MINIMAX ESTIMATION OF MISSING MASS

A Project Report

submitted by

NIKHILESH RAJARAMAN

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

MAY 2017

THESIS CERTIFICATE

This is to certify that the thesis titled **Minimax Estimation of Missing Mass**, submitted by **Nikhilesh Rajaraman**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Andrew Thangaraj
Research Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date: 7th May 2017

ACKNOWLEDGEMENTS

I thank my project guide, Prof. Andrew Thangaraj, for giving me the opportunity to work on this project and for his valuable guidance, ideas and feedback.

I also thank Ananda Theertha Suresh from Google Research and Alon Orlitsky from UCSD for collaborating with me on this project and providing their expert insights on the problem.

Finally, I would like to thank my parents for their constant encouragement and support.

ABSTRACT

KEYWORDS: Estimation; Missing Mass; Squared Error Loss; Good-Turing Estimator of Missing Mass; Poisson Sampling Model; Worst-case Risk; Linear Estimator of Missing Mass; Multinomial Sampling Model; Minimax Risk.

Missing mass estimation is a basic problem in statistics with wide practical applications including language modelling and ecology. This work considers the problem of minimax estimation of missing mass under a squared error loss. The popular Good-Turing estimator is shown to have a worst-case risk of $1/n$ under the Poisson sampling model and we prove that it is asymptotically optimal within an extended class of linear estimators. Under the multinomial sampling model, the worst-case risk of the Good-Turing estimator is shown to be between $0.6080/n$ and $0.6179/n$ asymptotically. The minimax risk under the multinomial model is shown to be asymptotically lower bounded by $0.25/n$, which shows that the Good-Turing estimator is order-optimal under this model.

TABLE OF CONTENTS

NOTATION

$\mathbb{I}(E)$	Indicator random variable for event E
$\mathcal{X}$	Sample space
k	Size of sample space
u, v	Elements of the sample space
p or P	Sampling distribution
$\mathcal{P}$	Set of probability distributions p belongs to
$p\left(u\right)$	Probability of symbol u under distribution p
n	Deterministic sample size
N	Poisson distributed sample size
X^n	Sample of size n
$\Phi_i\left(X^n\right)$	Number of symbols appearing i times in X^n
M_0	Missing mass
$\hat{M}_0$	Missing mass estimator
$\hat{M}_0^{GT}$	Good-Turing estimator of missing mass
R_n	Risk or worst-case risk of estimator
R_n^*	Minimax Risk

INTRODUCTION AND CHAPTER OUTLINES

Given a set of independent samples from a probability distribution, the missing mass of the samples is the total probability of all unseen symbols. A basic problem in statistics is to estimate the missing mass with no information about the distribution given only the samples - this has applications in several fields including language modelling and ecology. One of the earliest missing mass estimators is the Good-Turing estimator, developed by Irving Good and Alan Turing during World War II. The Good-Turing estimator uses the fraction of symbols that appear only once in the sample as an estimate of the missing mass.

There is a significant body of work that analyzes missing mass estimators; especially the Good-Turing estimator. However, a notable omission in the literature is an analysis of the worst-case risk of estimators and the minimax risk of the missing mass estimation problem. This work seeks to bridge that gap by providing the first conclusive results on the minimax risk of missing mass. Additionally, we also show some optimality guarantees for the Good-Turing estimator in the worst-case sense and give tight bounds on its worst-case risk.

Chapters 1 and 2 introduce the problem of minimax estimation of missing mass and any necessary background. Chapter 2 additionally introduces the Good-Turing estimator and contains a survey of prior work on missing mass estimation and its applications.

The main body of the work starts from chapter 3, which introduces linear estimators - a class of estimators that generalize the Good-Turing estimator. We analyze linear estimators under the Poisson sampling model and show that the Good-Turing estimator has a worst-case risk of $\frac{1}{n}$. Additionally, we prove that the Good-Turing estimator is asymptotically optimal among all linear estimators.

Having shown its optimality under the Poisson model, chapter 4 looks at the worst-case risk of the Good-Turing estimator under the multinomial sampling model. We show that the worst-case risk of the Good-Turing estimator lies between $0.6080/n$ and $0.6179/n$. This chapter consists of the work of the author's collaborator and guide, Andrew Thangaraj, and has been included for completeness.

Finally, chapter 5 puts the worst case risk of the Good-Turing estimator in perspective by showing lower bounds on the minimax risk. Two lower bounds are obtained using different methods; the first using Bayes risk gives a bound of $4/27n$ and the second using distribution estimation yields a $1/4n$ lower bound. Section 5.2 on the distribution estimation bound is by the author's collaborator, Ananda Theertha Suresh, and has been included for completeness.

CHAPTER 1

FUNDAMENTALS OF ESTIMATION

This chapter provides a brief review of the concepts in estimation theory used in this work. We first describe the basic setup of estimation problems. After this, we look at loss functions and define the risk of an estimator, which serves as a metric of evaluation. From this, we head to defining the worst-case risk of an estimator and the minimax risk for an estimation problem. Finally, we take a look at the multinomial and Poisson sampling models, which we use in chapter 3.

1.1 Statistical Estimation

Let p be an unknown probability distribution over a sample space $\mathcal{X}$, parametrized by parameters θ. Assume that θ is unknown, but deterministic. We observe a vector of n i.i.d samples from the distribution, which we denote $X^n = (X_1, X_2, \dots X_n)$.

We are interested in finding the value of a function $f\left(X^n, \theta\right)$ that depends on both the distribution and sample. Since θ is unknown, we can only use the observed sample X^n to estimate $f\left(X^n, \theta\right)$. A function $\hat{f}\left(X^n\right)$ that seeks to approximate the value of $f\left(X^n, \theta\right)$ is called an **estimator** of $f\left(X^n, \theta\right)$.

1.2 Loss and Risk

Having defined an estimator, we now need a means to evaluate the performance of an estimator and compare different estimators. To this end, we define two quantities : the loss and the risk of an estimator.

The **loss** $L_n\left(f, \hat{f}, \theta, X^n\right)$ is a function that measures the performance of an estimator for a particular sample and parameter vector. Several loss functions are commonly used in literature, including:

- Squared error or ℓ_2 loss : $L_n = \left(f\left(X^n, \theta\right) - \hat{f}\left(X^n\right)\right)^2$
- Absolute error or ℓ_1 loss : $L_n = \left|f\left(X^n, \theta\right) - \hat{f}\left(X^n\right)\right|$
- Zero-one loss : $L_n = 0$ if $f\left(X^n, \theta\right) = \hat{f}\left(X^n\right)$, $L_n = 1$ otherwise.

From the definition, it can be seen that the loss of an estimator depends on the particular sample X^n that is observed. To obtain a single metric that is independent of the sample, we need to aggregate the loss for each sample in some manner. One way of doing this is using the **risk** $R_n\left(f, \hat{f}, \theta\right)$, defined as:

$$R_n\left(f, \hat{f}, \theta\right) \triangleq \mathbb{E}_{X^n}\left[L_n\left(f, \hat{f}, \theta, X^n\right)\right] \tag{1.1}$$

The expectation is explicitly specified to be over X^n to emphasize that θ is not a random variable - this will be omitted in future sections.

1.3 Minimax Risk and Minimax Estimators

From the definition of risk in the previous section, it can be seen that the risk depends on the value of the parameter vector θ as well. However, as the true value of θ is unknown, we cannot compute the risk of an estimator in practice. So, we instead use the **worst-case risk** $R_n\left(f, \hat{f}\right)$ as a metric of performance, defined as:

$$R_n\left(f, \hat{f}\right) \triangleq \max_{\theta} R_n\left(f, \hat{f}, \theta\right) \tag{1.2}$$

Note that the same notation is used for both the risk and the worst-case risk; which one is being referred to will be made clear from either the context or the function parameters.

The lowest possible worst-case risk attainable by any estimator for a given problem is called the **minimax risk** $R_n^*(f)$, defined as:

$$R_n^*(f) \triangleq \min_{\hat{f}} R_n\left(f, \hat{f}\right) \tag{1.3}$$

$$= \min_{\hat{f}} \max_{\theta} R_n\left(f, \hat{f}, \theta\right) \tag{1.4}$$

The second equation explains why the term "minimax" is used to describe this risk. An estimator whose worst-case risk is equal to the minimax risk is called a **minimax estimator**.

The final goal of solving an estimation problem under the minimax framework is to find a minimax estimator. If this is not possible, the next best thing is to find an estimator whose worst-case risk is close to the minimax risk for the problem.

1.4 Sampling Models

The definition of risk in equation 1.1 involves an expectation over the vector X^n. To compute this, the distribution of X^n is required. However, we have available only the distribution p of each element of X^n - finding the distribution of X^n requires additional information on how the samples are generated. We look at two popular sampling models : the Multinomial model and the Poisson model.

Under the **Multinomial sampling model**, the number of samples n is a deterministic constant. The vector of samples X^n consists of n i.i.d samples from the underlying parametrized distribution p.

Under the **Poisson sampling model**, the number of samples is a random variable which follows a Poisson distribution. To emphasize this, we denote the number of samples as N instead, with $N \sim \text{Pois}(n)$. Here, n is the mean of the Poisson distribution used to generate the number of samples and is a deterministic constant. As before, the vector of samples X^N consists of N i.i.d samples from p. The Poisson sampling model arises naturally when the samples are generated from a Poisson process with a fixed duration of sampling.

One of the useful features of the Poisson sampling model is that the counts of different symbols in the sample X^N are independent. This is not true for the multinomial sampling model, for instance, since the sum of the counts of all symbols is n, which makes them dependent. We will make use of this property in chapter 3.

CHAPTER 2

MISSING MASS

In this chapter, we introduce the idea of missing mass. We start with its basic definition and look at some applications of missing mass in language modelling and ecology. We then turn to the problem of estimating missing mass from a sample; in particular, we look at the problem of finding good missing mass-estimators in the minimax sense. We then look at the Good-Turing missing mass estimator - one of the earliest developed estimators for this problem. The chapter concludes with a brief survey of some of the important prior results on missing mass estimation.

2.1 Definition

Let p be an unknown probability distribution over an *unknown* finite sample space $\mathcal{X}$. We observe n samples $X^n = (X_1, X_2, \ldots X_n)$ which are i.i.d according to p. The **missing mass** $M_0\left(X^n, p\right)$ is defined as:

$$M_0\left(X^n, p\right) \triangleq \mathbb{P}\left(\mathcal{X} \backslash \left\{X_1, X_2, \ldots X_n\right\}\right) \tag{2.1}$$

Essentially, the missing mass is the total probability of all symbols in $\mathcal{X}$ that have not been observed in the sample X^n.

To make this definition more analytically tractable, we define a random variable $N_u\left(X^n\right)$ which denotes the number of times a symbol $u \in \mathcal{X}$ appears in the sample X^n. Using this, we can rewrite the definition of missing mass as:

$$M_0\left(X^n, p\right) \;=\; \sum_{u \in \mathcal{X}} p\left(u\right) \mathbb{I}\left(N_u\left(X^n\right) = 0\right) \tag{2.2}$$

Here, $p\left(u\right)$ denotes the probability of symbol u according to the distribution p and $\mathbb{I}$ denotes the indicator function.

2.2 Applications of Missing Mass Estimation

Missing mass estimation has several applications in multiple fields. Most applications stem from the use of missing mass in estimating a distribution from samples - Gale and Sampson (1995) provides an account of this, for instance. Another theoretical application is explored by Vu *et al.* (2007), which looks at using sample coverage estimates to improve estimates of a distribution's entropy, where the coverage of a sample is $1 - M_0\left(X^n, p\right)$.

Language modelling is one field where missing mass sees use. Several works (Katz, 1987; Chen and Goodman, 1999; Church and Gale, 1991) use missing mass estimates to improve estimates of n-gram probability distributions by smoothing. Sproat *et al.* (1996) use missing mass estimators in the problem of word segmentation of Chinese text.

Another sphere where missing mass estimators see use is ecology, where missing mass helps provide estimates of the size of a population. Chao and Lee (1992) use sample coverage estimates to obtain estimates of the number of species in a population. Shen *et al.* (2003) use sample coverage estimators to predict the number of new species that will be observed when a further survey is conducted, which measures the viability of surveying.

2.3 Minimax Estimation of Missing Mass

Let $\hat{M}_0\left(X^n\right)$ be an estimator of the missing mass $M_0\left(X^n, p\right)$. As discussed in chapter 1, we define the squared loss risk $R_n\left(M_0, \hat{M}_0, p\right)$ of the estimator as:

$$R_n\left(M_0, \hat{M}_0, p\right) \triangleq \mathbb{E}_{X^n}\left[\left(M_0\left(X^n, p\right) - \hat{M}_0\left(X^n\right)\right)^2\right] \tag{2.3}$$

We will henceforth omit the M_0 argument from the risk and write it simply as $R_n\left(\hat{M}_0, p\right)$, since we will only be looking at estimators of missing mass. We can also define the worst case risk $R_n\left(\hat{M}_0\right)$ and the minimax risk R_n^*:

$$\begin{aligned} R_n\left(\hat{M}_0\right) &\triangleq \max_p \mathbb{E}_{X^n}\left[\left(M_0\left(X^n, p\right) - \hat{M}_0\left(X^n\right)\right)^2\right] && (2.4) \\ R_n^* &\triangleq \min_{\hat{M}_0} \max_p \mathbb{E}_{X^n}\left[\left(M_0\left(X^n, p\right) - \hat{M}_0\left(X^n\right)\right)^2\right] && (2.5) \end{aligned}$$

The above expressions are applicable when we work with the multinomial sampling model with a fixed n. Under the Poisson sampling model, the definitions remain mostly the same, except that the fixed length sample X^n is replaced with the random length sample X^N and the expectations are taken over both N and X^N.

2.4 The Good-Turing Estimator

Some of the earliest work on missing mass estimation was done by Irving Good and Alan Turing, which resulted in the so-called Good-Turing (GT) estimator for missing mass (Good, 1953). Their work describes a class of method of moments estimators for various population parameters that depend only on the sample size

n and the number of symbols that appear an equal number of times, as described below.

Let $\Phi_i(X^n)$ denote number of distinct symbols that have appeared i times in the sample X^n. We define it mathematically as:

$$\Phi_i(X^n) \triangleq \sum_{u \in \mathcal{X}} \mathbb{I}(N_u(X^n) = i) \tag{2.6}$$

Here, $N_u(X^n)$ is the number of times u appears in X^n, as defined in section 2.1. The **Good-Turing (GT) estimator** of missing mass $\hat{M}_0^{GT}(X^n)$ is defined as:

$$\hat{M}_0^{GT}(X^n) \triangleq \frac{1}{n}\Phi_1(X^n) \tag{2.7}$$

In later chapters, we will look at the worst-case risk of the GT estimator and use it as a benchmark for evaluating other estimators.

2.5 Prior Results on Missing Mass Estimation

We first look at some results that give performance guarantees for the Good-Turing estimator. McAllester and Schapire (2000) show that the absolute bias of the Good-Turing estimator is upper bounded by $1/n$, and use this to show stronger PAC bounds on the error of the Good-Turing estimator. Esty (1983) showed that the Good-Turing estimate follows a normal limit law in its convergence to the true missing mass. Several variations of the Good-Turing estimator for problems apart from missing mass estimation such as distribution estimation (Orlitsky and Suresh, 2015), unseen species estimation (Good and Toulmin, 1956; Orlitsky *et al.*, 2016) sequence estimation (Wagner *et al.*, 2007), rare event probability estimation (Ohannessian and Dahleh, 2012) have been analyzed.

Apart from the performance of estimators, some results show distribution-free concentration bounds on the missing mass itself. Multiple works (Berend and Kontorovich, 2013; McAllester and Ortiz, 2003) show an exponentially decaying concentration bound for the missing mass about its mean, while Ben-Hamou *et al.* (2017) show sub-Gaussian and sub-Gamma bounds for the lower and upper tails respectively.

CHAPTER 3

LINEAR ESTIMATORS OF MISSING MASS

We introduced the Good-Turing estimator for missing mass in the previous chapter. Now, we look at a class of estimators that naturally extend the Good-Turing estimator, which we call linear estimators. The chapter starts with a mathematical definition of linear estimators. We then look at the problem of computing the risk of a linear estimator. This is difficult to do under the multinomial sampling model, so we instead compute the risk under the Poisson sampling model, which is summarized in Lemma 3.2.1. We then use this expression for risk to show that the worst case risk of the Good-Turing estimator is $\frac{1}{n}$. Finally, we find a lower bound on the worst-case risk of an arbitrary linear estimator and show that the Good-Turing estimator is asymptotically optimal within the class of linear estimators, made precise in Theorem 3.5.1.

3.1 Definition

Recall the definition of $\Phi_i\left(X^n\right)$ and the Good-Turing estimator from section 2.4. A **linear estimator** generalizes the idea of the Good-Turing estimator; we denote and define a linear estimator as:

$$\begin{aligned} \hat{M}_0\left(X^n; \boldsymbol{c}\left(n\right)\right) &\triangleq \sum_{i=1}^{\infty} c_i\left(n\right)\Phi_i\left(X^n\right) && (3.1) \\ &= \sum_{i=1}^{\infty}\sum_{u\in\mathcal{X}} c_i\left(n\right)\mathbb{I}\left(N_u\left(X^n\right)=i\right) && (3.2) \end{aligned}$$

where $c_i(n), i \in \mathbb{N}$ are scaling factors that only depend on n and $\boldsymbol{c}(n) = (c_1(n), c_2(n), \ldots)$. Equation 3.2 follows from the definition of $\Phi_i(X^n)$ in equation 2.6. For any finite N, the sum absolutely converges if the set $\{|c_i(n)|, i = 1, 2, \ldots\}$ is bounded, so the estimate is well defined for such estimators and the order of summation can be interchanged - only such estimators will be considered in the following discussion. If there exists a $J \in \mathbb{N}$ such that $c_i(n) = 0$ for all $i > J$, the estimator is termed as an **order-J estimator**, which is denoted as $\hat{M}_0(X^n; c_1(n), c_2(n) \ldots, c_J(n))$. Note that the Good-Turing estimator is an order-1 linear estimator $\hat{M}_0\left(X^N; 1/n\right)$.

3.2 Risk of Linear Estimators and the Poisson Model

In this section, we look at computing the risk of a linear estimator $\hat{M}_0(X^n; \boldsymbol{c}(n))$ for a symbol distribution p. We do this using a **bias-variance decomposition**:

$$\begin{aligned} R_n\left(\hat{M}_0, p\right) &= \mathbb{E}\left[\left(M_0(X^n, p) - \hat{M}_0(X^n; \boldsymbol{c}(n))\right)^2\right] \\ &= \mathbb{E}\left[M_0(X^n, p) - \hat{M}_0(X^n; \boldsymbol{c}(n))\right]^2 \\ &\quad + \text{var}\left(M_0(X^n, p) - \hat{M}_0(X^n; \boldsymbol{c}(n))\right) \end{aligned} \tag{3.3}$$

The first term is the squared **bias** of $\hat{M}_0$ and the second term is its **variance**:

$$\begin{aligned} \text{Bias} &\triangleq \mathbb{E}\left[\sum_{u \in \mathcal{X}} \sum_{i=1}^{\infty} \left(c_i(n) \mathbb{I}(N_u(X^n) = i)\right)\right. \\ &\qquad \left. - p(u) \mathbb{I}(N_u(X^n) = 0)\right] \end{aligned} \tag{3.4}$$

$$\begin{aligned} \text{Variance} &\triangleq \text{var}\left[\sum_{u \in \mathcal{X}} \sum_{i=1}^{\infty} \left(c_i(n) \mathbb{I}(N_u(X^n) = i)\right)\right. \\ &\qquad \left. - p(u) \mathbb{I}(N_u(X^n) = 0)\right] \end{aligned} \tag{3.5}$$

The bias can be easily computed for any p using the linearity of expectation. However, we quickly run into a problem when trying to evaluate the variance. The variance consists of a summation of terms for each symbol $u \in \mathcal{X}$ which each depend on $N_u\left(X^n\right)$. However, for a fixed sample size n, the $N_u\left(X^n\right)$ are not independent, so we cannot express the variance of the sum as a sum of variances.

To solve this problem, we use the approach Orlitsky *et al.* (2016) adopted for a similar problem, which is to use the *Poisson* sampling model instead of the multinomial model. Under this model, we replace the number of samples n with a random variable $N \sim \text{Poisson}\left(n\right)$. This results in $N_u\left(X^n\right)$ being independent for each $u \in \mathcal{X}$, so we can express the variance of the sum as a sum of variances. We now proceed to evaluate the bias and variance under this model.

3.2.1 Computing the Bias

$$\begin{aligned} \text{Bias} &= \mathbb{E}\left[\sum_{u\in\mathcal{X}}\sum_{i=1}^{\infty}\left(c_i\mathbb{I}\left(N_u = i\right)\right) - p(u)\mathbb{I}\left(N_u = 0\right)\right] \\ &= \sum_{u\in\mathcal{X}}\sum_{i=1}^{\infty}\left(c_i\mathbb{P}\left(N_u = i\right)\right) - p(u)\mathbb{P}\left(N_u = 0\right) \qquad (3.6) \end{aligned}$$

Here, N_u is short for $N_u\left(X^n\right)$ and c_i is short for $c_i\left(n\right)$. Under the Poisson model, $N_u\left(X^N\right) \sim \text{Poisson}\left(np\left(u\right)\right)$, so $\mathbb{P}\left(N_u\left(X^N\right) = i\right) = \frac{1}{i!}\exp\left(-np\left(u\right)\right)\left(np\left(u\right)\right)^i$. Thus, we have:

$$\begin{aligned} \text{Bias} &= \sum_{u\in\mathcal{X}}\exp\left(-np\left(u\right)\right)\cdot\left(\sum_{i=1}^{\infty}\left(\frac{1}{i!}c_i\left(n\right)\left(np\left(u\right)\right)^i\right) - p(u)\right) \\ &= \sum_{u\in\mathcal{X}}\exp\left(-np\left(u\right)\right)\cdot\left(f_c\left(np\left(u\right)\right) - p(u)\right) \qquad (3.7) \end{aligned}$$

where $f_c\left(x\right) \triangleq \sum_{i=1}^{\infty} \frac{1}{i!} c_i\left(n\right) x^i$.

3.2.2 Computing the Variance

$$
\begin{aligned}
\text{Variance} &= \text{var}\left[\sum_{u\in\mathcal{X}}\sum_{i=1}^{\infty}\left(c_i\mathbb{I}\left(N_u=i\right)\right)-p(u)\mathbb{I}\left(N_u=0\right)\right] \\
&= \sum_{u\in\mathcal{X}}\text{var}\left[\sum_{i=1}^{\infty}\left(c_i\mathbb{I}\left(N_u=i\right)\right)-p(u)\mathbb{I}\left(N_u=0\right)\right] \\
&= \sum_{u\in\mathcal{X}}\mathbb{E}\left[\left(\sum_{i=1}^{\infty}\left(c_i\mathbb{I}\left(N_u=i\right)\right)-p(u)\mathbb{I}\left(N_u=0\right)\right)^2\right] \\
&\quad -\mathbb{E}\left[\sum_{i=1}^{\infty}\left(c_i\mathbb{I}\left(N_u=i\right)\right)-p(u)\mathbb{I}\left(N_u=0\right)\right]^2 \\
&= \sum_{u\in\mathcal{X}}\sum_{i=1}^{\infty}\left(c_i^2\mathbb{P}\left(N_u=i\right)\right)+p^2(u)\mathbb{P}\left(N_u=0\right) \\
&\quad -\left(\sum_{i=1}^{\infty}\left(c_i\mathbb{P}\left(N_u=i\right)\right)-p(u)\mathbb{P}\left(N_u=0\right)\right)^2 \qquad (3.8) \\
&= \sum_{u\in\mathcal{X}}\exp\left(-np\left(u\right)\right)\cdot\left(\sum_{i=1}^{\infty}\left(\frac{1}{i!}c_i^2\left(n\right)\left(np\left(u\right)\right)^i\right)+p^2(u)\right) \\
&\quad -\exp\left(-2np\left(u\right)\right)\cdot\left(\sum_{i=1}^{\infty}\left(\frac{1}{i!}c_i^2\left(n\right)\left(np\left(u\right)\right)^i\right)-p(u)\right)^2 \\
&= \sum_{u\in\mathcal{X}}\exp\left(-np\left(u\right)\right)\cdot\left(f_{c^2}\left(np\left(u\right)\right)+p^2(u)\right) \\
&\quad -\exp\left(-2np\left(u\right)\right)\cdot\left(f_c\left(np\left(u\right)\right)-p(u)\right)^2 \qquad (3.9)
\end{aligned}
$$

where $f_{c^2}\left(x\right) \triangleq \sum_{i=1}^{\infty} \frac{1}{i!} c_i^2\left(n\right) x^i$. 3.8 follows from the fact that $\mathbb{I}\left(N_u\left(X^N\right)=i\right)\mathbb{I}\left(N_u\left(X^N\right)=j\right)=0$ for $i\neq j$.

3.2.3 Putting it All Together

From equations 3.3,3.7 and 3.9, we get an expression for the risk of any linear estimator, summarized in the lemma below:

Lemma 3.2.1. *For any linear estimator $\hat{M}_0\left(X^n; \boldsymbol{c}\left(n\right)\right)$ under the Poisson sampling model, we have:*

$$\begin{aligned} R_n\left(\hat{M}_0, p\right) &= \left(\sum_{u\in\mathcal{X}} \exp\left(-np\left(u\right)\right)\cdot\left(f_c\left(np\left(u\right)\right)-p(u)\right)\right)^2 \\ &\quad+\sum_{u\in\mathcal{X}} \exp\left(-np\left(u\right)\right)\cdot\left(f_{c^2}\left(np\left(u\right)\right)+p^2(u)\right) \\ &\qquad -\exp\left(-2np\left(u\right)\right)\cdot\left(f_c\left(np\left(u\right)\right)-p(u)\right)^2 \end{aligned}$$

where $f_c\left(x\right) \triangleq \sum_{i=1}^{\infty}\frac{1}{i!}c_i\left(n\right)x^i$ *and* $f_{c^2}\left(x\right) \triangleq \sum_{i=1}^{\infty}\frac{1}{i!}c_i^2\left(n\right)x^i$.

3.3 Worst-case Analysis of the Good-Turing Estimator

We now have the tools to compute the worst-case risk $R_n\left(\hat{M}_0^{GT}\right)$ for the Good-Turing estimator under the Poisson sampling model. For the Good-Turing estimator $f_c\left(x\right)=\frac{x}{n}$ and $f_{c^2}\left(x\right)=\frac{x}{n^2}$. From Lemma 3.2.1, we get:

$$R_n\left(\hat{M}_0^{GT}, p\right) = \sum_{u\in\mathcal{X}} \exp\left(-np\left(u\right)\right)\cdot\left(\frac{p(u)}{n}+p^2(u)\right) \tag{3.10}$$

$$R_n\left(\hat{M}_0^{GT}\right) = \max_p \sum_{u\in\mathcal{X}} \exp\left(-np\left(u\right)\right)\cdot\left(\frac{p(u)}{n}+p^2(u)\right) \tag{3.11}$$

An interesting side-note is that the bias term vanishes for the Good-Turing estimator - meaning that it is unbiased under the Poisson model.

We first obtain a upper bound for $R_n\left(\hat{M}_0^{GT}\right)$. To do this, we obtain a distribution independent upper bound on the risk by using the inequality $\exp\left(-x\right) \leq \frac{1}{1+x}$, $x \geq 0$:

$$\begin{aligned}
R_n\left(\hat{M}_0^{GT}\right) &\leq \max_p \sum_{u\in\mathcal{X}} \frac{1}{1+np\left(u\right)} \cdot \left(\frac{p(u)}{n} + p^2(u)\right) \\
&= \max_p \frac{1}{n} \sum_{u\in\mathcal{X}} p(u) \\
&= \frac{1}{n}
\end{aligned}$$

$$\therefore R_n\left(\hat{M}_0^{GT}\right) \leq \frac{1}{n} \tag{3.12}$$

To obtain a corresponding lower bound, we evaluate $R_n\left(\hat{M}_0^{GT}, p^U\right)$ for the uniform distribution p^U over the set $\mathcal{X} = \{1, 2 \ldots k\}$. For this distribution, $p^U\left(u\right) = \frac{1}{k}$ for all $u \in \mathcal{X}$.

$$\begin{aligned}
R_n\left(\hat{M}_0^{GT}\right) &\geq R_n\left(\hat{M}_0^{GT}, p^U\right) \\
&= \sum_{u\in\mathcal{X}} \exp\left(-\frac{n}{k}\right) \cdot \left(\frac{1}{kn} + \frac{1}{k^2}\right) \\
&= \exp\left(-\frac{n}{k}\right) \cdot \left(\frac{1}{k} + \frac{1}{n}\right)
\end{aligned}$$

Letting $k \to \infty$, we get:

$$R_n\left(\hat{M}_0^{GT}\right) \geq \frac{1}{n} \tag{3.13}$$

From equations 3.13 and 3.12, we obtain an exact expression for the risk of the Good-Turing estimator under the Poisson model:

$$R_n\left(\hat{M}_0^{GT}\right) = \frac{1}{n} \tag{3.14}$$

3.4 Worst-case Analysis of Linear Estimators

Lemma 3.2.1 can also be used to obtain lower bounds on the worst-case risk of any linear estimator $\hat{M}_0\left(X^N; \boldsymbol{c}(n)\right)$. To do this, we once again bound $R_n\left(\hat{M}_0\right)$ by $R_n\left(\hat{M}_0, p^U\right)$, where p^U is the uniform distribution over $\mathcal{X} = \{1, 2 \dots k\}$.

$$\begin{aligned}
R_n\left(\hat{M}_0\right) &\geq R_n\left(\hat{M}_0, p^U\right) \qquad (3.15)\\
&= \left(\sum_{u\in\mathcal{X}} \exp\left(-\frac{n}{k}\right)\cdot\left(f_c\left(\frac{n}{k}\right) - \frac{1}{k}\right)\right)^2\\
&\quad + \sum_{u\in\mathcal{X}}\left(\exp\left(-\frac{n}{k}\right)\cdot\left(f_{c^2}\left(\frac{n}{k}\right)+\frac{1}{k^2}\right)\right.\\
&\qquad \left. - \exp\left(-2\frac{n}{k}\right)\cdot\left(f_c\left(\frac{n}{k}\right) - \frac{1}{k}\right)^2\right)\\
&= \left(k^2 - k\right)\exp\left(-\frac{2n}{k}\right)\cdot\left(f_c\left(\frac{n}{k}\right)-\frac{1}{k}\right)^2\\
&\quad + k\exp\left(-\frac{n}{k}\right)\cdot\left(f_{c^2}\left(\frac{n}{k}\right)+\frac{1}{k^2}\right) \qquad (3.16)
\end{aligned}$$

The RHS in 3.16 can be seen to be quadratic in the c_i. We can find a $\boldsymbol{c}^* = (c_1^*, c_2^*, \dots)$ that minimizes it by setting its gradient to 0:

$$\begin{aligned}
0 &= 2\left(k^2-k\right)\exp\left(-\frac{2n}{k}\right)\cdot\left(f_{c^*}\left(\frac{n}{k}\right)-\frac{1}{k}\right)\cdot\frac{\left(\frac{n}{k}\right)^i}{i!}\\
&\quad + k\exp\left(-\frac{n}{k}\right)\cdot\frac{\left(\frac{n}{k}\right)^i}{i!}\cdot 2c_i^* \qquad (3.17)
\end{aligned}$$

Solving equation 3.17 for $\boldsymbol{c}^*$, we get:

$$\begin{aligned}
c_i^* &= \frac{\exp\left(-\frac{n}{k}\right)}{A}\left(1-\frac{1}{k}\right) && (3.18)\\
f_{c^*}\left(\frac{n}{k}\right) &= \left(1-\frac{1}{A}\right)\cdot\frac{1}{k} && (3.19)\\
f_{(c^*)^2}\left(\frac{n}{k}\right) &= \exp\left(-\frac{n}{k}\right)\left(\frac{1}{A}-\frac{1}{A^2}\right)\left(\frac{1}{k}-\frac{1}{k^2}\right) && (3.20)
\end{aligned}$$

where $A = (k-1)\cdot\left(\exp\left(-\frac{n}{k}\right)-1\right)+1$. Plugging this into 3.16, we get:

$$\begin{aligned}
R_n\left(\hat{M}_0\right) \geq\; & \frac{(k-1)}{kA^2}\exp\left(-\frac{2n}{k}\right)+k\exp\left(-\frac{n}{k}\right)\cdot\\
& \left[\exp\left(-\frac{n}{k}\right)\left(\frac{1}{A}-\frac{1}{A^2}\right)\left(\frac{1}{k}-\frac{1}{k^2}\right)+\frac{1}{k^2}\right] && (3.21)
\end{aligned}$$

Finally, we evaluate the bound as $m \to \infty$. Noting that $\lim_{k\to\infty} A = n$, we get:

$$R_n\left(\hat{M}_0\right) \geq \frac{1}{n}\pm o\left(\frac{1}{n}\right) \qquad (3.22)$$

3.5 Summary

We summarize the main results of this chapter in the following theorem:

Theorem 3.5.1. *Under the Poisson sampling model, the Good-Turing estimator is asymptotically optimal within the class of linear estimators using the worst-case risk measure. Specifically, if $\hat{M}_0$ is an arbitrary linear estimator and $\hat{M}_0^{GT}$ is the Good-Turing estimator:*

$$R_n\left(\hat{M}_0\right) \geq R_n\left(\hat{M}_0^{GT}\right)(1\pm o(1)) = \frac{1}{n}(1\pm o(1))$$

This result motivates us to look at the Good-Turing estimator in greater detail. In the next chapter, we shall analyze the Good-Turing estimator under the multinomial sampling model as well.

CHAPTER 4

MULTINOMIAL ANALYSIS OF THE GOOD-TURING ESTIMATOR

In the last chapter, we analyzed the class of linear estimators under the Poisson sampling model and showed the optimality of the Good-Turing estimator within that class. We now analyze the Good-Turing estimator under the multinomial sampling model. The main difficulty in doing this is highlighted in section 3.2, which is that the symbol counts $N_u\left(X^n\right)$ are not independent. This requires modelling the joint distribution of $N_u\left(X^n\right)$ and $N_v\left(X^n\right)$ for all pairs of symbols u, v. We first reduce the risk to a form which simplifies analysis; this can be seen in Theorem 4.1.3. Using this, we derive upper and lower bounds of $0.6179/n$ and $0.6080/n$ respectively for the worst-case risk of the Good-Turing estimator, which is summarized in Theorem 4.4.1.

4.1 Risk of the Good-Turing Estimator

The analysis of McAllester and Schapire (2000) can be extended to characterize the risk of the Good-Turing estimator for missing mass. The risk of the Good-

Turing estimator $\hat{M}_0^{GT}(X^n)$ for a distribution p can be written down as follows:

$$\begin{aligned} R_n\left(\hat{M}_0^{GT}, p\right) &= \mathbb{E}\left[\left(\hat{M}_0^{GT}(X^n) - M_0(X^n, p)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{u \in \mathcal{X}} \frac{1}{n}\mathbb{I}(N_u = 1) - p(u)\mathbb{I}(N_u = 0)\right)\right. \\ &\quad \left.\cdot \left(\sum_{v \in \mathcal{X}} \frac{1}{n}\mathbb{I}(N_v = 1) - p(v)\mathbb{I}(N_v = 0)\right)\right] \\ &= \frac{1}{n^2}\sum_{u,v \in \mathcal{X}} \big(P_n(1,1) - 2np(u)P_n(0,1) \\ &\qquad +n^2 p(u)p(v)P_n(0,0)\big) \end{aligned} \tag{4.1}$$

where, in the final step, we use the fact that $\mathbb{E}\left(\mathbb{I}\left(X\right)\right) = \mathbb{P}\left(X\right)$ and define $P_n(i,j) \triangleq \mathbb{P}\left(N_u(X^n) = i, N_v(X^n) = j\right)$.

The probability $P_n(i,j)$ can be written down as:

$$P_n(i,j) = \begin{cases} \binom{n}{i\ j} p(u)^i p(v)^j (1 - p(u) - p(v))^{n-i-j}, \ u \neq v, \\ \\ \binom{n}{i} p(u)^i (1 - p(u)^{n-i}, \ u = v, i = j \end{cases} \tag{4.2}$$

where $\binom{n}{i\ j} = \frac{n!}{i!j!(n-i-j)!}$ and $\binom{n}{i} = \frac{n!}{i!(n-i)!}$. We split the summation in (4.1) into two cases: $u \neq v$ and $u = v$. Denoting $P(u,v) = p(u)p(v)(1 - p(u) - p(v))^{n-2}$, we have, for $u \neq v$:

$$\begin{aligned} p(u)p(v)P_n(0,0) &= (1 - p(u) - p(v))^2 P(u,v), \\ p(u)P_n(0,1) &= n(1 - p(u) - p(v))P(u,v), \\ P_n(1,1) &= n(n-1)P(u,v). \end{aligned}$$

For $u = v$, observe that $P_n(0,1) = 0$. Using the above observations, the summation in (4.1) simplifies to

$$\begin{aligned} R_n\left(\hat{M}_0^{GT}, p\right) &= \frac{1}{n}\sum_{\substack{u,v\in\mathcal{X}\\ v\neq u}} P(u,v)\Big[n\big(p(u)+p(v)\big)^2 - 1\Big] \\ &\quad + \frac{1}{n}\sum_{u\in\mathcal{X}}\Big[p(u)(1-p(u))^{n-1} + np(u)^2(1-p(u))^n\Big]. \end{aligned} \quad (4.3)$$

The following lemma is useful in bounding certain terms in the first summation above as a function of n, independent of the unknowns $\mathcal{X}$ and p.

Lemma 4.1.1. *For $i \geq 1$, $j \geq 1$,*

$$\sum_{u,v\in\mathcal{X}, u\neq v} p(u)^i p(v)^j (1-p(u)-p(v))^n \leq \frac{(i-1)!(j-1)!n!}{(n+i+j-2)!}.$$

Proof. Let X and Y be a pair of independent and identical random variables with marginal distribution p. Define a random variable $T(X,Y)$, whose value $T(u,v) = 0$ for $u = v$ and, for $u \neq v$,

$$T(u,v) = \binom{n+i+j-2}{i-1\;\; j-1} p(u)^{i-1} p(v)^{j-1} (1-p(u)-p(v))^n.$$

We see that $T(X,Y)$ is a probability for $X \neq Y$, and that it takes values in $[0,1]$ in all cases. Therefore, its expectation:

$$\begin{aligned} \mathbb{E}\left[T\left(X,Y\right)\right] &= \sum_{\substack{u,v\in\mathcal{X}\\ u\neq v}} p(u)p(v)T(u,v) \\ &= \sum_{\substack{u,v\in\mathcal{X}\\ u\neq v}} \binom{n+i+j-2}{i-1\;\; j-1} p(u)^i p(v)^j (1-p(u)-p(v))^n \\ &\leq 1 \end{aligned}$$

and the lemma follows. $\square$

A univariate version of (4.1.1) is useful as well:

Lemma 4.1.2. *For $i \geq 1$,*

$$\sum_{u \in \mathcal{X}} p(u)^i (1-p(u))^n \quad \leq \quad \frac{(i-1)!n!}{(n+i-1)!}.$$

Proof. For $X \sim p$, define $T(X) = \binom{n+i-1}{i-1} p(X)^{i-1}(1-p(X))^n$ and follow the proof of Lemma 4.1.1. $\square$

Using Lemma 4.1.1, observe that

$$\sum_{u,v \in \mathcal{X}, u \neq v} P(u,v)(p(u)+p(v))^2 \quad = \quad o(1/n) \tag{4.4}$$

Therefore, the risk can be written as

$$\begin{aligned} R_n\left(\hat{M}_0^{GT}, p\right) \quad &= \quad \frac{1}{n}\Bigg[\sum_{u \in \mathcal{X}} p(u)(1-p(u))^{n-1} - \sum_{\substack{u,v \in \mathcal{X} \\ v \neq u}} P(u,v) \\ &\quad + \sum_{u \in \mathcal{X}} np(u)^2(1-p(u))^n\Bigg] + o(1/n). \end{aligned} \tag{4.5}$$

The summation terms above can be rewritten as follows:

$$\begin{aligned}\sum_{u\in\mathcal{X}} p(u)(1-p(u))^{n-1} &= \mathbb{E}\bigg[\frac{\Phi_1(X^n)}{n}\bigg]. && (4.6)\\ \sum_{u\in\mathcal{X}} np(u)^2(1-p(u))^{n} &= \frac{2}{n-1}\sum_{u\in\mathcal{X}} P_n(2,0)(1-p(u))^2 \\ &\overset{(a)}{=} \frac{2}{n-1}\sum_{u\in\mathcal{X}} P_n(2,0) \pm o\left(\frac{1}{n}\right) \\ &= \mathbb{E}\left[\frac{2\Phi_2(X^n)}{n}\right] \pm o\left(\frac{1}{n}\right)s, && (4.7)\end{aligned}$$

where (a) follows using Lemma 4.1.2.

$$\begin{aligned}\sum_{\substack{u,v\in\mathcal{X}\\ v\neq u}} P(u,v) &= \frac{1}{n(n-1)}\sum_{\substack{u,v\in\mathcal{X}\\ v\neq u}} P_n(1,1) \\ &= \frac{1}{n(n-1)}\mathbb{E}\bigg[\sum_{\substack{u,v\in\mathcal{X}\\ v\neq u}} \mathbb{I}(N_u(X^n)=1)\mathbb{I}(N_v(X^n)=1)\bigg] \\ &= \mathbb{E}\bigg[\frac{1}{n(n-1)}\Phi_1(X^n)(\Phi_1(X^n)-1)\bigg] \\ &= \mathbb{E}\left[\frac{\Phi_1^2(X^n)}{n}\right] \pm o(1). && (4.8)\end{aligned}$$

Using the above expressions in (4.5), we get the following characterization of the risk.

Theorem 4.1.3. *The risk of the Good-Turing estimator under the multinomial sampling model satisfies:*

$$R_n\left(\hat{M}_0^{GT}, p\right) = \frac{1}{n}\mathbb{E}\left[\frac{2\Phi_2}{n} + \frac{\Phi_1}{n}\left(1-\frac{\Phi_1}{n}\right)\right] + o\left(\frac{1}{n}\right). \quad (4.9)$$

Note that the RHS in theorem 4.1.3 depends on the distribution p, since the expectations of Φ_1 and Φ_2 depend on p. This form, however, enables us to find bounds independent of p, which yields bounds on the worst-case risk $R_n\left(\hat{M}_0^{GT}\right)$.

4.2 Upper Bound on the Worst-Case Risk

To obtain an upper bound on the risk, we start with the following upper bound on one of the terms in (4.5):

$$\begin{aligned}\sum_{u\in\mathcal{X}} np(u)^2(1-p(u))^n &\leq \sum_{u\in\mathcal{X}} p(u)\left(np(u)e^{-np(u)}\right)\\ &\leq e^{-1}, \end{aligned} \tag{4.10}$$

where the first step follows because $1-x \leq e^{-x}$ for a fraction x, and the second step follows because $te^{-t} \leq e^{-1}$ for $t \geq 0$. Additionally, we have:

$$\begin{aligned}\mathbb{E}\left[\frac{\Phi_1}{n}\left(1-\frac{\Phi_1}{n}\right)\right] &\leq \mathbb{E}\left[\frac{1}{4}\right]\\ &= \frac{1}{4} \end{aligned} \tag{4.11}$$

Equation 4.11 follows from the fact that $x\,(1-x) \leq 1/4$ if $x \leq 1$. From equations 4.7, 4.9, 4.10 and 4.11, we obtain an upper bound on the worst-case risk of the Good-Turing estimator:

$$\begin{aligned}R_n\left(\hat{M}_0^{GT}\right) &= \max_p R_n\left(\hat{M}_0^{GT}, p\right)\\ &\leq \max_p \frac{e^{-1}+0.25}{n} \pm o\left(\frac{1}{n}\right)\end{aligned}$$

$$\therefore R_n\left(\hat{M}_0^{GT}\right) \leq \frac{0.25+e^{-1}}{n} \pm o\left(\frac{1}{n}\right) \tag{4.12}$$

The above constant $e^{-1} + 0.25 \approx 0.6179$ is not best possible as the bounds above are not tight. However, we show that the improvement is not significant through a lower bound on the worst-case risk in the next section.

4.3 Lower Bound on the Worst-Case Risk

A lower bound can be obtained for the worst case risk of the Good-Turing estimator by evaluating the risk for the uniform distribution p^U on $\mathcal{X}$. Let $|\mathcal{X}| = cn$ and $p^U(x) = \frac{1}{cn}$ for all $x \in \mathcal{X}$, where c is a positive constant. Using (4.5), we get

$$\begin{aligned} R_n\left(\hat{M}_0^{GT}, p^U\right) &= \frac{1}{n}\left[\frac{cn \cdot n}{(cn)^2}\left(1-\frac{1}{cn}\right)^n + \frac{cn}{cn}\cdot\left(1-\frac{1}{cn}\right)^{n-1}\right. \\ &\quad \left. -\left(\frac{cn}{cn}\cdot\left(1-\frac{1}{cn}\right)^{n-1}\right)^2\right] + o\left(\frac{1}{n}\right) \\ &\overset{(1)}{=} \frac{1}{n}\left(\left(\frac{1}{c}+1\right)\left(1-\frac{1}{cn}\right)^n - \left(1-\frac{1}{cn}\right)^{2n}\right) + o\left(\frac{1}{n}\right) \\ &\overset{(2)}{=} \frac{1}{n}\left(\left(\frac{1}{c}+1\right)e^{-\frac{1}{c}} - e^{-\frac{2}{c}}\right) + o\left(\frac{1}{n}\right) \end{aligned} \tag{4.13}$$

where the reasoning for the steps is as follows:

1. replacing $\left(1-\frac{1}{cn}\right)^{n-1}$ with $\left(1-\frac{1}{cn}\right)^n(1+o(1))$.
2. using the fact that $\left(1-\frac{1}{cn}\right)^n = e^{-1/c}(1+o(1))$.

The coefficient of $\frac{1}{n}$ in (4.13) is maximized at $c = \frac{1}{W(2)} \approx 1.1729$ to attain a maximum value of around 0.6080, where $W(\cdot)$ is the Lambert-W function. From this, we obtain a lower bound on the worst-case risk:

$$R_n\left(\hat{M}_0^{GT}\right) \geq \frac{0.6080}{n} \pm o\left(\frac{1}{n}\right) \tag{4.14}$$

4.4 Summary

We summarize the results of this chapter in this section. From (4.12) and (4.14), we have:

Theorem 4.4.1. *The worst-case risk of the Good-Turing estimator under the multinomial model satisfies the following bounds:*

$$\frac{0.6080}{n} \pm o\left(\frac{1}{n}\right) \leq R_n\left(\hat{M}_0^{GT}\right) \leq \frac{0.6179}{n} \pm o\left(\frac{1}{n}\right). \tag{4.15}$$

Therefore, the constant in equation (4.12) is fairly tight.

CHAPTER 5

LOWER BOUNDS ON MINIMAX RISK

In this chapter, we look at lower bounds on the minimax risk of missing mass estimation under the multinomial model. We explore two techniques to obtain bounds. The first uses the Bayes risk for missing mass with a chosen prior distribution as a lower bound to the minimax risk. Using a Dirichlet prior, an asymptotic lower bound of $4/27n$ is obtained. The second method reduces the problem of minimax estimation of missing mass to a distribution estimation problem, for which there are existing results on minimax risk. This method yields an asymptotic lower bound of $1/4n$.

5.1 Lower Bounds using Bayes Risk

Computing the minimax risk for missing mass estimation involves solving two optimization problems - a maximization over p for every $\hat{M}_0$ and then a minimization of the worst case risk over $\hat{M}_0$. One approach to simplifying this problem is to replace the maximization over p with an *expectation* over a family of distributions $\mathcal{P}$ using a prior distribution P over $\mathcal{P}$ - which results in the **Bayes Risk** for the prior P. This results in the minimization over $\hat{M}_0$ also becoming easier, as we can see in the following lemma:

Lemma 5.1.1. *Let P be a random variable over a family of distributions $\mathcal{P}$, having an alphabet $\mathcal{X} = \{1, 2, \dots k\}$. Then, we have the following lower bound for* *the minimax risk of missing mass:*

$$R_n^* \;\geq\; \mathbb{E}_{X^n}\left[\mathit{var}_{P|X^n}\left[M_0\left(X^n, P\right)\middle|\, X^n\right]\right] \tag{5.1}$$

Proof.

$$\begin{aligned}
R_n^* &= \min_{\hat{M}_0}\max_{p}\mathbb{E}_{X^n}\left[\left(M_0\left(X^n,p\right)-\hat{M}_0\left(X^n\right)\right)^2\right] \\
&\geq \min_{\hat{M}_0}\max_{p\in\mathcal{P}}\mathbb{E}_{X^n}\left[\left(M_0\left(X^n,p\right)-\hat{M}_0\left(X^n\right)\right)^2\right] \\
&\geq \min_{\hat{M}_0}\mathbb{E}_{P}\left[\mathbb{E}_{X^n|P}\left[\left(M_0\left(X^n,P\right)-\hat{M}_0\left(X^n\right)\right)^2\middle|\, P\right]\right] \\
&= \min_{\hat{M}_0}\mathbb{E}_{X^n}\left[\mathbb{E}_{P|X^n}\left[\left(M_0\left(X^n,P\right)-\hat{M}_0\left(X^n\right)\right)^2\middle|\, X^n\right]\right] &(5.2)\\
&= \mathbb{E}_{X^n}\left[\mathbb{E}_{P|X^n}\left[\left(M_0\left(X^n,P\right)-\mathbb{E}_{P|X^n}\left[M_0\left(X^n,P\right)\middle|\, X^n\right]\right)^2\middle|\, X^n\right]\right] &(5.3)\\
&= \mathbb{E}_{X^n}\left[\mathrm{var}_{P|X^n}\left[M_0\left(X^n,P\right)\middle|\, X^n\right]\right] &(5.4)
\end{aligned}$$

Here, 5.2 follows from the law of total expectation and 5.3 follows from the fact that $\mathbb{E}_{P|X^n}\left[\left(M_0\left(X^n,P\right)-\hat{M}_0\left(X^n\right)\right)^2\middle|\, X^n\right]$ is minimized when $\hat{M}_0\left(X^n\right) = \mathbb{E}_{P|X^n}\left[M_0\left(X^n,P\right)\middle|\, X^n\right]$ for every X^n. □

As it is presented, Lemma 5.1.1 can be used to obtain lower bounds for *any* minimax estimation problem. To obtain bounds specific to the missing mass estimation problem, we evaluate the conditional variance:

$$\begin{aligned}
R_n^* &\geq \mathbb{E}_{X^n}\left[\mathrm{var}_{P|X^n}\left[M_0\left(X^n,P\right)\middle|\, X^n\right]\right] \\
&= \mathbb{E}_{X^n}\left[\mathrm{var}_{P|X^n}\left[\sum_{u\in\mathcal{X}} P\left(u\right)\mathbb{I}\left(N_u\left(X^n\right)=0\right)\middle|\, X^n\right]\right] \\
&= \sum_{u,v\in\mathcal{X}}\mathbb{E}_{X^n}\left[\mathbb{I}\left(N_u\left(X^n\right)=0, N_v\left(X^n\right)=0\right)\right. \\
&\qquad\left.\cdot\mathrm{cov}_{P|X^n}\left[P\left(u\right), P\left(v\right)\middle|\, X^n\right]\right] &(5.5)
\end{aligned}$$

Equation 5.5 gives us a family of lower bounds for the minimax risk for missing mass estimation, depending on the choice of the prior distribution P. In the succeeding section, we will evaluate it for specific prior distributions.

5.1.1 Bayes Risk for a Dirichlet Prior

This section requires some background on Dirichlet and related probability distributions, an exposition can be found in Ng *et al.* (2011). Under the multinomial sampling model, each sample in X^n follows a categorical distribution. A natural choice for a prior would be to have P follow the *conjugate prior* to the categorical distribution, which is the **Dirichlet distribution**.

Specifically, let P follow a Dirichlet distribution $\mathrm{Dir}\,(k, \boldsymbol{\alpha})$, where k is the size of the underlying alphabet $\mathcal{X}$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_k)$ is a vector of hyperparameters. The likelihood $X_i|P$ follows a categorical distribution $\mathrm{Cat}\,(k, P)$. Thus, the posterior distribution $P|X^n$ will be a $\mathrm{Dir}\,(k, \boldsymbol{\alpha} + \boldsymbol{N}\,(X^n))$ distribution, where $\boldsymbol{N}\,(X^n) \triangleq (N_1\,(X^n)\,, N_2\,(X^n)\,, \ldots N_k\,(X^n))$ is the vector of symbol counts. Thus, the conditional covariance $\mathrm{cov}_{P|X^n}\,[P(u), P(v)|\, X^n]$ can be evaluated using the expression for the covariance of a Dirichlet distribution:

$$\begin{aligned} \mathrm{cov}_{P|X^n}\,[P(u), P(v)|\, X^n] \;=\; & \frac{1}{(\alpha_0 + n)^2\,(\alpha_0 + n + 1)} \cdot \\ & [\mathbb{I}\,(u = v)\,(\alpha_u + N_u\,(X^n))\,(\alpha_0 + n) \\ & - (\alpha_u + N_u\,(X^n))\,(\alpha_v + N_v\,(X^n))] \end{aligned} \tag{5.6}$$

Here, $\alpha_0 \triangleq \sum_{i=1}^{k} \alpha_i$. Equation 5.6 also makes use of the fact that $\sum_{i=1}^{k} N_i\,(X^n) =$

n. Using this in 5.5:

$$\begin{aligned} E_{X^n}\left[\mathrm{var}_{P|X^n}\left[M_0\left(X^n,P\right)\middle|\,X^n\right]\right] &= E_{X^n}\left[\sum_{u,v\in\mathcal{X}} \mathbb{I}\left(N_u\left(X^n\right)=0,N_v\left(X^n\right)=0\right)\cdot \right.\\ &\qquad \left.\frac{\mathbb{I}\left(u=v\right)\alpha_u\left(\alpha_0+n\right)-\alpha_u\alpha_v}{\left(\alpha_0+n\right)^2\left(\alpha_0+n+1\right)}\right] \qquad (5.7)\\ &= \sum_{u,v\in\mathcal{X}} \mathbb{P}\left(N_u\left(X^n\right)=0,N_v\left(X^n\right)=0\right)\cdot\\ &\qquad \frac{\mathbb{I}\left(u=v\right)\alpha_u\left(\alpha_0+n\right)-\alpha_u\alpha_v}{\left(\alpha_0+n\right)^2\left(\alpha_0+n+1\right)} \qquad (5.8)\end{aligned}$$

Equation 5.7 follows from the fact that $\mathbb{I}\left(N_u\left(X^n\right)=0\right)N_u\left(X^n\right)=0$. The joint distribution $\mathbb{P}\left(N_u\left(X^n\right)=0,N_v\left(X^n\right)=0\right)$ and the marginal distribution $\mathbb{P}\left(N_u\left(X^n\right)=0\right)$ (in the case $u=v$) are required to evaluate the RHS in 5.8. The conditional distribution $\boldsymbol{N}\left(X^n\right)|P$ is a multinomial distribution $\mathrm{Mult}\left(n,P\right)$. Since $P\sim\mathrm{Dir}\left(k,\boldsymbol{\alpha}\right)$, the marginal distribution of $\boldsymbol{N}\left(X^n\right)$ will be a Dirichlet-Multinomial distribution $\mathrm{DirMult}\left(n,\boldsymbol{\alpha}\right)$. From this, we can obtain the required joint and marginal distributions. In particular, $\left(N_u\left(X^n\right),n-N_u\left(X^n\right)\right)$ follows a $\mathrm{DirMult}\left(n,\alpha_u,\alpha_0-\alpha_u\right)$ distribution, while $\left(N_u\left(X^n\right),N_v\left(X^n\right),n-N_u\left(X^n\right)-N_v\left(X^n\right)\right)$ has a $\mathrm{DirMult}\left(n,\alpha_u,\alpha_v,\alpha_0\right)$ distribution. So, from the PMF of the Dirichlet-Multinomial distribution, we have:

$$\mathbb{P}\left(N_u\left(X^n\right)=0,N_v\left(X^n\right)=0\right) = \begin{cases} \frac{B(\alpha_0,n)}{B(\alpha_0-\alpha_u,n)} & \text{if } u=v\\ \frac{B(\alpha_0,n)}{B(\alpha_0-\alpha_u-\alpha_v,n)} & \text{if } u\neq v\end{cases} \qquad (5.9)$$

Here, $B(\cdot,\cdot)$ is the Beta function. Using this in 5.8:

$$\begin{aligned}
E_{X^n}\left[\mathrm{var}_{P|X^n}\left[M_0(X^n,P)|\,X^n\right]\right] &= \sum_{u=v}\mathbb{P}\left(N_u(X^n)=0\right)\frac{\alpha_u(\alpha_0+n)-\alpha_u^2}{(\alpha_0+n)^2(\alpha_0+n+1)} \\
&\quad -\sum_{u\neq v}\mathbb{P}\left(N_u(X^n)=0,N_v(X^n)=0\right)\cdot \\
&\quad \frac{\alpha_u\alpha_v}{(\alpha_0+n)^2(\alpha_0+n+1)} \qquad (5.10)\\
&= \sum_u\frac{B(\alpha_0,n)}{B(\alpha_0-\alpha_u,n)}\cdot\frac{\alpha_u(\alpha_0+n)-\alpha_u^2}{(\alpha_0+n)^2(\alpha_0+n+1)} \\
&\quad -\sum_{u\neq v}\frac{B(\alpha_0,n)}{B(\alpha_0-\alpha_u-\alpha_v,n)}\cdot \\
&\quad \frac{\alpha_u\alpha_v}{(\alpha_0+n)^2(\alpha_0+n+1)} \qquad (5.11)
\end{aligned}$$

Thus, from equations 5.5 and 5.11, we have:

$$\begin{aligned}
R_n^* &\geq \frac{B(\alpha_0,n)}{(\alpha_0+n)^2(\alpha_0+n+1)}\left(\sum_u\frac{\alpha_u(\alpha_0+n)-\alpha_u^2}{B(\alpha_0-\alpha_u,n)}\right. \\
&\quad \left.-\sum_{u\neq v}\frac{\alpha_u\alpha_v}{B(\alpha_0-\alpha_u-\alpha_v,n)}\right) \qquad (5.12)
\end{aligned}$$

The expression in equation 5.12 provides a family of lower bounds for the minimax MSE - one for each choice of the parameters k and $\boldsymbol{\alpha}$, which may both depend on n in turn.

Let $\boldsymbol{\alpha}=\left(\frac{1}{n},\frac{1}{n},\frac{1}{n},\ldots,\frac{1}{n}\right)$ and $k=cn^2$, where c is a positive constant. For this choice of parameters, $\alpha_0 = cn$. Using these parameters:

$$\begin{aligned}
R_n^* &\geq \frac{B(cn, n)}{(cn+n)^2(cn+n+1)}\left(\left(cn^2-1\right)\cdot\frac{\frac{1}{n}\cdot(cn+n)-\frac{1}{n^2}}{B\left(cn-\frac{1}{n}, n\right)}\right. \\
&\quad \left. -cn^2\left(cn^2-1\right)\cdot\frac{\frac{1}{n}\cdot\frac{1}{n}}{B\left(cn-\frac{2}{n}, n\right)}\right) \\
&\gtrsim \frac{1}{n}\cdot\frac{1}{(c+1)^3}\left(c(c+1)\cdot\frac{B(cn,n)}{B\left(cn-\frac{1}{n}, n\right)}\right. \\
&\quad \left. -c^2\cdot\frac{B(cn,n)}{B\left(cn-\frac{2}{n}, n\right)}\right) \\
&= \frac{1}{n}\cdot\frac{1}{(c+1)^3}\left(c(c+1)\cdot\frac{\Gamma(cn)}{\Gamma\left(cn-\frac{1}{n}\right)}\cdot\frac{\Gamma\left((c+1)n-\frac{1}{n}\right)}{\Gamma((c+1)n)}\right. \\
&\quad \left. -c^2\cdot\frac{\Gamma(cn)}{\Gamma\left(cn-\frac{2}{n}\right)}\cdot\frac{\Gamma\left((c+1)n-\frac{2}{n}\right)}{\Gamma((c+1)n)}\right) \\
&\gtrsim \frac{1}{n}\cdot\frac{1}{(c+1)^3}\left(c(c+1)\left(\frac{cn}{(c+1)n}\right)^{\frac{1}{n}} - c^2\left(\frac{cn}{(c+1)n}\right)^{\frac{2}{n}}\right) \\
&\gtrsim \frac{1}{n}\cdot\frac{1}{(c+1)^3}\left(c(c+1)-c^2\right) \\
&= \frac{1}{n}\cdot\frac{c}{(c+1)^3} \qquad (5.13)
\end{aligned}$$

where $x \gtrsim y$ means $x \geq y \pm o(y)$. The coefficient of $\frac{1}{n}$ attains a maximum value of $\frac{4}{27}$ when $c = \frac{1}{2}$. Thus, we have:

$$R_n^* \geq \frac{4}{27n} \pm o\left(\frac{1}{n}\right) \qquad (5.14)$$

5.2 Lower Bound using Distribution Estimation

A second approach to bound the minimax risk for missing mass estimation is to reduce the problem to that of estimating a distribution. Let $\mathcal{P}$ be the set of distributions over the set $\mathcal{X} = \{0,1\}$ such that for all $p \in \mathcal{P}$, $p\left(0\right) \geq \frac{1}{2}$. A known result (refer Lehmann and Casella (1998) for instance) states that the minimax ℓ^2 risk in estimating $p(0)$ is $\frac{1}{4n}$. More precisely, let $\hat{p}(X^n)$ be an estimator for $p(0)$ from a random sample X^n distributed according to p. Then, we have:

Lemma 5.2.1.

$$\min_{\hat{p}(0)} \max_{p \in \mathcal{P}} \mathbb{E}_{X^n \sim p} \left(p\left(0\right) - \hat{p}\left(X^n\right)\right)^2 \quad = \quad \frac{1}{4n} + o\left(\frac{1}{n}\right)$$

For an arbitrary positive integer k, let $\mathcal{P}_c$ be the set of distributions over the set $\mathcal{X} = \{0,1,2,\ldots k-1\}$, such that for any $p_c \in \mathcal{P}_c$, we have $p_c\left(0\right) \geq \frac{1}{2}$ and $p_c\left(i\right) = \frac{1-p_c(0)}{k}$ for all $i \geq 1$. We can use Lemma 5.2.1 to obtain minimax bounds in estimating $p_c\left(0\right)$ for this family of distributions as well. Let $\hat{p}_c(X^n)$ be an estimator for p_c from a random sample X^n distributed according to p_c. Let $\hat{p}_c(X^n, i)$ be the probability $\hat{p}_c$ assigns to the symbol i.

Lemma 5.2.2.

$$\min_{\hat{p}_c(0)} \max_{p_c \in \mathcal{P}_c} \mathbb{E}_{X^n \sim p_c} \left(p_c\left(0\right) - \hat{p}_c\left(X^n, 0\right)\right)^2 \quad \geq \quad \frac{1}{4n} + o\left(\frac{1}{n}\right)$$

Proof. Suppose we want to estimate an unknown distribution $p \in P$ and we have an estimator $\hat{p}_c$ for distributions in $\mathcal{P}_c$. Then we can use $\hat{p}_c$ to estimate p as follows. Take the observed sample distributed according to p, and if it is 0, keep it unchanged. If it is 1, then replace it with an uniformly sampled random variable over $\{1,2,\ldots k\}$. The result of this sampling process is a distribution p_c in $\mathcal{P}_c$

with $p_c(0) = p(0)$. Thus, any estimator for distributions in $\mathcal{P}_c$ can be reduced to an estimator for distributions in $\mathcal{P}$ and

$$\min_{\hat{p}(0)} \max_{p \in \mathcal{P}_c} \mathbb{E}_{X^n \sim p_c} \left(p_c(0) - \hat{p}_c(X^n, 0)\right)^2 \\ \geq \min_{\hat{p}(0)} \max_{p \in \mathcal{P}} \mathbb{E}_{X^n \sim p} \left(p(0) - \hat{p}(X^n)\right)^2$$

and the proof follows from Lemma 5.2.1. □

Lemma 5.2.3. *Let $k = e^n$. With probability at least $1 - 1/2^n$, the missing mass $M_0(X^n)$ satisfies*

$$M_0(X^n) = 1 - p(0) + O\left(ne^{-n}\right).$$

Proof. The probability of symbol 0 appearing at least once in X^n is $1 - (1 - p(0))^n \geq 1 - 1/2^n$. Furthermore, at most n distinct symbols from $1, 2, \ldots k-1$ can appear in X^n. Hence, with probability $1 - 1/2^n$, the observed mass $1 - M_0(X^n)$ satisfies

$$p(0) \leq 1 - M_0(X^n) \leq p(0) + (1 - p(0))\, ne^{-n}, \tag{5.15}$$

and hence follows the lemma. □

From Lemmas 5.2.2 and 5.2.3, we can obtain a lower bound on the minimax risk of missing mass estimation:

$$\begin{aligned} R_n^* &= \min_{\hat{M}_0} \max_{p} \mathbb{E}\left(M_0(X_n, p) - \hat{M}_0(X_n)\right)^2 \\ &\geq \min_{\hat{M}_0} \max_{p \in \mathcal{P}_c} \mathbb{E}\left(M_0(X_n, p) - \hat{M}_0(X_n)\right)^2 \\ &\geq \min_{\hat{M}_0} \max_{p \in \mathcal{P}_c} \left(1 - \frac{1}{2^n}\right) \mathbb{E}\left[p(0) - \left(1 - \hat{M}_0(X_n)\right) - O\left(ne^{-n}\right)\right]^2 \\ &= \frac{1}{4n} + o\left(\frac{1}{n}\right) \end{aligned}$$

$$\therefore R_n^* \quad \geq \quad \frac{1}{4n} + o\left(\frac{1}{n}\right) \tag{5.16}$$

This bound improves upon the $4/27n$ lower bound obtained in the previous section, at the cost of somewhat lower generalizability.

SUMMARY AND FUTURE WORK

We looked at the problem of missing mass estimation and the Good-Turing estimator. The Good-Turing estimator was shown to have a worst case risk of $1/n$ under the Poisson sampling model and was proven to be asymptotically optimal among all linear estimators. Under the multinomial sampling model, the Good Turing estimator was shown to have a worst case risk between $0.6080/n$ and $0.6179/n$. We also looked at multinomial lower bounds for the minimax risk; a bound of $4/27n$ is obtained using Bayes risk and can be improved to $0.25/n$ using distribution estimation. Combined with the earlier result on the Good-Turing estimator, we give the following guarantee on the minimax risk of missing mass estimation:

$$\frac{0.25}{n} \leq R_n^* \leq \frac{0.6179}{n}$$

Thus, the Good-Turing estimator is order-optimal for missing mass estimation under the multinomial model.

The following are some suggestions for further directions this work can be taken in:

1. While the Good-Turing estimator was shown to be optimal among linear estimators under the Poisson model, it is possible that it can be outperformed by another linear estimator under the multinomial model. The analysis of chapter 3 could possibly be extended to give bounds on the worst case risk for arbitrary linear estimators.
2. The Bayes risk bound could also be used to obtain lower bounds for the minimax risk under the Poisson model.

3. The lower bound on minimax risk using Bayes risk could be improved, either by choosing better hyperparameters(k and $\boldsymbol{\alpha}$) or by using a different prior distribution altogether.

4. Similar techniques could be used to obtain minimax results for other estimation problems. This includes distribution estimation and estimating M_k, which is the total probability of all symbols seen k times in the sample.

REFERENCES

1. **Ben-Hamou, A.**, **S. Boucheron**, , and **M. I. Ohannessian** (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, **23**(1), 249–287.

2. **Berend, D.** and **A. Kontorovich** (2013). On the concentration of the missing mass. *Electronic Communications in Probability*, **18**(3).

3. **Chao, A.** and **S.-M. Lee** (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, **87**(417), 210–217.

4. **Chen, S. F.** and **J. Goodman** (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**, 359âĂŞ394.

5. **Church, K. W.** and **W. A. Gale** (1991). A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech & Language*, **5**(1), 19–54.

6. **Esty, W. W.** (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics*, **11**(3), 905–912.

7. **Gale, W. A.** and **G. Sampson** (1995). Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, **2**(3), 217–237.

8. **Good, I. J.** (1953). The population frequencies of species and the estimation of population parameters. *Biometricka*, **40**(3/4), 237–264.

9. **Good, I. J.** and **G. H. Toulmin** (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**(1/2), 45–63.

10. **Katz, S.** (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **35**(3), 400 – 401.

11. **Lehmann, E. L.** and **G. Casella**, *Theory of Point Estimation*, chapter 5. Springer, 1998, 311–312.

12. **McAllester, D.** and **L. Ortiz** (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, **4**, 895–911.

13. **McAllester, D. A.** and **R. E. Schapire**, On the convergence rate of good-turing estimators. *In Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, COLT '00. Morgan Kaufmann Publishers Inc., 2000.

14. **Ng, K. W.**, **G.-L. Tian**, and **M.-L. Tang**, *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley & Sons, 2011. ISBN 978-1-119-99841-9.

15. **Ohannessian, M. I.** and **M. A. Dahleh**, Rare probability estimation under regularly varying heavy tails. *In Proceedings of the 25th Annual Conference on Learning Theory*, volume 23. 2012.

16. **Orlitsky, A.** and **A. T. Suresh**, Competitive distribution estimation: Why is good-turing good. *In* **C. Cortes**, **N. D. Lawrence**, **D. D. Lee**, **M. Sugiyama**, and **R. Garnett** (eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Curran Associates, Inc., 2015, 2143–2151.

17. **Orlitsky, A.**, **A. T. Suresh**, and **Y. Wu**, Optimal prediction of the number of unseen species. *In Proceedings of the National Academy of Sciences USA*, volume 113. 2016.

18. **Shen, T.-J.**, **A. Chao**, and **C.-F. Lin** (2003). Predicting the number of new species in further taxonomic sampling. *Ecology*, **84**(3), 798–804.

19. **Sproat, R.**, **W. Gale**, **C. Shih**, and **N. Chang** (1996). A stochastic finite-state word-segmentation algorithm for chinese. *Computational Linguistics*, **22**(3), 377–404.

20. **Vu, V. Q.**, **B. Yu**, and **R. E. Kass** (2007). Coverage-adjusted entropy estimation. *Statistics in Medicine*, **26**(21).

21. **Wagner, A. B.**, **P. Viswanath**, and **S. R. Kulkarni**, A better good-turing estimator for sequence probabilities. *In IEEE International Symposium on Information Theory*. 2007.

LIST OF PAPERS BASED ON THESIS

1. **Rajaraman, N.**, **A. Thangaraj**, and **A. T. Suresh**, Minimax risk for missing mass estimation. *To appear in IEEE International Symposium on Information Theory*. 2017.