A geometric perspective on guesswork

A Geometric Perspective on Guesswork

Ahmad Beirami,[†‡] Robert Calderbank,[†] Mark Christiansen,[§] Ken Duffy,[§] Ali Makhdoumi,[‡] Muriel Médard[‡]

[†]Department of Electrical and Computer Engineering, Duke University, USA
[‡]Research Laboratory of Electronics, Massachusetts Institute of Technology, USA
[§]Hamilton Institute, National University of Ireland Maynooth, Ireland

Emails: [†]{ahmad.beirami, robert.calderbank}@duke.edu, [‡]{beirami, makhdoum, medard}@mit.edu,
[§]{mark.christiansen, ken.duffy}@nuim.ie

***Abstract*—Guesswork is the position at which a random string drawn from a given probability distribution appears in the list of strings ordered from the most likely to the least likely. We define the tilt operation on probability distributions and show that it parametrizes an exponential family of distributions, which we refer to as the tilted family of the source. We prove that two sources result in the same guesswork, i.e., the same ordering from most likely to least likely on all strings, if and only if they belong to the same tilted family. We also prove that the strings whose guesswork is smaller than a given string are concentrated on the tilted family. Applying Laplace's method, we derive precise approximations on the distribution of guesswork on i.i.d. sources. The simulations show a good match between the approximations and the actual guesswork for i.i.d. sources.**

***Index Terms*—Ordering; Guesswork; One-to-One Codes; Rényi Entropy; Laplace's Method.**

I. Introduction

Let $X^n := X_1, \ldots, X_n$ denote a random n-string drawn from the parametric distribution $\mu_\theta^n(\cdot)$ on a finite alphabet $\mathcal{X}$. Order all the $|\mathcal{X}|^n$ strings of length n from the most likely to the least likely. Denote $G_\theta^n(X^n)$ as the guesswork random variable which is the order at which X^n appears in this ordered list. It is clear that $G_\theta^n(X^n)$ takes values in $\{1, \ldots, |\mathcal{X}|^n\}$. The goal of this paper is to analytically understand the probability mass function of $G_\theta^n(x^n)$, which finds applications in sequential decoding, computational security against brute-force attack, and source coding.

A. Related Work

The original motivation for the study of guesswork was to provide lower bounds on the computational complexity of sequential decoding [1] where the decoder sequentially examines several paths until it finds the correct coded sequence using some distance metric. Average guesswork was first studied by Massey [2] where he showed that guesswork is not related to Shannon entropy in general. Arıkan [3] considered guesswork on i.i.d. processes and proved that for long string lengths the moments are related to the Rényi entropy rate of the process. This has been generalized to ergodic Markov chains [4] and a wide range of stationary sources [5]. It has also been studied subject to an allowable distortion [6], and subject to constrained Shannon entropy [7]. Hanawal and Sundarasen [8] rederived the moments of guesswork assuming large deviations principle (LDP) holds. Christiansen and Duffy [9] established that guesswork satisfies LDP and completely characterized the rate function. They also provided an approximation to the distribution of guesswork.

Guesswork can also be used to quantify computational security against brute-force attack [10]. Suppose that a secret string is drawn from μ_θ^n on $\mathcal{X}^n$, which is used to secure a system which would only allow access if the correct secret string is provided, and does not reveal any information otherwise. If a brute-force attacker adversary wants to guess the secret string by query, $G_\theta^n(x^n)$ is exactly the number of guesses that the smartest attacker has to make until he finds the string x^n. Christiansen *et al.* [11] studied guesswork over the (weakly) typical set and proved that the exponent is strictly smaller than that of a uniform set with the same support size; they showed that the average guesswork of a password over an erasure channel does not relate to the average noise in general [12]; they also considered the setting where an attacker wants to guess one or more out of many secret strings drawn independently from not necessarily identical string-sources [10]. Finally, the idea of guesswork has been extended to the setup where the probability distribution is unknown [7], [13], [14].

In the context of source coding, it is known that the length of the optimal one-to-one source code for x^n (that need not satisfy Kraft's inequality) is within one bit of $\log G_\theta^n(x^n)$, and hence is related to the normalized zeroth moment of guesswork (see [14]). The source coding problem without prefix constraint dates back to Wyner [15] who showed that the average codeword length of one-to-one codes is upper bounded by the entropy. Alon and Orlistky derived a lower bound on the average codeword length in terms of the Shannon entropy [16], which was recently revisited for other moments of guesswork [17]. Szpankowski [18] derived

This work of A. Beirami, A. Makhdoumi, and M. Médard was supported in part by the Air Force Office of Scientific Research (AFOSR) under award No FA 9550-14-1-043; the work of R. Calderbank was supported in part by AFOSR under award No FA 9550-13-1-0076; the work of K. Duffy and M. Médard was supported in part by Netapp through a Faculty Fellowship.

the asymptotic average codeword length of one-to-one codes on binary memoryless sources, which was subsequently generalized to finite-alphabet i.i.d. processes [19], [20], and later studied under a universal setup [21]–[23].

B. Problem Setup

We study guesswork on i.i.d. processes over a finite alphabet. Let $\mathcal{X} = \{a_1, \ldots, a_{|\mathcal{X}|}\}$ be a finite alphabet of size $|\mathcal{X}|$. An i.i.d. source is defined using the set of probabilities $\theta_i = P[X = a_i]$ where $\sum_{i=1}^{|\mathcal{X}|} \theta_i = 1$. We refer to $\theta = (\theta_1, \ldots, \theta_{|\mathcal{X}|})$ as the source parameter vector, which is an element of the $d = (|\mathcal{X}| - 1)$ dimensional simplex of all stochastic vectors of size $|\mathcal{X}|$. Denote Λ as the (open) set of all parameter vectors θ such that $\theta_i > 0$ for all $i \in [|\mathcal{X}|]$.[1] We denote the parameter vector that results in uniform i.i.d. symbols by u, i.e., $u := (1/|\mathcal{X}|, \ldots, 1/|\mathcal{X}|) \in \Lambda$. We further define $\tilde{\Lambda}$ as

$$\tilde{\Lambda} := \{\theta \in \Lambda |\ \theta_i \neq \theta_j \text{ for all } i \neq j\}.$$

Our main tool in analysis will be the tilt operation defined on any $\theta \in \Lambda$, which is formally defined in the following:

Definition 1 (tilted θ of order α): For any $\alpha \in \mathbb{R}$, define $\theta^\alpha (\in \Lambda)$ as "tilted θ of order α" given by

$$\theta^\alpha := \frac{(\theta_1^\alpha, \ldots, \theta_{|\mathcal{X}|}^\alpha)}{\sum_{i=1}^{|\mathcal{X}|} \theta_i^\alpha}.$$

Note that tilted θ of order 1 is the source parameter vector θ itself, i.e., $\theta^1 = \theta$. We further use the notation $(\theta^\alpha)_i$ to denote the i-th element of the vector θ^α, i.e.,[2]

$$(\theta^\alpha)_i = \frac{\theta_i^\alpha}{\sum_{i=1}^{|\mathcal{X}|} \theta_i^\alpha}.$$

Definition 2 (tilted family of θ): Let the set $\Gamma_\theta^+ \in \Lambda$ denote the tilted family of θ and be given by

$$\Gamma_\theta^+ := \{\theta^\alpha\}_{\alpha \in \mathbb{R}^+}.$$

Note that the tilted family of θ is indeed an exponential family (see [24]). Also, the surfaces of equiprobable types form a linear family of distributions.

As an example, consider an i.i.d. source on a ternary alphabet ($|\mathcal{X}| = 3$). Γ_θ^+ is depicted in Fig. 1 for $\theta = (0.2, 0.3, 0.5)$. As can be seen, the tilted family of θ is a parametric curve (parametrized by α) that lives in the 2-dimensional simplex of stochastic vectors. The curve starts from the maximum entropy point u (for $\alpha = 0$) and then moves to one of the zero-entropy corner points of the simplex corresponding to the most likely symbol as $\alpha \to \infty$. We will show that these properties hold generally for any finite alphabet size and any parameter vector that lives in $\tilde{\Lambda}$.

In this paper, we establish two operational meanings of Γ_θ^+ and show that this set plays key roles in determining the behavior of the probability mass function (PMF) of guesswork. The organization of the paper and our contributions are summarized below.

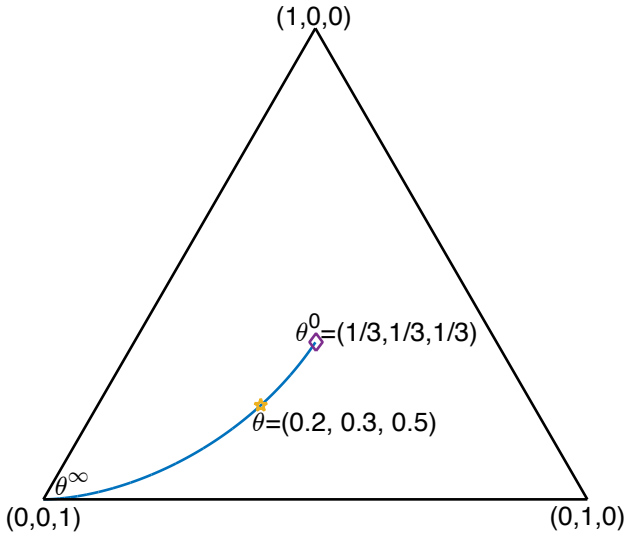

Fig. 1: The black lines depict the boundaries of the simplex of all ternary probability vectors. The yellow pentagon marker is the ternary source parameter vector $\theta = (0.2, 0.3, 0.5)$. The purple diamond marker corresponds to the uniform parameter vector $u = \theta^0 = (1/3, 1/3, 1/3)$. The blue curve depicts Γ_θ^+, i.e., the tilted family of θ.

- In Section II, we define the optimal ordering and provide the necessary definitions and notations.
- In Section III, we define the equivalent order class of a parameter vector θ as the set of all i.i.d. sources on a finite alphabet that lead to the same optimal ordering for all strings of all lengths. We show that this set coincides with Γ_θ^+.
- In Section IV, we define the dominating type of certain order as the type whose number of elements is the largest among all types that are more likely than a given type. We show that the dominating type asymptotically converges to a certain member of Γ_θ^+.
- In Section V, using the findings about the dominating type and Laplace's method (the principle of the largest term), we derive approximations on the PMF of the guesswork.

II. Optimal Ordering

Denote $x_k^{n+k-1} = x_k x_{k+1} \ldots x_{n+k-1} \in \mathcal{X}^n$ as a n-string over $\mathcal{X}$. Further, let $x^n = x_1^n$ and for $i > n$, $x_i^n = \varnothing$, where $\varnothing$ denotes the null string. In this paper, we focus on i.i.d. (string) sources on the finite alphabet $\mathcal{X}$.

Let μ_θ^1 denote the probability measure on $\mathcal{X}$ associated with the source parameter vector θ, i.e., $\mu_\theta^1(a_i) = \theta_i$. Denote μ_θ^n as the i.i.d. probability measure on $\mathcal{X}^n$, i.e., $\mu_\theta^n(x^n) = \prod_{i=1}^n \mu_\theta(x_i)$. Let $r_i(x^n)$ denote the count of the symbol a_i

[1] for any $n \in \mathbb{N}$, we define $[n] := \{1, \ldots, n\}$.

[2] This is in contrast to θ_i^α which means θ_i (the i-th element of θ) exponentiated to the power of α.

in string x^n. Formally, $r_i(x^n) = \sum_{i=1}^n \mathbb{I}_{\{a_i\}}(x_i)$ where $\mathbb{I}_{\mathcal{A}}$ is the indicator function of set $\mathcal{A} \subset \mathcal{X}$. Further, the vector

$$\mathbf{t}(x^n) := \frac{1}{n}\mathbf{r}(x^n) = \left(\frac{r_1(x^n)}{n}, \ldots, \frac{r_{|\mathcal{X}|}(x^n)}{n}\right)$$

denotes the type of the string x^n. Therefore,

$$\begin{aligned}\mu_\theta^n(x^n) &= \prod_{i=1}^{|\mathcal{X}|} \theta_i^{nt_i(x^n)} \\ &= \exp(-n\sum_{i=1}^{|\mathcal{X}|} t_i(x^n)\log\frac{1}{\theta_i}) \\ &= \exp(-n[H(\mathbf{t}(x^n)) + D(\mathbf{t}(x^n)||\theta)]) \\ &= \exp(-n[H(\mathbf{t}(x^n), \theta)],\end{aligned}$$

where $\exp(\cdot)$ denotes exponentiation in base 2, and $H(\cdot)$ is the $|\mathcal{X}|$-ary entropy function defined as

$$H(\theta) := \sum_{i=1}^{|\mathcal{X}|} \theta_i \log\frac{1}{\theta_i}, \tag{1}$$

and $D(\cdot||\cdot)$ is the $|\mathcal{X}|$-ary relative entropy function defined as

$$D(\theta||\gamma) := \sum_{i=1}^{|\mathcal{X}|} \theta_i \log\frac{\theta_i}{\gamma_i}, \tag{2}$$

and

$$\begin{aligned}H(\theta,\gamma) :=& H(\theta) + D(\theta||\gamma) \\ =& \sum_{i=1}^{|\mathcal{X}|} \theta_i \log\frac{1}{\gamma_i}\end{aligned} \tag{3}$$

By definition $H(\theta,\theta) = H(\theta)$.

We also need to define the Rényi entropy function of order ρ denoted by $H_\rho(\cdot)$ as given by

$$H_\rho(\theta) = \frac{1}{1-\rho}\log\left(\sum_{i=1}^{|\mathcal{X}|}\theta_i^\rho\right). \tag{4}$$

Note that $\lim_{\rho\to 1} H_\rho(\theta) = H(\theta)$ for all $\theta \in \Lambda$.

We refer to $\{\mu_\theta^n\}_{n=1}^\infty$ as a string-source (or in short source) with parameter vector θ. We use the notation $\{\mu_\theta^n\}$ to denote $\{\mu_\theta^n\}_{n=1}^\infty$ as well. Further, let $T_{\mathbf{t}(x^n)}$ denote the type class of x^n, i.e.,

$$T_{\mathbf{t}(x^n)} := \{y^n \in \mathcal{X}^n |\ \mathbf{t}(y^n) = \mathbf{t}(x^n)\}.$$

We further denote Δ_n as the set of all type-classes of strings of length n, i.e.,

$$\Delta_n = \{T_{\mathbf{t}(x^n)} |\ x^n \in \mathcal{X}^n\}.$$

Let $S_h \in \Lambda$ denote the manifold of all parameter vectors with constant entropy h, i.e.,

$$S_h := \{\theta \in \Lambda |\ H(\theta) = h\}.$$

As described in Section I, understanding the guesswork problem requires understanding of the ordering of the strings from most likely to the least likely, which we shall call the optimal ordering.

Definition 3 (ordering): Any one-to-one function $G^n : \mathcal{X}^n \to [|\mathcal{X}|^n]$ is called an ordering (on n-strings).

Definition 4 (optimal ordering for θ): An ordering G^n is said to be optimal for μ_θ^n (or in short for θ) if for any other ordering G'^n, and any $i \in [|\mathcal{X}|^n]$

$$P_\theta[G^n(X^n) \le i] \ge P_\theta[G'^n(X^n) \le i]. \tag{5}$$

It is straightforward to see that (5) is equivalent to the following:

$$\mu_\theta^n(x^n) > \mu_\theta^n(y^n) \Rightarrow G^n(x^n) < G^n(y^n). \tag{6}$$

Note that optimal ordering is not unique since swapping any two equally likely strings in an optimal ordering would result in another optimal ordering. In short, any arbitrary rule for breaking the ties in likelihood of strings results in an optimal ordering.

Let $\{e_i\}_{i=1}^{|\mathcal{X}|}$ denote the standard basis for $\mathbb{R}^{|\mathcal{X}|}$. Throughout the paper, we let $X^n : \Omega \mapsto \mathcal{X}^n$ be a random n-string drawn from μ_θ^n.

III. Order Class of the Source

In this section, we characterize the set of all i.i.d. sources on alphabet $\mathcal{X}$ that induce the same optimal ordering of the strings (from most likely to the least likely) and characterize the locus of all the source parameters in the simplex of probability vectors that have the same order as any given parameter vector θ.

Definition 5 (order equivalent): We say two string sources $\{\mu_\theta^n\}$ and $\{\mu_\gamma^n\}$ are order equivalent (or in short θ and γ are order equivalent) and denote by $\theta \equiv \gamma$ if and only if for all $n \in \mathbb{N}$ if G^n is an optimal ordering for θ it is also an optimal ordering for γ.

Definition 6 (equivalent order class): The equivalent order class of a parametric string source $\{\mu_\theta^n\}$ is denoted by C_θ and is given by

$$C_\theta = \{\gamma \in \Lambda |\ \gamma \equiv \theta\}.$$

Note that the optimal ordering on n-strings (for any $n \in \mathbb{N}$) drawn from the parametric sources $\gamma \in C_\theta$ is the same. Therefore, they result in the same guessing procedure and also the same one-to-one source code. Hence, we are interested in finding the equivalent order class of θ.

Theorem 1: For any $\theta \in \tilde{\Lambda}$, the equivalent order class of θ is given by

$$C_\theta = \Gamma_\theta^+.$$

To prove the theorem, we need to state a few lemmas.

Lemma 1: For all $\theta \in \tilde{\Lambda}$, θ^α is continuous in $\alpha \in \mathbb{R}^+$; and $\lim_{\alpha\to 0}\theta^\alpha = u$, where $u = (\frac{1}{|\mathcal{X}|}, \ldots, \frac{1}{|\mathcal{X}|})$; and $\lim_{\alpha\to\infty}\theta^\alpha = e_m$ where $m = \arg\max_{1\le i\le|\mathcal{X}|}\theta_i$.

Proof: The continuity is clear from the definition of θ^α and the limit as $\alpha \to 0$ is straightforward. The last part of the claim follows because $\theta \in \tilde{\Lambda}$ implies $m = \arg\max_{1\le i\le|\mathcal{X}|} \theta_i$ is unique. ∎

Lemma 2: For any $\theta \in \tilde{\Lambda}$, the equivalent order class of θ contains Γ_θ^+, i.e., $\Gamma_\theta^+ \subseteq C_\theta$.

Proof: We need to prove that for any $\theta^\alpha \in \Gamma_\theta$, we have $\theta^\alpha \equiv \theta$, i.e., $\theta^\alpha \in C_\theta$. To this end, we need to show that for any G_θ^n that is an optimal ordering on n-strings for θ, we have

$$\mu_{\theta^\alpha}^n(x^n) < \mu_{\theta^\alpha}^n(y^n) \Rightarrow G_\theta^n(x^n) > G_\theta^n(y^n).$$

It suffices to show that for any $\alpha \in \mathbb{R}^+$, we have

$$\mu_\theta(x^n) > \mu_\theta(y^n) \Leftrightarrow \mu_{\theta^\alpha}(x^n) > \mu_{\theta^\alpha}(y^n).$$

We show here

$$\begin{aligned}
&\mu_\theta(x^n) > \mu_\theta(y^n)\\
&\Leftrightarrow \prod_{i=1}^{|\mathcal{X}|} \theta_i^{r_i(x^n)} > \prod_{i=1}^{|\mathcal{X}|} \theta_i^{r_i(y^n)}\\
&\Leftrightarrow \prod_{i=1}^{|\mathcal{X}|} \theta_i^{\alpha r_i(x^n)} > \prod_{i=1}^{|\mathcal{X}|} \theta_i^{\alpha r_i(y^n)} \qquad (7)\\
&\Leftrightarrow \prod_{i=1}^{|\mathcal{X}|} \left(\frac{\theta_i^\alpha}{\sum_{j=1}^{|\mathcal{X}|}\theta_j^\alpha}\right)^{r_i(x^n)} > \prod_{i=1}^{|\mathcal{X}|} \left(\frac{\theta_i^\alpha}{\sum_{j=1}^{|\mathcal{X}|}\theta_j^\alpha}\right)^{r_i(y^n)} \qquad (8)\\
&\Leftrightarrow \prod_{i=1}^{|\mathcal{X}|} (\theta^\alpha)_i^{r_i(x^n)} > \prod_{i=1}^{|\mathcal{X}|} (\theta^\alpha)_i^{r_i(y^n)}\\
&\Leftrightarrow \mu_{\theta^\alpha}(x^n) > \mu_{\theta^\alpha}(y^n),
\end{aligned}$$

where (7) holds because $\alpha > 0$, and (8) holds because $\sum_{i=1}^{|\mathcal{X}|} r_i(x^n) = n$. This completes the proof. ∎

Now, we are ready to provide the proof of the theorem.

Proof of Theorem 1: To prove the theorem, we show that for $\theta \in \tilde{\Lambda}$ and any $0 < h < \log|\mathcal{X}|$, the intersection of S_h and C_θ denoted by $C_\theta \cap S_h$ contains one and only one parameter vector. We further show that $C_\theta \cap S_h \in \Gamma_\theta$.

Let $\gamma, \lambda \in C_\theta \cap S_h$ be two parameter vectors that are contained in the intersection of S_h and C_θ. Therefore,

$$\sum_{i=1}^n \gamma_i \log\frac{1}{\gamma_i} = \sum_{i=1}^n \lambda_i \log\frac{1}{\lambda_i} = h.$$

Let $\mathbf{r}(n\gamma)$ be defined as

$$\mathbf{r}(n\gamma) = \arg \min_{\substack{r_i\in\mathbb{Z}^+\\ \sum_{i=1}^{|\mathcal{X}|} r_i = n}} ||\mathbf{r} - n\gamma||_1.$$

Observe that $\mathbf{r}(n\gamma)$ by definition is a type on n-strings. Further, $|r_i(n\gamma) - n\gamma_i| < 1$. We define $\mathbf{r}(\lambda, n)$ similarly. Let $x^n \in T_{\mathbf{r}(n\gamma)}$ and $y^n \in T_{\mathbf{r}(n\lambda)}$ be two strings drawn from type classes of $\mathbf{r}(n\gamma)$ and $\mathbf{r}(n\lambda)$, respectively. Therefore,

$$\begin{aligned}
\log\frac{1}{\mu_\gamma(x^n)} &= \sum_{i=1}^{|\mathcal{X}|} r_i(n\gamma)\log\frac{1}{\gamma_i}\\
&= \sum_{i=1}^{|\mathcal{X}|} n\gamma_i\log\frac{1}{\gamma_i} + \sum_{i=1}^{|\mathcal{X}|}(r_i(n\gamma) - n\gamma_i)\log\frac{1}{\gamma_i}\\
&\le nh + \sum_{i=1}^{|\mathcal{X}|}\log\frac{1}{\gamma_i}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\log\frac{1}{\mu_\gamma(y^n)} &= \sum_{i=1}^{|\mathcal{X}|} r_i(n\lambda)\log\frac{1}{\gamma_i}\\
&= \sum_{i=1}^{|\mathcal{X}|} n\lambda_i\log\frac{1}{\gamma_i} + \sum_{i=1}^{|\mathcal{X}|}(r_i(n\lambda) - n\lambda_i)\log\frac{1}{\gamma_i}\\
&\ge nh + nD(\lambda||\gamma) - \sum_{i=1}^{|\mathcal{X}|}\log\frac{1}{\gamma_i}.
\end{aligned}$$

Therefore, if $\lambda \neq \gamma$, for sufficiently large n, we have $\mu_\gamma(x^n) > \mu_\gamma(y^n)$ and hence $G_\gamma(x^n) < G_\gamma(y^n)$. By repeating similar arguments we can show that for sufficiently large n, we have $\mu_\lambda(x^n) < \mu_\lambda(y^n)$, which leads to $G_\lambda(x^n) > G_\lambda(y^n)$. Therefore, γ and λ are not order equivalent, i.e., $\gamma \not\equiv \lambda$, which is a contradiction. Hence, the intersection of C_θ and S_h is unique.

Next, we need to show that $C_\theta \cap S_h$ is not empty and is contained in Γ_θ. This is carried out by invoking Lemma 1 and noting that $H(u) = \log|\mathcal{X}|$ and $H(e_m) = 0$. ∎

We find that the equivalent order class of a parametric string source with parameter vector $\theta \in \tilde{\Lambda}$ is characterized by a set of parameters that lie on a parametric curve in the simplex of probability vectors, which starts (for $\alpha = 0^+$) from the maximum entropy point corresponding to uniform distribution and ends (for $\alpha \to \infty$) in one of the zero-entropy corner points in the simplex.

IV. Dominating Type

In this section, we provide another operational meaning of the equivalent order class of the source θ by defining the dominating type, which is the type with maximum cardinality among all types whose elements are more likely than a given string.

Definition 7 (dominating type of order α): $\mathbf{t}_{\theta,\alpha}^n$ is called the dominating type of order α and is defined as

$$\mathbf{t}_{\theta,\alpha}^n := \begin{cases} \arg\max\limits_{\mathbf{t}\in\Delta_n} & H(\mathbf{t})\\ \text{s.t.} & H(\mathbf{t},\theta) \le H(\theta^\alpha,\theta)\end{cases}, \qquad (9)$$

where $H(\cdot,\cdot)$ is defined in (3).

Observe that $\mathbf{t}_{\theta,\alpha}^n$ is the type with the largest cardinality among all n-strings whose probability is larger than or equal

to a constant specified through

$$\frac{1}{n}\log\frac{1}{\mu_\theta^n(x^n)} \le H(\theta^\alpha,\theta).$$

Definition 8 (dominating class of positive orders): The dominating class of all positive orders of the source is defined as

$$\{\mathbf{t}_{\theta,\alpha}^n\}_{\alpha\in\mathbb{R}^+}.$$

The theorem below shows the connection between the dominating class of positive orders and the tilted family of θ, which suggests a second operational meaning to Γ_θ^+.

Theorem 2: The dominating type of order α converges to θ^α as $n\to\infty$. In particular,

$$||\mathbf{t}_{\theta,\alpha}^n - \theta^\alpha||_2 \le \frac{1}{n}$$

Further, the dominating class of positive orders of source θ uniformly converges to Γ_θ^+ as $n\to\infty$.

We need the following lemmas to prove the theorem.

Lemma 3: $H(\theta^\alpha)$ is a decreasing function of α, which starts from $\log|\mathcal{X}|$ at $\alpha=0$ and vanishes as $\alpha\to\infty$.

Proof: We need to show that $\frac{dH(\theta^\alpha)}{d\alpha}<0$ for all $\alpha\in\mathbb{R}^+$. On the other hand, it suffices to show that $\left.\frac{dH(\theta^\alpha)}{d\alpha}\right|_{\alpha=1}<0$ due to the properties of the parametric curve that specifies the tilted family.

$$\begin{aligned}
\left.\frac{dH(\theta^\alpha)}{d\alpha}\right|_{\alpha=1} &= \sum_{i=1}^{|\mathcal{X}|}\left.\frac{d(\theta^\alpha)_i}{d\alpha}\right|_{\alpha=1}(\log\frac{1}{\theta_i}-\log e)\\
&= \sum_{i=1}^{|\mathcal{X}|}\frac{\theta_i}{\log e}(H(\theta)-\log\frac{1}{\theta_i})(\log\frac{1}{\theta_i}-\log e)\\
&= E\left\{\frac{1}{\log e}(H(\theta)-\log\frac{1}{\theta_i})(\log\frac{1}{\theta_i}-\log e)\right\}\\
&\le \frac{1}{\log e}(H(\theta)-E\{\log\frac{1}{\theta_i}\})(E\{\log\frac{1}{\theta_i}\}-\log e) \qquad (10)\\
&= 0,
\end{aligned}$$

where (10) follows from concavity of $(a-x)(x-b)$ with respect to x and Jensen's inequality. Note that the inequality in (10) is strict unless $\theta_i=\frac{1}{e}$ (which is impossible) or $\theta_i=\frac{1}{|\mathcal{X}|}$, which is the uniform distribution and is only obtained at $\alpha=0$. Thus, the inequality is strict for $\alpha\in\mathbb{R}^+$. ∎

Lemma 4: Let $\xi_{\theta,\beta}$ be defined as

$$\xi_{\theta,\beta} = \arg\max_\xi\left\{H(\xi)-\frac{1}{\beta}D(\xi||\theta)\right\}. \qquad (11)$$

Then, $\xi_{\theta,\beta}=\theta^{1/(1+\beta)}$.

Proof: We have

$$\begin{aligned}
H(\xi)-\frac{1}{\beta}D(\xi||\theta) &= \sum_{i=1}^{|\mathcal{X}|}\xi_i\log\frac{1}{\xi_i}-\frac{1}{\beta}\sum_{i=1}^{|\mathcal{X}|}\xi_i\log\frac{\xi_i}{\theta_i}\\
&= \frac{1+\beta}{\beta}\sum_{i=1}^{|\mathcal{X}|}\xi_i\log\frac{\theta_i^{\frac{1}{1+\beta}}}{\xi_i}\\
&\le \frac{1+\beta}{\beta}\log\sum_{i=1}^{|\mathcal{X}|}\theta_i^{\frac{1}{1+\beta}} \qquad (12)\\
&= H_{1/(1+\beta)}(\theta),
\end{aligned}$$

where (12) is due to the log-sum inequality, and where $H_{1/(1+\beta)}(\theta)$ is the Rényi entropy of order $1/(1+\beta)$ of the source as defined in (4). Further, equality in (12) is achieved if and only if $\xi_i=C\theta_i^{\frac{1}{1+\beta}}$. This also implies that $\xi_{\theta,\beta}=\theta^{1/(1+\beta)}$, which completes the proof. ∎

Note that the solution to this minimization is related to the rate function of the LDP as also formulated by Hanawal and Sundarasen [8]. Next, we state the proof of the main result in this section.

Proof of Theorem 2: The theorem is proved by relaxing the integer constraint on the types, i.e., let

$$\phi_{\theta,\alpha} := \begin{cases}\arg\max\limits_{\phi} & H(\phi)\\ \text{s.t.} & H(\phi,\theta)=H(\theta^\alpha,\theta)\end{cases}$$

Since $H(\cdot)$ is a concave function, this relaxed version of the problem can be solved using Lagrange multipliers, from which and Lemma 4, it is deduced that the maximizer $\phi_{\theta,\alpha}$ satisfies $\phi_{\theta,\alpha}\in\Gamma_\theta^+$ and hence $\phi_{\theta,\alpha}=\theta^\alpha$. Now, by combining the above with Lemma 3, we conclude that:

$$\theta^\alpha = \begin{cases}\arg\max\limits_{\phi} & H(\phi)\\ \text{s.t.} & H(\phi,\theta)\ge H(\theta^\alpha,\theta)\end{cases}$$

Therefore, the maximizer of the original problem is within Euclidean distance $\frac{1}{n}$ of the relaxed problem, and hence, as $n\to\infty$, it converges to θ^α. ∎

V. Approximation of the Optimal Ordering Distribution

In this section, we present an approximation on the probability distribution of $G_\theta^n(X^n)$, where $G_\theta^n(\cdot)$ is an optimal ordering on n-strings for all $n\in\mathbb{N}$. In Theorem 2, we determined that the type that dominates guesswork is "close" to the tilted family of θ, i.e., guesswork is concentrated on the tilted family. In this section, we further use Laplace's method (the principle of the dominating term) to approximate the guesswork distribution within a multiplicative factor of $(1+O(\frac{1}{\sqrt{n}}))$.

We will argue that as n gets bigger, the size of dominating type is going to dominate the number of strings whose probability is larger than a certain limit. This in turn helps us approximate the likelihood of the strings ordered by the optimal guessing order.

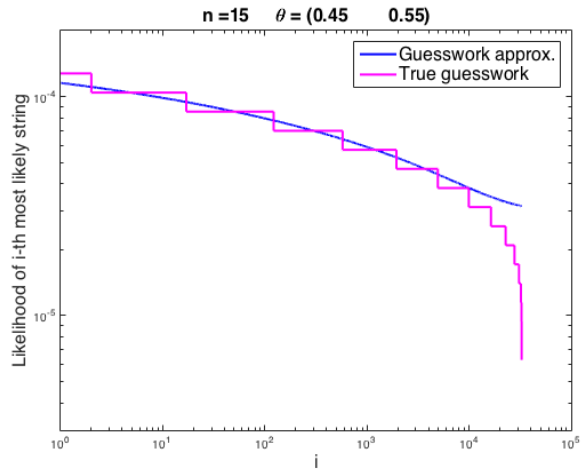

Fig. 2: The approximation on the distribution of the guesswork for a binary memoryless source ($|\mathcal{X}| = 2$) and $n = 15$.

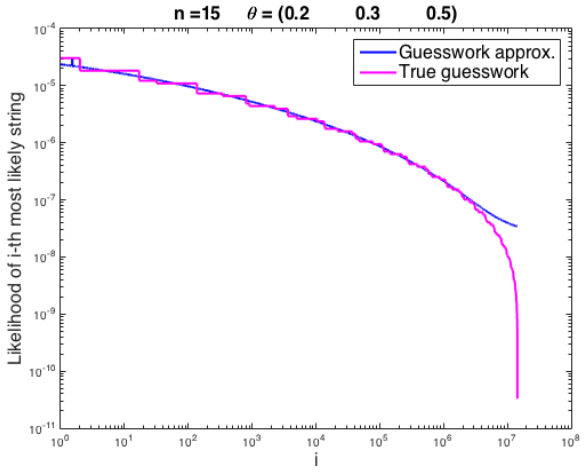

Fig. 3: The approximation on the distribution of the guesswork for a ternary memoryless source ($|\mathcal{X}| = 3$) and $n = 15$.

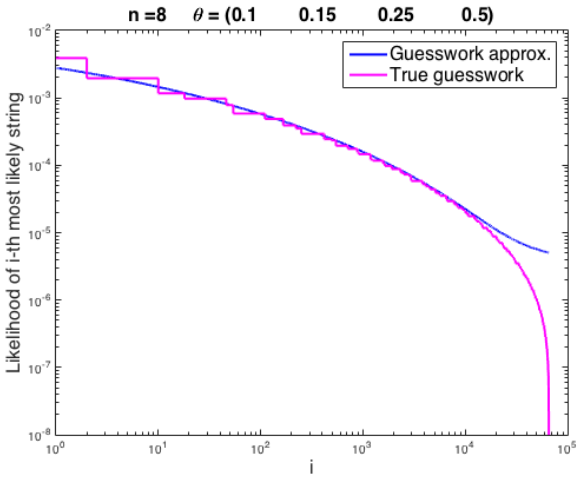

Fig. 4: The approximation on the distribution of the guesswork for a 4-ary memoryless source ($|\mathcal{X}| = 4$) and $n = 8$.

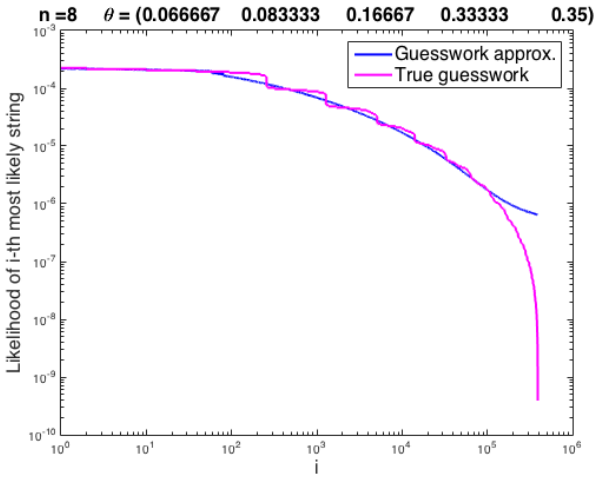

Fig. 5: The approximation on the distribution of the guesswork for a 5-ary memoryless source ($|\mathcal{X}| = 5$) and $n = 8$.

Definition 9 (projection of x^n on the tilted family of θ): Let $\Pi(x^n)$ denote the projection of x^n on the tilted family of θ as given by

$$\Pi(x^n) = \{\phi \in \Gamma_\theta^+ : H(\phi, \theta) = H(\mathbf{t}(x^n), \theta), \tag{13}$$

where $\mathbf{t}(x^n)$ is the type of x^n. Further, define $\alpha(x^n)$

$$\alpha(x^n) = \arg_\beta\{\Pi(x^n) = \theta^\beta\}.$$

Observe that by definition $\Pi(x^n) = \theta^{\alpha(x^n)}$.

Lemma 5: $H(\theta^\alpha, \theta)$ is a decreasing function of α for all $\alpha \in \mathbb{R}^+$. In other words, if $\alpha_1 > \alpha_2 > 0$, then $H(\theta^{\alpha_1}, \theta) < H(\theta^{\alpha_2}, \theta)$.

Proof: We need to show that $\frac{d}{d\alpha} H(\theta^\alpha, \theta) < 0$. On the other hand, it suffices to show that

$$\sum_{i=1}^{|\mathcal{X}|} \frac{d(\theta^\alpha)_i}{d\alpha}\bigg|_{\alpha=1} \log\frac{1}{\theta_i} < 0.$$

We have

$$\begin{aligned}\sum_{i=1}^{|\mathcal{X}|} \frac{d(\theta^\alpha)_i}{d\alpha}\bigg|_{\alpha=1} \log\frac{1}{\theta_i} &= \sum_{i=1}^{|\mathcal{X}|} \frac{\theta_i}{\log e}(H(\theta) - \log\frac{1}{\theta_i})(\log\frac{1}{\theta_i}) \\ &= E\left\{\frac{1}{\log e}(H(\theta) - \log\frac{1}{\theta_i})(\log\frac{1}{\theta_i})\right\} \\ &\le \frac{(H(\theta) - E\{\log\frac{1}{\theta_i}\})(E\{\log\frac{1}{\theta_i}\})}{\log e} \qquad (14) \\ &= 0,\end{aligned}$$

where the reasoning for (14) is similar to that of (10) in the proof of Lemma 3. ∎

Lemma 6: For any x^n such that $\mu_\theta^n(x^n) > \frac{1}{|\mathcal{X}|^n}$, $\Pi(x^n)$ exists and is unique.

Proof: The existence follows from Lemma 5 and the uniqueness is proved by following the same lines of the proof of Theorem 1. ∎

Let b_θ be the rate of change of the entropy on the tilted family as given by

$$b_\theta := \frac{\frac{dH(\theta^\beta)}{d\beta}\Big|_{\beta=1}}{\left|\left|\frac{d\theta^\beta}{d\beta}\Big|_{\beta=1}\right|\right|_2},$$

where $d\theta^\beta$ is the element-wise differential of θ^β (with respect to the parameter β).

Further, for any $x^n \in \mathcal{X}^n$ and $\Pi = \Pi(x^n)$ and $\alpha = \alpha(x^n)$ that are the projection coefficients defined in Definition 9, define $\overline{G}^n(\Pi)$ and $\underline{G}^n(\Pi)$ as:

$$\overline{G}^n(\Pi) := \frac{1}{1-\exp\left(b_{\theta^\alpha}\right)}\sqrt{\frac{1}{2\pi n}}\ 2^{nH(\Pi)-b_{\theta^\alpha}}, \tag{15}$$

$$\underline{G}^n(\Pi) := \frac{1}{1-\exp\left(b_{\theta^\alpha}\right)}\sqrt{\frac{1}{2\pi n}}\ 2^{nH(\Pi)+b_{\theta^\alpha}}. \tag{16}$$

Here is our main result on the approximation of the distribution of guesswork.

Theorem 3: Let $G_\theta^n(\cdot)$ be any optimal ordering for θ on n-strings. Then, for any x^n:

$$\underline{G}^n(\Pi)(1+O(\frac{1}{\sqrt{n}})) \le G_\theta^n(x^n) \le \overline{G}^n(\Pi)(1+O(\frac{1}{\sqrt{n}})), \tag{17}$$

where $\Pi = \Pi(x^n)$ is the projection of x^n on the tilted family of θ.

Sketch of the proof: Write down the Taylor expansion of $G_\theta^n(x^n)$ around $\Pi(x^n)$ and notice that the Euclidean distance between $\Pi(x^n)$ and the dominating type is bounded by $\frac{1}{n}$ according to Theorem 2. Then using Laplace's method for summing all the terms that are more probable than x^n we arrive at the desired result of the theorem. ∎

Note that in the binary case where $|\mathcal{X}| = 2$, Theorem 3 is a straightforward deduction of Lemma 1 of Szpankowski [18]. On the other hand, the proof of Szpankowski is not readily extendible to $|\mathcal{X}| > 2$ due to the complications that arise in identifying the optimal ordering of the strings. Further observe that these bounds are tight up to the multiplicative $(1+O(\frac{1}{\sqrt{n}}))$ factor in the sense that either end could be achieved by some particular optimal ordering.

Next, we use this result to approximate the distribution of (any) optimal ordering $G_\theta^n(X^n)$. The approximation is as follows:

$$P_\theta\left[G_\theta^n(X^n) \approx \frac{1}{\sqrt{2\pi n}(1-\exp(\frac{b_{\theta,\alpha}}{a_{\theta,\alpha}}))}2^{nH(\theta^\alpha)}\right] \tag{18}$$

$$\approx \frac{P_\theta\left[\underline{G}^n(\theta^\alpha) \le G_\theta^n(X^n) \le \overline{G}^n(\theta^\alpha)\right]}{\binom{n}{n\theta^\alpha}}$$

$$\approx \frac{P_\theta[\Pi(X^n) = \theta^\alpha]}{\binom{n}{n\theta^\alpha}}$$

$$\approx \exp(-nH(\theta^\alpha,\theta)), \tag{19}$$

where in (18) the RHS is the geometric mean of $\overline{G}^n(\theta^\alpha)$ and $\underline{G}^n(\theta^\alpha)$.

Although the above equation can be used to obtain an approximation for $P[G_\theta^n(X^n) = i]$ for any $i \in [|\mathcal{X}|^n]$, this approximation is only valid for strings for which $G_\theta^n(x^n) \ll |\mathcal{X}|^n$. To show the accuracy of this approximations, we have run several experiments and report four here (Figs. 2-5). The approximation in (19) has been plotted as the solid blue curve and the true probability distribution of (any) optimal ordering is shown in pink. As can be seen, this is a good approximation for small i and has no predictive value for $i \approx |\mathcal{X}|^n$. On the other hand, when compared to the approximation on the distribution derived in [9], this approximation is much more accurate for small values of n primarily due to the exact asymptotic expansion of the pre-factor in the large deviation estimates.

VI. Conclusion

In this paper, we provided a geometric perspective on the guesswork problem. We defined a tilt operation on stochastic vectors and referred to the collection of all tilts of positive orders as the tilted family of the source parameter vector. We established two operational meanings of the tilted family by demonstrating that this set coincides with the equivalent order class of an i.i.d. source who lead to the same ordering on all strings of all lengths. We also showed that all types that dominate the guesswork lie close to the tilted family. Using these results in conjunction with Laplace's method, we provided an asymptotic approximation on the probability distribution of the guesswork, which is in good agreement with the actual distribution even for small string lengths.

Acknowledgement

The authors thank Erdal Arıkan for informative discussions about guesswork and sequential decoding. The authors also thank Jossy Sayir for suggesting the symmetric plotting of the probability simplex on an equilateral triangle.

References

[1] I. Jacobs and E. Berlekamp, "A lower bound to the distribution of computation for sequential decoding," *IEEE Trans. Info. Theory*, vol. 13, no. 2, pp. 167–174, April 1967.

[2] J. L. Massey, "Guessing and entropy," in *1994 IEEE International Symposium on Information Theory Proceedings*. IEEE, 1994, p. 204.

[3] E. Arıkan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 99–105, Jan. 1996.

[4] D. Malone and W. G. Sullivan, "Guesswork and entropy," *IEEE Trans. Inf. Theory*, vol. 50, no. 3, pp. 525–526, Mar. 2004.

[5] C. E. Pfister and W. G. Sullivan, "Renyi entropy, guesswork moments, and large deviations," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2794–2800, Nov. 2004.

[6] E. Arıkan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[7] A. Beirami, R. Calderbank, K. Duffy, and M. Médard, "Quantifying computational security subject to source constraints, guesswork and inscrutability," in *2015 IEEE International Symposium on Information Theory Proceedings (ISIT)*, Jun. 2015.

[8] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 70–78, Jan. 2011.

[9] M. M. Christiansen and K. R. Duffy, "Guesswork, large deviations, and Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 796–802, Feb. 2013.

[10] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, "Quantifying the computational security of multi-user systems," *accepted in IEEE Trans. Info. Theory*, 2015.

[11] ——, "Brute force searching, the typical set and guesswork," in *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, July 2013, pp. 1257–1261.

[12] ——, "Guessing a password over a wireless channel (on the effect of noise non-uniformity)," in *2013 Asilomar Conference on Signals, Systems and Computers*, 2013, pp. 51–55.

[13] R. Sundaresan, "Guessing under source uncertainty," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 269–287, Jan. 2007.

[14] O. Kosut and L. Sankar, "Asymptotics and non-asymptotics for universal fixed-to-variable source coding," *arXiv preprint arXiv:1412.4444*, 2014.

[15] A. D. Wyner, "An upper bound on the entropy series," *Information and Control*, vol. 20, no. 2, pp. 176–181, 1972.

[16] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1670–1672, Sept. 1994.

[17] T. Courtade and S. Verdú, "Cumulant generating function of codeword lengths in optimal lossless compression," in *2014 IEEE International Symposium on Information Theory (ISIT)*, July 2014, pp. 2494–2498.

[18] W. Szpankowski, "A one-to-one code and its anti-redundancy," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4762–4766, Oct. 2008.

[19] W. Szpankowski and S. Verdú, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4017–4025, Jul. 2011.

[20] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.

[21] O. Kosut and L. Sankar, "Universal fixed-to-variable source coding in the finite blocklength regime," in *2013 IEEE International Symposium on Information Theory Proceedings (ISIT '13)*, Jul. 2013, pp. 649–653.

[22] ——, "New results on third-order coding rate for universal fixed-to-variable source coding," in *2014 International Symposium on Information Theory (ISIT 2014)*, Jul. 2014, pp. 2689–2693.

[23] A. Beirami and F. Fekri, "Fundamental limits of universal lossless one-to-one compression of parametric sources," in *2014 IEEE Information Theory Workshop (ITW '14)*, Nov. 2014, pp. 212–216.

[24] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.