Bayesian Learning using a Dirichlet Prior for Regression and Classification

Paul Rademacher

January 17, 2020

Contents

Todo list

use todonotes package instead of my initials?

equation numbers to final line!

brackets for expectation ops?

line break symbol format, before/after?

learners (full) or decision functions (range)?

likelihood function terminology

Dirichlet localization or concentration?

fix arguments for theta, nbar dists.

suppress arguments where sensible? eg Pd = theta

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

is Θ redundant given $\mathcal{P}$?

DIM and PR operator from AMS?

multinomial random process?

introduce tilde alpha to match tilde theta?

use x,y subs instead of prime/tilde?

change from nbar to emprical RV???

investigate N lim for nbar given theta, bayes risk vs model support

SEMI-SUPERVISED - generalize to joint decisions!!! training/test!

priors = sparse conditionals; w/ sufficient statistics

NFLT investigation? try matlab examples

use clairvoyant/irreducible terms and symbols

empirical risk terms/discussion?

check notation throughout for expectation/variance operators, proper arguments

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

jeffrey prior, fisher info?

PGR: 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 stastical 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 underperform 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

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 $\mathcal{Y}^{\mathcal{X}}$.

Various mappings will be defined for which the domain and/or the range [16] 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 different groups of arguments. A set of finite sequences $\{1, \ldots, N\} \mapsto \mathcal{S}$ will be represented as $\mathcal{S}^N$ for brevity.

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?

Random elements, variables, and processes

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) [13], denoted $\mathrm{P}_{\mathrm{x}} \in \mathcal{P}(\mathcal{X})$. 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 $\mathrm{p}_{\mathrm{x}} \in \mathcal{P}(\mathcal{X})$.

explicit PMF/PDF formula with P of events?

For notational simplicity, probability distributions are occasionally represented as $\mathrm{P}(\mathrm{x})$; in such instances, the formal notation can be recovered by replacing the distribution with $\mathrm{P}_{\mathrm{x}}(x)$ and all instances of the roman symbol x with the italic symbol x.

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. These distributions may be compactly represented as $\mathrm{P}(\mathrm{x}\,|\,\mathrm{z})$.

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?

Remove subscript suppression convention?

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 Dirichlet distribution with parameters $\alpha \in \mathcal{A}$; the PDF will be notated as $\mathrm{Dir} : \mathcal{A} \mapsto \mathcal{P}(\mathcal{X})$, where the range is the set of valid PDF's. More compactly, the notation $\mathrm{x} \sim \mathrm{Dir}(\alpha)$ implies that $\mathrm{P}_{\mathrm{x}} = \mathrm{Dir}(\cdot; \alpha)$. Other distribution functions repeatedly used include Multi, DM, DP, and DMP, representing the multinomial distribution, the Dirichlet-multinomial distribution, the Dirichlet process, and the Dirichlet-Multinomial process.

Expectation Operators

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

$$\mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right] = \sum_{x} \mathrm{P}_{\mathrm{x}}(x) g(x) \;, \tag{1.1}$$

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

$$\mathrm{C}_{\mathrm{x}}\left[g(\mathrm{x})\right] = \mathrm{E}_{\mathrm{x}}\left[\Big(g(\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right]\Big)^2\right] . \tag{1.2}$$

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 condtional distribution as well. In this case, the expectation operator is a function of the observed value of z, such that

$$\mathrm{E}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](z) = \sum_{x} \mathrm{P}_{\mathrm{x}\,|\,\mathrm{z}}(x|z) g(x) \;. \tag{1.3}$$

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 $\Sigma_{\mathrm{x}\,|\,\mathrm{z}}$ operators.

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

$$\mathrm{E}_{\mathrm{x}}\left[g(\mathrm{x})\right](y) = \sum_{x} \mathrm{P}_{\mathrm{x}}(x) g(y; x) \; . \tag{1.4}$$

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

$$\begin{aligned} &\mathrm{C}_{\mathrm{x}}\left[g(\mathrm{x})\right](y, y') \\ &\quad = \mathrm{E}_{\mathrm{x}}\left[\Big(g(y;\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(y;\mathrm{x})\right]\Big)\Big(g(y';\mathrm{x}) - \mathrm{E}_{\mathrm{x}}\left[g(y';\mathrm{x})\right]\Big)\right] \; . \end{aligned} \tag{1.5}$$

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 $\mu_{\mathrm{x}}(y)$ 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 extends 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.

Again, it is common for the conditional random element z to be left as a random quantity instead of being explicitly defined. In such cases, the italic z is dropped from the arguments and the formulae $\mathrm{E}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](y)$, $\mathrm{C}_{\mathrm{x}\,|\,\mathrm{z}}\left[g(\mathrm{x})\right](y, y')$, $\mu_{\mathrm{x}\,|\,\mathrm{z}}(y)$, and $\Sigma_{\mathrm{x}\,|\,\mathrm{z}}(y, y')$ imply dependence on z.

ABOVE NOTATION CREATES AMBIGUITY!!!!!!! Check for uses...

As for probability distributions, the subscript notation of these operators may be suppressed. In such cases, the expectations are to be performed with respect to the joint distribution of all random elements (roman font) found in the argument. For example, $\mathrm{E}\left[f(\mathrm{y}, \mathrm{x})|\,\mathrm{z}\right]$ compactly represents $\mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\mathrm{z}}\left[f(\mathrm{y}, \mathrm{x})\right]$.

Special Functions

Certain specialized functions are detailed next. Both the Dirac and Kronecker delta functions will be used throughout. 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

$$\int_{\mathcal{X}} \delta(x)\mathrm{d}x = 1 \ . \tag{1.6}$$

Consequently, it also satisfies

$$\int_{\mathcal{X}} g(x)\delta(x)\mathrm{d}x = g(0) \ . \tag{1.7}$$

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

$$\delta[x, x'] = \begin{cases} 1 & \text{if } x = x', \\ 0 & \text{if } x \neq x'. \end{cases} \tag{1.8}$$

PGR: 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

$$\mathcal{M}(g) = \frac{\left(\sum_{x\in\mathcal{X}} g(x)\right)!}{\prod_{x\in\mathcal{X}} g(x)!} \ . \tag{1.9}$$

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

$$\beta(g) = \frac{\prod_{x\in\mathcal{X}} \Gamma\big(g(x)\big)}{\Gamma\left(\sum_{x\in\mathcal{X}} g(x)\right)} \ . \tag{1.10}$$

Note that the countable domains of the input functions may have an infinite number of elements. These functions will also be used to operate on a subset of a functions' domain $\mathcal{S} \subset \mathcal{X}$ and its corresponding image. Set notation for a function is used to express the argument, so that $\mathcal{M}\Big(\big\{g(x) : x \in S\big\}\Big)$ and $\beta\Big(\big\{g(x) : x \in S\big\}\Big)$.

define subset functions, ditch set inputs? appendix?

Chapter 2

Problem Statement

2.1 Data Model and Objective

italic theta font before Bayes?

Consider an observable random element $\mathrm{x} \in \mathcal{X}$ and and unobservable random element $\mathrm{y} \in \mathcal{Y}$ which are jointly distributed according to an unknown probability distribution $\theta \in \Theta = \mathcal{P}(\mathcal{Y} \times \mathcal{X})$, 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.

PGR: consider definition/equivalence of D and Y,X. Think indexing.

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 $\mathrm{D} \equiv (\mathrm{Y}, \mathrm{X})$. The N data pairs are conditionally independent from one another and are identically distributed as $\mathrm{P}_{\mathrm{D}_n\,|\theta} = \theta$. The samples are also conditionally independent from (y, x). Thus,

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

The task in supervised machine learning is to design a decision 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 learning functions are non-parametric and there are no restrictions on the set of achievable functions $\mathcal{F}$.

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 conditional expected loss, or conditional "risk",

$$\begin{aligned} \mathcal{R}_\Theta(f;\theta) &= \mathrm{E}_{\mathrm{y,x,D}\,|\theta}\Big[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\Big] \\ &= \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\Big[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\Big]\right] \\ &= \mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\Big[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\Big]\right]\right]. \end{aligned} \tag{2.2}$$

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 is unknown and the designer can never be precisely sure how well any decision function performs.

2.1.1 Clairvoyant Decision

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

PGR: use marginal/conditional thetas?

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

$$f_\Theta(\theta) = \arg\min_{f \in \mathcal{F}} \mathcal{R}_\Theta(f;\theta)\,. \tag{2.3}$$

For a given set of observations x and D, the function $f_\Theta(\theta) \in \mathcal{F}$ selects the decision $h = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h,\mathrm{y})\right]$. Note the conditional independence of y from D in (2.2) - the knowledge of θ renders the training data D useless. As such, the range of the clairvoyant function is recast as $f_\Theta : \Theta \mapsto \mathcal{H}^{\mathcal{X}}$ and the decisions are

$$f_\Theta(\mathrm{x};\theta) = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h,\mathrm{y})\right]\,. \tag{2.4}$$

The corresponding clairvoyant risk for a given model θ is

$$\begin{aligned}\mathcal{R}^*_\Theta(\theta) &\equiv \mathcal{R}_\Theta\left(f_\Theta(\theta);\theta\right) \\ &= \min_{f\in\mathcal{F}} \mathcal{R}_\Theta(f;\theta) \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h,\mathrm{y})\right]\right] .\end{aligned} \tag{2.5}$$

2.1.2 Bayes Decision

To design an optimal decision function $f \in \mathcal{F}$, an operator must be chosen to remove the dependency of the conditional risk 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

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta\left[\,\mathcal{R}_\Theta(f;\theta)\right] \\ &= \mathrm{E}_{\mathrm{y,x,D}}\left[\,\mathcal{L}(f(\mathrm{x};\mathrm{D}),\mathrm{y})\right] \\ &= \mathrm{E}_{\mathrm{x,D}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\,\right)\right]\right] \\ &= \mathrm{E}_{\mathrm{D}}\left[\,\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\,\right)\right]\right]\right]\end{aligned} \tag{2.6}$$

and y, x, and D are treated as jointly distributed random elements. Observe that the Bayesian predictive distributions can be represented as $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{D}}\left[\,\mathrm{P}_{\mathrm{x}\,|\theta}\,\right]$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} = \mathrm{E}_{\theta|\,\mathrm{x,D}}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\,\right]$, the expected values of the corresponding clairvoyant distributions with respect to the model posteriors $\mathrm{p}_{\theta|\,\mathrm{D}}$ and $\mathrm{p}_{\theta|\,\mathrm{x,D}}$, respectively.

PGR: BELOW, add forumla for f(D)?

Finally, express the optimal learning function

$$f^* = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f) \; . \tag{2.7}$$

The decision expressed by the learning function f^* given observed values of x and D is

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] \\ &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h,\mathrm{y})\right]\right]. \end{aligned} \tag{2.8}$$

Thus, the Bayesian approach uses the model posterior $\mathrm{p}_{\theta|\,\mathrm{x},\mathrm{D}}$ to integrate out the dependency on the model given the observable random elements. The minimum Bayes risk is

$$\begin{aligned} \mathcal{R}^* &\equiv \mathcal{R}(f^*) \\ &= \min_{f\in\mathcal{F}} \mathcal{R}(f) \\ &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right]\right] \\ &= \mathrm{E}_{\mathrm{D}}\left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right]\right]\right]. \end{aligned} \tag{2.9}$$

2.1.2.1 Irreducible Risk

PGR: RECONSIDER IRREDUCIBLE TERMINOLOGY!!!!!!

PGR: bound met in limit of N? full support? bounded prior? change/move to clairvoyant discussion?

The clairvoyant risk (2.5) 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 $\mathrm{E}_\theta\left[\,\mathcal{R}^*_\Theta(\theta)\right] \leq \mathcal{R}(f) \quad \forall f \in \mathcal{F}$; the expected value of the clairvoyant risk will thus be referred to as the "irreducible" risk.

It is important to note that this inequality holds for any number of training samples N and that the irreducible risk does not depend on N. Thus, even with unlimited training data, no learning function can provide a Bayes risk lower than this value.

2.2 Sufficient Statistic: the Empirical PMF

PGR: MOVE SECTION before Bayesian approach??

PGR: continuous? DMP?

PGR: change to emp PMF RP ψ?

PGR: trends in limit of N, use mean/cov?? EMPIRICAL RISK, bounded prior

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

$$\begin{aligned} \mathrm{P}_{\mathrm{D}\,|\,\theta}\left(D|\theta\right) &= \prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\,\theta}\left(D_n|\theta\right) \\ &= \prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{N}(y,x;D)} \ , \end{aligned} \tag{2.10}$$

where the dependency on the training data D is expressed though a transform function $\bar{N} : \mathcal{D} \mapsto \bar{\mathcal{N}}$, where the range is

$$\bar{\mathcal{N}} = \left\{ \bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \bar{n}(y,x) = N \right\} \tag{2.11}$$

and the function is defined as

$$\begin{aligned} \bar{N}(y,x;D) &= \sum_{n=1}^{N} \delta\big[(y,x), D_n\big] \\ &= \sum_{n=1}^{N} \delta\left[y, Y_n\right] \delta\left[x, X_n\right] \ . \end{aligned} \tag{2.12}$$

This function counts the number of occurrences of the pair (y,x) in the training set D.

PGR: show SS via Kay, data likelihood??!!

Note that $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ depends on the training data D only through the transform $\bar{N}$; $\bar{N}(\mathrm{D})$ is thus a sufficient statistic [2] for the model θ. Consequently, other distributions of interest P_{D}, $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ will also depend on D via $\bar{N}(\mathrm{D})$. As such, it is useful to define a new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$.

Frequently, the corresponding distributions $\mathrm{P}_{\bar{\mathrm{n}}}$, $\mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ will be used to find the optimal decision functions and the minimum risk. Note that $\mathcal{M}\left(\bar{N}(D)\right)\mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta) = \mathrm{P}_{\bar{\mathrm{n}}\,|\,\theta}\left(\bar{N}(D)|\theta\right)$, where $\mathcal{M}$ is the multinomial operator. Also note that $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}(D) = \mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}\left(\bar{N}(D)\right)$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(x,D) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left(x, \bar{N}(D)\right)$.

The cardinality of the random process' domain is $|\bar{\mathcal{N}}| = \mathcal{M}\left(\{N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\}\right)$; 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 $|\bar{\mathcal{N}}| \leq |\,\mathcal{D}\,|$ and the sufficient statistic compactly represents the valuable information in the training data. Also, observe that the set $\{\bar{n}/N : \bar{n} \in \bar{\mathcal{N}}\} \subset \Theta$ and thus that the empirical distribution $\bar{N}(\mathrm{D})/N$ assumes one of a finite number of the elements from Θ.

PGR: sufficient statistic savings in memory bits???

Conditioned on the model θ, the PMF of $\bar{\mathrm{n}}$ is a multinomial distribution

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}\,|\theta}(\bar{n}|\theta) &= \sum_{D:\bar{N}(D)=\bar{n}} \mathrm{P}_{\mathrm{D}\,|\theta}(D|\theta) \\ &= \big|\{D : \bar{N}(D) = \bar{n}\}\big| \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{n}(y,x)} \\ &= \mathcal{M}(\bar{n}) \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{n}(y,x)} \\ &= \mathrm{Multi}\left(\bar{n}; N, \theta\right) , \end{aligned} \tag{2.13}$$

where the multinomial operator $\mathcal{M}$ is used.

The first and second joint moments of this multinomial distribution are [18]

$$\mu_{\bar{\mathrm{n}}|\theta} = N\theta \tag{2.14}$$

and

$$\begin{aligned} &\mathrm{E}_{\bar{\mathrm{n}}|\theta}\left[\bar{\mathrm{n}}(y,x)\bar{\mathrm{n}}(y',x')\right] \\ &\quad = N\big(\theta(y,x)\delta[y,y']\delta[x,x'] + (N-1)\theta(y,x)\theta(y',x')\big) \end{aligned} \tag{2.15}$$

and the covariance function is

$$\Sigma_{\bar{\mathrm{n}}\,|\theta}(y,x,y',x') = N\big(\theta(y,x)\delta[y,y']\delta[x,x'] - \theta(y,x)\theta(y',x')\big) . \tag{2.16}$$

PGR: trends with N? delta? Figures??

Also, observe that the maximum likelihood estimate of θ given the training statistic is [14],

$$\begin{aligned} \theta_{\mathrm{ML}}\big(\bar{n}\big) &= \arg\max_{\theta\in\Theta} \mathrm{P}_{\bar{\mathrm{n}}\,|\theta}(\bar{n}|\theta) \\ &= \frac{\bar{n}}{N} , \end{aligned} \tag{2.17}$$

the empirical distribution.

2.3 Marginal/Conditional Distributions for Observed/Unobserv Elements

PGR: move section?

PGR: clean up

2.3.1 Marginal and Conditional Distributions of θ

PGR: are they independent??? No?!

PGR: bijection!!

As only y is unobservable, it will be useful to decompose the model distribution as $\theta \equiv (\theta', \tilde{\theta})$. First, introduce the marginal distribution $\theta' \equiv \sum_{y \in \mathcal{Y}} \theta(y, \cdot) \in \mathcal{P}(\mathcal{X})$; note that the summation is replaced by an integral when y is a continuous random variable. Next, introduce the conditional distributions $\tilde{\theta} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$ defined as $\tilde{\theta}(x) \equiv \theta(\cdot, x)/\theta'(x)$.

This decomposition enables the clairvoyant distributions to be represented as $\mathrm{P}_{\mathrm{x}\,|\theta} = \mathrm{P}_{\mathrm{x}\,|\theta'} = \theta'$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\tilde{\theta}} = \tilde{\theta}(\mathrm{x})$; these distributions will be of recurring importance.

PGR: conditional theta condition on x necessary above? Yes?!

PGR: marginal theta conditional on X, not full D above???? No!?

2.3.2 Marginal and Conditional Distributions of Training Data

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

$$\begin{aligned} \mathrm{P}_{\mathrm{X}\,|\theta}\left(X|\theta\right) &= \prod_{n=1}^{N} \mathrm{P}_{\mathrm{X}_n\,|\theta}\left(X_n|\theta\right) \\ &= \prod_{x \in \mathcal{X}} \theta'(x)^{N'(x;X)} \,, \end{aligned} \tag{2.18}$$

where the dependency on θ is only through the marginal model θ'. Additionally, note that the dependency on the training observations X is expressed though a "marginal" counting function $N' : \mathcal{X} \mapsto \mathcal{N}'$ with range

$$\mathcal{N}' = \left\{ n' \in {\mathbb{Z}_{\geq 0}}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} n'(x) = N \right\} \,, \tag{2.19}$$

defined as

$$N'(X) = \sum_{n=1}^{N} \delta\big[\cdot, X_n\big] \equiv \sum_{y \in \mathcal{Y}} \bar{N}(y, \cdot; D) \,. \tag{2.20}$$

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

$$\begin{aligned} \mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\theta}\left(Y|X,\theta\right) &= \prod_{n=1}^{N} \frac{\mathrm{P}_{\mathrm{Y}_n,\mathrm{X}_n\,|\,\theta}\left(Y_n,X_n|\theta\right)}{\mathrm{P}_{\mathrm{X}_n\,|\,\theta}\left(X_n|\theta\right)} \\ &= \prod_{x\in\mathcal{X}} \left[\prod_{y\in\mathcal{Y}} \tilde{\theta}(y;x)^{\bar{N}(y,x;Y,X)}\right] , \end{aligned} \tag{2.21}$$

which is dependent on the model θ only through the conditional models $\tilde{\theta}(x)$.

As before, the dependency on the training data can be simplified using a sufficient statistic. Introduce the "marginalized" random process n' over the set $\mathcal{X}$, defined as $\mathrm{n}' \equiv \sum_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,\cdot) \equiv N'(\mathrm{X}) \in \mathcal{N}'$. By the aggregation property of Multinomial random processes [9], the aggregation conditioned on the model θ is distributed as $\mathrm{n}'\,|\theta \sim \mathrm{Multi}(N,\theta')$.

Also of interest is the distribution of $\bar{\mathrm{n}}$ conditioned on its aggregation n'. Using the multinomial distribution properties proven in Appendix A.2, it can be shown that when conditioned on the model θ as well, the PMF of $\bar{\mathrm{n}}$ is

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}',\theta}(\bar{n}|n',\theta) &= \prod_{x\in\mathcal{X}} \left[\mathcal{M}\left(\bar{n}(\cdot,x)\right) \prod_{y\in\mathcal{Y}} \tilde{\theta}(y;x)^{\bar{n}(y,x)}\right] \\ &= \prod_{x\in\mathcal{X}} \mathrm{Multi}\left(\bar{n}(\cdot,x); n'(x), \tilde{\theta}(x)\right) , \end{aligned} \tag{2.22}$$

over the domain $\left\{\bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}} \bar{n}(y,\cdot) = n'\right\}$. Observe that conditioning on the aggregation renders the function segments $\bar{\mathrm{n}}(\cdot,x)$ independent of one another and that they are also Multinomial, such that $\bar{\mathrm{n}}(\cdot,x)|\,\mathrm{n}'(x),\theta \sim \mathrm{Multi}\left(\mathrm{n}'(x),\tilde{\theta}(x)\right)$. Furthermore, the dependency on θ is expressed through the conditional model $\tilde{\theta}$.

2.3.3 Model Posteriors

PGR: clean, discuss

The Bayesian predicitive distributions analagous to the clairvoyant distributions can now be simplified as $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} = \mu_{\theta'\,|\,\mathrm{D}}$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mu_{\tilde{\theta}(\mathrm{x})\,|\,\mathrm{x},\mathrm{D}}$.

PGR: use $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\tilde{\theta}|\,\mathrm{x},\mathrm{D}}\left[\tilde{\theta}(\mathrm{x})\right]$ instead???

$$\begin{aligned} \mathrm{p}_{\theta',\tilde{\theta}|\,\mathrm{Y,X}}(\theta',\tilde{\theta}|Y,X) &= \frac{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}(Y|X,\tilde{\theta})}{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X)} \frac{\mathrm{P}_{\mathrm{X}\,|\theta'}(X|\theta')}{\mathrm{P}_{\mathrm{X}}(X)}\, \mathrm{p}_{\theta',\tilde{\theta}}(\theta',\tilde{\theta}) \\ &= \mathrm{p}_{\tilde{\theta}|\,\mathrm{Y,X}}(\tilde{\theta}|Y,X)\, \mathrm{p}_{\theta'|\,\mathrm{X}}(\theta'|X) \frac{\mathrm{p}_{\tilde{\theta}|\theta'}(\tilde{\theta}|\theta')}{\mathrm{p}_{\tilde{\theta}|\,\mathrm{X}}(\tilde{\theta}|X)} \end{aligned} \tag{2.23}$$

$$\begin{aligned} \mathrm{p}_{\theta',\tilde{\theta}|\,\bar{\mathrm{n}}}(\theta',\tilde{\theta}|\bar{n}) &= \frac{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}\left(\bar{n}|\sum_y \bar{n}(y,\cdot),\tilde{\theta}\right)}{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}\left(\bar{n}|\sum_y \bar{n}(y,\cdot)\right)} \frac{\mathrm{P}_{\mathrm{n}'\,|\theta'}\left(\sum_y \bar{n}(y,\cdot)|\theta'\right)}{\mathrm{P}_{\mathrm{n}'}\left(\sum_y \bar{n}(y,\cdot)\right)}\, \mathrm{p}_{\theta',\tilde{\theta}}(\theta',\tilde{\theta}) \\ &= \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}},\mathrm{n}'}\left(\tilde{\theta}|\bar{n},\sum_y \bar{n}(y,\cdot)\right) \mathrm{p}_{\theta'|\,\mathrm{n}'}\left(\theta'|\sum_y \bar{n}(y,\cdot)\right) \frac{\mathrm{p}_{\tilde{\theta}|\theta'}(\tilde{\theta}|\theta')}{\mathrm{p}_{\tilde{\theta}|\,\mathrm{n}'}\left(\tilde{\theta}|\sum_y \bar{n}(y,\cdot)\right)} \end{aligned} \tag{2.24}$$

Note that $\mathrm{p}_{\tilde{\theta}|\,\mathrm{X}} = \mathrm{E}_{\theta'|\,\mathrm{X}}\left[\mathrm{p}_{\tilde{\theta}|\theta'}\right]$ and $\mathrm{p}_{\tilde{\theta}|\,\mathrm{n}'} = \mathrm{E}_{\theta'|\,\mathrm{n}'}\left[\mathrm{p}_{\tilde{\theta}|\theta'}\right]$. Also, note that $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}} = \mathrm{E}_{\theta'|\,\mathrm{X}}\left[\mathrm{E}_{\tilde{\theta}|\theta'}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}\right]\right]$ and $\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'} = \mathrm{E}_{\theta'|\,\mathrm{n}'}\left[\mathrm{E}_{\tilde{\theta}|\theta'}\left[\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}\right]\right]$.

$$\begin{aligned} \mathrm{p}_{\theta',\tilde{\theta}|\,\mathrm{Y,X,x}}(\theta',\tilde{\theta}|Y,X,x) &= \frac{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}(Y|X,\tilde{\theta})}{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X,x}}(Y|X,x)} \frac{\mathrm{P}_{\mathrm{X,x}\,|\theta'}(X,x|\theta')}{\mathrm{P}_{\mathrm{X,x}}(X,x)}\, \mathrm{p}_{\theta',\tilde{\theta}}(\theta',\tilde{\theta}) \\ &= \mathrm{p}_{\tilde{\theta}|\,\mathrm{Y,X}}(\tilde{\theta}|Y,X)\, \mathrm{p}_{\theta'|\,\mathrm{X,x}}(\theta'|X,x) \frac{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X)}{\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X,x}}(Y|X,x)} \frac{\mathrm{p}_{\tilde{\theta}|\theta'}(\tilde{\theta}|\theta')}{\mathrm{p}_{\tilde{\theta}|\,\mathrm{X}}(\tilde{\theta}|X)} \end{aligned} \tag{2.25}$$

$$\begin{aligned} \mathrm{p}_{\theta',\tilde{\theta}|\,\bar{\mathrm{n}},\mathrm{x}}(\theta',\tilde{\theta}|\bar{n},x) &= \frac{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}\left(\bar{n}|\sum_y \bar{n}(y,\cdot),\tilde{\theta}\right)}{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\mathrm{x}}\left(\bar{n}|\sum_y \bar{n}(y,\cdot),x\right)} \frac{\mathrm{P}_{\mathrm{n}',\mathrm{x}\,|\theta'}\left(\sum_y \bar{n}(y,\cdot),x|\theta'\right)}{\mathrm{P}_{\mathrm{n}',\mathrm{x}}\left(\sum_y \bar{n}(y,\cdot),x\right)}\, \mathrm{p}_{\theta',\tilde{\theta}}(\theta',\tilde{\theta}) \\ &= \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}},\mathrm{n}'}\left(\tilde{\theta}|\bar{n},\sum_y \bar{n}(y,\cdot)\right) \mathrm{p}_{\theta'|\,\mathrm{n}',\mathrm{x}}\left(\theta'|\sum_y \bar{n}(y,\cdot),x\right) \\ &\quad \frac{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}\left(\bar{n}|\sum_y \bar{n}(y,\cdot)\right)}{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\mathrm{x}}\left(\bar{n}|\sum_y \bar{n}(y,\cdot),x\right)} \frac{\mathrm{p}_{\tilde{\theta}|\theta'}(\tilde{\theta}|\theta')}{\mathrm{p}_{\tilde{\theta}|\,\mathrm{n}'}\left(\tilde{\theta}|\sum_y \bar{n}(y,\cdot)\right)} \end{aligned} \tag{2.26}$$

Note that $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X,x}} = \mathrm{E}_{\theta'|\,\mathrm{X,x}}\left[\mathrm{E}_{\tilde{\theta}|\theta'}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}\right]\right]$ and also that $\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\mathrm{x}} = \mathrm{E}_{\theta'|\,\mathrm{n}',\mathrm{x}}\left[\mathrm{E}_{\tilde{\theta}|\theta'}\left[\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}\right]\right]$.

2.4 Applications to Common Loss Functions

PGR: equation label conflicts??

PGR: marginal/condtitional???

PGR: sum vs int???

PGR: move before nbar, keep D?

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 conditional risk (2.2) is assessed, clairvoyant learners (2.4) are found, and the clairvoyant risk (2.5) is expressed.

2.4.1 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 learning function's estimate is allowed to assume real numbers; thus, $\mathcal{H} = \mathbb{R} \supseteq \mathcal{Y}$.

The loss function is defined as

$$\mathcal{L}(h, y) = (h - y)^2 \,. \tag{2.27}$$

Substituting the squared-error loss into (2.2), the conditional squared-error risk is

$$\begin{aligned} \mathcal{R}_\Theta(f; \theta) &= \mathrm{E}_{\mathrm{D}\,|\theta} \left[\mathrm{E}_{\mathrm{y,x}\,|\theta} \left[\big(f(\mathrm{x}; \mathrm{D}) - \mathrm{y} \big)^2 \right] \right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta} \left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta} \left[\mathrm{E}_{\mathrm{D}\,|\theta} \left[\big(f(\mathrm{x}; \mathrm{D}) - \mathrm{y} \big)^2 \right] \right] \right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta} \left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta} \left[(\mathrm{y} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta})^2 \right] \right] + \mathrm{E}_{\mathrm{x,D}\,|\theta} \left[\big(f(\mathrm{x}; \mathrm{D}) - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta} \big)^2 \right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta} \left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta} \right] + \mathrm{E}_{\mathrm{x,D}\,|\theta} \left[\big(f(\mathrm{x}; \mathrm{D}) - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta} \big)^2 \right] , \end{aligned} \tag{2.28}$$

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 Bayesian estimate and the true mean $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$.

2.4.1.1 Clairvoyant Estimation

To find the clairvoyant estimator, the squared-error loss is substituted into (2.4); 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 observed value x, such that

$$\begin{aligned} f_\Theta(\mathrm{x};\theta) &= \arg\min_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(h-\mathrm{y})^2\right] \\ &= \mu_{\mathrm{y}\,|\,\mathrm{x},\theta} = \sum_{y\in\mathcal{Y}} y\tilde{\theta}(y;\mathrm{x}) \ . \end{aligned} \tag{2.29}$$

Substituting the loss and clairvoyant function into (2.5), the resulting clairvoyant risk is

$$\begin{aligned} \mathcal{R}^*_\Theta(\theta) &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(\mathrm{y}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})^2\right]\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \ . \end{aligned} \tag{2.30}$$

Observe that the general conditional risk (2.28) 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 clairvoyant squared-error; the second term is dependent on the difference between the general estimate and the clairvoyant estimate.

Figure 2.1 displays the clairvoyant risk for predictive models $\tilde{\theta}(x)$ independent of x.

2.4.1.2 Bayesian Estimation

Optimal Estimate: the Posterior Mean PGR: plots?

To find the optimal estimator, the squared-error loss is substituted into (2.8). 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

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[(h-\mathrm{y})^2\right] \\ &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \ . \end{aligned} \tag{2.31}$$

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 (2.8), the optimal Bayes estimator is the condtional expected value of the clairvoyant estimate with respect to the model posterior distribution.

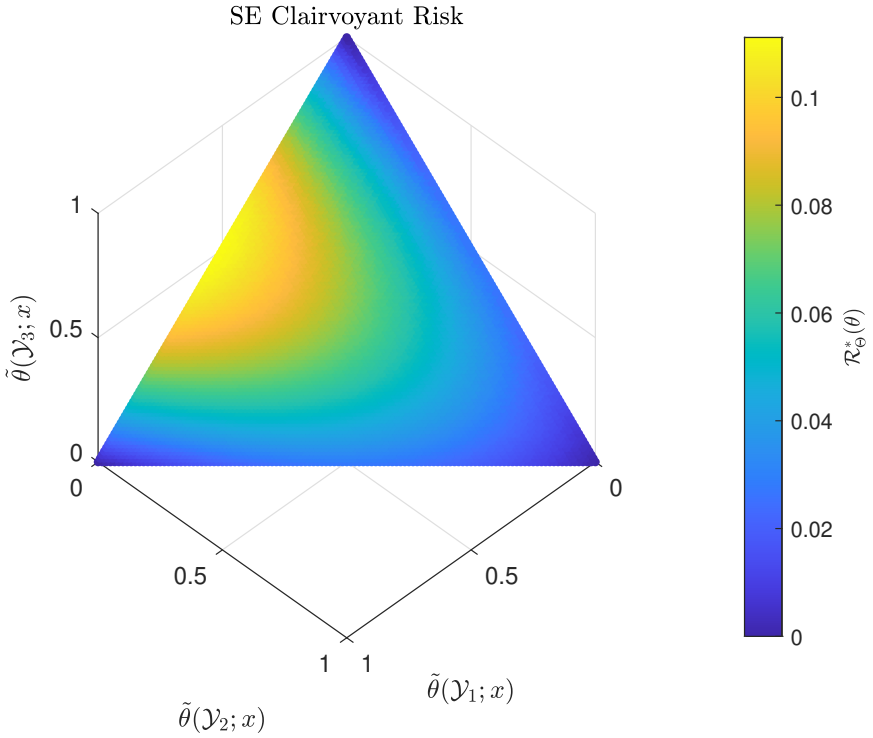

Figure 2.1: Clairvoyant Risk for the Squared-Error Loss Function, constant $\tilde{\theta}(x)$

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

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{y,x}\,|\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\big)^2\right]\right]\right] \\ &= \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\big)^2\right]\right] \\ &= \mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] + \mathrm{E}_{\mathrm{x,D},\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\right] \\ &= \mathrm{E}_{\mathrm{x,D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x,D}}\right] + \mathrm{E}_{\mathrm{x,D}}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x,D}}\big)^2\right].\end{aligned} \tag{2.32}$$

Substituting the optimal estimator (2.31) into Equation (2.32), the minimum Bayes risk is the expected conditional variance

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_{\mathrm{x,D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \\ &= \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{C}_{\theta|\,\mathrm{x,D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\right] \\ &= \mathrm{E}_\theta\left[\mathcal{R}^*_\Theta(\theta)\right] + \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{C}_{\theta|\,\mathrm{x,D}}\left[f_\Theta(\mathrm{x};\theta)\right]\right].\end{aligned} \tag{2.33}$$

The first term is the 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 $\mathrm{p}_{\theta|\,\mathrm{x,D}}$

PGR: assess irreducible risk by performing theta agg conditioning???

2.4.2 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

$$\mathcal{L}(h, y) = 1 - \delta[h, y] \; . \tag{2.34}$$

Applying the 0-1 loss, the conditional risk (2.2) for a general classifier is

$$\begin{aligned} \mathcal{R}_\Theta(f;\theta) &= 1 - \mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{y,x}\,|\theta}\left[\delta\big[f(\mathrm{x};\mathrm{D}),\mathrm{y}\,\big]\right]\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \mathrm{E}_{\mathrm{D}\,|\theta}\left[\theta\big(f(x;\mathrm{D}),x\big)\right] \\ &= 1 - \mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left(f(\mathrm{x};\mathrm{D})|\,\mathrm{x},\theta\right)\right]\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \theta'(x)\,\mathrm{E}_{\mathrm{D}\,|\theta}\left[\tilde{\theta}\big(f(x;\mathrm{D});x\big)\right] \; . \end{aligned} \tag{2.35}$$

2.4.2.1 Clairvoyant Hypothesis

To find the clairvoyant classifier, the 0-1 loss is substituted into (2.4); given an observation x, the optimum hypothesis is simply the value y that maximizes the conditional model $\tilde{\theta}(x)$,

$$\begin{aligned} f_\Theta(\mathrm{x};\theta) &= \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[1 - \delta[h,y]\right] \\ &= \arg\max_{h\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(h|\,\mathrm{x},\theta) \\ &= \arg\max_{y\in\mathcal{Y}} \tilde{\theta}(y;\mathrm{x}) \\ &= \arg\max_{y\in\mathcal{Y}} \theta(y,\mathrm{x}) \; . \end{aligned} \tag{2.36}$$

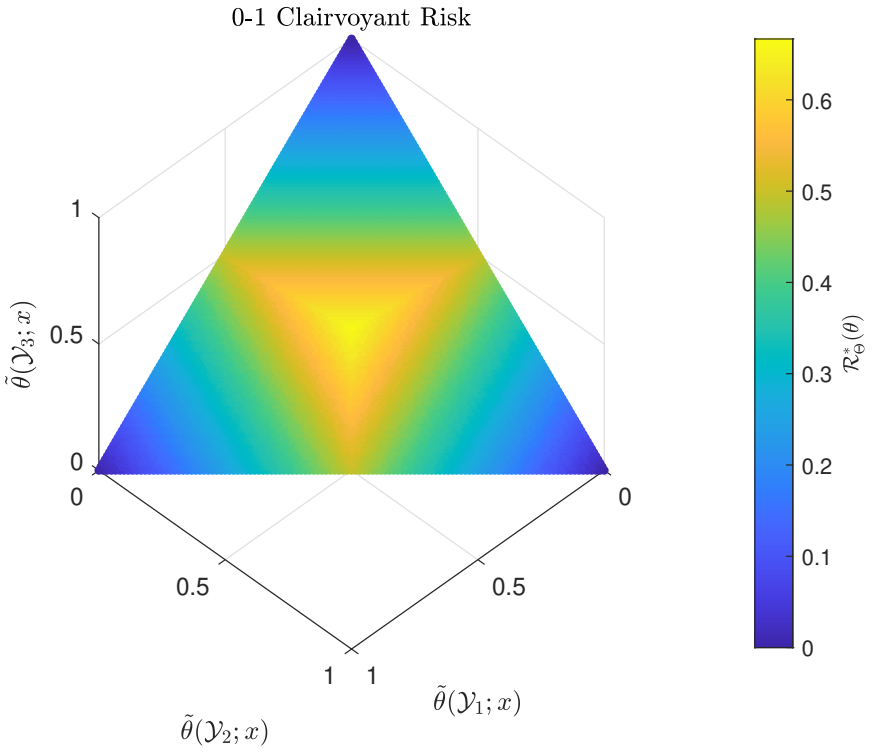

Figure 2.2: Clairvoyant Risk for the 0–1 Loss Function, constant $\tilde{\theta}(x)$

Substituting the 0-1 loss and clairvoyant hypothesis into (2.5), the resulting clairvoyant risk is

$$\begin{aligned} \mathcal{R}^*_\Theta(\theta) &= 1 - \mathrm{E}_{\mathrm{x}\,|\theta}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(y|\,\mathrm{x},\theta)\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \theta'(x) \max_{y\in\mathcal{Y}} \tilde{\theta}(y;x) \\ &= 1 - \sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \theta(y,x)\ . \end{aligned} \tag{2.37}$$

Figure 2.2 displays the clairvoyant risk for predictive models $\tilde{\theta}(x)$ independent of x. Intuitively, the models that are more concentrated lead to lower probability of error.

2.4.2.2 Bayesian Classification

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

To determine the optimal learning function, the 0-1 loss from Equation (2.34) is substi-

tuted into Equation (2.8) to find

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[1-\delta[h,\mathrm{y}]\right] \\ &= \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\ . \end{aligned} \tag{2.38}$$

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

Minimum Risk: Probability of Error Using the 0-1 loss, the Bayes probability of error (2.9) is

$$\mathcal{R}(f) = 1 - \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left(f(\mathrm{x};\mathrm{D})|\,\mathrm{x},\mathrm{D}\right)\right]\ . \tag{2.39}$$

Substituting the optimal learner (2.38) into the general risk (2.39), the minimum probability of error is

$$\mathcal{R}^* = 1 - \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\right]\ . \tag{2.40}$$

Chapter 3

Finite Dirichlet Model

This chapter demonstrates the optimal decision 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 decision function, the joint PMF $\mathrm{P}_{\mathrm{y,x,D}}$ 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 decision function.

Other distributions of interest will be provided, such as the training data PMF P_{D} and the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ used to form a decision given specific observations.

3.1.1 Model PDF, p_{θ}

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

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

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

The parameter α controls around which models θ the PDF concentrates and how strongly. For convenience, introduce the concentration parameter $\alpha_0 \equiv \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \alpha(y,x)$.

The first and second joint moments of the model are

$$\mu_\theta = \frac{\alpha}{\alpha_0} \tag{3.2}$$

and

$$\mathrm{E}_\theta\left[\theta(y,x)\theta(y',x')\right] = \frac{\alpha(y,x)\alpha(y',x') + \alpha(y,x)\delta[y,y']\delta[x,x']}{\alpha_0(\alpha_0+1)} \,. \tag{3.3}$$

Observe that $\mathrm{P}_{\mathrm{y,x}} = \mu_\theta = \alpha/\alpha_0$. The covariance is

$$\begin{aligned}\Sigma_\theta(y,x,y',x') &= \mathrm{E}_\theta\left[\big(\theta(y,x) - \mu_\theta(y,x)\big)\big(\theta(y',x') - \mu_\theta(y',x')\big)\right] \\ &= \frac{\mu_\theta(y,x)\delta[y,y']\delta[x,x'] - \mu_\theta(y,x)\mu_\theta(y',x')}{\alpha_0+1} \,.\end{aligned} \tag{3.4}$$

Also, for $\alpha(y,x) > 1$, the maximizing value of the distribution is

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

This can be easily shown by maximizing the logarithm of the distribution using the method of Lagrange multipliers, as demonstrated in B.1.

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

$$\mathrm{p}_\theta(\theta) \to \delta\left(\theta - \frac{\alpha}{\alpha_0}\right) \,. \tag{3.6}$$

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

$$\mathrm{p}_\theta(\theta) \to \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0} \delta\big(\theta - \delta[\cdot,y]\delta[\cdot,x]\big) \,, \tag{3.7}$$

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 $\int_\Theta \delta(\theta)\mathrm{d}\theta = 1$.

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

These trends are demonstrated with Figure 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 $\alpha_0 = 2.99$, $\alpha < 1$ and 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 parameterizing function is $\alpha(y, x) = 1$, the distribution becomes a uniform PDF and is represented as

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

Note that the concentration parameter is $\alpha_0 = |\mathcal{Y}||\mathcal{X}|$ and $\mathrm{P}_{\mathrm{y,x}} = \big(|\mathcal{Y}||\mathcal{X}|\big)^{-1}$ is also uniform. Figure 3.2 shows the uniform distribution amplitude for $|\mathcal{Y}| = 3$ and $|\mathcal{X}| = 1$.

3.1.1.1 Marginal and Conditional Distributions

PGR: move/add Dir figs here?

The marginal distribution Θ' and the conditional distribution $\tilde{\Theta}$ will also be of interest. By the aggregation property [7], Θ' is a Dirichlet random process parameterized by $\alpha' : \mathcal{X} \mapsto \mathbb{R}^+$, where $\alpha' \equiv \sum_{y \in \mathcal{Y}} \alpha(y, \cdot)$. Note that $\mathrm{P_x} = \mu_{\Theta'} = \alpha'/\alpha_0$.

PGR: introduce tilde alpha to match tilde theta?

Also of interest is the distribution of the predictive model $\tilde{\Theta}$ conditioned on the marginal Θ'. As demonstrated in Appendix A.1, these random processes are jointly distributed as

$$\begin{aligned} \mathrm{p}_{\tilde{\Theta}|\Theta'}\left(\tilde{\theta}|\theta'\right) &= \prod_{x \in \mathcal{X}} \left[\beta\big(\alpha(\cdot, x)\big)^{-1} \prod_{y \in \mathcal{Y}} \tilde{\theta}(y; x)^{\alpha(y,x)-1} \right] \\ &= \prod_{x \in \mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x); \alpha(\cdot, x)\right) , \end{aligned} \tag{3.9}$$

a product of Dirichlet distributions defined on $\tilde{\theta} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$. As shown, the processes $\tilde{\Theta}(x)$ are Dirichlet with parameterizing functions $\alpha(\cdot, x)$, independent of one another, and independent of the marginal distribution Θ'. Observe that the values $\alpha'(x)$ represent the concentration parameters for the individual Dirichlet processes; also, note that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \mu_{\tilde{\Theta}(\mathrm{x})} = \alpha(\cdot, \mathrm{x})/\alpha'(\mathrm{x})$.

PGR: use conditional independence property to simplify throughout?!? In loss app sections?!?!

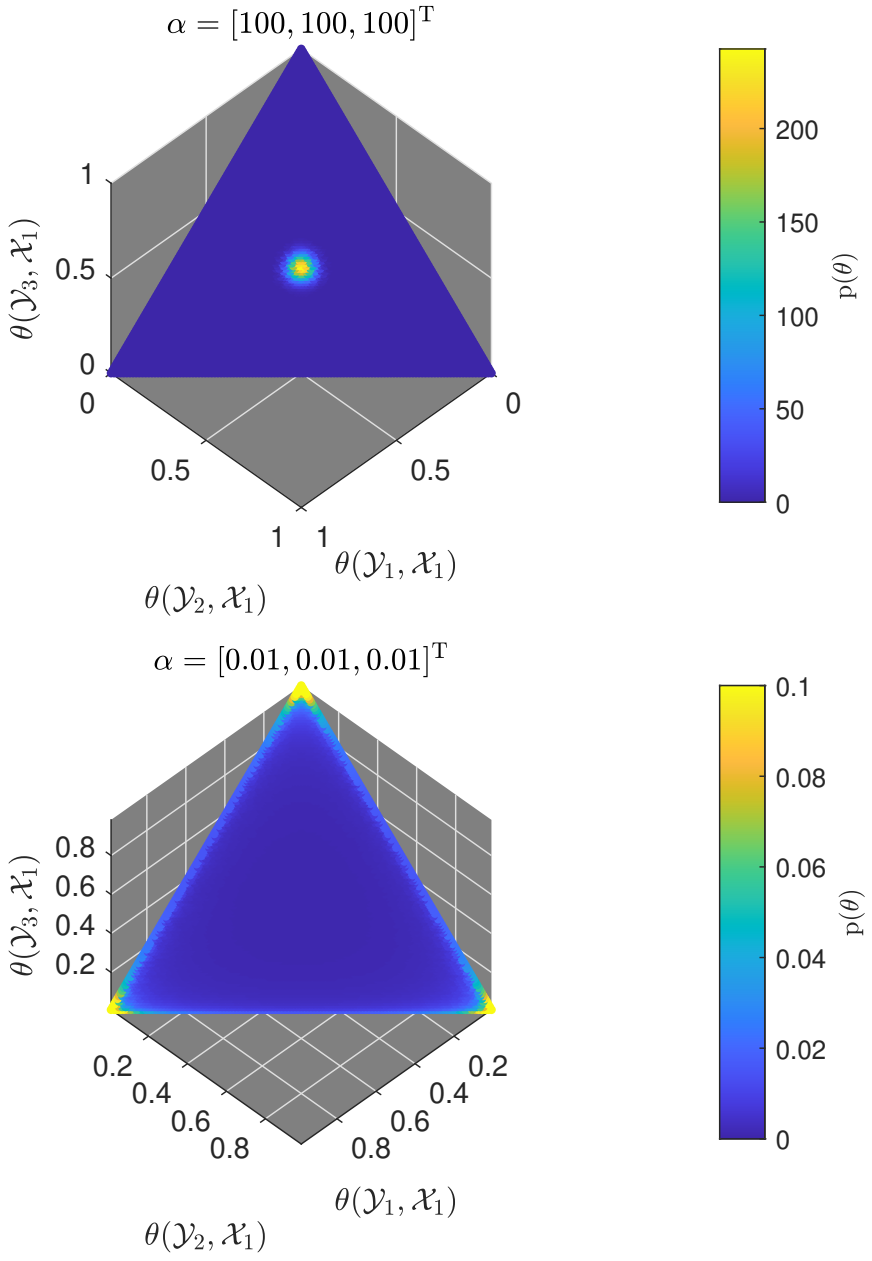

Figure 3.1: Model prior PDF for different concentrations α_0

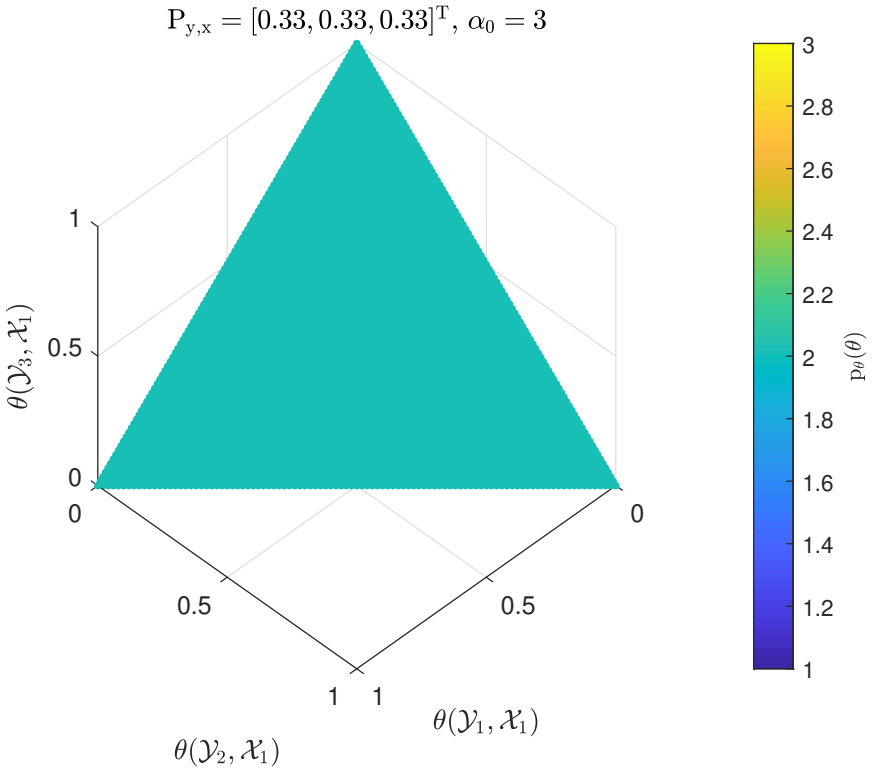

Figure 3.2: Uniform model prior PDF, $|\,\mathcal{Y}\,| = 3$, $|\,\mathcal{X}\,| = 1$

PGR: implications of independence for posterior learning?

3.1.2 Training Set PMF, P_{D}

PGR: EVIDENCE TERMINOLOGY??

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

As the conditional distribution $\mathrm{P}_{\mathrm{D}\,|\boldsymbol{\theta}}$ is of exponential form, it can be readily shown that the marginal distribution of the training data is [11]

$$\begin{aligned}\mathrm{P}_{\mathrm{D}}(D) &= \mathrm{E}_{\boldsymbol{\theta}}\left[\prod_{n=1}^{N}\mathrm{P}_{\mathrm{D}_n\,|\boldsymbol{\theta}}\left(D_n|\boldsymbol{\theta}\right)\right] \\ &= \mathrm{E}_{\boldsymbol{\theta}}\left[\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\boldsymbol{\theta}(y,x)^{\bar{N}(y,x;D)}\right] \\ &= \beta(\boldsymbol{\alpha})^{-1}\beta\left(\boldsymbol{\alpha}+\bar{N}(D)\right)\ .\end{aligned} \tag{3.10}$$

Note that values of the PMF P_{D} are equivalent to joint moments of the model $\boldsymbol{\theta}$.

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 concetrates at its mean and the training data D distribution is

$$\begin{aligned}\mathrm{P}_{\mathrm{D}}(D) &\to \mathrm{E}_{\boldsymbol{\theta}}\left[\prod_{n=1}^{N}\boldsymbol{\theta}(Y_n,X_n)\right] \\ &= \prod_{n=1}^{N}\frac{\boldsymbol{\alpha}\big(Y_n,X_n\big)}{\alpha_0}\ .\end{aligned} \tag{3.11}$$

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

$$\mathrm{P}_{\mathrm{D}}(D) \to \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\frac{\boldsymbol{\alpha}(y,x)}{\alpha_0}\prod_{n=1}^{N}\delta\big[D_n,(y,x)\big] \tag{3.12}$$

and the training data are identical.

Next, the distribution of the sufficient statistic $\bar{\mathrm{n}}$ will be represented. As a Dirichlet distribution characterizes the parameters of the multinomial distribution $\mathrm{P}_{\mathrm{D}\,|\boldsymbol{\theta}}$, the marginal PMF of $\bar{\mathrm{n}}$ is a Dirichlet-Multinomial distribution [9] parameterized by $\boldsymbol{\alpha}$,

$$\begin{aligned}\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) &= \mathcal{M}(\bar{n})\beta(\boldsymbol{\alpha})^{-1}\beta(\boldsymbol{\alpha}+\bar{n}) \\ &= \mathrm{DM}\left(\bar{n};N,\boldsymbol{\alpha}\right)\ .\end{aligned} \tag{3.13}$$

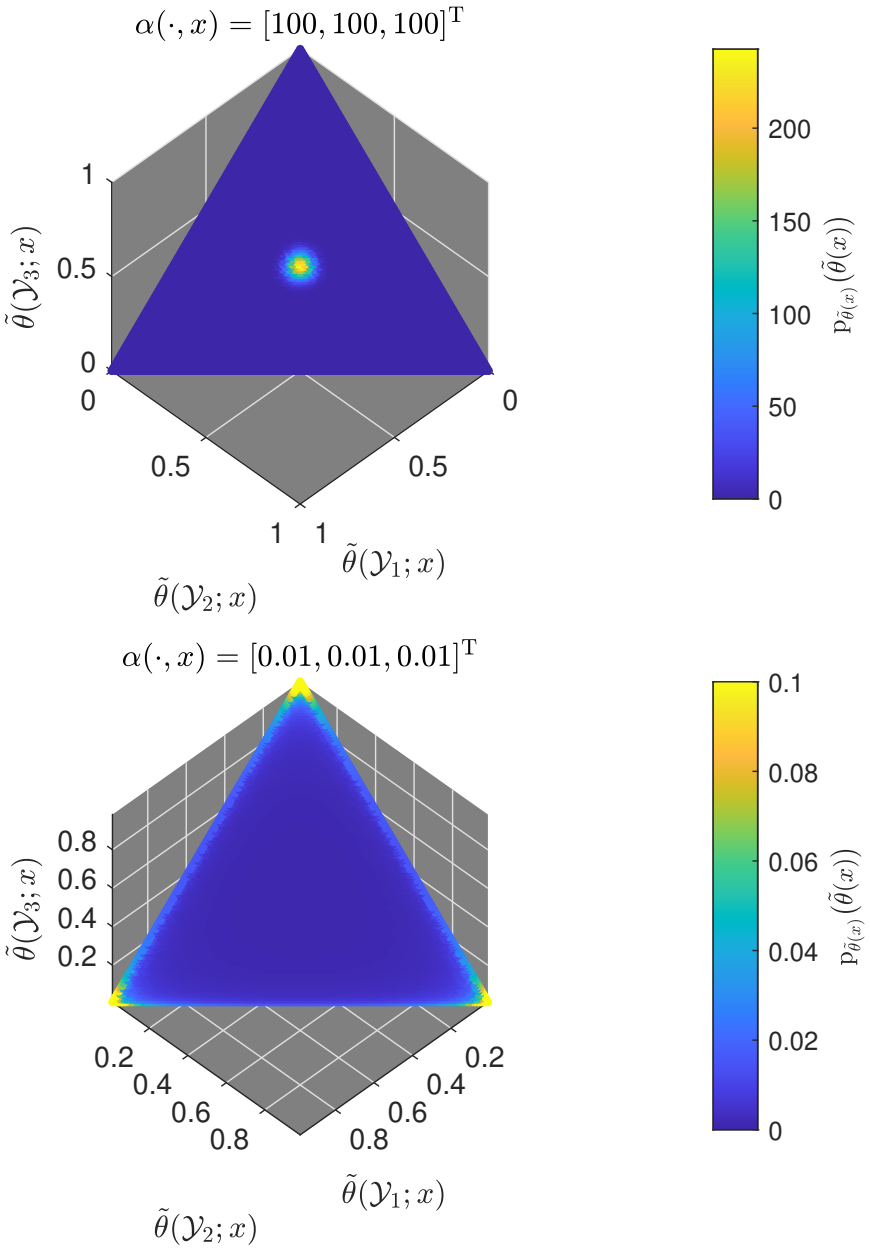

Figure 3.3: Model prior PDF for different concentrations $\alpha'(x)$

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

$$\mu_{\bar{\mathrm{n}}} = N\frac{\alpha}{\alpha_0} = N\mu_\theta \tag{3.14}$$

and

$$\begin{aligned}
&\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y,x)\bar{\mathrm{n}}(y',x')\right] \\
&\quad = \frac{N}{\alpha_0(\alpha_0+1)}\Big((\alpha_0+N)\alpha(y,x)\delta[y,y']\delta[x,x'] + (N-1)\alpha(y,x)\alpha(y',x')\Big) \\
&\quad = \frac{N}{\alpha_0+1}\Big((\alpha_0+N)\mu_\theta(y,x)\delta[y,y']\delta[x,x'] + \alpha_0(N-1)\mu_\theta(y,x)\mu_\theta(y',x')\Big)\ .
\end{aligned} \tag{3.15}$$

The covariance function is

$$\begin{aligned}
\Sigma_{\bar{\mathrm{n}}}(y,x,y',x') &= \frac{N(\alpha_0+N)}{\alpha_0+1}\big(\mu_\theta(y,x)\delta[y,y']\delta[x,x'] - \mu_\theta(y,x)\mu_\theta(y',x')\big) \\
&= N(\alpha_0+N)\Sigma_\theta(y,x,y',x')\ .
\end{aligned} \tag{3.16}$$

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 and thus $\bar{\mathrm{n}}$ is characterized by a multinomial distribution,

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \to \mathcal{M}(\bar{n}) \prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}} \left(\frac{\alpha(y,x)}{\alpha_0}\right)^{\bar{n}(y,x)} \tag{3.17}$$

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

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \to \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0}\delta\big[\bar{n}, N\delta[\cdot,y]\delta[\cdot,x]\big]\ . \tag{3.18}$$

PGR: formal proofs for limiting PMFs? stirling/gautschi?

Figure 3.4 displays the distribution of $\bar{\mathrm{n}}$ for $N = 10$ and different model concentrations α_0. Observe that for large α_0, the distribution approaches a multinomial distribution $\bar{\mathrm{n}} \sim \mathrm{Multi}(N, \alpha/\alpha_0)$. Figure 3.5 shows how a specific model prior influences the data PMF differently for different N. Observe that as the number of training samples increases, the PMF $\mathrm{P}_{\bar{\mathrm{n}}}$ tends toward $\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \approx N^{1-|\mathcal{Y}||\mathcal{X}|}\,\mathrm{p}_\theta(\bar{n}/N)$; this can be proven using Gautschi's inequality [19].

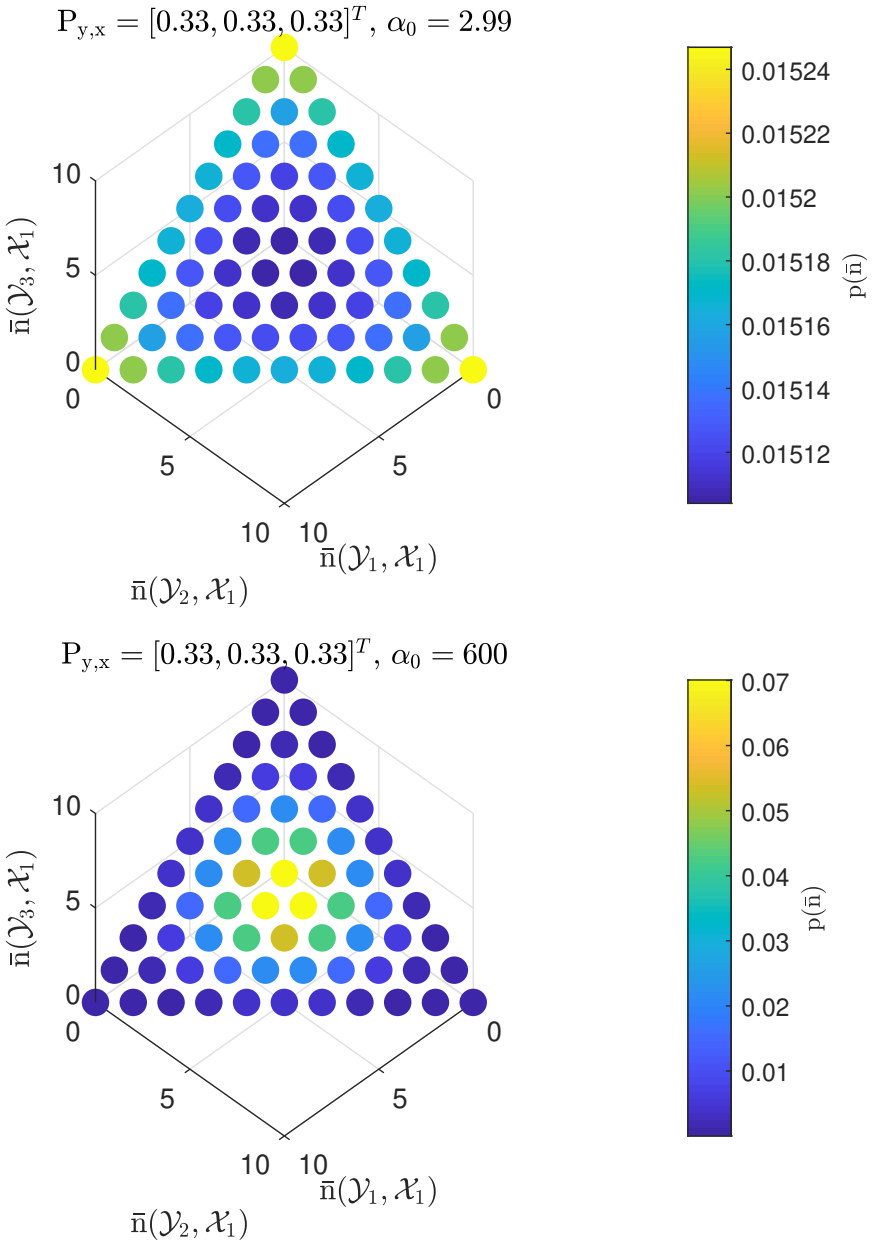

Figure 3.4: $P(\bar{n})$ for different prior concentrations α_0

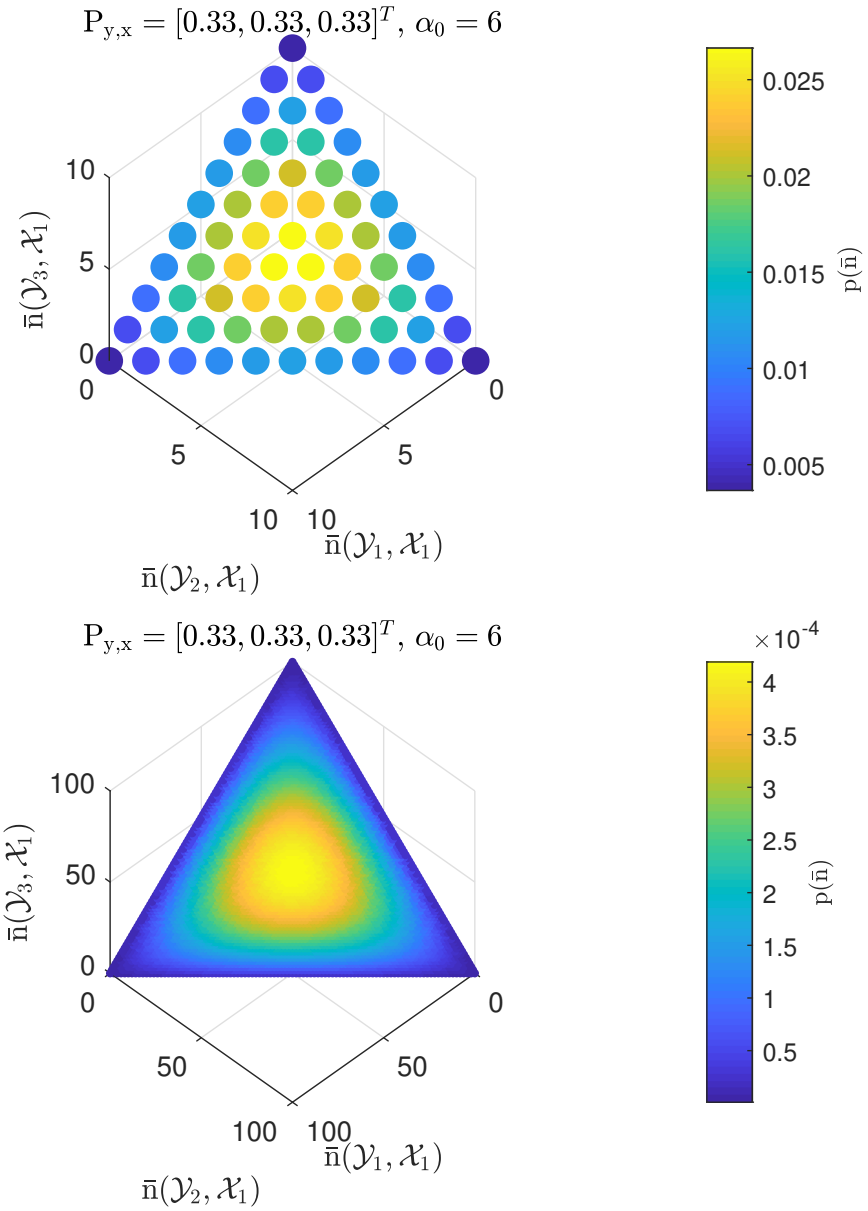

Figure 3.5: $\mathrm{P}(\bar{\mathrm{n}})$ for different training set sizes N

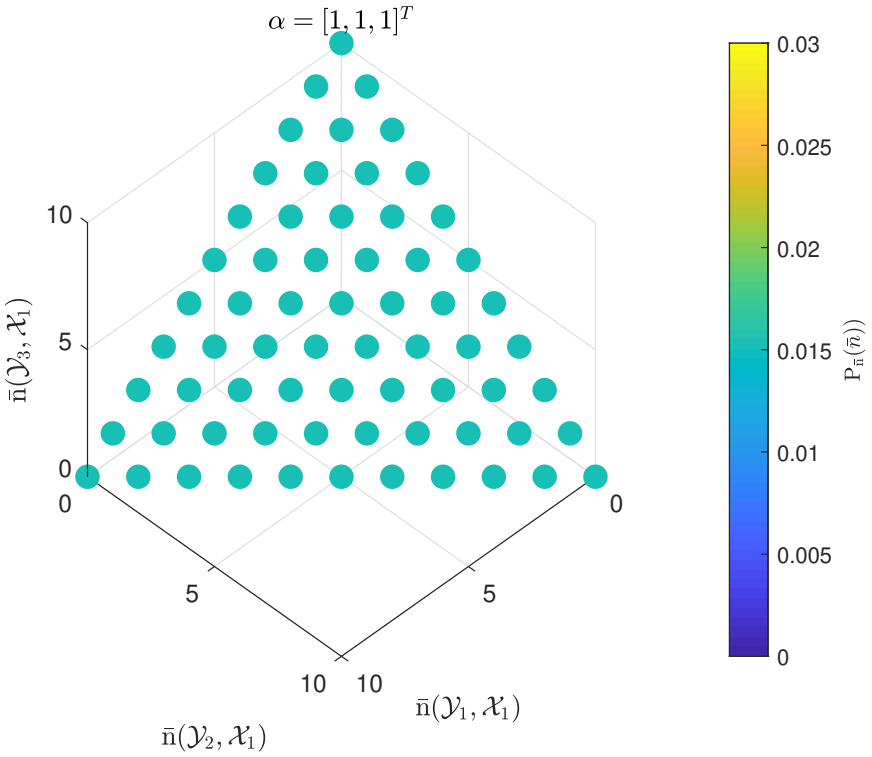

Figure 3.6: P(n̄) for uniform prior

Uniform Prior For the uniform distribution, $\alpha(y, x) = 1$,

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

and

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

The distribution of n̄ is uniform over the set $\bar{\mathcal{N}}$. The PMF for D depends on the training data only through the multinomial coefficient; consequently, more "concentrated" training sets are more probable.

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', the marginal distribution of X can be expressed as

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

As the model marginal θ' and conditional $\tilde{\theta}$ are independent, the distribution $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}$ can be represented as

$$\begin{aligned} \mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X) &= \mathrm{E}_{\tilde{\theta}}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}\right](Y|X) \\ &= \prod_{x\in\mathcal{X}}\mathrm{E}_{\tilde{\theta}(x)}\left[\prod_{y\in\mathcal{Y}}\tilde{\theta}(y;x)^{\bar{N}(y,x;Y,X)}\right] \\ &= \prod_{x\in\mathcal{X}}\beta\big(\alpha(\cdot,x)\big)^{-1}\beta\big(\alpha(\cdot,x)+\bar{N}(\cdot,x;Y,X)\big)\ . \end{aligned} \tag{3.22}$$

The corresponding distributions for the sufficient statistics will be expressed as well. Recall that $\mathrm{n}'|\theta \sim \mathrm{Multi}(N,\theta')$; by the aggregation property of Dirichlet-Multinomial functions [9], the random process is distributed as $\mathrm{n}' \sim \mathrm{DM}(N,\alpha')$.

Also of interest is the distribution of $\bar{\mathrm{n}}$ conditioned on its aggregation n'. Using the Dirichlet-Multinomial properties presented in Appendix A.3, it can be shown that

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}'}(\bar{n}|n') &= \prod_{x\in\mathcal{X}}\Big[\mathcal{M}\left(\bar{n}(\cdot,x)\right)\beta\big(\alpha(\cdot,x)\big)^{-1}\beta\big(\alpha(\cdot,x)+\bar{n}(\cdot,x)\big)\Big] \\ &= \prod_{x\in\mathcal{X}}\mathrm{DM}\left(\bar{n}(\cdot,x);n'(x),\alpha(\cdot,x)\right) \end{aligned} \tag{3.23}$$

over the domain $\left\{\bar{n}\in{\mathbb{Z}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}}\bar{n}(y,\cdot)=n'\right\}$. Observe that conditioning on the aggregation renders the function segments $\bar{\mathrm{n}}(\cdot,x)$ independent of one another and that they are also Dirichlet-Multinomial, such that $\bar{\mathrm{n}}(\cdot,x)|\,\mathrm{n}'(x) \sim \mathrm{DM}\left(\mathrm{n}'(x),\alpha(\cdot,x)\right)$.

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

As shown in Equation (2.8), 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 [18]; thus, the model posterior PDF given the training data is

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

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

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,

$$\theta_{\mathrm{MAP}}(D) = \arg\max_{\theta\in\Theta} \mathrm{P}_{\theta|\,\mathrm{D}}(\theta|D) = \frac{\bar{N}(D) + \alpha - 1}{N + \alpha_0 - |\,\mathcal{Y}\,||\,\mathcal{X}\,|}\,. \tag{3.25}$$

This maximizing value is only valid when $\bar{N}(D) > 1$. For the uniform model prior, the maximizing value of the posterior is the empirical PMF $\bar{N}(D)/N$.

PGR: MAP discussion out of place??

Also, the concentration parameter increases proportionately with increasing volumes of training data; consequently, as $N \to \infty$, the posterior converges to $\mathrm{p}_{\theta|\,\mathrm{D}} \to \delta\big(\,\cdot - \bar{N}(\mathrm{D})/N\big)$. Thus, as more data is collected, the model can be more positively identified and used to formulate minimum risk decisions. Conversely, as $\alpha_0 \to \infty$, the prior model certainty is stronger and the posterior tends toward $\mathrm{p}_{\theta|\,\mathrm{D}} \to \delta(\cdot - \alpha/\alpha_0)$, independent of the training data.

Figure 3.7 shows the influence of the training data on the model distribution; after conditioning on the training data (via $\bar{\mathrm{n}}$), the PDF concentration shifts away from the models favored by the prior knowledge and towards other models that better account for the observations.

The joint PMF of y and x conditioned on the training data is expressed as [12]

$$\begin{aligned} \mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{D}} = \mu_{\theta|\,\mathrm{D}} &= \frac{\alpha + \bar{N}(\mathrm{D})}{\alpha_0 + N} \\ &= \left(\frac{\alpha_0}{\alpha_0 + N}\right)\frac{\alpha}{\alpha_0} + \left(\frac{N}{\alpha_0 + N}\right)\frac{\bar{N}(\mathrm{D})}{N} \\ &= \left(\frac{\alpha_0}{\alpha_0 + N}\right)\mathrm{P}_{\mathrm{y,x}} + \left(\frac{N}{\alpha_0 + N}\right)\frac{\bar{N}(\mathrm{D})}{N}\,. \end{aligned} \tag{3.26}$$

This is a mixture distribution of the prior expectation $\mu_{\theta} = \alpha/\alpha_0$ and the empirical distribution $\bar{N}(\mathrm{D})/N$. The more informative the model prior (i.e. larger α_0), the more the

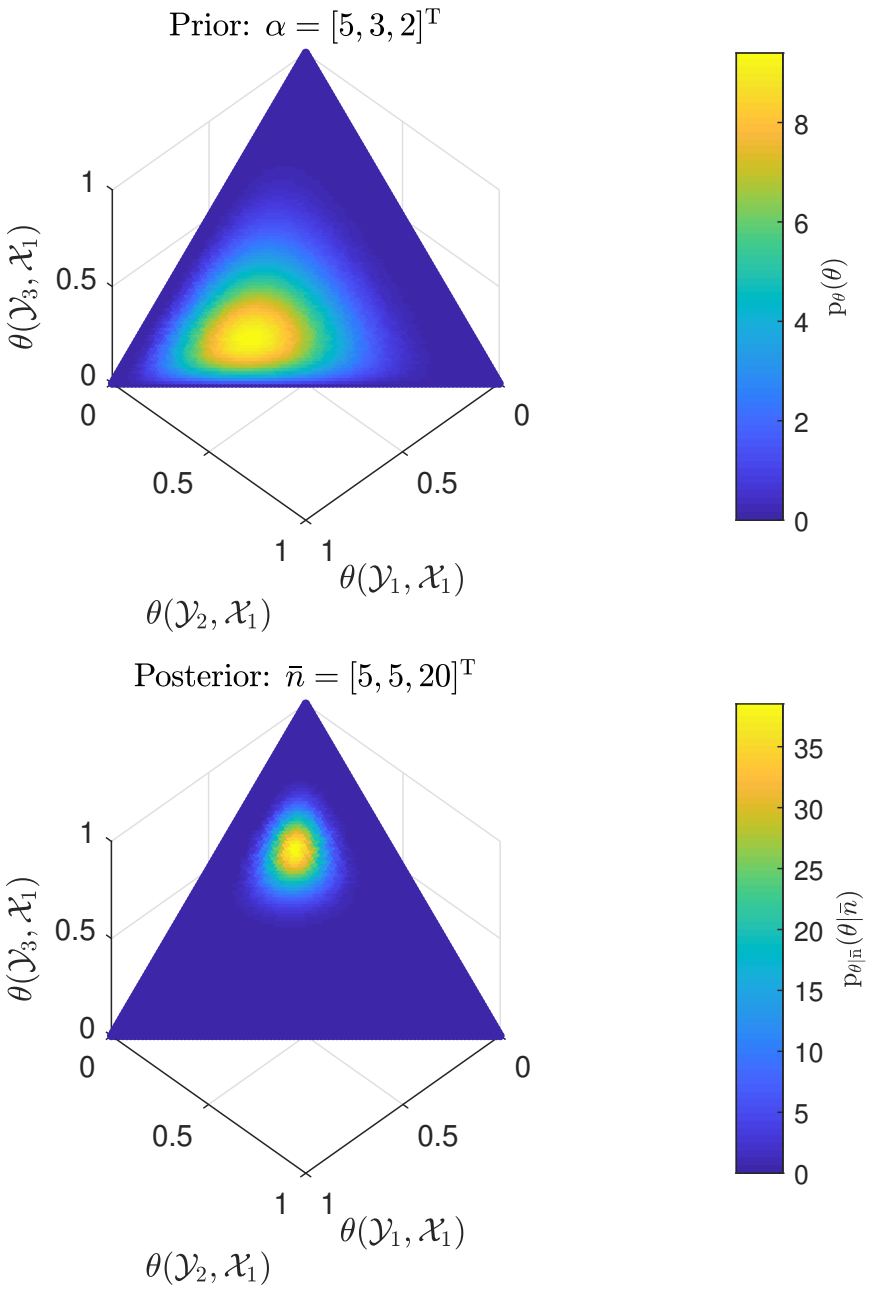

Figure 3.7: Model θ PDF, prior and posterior

prior mean is favored; the more data, the more the empirical PMF is favored. The marginal distribution for x given D is

$$\begin{aligned} \mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} &= \frac{\alpha' + N'(\mathrm{D})}{\alpha_0 + N} \\ &= \left(\frac{\alpha_0}{\alpha_0 + N}\right)\frac{\alpha'}{\alpha_0} + \left(\frac{N}{\alpha_0 + N}\right)\frac{N'(\mathrm{D})}{N} \\ &= \left(\frac{\alpha_0}{\alpha_0 + N}\right)\mathrm{P}_{\mathrm{x}} + \left(\frac{N}{\alpha_0 + N}\right)\frac{N'(\mathrm{D})}{N}\ . \end{aligned} \tag{3.27}$$

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

$$\begin{aligned} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} &= \frac{\alpha(\cdot,\mathrm{x}) + \bar{N}(\cdot,\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\bar{N}(\cdot,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\bar{N}(\cdot,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})}\ . \end{aligned} \tag{3.28}$$

The last representation views the distribution as a convex combination of two conditional distributions. The first distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$ 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 PMF and depends on D, not on α.

The weighting factors $\alpha'(x)$ and $N'(\mathrm{x};\mathrm{D})$ are the concentration of the conditional prior $\tilde{\theta}(x)$ and the number of training samples satisfying $X_n = \mathrm{x}$. As $N'(\mathrm{x};\mathrm{D})/\alpha'(x) \to 0$, the PMF tends toward the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$, which only depends on the model parameter α. As $N'(\mathrm{x};\mathrm{D})/\alpha'(x) \to \infty$, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ tends towards the empirical conditional distribution.

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

$$\begin{aligned} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} &= \frac{\bar{N}(\cdot,\mathrm{x};\mathrm{D}) + 1}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|} + \\ &\qquad \left(\frac{N'(\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|}\right)\frac{\bar{N}(\cdot,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})}\ . \end{aligned} \tag{3.29}$$

Now the prior PMF contribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$ is a uniform distribution over the $|\,\mathcal{Y}\,|$ possible outputs. The weighting factors are controlled by $\alpha'(\mathrm{x}) = |\,\mathcal{Y}\,|$; the more possible outcomes $|\,\mathcal{Y}\,|$ there are for a given training set size, the more the conditional distribution tends toward the uniform PMF implied by the model prior.

3.1.3.1 Representation using the complete model posterior

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 $\mathrm{p}_{\theta'|\,\mathrm{D}}$ and $\mathrm{p}_{\tilde{\theta}|\,\mathrm{x,D}}$, respectively. As the Dirichlet assumption renders θ' and $\tilde{\theta}$ independent, it can be shown that $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}} = \mathrm{E}_{\tilde{\theta}}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}\right]$ and thus that θ' is conditionally independent of Y given X. Furthermore, the Dirichlet distribution $\mathrm{p}_{\theta'}$ is the conjugate prior for $\mathrm{P}_{\mathrm{X}\,|\theta'}$. As a result, $\theta'|\,\mathrm{D} \sim \mathrm{Dir}(\alpha' + N'(\mathrm{X}))$ and

$$\begin{aligned} \mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}(D) &= \mu_{\theta'|\,\mathrm{D}}(D) = \mu_{\theta'|\,\mathrm{X}}(X) \\ &= \frac{\alpha' + N'(X)}{\alpha_0 + N} \,. \end{aligned} \tag{3.30}$$

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

$$\begin{aligned} \mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}(\bar{n}) &= \mu_{\theta'|\,\bar{\mathrm{n}}}(\bar{n}) = \mu_{\theta'|\,\mathrm{n}'}\left(\sum_y \bar{n}(y,\cdot)\right) \\ &= \frac{\alpha' + \sum_y \bar{n}(y,\cdot)}{\alpha_0 + N} \,, \end{aligned} \tag{3.31}$$

where the dependency on $\bar{\mathrm{n}}$ is expressed only through the marginal random process n'.

The posterior $\mathrm{p}_{\tilde{\theta}|\,\mathrm{x,D}}$ can be simplified by noting that the independence of θ' and $\tilde{\theta}$ implies $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X,x}} = \mathrm{E}_{\tilde{\theta}}\left[\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}\right] = \mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X}}$. Consequently, $\tilde{\theta}$ is conditionally independent of x given D. Thus, as $\mathrm{p}_{\tilde{\theta}}$ is a conjugate prior for $\mathrm{P}_{\mathrm{Y}\,|\,\mathrm{X},\tilde{\theta}}$ the posterior distribution is

$$\begin{aligned} \mathrm{p}_{\tilde{\theta}|\,\mathrm{D,x}}\left(\tilde{\theta}|D,x\right) = \mathrm{p}_{\tilde{\theta}|\,\mathrm{D}}\left(\tilde{\theta}|D\right) &= \prod_{x'\in\mathcal{X}} \mathrm{p}_{\tilde{\theta}(x')|\,\mathrm{D}}\left(\tilde{\theta}(x')|D\right) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x'); \alpha(\cdot,x') + \bar{N}(\cdot,x';D)\right) \end{aligned} \tag{3.32}$$

and the distinct model conditional PMF's are independent from one another. A similar treatment demonstrates that

$$\begin{aligned} \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}},\mathrm{x}}\left(\tilde{\theta}|\bar{n},x\right) = \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}}}\left(\tilde{\theta}|\bar{n}\right) &= \prod_{x'\in\mathcal{X}} \mathrm{p}_{\tilde{\theta}(x')|\,\bar{\mathrm{n}}(\cdot,x')}\left(\tilde{\theta}(x')|\bar{n}(\cdot,x')\right) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x'); \alpha(\cdot,x') + \bar{n}(\cdot,x')\right) \,. \end{aligned} \tag{3.33}$$

Observe that when the conditioning is performed using the sufficient statistic, the independent conditional models $\tilde{\theta}(x)$ are only dependent on their corresponding subset of the empirical PMF, $\bar{n}(\cdot, x)$.

The Bayes predictive PMF can thus be expressed as

$$\begin{aligned} P_{y\,|\,x,D}(x, D) &= \mu_{\tilde{\theta}(x)|\,x,D}(x, D) = \mu_{\tilde{\theta}(x)|\,D}(D) \\ &= \frac{\alpha(\cdot, x) + \bar{N}(\cdot, x; D)}{\alpha'(x) + N'(x; D)} \end{aligned} \tag{3.34}$$

or, via the sufficient statistic,

$$\begin{aligned} P_{y\,|\,x,\bar{n}}(x, \bar{n}) &= \mu_{\tilde{\theta}(x)|\,x,\bar{n}}(x, \bar{n}) = \mu_{\tilde{\theta}(x)|\,\bar{n}(\cdot,x)}\big(\bar{n}(\cdot, x)\big) \\ &= \frac{\alpha(\cdot, x) + \bar{n}(\cdot, x)}{\alpha'(x) + \sum_y \bar{n}(y, x)} \ . \end{aligned} \tag{3.35}$$

A consequence of the Dirichlet prior is that the predictive PMF for a given value of x only depends on the corresponding training data $\bar{n}(\cdot, x)$, such that $P_{y\,|\,x,\bar{n}}(x, \bar{n}) = P_{y\,|\,x,\bar{n}(\cdot,x)}\big(x, \bar{n}(\cdot, x)\big)$. This is intuitive considering the independence of the conditional models $\tilde{\theta}(x)$ from one another.

3.2 Model Estimation Perspective

PGR: generalize, move before Dirichlet?

PGR: use poster table??

It is instructive to treat the distribution $P_{y\,|\,x,\bar{n}}$ as an estimate of the unknown conditional PMF $P_{y\,|\,x,\theta} \equiv \tilde{\theta}(x)$ and investigate the effects of informative prior knowledge. For a given x and corresponding number of training samples $n'(x)$, the expected value of the estimate condtioned on the true model θ is

$$\begin{aligned} E_{\bar{n}\,|\,n',\theta}\left[P_{y\,|\,x,\bar{n}}\right] &= E_{\bar{n}(\cdot,x)|\,n'(x),\tilde{\theta}(x)}\left[P_{y\,|\,x,\bar{n}(\cdot,x)}\right] \\ &= \left(\frac{\alpha'(x)}{\alpha'(x) + n'(x)}\right)\frac{\alpha(\cdot, x)}{\alpha'(x)} + \left(\frac{n'(x)}{\alpha'(x) + n'(x)}\right)\tilde{\theta}(x) \ , \end{aligned} \tag{3.36}$$

where the properties of a multinomial distribution conditioned on its aggregation have been used. The result is a convex combination of the conditional data-independent distribution

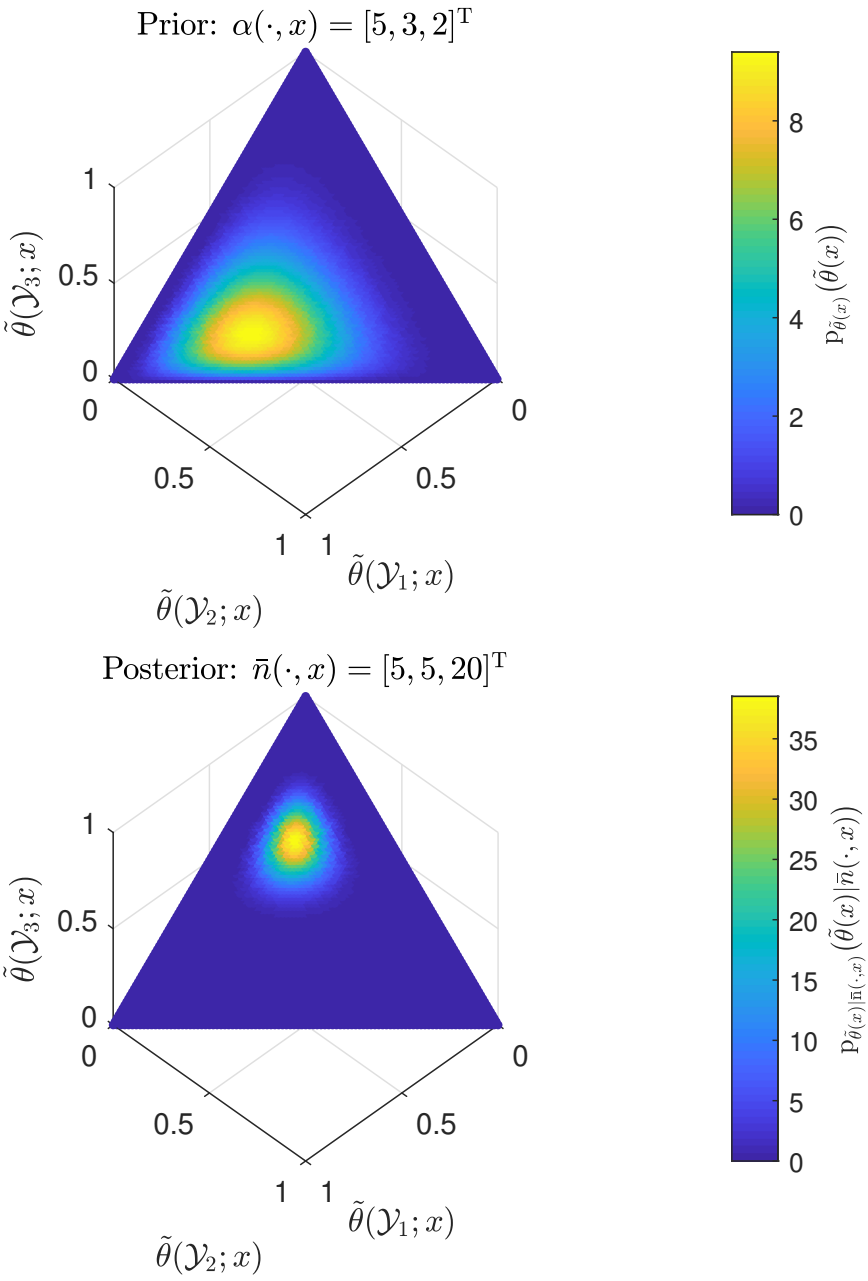

Figure 3.8: Model PDF, prior and posterior

$\alpha(\cdot, x)/\alpha'(x)$ and the true conditional distribution $\tilde{\theta}(\mathrm{x})$. The convex coefficients are dependent on the "marginal" values α' and n'; note that as the number of matching training samples $\mathrm{n}'(x)$ increases relative to $\alpha'(x)$, the estimate tends towards the true conditional PMF.

PGR: suppress delta dependency on nbar and theta?

To aid characterization of the estimator, define the random process $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \in \mathbb{R}^{\mathcal{Y}}$. For a given x and corresponding number of training samples $\mathrm{n}'(\mathrm{x})$, the bias of the conditional PMF estimate is

$$\begin{aligned} \mathrm{Bias}(\mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta)\right] \\ &= \frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}\left(\frac{\alpha(\cdot, \mathrm{x})}{\alpha'(\mathrm{x})} - \tilde{\theta}(\mathrm{x})\right) \end{aligned} \tag{3.37}$$

and its covariance function is

$$\begin{aligned} \mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{C}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot\,|\,\mathrm{x}, \bar{\mathrm{n}})\right](y, y') \\ &= \frac{\Sigma_{\bar{\mathrm{n}}(\cdot,\mathrm{x})|\,\mathrm{n}'(\mathrm{x}),\tilde{\theta}(\mathrm{x})}(y, y')}{\left(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\right)^2} \\ &= \frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\right)^2}\left(\tilde{\theta}(y; \mathrm{x})\delta[y, y'] - \tilde{\theta}(y; \mathrm{x})\tilde{\theta}(y'; \mathrm{x})\right) , \end{aligned} \tag{3.38}$$

where the properties of multinomial random processes have been used. Note that the bias is proportionate to the difference between the true conditional model and the data-independent estimate. The scaling factor tends from one to zero as $\mathrm{n}'(x)/\alpha'(x)$ tends from zero to infinity; as such, more informative priors (large $\alpha'(x)$) will lead to PMF estimates that are prone to bias. Conversely, the variance of the PMF estimate tends to zero as $\alpha'(x) \to \infty$.

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

$$\begin{aligned} \mathcal{E}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(y; \mathrm{x}, \bar{\mathrm{n}}, \theta)\Delta(y'; \mathrm{x}, \bar{\mathrm{n}}, \theta)\right] \\ &= \mathrm{Bias}(y; \mathrm{x}, \mathrm{n}', \theta)\mathrm{Bias}(y'; \mathrm{x}, \mathrm{n}', \theta) + \mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta) . \end{aligned} \tag{3.39}$$

As $\mathrm{n}'(x) \to \infty$, this function tends to zero and thus the underlying model $\tilde{\theta}(x)$ is determined precisely. A more practical case is estimation with a finite volume of training data. Specification of the Dirichlet model prior can be interpreted as providing a distribution estimate $\alpha(\cdot, x)/\alpha'(x)$ and a confidence level $\alpha'(x)$. Higher confidence reduces error due to

the variance of the estimator, but increases the error due to bias between the true model and its estimate; low confidence renders the estimate unbiased, but maximizes the estimator variance.

Also of interest, the conditional expectation of $\mathcal{E}(\cdot,\cdot;\mathrm{x},\mathrm{n}',\theta)$ is

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x},\mathrm{n}'\,|\,\theta}\left[\mathcal{E}(y,y';\mathrm{x},\mathrm{n}',\theta)\right] \qquad (3.40)\\
&\quad=\mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta}\left[\left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)^2\right]\left(\frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})}-\tilde{\theta}(y;\mathrm{x})\right)\left(\frac{\alpha(y',\mathrm{x})}{\alpha'(\mathrm{x})}-\tilde{\theta}(y';\mathrm{x})\right)\right]\\
&\qquad+\mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta}\left[\frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\right]\left(\tilde{\theta}(y;\mathrm{x})\delta[y,y']-\tilde{\theta}(y;\mathrm{x})\tilde{\theta}(y';\mathrm{x})\right)\right]\ .
\end{aligned}$$

To exemplify how the model estimate $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ approximates $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$, consider a scenario with $|\,\mathcal{Y}\,|=10$. The data-independent PMF $\alpha(\cdot,x)/\alpha'(x)$ and true model $\tilde{\theta}(x)$ are shown in Figure 3.9 - note the significant mismatch.

PGR: ABOVE - STATE XCAL cardinality = 1

Figures 3.10 and 3.11 show how the bias and variance of the estimate change for different values of $\mathrm{n}'(x)$ and $\alpha'(x)$. The plot markers represent the conditional mean of the estimator, $\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|\,\mathrm{x},\bar{\mathrm{n}})\right]$; the upper and lower error bars represent the square-root of the expected squared deviation above and below the conditional mean, respectively. Each individual plot heading provides the error $\sqrt{\sum_{y\in\mathcal{Y}}\mathcal{E}(y,y;\mathrm{x},\mathrm{n}',\theta)}$ to assess the quality of the PMF estimate.

Observe that for $\mathrm{n}'(x)=1$, the high variance of the $\alpha'(x)=0.1$ estimate (favoring the empirical PMF) renders it worse than the $\alpha_0=10$ estimate; in fact, the variance is so high that the error exceeds that of the data-independent estimate $\alpha(\cdot,x)/\alpha'(x)$ (Figure 3.9). Conversely, for $\mathrm{n}'(x)=10$, the confidence of the $\alpha'(x)=10$ estimate leads to high bias and the $\alpha'(x)=0.1$ estimate is superior. For $\mathrm{n}'(x)=100$, both the $\alpha'(x)=0.1$ and $\alpha'(x)=10$ estimates begin converging to the true distribution - this is guaranteed due to the full support of the Dirichlet prior.

PGR: full support discussion?

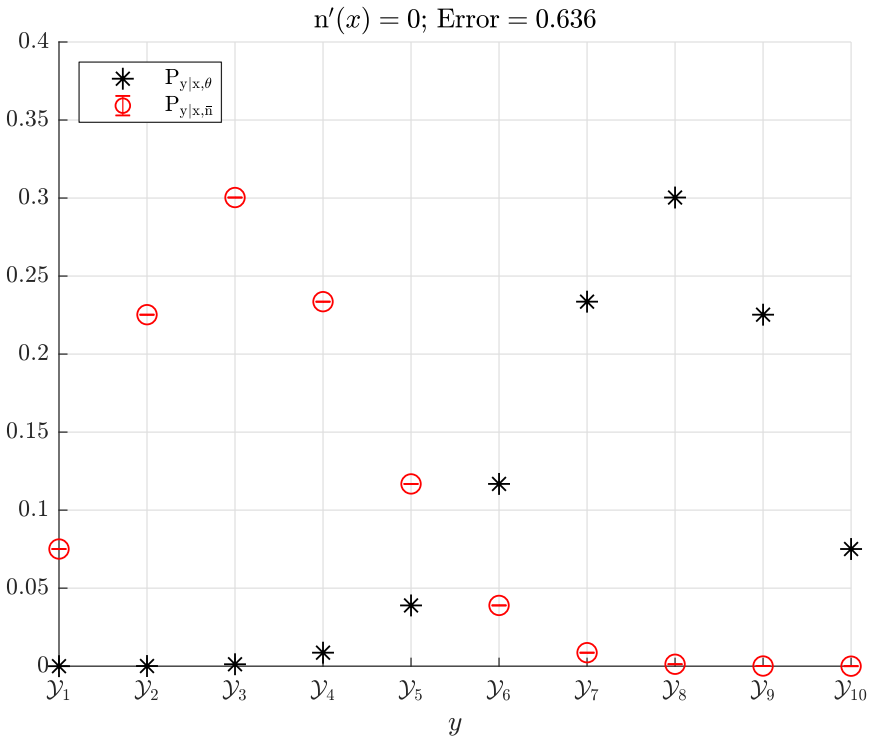

Figure 3.9: Model θ estimate, no training data

3.3 Applications to Common Loss Functions

PGR: REMOVE GENERAL, RELOCATED MATERIAL!

PGR: equations, plots for specific theta results? informative/non-informative tradeoff??? alpha/theta mismatch results???

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 conditional risk $\mathcal{R}_\Theta(f^*;\theta)$ is analyzed.

PGR: add formula for f(D)??? EMPIRICAL RISK DISCUSS, REGULARIZING weight

It is useful to substitute the Bayes predictive distribution using the Dirichlet prior (3.28)

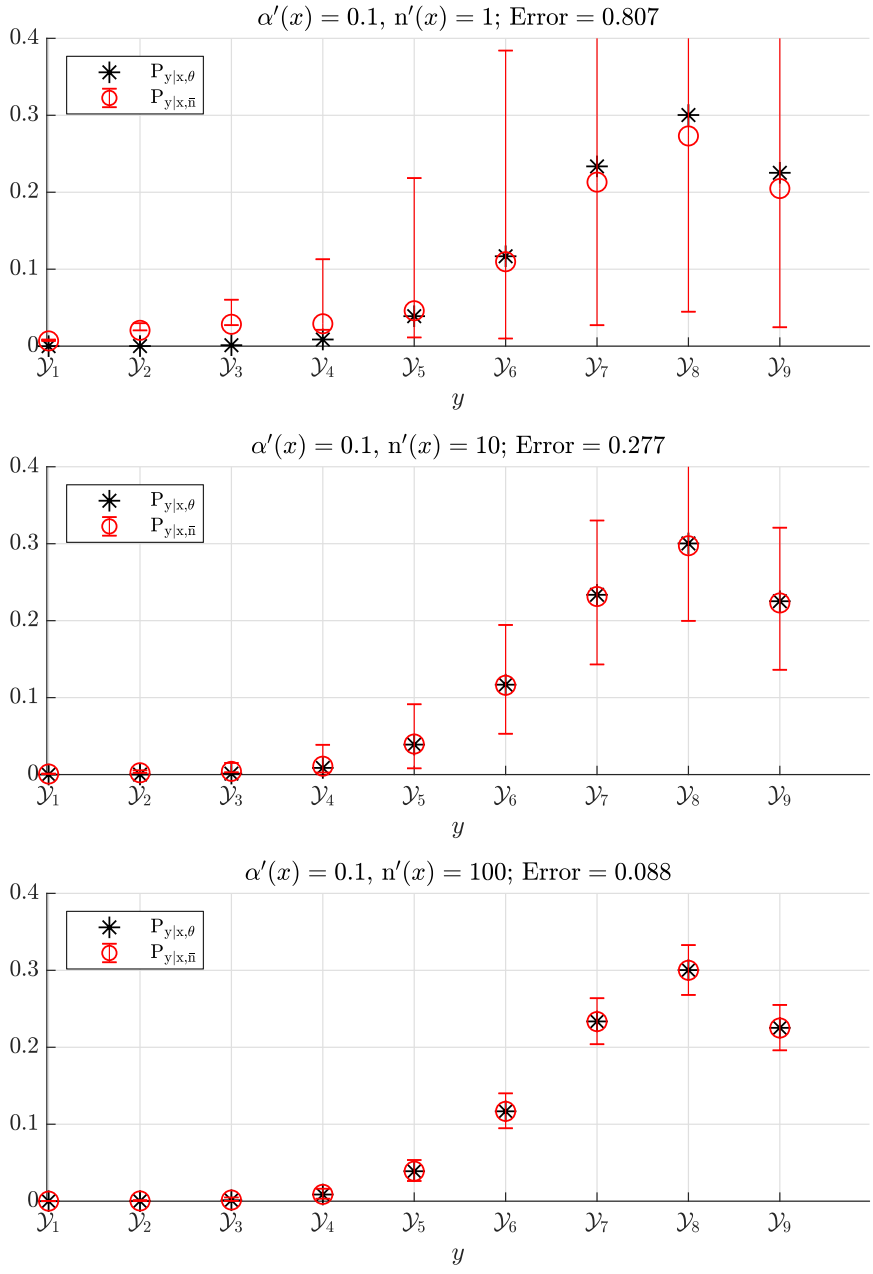

Figure 3.10: Model θ estimates, $\alpha_0 = 0.1$

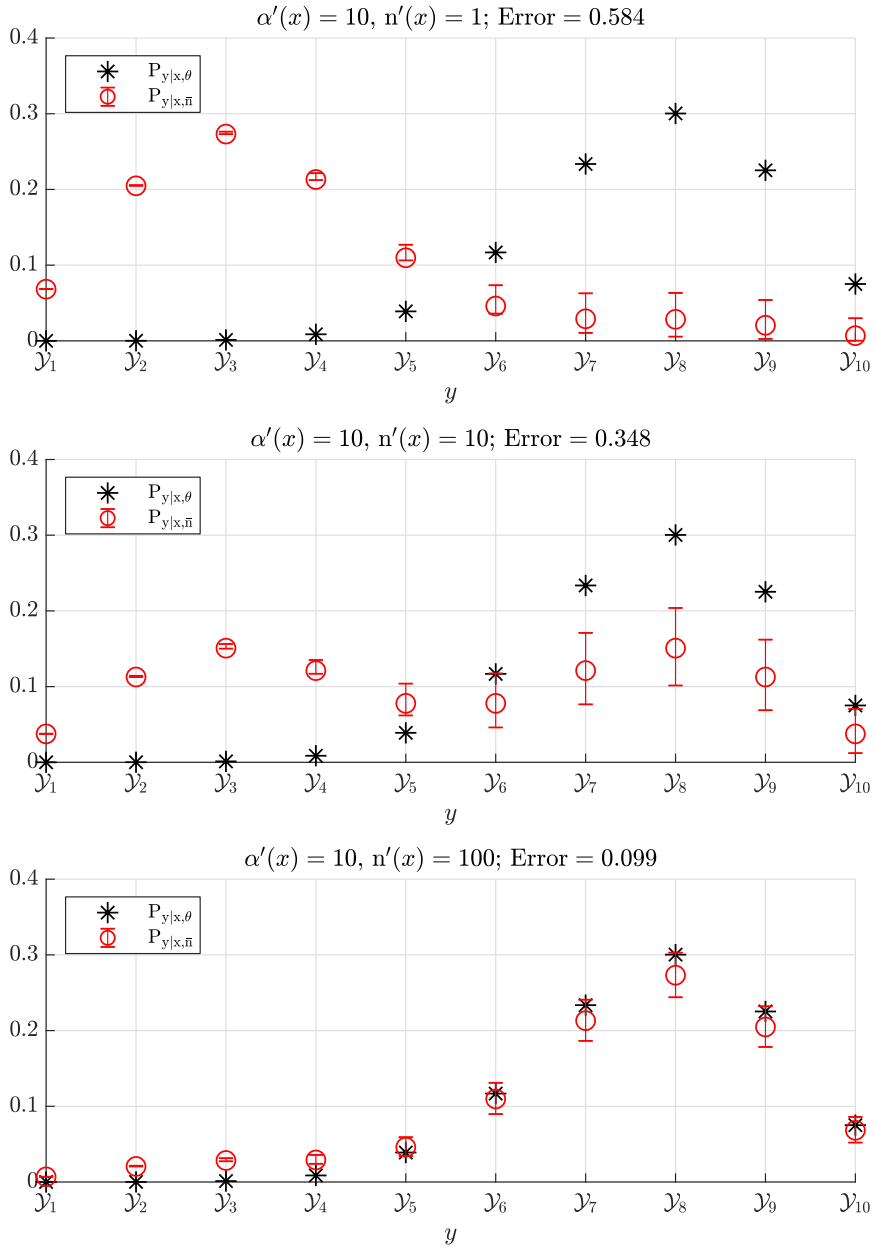

Figure 3.11: Model θ estimates, $\alpha_0 = 10$

into Equation (2.8), expressing the decision for a given input x and training set D as

$$
\begin{aligned}
f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] \qquad (3.41)\\
&= \arg\min_{h\in\mathcal{H}} \frac{\sum_{y\in\mathcal{Y}} \alpha(y,\mathrm{x})\,\mathcal{L}(h,y) + \sum_{y\in\mathcal{Y}} \bar{N}(y,\mathrm{x};\mathrm{D})\,\mathcal{L}(h,y)}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\\
&= \arg\min_{h\in\mathcal{H}} \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \sum_{y\in\mathcal{Y}} \frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})}\,\mathcal{L}(h,y) + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \sum_{y\in\mathcal{Y}} \frac{\bar{N}(y,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})}\,\mathcal{L}(h,y)\\
&= \arg\min_{h\in\mathcal{H}} \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathcal{L}(h,\mathrm{y})\right] + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \frac{\sum_{n=1}^{N} \delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathcal{L}\left(h,\mathrm{Y}_n\right)}{\sum_{n=1}^{N} \delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]}\,.
\end{aligned}
$$

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 $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$, which reflects the prior knowledge imparted by the model parameter α. 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 parameter $\alpha'(\mathrm{x})$ and the number of matching training samples $N'(\mathrm{x};\mathrm{D})$ 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???

PGR: add Dir conditional risk and analysis!!!

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 Optimal Estimate: the Posterior Mean

PGR: plots?

Substituting in the Bayes predictive distribution for a Dirichlet prior (3.28) into (2.31),

the optimal Bayesian estimate is

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\sum_{y\in\mathcal{Y}} y\frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\sum_{y\in\mathcal{Y}} y\frac{\bar{N}(y,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\mu_{\mathrm{y}\,|\,\mathrm{x}} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathrm{Y}_n}{N'(\mathrm{x};\mathrm{D})}\ . \end{aligned} \tag{3.42}$$

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 $\alpha'(\mathrm{x})$) 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

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \left(\frac{|\,\mathcal{Y}\,|}{N'(\mathrm{x};\mathrm{D})+|\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y + \left(\frac{N'(\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})+|\,\mathcal{Y}\,|}\right)\sum_{y\in\mathcal{Y}} y\frac{\bar{N}(y,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N'(\mathrm{x};\mathrm{D})+|\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y + \left(\frac{N'(\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})+|\,\mathcal{Y}\,|}\right)\frac{\sum_{n=1}^{N}\delta\big[\,\mathrm{x},\mathrm{X}_n\,\big]\,\mathrm{Y}_n}{N'(\mathrm{x};\mathrm{D})}\ . \end{aligned} \tag{3.43}$$

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 $\mathcal{Y}$.

3.3.1.2 Minimum Risk: the Expected Posterior Variance

PGR: determine irreducible risk separately, before??

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 $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$, the minimum risk can also be represented as $\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right]$; as such, the expectations are performed over $\bar{\mathrm{n}}$. Decompose the conditional variance as

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} \tag{3.44}$$

and assess the expected values of these terms separately using distributions derived from the

Dirichlet prior. The first term is simply

$$\begin{aligned}\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2]\right] = \mathrm{E}_{\mathrm{y}}[\mathrm{y}^2] &= \sum_{y\in\mathcal{Y}} y^2 \left(\sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0}\right) \\ &= \mathrm{E}_{\mathrm{x}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2]\right] = \sum_{x\in\mathcal{X}} \frac{\alpha'(x)}{\alpha_0} \sum_{y\in\mathcal{Y}} y^2 \frac{\alpha(y,x)}{\alpha'(x)}\ ,\end{aligned}$$

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

$$\begin{aligned}\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{x}}\left[\frac{\big(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{y\in\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\big)^2}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\big)^2}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\bar{\mathrm{n}}}\left[\frac{\alpha_0\big(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{y\in\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\big)^2}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\big)(\alpha_0+N)}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{n}'}\left[\frac{\alpha_0\,\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}\left[\big(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{y\in\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\big)^2\right]}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\big)(\alpha_0+N)}\right]\right] \\ &= \ldots \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\,\mathrm{E}_{\mathrm{n}'}\left[\mathrm{n}'(\mathrm{x})\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \big(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})+1\big)\alpha'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\right]}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x})+1\big)(\alpha_0+N)}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{N\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \big(\alpha_0\alpha'(\mathrm{x}) + N\alpha'(\mathrm{x}) + \alpha_0\big)\mu^2_{\mathrm{y}\,|\,\mathrm{x}}}{\big(\alpha'(\mathrm{x})+1\big)(\alpha_0+N)}\right]\ .\end{aligned} \tag{3.45}$$

PGR: provide additional steps?

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

PGR: move to appendix???

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

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\alpha'(\mathrm{x}) + N\alpha'(\mathrm{x}) + \alpha_0}{\big(\alpha'(\mathrm{x})+1\big)(\alpha_0+N)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + (\alpha_0+N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + \alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]\ .\end{aligned} \tag{3.46}$$

The minimum risk is the expected value of the scaled conditional variance with respect to $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$. The expectation is taken with respect to the prior marginal distribution $\mathrm{P}_{\mathrm{x}} = \alpha'/\alpha_0$.

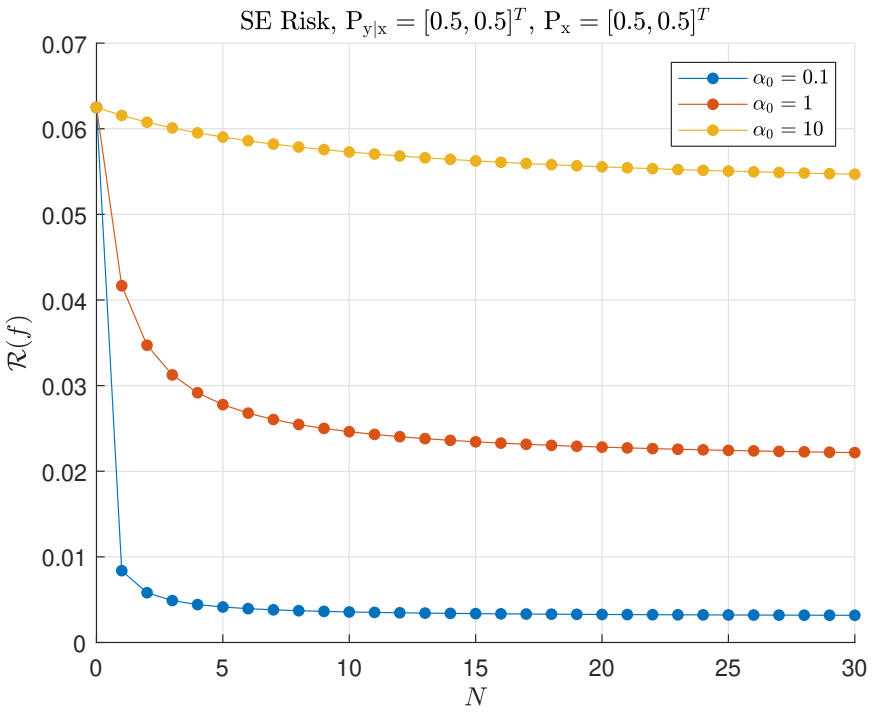

Figure 3.12: Minimum SE Risk for different training set sizes N

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 $\mathcal{R}^* \to \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$; note that this is equivalent to the 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 $\mathcal{R}^* = \Sigma_{\mathrm{y}\,|\,\mathrm{x}}\,\mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+(\alpha_0+N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}}\right]$. Figures 3.12 and 3.13 display how the risk changes with N and α_0 when $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$ and P_{x} are fixed.

It may not seem intutitve for the risk to decrease when α_0 is smaller – the variance of the model θ increases and the prior knowledge is less definitive. This is a result of the Dirichlet PDF weight shifting towards the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models which have ℓ_0 norms satisfying

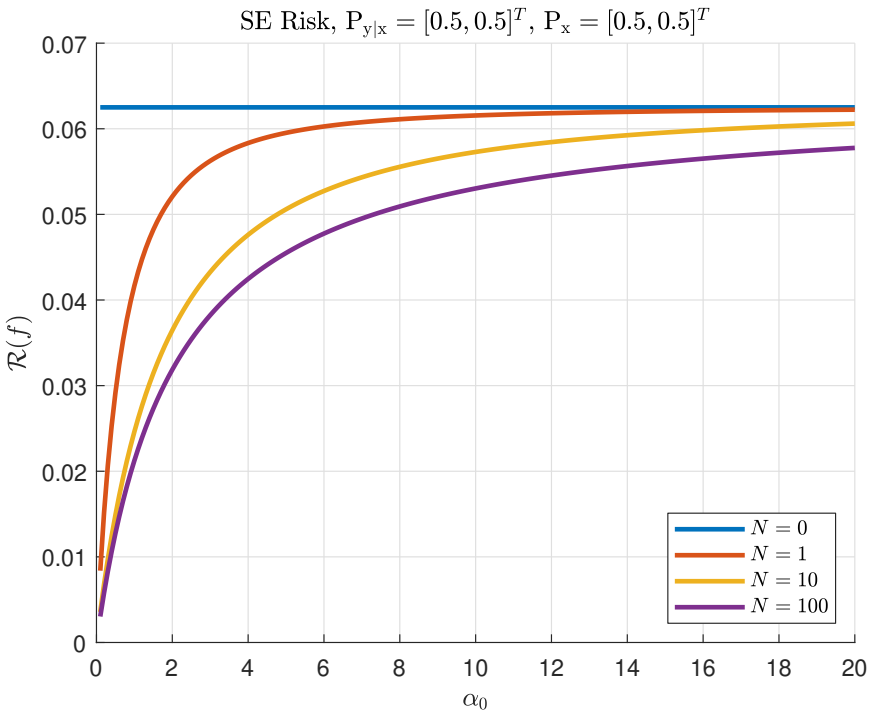

Figure 3.13: Minimum SE Risk for different prior concentrations α_0

$\|\theta\|_0 = 1$. Although these PMF's are maximally separated (and uncorrelated), they all have zero variance. The optimal learner (3.42) 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 $\mathrm{P}_{\mathrm{y}|\mathrm{x}}$. Figure 3.14 demonstrates how the squared-error tends towards zero for PMFs that have ℓ_0-norm equal to one.

Next, consider the effect of the marginal distribution P_{x}. Figure 3.15 demonstrates how the risk changes with this marginal PMF. Observe that the risk is maximal at the distributions satisfying $\|\mathrm{P}_{\mathrm{x}}\|_0 = 1$; the scaling factor for the conditional variance $\Sigma_{\mathrm{y}|\mathrm{x}}$ becomes $\frac{1+(\alpha_0+N)^{-1}}{1+\alpha_0^{-1}}$. Conversely, for $\mathrm{P}_{\mathrm{x}} = 1/|\mathcal{X}|$ the scaling factor becomes $\frac{|\mathcal{X}|^{-1}+(\alpha_0+N)^{-1}}{|\mathcal{X}|^{-1}+\alpha_0^{-1}}$ and the risk is minimal. Figures 3.16 and 3.17 show how different marginals P_{x} affect the risk as a function of N and α_0, respectively.

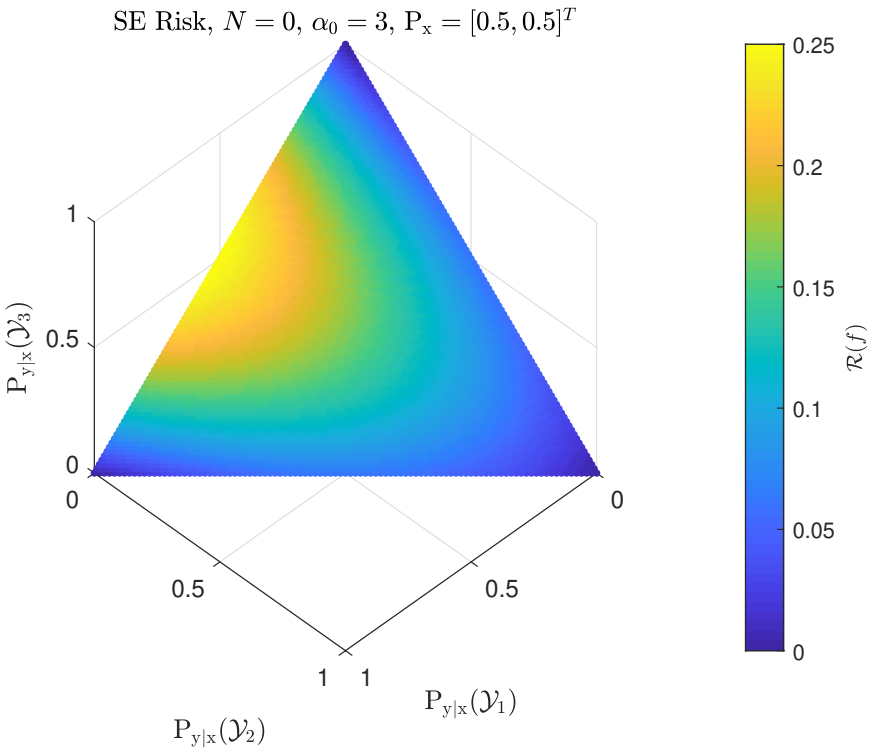

Figure 3.14: Minimum SE Risk for different prior means $\mathrm{P}_{\mathrm{y\,|\,x}}$

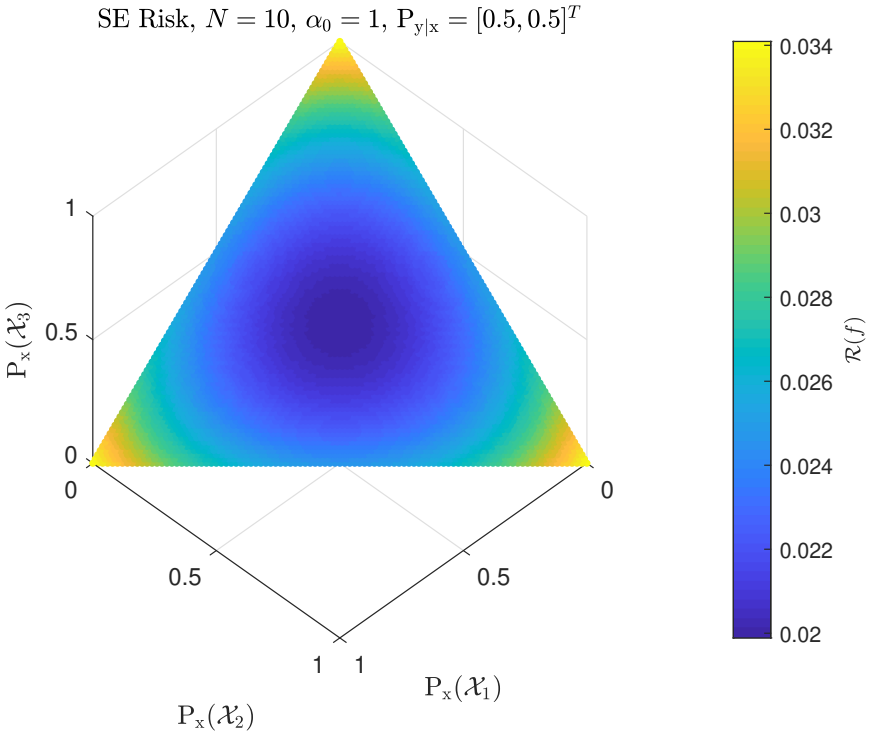

Figure 3.15: Minimum SE Risk for different prior means $\mathrm{P_x}$

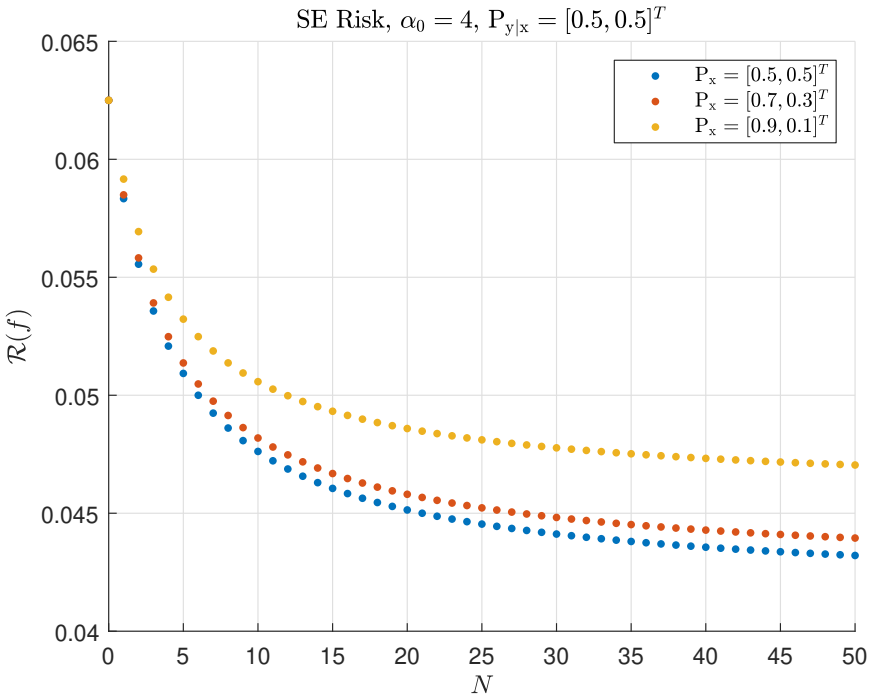

Figure 3.16: Minimum SE Risk for different training set volumes N

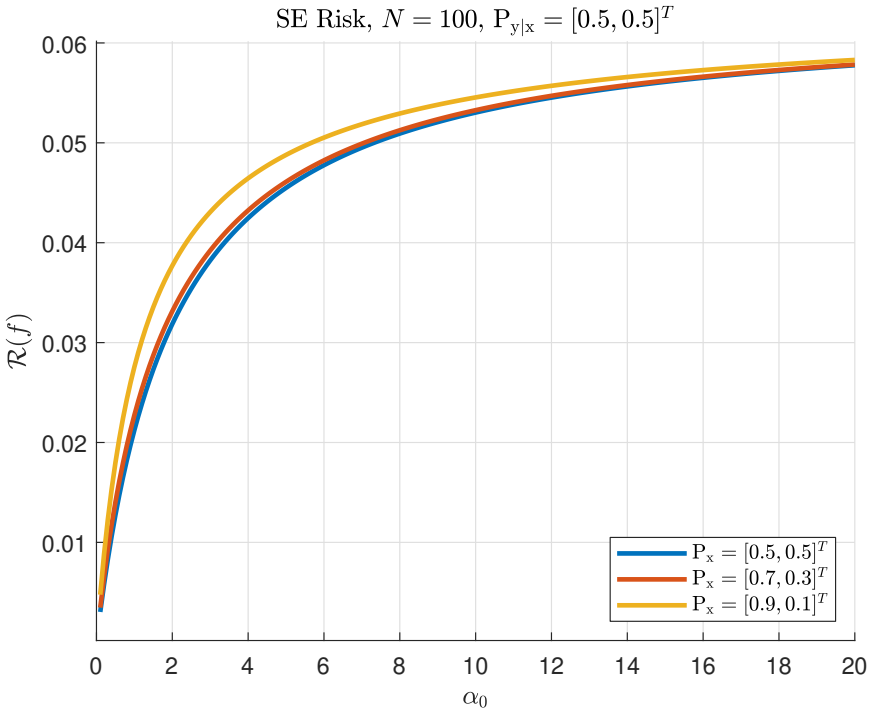

Figure 3.17: Minimum SE Risk for different prior concentrations α_0

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

$$\begin{aligned}\mathcal{R}^* &= \frac{|\,\mathcal{Y}\,|\big(N/|\,\mathcal{X}\,|+|\,\mathcal{Y}\,|+1\big)}{\big(|\,\mathcal{Y}\,|+1\big)\big(N/|\,\mathcal{X}\,|+|\,\mathcal{Y}\,|\big)}\left[\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}}y^2\right)-\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}}y\right)^2\right] \\ &= \frac{1+\big(N/|\,\mathcal{X}\,|+|\,\mathcal{Y}\,|\big)^{-1}}{1+|\,\mathcal{Y}\,|^{-1}}\left[\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}}y^2\right)-\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}}y\right)^2\right]\,.\end{aligned} \tag{3.47}$$

Since all possible values of x are equally probable and the conditional probability $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$ 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 ($N = 0$), the scaling is unity; as $N/|\,\mathcal{X}\,| \to \infty$, the scaling factor is $\big(1+|\,\mathcal{Y}\,|^{-1}\big)^{-1}$.

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

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x}} = \frac{|\,\mathcal{Y}\,|^2-1}{12|\,\mathcal{Y}\,|^2} = \frac{1-|\,\mathcal{Y}\,|^{-2}}{12} \tag{3.48}$$

and the minimum risk is expressed as

$$\begin{aligned}\mathcal{R}^* &= \frac{\big(1-|\,\mathcal{Y}\,|^{-1}\big)\Big(1+\big(N/|\,\mathcal{X}\,|+|\,\mathcal{Y}\,|\big)^{-1}\Big)}{12} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N/|\,\mathcal{X}\,|+|\,\mathcal{Y}\,|}\right)\frac{1-|\,\mathcal{Y}\,|^{-2}}{12}+\left(\frac{N/|\,\mathcal{X}\,|}{N/|\,\mathcal{X}\,|+|\,\mathcal{Y}\,|}\right)\frac{1-|\,\mathcal{Y}\,|^{-1}}{12}\,.\end{aligned} \tag{3.49}$$

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 $\alpha'(\mathrm{x})$ and on $N'(\mathrm{x};\mathrm{D})$. For the uniform prior, $\alpha'(\mathrm{x}) = |\,\mathcal{Y}\,|$ and $\mathrm{E}_{\mathrm{D}}\left[N'(\mathrm{D})\right] = N/|\,\mathcal{X}\,|$.

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. Figure 3.18 illustrates the difference between these cases.

PGR: additional figures for uniform case?

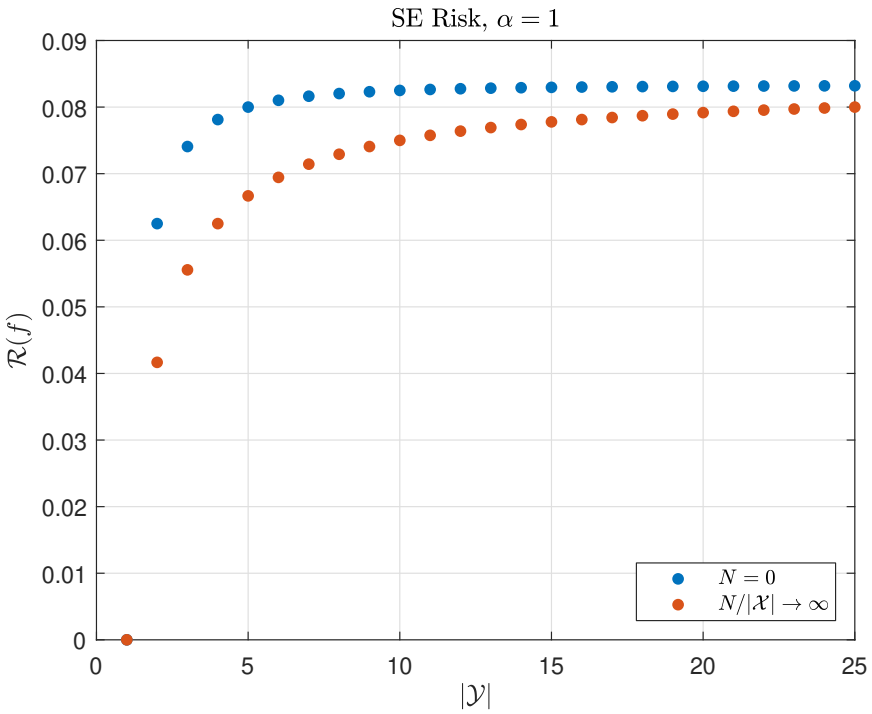

Figure 3.18: Minimum SE Risk, Uniform Prior, zero and infinite training data

3.3.1.3 Conditional Squared-Error for a Dirichlet-based Estimator

Having derived the optimal estimator based on a Dirichlet model prior, it is important to consider the conditional risk $\mathcal{R}_\Theta(f^*;\theta)$ and analyze how different prior parametrizations α influence the squared-error for different models θ. Starting from the conditional squared-error risk (2.28) and substituting the Bayesian estimator (2.31), the formula simplifies to

$$\begin{aligned}\mathcal{R}_\Theta(f^*;\theta) &= \mathcal{R}^*_\Theta(\theta) + \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(f^*(\mathrm{x};\mathrm{D}) - f_\Theta(\mathrm{x};\theta)\right)^2\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x,D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right].\end{aligned} \tag{3.50}$$

Defining the excess conditional risk $\mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta) \equiv \mathcal{R}_\Theta(f;\theta) - \mathcal{R}^*_\Theta(\theta)$, the second term above is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x,D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, the average squared bias between the Bayesian predictive mean and the true predictive mean.

PGR: general above? move before dir? simplify form in terms of Ptheta?

PGR: define excess clairvoyant/bayesian risk up front before dir?

Evaluation of the excess risk for an estimator based on the Dirichlet prior will be performed using the sufficient statistic $\bar{\mathrm{n}}$ in place of the training set D. Using the random

process $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \in \mathbb{R}^{\mathcal{Y}}$ introduced in 3.2, the term is expressed as

$$\begin{aligned}
\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right] \\
&= \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y'\,\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}\,|\theta}\left[\Delta(y;\mathrm{x},\bar{\mathrm{n}},\theta)\Delta(y';\mathrm{x},\bar{\mathrm{n}},\theta)\right] \\
&= \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y'\,\mathrm{E}_{\mathrm{x},\mathrm{n}'\,|\theta}\left[\mathcal{E}(y,y';\mathrm{x},\mathrm{n}',\theta)\right] \\
&= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\,\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\right]\right] \\
&\qquad + \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2 \mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)^2\right]\right],
\end{aligned} \tag{3.51}$$

where the function $\mathcal{E}$ is defined in (3.39).

The excess conditional risk can thus be represented as the conditional expectation (with respect to $\mathrm{P}_{\mathrm{x}\,|\theta}$) of a sum of two functions of x. The first function measures the additional variance beyond that of the clairvoyant estimator (i.e. the clairvoyant squared-error); like the clairvoyant risk, it depends on $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}$, the conditional variance of the clairvoyant estimate for a given observation of x. The second function is dependent on the squared bias between the clairvoyant estimate $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ and the data-independent estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$. This term alone is influenced by the data-independent Bayes predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$.

The two second-order (in terms of y) terms are scaled by factors dependent on the prior concentrations $\alpha'(x)$ and on $\theta'(x)$ and N via conditional expectations with respect to $\mathrm{n}'(x)$. Note that by the aggregation property of multinomial distributions, the random variable $\mathrm{n}'(x)|\theta'(x) \sim \mathrm{Bi}\left(N, \theta'(x)\right)$. Closed-forms have not been found for the function expectations of binomial random variables above.

PGR: binomial citations?

It is instructional to consider the trends in the conditional squared-error risk (3.51) for different volumes of training data N and for different selections of α.

First consider how the excess risk changes with the training volume N. For $N = 0$, it is evident that the excess risk is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, the expected squared bias between the clairvoyant and data-independent estimators. As N tends to infinity, the binomial distribution controlling the scaling factors concentrates at $\mathrm{n}'(x) \approx N\theta'(x)$; as such, the two expectations of interest tend to zero and thus $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to 0$. This desirable

property of the estimator 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 depedency of the excess conditional risk on N is that, depending on the learner parameterization, there may be a local maximum. Consider the trivial case of $|\mathcal{X}| = 1$ - treating N as a real number, there would be a maximum at

$$N \equiv \alpha'(x)\left(1 - 2\alpha'(x)\frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}\right) . \tag{3.52}$$

Note that as the squared-difference between the prior mean and true mean increases, 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) compromize the estimator performance. Also consider the effect of prior concentration; informative priors with sufficiently high $\alpha'(x)$ will not have the local maxima.

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

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta}\left[\frac{\frac{1}{\alpha'(\mathrm{x})} - \frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}}{4\left(1 - \frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}\right)}\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] . \tag{3.53}$$

PGR: better form above???

Figures 3.19 and 3.20 exemplify the excess conditional squared-error as a function of N for estimators based on Dirichlet priors of varying concentration $\alpha'(x)$. 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 distributions $\alpha(\cdot, x)/\alpha'(x)$ and their concentrations $\alpha'(x)$.

First consider the conditional prior PMF's $\alpha(\cdot, x)/\alpha'(x)$; as shown, they manifest themselves in the risk through the squared estimator bias. It is clear that regardless of how the values $\alpha'(x)$ are chosen, the best selections for these conditional priors must have first moments matching those of the corresponding clarivoyant predictive distrbutions $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ for

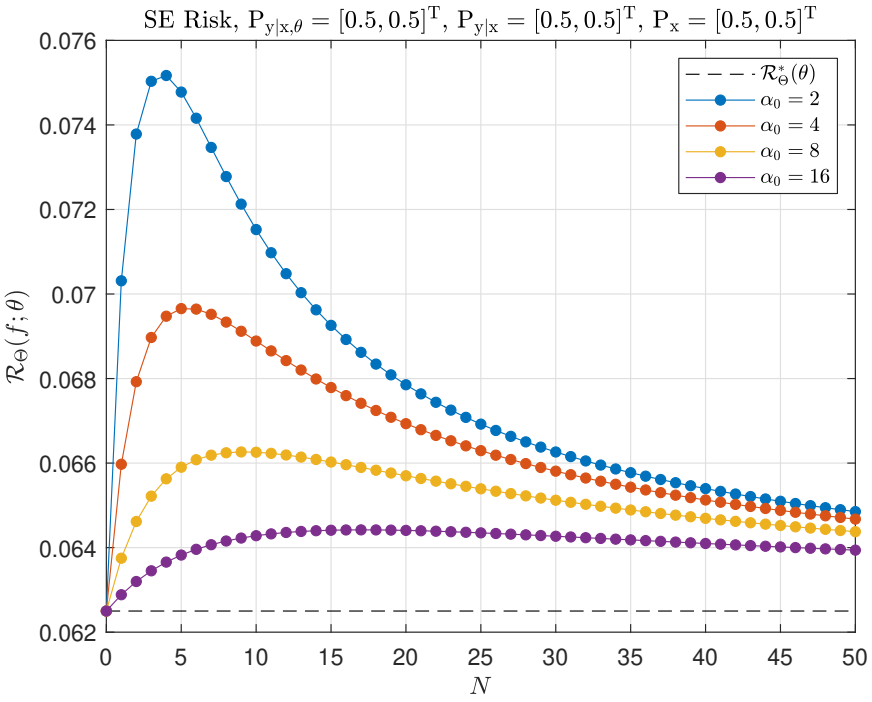

Figure 3.19: Conditional SE Risk versus N, unbiased Dirichlet estimators of varying concentration

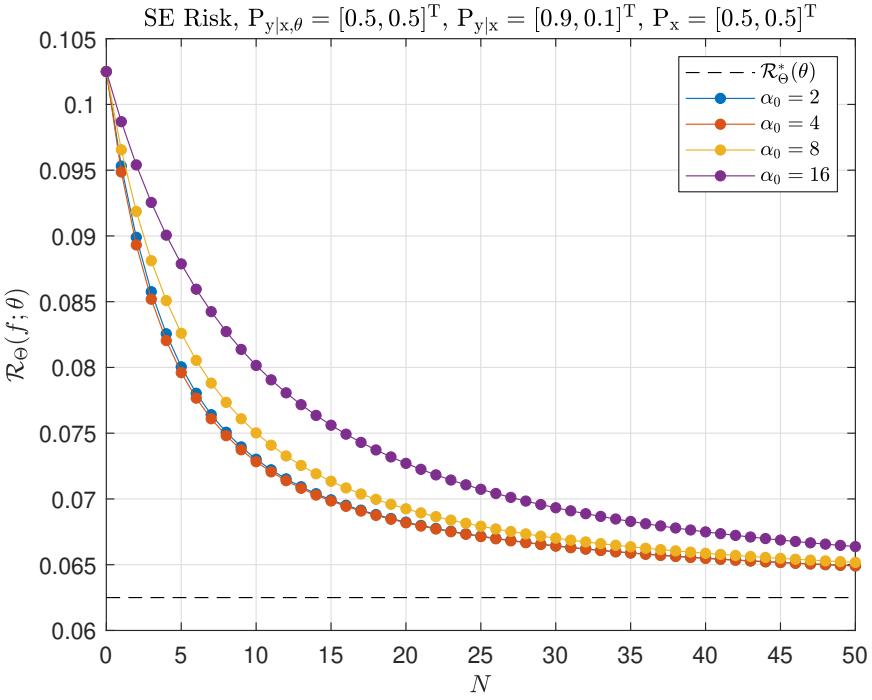

Figure 3.20: Conditional SE Risk versus N, biased Dirichlet estimators of varying concentration

each $x \in \mathcal{X}$. Such estimators are unbiased; as a result, the excess conditional risk is thus equivalent to the first term in (3.51), measuring additional variance due to model uncertainty.

The other user-selected Dirichlet parameters $\alpha'(x)$ are the concentration parameters for the corresponding conditional distributions; they control important bias/variance trade-offs via the two scaling factors in (3.51). First, consider the asymptotic trends.

Consider how the excess risk tends as the priors become maximally concentrated. As the parameters $\alpha'(x) \to \infty$, 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 conditional squared-error between the means of the Bayesian predicitve PMF and the clairvoyant predictive PMF. This is intuitive given that the estimator tends toward a data-independent solution; analagous to the discussion in Section 3.2, the estimator may be biased, but will have no variance due to the training data statistics.

Conversely, if concentrations $\alpha'(x) \to 0$ are chosen, the Bayesian estimate tends to the empirical mean, independent of $\alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$, and the excess risk tends to

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

Note that the first term's scaling factor is proportionate to the first inverse moment of a positive binomial random variable [17]. The second term's scaling factor tends to $\mathrm{P}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left(0|\theta'(x)\right)$, the probability that no training samples are observed matching the value x. As N increases, this term tends to zero, the risk due to the prior estimate bias decreases, and the excess risk becomes a function of θ only.

Of further interest are the values $\alpha'(x)$ that minimize the excesss squared-error for a given prior conditional distribution $\alpha(\cdot,x)/\alpha'(x)$. With the asymptotic values of the excess risk known, all that remains is to determine any local minima. Since the excess risk is a sum of $|\,\mathcal{X}\,|$ terms of identical form, each dependent on their own concentration $\alpha'(x)$, only one component needs to be minimized.

PGR: add the derivative details below???

Calculating the first derivative with respect to $\alpha'(x)$, it can be shown that for $N > 0$ and

$\theta'(x) > 0$, only one stationary point exists, at

$$\alpha'(x) \equiv \frac{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2} \,. \tag{3.54}$$

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

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\frac{1}{\mathrm{n}'(\mathrm{x})\Sigma^{-1}_{\mathrm{y}\,|\,\mathrm{x},\theta} + \left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^{-2}}\right]\right] \,, \tag{3.55}$$

which can be easily shown to be less than both the asymptotic values for $\alpha'(x) \to 0$ and $\alpha'(x) \to \infty$. Thus the concentation values (3.54) provide the minimum excess risk for the given prior conditional distributions.

Note that the minimizing concentration values $\alpha'(x)$ are inversely proportional to the squared-bias of the prior conditional mean. This is sensible; the better the match between the true and prior predictive distributions, the more confidence should be expressed. Also, low concentrations are preferable when the model has low conditional variance; these models can be quickly identified with learners prioritizing the emprical PMF estimate over the prior estimate. Additionally, note that these values $\alpha'(x)$ do not depend on the training volume N.

Figures 3.21 and 3.22 show how the excess conditional 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???

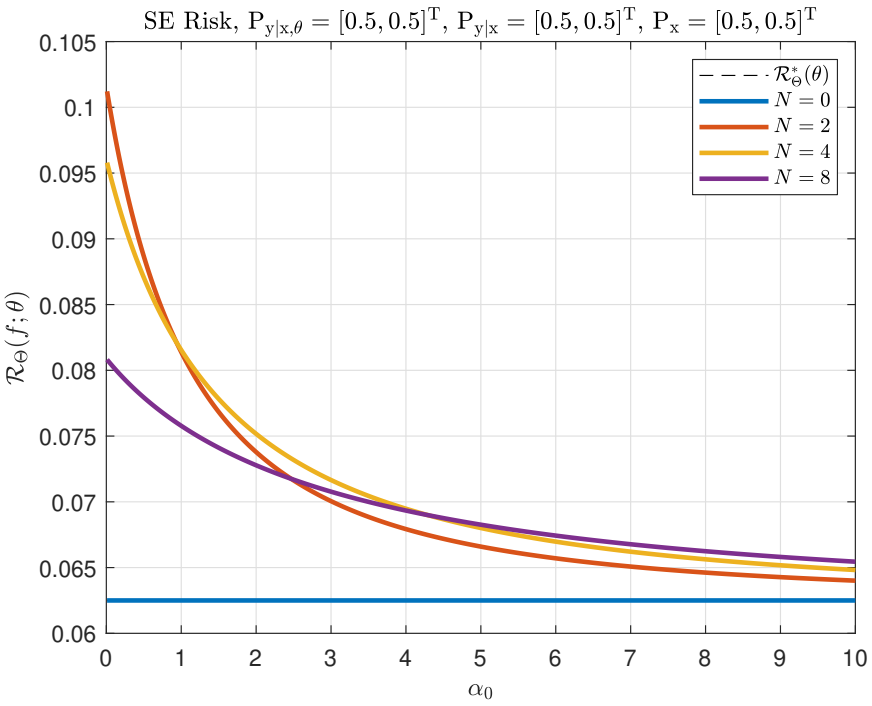

Figure 3.21: Conditional SE Risk versus $\alpha'(x)$, unbiased Dirichlet estimator using varying training set volumes

PGR: newpage

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 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 (2.34) is substi-

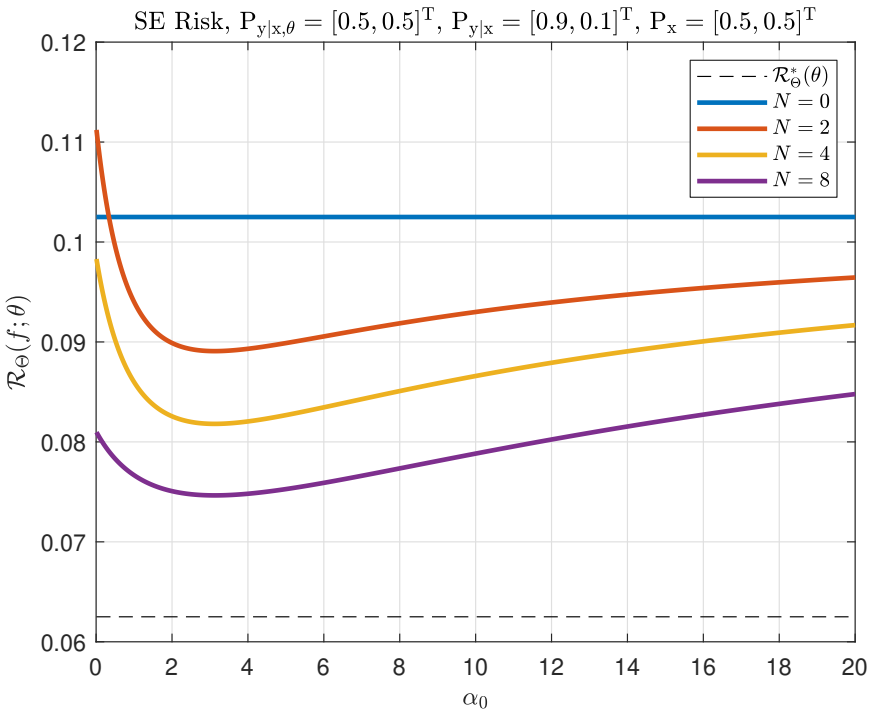

Figure 3.22: Conditional SE Risk versus $\alpha'(x)$, biased Dirichlet estimator using varying training set volumes

tuted into Equation (3.41) and Equation (2.8) to find

$$\begin{aligned} f^*(x;D) &= \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|x,D) \\ &= \arg\max_{y\in\mathcal{Y}} \frac{\alpha(y,x) + \bar{N}(y,x;D)}{\alpha'(x) + N'(x;D)} \\ &= \arg\max_{y\in\mathcal{Y}} \left(\alpha(y,x) + \bar{N}(y,x;D)\right) \ . \end{aligned} \tag{3.56}$$

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 $\alpha(\cdot, \mathrm{x})$.

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

$$f^*(x;D) = \arg\max_{y\in\mathcal{Y}} \bar{N}(y,x;D) \ , \tag{3.57}$$

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.

3.3.2.2 Minimum Risk: Probability of Error

PGR: GENERATE NON-SIM FIGS!!!!!!!

PGR: DIR SIM FIGS COMMENTED!!!

PGR: no closed-forms found???

Evaluating the minimum risk (2.40) using the distributions derived from the Dirichlet prior, the Bayes minimum probability of error is

$$\begin{aligned}\mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \\ &= 1 - \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\frac{\max_{y\in\mathcal{Y}}\left(\alpha(y,x)+\bar{\mathrm{n}}(y,x)\right)}{\alpha'(x)+\mathrm{n}'(x)}\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \frac{\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}}\left(\alpha(y,x)+\bar{\mathrm{n}}(y,x)\right)\right]}{\alpha_0+N}\,.\end{aligned} \tag{3.58}$$

Figures 3.23 and 3.24 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 $\mathcal{R}^* = 1 - \sum_{x\in\mathcal{X}} \frac{\max_{y\in\mathcal{Y}}\alpha(y,x)}{\alpha_0}$. 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 $\mathcal{R}^* \to 1 - \sum_{x\in\mathcal{X}} \frac{\max_{y\in\mathcal{Y}}\alpha(y,x)}{\alpha_0}$. These trends can be visualized in Figures 3.25 and 3.26.

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???

PGR: use Mcal not binom!

PGR: add nmax CDF fig!

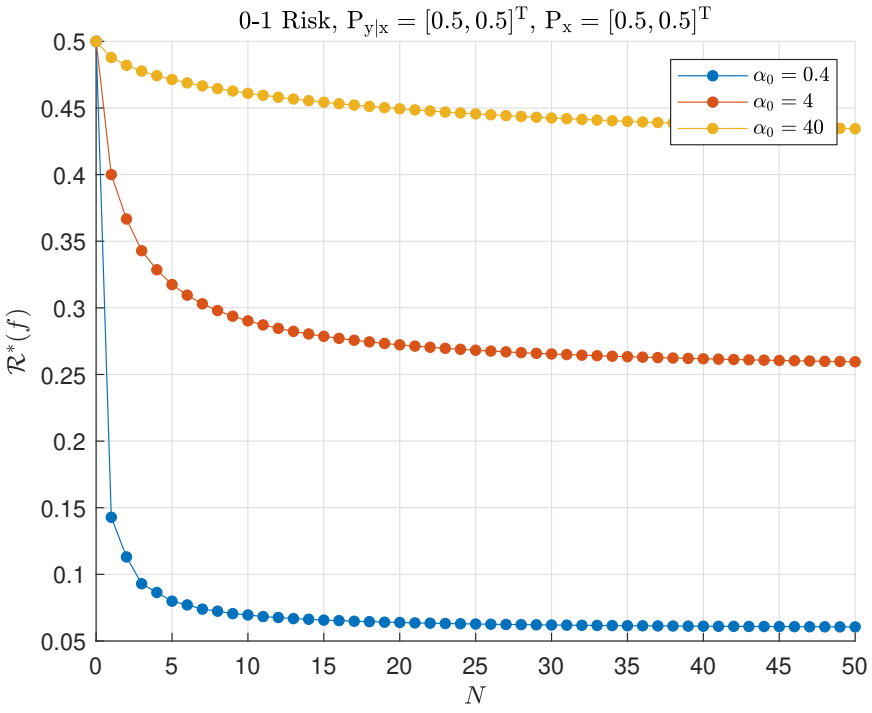

Figure 3.23: Minimum 0-1 Risk for different training data volumes N

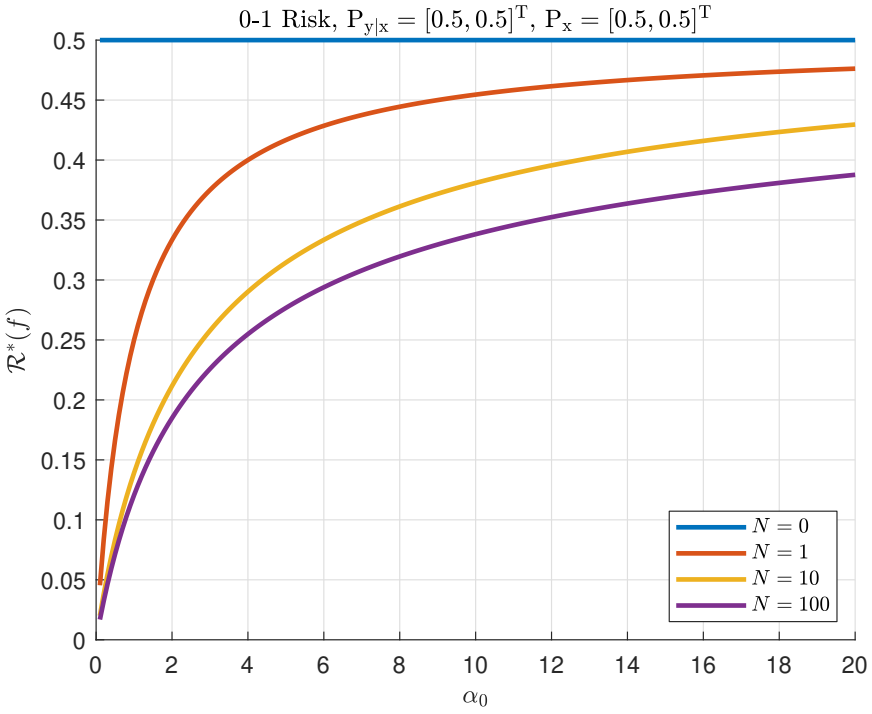

Figure 3.24: Minimum 0-1 Risk for different prior concentrations α_0

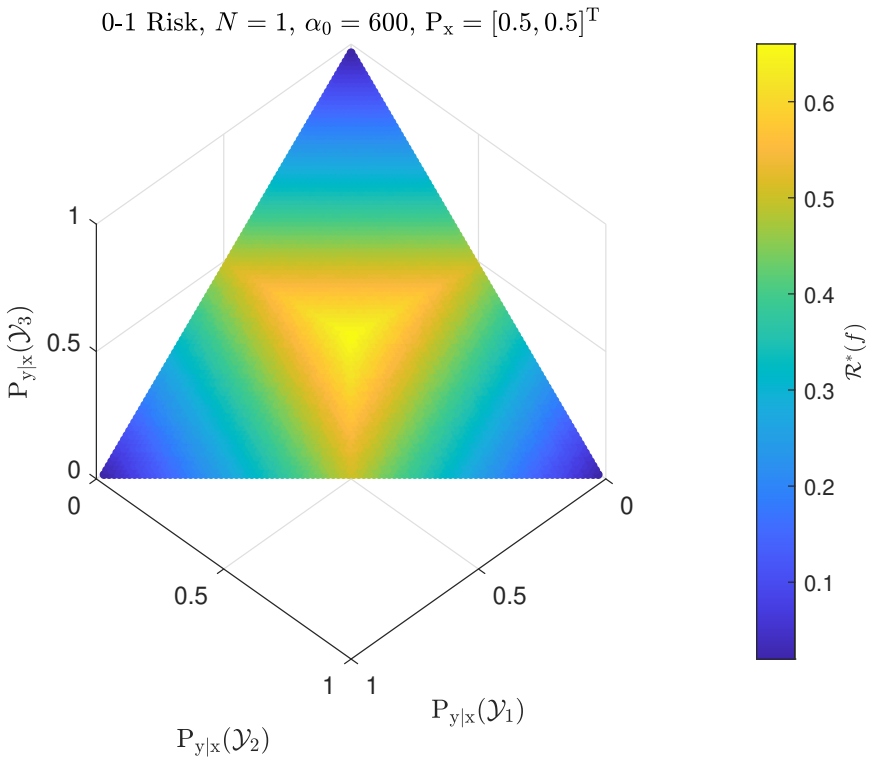

Figure 3.25: Minimum 0-1 Risk for different prior means $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$

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

$$\begin{aligned}\mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \frac{\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)\right] + 1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,| + N} \\ &= 1 - \frac{1 + |\,\mathcal{X}\,|^{-1}\sum_{x\in\mathcal{X}} \mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)\right]}{|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|}\,.\end{aligned} \tag{3.59}$$

The expectation operates on the maximum value from a subset of a uniform Dirichlet-Multinomial random process. Via the Dirichlet-Multinomial aggregation property [9], a consequence of the the uniform PMF $\mathrm{P}_{\bar{\mathrm{n}}}$ is that the individual segments $\bar{\mathrm{n}}(\cdot, x)$ are identically distributed; thus, the expectation will be same for every value x.

To evaluate this expectation, new random variables $\bar{\mathrm{n}}_{\max}(x) \equiv \max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)$ are introduced and characterized by their identical PMF. To this end, the probability of the event $\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) = \mathrm{P}\left(\cup_{y\in\mathcal{Y}}\{\bar{\mathrm{n}}(y,x) \geq n\}\right)$ will be determined. As the distribution of $\bar{\mathrm{n}}$ is uniform, the event probability is proportionate to the cardinality of the set

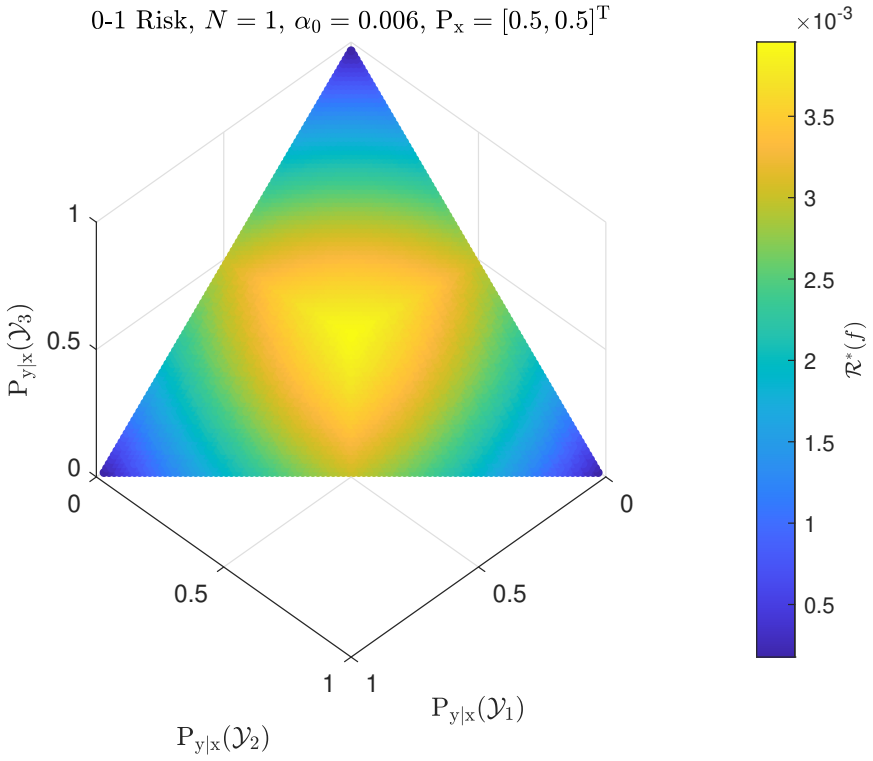

Figure 3.26: Minimum 0-1 Risk for different prior means $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$

$\cup_{y\in\mathcal{Y}}\{\bar{n} : \bar{n}(y,x) \geq n\}$. Using the inclusion-exclusion principle [5], the cardinality is represented as

$$\begin{aligned} &\left|\cup_{y\in\mathcal{Y}}\left\{\bar{n} : \bar{n}(y,x) \geq n\right\}\right| \qquad (3.60)\\ &= \begin{cases} \binom{N+|\mathcal{Y}||\mathcal{X}|-1}{|\mathcal{Y}||\mathcal{X}|-1} & \text{if } n < 0, \\ \sum_{m=1}^{|\mathcal{Y}|}\binom{|\mathcal{Y}|}{m}(-1)^{m-1}\binom{N-mn+|\mathcal{Y}||\mathcal{X}|-1}{|\mathcal{Y}||\mathcal{X}|-1}H\left(\left\lfloor\frac{N}{m}\right\rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N, \end{cases} \end{aligned}$$

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

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 $\{\bar{n} : \bar{n}(y,x) \geq n\}$. 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 $n > \left\lfloor \frac{N}{m} \right\rfloor$, 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 $|\bar{\mathcal{N}}|$ - the only difference is the number of points characterizing the $|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1$ dimensional region.

The probability of interest can thus be expressed as

$$\begin{aligned} \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) &= \binom{N + |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1}^{-1} \left| \cup_{y \in \mathcal{Y}} \{\bar{n} : \bar{n}(y, x) \geq n\} \right| \qquad (3.61) \\ &= \begin{cases} 1 & \text{if } n < 0, \\ \sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right) H\left(\left\lfloor \frac{N}{m} \right\rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N. \end{cases} \end{aligned}$$

PGR: use Mcal op?

PGR: Heaviside reference?

As the PMF of $\bar{\mathrm{n}}_{\max}(x)$ has support on $n \in [0, \ldots, N]$, the expectation over $\bar{\mathrm{n}}$ is evaluated as

$$\begin{aligned} \mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}_{\max}(x)\right] &= \sum_{n=0}^{N} n \Big(\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) - \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n+1\right) \Big) \qquad (3.62) \\ &= -1 + \sum_{n=0}^{N} \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) \\ &= -1 + \sum_{m=1}^{|\mathcal{Y}|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right) \end{aligned}$$

and the minimum 0-1 risk is

$$\mathcal{R}^* = 1 - \frac{\sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right)}{|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|} \,. \qquad (3.63)$$

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

$$\sum_{m=0}^{M} \binom{M}{m} (-1)^m g(m) = 0 \,, \qquad (3.64)$$

where g is a polynomial function of degree less than M [8], it can be shown that for $N = 0$, the minimum 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 irreducible risk explicitly from theta?

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

$$\begin{aligned}
&\lim_{N\to\infty} \left(|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|\right)^{-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \left(1 - \frac{mn}{N+l}\right) \qquad (3.65)\\
&\quad = \lim_{N/m\to\infty} \frac{|\,\mathcal{X}\,|}{m} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \left(1 - \frac{mn}{N}\right)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \frac{m}{N}\\
&\quad = \frac{|\,\mathcal{X}\,|}{m} \int_0^1 (1-t)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \mathrm{d}t\\
&\quad = \frac{1}{m|\,\mathcal{Y}\,|}\,.
\end{aligned}$$

The irreducible 0-1 risk for the uniform prior tends toward

$$\begin{aligned}
\mathcal{R}^* &\to 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\,\mathcal{Y}\,|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} m^{-1} \qquad (3.66)\\
&= 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\,\mathcal{Y}\,|} m^{-1}\,,
\end{aligned}$$

providing a lower bound for the achievable 0-1 Bayes risk. The above formulation has made use of the alternating summation identity from [15] 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 irreducible risk does not depend on the cardinality $|\,\mathcal{X}\,|$.

PGR: harmonic reference?

Figure 3.27 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 is sensible that the probability of error should increase when more classes have to be considered. Figure 3.28 illustrates the minimum risk with multiple sequences for different cardinalities $|\,\mathcal{X}\,|$. Note that risk increases with $|\,\mathcal{X}\,|$. Considering $\mathrm{E_D}\left[N'(\mathrm{D})\right] = \mu_{\mathrm{n}'} = N/|\,\mathcal{X}\,|$, this should be intuitive - each conditional empirical distribution $\bar{N}(\cdot, x; D)/N'(x; D)$ is forced to approximate $\tilde{\theta}(x)$ with less 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 Figure 3.29, it is shown that the minimal risk can be approximated

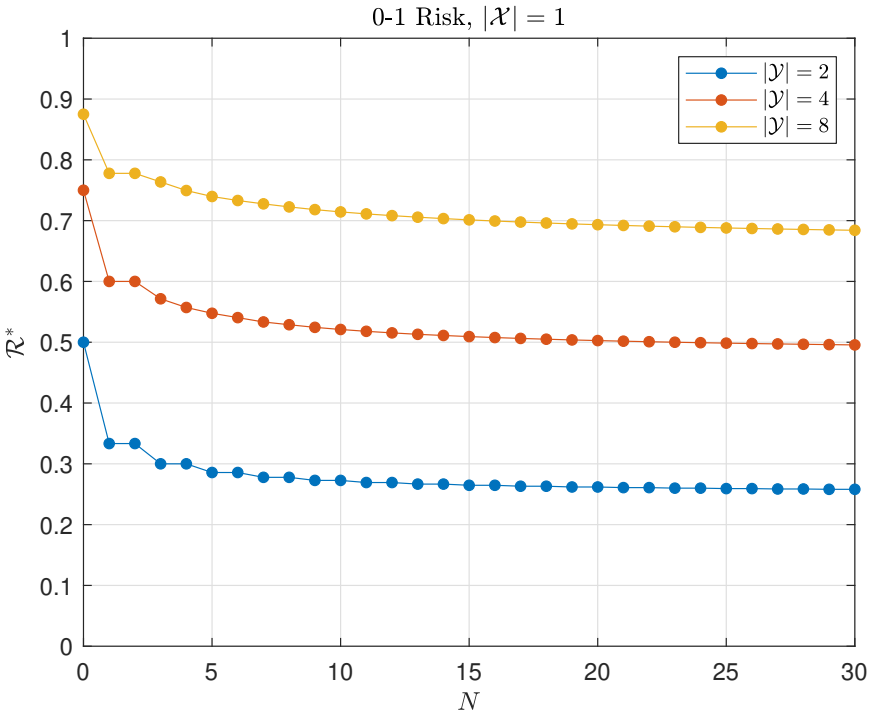

Figure 3.27: Minimum 0-1 Risk vs training set volume N

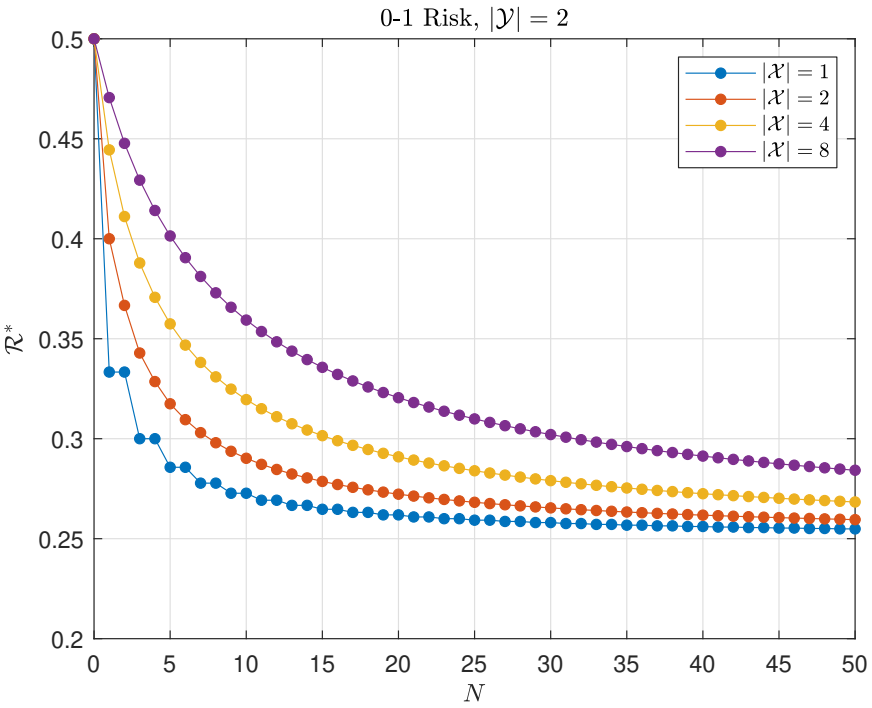

Figure 3.28: Minimum 0-1 Risk vs training set volume N

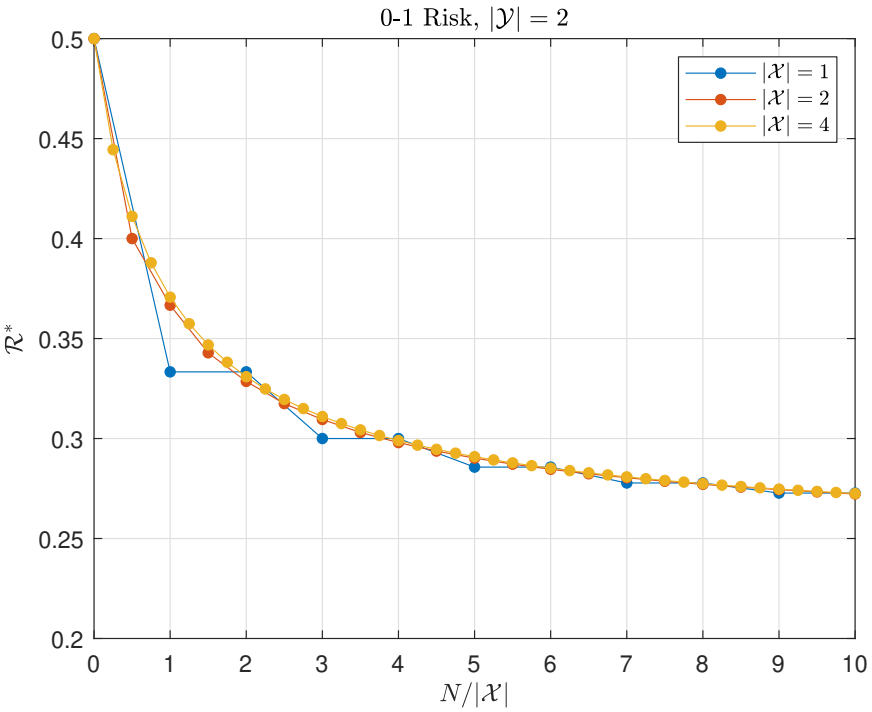

Figure 3.29: Minimum 0-1 Risk vs $N/|\mathcal{X}|$

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 $N \to \infty$ minimum risk as a function of $|\mathcal{Y}|$; both formulas are independent of $|\mathcal{X}|$. Figure 3.30 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 unity and the error reduction for $N \to \infty$ decreases.

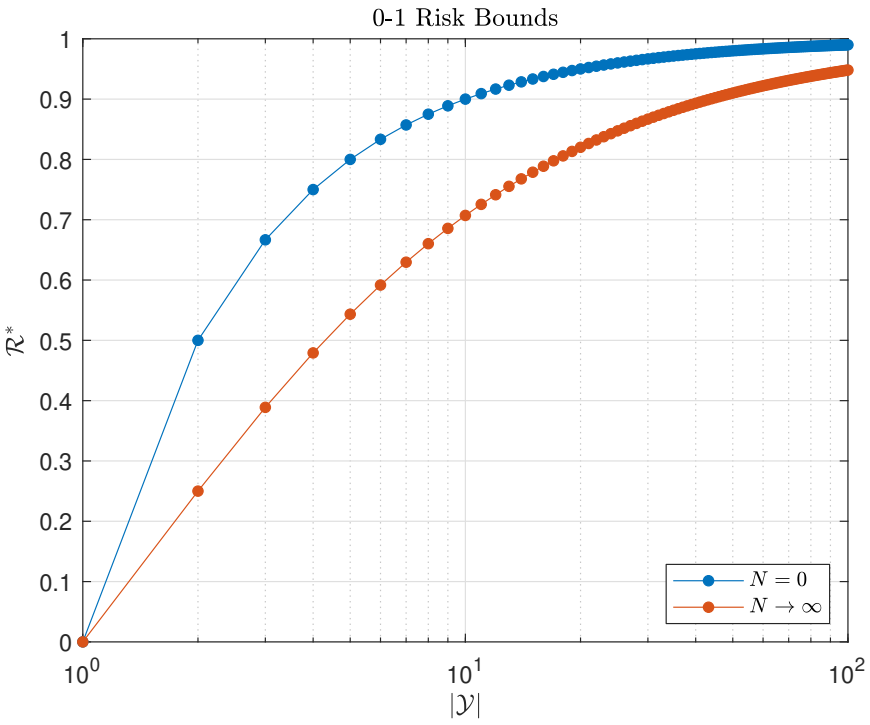

Figure 3.30: Minimum 0-1 Risk vs $|\mathcal{Y}|$

PGR: newpage

3.3.2.3 Conditional Probability of Error for a Dirichlet-based Classifier

PGR: INCOMPLETE

PGR: comment on alpha0 versus alphax simplification

Substituting the optimal Dirichlet-based classifier into the formula for the conditional probability of error 2.35, the risk is

$$\mathcal{R}_\Theta(f;\mathsf{\theta}) = 1 - \sum_{x\in\mathcal{X}} \mathsf{\theta}'(x)\, \mathrm{E}_{\bar{\mathsf{n}}\,|\mathsf{\theta}} \left[\tilde{\mathsf{\theta}}\Big(\arg\max_{y\in\mathcal{Y}} \big(\bar{\mathsf{n}}(y,x) + \alpha(y,x) \big); x \Big) \right] . \tag{3.67}$$

Figures 3.31 and 3.32 show how the conditional 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 $\tilde{\theta}(x)$ and the prior mean $\alpha(\cdot,x)/\alpha'(x)$ are the same.

Also, it is important to consider how a given classifier performs for varying models θ.

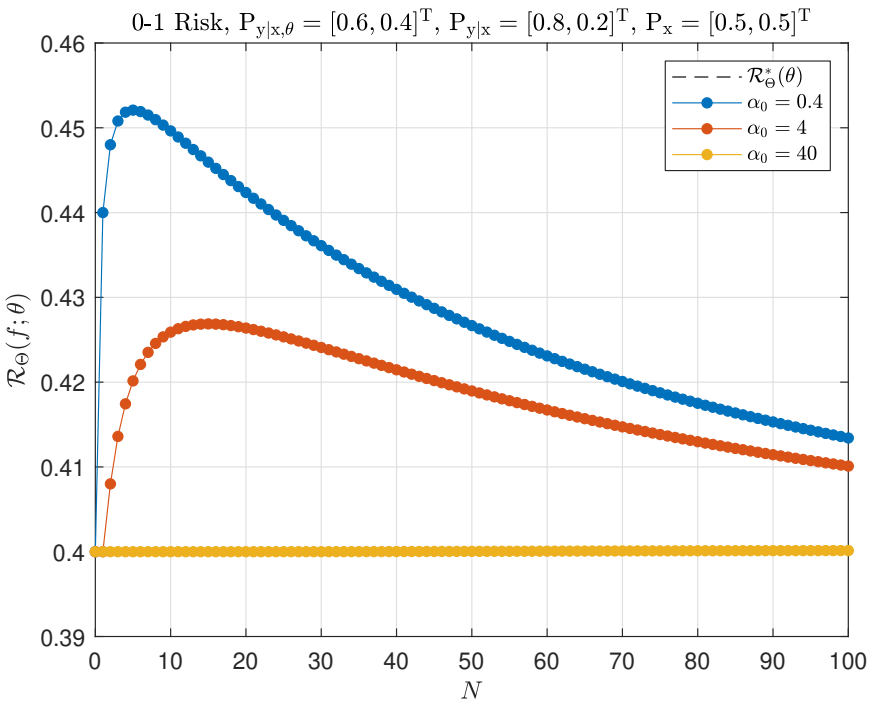

Figure 3.31: Excess conditional probability of error, well-matched informative Dirichlet-based classifier

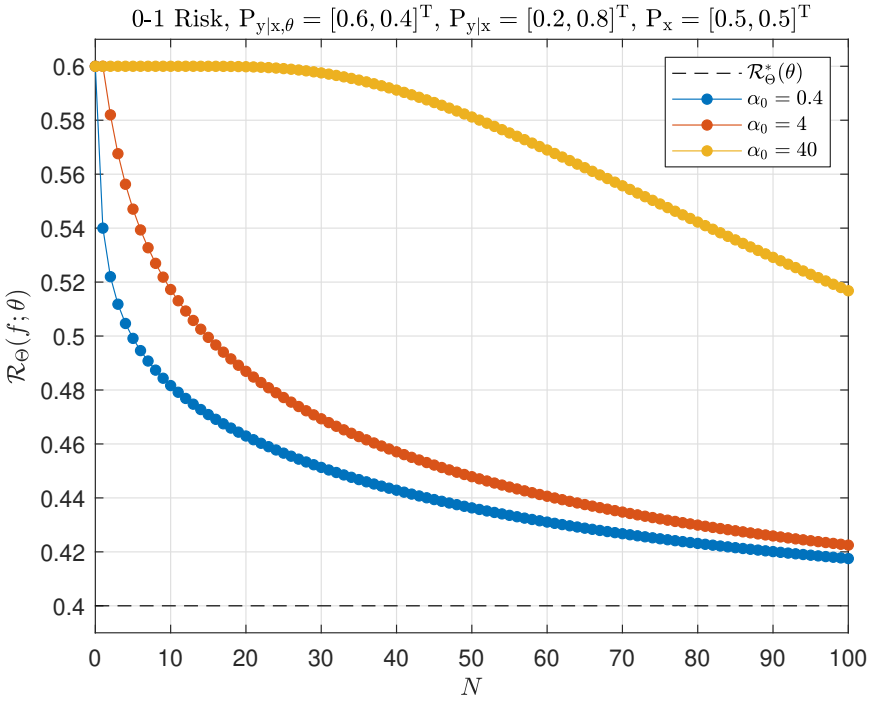

Figure 3.32: Excess conditional probability of error, poorly-matched informative Dirichlet-based classifier

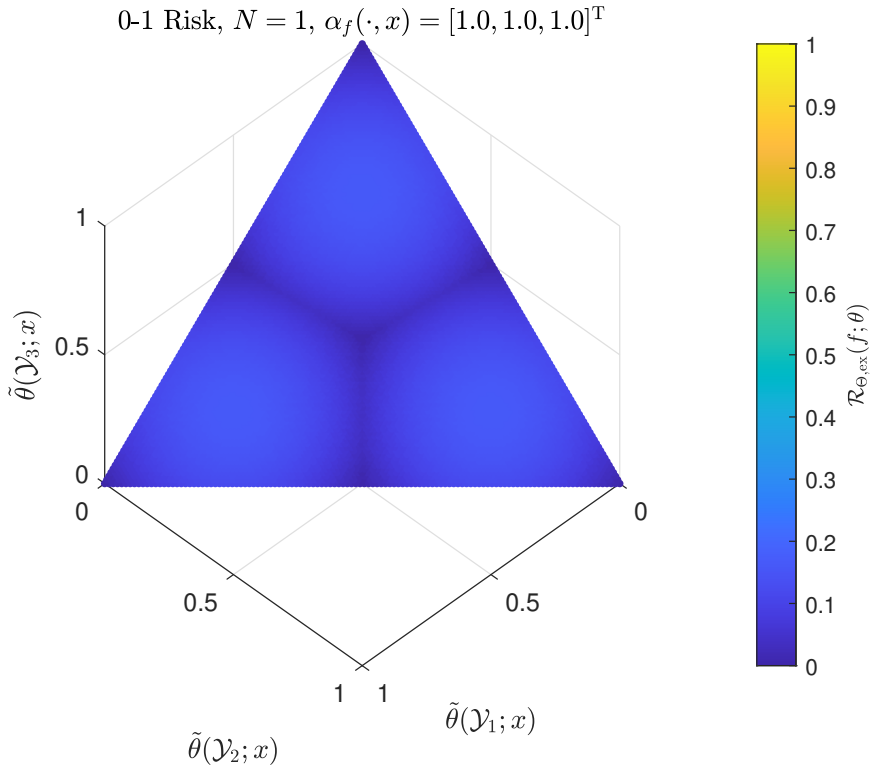

Figure 3.33: Excess conditional probability of error, conditional majority decision

Figures 3.33 and 3.34 demonstrate the excess conditional 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 clairvoyant risk $\mathcal{R}^*_\Theta(\theta)$ is achieved. This a fundamental trade-off between Bayesian learners based on non-informative versus informative priors.

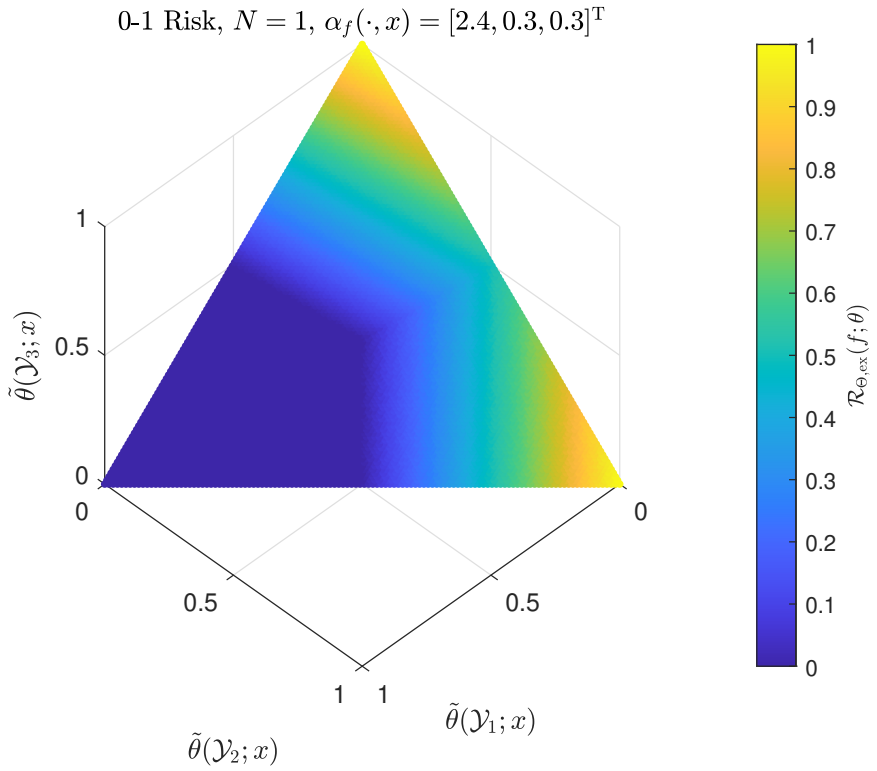

Figure 3.34: Excess conditional probability of error, informative Dirichlet-based classifier

Chapter 4

Extention to Infinite-Dimensional Spaces - Countably Infinite

PGR: MERGE with finite chapter???

4.1 Intro

This chapter extends previous results for applications where the space $\mathcal{Y}$ is countably infinite, that is $|\mathcal{Y}| = \aleph_0$. Specifically, the model prior distribution will be characterized by a discrete-domain Dirichlet process.

4.2 Basic Model

4.2.1 Probability Distributions

PGR: ???

4.2.1.1 Model PDF, $p(\theta)$

PGR: Valid model representation? Marginals instead?

$$\mathrm{p}(\theta) = \beta(\alpha)^{-1} \prod_{y \in \mathcal{Y}} \theta(y)^{\alpha(y)-1} \ , \tag{4.1}$$

$$\beta(\alpha) = \frac{\prod_{y \in \mathcal{Y}} \Gamma\big(\alpha(y)\big)}{\Gamma\left(\sum_{y \in \mathcal{Y}} \alpha(y)\right)} \ . \tag{4.2}$$

The first and second joint moments of the model are

$$\mu_\theta(y) = \mathrm{E}_\theta\left[\theta(y)\right] = \frac{\alpha(y)}{\alpha_0} \tag{4.3}$$

and

$$\mathrm{E}_\theta\left[\theta(y)\theta(y')\right] = \frac{\alpha(y)\alpha(y') + \alpha(y)\delta[y,y']}{\alpha_0(\alpha_0+1)} \ . \tag{4.4}$$

4.2.1.2 Training Data PMF, P(D)

$$\mathrm{P}(\mathrm{D}\,|\theta) = \prod_{y \in \mathcal{Y}} \theta(y)^{\bar{N}(y;\mathrm{D})} \ . \tag{4.5}$$

$$\mathrm{P}(\bar{\mathrm{n}}\,|\theta) = \mathcal{M}(\bar{\mathrm{n}}) \prod_{y \in \mathcal{Y}} \theta(y)^{\bar{\mathrm{n}}(y)} \ , \tag{4.6}$$

Thus,

$$\mathrm{P}(\bar{\mathrm{n}}) = \mathcal{M}(\bar{\mathrm{n}})\beta(\alpha)^{-1}\beta(\alpha+\bar{\mathrm{n}}) \ . \tag{4.7}$$

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

$$\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y)\right] = N\frac{\alpha(y)}{\alpha_0} \tag{4.8}$$

and

$$\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y)\bar{\mathrm{n}}(y')\right] = \frac{N}{\alpha_0(\alpha_0+1)}\big((\alpha_0+N)\alpha(y)\delta[y,y'] + (N-1)\alpha(y)\alpha(y')\big) \ . \tag{4.9}$$

Also,

$$\mathrm{P}(\mathrm{D}) = \beta(\alpha)^{-1}\beta\big(\alpha + \bar{N}(\mathrm{D})\big) \ . \tag{4.10}$$

4.2.1.3 Output conditional PMF, $\mathrm{P}(\mathrm{y}\,|\,\mathrm{D})$

$$\begin{aligned}\mathrm{p}(\theta|\,\mathrm{D}) &= \frac{\mathrm{P}(\mathrm{D}\,|\theta)\,\mathrm{p}(\theta)}{\mathrm{P}(\mathrm{D})} \\ &= \beta\left(\alpha + \bar{N}(\mathrm{D})\right)^{-1}\prod_{y\in\mathcal{Y}}\theta(y)^{\alpha(y)+\bar{N}(y;\mathrm{D})-1}\end{aligned} \tag{4.11}$$

$$\mathrm{p}(\theta|\,\bar{\mathrm{n}}) = \beta\left(\alpha + \bar{\mathrm{n}}\right)^{-1}\prod_{y\in\mathcal{Y}}\theta(y)^{\alpha(y)+\bar{\mathrm{n}}(y)-1}\ , \tag{4.12}$$

The PMF of interest is

$$\begin{aligned}\mathrm{P}(\mathrm{y}\,|\,\mathrm{D}) &= \mathrm{E}_{\theta|\,\mathrm{D}}\left[\theta(\mathrm{y})\right] \\ &= \frac{\alpha(\mathrm{y}) + \bar{N}(\mathrm{y};\mathrm{D})}{\alpha_0 + N} \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\frac{\alpha(\mathrm{y})}{\alpha_0} + \left(\frac{N}{\alpha_0+N}\right)\frac{\bar{N}(\mathrm{y};\mathrm{D})}{N}\end{aligned} \tag{4.13}$$

$$\begin{aligned}\mathrm{P}(\mathrm{y}\,|\,\bar{\mathrm{n}}) &= \mathrm{E}_{\theta|\,\bar{\mathrm{n}}}\left[\theta(\mathrm{y})|\,\bar{\mathrm{n}}\right] \\ &= \frac{\alpha(\mathrm{y}) + \bar{\mathrm{n}}(\mathrm{y})}{\alpha_0 + N} \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\frac{\alpha(\mathrm{y})}{\alpha_0} + \left(\frac{N}{\alpha_0+N}\right)\frac{\bar{\mathrm{n}}(\mathrm{y})}{N}\end{aligned} \tag{4.14}$$

4.3 Application to Common Loss Functions

PGR: Definitely regression. Classification sensible for countably infinite?

PGR: Results identical to finite Dirichlet

$$\begin{aligned}&\mathrm{E}_{\mathrm{y}\,|\,\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] = \sum_{y\in\mathcal{Y}}\mathcal{L}(h,y)\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{D}}(y|\,\mathrm{D}) \\ &= \frac{\sum_{y\in\mathcal{Y}}\alpha(y)\,\mathcal{L}(h,y) + \sum_{y\in\mathcal{Y}}\bar{N}(y;\mathrm{D})\,\mathcal{L}(h,y)}{\alpha_0+N} \\ &= \frac{\sum_{y\in\mathcal{Y}}\alpha(y)\,\mathcal{L}(h,y) + \sum_{n=1}^{N}\mathcal{L}\left(h,\mathrm{D}_n\right)}{\alpha_0+N} \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\sum_{y\in\mathcal{Y}}\mathcal{L}(h,y)\frac{\alpha(y)}{\alpha_0} + \left(\frac{N}{\alpha_0+N}\right)N^{-1}\sum_{n=1}^{N}\mathcal{L}\left(h,\mathrm{D}_n\right)\ .\end{aligned} \tag{4.15}$$

4.4 General Model

Extension to output and input spaces $\mathcal{Y}$ and $\mathcal{X}$ can have an infinite number of elements.

4.5 Applications: General Model

PGR: Definitely regression. Classification sensible for countably infinite?

PGR: Results identical to finite Dirichlet

Chapter 5

Extention to Infinite-Dimensional Spaces - Uncountably Infinite

PGR: SPECIFY EUCLIDEAN/HILBERT??

PGR: account for impulsive alpha?

5.1 Intro

This chapter extends further to the case where $\mathcal{Y}$ is a Euclidean space and the model θ is a continuous-domain Dirichlet process.

5.2 Basic Model

5.2.1 Probability Distributions

PGR: ???

5.2.1.1 Model θ Characterization

The model is now a continuous-domain Dirichlet process $\theta \sim \mathrm{DP}(\alpha)$. The concentration parameter is $\alpha_0 \equiv \int_{y \in \mathcal{Y}} \alpha(y) \mathrm{d}y$. By definition, for any partition of the set $\mathcal{Y}$, $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$, we can generate a Dirichlet random process $\phi(z) \equiv \int_{\mathcal{S}(z)} \theta(y) \mathrm{d}y$ with parameterizing

function $\lambda(z) \equiv \int_{\mathcal{S}(z)} \alpha(y)\mathrm{d}y$. This is commonly referred to as the aggregation property. The PDF for the aggregation is thus

$$\mathrm{P}_{\phi}(\phi) = \beta(\lambda)^{-1} \prod_{z \in \mathcal{Z}} \phi(z)^{\lambda(z)-1} \,. \tag{5.1}$$

As detailed in Appendix A.4, the first and second moments of the Dirichlet process θ are

$$\mu_{\theta} = \frac{\alpha}{\alpha_0} \tag{5.2}$$

and

$$\mathrm{E}_{\theta}\left[\theta(y)\theta(y')\right] = \frac{\alpha(y)\alpha(y') + \alpha(y)\delta(y-y')}{\alpha_0(\alpha_0+1)} \,. \tag{5.3}$$

Again, $\mathrm{p}_{\mathrm{y}} = \mathrm{p}_{\mathrm{D}_n} = \mu_{\theta}$ for all n.

5.2.1.2 Output conditional PDF, $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{D}}$

In Appendix A.5, it was shown that if the model $\theta \sim \mathrm{DP}(\alpha)$ is a Dirichlet process, then the model conditioned on the training data D is also a Dirichlet process with parameterizing function $\alpha + \bar{N}(\mathrm{D})$, where $\bar{N}(y;\mathrm{D}) = \sum_{n=1}^{N} \delta\left(y - \mathrm{D}_n\right)$ generalizes for a continuous domain. Thus, the conditional PDF of interest can be formulated as

$$\begin{aligned} \mathrm{p}_{\mathrm{y}\,|\,\mathrm{D}} &= \mu_{\theta|\,\mathrm{D}} \\ &= \frac{\alpha + \bar{N}(\mathrm{D})}{\alpha_0 + N} = \frac{\alpha + \sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{D}_n\big)}{\alpha_0 + N} \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\frac{\alpha}{\alpha_0} + \left(\frac{N}{\alpha_0+N}\right)\frac{\sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{D}_n\big)}{N} \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\mathrm{p}_{\mathrm{y}} + \left(\frac{N}{\alpha_0+N}\right)\frac{\sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{D}_n\big)}{N} \,. \end{aligned} \tag{5.4}$$

With the generalization to a Euclidean space $\mathcal{Y}$, the training data dependent component of the PDF is formulated with Dirac delta functions.

5.2.1.3 Training Data PDF, $\mathrm{p}(\mathrm{D})$

To represent the training data distribution, note that the Dirichlet process conditional model also provides

$$\mathrm{p}_{\mathrm{D}_{n+1}\,|\,\mathrm{D}_n,\ldots,\mathrm{D}_1} = \frac{\alpha + \sum_{i=1}^{n} \delta\big(\,\cdot - \mathrm{D}_i\big)}{\alpha_0 + N} \tag{5.5}$$

and thus the complete PDF is

$$\begin{aligned} \mathrm{p_D}(D) &= \mathrm{p_{D_1}}\left(D_1\right) \prod_{n=2}^{N} \mathrm{p_{D_n\,|\,D_{n-1},\ldots,D_1}}\left(\mathrm{D}_n \,|\, \mathrm{D}_{n-1}, \ldots, \mathrm{D}_1\right) \\ &= \frac{\alpha\big(D_1\big)}{\alpha_0} \prod_{n=2}^{N} \frac{\alpha\big(D_n\big) + \sum_{i=1}^{n-1} \delta\big(D_n - D_i\big)}{\alpha_0 + n - 1} \,. \end{aligned} \tag{5.6}$$

Additionally, note that since $\mathrm{p_{D_n\,|\theta}} = \theta$ is independent of sample index n, the PDF does not vary when the input arguments are permuted. Furthermore, all marginal distributions of D will have the same form, regardless of which training samples D_n are used.

Using these properties, the first and second joint moments of D are found to be

$$\begin{aligned} \mu_{\mathrm{D}_n} &= \int_{\mathcal{Y}} y\, \mathrm{p_{D_n}}(y)\mathrm{d}y = \int_{\mathcal{Y}} y\, \mathrm{E}_\theta\left[\,\mathrm{P_{D_n\,|\theta}}(y)\right]\mathrm{d}y \\ &= \int_{\mathcal{Y}} y \mu_\theta(y)\mathrm{d}y \\ &= \int_{\mathcal{Y}} y \frac{\alpha(y)}{\alpha_0}\mathrm{d}y \equiv \mu_{\mathrm{y}} \end{aligned} \tag{5.7}$$

and

$$\begin{aligned} \mathrm{E_D}\left[\,\mathrm{D}_n^2\,\right] &= \int_{\mathcal{Y}} y^2\, \mathrm{p_{D_n}}(y)\mathrm{d}y = \int_{\mathcal{Y}} y^2\, \mathrm{E}_\theta\left[\,\mathrm{P_{D_n\,|\theta}}(y)\right]\mathrm{d}y \\ &= \int_{\mathcal{Y}} y^2 \mu_\theta(y)\mathrm{d}y \\ &= \int_{\mathcal{Y}} y^2 \frac{\alpha(y)}{\alpha_0}\mathrm{d}y = \mathrm{E_y}[\mathrm{y}^2] \,, \end{aligned} \tag{5.8}$$

$$\begin{aligned} \mathrm{E_D}\left[\,\mathrm{D}_n\,\mathrm{D}_{n'}\,\right] &= \int_{\mathcal{Y}}\int_{\mathcal{Y}} yy'\, \mathrm{p_{D_n,D_{n'}}}(y,y')\mathrm{d}y\mathrm{d}y' \\ &= \int_{\mathcal{Y}}\int_{\mathcal{Y}} yy'\, \mathrm{E}_\theta\left[\,\mathrm{p_{D_n\,|\theta}}(y)\,\mathrm{p_{D_{n'}\,|\theta}}(y')\right]\mathrm{d}y\mathrm{d}y' \\ &= \int_{\mathcal{Y}}\int_{\mathcal{Y}} yy'\, \mathrm{E}_\theta\left[\theta(y)\theta(y')\right]\mathrm{d}y\mathrm{d}y' \\ &= \int_{\mathcal{Y}}\int_{\mathcal{Y}} yy' \frac{\alpha(y)\alpha(y') + \alpha(y)\delta(y-y')}{\alpha_0(\alpha_0+1)}\mathrm{d}y\mathrm{d}y' \\ &= \frac{\alpha_0\mu_{\mathrm{y}}^2 + \mathrm{E_y}[\mathrm{y}^2]}{\alpha_0+1} \,. \end{aligned} \tag{5.9}$$

Combining,

$$\mathrm{E_D}\left[\,\mathrm{D}_n\,\mathrm{D}_{n'}\,\right] = \mathrm{E_y}[\mathrm{y}^2] - \big(1 - \delta[n,n']\big)\frac{\alpha_0}{\alpha_0+1}\Sigma_{\mathrm{y}} \,. \tag{5.10}$$

PGR: move proofs to appendix???

PGR PGR: Dirichlet-Multinomial Process perspective???

Define the Dirichlet-Multinomial process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$. The mean and correlation functions below are found in Appendix A.6. The mean function is

$$\mu_{\bar{\mathrm{n}}} = N\frac{\alpha}{\alpha_0} \tag{5.11}$$

and the correlation function is

$$\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y)\bar{\mathrm{n}}(y')\right] = \frac{N}{\alpha_0(\alpha_0+1)}\left[(N-1)\alpha(y)\alpha(y') + (\alpha_0+N)\alpha(y)\delta(y-y')\right]. \tag{5.12}$$

5.3 Application to Common Loss Functions

PGR: Discuss continuous notation

$$\begin{aligned}
&\mathrm{E}_{\mathrm{y}\,|\,\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] = \int_{\mathcal{Y}} \mathcal{L}(h,y)\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{D}}(y|\,\mathrm{D})\mathrm{d}y \qquad (5.13)\\
&= \frac{\int_{\mathcal{Y}} \alpha(y)\,\mathcal{L}(h,y)\mathrm{d}y + \int_{\mathcal{Y}} \sum_{n=1}^{N} \delta\big(y-\mathrm{D}_n\big)\,\mathcal{L}(h,y)\mathrm{d}y}{\alpha_0+N}\\
&= \frac{\int_{\mathcal{Y}} \alpha(y)\,\mathcal{L}(h,y)\mathrm{d}y + \sum_{n=1}^{N} \mathcal{L}\left(h,\mathrm{D}_n\right)}{\alpha_0+N}\\
&= \left(\frac{\alpha_0}{\alpha_0+N}\right)\int_{\mathcal{Y}} \frac{\alpha(y)}{\alpha_0}\,\mathcal{L}(h,y)\mathrm{d}y + \left(\frac{N}{\alpha_0+N}\right) N^{-1}\sum_{n=1}^{N} \mathcal{L}\left(h,\mathrm{D}_n\right)\\
&= \left(\frac{\alpha_0}{\alpha_0+N}\right)\mathrm{E}_{\mathrm{y}}\left[\,\mathcal{L}(h,\mathrm{y})\right] + \left(\frac{N}{\alpha_0+N}\right) N^{-1}\sum_{n=1}^{N} \mathcal{L}\left(h,\mathrm{D}_n\right)
\end{aligned}$$

5.3.1 Regression: the Squared-Error Loss

$$\mathcal{L}(h,y) = (h-y)^2\,. \tag{5.14}$$

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

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\theta}\left[\left(f(\mathrm{D})-\mathrm{y}\right)^2\right]\right]\right] \\ &= \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{y}\,|\theta}\left[(\mathrm{y}-\mu_{\mathrm{y}\,|\theta})^2\right]\right] + \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\left(f(\mathrm{D})-\mu_{\mathrm{y}\,|\theta}\right)^2\right]\right] \\ &= \mathrm{E}_{\theta}\left[\Sigma_{\mathrm{y}\,|\theta}\right] + \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\left(f(\mathrm{D})-\mu_{\mathrm{y}\,|\theta}\right)^2\right]\right]\end{aligned} \tag{5.15}$$

5.3.1.1 Optimal Learner

The optimal function is again the expected value of the output conditional PMF,

$$\begin{aligned}f^*(\mathrm{D}) &= \arg\min_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{D}}\left[(h-\mathrm{y})^2\right] \\ &= \mu_{\mathrm{y}\,|\,\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{D}}[\mu_{\mathrm{y}\,|\theta}] \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\int_{\mathcal{Y}}\frac{\alpha(y)}{\alpha_0}y\mathrm{d}y + \left(\frac{N}{\alpha_0+N}\right)\frac{1}{N}\sum_{n=1}^{N}\mathrm{D}_n \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\mu_{\mathrm{y}} + \left(\frac{N}{\alpha_0+N}\right)\frac{1}{N}\sum_{n=1}^{N}\mathrm{D}_n \\ &= \left(\frac{\alpha_0}{\alpha_0+N}\right)\mu_{\mathrm{y}} + \left(\frac{N}{\alpha_0+N}\right)\int_{\mathcal{Y}} y\frac{\bar{N}(y;\mathrm{D})}{N}\mathrm{d}y\ .\end{aligned} \tag{5.16}$$

5.3.1.2 Minimum Risk

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_{\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{D}}\right] \\ &= \mathrm{E}_{\theta}\left[\Sigma_{\mathrm{y}\,|\theta}\right] + \mathrm{E}_{\mathrm{D}}\left[\mathrm{E}_{\theta|\,\mathrm{D}}\left[\left(\mu_{\mathrm{y}\,|\theta}-\mathrm{E}_{\theta|\,\mathrm{D}}[\mu_{\mathrm{y}\,|\theta}]\right)^2\right]\right] \\ &= \mathrm{E}_{\theta}\left[\Sigma_{\mathrm{y}\,|\theta}\right] + \mathrm{E}_{\mathrm{D}}\left[\,\mathrm{C}_{\theta|\,\mathrm{D}}[\mu_{\mathrm{y}\,|\theta}]\right] \\ &= \mathrm{E}_{\bar{\mathrm{n}}}\left[\Sigma_{\mathrm{y}\,|\,\bar{\mathrm{n}}}\right]\ .\end{aligned} \tag{5.17}$$

The conditional variance is expanded as

$$\Sigma_{\mathrm{y}\,|\,\mathrm{D}} = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{D}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{D}} \tag{5.18}$$

and the expectations of the two terms are evaluated separately.

$$\mathrm{E}_{\mathrm{D}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{D}}[\mathrm{y}^2]\right] = \mathrm{E}_{\mathrm{y}}[\mathrm{y}^2] \tag{5.19}$$

$$\begin{aligned}
&\mathrm{E}_{\mathrm{D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{D}}\right] \\
&\quad= \frac{\alpha_0^2\mu_{\mathrm{y}}^2 + 2\alpha_0\mu_{\mathrm{y}}\sum_{n=1}^{N}\mu_{\mathrm{D}_n} + \sum_{n=1}^{N}\sum_{n'=1}^{N}\mathrm{E}_{\mathrm{D}}\left[\,\mathrm{D}_n\,\mathrm{D}_{n'}\,\right]}{(\alpha_0+N)^2} \\
&\quad= \frac{\alpha_0^2\mu_{\mathrm{y}}^2 + 2\alpha_0 N\mu_{\mathrm{y}}^2 + N^2\,\mathrm{E}_{\mathrm{y}}[\mathrm{y}^2] - N(N-1)\alpha_0(\alpha_0+1)^{-1}\Sigma_{\mathrm{y}}}{(\alpha_0+N)^2} \\
&\quad= \mu_{\mathrm{y}}^2 + \frac{N}{(\alpha_0+1)(\alpha_0+N)}\Sigma_{\mathrm{y}}
\end{aligned} \tag{5.20}$$

PGR: DMP PERSPECTIVE???

$$\begin{aligned}
&\mathrm{E}_{\mathrm{D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{D}}\right] = \mathrm{E}_{\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\bar{\mathrm{n}}}\right] \\
&\quad= \frac{\alpha_0^2\mu_{\mathrm{y}}^2 + 2\alpha_0\mu_{\mathrm{y}}\int_{\mathcal{Y}} y\mu_{\bar{\mathrm{n}}}(y)\big]\mathrm{d}y + \int_{\mathcal{Y}}\int_{\mathcal{Y}} yy'\,\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y)\bar{\mathrm{n}}(y')\right]\mathrm{d}y\mathrm{d}y'}{(\alpha_0+N)^2} \\
&\quad= \frac{\alpha_0^2\mu_{\mathrm{y}}^2 + 2\alpha_0 N\mu_{\mathrm{y}}^2 + N(N-1)\alpha_0(\alpha_0+1)^{-1}\mu_{\mathrm{y}}^2 + N(\alpha_0+N)(\alpha_0+1)^{-1}\,\mathrm{E}_{\mathrm{y}}[\mathrm{y}^2]}{(\alpha_0+N)^2} \\
&\quad= \frac{\alpha_0(\alpha_0+N+1)\mu_{\mathrm{y}}^2 + N\,\mathrm{E}_{\mathrm{y}}[\mathrm{y}^2]}{(\alpha_0+1)(\alpha_0+N)}
\end{aligned} \tag{5.21}$$

PGR: nicer algebra with DMP!

PGR: DMP

The minimal risk is again

$$\begin{aligned}
\mathcal{R}^* &= \left(1 - \frac{N}{(\alpha_0+1)(\alpha_0+N)}\right)\Sigma_{\mathrm{y}} \\
&= \frac{\alpha_0(\alpha_0+N+1)}{(\alpha_0+1)(\alpha_0+N)}\Sigma_{\mathrm{y}} \\
&= \frac{1+(\alpha_0+N)^{-1}}{1+\alpha_0^{-1}}\Sigma_{\mathrm{y}}\,.
\end{aligned} \tag{5.22}$$

As for Dirichlet distributions for functions with countable domains, the minimal risk is dependent on the model only through the concentration parameter α_0 and the variance of the expected distribution $\mathrm{P}_{\mathrm{y}} = \mu_{\theta}$.

5.4 General Model

PGR: change dirac deltas to kronecker in fractions, no divide by zero???

5.4.1 Model Extension

This section adds the input space $\mathcal{X}$, now considered to be uncountably infinite; that is $|\,\mathcal{X}\,| \geq \aleph_1$. The model distribution is a Dirichlet process over the space $\mathcal{Y} \times \mathcal{X}$.

5.4.2 General Probability Distributions

PGR

5.4.2.1 Model θ Characterization

The model is characterized by a Dirichlet process $\theta \sim \mathrm{DP}(\alpha)$ with parameter function $\alpha : \mathcal{Y} \times \mathcal{X} \mapsto \mathbb{R}^+$. The concentration parameter generalizes to $\alpha_0 = \int_{\mathcal{Y}} \int_{\mathcal{X}} \alpha(y, x) \mathrm{d}x\mathrm{d}y$. Using the aggregation property, any partition of the set $\mathcal{Y} \times \mathcal{X}$, $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$ is a Dirichlet random process $\phi(z) = \iint_{\mathcal{S}(z)} \theta(y, x) \mathrm{d}x\mathrm{d}y$ with parameterizing function $\lambda(z) = \iint_{\mathcal{S}(z)} \alpha(y, x) \mathrm{d}x\mathrm{d}y$. The PDF for the aggregation is

$$\mathrm{p}_\phi(\phi) = \beta(\lambda)^{-1} \prod_{z \in \mathcal{Z}} \phi(z)^{\lambda(z)-1} \,. \tag{5.23}$$

The expected value of a Dirichlet process generalizes to $\mathrm{DP}(\alpha)$

$$\mu_\theta = \frac{\alpha}{\alpha_0} \tag{5.24}$$

and the correlation function is

$$\mathrm{E}_\theta \left[\theta(y, x)\theta(y', x')\right] = \frac{\alpha(y, x)\alpha(y', x') + \alpha(y, x)\delta(y - y')\delta(x - x')}{\alpha_0(\alpha_0 + 1)} \,. \tag{5.25}$$

5.4.2.2 Output conditional PDF, $\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}$

The properties of the Dirichlet distribution proven in Appendix A.5 generalize, such that the model conditioned on the training data D is Dirichlet with parameterizing function $\alpha + \bar{N}(\mathrm{D})$, where $\bar{N}(y, x; \mathrm{D}) = \sum_{n=1}^{N} \delta\left(y - \mathrm{Y}_n\right)\left(x - \mathrm{X}_n\right)$. Recall that $\mathrm{D}_n = \left(\mathrm{Y}_n, \mathrm{X}_n\right)$ with $Y \in \mathcal{Y}^N$ and $X \in \mathcal{X}^N$.

The PDF $\mathrm{p}_{\mathrm{y,x\,|\,D}}$ is thus

$$\begin{aligned}\mathrm{p}_{\mathrm{y,x\,|\,D}} &= \mu_{\theta|\,\mathrm{D}} \\ &= \frac{\alpha + \bar{N}(\mathrm{D})}{\alpha_0 + N} \\ &= \frac{\alpha + \sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{Y}_n\,\big)\delta\big(\,\cdot - \mathrm{X}_n\,\big)}{\alpha_0 + N} \\ &= \left(\frac{\alpha_0}{\alpha_0 + N}\right)\frac{\alpha}{\alpha_0} + \left(\frac{N}{\alpha_0 + N}\right)\frac{\sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{Y}_n\,\big)\delta\big(\,\cdot - \mathrm{X}_n\,\big)}{N}\end{aligned} \tag{5.26}$$

and the conditional PDF of interest is

$$\begin{aligned}\mathrm{p}_{\mathrm{y\,|\,x,D}} &= \frac{\alpha(\cdot,\mathrm{x}) + \bar{N}(\cdot,\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})} \\ &= \frac{\alpha(\cdot,\mathrm{x}) + \sum_{n=1}^{N} \delta\big(\,\cdot - \mathrm{Y}_n\,\big)\delta\big(\,\mathrm{x} - \mathrm{X}_n\,\big)}{\alpha'(\mathrm{x}) + \sum_{n=1}^{N} \delta\big(\,\mathrm{x} - \mathrm{X}_n\,\big)} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\bar{N}(\cdot,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})}\,,\end{aligned} \tag{5.27}$$

where $\alpha'(x) = \int_{\mathcal{Y}} \alpha(y,x)\mathrm{d}y$ and $N'(x;\mathrm{D}) = \sum_{n=1}^{N} \delta\big(x - \mathrm{X}_n\,\big)$.

The conditional distribution when $\mathcal{X}$ is a Euclidean space has notable differences from its form for a countable set $\mathcal{X}$. Specifically, as $N'(x;\mathrm{D}) \in [0,\infty)$, and in fact will either be zero or tend towards infinity, the coefficients dictating the convex combination of distributions will be zero or one (assuming a non-impulsive model parameter α). Thus, the distribution for a given observation x will be either strictly dependent on either the training data or the prior knowledge regarding θ.

5.4.2.3 Training Data PDF, p_{D}

By the Dirichlet process properties,

$$\mathrm{p}_{\mathrm{D}_{n+1}\,|\,\mathrm{D}_n,\ldots,\mathrm{D}_1} = \frac{\alpha(\cdot,\cdot) + \sum_{n'=1}^{n} \delta\big(\,\cdot - \mathrm{Y}_i\,\big)\delta\big(\,\cdot - \mathrm{X}_i\,\big)}{\alpha_0 + N} \tag{5.28}$$

and thus

$$\begin{aligned}\mathrm{p}_{\mathrm{D}}(D) &= \mathrm{p}\left(D_1\right)\prod_{n=2}^{N}\mathrm{p}\left(D_n|D_{n-1},\ldots,D_1\right) \\ &= \frac{\alpha\big(Y_1,X_1\big)}{\alpha_0}\prod_{n=2}^{N}\frac{\alpha\big(Y_n,X_n\big) + \sum_{i=1}^{n-1}\delta\big(Y_n - Y_i\big)\delta\big(X_n - X_i\big)}{\alpha_0 + n - 1}\end{aligned} \tag{5.29}$$

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 the PDF for X can be represented as

$$\begin{aligned} \mathrm{p}_{\mathrm{X}}(X) &= \mathrm{E}_{\theta}\left[\mathrm{p}_{\mathrm{X}\,|\,\theta}(X|\theta)\right] = \mathrm{E}_{\theta}\left[\prod_{n=1}^{N}\theta'(X_n)\right] \\ &= \frac{\alpha'(X_1)}{\alpha_0}\prod_{n=2}^{N}\frac{\alpha'(X_n)+\sum_{i=1}^{n-1}\delta(X_n-X_i)}{\alpha_0+n-1} \end{aligned} \tag{5.30}$$

since, by the aggregation principle, $\theta' = \int_{\mathcal{Y}}\theta(y,\cdot)\mathrm{d}y$ is a Dirichlet process with parameter function $\alpha' : \mathcal{X} \mapsto \mathbb{R}^{+}$.

Additionally, by the invariance principle, the PDF's for the first-degree marginals are

$$\mathrm{p}_{\mathrm{X}_n} = \frac{\alpha'}{\alpha_0}\ . \tag{5.31}$$

which equivalent to p_{x}.

PGR: express conditional below using Dir aggregation conditional independence properties?

The conditional distribution of intererest is

$$\mathrm{p}_{\mathrm{Y}\,|\,\mathrm{X}}(Y|X) = \frac{\alpha(Y_1,X_1)}{\alpha'(X_1)}\prod_{n=2}^{N}\frac{\alpha(Y_n,X_n)+\sum_{i=1}^{n-1}\delta(Y_n-Y_i)\delta(X_n-X_i)}{\alpha'(X_n)+\sum_{i=1}^{n-1}\delta(X_n-X_i)} \tag{5.32}$$

Marginalized conditional PDF's for the first and second samples are found. Observe that the marginal distribution for the first $N-1$ values of Y is

$$\begin{aligned} &\mathrm{p}_{\mathrm{Y}_1,\ldots,\mathrm{Y}_{N-1}\,|\,\mathrm{X}}\left(Y_1,\ldots,Y_{N-1}|X\right) \\ &= \int_{\mathcal{Y}}\frac{\alpha(Y_1,X_1)}{\alpha'(X_1)}\prod_{n=2}^{N}\frac{\alpha(Y_n,X_n)+\sum_{i=1}^{n-1}\delta(Y_n-Y_i)\delta(X_n-X_i)}{\alpha'(X_n)+\sum_{i=1}^{n-1}\delta(X_n-X_i)}\mathrm{d}Y_N \\ &= \frac{\alpha(Y_1,X_1)}{\alpha'(X_1)}\prod_{n=2}^{N-1}\frac{\alpha(Y_n,X_n)+\sum_{i=1}^{n-1}\delta(Y_n-Y_i)\delta(X_n-X_i)}{\alpha'(X_n)+\sum_{i=1}^{n-1}\delta(X_n-X_i)} \end{aligned} \tag{5.33}$$

which is independent of X_N. Repeated integrations and an application of the permutation invariance principle can show that when conditioned on X any subset of training data values $\mathrm{Y}_1,\ldots,\mathrm{Y}_N$ will only be dependent on the corresponding values X_n. The first and second order conditional distributions are

$$\mathrm{P}_{\mathrm{Y}_n\,|\,\mathrm{X}_n}(y|x) = \frac{\alpha(y,x)}{\alpha'(x)} = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}(y|x) \tag{5.34}$$

and

$$\begin{aligned} &\mathrm{P}_{\mathrm{Y}_n,\mathrm{Y}_{n'}\,|\,\mathrm{X}_n,\mathrm{X}_{n'}}(y,y'|x,x') \\ &\quad = \frac{\alpha(y,x)\alpha(y',x') + \alpha(y,x)\delta(y-y')\delta(x-x')}{\alpha'(x)\alpha'(x') + \alpha'(x')\delta(x-x')} \end{aligned} \tag{5.35}$$

and the first and second order moments of interest are

$$\mu_{\mathrm{Y}_n\,|\,\mathrm{X}} = \mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\,\mathrm{X}_n\,\big)\;, \tag{5.36}$$

$$\mathrm{E}_{\mathrm{Y}_n\,|\,\mathrm{X}}\left[\,\mathrm{Y}_n^2\,\right] = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\big(\,\mathrm{X}_n\,\big)\;, \tag{5.37}$$

and

$$\begin{aligned} &\mathrm{E}_{\mathrm{Y}_n\,|\,\mathrm{X}}\left[\,\mathrm{Y}_n\,\mathrm{Y}_{n'}\,\right] \\ &\quad = \frac{\alpha'\big(\,\mathrm{X}_n\,\big)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\,\mathrm{X}_n\big)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\,\mathrm{X}_{n'}\,\big) + \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathrm{y}^2\,\right]\big(\,\mathrm{X}_n\big)\delta\big(\,\mathrm{X}_n - \mathrm{X}_{n'}\,\big)}{\alpha'\big(\,\mathrm{X}_n\,\big) + \delta\big(\,\mathrm{X}_n - \mathrm{X}_{n'}\,\big)}\;. \end{aligned} \tag{5.38}$$

PGR: formalize permutation invariance principle???

PGR: Add Y given X equations (with Betas) for discrete case in previous chapters?

PGR: Dirichlet-Multinomial Process perspective

We have the DMP $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$ with mean and correlation functions

$$\mu_{\bar{\mathrm{n}}} = N\frac{\alpha}{\alpha_0} \tag{5.39}$$

and

$$\begin{aligned} &\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y,x)\bar{\mathrm{n}}(y',x')\right] \\ &\quad = \frac{N}{\alpha_0(\alpha_0+1)}\big[(N-1)\alpha(y,x)\alpha(y',x') + (\alpha_0+N)\alpha(y,x)\delta(y-y')\delta(x-x')\big]\;. \end{aligned} \tag{5.40}$$

Observe that by the aggregation principle, $\mathrm{n}' = \int_{\mathcal{Y}} \bar{\mathrm{n}}(y,\cdot)\mathrm{d}y \equiv \sum_{n=1}^{N}\delta\big(\,\cdot - \mathrm{X}_n\,\big)$ is a DMP over the set $\mathcal{X}$ with parametrizing function $\alpha' : \mathcal{X} \mapsto \mathbb{R}^+$.

Additionally, the 1-dimensional subsets conditioned on the marginalized DMP are characterized as

$$\frac{\bar{\mathrm{n}}(\cdot,x)}{\delta(0)}\Big|\,\mathrm{n}'(x) \sim \mathrm{DMP}\left(\frac{\mathrm{n}'(x)}{\delta(0)}, \frac{\alpha(\cdot,x)}{\delta(0)}\right) \tag{5.41}$$

PGR: add proof???

5.5 Applications: General Model

PGR: COPIED, incomplete

$$
\begin{aligned}
&\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] = \int_{\mathcal{Y}} \mathcal{L}(h,y)\,\mathrm{p}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\mathrm{d}y \qquad (5.42)\\
&= \frac{\int_{\mathcal{Y}} \alpha(y,\mathrm{x})\,\mathcal{L}(h,y)\mathrm{d}y + \int_{\mathcal{Y}} \bar{N}(y,\mathrm{x};\mathrm{D})\,\mathcal{L}(h,y)\mathrm{d}y}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\\
&= \frac{\int_{\mathcal{Y}} \alpha(y,\mathrm{x})\,\mathcal{L}(h,y)\mathrm{d}y + \sum_{n=1}^{N} \delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\,\mathcal{L}\left(h,\mathrm{D}_n\right)}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\\
&= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\,\mathcal{L}(h,\mathrm{y})\right] + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)\,\mathcal{L}\left(h,\mathrm{Y}_n\right)}{\sum_{n=1}^{N}\delta\big(\,\mathrm{x}-\mathrm{X}_n\,\big)}\,.
\end{aligned}
$$

5.5.1 Regression: the Squared-Error Loss

$$
\mathcal{L}(h,y) = (h-y)^2\,. \qquad (5.43)
$$

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

$$
\begin{aligned}
\mathcal{R}(f) &= \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{y,x}\,|\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\,\big)^2\right]\right]\right] \qquad (5.44)\\
&= \mathrm{E}_{\mathrm{x},\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(\mathrm{y}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})^2\right]\right] + \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\right]\right]\\
&= \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\right]\right]
\end{aligned}
$$

5.5.1.1 Optimal Learner

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

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \\ &= \frac{\int_{\mathcal{Y}} y\big(\alpha(y,\mathrm{x}) + \bar{N}(y,\mathrm{x};\mathrm{D})\big)\mathrm{d}y}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\int_{\mathcal{Y}} y\frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})}\mathrm{d}y \\ &\quad + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\,\mathrm{Y}_n}{\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\mu_{\mathrm{y}\,|\,\mathrm{x}} \\ &\quad + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right)\frac{\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\,\mathrm{Y}_n}{\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)}\,. \end{aligned} \tag{5.45}$$

5.5.1.2 Minimum Risk

Generalizing from the basic model discussion, we again have

$$\begin{aligned} \mathcal{R}^* &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] = \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\ &= \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\,\mathrm{C}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\right], \end{aligned} \tag{5.46}$$

where we choose to perform the expectation over $\bar{\mathrm{n}}$.

The conditional varance is now

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left[\,\mathrm{y}^2\,\right] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\,. \tag{5.47}$$

and the two terms are independently evaluated.

$$\begin{aligned} \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathrm{y}^2\,\right]\right] &= \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left[\,\mathrm{y}^2\,\right]\right] \\ &= \mathrm{E}_{\mathrm{y}}[\mathrm{y}^2] = \int_{\mathcal{Y}} y^2\int_{\mathcal{X}}\frac{\alpha(y,x)}{\alpha_0}\mathrm{d}x\mathrm{d}y \\ &= \mathrm{E}_{\mathrm{x}}\left[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2]\right] = \int_{\mathcal{X}}\frac{\alpha'(x)}{\alpha_0}\int_{\mathcal{Y}} y^2\frac{\alpha(y,x)}{\alpha'(x)}\mathrm{d}y\mathrm{d}x\,. \end{aligned} \tag{5.48}$$

PGR: D PERSPECTIVE

$$
\begin{aligned}
\mathrm{E}_{\mathrm{x,D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] &= \mathrm{E}_{\mathrm{x}}\,\mathrm{E}_{\mathrm{D}\,|\,\mathrm{x}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \\
&= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{D}\,|\,\mathrm{x}}\left[\left(\frac{\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)}{\alpha'(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)}\right)^2\right]\right] \\
&= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{D}}\left[\frac{\alpha_0\left(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)^2}{\alpha'(\mathrm{x})\left(\alpha'(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)(\alpha_0+N)}\right]\right] \\
&= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{X}}\left[\frac{\alpha_0\,\mathrm{E}_{\mathrm{Y}\,|\,\mathrm{X}}\left[\left(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)^2\right]}{\alpha'(\mathrm{x})\left(\alpha'(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)(\alpha_0+N)}\right]\right]
\end{aligned}
\tag{5.49}
$$

Evaluating the expectation over Y given X, we have

$$
\begin{aligned}
&\mathrm{E}_{\mathrm{Y}\,|\,\mathrm{X}}\left[\left(\alpha'(\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{x})^2\mu^2_{\mathrm{y}\,|\,\mathrm{x}} + 2\alpha'(\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'\big(\mathrm{X}_n\big)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\mathrm{X}_n\big)\alpha'\big(\mathrm{X}_{n'}\big)\mu_{\mathrm{y}\,|\,\mathrm{x}}\big(\mathrm{X}_{n'}\big) + \alpha'\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'\big(\mathrm{X}_n\big)\alpha'\big(\mathrm{X}_{n'}\big) + \alpha'\big(\mathrm{X}_n\big)\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'(\mathrm{x})^2\mu^2_{\mathrm{y}\,|\,\mathrm{x}} + 2\alpha'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\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'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}} + \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\delta(0)}{\alpha'(\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'(\mathrm{x}) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)}{\alpha'(\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'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\left(\alpha'(\mathrm{x}) + \delta(0) + \sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)\right)
\end{aligned}
\tag{5.50}
$$

PGR: Y given X PDFs, moments??? In PDF section, or in Appendix?

Plugging,

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x,D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \qquad (5.51)\\
&= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{X}}\left[\frac{\alpha_0\,\mathrm{E}_{\mathrm{Y}\,|\,\mathrm{X}}\left[\left(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}+\sum_{n=1}^{N}\mathrm{Y}_n\,\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)^2\right]}{\alpha'(\mathrm{x})\left(\alpha'(\mathrm{x})+\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)(\alpha_0+N)}\right]\right]\\
&= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\,\mathrm{E}_{\mathrm{X}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\delta(0)\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)+\alpha'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\left(\alpha'(\mathrm{x})+\delta(0)+\sum_{n=1}^{N}\delta\big(\mathrm{x}-\mathrm{X}_n\big)\right)\right]}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x})+\delta(0)\big)(\alpha_0+N)}\right]
\end{aligned}$$

Evaluating the expectation over X,

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

Plugging,

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x,D}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \qquad (5.52)\\
&\quad= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\delta(0)N+\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\big(\alpha_0\alpha'(\mathrm{x})+\alpha_0\delta(0)+N\alpha'(\mathrm{x})\big)}{\big(\alpha'(\mathrm{x})+\delta(0)\big)(\alpha_0+N)}\right]
\end{aligned}$$

Combining with the second moment produces the risk,

$$\begin{aligned}
\mathcal{R}^* &= \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\mathrm{y}^2\right]-\mu^2_{\mathrm{y}\,|\,\mathrm{x,D}}\right] \qquad (5.53)\\
&= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\alpha'(\mathrm{x})+\alpha_0\delta(0)+N\alpha'(\mathrm{x})}{(\alpha_0+N)\big(\alpha'(\mathrm{x})+\delta(0)\big)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]\\
&= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+(\alpha_0+N)^{-1}\delta(0)}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}\delta(0)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]
\end{aligned}$$

PGR: Discuss Dirac deltas!!!

PGR: DMP PERSPECTIVE???

To perform the expectation over the Dirichlet-Multinomial process $\bar{\mathrm{n}} \sim \mathrm{DMP}(\alpha)$, split the expectation into an expectation over the marginal DMP $\mathrm{n}' \sim \mathrm{DMP}(\alpha')$ and a conditional expectation over $\bar{\mathrm{n}}$ given n'. The characterization of the conditional DMP is found in Appendix A.3.

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\
&\quad= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{x}}\left[\left(\frac{\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \int_{\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\mathrm{d}y}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}\right)^2\right]\right] \\
&\quad= \mathrm{E}_{\mathrm{x}}\,\mathrm{E}_{\mathrm{n}'}\left[\frac{\alpha_0\,\mathrm{E}_{\bar{\mathrm{n}}|\,\mathrm{n}'}\left[\left(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \int_{\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\mathrm{d}y\right)^2\right]}{\alpha'(\mathrm{x})\big(\alpha'(x) + \mathrm{n}'(x)\big)(\alpha_0 + N)}\right]
\end{aligned} \tag{5.54}$$

Evaluating the conditional expectation,

$$\begin{aligned}
&\mathrm{E}_{\bar{\mathrm{n}}|\,\mathrm{n}'}\left[\left(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \int_{\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\mathrm{d}y\right)^2\right] \\
&= \alpha'(\mathrm{x})^2\mu^2_{\mathrm{y}\,|\,\mathrm{x}} + 2\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}}\int_{\mathcal{Y}} y\delta(0)\frac{\mathrm{n}'(\mathrm{x})}{\delta(0)}\frac{\delta(0)^{-1}\alpha(y,\mathrm{x})}{\delta(0)^{-1}\alpha'(\mathrm{x})}\mathrm{d}y \\
&\quad+ \delta(0)^2\frac{\delta(0)^{-1}\,\mathrm{n}'(\mathrm{x})}{\big(\delta(0)^{-1}\alpha'(\mathrm{x})\big)\big(\delta(0)^{-1}\alpha'(\mathrm{x})+1\big)}\int_{\mathcal{Y}}\int_{\mathcal{Y}} yy'\Bigg[\left(\frac{\mathrm{n}'(\mathrm{x})}{\delta(0)} - 1\right)\frac{\alpha(y,\mathrm{x})}{\delta(0)}\frac{\alpha(y',\mathrm{x})}{\delta(0)} \\
&\quad+ \left(\frac{\alpha'(\mathrm{x})}{\delta(0)} + \frac{\mathrm{n}'(\mathrm{x})}{\delta(0)}\right)\frac{\alpha(y,\mathrm{x})}{\delta(0)}\delta(y-y')\Bigg]\mathrm{d}y\mathrm{d}y' \\
&= \alpha'(\mathrm{x})^2\mu^2_{\mathrm{y}\,|\,\mathrm{x}} + 2\alpha'(\mathrm{x})\,\mathrm{n}'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}} \\
&\quad+ \frac{\mathrm{n}'(\mathrm{x})}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x})+\delta(0)\big)}\Big[\big(\mathrm{n}'(\mathrm{x}) - \delta(0)\big)\alpha'(\mathrm{x})^2\mu^2_{\mathrm{y}\,|\,\mathrm{x}} + \delta(0)\big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\big)\alpha'(\mathrm{x})\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\Big] \\
&= \frac{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}{\alpha'(\mathrm{x}) + \delta(0)}\Big[\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x}) + \delta(0)\big) + \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\delta(0)\,\mathrm{n}'(\mathrm{x})\Big]
\end{aligned} \tag{5.55}$$

Plugging,

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\
&\quad= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\,\mathrm{E}_{\mathrm{n}'}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x}) + \delta(0)\big) + \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\delta(0)\,\mathrm{n}'(\mathrm{x})\right]}{\alpha'(\mathrm{x})\big(\alpha'(\mathrm{x}) + \delta(0)\big)(\alpha_0 + N)}\right] \\
&\quad= \mathrm{E}_{\mathrm{x}}\left[\frac{\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\big(\alpha_0\alpha'(\mathrm{x}) + N\alpha'(\mathrm{x}) + \delta(0)\alpha_0\big) + \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\left[\mathrm{y}^2\right]\delta(0)N}{(\alpha_0 + N)\big(\alpha'(\mathrm{x}) + \delta(0)\big)}\right]
\end{aligned} \tag{5.56}$$

Combining produces the risk,

$$\begin{aligned}
\mathcal{R}^* &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\alpha'(\mathrm{x}) + \alpha_0\delta(0) + N\alpha'(\mathrm{x})}{(\alpha_0 + N)\big(\alpha'(\mathrm{x}) + \delta(0)\big)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] \\
&= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}(\mathrm{x}) + (\alpha_0 + N)^{-1}\delta(0)}{\mathrm{P}(\mathrm{x}) + \alpha_0^{-1}\delta(0)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]
\end{aligned} \tag{5.57}$$

Appendix A

PGR: notation/formatting

PGR: remove beta set arguments by introducing alpha sub z?

A.1 Dirichlet random process conditioned on its aggregation

This section details an important property of Dirichlet distributed random processes. The following development first considers Dirichlet random processes over a countable domain and then generalizes for continuous-domain Dirichlet processes.

First, define the PDF of a Dirichlet aggregation [7]. Let the random process $\theta \in \Theta = \mathcal{P}(\mathcal{Y})$ be Dirichlet over the countable set $\mathcal{Y}$ with parameterizing function $\alpha \in \mathbb{R}^{+\mathcal{Y}}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$; the aggregation $\theta' \in \mathcal{P}(\mathcal{Z})$, $\theta'(z) \equiv \sum_{y \in \mathcal{S}(z)} \theta(y)$ is thus also Dirichlet and has a parameterizing function $\alpha' \in \mathbb{R}^{+\mathcal{Z}}$, $\alpha'(z) \equiv \sum_{y \in \mathcal{S}(z)} \alpha(y)$.

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

$$\begin{aligned} \mathrm{p}_{\theta|\theta'}(\theta|\theta') &= \frac{\beta(\alpha') \prod_{y \in \mathcal{Y}} \theta(y)^{\alpha(y)-1}}{\beta(\alpha) \prod_{z \in \mathcal{Z}} \theta'(z)^{\alpha'(z)-1}} \\ &= \prod_{z \in \mathcal{Z}} \left[\beta\Big(\big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\Big)^{-1} \frac{\prod_{y \in \mathcal{S}(z)} \theta(y)^{\alpha(y)-1}}{\theta'(z)^{\alpha'(z)-1}} \right] \\ &= \prod_{z \in \mathcal{Z}} \left[\frac{\theta'(z)^{1-|\mathcal{S}(z)|}}{\beta\Big(\big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\Big)} \prod_{y \in \mathcal{S}(z)} \left(\frac{\theta(y)}{\theta'(z)} \right)^{\alpha(y)-1} \right], \end{aligned} \tag{A.1}$$

which is defined for $\left\{\theta \in {\mathbb{R}_{\geq 0}}^{\mathcal{Y}} : \sum_{y\in\mathcal{S}(z)} \theta(y) = \theta'(z), \quad \forall z \in \mathcal{Z}\right\}$.

Observe that the partitioned segements are conditionally independent; introduce subscript notation to refer to the function segment $\theta_z = \left\{\theta(y) : y \in \mathcal{S}(z)\right\}$. The PDF $\mathrm{p}_{\theta|\theta'}$ can now be decomposed as $\mathrm{p}_{\theta|\theta'}(\ldots,\theta_z,\ldots|\theta') = \prod_{z\in\mathcal{Z}} \mathrm{p}_{\theta_z|\theta'(z)}\left(\theta_z|\theta'(z)\right)$

Next, normalize the segments of θ to form $\tilde{\theta} = \left(\ldots,\tilde{\theta}_z,\ldots\right)$, where $\tilde{\theta}_z \equiv \theta_z/\theta'(z)$, and formulate the conditional PDF

$$\begin{aligned}\mathrm{p}_{\tilde{\theta}|\theta'}\left(\tilde{\theta}|\theta'\right) &= \prod_{z\in\mathcal{Z}}\left[\frac{\prod_{y\in\mathcal{S}(z)}\tilde{\theta}_z(y)^{\alpha(y)-1}}{\beta\Big(\left\{\alpha(y) : y\in\mathcal{S}(z)\right\}\Big)}\right] \\ &= \prod_{z\in\mathcal{Z}} \mathrm{Dir}\left(\tilde{\theta}_z; \left\{\alpha(y) : y\in\mathcal{S}(z)\right\}\right).\end{aligned} \tag{A.2}$$

which is defined for $\tilde{\theta} \in \prod_{z\in\mathcal{Z}}\left\{\tilde{\theta}_z \in \mathcal{P}\left(\mathcal{S}(z)\right)\right\}$. Thus after conditioning, the normalized segments $\tilde{\theta}_z$ are Dirichlet distributed, independent of one another, and independent of the aggregation θ'.

PGR: discuss transform Jacobian and dimensionality?

This principle holds for continuous-domain Dirichlet processes θ as well - the segments $\tilde{\theta}_z$ are now continuous-domain Dirichlet processes.

A.2 Multinomial Distribution Properties

A.2.1 Aggregation

A characteristic of a Multinomial random process is that its aggregations are also Multinomial [9]. Consider a random process $\bar{\mathrm{n}} \sim \mathrm{Multi}(N,\theta)$ over the set $\mathcal{Y}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots,\mathcal{S}(z),\ldots\}$, $z\in\mathcal{Z}$; the transformed random process $\mathrm{n}'(z) \equiv \sum_{y\in\mathcal{S}(z)} \bar{\mathrm{n}}(y)$ is distributed as $\mathrm{n}' \sim \mathrm{Multi}(N,\theta')$ with parameterizing function $\theta'(z) = \sum_{y\in\mathcal{S}(z)}\theta(y)$.

To prove this principle, define the subset $\tilde{\mathcal{N}} = \left\{\bar{n} \in \mathcal{N} : \sum_{y\in\mathcal{S}(z)} \bar{n}(y) = n'(z), \quad \forall z \in\right.$

$\mathcal{Z}\,\} \subseteq \bar{\mathcal{N}}$, where the original random process $\bar{\mathrm{n}} \in \bar{\mathcal{N}}$. Next, observe

$$\begin{aligned} \mathrm{P}_{\mathrm{n}'}(n') &= \sum_{\bar{n}\in\tilde{\mathcal{N}}} \mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \\ &= \mathcal{M}(n') \prod_{z\in\mathcal{Z}} \sum_{\substack{n'(z)= \\ \sum_{y\in\mathcal{S}(z)}\bar{n}(y)}} \mathcal{M}\left(\{\bar{n}(y) : y\in\mathcal{S}(z)\}\right) \prod_{y\in\mathcal{S}(z)} \theta(y)^{\bar{n}(y)} \\ &= \mathcal{M}(n') \prod_{z\in\mathcal{Z}} \theta'(z)^{n'(z)} = \mathrm{Multi}(n'; N, \theta')\ , \end{aligned} \tag{A.3}$$

where the multinomial theorem [8] has been used.

A.2.2 Conditioned on its Aggregation

If the multinomial random process $\bar{\mathrm{n}}$ is conditioned on its aggregation over the partition $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$, the distinct segements $\bar{\mathrm{n}}(y)$, $y \in \mathcal{S}(z)$ become independent multinomial random processes,

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}'}(\bar{n}|n') &= \frac{\mathcal{M}(\bar{n}) \prod_{y\in\mathcal{Y}} \theta(y)^{\bar{n}(y)}}{\mathcal{M}(n') \prod_{z\in\mathcal{Z}} \theta'(z)^{n'(z)}} \\ &= \prod_{z\in\mathcal{Z}} \left[\mathcal{M}\left(\left\{\bar{n}(y) : y\in\mathcal{S}(z)\right\}\right) \prod_{y\in\mathcal{S}(z)} \left(\frac{\theta(y)}{\theta'(z)}\right)^{\bar{n}(y)} \right] \\ &= \prod_{z\in\mathcal{Z}} \left[\mathcal{M}\left(\left\{\bar{n}(y) : y\in\mathcal{S}(z)\right\}\right) \prod_{y\in\mathcal{S}(z)} \tilde{\theta}_z(y)^{\bar{n}(y)} \right] \\ &= \prod_{z\in\mathcal{Z}} \mathrm{Multi}\left(\left\{\bar{n}(y) : y\in\mathcal{S}(z)\right\}; n'(z), \left\{\tilde{\theta}_z(y) : y\in\mathcal{S}(z)\right\}\right)\ , \end{aligned} \tag{A.4}$$

on the domain $\left\{\bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}} : \sum_{y\in\mathcal{S}(z)} \bar{n}(y) = n'(z), \quad \forall z \in \mathcal{Z}\right\}$. Observe that the segment over the set $\mathcal{S}(z)$ sums to $\mathrm{n}'(z)$ and is parameterized by the normalized segment $\tilde{\theta}_z \equiv \left\{\theta(y)/\theta'(z) : y \in \mathcal{S}(z)\right\}$.

A.3 Dirichlet-Multinomial random process conditioned on its aggregation

A defining characteristic of a Dirichlet-Multinomial random process is that its aggregations are also Dirichlet-Multinomial [9]. Consider a DM random process $\bar{\mathrm{n}} \sim \mathrm{DM}(N, \alpha)$ over the

set $\mathcal{Y}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$; the transformed random process $\mathrm{n}'(z) \equiv \sum_{y \in \mathcal{S}(z)} \bar{\mathrm{n}}(y)$ is neccessarily Dirichlet-Multinomial with parameterizing function $\alpha'(z) = \sum_{y \in \mathcal{S}(z)} \alpha(y)$.

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

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}'}(\bar{n}|n') &= \frac{\mathcal{M}(\bar{n})\beta(\alpha)^{-1}\beta(\alpha+\bar{n})}{\mathcal{M}(n')\beta(\alpha')^{-1}\beta(\alpha'+n')} \\ &= \left(\prod_{z\in\mathcal{Z}} \frac{\Gamma\big(\alpha'(z)+n'(z)\big)}{n'(z)!\Gamma\big(\alpha'(z)\big)}\right)^{-1} \left(\prod_{y\in\mathcal{Y}} \frac{\Gamma\big(\alpha(y)+\bar{n}(y)\big)}{\bar{n}(y)!\Gamma\big(\alpha(y)\big)}\right) \\ &= \prod_{z\in\mathcal{Z}} \left[\frac{n'(z)!\Gamma\big(\alpha'(z)\big)}{\Gamma\big(\alpha'(z)+n'(z)\big)} \prod_{y\in\mathcal{S}(z)} \frac{\Gamma\big(\alpha(y)+\bar{n}(y)\big)}{\bar{n}(y)!\Gamma\big(\alpha(y)\big)}\right] \\ &= \prod_{z\in\mathcal{Z}} \mathrm{DM}\left(\big\{\bar{n}(y) : y \in \mathcal{S}(z)\big\}; n'(z), \big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\right), \end{aligned} \tag{A.5}$$

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

A.4 First and Second moments of a Dirichlet Process

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

$$\mu_\theta = \frac{\alpha}{\alpha_0}\,, \tag{A.6}$$

where $\alpha_0 = \int_{\mathcal{Y}} \alpha(y)\mathrm{d}y$.

The defining characteristic of Dirichlet processes is that their aggregations are also Dirichlet. Define the partition of $\mathcal{Y}$, $\big\{\,\mathcal{S}(y), S^c(y)\big\}$ where $\mathcal{S}(y) = \{t \in \mathcal{Y} : t \leq y\}$. The transform random variable $\theta'(y) \equiv \int_{-\infty}^{y} \theta(t)\mathrm{d}t$ is thus a Beta random variable with parameters $\lambda = \int_{-\infty}^{y} \alpha(t)\mathrm{d}t$ and $\lambda^c = \int_{y}^{\infty} \alpha(t)\mathrm{d}t$. Using the formula for the expected value of a beta random variable [13], note that

$$\begin{aligned} \mu_{\theta'} &= \frac{\lambda}{\lambda+\lambda^c} \\ &= \frac{\int_{-\infty}^{y} \alpha(t)\mathrm{d}t}{\alpha_0} = \int_{-\infty}^{y} \mu_\theta(t)\mathrm{d}t\,. \end{aligned} \tag{A.7}$$

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

Next, the correlation function is shown to be

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

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 parameters $\Big(\int_{-\infty}^{y_1}\alpha(t)\mathrm{d}t, \int_{y_1}^{y_2}\alpha(t)\mathrm{d}t, \int_{y_2}^{\infty}\alpha(t)\mathrm{d}t\Big)$.

Define the function

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

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

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

Finally,

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

A.5 Proof: Continuous Model Posterior Distribution is Dirichlet Process

In this section, it is shown that if the model $\theta \sim \mathrm{DP}(\alpha)$ is a Dirichlet process, then the model conditioned on the training data D is also a Dirichlet process with parameterizing function

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

The defining characteristic of Dirichlet processes is that their aggregations are also Dirichlet. Consider a DP over the set $\mathcal{Y}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$; the transformed random process $\theta'(z) \equiv \int_{\mathcal{S}(z)} \theta(y)\mathrm{d}y$ is neccessarily Dirichlet with parameterizing function $\alpha'(z) \equiv \int_{\mathcal{S}(z)} \alpha(y)\mathrm{d}y$.

To prove the hypothesis, it is required that

$$\theta' | \, \mathrm{D} \sim \mathrm{Dir}\left(\alpha' + \bar{N}(\mathrm{D})\right), \tag{A.12}$$

where $\bar{N}(z; \mathrm{D}) = \int_{\mathcal{S}(z)} \sum_{n=1}^{N} \delta\big(y - \mathrm{D}_n\big)\mathrm{d}y = \sum_{n=1}^{N} \chi\big(\mathrm{D}_n; \mathcal{S}(z)\big)$. χ is the indicator function

$$\chi(x; S) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{if } x \notin S. \end{cases} \tag{A.13}$$

PGR: SUFFICIENT STATISTIC FOR D!!!?

To prove the hypothesis, exploit the results of Appendix A.1 to represent the training data distribution conditioned on the aggregation θ'. The conditional distribution of interest is

$$\begin{aligned} \mathrm{p}_{\mathrm{D}\,|\theta'}(D|\theta') &= \mathrm{E}_{\theta|\theta'}\left[\,\mathrm{p}_{\mathrm{D}\,|\theta}(D|\theta)\right] = \mathrm{E}_{\theta|\theta'}\left[\prod_{n=1}^{N} \theta\big(D_n\big)\right] \\ &= \prod_{z\in\mathcal{Z}} \mathrm{E}_{\theta_z|\theta'(z)}\left[\prod_{n=1}^{N} \theta_z\big(D_n\big)^{\chi\left(D_n;\mathcal{S}(z)\right)}\right] \\ &= \left(\prod_{z\in\mathcal{Z}}\prod_{n=1}^{N} \theta'(z)^{\chi\left(D_n;\mathcal{S}(z)\right)}\right) \prod_{z\in\mathcal{Z}} \mathrm{E}_{\tilde{\theta}_z}\left[\prod_{n=1}^{N} \tilde{\theta}_z\big(D_n\big)^{\chi\left(D_n;\mathcal{S}(z)\right)}\right] \\ &= \left(\prod_{z\in\mathcal{Z}} \theta'(z)^{\bar{N}(z;D)}\right) \prod_{z\in\mathcal{Z}} \mathrm{E}_{\tilde{\theta}_z}\left[\prod_{n=1}^{N} \tilde{\theta}_z\big(D_n\big)^{\chi\left(D_n;\mathcal{S}(z)\right)}\right], \end{aligned} \tag{A.14}$$

Observe that the dependency on θ' is polynomial. The training data marginal distribution is

$$\begin{aligned} \mathrm{p}_{\mathrm{D}}(D) &= \mathrm{E}_{\theta'}\left[\prod_{z\in\mathcal{Z}} \theta'(z)^{\bar{N}(z;D)}\right] \prod_{z\in\mathcal{Z}} \mathrm{E}_{\tilde{\theta}_z}\left[\prod_{n=1}^{N} \tilde{\theta}_z\big(D_n\big)^{\chi\left(D_n;\mathcal{S}(z)\right)}\right] \\ &= \frac{\beta\big(\alpha' + \bar{N}(D)\big)}{\beta(\alpha')} \prod_{z\in\mathcal{Z}} \mathrm{E}_{\tilde{\theta}_z}\left[\prod_{n=1}^{N} \tilde{\theta}_z\big(D_n\big)^{\chi\left(D_n;\mathcal{S}(z)\right)}\right] \end{aligned} \tag{A.15}$$

and thus the distribution of interest is

$$\begin{aligned} \mathrm{p}_{\theta'|\,\mathrm{D}}(\theta'|D) &= \frac{\prod_{z\in\mathcal{Z}} \theta'(z)^{\alpha'(z)+\bar{N}(z;D)-1}}{\beta\big(\alpha'+\bar{N}(D)\big)} \\ &= \mathrm{Dir}\left(\theta';\alpha'+\bar{N}(\cdot;\mathrm{D})\right) \ . \end{aligned} \tag{A.16}$$

This proves the hypothesis.

A.6 The Dirichlet-Multinomial Process

PGR: multinomial process too???

A.6.1 Definition

This section introduces a new random process, referred to as the Dirichlet-Multinomial process (DMP). It is the generalization of the Dirichlet-Multinomial distribution for i.i.d. samples drawn from a PDF; the underlying distribution is characterized by a Dirchlet process with parameter α. The Dirichlet-Multinomial process assumes functions from the set $\left\{\bar{n}\in{\mathbb{R}_{\geq 0}}^{\mathcal{Y}} : \int_{\mathcal{Y}} \bar{n}(y)\mathrm{d}y = N\right\}$ and is parameterized by a function $\alpha : \mathcal{Y} \mapsto \mathbb{R}^+$.

Analagous to the Dirichlet and Dirichlet-Multinomial distributions for countable spaces, the Dirichlet-Multinomial process inherits the aggregation property from the Dirichlet process prior. That is, for a Dirichlet-Multinomial process $\bar{\mathrm{n}} \in \left\{\bar{n}\in{\mathbb{R}_{\geq 0}}^{\mathcal{Y}} : \int_{\mathcal{Y}} \bar{n}(y)\mathrm{d}y = N\right\}$ and a partition of $\mathcal{Y}$, $\{\ldots,\mathcal{S}(z),\ldots\}$, $z\in\mathcal{Z}$, the transformed random process $\mathrm{n}'(z) \equiv \int_{\mathcal{S}(z)} \bar{\mathrm{n}}(y)\mathrm{d}y$ is neccessarily Dirichlet-Multinomial with parameterizing function $\alpha'(z) \equiv \int_{\mathcal{S}(z)} \alpha(y)\mathrm{d}y$.

PGR: tilde not prime?

A.6.2 Proof that $\sum_{n=1}^{N} \delta\big(y-\mathrm{D}_n\big)$ is a DMP

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

Observe that $\mathrm{n}'(z) \equiv \sum_{n=1}^{N} \chi\big(\mathrm{D}_n;\mathcal{S}(z)\big)$, where χ is the indicator function

$$\chi(x;S) = \begin{cases} 1 & \text{if } x\in S, \\ 0 & \text{if } x\notin S. \end{cases} \tag{A.17}$$

and note that $\mathrm{P}\left(\chi\big(\mathrm{D}_n;\mathcal{S}(z)\big)=1\big|\theta\right)=\int_{\mathcal{S}(z)}\theta(y)\mathrm{d}y$. As such, n' conditioned on the model θ is characterized by a multinomial distribution

$$\mathrm{P}_{\mathrm{n}'|\theta}(n'|\theta)=\mathcal{M}(n')\prod_{z\in\mathcal{Z}}\left(\int_{\mathcal{S}(z)}\theta(y)\mathrm{d}y\right)^{n'(z)}=\mathrm{Multi}\left(n';N,\theta'(z)\right)\ , \tag{A.18}$$

where $\theta'(z)\equiv\int_{\mathcal{S}(z)}\theta(y)\mathrm{d}y$, $z\in\mathcal{Z}$.

By the aggregation property of the Dirichlet process θ, the parameters of this multinomial distribution are characterized as $\theta'\sim\mathrm{Dir}\left(\alpha'\right)$, and thus $\tilde{\mathrm{n}}$ is drawn from a Dirichlet-Multinomial PMF with the same parameters α'. As this holds for any partition of $\mathcal{Y}$, $\bar{\mathrm{n}}$ is a Dirichlet-Multinomial Process.

A.6.3 Mean and Correlation Functions

In this subsection the mean and correlation functions of a DMP are expressed. The mean function is

$$\begin{aligned}\mu_{\bar{\mathrm{n}}}(y)&=\sum_{n=1}^{N}\mathrm{E}_{\mathrm{D}_n}\left[\delta\big(y-\mathrm{D}_n\big)\right]\\&=\sum_{n=1}^{N}\mathrm{P}_{\mathrm{D}_n}(y)\\&=N\frac{\alpha(y)}{\alpha_0}\ .\end{aligned} \tag{A.19}$$

The correlation function is

$$\begin{aligned}
\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y)\bar{\mathrm{n}}(y')\right] &= \sum_{n=1}^{N} \mathrm{E}_{\mathrm{D}_n}\left[\delta\big(y-\mathrm{D}_n\big)\right] \\
&= \sum_{n=1}^{N}\sum_{n'=1}^{N} \mathrm{E}_{\mathrm{D}_n,\mathrm{D}_{n'}}\left[\delta\big(y-\mathrm{D}_n\big)\delta\big(y-\mathrm{D}_{n'}\big)\right] \\
&= \sum_{n} \mathrm{E}_{\mathrm{D}_n}\left[\delta\big(y-\mathrm{D}_n\big)\delta\big(y'-\mathrm{D}_n\big)\right] + \\
&\quad \sum_{n\neq n'} \mathrm{E}_{\mathrm{D}_n,\mathrm{D}_{n'}}\left[\delta\big(y-\mathrm{D}_n\big)\delta\big(y'-\mathrm{D}_{n'}\big)\right] \\
&= \sum_{n} \int_{\mathcal{Y}} \frac{\alpha(\tilde{y})}{\alpha_0}\delta(y-\tilde{y})\delta(y'-\tilde{y}) + \\
&\quad \sum_{n\neq n'} \int_{\mathcal{Y}}\int_{\mathcal{Y}} \frac{\alpha(\tilde{y})\alpha(\tilde{y}')+\alpha(\tilde{y})\delta(\tilde{y}-\tilde{y}')}{\alpha_0(\alpha_0+1)}\delta(y-\tilde{y})\delta(y-\tilde{y}') \\
&= N\frac{\alpha(y)}{\alpha_0}\delta(y-y') + N(N-1)\frac{\alpha(y)\alpha(y')+\alpha(y)\delta(y-y')}{\alpha_0(\alpha_0+1)} \\
&= \frac{N}{\alpha_0(\alpha_0+1)}\big[(N-1)\alpha(y)\alpha(y') + (\alpha_0+N)\alpha(y)\delta(y-y')\big] \,.
\end{aligned} \tag{A.20}$$

A.6.4 Continuous aggregation

If $\bar{\mathrm{n}}$ is a Dirichlet-Multinomial process over a Euclidean space $\mathcal{Y}$, then conditioning on its discrete aggregation n' produces independent Dirichlet-Multinomial processes $\big\{\bar{\mathrm{n}}(y) : y \in \mathcal{S}(z)\big\} \sim \mathrm{DMP}\left(\mathrm{n}'(z), \big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\right)$ over the partition spaces $\mathcal{S}(z)$.

The previous result can be extended to conditioning on a continuous aggregation. Define $\bar{\mathrm{n}} \sim \mathrm{DMP}(N,\alpha)$ over the set $\mathcal{Y}\times\mathcal{X}$ and the aggregation DMP $\mathrm{n}' = \int_{\mathcal{Y}} \bar{\mathrm{n}}(y,\cdot)\mathrm{d}y$ over set $\mathcal{X}$ with parameterizing function $\alpha' = \int_{\mathcal{Y}} \alpha(y,\cdot)\mathrm{d}y$.

Use the aggregation propery to introduce a Dirichlet-Multinomial process $\tilde{\mathrm{n}}(y;k) = \int_{\Delta k}^{\Delta(k+1)} \bar{\mathrm{n}}(y,x)\mathrm{d}x$ with parameter $\tilde{\alpha}(y;k) = \int_{\Delta k}^{\Delta(k+1)} \alpha(y,x)\mathrm{d}x$. Additionally, introduce its own aggregation, a Dirichlet-Multinomial random process $\dot{n}(k) = \int_{\mathcal{Y}} \tilde{n}(y,k)\mathrm{d}y$ with parameter $\dot{\alpha}(k) = \int_{\mathcal{Y}} \tilde{\alpha}(y,k)\mathrm{d}y$. By the conditioning property for discrete aggregations demonstrated previously, $\tilde{\mathrm{n}}(\cdot,k)|\dot{\mathrm{n}}(k) \sim \mathrm{DMP}\left(\dot{\mathrm{n}}(k), \tilde{\alpha}(\cdot,k)\right)$ are independent DMP's.

Note that as $\Delta \to 0$, $\tilde{\mathrm{n}}(y,k) \approx \Delta\bar{\mathrm{n}}(y,\Delta k)$, $\tilde{\alpha}(y,k) \approx \Delta\alpha(y,\Delta k)$, and $\dot{\mathrm{n}}(k) \approx \Delta\,\mathrm{n}'(\Delta k)$. Letting $x \equiv \Delta k$, the statistics of the DMP conditioned on its continuous aggregation can be

represented as

$$\Delta\bar{\mathrm{n}}(\cdot, x)|\Delta\,\mathrm{n}'(k) \sim \mathrm{DMP}\left(\Delta\,\mathrm{n}'(k), \Delta\alpha(\cdot, x)\right). \tag{A.21}$$

Appendix B

PGR: Many sections redundant given Dirichlet perspective...

PGR: needs notation/formatting scrub

B.1 Maximum *a Posteriori* estimate of θ given D

PGR: delete? been done...

PGR: REDO USING EASY LOG-LIKELIHOOD and LAGRANGE!!!!!!! Generalize for Dirichlet!!!

To determine the MAP estimate of the model PMF θ given the training data D,

$$\hat{\theta}_{MAP}(D) = \arg\max_{\theta \in \Theta} \mathrm{P}_{\theta|\,\mathrm{D}}(\theta|D) \;, \tag{B.1}$$

we perform contrained optimization. Note that the set $\Theta = \left\{ \theta \in {\mathbb{R}_{\geq 0}}^M : \sum_{m=1}^{M} \theta_m = 1 \right\}$ implies both equality and inequality constraints...

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] Thomas P. Minka. *Bayesian inference, entropy, and the multinomial distribution*. Tech. rep. Microsoft Research, 2003.

[12] Kevin P. Murphy. *Binomial and multinomial distributions*. Tech. rep. University of British Columbia, 2006.

[13] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. Fourth. McGraw-Hill, 2002.

[14] 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.

[15] Steven Roman. "The Logarithmic Binomial Formula". In: *American Mathematical Monthly* 99.7 (1992).

[16] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* USA: McGraw-Hill, Inc., 1987. ISBN: 0070542341.

[17] 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.

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

[19] J. G. Wendel. "Note on the Gamma Function". In: *The American Mathematical Monthly* 55.9 (1948), pp. 563–564.