ELI UNDERGRADUATE RESEARCH REPORT

Cornell ECE Early Career Research Scholars Program
School of Electrical and Computer Engineering
Cornell University

Good-Turing Estimator and its Application to Data Analytics for Real-Time Electricity Markets

Author:
Alice Xinyun CHEN
xc278@cornell.edu

Research advisor:
Prof. Lang TONG
ltong@ece.cornell.edu

August 16, 2016

Abstract—**Getting accurate electricity price estimates is critical to electricity market participants. However, electricity prices can be highly unpredictable, especially with the rapid development of renewable energy in recent years. Estimating its probability distribution is also hard because not much data is useful to be included in the sample. This report looks at the Good-Turing estimator and one of its variants, the Simple Good-Turing (SGT) estimator, which works particularly well when estimating the probability distribution of a random event given only sparse data. A detailed (and new) proof of the central formula in the Good-Turing estimator as well as the intuition for the SGT estimator is given, and the performances of the SGT, empirical, and Laplace estimators are compared for samples of different kinds. The fact that the SGT estimator is doing better overall suggests that the Good-Turing estimator and its variants may yield relatively more accurate results when applied to real electricity prices.**

I. Introduction

Forecasting electricity prices is crucial for participants in electricity markets. Market participants use the forecast prices to adjust their bidding strategies, minimizing the risk and maximizing the profits. For example, convergence bidding is a financial mechanism that allows market participants to arbitrage price differences between the day-ahead market and the real-time market without physically consuming or producing energy. If one always gets a good estimate of the difference between the day-ahead and real-time local marginal prices (LMP), then one can make a lot of money from convergence bidding alone. However, due to the high cost of storing electricity and the unpredictability of electricity production (especially with the rapid development of renewable energy), LMP's are rather hard to predict. Currently, most LMP forecast schemes provide only point forecast that gives a single quantity as the prediction, which can be quite inaccurate for systems with high levels of uncertainty. Taking that into account, estimating the *probability distribution* of LMP's at a particular place at a particular time, rather than the LMP's themselves, can be more instructional for market participants.

Usually, in order to estimate the probability distribution of any random event, abundant data is needed to produce sufficiently accurate estimates. But in the case of LMP's, the data is rather sparse even though LMP's from as far back as 1998 are readily available online - the market has changed so much since then that only the most recent LMP's are useful. Moreover, LMP's at different times in a day and in different days of a week can have very different probability distributions and therefore cannot be estimated as a whole. Now the question becomes: given sparse data only, how do we accurately estimate the probability distribution of a random event? In other words, we need an estimator (a rule for calculating an estimate of a given quantity based on observed data) that performs well when many possible values of the quantity are not observed.

The Good-Turing estimator was first introduced by Irving J. Good in [2] to do just that. In this report, we will first introduce and define the Good-Turing estimator along with two other more intuitive and frequently used estimators, namely, the empirical estimator and the Laplace estimator, and then flesh out a proof for the central formula used in the Good-Turing estimator, which we hope is clearer, more concise, and assumes less than the one given in [2]. An easy-to-use variant of the Good-Turing estimator, namely, the Simple Good-Turing (SGT) estimator proposed by William A. Gale and Geoffrey Sampson in [1] is then introduced and tested on known distributions as well as simulated LMP's with unknown distribution. Results show that the SGT estimator is performing relatively well in all cases we have tried compared with the other two estimators we have introduced and thus worth further investigation in its applicability to data analytics for LMP's.

II. The Three Estimators

Now let's forget about all the terminologies of the electricity market for the moment and re-introduce the problem this way:

Figure 1. All that we know about the jungle - data from a sample of size 28

Suppose we want to know the population frequency of each species in a jungle. Assume that Fig. 1 is all we know about the underlying population (quite sparse data). Because of that, we will not be able tell the difference between two species that are represented the same number of times in the sample (for example, rhino and cheetah) and thus might as well assign the same population frequency to both. Naturally, the first step would be to organize the information in the picture into numbers. Counting the number of occurrences of each species in the sample (illustrated in Fig. 2), we have the following organized data:

- 1 rhino, 1 cheetah, 1 orangutan, 1 snake
- 2 gazelles, 2 ducks, 2 lions, 2 elephants, 2 giraffes, 2 zebras
- 3 parrots, 3 flamingos
- 6 butterflies

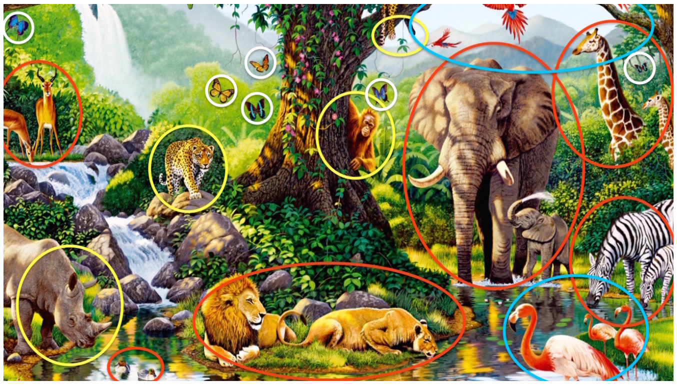

Figure 2. Counting...

Rearrange the information above to get Table I:

Table I
r-n_r PAIRS FOR THE JUNGLE SAMPLE

r	1	2	3	6
n_r	4	6	2	1

where r is the number of times a species is represented in the sample, and n_r is the number of species each represented r times in the sample. Also, since we do not know the difference between rhino and cheetah, let's denote the population frequency of both as q_1, and the empirical estimate as $\hat{q}_r^e$, the Laplace estimate as $\hat{q}_r^l$, and the Good-Turing estimate of q_1 as $\hat{q}_1^{gt}$. Generalizing from 1 to r gives notations for all population frequencies and the corresponding estimates.

Now we are ready to introduce the three estimators and see what each of them gives us as the population frequencies:

$$\hat{q}_r^e = \frac{r}{N} \tag{1}$$

$$\hat{q}_r^l = \frac{r+1}{N+s} \tag{2}$$

$$\hat{q}_r^{gt} = \begin{cases} \frac{n_1}{N}, & \text{if } r = 0 \\ \frac{(r+1)n_{r+1}}{Nn_r}, & \text{if } r > 0 \end{cases} \tag{3}$$

Note that in Formula (2), s, the number of species in the jungle, has to be known in order to apply the Laplace estimator (let's take s to be 14 for the purpose of this example - only one species is unseen), but knowledge of s is not necessary for either the empirical or the Good-Turing estimator. Moreover, in Formula (3), $\hat{q}_0^{gt}$ is the *total* probability of all unseen species, not that of *each* unseen species. By defining the Good-Turing estimator this way, we will not need to use the value of n_0 (the number of unseen species), the knowledge of which is equivalent to the knowledge of s since $s = n_0 + n_1 + \ldots$ is always true. Applying the formulas above, we get Table II:

Table II
ESTIMATED POPULATION FREQUENCIES OF THE JUNGLE SPECIES REFERENCED BY r

r	n_r	$\hat{q}_r^e$	$\hat{q}_r^l$	$\hat{q}_r^{gt}$
0	(1)	0/28	1/42	4/28[a]
1	4	1/28	2/42	3/28
2	6	2/28	3/42	1/28
3	2	3/28	4/42	0/28
5	0	0/28	0/42	1.2/28[b]
6	1	6/28	7/42	0/28

[a] This is the *total* probability for all unseen species.

[b] Since there are no species represented five times in the sample, this probability would be "lost" when translating $\hat{q}_r$'s to population frequencies of each species. We will show how to "smooth" the observed data to tackle this apparent source of inaccuracy in the fourth section.

From Table II, it is easy to see the rationale behind each estimator. The empirical estimator assigns a population frequency directly proportional to the number of occurrences of each species. It is perhaps the most intuitive and in fact closest one to truth when the sample size is large enough so that almost every possible species is represented a sufficient number of times in the sample. When given sparse data, however, it simply assigns zero probability to all unseen species, which can be a notable source of inaccuracy. The Laplace estimator tries to avoid assigning zero probability to unseen species by adding one to the number of occurrences of each species (including the unseen ones), and thus unseen species would have numbers of occurrences of 1 and therefore nonzero probabilities associated with them. The downside of the Laplace estimator, however, is that the number of species has to be known, and that the decision to add 1 to the numbers of occurrences instead of 1/2 or 2 is rather random. With the Good-Turing estimator, both problems are solved: the number of species in the sample is not needed in the estimating process, and the probability of unseen species is estimated using the number of species that are represented once in the sample divided by the sample size. A proof of Formula (3) is in the next section.

III. Proof

We will prove Formula (3) in the previous section under the assumption that there are an unknown, finite[1] number of species (denoted by s). Since the case when $r = 0$ directly comes from rearranging the formula in the second case, we only need to prove

$$\hat{q}_r^{gt} = \frac{(r+1)n_{r+1}}{Nn_r} \tag{4}$$

for all nonnegative integers r. Before diving into the details of the proof, though, let's define the problem more precisely.

In a jungle of infinitely many animals, each individual belongs to one and only one species among all s species. s is finite. Suppose that the population frequencies of all species are $p_1, p_2, \ldots, p_s$, respectively[2]. It is clear that $\sum\limits_{\mu=1}^{s} p_\mu = 1$.

Now pick a species x out of the s species equiprobably, and let M denote the random event that the species x is the μth species (and therefore having a population frequency of p_μ). Then draw a sample of size N with replacement. Count the number of times each species is represented in the sample and we get random variables $r_1, r_2, \ldots, r_s$, where r_μ is the number of times the μth species is represented in the particular sample, $r_\mu \in \{0, 1, 2, \ldots, N\}$. Let R denote the random event that the species x is represented r times in the sample. Next, count the number of species represented the same number of times and we get random variables $n_0, n_1, n_2, \ldots, n_N$, where n_r is the number of species each represented r times in the sample. It is clear that $\sum\limits_{\mu=1}^{s} r_\mu = \sum\limits_{r=1}^{N} rn_r = N$.

We want to prove that Formula (4) produces a good estimation of random variables $q_1, q_2, \ldots, q_N$, where q_r is the population frequency of an arbitrary species that is represented r times in the sample. The way we prove that is to show that

$$E(q_r) \approx \hat{q}_r^{gt} \tag{5}$$

for all nonnegative integers r. $E(q_r)$ is the expectation of q_r. Note that when using the Good-Turing estimator to estimate population frequencies, $p_1, p_2, \ldots, p_s$ are unknown and used only in intermediate deductions in the proof, and s and n_0 can be unknown. All relevant notations are summarized below for the readers' convenience:

- x: the species we choose from all s species equiprobably
- s: the number of distinct species in the entire population
- p_μ: the population frequency of the μth species with $\mu \in \{1, 2, \ldots, s\}$
- N: the sample size, i.e., the number of individuals in the sample
- r: the number of times a species is represented in the sample; $r \in \{0, 1, 2, \ldots, N\}$
- n_r: the number of species each represented r times in the sample with $r \in \{0, 1, 2, \ldots, N\}$
- q_r: the population frequency of an arbitrary species that is represented r times in the sample with $r \in \{0, 1, 2, \ldots, N\}$
- $\hat{q}_r^{gt}$: the Good-Turing estimate of q_r with $r \in \{0, 1, 2, \ldots, N\}$
- M: the random event that the chosen species x is the μth species
- R: the random event that the chosen species x is represented r times in the sample

Proof. $E(q_r)$ is the expectation of q_r, the expected population frequency of an arbitrary species that is

[1] The Good-Turing estimator should still work fine when s is infinite, though the proofs are more rigorous when it is finite.

[2] This means that the μth species has a population frequency of p_μ. Also, in [2], p_μ's are assumed to be unequal, but in our proof, we find this assumption unnecessary.

represented r times in the sample, i.e., the expected population frequency of x if x is represented r times in the sample. Since the population frequency of x must be one of $1, 2, \ldots, s$, and the probability for each case is $p_1, p_2, \ldots, p_s$, it is easy to see that

$$E(q_r) = \sum_{\mu=1}^{s} P(M|R)p_\mu \tag{6}$$

Now we need to calculate $P(M|R)$. By Bayes' Rule,

$$\begin{aligned} P(M|R) &= \frac{P(R|M)P(M)}{P(R)} && (7) \\ &= \frac{P(R|M)P(M)}{\sum\limits_{\mu=1}^{s} P(R|M)P(M)} && (8) \end{aligned}$$

where $P(R|M)$ is the probability that the μth species is represented r times in the sample, and $P(M)$ is the probability that x is the μth species. It is easy to see that:

$$\begin{aligned} P(R|M) &= \binom{N}{r} p_\mu^r (1-p_\mu)^{N-r} && (9) \\ P(M) &= \frac{1}{s} && (10) \end{aligned}$$

Plugging Equations (9) and (10) into (8), we get:

$$\begin{aligned} P(M|R) &= \frac{\binom{N}{r} p_\mu^r (1-p_\mu)^{N-r} (\frac{1}{s})}{\sum\limits_{\mu=1}^{s} \binom{N}{r} p_\mu^r (1-p_\mu)^{N-r} (\frac{1}{s})} && (11) \\ &= \frac{p_\mu^r (1-p_\mu)^{N-r}}{\sum\limits_{\mu=1}^{s} p_\mu^r (1-p_\mu)^{N-r}} && (12) \end{aligned}$$

Plugging Equation (12) back into (6), we get:

$$\begin{aligned} E(q_r) &= \sum_{\mu=1}^{s} \frac{p_\mu^r (1-p_\mu)^{N-r}}{\sum\limits_{\mu=1}^{s} p_\mu^r (1-p_\mu)^{N-r}} p_\mu && (13) \\ &= \frac{\sum\limits_{\mu=1}^{s} p_\mu^{r+1} (1-p_\mu)^{N-r}}{\sum\limits_{\mu=1}^{s} p_\mu^r (1-p_\mu)^{N-r}} && (14) \end{aligned}$$

To express Equation (14) in terms of data obtainable from the sample, let's calculate $E_N(n_r)$, namely, the expectation of the number of species each represented r times in a sample of size N:

$$\begin{aligned} E_N(n_r) &= \sum_{\mu=1}^{s} 1P(R|M) + 0(1-P(R|M)) && (15) \\ &= \sum_{\mu=1}^{s} \binom{N}{r} p_\mu^r (1-p_\mu)^{N-r} && (16) \end{aligned}$$

Replacing N with $N+1$ and r with $r+1$, we get $E_{N+1}(n_{r+1})$, namely, the expectation of the number of species each represented $r+1$ times in a sample of size $N+1$:

$$E_{N+1}(n_{r+1}) = \sum_{\mu=1}^{s} \binom{N+1}{r+1} p_\mu^{r+1} (1-p_\mu)^{N-r} \tag{17}$$

Rearranging Equations (16) and (17), we get:

$$\begin{aligned} \sum_{\mu=1}^{s} p_\mu^r (1-p_\mu)^{N-r} &= \frac{E_N(n_r)}{\binom{N}{r}} && (18) \\ \sum_{\mu=1}^{s} p_\mu^{r+1} (1-p_\mu)^{N-r} &= \frac{E_{N+1}(n_{r+1})}{\binom{N+1}{r+1}} && (19) \end{aligned}$$

Plugging Equations (18) and (19) back into (14), we get:

$$\begin{aligned} E(q_r) &= \frac{\frac{E_{N+1}(n_{r+1})}{\binom{N+1}{r+1}}}{\frac{E_N(n_r)}{\binom{N}{r}}} && (20) \\ &= \frac{\binom{N}{r} E_{N+1}(n_{r+1})}{\binom{N+1}{r+1} E_N(n_r)} && (21) \\ &= \frac{(r+1)E_{N+1}(n_{r+1})}{(N+1)E_N(n_r)} && (22) \\ &\approx \frac{(r+1)n_{r+1}}{Nn_r} && (23) \\ &= \hat{q}_r^{gt} && (24) \end{aligned}$$

Notice that Equation (24) is exactly the same as Equation (5)[3], which we set out to prove. Thus, the Good-Turing estimator is theoretically sound and does not rest on any random decisions. □

[3] The detailed rationale for the transition from (22) to (23) can be found in [2].

IV. Simple Good-Turing Estimator

As illustrated in the second section, estimates given by the Good-Turing estimator is not *good* enough for large r's since gaps in r's (where n_r's are zero) can result in “loss” of probability and zero probability assigned to species that are actually represented the most times. That is why smoothing is needed. Various smoothing techniques have been discussed by Good himself and other people, among which a simple and relatively good one is the Simple Good-Turing (SGT) estimator proposed by William A. Gale and Geoffrey Sampson in [1]. Since a step-by-step procedure of the SGT estimator can be found in [1], in this report, we will focus on giving the main intuition behind the smoothing with graph illustrations generated from our simulated LMP's (Fig. 3).

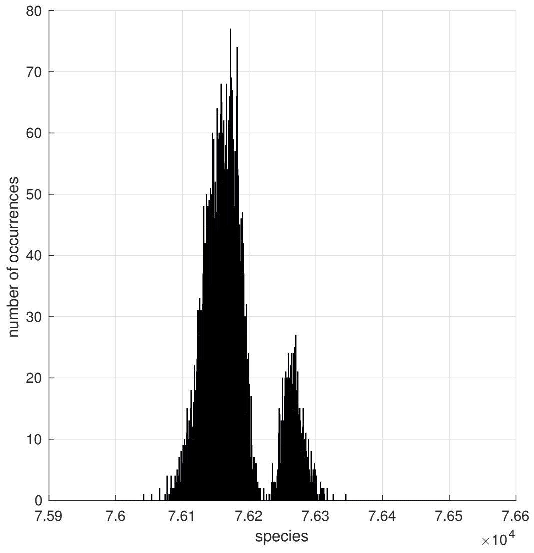

Figure 3. Number of occurrences vs. species

A random sample drawn from a large population usually contains many many species each represented once up to a few times in the sample, and a few species each represented many many times in the sample (Fig. 4 represents 5000 out of the 10000 simulated LMP's at one time at one place). As r gradually increments by 1, it is less and less likely to find any species represented exactly r times in the sample at all, and thus the n_r's would alternate between small numbers, most likely 0 and 1, until the species with the greatest number of occurrences in the sample is included.

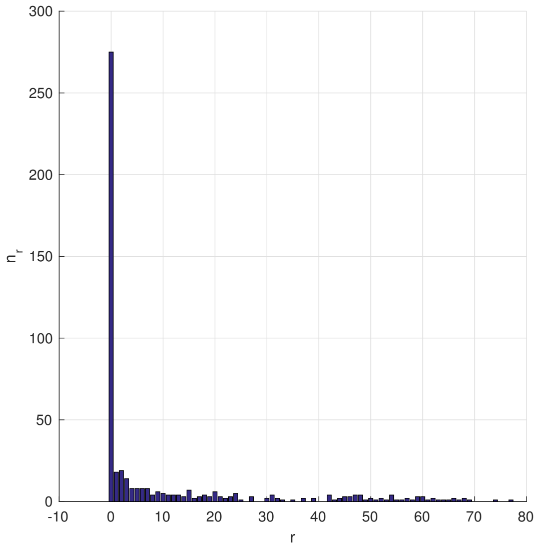

Figure 4. n_r and smoothed n_r vs. r

Plotting n_r against r on a logarithmic scale would show that nonzero n_r's for large r's have a lower limit of 1 and thus need to be rounded down so that the nonzero n_r's can get averaged with the zero n_r's after the best fitting line is found (in Fig. 5, the green dots represent n_r's before smoothing, the blue circles represent rounded-down n_r's, and the red line represents the smoothed value for all positive integers r).

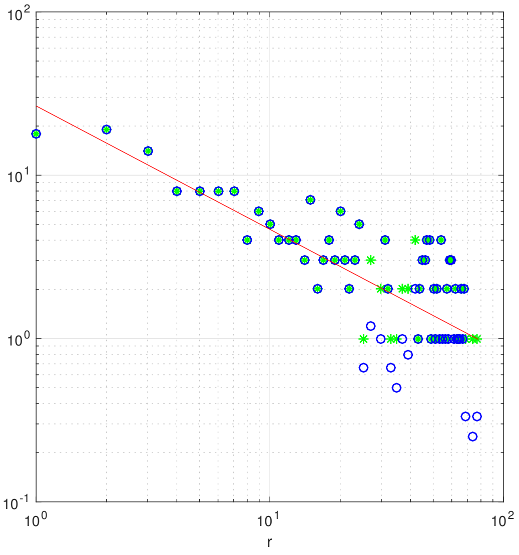

Figure 5. n_r vs. r

Apart from averaging the non-zero n_r's with the zero ones for large r's and finding the best fitting line for the averaged data, the SGT estimator also preserves the more accurate original n_r for small r's by not using the smoothed value until after the difference between the original and the smoothed is below a certain threshold[4]. All these rather simple and intuitive steps, when combined together, yield quite good estimates.

V. Results and Discussions

Since no one knows the real probability distribution of LMP's, directly applying the SGT estimator to day-ahead and real-time LMP's would not tell us much about the estimator's performance. Therefore, we investigated two ways of testing the performance of the SGT estimator against the other two estimators introduced in the second section.

One is to use the estimators on random numbers generated from certain known distributions (normal, uniform, and Poisson distributions). For each of the three different distributions, we choose four different sets of properties (e.g. mean and deviation), and for each set, we generate 100 random numbers and use the estimators to estimate the underlying distribution. We then calculate the l_1-distance (sum of absolute values of differences) between the true probabilities and the estimated ones. Finally, we compare the average l_1-distance over 1000 trials.

The other is to use the estimators on simulated LMP's whose "true" distribution is approximated by applying the empirical estimator on all 10000 LMP's at one time at one place[5], assuming that a sample size of 10000 is sufficiently large. Then similarly, we choose four different sets of distinct time and location pairs, and for each set, we partition the 10000 LMP's into 100 unordered samples of size 100 and take the average l_1-distances over the 100 trials to compare.

When the distribution is continuous, we round the numbers to the fourth digit and scale them so that they are integers and easier to work with. Results from both methods are summarized in the following tables:

Table III
l_1-Distances for Normal Distribution ($\mu = 0$)

case	σ (SD)	SGT	Empirical	Laplace
1	3	1.5177	2.2364	2.1165
2	4	1.4833	2.2178	2.0649
3	5	1.4629	2.1913	2.0122
4	6	1.4766	2.2105	2.0591

Table IV
l_1-Distances for Uniform Distribution

case	N (max)	SGT	Empirical	Laplace
1	50	0.4469	0.5370	0.3580
2	100	0.3466	0.7314	0.3657
3	150	0.3031	1.0250	0.4100
4	200	0.2690	1.2111	0.4037

Table V
l_1-Distances for Poisson Distribution

case	λ (mean)	SGT	Empirical	Laplace
1	50	0.4414	0.4503	0.4581
2	150	0.6489	0.5944	0.8464
3	350	0.8324	0.7239	0.12024
4	500	0.9225	0.7981	1.3351

For the normal distribution, the SGT estimator consistently does significantly better than the other two. For the uniform distribution, except for the first case when N is small, the SGT estimator consistently does the best among all three. But for the Poisson distribution, the SGT estimator only does the best when λ is small. The empirical estimator is usually the best in this case, although the SGT estimator is notably closer to the empirical estimate than the Laplace estimator. Therefore, tests on known distributions so far show that the SGT estimator should be the better choice among all three estimators.

Table VI
l_1-Distances for LMP

case	time, place	SGT	Empirical	Laplace
1	2, 500	0.9817	0.9890	1.1079
2	2, 3120	0.9554	0.9429	1.1621
3	85, 1	0.8686	0.9290	0.9714
4	90, 2000	1.2048	1.1878	1.4909
Avg.		1.0026	1.0122	1.1581

For the simulated LMP's, the SGT estimator wins in two of the the four sets and the empirical estimator wins in the other two possibly due to different

[4] ... or when the first gap appears, which usually happens after the threshold requirement is met.

[5] The simulated LMP data generated by Weisi Deng contains 94 different times and 3120 different places, and 10000 LMP's for each of the 94×3120 combinations of times and places.

underlying LMP patterns of the four sets. After taking the average l_1-distances over all four cases, the SGT estimator seems to be the best. However, the results here is by no means conclusive, since more extensive and thorough tests are required to further explore the "goodness" of the SGT estimator under most distributions, and if it is only "good" under certain distributions, what those distributions are.

VI. Conclusion

Estimating the probability distribution of random events from sparse data can be highly applicable to convergence bidding and other electricity market-related operations. The Good-Turing estimator was introduced to help solve the problem. It accounts for the probability of unobserved species, and one of its variants, the Simple Good-Turing estimator, does seem to perform better overall than the empirical and Laplace estimators on random samples from most known distributions and simulated data. The distributions covered in this report are by no means exhaustive. Also, more can be explored about how to apply the SGT estimator (and perhaps other variants of the Good-Turing estimator) to real local marginal prices and how the results can help market participants adjust their bidding strategies.

Acknowledgments

The author is grateful to Prof. Lang Tong for giving general guidance and challenging her to formulate and prove things rigorously throughout the summer. The author would also like to give special thanks to her mentor, Sevi Baltaoglu, for patiently answering all her questions and helping to refine her proof. Weisi Deng also gave the author advice on how to lay out this report, generated the 16.2 GB simulated LMP's using her programs, and went out of her way to help the author to run MATLAB programs on the desktop. Yuting Ji, Daniel Munoz-Alvarez, Eugene Zhe Yu, and other members of the Tong Group at Cornell ECE also helped and taught the author in many meaningful ways.

This research was funded by the Cornell ECE Early Career Research Scholars Program and administered through Engineering Learning Initiatives at Cornell.

References

[1] William A. Gale and Geoffrey Sampson. Good-Turing Frequency Estimation Without Tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[2] Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.