Design and Analysis of Statistical Learning Algorithms which Control False Discoveries

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Arijit Das
aus Budge Budge

Berichterstatter/in:	Prof. Dr. Achim Tresch
	Prof. Dr. Andreas Beyer
Prüfungsvorsitzender:	Prof. Dr. Stanislav Kopriva

Tag der mündlichen Prüfung: Freitag, 20. Juli 2018

Abstract

In this thesis, general theoretical tools are constructed which can be applied to develop machine learning algorithms which are consistent, with fast convergence and which minimize the generalization error by asymptotically controlling the rate of false discoveries (FDR) of features, especially for high dimensional datasets. Even though the main inspiration of this work comes from biological applications, where the data is extremely high dimensional and often hard to obtain, the developed methods are applicable to any general statistical learning problem.

In this work, the various machine learning tasks like hypothesis testing, classification, regression, etc are formulated as risk minimization algorithms. This allows such learning tasks to be viewed as optimization problems, which can be solved using first order optimization techniques in case of large data scenarios, while one could use faster converging second order techniques for small to moderately sized data sets. Further, such a formulation allows us to estimate the first order convergence rates of an empirical risk estimator for any arbitrary learning problem, using techniques from large deviation theory.

In many scientific applications, robust discovery of factors affecting an outcome or a phenotype, is more important than the accuracy of predictions. Hence, it is essential to find an appropriate approach to regularize an under-determined estimation problem and thereby control the generalization error. In this work, the use of local probability of false discovery is explored as such a regularization parameter, which forces the optimized solution towards functions with a lower probability to be a false discovery. Again, techniques from large deviation theory and the Gibbs principle allow the derivation of an appropriately regularized cost function.

These two theoretical results are then used to develop concrete applications. First, the problem of multi-classification is analyzed, which classifies a sample from an arbitrary probability measure into a finite number of categories, based on a given training data set. A general risk functional is derived, which can be used to learn Bayes optimal classifiers controlling the false discovery rate.

Secondly, the problem of model selection in the regression context is considered, aiming to select a subset of given regressors which explains most of the observed variation i.e. perform

ANOVA. Again, using techniques mentioned above, a risk function is derived which when optimized, controls the rate of false discoveries. This technique is shown to outperform the popular LASSO algorithm, which can be proven to not control the FDR, but only the FWER.

Finally, the problem of inferring under-sampled and partially observed non-negative discrete random variables is addressed, which has applications to analyzing RNA sequencing data. By assuming infinite divisibility of the underlying random variable, its characterization as being a discrete Compound Poisson Measure (DCP), is derived. This allows construction of a non-parametric Bayesian model of DCPs with a Pitman-Yor Mixture process prior, which is shown to allow for consistent inference under Kullback-Liebler and Renyi divergences even in the under-sampled regime.

Abstract

In dieser Arbeit werden allgemeine theoretische Methoden entwickelt, die angewendet werden können um maschinelle Lernalgorithmen zu generieren die konsistent sind, schnelle Konvergenz zeigen und den Generalisierungsfehler minimieren, indem die False Discovery Rate (FDR) insbesondere für hochdimensionale Datensätze gesteuert wird. Obwohl die Hauptinspiration dieser Arbeit von biologischen Anwendungen herrührt, bei denen die Daten extrem hochdimensional und oft schwer zu erhalten sind, sind die entwickelten Methoden auf alle allgemeinen statistischen Lernprobleme anwendbar.

In dieser Arbeit werden die verschiedenen maschinellen Lernaufgaben wie Hypothesentest, Klassifizierung, Regression usw. als Risikominimierungsalgorithmen formuliert. Auf diese Weise können solche Lernaufgaben als Optimierungsprobleme angesehen werden, die im Fall von großen Datenmengen mit Optimierungstechniken erster Ordnung gelöst werden können, während für kleine bis mittelgroße Datenmengen Techniken zweiter Ordnung mit schnellerer Konvergenz verwendet werden könnten. Darüber hinaus ermöglicht eine solche Formulierung die Schätzung der Konvergenzraten erster Ordnung eines empirischen Risikoschätzers für jedes beliebige Lernproblem unter Verwendung von Techniken aus der Theorie der großen Abweichungen.

In wissenschaftlichen Anwendungen ist eine robuste Detektion von Faktoren, die das Ergebnis beeinflussen, wichtiger als die Genauigkeit von Vorhersagen. Daher ist es wichtig, einen geeigneten Ansatz zu finden, um ein unterbestimmtes Schätzproblem zu regulieren und dadurch den Generalisierungsfehler zu kontrollieren. In dieser Arbeit wird die Verwendung der lokalen Wahrscheinlichkeit einer False Discovery als ein solcher Regularisierungsparameter untersucht, der die optimierte Lösung in Richtung von Funktionen mit einer geringeren Wahrscheinlichkeit einer False Discovery zwingt. Auch hier erlauben Techniken der Theorie der großen Abweichungen und des Gibbs-Prinzips die Ableitung einer angemessen regulierten Kostenfunktion.

Diese beiden theoretischen Ergebnisse werden anschließend verwendet, um konkrete Anwendungen zu entwickeln. Zunächst wird das Problem der Multi-Klassifikation analysiert, das basierend auf einem gegebenen Trainingsdatensatz eine Stichprobe aus einem beliebigen Wahrscheinlichkeitsmaß in eine endliche Anzahl von Kategorien einordnet. Es wird ein

allgemeines Risikofunktional abgeleitet, das verwendet werden kann, um optimale Bayes-Klassifikatoren zu lernen, die die False Discovery Rate steuern.

Zweitens wird das Problem der Modellauswahl im Regressionskontext betrachtet, das darauf abzielt, eine Untergruppe gegebener Regressoren auszuwählen, die den Großteil der beobachteten Variation erklärt, d. H. ANOVA durchführt. Unter Verwendung der oben erwähnten Techniken wird wiederum eine Risikofunktion abgeleitet, die, wenn sie optimiert ist, die False Discovery Rate steuert. Diese Methode ermöglicht den Nachweis, dass der häufig verwendete LASSO-Algorithmus nicht FDR sondern nur FWER steuert.

Schließlich wird das Problem der Ableitung von unterabgetasteten und teilweise beobachteten nicht-negativen diskreten Zufallsvariablen behandelt, die Anwendungen zur Analyse von RNA-Sequenzierungsdaten haben. Durch Annahme einer unendlichen Teilbarkeit der zugrundeliegenden Zufallsvariablen wird ihre Charakterisierung als diskretes zusammengesetztes Poisson-Maß (DCP) abgeleitet. Dies ermöglicht die Konstruktion eines nicht-parametrischen Bayes-Modells von DCPs mit einem Pitman-Yor-Mixture-Prozess, der gezeigt hat, dass konsistente Inferenz unter Kullback-Leibler- und Renyi-Divergenzen möglich ist, selbst wenn der Träger unterabgetastet ist.

Contents

List of Figures

Chapter 1

Design and Analysis of Statistical Algorithms

1.1 Introduction

With increasing generation and access to large amounts of data in the last decades, statistical learning theory has become an imperative in providing a general principled framework for automating the process of gaining knowledge, making decisions and constructing models to make predictions from a given set of data. Unlike artificial intelligence, the idea is not to explain or generate "intelligent behavior", its goal is more modest: it just wants to discover mechanisms for statistically consistent inductive inference with the ability to generalize. These principles allow us to precisely define what can and cannot be learned under different situations, whether the algorithm is stable, how much data is necessary to achieve certain performance targets, as well as to automate the design of such learning algorithms [120].

Such a theory starts by characterizing the space of all finitely approximable mathematical objects, which one can actually describe in the real world. Such objects turn out to be characterized by having the Polish topology, i.e. being a complete separable metric space [77]. This result indicates that in order to build a learning algorithm about an object they necessarily need to have this property. Functions on these spaces can then be defined which could represent certain probability measures on it or even the class of possible algorithms under consideration, which we call the Hypothesis space. These functions can range from the space of indicator functions, which are useful for classification tasks and linear models for regression and prediction tasks at one end, to the state of the art deep learning models (like convolutional neural networks) at the other, which have been applied to a wide variety of learning problems recently.

The performance of elements of the Hypothesis space is described using integrable non-negative real valued functions, known as loss functions. They are characteristic of the population distribution of the observed data and the specific learning problem like classification, hypothesis testing, regression, etc. Recently algorithms like GAN [12, 13, 116], VB Generative Models [88, 97, 100], Inverse Autoregressive Flow [73, 74, 117, 119] etc have been introduced wherein the loss functions are directly learned from the data thereby providing state of the art performance in sampling algorithms from extremely complicated distributions like images, audio, etc.

In the risk minimization framework of statistical learning theory, the goal is to find a function which minimizes the Risk i.e.

$$f^* = \arg \inf_{f \in \mathscr{H}} R(f)$$

where risk is defined to be the expected loss function. When such an expected value exists and an optimal solution exists, then the problem is said to be learnable. A "learning algorithm" therefore can then be defined as the iterative optimization algorithm which constructs a sequence of functions $f_1, f_2, \ldots$ which converges to the optimal solution $f^* = \lim_{m \to \infty} f_m$. The performance of different learning algorithms can then be characterized in terms of their rate of convergence and run-time complexity.

In the case of convex loss functions, there exists many of the shelf convex optimization algorithms which can be applied in an essentially black-box manner to solve the required learning problem. Choice among these algorithms depends on the required accuracy of solutions, run-time complexity of algorithms among other considerations [8, 44, 45, 48][8, 44, 45, 48]. For example for low dimensional problems when even with higher computational complexity we get acceptable run times, the second order Newton and Quasi-Newton methods [44, 45, 47, 61, 67] provides a provably quadratic convergence to the correct solution. However for high dimensional problems however, one usually settles for the class of first order gradient descent methods [64, 68, 69, 80], including their accelerated versions.

For a finite number of training examples it is always possible to build a function which fits the data exactly, i.e. have essentially zero risk. However such a function may not perform as well for unseen instances especially in the presence of noise (i.e. have the problem of overfitting). Therefore one not only needs to find a function which minimizes risk within the hypothesis space, but at the same time the difference between the estimated risk based on the observed data to the true population risk, a quantity known as the Generalization error, needs to minimized. Then one can say that the function overfits/underfits iff the generalization error is significantly negative/positive.

The generalization error itself can be decomposed into two familiar components [58, 64, 72]. The first term is called the approximation error, which measures how well functions in the Hypothesis space can approach the target. The second term known as the estimation error, is a random quantity which depends on the sampling process but is independent of the target. It measures how close the estimated function is to the best possible choice in the Hypothesis space. These terms are analogous to the classical concepts of bias and variance respectively which are usually associated to the problem of regression with a square error loss.

This allows us to define a learning algorithm to be consistent if and only if the estimation error converges to zero as the number of observed instances increases. The conditions which characterizes the consistency of an algorithm is independent of the target function and are of central interest in machine learning theory. However, estimating the approximation error requires specific assumptions about the target, such as having certain regularities, degrees of differentiability, etc as otherwise even a consistent learning algorithm might have an arbitrarily slow rate of convergence.

Clearly, enforcing more assumptions on a class of functions reduces its "capacity" [15, 57, 58, 72, 125] to represent different functions which inevitably would lead to higher approximation error for an arbitrary target function. This is the well known dichotomy between controlling the trade-off between approximation (bias) and estimation (variance) error. The Approximation error is determined purely by the choice of the Hypothesis space of functions $\mathscr{H}$ and is independent of the population measure. However, the complexity of $\mathscr{H}$ in terms of its degrees of differentiability, types of singularities, Group invariance and equivariance, etc effects the rate of convergence of the optimization algorithms. Higher the complexity/regularity/capacity of $\mathscr{H}$, slower the rate of convergence one might expect, independent of the learning problem.

In the context of simple hypothesis testing, this tradeoff is analogous to Neyman Pearson Optimality where one wants to minimize Type II error while controlling Type I error. Here one can see Type II error as the risk associated with the choice of a test and we want to choose the one which minimizes it. While the Type I error in analogous to the generalization error of a test which we want to control at a certain level.

Therefore Hypothesis spaces with higher capacity can represent target functions of higher complexity. However, rather unintuitively, there is no universal way of measuring the complexity of the elements in the Hypothesis space, a fact which has been formalized in what is known as the No Free Lunch theorem [120]. It is one of the most important theorem in the theory of statistical learning and essentially says that if there is no a priori restriction on the possible phenomena that are expected, it is impossible to generalize and no admissible estimator exists (i.e. one which outperforms others, under all situations). It simply tells us that in order to be able to learn successfully with guarantees on the behavior of the learning

algorithm, we need to make assumptions on the underlying distribution under consideration.

In this work we are interested in the relation of False Discoveries in controlling the generalization error. The probability of False Discovery is defined to be the posterior probability of the chosen label being wrong (for classification) or the chosen element in the dictionary being wrong (for regression) to represent a certain function, given the observed data.

Therefore certain reasonable assumptions are absolutely necessary to construct non vacuous results for a particular learning problem. Usually one assumes that the observed samples are independent and identically distributed, which is reasonable in many cases, though not all. Further, since at the time of training the underlying distribution is not known, one might need to assume that the training data was sampled from a certain family of distributions (e.g. exponential family) or more generally the distribution has a certain tail behavior (e.g. exponential or polynomial decay). One could also consider that the function of interest is not deterministic but the results are a sample from a conditional distribution. This relaxation is important in the case of noisy labels in a classification problem.

Based on the framework described here, the following questions and applications were considered as a part of the thesis. Firstly one is interested in determining the necessary and sufficient conditions for uniform consistency of an algorithm, i.e. the conditions under which

$$\Pr(\sup_{f\in\mathscr{H}} \mid \hat{R}_{\text{emp}}(f) - R(f) \mid > \varepsilon)$$

the probability of estimation error going to zero increases with the number of observations, independent of the probability measure the data was sampled from. In this work, by modeling the distribution of the non-negative real valued losses using a Levy process, we derived the necessary and sufficient conditions for a finite risk functional to exist. This representation also allowed us to model the distribution of losses using gamma and stable family of distributions which lead to efficient estimators of risk, especially useful when one does not have access to a large amount of data, which is common in biological applications.

Secondly by applying the Cramer's theorem, the exact asymptotic tail behavior was also calculated which allows us to answer the question about the consistency of a machine learning algorithm for an extremely large class of data distributions. This allows us to construct a universal compound Hypothesis tests which optimize for the the posterior analogues of Type I and Type II error, i.e the local probability of false discovery and the local probability of false non-discovery.

Thirdly we apply this result to the regularization problem of controlling the bias-variance tradeoff. Since any statistical learning problem can then be formulated as risk minimization

which controls the generalization error by controlling the tail error probabilities i.e.

$$U(f) = \widehat{R}(f) + \gamma \text{Reg}(f)$$

where $\gamma > 0$, then

$$f^* = \arg\min_{f \in \mathscr{H}} U(f)$$

However the optimization of such a risk functional leads to a sequence of functions $f_1, f_2, ...$ which converges to the optimal solution $f^* = \lim_{m\to\infty} f_m$, hence the convergence is point-wise or strong convergence which may lead to instabilities. In this work we instead consider weak convergence, which is more robust and exhibits faster rates of convergence. We start by showing that the conditional distribution of the risk given the probability of false discoveries is given by a Gibbs type measure. This provides a general framework for constructing objectives, once one decides on the loss function. Then any statistical learning algorithm can be constructed in terms of a sequence of measures $\rho_1, \rho_2, ... \in \mathscr{M}_+^1(\mathscr{H})$ such that

$$\mathbb{E}_{\rho_m}[f] \to \mathbb{E}_{\rho^*}[f]$$

where $\rho^* \in \mathscr{M}_+^1(\mathscr{H})$ is given by the Gibbs Measure i.e.

$$\rho^* = \arg\min_{\rho \in \mathscr{M}_+^1(\mathscr{H})} F_\beta[\rho]$$

where

$$F_\beta[\rho] = \mathbb{E}_\rho[U(f)] + \beta^{-1} H(\rho)$$

is known as the Free Energy functional with $\beta > 0$ (temperature) and $H(\rho)$ the entropy of ρ. Under Laplacian prior over the parameter space we show that this formulation reduces to the SLOPE regression, which was recently shown to control FDR at a given rate [34]. In this thesis, we apply this framework to a GWAS dataset and compare its FDR performance over LASSO regression.

Fourthly we look at the problem of multi-classification which control False Discoveries. We construct a general family of loss functions which can be used to construct Bayes optimal classifiers and showcase its performance on Diabetic Retinopathy Detection dataset, in which the task was to classify retinal image from patients into five categories which range from no signs of diabetes to proliferate levels. In this case we also employ the use of Convolutional Neural Networks to extract features from the images.

Finally we look at the problem of inferring the distribution of non-negative discrete random variables which are applied to the study of RNA-Seq data. By assuming infinite divisibility

of the underlying random variable, its characterization as being a discrete Compound Poisson Measure (DCP), is derived. This allows construction of a non-parametric Bayesian model of DCPs with a Pitman-Yor Mixture process prior, which is shown to allow for consistent inference under Kullback-Liebler and Renyi divergences especially in the under-sampled regime.

1.2 Metric Measure Spaces

1.2.1 Polish Space

Axiomatically defined mathematical objects like real numbers, functions, probability measures, etc may have an infinite length representation, like in the case of irrational numbers. However, when implementing an algorithm in practice, one can represent and manipulate only a finite number of operations in a finite amount of time. Hence when designing machine learning algorithms we need to restrict ourselves to the study of only those topological spaces, like those consisting of real numbers, probability measures, functions, etc that can be arbitrarily well approximated by finite representations i.e. be computable.

The concept of finite approximability or computability can be defined [2, 16, 65, 101, 123] in terms of the existence of a computational model, which is a directed sequence of partially ordered sets which are consistent, continuous and observable using a dense subset. To get an intuitive idea, lets consider the case of real numbers.

Since there exists a dense subset of real numbers, the set of rationals $\mathbb{Q}$, each real number $x \in \mathbb{R}$ can be identified as a collection of intervals, say $\{[p_i, q_i]\}_{i=1}^{\infty}$ where $p_i, q_i \in \mathbb{Q}$, such that $x = \limsup p_i$ and $x = \liminf q_i$. When such a sequence of intervals of decreasing length exist, then it is said that the elements of the real number line are *observable*. Now, if the intersection of these intervals defines the required real number $\{x\} = \cap_{i=1}^{\infty} [p_i, q_i]$, then they are said to be *consistent*. Further in such a representation, smaller intervals contain more information about the number, one is trying to approximate. So if $I_i := [p_i, q_i] \supset I_j := \left[p_j, q_j\right]$, then the interval I_j carries more information than the interval I_i, and we represent it by writing $I_i \leq I_j$. If for the sequence $I_1, I_2, \ldots$ we have $I_1 \leq I_2 \leq \cdots$ then the sequence of intervals is said to *continuous*. When all these conditions are satisfied by the sequence of intervals $\{[p_i, q_i]\}_{i=1}^{\infty}$, it is said to be a computational model for the real number x.

Generalizing these concepts to an abstract topological space, it turns out that that there exists a directed sequence of partially ordered sets which are consistent, continuous and observable using a dense subset, i.e. a computational model on a topological space if and only if it is a Polish space [77] i.e. a completely separable metrizable space. In this work we denote $\mathscr{X} := (X, \rho)$ as a Polish space, where X is a complete separable metric space along with the

associated metric, which is a bivariate function $\rho : X \times X \to \mathbb{R}_+$ such that for all $x, y \in X$, it vanishes if and only if $x = y$, is symmetric and satisfies the triangle inequality. Common examples of Polish spaces are $\mathbb{Z}^n$, $\mathbb{R}^n$, $\mathbb{C}^n$, any separable Banach Space, Hilbert space of functions, etc along with their natural metric (like the Hamming distance, Euclidean metric, etc), and encompass practically all examples one encounters in real world data sets.

Thus the first step in designing machine learning algorithms is to define an appropriate metric on the sample space of interest. It is important to note that, if the data set to be analyzed cannot be modeled as a Polish space, i.e. a metric can not be defined which induces a Polish topology. Then an appropriate computational model does not exist to approximate elements of that space and consequently a learning algorithm can not be defined which can approximate statistics, measures, functions, etc, a fact which is important to keep in mind while designing learning algorithms.

Example 1.1. In Natural Language Processing, conceptually it is hard to define a quantitive distance between words. However recent word vector embedding approaches have been extremely successful, as they approximate the words in a natural language in terms of embeddings in vector space, which is a Polish space. [90, 112]

1.2.2 Radon Measures

A measure is a non-negative function which is a generalization of the concept of size, like length or volume, for elements of the topology τ_X [1] of some topological space X, such that the size or measure of the union of disjoint sets in τ_X is the sum of their individual measures. Even though one would like to assign a measure to every element of the topology, in general this is not possible. For example consider length as the measure of subsets of the real line, then using the axiom of choice one can show there exist sets for which no size exists [52, 54]. The subsets whose length or size can be measured, are known as the measurable sets.

For a Polish metric space $\mathscr{X}$, this collection of measurable subsets are defined by the Borel σ-algebra $\mathscr{B}_{\mathscr{X}}$, which is the smallest family of subsets from τ_X which contains the empty set, contains all closed sets and is closed under countable unions, as well as their countable intersections and their relative complements. A non-negative measure $\mu : \mathscr{B}_{\mathscr{X}} \to \mathbb{R}_+$ can then be defined on $\mathscr{B}_{\mathscr{X}}$ to satisfy certain conditions, necessary for the application at hand.

While designing learning algorithms we need to deal with measures that are either discrete or continuous or both within the same framework. This is possible by relying on the set of

[1] A topology τ_X, is a family of subsets of some set X such that both the empty set and X are elements of τ_X, any union of elements of τ_X is an element of τ_X and any intersection of finitely many elements of τ_X is an element of τ_X.

non-negative Radon measures $\mathscr{M}_+(\mathscr{X})$ on the space $\mathscr{X}$. A Radon measure is a measure on $\mathscr{B}_{\mathscr{X}}$, which is inner regular (i.e. tight) [2], outer regular [3] and locally finite [4].

Examples of non-negative Radon measures include, all Borel Probability measures on a Polish spaces, Dirac measures on any topological space, Haar measure on any locally compact topological group, etc. Unlike the Lebesgue measure, a Radon measure on a single point is not necessarily of measure 0. This is particularly useful when working with Lévy processes which we use extensively in this work to construct a non-parametric Bayesian risk estimator in a later section.

Radon measures are also finitely approximable, and next we discuss how to construct a directed sequence of partially ordered sets which are consistent, continuous and observable using a dense subset. Inner and Outer regularity of Radon measures imply that the measure of any Borel set $A \in \mathscr{B}_{\mathscr{X}}$ can be lower and upper bounded via sequences of compact $K_1, K_2, \ldots \subset A \in \mathscr{B}_{\mathscr{X}}$ and open $A \subset U_1, U_2, \ldots \in \mathscr{B}_{\mathscr{X}}$ sets respectively. Let the limit supremum and limit infimum be defined as

$$\begin{aligned}\limsup_{n\to\infty} K_n &= \bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} K_m \\ &= \{x : x \in K_m \text{ infinitely often}\} = \{K_m \text{ i.o.}\}\end{aligned}$$

$$\begin{aligned}\liminf_{n\to\infty} U_n &= \bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} U_m \\ &= \{x : x \in U_m \text{ for all but finitely many m's}\}\end{aligned}$$

then clearly both of these sequences are *continuous*. Further, Fatou's lemma implies that for some measure $\mu \in \mathscr{M}_+(\mathscr{X})$

$$\liminf_{n\to\infty} \mu(U_n) \geq \mu(\liminf_{n\to\infty} U_n) \geq \mu(A) \geq \mu(\limsup_{n\to\infty} K_n) \geq \limsup_{n\to\infty} \mu(K_n)$$

then due to inner and outer regularity, both the outer and inner measures are equal, and hence we have

$$\liminf_{n\to\infty} \mu(U_n) = \limsup_{n\to\infty} \mu(K_n)$$

Therefore the sequence of compact and open sets are *consistent* and we can define the measure

[2] for any Borel set $B \in \mathscr{B}_{\mathscr{X}}$, $\mu(B)$ is the supremum of $\mu(K)$ over all compact subsets K of B
[3] for any Borel set $B \in \mathscr{B}_{\mathscr{X}}$, $\mu(B)$ is the infimum of $\mu(U)$ over all open sets U containing B
[4] if every point of $\mathscr{X}$ has a neighborhood U for which $\mu(U)$ is finite

of A as the limit of these sequence of sets as

$$\mu(A) \overset{\text{a.s.}}{=} \lim_{n\to\infty} \mu(U_n) = \lim_{n\to\infty} \mu(K_n)$$

Finally since $\mathscr{X}$ is Polish, there exists a countable dense subset $D = \{x_1, x_2, ...\} \subset A$ such that

$$K := \bigcap_{m=1}^{\infty} \bigcup_{k=1}^{n_m} B(x_k, 1/m)$$

where $B(x,\delta)$ is a ball of radius δ at x. Note that $\mu(B(x,\delta)) < \infty$ due to local finiteness. Then K is closed and for each $\delta > 0$ and $m > 1/\delta$,

$$K \subset \bigcup_{k=1}^{n_m} B(x_k, 1/m) \subset \bigcup_{k=1}^{n_m} B(x_k, \delta)$$

This means that K is covered by finitely many balls of radius less than or equal to $\delta > 0$. Then for any sequence in K, whose limit is in A (always true if $A \subset \mathscr{X}$ is complete), one can always construct a Cauchy subsequence with limit in K. Now since K is also closed, K is therefore compact. Then for any $\varepsilon > 0$ and each $m \geq 1$, there exists an n_m such that

$$\begin{aligned}\mu(A\backslash K) &= \mu\left(\bigcup_{m=1}^{\infty}\left(A\backslash\bigcup_{k=1}^{n_m} B(x_k, 1/m)\right)\right) \leq \sum_{m=1}^{\infty} \mu\left(A\backslash\bigcup_{k=1}^{n_m} B(x_k, 1/m)\right) \\ &= \sum_{m=1}^{\infty}\left(\mu(A) - \mu\left(\bigcup_{k=1}^{n_m} B(x_k, 1/m)\right)\right) < \sum_{m=1}^{\infty} 2^{-m}\varepsilon = \varepsilon\end{aligned}$$

This means that for any $\varepsilon > 0$ and every $\lambda \geq 1$, there exists an $n(\lambda,\varepsilon) < \infty$, such that

$$K_{\lambda,\varepsilon} = \bigcup_{k=1}^{n(\lambda,\varepsilon)} B\left(x_k, \frac{1}{\log\lambda}\right) \implies \mu(K_{\lambda,\varepsilon}) > \mu(A) - \lambda^{-1}\varepsilon$$

for some $x_1, x_2, ..., x_{n(\lambda,\varepsilon)} \overset{\text{i.i.d.}}{\sim} \mu(A)$. Thus there exists a dense subset of balls which can be used to define the required directed sequence of partially ordered sets and therefore a computational model exists. As we saw previously, this means that Radon measures on a Polish space are finitely approximable and we can define a metric on $\mathscr{M}_+(\mathscr{X})$ which makes it as well into a Polish space.

1.2.2.1 Integral Probability Metrics

Let $\mu, \nu \in \mathscr{M}_+(\mathscr{X})$ be a pair of non-negative Radon measures, then the integral probability metrics (IPMs) [109, 110], are distances between measures that have a variational form and can be written as a supremum over the mean discrepancies of functions restricted to a specific function class i.e.

$$\rho_{\mathscr{F}}(\mu, \nu) \coloneqq \sup_{h \in \mathscr{F}} \mid \Delta(h, \mu, \nu) \mid$$

where $\mathscr{F}$ is a class of real-valued bounded measurable functions on $\mathscr{X}$ and $\Delta : \mathscr{F} \to \mathbb{R}$ is the mean discrepancy.

Intuitively an IPM between two measures, looks for a witness function h, called the critic, which maximally discriminates between the two measures according to a certain mean discrepancy Δ. By defining different classes of functions $\mathscr{F}$, from which the critic comes from and different mean discrepancies one defines different IPMs and in certain cases the above variational formulation has a closed form expression.

The reason we are interested in studying a wide variety of distances, as we shall soon see that this family generates, is so as to study different notions of convergence of sequences of measures. A sequence of measures $\{\mu_n\}_{n \in \mathbb{N}}$ converges if and only if there is a distribution μ_∞ such that $\rho_{\mathscr{F}}(\mu_n, \mu_\infty)$ tends to zero. The topology of associated convergence depends on the metric ρ and in specific cases one might require a weaker or a stronger topology so that it easier or harder, respectively for a sequence of distribution to converge.

In a learning algorithm these sequences of measures are usually parameterized, say by the corresponding sequence of parameters $\{\theta_n\}_{n \in \mathbb{N}}$. Then the mapping $\theta_n \to \mu_{\theta_n}$ is said to be continuous if when $\theta_n \to \theta$ then $\mu_{\theta_n} \to \mu_\theta$ which is clearly a desirable property for any learning algorithm. Now since the notion of convergence of the sequences of measures depends on the metric considered between them, weaker the metric, easier it is to define a continuous map between the parameters and the sequences of measures. Thus for the machine learning problem of interest we would like to choose such families of measures which are continuous with respect to the parameters for the particular metric of interest.

φ-Divergences Csiszár's φ-Divergences [25, 37, 71, 75, 83, 86, 127] are one of the most common families of distances/divergences between probability measures. In this case the class of critics or witness functions is given by

$$\mathscr{F} = \left\{ f : \mathscr{X} \to \mathbb{R}, f \in \mathrm{dom}_{\varphi^*} \right\}$$

and mean discrepancy is given by

$$\Delta(h,\nu,\mu) = \int_{\mathscr{X}} h d\mu - \int_{\mathscr{X}} \varphi^* \circ h d\nu$$

where φ^* is the Fenchel conjugate of φ. In this case the divergence has a closed form representation as

$$\rho_{\mathscr{F}}(\mu,\nu) := \begin{cases} \int_{\mathscr{X}} \varphi\left(\frac{d\nu}{d\mu}\right) d\mu & \nu \ll \mu \\ +\infty & \text{otherwise} \end{cases}$$

where $\frac{d\nu}{d\mu}$ is the Radon-Nikodym derivative which is defined for both continuous and discrete measures. Well-known distance/divergence measures obtained by appropriately choosing φ, for example

1. Kullback-Liebler (KL) divergence:

$$\varphi(t) = t \log t$$

2. Hellinger distance:

$$\varphi(t) = \left(\sqrt{t}-1\right)^2$$

3. Total Variation Distance:

$$\varphi(t) = \mid t-1 \mid$$

4. χ^2-divergence:

$$\varphi(t) = (t-1)^2$$

Wasserstein-p Metric In this case [12, 22, 43, 53] the class of critics or witness functions is given by the space of all functions with bounded Lipschitz constants

$$\mathscr{F} = \left\{ f : \mathscr{X} \to \mathbb{R}, \parallel f \parallel_{\text{Lip}} \leq 1 \right\}$$

and mean discrepancy is given by

$$\Delta(h,\nu,\mu) = \int_{\mathscr{X}} h d\mu - \int_{\mathscr{X}} h d\nu$$

In this case the distance has a closed form representation in terms of a Wasserstein-1 metric

$$W_1(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X}} \parallel x-y \parallel_1 d\pi(x,y)$$

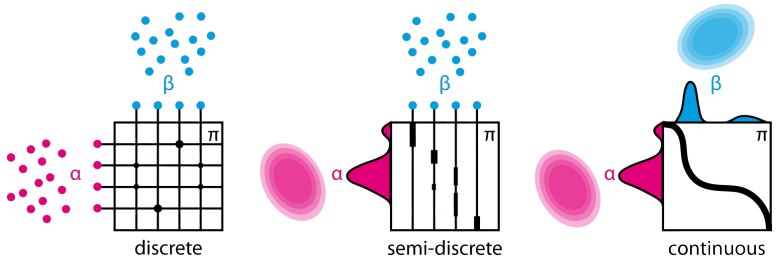

Figure 1.1: Wasserstein Metric

where $\Pi(\mu,\nu)$ is the space of all joint measures defined on $\mathscr{X}\times\mathscr{X}$ with μ and ν as marginals. However, in general a Wasserstein-p metric for $p\geq 1$ can be defined as

$$W_p(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}} \| x-y \|_p^p \, d\pi(x,y)$$

MMD Maximum Mean Discrepancy [59] is a metric between measures defined using kernel mean map in a Reproducing Kernel Hilbert Space (RKHS). In this case the class of critics or witness functions is given by the space of all functions with bounded norm in the RKHS generated by a characteristic kernel K,

$$\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R}, \| f \|_{\mathscr{H}_K} \leq 1\}$$

and mean discrepancy is given by

$$\begin{aligned}\Delta(h,\nu,\mu) &= \int_{\mathscr{X}} h d\mu - \int_{\mathscr{X}} h d\nu \\ &= \langle h, \int_{\mathscr{X}} K(\cdot,x) d\mu \rangle_{\mathscr{H}_K} - \langle h, \int_{\mathscr{X}} K(\cdot,x) d\nu \rangle_{\mathscr{H}_K}\end{aligned}$$

In this case the distance has a closed form representation in terms of mean kernel maps

$$\mathrm{MMD}(\mu,\nu) := \| \mathbb{E}_\mu [K_x] - \mathbb{E}_\nu [K_x] \|_{\mathscr{H}_K}$$

Depending on the Kernel map K used, MMD defines a very general class of L^2 losses.

Stein Discrepancy In this case the class [20, 22] of critics or witness functions is given by smooth functions which vanish at the boundary

$$\mathscr{F} = \left\{ f : \mathscr{X} \to \mathbb{R}^d, f \in \mathscr{C}(\mathscr{X}), \partial f = 0 \right\}$$

and mean discrepancy is given by

$$\Delta(h, \nu, \mu) = \int_{\mathscr{X}} T(d\mu) h d\nu$$

where

$$T(d\mu) = (\nabla_x \log(d\mu))^T + \nabla_x$$

and μ and ν are continuous differentiable measures. In this case the distance does not have a closed form solution. Add the closed form solution in case of a RKHS.

1.3 Hypothesis Space

In any machine learning problem we are given a sample space $(\mathscr{X}, \mu)$ where $\mu \in \mathscr{M}_+(\mathscr{X})$ and an associated target space $(\mathscr{Y}, \nu)$ with $\nu \in \mathscr{M}_+(\mathscr{Y})$, containing the possible decisions, predictions, etc that we are interested in making. For example, in a binary classification problem we might have samples from $(\mathscr{X} = \left(\mathbb{R}^d, \text{Euc}\right), \mu)$ for some probability measure $\mu \in \mathscr{M}_+^1(\mathscr{X})$ and we would like to find a map which chooses one of the elements of the binary space $(\mathscr{Y} = (H_0 = \mu_0, H_1 = \mu_1), \nu = \text{Bin}(p, 1-p))$ depending on its distance to μ. In a Regression problem, we usually have paired samples from $(\mathscr{X} = \left(\mathbb{R}^d, \text{Euc}\right), \mu)$ and $(\mathscr{Y} = (\mathbb{R}, \text{Euc}), \nu)$ where $\mu, \nu \in \mathscr{M}_+^1(\mathscr{X})$, and the learning task is to find a function which transforms $\mu \to \nu$.

In both examples, we see that the learning task can be defined in terms of a map from one space to the other. In this work we define the space of all such functions of the type $f : \mathscr{X} \to \mathscr{Y}$ which map elements from the measurable space of $\mathscr{X}$ to that of $\mathscr{Y}$ as the Hypothesis space, $\mathscr{H}$. Such functions induce a push-forward operator $f_\sharp$ transforming an entire measure on $\mathscr{X}$ to a new Radon measure on $\mathscr{Y}$ i.e.

$$f_\sharp : (\mathscr{X}, \mu) \mapsto (\mathscr{Y}, \nu)$$

where $\nu = f_\sharp \mu \in \mathscr{M}_+(\mathscr{Y})$ is called the push-forward measure for every $\mu \in \mathscr{M}_+(\mathscr{X})$.

A push forward measure can be defined as a Radon measure $\nu \in \mathscr{M}_+(\mathscr{Y})$ which satisfies

$$\int_{f^{-1}(B)} h \circ f d\mu = \int_B h d\nu \tag{1.1}$$

for any $B \in \mathscr{B}_{\mathscr{Y}}$ and for all $h \in L^1(\mathscr{Y})$, the space of Lebesgue integrable functions on $\mathscr{Y}$.

Representing complicated functions as superpositions of basic transforms of simpler functions has been a subject of study in Harmonic analysis since the introduction of the Fourier transform, centuries ago. Such a representation allows us to extract information from observed

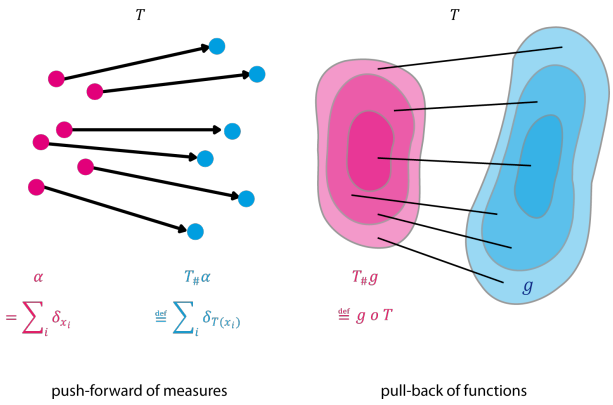

Figure 1.2: Push-Forward Measures

signals or functions by transforming the function from its original domain into a new domain, with the purpose of extracting the characteristic information which is otherwise not readily observable in its original form.

Formally this means that we define a Hypothesis space as a space of functions where for any element say $f \in \mathscr{H}$, we can extract a series of coefficients, based on the inner product between the function $f(x)$ and a set called the dictionary, consisting of template functions $\Phi(x) := \{\phi_\lambda(x)\}_{\lambda \in \Lambda}$ as

$$\theta_\lambda = \langle f, \psi_\lambda \rangle := \int_{\mathscr{X}} f(x) \phi_\lambda^*(x) dx$$

where $(\cdot)^*$ stands for the complex conjugate. The inner product in essence describes the "similarity" between $f(x)$ and the dictionary $\{\phi_\lambda(x)\}_{\lambda \in \Lambda}$, where a higher inner product signifies a higher similarity between the function and that particular element of the dictionary. These coefficients represent features, that can be extracted from the given data using the template functions.

Such representations under certain assumptions on the template functions $\Phi(x)$, also admit a Calderon-type reproducing [41] formula given by

$$f(x) = \int_\Lambda \theta_\lambda \phi_\lambda(x) d\lambda =: \langle \theta, \Phi(x) \rangle \tag{1.2}$$

where Λ is the parameter space spanning the dictionary.

Thus by studying properties of these template functions we can construct a variety of new representations and push-forward maps for our machine learning application. However before we do that, we introduce an alternate approach to modeling the push-forward measure, which is useful especially in probabilistic modeling approaches.

Consider the definition of a push-forward measure from a different perspective. For any

$\mu \in \mathscr{M}_+(\mathscr{X})$ we have $\nu \in \mathscr{M}_+(\mathscr{Y})$ such that for some $f \in \mathscr{H}$

$$\nu(B) = \mu\left(\{x \in \mathscr{X} : f(x) \in B\}\right) = \mu\left(f^{-1}(B)\right)$$

for any Borel set $B \in \mathscr{B}_{\mathscr{Y}}$ where $f^{-1}(B) \subset \mathscr{B}_{\mathscr{X}}$. It is easy to see that $f_\sharp$ preserves positivity and total mass of the measure, so that if $\mu \in \mathscr{M}_+^1(\mathscr{X})$ is a Borel probability measure, then so is $f_\sharp\mu \in \mathscr{M}_+^1(\mathscr{Y})$.

Now the push-forward measure ν is absolutely continuous with respect to μ, since for any $B \in \mathscr{B}_{\mathscr{Y}}$

$$\mu\left(f^{-1}(B)\right) = 0 \implies \nu(B) = 0$$

which by the Radon-Nikodym theorem means that a measurable function $g : \mathscr{X} \to \mathbb{R}_+$ exists, such that

$$\nu(B) = \int_B d\nu = \int_{f^{-1}(B)} g d\mu$$

where $g := \frac{d\nu}{d\mu}$ is known as the Radon-Nikodym derivative of ν w.r.t. μ.

The Radon-Nikodym derivative is defined for both continuous and discrete Radon measures, and can therefore either represent a measure density or a discrete measure respectively. Therefore defining classes of push-forward maps is equivalent to considering families of induced Radon-Nikodym densities on $\mathscr{X}$, which are independent of the base measure μ.

Now naturally we want our Radon-Nikodym densities to be finitely approximable, which means that it should be a function of either bounded or slowly (logarithmic or polynomially slow) increasing number of features, as the sample size increases. The set of features or more formally statistics, calculated from the observed data is said to be sufficient if the conditional expectation given these statistics, is independent of the observed data. Thus intuitively, sufficient statistics capture all the necessary statistical information from observed data in order to define the conditional density.

According to the Pitman–Koopman–Darmois theorem [9, 49, 87, 102], the Exponential family of distributions is exactly that family for which the dimension of the sufficient statistic (or features) remains bounded as the sample size increases, while the domain remaining fixed for the parameters being estimated. The Exponential family in addition is also the family of distributions with maximum-entropy, under given constraints on the expected values of these sufficient statistics (or features).

Thus if we assume that only a bounded number of features from the sample effects the the target measure, then the relative Radon-Nikodym density of the push-forward measure can be

defined as a member of the Exponential family by

$$\frac{d\nu}{d\mu} := e^{\langle \theta, \Phi(x) \rangle - A(\theta)} \tag{1.3}$$

where

$$A(\theta) = \log \left(\int_{\mathscr{X}} \exp \left(\langle \theta, \Phi(x) \rangle \right) d\mu \right) < \infty$$

and $f(x) = \langle \theta, \Phi(x) \rangle \in \mathscr{H}$ as before (1.2).

The space of values of $\theta \in \mathscr{H}$ for which $A(\theta) < \infty$, is known as the natural parameter space and is always convex. The finiteness of $A(\theta)$ represents the first assumption that we make on the elements in the Hypothesis space. The function $A(\theta)$ is the known as the log-moment or cumulant generating function and all the moments of the push-forward measure can be derived simply by differentiating $A(\theta)$.

To summarize we have shown there two different approaches to use a defined Hypothesis space. In the first case we use the elements in the space to model possible push-forward maps that we are interested in, while in the other case we use these elements to define an exponential family of relative push-forward densities. In either case the Hypothesis space is defined based on a sequence of template functions which extracts features or sufficient statistics from the sample data. Therefore while designing new Hypothesis spaces, the only freedom remaining from the users perspective is to design such template s or data representations which in different contexts generates features, sufficient statistics or in the infinite dimensional case, the basis for functional spaces.

1.3.1 Data Representations

Choice of the set of template functions or the data representation is extremely important in the performance of the developed machine learning algorithm. Hence for complex data sets, much effort needs to be put into designing preprocessing pipelines and templates which can extract relevant features for a specific application.

However, with the advent of complex large data sets like in Natural Language Processing, Image Processing, Genetics, etc extracting new novel features has become extremely important in order to perform complicated tasks. However, this aspect is rather labor-intensive and highlights the weakness of fixed basis functions: their inability to extract and organize the discriminative information from the data. Therefore in many cases handmade features using prior expert knowledge has proven to be useful in many cases.

Historically in the development of most learning algorithms, fixed template functions like Linear, Fourier, Wavelet and Kernel basis were used with much success. However, with in-

creasing access large amounts of different data sets, it has become an imperative to expand the flexibility of these feature extractors by learning them directly from the data itself. The newest class of such representations are known as the deep learning methods which are formed by the composition of multiple non-linear transformations, with the goal of yielding more abstract – and ultimately more useful – representations. This allows the algorithm to become less dependent on human input and biases, thereby allowing novel applications.

Therefore it is extremely important to learn representations of the data that allow easier extraction of useful information while building Hypothesis spaces as described earlier. In the case of probabilistic models, which in our case is modeled by conditional exponential family of densities, good representations captures the posterior distribution of the underlying explanatory factors for the observed input.

While designing or learning Hypothesis spaces we still need to make certain basic assumptions which allow us to define a plausible model for these spaces of functions. In general all function spaces considered in this work consist of vector valued square integrable functions $L^2(\mathscr{X})$. However this space is quite big and further assumptions are necessary in order to design a subset of well behaved functions. We enumerate some of these assumptions and the models they lead to in the next subsections.

1.3.1.1 Smoothness

Degree of smoothness is to do with the degree of differentiability that can be assumed for the functions in the Hypothesis space. The most non-trivial model of a Hypothesis space with smoothness as the main characteristic is the Reproducing Kernel Hilbert Space.

Let (X,ρ,μ) be a Polish metric measure space of negative type. Polish spaces of negative type characterize those metric spaces which can be isometrically imbedded into a Hilbert space. In [5, 14] metric induced kernels were introduced, where for every $\rho \in L^1(X\times X,\mu\times\mu;\mathbb{R}_+)$, and any $x_0 \in X$, the symmetric integrable function $\varphi \in L^1(X\times X,\mu\times\mu;\mathbb{R}_+)$ defined as

$$\varphi(x,y) = \frac{1}{2}\left(\rho(x,x_0)+\rho(y,x_0)-\rho(x,y)\right)$$

is positive definite if and only if ρ is negative definite. Then from Moore-Aronszajn's theorem, for every such positive definite function φ, there exists a unique Hilbert space $\mathscr{H}_\varphi$ of functions on X for which φ is a reproducing kernel. This means that the mapping $x \mapsto \varphi_x$ from X to $\mathscr{H}_\varphi$ is injective and defines the inner product

$$\varphi(x,y) = \langle \varphi_x, \varphi_y \rangle_{\mathscr{H}_\varphi} = \int_X \varphi_x(z)\varphi_y(z)dz \quad \text{for all } x,y \in X$$

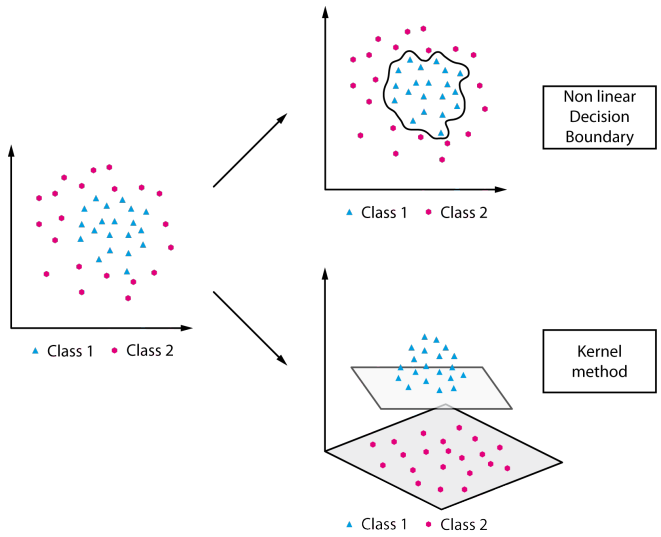

Figure 1.3: Reproducing Kernel Hilbert Space Classification

on $\mathscr{H}_\varphi$ using the Reisz representation theorem. Further for all $f \in \mathscr{H}_\varphi$, we have

$$\langle f, \varphi_x \rangle_{\mathscr{H}_\varphi} = f(x)$$

which is known as the reproducing property. The Hilbert space $\mathscr{H}_\varphi$, by the reproducing property consists of real valued functions of the form

$$f(x) = \langle \sum_i \theta_i \varphi_{x_i}, \varphi_x \rangle = \sum_i \theta_i \varphi(x_i, x) \text{ for } \theta_i \in \mathbb{R}, i \in \mathbb{N}$$

where $\mathscr{X}_n = \{x_1, \ldots, x_n\} \overset{\text{i.i.d.}}{\sim} \mu$ such that $\sum_i \theta_i^2 \varphi(x_i, x_i) < \infty$. Thus in this case the feature map is given by $\Phi(x) = [\{\varphi(x_i, x)\}_{i=1}^n]^T$.

By choosing different metrics on the Polish space or equivalently different kernel functions φ we get different models of RKHS. For example for $\varphi(x,y) = \langle x, y \rangle$, we get the linear kernel and the Hypothesis space corresponds to all linear models. Note that in this case the associated metric on the Polish space is simply the Euclidean metric. In the case when $\varphi(x,y) = \exp -\frac{(x-y)^2}{2}$, we have the popular Gaussian Kernel and the functions defined in the corresponding Hypothesis space are infinitely differentiable.

In the RKHS model of Hypothesis space, generalization is achieved via local interpolation between neighboring training examples. Although smoothness is an useful assumption, it is insufficient to deal with the curse of dimensionality, especially when there are discontinuities in the target function. Since in that case the number of non-zero coefficients may grow quadratically with the number of samples. Therefore learning algorithms based only on the

smoothness assumption exploit the principle of local generalization and rely only on examples to explicitly map out the discontinuities of the target function.

For complicated target functions with many discontinuities, such RKHS models cannot capture enough of the complexity of interest, unless provided with the appropriate feature space. Therefore there is a need for flexible and non-parametric models of Hypothesis spaces which do not rely exclusively on the smoothness assumption.

However the smoothness-based models are still useful on top of features learned from more complicated models. They provide a modular approach to applying well established algorithms using novel feature learning algorithms. For example a combination of using a deep neural network to learn features and applying RKHS based algorithms in this feature space is equivalent to learning the kernel φ.

1.3.1.2 Distributed representations

We would also like the a reasonably sized set of features extracted to be expressive, that is explain a sizable portion of the variation. These features should also generalize to a huge number of possible input configurations. For example in a clustering application, a counting argument helps us assess the expressiveness of the learned features. Traditional clustering algorithms like Gaussian mixtures, nearest-neighbor algorithms, decision trees, or Gaussian Support Vector Machines all require $O(N)$ parameters (and/or $O(N)$ examples) to distinguish $O(N)$ input configurations and naively it seems like a reasonably tight result.

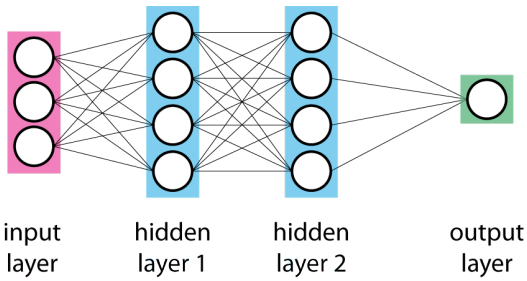

Figure 1.4: Restricted Boltzman Machines

However, up to $O(2^k)$ input regions (k is the sparsity parameter) can be represented by Restricted Boltzmann Machines (RBMs) [51], sparse coding [4], auto-encoders or multi-layer neural networks using only $O(N)$ parameters. These are all distributed or sparse representations. The generalization of clustering to distributed representations is known as multi-clustering, where either several clusterings take place in parallel or the same clustering is applied on different parts of the input.

The exponential gain comes about because each feature can be re-used in multiple examples that are not necessarily neighbors of each other, which is the case with local generalization. In all the classical single layer approaches different regions of the input space are essentially independent with their own private set of parameters, e.g., as in decision trees, nearest-neighbors, Gaussian SVMs, k-means, etc.

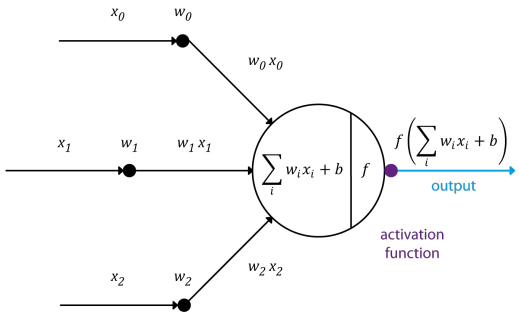

Figure 1.5: Neuron

1.3.1.3 Depth and abstraction

We saw how distributed representations provide an exponential increase in the capacity for representation of features. In deep representations the explanatory factors or features are in addition hierarchically organized, i.e. more abstract features are constructed using less abstract ones. This architecture promotes the re-use of features and leads to the construction of progressively more abstract features at higher layers of representation.

The notion of feature re-use, the main explanation behind the power of distributed representations, also explains the advantages behind deep learning, i.e., constructing multiple levels of representation or learning a hierarchy of features. Crucially deep representations have exponentially large number of paths from the data to the final feature, with respect to its depth.

Typically deep representations consist of a sequence of nodes which typically consist of computations like weighted sum, product, affine transforms, monotonic point wise non-linearity, computation of a kernel, or logic gates. Theoretically families of functions defined using a deep representation can be exponentially more efficient than one that is insufficiently deep [82, 85, 91].

Clearly when the same family of functions can be represented using fewer features, we should expect to be able to learn the parameters using fewer examples, yielding improvements in both computational and statistical efficiency (less parameters to learn, and re-use of these parameters over many different kinds of inputs). One of the most common and successful examples of deep representations is known as the Convolutional Neural Network (CNNs).

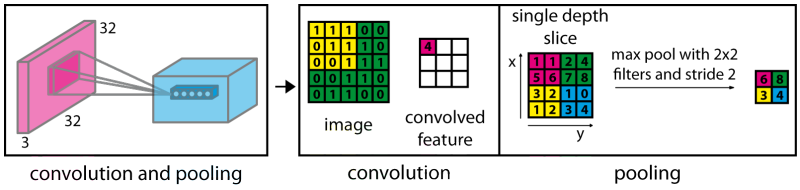

Figure 1.6: Convolutional and Pooling Layers

CNNs consist of a sequence of nodes which are made up of a convolution by a tensor (3 dimensional for color images) filter, followed by a point wise non-linearity, like ReLU which simply filters out all non-positive values. In these networks one builds abstractions explicitly, via a pooling mechanism. More abstract concepts are generally invariant to most local changes of the input. That makes the representations that capture these concepts generally highly non-linear functions of the raw input.

1.3.1.4 Invariance and Disentangling Factors of Variation

When constructing deep representations we would like the features generated to define explanatory factors towards the variation seen in the data. Since many complex real world data sets arise from complex interactions of essentially independent sources of variation, we would like our features also to be as disentangled as possible. This goal is different from a related distinct goal of learning invariant features which reduce sensitivity in the direction of invariance. Thus invariant features remove uninformative information from the data set, i.e. which do not contribute significantly to the observed variation.

Clearly it is often difficult to determine a priori which set of features and variations will ultimately be relevant to the task at hand. Further, as is often the case in the context of deep learning methods, the feature set being trained may be destined to be used in multiple tasks that may have distinct subsets of relevant features.

However, in many real world high dimensional data sets, the probability mass concentrates on a manifold of much lower dimensionality. The Johnson–Lindenstrauss lemma [70] states that a small set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved. The map used is at least Lipschitz, and could even be taken to be an orthogonal projection. This fact is explicitly exploited in some of the auto-encoder algorithms [28, 73] and other manifold-inspired algorithms.

Lower dimensional embeddings imply that local variations in these manifolds tend to preserve the categorical variables associated to them in a classification problem and regions between different classes tend to be well separated and not overlap much. This fact is exploited

in high dimensional data visualization algorithms like the t-SNE.

Considerations such as these lead us to the conclusion that the most robust approach to feature learning should be perform some form of dimensionality reduction, especially for high dimensional data sets so that the local directions of variation least represented in the training data should be first to be pruned out. Then the aim should be to disentangle as many factors as possible, discarding as little information about the data as is practical.

1.4 Risk Minimization Framework

In the previous section we saw the properties which depend on the application, we need to take into consideration, while designing the Hypothesis space. In this section we want to establish a framework for defining performance objectives which we could use to define a learning algorithm. In general we quantify the performance of an element in the Hypothesis space using a non-negative real valued functional, known as the Risk functional

$$R : \mathscr{H} \times ((\mathscr{X}, \mu), (\mathscr{Y}, \nu)) \to \mathbb{R}^+$$

In general the risk can be defined in terms of the distances between the push-forward and the target measure as we shall see both in the supervised and unsupervised learning scenarios. This allows us to then define a learning problem in terms of the following optimization problem

$$f^* = \arg\min_{f \in \mathscr{H}} R(f, \mu, \nu)$$

If a solution exists then we can say the required task is learnable in the given Hypothesis space and the optimization algorithm which solves the problem is then known as the learning algorithm. A "learning algorithm" therefore is the iterative optimization algorithm which constructs a sequence of functions $f_1, f_2, \ldots$ which converges to the optimal solution $f^* = \lim_{m \to \infty} f_m$. One can then quantify the performance of such a learning algorithm in terms of its rate of convergence, complexity, etc.

When the risk functional is a convex function, there exists many of the shelf convex optimization algorithms which can be applied in an essentially black-box manner to solve the required learning problem. For example in the case of low dimensional problems, the second order Newton and Quasi-Newton methods provides a provably quadratic convergence to the correct solution, with acceptable runtimes even with polynomial computational complexity. In the case of high dimensional problems however, one usually settles for the class of first order gradient descent methods, which usually have first order convergence but also linear time

complexity.

1.4.1 Supervised Learning

In a supervised learning framework we have access to paired samples from both $(\mathscr{X},\mu)$ as well as $(\mathscr{Y},\nu)$. Then the aim of the algorithm is to find a push-forward map f, whose action on μ induces a push-forward measure closest to ν. Previously we defined a class of integral probability metrics which include a wide variety of commonly known metrics between measures, and therefore can be used to define a general class of risk functionals between the push-forward measure $f_\sharp\mu$ and the target measure ν

$$R(f,\mu,\nu) \coloneqq \rho_{\mathscr{F}}\left(f_\sharp\mu,\nu\right) \coloneqq \sup_{h\in\mathscr{F}} \mid \Delta(h,f_\sharp\mu,\nu) \mid$$

As we saw before, different choices of the functional classes $\mathscr{F}$ and mean discrepancy Δ, lead to different risk functionals. For example consider the φ-Divergences which defines the risk functional by

$$R(f,\mu,\nu) = \rho_\varphi\left(f_\sharp\mu,\nu\right) \coloneqq \begin{cases} \int_{\mathscr{X}} \varphi\left(\frac{d\nu}{df_\sharp\mu}\right) df_\sharp\mu & \nu \ll f_\sharp\mu \\ +\infty & \text{otherwise} \end{cases}$$

for different choices of the function φ.

These divergences are especially appropriate in the case where we have a model for the target distribution, for example in the case of binary classification. In this case, the target space $(\mathscr{Y}=\{0,1\},\nu)$ is binary and the task is to construct a map based on the training data, which assigns a category to each new observation in $(\mathscr{X},\mu)$. In this case the target measure ν, is simply the Bernoulli distribution conditional on the measure μ in the sample space $(\mathscr{X},\mu)$ i.e.

$$d\nu = p(x)d\mu$$

where $p(x)$ represents the observed conditional distribution. Then by defining the Hypothesis space $\mathscr{H}$ such that it induces a Bernoulli conditional density, we have

$$df_\sharp\mu = \frac{1}{1+e^{-\langle\theta,\Phi(x)\rangle}}d\mu$$

When we choose $\varphi(t) = t\log t$, i.e. the KL divergence as the Risk functional we get

$$\begin{aligned}\mathrm{KL}(f_\sharp\mu,\nu) = &-\int_{\mathscr{X}} p(x)\log\frac{1}{1+e^{-\langle\theta,\Phi(x)\rangle}}d\mu(x)\\ &-\int_{\mathscr{X}}(1-p(x))\log\left(1-\frac{1}{1+e^{-\langle\theta,\Phi(x)\rangle}}\right)d\mu(x)\\ &+\int_{\mathscr{X}} p(x)\log p(x)d\mu(x)\end{aligned}$$

which is the traditional risk functional for binary classification used in many modern applications like for image classification, where the data representation Φ is constructed using Convolutional Neural Networks [79, 91]. It is important to note that the integrals clearly need to be estimate based on samples from μ.

Alternatively, if we are looking at the regression problem, then the general class of Wasserstein distances might be preferable, where the risk functional is given by

$$R(f,\mu,\nu) = W_p\left(f_\sharp\mu,\nu\right) \coloneqq \inf_{\pi\in\Pi(f_\sharp\mu,\nu)}\int_{\mathscr{X}\times\mathscr{Y}} \| y-f(x)\|_p^p\, d\pi(f(x),y)$$

where $\Pi(f_\sharp\mu,\nu)$ is the space of all joint measures defined on $\mathscr{Y}\times\mathscr{Y}$ with $f_\sharp\mu$ and ν as marginals. For example the case of $p=2$ gives us the common mean squared loss. In the case of $p=1$ if we apply a strong entropy regularizer on the joint distribution, then the distance can be shown to converge to the energy distance. Further, if we define the inner product using of a positive definite kernel, then we end up with the MMD distance, allowing us to perform even kernel regression in the same framework.

1.4.2 Unsupervised Learning

1.4.2.1 Auto-Encoder Framework

A large class of unsupervised learning algorithm can be described in terms of the Auto-Encoder framework [7, 28]. These algorithms only have access to samples $\mathscr{X}_n$ from $(\mathscr{X},\mu)$ and we are interested in a push-forward map f, whose action on μ induces a push-forward measure with a relative density $\frac{df_\sharp\mu}{d\mu}$, in either the Exponential family or their mixtures, which is closest to μ based on an appropriate metric. This formulation is simply a projection of the arbitrary measure μ onto the Exponential Family of distributions. Such a push-forward map in known as the encoder.

From the encoded distribution, there is a pull-back map, say g which pulls the measure $f_\sharp\mu$ back to μ, i.e. $g_\sharp f_\sharp\mu \to \mu$. However the functions are parameterized such that they are not invertible, which is achieved by regularizing the parameters of the functions to be for example

have lower dimension or be sparse. When these functions are linear but not invertible, we get the Principle Components Analysis algorithm, which if they are further forced to be sparse we get what is known as sparse coding. In general the class of φ-Divergences are particularly useful in deriving many popular algorithms as special cases in this framework.

1.4.2.2 Generative Models

Generative models are a new class of unsupervised algorithms [11, 18, 46] which can be defined in terms of a latent variable formulation. Let η be a fixed measure on some continuous latent space $\mathscr{Z}$, then the aim of the algorithm is to learn a push-forward map f, whose action on η induces a push-forward measure closest to μ i.e.

$$f_\sharp \eta \to \mu$$

These class of models are known as Generative Adversarial models when the distance between them is defined in terms of a discriminator network which is also learned from the data. They have proven to be extremely successful in sampling from rather complicated distributions like images, audio, etc.

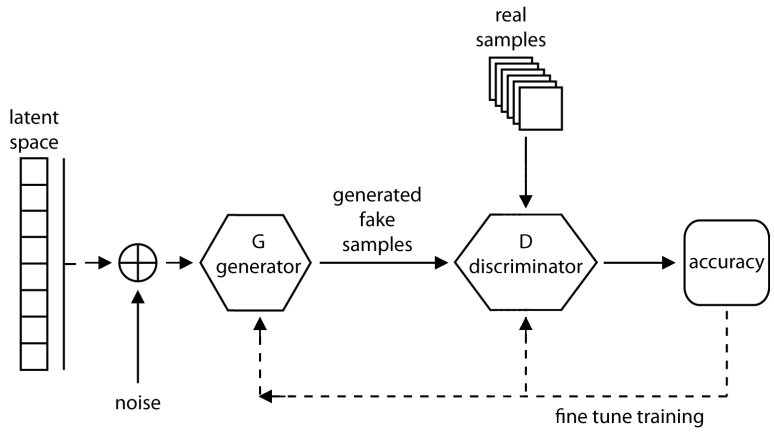

Figure 1.7: Generative Adversarial Network

Variational Auto-Encoders (VAEs) [73] and Generative Adversarial Networks (GANs) [103] are well known examples of this approach. Because VAEs focus on the approximate likelihood of the examples, they share the limitation of the standard models and need to fiddle with additional noise terms. GANs offer much more flexibility in the definition of the objective function, including φ-Divergences, and the Wasserstein Divergences as well as some exotic combinations. On the other hand, training GANs is well known for being delicate and unstable, for reasons theoretically investigated in [15].

1.5 Risk Estimation

Till now we saw how to define a learning problem, where start by constructing a Hypothesis space which is suitable for our application, either based on prior knowledge or theoretical understanding of the problem. In some cases a simple linear model is enough while for others even the cutting edge Convolutional Neural Networks struggles to achieve acceptable performance. We also saw that generative adversarial models, use families of metrics and choose the one which maximizes the discriminative power between functions in the Hypothesis space. This approach has achieved great sampling performance for extremely complicated distributions, like images, audio, etc.

However, in all these cases, from the classical linear regression to the state of the art generative adversarial networks, the risk functional always needs to be estimated based on the observed samples $\mathscr{X}_n$ and $\mathscr{Y}_n$ from $(\mathscr{X},\mu)$ and $(\mathscr{Y},\nu)$ respectively, whose estimator we denote by $\widehat{R}(f)$.

Since in general our risk is defined in terms of the mean discrepancy function Δ, which itself can be written in terms of being the expected value of a certain loss function L

$$\Delta = \mathbb{E}_{\mu,\nu}\left[L(f,x,y)\right]$$

whose push-forward measure is defined for each $f \in \mathscr{H}$ as

$$L_\sharp(f) : (\mathscr{X},\mu)\times(\mathscr{Y},\nu) \to \left(\mathbb{R}_+,\lambda_f\right)$$

where $\lambda_f \in \mathscr{M}_+(\mathbb{R}_+)$. Then assuming that we have observed i.i.d. samples $\mathscr{X}_n$ and $\mathscr{Y}_n$, we can construct i.i.d. samples from the push-forward measure η_f

$$\mathscr{L}_n(f) \coloneqq \{l_i = L(f,x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (\mathbb{R}_+,\lambda_f)$$

where $l = L(f,x,y)$ and the risk associated with the function $f \in \mathscr{H}$ can then be defined as the first moment of the push-forward measure λ_f

$$R(f) \coloneqq \int_{\mathscr{X}} L(f,x,y)d\left(\mu\times\nu\right) = \int_{\mathbb{R}_+} l\lambda_f(dl) < \infty$$

If no functions exist in the space of Hypothesis functions with finite risk, then either the Loss function, the Hypothesis space $\mathscr{H}$ or the problem itself is ill posed.

1.5.1 Generalization Error

Consider a learning algorithm which generates a sequence of functions $f_1, f_2, \ldots \in \mathscr{H}$ in the Hypothesis space which converges to $f^* = \lim_{m\to\infty} f_m$ where $f^* = \arg\inf_{f\in\mathscr{H}} R(f)$. Further let $R^* = \inf_{f\in\mathscr{B}(\mathscr{X},\mathscr{Y})} R(f)$ be the inestimable/abstract "true" minimal risk calculated among all Borel measurable functions $\mathscr{B}(\mathscr{X}) \to \mathscr{B}(\mathscr{Y})$, which is known as the Bayes risk.

The generalization error of the learning algorithm is a random variable (stochastic process) defined by

$$\underbrace{\left[\widehat{R}(f_m) - R^*\right]}_{\text{Generalization Error}} = \underbrace{[R(f^*) - R^*]}_{\text{Approximation Error}} + \underbrace{\left[\widehat{R}(f_m) - R(f^*)\right]}_{\text{Estimation Error}}$$

and a learning algorithm is said to generalize well if it has a low generalization error which can be decomposed into two parts known as the Approximation error and the Estimation error.

Approximation error [95, 120, 121, 130] is the measure of how close the risk of functions in the Hypothesis space can get to the risk of the target function. Clearly $R(f^*) \geq R^*$ as the Hypothesis space may not contain the Borel measurable function which has the globally minimal risk. Thus the Approximation error is determined purely by the choice of the Hypothesis space of functions $\mathscr{H}$ and is independent of the population measure. However, the complexity of $\mathscr{H}$ in terms of its degrees of differentiability, types of singularities, Group invariance and equivariance, etc effects the rate of convergence of the optimization algorithms. Higher the complexity/regularity/capacity of $\mathscr{H}$, slower the rate of convergence one might expect, independent of the learning problem.

Estimation error on the other hand is a random variable which is dependent on the sampling process and the properties of the estimator but is independent of the target function. In practice the estimation error can be easily calculated based on the properties of the estimator used, however the approximation error is harder to determine as it depends on the assumptions one can make on the target function and the Hypothesis space of functions that one chooses. If the size of the expected generalization error i.e. the bias of the risk estimator, is positive then the learned function corresponding to the minimal risk is said to underfit, while if it is negative it is said to overfit.

A concept which is closely related to the generalization is the one of consistency which is only dependent on the estimation error. A statistical learning algorithm thus is said to be consistent if and only if the estimation error converges to zero as the number of observations increase. That is the learning algorithm generating a sequence of solutions $f_1, f_2, \ldots$ is consistent with respect to $(\mathscr{X}, \mu)$ and $(\mathscr{Y}, \nu)$ given $\mathscr{H}$, if for all $\varepsilon > 0$

$$\Pr\left(R(f_m) - R(f^*) > \varepsilon\right) \to 0 \text{ as } m \to \infty$$

it is Bayes consistent with respect to $(\mathscr{X},\mu)$ and $(\mathscr{Y},\nu)$ if for all $\varepsilon > 0$

$$\Pr\left(R(f_m) - R^* > \varepsilon\right) \to 0 \text{ as } m \to \infty$$

and Universally consistent with respect to $\mathscr{H}$ if it is consistent independent of $(\mathscr{X},\mu)$ and $(\mathscr{Y},\nu)$.

There is one important fact to note about these definitions do not depend on the estimator of risk and are only concerned with the true risk $R(f_m)$. Clearly the measure of quality of an element of the Hypothesis space is the true risk, and we want the true risk to become as good as possible. The difficulty of this task stems from the fact that the quantity we want to minimize actually cannot be evaluated, as the associated measures μ and ν are unknown. However we do have access to samples $\mathscr{X}_n$ and $\mathscr{Y}_n$ from which we can try to infer a function f whose risk is close to the best possible risk. Thus if we can construct a good efficient estimator of risk $\hat{R}(f_m)$, then its consistency would imply consistency of the true risk, which is known as the induction principle.

1.5.2 Empirical Risk Estimator

Clearly the most straightforward way to proceed is to approximate the true risk by the empirical risk computed on the training data. Instead of looking for a function which minimizes the true risk $R(f)$, given some training data $(\mathscr{X}_n,\mathscr{Y}_n)$, the Hypothesis space $\mathscr{H}$ and a loss function L, we have i.i.d. samples from the push-forward loss measure λ_f

$$\mathscr{L}_n(f) \coloneqq \{l_i = L(f,x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (\mathbb{R}_+,\lambda_f)$$

which allows us to write the empirical risk as

$$\hat{R}_{\text{emp}}(f) \coloneqq \frac{1}{n}\sum_{i=1}^n l_i$$

Then if the optimal solution given by

$$f^* \coloneqq \arg\min_{f\in\mathscr{H}} \hat{R}_{\text{emp}}(f)$$

is uniformly consistent i.e.

$$\Pr\left(\sup_{f\in\mathscr{H}} \mid \hat{R}_{\text{emp}}(f) - R(f) \mid > \varepsilon\right) \to 0$$

as the number of observations increase, then this approach is called the empirical risk minimization induction principle, abbreviated by ERM [120]. Now we will not only calculate this probability for each given element of the Hypothesis space, but as well calculate its rate of convergence.

Now given $l_1, l_2, \ldots$ i.i.d. real-valued random variables with finite expectation, which we have assumed to be true for the problem to be learnable, let

$$S_n := \sum_{i=1}^{n} l_i$$

and λ_n be the law of S_n, then the Weak Law of Large Numbers [92, 93] asserts that the empirical sum S_n converges in distribution to $n\mathbb{E}[l_1]$ i.e. we can say that for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(\frac{1}{n} S_n \geq \mathbb{E}[l_1] + \varepsilon) = 0$$

In fact, if $\mathbb{E}\left[l_1^2\right] < \infty$, we have the Central Limit Theorem [Ref], and a consequence is that

$$\lim_{n\to\infty} \mathbb{P}(\frac{1}{n} S_n \geq \mathbb{E}[l_1] + n^{1-\alpha}) = 0$$

whenever $\alpha > \frac{1}{2}$.

In the above statement we are considering only certain sets $[a,\infty), a > \mathbb{E}[l_1]$, though we could equally well have considered $(-\infty, a], a < \mathbb{E}[l_1]$. Then we can also consider intervals of type $[a,b], \mathbb{E}[l_1] < a < b$, in which case $\lambda_n([a,b]) = \lambda_n([a,\infty)) - \lambda_n((b,\infty))$, and we might as well assume that λ_n is sufficiently continuous, at least in the limit, that we can replace the open interval bound with a closed one. Then another way of stating the Central Limit Theorem would be to say for any closed set $F \subset \mathbb{R}$ such that $\mathbb{E}[l_1] \notin F$ then

$$\lim_{n\to\infty} \lambda_n(F) = 0$$

Example 1.2. In a concrete example, if we toss a coin some suitably large number of times, the probability that the proportion of heads will be substantially greater or smaller than $\frac{1}{2}$ tends to zero. So the probability that at least $\frac{3}{4}$ of the results are heads tends to zero. But the question is, how fast? Consider first four tosses, then eight. A quick addition of the relevant terms in

the binomial distribution gives:

$$\mathbb{P}\left(\text{At least } \tfrac{3}{4} \text{ out of four tosses are heads}\right) = \frac{1}{16} + \frac{4}{16} = \frac{5}{16}$$
$$\mathbb{P}\left(\text{At least } \tfrac{3}{4} \text{ out of twelve tosses are heads}\right) = \frac{1}{2^{12}} + \frac{12}{2^{12}} + \frac{66}{2^{12}} + \frac{220}{2^{12}} = \frac{299}{2^{12}}$$

There are two observations to be made. The first is that the probability of the second case is substantially smaller than the first – the decay appears to be relatively fast. The second observation is that $\frac{220}{2^{12}}$ is substantially larger than the rest of the sum. So by far the most likely way for at least $\frac{3}{4}$ out of twelve tosses to be heads is if exactly $\frac{3}{4}$ are heads. Cramer's theorem applies to a general i.i.d. sequence of random variables, provided the tail is not too heavy. It shows that the probability of any such large deviation event decays exponentially with n, and identifies the exponent.

In order to state the Cramer's theorem we need the logarithmic moment generating function, which for the random variable l_1 is defined as

$$\Lambda_f(\eta) \coloneqq \log \mathbb{E}\left[e^{\eta l_1}\right]$$

and its convex conjugate defined via the Legendre-Fenchel transform of $\Lambda_f(\eta)$:

$$\Lambda_f^*(x) \coloneqq \sup_{\eta \in \mathbb{R}} \left\{\eta x - \Lambda_f(\eta)\right\}$$

Let $\mathscr{D}_\Lambda \coloneqq \left\{\eta : \Lambda_f(\eta) < \infty\right\}$ and $\mathscr{D}_{\Lambda^*} \coloneqq \left\{x : \Lambda_f^*(x) < \infty\right\}$ then the Cramer's theorem is stated as follows.

Theorem 1.1. *Cramer's Theorem [55]: Let $l_i \in \mathbb{R}$ be i.i.d. real-valued random variables which satisfy $\mathbb{E}\left[e^{\lambda l_1}\right] < \infty$ for every $\eta \in \mathbb{R}$. Then Λ_f^* is called the rate function and*

1. For any closed set $F \subset \mathbb{R}$

$$\limsup_{n\to\infty} \frac{1}{n} \log \lambda_n(F) \leq -\inf_{y \in F} \Lambda_f^*(y)$$

   (a) For any open set $G \subset \mathbb{R}$

$$\liminf_{n\to\infty} \frac{1}{n} \log \lambda_n(G) \geq -\inf_{y \in G} \Lambda_f^*(y)$$

Theorem 1.2. *Here μ_n is the law of S_n.*

The proof of Cramer's theorem splits into an upper bound and a lower bound. The former is relatively straightforward, applying Markov's inequality to $e^{\eta S_n}$, then optimizing over the choice of η. This idea is referred to by various sources as the exponential Chebyshev inequality or a Chernoff bound [87, 99]. The lower bound is more challenging. We re-weight the distribution function $F(x)$ of l_1 by a factor $e^{\eta l}$, then choose η so that the large deviation event is in fact now within the treatment of the central limit theorem, from which suitable bounds are obtained.

Therefore we that for S_n with some law, i.e. a measure λ_n on $\mathbb{R}$. The law of large numbers asserts that as $n \to \infty$, these measures are increasingly concentrated at a single point in $\mathbb{R}$, which in this case is $\mathbb{E}[l_1]$. Cramer's theorem then asserts that the measure of certain sets not containing this point of concentration decays exponentially in n, and quantifies the exponent, by a so-called rate function, which is obtained via a Legendre transform of the log moment generating function of the underlying distribution. Then informally Cramer's theorem asserts that

$$\lambda_n([a,\infty)) \sim e^{-n\Lambda^*(a)}$$

which also gives us an universal estimate for sample size dependent p-values. Such a principle is extremely useful in deriving asymptotic rates of convergence in various situations, however in this work we apply it to prove consistency of the empirical risk estimator.

Corollary 1.1. *The performance of the empirical estimator for all $f \in \mathscr{H}$ is given by*

$$\limsup_{n\to\infty} \frac{1}{n} \log \lambda_n \left(\sup_{f\in\mathscr{H}} \mid \hat{R}_{emp}(f) - R(f) \mid > \varepsilon \right) \leq -2 \sup_{f\in\mathscr{H}} \Lambda_f^*(\varepsilon)$$

In the following we will show how to model the push-forward loss measure and estimate the rate function Λ_f^* necessary to calculate the above defined rate of convergence.

1.5.2.1 Subordinators and Levy processes

In this section we introduce a technique to calculate the rate function that we discussed with respect to the Cramer's theorem which we can use to calculate the tail probabilities of the empirical risk estimator.

As before we are given some training data $(\mathscr{X}_n, \mathscr{Y}_n)$, a Hypothesis space $\mathscr{H}$ and a loss function L, we have i.i.d. samples from the push-forward loss measure λ_f

$$\mathscr{L}_n(f) \coloneqq \{l_i = L(f, x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (\mathbb{R}_+, \lambda_f)$$

Now it is not necessary that that the push forward measure λ_f for each element f of the Hy-

pothesis space, is a probability measure. Hence we need to make the following assumptions. Let λ_f be a general measure on $]0,\infty[$ such that

$$\int_{]0,\infty[} \min(\delta, x) d\lambda_f(x) < \infty$$

for some $\delta > 0$ and

$$\lambda_f(]0,\infty[) = \infty$$

The fact that for every $x > 0$ the tail-intensity $\Lambda(]x,\infty[)$ is finite implies that there are only finitely many atoms in $]x,\infty[$. On the other hand, there are infinitely many atoms in $]0,\infty[$ since $\Lambda(]0,\infty[) = \infty$. We may thus rank these atoms is decreasing order, and denote them

$$l_1^* \geq l_2^* \geq \cdots \geq 0$$

as the ranked sequence.

Next, we point out that the integral condition above ensures the summability of the series $\sum_1^\infty l_i$. Indeed, by the first-moment formula, we see that the series $\sum_1^\infty l_i \mathbb{I}_{\{l_i \leq \delta\}}$ converges almost surely and since there are only finitely many atoms in $]\delta,\infty[$, we have

$$\zeta(1) := \sum_1^\infty l_i < \infty$$

almost surely. Conversely, it can also be checked that the series $\sum_1^\infty l_i$ diverges almost surely whenever the integral condition fails. This means that the above integral condition provides the necessary and sufficient condition for the Risk functional to be finite for each function f and thereby characterizes learnability.

Now we consider certain increasing processes which is known as a Subordinator [23, 104]. Introduce an independent sequence $U_1, U_2 \ldots$ of i.i.d. uniform variables on $[0,1]$, and then define the increasing process

$$\zeta(t) := \sum_{i=1}^\infty l_i \mathbb{I}_{\{U_i \leq t\}} = \sum_{U_i \leq t} l_i, \;\; t \in [0,1] \tag{1.4}$$

In other words, the collection of jump times and jump sizes of purely discontinuous increasing process ζ is $\{(U_i, l_i), i \in \mathbb{N}\}$.

Now increasing process $(\zeta(t), 0 \leq t \leq 1)$ has independent and stationary increments. This means that for every $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$, the variables $\zeta(t_1) - \zeta(t_0), \ldots, \zeta(t_{n+1}) - \zeta(t_n)$ are independent, and $\zeta(t_{i+1}) - \zeta(t_i)$ has the same law as $\zeta(t_{i+1} - t_i)$. Then the process $(\zeta(t), 0 \leq t \leq 1)$ is known as a subordinator on the time interval $[0,1]$. The Levy-Khintchine

formula (1.5) shows that $\zeta(t)$ is a subordinator if and only if ν_f, the push-forward measure defined on $\mathbb{R}_+$ is a completely random measure i.e. a Levy measure. Which allows us to define the function Λ_f as the Laplace exponent and the representation of the subordinator itself (1.4) as the Levy-Ito decomposition.

Theorem 1.3. *Levy–Khintchine formula [23]: The Laplace-Stieltjes transform of a subordinator on $\mathbb{R}_+$ has a unique representation of the form*

$$\log \mathbb{E}\left[\exp\left(-\varepsilon\zeta(t)\right)\right] = -\varepsilon t\theta - t\int_{\mathbb{R}_+}\left[1-e^{-\varepsilon l}\right]\lambda_f(dl) = -t\Lambda_f(\varepsilon) \tag{1.5}$$

for all $\varepsilon,\theta \in \mathbb{R}_+$ and where the push-forward measure λ_f on $\mathbb{R}_+$ satisfies, for some $\delta > 0$

$$\int_{\mathbb{R}_+} \min(\delta, l)\,\lambda_f(dl) < \infty$$

This means that under the very weak integrability condition of the push forward measure λ_f, which essentially means that if the points as sampled from it are concentrated in a compact subset, then λ_f is a Levy measure. Many of the well studied measures belong to the family of Levy measures, including Compound Poisson Processes , Gaussian Processes , Gamma Processes, Stable Processes, etc. Here we use Gamma and Stable sub-families to get model our push-forward loss measure and thereby estimate the rate function corresponding each element of the Hypothesis space.

Gamma subordinators The Gamma subordinator [23] is suitable when the tail of the push-forward loss measure decreases exponentially fast. Let $\theta, c > 0$ be two fixed real numbers. The subordinators $(\zeta(t), 0 \le t \le 1)$ corresponding to the Levy measure

$$d\lambda_f(x) = \theta x^{-1} e^{-cx} dx, \;\; x > 0$$

is called a gamma subordinator with parameter (θ, c). Its Laplace exponent is given by

$$\Lambda_f(\varepsilon) = \theta\int_0^\infty (1-e^{-\varepsilon x})x^{-1}e^{-cx}dx = \theta\log(1+\varepsilon/c), \;\; q \ge 0$$

note that $\zeta(t)$ has the gamma distribution with parameter $(\theta t, c)$. Here again, the parameter c will have a very minor role, due to the easy fact that $c\gamma(\cdot)$ is a gamma subordinator with parameter $(\theta, 1)$. In this direction, it might be also interesting to point out that for every $a \in]0,1[$, $(\gamma(at), 0 \le t \le 1)$ is a gamma subordinator with parameter $(a\theta, c)$.

Then the rate function corresponding to a Gamma subordinator if given by the Legendre-

Fenchel transform of $\Lambda_f(\varepsilon)$:

$$\begin{aligned}\Lambda_f^*(x) &= \sup_{\varepsilon\in\mathbb{R}} \left\{\varepsilon x - \Lambda_f(\varepsilon)\right\} \\ &= \sup_{\varepsilon\in\mathbb{R}} \left\{\varepsilon x - \theta \log(1+\varepsilon/c)\right\} \\ &= \theta - \theta\log\theta - cx + \theta\log(cx)\end{aligned}$$

Stable Subordinators The Stable subordinator [23] is suitable when the tail of the push-forward loss measure decreases polynomially fast. Let $\alpha \in]0,1[$ and $c>0$ are fixed parameters. Then subordinators $(\zeta(t), 0 \le t \le 1)$ corresponding to the Stable Levy measure

$$d\lambda_f(x) = \frac{c\alpha}{\Gamma(1-\alpha)} x^{-1-\alpha}, \;\; x>0$$

has a Laplace exponent given by

$$\Lambda_f(\varepsilon) = c\varepsilon^\alpha = \frac{c\alpha}{\Gamma(1-\alpha)} \int_0^\infty (1-e^{-\varepsilon x}) x^{-1-\alpha} dx$$

It is known as a stable subordinator with index α. The parameter $c>0$ has a very minor role, as changing ζ into $k\zeta$ merely amounts to change c into $k^\alpha c$. In particular, the following definition does not depend on c.

Similarly using the rate function corresponding to a Gamma subordinator if given by the Legendre-Fenchel transform of $\Lambda_f(\varepsilon)$:

$$\begin{aligned}\Lambda_f^*(x) &= \sup_{\varepsilon\in\mathbb{R}} \left\{\varepsilon x - \Lambda_f(\varepsilon)\right\} \\ &= \sup_{\varepsilon\in\mathbb{R}} \left\{\varepsilon x - c\varepsilon^\alpha\right\} \\ &= \left(\frac{x}{\alpha c}\right)^{\frac{1}{\alpha-1}}\end{aligned}$$

Thus using the observed samples from the push-forward loss measure $\mathscr{L}_n(f)$, we should be able to estimate the parameters of either the Gamma or the Stable distribution, which would then allow us to write the estimate of the rate function corresponding to each element of the hypothesis space.

1.5.3 Bias-Variance Tradeoff

Now we know how to calculate the performance of each element of the Hypothesis space using an empirical estimator. However as a whole the performance of the Hypothesis space, in terms of the rate of convergence to the true solution depends on the assumptions on it, which reduces its "capacity" to represent different functions thereby increasing the approximation error for an arbitrary target function, however leads to faster convergence.

Originally coined in the case of least squared regression, this is the well known dichotomy between controlling the trade-off between approximation (bias) and estimation (variance) error. In statistics, estimation error is also called the variance which measures the variation of the risk of the function f_n estimated on the sample, and the approximation error measures the "bias" introduced in the model by choosing too small a function class.

Intuitively speaking, Hypothesis spaces with higher capacity can represent target functions of higher complexity. However, rather unintuitively, there is no universal way of measuring the complexity of the elements in the Hypothesis space, a fact which has been formalized in what is known as the No Free Lunch theorem [126].

For example in the context of simple hypothesis testing this tradeoff is analogous to Neyman Pearson Optimality where one wants to minimize Type II error while controlling Type I error. Here one can see Type II error as the risk associated with the choice of a test and we want to choose the one which minimizes it. While the Type I error characterizes the generalization error of the test which we want to control at a certain level.

Here we see that Type I error is one possible way of controlling the generalization error which results in the notion of Neyman Pearson optimality of hypothesis tests. An associated concept is the probability of False Discovery, which is defined to be the posterior probability of the chosen label being wrong (for classification) or the chosen element in the dictionary being wrong (for regression) to represent a certain function, given the observed data. In the context of generalization error, we want to control the rate of false discoveries while searching through the Hypothesis space for the optimal solution. We develop this concept in the next section.

1.5.3.1 Probability of False Discovery

In order to control the generalization error, for each function present in the Hypothesis space, we would like to construct a Hypothesis test of whether it belongs to a reasonable subclass of functions which could be the plausible solution. This is a common approach especially for regression, where a normally distributed prior on the parameters leads to ridge regression [5, 60, 84], while a Laplacian prior leads to Lasso [27, 34, 118, 124] in the case of square

error loss. In this section we would like to generalize this strategy to any arbitrary family of measures and any machine learning application.

This problem can be seen as a compound hypothesis testing problem, where only partial information about the null hypothesis is known a priori. One possible way to model the situation is by a composite null hypothesis $H_0 := \mu_0 \in \cup_{\theta\in\Theta}\mu_\theta$ for $\mu_\theta \in \mathscr{P}(\mathscr{H})$ with $\theta \in \Theta$ represents the family of measures that is independent of n (the number of observations). Then the testing problem can be stated as

$$H_0 = \mu_0 \in \cup_{\theta\in\Theta}\mu_\theta \text{ vs } H_1 = \mu_1 \in \mathscr{P}(\mathscr{H})\setminus \cup_{\theta\in\Theta}\mu_\theta \tag{1.6}$$

This case is also known as the Detection problem in literature, where one would be interested in determining whether some phenomena is present or not based on the given observations.

We start by constructing a model which can be characterized by a binary random variable $S \in \{0,1\}$ called the hypothesis random variable which corresponds to

$$S = \begin{cases} 0 & \text{if } H_0 \\ 1 & \text{if } H_1 \end{cases}$$

along with the samples from the push-forward loss measure $\mathscr{L}_n(f)$. Then a statistical decision test $\hat{S}_n(\mathscr{L}_n(f))$ is a sequence of Borel measurable (w.r.t. the product σ-field) maps $\hat{S}_n : \mathscr{L}_n(f) \to S$, with the interpretation that when $\mathscr{L}_n(f)$ is observed

$$\hat{S}_n(\mathscr{L}_n(f)) = S$$

maps $\mathscr{L}_n(f)$ to the hypothesis random variable S, i.e. makes a decision whether the function under consideration f should be accepted or rejected.

The probability $P(S)$ of hypothesis S is referred to as the *a priori* probability of the hypothesis S. This a priori probability is not known in most circumstances and usually classical results in testing theory can be deduced by assigning a probability of $1/2$ to each hypothesis. However in many real world phenomena, such an assumption is not reasonable. For example in the case of rare events, $P(S=0) \gg P(S=1)$, i.e. a priori we know that the probability of the alternate hypothesis being true is very small compared to the null hypothesis, even if we don't actually know their values. Thus by studying various models of this prior probability, we gain insight into constructing better testing algorithms especially in such situations.

1.5.3.2 Testing Criteria

The performance of a statistical decision test $\hat{S}_n(\mathscr{L}_n(f))$ can be determined using a variety of error criteria and subsequently can be optimized with respect to them to construct different algorithms.

First we consider maximizing the probability of making the correct decision. Thus we want to maximize the following posterior probability

$$P(S \mid \mathscr{L}_n(f)) = \frac{P(\mathscr{L}_n(f) \mid S)P(S)}{P(\mathscr{L}_n(f))}$$

which gives us the MAP rule

$$\hat{S}_n(\mathscr{L}_n(f)) = \arg\max_S P(\mathscr{L}_n(f) \mid S)P(S) \tag{1.7}$$

The MAP rules reduces to the classical maximum likelihood test when $P(S=0) = 1/2$. Then we can define the classical errors i.e. the pair of Type I and II errors which can be described as follows. Type I error is defined as the probability of the statistical decision test rejecting the null hypothesis, when the null is actually true i.e.

$$\alpha_n(\hat{S}_n(\mathscr{L}_n(f))) \coloneqq P(\hat{S}_n(\mathscr{L}_n(f)) = 1 \mid S = 0) \tag{1.8}$$

and the Type II error is the probability of the statistical decision test accepting the null hypothesis when the alternate is true

$$\beta_n(\hat{S}_n(\mathscr{L}_n(f))) \coloneqq P(\hat{S}_n(\mathscr{L}_n(f)) = 0 \mid S = 1) \tag{1.9}$$

Now $\beta_n(\hat{S}_n(\mathscr{L}_n(f)))$ may always be minimized by choosing $\hat{S}_n(\mathscr{L}_n(f)) \equiv 1$ at the expense of $\alpha_n(\hat{S}_n(\mathscr{L}_n(f))) = 1$.

The Neyman Pearson criterion for optimality involves looking for a test $\hat{S}_n(\mathscr{L}_n(f))$ that minimizes $\beta_n(\hat{S}_n(\mathscr{L}_n(f)))$ subject to the constraint $\alpha_n(\hat{S}_n(\mathscr{L}_n(f))) \leq \eta$ for some $0 < \eta < 1$. This criterion is satisfied by the MAP rule and can be reformulated in terms of a log-likelihood ratio test as defined in the next section. However, such a criterion can only be met in case of a Simple Hypothesis test, when both the alternative measures are known.

The Hoeffding criterion [19, 26, 96, 105, 108, 122, 129] for optimality for a test $\hat{S}_n$, is defined as: if among all tests that satisfy

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n(\hat{S}_n) \leq -\eta$$

the test $\hat{S}_n$ has the maximal exponential rate of error, i.e. uniformly over all possible alternate laws,

$$-\limsup_{n\to\infty}\frac{1}{n}\log\beta_n(\hat{S}_n)$$

is maximal, then $\hat{S}_n$ is called optimal. This criterion applies to a more general situation when only the null measure is known, or even only partially known and is optimized by the generalized log-likelihood ratio test. This optimality criterion can also be modified and stated in terms of controlling the analogous posterior errors, the FDR and FNR, as defined below.

Then the probability of False Discovery [19, 34, 40, 98, 124] can be defined as the posterior error rate analogous to Type I error,

$$\begin{aligned}\mathrm{FDR}(\hat{S}_n(\mathscr{L}_n(f))) &= P(S=0 \mid \hat{S}_n(\mathscr{L}_n(f))=1)\\ &= \left(1+\frac{1-\beta_n(\hat{S}_n(\mathscr{L}_n(f)))}{\alpha_n(\hat{S}_n(\mathscr{L}_n(f)))}\frac{P(S=1)}{P(S=0)}\right)^{-1}\end{aligned} \tag{1.10}$$

while the probability of False Non-discovery, the posterior error analogous to the Type II error,

$$\begin{aligned}\mathrm{FNR}(\hat{S}_n(\mathscr{L}_n(f))) &= P(S=1 \mid \hat{S}_n(\mathscr{L}_n(f))=0)\\ &= \left(1+\frac{1-\alpha_n(\hat{S}_n(\mathscr{L}_n(f)))}{\beta_n(\hat{S}_n(\mathscr{L}_n(f)))}\frac{P(S=0)}{P(S=1)}\right)^{-1}\end{aligned} \tag{1.11}$$

Clearly these error rates depend on the prior probabilities of the two Hypothesis, which is usually not known for a given application and hence needs to be inferred from data. Usually in literature these error rates are defined in the context of multiple testing problem, since in that case one has access to multiple p-values which can be used to estimate these prior probabilities. In this work by defining them as posterior error rates, we are able to define a generalized likelihood ratio test with data adaptive threshold which is able to control the FDR probability.

1.5.3.3 Generalized Likelihood Ratio Test

The major deficiency of the likelihood ratio test lies in the fact that it requires perfect knowledge of the measures μ_0 and μ_1, both in forming the likelihood ratio and in computing the threshold γ. Thus, it is not applicable in situations where the alternative hypotheses consist of a family of probability measures, for example when only partial information about the null probability measure is given, which might be that it belongs to a certain family of measures

$$H_0 = \mu_0 \in \cup_{\theta\in\Theta}\mu_\theta \text{ vs } H_1 \neq \mu_0 \in \cup_{\theta\in\Theta}\mu_\theta$$

To overcome these difficulties, the error criterion has to be modified, since the requirement of uniformly small β_n over a large class of plausible laws μ_1 may be too strong and it may be that no test can satisfy such a condition. It is reasonable therefore to search for a criterion that involves asymptotic limits.

Then the optimality condition (due to Hoeffding) under both scenarios for a test $\hat{S}_n$ (for a given threshold $\gamma > 0$) could be considered as, if among all tests that satisfy

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n(\hat{S}_n) \leq -\gamma$$

the test $\hat{S}_n$ has the maximal exponential rate of error, i.e. uniformly over all possible alternate laws,

$$-\limsup_{n\to\infty} \frac{1}{n} \log \beta_n(\hat{S}_n)$$

is maximal. To present the optimal test, we need certain definitions. Let $(\hat{S}_n^0, \hat{S}_n^1)$ be partitions induced on $\mathscr{H}$ by the test $\hat{S}_n$ such that

$$\hat{S}_n^0 := \left\{ \mathscr{L}_n(f) \sim \mathscr{P}(\mathscr{H}) : \hat{S}_n(\mathscr{L}_n(f)) = 0 \right\} \text{ and } \hat{S}_n^1 = \mathscr{H} \backslash \hat{S}_n^0$$

Since in the general framework discussed here, point-wise bounds on error probabilities are not available, smooth versions of the maps $\hat{S}_n$ are considered. Specifically, for each $\delta > 0$, let $\hat{S}_n^\delta$ denote the δ-smoothing of the map $\hat{S}_n$ defined via

$$\hat{S}_n^{0,\delta} := \left\{ \mathscr{L}_n(f) \sim \mathscr{P}(\mathscr{H}) : d(\mathscr{L}_n(f), \hat{S}_n^0) < \delta \right\} \text{ and } \hat{S}_n^{1,\delta} = \mathscr{H} \backslash \hat{S}_n^{0,\delta}$$

i.e., the original partition is smoothed by using $\hat{S}_n^{0,\delta}$, the open δ-blowup of the set $\hat{S}_n^0$. Finally we can define the δ-smoothed rate function as

$$J_\delta(y) := \inf_{x \in B_{2\delta,y}} \Lambda^*(x) \tag{1.12}$$

where $B_{2\delta,z}(\mathbf{x}) = \{\mathbf{x} \in X^n : d(z,\mathbf{x}) \leq 2\delta\}$. For the δ-smoothed version of the rate function for the partially known null hypothesis we have a natural candidate as

$$J_\delta(y) := \inf_{\theta \in \Theta} \inf_{x \in B_{2\delta,y}} \Lambda_\theta^*(x) \tag{1.13}$$

Then the optimal tests under both scenarios is given by the following theorem.

Theorem 1.4. *Dembo-Zetouni Theorem [55]: For any $\delta > 0$, any $\gamma \geq 0$ and for all $\theta \in \Theta$ if*

applicable let

$$\hat{S}_n^{*,\delta}(\gamma) = \begin{cases} 0 & \text{if } J_\delta(\mathscr{L}_n(f)) < \gamma \\ 1 & \text{otherwise} \end{cases}$$

such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n(\hat{S}_n^{*,\delta}(\gamma)) \leq -\gamma$$

then for any other test $\hat{S}_n^{\delta}(\eta)$ *where*

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n(\hat{S}_n^{c\delta}(\gamma)) \leq -\gamma$$

where $c > 0$ *is an arbitrary constant, then we have*

$$-\limsup_{n\to\infty} \frac{1}{n} \log \beta_n(\hat{S}_n^{*\delta}(\eta)) \geq -\limsup_{n\to\infty} \frac{1}{n} \log \beta_n(\hat{S}_n^{\delta}(\eta))$$

This theorem provides an extremely general way of constructing a statistical test which would be true even if the i.i.d. assumption on the data made above is relaxed, a case which however we do not study in this work. Finally applying the Cramer's theorem we can calculate the p-value corresponding to the test $\hat{S}_n^{*,\delta}$ quite easily by

$$\limsup_{n\to\infty} \frac{1}{n} \log(\text{p-value}(\mathscr{L}_n(f))) = \lim_{n\to\infty} \frac{1}{n} \log P_{\mu_0}((J_\delta(\mathscr{L}_n(f)), \infty)) = -J_\delta(\mathscr{L}_n(f))$$

1.5.3.4 FDR Control

Till now we used the Type I and II errors to develop our notion of optimality and corresponding tests. Now instead, consider the problem of constructing an analogous optimal test based on controlling the pair FDR and FNR. As we have seen previously that both quantities do not go to zero together for a fixed threshold level. However, if we fix FDR or FNR at a certain level, then one can calculate the other quantity. In practice therefore it is necessary to choose which error is more important to control, in a given real world application.

Hence for a given FDR [1.10], by using some basic algebra we can calculate FNR as

$$\text{FNR}(\alpha_n, \beta_n, \gamma) = \left(1 + \frac{e^{\gamma}}{\beta_n} - \frac{\text{FDR}(\alpha_n, \beta_n, \gamma)}{1 - \text{FDR}(\alpha_n, \beta_n, \gamma)} \frac{1 - \beta_n}{\beta_n}\right)^{-1}$$

or equivalently for a given value of FNR [1.11], we can calculate the corresponding value of FDR as

$$\text{FDR}(\alpha_n, \beta_n, \gamma) = \left(1 + \frac{e^{-\gamma}}{\alpha_n} - \frac{\text{FNR}(\alpha_n, \beta_n, \gamma)}{1 - \text{FNR}(\alpha_n, \beta_n, \gamma)} \frac{1 - \alpha_n}{\alpha_n}\right)^{-1}$$

Following the idea of Hoeffding, we develop a criterion for an optimal test $\hat{S}_n$, as to select an adaptive threshold, which fixes either one of FDR or FNR while minimizing the other, among all other tests asymptotically. For example, in this work we call $\hat{S}_n$ optimal if among all tests that satisfy

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{FDR}(\hat{S}_n) \leq -\eta$$

the test $\hat{S}_n$ has the maximal exponential rate of error, i.e. uniformly over all possible alternate measures

$$-\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{FNR}(\hat{S}_n)$$

is maximal. Then the following theorem describes a way to build such a statistical test.

Theorem 1.5. *For every $\mathscr{L}_n(f)$ and any $\delta > 0$, if the statistical test is given by*

$$\hat{S}_n^{*,\delta}(\mathscr{L}_n(f)) = \begin{cases} 0 & \text{if } J_\delta(\mathscr{L}_n(f)) < \mathbb{E}_{\mu_0}\left[J_\delta(\mathscr{L}_n(f))\right] + \eta \\ 1 & \text{otherwise} \end{cases}$$

where J_δ is given by [1.12 or 1.13], then the FDR rate is controlled at

$$\limsup_{n\to\infty} \frac{1}{n} \log FDR(\hat{S}_n^{*,\delta}) \leq -\eta$$

while having the maximal exponential rate of error, i.e. uniformly over all possible alternate measures

$$-\limsup_{n\to\infty} \frac{1}{n} \log FNR(\hat{S}_n^{*,\delta})$$

Proof. In order to prove the theorem we start from the formulation of the Dembo-Zetouni theorem, i.e. for any $\delta > 0$, any $\gamma \geq 0$ and for all $\theta \in \Theta$ if applicable

$$\hat{S}_n^{*,\delta}(\gamma) = \begin{cases} 0 & \text{if } J_\delta(\mathscr{L}_n(f)) < \gamma \\ 1 & \text{otherwise} \end{cases}$$

since it asymptotically has the maximum power. Now for such a test, the FDR is defined by [1.10], which allows us to setup the optimality condition, for any $\eta \geq 0$ as

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{FDR}(\hat{S}_n^{*,\delta}(\mathscr{L}_n(f))) \leq -\eta$$

then we have

$$\begin{aligned}
-\eta &\geq \limsup_{n\to\infty} \frac{1}{n} \log \mathrm{FDR}(\hat{S}_n^{*,\delta}(\mathscr{L}_n(f))) \\
&= -\limsup_{n\to\infty} \frac{1}{n} \log \left(1 + \frac{1-\beta_n(\hat{S}_n^{*,\delta}(\mathscr{L}_n(f)))}{\alpha_n(\hat{S}_n^{*,\delta}(\mathscr{L}_n(f)))} e^{-\gamma}\right) \\
&= -\limsup_{n\to\infty} \frac{1}{n} \log \left(\frac{1-\beta_n(\hat{S}_n^{*,\delta}(\mathscr{L}_n(f)))}{\alpha_n(\hat{S}_n^{*,\delta}(\mathscr{L}_n(f)))}\right) := \Gamma(\hat{S}_n^{*,\delta})
\end{aligned}$$

For any arbitrary $\mu_1 \in \mathscr{P}(X)\backslash\mu_0$, we want to select the threshold γ such that

$$\begin{aligned}
\sup_{\mu_1\in\mathscr{P}(X)\backslash\mu_0} \Gamma(\hat{S}_n^{*,\delta}) &= \sup_{\mu_1\in\mathscr{P}(X)\backslash\mu_0} \limsup_{n\to\infty} \frac{1}{n} \log \left(\frac{P_{\mu_1}(J_\delta(\mathscr{L}_n(f)) \in (\gamma,\infty))}{P_{\mu_0}(J_\delta(\mathscr{L}_n(f)) \in (\gamma,\infty))}\right) \\
&= \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mu_0}\left(e^{nJ_\delta(\mathscr{L}_n(f))} 1_{J_\delta(\mathbf{x})\in(\gamma,\infty)}\right) \\
&\quad - \limsup_{n\to\infty} \frac{1}{n} \log P_{\mu_0}(J_\delta(\mathscr{L}_n(f)) \in (\gamma,\infty))
\end{aligned}$$

Since the test rejects the null hypothesis in the case of FDR, we have $\gamma < J_\delta(\mathscr{L}_n(f))$, and hence

$$\begin{aligned}
\sup_{\mu_1\in\mathscr{P}(X)\backslash\mu_0} \Gamma(\hat{S}_n^{*,\delta}) &\leq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mu_0}\left[P_{\mu_0}((J_\delta(\mathscr{L}_n(f)),\infty))\right] \\
&\quad - \lim_{n\to\infty} \frac{1}{n} \log P_{\mu_0}((J_\delta(\mathscr{L}_n(f)),\infty)) \\
&= \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\text{p-value}(\mathscr{L}_n(f))\right] - \liminf_{n\to\infty} \frac{1}{n} \log\left(\text{p-value}(\mathscr{L}_n(f))\right)
\end{aligned}$$

Since the p-value corresponding to the test $\hat{S}_n^{*,\delta}$ is given by

$$\liminf_{n\to\infty} \frac{1}{n} \log\left(\text{p-value}(\mathscr{L}_n(f))\right) = \lim_{n\to\infty} \frac{1}{n} \log P_{\mu_0}((J_\delta(\mathscr{L}_n(f)),\infty)) = -J_\delta(\mathscr{L}_n(f))$$

using the Jensen's Inequality, we can estimate an upper bound on the expected p-value as

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\text{p-value}(\mathscr{L}_n(f))\right] \leq \mathbb{E}\left[\liminf_{n\to\infty} \frac{1}{n} \log\left(\text{p-value}(\mathscr{L}_n(f))\right)\right] = -\mathbb{E}\left[J_\delta(\mathscr{L}_n(f))\right]$$

This means that the optimal threshold γ corresponds to satisfying

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{FDR}(\mathscr{L}_n(f)) \geq \liminf_{n\to\infty} \frac{1}{n} \log\left(\text{p-value}(\mathscr{L}_n(f))\right) - \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mu_0}\left[\text{p-value}(\mathscr{L}_n(f))\right]$$

in other words

$$\limsup_{n\to\infty} \frac{1}{n} \log \text{FDR} \geq -J_\delta(\mathscr{L}_n(f)) + \mathbb{E}_{\mu_0}\left[J_\delta(\mathscr{L}_n(f))\right]$$

which means for any $\mathbf{x}$, we reject null hypothesis if

$$J_\delta(\mathscr{L}_n(f)) - \mathbb{E}_{\mu_0}\left[J_\delta(\mathscr{L}_n(f))\right] - \eta \geq 0 \tag{1.14}$$

then the FDR is controlled at the level $e^{-n\eta}$. □

Example 1.3. Benjamini-Hochberg Procedure: Let X denote the orthogonal design matrix with unit norm, whose columns are perpendicular to each other (note that this implies $p \leq n$). Further let the errors be i.i.d. N (0, 1), then

$$\tilde{y} = X'y \sim N(\beta, I_p)$$

where I_p is the $p \times p$ identity matrix. For testing the p hypotheses i.e. $H_i : \beta_i = 0$, the Benjamini-Hochberg Procedure (BHq) step-up procedure proceeds as follows:

1. Sort the entries of $\tilde{y}$ in decreasing order of magnitude, $|\tilde{y}|_{(1)} \geq |\tilde{y}|_{(2)} \geq \cdots |\tilde{y}|_{(p)}$ (this yields corresponding ordered hypotheses $H_{(1)}, ..., H_{(p)}$).

   (a) Find the largest index i such that

   $$|\tilde{y}|_{(i)} > \Phi^{-1}(1 - q_i),\ q_i = q\frac{i}{2p}$$

   where $\Phi^{-1}(\alpha)$ is the αth quantile of the standard normal distribution and q is a parameter in $[0, 1]$. Call this index i_{SU}. (For completeness, the BHq procedure is traditionally expressed via the inequality $|\tilde{y}|_{(i)} \geq \Phi^{-1}(1 - q_i)$ but this does not change anything since $\tilde{y}$ is a continuous random variable.)

   (b) Reject all $H_{(i)}$'s for which $i \leq i_{\text{SU}}$ (if there is no i then make no rejection).

Example 1.4. This procedure is adaptive in the sense that a hypothesis is rejected if and only if its z-value is above a data-dependent threshold. In their seminal paper [29], Benjamini and Hochberg proved that this procedure controls the FDR. Letting V (resp. R) be the total number of false rejections (resp. total number of rejections), we have

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = q\frac{p_0}{p}$$

where p_0 is the number of true null hypothesis, $p_0 = \mid \{i : \beta_i = 0\} \mid$, so that $p = p_0 + \parallel \beta \parallel_{l_0}$. This always true, no matter the value of the mean vector β. In the general case the $\tilde{y}$ are given exactly by $J_\delta(\cdot)$.

Therefore we see that by modeling the samples from the push-forward loss measure, using the subordinator formulation, we were able to not only estimate convergence rates for the empirical risk estimator under a large class of machine learning problems, but also associate to each element of the Hypothesis space a probability of False discovery which allows us to construct an alternative approach to control the generalization error.

Chapter 2

Model Selection by Adapting to unknown Sparsity

2.1 Introduction

Representing complicated functions as superpositions of basic transforms of simpler functions has been a subject of study in Harmonic analysis since the introduction of the Fourier transform, centuries ago [31, 32, 42]. Such a representation allows us to extract information from observed signals or functions by transforming the function from its original domain into a new domain, with the purpose of extracting the characteristic information which is otherwise not readily observable in its original form.

Consider a sequence of paired i.i.d. random variables $(X_1,Y_1),...,(X_n,Y_n)$ on a Polish space $(\mathscr{X},\mu)\times(\mathscr{Y},\nu)$ with a Borel probability measures, $\mu\in\mathscr{M}_+^1(\mathscr{X})$ and $\nu\in\mathscr{M}_+^1(\mathscr{Y})$. Let $f\in L^2(\mathscr{X})$ be a square integrable function $f:\mathscr{X}\to\mathscr{Y}$, then the synthesis problem is the process of building a representation

$$y=f(x)=\langle\theta,\Phi(x)\rangle+\varepsilon$$

where $\Phi(x)=\{\phi_\lambda(x)\}_{\lambda\in\Lambda}$ is a set of template functions called the dictionary. The dictionary provides an over-complete representation, i.e. represents way more "features" than necessary to represent the function. Then the aim of any learning algorithm is to find a sparse parameter vector $\theta\in\mathbb{R}^{\#(\Lambda)}$ which minimizes the norm of the error $\varepsilon\in\mathscr{Y}$ i.e.

$$\arg\min_\theta \mathbb{E}_{\mu\times\nu}\parallel y-\langle\theta,\Phi(x)\rangle\parallel_2^2+\mathrm{Reg}(\theta)$$

where $\mathrm{Reg}(\theta)$ is a regularizer and is a function of the parameter θ.

When the regularizer is equal to the L^1 norm, we get the standard risk functional corresponding to the LASSO algorithm, while for the L^2 norm we get the one corresponding to ridge regression. In this work, we want to find a regularizer, which controls the sparsity of θ by controlling the asymptotic FDR rate of elements in the dictionary $\Phi(x)$.

In order to derive the regularizer, we start by using the Gibbs principle to define the joint distribution of the empirical risk and the probability of false discovery for each element of the dictionary which we calculated earlier using the compound Hypothesis rate function denoted by $J_\delta(\theta)$. This provides a general framework for constructing regularized objectives once one chooses the loss function. Thus learning algorithm would construct a sequence of measures $\rho_1, \rho_2, ... \in \mathscr{M}_+^1(\mathscr{H})$ on the parameter space, such that

$$\mathbb{E}_{\rho_m}[f] \to \mathbb{E}_{\rho^*}[f]$$

where $\rho^* \in \mathscr{M}_+^1(\mathscr{H})$ is given by the Gibbs Measure i.e.

$$\rho^* = \arg \min_{\rho \in \mathscr{M}_+^1(\mathscr{H})} F_\beta[\rho]$$

where

$$F_\beta[\rho] = \mathbb{E}_\rho\left[\mathbb{E}_{\mu\times\nu} \parallel y - \langle \theta, \Phi(x)\rangle \parallel_2^2 + \gamma_\theta J_\delta(\theta)\right] + \beta^{-1} H(\rho)$$

is known as the Free Energy functional with $\beta > 0$ (temperature) and $H(\rho)$ the entropy of ρ. This allows us to show that if γ_θ in proportional on the first order statistic distribution, then ρ^* controls the rate of False Discoveries. Therefore it is possible that $\exists \rho$ such that

$$\mathbb{E}_{\rho^*}[\Phi(f)] > \underset{\text{Jensen's Inequality}}{\Phi\left(\mathbb{E}_{\rho^*}[f]\right)} > \Phi\left(\mathbb{E}_\rho[f]\right)$$

but then ρ would have a higher rate of False Discovery.

We confirm the rather theoretical result by deriving the regularizer under a Laplacian prior on the parameter vector. This allows us to derive that

$$\text{Reg}(\theta) = \sum_\lambda \gamma_\lambda \mid \theta_{(\lambda)} \mid$$

where $\gamma_1 \geq \gamma_2 \geq \cdots$ and $\mid \theta_{(1)} \mid \geq \mid \theta_{(2)} \mid \geq \cdots$ are the order statistics of the magnitudes of the coefficients is the recently derived SLOPE regularizer [27, 34, 106, 111] which was shown to control FDR at a given level by choosing γ_i's appropriately.

2.1.1 Loss Functions which control FDR

As discussed earlier, the risk for a general machine learning problem is defined in terms of the mean discrepancy function Δ, which itself can be written in terms of being the expected value of a certain loss function L

$$\Delta = \mathbb{E}_{\mu,\nu}\left[L(f,x,y)\right]$$

whose allows us to write the push-forward measure for each $f \in \mathscr{H}$ as

$$L_\sharp(f) : (\mathscr{X},\mu) \times (\mathscr{Y},\nu) \to \left(\mathbb{R}_+, \lambda_f\right)$$

where $\lambda_f \in \mathscr{M}_+(\mathbb{R}_+)$. Then assuming that we have observed i.i.d. samples $\mathscr{X}_n$ and $\mathscr{Y}_n$, we can construct i.i.d. samples from the push-forward measure λ_f

$$\mathscr{L}_n(f) \coloneqq \{l_i = L(f,x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (\mathbb{R}_+, \lambda_f)$$

where $l = L(f,x,y)$ and let $L_n \in \mathscr{M}_+(\mathbb{R}_+)$ denote the empirical measure associated with these variables.

Given a functional $\Phi : \mathscr{M}_+(\mathbb{R}_+) \to \mathbb{R}$ (the energy functional), we are interested in computing the law of l_1 under the constraint $\Phi(L_n) \in D$, where D is some Borel subset of $\mathbb{R}$ representing the partition of the non-negative real line corresponding to the compound hypothesis test controlling the FDR at a predefined level.

For every measurable set $A \subset \mathscr{M}_+(\mathbb{R}_+)$ such that $\{L_n \in A\}$ is of positive probability, and every bounded measurable function $g : \mathbb{R}_+ \to \mathbb{R}$, due to exchangeability of the l_i's

$$\begin{aligned}
\mathbb{E}\left[g(l_1) \mid L_n \in A\right] &= \mathbb{E}\left[f(l_i) \mid L_n \in A\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_i g(l_i) \mid L_n \in A\right] \\
&= \mathbb{E}\left[\langle \mathbf{g}, L_n\rangle \mid L_n \in A\right]
\end{aligned}$$

Thus for $A \coloneqq \{\nu : \Phi(\nu) \in D\}$, computing the conditional law of l_1 under the conditioning $\{\Phi(L_n) \in D\} = \{L_n \in A\}$ is equivalent to the computation of the conditional expectation of L_n under the same constraint.

2.1.2 Existence of a Conditional Measure

First we want to show that we can define a push-forward loss measure conditional on the Hypothesis that it has a certain prior distribution. Let $\mathscr{M}_+(\mathbb{R}_+)$ be equipped with the τ-topology

and the cylinder σ-field $\mathscr{B}^{cy}$ (define). For any $\lambda \in \mathscr{M}_+(\mathbb{R}_+)$, let $\lambda^n \in \mathscr{M}_+(\mathbb{R}_+^n)$ denote the induced product measure on $\mathbb{R}_+^n$ and let Q_n be the measure induced by λ^n in $(\mathscr{M}_+(\mathbb{R}_+), \mathscr{B}^{cy})$ through L_n. Let $A_\delta \in \mathscr{B}^{cy}$, $\delta > 0$ be nested measurable sets, i.e. $A_\delta \subseteq A_{\delta'}$ if $\delta < \delta'$. Let F_δ be nested closed sets such that $A_\delta \subseteq F_\delta$. Define $F_0 = \cap_{\delta>0} F_\delta$ and $A_0 = \cap_{\delta>0} A_\delta$ so that $A_0 \subseteq F_0$.

Assumption 2.1. *There exists a $\rho_* \in A_0$ (not necessarily unique) satisfying*

$$KL(\nu_* \mid \lambda) = \inf_{\nu \in F_0} KL(\rho \mid \lambda) := I_F < \infty$$

and for all $\delta > 0$

$$\lim_{n\to\infty} \rho_*^n (\{L_n \in A_\delta\}) = 1$$

where KL represents the KL-divergence.

Think of the following situation as representative: $A_\delta = \{\rho :| \Phi(\rho) |\leq \delta\}$, where $\Phi : \mathscr{M}_+(\mathbb{R}_+) \to [-\infty, \infty]$ is only lower semicontinuous, and thus A_δ is neither open or closed. The nested, closed sets F_δ are then chosen a $F_\delta = \{\rho : \Phi(\rho) \leq \delta\}$ with $F_0 = \{\rho : \Phi(\rho) \leq 0\}$, while $A_0 = \{\rho : \Phi(\rho) = 0\}$. We are then interested in the conditional distribution of l_1 under a constraint of the form $\Phi(L_n) = 0$ (for example a specified average energy).

Theorem 2.1. *[55]Under Assumption 2.1: $\mathscr{M} := \{\rho \in F_0 : KL(\rho \mid \lambda) = I_F\}$ is a non-empty, compact set. Further, for any $\Gamma \in \mathscr{B}^{cy}$ with $\mathscr{M} \subset \Gamma^\circ$,*

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log \lambda^n (L_n \notin \Gamma \mid L_n \in A_\delta) < 0$$

Proof. Note that $A_0 \subseteq F_0$, so $\rho_* \in \mathscr{M}$ by assumption 2.1. Moreover, $I_F < \infty$ implies that $\mathscr{M}$ being the intersection of the closed set F_0 and the compact set $\{\rho : H(\rho \mid \lambda) \leq I_F\}$, is a compact set. Clearly,

$$\begin{aligned}
&\limsup_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log \lambda^n (L_n \notin \Gamma \mid L_n \in A_\delta) \\
&\quad \leq \lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log Q_n (\Gamma^c \cap A_\delta) - \lim_{\delta\to 0} \liminf_{n\to\infty} \frac{1}{n} \log Q_n (A_\delta)
\end{aligned}$$

Let $G := \Gamma^\circ$. Then, since $\Gamma^c \cap A_\delta \subset G^c \cap F_\delta$, with $G^c \cap F_\delta$ being closed set, the upper bound of Sanov's theorem [Ref] yields

$$\begin{aligned}
&\lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log Q_n (\Gamma^c \cap A_\delta) \\
&\quad \leq \lim_{\delta\to 0} \left\{ - \inf_{\rho \in G^c \cap F_\delta} \mathrm{KL}(\rho \mid \lambda) \right\} = - \inf_{\rho \in G^c \cap F_0} H(\rho \mid \lambda) < -I_F
\end{aligned}$$

where the equality follows from the nested, closed sets $G^c \cap F_\delta$, and the strict inequality follows from the closedness of $G^c \cap F_0$ and the definition of $\mathscr{M}$. □

Lemma 2.1. *Under assumption 2.1, for all* $\delta > 0$

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n(A_\delta) \geq -I_F$$

Proof. Since A_δ in general may contain no neighborhood of points from $\mathscr{M}$, the lower bound of Sanov's theorem cannot be used directly. Instead, a direct computation of the lower bound via the change of measure argument will be used in conjunction with the fact that for all $\delta > 0$

$$\lim_{n\to\infty} \rho_*^n(\{L_n \in A_\delta\}) = 1$$

Let ν_* be as in Assumption 2.1. Since $H(\rho_* \mid \lambda) < \infty$, the Radon-Nikodym derivative $f = d\rho_*/d\lambda$ exists. Fix $\delta > 0$ and define the sets

$$\Gamma_n := \left\{ \mathbf{l} \in \mathbb{R}_+^n : g_n(\mathbf{l}) := \prod_{i=1}^n f(l_i) > 0, L_n \in A_\delta \right\}$$

Which implies that $\lim_{n\to\infty} \rho_*^n(\Gamma_n) \to 1$. Hence,

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log Q_n(A_\delta) &\geq \liminf_{n\to\infty} \frac{1}{n} \log \int_{\Gamma_n} \frac{1}{g_n(\mathbf{l})} \rho_*^n(d\mathbf{l}) \\
&= \liminf_{n\to\infty} \frac{1}{n} \log \left(\frac{1}{\nu_*^n(\Gamma_n)} \int_{\Gamma_n} \frac{1}{g_n(\mathbf{l})} \rho_*^n(d\mathbf{l}) \right)
\end{aligned}$$

Therefore, by Jensen's inequality

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log Q_n(A_\delta) &\geq -\limsup_{n\to\infty} \frac{1}{n\nu_*^n(\Gamma_n)} \int_{\Gamma_n} \log\left(g_n(\mathbf{l})\right) \rho_*^n(d\mathbf{l}) \\
&= -H(\rho_* \mid \lambda) + \liminf_{n\to\infty} \frac{1}{n} \int_{(\Gamma_n)^c} \log(f_n(\mathbf{l})) \rho_*^n(d\mathbf{l})
\end{aligned}$$

Note that

$$\int_{(\Gamma_n)^c} \log(f_n(\mathbf{l})) \rho_*^n(d\mathbf{l}) = \int_{(\Gamma_n)^c} f_n(\mathbf{l}) \log(f_n(\mathbf{l})) \lambda^n(d\mathbf{l}) \geq C$$

where $C = \inf_{x\geq 0}\{x \log x\} > -\infty$. Since $H(\rho_* \mid \lambda) = I_F$, the proof is complete. □

The following corollary shows that if ρ_* of assumption 2.1 is unique, then $\lambda^n_{\mathscr{L}_k(f)|A_\delta}$, the law of $\mathscr{L}_k(f)$ conditional upon the event $\{L_n \in A_\delta\}$, is approximately a product measure.

Corollary 2.1. *If $\mathscr{M} = \{\rho_*\}$ then $\lambda^n_{\mathscr{L}_k(f)|A_\delta} \to (\rho_*)^k$ weakly in $\mathscr{M}^1_+(\mathbb{R}^k_+)$ for $n \to \infty$ followed by $\delta \to 0$.*

Proof. Assume $\mathscr{M} = \{\rho_*\}$ and fix $\phi_j \in C_b(\Sigma)$, $j = 1, ..., k$. By the invariance of $\lambda^n_{\mathscr{L}_n(f)|A_\delta}$ with respect to permutations of $\{l_1, ..., l_n\}$,

$$\langle \prod_{j=1}^k \phi_j, \mu^n_{\mathscr{L}_k(f)|A_\delta} \rangle = \frac{(n-k)!}{n!} \sum_{i_1 \neq \cdots \neq i_k} \int_{\Sigma^n} \prod_{j=1}^k \phi_j(l_{i_j}) \lambda^n_{\mathscr{L}_k(f)|A_\delta}(d\mathbf{l})$$

Since,

$$\mathbb{E}\left[\prod_{j=1}^k \langle \phi_j, L_n \rangle \mid L_n \in A_\delta \right] = \frac{1}{n^k} \sum_{i_1, ..., i_k} \int_{\mathbb{R}^n_+} \prod_{j=1}^k \phi_j(l_{i_j}) \lambda^n_{\mathscr{L}_k(f)|A_\delta}(d\mathbf{l})$$

and ϕ_j are bounded functions, it follows that

$$| \langle \prod_{j=1}^k \phi_j, \lambda^n_{\mathscr{L}_k(f)|A_\delta} \rangle - \mathbb{E}\left[\prod_{j=1}^k \langle \phi_j, L_n \rangle \mid L_n \in A_\delta \right] | \leq C \left(1 - \frac{n!}{n^k (n-k)!} \right) \overset{n \to \infty}{\to} 0$$

For $\mathscr{M} = \{\nu_*\}$, Theorem 2.1 implies that for any $\eta > 0$

$$\mu^n \left(| \langle \phi_j, L_n \rangle - \langle \phi_j, \rho_* \rangle | > \eta \mid L_n \in A_\delta \right) \to 0$$

as $n \to \infty$ followed by $\delta \to 0$. Since $\langle \phi_j, L_n \rangle$ are bounded

$$\mathbb{E}\left[\prod_{j=1}^k \langle \phi_j, L_n \rangle \mid L_n \in A_\delta \right] \to \langle \prod_{j=1}^k \phi_j, (\rho_*)^k \rangle$$

so that

$$\limsup_{\delta \to 0} \limsup_{n \to \infty} \langle \prod_{j=1}^k \phi_j, \mu^n_{\mathscr{L}_k(f)|A_\delta} - (\rho_*)^k \rangle = 0$$

Recall that $C_b(\mathbb{R}_+)^k$ is convergence determining for $\mathscr{M}^1_+(\mathbb{R}^k_+)$, hence it follows that $\mu^n_{\mathscr{L}_k(f)|A_\delta} \to (\rho_*)^k$ weakly in $\mathscr{M}^1_+(\mathbb{R}^k_+)$. □

2.1.3 Free Energy Functional

Define the functional $\Phi : \mathscr{M}^1_+(\mathbb{R}_+) \to [-1, \infty]$ by

$$\Phi(\rho) = \langle U, \rho \rangle - 1$$

and consider the constraint

$$\left\{L_n^{\mathbf{Y}} \in A_\delta\right\} := \left\{ \mid \Phi(L_n^{\mathbf{Y}}) \mid \leq \delta \right\} = \left\{ \mid \frac{1}{n}\sum_{i=1}^{n} U(Y_i) - 1 \mid \leq \delta \right\}$$

By formally solving the optimization problem

$$\inf_{\{\nu:\langle U,\rho\rangle=1\}} H(\rho \mid \lambda)$$

one is led to conjecture that ν_* of assumption 1 should be a Gibbs measure, namely one of the measures γ_β, where

$$\frac{d\gamma_\beta}{d\mu} = \frac{e^{-\beta U(f)}}{Z_\beta}$$

and Z_β the partition function, is the normalizing constant

$$Z_\beta = \int_\Sigma e^{-\beta U(f)} \lambda_f(df)$$

Throughout this section, $\beta \in (\beta_\infty, \infty)$ where $\beta_\infty := \inf\left\{\beta : Z_\beta < \infty\right\}$.

Lemma 2.2. *[55] Assume that* $\mu\left(\{x : U(x) > 1\}\right) > 0$, $\mu\left(\{x : U(x) < 1\}\right) > 0$ *and either* $\beta_\infty = -\infty$ *or*

$$\lim_{\beta \searrow \beta_\infty} \langle U, \gamma_\beta \rangle > 1$$

Then there exists a unique $\beta^* \in (\beta_\infty, \infty)$ *such that* $\langle U, \gamma_{\beta^*} \rangle = 1$.

Theorem 2.2. *[55] Let* U, μ *and* β^* *be as in the previous lemma. If either* U *is bounded or* $\beta^* \geq 0$, *then 2.1 applies, with* $\mathscr{M}$ *consisting of a unique Gibbs measure* γ_{β^*}.

Proof. Note that by the monotone convergence theorem, $\langle U, \cdot \rangle = \sup_n \langle U \wedge n, \cdot \rangle$. Since $U \wedge n \in B(\Sigma)$, it follows that $\Phi(\cdot) = \langle U, \cdot \rangle - 1$ is a τ-lower semicontinuous functional. Hence, $F_\delta := \{\nu : \langle U, \nu \rangle \leq 1 + \delta\}, \delta > 0$, are nested closed sets, whereas $F_0 = \{\nu : \langle U, \nu \rangle \leq 1\}$ is convex, closed set. By the previous lemma, $\gamma_{\beta^*} \in F_0$, and by a direct computation

$$H(\gamma_{\beta^*} \mid \mu) = -\beta^* \langle U, \gamma_{\beta^*} \rangle - \log Z_{\beta^*} < \infty$$

implying that $I_F < \infty$. Since $H(\cdot \mid \mu)$ is strictly convex within its I_F level set, it follows that $\mathscr{M}$ contains precisely one probability measure, denoted ν_0. A direct computation, using the

equivalence of μ and γ_{β^*}, yields that

$$\begin{aligned} -H(\nu_0 \mid \gamma_{\beta^*}) &\geq -H(\nu_0 \mid \gamma_{\beta^*}) + \left[H(\nu_0 \mid \mu) - H(\gamma_{\beta^*} \mid \mu)\right] \\ &= \beta^* \left(\langle U, \gamma_{\beta^*}\rangle - \langle U, \nu_0\rangle\right) = \beta^* \left(1 - \langle U, \nu_0\rangle\right) \end{aligned}$$

where the preceding inequality is implied by $\nu_0 \in \mathscr{M}$ and $\gamma_{\beta^*} \in F_0$. For $\beta^* \geq 0$, it follows that $H(\nu_0 \mid \gamma_{\beta^*}) \leq 0$, since $\langle U, \nu_0\rangle \leq 1$. Hence, $\nu_0 = \gamma_{\beta^*}$ and consequently, $\mathscr{M} = \{\gamma_{\beta^*}\}$. Now, Assumption 1 holds for $\nu_* = \gamma_{\beta^*} \in A_0$ as the limit for all $\delta > 0$

$$\lim_{n\to\infty} \nu_*^n \left(\left\{L_n^{\mathbf{Y}} \in A_\delta\right\}\right) = 1$$

follows from the weak law of large numbers. Consequently Theorem 2.1 holds. When U is bounded, then A_δ are closed sets. Therefore, in this case $F_\delta = A_\delta$ can be chosen to start with, yielding $\langle U, \nu_0\rangle = 1$. Consequently, when U is bounded, $\nu_0 = \gamma_{\beta^*} = \nu_*$ even for $\beta^* < 0$. □

Proof. Lemma 2.2:$\log Z_\beta$ is a C^∞ function in (β_∞, ∞). By dominated convergence,

$$\langle U, \gamma_\beta\rangle = -\frac{d}{d\beta} \log Z_\beta$$

and is finite for all $\beta > \beta_\infty$. Then

$$\frac{d}{d\beta}\langle U, \gamma_\beta\rangle = -\int_\Sigma \left(U - \langle U, \gamma_\beta\rangle\right)^2 d\gamma_\beta < 0$$

where the strict inequality follows, since by our assumptions, U cannot be constant μ a.e. Hence, $\langle U, \gamma_\beta\rangle$ is strictly decreasing and continuous as a function of $\beta \in (\beta_\infty, \infty)$. Thus, it suffices to show that

$$\lim_{\beta\to\infty} \langle U, \gamma_\beta\rangle < 1$$

and that when $\beta_\infty = -\infty$,

$$\lim_{\beta\to-\infty} \langle U, \gamma_\beta\rangle > 1$$

To see this, note that by assumption, there exists a $0 < u_0 < 1$ such that $\mu\left(\{x : U(x) < u_0\}\right) > 0$. Now for $\beta > 0$

$$\int_\Sigma e^{-\beta U(x)} \mu(dx) \geq e^{-\beta u_0} \mu\left(\{x : U(x) \in [0, u_0)\}\right)$$

and

$$\begin{aligned}\int_{\Sigma}(U(x)-u_0)e^{-\beta U(x)}\mu(dx)&\\ \leq & e^{-\beta u_0}\int_{\Sigma}(U(x)-u_0)\,1_{\{U(x)>u_0\}}e^{-\beta(U(x)-u_0)}\mu(dx)\\ \leq & \frac{e^{-\beta u_0}}{\beta}\sup_{y\geq 0}\left\{ye^{-y}\right\}\end{aligned}$$

Hence, for some $C<\infty$

$$\langle U,\gamma_\beta\rangle = u_0+\frac{\int_\Sigma (U(x)-u_0)\,e^{-\beta U(x)}\mu(dx)}{\int_\Sigma e^{-\beta U(x)}\mu(dx)}\leq u_0+\frac{C}{\beta}$$

which implies

$$\lim_{\beta\to\infty}\langle U,\gamma_\beta\rangle<1$$

For the other case, when $\beta_\infty=-\infty$, choose $u_2>u_1>1$ such that $\mu\left(\{x:U(x)\in[u_2,\infty)\}\right)>0$. Note that for all $\beta\leq 0$

$$\begin{aligned}\frac{1}{\gamma_\beta\left(\{x:U(x)\in[u_1,\infty)\}\right)}&=1+\frac{\int_\Sigma 1_{\{x:U(x)\in[0,u_1)\}}e^{-\beta U(x)}\mu(dx)}{\int_\Sigma 1_{\{x:U(x)\in[u_1,\infty)\}}e^{-\beta U(x)}\mu(dx)}\\ &\leq 1+\frac{e^{\beta(u_2-u_1)}}{\mu\left(\{x:U(x)\in[u_2,\infty)\}\right)}\end{aligned}$$

implying that

$$\liminf_{\beta\to-\infty}\langle U,\gamma_\beta\rangle\geq u_1\liminf_{\beta\to-\infty}\gamma_\beta\left(\{x:U(x)\in[u_1,\infty)\}\right)\geq u_1$$

and consequently

$$\lim_{\beta\to-\infty}\langle U,\gamma_\beta\rangle>1$$

□

Corollary 2.2. *The Free Energy functional is then given by where*

$$F_\beta[\rho]=\mathbb{E}_\rho\left[\mathbb{E}_{\mu\times\nu}\parallel y-\langle\theta,\Phi(x)\rangle\parallel_2^2+\gamma_\theta J_\delta(\theta)\right]+\beta^{-1}H(\rho)$$

where H is the entropy functional and the optimal measure

$$\rho^*=\arg\min_{\rho\in\mathscr{M}_+^1(\mathscr{H})}F_\beta[\rho]$$

is called the Gibbs Measure.

2.2 Nonlinear Approximations

In the last section we studied how to construct an appropriate cost function which when optimized controls the rate of False Discovery. This is equivalent to studying nonlinear approximations of functions using dictionaries which are over-complete or redundant. In this case these decompositions are not unique, because not all elements of the dictionary are independent. However the non-uniqueness gives the possibility of adaptation, i.e. choosing among the various alternatives the one which minimizes some error for a particular machine learning problem.

In an adaptive representation the aim is to simultaneously achieve sparsity (i.e. fewest significant coefficients) and supper-resolution (i.e. smaller error than nonadaptive methods) using an efficient (i.e. essentially linear time) algorithm. In a practical application one is only able to store finitely many coefficients to represent a function and hence we are usually interested in a m-term approximation. There are two ways of attaining this approximation, Synthesis and Analysis.

2.2.1 Synthesis

The operation of building up a function by superposing elements of the dictionary is called Synthesis [31, 33, 35, 39]. Let Σ_m be the manifold consisting of all functions of the form

$$\Sigma_m := \left\{ f_m^{\mathscr{D}}(x) = \sum_{\psi_\lambda \in \mathscr{D}} \theta_\lambda \phi_\lambda(x) =: \langle \theta, \Phi(x) \rangle : x \in \mathbb{R}^d, \| \theta \|_0 = m \right\} \tag{2.1}$$

where the dictionary is defined by

$$\Phi(x) := \{ \phi_\lambda(x) : \lambda \overset{\text{i.i.d.}}{\sim} \Lambda \}$$

where the parameter λ is uniformly and independently sampled from Λ, which is the parameter space of the dictionary. Many traditional basis learning algorithms like PCA, Kernel Regression, Basis Pursuit, etc can be formulated as a synthesis problem.

Then the synthesis error for the function $f \in L^p(\mathbb{R}^d)$ using an element from Σ_m is given by

$$\varepsilon_m(f)_{L^p} = \inf_{f_m^\Phi \in \Sigma_m} \| f - f_m^\Phi \|_{L^p} \tag{2.2}$$

Here f_m^{Φ} is the approximation obtained by keeping only those elements of the dictionary corresponding to the m largest coefficients. Now such a m-term approximation to a function f is clearly a nonlinear approximation. Since if f has the best m-term approximation f_m^{Φ} and g has g_m^{Φ}, then $f_m^{\Phi} + g_m^{\Phi}$ is not in general made up of m distinct elements of the dictionary, but a subset of Σ_{2m}.

2.2.2 Analysis

Alternatively the operation of associating with each function f, a vector of coefficients attached to the elements of the dictionary is called Analysis and the coefficients are given by

$$\theta = \{\langle f, \psi_\lambda \rangle\}_{\psi_\lambda \in \mathscr{D}} \tag{2.3}$$

which is possibly an infinite dimensional vector, depending on the cardinality of the dictionary. A continuous and differentiable dictionary lends itself well to the back-propagation algorithm in learning the set of coefficients and their corresponding dictionary elements. For a m-term approximation, one starts from an arbitrary collection of m-terms from the dictionary and then using a gradient descent type algorithm, one can optimize a certain cost function corresponding to the machine learning problem at hand. Thus the analysis approach allows us to generate a new class of algorithms which have become popular recently, like in deep convolutional networks, recurrent neural networks, etc.

Synthesis and Analysis are very different operations and care must be taken to distinguish them. One should avoid assuming that the analysis operation gives us coefficients that can be used to synthesize f. One does not uniquely and automatically solve the synthesis problem by applying the analysis operator and analysis is in general clearly not sparsity preserving as every non zero inner product is potentially a member of the solution. Hence care must be taken in defining appropriate regularizers to enforce sparsity.

2.2.3 Fourier Basis

The Fourier transform is probably still the most widely applied linear transform for the representation of functions in various machine learning and signal processing applications [62]. In one dimensions, it reveals the frequency composition of a function, say a time series by transforming it from the time domain into the frequency domain. Using the inner product notation, the function $f(x)$ $x \in \mathbb{R}^d$ is decomposed into inner products on the template functions given by $\{e^{i2\pi\langle\lambda,x\rangle}\}_{\lambda\in\mathbb{R}^d}$

$$\theta_\lambda = \langle f, e^{i2\pi\langle\lambda,x\rangle}\rangle = \int_{x\in\mathbb{R}^d} f(x)e^{-i2\pi\langle\lambda,x\rangle}dx$$

Now under the condition that

$$\int_{x\in\mathbb{R}^d} \mid f(x) \mid^2 dx < \infty$$

we can write the reproducing formula as

$$f(x) = \int_{\lambda\in\mathbb{R}^d} \langle f, e^{i2\pi\langle\lambda,x\rangle}\rangle e^{i2\pi\langle\lambda,x\rangle} d\omega$$

This means the coefficients preserve the norm of the function,

$$\int_{\lambda\in\mathbb{R}^d} \mid f(x) \mid^2 dx = \int_{\lambda\in\mathbb{R}^d} \mid \theta_\lambda \mid^2 d\lambda$$

a result which is known as the Plancheral theorem.

2.2.3.1 *m*-term Approximation

We can then use the Fourier transform to construct the best m-term approximation. Consider the space of functions

$$\mathscr{B}(C) := \left\{ f \in L^2([0,1]^d) : \int_{\mathbb{R}^d} \mid \lambda \mid\mid \hat{f}(\lambda) \mid d\lambda \leq C \right\}$$

then we are interested in the best m-term approximation of the elements of $\mathscr{B}(C)$ using the dictionary $\mathscr{D}_F = \left\{ e^{i2\pi\langle k,x\rangle} \right\}_{k\in\mathbb{Z}^d}$ which forms an orthonormal basis. Then we have the following minimax bound on the synthesis error

$$\sup_{f\in\mathscr{B}(C)} \inf_{f_m^F\in\Sigma_m} \| f - f_m^F \|_{L^2} \leq C \cdot m^{-1/2-1/d} \tag{2.4}$$

where Σ_m is the nonlinear manifold of m-term approximations [2.1]. Roughly this follows from the equivalence (since f is compactly supported)

$$\int_{\mathbb{R}^d} \mid \lambda \mid\mid \hat{f}(\lambda) \mid d\lambda \approx \sum_{k\in\mathbb{Z}^d} \mid k \mid\mid \hat{f}(2\pi k) \mid$$

which is a simple consequence of a famous theorem about the sampling of band-limited functions due to Polya and Plancherel. Therefore, $f \in \mathscr{B}(C)$ implies that the Fourier coefficients $\theta_k(f)$ of f obey

$$\sum_{k\in\mathbb{Z}^d} \mid k \mid\mid \theta_k(f) \mid \leq C$$

Which means that there is bound on the decay of the coefficient sequence of f. Skipping technical details and letting $\mid \theta_{(k)}(f) \mid$ be the kth largest entry in the sequence we have

$$\sum_{k>m} \mid \theta_{(k)}(f) \mid^2 \leq C \cdot m^{-(1+2/d)}$$

which implies [2.4].

2.2.3.2 Limitations of the Fourier Basis

However there are certain limitations with the use of Fourier transforms. First, the Fourier coefficients of a function are essentially associated to the moments of the function with respect to the sampling measure. Even though extremely useful as an aggregate information of the function, it does not reveal its local behavior.

Secondly, for functions on compact domains, one needs to extend the functions periodically to calculate its discrete Fourier transform. This however leads to discontinuities in the function and effects the Fourier coefficients. This effect is called the Leakage. Applying a window to the function to force it to contain a full period can prevent leakage from happening. However, the window itself may contribute frequency information to the function.

Thirdly, violations of the Shannon's sampling theorem causes the actual frequency component to appear at different locations in the frequency spectrum, and is called Aliasing. This can be solved by ensuring the sampling frequency to be at least twice as large as the maximum frequency component contained in the function. However, this requires that the maximal frequency component be known a priori.

Finally, even though the Fourier transform modulus $\mid \widehat{f} \mid$ is translation invariant [1], deformations lead to well-known instabilities at high frequencies. This is illustrated with a small scaling operator, $L_\tau f(x) = f(x-\tau(x)) = f((1-s)x)$ for $\tau(x) = sx$ and $\| \nabla\tau \|_\infty = \mid s \mid < 1$. If $f(x) = e^{i2\pi\langle\lambda,x\rangle}\varphi(x)$, then scaling by $1-s$ translates the central frequency λ to $(1-s)\lambda$. If φ is regular with a fast decay, then

$$\| \mid \widehat{L_\tau f} \mid - \mid \widehat{f} \mid \| \sim \mid s \mid \| \lambda \| \| \varphi \| = \| \nabla\tau \|_\infty \mid \lambda \| \| f \|$$

Since $\mid \lambda \mid$ can be arbitrarily large, $\Phi(f) = \mid \widehat{f} \mid$ is not Lipschitz continuous with respect to scaling at high frequencies.

The frequency displacement from λ to $(1-s)\lambda$ has a smaller impact if sinusoidal waves are replaced by localized functions having a Fourier support that is wider at high frequencies. Such localized functions can also reveal local behavior of the function, as well as solve the

[1] (for $c \in \mathbb{R}^d$, the translation $L_c f(x) = f(x-c)$ satisfies $\widehat{L_c f}(\omega) = e^{-i2\pi\langle c,\omega\rangle}\hat{f}(\omega)$ and hence $\mid \widehat{L_c f}(\omega) \mid = \mid \widehat{f} \mid$)

problem of leaking since they can be applied on compact intervals. The wavelet basis provides such a set of localized functions.

2.2.4 Wavelet Basis

A straightforward solution to overcoming the limitations of the Fourier transform is to introduce an analysis ball of a certain radius that glides though the domain of the function to perform a localized Fourier transform. In one dimension such an algorithm is called the Short Time Fourier Transform. It solves many of the problems discussed earlier, but still requires the ball to be of constant radius and by Balian-Low Theorem [30, 33] there is no good local orthogonal Fourier basis.

A wavelet basis on the other hand enables windows of variable radius, and is constructed by scaling (i.e. dilation and contraction) and shift (or any other group action $g \in G$, where G is a group defined on the domain like translation, reflection, rotation [2], etc or combinations thereof) of a mother wavelet $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ under the condition that ψ satisfies

$$K_\psi = \int_{\mathbb{R}^d} \frac{\| \hat{\psi}(\omega) \|^2}{\omega^d} d\omega < \infty \tag{2.5}$$

where $\hat{\psi}$ is the Fourier transform of ψ. This condition says that the function ψ is oscillatory and has vanishing moments up to about $d/2$. Then the dictionary of wavelets can be written as

$$\mathscr{D}_W = \left\{ \psi_\lambda := a^{jd/2} \psi\left(a^j \cdot g^{-1} \cdot x\right) : \lambda \in \Lambda = \{a > 1, j \in \mathbb{Z}, g \in G\} \right\} \tag{2.6}$$

which is known as a semi-discrete frame, due to the discrete nature of the scaling exponent. The Fourier transform of the elements of the dictionary have a fairly simple form given by

$$\widehat{\psi}_\lambda(\omega) = \widehat{\psi}(a^{-j} g^{-1} \omega)$$

which allows us to model the mother wavelet in the Fourier domain where many of the associated operations like convolutions, differential calculus, etc become simple linear algebraic equations.

Using the inner product notation, we can now calculate the coefficients of the wavelet transform of the function $f \in L^2(\mathbb{R}^d)$ as

$$\theta_\lambda = \langle f, \psi_\lambda \rangle = a^{-jd} \int_{\mathbb{R}^d} f(x) \psi^*\left(a^j \cdot g^{-1} \cdot x\right) dx \tag{2.7}$$

[2] If d is even, then G is a subgroup of $\mathrm{SO}(d)$; if d is odd, then G is a finite subgroup of $\mathrm{O}(d)$.

and as before we have Calderon's reproducing formula as [3]

$$f(x) = \int_{\Lambda} \theta_{\lambda} \psi_{\lambda}(x) d\lambda$$

along with the Plancheral theorem

$$\int_{\mathbb{R}^d} | f(x) |^2 dx = \int_{\Lambda} | \theta_{\lambda} |^2 d\lambda$$

which essentially means that the wavelet transform is a norm preserving linear transform.

Example 2.1. Let $d = 1$ and $G = \{-1, 1\}$ which represents the reflection group. To build a complex wavelet ψ concentrated on a single frequency band, let

$$\psi(x) = e^{i\eta \cdot x} \theta(x)$$

where $\widehat{\theta}(\omega)$ is a real function concentrated in a low frequency ball at $\omega = 0$ whose radius is of the order π. Then the Fourier transform of ψ is given by

$$\widehat{\psi}(\omega) = \widehat{\theta}(\omega - \eta)$$

where we set

$$\widehat{\psi}(\omega) = 0 \text{ for } \omega < 0$$

As a result, $\widehat{\psi}(\omega)$ is real and concentrated in a frequency ball of the same radius but centered at $\omega = \eta$ satisfying [2.5]. To simplify notation, we denote $\lambda = a^j g \in a^{\mathbb{Z}} \times G$, with $| \lambda | = a^j$. After dilation and reflection,

$$\widehat{\psi}_{\lambda}(\omega) = \widehat{\theta}(\lambda^{-1}\omega - \eta)$$

covers a ball centered at $\lambda_j \eta$ with a radius proportional to $| \lambda | = a^j$. The index λ thus specifies the frequency localization and spread of $\widehat{\psi}_{\lambda}$. For example when $\hat{\theta}(\omega) = \exp(-\omega^2/2)$ the wavelet generated ψ is known as the Morlet wavelet.

2.2.4.1 *m*-term Approximation

As before like for the Fourier series, we can construct the best n-term approximation using the dictionary of wavelet bases $\mathscr{D}_W$. The importance of wavelets is not so much that they provide a simultaneous (approximate) time and frequency localization which yields a description of many spaces, such as Sobolev, Besov and Triebel-Lizorkin spaces, etc [24, 30, 31]in terms

[3] Here an integral over integers is simply defined as the summation, to keep the notation clean.

of the coefficients in wavelet decompositions and a description of important classes of operators, such as the Calderon-Zygmund operators, in terms of almost diagonal matrices. This means the space of functions under consideration go beyond just the Hilbert space of square integrable functions but to practically any $L^p(\mathbb{R}^d)$ function for $0 < p < \infty$.

Let Σ_m be the nonlinear manifold of m-term approximations [2.1] then we can approximate functions $f \in L^p(\mathbb{R}^d)$ by elements from Σ_m with the synthesis error

$$\varepsilon_m(f)_{L^p} = \inf_{f_m^W \in \Sigma_m} \| f - f_m^W \|_{L^p}$$

Then for each $0 < p < \infty$ and for a certain range of α,

$$\sum_{m=1}^{\infty} \left[m^{\alpha/d} \varepsilon_m(f)_{L^p} \right]^{\tau} \frac{1}{m} < \infty \iff f \in B_\tau^\alpha$$

where B_τ^α are the Besov spaces with $\tau := (\alpha/d + 1/p)^{-1}$ with the norm given by

$$\| f \|_{B_\tau^\alpha} = \left(\| \left[(1+ | \widetilde{\omega} |^2)^{\frac{\lfloor \alpha \rfloor}{2}} \hat{f} \right] \|_{L^p(\mathbb{R}^d)}^{\tau} + \int_0^\infty | \frac{c_p^2(f^{(\lfloor \alpha \rfloor)}, t)}{t^\alpha} |^{\frac{1}{t}} dt \right)$$

where $\tilde{f}$ is the inverse Fourier transform of f and

$$c_p^2(f^{(\lfloor \alpha \rfloor)}, t) = \sup_{|h| \le t} \| \Delta_h^2 f \|_p$$

with $\Delta_h f(x) = f(x-h) - f(x)$. The Besov norm implies that the Besov space consists of $L^p(\mathbb{R}^d)$ functions whose derivates upto $\lfloor \alpha \rfloor$ degree have a finite norm, along with having only finitely valued jumps. Therefore Besov space covers a wider variety of functions observable in a real world application.

A consequence of the above result implies the following inequalities hold for some $\beta > \alpha$

$$\varepsilon_m(f)_{L^p} \le C m^{-\beta/d} \| f \|_{B_\tau^\beta}, \; f \in B_\tau^\beta \tag{2.8}$$

and

$$\| f_m^W \|_{B_\tau^\beta} \le C m^{\beta/d} \| f_m^W \|_{L^p(\mathbb{R}^d)}, \; f_m^W \in \Sigma_m \tag{2.9}$$

The estimate [2.8] is called the Jackson Estimate, while [2.9] is called the Bernstein estimate. From the Jackson estimate we can derive the direct comparison to the minimax synthesis error

estimate for Fourier transforms on compact domain as

$$\sup_{f \in L^p(\mathbb{R}^d)} \inf_{f_m^W \in \Sigma_n} \| f - f_m^W \|_{L^p} \leq C m^{-1-\alpha/d}$$

Which means that using the wavelets as defined here, the m-term approximation provides a faster rate of convergence to a bigger class of target functions than a Fourier series approximation.

2.2.4.2 Limitations of Wavelet Bases

Wavelet bases as we saw above are optimal for point like singularities like jumps and point like noise. However higher dimensional singularities like boundaries cannot be well represented and leads to higher number of coefficients. Further, in both Fourier as well as Wavelet bases we have discussed till now are orthogonal bases and in which case synthesis and analysis coefficients can be used interchangeably. However, since we are interested in an adaptive representation which simultaneously achieves sparsity and super-resolution, it is necessary to introduce a redundant basis, like the Littlewood Paley wavelets.

2.3 GWAS Data

Genome-wide association studies (GWAS) can be used to map an entire species genome for association of a trait of interest and millions of SNPs [81, 94, 107, 114]. By using the genotype of each SNP as predictor of the trait of interest, GWAS fits p independent univariate linear models from p SNPs and n samples. The coefficient estimate β of the corresponding SNP is used to determine the significance of association (P-value) in each of the p tests. The resulting values are adjusted using multiple hypothesis testing methods such as Bonferroni, Benjamini-Hochberg to control for FDR [40, 81].

It is essential to distinguish between association and linkage mapping, or quantitative trait loci (QTL) mapping. Association mapping uses high-density SNP genotyping of unrelated individuals, the output of this experiment are point mutations in the genome. Linkage mapping requires controlled breeding experiments and relies on the segregation of fewer markers, here the output are chromosomal regions.

In a GWAS study usually there are three types of datasets generated, which together provide the necessary information. The .raw file contains all the SNPs, the .fam file contains information about the observations including the phenotype and usually consists of six columns. The .map file on the other hand contains the mapping information about the SNPs but is not

necessary even though recommended, as their exclusion leads to less informative plots and summaries.

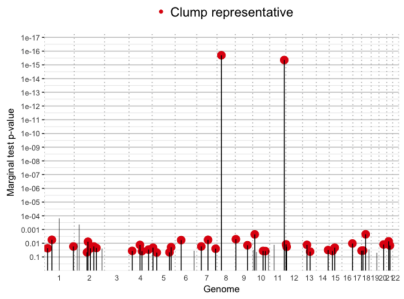

Figure 2.1: Clump Distribution

For large datasets SNPs are usually filtered with their marginal test p-values. All SNPs with p-values over a certain threshold are truncated. Among these filtered SNPs one performs clumping which for a given parameter $\rho \in (0,1)$ is as follows:

1. Each SNP corresponds to a feature and when performing regression, we get the tail-error probability for the corresponding sequence of coefficients.

2. Fit an infinite mixture model which is a non-parametric Bayesian clustering with respect to the correlation coefficient between the different SNPs. Here the distance is defined to be inverse of the strength of its correlation (say Pearson correlation).

3. The clusters obtained are called clumps.

Finally one performs LASSO or in our case SLOPE regression on these clustered SNPs to identify the important clumps or clusters of SNPs which have an effect on the phenotype, while controlling the False Discovery Rate.

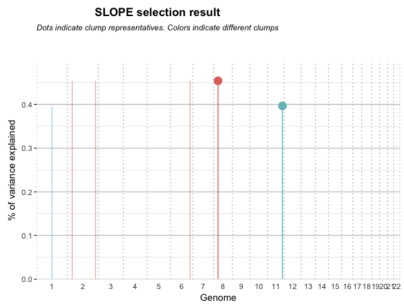

Figure 2.2: Selection of Different Centers of Influence

Chapter 3

Optimal Classification by Controlling rate of False Discovery

3.1 Introduction

In the previous chapter we studied the tools necessary to design and develop arbitrary machine learning algorithms from a theoretical point of view. Now we consider a practical application in which the multi-classification problem is tackled. The problem of classification or discrimination arises whenever one wants to associate a given sample from an arbitrary probability measure to one of a finite number of categories, based on a given training data set.

More precisely, let $\mu_1, ..., \mu_K \in \mathscr{M}_+^1(\mathbb{R}^d)$ be Borel Probability measures defined on the Euclidean space, $\mathbb{R}^d$. Then any observation is assumed to have been sampled in the following way: randomly select one of $k = 1, ..., K$ with probability π_k, then the take a sample from $\mathbb{R}^d$, according to the measure μ_k. For notational convenience, let $\bar{\mu} = (\mu_1, ..., \mu_K)$ and $\pi = (\pi_1, ..., \pi_K) \in \mathscr{S}_K$, which is the unit simplex in $\mathbb{R}^K$ whose elements are nonnegative and sum to 1. We assume that the underlying measures corresponding to the various classes are unknown but a paired training samples of size n is available, which we denote by $(\mathscr{X}_n, \mathscr{Y}_n)$. Further we also assume that these K sets of samples are independent of each other.

In this setting, the Hypothesis space $\mathscr{H}$, consists of measurable functions which define the classification rule for each new data point. These functions, thus depending on the prior probability vector π and the set of class measures $\bar{\mu}$, map a given data point $\mathbf{x} \in \mathbb{R}^d$ to a label between $\{1, ..., K\}$ i.e.

$$f \in \mathscr{H} : \mathbb{R}^d \times \mathscr{S}_K \times \left(\mathscr{M}_+^1(\mathbb{R}^d)\right)^K \to \mathscr{Y} = \{1, ..., K\}$$

Clearly in a real world application we do not have access to the measures corresponding to

each class, and the classification problem is then to estimate the function $\hat{f}$ using observations $\mathscr{X}_n^k$ from each μ_k. If the prior probability vector π is assumed to be known, the parametrized form of the Hypothesis space used for classification depends purely on the class measures $\bar{\mu}$. Recently, for complex data sets like images, audio, etc Convolutional Neural Networks have had much success in modeling these class measures and thereby estimating the corresponding classification rule.

In order to define loss functions on the Hypothesis space which characterize the performance of its elements, it is useful to consider the problem from a decision theoretic point of view. An arbitrary loss function L is a non-negative real valued function

$$L : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$$

where $L(i, j)$ denotes the loss when for a given data point, one guesses i, but the classification rule maps to j.

These $L(i, j)$ are allowed to be different to quantify any feelings the experimenter may have about one type of mistake being worse than another. For example, in the diagnosis of disease, it can be worse to classify a sick person as healthy, than to make the opposite error. The only assumption needed regarding the loss L is

$$\max_i L(i,i) < \min_{i \neq j} L(i,j)$$

and it is convenient to define

$$\underline{L} = \min_{i \neq j} L(i,j) - \max_i L(i,i) > 0 \text{ and } \bar{L} = \max_{i,j} \mid L(i,j) \mid$$

Now for any observation $\mathbf{x} \in \mathbb{R}^d$ and $\pi \in \mathscr{S}_K$, the posterior probability of the class $i \in \{1, ..., K\}$, is given by

$$\Pr\left(f(\mathbf{x}, \pi, \bar{\mu}) = i \mid X = \mathbf{x}\right) = \frac{\pi_i \rho_i(\mathbf{x})}{\sum_j \pi_j \rho_j(\mathbf{x})} \tag{3.1}$$

where $\rho_i = \frac{d\mu_i}{d\lambda}$ represents the Radon-Nikodym density of the ith class with respect to some common base measure λ. We can then write the corresponding risk functional as a conditional expectation with respect to the event $X = \mathbf{x}$, for any decision rule $f \in \mathscr{H}$ as

$$R(f = k, \pi, \mathbf{x}) = \sum_i L(k,i) \Pr\left(f = i \mid X = \mathbf{x}\right) \tag{3.2}$$

where $R(f(\mathbf{x}, \pi, \bar{\mu}) = k, \pi, \mathbf{x})$ is denoted by $R(f = k, \pi, \mathbf{x})$. The Bayes optimal risk can then be defined for each $\mathbf{x} \in \mathbb{R}^d$ and each $\pi \in \mathscr{S}_K$, as:

$$R\left(\hat{f}_{\text{Bayes}}, \mathbf{x}, \pi\right) = \min_{k=1,\ldots,K} R(k, \mathbf{x}, \pi) \tag{3.3}$$

When the minimum is not unique, the manner in which ties are broken is irrelevant. The loss function appearing most often in the literature is $0-1$ loss, which is suitable for Binary Hypothesis testing and classification and is defined as

$$L(i, j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

In this case the Bayes optimal classifier is given by

$$\begin{aligned} \hat{f}_{\text{Bayes}} &= \arg\min_k \sum_{i \neq k} \Pr(f = i \mid X = \mathbf{x}) \\ &= \arg\min_k \Pr(f \neq k \mid X = \mathbf{x}) \\ &= \arg\max_k \Pr(f = k \mid X = \mathbf{x}) \end{aligned}$$

which is the classical MAP (Maximum A Posteriori) rule for binary decisions or testing. In the rest of the work we show that the consistency of the Bayes optimal classifier depends on the consistency of the relative density estimator for each class. Subsequently we review the Gaussian Mixture Model, the Reproducing Kernel Hilbert Space and the Convolutional Neural Network as density estimators, each of which can be used to build a classifier depending on the complexity of the given data set. Then we look at a family of performance measures and show how to generate loss functions from them which encompass a huge class of examples studied in the literature which may be suitable under different situations. Finally we provide an algorithm to learn the Bayes Optimal Classifier in all these scenarios and apply it to xyz data set.

3.2 Hypothesis Space

Clearly when both π and $\bar{\mu}$ are known, then it is simple to compute the Bayes classification rule. In this work we assume that π is known, however $\bar{\mu}$ is unknown. This is not such a restrictive assumption, since when π is unknown, its estimators converge much faster than the estimators of $\bar{\mu}$. This assumption however tells us that a Bayes Optimal Classifier is only

dependent on the estimator of $\bar{\mu}$ and thus the Hypothesis space essentially only needs to model these class measures.

Since we only have paired samples $(\mathscr{X}_n, \mathscr{Y}_n)$ from these class measures $\bar{\mu}$, we need to ensure any classifier $\hat{f}_n$ which is defined as

$$\hat{f}_n \in \mathscr{H} : \mathbb{R}^d \times \mathscr{S}_K \times \mathscr{X}_n \to Y \in \mathscr{Y}_n$$

at least asymptotically converges to $\hat{f}_{\text{Bayes}}$, i.e.

$$\lim_{n\to\infty} R\left(\hat{f}_n, \boldsymbol{\pi}, \mathbf{x}\right) = R(\hat{f}_{\text{Bayes}}, \boldsymbol{\pi}, \mathbf{x})$$

in a certain mode of convergence. Such classifiers are said to be Bayes Risk Consistent [38, 56, 78].

To define the mode of convergence, first fix a compact set $\mathscr{C} \subset \mathbb{R}^d$, which has nonempty interior. Then the estimator $\hat{f}_n$ is said to be Bayes Risk Consistent if

$$\lim_{n\to\infty} \int_{\mathscr{S}_K} \int_{\mathscr{C}} \left[R\left(\hat{f}_n, \boldsymbol{\pi}, \mathbf{x}\right) - R(\hat{f}_{\text{Bayes}}, \boldsymbol{\pi}, \mathbf{x})\right] d\mathbf{x} d\boldsymbol{\pi} = 0 \tag{3.4}$$

Absolute values are not required because, for $k = 1, \ldots, K$ and for each $\mathbf{x} \in \mathbb{R}^d$ and each $\boldsymbol{\pi} \in \mathscr{S}_K$

$$R\left(\hat{f}_n, \boldsymbol{\pi}, \mathbf{x}\right) \geq R(\hat{f}_{\text{Bayes}}, \boldsymbol{\pi}, \mathbf{x})$$

Here integrating with respect to $\boldsymbol{\pi}$ and $\mathbf{x}$ removes the effect of local irregularities to the convergence rate.

Further if we expand the consistency condition (3.4) above we see that it is equivalent to

$$\lim_{n\to\infty} \int_{\mathscr{S}_K} \int_{\mathscr{C}} \left[\min_k \sum_{i=1}^{K} L(k,i) \left[\frac{\pi_i \hat{\rho}_i(\mathbf{x})}{\sum_j \pi_j \hat{\rho}_j(\mathbf{x})} - \frac{\pi_i \rho_i(\mathbf{x})}{\sum_j \pi_j \rho_j(\mathbf{x})}\right]\right] d\mathbf{x} d\boldsymbol{\pi} = 0$$

which means that the rate of convergence of this quantity is exclusively dependent on the rate of convergence of the density estimate $\hat{\rho}_i \to \rho_i$ as $n \to \infty$. Therefore comparing different classifiers, is equivalent to comparing the rate of convergence of the corresponding probability density estimators [Ref]. There is a constant $c_3 > 0$ and a density estimator $\hat{f}_N(x, X^1, \ldots, X^N)$ so that, when $r = {}^{2p}/_{(2p+d)}$

$$\lim_{N\to\infty} \sup_{\mathbf{f}\in\mathscr{F}_1} \Pr\left[\int_{\mathscr{C}} \left[\hat{f}_N(\mathbf{x}) - f(\mathbf{x})\right]^2 d\mathbf{x} > c_3 N^{-r}\right] = 0$$

The performance of density estimators depend on the assumptions on the target distribution

and the properties of the Hypothesis space. In the previous chapter we showed that the relative densities of the form $\rho_i = \frac{d\mu_i}{d\lambda}$, which are dependent on the data only through bounded number of features, belong to the exponential family i.e. are of the form

$$d\mu_i(\mathbf{x}) = e^{\langle \theta_i, \Phi(\mathbf{x}) \rangle - A(\theta_i)} d\lambda$$

We also discussed the nature of the feature extraction map $\Phi(\mathbf{x})$, which can be chosen depending up on the data set under consideration. For simple low dimensional data sets, if class densities can be modeled as Gaussian distribution, then

$$\Phi(\mathbf{x}) = (\langle 1, \mathbf{x} \rangle, \langle \mathbf{x}, \mathbf{x} \rangle)^T$$

and we get the Gaussian mixture model for classification. Other formulations of sufficient statistics, lead to a different mixture models with each component being in the exponential family.

We might also consider a Reproducing Kernel Hilbert space embedding approach where for $\mathbf{x}_i \in \mathscr{X}_n^k$,

$$\Phi(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), ..., K(\mathbf{x}_n, \mathbf{x}))^T$$

is the kernel map corresponding to some characteristic kernel K, like the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})\right)$. In this case as the number of parameters increases with the number of data points, the classification algorithm derived is known as a non-parametric method. However, due to the properties of the Kernel, points which are far away from each other do not influence one another and hence the number of non-zero parameters in the end is not unbounded.

Since 2012 [79] Convolutional Neural Networks have provided state of the art results for classification of complicated data sets, like images, audio, etc which provide a deep distributed representation of features. Distributed representations allow an exponential increase in the capacity for representation of features by re-using multiple examples that are not necessarily neighbors of each other, which is the case for example in the RKHS model mentioned above.

In deep models the features in addition are hierarchically organized, i.e. more abstract features are constructed using less abstract ones. This architecture further promotes the re-use of features and leads to the construction of progressively more abstract features at higher layers of representation. Crucially deep representations have exponentially large number of paths from the data to the final feature, with respect to its depth. Clearly when the same family of functions can be represented using fewer features, we should expect to be able to learn the parameters using fewer examples, yielding improvements in both computational and statistical

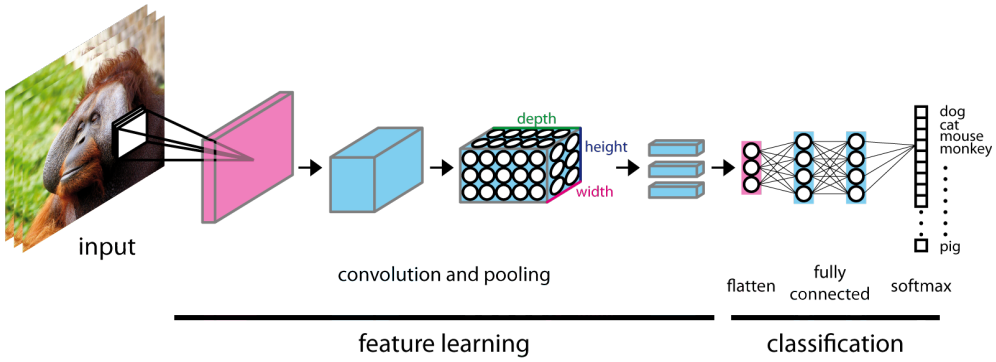

Figure 3.1: Convolutional Neural Network Classifier

efficiency (less parameters to learn, and re-use of these parameters over many different kinds of inputs).

3.3 Loss Functions

Consider the confusion matrix corresponding to any arbitrary classifier f, denoted by $C[f] \in [0,1]^{K\times K}$, with entries defined by

$$C_{ij}[f] = \Pr(f(\mathbf{x}) = i, y = j)$$

for all $(\mathbf{x}, y) \in (\mathscr{X}_n, \mathscr{Y}_n)$ where $\sum_{i,j} C_{ij}[f] = 1$. One can then define a general performance measure of a classifier as a non-negative valued differentiable function

$$\psi : C[f] \to \mathbb{R}_+$$

of the confusion matrix, which is monotonically increasing along the diagonal and monotonically non-increasing along the off-diagonal elements.

In this case higher values of the ψ correspond to better performance. Then the optimal ψ-performance over all feasible confusion matrices, which is a convex set is given by its gradient, which when normalized into $[0,1]^{K\times K}$ is defined as the necessary loss function

$$L := -\widetilde{\nabla\psi(C)}$$

Clearly the classifier for which the loss is minimized, also maximizes the the performance measure ψ. This formulation captures both common loss-based performance measures, which are effectively linear functions of the entries of the confusion matrix as well as the more

complicated performance measures like the G-mean, micro F_1 measure, etc.

Binary Classification Consider binary classification, where $K = 2$ and the labels are indexed as $Y = \{0,1\}$. Then the confusion matrix of a binary classifier can be defined as

$$C = \begin{bmatrix} C_{0,0} = \text{True Negative} & C_{0,1} = \text{False Positive} \\ C_{1,0} = \text{False Negative} & C_{1,1} = \text{True Positive} \end{bmatrix}$$

Then a large class of existing binary performance measures can be written as

$$\psi(C) = \frac{\langle A, C \rangle}{\langle B, C \rangle}$$

for some $A, B \in \mathbb{R}^{K \times K}$ with $\langle B, C \rangle > 0$. Then the corresponding loss function is given by

$$L = -\widetilde{\nabla \psi(C)} = -\left(\widetilde{A - tB}\right)$$

where $t \in \mathbb{R}_+$ is a non-negative scalar quantity.

In this framework many of the traditional loss functions and performance measures can be defined. For example consider $\psi^{0-1}(C) = \sum_{ii} C_{ii}$, then $L = -\widetilde{\nabla \psi(C)} = \left[\delta_{i \neq j}\right]_{i,j}$ which is the standard $0-1$ loss for classical classification accuracy. The 'balanced accuracy' or AM measure [56] given by

$$\psi^{\text{AM}}(C) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)$$

or the F_β measure ($\beta > 0$) [66, 78] given by

$$\psi^{F_\beta}(C) = \left(\frac{(1+\beta^2)\text{TP}}{(1+\beta^2)\text{TP} + \beta^2 \text{FN} + \text{FP}}\right)$$

can all be expressed. Further the case of ordinal regression [Ref] can also be expressed similarly

$$\psi^{\text{ord}}(C) = \sum_{ij}(1 - \frac{1}{n-1} \mid i-j \mid)C_{ij}$$

where the corresponding loss is given by $L_{ij}^{\text{ord}} = \frac{1}{n-1} \mid i-j \mid$.

Multi-classification The G-mean measure [66, 78] is used to evaluate both binary and multi-class classifiers in settings with class imbalance and is given by

$$\psi^{\mathrm{GM}}(C) = \left(\prod_{i=1}^{K} \frac{C_{ii}}{\sum_{j=1}^{K} C_{ij}} \right)^{1/K}$$

The micro F_1-measure [66, 78] is also a widely used measure to evaluate multi-class classifiers in information retrieval and information extraction applications and is given by

$$\psi^{\text{micro-}F_1}(C) = \frac{2\sum_{i=2}^{n} C_{ii}}{2 - \sum_{i=1}^{n} C_{1i} - \sum_{i=1}^{n} C_{i1}}$$

3.4 Diabetic Retinopathy Dataset

Diabetes can lead to an eye disease called diabetic retinopathy (DR), which causes high blood sugar levels leading to damage to blood vessels in the retina. Vision can be impaired by growth of abnormal new blood vessels, or swelling, leaking and closing of existing ones.

The public Diabetic Retinopathy dataset provided by Kaggle.com consists of high-resolution retina images taken under a variety of imaging conditions. Each subject is assigned an ID, and a left and right field image is provided for every subject. A scale of 0 to 4 (0 - No DR, 1 – Mild, 2 – Moderate, 3 – Severe, 4 - Proliferative DR) was defined by a clinician, rating the presence of diabetic retinopathy of the images.

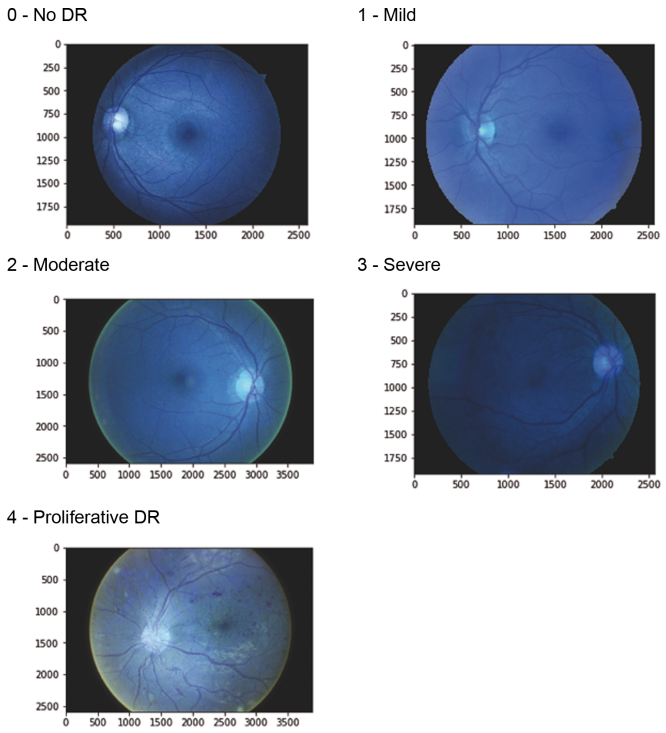

Figure 3.2: Gold Standard Classification of Diabetic Retinopathy

The task given by Kaggle was to create an automated analysis system, which can be used to assign a score to the provided images based on this scale. The images provided can differ in visual appearance, as they were recorded with different models of cameras. Some images can be displayed anatomically, e.g. for the right eye with the macula on the left and the optic nerve on the right. Other images can be inverted, as they would be seen in a live eye exam.

There are two characteristics to identify the inverted images: either the macula is slightly higher than the midline through the optic nerve, or there is no notch on the side of the image (square, triangle, or circle). Both, the images and their labels might show noise. In addition, the images can contain artifacts, be out of focus, underexposed, or overexposed. Thus, it is essential that the chosen algorithm is robust and not sensitive to noise and variation.

Preprocessing Images To remove unnecessary variation among the images, preprocessing is necessary. We resize, crop and normalize all the images so that the eye is always in the center of the image.

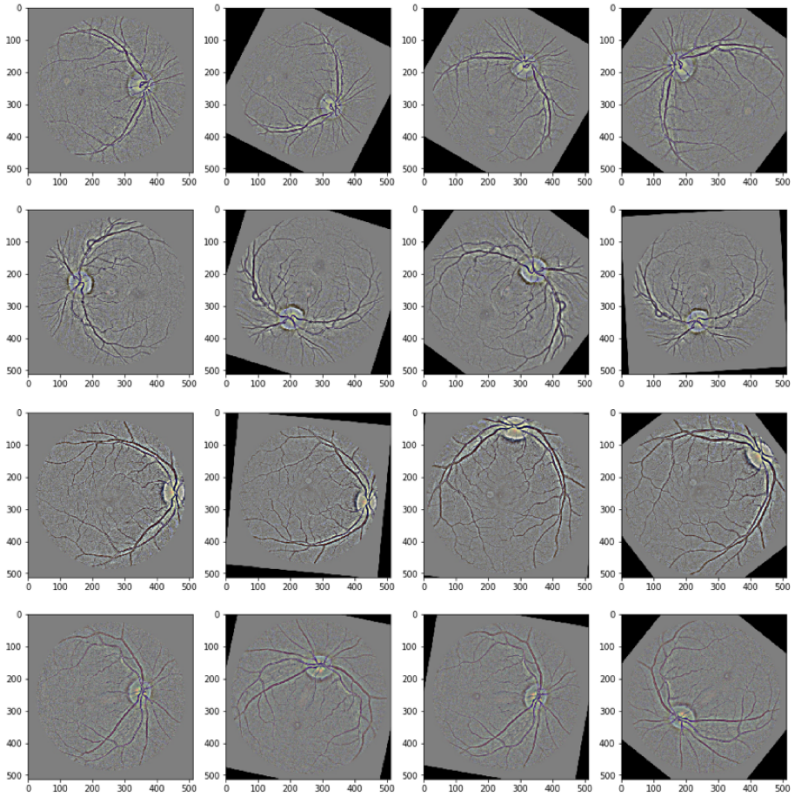

Figure 3.3: Augmented and Preprocessed Images

Data Augmentation Since deep learning models are very flexible, we need large datasets to avoid overfitting. In this example, we randomly rotate, flip and scale the images, and create extra images virtually. This data augmentation technique allows the model to learn invariance to these random transformations.

Convolutional Neural Network We use a 7 layer convolutional neural network, with architecture as shown in the figure which is followed by a dense layer which performs the classification. We use

$$\psi(C) = \frac{\langle A, C \rangle}{\langle B, C \rangle}$$

as the classification performance measure where C is the confusion matrix.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 510, 510, 32)	896
conv2d_2 (Conv2D)	(None, 508, 508, 32)	9248
max_pooling2d_1 (MaxPooling2	(None, 254, 254, 32)	0
conv2d_3 (Conv2D)	(None, 252, 252, 64)	18496
conv2d_4 (Conv2D)	(None, 250, 250, 64)	36928
max_pooling2d_2 (MaxPooling2	(None, 125, 125, 64)	0
conv2d_5 (Conv2D)	(None, 123, 123, 96)	55392
conv2d_6 (Conv2D)	(None, 121, 121, 96)	83040
max_pooling2d_3 (MaxPooling2	(None, 60, 60, 96)	0
conv2d_7 (Conv2D)	(None, 58, 58, 128)	110720
conv2d_8 (Conv2D)	(None, 56, 56, 128)	147584
max_pooling2d_4 (MaxPooling2	(None, 28, 28, 128)	0
conv2d_9 (Conv2D)	(None, 26, 26, 192)	221376
conv2d_10 (Conv2D)	(None, 24, 24, 192)	331968
max_pooling2d_5 (MaxPooling2	(None, 12, 12, 192)	0
conv2d_11 (Conv2D)	(None, 10, 10, 256)	442624
conv2d_12 (Conv2D)	(None, 8, 8, 256)	590080
max_pooling2d_6 (MaxPooling2	(None, 4, 4, 256)	0
conv2d_13 (Conv2D)	(None, 2, 2, 256)	590080
max_pooling2d_7 (MaxPooling2	(None, 1, 1, 256)	0
flatten_1 (Flatten)	(None, 256)	0
dense_1 (Dense)	(None, 256)	65792
dense_2 (Dense)	(None, 5)	1285

Total params: 2,705,509
Trainable params: 2,705,509
Non-trainable params: 0

Figure 3.4: Convolutional Neural Network Architecture

Stochastic Gradient Descent Learning We optimize the risk functional based on the convolutional network and the above defined cost function using the standard stochastic gradient descent algorithm with periodically decreasing learning rates.

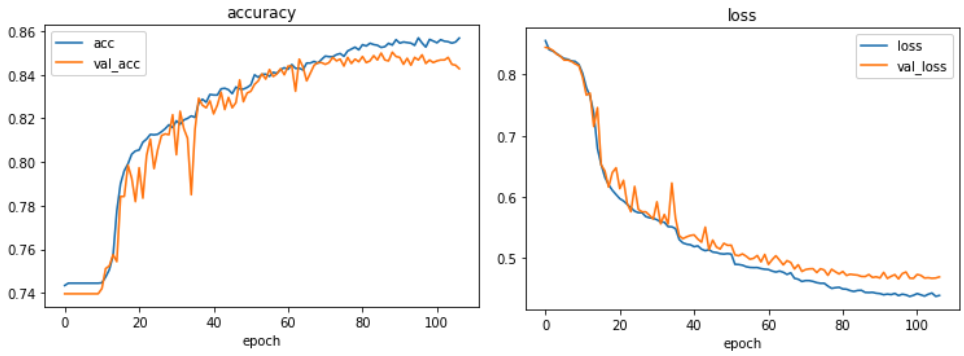

Figure 3.5: Accuracy and Loss

Result We see from the normalized confusion matrix that not all classes perform equally well, as some types of images are harder to extract features from than others. For example the neural network is not able to distinguish well between no diabetic retinopathy and mild levels however is able to distinguish fairly well among well separated categories.

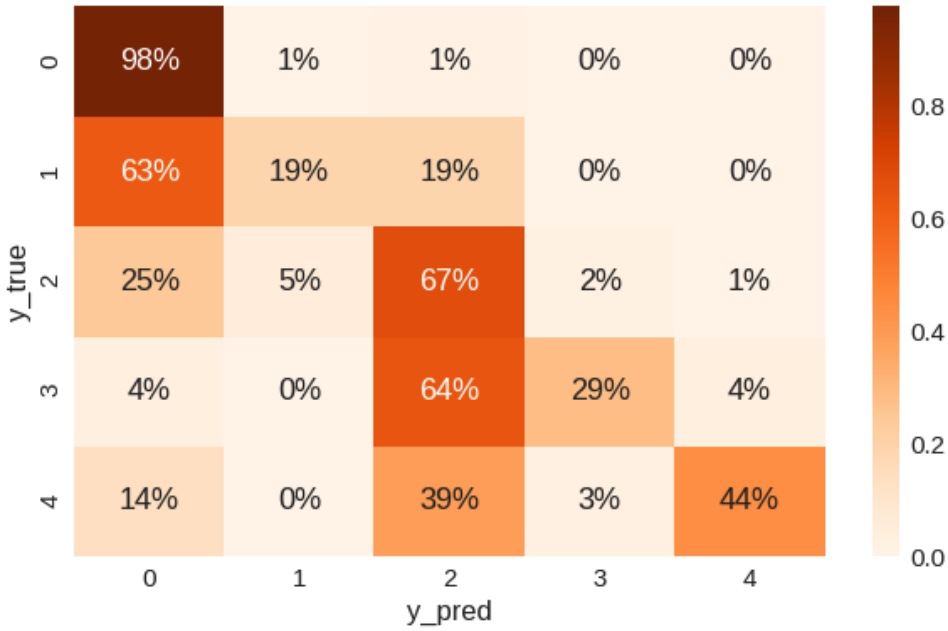

Figure 3.6: Normalized Confusion Matrix

Chapter 4

Non-parametric Bayesian Inference of Count Data

4.1 Introduction

RNA is a key intermediate between the genome and the proteome and has proven to be useful in providing a means to study gene expression. Typically 90% of the RNA is of the type ribosomal RNA (rRNA) which is essential for protein synthesis, while messenger RNA (mRNA) which is only 1-2% conveys the genetic information which directs synthesis of specific proteins. Therefore usually in a RNA-seq experiment only mRNAs are studied [50].

In the experiment mRNA is isolated, reverse transcribed into complementary (cDNA) and then shattered into small pieces. These are then sequenced, giving a list of short sequences called reads which are subsequently mapped back (using an appropriate algorithm) onto the reference genome. Finally, for a set of regions of interest on the genome, such as genes, exons, or junctions, we count the number of reads mapped unambiguously to each of them, and use this count as a measure of expression for the region. The number of reads observed however also depends on the sequencing depth, which is the total number of reads sequenced and may vary with the experiment. Thus two technical replicates would have different counts if the sequencing depth is different, and hence needs to be corrected for while analyzing the data [50].

In this work we propose a novel nonparametric Bayesian approach to modeling RNA-seq data. We want to model the process of generating the expression level for each gene corresponding to every sample in the experiment as follows:

1. Assume that N_i cells were sampled from the tissue of interest, where $N_i \sim \text{Poisson}(\lambda_i)$ for the ith sample.

2. The contribution of a single cell to the read counts mapped unambiguously to a particular gene is an arbitrary discrete random variable $X^{(j)} \sim (\alpha_1, \alpha_2, ...)^{(j)} \in [0,1)^\infty$ where $\mathbb{P}(X^{(j)} = k) = \alpha_k$, and $\sum_{l=1}^{\infty} \alpha_l = 1$ which we want to estimate.

3. Then the observed count data for each gene j and sample i is $Y_{ij} = \sum_{l=1}^{N_i} X_l^{(j)}$

Usually in the literature Y_{ij}'s are modeled as either Poisson or Negative Binomial distribution [3, 21, 22, 36]. In our approach we make the least necessary assumption, which is that the observed counts are from an infinitely divisible non-negative discrete measure, in the under-sampled regime, where we do not expect to observe all possible outcomes of the discrete random variables. Infinitely divisible measures are characterized as those that can be expressed as a sum of an arbitrary number of independent and identically distributed random variables, which is a reasonable description of the data generating process.

Clearly the main inference problem in such a scenario, would be to ensure consistency of any derived estimator with reasonably fast convergence rates, i.e. low sample complexity (the number of samples necessary to achieve an ε error with high probability). Further, as the number of dimensions are in the order $\sim 10^5$ to 10^6, any proposed inference algorithm needs to be scalable.

In the under-sampled regime, it would be unreasonable to expect strong (i.e. point-wise) consistency of the inferred marginal distributions, however here we show that provable convergence can be achieved in the topologies defined by Kullback-Liebler and Renyi type divergences. This means that the inferred marginal distributions provide the same information as the true marginal distributions, with arbitrary accuracy. In these topologies the sample complexity [1] can be shown to be $O(k/\log k)$ for Kullback-Liebler divergence and $O(k^{1-1/\alpha})$ for Renyi divergence of order $1 < \alpha \in \mathbb{N}$, where k is the size of the support. Therefore for certain topologies, even sub-linear observation of the support of the random variable is enough for consistent inference.

Our work on modeling the marginal distributions starts from a characterization due to Feller [6], that every infinitely divisible non-negative discrete random variable belongs to the family of discrete Compound Poisson Measures. One could see this family of measures as the discrete analogue of the stable family of measures in the continuous case, which includes for example the Normal distribution. Since we only partially observe each high dimensional sample, we cannot assume independence of the marginal observations, however we can assume their exchangeability. By applying the de-Finetti theorem, we can express the marginal distribution from such a sample, as a mixture of discrete Compound Poisson Measures with a certain mixing measure, which we model here by the Pitman-Yor Mixture process.

The Pitman-Yor process provides an attractive family of priors in this setting, since not only are the induced posterior distributions of functionals have analytically tractable moments,

but also that the induced posterior distributions have power-law tails, a common feature in many real world data sets. However, for fixed hyper-parameters it imposes a narrow prior, leading to bias and overly narrow confidence intervals especially in the under-sampled regime.

A continuous mixing over the hyper-parameters flattens the prior and helps ensure positive probability over the true discrete measure, thereby providing a consistent posterior distribution. Using similar arguments, the parameter of the Poisson distribution is modeled by a continuous mixing of Gamma priors. We show that such a model achieves the optimal Bayesian rate of convergence in the topologies under consideration, in the under-sampled regime.

4.2 Discrete Compound Poisson measures

Let $Z_1, Z_2, ..., Z_n$ be non-negative integer valued random variables then we are interested in the law of their sum i.e. the random variable

$$S_n = \sum_{i=1}^{n} Z_i$$

It is convenient to write each $Z_i = B_i X_i$ of two independent random variables, where $B_i \sim$ Bernoulli(p_i) and X_i is an arbitrary random variable which takes values in $\mathbb{N}$. This can be done uniquely and without loss of generality, by taking $p_i = \Pr(Z_i \neq 0)$ and X_i having distribution $Q_i(k) = {}^{\Pr(Z_i=k)}/_{p_i}$ for $k \geq 1$, so that Q_i is simply the conditional distribution of Z_i given that $\{Z_i \geq 1\}$.

The simplest example, which is in a sense, the very definition of the compound Poisson distribution, is when the $\{Z_i\}$ are i.i.d. with each $Z_i = B_i X_i$ being the product of a Bernoulli$({}^{\lambda}/_{n})$ random variable B_i and X_i with an arbitrary distribution Q on $\mathbb{N}$. Then,

$$S_n = \sum_{i=1}^{n} B_i X_i \stackrel{\mathrm{d}}{=} \sum_{i=1}^{N(n)} X_i$$

where $N(n) = \sum_{i=1}^{n} B_i$ has a Binomial$(n, {}^{\lambda}/_{n})$ distribution, and $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. Since $N(n)$ converges to Po(λ) as $n \to \infty$, it is easily seen that P_{S_n} will converge to the distribution of,

$$\sum_{i=1}^{N} X_i$$

where $N \sim$ Po(λ) is independent of the $\{X_i\}$. This expression is precisely the definition of the Discrete Compound Poisson distribution [6] with parameters λ and Q, denoted by DCP(λ, Q). The probability generating functions for random sums of random variables is given by their

composition, hence in this case we have

$$\begin{aligned} G_{S_n}(z) &= G_N(G_X(z)) \\ &= \exp\left(\lambda \sum_{k=0}^{\infty} Q(k)(z^k - 1)\right), \quad (|\, z \,| \leq 1) \end{aligned}$$

from which we can write

$$\Pr(S_n = k) = \frac{G_{S_n}^{(k)}(0)}{k!} =: C_{\lambda,Q}(k)$$

where $G_{S_n}^{(k)}(0)$ is the kth derivative of the probability generating function at $z = 0$.

Example 4.1. Negative Binomial distribution is a DCP. Let $\mathbb{P}(X_l = k) = \frac{-1}{\log(1-p)} \frac{p^k}{k}, k = 0, 1, 2, \ldots$ i.e. have the logarithmic distribution with $p \in (0,1)$, and $N \sim \text{Poisson}(\lambda)$ with $\lambda = -r\log(1-p)$. Then the random sum

$$S_N = \sum_{i=1}^{N} X_i$$

is Negative Binomial(r, p) distributed. This result can be easily proved. Since moment generating function of Poisson distribution is $G_N(z) = \exp(\lambda(z-1))$ and for logarithmic distribution it is $G_X(z) = \frac{\log(1-pz)}{\log(1-z)}, |\, z \,| < \frac{1}{p}$. Hence we obtain the distribution of the sum as

$$\begin{aligned} G_Y(z) &= G_N(G_X(z)) \\ &= \exp(-r(\log(1-pz) - \log(1-p))) \\ &= \left(\frac{1-p}{1-pz}\right)^r, |\, z \,| < \frac{1}{p} \end{aligned}$$

which is the probability generating function for Negative Binomial(r, p) distributed random variable.

4.2.1 Convergence of General Sums

Now in the general case even if the summands $\{Z_i\}_{i=1}^n$ are not i.i.d., it is often the case that the distribution P_{S_n} of S_n can be accurately approximated by a discrete Compound Poisson distribution. Intuitively, the minimal requirements for such an approximation is if none of the $\{Z_i\}$ dominate the sum, i.e. the parameters $p_i = \Pr(Z_i \neq 0)$ are all appropriately small and if the $\{Z_i\}$ are only weakly dependent.

We will measure the closeness between P_{S_N} and an appropriately chosen compound Poisson measure $DCP(\lambda, \bar{Q})$ in terms of the Kullback-Liebler divergence $KL(P_{S_N} \mid DCP(\lambda, \bar{Q}))$,

defined as usual by

$$KL(P \mid Q) = \sum_{s \in S} P(s) \log \left[\frac{P(s)}{Q(s)} \right]$$

for any pair of probability distributions P and Q, on the same countably infinite set S. Although not a proper metric, relative entropy is an important measure of closeness between probability distribution and it can be used to obtain total variation bounds via Pinsker's inequality

$$\| P - Q \|_{TV}^2 \leq 2KL(P \mid Q)$$

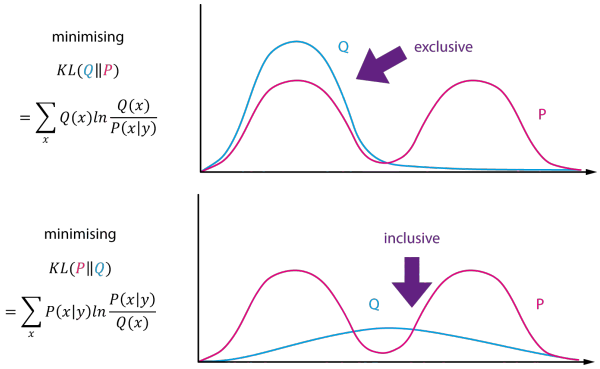

Figure 4.1: Behavior of KL Divergence

Mutual Information Bound Now we see how to bound the $KL(P_{S_N} \mid DCP(\lambda, \bar{Q}))$. Suppose S_N is the sum $\sum_{i=1}^N X_i$ of N possibly dependent random variables X_i with values in $\{0,1,2,...\}$. Then if $Z_1,...,Z_N$ are independent compound Poisson random variables with each $Z_i \sim \text{CPD}(p_i, Q_i)$ where $p_i = P[X_i \neq 0]$ by the basic infinite divisibility property of the compound Poisson law, the distribution of $T_N = \sum_{i=1}^N Z_i$ is $CPD(\lambda, \bar{Q})$ where $\lambda = \sum_{i=1}^N p_i$, $Q_i(k) = {}^{P[X_i=k]}/_{p_i}$, $k \geq 1$ and

$$\bar{Q} = \sum_{i=1}^N \frac{p_i}{\lambda} Q_i$$

is a mixture distribution. By the data processing inequality for relative entropy we have

$$\begin{aligned} KL(P_{S_N} \mid DCP(\lambda, \bar{Q})) =& KL(P_{S_N} \mid P_{T_N}) \\ \leq& KL(P_{X_1,...,X_N} \mid P_{Z_1,...,Z_n}) \end{aligned}$$

where $P_{X_1,\ldots,X_N}$ denotes the joint distribution of the X_i's and similarly for Z_i's. Applying the chain rule for relative entropy gives,

$$KL(P_{S_N} \mid DCP(\lambda, \bar{Q})) \leq \sum_{i=1}^{N} KL(P_{X_i} \mid P_{Z_i}) + \sum_{i=1}^{N} H(X_i) - H(X_1, \ldots, X_N)$$

where $H(X) = -\sum_{\mathbb{Z}} P(i) \log P(i)$. Finally we need to bound the term

$$KL(P_{X_i} \mid P_{Z_i}) = \sum_{j=0}^{\infty} P_{X_i}(j) \log \frac{P_{X_i}(j)}{P_{Z_i}(j)}$$

Let Q_i^{*k} be the k-fold convolution of Q_i, i.e. the law of the sum of i.i.d. random variables with common distribution Q_i. Then

$$P_{X_i}(j) = \sum_{k=1}^{\infty} e^{-p_i} \frac{p_i^k}{k!} Q_i^{*k}(j) \geq e^{-p_i} p_i Q_i(j)$$

which implies that

$$P_{X_i}(j) \log \frac{P_{X_i}(j)}{P_{Z_i}(j)} \leq [p_i Q_i(j)] \log \frac{p_i Q_i(j)}{e^{-p_i} p_i Q_i(j)} = Q_i(j) p_i^2$$

and hence by summing over all j, we get

$$KL(P_{X_i} \mid P_{Z_i}) \leq p_i^2$$

and therefore we can say

$$\mathrm{KL}(P_{S_N} \mid \mathrm{DCP}(\lambda, \bar{Q})) \leq \sum_{i=1}^{N} p_i^2 + \left[\sum_{i=1}^{N} H(X_i) - H(X_1, \ldots, X_N) \right] \tag{4.1}$$

Note that second term is the Mutual Information among the random variables $\{X_i\}_{i=1}^N$.

Log-Sobolev Bound The bound obtained above (4.1) in terms of its mutual information are enough to prove convergence, but the rate of convergence is sub-optimal and not easily estimable based on partially observed data. In order to build bounds with faster convergence rates, we need to look at a notion of information bounds which are more locally defined, which allow us to exploit parts of the distribution which are well observed.

Consider the notion of size-biased sampling [6], which is a type of nonrandom sampling in which the probability of sampling an object is proportional to the size of the object. Then

it can be shown for any distribution P on $\mathbb{Z}_+$ with mean λ, the corresponding size-biased distribution $P^{\#}$ is given by

$$P^{\#}(y) = \frac{(y+1)P(y+1)}{\lambda},\ y \geq 0$$

Recall that a distribution P on $\mathbb{Z}_+$ satisfies the recursion,

$$(k+1)P(k+1) := \lambda P(k)$$

iff $P = \text{Po}(\lambda)$, it is immediate that $P = \text{Po}(\lambda)$ if and only if $P = P^{\#}$. Building in this way, we can say that $P_{S_N} \sim \text{DCP}(\lambda, Q)$ if and only if $N \sim Po(\lambda)$, i.e. $P_{S_N} \sim \text{DCP}(\lambda, Q)$ if and only if $P = P^{\#}$.

Consider the logarithmic Sobolev inequality [76] for a Poisson distribution, which is given by

$$KL(P \mid Po(\lambda)) \leq \lambda \mathbb{E}\left[\left(\frac{(X+1)P(X+1)}{\lambda P(X)} - 1\right)^2\right]$$

for any distribution P on $\mathbb{Z}_+$ and any $\lambda > 0$. Also note that using the characteristic function for $DCP(\lambda, \bar{Q})$, and same notion as before, there exists an alternate representation [3, 128] given by

$$\sum_{j=1}^{\infty} jZ_j$$

where the Z_j are independent Poisson random variables with each $Z_j \sim Po(\lambda \bar{Q}(j))$. Then applying the log-Sobolev inequality on the distribution of the above sum yields

$$KL(P_{S_N} \mid DCP(\lambda, \bar{Q})) \leq J_{\lambda,\bar{Q}}(X)$$

where

$$J_{\lambda,Q}(X) = \lambda \sum_{j=1}^{\infty} \bar{Q}(j)\mathbb{E}_X\left[\left(\frac{P_{S_N}(X+j)}{P_{S_N}(X)}\frac{C_{\lambda,\bar{Q}}(X)}{C_{\lambda,\bar{Q}}(X+j)} - 1\right)^2\right]$$

is known as the DCP-Fisher information or the Log-Sobolev Bound [6]. Here $C_{\lambda,\bar{Q}}$ denotes the probability measure from the $DCP(\lambda, \bar{Q})$ distribution. For any random variable X, then it is clear that

$$J_{\lambda,Q}(X) = 0 \ \text{ iff } \ X \sim DCP(\lambda, \bar{Q})$$

.

4.3 Bayesian Nonparametric Inference

In the previous section we studied properties of the DCP measure and constructed a loss function given by the Log-Sobolev bound which we want to use to infer the unknown discrete distribution Q from which the RNA-seq data was generated according to the data generating model discussed in the introduction.

Consider samples $\mathbf{x} \coloneqq \{x_j\}_{j=1}^{N}$ drawn i.i.d. from an unknown discrete distribution $\pi \coloneqq \{\pi_i\}_{i=1}^{\infty}$ where

$$p(x_j = i) = \pi_i \text{ and } \sum_i \pi_i = 1$$

then the aim of our inference algorithm would be to estimate the discrete distribution Q and the parameter $\lambda > 0$ which satisfies

$$\hat{\lambda}, \hat{Q} = \arg\min_{\lambda, Q} J_{\lambda, Q}(\pi)$$

where

$$J_{\lambda, Q}(\pi) = \lambda \sum_{j=1}^{\infty} Q(j) \mathbb{E}_X \left[\left(\frac{\pi_{k+j}}{\pi_k} \frac{C_{\lambda, Q}(k)}{C_{\lambda, Q}(k+j)} - 1 \right)^2 \right]$$

is the Log-Sobolev bound of the random variable X.

Here we are interested in the under sampled regime, as with finite samples most of the number line remains unobserved. A naive approach would be to use empirical estimates of π and then optimizing the Log-Sobolev bound to get the required quantities. However in this regime, such an approach results in severely biased estimators, which leads us to the application of non-parametric Bayesian techniques.

4.3.1 Plugin Estimator

The simplest approach would be to estimate the distribution π and then plugin to the Log-Sobolev bound and optimize for λ and Q, using a gradient descent or a MCMC based optimizer. The empirical distribution $\hat{\pi} = (\hat{\pi}_1, ..., \hat{\pi}_{\mathscr{A}})$ is computed by normalizing the observed counts $\mathbf{n} \coloneqq (n_1, ..., n_{\mathscr{A}})$ of each symbol

$$\hat{\pi}_k = {}^{n_k}/_{N},\ n_k = \sum_{i=1}^{N} 1_{\{x_i = k\}}$$

for each $k \in \mathscr{X}$. Plugging this estimate for π, we obtain the so-called "plugin" estimator

$$\hat{J}_{\lambda, Q}^{\text{plugin}} = J_{\lambda, Q}(\hat{\pi})$$

which is also the maximum likelihood estimator under the categorical (or multinomial) likelihood. Then

$$\hat{\lambda}^{\text{plugin}}, \hat{Q}^{\text{plugin}} = \arg\min_{\lambda,Q} \hat{J}^{\text{plugin}}_{\lambda,Q}(\hat{\pi})$$

can be calculated.

However as appealing its simplicity may be, the performance of such an estimator is highly unstable especially due to the under-sampled regime. As there are many unobserved locations, an empirical estimate of π, has many gaps and therefore most terms in the Log-Sobolev bound cannot be calculated. Further if the mutual information bound (4.1) is used, estimates behave like entropy estimators, in which case it is well known that the empirical estimator is substantially negatively biased [63, 89, 113, 115]. The biased estimates of the information bounds therefore lead to unstable optimization of the underlying measure Q that we are interested in.

Even though some results exist, which attempt to remove bias for the mutual information estimator, when the support of the discrete measure is finite and known, for example using series expansions of the entropy functional in (Panzeri and Treves, 1996; Grassberger, 2008), or by minimizing an upper bound over a class of linear estimators (Paninski, 2003), and a James-Stein estimator (Hausser and Strimmer, 2009). The performance of these results are hampered by the under-sampled regime.

4.3.2 Bayesian Estimation of Log-Sobolev Bound

The Bayesian approach to estimation involves formulating a prior over distributions π, and constructing the posterior distribution of $J_{\lambda,Q}$ using the Bayes theorem. Then Bayes' least squares (BLS) estimators take the form

$$\hat{J_{\lambda,Q}}(\mathbf{x}) = \mathbb{E}\left[J_{\lambda,Q} \mid \mathbf{x}\right] = \int J_{\lambda,Q}(\pi) p(\pi \mid \mathbf{x}) d\pi$$

where $p(\pi \mid \mathbf{x})$ is the posterior over π under some prior $p(\pi)$ and discrete likelihood $p(\mathbf{x} \mid \pi)$. To the extent that $p(\pi)$ expresses our true uncertainty over the unknown distribution that generated the data, this estimate is optimal (in a least squares sense) and the corresponding credible intervals capture our uncertainty about $J_{\lambda,Q}$ given the data.

4.3.2.1 Pitman-Yor Process Priors

A very general class of priors over unknown or countably infinite discrete distributions have been defined in terms of stochastic processes called the Dirichlet Process (DP) and Pitman-Yor process (PYP) whose samples are countably infinite discrete distributions (Ferguson, 1973;

Pitman and Yor, 1997). Such a sample may be written as

$$\sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}$$

where $\pi := \{\pi_i\}_{i=1}^{\infty}$ denotes a countably infinite set of 'weights' on a set of atoms $\{\phi_i\}$ drawn from some base probability measure, where δ_{ϕ_i} is a delta function on the atom ϕ_i [1]. The DP and PYP defines a prior distribution on the infinite-dimensional simplex representing the space of all discrete probability measures.

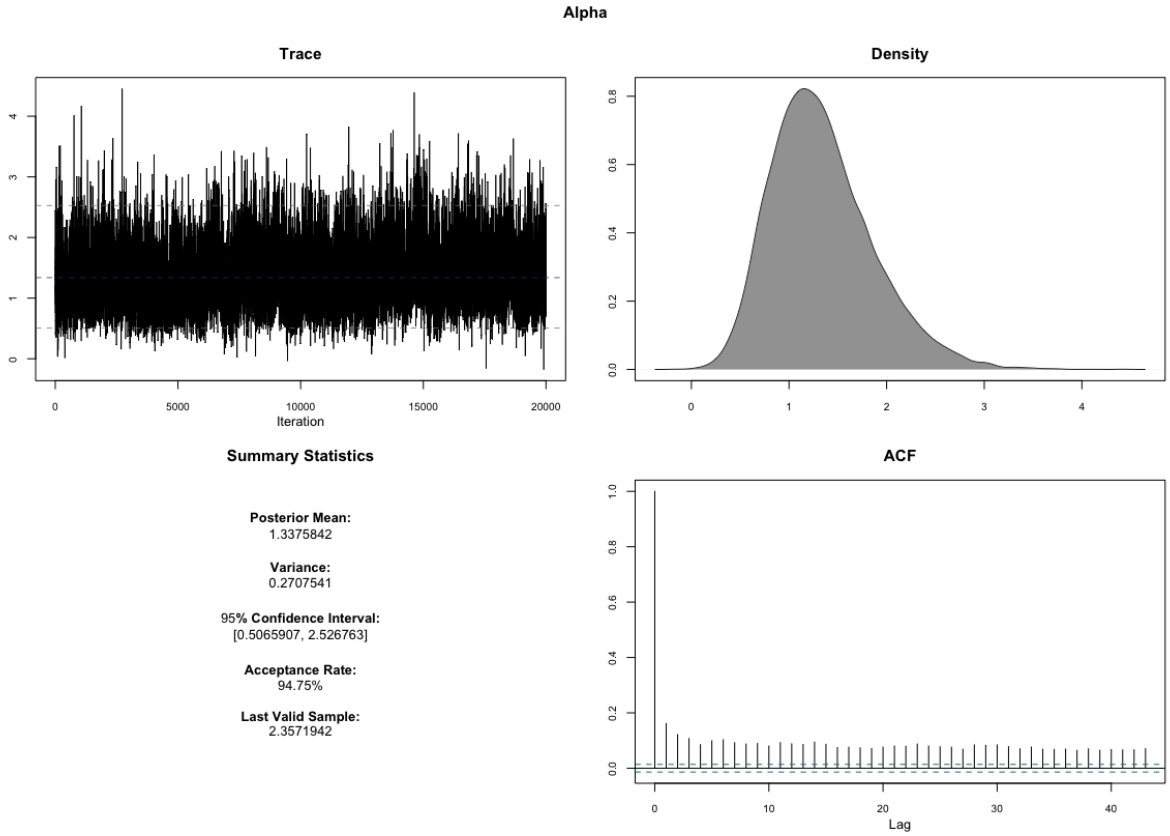

Figure 4.2: Pitman-Yor Process

For distributions with known finite alphabet size K, the Dirichlet distribution provides an obvious choice of prior due to its conjugacy with the categorical distribution. It takes the form

$$p_{\mathrm{Dir}}(\pi) \propto \prod_{i=1}^{K} \pi_i^{a-1}$$

[1] Here, we will assume the base measure is non-atomic, so that the atoms ϕ_i's are distinct with probability one.

for π on the $\mathscr{A}$-dimensional simplex ($\pi_i \geq 1, \sum \pi_i = 1$), where $a > 0$ is a "concentration" parameter.

The Dirichlet Process distribution over π results from a limit of the Dirichlet distribution where alphabet size grows and concentration parameter shrinks: $K \to \infty$ and $a \to 0$ s.t. $aK \to \alpha$. The PYP distribution over π generalizes the DP to allow power-law tails, and includes DP as a special case (Kingman, 1975; Pitman and Yor, 1997). For $\mathrm{PY}(d, \alpha)$ with discount parameter $d \in [0, 1)$ and concentration parameter $\alpha > -d$, the tails approximately follow a power-law:

$$\pi_i \propto i^{-1/d}$$

When $d = 0$, this reduces to the Dirichlet process, $\mathrm{DP}(\alpha)$ which on the other hand have exponentially small tails.

To gain intuition for the DP and PYP, it is useful to consider typical samples π with weights $\{\pi_i\}$ sorted in decreasing order of probability, so that $\pi_{(1)} > \pi_{(2)} > \cdots$. The concentration parameter α controls how much of the probability mass is concentrated in the first few samples, that is, in the head instead of the tail of the sorted distribution. For small α the first few weights carry most of the probability mass whereas, for large α, the probability mass is more spread out so that π is more uniform. As noted above the discount parameter d controls the shape of the tail. Larger d gives heavier power-law tails, while $d = 0$ yields exponential tails.

We can draw samples $\pi \sim \mathrm{PY}(d, \alpha)$ using an infinite sequence of independent Beta random variables in a process known as "stick-breaking" (Ishwaran and James, 2001)

$$\beta_i \sim \mathrm{Beta}(1 - d, \alpha + id), \quad \tilde{\pi}_i = \beta_i \prod_{j=1}^{i-1} (1 - \beta_j) \tag{4.2}$$

where $\tilde{\pi}_i$ is known as the i'th size biased permutation from π (Pitman, 96). The $\tilde{\pi}_i$ sampled in this manner are not strictly decreasing, but decreases on average such that $\sum_{i=1}^{\infty} \tilde{\pi}_i = 1$ with probability 1 (Pitman and Yor, 1997).

Posterior Distribution A useful property of PYP priors (for multinomial observations) is that the posterior $p(\pi \mid \mathbf{x}, d, \alpha)$ takes the form of a mixture of a Dirichlet distribution (over the observed symbols) and a Pitman-Yor process (over the unobserved symbols) (Ishwaran and James, 2003). Let K be the number of unique symbols observed in N samples, i.e., $K = \sum_i 1_{\{n_i > 0\}}$. Further, let $\alpha_i = n_i - d$, $N = \sum_i n_i$, and $A = \sum_i \alpha_i = \sum_i n_i - Kd = N - Kd$. Now, following Ishwaran and colleagues (Ishwaran and Zarepour, 2002), we write the posterior distribution of π as an infinite random vector $\pi \mid \mathbf{x}, d, \alpha = (p_1, p_2, ..., p_K, p_* \pi')$, where

$$(p_1, p_2, ..., p_K, p_* \pi') \sim \mathrm{Dir}(n_1 - d, ..., n_K - d, \alpha + Kd) \tag{4.3}$$

$$\pi' \coloneqq (\pi_1, \pi_2, \pi_3, ...) \sim \mathrm{PY}(d, \alpha + Kd)$$

4.3.2.2 Expectations over Pitman-Yor Process Priors and Posteriors

A key virtue of Pitman-Yor process priors is invariance under size-biased sampling, a property which we exploited earlier to derive the information bound $J_{\lambda,Q}$. Here this property allows us to convert expectations over π on the infinite-dimensional simplex (which are required for computing the mean of $J_{\lambda,Q}$ from the given data) into one or two-dimensional integrals with respect to the distribution of the first two size-biased samples (Perman et al., 1992; Pitman, 1996).

Proposition 4.1. *(Expectations with first two sized-biased samples [Pitman and Yor, 1997]) For $\pi \sim PY(d, \alpha)$*

$$\mathbb{E}_{(\pi|d,\alpha)}\left[\sum_{i=1}^{\infty} f(\pi_i)\right] = \mathbb{E}_{(\tilde{\pi}_1|d,\alpha)}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right]$$

$$\mathbb{E}_{(\pi|d,\alpha)}\left[\sum_{i,j\neq i} g(\pi_i, \pi_j)\right] = \mathbb{E}_{(\tilde{\pi}_1,\tilde{\pi}_2|d,\alpha)}\left[\frac{g(\tilde{\pi}_1, \tilde{\pi}_2)}{\tilde{\pi}_1\tilde{\pi}_2}(1-\tilde{\pi}_1)\right]$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are the first two sized biased samples from π.

The direct consequence of this proposition is that the integrals over the infinite-dimensional simplex becomes tractable and, as a result, we obtain closed-form solutions for expected value of $J_{\lambda,Q}(\pi)$ under both prior and posterior distributions

$$J_{\lambda,Q}(\pi) = \lambda \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \pi_k Q(j)\left(\frac{\pi_{k+j}}{\pi_k}\frac{C_{\lambda,Q}(k)}{C_{\lambda,Q}(k+j)} - 1\right)^2$$

$$\mathbb{E}\left[J_{\lambda,Q} \mid d, \alpha\right] = \psi_0(\alpha+1) - \psi_0(1-d) \tag{4.4}$$

$$\mathbb{E}\left[J_{\lambda,Q} \mid \mathbf{x}, d, \alpha\right] = \psi_0(\alpha+N+1) - \frac{\alpha+Kd}{\alpha+N}\psi_0(1-d) - \frac{1}{\alpha+N}\left[\sum_{i=1}^{K}(n_i-d)\psi_0(n_i-d+1)\right] \tag{4.5}$$

4.3.2.3 Pitman-Yor Mixture Process priors

The prior and posterior expectations computed earlier provide a class of estimators for the information bound $J_{\lambda,Q}$ for distributions with possibly countably infinite support. However

Dirichlet and Pitman-Yor Process priors for a fixed (d,α), are highly informative in the under-sampled regime and induces a rather narrow prior distribution over $J_{\lambda,Q}$. This inevitably leads to undesirably narrow posterior credible intervals [1, 10, 17, 21], reflecting the narrow prior uncertainty rather than strong evidence from the data, thereby giving incorrect answers with high confidence!

We address this problem by introducing a mixture prior $p(d,\alpha)$ on $\mathrm{PY}(d,\alpha)$ under which the implied prior on $J_{\lambda,Q}$ is flat. One way of constructing a flat mixture prior is by setting $p(d,\alpha)$ proportional to the derivative of the expected entropy [10]. We use the entropy to define this prior as it regularizes the class of measures that we learn, to have finite entropy. The two parameters (d,α) are explicitly controlled by re-parameterizing Pitman-Yor Process prior as follows

$$h = \psi_0(\alpha+1) - \psi_0(1-d),\ \gamma = \frac{\psi_0(1) - \psi_0(1-d)}{\psi_0(\alpha+1) - \psi_0(1-d)}$$

where $h > 0$ is equal to the expected prior entropy, and $\gamma \in [0,\infty)$ captures prior beliefs about the tail behavior.

For $\gamma = 0$, we have the DP (i.e. $d = 0$, giving π with exponential tails), while for $\gamma = 1$ we have a $\mathrm{PY}(d,0)$ process (i.e., $\alpha = 0$, yielding π with power-law tails). In the limit where $\alpha \to -1$ and $d \to 1$, $\gamma \to \infty$. Where required, the inverse transformation to standard Pitman-Yor Process parameters is given by:

$$\alpha = \psi_0^{-1}(h(1-\gamma) + \psi_0(1)) - 1,\ d = 1 - \psi_0^{-1}(\psi_0(1) - h\gamma)$$

where $\psi_0^{-1}(\cdot)$ denotes the inverse digamma function.

We can then construct an approximately flat improper distribution on $[0,\infty]$ by setting $p(h,\gamma) = q(\gamma)$ for all h, where q is any density on $[0,\infty)$. We call this the Pitman-Yor process mixture (PYM) prior. The induced prior on $J_{\lambda,Q}$ is thus

$$p(J_{\lambda,Q}) = \int\int p(J_{\lambda,Q} \mid \pi) p(\pi \mid \gamma, h) p(\gamma, h) d\gamma dh$$

where $p(\pi \mid \gamma, h)$ denotes a Pitman-Yor Process prior on π with parameters γ, h.

In this work we compare only two choices for $q(\gamma)$, the gamma family which has exponential tails and the stable family which has power law tails thereby providing different prior beliefs one might have regarding the data set at hand. PYM mixture priors resulting from different choices of $q(\gamma)$ are all approximately flat on $J_{\lambda,Q}$, but each favors distributions with different tail behavior; the ability to select $q(\gamma)$ greatly enhances the flexibility of PYM, allowing the practitioner to adapt it to there own data.

The Pitman-Yor Mixture Estimator Now that we have determined a prior on the infinite simplex, we turn to the problem of inference given observations x. The Bayes least squares entropy estimator under the mixture prior $p(d,\alpha)$, the Pitman-Yor Mixture (PYM) estimator, takes the form

$$\hat{J}^{\mathrm{PYM}}_{\lambda,Q} = \mathbb{E}\left[J_{\lambda,Q} \mid \mathbf{x}\right] = \int \mathbb{E}\left[J_{\lambda,Q} \mid \mathbf{x}, d, \alpha\right] \frac{p(\mathbf{x} \mid d,\alpha)p(d,\alpha)}{p(\mathbf{x})} \mathrm{d}(d,\alpha) \tag{4.6}$$

where $\mathbb{E}\left[J_{\lambda,Q} \mid \mathbf{x}, d, \alpha\right]$ is the expected posterior entropy for a fixed (d,α). The quantity $p(\mathbf{x} \mid d,\alpha)$ is the evidence, given by

$$p(\mathbf{x} \mid d,\alpha) = \frac{\left(\prod_{l=1}^{K-1}(\alpha + ld)\right)\left(\prod_{i=1}^{K}\Gamma(n_i - d)\right)\Gamma(1+\alpha)}{\Gamma(1-d)^K\Gamma(\alpha+N)}$$

4.3.3 Optimizing the Log-Sobolev Bound

Because of the improper prior $p(d,\alpha)$, and because by [4.6] it must be integrated over all $\alpha > 0$, it is not obvious that the PYM estimate $\hat{J}^{\mathrm{PYM}}_{\lambda,Q}$ is computationally tractable. In principle the PYM integral over α is supported on the range $[0,\infty)$. In practice, however, the posterior concentrates on a relatively small region of parameter space. It is generally unnecessary to consider the full integral over a semi-infinite domain. Instead, we select a subregion of $[0,1] \times [0,\infty)$ which supports the posterior up to ε probability mass.

The posterior is usually unimodal in each variable α and d separately, however if there are multiple modes, they must lie on a strictly decreasing line of d as a function of α and, in practice, we find the posterior to be unimodal. We compute the hessian at the MAP parameter value, $(d_{\mathrm{MAP}}, \alpha_{\mathrm{MAP}})$. Using the inverse hessian as the covariance of a Gaussian approximation to the posterior, we select a grid spanning ± 6 std. We use numerical integration (Gauss-Legendre quadrature) on this region to compute the integral. When the hessian is rank-deficient (which may occur, for instance, when the $\alpha_{\mathrm{MAP}} = 0$ or $d_{\mathrm{MAP}} = 0$), we use Gauss-Legendre quadrature to perform the integral in d over $[0,1)$, but employ a Fourier-Chebyshev numerical quadrature routine to integrate α over $[0,\infty)$ (Boyd, 1987).

Thereby allowing us to estimate $\hat{J}^{\mathrm{PYM}}_{\lambda,Q}$ for each value of λ and Q. Here again we choose a Pitman-Yor Process model on the distribution Q

$$Q \coloneqq (Q(1), Q(2), Q(2), \ldots) \sim \mathrm{PY}(a,b)$$

and then the aim of the learning algorithm is to optimize for the value of a,b which minimize the expected value of $\hat{J}^{\mathrm{PYM}}_{\lambda,Q}$. Since by stick breaking process, as described by [4.2] we have a straightforward algorithm for sampling distributions $\pi \sim \mathrm{PY}(d,\alpha)$ and $Q \sim \mathrm{PY}(a,b)$. With

enough stick-breaking samples, it is always possible to approximate π and Q to arbitrary accuracy which then allows us to finally get the required

$$\hat{\lambda}^{\mathrm{PYM}}, \hat{Q}^{\mathrm{PYM}} = \arg\min_{\lambda, Q} \hat{J}^{\mathrm{PYM}}_{\lambda, Q}(\pi)$$

via a simple gradient descent algorithm on the parameters (a, b).

Appendix A

Polish Spaces

In this section we expound these properties so as to get a deeper understanding, however they are not critical to the rest of the work. We start with an example. Consider the case of real numbers, and identify each real number $x \in \mathbb{R}$ as a collection of intervals, say $\{[p_i, q_i]\}_{i=1}^{\infty}$ where $p_i, q_i \in \mathbb{Q}$ are rational numbers, such that $x = \limsup p_i$ and $x = \liminf q_i$. Then there intersection defines the real number $\{x\} = \cap_{i=1}^{\infty} [p_i, q_i]$. In such a representation, smaller an interval more information one has about the number, one is trying to approximate. So if $x := [p_i, q_i] \supset y := \left[p_j, q_j\right]$, then the interval y carries more information than the interval x, and we represent it by writing $x \leq y$.

The aim is to generalize this idea of a system of approximating intervals to arbitrary topological spaces. A topological space is an ordered pair (X, τ_X), where X is some set and τ_X is a collection of subsets of X, satisfying the following axioms

1. The empty set $\emptyset$ and X are closed (i.e. contains all its limit points)

2. The intersection of any (finite or infinite) number of members of τ_X, is closed

3. The union of finitely many elements of τ_X, is also closed

The elements of τ_X are called closed sets and the collection τ_X is a called a topology on X. Now consider, partially ordered sets $(P_{\tau_X}, \leq)$ or posets, as a sequence of compact (i.e. closed and bounded) subsets, which are elements of the topology τ_X, equipped with a transitive, reflexive, and antisymmetric partial order $\leq$, defined by the reverse set inclusion, $\supseteq$. Here antisymmetry means that if $x \leq y$ and $y \leq x$ then $x = y$. This choice of partial order implies smaller sets provide more information than larger sets. We can then use these posets as a model for the topological space (X, τ_X), provided a few more properties are satisfied, which we study now.

Firstly, we need to ensure consistency, i.e. the sequence of compact subsets should converge to an unique object. One way to interpret consistency would be using the idea of a "knowledge closed" topology. A set $C \subseteq P_{\tau_X}$ is knowledge closed, or closed in the Scott topology [Scott, 1970] if

1. C is a lower set: if $p \in C$ and $q \leq p$ then $q \in C$ i.e. if we know that $x \in p \subseteq q$ then we know that $x \in q$.

2. If $\mathscr{D} \subseteq C$ is directed, then $\vee\mathscr{D} \in C$ i.e. if we know that $x \in p$ for each $p \in \mathscr{D}$, then we know that $x \in \vee\mathscr{D}$.

To ensure Scott topology on the poset P_{τ_X}, it is enough to assume that P_{τ_X} is a directed complete poset (dcpo), that is if $\mathscr{D} \subseteq P_{\tau_X}$ is a directed subset and bounded above, then $\mathscr{D}$ has a supremum, denoted as $\vee\mathscr{D}$. Thus the notion of dcpo ensures that increasingly smaller subsets indeed approximates an unique object.

Secondly, from the poset it should be possible to "observe" the possible approximations of the object under consideration. For example, for the dcpo $(P_{\tau_{\mathbb{R}}}, \leq)$, if r is an endpoint of the interval $x \in P_{\tau_{\mathbb{R}}}$, no magnification of the real line would make it possible to see whether r is actually in x or not. Similarly, for another interval $y \in P_{\tau_{\mathbb{R}}}$, if either the left or right endpoints of x and y are identical, it will not be possible under any magnification of the real line to see whether one of the intervals contains the other.

Lemma A.1. One can determine for $x, y \in P_{\tau_X}$, that $x \supset y$ (a relationship denoted by $x \ll y$ and read as x is way below y) if and only if whenever $y \leq \vee\mathscr{D}$, where $\mathscr{D} \subseteq P_{\tau_X}$ is a directed subset, then for some $r \in \mathscr{D}$, $x \leq r$.

Proof. Assuming $x \supset y$, let $y \supseteq \vee\mathscr{D}$, where $\mathscr{D} \subseteq P_{\tau_X}$ is a directed subset. Then $(X \backslash \mathrm{int}(x)) \cap y = \emptyset$ implies $(X \backslash \mathrm{int}(x)) \cap \vee\mathscr{D} = \emptyset$, so by the compactness of elements of $\mathscr{D}$, $(X \backslash \mathrm{int}(x)) \cap \vee F = \emptyset$ for some finite set $F \subseteq \mathscr{D}$. Since $\mathscr{D}$ is directed, there is an $r \in \mathscr{D}$ such that $r \subseteq \vee F$, whence $(X \backslash \mathrm{int}(x)) \cap r = \emptyset$; that is $\mathrm{int}(x) \supseteq r$, so in particular $x \leq r$.

Conversely, for $x, y \in P_{\tau_X}$ such that $x, y \leq \vee\mathscr{D}$, where $\mathscr{D} \subseteq P_{\tau_X}$ is some directed subset. Now since $x \leq r$ for some $r \in \mathscr{D}$, this means that by the compactness of elements of $\mathscr{D}$ there is some finite set $F \subseteq \mathscr{D}$ such that $r \subseteq \vee F$. Now since $y \supseteq \vee\mathscr{D}$ and $\vee\mathscr{D} \supseteq \vee F$, we have $x \leq r \leq y$. In other words $\mathrm{int}(x) \supseteq y$. This is exactly when one can determine for $x, y \in P_{\tau_X}$, that $x \supset y$. □

Thus we can observe the possible approximations only when for a sequence of approximations $x_1, x_2, \ldots \in P_{\tau_X}$ such that

$$x_1 \ll x_2 \ll \cdots$$

Thirdly, we need to ensure that sufficient information needed to compute any object is available in the objects way below it. This property is described by the concept of continuity. A continuous dcpo is a dcpo P_{τ_X} such that for every $x \in P_{\tau_X}$, $\Downarrow x$ is directed and $x = \vee(\Downarrow x)$ where

$$\Downarrow x = \{a \in P_{\tau_X} : y \ll a \text{ for some } y \in \{x\}\}$$

For example, clearly for $[p,q] \in P_{\tau_{\mathbb{R}}}$ we have

$$\vee(\Downarrow [p,q]) = \cap\{[r,s] : r < p \leq q < s\} = [p,q]$$

so $P_{\tau_{\mathbb{R}}}$ is a continuous dcpo.

Note that $\ll$ satisfies a transitivity condition and it is stronger than $\leq$, since if $x \ll y$ then $x \leq y$ and if $w \leq x \ll y \leq z$ then $w \ll z$. These properties imply that $\Downarrow(\Downarrow x) = \Downarrow x$ for every $x \in P_X$, where P_X is a continuous dcpo. This fact can be reinterpreted as an interpolation property, i.e. if $x \ll y$ then there exists a $z \in P_{\tau_X}$ such that $x \ll z \ll y$. In the case of $P_{\tau_{\mathbb{R}}}$ the interpolation property is obvious.

Finally, we need a countable dense subset, called a basis, whose elements could be used to recursively approximate the elements of X. Assuming P_{τ_X} has the interpolation property, then D is a basis for P_{τ_X} if and only if D is $\ll$-dense in P_{τ_X} in the sense that if $x \ll y$ then there exists a $d \in D$ such that $x \ll d \ll y$. A poset P_{τ_X} is ω-continuous provided it is a bounded continuous dcpo and has a countable basis. Notice that if D is a basis for dcpo P_{τ_X} then

$$x = \vee(D \cap \Downarrow x) \quad \text{for every } x \in P_{\tau_X}$$

Clearly the family of all intervals with rational endpoints form a countable $\ll$-dense subset of $P_{\tau_{\mathbb{R}}}$.

The property that $x = \vee(D \cap \Downarrow x)$ means that x is uniquely determined by $\mathscr{F}_x = D \cap \Downarrow x$ which is a filter in D. This means that for every object $x \in \max(P_X)$, where

$$\max(P_X) = \{\vee\mathscr{D} : \forall\mathscr{D}, \text{ the directed subsets of } P_{\tau_X}\}$$

one could define a "continuous learning process" if the poset is ω-continuous by encoding the incoming information using the elements from the countable set D which is $\ll$-dense in P_{τ_X}.

Definition A.1. A bounded complete computational model of a topological space (X, τ_X) is a ω-continuous dcpo $(P_{\tau_X}, \leq)$ together with a bijection $\phi : X \to \max(P_{\tau_X})$, such that

1. ϕ is a homeomorphism between (X, τ_X) and $\max(P_{\tau_X})$ considered with the subspace Scott topology inherited from P_{τ_X}.

(a) For every $x \in P_{\tau_X}$ the set $\phi^{-1}(\{y \in \max(P_{\tau_X}) : x \leq y\})$ is τ_X-closed.

Now we state a theorem without proof, since it is beyond the scope of the current work, but it provides the basis for restricting ourselves to only the study of Polish spaces when designing learning algorithms.

Theorem A.1. [Lawson, 1996] A topological space (X, τ_X) has a bounded complete computational model if and only if it is Polish, i.e. is a completely separable metrizable space.

Appendix B

Discrete Gibbs Principle

Let $Y_1, Y_2, ..., Y_n$ be a sequence of i.i.d. random variables with strictly positive law μ on the finite alphabet Σ. Let $X_k = f(Y_k)$ for some deterministic $f : \Sigma \to \mathbb{R}$. Given a set $A \subset \mathbb{R}$ and a constraint of the type $\hat{S}_n \in A$, what is the conditional law of Y_1 when n is large? In other words, what are the limit points, as $n \to \infty$ of the conditional probability vector

$$\mu_n^*(a_i) := \Pr_\mu \left(Y_1 = a_1 \mid \hat{S}_n \in A \right), \; i = 1, ..., \mid \Sigma \mid$$

Recall that $\hat{S}_n := \frac{1}{n} \sum_{j=1}^n X_j = \langle \mathbf{f}, L_n^{\mathbf{Y}} \rangle$, where $\mathbf{f} = \left(f(a_1), \cdots, f(a_{|\Sigma|}) \right)$, and note that under the conditioning $\hat{S}_n \in A$, Y_j are identically distributed, although not independent. Therefore, for every function $\phi : \Sigma \to \mathbb{R}$

$$\begin{aligned} \langle \phi, \mu_n^* \rangle &= \mathbb{E} \left[\phi(Y_1) \mid \hat{S}_n \in A \right] \\ &= \mathbb{E} \left[\frac{1}{n} \sum_{j=1}^n \phi(Y_j) \mid \hat{S}_n \in A \right] = \mathbb{E} \left[\langle \phi, L_n^{\mathbf{Y}} \rangle \mid \langle \mathbf{f}, L_n^{\mathbf{Y}} \rangle \in A \right] \end{aligned}$$

where $\phi = \left(\phi(a_1), \cdots, \phi(a_{|\Sigma|}) \right)$. Hence, with $\Gamma := \{ \nu : \langle \mathbf{f}, \nu \rangle \in A \}$

$$\mu_n^* = \mathbb{E} \left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in \Gamma \right]$$

Using this identity, the following characterization of the limit points of $\{\mu_n^*\}$ applies to any non-empty set Γ for which

$$I_\Gamma := \inf_{\nu \in \Gamma^\circ} H(\nu \mid \mu) = \inf_{\nu \in \bar{\Gamma}} H(\nu \mid \mu)$$

Theorem B.1. *Gibb's Principle [55]: Let*

$$\mathscr{M} := \left\{\nu \in \bar{\Gamma} : H(\nu \mid \mu) = I_\Gamma\right\}$$

1. All the limit points of $\{\mu_n\}^*$ belong to $\bar{\text{co}}(\mathscr{M})$, the closure of the convex hull of $\mathscr{M}$.

   (a) When Γ is a convex set of non-empty interior, the set $\mathscr{M}$ consists of a single point to which μ_n^* converge as $n \to \infty$.

Proof. Since $|\Sigma| < \infty$, $\bar{\Gamma}$ is a compact set and thus $\mathscr{M}$ is non-empty. For every $U \subset M_1(\Sigma)$

$$\begin{aligned}\mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in \Gamma\right] - \mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in U \cap \Gamma\right] \\ = \Pr\left(L_n^{\mathbf{Y}} \in U^c \mid L_n^{\mathbf{Y}} \in \Gamma\right) \\ \left\{\mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in U^c \cap \Gamma\right] - \mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in U \cap \Gamma\right]\right\}\end{aligned}$$

Since $\mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in U \cap \Gamma\right]$ belongs to $\text{co}(U)$, while $\mu_n^* = \mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in \Gamma\right]$, it follows that

$$\begin{aligned}&d_V(\mu_n^*, \text{co}(U)) \\ &\quad \le \Pr\left(L_n^{\mathbf{Y}} \in U^c \mid L_n^{\mathbf{Y}} \in \Gamma\right) d_V\left(\mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in U^c \cap \Gamma\right], \mathbb{E}\left[L_n^{\mathbf{Y}} \mid L_n^{\mathbf{Y}} \in U \cap \Gamma\right]\right) \\ &\quad \le \Pr\left(L_n^{\mathbf{Y}} \in U^c \mid L_n^{\mathbf{Y}} \in \Gamma\right)\end{aligned}$$

where the last inequality is due to the bound $d_V(\cdot,\cdot) \le 1$. With $\mathscr{M}^\delta := \{\nu : d_V(\nu, \mathscr{M}) < \delta\}$, it is proved shortly that for every $\delta > 0$,

$$\lim_{n\to\infty} \Pr\left(L_n^{\mathbf{Y}} \in \mathscr{M}^\delta \mid L_n^{\mathbf{Y}} \in \Gamma\right) = 1$$

with an exponential (in n) rate of convergence. Consequently for $U = \mathscr{M}^\delta$ results in $d_V(\mu_n^*, \text{co}(\mathscr{M}^\delta)) \to 0$. Since d_V is a convex function on $M_1(\Sigma) \times M_1(\Sigma)$, each point in $\text{co}(\mathscr{M}^\delta)$ is within variational distance δ of some point in $\text{co}(\mathscr{M})$. With $\delta > 0$ being arbitrarily small, limit points of μ_n^* are necessarily in the closure of $\text{co}(\mathscr{M})$.

Using Sanov's theorem we have

$$I_\Gamma = -\lim_{n\to\infty} \frac{1}{n} \log \Pr\left(L_n^{\mathbf{Y}} \in \Gamma\right)$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log\Pr\left(L_n^{\mathbf{Y}}\in\left(\mathscr{M}^{\delta}\right)^{c}\cap\Gamma\right)\leq-\inf_{\nu\in\left(\mathscr{M}^{\delta}\right)^{c}\cap\Gamma}H(\nu\mid\mu)$$
$$\leq-\inf_{\nu\in\left(\mathscr{M}^{\delta}\right)^{c}\cap\bar{\Gamma}}H(\nu\mid\mu)$$

Observe that $\mathscr{M}^{\delta}$ are open sets and, therefore $\left(\mathscr{M}^{\delta}\right)^{c}\cap\bar{\Gamma}$ are compact sets. Thus for some $\tilde{\nu}\in\left(\mathscr{M}^{\delta}\right)^{c}\cap\bar{\Gamma}$,

$$\inf_{\nu\in\left(\mathscr{M}^{\delta}\right)^{c}\cap\bar{\Gamma}}H(\nu\mid\mu)=H(\tilde{\nu}\mid\mu)>I_{\Gamma}$$

hence

$$\limsup_{n\to\infty}\Pr\left(L_n^{\mathbf{Y}}\in\left(\mathscr{M}^{\delta}\right)^{c}\mid L_n^{\mathbf{Y}}\in\Gamma\right)$$
$$=\limsup_{n\to\infty}\left\{\frac{1}{n}\log\Pr\left(L_n^{\mathbf{Y}}\in\left(\mathscr{M}^{\delta}\right)^{c}\cap\Gamma\right)-\frac{1}{n}\log\Pr\left(L_n^{\mathbf{Y}}\in\Gamma\right)\right\}<0$$

□

s

References

[1] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating renyi entropy.

[2] Nathanael L Ackerman, Cameron E Freer, and Daniel M Roy. On the computability of conditional probability. May 2010.

[3] R M Adelson. Compound poisson distributions. *OR*, 17(1):73–75, 1966.

[4] Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[5] Ahmed El Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 775–783, Cambridge, MA, USA, 2015. MIT Press.

[6] David Aldous and J Michael Steele. The objective method: Probabilistic combinatorial optimization and local weak convergence. In Harry Kesten, editor, *Probability on Discrete Structures*, Encyclopaedia of Mathematical Sciences, pages 1–72. Springer Berlin Heidelberg, 2004.

[7] Jaan Altosaar, Rajesh Ranganath, and David M Blei. Proximity variational inference. May 2017.

[8] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. June 2016.

[9] Michael Arbel and Arthur Gretton. Kernel conditional exponential family. November 2017.

[10] Evan Archer. Bayesian entropy estimation for countable discrete distributions.

[11] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. January 2017.

[12] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. January 2017.

[13] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. January 2017.

[14] N Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[15] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). March 2017.

[16] Jeremy Avigad. The metamathematics of ergodic theory. *Annals of Pure and Applied Logic*, 157(2):64–76, February 2009.

[17] John C Baez, Tobias Fritz, and Tom Leinster. A characterization of entropy in terms of information loss. June 2011.

[18] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins*, 79(4):1061–1078, April 2011.

[19] Rina Foygel Barber and Aaditya Ramdas. The p-filter: multi-layer FDR control for grouped hypotheses. December 2015.

[20] A D Barbour. Stein's method for diffusion approximations. *Probability Theory and Related Fields*, 84(3):297–322, 1990.

[21] A D Barbour, Oliver Johnson, Ioannis Kontoyiannis, and Mokshay Madiman. Compound poisson approximation via information functionals. *Electronic Journal of Probability*, 15:1344–1369, 2010.

[22] Andrew D Barbour, Vydas Cekanavicius, and Aihua Xia. On stein's method and perturbations. February 2007.

[23] Ole Barndor Nielsen MaPhy. Probability densities and levy densities.

[24] Adriaan Barri, Ann Dooms, and Peter Schelkens. The near shift-invariance of the dual-tree complex wavelet transform revisited. *Journal of mathematical analysis and applications*, 389(2):1303–1314, May 2012.

[25] Pierre Baudot and Daniel Bennequin. The homological nature of entropy. *Entropy*, 17 (5):3253–3318, May 2015.

[26] V P Belavkin. Multiple optimal quantum statistical hypothesis testing. *Stochastics. An International Journal of Probability and Stochastic Processes*.

[27] Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. May 2016.

[28] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: a review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, August 2013.

[29] Yoav Benjamini. Discovering the false discovery rate: False discovery rate. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 72(4):405–416, August 2010.

[30] Swanhild Bernstein. Wavelets in clifford analysis. In Daniel Alpay, editor, *Operator Theory*, pages 1–25. Springer Basel, 2014.

[31] G Beylkin, R Coifman, and V Rokhlin. Fast wavelet transforms and numerical algorithms I. *Communications on Pure and Applied Mathematics*, 44(2):141–183, March 1991.

[32] G Beylkin, R Coifman, and V Rokhlin. Fast wavelet transforms and numerical algorithms. In Christopher Heil and David F Walnut, editors, *Fundamental Papers in Wavelet Theory*. Princeton University Press, Princeton, January 2009.

[33] Gregory Beylkin and James M Keiser. On the adaptive numerical solution of nonlinear partial differential equations in wavelet bases. *Journal of computational physics*, 132 (2):233–259, April 1997.

[34] Malgorzata Bogdan, Ewout van den Berg, Weijie Su, and Emmanuel Candes. Statistical estimation and testing via the sorted L1 norm. October 2013.

[35] Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. May 2017.

[36] Michael V Boutsikas and Eutichia Vaggelatou. A new method for obtaining sharp compound poisson approximation error estimates for sums of locally dependent random variables. October 2010.

[37] Jop Briët and Peter Harremoës. Properties of classical and quantum Jensen-Shannon divergence. June 2008.

[38] Joan Bruna and Stéphane Mallat. Classification with scattering operators. November 2010.

[39] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, August 2013.

[40] Damian Brzyski, Christine B Peterson, Piotr Sobczyk, Emmanuel J Candès, Malgorzata Bogdan, and Chiara Sabatti. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1):61–75, January 2017.

[41] Calderon. SINGULAR INTEGRALS.

[42] Emmanuel J Candes and Laurent Demanet. The curvelet representation of wave propagators is optimally sparse. July 2004.

[43] Eric Carlen and Wilfrid Gangbo. Constrained steepest descent in the 2-wasserstein metric. *Annals of mathematics*, 157(3):807–846, May 2003.

[44] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for Non-Convex optimization. November 2016.

[45] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. December 2010.

[46] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. June 2016.

[47] Otis Chodosh. Optimal transport and ricci curvature: Wasserstein space over the interval. May 2011.

[48] Stéphan Clémençon, Aurélien Bellet, and Igor Colin. Scaling-up empirical risk minimization: Optimization of incomplete u-statistics. January 2015.

[49] Taco S Cohen and Max Welling. Harmonic exponential families on manifolds. May 2015.

[50] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17:13, January 2016.

[51] Maria Angelica Cueto, Jason Morton, and Bernd Sturmfels. Geometry of the restricted boltzmann machine. August 2009.

[52] Giuseppe Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer Science & Business Media, August 2006.

[53] Francesco D'Andrea and Pierre Martinetti. A view on optimal transport from noncommutative geometry. June 2009.

[54] Arijit Das and Achim Tresch. Computable learning via metric measure theory.

[55] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications:*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2010.

[56] John C Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence, and surrogate risk. March 2016.

[57] David Duvenaud, Oren Rippel, Ryan P Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. February 2014.

[58] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. March 2017.

[59] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. May 2015.

[60] Ahmed El Alaoui and Michael W Mahoney. Fast randomized kernel methods with statistical guarantees. November 2014.

[61] R. V. Gamkerlidze, editor. *Analysis II: Convex Analysis and Approximation Theory*, volume 14 of *Encyclopaedia of Mathematical Sciences*. Springer-Verlag, New York, 1990.

[62] Robert X Gao and Ruqiang Yan. From fourier transform to wavelet transform: A historical perspective. In Robert X Gao and Ruqiang Yan, editors, *Wavelets: Theory and Applications for Manufacturing*, pages 17–32. Springer US, Boston, MA, 2011.

[63] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax Rate-Optimal estimation of divergences between discrete distributions. May 2016.

[64] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. September 2015.

[65] Mathieu Hoyrup, Cristobal Rojas, and Klaus Weihrauch. Computability of the Radon-Nikodym derivative. December 2011.

[66] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q Weinberger. Multi-Scale dense networks for resource efficient image classification. March 2017.

[67] Martin Jaggi. Revisiting {frank-wolfe}: Projection-Free sparse convex optimization. In *Proceedings of The 30th International Conference on Machine Learning*, pages 427–435, 2013.

[68] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. November 2017.

[69] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction.

[70] William B Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, June 1986.

[71] B H Juang and L R Rabiner. A probabilistic distance measure for hidden markov models. *AT T Technical Journal*, 64(2):391–408, February 1985.

[72] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. October 2017.

[73] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. June 2016.

[74] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. June 2016.

[75] B J K Kleijn and Y Y Zhao. Criteria for posterior consistency. August 2013.

[76] I Kontoyiannis and M Madiman. Measure concentration for compound poisson distributions. June 2005.

[77] Ralph Kopperman, Hans-Peter A Künzi, and Paweł Waszkiewicz. Bounded complete models of topological spaces. *Topology and its applications*, 139(1):285–297, April 2004.

[78] Oluwasanmi Koyejo and Nagarajan Natarajan. Consistent binary classification with generalized performance metrics.

[79] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks.

[80] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm.

[81] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic acids research*, November 2016.

[82] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, September 2017.

[83] Pierre Moulin and Patrick R Johnstone. Kullback-Leibler divergence and the central limit theorem.

[84] Cameron Musco and Christopher Musco. Provably useful kernel matrix approximation in linear time. May 2016.

[85] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. May 2017.

[86] Xuanlong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. September 2011.

[87] Frank Nielsen. Chernoff information of exponential families. February 2011.

[88] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc., 2016.

[89] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15 (6):1191–1253, 2003.

[90] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. June 2016.

[91] Ankit B Patel, Tan Nguyen, and Richard G Baraniuk. A probabilistic theory of deep learning. April 2015.

[92] Mathew D Penrose and J E Yukich. Weak laws of large numbers in geometric probability. *The annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 13(1):277–303, January 2003.

[93] Mathew D Penrose and J E Yukich. Limit theory for point processes in manifolds. April 2011.

[94] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean, and Mark A DePristo. Creating a universal SNP and small indel variant caller with deep neural networks. March 2018.

[95] Mailbox R2 and 4259 Nagatsuta. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research: JMLR*, 11:3571–3594, 2010.

[96] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. Optimal rates and tradeoffs in multiple testing. May 2017.

[97] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. November 2015.

[98] Aaditya Ramdas, Rina Foygel Barber, Martin J Wainwright, and Michael I Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. March 2017.

[99] Benjamin Recht. A simpler approach to matrix completion. *Journal of machine learning research: JMLR*, 12(Dec):3413–3430, 2011.

[100] Yong Ren, Jialian Li, Yucen Luo, and Jun Zhu. Conditional generative Moment-Matching networks.

[101] Daniel M Roy. Computability, inference and modeling in probabilistic programming.

[102] Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. Exponential family embeddings. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 478–486. Curran Associates, Inc., 2016.

[103] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. June 2016.

[104] Ken-Iti Sato. STOCHASTIC INTEGRALS WITH RESPECT TO LEVY PROCESSES AND INFINITELY DIVISIBLE DISTRIBUTIONS.

[105] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of statistics*, 41(5):2263–2291, October 2013.

[106] Amir Sepehri. The bayesian SLOPE. August 2016.

[107] S Singh, Y Yang, B Poczos, and J Ma. Predicting Enhancer-Promoter interaction from genomic sequence with deep neural networks. *bioRxiv*, 2016.

[108] I Smith and A Ferrari. Generalizations related to hypothesis testing with the posterior distribution of the likelihood ratio. June 2014.

[109] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R G Lanckriet. On integral probability metrics, φ-divergences and binary classification. January 2009.

[110] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R G Lanckriet. On the empirical estimation of integral probability metrics. *Electronic journal of statistics*, 6:1550–1599, 2012.

[111] Weijie Su and Emmanuel Candes. SLOPE is adaptive to unknown sparsity and asymptotically minimax. March 2015.

[112] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from Tree-Structured long Short-Term memory networks. February 2015.

[113] V Y F Tan, A Anandkumar, L Tong, and A S Willsky. A large-deviation analysis for the maximum likelihood learning of tree structures. In *2009 IEEE International Symposium on Information Theory*, pages 1140–1144, June 2009.

[114] Amalio Telenti, Levi C T Pierce, William H Biggs, Julia di Iulio, Emily H M Wong, Martin M Fabani, Ewen F Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A Perkins, Franz J Och, Yaron Turpaz, and J Craig Venter. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11901–11906, October 2016.

[115] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. March 2015.

[116] Ilya Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. AdaGAN: Boosting generative models. January 2017.

[117] Jakub M Tomczak and Max Welling. Improving variational Auto-Encoders using convex combination linear inverse autoregressive flow. June 2017.

[118] Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. October 2009.

[119] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. September 2016.

[120] V N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 10(5):988–999, 1999.

[121] Ulrike von Luxburg and Bernhard Schoelkopf. Statistical learning theory: Models, concepts, and results. October 2008.

[122] Shun Watanabe. Neyman-Pearson test for Zero-Rate multiterminal hypothesis testing. November 2016.

[123] Klaus Weihrauch. Computability on computable metric spaces. *Theoretical computer science*, 113(2):191–210, June 1993.

[124] Asaf Weinstein, Rina Barber, and Emmanuel Candes. A power and prediction analysis for knockoffs with lasso statistics. December 2017.

[125] S Wisdom, T Powers, J Hershey, J Le Roux, and others. Full-Capacity unitary recurrent neural networks. *Advances in neural information processing systems*, 2016.

[126] D H Wolpert and W G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997.

[127] Yuefeng Wu and Subhashis Ghosal. Kullback leibler property of kernel mixture priors in bayesian density estimation. October 2007.

[128] Huiming Zhang, Lili Chu, and Yu Diao. Some properties of the generalized stuttering poisson distribution and its applications. July 2012.

[129] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-Scale kernel methods for independence testing. June 2016.

[130] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and applied mathematics*, 220(1–2):456–463, October 2008.

Acknowledgements

I would like to thank Achim Tresch for giving me the opportunity to work on my thesis. He supported me throughout the time of my PhD and helped me develop the soft skills necessary for an academic career. Furthermore, I would like to thank Andreas Beyer and his group for organizing the weekly talks which introduced me to various aspects of Computational Biology. I am also grateful to Korbinian Strimmer for being an unofficial part of my thesis committee and providing meaningful suggestions to improve my thesis. Finally, I would like to thank Lisa Marie Stephan for supporting me both personally and professionally throughout the years of late nights and for always pushing me to be more efficient.

Declaration

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig ange- fertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen , die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie abgesehen von unten angegebenen Teilpublikationen noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Achim Tresch betreut worden.

Teilpublikationen liegen nicht vor.

Arijit Das
2018

Arijit Das

Curriculum Vitae

Education

2013–2018 **PhD Computational Biology**, *Max Planck Institute for Plant Breeding Research*, Cologne, *Germany*.

2007–2009 **Masters of Statistics**, *Indian Institute of Technology*, Kanpur, *India*.

2004–2007 **Bachelor of Statistics**, *University of Delhi*, Delhi, *India*.

PhD Thesis

Title *Design and Analysis of Statistical Learning Algorithms which Control False Discoveries*

Supervisors Professor Achim Tresch

Professional Experience

2011–2012 **Research Associate**, Trinity College Dublin, Dublin, Ireland.
Developed Variational Bayes Algorithms for Telecommunication applications.

2009–2011 **Research Engineer**, INRIA, Bordeaux, France.
Developed Time Series models to predict Electricity prices for millions of customers of Electricite de France (EDF).

Publications

2017 Learning Interactions in the Human Genome, Arijit Das and Achim Tresch, under preparation

2017 Locally Lie Group Invariant Reproducing Kernels, Arijit Das and Achim Tresch, submitted to Journal of Machine Learning Research

2017 Locally Accelerated Coordinate Descent Method for Convex Optimization, Arijit Das and Achim Tresch, submitted to Journal of Machine Learning Research

2017 Push-forward Metric Measures, Arijit Das and Achim Tresch, submitted to Annals of Statistics

Kämmergasse 12 – Köln, NRW 50676
(049) 176 8437 6170 • *das@mpipz.mpg.de*

2017 Local Dependence Approximation of Kernel Matrices, Arijit Das and Achim Tresch, submitted to NIPS Conference 2017

2014 Computable Learning via Metric Measure Theory, Arijit Das and Achim Tresch, NIPS 2014 Modern Nonparametrics 3: Automating the Learning Pipeline Workshop

2011 A Variational Bayes Approach to Decoding in a Phase-Uncertain Digital Receiver, Arijit Das and Anthony Quinn, ISSC 2011, Trinity College Dublin

2010 Méthodes d'estimation récursive pour la prévision de la consommation d'électricité: La prévision de courbes de charge par estimation fonctionnelle non paramétrique, Bernard Bercu, Francois Caron, Arijit Das, Frederic Proia, INRIA Research Report

2010 Méthodes d'estimation récursive pour la prévision de la consommation d'électricité : La prévision de courbes de charge par séries chronologiques linéaires, Bernard Bercu, Francois Caron, Arijit Das, Frederic Proia, INRIA Research Report