entropy

MDPI

Article

Robust Universal Inference

Amichai Painsky [1,*] and Meir Feder [2]

[1] The Industrial Engineering Department, Tel Aviv University, Tel Aviv 6997801, Israel
[2] The School of Electrical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel; meir@eng.tau.ac.il
* Correspondence: amichaip@tauex.tau.ac.il

Abstract: Learning and making inference from a finite set of samples are among the fundamental problems in science. In most popular applications, the paradigmatic approach is to seek a model that best explains the data. This approach has many desirable properties when the number of samples is large. However, in many practical setups, data acquisition is costly and only a limited number of samples is available. In this work, we study an alternative approach for this challenging setup. Our framework suggests that the role of the train-set is not to provide a single estimated model, which may be inaccurate due to the limited number of samples. Instead, we define a class of "reasonable" models. Then, the worst-case performance in the class is controlled by a minimax estimator with respect to it. Further, we introduce a robust estimation scheme that provides minimax guarantees, also for the case where the true model is not a member of the model class. Our results draw important connections to universal prediction, the redundancy-capacity theorem, and channel capacity theory. We demonstrate our suggested scheme in different setups, showing a significant improvement in worst-case performance over currently known alternatives.

Keywords: minimax estimation; minimax risk; statistical inference; estimation theory; universal prediction

check for updates

Citation: Painsky, A.; Feder, M. Robust Universal Inference. *Entropy* **2021**, *23*, 773. https://doi.org/10.3390/e23060773

Academic Editor: Sangun Park

Received: 2 May 2021
Accepted: 15 June 2021
Published: 18 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

One of the major challenges in statistics and machine learning is making predictions and inference from a limited number of samples. This problem is mostly evident in modern statistics (big data), where the dimension of the problem is very large compared to the number of samples in hand, or in cases where data acquisition is relatively costly, and only a small number of samples is available (such as in complicated clinical trails). The standard approach in many applications is to seek a model that best explains the data. For example, empirical risk minimization (ERM) [1] is a commonly used criterion in predictive modeling. Minimizing the empirical risk has many desirable properties. Under different loss functions, we may attain consistency, unbiasedness, and other favorable attributes. In parametric estimation, perhaps the most popular approach is maximum likelihood. Here, again, we seek parameters that maximize the likelihood of the given set of observations.

However, what happens if our specific instance of data does not represent the true model well enough (as happens in high-dimensional problems)? Is it still desirable to choose the single model that best explains it?

In this work, we study an alternative approach for this challenging setup. Here, instead of choosing a model that best describes the data, we define a class of models that describe it with high confidence. Then, we seek a scheme that minimizes the worst-case loss in the class. This way, we control the performance over a class of reasonable models and provide explicit worst-case guarantees, even when the given data fail to accurately represent the true model. This scheme is, in fact, a data-driven approach of minimax estimation, as later discussed. Further, we show it provides worst-case guarantees for the expected regret of future samples. This property makes our framework applicable both for inference and prediction tasks.

One of the major challenges of our suggested scheme is to characterize the model class. In this work, we consider a class of models that corresponds to a confidence region of the unknown parameters. This way we provide a PAC-like generalization bound, as the true model is a member of this class with high confidence. Then, we introduce a more robust approach which drops the model class assumptions and provides stronger performance guarantees for the derived estimator. We demonstrate our suggested approach in classical inference problems and more challenging large alphabet probability estimation. We further study a real-world example, motivated from the medical domain. Our suggested approach introduces favorable worst-case performance, for every given instance of data, at a typically low cost on the average. This demonstrates an "insurance-like" trade-off; we pay a small cost on the average to avoid a great loss if "something bad happens" (that is, the observed samples do not represent the true model well enough).

The rest of this manuscript is organized as follows. In Section 2, we review related work to our problem. We introduce our suggested framework and some of its basic properties in Section 3. Then, we extend the framework to a more robust estimation scheme in Section 4. We demonstrate our suggested scheme in several setups. In Section 5, we study the unknown normal mean problem, while in Section 6 we focus on multinomial probability estimation. We consider a more challenging large alphabet probability estimation problem in Section 7. Finally, we study a real-work breast cancer problem in Section 8. We conclude with a discussion in Section 9.

2. Previous Work

Minimax estimation has been extensively studied over the years. Here, we briefly review the more relevant results for our work. Let $x^n \sim p_\theta^n$ be a collection of n i.i.d. samples, drawn from a distribution p_θ, where θ is a fixed and unknown parameter. Let Θ be a given class of parameters. Assume that $\theta \in \Theta$. Let $\hat{\theta} \triangleq \hat{\theta}(x^n)$ be an estimator of θ from x^n. Let $R(\theta, \hat{\theta})$ be a risk function which measures the expected error between the true parameter θ and its corresponding estimate $\hat{\theta}$. For example, $R(\theta, \hat{\theta}) = \mathbb{E}_{x^n \sim p_\theta^n}(\theta - \hat{\theta}(x^n))^2$ is the mean squared error. The minimax risk [2] is defined as

$$r_{mm} = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}). \tag{1}$$

A minimax estimator $\hat{\theta}_{mm}$ satisfies $\sup_{\theta \in \Theta} R(\theta, \hat{\theta}_{mm}) = r_{mm}$, if such exists. In words, $\hat{\theta}_{mm}$ minimizes the worst-case risk for a given class of parameters Θ. Finding the minimax estimator is, in general, not an easy task. However, the optimal solution is characterized by several important properties.

Let $\hat{\theta}_\pi = \int_{\theta \in \Theta} \theta \pi(\theta) d\theta$ be a Bayes estimator with respect to some prior $\pi(\theta)$ over Θ. In words, $\hat{\theta}_\pi$ is a weighted average of $\theta \in \Theta$, according to a given weight function $\pi \triangleq \pi(\theta)$. Let $r_\pi = \int_{\theta \in \Theta} R(\theta, \hat{\theta}_\pi) d\pi(\theta)$ be the average risk with respect to π. One of the basic results in the minimax theory suggests that if $r_\pi = r_{mm}$, then $\hat{\theta}_\pi$ is a minimax estimator and π is a least favorable prior (satisfying $r_\pi \geq r_{\pi'}$ for any π') [2]. Importantly, if a Bayes estimator has a constant risk, it is minimax. Note that this is not a necessary condition.

For example, consider the problem of estimating the mean of a d-dimensional Gaussian vector. Here, it can be shown that the maximum likelihood estimator (MLE) is also the minimax estimator with respect to the squared error. Interestingly, in this example, the MLE is known to be inadmissible for $d > 2$; assuming that the mean is finite, the famous James–Stein estimator [3] dominates the MLE, as it achieves a lower mean squared error (where the phenomenon is more evident as the mean is closer to zero) [2,3]. Additional examples for the minimax estimators are provided in [2].

The minimax formulation was studied in a variety of setups. In [4], the authors considered minimax estimation of parameters over L^p loss and provided key analytical results. These results were further studied and generalized (for example, see in [5]). Bickel studied minimax estimation of the normal mean when the parameter space is restricted [6].

Later, Marchand and Perron considered the case where the norm of the normal mean is bounded [7]. The minimax approach is also applicable to supervised learning problems. In [8], the authors considered minimax classification with fixed first- and second-order moments. Eban et al. developed a classification approach by minimizing the worst-case hinge loss subject to fixed low-order marginals [9]. Razaviyayn et al. fitted a model that minimizes the maximal correlation under fixed pairwise marginals to design a robust classification scheme [10]. Farnia et al. described a minimax approach for supervised learning by generalizing the maximum entropy principle [11].

The minimax approach has many applications, as it defines a conservative estimate for a given class of models. A variety of examples spans different fields including optimization [12], signal processing [13,14], communications [15], and others [16].

It is important to emphasize that the minimax problem (1), and its corresponding solution, heavily depend on the assumption that the unknown parameter θ is a member of the given class of parameters Θ. However, what happens if this assumption is false, and $\theta \notin \Theta$ (as discussed, for example, in [17–19])? Furthermore, how do we choose Θ in practice? If we choose Θ to be too large, we might control a class of models that are unreasonable. On the other hand, if Θ is too small, we may violate the assumption that $\theta \in \Theta$. Finally, notice that the minimax problem is typically concerned with the expected worst-case performance (the risk). However, in many real-world applications we are given a single instance of data, which may be quite costly to acquire. Therefore, we require worst-case performance guarantees for this specific instance of data. In this work, we address these concerns and suggest a robust, data-driven, universal estimation scheme for a given set of observations.

3. The Suggested Inference Scheme

For the purpose of our presentation, we use the following additional notations. Let Θ_r be a *restricted* class of parameters and denote $\mathcal{P}(\Theta_r)$ as the corresponding *restricted* class of parametric distributions. Assume that the true model p_θ is a member of $\mathcal{P}(\Theta_r)$ (or alternatively, $\theta \in \Theta_r$). For example, $p_\theta = \mathcal{N}(\theta, 1)$ is a normal distribution with an unknown mean θ and a unit variance, while $\mathcal{P}(\Theta_r)$ is a set of all normal distributions with $\theta \in \Theta_r = [\theta_a, \theta_b]$ (henceforth, restricted to $[\theta_a, \theta_b]$) and a unit variance. Let $\mathcal{P} = \{p | \, p(x) \geq 0, \; \int p(x)dx = 1\}$ be the class of all probability measures. Let $q \triangleq q(\cdot|x^n)$ be a probability measure which estimates p_θ given the samples x^n. Notice that as opposed to the presentation in (1), the estimator q is with respect to the entire probability distribution p_θ, and not just unknown parameter θ. We measure the estimate's accuracy using the Kullback–Leibler (KL) divergence between the true underlying distribution p_θ and q, formally defined as $D_{\mathrm{KL}}(p_\theta || q) = \int p_\theta(x) \log \frac{p_\theta(x)}{q(x)} dx$. The KL divergence is a widely used measure for the discrepancy between two probability distributions, with many desirable properties [20]. In addition, the KL divergence serves as an upper bound for a collection of popular discrepancy measures (for example, the Pinsker inequality [21] and the universality results in [22,23]). In this sense, by minimizing the KL divergence, we implicity bound from above a large set of common performance merits.

Ultimately, our goal is to find an estimate q that minimizes $D_{\mathrm{KL}}(p_\theta || q)$ for the unknown θ. Thus, we consider a minimax formulation

$$\min_{q} \sup_{p_\theta \in \mathcal{P}(\Theta_r)} D_{\mathrm{KL}}(p_\theta || q) \tag{2}$$

where $q \in \mathcal{P}$ is the minimizer of the worst-case divergence over the class $\mathcal{P}(\Theta_r)$, if such exists. In words, q minimizes the worst possible divergence, over the restricted model class $\mathcal{P}(\Theta_r)$. To avoid an overload of notation, we assume that $q \in \mathcal{P}$ throughout the text, unless otherwise stated. We observe several differences between (1) and (2). First, the formulation in (1) considers an estimate $\hat{\theta}$, and by that implicitly restricts the solution to be a parametric distribution $p_{\hat{\theta}}$ of the same family as p_θ. On the other hand, (2) considers the entire distribution and does not impose any restrictions on the solution. Second, the standard minimax formulation (1) focuses on the risk. Our approach considers the estimation accuracy for every given instance of data ($q \triangleq q(\cdot|x^n)$, as defined above). Third, notice that $D_{\mathrm{KL}}(p_\theta||q)$ can also be viewed as the expected log-loss regret, where the expectation is with respect to a future sample, $D_{\mathrm{KL}}(p_\theta||q) = \mathbb{E}_{x\sim p_\theta}(l(x,q) - l(x,p_\theta))$ and $l(x,q) = -\log(q(x))$ is the logarithmic loss. This means that while (1) focuses on the expected loss with respect to the given samples, (2) considers the expected performance of future samples. We further discussed these points in Sections 5, 6 and 8.

In practice, one of the major challenges in applying any minimax formulation is the choice (or design) of the parametric class. Specifically, using the notations of (2), choosing a larger class Θ_r is more likely to include the true model θ, but may also include unreasonable worst-case models (for example, $\Theta_r = \mathbb{R}$ in the unknown normal mean example above). On the other hand, choosing a more restrictive Θ_r may violate our assumption that $\theta \in \Theta_r$. Therefore, a trade-off between the two seems inevitable. In the following, we focus on the design and characterization of a set Θ_r that depends on the given samples, $\Theta_r(x^n)$. In other words, we use the train-set x^n to construct a minimal-size restricted model class $\Theta_r(x^n)$ that contains the true parameter θ with high confidence. Then, we solve the minimax problem (2) with respect to it and attain an estimator that minimizes the maximal divergence in the class (or equivalently, the expected log-loss regret for future samples).

Designing and Controlling the Restricted Model Class

Our first objective is to construct a minimal-size $\Theta_r(x^n)$ such that $\theta \in \Theta_r(x^n)$ with high confidence. For this purpose, we turn to classical statistics and construct a confidence region for the desired parameter θ. A confidence region of level $100(1-\alpha)\%$ is defined as a region $\mathcal{T}$ such that $P(\theta \in \mathcal{T}) = 1-\alpha$. Notice that $\mathcal{T}$ is random and depends on the samples x^n, while θ is an unknown (non-random) parameter. Further, notice that a confidence region $\mathcal{T}$ is data-dependent and does not require knowledge of the true parameter θ. Obviously, there are many ways to define $\mathcal{T}$ to satisfy the above. We are interested in a confidence region that has a minimal expected volume. For example, consider n i.i.d. samples from $\mathcal{N}(\theta,1)$, as discussed above. Let $\bar{x}$ be the sample mean. Then, the minimal-size $100(1-\alpha)\%$ confidence interval is $[\bar{x} \pm z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}}]$, where z_α is the upper 100α percentile of a standard normal distribution [24].

Given the restrictive model class, we would like to solve the minimax problem defined in (2). A general form of this problem was extensively studied over the years, mainly in the context of universal compression and universal prediction [15]. There, $D_{\mathrm{KL}}(p_\theta||q)$ is the expected number of extra bits (over the optimal code-book), required to code samples from p_θ using a code designed for q. The celebrated redundancy-capacity theorem demonstrates a basic connection between the desired formulation (2) and channel capacity theory. Let $T \sim \pi$ be a source variable, X be a target variable and $\mathcal{P}(\Theta_r)$ be the set of transition probabilities from X to T. In other words, T is a message, transmitted through a noisy channel, characterized by $\mathcal{P}(\Theta_r)$. The received (noisy) message is denoted by X. Let $I(T;X)$ be the mutual information between T and X, and $C(\Theta_r) \triangleq \sup_\pi I(T;X)$ be the corresponding channel capacity. The redundancy-capacity theorem [25–27] suggests that for $C(\Theta_r) < \infty$, the minimax formulation presented in (2) is equivalent to

$$\sup_{\pi(\theta)} \int_{\theta\in\Theta_r} \pi(\theta) D_{\mathrm{KL}}(p_\theta||q_\pi) d\theta = \sup_\pi I(T;X) \triangleq C(\Theta_r) \tag{3}$$

where $\pi(\theta)$ is a weight function for every $\theta \in \Theta_r$ and $q_\pi = \int_{\theta\in\Theta_r} \pi(\theta) p_\theta d\theta$ is a *mixture distribution*. In words, solving (2) is equivalent to solving a channel capacity problem. Furthermore, the source distribution which maximizes the mutual information between the source and the target (and henceforth achieves the channel capacity) is a mixture over $\mathcal{P}(\Theta_r)$. This solution is quite similar to the solution of (1); in both cases, we obtain a Bayes estimator over the given class, while the least favorable prior (or equivalently, the capacity achieving prior), if such exists, attains the maximal average risk (the channel capacity). See examples in Sections 5 and 6 for further detail. It is important to emphasize that while θ is a fixed and unknown parameter, $\pi(\theta)$ is a weight function for every $\theta \in \Theta_r$, and q_π is a weighted average over $\mathcal{P}(\Theta_r)$.

The redundancy-capacity theorem shows that the solution to the minimax problem (2) is achieved by solving the channel capacity problem (3). We apply the capacity-redundancy theorem to our suggested class $\Theta_r(x^n)$ to attain the desired solution. Theorem 1 below summarizes our parametric inference approach.

Theorem 1. *Let $x^n \sim p_\theta^n$ be a collection of n i.i.d. samples, drawn from an unknown distribution p_θ. Let $\Theta_r(x^n)$ be a $100(1-\alpha)\%$ confidence region for the parameter θ. Assume that $C(\Theta_r(x^n)) < \infty$. Then, with probability $1-\alpha$ (over the samples), $C(\Theta_r(x^n))$ is the minimal worst-case divergence and q_π is the corresponding minimax estimator, denoted as the mixture model.*

Theorem 1 establishes a PAC-like generalization bound for parametric inference. It defines the worst-case expected performance of future samples (with respect to a logarithmic loss, as discussed above), at a confidence level of $1-\alpha$ over the drawn samples. Specifically, with probability $1-\alpha$ we have that $\mathbb{E}_{x\sim p_\theta}(l(x,q_\pi) - l(x,p_\theta)) \le C(\Theta_r(x^n))$ for the entire parametric class. It is important to emphasize that the resulting minimax estimator q_π is data-dependent, as it is a mixture over the data-driven restricted model class.

Solving the channel capacity problem is, in general, not an easy task. However, there exist several cases where the solution to (3) holds a closed-form expression, or an efficient computational routine. We demonstrate basic examples in Sections 5 and 6.

4. A Generalized Inference Scheme beyond the Restricted Class

In the previous section, we derive a minimax solution to (2) under the assumption that $p_\theta \in \mathcal{P}(\Theta_r)$ (equivalently, $\theta \in \Theta_r$), with high confidence. Unfortunately, it does not provide any guarantee for the event where $p_\theta \notin \mathcal{P}(\Theta_r)$. We now consider a general setup where p_θ is not necessarily in $\mathcal{P}(\Theta_r)$ as well as introduce a more robust approach which addresses this case.

Let $\mathcal{P}(\Theta)$ be a (non-restricted) model class that is known to contain the true parametric model p_θ. Here, we define Θ as the set of all possible parameter values, such that $\theta \in \Theta$. For example, $\mathcal{P}(\Theta)$ is a class of all normal distributions with an unknown mean and a unit variance ($\Theta = \mathbb{R}$), in the normal mean example above. As before, we would like to find a distribution q that minimizes $D_{\text{KL}}(p_\theta||q)$. Simple calculus shows that

$$D_{\text{KL}}(p_\theta||q) = D_{\text{KL}}(p_\theta||p_{\theta'}) + \int p_\theta(x) \log \frac{p_{\theta'}(x)}{q(x)} dx \tag{4}$$

for any choice of $p_{\theta'}$. Specifically, (4) holds for any model in the restricted model class, $p_{\theta'} \in \mathcal{P}(\Theta_r)$. Notice that in this case, the first term of (4) is an error induced by the restrictive model class, independent of the choice of q. The second term is the residual, which depends on q. Notice that the first term only depends on p_θ and the reference distribution $p_{\theta'} \in \mathcal{P}(\Theta_r)$. This means that by choosing a model class $\mathcal{P}(\Theta_r)$ that is too "far" (or from the true distribution), we face a large overhead term that is independent of the estimator q. On the other hand, the second term depends on q, and may be universally bounded. In other words, we are interested in a universal bound of the form

$$\min_q \sup_{p_\theta \in \mathcal{P}(\Theta)} \sup_{p_{\theta'} \in \mathcal{P}(\Theta_r)} \left(\int p_\theta(x) \log \frac{p_{\theta'}(x)}{q(x)} dx \right). \tag{5}$$

Interestingly, notice that (5) may be equivalently written as

$$\min_q \sup_{p_\theta \in \mathcal{P}(\Theta)} \sup_{p_{\theta'} \in \mathcal{P}(\Theta_r)} \left(\int p_\theta(x) \log \frac{p_\theta(x)}{q(x)} dx - \int p_\theta(x) \log \frac{p_\theta(x)}{p_{\theta'}(x)} dx \right) = \tag{6}$$

$$\min_q \sup_{p_\theta \in \mathcal{P}(\Theta)} \left(D_{\mathrm{KL}}(p_\theta || q) - \inf_{p_{\theta'} \in \mathcal{P}(\Theta_r)} D_{\mathrm{KL}}(p_\theta || p_{\theta'}) \right).$$

This means that (5) is just a constrained variant of (2). Therefore, similarly to (2), we would like to represent (5) as a (constrained) channel capacity problem.

Definition 1. *Let $p_\theta \in \mathcal{P}(\Theta)$ be an unknown probability distribution. Let $\mathcal{P}(\Theta_r)$ be a restrictive model class (that does not necessarily contain p_θ). Assume that Θ_r is bounded. Define*

$$F(\Theta, \Theta_r) \triangleq \sup_{\pi(\theta)} \int_{\theta \in \Theta} \pi(\theta) \left(D_{\mathrm{KL}}(p_\theta || q_\pi) - \min_{\theta' \in \Theta_r} D_{\mathrm{KL}}(p_\theta || p_{\theta'}) \right) d\theta = \tag{7}$$

$$\sup_{\pi(\theta)} \left(I(T; X) - \mathbb{E}_{\pi(\theta)} \min_{\theta' \in \Theta_r} D_{\mathrm{KL}}(p_\theta || p_{\theta'}) \right).$$

As in (2), $I(T;X)$ is the mutual information between a source variable $T \sim \pi$, and a target variable X, that is characterized by the transition probabilities $\mathcal{P}(\Theta)$. The constraint is simply the expected divergence (with respect to π) between p_θ and its closest projection in Θ_r. The term q_π is a mixture distribution over $\mathcal{P}(\Theta)$, according to the prior π. Notice that here, the mixture is over the non-restricted model class, as opposed to (2), where the mixture is over $\mathcal{P}(\Theta_r)$. We denote this distribution, q_π, as the projected mixture distribution.

Theorem 2. *Let $p_\theta \in \mathcal{P}(\Theta)$ be an unknown probability distribution. Let $\mathcal{P}(\Theta_r)$ be a restrictive model class (that does not necessarily contain p_θ). Assume that Θ_r is bounded. Then, for $F(\Theta, \Theta_r) < \infty$ the following holds:*

$$\min_q \sup_{p_\theta \in \mathcal{P}(\Theta)} \sup_{p_{\theta'} \in \mathcal{P}(\Theta_r)} \left(\int p_\theta(x) \log \frac{p_{\theta'}(x)}{q(x)} dx \right) = F(\Theta, \Theta_r). \tag{8}$$

A proof of Theorem 2 is provided in Appendix A. Theorem 2 establishes a redundancy-capacity result, similarly to (2). It shows that (5) may be obtained by solving a constrained channel capacity problem, and the distribution which achieves it is, again, a mixture distribution. This result is further discussed in [28] in a different (asymptotic) setup.

In addition, notice that for a bounded Θ_r the formulation in (5) may be equivalently written as

$$\min_q \sup_{p_\theta \in \mathcal{P}(\Theta)} \int p_\theta(x) \log \frac{p^*_\theta(x)}{q(x)} dx,$$

where $p^*_\theta = \mathrm{argmin}_{p_{\theta'} \in \mathcal{P}(\Theta_r)} D_{\mathrm{KL}}(p_\theta || p_{\theta'})$. In other words, $F(\Theta, \Theta_r)$ is also the optimal universal minimizer of the second term of (4), for a specific (greedy) choice of $p_{\theta'} \in \mathcal{P}(\Theta_r)$ that minimizes the first term. This result may be viewed as a "triangle inequality" for the KL divergence: given a reference set $\mathcal{P}(\Theta_r)$, the KL divergence $D_{\mathrm{KL}}(p_\theta || q)$ is bounded from above by the closest projection in $\mathcal{P}(\Theta_r)$ to p_θ, plus an overhead-term $F(\Theta, \Theta_r)$. It is important to emphasize that $F(\Theta, \Theta_r)$ is not new to the universal coding literature. In fact, it was introduced in [17] as relative redundancy, in the context of robust codes for universal compression. However, it was mostly studied in an asymptotic regime, where n i.i.d. variables X^n are simultaneously compressed. However, it was mostly studied in an asymptotic regime, where n i.i.d. variables X^n are simultaneously compressed [28].

Similarly to the channel capacity problem, the term $F(\Theta, \Theta_r)$ holds a closed-form analytical expression only in several special cases. Therefore, we introduce a simple iterative algorithm, which provides an optimal solution to it (as indicated in [28]). Our suggested routine is similar in spirit to the Blahut–Arimoto algorithm [21], which is typically applied to intractable channel capacity problems.

Theorem 3. *Let $\mathcal{P}(\Theta)$ and $\mathcal{P}(\Theta_r)$ be two model classes. Let $F(\Theta, \Theta_r)$ follow the definition above. Assume that Θ_r is bounded. Then, for $F(\Theta, \Theta_r) < \infty$ the following holds:*

$$F(\Theta, \Theta_r) = \sup_{\phi(\theta), \psi(\theta, x)} \int_{\theta \in \Theta} \int_x \phi(\theta) p_\theta(x) \log \frac{\psi(\theta, x)}{\phi(\theta)} \frac{p^*_\theta(x)}{p_\theta(x)} dx d\theta \tag{9}$$

where $\phi(\theta)$ and $\psi(\theta, x)$ are probability distributions (over the variable θ, for any given x), and $p^*_\theta(x) = \mathrm{argmin}_{\theta' \in \Theta_r} D_{\mathrm{KL}}(p_\theta || p_{\theta'})$. Further, the solution to (9) may be attained by the following iterative projection algorithm:

1. *For a fixed $\phi(\theta)$, we set $\psi(\theta, x) = \frac{\phi(\theta) p_\theta(x)}{\int_{\theta \in \Theta} \phi(\theta) p_\theta(x) d\theta}$*
2. *For a fixed $\psi(\theta, x)$, we set $\phi(\theta) = \frac{\prod_x \tilde{\psi}(\theta, x)^{p_\theta(x)}}{\int_{\theta \in \Theta} \prod_x \tilde{\psi}(\theta, x)^{p_\theta(x)} d\theta}$ where $\tilde{\psi}(\theta, x) = \psi(\theta, x) \frac{p^*_\theta(x)}{p_\theta(x)}$.*

Finally, the distribution q that achieves $F(\Theta, \Theta_r)$ is given by $q_\Theta = \int_{\theta \in \Theta} \phi^*(\theta) p_\theta d\theta$, where $\phi^*(\theta)$ is $\phi(\theta)$ at the final iteration of the algorithm.

A proof for this theorem is provided in Appendix B.

In many practical cases, the choice of a model class $\mathcal{P}(\Theta)$ is not a trivial task. For example, consider a real-world setup where a domain expert suggests that the underlying model follows a Normal distribution with an unknown mean. However, we would like to design a scheme that does not heavily rely on this assumption. Therefore, we may consider the general case where $\mathcal{P}(\Theta)$ is the simplex of all possible probability distributions. This important special case described in the following section.

The Normalized Maximum Likelihood

Consider a model class $\mathcal{P}(\Theta) = \mathcal{P} = \{p | \, p(x) \geq 0, \, \int p(x) dx = 1\}$. Here, the solution to (7) holds a closed form expression.

Theorem 4. *Let $p_\theta \in \mathcal{P}$ be an unknown probability distribution where $\mathcal{P} = \{p | \, p(x) \geq 0, \, \int p(x) dx = 1\}$. Let $\mathcal{P}(\Theta_r)$ be a restrictive model class. Assume that Θ_r is bounded and $Z \triangleq \int \max_{p_{\theta'} \in \mathcal{P}(\Theta_r)} p_{\theta'}(x) dx$. Let $\Gamma(\Theta_r) \triangleq \log(Z)$. For $\Gamma(\Theta_r) < \infty$,*

$$\min_q \sup_{p_\theta \in \mathcal{P}} \sup_{p_{\theta'} \in \mathcal{P}(\Theta_r)} \int p_\theta(x) \log \frac{p_{\theta'}(x)}{q(x)} dx \triangleq \Gamma(\Theta_r)$$

and the model q that achieves the minimum is the normalized maximum likelihood (NML) [29] , $q_{nml}(x) = \max_{p_{\theta'} \in \mathcal{P}(\Theta_r)} p_{\theta'}(x) / Z$

This theorem is an immediate application of Shtarkov's NML result [29]. It suggests that given a model class $\mathcal{P}(\Theta_r)$ which does not necessarily contain the true model p, the NML estimator q_{nml} minimizes the worst-case regret over all possible distributions and guarantees an overhead of at most $\Gamma(\Theta_r)$ bits, compared to the best model in the class $\mathcal{P}(\Theta_r)$, for any possible p. Further, it is shown that this result is tight, in the sense that there exist probability distributions $p \in \mathcal{P}(\Theta)$ and $p_{\theta'} \in \mathcal{P}(\Theta_r)$ that achieve the $\Gamma(\Theta_r)$ term. Notice that for every $\mathcal{P}(\Theta)$ and $\mathcal{P}(\Theta_r)$ that satisfy the conditions above, we have that $C(\Theta_r) \leq F(\Theta, \Theta_r) \leq \Gamma(\Theta_r)$. This means that under more restrictive assumptions we attain tighter worst-case performance guarantees, as expected. We now demonstrate our suggested methods in synthetic and real-world problems.

5. The Normal Distribution

Let us first study the classical unknown mean problem in the Gaussian case. Consider n i.i.d. samples, drawn from a d-dimensional multivariate normal distribution with an unknown mean μ and a known covariance matrix Σ. The $100(1-\alpha)\%$ confidence region for μ is $\mathcal{M}_r = \{\mu|(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) \leq \frac{1}{n}\chi^2_d(1-\alpha)\}$, where $\bar{x}$ is the sample mean, and χ^2_d is a Chi-squared distribution with d degrees of freedom. First, we would like to solve the minimax problem (2) with respect to $\mathcal{M}_r$. We apply the redundancy-capacity theorem (3) and define a corresponding channel, $X = M + Z$, where M is a random vector, taking values over the domain $\mathcal{M}_r$, while $Z \sim \mathcal{N}(0,\Sigma)$, independent of M (see [30] for detail). This formulation is also known as an amplitude-constrained capacity problem. We show (Appendix C) that it is equivalent to the generic case where $Z \sim \mathcal{N}(0, I_d)$ and $M \in \mathcal{M}'_r$ where $\mathcal{M}'_r = \{\mu|\mu^T\mu \leq \frac{1}{n}\chi^2_d(1-\alpha)\}$. Notice that the domain of M is now restricted to a d-dimensional ball (defined by $\mathcal{M}'_r$) and our goal is to find the capacity achieving distribution of M. It has been shown [31] that the solution to this problem is achieved when M is supported on a finite number of concentric spheres. Recently, the authors of [32] studied the necessary conditions under which the solution is a single sphere, centered at the origin. Specifically, they derived the largest radius r_d for which the capacity achieving distribution is uniform on the sphere of the d-dimensional ball. This means that if the radius defined by $\mathcal{M}'_r$ is smaller than r_d, then the solution to our minimax problem is immediate. Applying Dytso et al. analysis to our problem, we attain the following result.

Theorem 5. *Let x^n be a collection of n i.i.d. samples from a d-dimensional multivariate normal distribution with an unknown mean μ and a known covariance matrix Σ. Let $\mathcal{M}_r$ be a $100(1-\alpha)\%$ confidence region for μ. Let r_d be the largest radius for which the capacity achieving distribution is uniform on the sphere of a d-ball, as defined in Table 1 of [32]. Then, for any $n \geq \chi^2_d(1-\alpha)/r^2_d$, the solution to the minimax problem (2) over the confidence region $\mathcal{M}_r$ is attained by a uniform mixture of Gaussians with means on the confidence region, $q_\pi \propto \int_{\mu\in\mathcal{O}(\mathcal{M}_r)} \mathcal{N}(\mu,\Sigma)d\mu$ where $\mathcal{O}(\mathcal{M}_r) = \{\mu|(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) = \frac{1}{n}\chi^2_d(1-\alpha)\}$.*

For example, let $\alpha = 0.05$ and $d = 2$. We have that $r_d = 2.454$ (as appears in Table 1 of [32]), and the solution to (2) over $\mathcal{M}_r$ is given by $q_\pi \propto \int_{\mu\in\mathcal{O}(\mathcal{M})} \mathcal{N}(\mu,\Sigma)d\mu$, for every $n \geq 1$. The left chart of Figure 1 illustrates the shape of q_π in this case. This Gaussian mixture shape may seem counterintuitive at a first glance, as x^n are known to be drawn from a normal distribution. However, the reason is quite clear. Our inference criterion strives to control a set of Gaussian models. Therefore, the optimal solution is not necessarily the most likely model in the set, but a mixture of models.

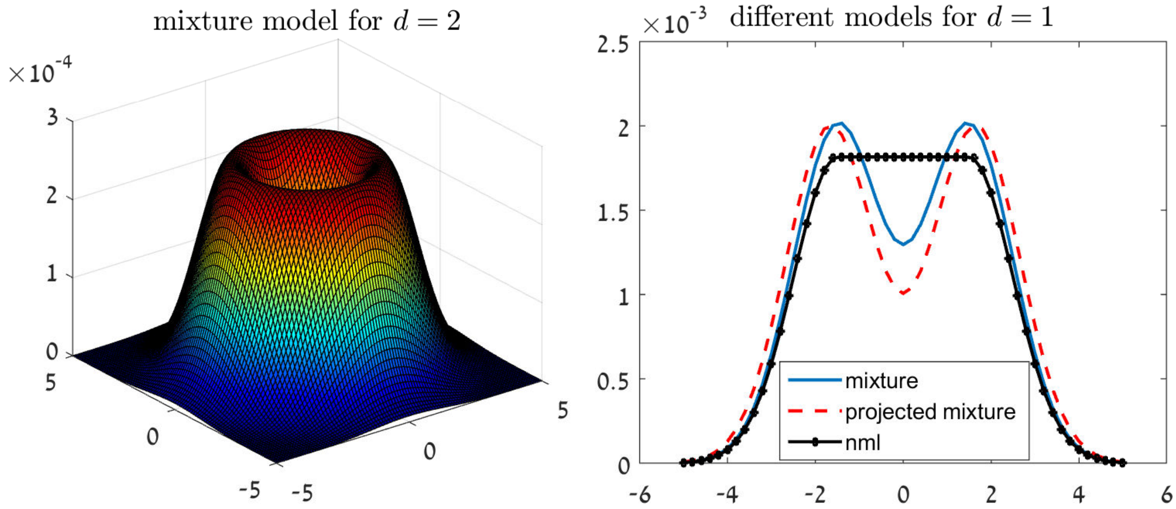

Figure 1. The shape of our suggested solutions in the unknown normal mean problem. (**Left**)—the mixture distribution q_π for $d = 2$. (**Right**)—all methods for $d = 1$ and an example confidence interval of $[-1.5, 1.5]$.

Let us now turn to the projected mixture distribution and the NML. The right chart of Figure 1 demonstrates the shape of these estimators for $d = 1$ and $\mathcal{M}_r = [-1.5, 1.5]$.

First, we notice that the projected mixture is again a Gaussian mixture, with means outside of the confidence interval. On the other hand, the NML solution is not a Gaussian mixture; simple calculus shows that

$$q_{nml}(x) \propto \begin{cases} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x+a)^2) & x < -a \\ \frac{1}{\sqrt{2\pi}} & -a \leq x \leq a \\ \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x-a)^2) & x > a \end{cases} \tag{10}$$

for a symmetric confidence interval $[-a, a]$.

Let us illustrate the performance of our suggested methods. We draw n i.i.d. samples from a standard normal distribution $p_\mu \sim \mathcal{N}(0,1)$ where the mean $\mu = 0$ is unknown and the variance is known. We apply our suggested methods (with $\alpha = 0.05$) and evaluate the KL divergence from the true distribution, $D_{\text{KL}}(p_\mu || q(\cdot|x^n))$. We compare our results with the performance of the MLE, $D_{\text{KL}}(p||q_{\text{mle}}(x^n))$, where $q_{\text{mle}}(x^n) = \mathcal{N}(\bar{x}, 1)$. Notice that the MLE is also known to be the minimax solution to (1) in this setting. We repeat this experiment k = 10,000 times, for different sample sizes n. For each n we evaluate the mean $\mathbb{E}_{x^n \sim p_\mu^n} D_{\text{KL}}(p_\mu || q(\cdot|x^n))$, the variance $\text{var}_{x^n \sim p_\mu^n} D_{\text{KL}}(p_\mu || q(\cdot|x^n))$ and the worst-case $\max_{x^n \in \mathcal{X}_k} D_{\text{KL}}(p_\mu || q(\cdot|x^n))$, where $\mathcal{X}_k$ is the set of k random draws of x^n from p_μ^n. Figure 2 demonstrates our results. Notice that we lose some accuracy, on the average, with all of our methods, compared to the MLE. On the other hand, the variance of the MLE is significantly greater, which suggests that it is less reliable for a given instance of data. Finally, we notice a significant gain in the worst-case performance. This behavior is not surprising: our approach strives to control the worst-case performance for each given draw. In this sense, we may view our approach as an "insurance policy"—we pay a small cost on the average, but attain a more stable estimator and gain significantly if "something bad happens" (that is, we observe x^n that do not represent the true model well enough). Notice that this phenomenon is more evident when the inference problem is more challenging (smaller n's). As we compare our suggested models to each other, we notice that the mixture distribution is the most conservative (that is, smallest cost and smallest gain), while the projected mixture is the least conservative. The reason is quite clear: in about $(1-\alpha)$ of the draws, the true parameter lies within the confidence region, and the mixture distribution is closer to it. This implies better performance on the average and worse performance in the extremes. Interestingly, the NML behaves as a compromise between the two. This is mostly as the NML does not assume that $\mu \in \mathcal{M}_r$ (better than the mixture in the worst-case). However, it also unnecessarily controls non-Gaussian models (worse than the projected mixture).

Let us now illustrate our suggested approach in a high-dimensional setting, $p = \mathcal{N}(\underline{1}, I_d)$. Figure 3 compares the mixture estimator (which demonstrates a reasonable compromise between mean and worst-case performance) with the MLE and the James–Stein (JS) estimator. As we can see, the JS estimator slightly outperforms MLE on the average (as discussed in [3]), while the mixture distribution is very close to them. However, as we focus on the variance and the worst-case performance, the mixture distribution demonstrates a significant improvement, as expected. It is important to mention that in a zero mean case, the JS estimator achieves a significantly lower mean error (as discussed in [2]) and a remarkable increase in variance and worst-case performance. These results are omitted for brevity.

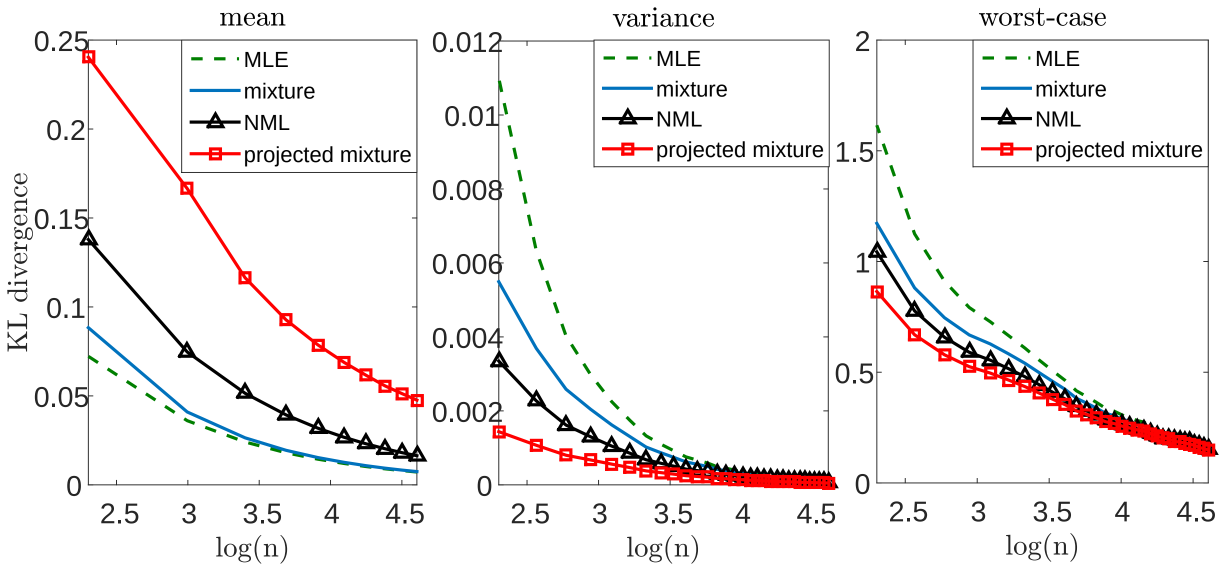

Figure 2. Mean, variance, and worst-case performance in the Gaussian unknown mean problem. We draw n samples from $p_\mu \sim \mathcal{N}(0,1)$ and compute $D_{\text{KL}}(p_\mu||q(\cdot|x^n))$. We repeat this experiment 10,000 times and evaluate the mean (**left**), variance (**middle**), and worst-case (**right**) performance.

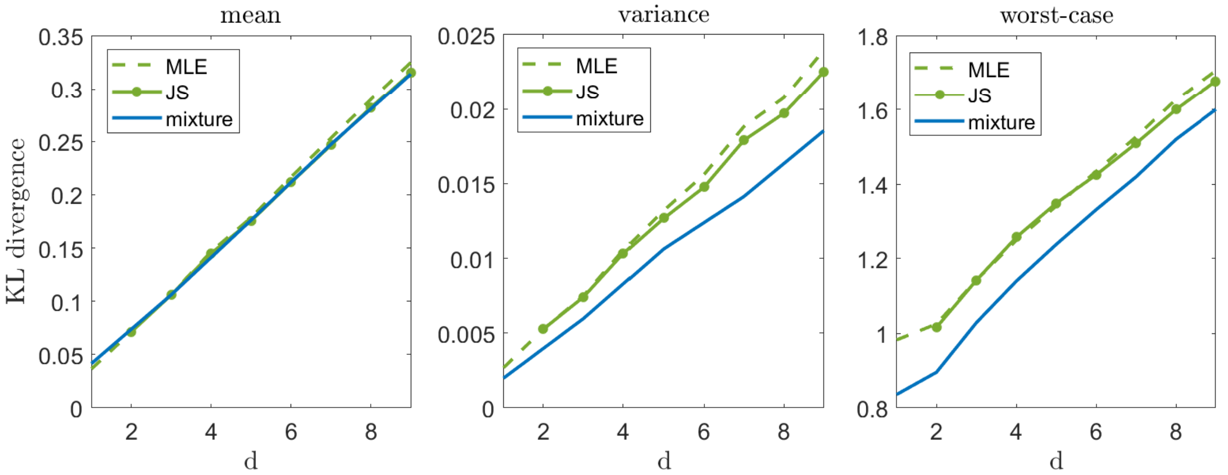

Figure 3. Mean, variance, and worst-case performance in high-d Gaussian unknown mean problem. In each experiment we draw $n = 20$ samples from $p = \mathcal{N}(\underline{1}, I_d)$ and compute $D_{\text{KL}}(p||q(\cdot|x^n))$. We repeat this experiment 10,000 times and evaluate the mean (**left**), variance (**middle**), and worst-case (**right**) performance.

6. The Multinomial Distribution

We now turn to an additional important example of finite alphabet distributions. Let x^n be n i.i.d. draws from a multinomial distribution over an alphabet size m. Notice that here, the parametric family spans the entire simplex. Therefore, we omit the parametric subscript θ to avoid an overload of notation, and regard p as the unknown vector of parameters. As discussed above, we would first like to construct a minimal-volume confidence region for p, denoted as $\mathcal{P}_r$. Unfortunately, there exists no closed-form solution in the multinomial case. Therefore, we turn to an approximate confidence region suggested in [33]. As many other approximation techniques [34,35], Sison and Glaz derive a rectangular region $\mathcal{P}_{sg} = \{p | \; p_l(i) \leq p(i) \leq p_u(i) \; \forall i = 1, \ldots, m\}$ which demonstrates a smaller expected volume compared to alternatives. Our first step is to define a subset of Sison and Glaz region, $\mathcal{P}_r \subset \mathcal{P}_{sg}$, which corresponds to valid probability distributions, $\mathcal{P}_r = \{p | p \in \mathcal{P}_{sg}, \sum p(i) = 1\}$. Notice that $\mathcal{P}_r$ is a convex set, and denote its set of vertexes as $\mathcal{V}(\mathcal{P}_r)$. We show (Appendix D) that the solution to (2) over $\mathcal{P}_r$ is attained by solving (3) over $\mathcal{V}(\mathcal{P}_r)$. This means that instead of considering the entire class $\mathcal{P}_r$, we only need to focus on the discrete set $\mathcal{V}(\mathcal{P}_r)$.

Unfortunately, there exists no closed-form solution to (3) in this setting. However, as the cardinality of π is finite (as we optimize over $\mathcal{V}(\mathcal{P}_r)$), we may apply the Blahut–Arimoto algorithm [21] and attain a numerical solution, at a relatively small computational cost. Finally, we derive the projected mixture and the NML. As mentioned above, the parametric family spans the entire simplex. This means that the two methods are identical, and obtained by applying the NML over $\mathcal{P}_r$.

We now demonstrate our suggested approach. Let x^n be i.i.d. draws from a Zipf's law distribution over an alphabet size $m = 5$ and a parameter $s = 1.01$, $p(i) \propto i^{-s}$. The Zipf's law distribution is a commonly used benchmark distribution, mostly in modeling of natural (real-world) quantities. It is widely used in physical and social sciences, linguistics, economics, and many other fields [36–38]. As in Section 5, we compare our suggested methods to the MLE. In addition, we consider the popular Laplace estimator, which adds a single count to all events, followed by a MLE. In our experiments we focus on an enhanced variant of Laplace [39], which adds $1/2$ to all events, $q_{\text{lap}}(n_i) \propto n_i + 1/2$, where n_i is the number of appearances of the ith symbol in x^n. This variant holds important universality properties and is widely known as the Krichevsky–Trofimov estimator [39,40].

We repeat each experiment k = 10,000 times and report the estimated mean, variance, and worst-case performance, as in the Gaussian case. Figure 4 demonstrates the results we achieve. We omit the MLE as it typically results in an unbounded divergence (in cases where at least a single symbol fails to appear).

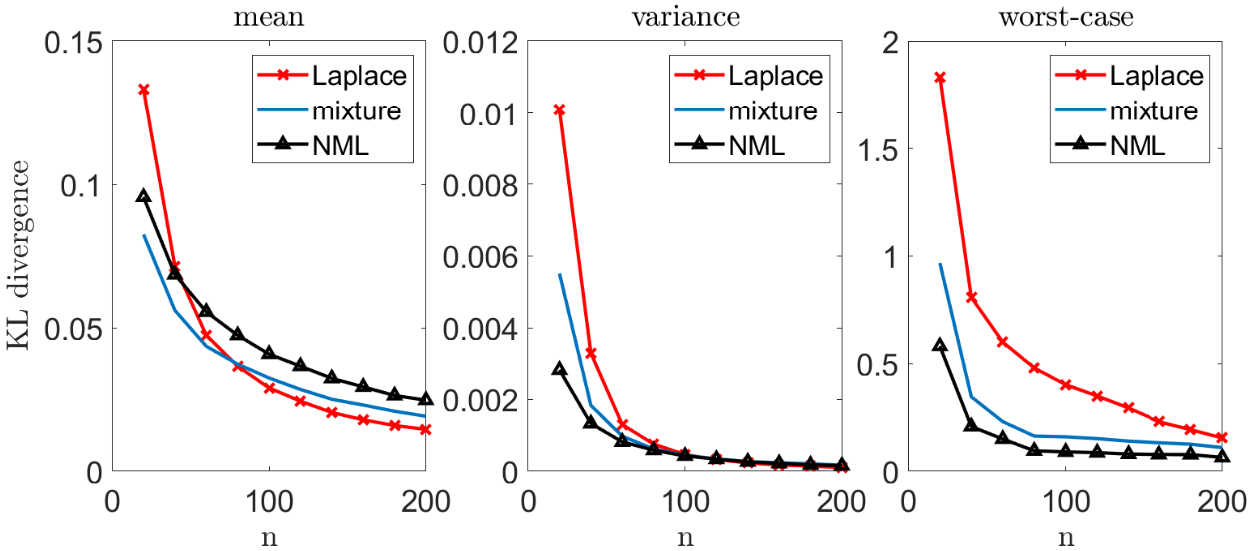

Figure 4. Multinomial inference. In each experiment, we draw n samples from a Zipf's law distribution with $m = 5$ and $s = 1.01$. We evaluate $D_{\text{KL}}(p||q(\cdot|x^n))$ for different estimators. We repeat this experiment 10,000 times and report the mean (**left**), variance (**middle**), and worst-case (**right**) performance.

As in the unknown normal mean problem, we notice that in more challenging setups (small n), our worst-case gain is quite remarkable. This gain narrows down as n increases, and all the estimators converge to the same solution. In addition, we observe a significant gain in expectation when n is small. It is important to emphasize that when the underlying distribution is easier to infer (all $p(x)$ are bounded away from zero, as with the uniform distribution), the advantage of using the minimax approach is less evident (similarly to the large n regime in the Zipf's law example).

7. Large Alphabet Probability Estimation

In the large alphabet regime, we study a multinomial distribution where $m >> n$. This problem has been extensively studied over the years, with many applications ranging from language processing to biological studies [41]. Here, traditional methods like MLE are typically ineffective, as they assign a zero probability to unseen events. Several alternatives have been suggested over the years. In his seminal work, Laplace [42] suggested to add a single count to all events, followed by a maximum likelihood estimator. The work of Laplace was later generalized to a class of add-constant estimators [39], with the important special case of the Krichevsky–Trofimov estimator (as discussed in Section 6). A significant milestone in the history of large alphabet probability estimation was established in the work of Good and Turing [43]. Their approach suggests that unseen events shall be assigned a probability proportional to the number of events that appear once. To this day, Good–Turing estimators are the most commonly used methods in practical problems (see, for example, Section 1.4 in [41]). Despite the great interest in large alphabet estimation, provably-optimal schemes remain elusive [41]. Moreover, the accuracy of existing methods do not allow us to construct practical confidence regions. In fact, Paninski [44] showed that in the large alphabet setup, the minimal expected worst-case divergence is unbounded, and grows like $\log(m/n)$. Therefore, it is quite difficult to define a small enough restrictive model class that contains p with high confidence. In this case, we introduce an alternative approach for the design of $\mathcal{P}_r$, followed by an NML estimator.

The Leave-One-Out Hypothesis Class

Define the convergence rate of $q(\cdot|x^n)$ as $\Delta_p(n) = \mathbb{E}(D_{\text{KL}}(p||q(\cdot|x^n)) - D_{\text{KL}}(p||q(\cdot|x^{n+1})))$. We say that an estimator $q(\cdot;x^n)$ is *proper* if it satisfies, for every p,

A. $\mathbb{E}D_{\text{KL}}(p||q(\cdot|x^n)) < \infty$ for all $n \geq 0$

B. $\Delta_p(n) \geq 0$ for all $n \geq n_0$

C. $\Delta_p(n)$ is monotonically non-increasing for all $n \geq n_0$

The first condition states that the expected loss is finite for any n. The second condition indicates that asymptotically, adding more samples only improves the expected accuracy. The third condition says that the rate of the improvement is non-increasing in the number of samples. For example, the improvement from 100 to 101 samples is greater than the improvement from 1000 to 1001 samples, on the average. We now define the *leave-one-out* model class. Let $x^{n-1}_{[-i]} = \{x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\}$ be the leave-one-out set of x^n, excluding the i^{th} sample. Let $q(\cdot|x^{n-1}_{[-i]})$ be the corresponding proper estimate. The leave-one-out (loo) model class is defined as $\mathcal{P}_{loo} = \{q(\cdot|x^{n-1}_{[-i]})\}_{i=1}^n$. Theorem 6 below establishes that on the average, the accuracy of the best model in $\mathcal{P}_{loo}$ is bounded from above by accuracy of $q(\cdot|x^n)$, plus an additional vanishing overhead term.

Theorem 6. *Let q be a proper estimator. Then,*

$$\mathbb{E}\left(\min_i D_{\text{KL}}\left(p||q\left(\cdot|x^{n-1}_{[-i]}\right)\right)\right) \leq \mathbb{E}(D_{\text{KL}}(p||q(\cdot|x^n))) + o\left(\frac{1}{n}\right). \tag{11}$$

A proof for this Theorem is provided in Appendix E. Notice that the inequality is due to the convexity of the different operators. This means that typically, we expect the inequality to be strict. In other words, given a proper estimator q, Theorem 6 shows that on the average, there exists at least a single model in $\mathcal{P}_{loo}$ that is better than $q(\cdot|x^n)$, up to a vanishing overhead term of $o\left(\frac{1}{n}\right)$. In Appendix F we show that any add-constant (Laplace) estimator satisfies (11). Further, our experiments indicate that the same property holds for the Good–Turing estimator. This motivates the use of these estimators in the design of $\mathcal{P}_r = \mathcal{P}_{loo}$, as suggested by (4).

Let us now demonstrate our suggested scheme. In each experiment, we draw n samples from a multinomial distribution over an alphabet size $m = 1000$. We apply the

Krichevsky–Trofimov estimator, $q(n_i) \propto n_i + 1/2$, and a Good–Turing estimator, following the implementation of Gale [45]. We compare these estimators to our suggested scheme; we construct a loo model class using Good–Turing, followed by an NML estimator. In addition, we compare the NML with a simple uniform average over the loo model class. A comprehensive description of our suggested scheme is provided in Appendix G. To emphasize the difference between the suggested schemes, we compare each estimator with a *natural oracle* $p_{nat}(x^n)$; an estimator who knows the true model p, but is restricted to assign the same probability to symbols that appear the same number of times in x^n. The performance of this oracle serves as a lower bound [41]. Figures 5 and 6 demonstrate the results we achieve for a Zipf's law distribution $p(i) \propto i^{-s}$ with a parameter value of $s = 1.01$ (left) and $s = 1.5$ (center). In addition, we consider a geometric distribution $p(i) = (1-s)^{i-1}s$ with $s = 0.05$ (right). We report the expected difference (regret) between $D_{\mathrm{KL}}(p||q(\cdot|x^n))$ and $D_{\mathrm{KL}}(p||p_{nat}(\cdot|x^n))$ in Figure 5, while the worst-case regret is presented in Figure 6. We omit the uncompetitive performance of the Krichevsky–Trofimov estimator.

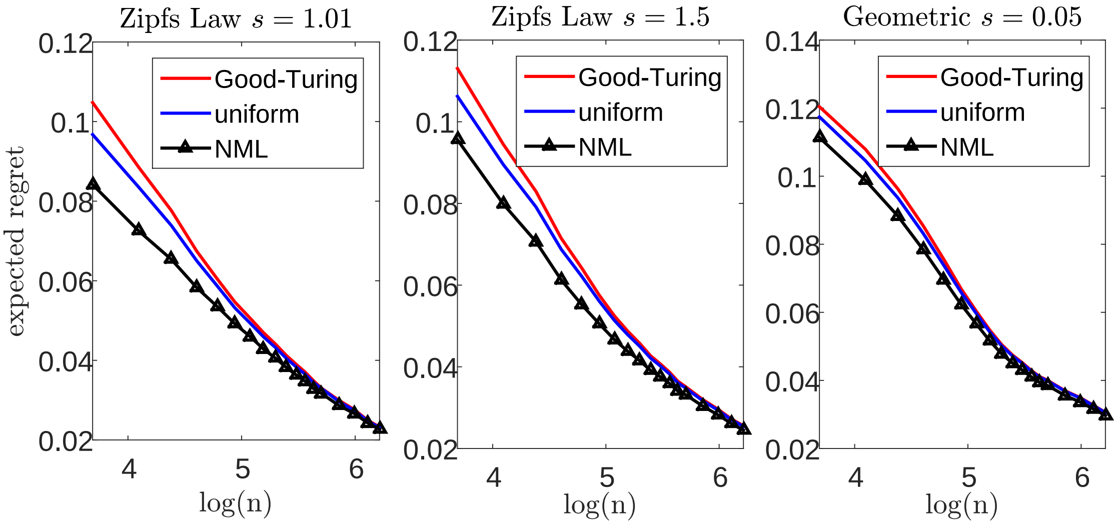

Figure 5. Large alphabet probability estimation- the expected regret (difference) between $D_{\mathrm{KL}}(p||q(\cdot|x^n))$ and the performance of the natural oracle, $D_{\mathrm{KL}}(p||p_{nat}(\cdot|x^n))$.

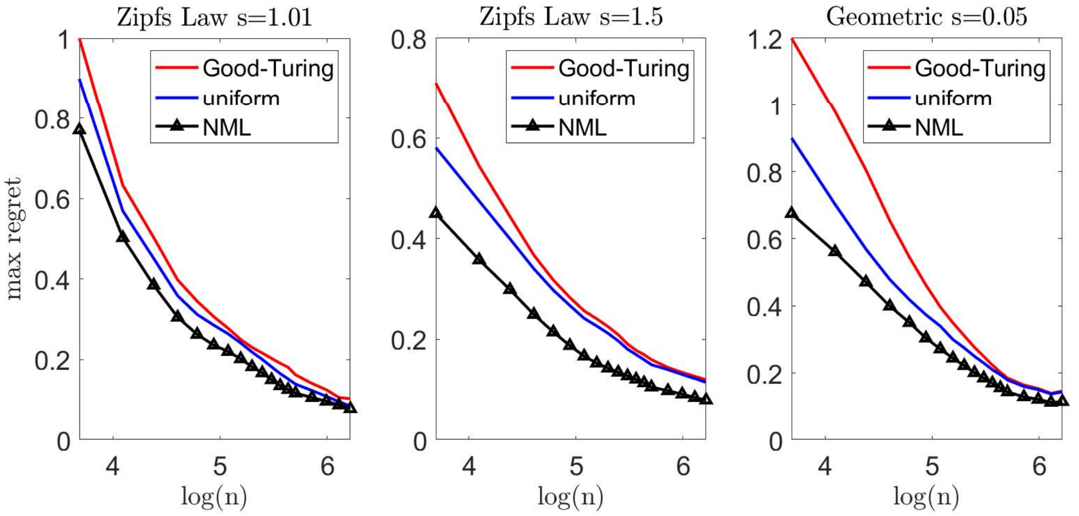

Figure 6. Large alphabet probability estimation—the worst-case regret between $D_{\mathrm{KL}}(p||q(\cdot|x^n))$ and the performance of the natural oracle, $D_{\mathrm{KL}}(p||p_{nat}(\cdot|x^n))$.

As we observe Figure 5, we notice that our suggested NML method outperforms Good–Turing when the alphabet size is relatively small. As n increases, the improvement becomes less evident as the restrictive model class converges to $q(\cdot|x^n)$. Further, we notice that a uniform average over the loo model class is also favorable, but demonstrates a slighter improvement. Finally, we compare the worst-case performance in Figure 6. Here, again, we notice a significant improvement as in the previous experiments. For example, for $n = 40$ and a Zipf's law distribution ($s = 1.5$), the Good–Turing results in a regret of 0.72 bits while the uniform mixture is 0.58 bits and the NML is only 0.43 bits.

8. Real-World Example

Let us now introduce a real-world example. The Wisconsin breast cancer study (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), accesed on 14 June 2021) considers 569 diagnosed tumors, of which 357 are benign (B) and 212 are malignant (M) [46] . Each tumor is characterized by 32 features, including its size, texture, surface, and more. We would like to study the radius of benign tumors and assess its probability function. This probability is of high interest as it allows us, for example, to control type-I error in a future hypothesis testing (the probability of deciding a tumor is malignant, given that it is benign).

The medical domain knowledge suggests that the size of the tumor follows a normal distribution, with different parameters for the B and M tumors. Therefore, the standard approach is to estimate the parameters from the given data. For simplicity, we assume the true variance is known (estimated from the entire population) and focus on the unknown mean.

As in the previous sections, we study the performance of different estimation schemes. We draw n samples from the B class, and apply the MLE and the suggested NML scheme. Notice that we focus on the NML as it is the most robust approach for the modeling assumption (and henceforth most suitable for such a clinical trail). We repeat this experiment 10,000 times for every value of n and report the mean, variance, and worst-case KL divergence between the "true empirical distribution" (based on all the B samples that we have) and each estimator that we examine. Figure 7 demonstrates the results we achieve. As we can see, our suggested approach attains a significantly better worst-case results.

It is important to emphasize that the MLE is the solution to the classical minimax estimation scheme (1), under the assumption that the data is generated from normal distribution (see Section 2). Our approach with the NML relaxes this strong restriction and attains a significant improvement in the worst-case performance.

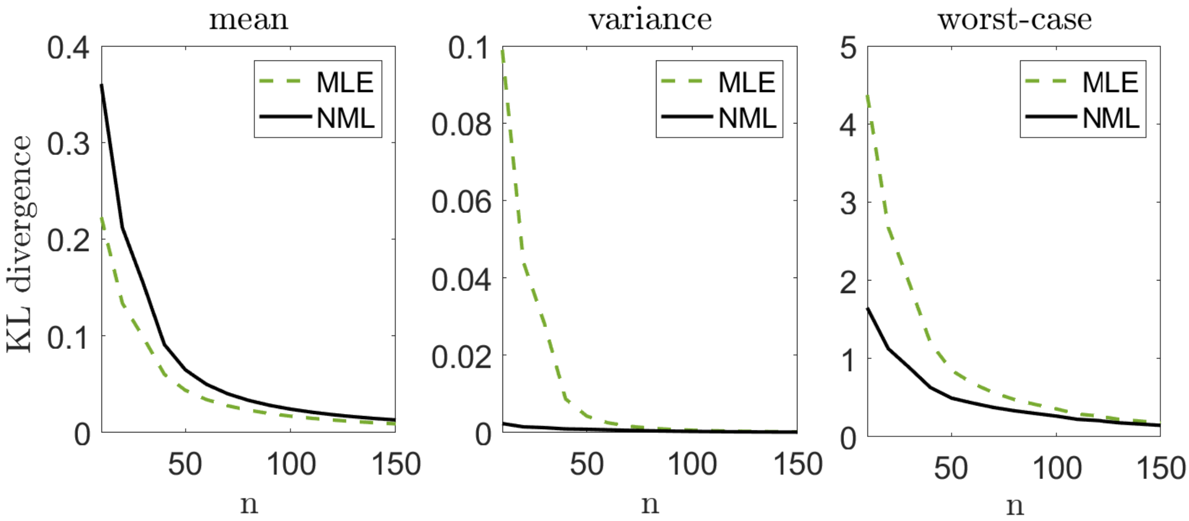

Figure 7. Breast cancer tumor study. Mean, variance, and worst-case performance of different estimators, based on n samples.

9. Discussion and Conclusions

In this work, we study a minimax inference framework. Our suggested scheme considers a class of models, defined by the parametric confidence region of the given samples. Then, we control the worst-case performance within this class. Our formulation relaxes some strong modeling assumptions of the classical minimax framework and considers a robust inference scheme for the complete unknown distribution. The attained solution draws fundamental connections to basic concepts in information theory. We demonstrate the performance of our suggested framework in classical inference problems, including normal and multinomial distributions. In addition, we demonstrate our suggested scheme on more challenging large alphabet probability estimation problems. Finally, we study a real-world breast cancer example. In all of these settings we introduce a significant improvement in the worst-case, at a typically low cost on the average. This demonstrates an "insurance-like" trade-off; we pay a small cost on the average to avoid a great loss if "something bad happens" (that is, the observed samples do not represent the true model well enough).

It is important to emphasize that our suggested scheme is not limited to confidence region model classes. In fact, in many cases, exact confidence regions are difficult to attain, or result in model classes that are too large to control (for example, large alphabet problems with many unseen symbols). In these cases, we consider alternative forms of "reasonable" classes of models. One possible solution is the leave-one-out (LOO) class, discussed in Section 7. Additional alternatives are bootstrap confidence regions, Markov Chain Monte Carlo (MCMC) sampling and others.

Finally, our suggested scheme may be generalized to a supervised learning framework. For example, consider a linear regression problem. The standard approach is to estimate the regression coefficients that best explain the data (typically by least squares analysis). However, notice we may also construct confidence intervals for the sought coefficients. This way, we can define a restricted model class (similarly to Section 3), and seek minimax estimates with respect to it. This idea may be generalized to more complex learning schemes such as deep neural networks. Specifically, we may construct a restricted model class as the vicinity of some class of parameters that the network converges to, and control the corresponding worst-case performance. We consider this direction for our future work.

Author Contributions: Conceptualization, A.P. and M.F.; methodology, A.P. and M.F.; software, A.P.; validation, A.P. and M.F.; formal analysis, A.P. and M.F.; investigation, A.P.; resources, M.F.; data curation, A.P.; writing–original draft preparation, A.P.; writing–review and editing, A.P. and M.F.; visualization, A.P.; supervision, M.F.; project administration, M.F.; funding acquisition, A.P. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Yury Polyanskiy for helpful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. A Proof of Theorem 2

Let $p_\theta \in \mathcal{P}(\Theta)$ be an unknown probability distribution. Let $\mathcal{P}(\Theta_r)$ be a restrictive model class that corresponds to a restricted set of parameters Θ_r. We would like to solve the minimax problem

$$\min_q \sup_{\theta' \in \Theta_r} \sup_{\theta \in \Theta} \int p_\theta(x) \log \frac{p_{\theta'}(x)}{q(x)} dx = \min_q \sup_{\theta \in \Theta} \left(D_{\mathrm{KL}}(p_\theta || q) - f_{\Theta_r}(p_\theta) \right), \tag{A1}$$

where $f_{\Theta_r}(p_\theta) \triangleq \inf_{\theta' \in \Theta_r} D_{\mathrm{KL}}(p_\theta || p_{\theta'})$. Let us now define an equivalent problem to (A1),

$$\min_q \sup_{\pi(\theta)} \int_{\theta \in \Theta} \pi(\theta)(D_{\mathrm{KL}}(p_\theta || q) - f_{\Theta_r}(p_\theta))d\theta. \tag{A2}$$

where $\pi(\theta)$ is a weight function satisfying $\pi(\theta) \geq 0$ and $\int_{\theta \in \Theta} \pi(\theta)d\theta = 1$. Notice that the equivalence holds since for a fixed q, a weight function which puts all probability mass on the worst θ is a least favorable function. Let us now change the order of the minimum and supremum, similarly to the redundancy-capacity theorem (3).

Let M_Θ be the collection of all measures (on X) that can be obtained as mixtures of the p_θ measures. Let $\bar{M}_\Theta$ be the closure of M_Θ. Define

$$\begin{aligned} \psi(q, \pi(\theta)) &= \int_{\theta \in \Theta} \pi(\theta)(D_{\mathrm{KL}}(p_\theta || q) - f_{\Theta_r}(p_\theta))d\theta \\ \bar{V} &= \min_q \sup_{\pi(\theta)} \psi(q, \pi(\theta)) \\ \underline{V} &= \sup_{\pi(\theta)} \min_q \psi(q, \pi(\theta)) \\ \tilde{V} &= \sup_{\pi(\theta)} \min_{q \in \bar{M}_\Theta} \psi(q, \pi(\theta)). \end{aligned} \tag{A3}$$

Notice that if $F(\Theta, \Theta_r) < \infty$, then $\tilde{V} = F(\Theta, \Theta_r)$. This holds as for every π, the mixture $q \in \bar{M}_\Theta$, which minimizes $\psi(q, \pi(\theta))$ is q_π (see, for example, [27]). Therefore, we would like to show that $\bar{V} = \tilde{V}$. Here, we follow the steps of [27] and Sion's minimax theorem [47].

Theorem A1 (**Sion's Minimax Theorem [47]**)**.** *Let $\mathcal{U}$ be a compact convex subset of a linear topological space and $\mathcal{V}$ be a convex subset of a linear topological space. If $f(u, v)$ is a real-valued function on $\mathcal{U} \times \mathcal{V}$ with*

1. *$f(u, \cdot)$ is upper semi-continuous and quasi-concave on $\mathcal{V}$ for all $u \in \mathcal{U}$*
2. *$f(\cdot, v)$ is lower semi-continuous and quasi-convex on $\mathcal{U}$ for all $v \in \mathcal{V}$*

then, $\min_{u \in \mathcal{U}} \sup_{v \in \mathcal{V}} f(u, v) = \sup_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} f(u, v)$.

Let us first assume that $\mathcal{P}(\Theta)$ is *uniformally tight*. In other words, for every $\epsilon > 0$ there exists a compact set $K \subseteq X$ such that $p_\theta(K) > 1 - \epsilon$ for all $p_\theta \in \mathcal{P}(\Theta)$. Haussler showed that if $\mathcal{P}(\Theta)$ is uniformally tight, then it is totally bounded, and thus $\bar{M}_\Theta$ is compact [27]. Therefore, for $\tilde{V} < \infty$ we have that

$$\begin{aligned} \bar{V} = \min_q \sup_{\pi(\theta)} \psi(q, \pi(\theta)) &\overset{(a)}{\leq} \min_{q \in \bar{M}_\Theta} \sup_{\pi(\theta)} \psi(q, \pi(\theta)) \overset{(b)}{=} \\ \sup_{\pi(\theta)} \min_{q \in \bar{M}_\Theta} \psi(q, \pi(\theta)) &\overset{(c)}{=} \sup_{\pi(\theta)} \min_q \psi(q, \pi(\theta)) = \underline{V} \end{aligned} \tag{A4}$$

where:

(a) follows from definition
(b) follows from Sion's minimax theorem
(c) for every $\pi(\theta)$, the distribution q which minimizes $\psi(q, \pi(\theta))$ is a mixture distribution [27]. Notice that $f_{\Theta_r}(p_\theta)$ does not depend on q.

This means that $\bar{V} \leq \tilde{V} = \underline{V}$. On the other hand, it is easy to verify that $\bar{V} \geq \underline{V}$ due to the max-min inequality [48]. This means that $\bar{V} = \tilde{V}$ as desired.

Let us now assume that $\mathcal{P}(\Theta)$ that is not uniformally tight. Haussler showed that in this case, $\sup_{\pi(\theta)} \min_q \int_{\theta \in \Theta} \pi(\theta) D_{\mathrm{KL}}(p_\theta || q)d\theta = \infty$ (Lemma 4 in [27]). Therefore, given that $f_{\Theta_r}(p_\theta) < \infty$ for all $\theta \in \Theta$ (as Θ_r is bounded), we have that $\underline{V} = \infty$. However, this contradicts $\tilde{V} < \infty$.

Appendix B. A Proof of Theorem 3

Assume that Θ_r is bounded. Let $p^*_\theta(x) = \text{argmin}_{\theta' \in \Theta_r} D_{\text{KL}}(p_\theta || p_{\theta'})$. Then,

$$F(\Theta, \Theta_r) \triangleq \sup_{\pi(\theta)} \int_{\theta \in \Theta} \pi(\theta) \left(D_{\text{KL}}(p_\theta || q_\pi) - \inf_{\theta' \in \Theta_r} D_{\text{KL}}(p_\theta || p_{\theta'}) \right) d\theta = \tag{A5}$$

$$\sup_{\pi(\theta)} \int_{\theta \in \Theta} \pi(\theta)(D_{\text{KL}}(p_\theta || q_\pi) - D_{\text{KL}}(p_\theta || p^*_\theta)) d\theta =$$

$$\sup_{\pi(\theta)} \int_{\theta \in \Theta} \pi(\theta) \left(\int_x p_\theta(x) \log \frac{p^*_\theta(x)}{q_\pi(x)} \right) d\theta.$$

Similarly to the channel capacity problem, this optimization does not hold a closed form solution in the general case. Therefore, we introduce an alternating projection algorithm, similar in spirit to the Blahut–Arimoto algorithm [49,50]. For this purpose, we apply the well-known alternating maximization theorem (Lemma 9.4 and 9.5 in [51]).

Lemma A1 **(The Alternating Maximization Theorem [51]).** *Let $f(u_1, u_2)$ be a real, concave and bounded-from-above function that is continuous and has continuous partial derivatives. Let $\mathcal{U}_1$ and $\mathcal{U}_2$ be two convex sets. Consider an optimization problem*

$$\sup_{u_1 \in \mathcal{U}_1,\, u_2 \in \mathcal{U}_2} f(u_1, u_2) = f^*. \tag{A6}$$

Denote $c_2(u_1) = \sup_{u_2 \in \mathcal{U}_2} f(u_1, u_2)$ and $c_1(u_2) = \sup_{u_1 \in \mathcal{U}_1} f(u_1, u_2)$. The alternating maximization algorithm is an iterative process where in each iteration k we maximize over one of the variables. Let (u_1^0, u_2^0) be an arbitrary starting point in $\mathcal{U}_1 \times \mathcal{U}_2$. For $k \geq 0$ let $(u_1^k, u_2^k) = (c_1(u_2^{k-1}), c_2(c_1(u_2^{k-1})))$ and let $f^k = f(u_1^k, u_2^k)$. Assume that $c_2(u_1) \in \mathcal{U}_2$ and $c_1(u_2) \in \mathcal{U}_1$ are unique for all $u_1 \in \mathcal{U}_1$ and $u_2 \in \mathcal{U}_2$, then $\lim_{k \to \infty} f^k = f^*$.

Let us reformulate (A5) according to the requirements of Lemma A1. First, we multiply the numerator and the denominator in the log by $\psi^*(\theta, x) = \frac{\pi(\theta) p_\theta(x)}{q_\pi(x)}$. We attain

$$\sup_{\pi(\theta)} \int_{\theta \in \Theta} \pi(\theta) \int_x p_\theta(x) \log \frac{\psi^*(\theta, x)}{\pi(\theta)} \frac{p^*_\theta(x)}{p_\theta(x)} dx d\theta. \tag{A7}$$

Define the following maximization problem:

$$\sup_{\phi(\theta) \in \mathcal{A}_1} \sup_{\psi(\theta,x) \in \mathcal{A}_2(x)} \int_{\theta \in \Theta} \phi(\theta) \int_x p_\theta(x) \log \frac{\psi(\theta, x)}{\phi(\theta)} \frac{p^*_\theta(x)}{p_\theta(x)} dx d\theta \tag{A8}$$

where $\mathcal{A}_1 = \{\phi(\theta) | \int_{\theta \in \Theta} \phi(\theta) d\theta = 1,\ \phi(\theta) \geq 0\}$ and $\mathcal{A}_2(x) = \{\psi(\theta, x) | \int_{\theta \in \Theta} \psi(\theta, x) d\theta = 1,\ \psi(\theta, x) \geq 0\}$ are convex sets. We now show that (A8) satisfies the conditions of the alternating maximization algorithm. First, we notice that our objective is real, concave, and bounded from above (as $F(\Theta, \Theta_r) < \infty$). Lemmas A2 and A3 below show that there exists a unique maximum for every $\phi(\theta) \in \mathcal{A}_1$ and $\psi(\theta, x) \in \mathcal{A}_2$, similarly to the Blahut–Arimoto algorithm for the channel capacity problem.

Lemma A2.

$$\underset{\psi(\theta,x) \in \mathcal{A}_2(x)}{\text{argmax}} \int_{\theta \in \Theta} \phi(\theta) \int_x p_\theta(x) \log \frac{\psi(\theta, x)}{\phi(\theta)} \frac{p^*_\theta(x)}{p_\theta(x)} dx d\theta = \frac{\phi(\theta) p_\theta(x)}{\int_{\theta' \in \Theta} \phi(\theta') p_{\theta'}(x) d\theta'}.$$

Proof. Define $w(x) = \int_{\theta' \in \Theta} \phi(\theta') p_{\theta'}(x) d\theta'$. Further, define $\psi'(\theta, x) = \frac{\phi(\theta) p_\theta(x)}{w(x)}$. We have that

$$\begin{aligned}&\int_{\theta\in\Theta}\phi(\theta)\int_x p_\theta(x)\log\frac{\psi'(\theta,x)}{\phi(\theta)}\frac{p^*_\theta(x)}{p_\theta(x)}dxd\theta - \int_{\theta\in\Theta}\phi(\theta)\int_x p_\theta(x)\log\frac{\psi(\theta,x)}{\phi(\theta)}\frac{p^*_\theta(x)}{p_\theta(x)}dxd\theta = \\ &\int_{\theta\in\Theta}\phi(\theta)\int_x p_\theta(x)\log\frac{\psi'(\theta,x)}{\psi(\theta,x)}dxd\theta = \int_{\theta\in\Theta}\int_x w(x)\psi'(\theta,x)\log\frac{\psi'(\theta,x)}{\psi(\theta,x)}dxd\theta = \\ &\int_x w(x) D_{\mathrm{KL}}(\psi'(\theta,x)||\psi(\theta,x)) \geq 0\end{aligned} \tag{A9}$$

where the second equality follows from the definition of $\psi'(\theta, x)$ above. □

Lemma A3.

$$\underset{\phi(\theta)\in\mathcal{A}_1}{\operatorname{argmax}} \int_{\theta\in\Theta}\phi(\theta)\int_x p_\theta(x)\log\frac{\psi(\theta,x)}{\phi(\theta)}\frac{p^*_\theta(x)}{p_\theta(x)}dxd\theta = \frac{\prod_x \tilde{\psi}(\theta,x)^{p_\theta(x)}}{\int_{\theta'\in\Theta}\prod_x \tilde{\psi}(\theta',x)^{p_{\theta'}(x)}d\theta'}$$

where $\tilde{\psi}(\theta, x) = \psi(\theta, x)\frac{p^*_\theta(x)}{p_\theta(x)}$

Proof. We apply calculus of variations and attain the optimality condition. Define the Lagrangian as

$$\begin{aligned}\mathcal{L} = \int_{\theta\in\Theta}\phi(\theta)\int_x p_\theta(x)\log\frac{\psi(\theta,x)}{\phi(\theta)}\frac{p^*_\theta(x)}{p_\theta(x)}dxd\theta - \lambda\left(\int_{\theta\in\Theta}\phi(\theta)d\theta - 1\right) = \\ \int_{\theta\in\Theta}\phi(\theta)\left(\int_x p_\theta(x)\log\frac{\psi(\theta,x)}{\phi(\theta)}\frac{p^*_\theta(x)}{p_\theta(x)}dx - \lambda\right)d\theta + \lambda\end{aligned} \tag{A10}$$

Then, the Euler–Lagrange condition requires that the partial derivative of the integrand with respect to $\phi(\theta)$ is zero. This yields

$$\int_x p_\theta(x)\left(\log\frac{\tilde{\psi}(\theta,x)}{\phi(\theta)} - 1\right)dx - \lambda = 0$$

where $\tilde{\psi}(\theta, x) = \psi(\theta, x)\frac{p^*_\theta(x)}{p_\theta(x)}$. Therefore, $\phi(\theta) \propto \prod_x \tilde{\psi}(\theta, x)^{p_\theta(x)}$ as desired. □

Appendix C

Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a d-dimensional Gaussian vector where μ is unknown and Σ is known. Let x^n be a collection of n i.i.d. draws from X. Define a $100(1-\alpha)\%$ confidence region for μ as a collection $\mathcal{M}_r = \{\mu | (\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) \leq \frac{1}{n}\chi^2_d(1-\alpha)\}$, as defined in Section 5. As previously established (and shown in [30]), our minimax problem is equivalent to a channel capacity problem $X = M + Z$ where $M \in \mathcal{M}_r$ and $Z \sim \mathcal{N}(0, \Sigma)$, independent of M.

Let $\Sigma = USU^T$ be the singular value decomposition (SVD) of Σ and $A^T = S^{-0.5}U^T$ be the diagonalizing transformation of Z, such that $Cov(S^{-0.5}U^TZ) = I_d$. Define $X' = A^TX = A^TM + A^TZ$. Notice that $I(M; X) = I(A^TM; A^TX)$, given that A is invertible. Let us study $I(A^TM; A^TX)$. We have that $Z' = A^TZ \sim \mathcal{N}(0, I_d)$ and

$$(A^T\bar{x} - A^T\mu)^T(A^T\bar{x} - A^T\mu) = (\bar{x}-\mu)^TAA^T(\bar{x}-\mu) = (\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) \leq \frac{1}{n}\chi^2_d(1-\alpha).$$

This means that $\mu' = A^T\mu$ satisfies $\mu' \in \mathcal{M}'_r$ where

$$\mathcal{M}'_r = \{\mu' | (A^T\bar{x} - \mu')^T(A^T\bar{x} - \mu') \leq \frac{1}{n}\chi^2_d(1-\alpha)\}.$$

Therefore, the channel $X' = M' + Z'$ is again a standard amplitude constrained Gaussian channel, where the input is now a sphere around $A^T\bar{x}$, instead of $\bar{x}$.

Finally, we define $X'' = X' - A^T\bar{x} = M' - A^T\bar{x} + Z'$. Let $M'' = M' - A^T\bar{x}$ and $\mu'' = \mu' - A^T\bar{x}$. We have that $M'' \in \mathcal{M}''_r$ where $\mathcal{M}''_r = \{\mu' | \mu''^T \mu'' \leq \frac{1}{n}\chi^2_d(1-\alpha)\}$. Further, we have that $I(M;X) = I(M'';X'')$. This means that our original minimax problem is equivalent to a standard, centered, amplitude-constrained channel capacity problem.

Appendix D

Let q_π be convex combination of $\mathcal{V}(\mathcal{P}_r)$. Assume that q_π satisfies $D_{\text{KL}}(p_j||q_\pi) = c$ for all $p_j \in \mathcal{V}(\mathcal{P}_r)$. Notice that every $p^* \in \mathcal{P}_r$ can be described as a convex combination of the vertexes $\mathcal{V}(\mathcal{P}_r)$ (since $\mathcal{P}_r$ is a convex set). Therefore, we have that $D_{\text{KL}}(p^*||q_\pi) = D_{\text{KL}}(\sum a_j p_j||q_\pi) \leq \sum a_j D_{\text{KL}}(p_j||q_\pi) = c$. This means that q_π satisfies that optimality conditions over the set $\mathcal{P}_r$. In other words, the solution to the minimax problem (2) over the confidence region $\mathcal{P}_r$ is attained by solving the capacity-redundancy problem (3) over its finite set of vertexes $\mathcal{V}(\Theta_r)$.

Appendix E. A Proof for Theorem 6

Let us first introduce the following Lemma.

Lemma A4. *Let q be a proper estimator. Then, $\Delta_p(n) \leq o\left(\frac{1}{n}\right)$.*

Proof. Applying a simple mathematical induction,

$$\begin{aligned} \mathbb{E}D_{\text{KL}}(p||q(\cdot|x^n)) =& \Delta_p(n) + \mathbb{E}D_{\text{KL}}(p||q(\cdot|x^{n+1})) = \\ & \Delta_p(n) + \Delta_p(n+1) + \mathbb{E}D_{\text{KL}}(p||q(\cdot|x^{n+2})) = \\ & \lim_{m\to\infty}\left(\sum_{k=n}^{m}\Delta_p(k) + \mathbb{E}D_{\text{KL}}(p||q(\cdot|x^m))\right) < \infty \end{aligned} \tag{A11}$$

where the finiteness is due to Condition A. Therefore, we necessarily have that the series $\sum_{k=n}^{m}\Delta_p(k) < \infty$ converges. Conditions B and C state that $\Delta_p(k)$ is non-negative and monotonically non-increasing. Therefore, simple calculus shows that $\lim_{m\to\infty} m\Delta_p(m) = 0$, which implies that $\Delta_p(m) \leq o\left(\frac{1}{m}\right)$. □

Now, define the *leave-one-out* model class as $\mathcal{P}_{loo} = \left\{q\left(\cdot|x^{n-1}_{[-i]}\right)\right\}_{i=1}^{n}$. We have that

$$\begin{aligned} &\mathbb{E}\left(\min_i D_{\text{KL}}\left(p||q\left(\cdot|x^{n-1}_{[-i]}\right)\right)\right) \leq \min_i \mathbb{E}\left(D_{\text{KL}}\left(p||q\left(\cdot|x^{n-1}_{[-i]}\right)\right)\right) = \\ &\mathbb{E}\left(D_{\text{KL}}\left(p||q\left(\cdot|x^{n-1}\right)\right)\right) = \mathbb{E}(D_{\text{KL}}(p||q(\cdot|x^n))) + \Delta_p(n) \leq \\ &\mathbb{E}(D_{\text{KL}}(p||q(\cdot|x^n))) + o\left(\frac{1}{n}\right) \end{aligned} \tag{A12}$$

where the first inequality is due to Jensen inequality and the second inequality is due to Lemma A4.

Appendix F

Define the expected convergence rate of an estimator $q(\cdot|x^n)$ as

$$\Delta_p(n) = \mathbb{E}\Big(D_{\text{KL}}(p||q(\cdot|x^n)) - D_{\text{KL}}(p||q(\cdot|x^{n+1}))\Big). \tag{A13}$$

The add-constant estimator follows $q(i|x^n) = q(n_i) = \frac{n_i+\beta}{n+m\beta}$ where β is the added constant. We have

$$\begin{aligned}\mathbb{E}D_{\mathrm{KL}}(p||q(\cdot|x^n)) =& \mathbb{E}\sum_i p(i)\log\frac{p(i)}{q(n_i)} = \\ & H(p) - \sum_i p(i)(\mathbb{E}\log(n_i+\beta) - \log(n+m\beta)) = \\ & H(p) + \log(n+m\beta) - \sum p(i)\mathbb{E}\log(n_i+\beta).\end{aligned} \tag{A14}$$

For $X \sim \mathrm{Bin}(n,\theta)$ we have that $\mathbb{E}\log(X+a) = \log(\theta n + a) - \frac{1-\theta}{2\theta n} + O\left(\frac{1}{n^2}\right)$. Further, $n_i = \sum_{j=1}^{n} \mathbb{1}\{x_j = 1\} \sim \mathrm{Bin}(n, p(i))$. Therefore, $\mathbb{E}\log(n_i+\beta) = \log(np(i)+\beta) - \frac{1-p(i)}{2np(i)} + O\left(\frac{1}{n^2}\right)$, leading to

$$\begin{aligned}\mathbb{E}D_{\mathrm{KL}}(p||q(\cdot|x^n)) = H(p) + \log(n+m\beta) - \sum p(i)\log(np(i)+\beta) + \\ \frac{m-1}{2n} + O\left(\frac{1}{n^2}\right).\end{aligned} \tag{A15}$$

Plugging this result to (A13), we obtain

$$\begin{aligned}\Delta_p(n) =& \sum_i p(i)\log\left(\frac{n+m\beta}{np(i)+\beta}\cdot\frac{(n+1)p(i)+\beta}{(n+1)+m\beta}\right) + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right) < \\ & \sum_i p(i)\log\frac{(n+1)p(i)+\beta}{np(i)+\beta} + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right) = \\ & -\sum_i p(i)\log\left(1 - \frac{p(i)}{(n+1)p(i)+\beta}\right) + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right) \leq \\ & \sum_i p(i)\frac{p(i)}{(n+1)p(i)+\beta}\cdot\frac{(n+1)p(i)+\beta}{(n+1)p(i)+\beta-p(i)} + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right) = \\ & \sum_i \frac{p^2(i)}{np(i)+\beta} + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right) \leq \sum_i \frac{p^2(i)}{np(i)} + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right) = \\ & \frac{1}{n} + \frac{m-1}{2n(n+1)} + O\left(\frac{1}{n^2}\right)\end{aligned} \tag{A16}$$

where the first inequality is due to $n + m\beta < n + 1 + m\beta$ and the second inequality is $-\log(1-x) \leq \frac{x}{1-x}$ for all $0 < x < 1$. Finally, we have

$$\begin{aligned}& \mathbb{E}\left(\min_i D_{\mathrm{KL}}\left(p||q\left(\cdot|x^{n-1}_{[-i]}\right)\right)\right) \leq \min_i \mathbb{E}D_{\mathrm{KL}}\left(p||q\left(\cdot|x^{n-1}_{[-i]}\right)\right) = \\ & \mathbb{E}D_{\mathrm{KL}}\left(p||q\left(\cdot|x^{n-1}\right)\right) = \Delta_p(n-1) + \mathbb{E}D_{\mathrm{KL}}(p||q(\cdot|x^n)) \leq \\ & \mathbb{E}D_{\mathrm{KL}}(p||q(\cdot|x^n)) + \frac{1}{n-1} + \frac{m-1}{2n(n-1)} + O\left(\frac{1}{n^2}\right)\end{aligned} \tag{A17}$$

where the first inequality is due to Jensen inequality, the first equality is since x^n are i.i.d. draws, and the last inequality is due to (A16).

Appendix G

One of the basic properties of most widely used finite alphabet estimators (MLE, add-constant, Good–Turing, and others) is the *natural* assumption; symbols that appear the same number of times are assigned the same probability estimate. This is not surprising, as it is easy to show that natural estimators maximize the expected estimation accuracy for a given set of samples (for example, see the work of Orlitsky and Suresh [41]). A natural estimator has $k-1$ degrees of freedom where k is the number of symbols with a unique frequency values n_i (the cardinality of the *frequency of frequencies*, as denoted by Good [43]).

Our suggested estimation scheme is also in the natural domain. First, we choose a natural estimator (for example, Good–Turing). Then, given a set of samples x^n, we identify sets of symbols with the same frequency values n_i. Denote the number of unique frequencies as k. Define the mass of all the symbols with the same frequency r as $p^{mass}(r|x^n) = \sum_{n_i=r} p(i)$ for $r = 0, ..., k$. We construct a leave-one-out (loo) model class by first excluding a single sample at a time and applying our chosen estimator, $q(i|x^{n-1})$. Then, we set $\hat{p}^{mass}_{loo}(r|x^n) = \sum_{n_i=r} q(i|x^{n-1})$. In words, the loo estimator of the r^{th} mass is attained by excluding a single sample, applying the chosen estimator, and accumulating the estimates of all the symbols of the original mass r. Finally, the estimate of a single symbol in a mass simply $\hat{p}_{loo}(i|x^n) = \hat{p}^{mass}_{loo}(n_i|x^n)/\sum_j \mathbb{1}\{n_j = n_i\}$. Notice that the size of the model class is k (and not n).

For example, consider the set $x^n = \{1,1,1,2,2,2,3,3\}$ and a maximum likelihood estimator (which is simpler to illustrate then Good-Turing). We would like to find the loo class given x^n. We seek loo estimates for symbols that appear twice, $p^{mass}(r=2|x^n)$, for $x=3$, and three times, $p^{mass}(r=3|x^n)$, for $x=1,2$. We first remove the first symbol $x_1 = 1$, and get $q(X=1|x^{n-1}_{[-1]}) = q(X=3|x^{n-1}_{[-1]}) = 2/7$. This leads to $\hat{p}^{mass}_{loo}(r=2|x^n) = 2/7$ and $\hat{p}_{loo}(r=3|x^n) = 5/7$. Notice that we get the same estimates as we remove $x_4 = 2$. Finally, by removing $x_7 = 3$ we attain $q\left(X=1|x^{n-1}_{[-7]}\right) = q\left(X=2|x^{n-1}_{[-7]}\right) = 3/7$, leading to $\hat{p}^{mass}_{loo}(r=2|x^n) = 1/7$ and $\hat{p}^{mass}_{loo}(r=3|x^n) = 6/7$. Therefore, our corresponding loo model class $\hat{p}_{loo}$ consists of two estimates, $[2.5/7, 2.5/7, 2/7]^T$ and $[3/7, 3/7, 1/7]^T$ and its cardinality is $k=2$.

References

1. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
2. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
3. Stein, C.M. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* **1981**, *9*, 1135–1151. [CrossRef]
4. Ghosh, M. Uniform approximation of minimax point estimates. *Ann. Math. Stat.* **1964**, *35*, 1031–1047. [CrossRef]
5. Donoho, D.L.; Liu, R.C.; MacGibbon, B. Minimax risk over hyperrectangles, and implications. *Ann. Stat.* **1990**, *18*, 1416–1437. [CrossRef]
6. Bickel, P. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Stat.* **1981**, *9*, 1301–1309. [CrossRef]
7. Marchand, É.; Perron, F. On the minimax estimator of a bounded normal mean. *Stat. Probab. Lett.* **2002**, *58*, 327–333. [CrossRef]
8. Lanckriet, G.R.; Ghaoui, L.E.; Bhattacharyya, C.; Jordan, M.I. A robust minimax approach to classification. *J. Mach. Learn. Res.* **2002**, *3*, 555–582.
9. Eban, E.; Mezuman, E.; Globerson, A. Discrete chebyshev classifiers. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; PMLR: Beijing, China, 2014; pp. 1233–1241.
10. Razaviyayn, M.; Farnia, F.; Tse, D. Discrete rényi classifiers. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 3276–3284.
11. Farnia, F.; Tse, D. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 4240–4248.
12. Bennett, W.R. Spectra of Quantized Signals. *Bell Syst. Tech. J.* **1948**, *27*, 446–472. [CrossRef]
13. Nisar, M.D. *Minimax Robustness in Signal Processing for Communications*; Shaker Verlag GmbH: Aachen, Germany, 2011.
14. Kassam, S.A.; Poor, H.V. Robust techniques for signal processing: A survey. *Proc. IEEE* **1985**, *73*, 433–481. [CrossRef]
15. Merhav, N.; Feder, M. Universal prediction. *IEEE Trans. Inf. Theory* **1998**, *44*, 2124–2147. [CrossRef]
16. Verdu, S.; Poor, H. On minimax robustness: A general approach and applications. *IEEE Trans. Inf. Theory* **1984**, *30*, 328–340. [CrossRef]
17. Takeuchi, J.I.; Barron, A.R. Robustly minimax codes for universal data compression. In Proceedings of the ISITA, Mexico City, Mexico, 14–16 October 1998.
18. Rissanen, J. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Trans. Inf. Theory* **2001**, *47*, 1712–1717. [CrossRef]
19. Grünwald, P. The safe bayesian. In *International Conference on Algorithmic Learning Theory*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 169–183.
20. Harremoës, P.; Tishby, N. The information bottleneck revisited or how to choose a good distortion measure. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 566–570.

21. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
22. Painsky, A.; Wornell, G. On the universality of the logistic loss function. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 936–940.
23. Painsky, A.; Wornell, G.W. Bregman Divergence Bounds and Universality Properties of the Logarithmic Loss. *IEEE Trans. Inf. Theory* **2019**, *66*, 1658–1673. [CrossRef]
24. Altman, D.; Machin, D.; Bryant, T.; Gardner, M. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
25. Gallager, R.G. *Information Theory and Reliable Communication*; Springer: Berlin/Heidelberg, Germany, 1968; Volume 588.
26. Kemperman, J. On the Shannon capacity of an arbitrary channel. In *Indagationes Mathematicae (Proceedings)*; Elsevier: North-Holland, The Netherlands, 1974; Volume 77, pp. 101–115.
27. Haussler, D. A general minimax result for relative entropy. *IEEE Trans. Inf. Theory* **1997**, *43*, 1276–1280. [CrossRef]
28. Feder, M.; Polyanskiy, Y. Sequential prediction under log-loss and misspecification. *arXiv* **2021**, arXiv: 2102.00050.
29. Shtarkov, Y.M. Universal sequential coding of single messages. *Probl. Inform. Transm.* **1988**, *23*, 3–17.
30. Raginsky, M. On the information capacity of Gaussian channels under small peak power constraints. In Proceedings of the 2008 46th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 23–26 September 2008; pp. 286–293.
31. Chan, T.H.; Hranilovic, S.; Kschischang, F.R. Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs. *IEEE Trans. Inf. Theory* **2005**, *51*, 2073–2088. [CrossRef]
32. Dytso, A.; Al, M.; Poor, H.V.; Shitz, S.S. On the capacity of the peak power constrained vector Gaussian channel: An estimation theoretic perspective. *IEEE Trans. Inf. Theory* **2019**, *65*, 3907–3921. [CrossRef]
33. Glaz, J.; Sison, C.P. Simultaneous confidence intervals for multinomial proportions. *J. Stat. Plan. Inference* **1999**, *82*, 251–262. [CrossRef]
34. Goodman, L.A. Simultaneous confidence intervals for contrasts among multinomial populations. *Ann. Math. Stat.* **1964**, *35*, 716–725. [CrossRef]
35. Quesenberry, C.P.; Hurst, D. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* **1964**, *6*, 191–195. [CrossRef]
36. Powers, D.M. Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*; ACL: New York, NY, USA, 1998 .
37. Okuyama, K.; Takayasu, M.; Takayasu, H. Zipf's law in income distribution of companies. *Phys. A Stat. Mech. Appl.* **1999**, *269*, 125–131. [CrossRef]
38. Saichev, A.I.; Malevergne, Y.; Sornette, D. *Theory of Zipf's Law and Beyond*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009; Volume 632.
39. Krichevsky, R.; Trofimov, V. The performance of universal encoding. *IEEE Trans. Inf. Theory* **1981**, *27*, 199–207. [CrossRef]
40. Witten, I.H.; Bell, T.C. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inf. Theory* **1991**, *37*, 1085–1094. [CrossRef]
41. Orlitsky, A.; Suresh, A.T. Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2015; pp. 2143–2151.
42. Laplace, P.S. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the Fifth French Edition of 1825 with Notes by the Translator*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1825; Volume 13.
43. Good, I.J. The population frequencies of species and the estimation of population parameters. *Biometrika* **1953**, *40*, 237–264. [CrossRef]
44. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
45. Gale, W.A.; Sampson, G. Good-turing frequency estimation without tears. *J. Quant. Linguist.* **1995**, *2*, 217–237. [CrossRef]
46. Mangasarian, O.L.; Street, W.N.; Wolberg, W.H. Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.* **1995**, *43*, 570–577. [CrossRef]
47. Sion, M. On general minimax theorems. *Pac. J. Math.* **1958**, *8*, 171–176. [CrossRef]
48. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
49. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473. [CrossRef]
50. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20. [CrossRef]
51. Yeung, R.W. *Information Theory and Network Coding*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.