such, ψ_{m} is Dirichlet-Empirical for all domain partitions and ψ is a Dirichlet-Empirical process.

B.4.3 Proof: Model Posterior Process is Dirichlet

In this section, it is shown that for i.i.d. data D distributed as $\mathrm{p}_{\mathrm{D}\,|\,\theta} = \bigotimes_{n=1}^{N} \theta$ with parameterizing distribution $\theta \sim \mathrm{DP}(\alpha_0, \alpha)$, then the model conditioned on the training data is also a Dirichlet process with concentration $\alpha_0 + N$ and mean function

$$\mu_{\theta|\,\mathrm{D}} = \left(\frac{\alpha_0}{\alpha_0 + N}\right) \alpha + \left(\frac{N}{\alpha_0 + N}\right) \Psi(\mathrm{D}) \;, \tag{B.17}$$

where $\Psi(\mathrm{D}) = N^{-1} \sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{D}_n\,\big)$.

A defining characteristic of Dirichlet processes is that their aggregations are also Dirichlet. Consider a DP over the set $\mathcal{Y}$. Define an arbitrary countable partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}_z, \ldots\}$, $z \in \mathcal{Z}$ and the corresponding function partitions $\theta_z \in {\mathbb{R}_{\geq 0}}^{\mathcal{S}_z}$, such that $\theta = (\ldots, \theta_z, \ldots)$, and $\alpha_z \in {\mathbb{R}^{+}}^{\mathcal{S}_z}$, such that $\alpha = (\ldots, \alpha_z, \ldots)$. The transformed random process $\theta_{\mathrm{m}} \in \mathcal{P}(\mathcal{Z})$, $\theta_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z} \theta_z(y) \mathrm{d}y$, is necessarily Dirichlet with concentration α_0 and a mean function $\alpha_{\mathrm{m}} \in {\mathbb{R}^{+}}^{\mathcal{Z}}$, $\alpha_{\mathrm{m}}(z) \equiv \int_{\mathcal{S}_z} \alpha_z(y) \mathrm{d}y$.

To prove the hypothesis, it must be shown that

$$\theta_{\mathrm{m}}\,|\,\mathrm{D} \sim \mathrm{Dir}\left(\alpha_0 + N, \mu_{\theta_{\mathrm{m}}\,|\,\mathrm{D}}\right) \;, \tag{B.18}$$

where

$$\mu_{\theta_{\mathrm{m}}\,|\,\mathrm{D}} = \left(\frac{\alpha_0}{\alpha_0 + N}\right) \alpha_{\mathrm{m}} + \left(\frac{N}{\alpha_0 + N}\right) \Psi_{\mathrm{m}}(D) \tag{B.19}$$

and $\Psi_{\mathrm{m}}(z; D) = \int_{\mathcal{S}_z} \Psi(y; D) \mathrm{d}y = N^{-1} \sum_{n=1}^{N} \chi(D_n; \mathcal{S}_z)$.

To demonstrate this property, exploit the results of Appendix A.2 to represent the training data distribution conditioned on the aggregation θ_{m}. Introduce the normalized functions $\theta_{\mathrm{c}}(z) \equiv \theta_z / \,\theta_{\mathrm{m}}(z) \in \mathcal{P}(\mathcal{S}_z)$, which are continuous-domain Dirichlet processes, independent from one another, and independent from the aggregation process θ_{m}.

The conditional distribution of interest is

$$\begin{aligned}
\mathrm{p}_{\mathrm{D}\,|\,\theta_\mathrm{m}}(D|\,\theta_\mathrm{m}) &= \mathrm{E}_{\theta_\mathrm{c}\,|\,\theta_\mathrm{m}}\left[\mathrm{p}_{\mathrm{D}\,|\,\theta_\mathrm{m},\theta_\mathrm{c}}(D|\,\theta_\mathrm{m},\theta_\mathrm{c})\right] \\
&= \mathrm{E}_{\theta_\mathrm{c}\,|\,\theta_\mathrm{m}}\left[\prod_{n=1}^{N}\prod_{z\in\mathcal{Z}}\left(\theta_\mathrm{m}(z)\,\theta_\mathrm{c}(D_n;z)\right)^{\chi(D_n;\mathcal{S}_z)}\right] \\
&= \left(\prod_{z\in\mathcal{Z}}\prod_{n=1}^{N}\theta_\mathrm{m}(z)^{\chi(D_n;\mathcal{S}_z)}\right)\prod_{z\in\mathcal{Z}}\mathrm{E}_{\theta_\mathrm{c}(z)}\left[\prod_{n=1}^{N}\theta_\mathrm{c}(D_n;z)^{\chi(D_n;\mathcal{S}_z)}\right] \\
&= \left(\prod_{z\in\mathcal{Z}}\theta_\mathrm{m}(z)^{\Psi_\mathrm{m}(z;D)}\right)^{N}\prod_{z\in\mathcal{Z}}\mathrm{E}_{\theta_\mathrm{c}(z)}\left[\prod_{n=1}^{N}\theta_\mathrm{c}(D_n;z)^{\chi(D_n;\mathcal{S}_z)}\right].
\end{aligned} \tag{B.20}$$

Observe that the dependency of this likelihood function on θ_m is polynomial. Thus, θ_m is a conjugate prior for D and the training data marginal distribution is

$$\begin{aligned}
\mathrm{p}_\mathrm{D}(D) &= \mathrm{E}_{\theta_\mathrm{m}}\left[\left(\prod_{z\in\mathcal{Z}}\theta_\mathrm{m}(z)^{\Psi_\mathrm{m}(z;D)}\right)^{N}\right]\prod_{z\in\mathcal{Z}}\mathrm{E}_{\theta_\mathrm{c}(z)}\left[\prod_{n=1}^{N}\theta_\mathrm{c}(D_n;z)^{\chi(D_n;\mathcal{S}_z)}\right] \\
&= \frac{\beta\big(\alpha_0\,\alpha_\mathrm{m}+N\,\Psi_\mathrm{m}(D)\big)}{\beta(\alpha_0\,\alpha_\mathrm{m})}\prod_{z\in\mathcal{Z}}\mathrm{E}_{\theta_\mathrm{c}(z)}\left[\prod_{n=1}^{N}\theta_\mathrm{c}(D_n;z)^{\chi(D_n;\mathcal{S}_z)}\right]
\end{aligned} \tag{B.21}$$

and the distribution of interest is

$$\begin{aligned}
\mathrm{p}_{\theta_\mathrm{m}\,|\,\mathrm{D}}(\theta_\mathrm{m}\,|D) &= \frac{\prod_{z\in\mathcal{Z}}\theta_\mathrm{m}(z)^{\alpha_0\,\alpha_\mathrm{m}(z)+N\,\Psi_\mathrm{m}(z;D)-1}}{\beta\big(\alpha_0\,\alpha_\mathrm{m}+N\,\Psi_\mathrm{m}(D)\big)} \\
&= \mathrm{Dir}\left(\theta_\mathrm{m};\alpha_0+N,\mu_{\theta_\mathrm{m}\,|\,\mathrm{D}}\right).
\end{aligned} \tag{B.22}$$

This proves the hypothesis.

B.4.3.1 Prior conjugacy PGR??

FIX? LOCATION?

The likelihood function of the data D given the model θ is

$$\begin{aligned}
\mathrm{p}_{\mathrm{D}\,|\theta}\left(D|\theta\right) &= \prod_{n=1}^{N}\mathrm{p}_{\mathrm{D}_n\,|\theta}\left(D_n|\theta\right) = \prod_{n=1}^{N}\theta(D_n) \\
&= \exp\left(\sum_{n=1}^{N}\ln\left(\theta(D_n)\right)\right) \\
&= \exp\left(\iint_{\mathcal{Y}\times\mathcal{X}} N\Psi(y,x;D)\ln\left(\theta(y,x)\right)\mathrm{d}y\mathrm{d}x\right) \\
&\equiv \prod_{\mathcal{Y}\times\mathcal{X}}\left(\theta(y,x)^{N\Psi(y,x;D)}\right)^{\mathrm{d}y\mathrm{d}x},
\end{aligned} \tag{B.23}$$

a function only dependent on the data through the empirical statistic $\Psi(\mathrm{D})$. Also, as shown in Appendix B.1, the random process $\psi \equiv \Psi(\mathrm{D})$ given θ is an Empirical process.

As a result, $\theta|\{\mathrm{D} = D\} \sim \theta|\{\psi = \Psi(D)\}$. This is a natural generalization of the results for the discrete-domain model process. In general, when an Empirical process $\psi|\theta \sim \mathrm{EP}(N, \theta)$ has a mean function which is characterized by a Dirichlet process $\theta \sim \mathrm{DP}(\alpha_0, \alpha)$, the posterior $\theta|\psi \sim \mathrm{DP}(\alpha_0 + N, \mu_{\theta|\psi})$, where

$$\mu_{\theta|\psi} = \left(\frac{\alpha_0}{\alpha_0 + N}\right)\alpha + \left(\frac{N}{\alpha_0 + N}\right)\psi \ . \tag{B.24}$$

Additionally, note that $\theta_{\mathrm{m}}|\{\mathrm{D} = D\} \sim \theta_{\mathrm{m}}|\{\psi_{\mathrm{m}} = \Psi_{\mathrm{m}}(D)\}$. The model aggregation process is only dependent on the aggregation of the empirical distribution. These properties result from the independence of θ_{m} from θ_{c}, a property that Dirichlet processes hold.

Appendix C

Bayesian generalized linear regression

introduce and use weighted inner product notation? basis is tuple of functionals?

comment on low-dim and redefinition of theta

Swap transposes??

For generalized linear regression, the space of data-generating models $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ considered is restricted to a finite-dimensional space $\theta \in \Theta = \mathbb{R}^K$. The observed data distribution $\mathrm{p}_{\mathrm{x}\,|\theta} = \mathrm{p}_{\mathrm{x}}$ is fixed. Note that the space of the data probability distributions considered is a strict subset of the entire function space. The conditional mean has the form $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta} = \phi(\mathrm{x})^\intercal\theta$, where $\phi \in \{\mathcal{Y}^{\mathcal{X}}\}^K$ is a vector of basis functions; the notation $\phi(x) \equiv [\phi_1(x), \ldots, \phi_K(x)]^\intercal$ is used for brevity. Additionally, the conditional variance is fixed and independent of x, such that $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \Sigma_{\mathrm{y}}$.

Note that the clairvoyant estimator (2.41) is $f_\Theta(\mathrm{x};\theta) = \phi(\mathrm{x})^\intercal\theta$ and the irreducible squared error (2.42) is $\mathcal{R}^*_\Theta(\theta) = \Sigma_{\mathrm{y}}$, which is independent of the true weights. To determine the optimal Bayesian estimator (2.43), note that the weights are conditionally independent of the novel observation x given the data D, resulting in

$f^*(\mathrm{x};\mathrm{D}) = \phi(\mathrm{x})^\intercal \mu_{\theta|\,\mathrm{D}}$. Plugging into (2.45), the minimum Bayesian squared error is

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{C}_{\theta|\,\mathrm{x},\mathrm{D}}\left[f_\Theta(\mathrm{x};\theta)\right]\right] \\ &= \Sigma_\mathrm{y} + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\phi(\mathrm{x})^\intercal \Sigma_{\theta|\,\mathrm{D}}\phi(\mathrm{x})\right] \\ &= \Sigma_\mathrm{y} + \mathrm{E}_\mathrm{x}\left[\phi(\mathrm{x})^\intercal\, \mathrm{E}_\mathrm{D}\left[\Sigma_{\theta|\,\mathrm{D}}\right]\phi(\mathrm{x})\right], \end{aligned} \tag{C.1}$$

noting that since x is independent of θ, it is independent of the data D as well.

C.1 Normal distribution assumptions

joint sufficient statistics?

Commonly, the true predictive model is assumed to be Normal, such that $\mathrm{y}\,|\,\mathrm{x},\theta \sim \mathcal{N}\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}, \Sigma_\mathrm{y}\right)$. Additionally, the weight prior is assumed to be Normal, such that $\mathrm{p}_\theta = \mathcal{N}\left(\mu_\theta, \Sigma_\theta\right)$. Using linear algebra, it is can be shown [19] that $\theta|\,\mathrm{D} \sim \mathcal{N}\left(\mu_{\theta|\,\mathrm{D}}, \Sigma_{\theta|\,\mathrm{D}}\right)$, where

$$\Sigma_{\theta|\,\mathrm{D}} \equiv \left(\Sigma_\theta^{-1} + \sum_{n=1}^{N} \phi(\mathrm{X}_n)\Sigma_\mathrm{y}^{-1}\phi(\mathrm{X}_n)^\intercal\right)^{-1} \tag{C.2}$$

and

$$\mu_{\theta|\,\mathrm{D}} \equiv \Sigma_{\theta|\,\mathrm{D}}\left(\Sigma_\theta^{-1}\mu_\theta + \sum_{n=1}^{N}\phi(\mathrm{X}_n)\Sigma_\mathrm{y}^{-1}\,\mathrm{Y}_n\right). \tag{C.3}$$

Note that the posterior mean of θ is a convex combination of the prior mean μ_θ and the maximum-likelihood estimate $\theta_{\mathrm{ML}}(\mathrm{D}) = \left(\sum_{n=1}^{N}\phi(\mathrm{X}_n)\Sigma_\mathrm{y}^{-1}\phi(\mathrm{X}_n)^\intercal\right)^{-1}\sum_{n=1}^{N}\phi(\mathrm{X}_n)\Sigma_\mathrm{y}^{-1}\,\mathrm{Y}_n$, which is simply weighted least-squares.

discuss trends

It can be further shown that the Bayesian predictive distribution is also Normal, with $\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \phi(\mathrm{x})^\intercal\mu_{\theta|\,\mathrm{D}}$ and $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \Sigma_\mathrm{y} + \phi(\mathrm{x})^\intercal\Sigma_{\theta|\,\mathrm{D}}\phi(\mathrm{x})$.

Bibliography

[1] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Second. Springer, 1980.

[2] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2000.

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

[4] George E.P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

[5] Richard A. Brualdi. *Introductory Combinatorics*. Fifth. Pearson, 2010.

[6] William Feller. *An Introduction to Probability Theory and Its Applications*. Second. Vol. 2. Probability and Mathematical Statistics. New York, New York: John Wiley & Sons, 1971.

[7] Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems". In: *The Annals of Statistics* 1.2 (1973), pp. 209–230.

[8] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics*. Second. Reading, Massachusetts: Addison-Wesley, 1994.

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

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

[11] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Vol. 1. Signal Processing Series. Upper Saddle River, New Jersey: Prentice-Hall, 1993.

[12] Thomas P. Minka. *Bayesian inference, entropy, and the multinomial distribution*. Tech. rep. Microsoft Research, 2003.

[13] Kevin P. Murphy. *Binomial and multinomial distributions*. Tech. rep. University of British Columbia, 2006.

[14] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. Fourth. McGraw-Hill, 2002.

[15] C. Radhakrishna Rao. "Maximum Likelihood Estimation for the Multinomial Distribution". In: *Sankhyā: The Indian Journal of Statistics (1933-1960)* 18.1/2 (1957), pp. 139–148.

[16] Steven Roman. "The Logarithmic Binomial Formula". In: *American Mathematical Monthly* 99.7 (1992).

[17] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* USA: McGraw-Hill, Inc., 1987. ISBN: 0070542341.

[18] Frederick F. Stephan. "The Expected Value and Variance of the Reciprocal and other Negative Powers of a Positive Bernoullian Variate". In: *The Annals of Mathematical Statistics* 16.1 (1945), pp. 50–61.

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

[20] J. G. Wendel. "Note on the Gamma Function". In: *The American Mathematical Monthly* 55.9 (1948), pp. 563–564.