Relaxing the Kolmogorov Structure Function for Realistic Computational Constraints

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Given a high-dimensional and noisy data source, quantifying the complexity of datasets or individual datapoints is a central challenge in machine learning, with applications including data collection, active learning, finding hidden subpopulations, detecting outliers or mislabeled data, compression, and generalization bounds. Shannon entropy (Shannon, 1948) and Kolmogorov complexity (**?**Kolmogorov, 1965) quantify the complexity of an object through its shortest possible compression. While such measures provide a rigorous foundation for understanding the true irreducible complexity of a dataset, their assumption of a computationally unconstrained compressor makes them infeasible to compute in practice. Recent works (Xu et al., 2020; Ethayarajh et al., 2022) incorporate computational constraints into measures of information by assuming that the compressor belongs to a fixed family of predictive functions. While appealing, a single fixed family of functions may not show the complete story about the information inside data because different aspects of the data become decodable at different scales of computation.

One way of viewing the process of learning is as a cascading sequence of rules and exceptions, in which one first learns simple and broad predictive rules, then the commonalities among exceptions to those rules, the exceptions to the exceptions, and so on. This viewpoint suggests a natural ordering of datapoints within any dataset, ranging from the most typical and representative examples to the hardest edge cases. By fitting models with different capacities to a given dataset, we can observe the amount and content of information corresponding to each level of the rule-exception hierarchy. The Kolmogorov Structure Function (KSF) (Kolmogorov, 1974) encodes this intuition by measuring, for each integer k, the fit of the best k-length description of a given string x. However, the KSF is infeasible to compute in practice, because the set of all k-bit descriptions is an infeasibly large search space for even moderately high values of k.

We propose the Constrained Structure Function (CSF): a practical and scalably computable function of any dataset that inherits many of the desirable properties of the KSF. Given a sequence of increasingly large hypothesis classes, the CSF measures the change in fit quality with increasing capacity. The CSF has an additional property not present in the KSF: that is useful for understanding the structure of datasets consisting of interchangeable datapoints. Under a mild factorization assumption about hypothesis classes that is satisfied by all standard neural networks for supervised learning, the CSF of a dataset can be expressed as the sum of datapoint-wise functions. The contribution of one datapoint to the dataset's CSF can show the properties of that datapoint in the context of the dataset.

2 Kolmogorov Structure Function

The Kolmogorov complexity (KC) (Kolmogorov, 1965) of a datapoint x is defined as the length of the shortest binary program that outputs x and halts. This quantity can be seen as encoding all irreducible bits of information inside x. As a running example, consider a binary string x of length $N = n^2$ that encodes an $n \times n$ black-and-white image. The image has a white background represented by zeros and a gray circle covering m pixels, where in order to represent gray using binary

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

pixels, we sample each of the m pixels according to an independent coin flip ($\sim$ Bern(0.5)). Even though x is a fixed string, the values of its gray pixels can be considered random in the sense that these pixel values cannot be compressed with high probability. The expected *Kolmogorov complexity* of this string is $m+O(1)$, where we use a constant number of bits to describe the background and circle, and m bits to describe the specific pixel values inside the circle. Note how in this example, the underlying structure inside the image informs the flow of the shortest program that describes it, and thus its description length.

The *Kolmogorov structure function* (KSF) is a generalization of the Kolmogorov complexity which reveals the internal structure of an object through the change in its description length at multiple scales. The only written record of this concept by Kolmogorov himself is the following abstract (Kolmogorov, 1974; Vereshchagin & Vitányi, 2004), translated from Russian by L. A. Levin:

> "To each constructive object corresponds a function $\Phi_x(k)$ of a natural number k—the log of minimal cardinality of x-containing sets that allow definitions of complexity at most k. If the element x itself allows a simple definition, then the function Φ drops to 1 even for small k. Lacking such definition, the element is "random" in a negative sense. But it is "probabilistically random" only when function Φ having taken the value Φ_0 at a relatively small $k = k_0$, then changes approximately as $\Phi(k) = \Phi_0 - (k - k_0)$."

By definition, the KSF of any object x is a decreasing function of the compression capacity k: $\Phi_x(k) < \Phi_x(k+1)$. The rate at which the KSF decreases, i.e. the value of $\Phi_x(k) - \Phi_x(k+1)$, is the rate at which specifying one more bit of information improves the fit to x.

Returning to the image example, the strategy of first describing the background and circle and then the "gray" pixel values maps onto the KSF as follows. We first provide a short description of the background, which narrows down the number of possible binary strings from 2^N to 2^m. Therefore, the KSF of the image steeply drops from N to m given a small constant number of bits. Compressing further corresponds to specifying the index of x among a lexicographical ordering of the 2^m possible images. The KSF from this point on has a moderate slope of -1 because each additional bit of information can only halve the number of possible images.

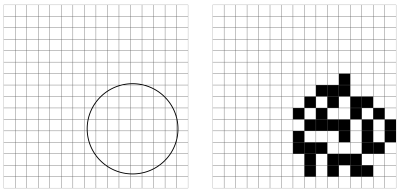

Figure 1: Example image with circle region represented by random values.

A core property of the KSF is that it shows internal structure: all of the regularity inside x lies in the fact that it is an element of the larger set of size 2^m, and the slope of the KSF reflects this. Note that a "vertical slice", i.e., a single function value of the KSF at some complexity level k, will not reveal such structure, and considering the entire function's shape can show useful properties of the data.

We formally define the KSF below, with notation following Vereshchagin & Vitányi (2004).

Definition 1 (Kolmogorov structure function, KSF)**.** *Let $K(\cdot)$ be the Kolmogorov complexity: the length of the shortest program that outputs the given object and halts. Let x be data and $\alpha \in \mathbb{N}^+$ an integer which bounds the complexity of distributions to consider. The structure function of x maps the complexity α to the best-fit distribution P as:*

$$h_x : \alpha \mapsto \min_P \{-\log P(x) : K(P) \leq \alpha\}. \tag{1}$$

Note that we consider distributions $P(x)$ instead of discrete sets that contain x. This generalization allows us to consider continuous input spaces; the set-based definition is a special case where the P is restricted to uniform categorical distributions. In fact, for discrete input spaces, the two definitions are equivalent up to a constant factor; see Appendix A.2 for more discussion.

3 The Constrained Structure Function

3.1 Relaxing the KSF via a Restricted Predictive Family

We now formally define the Constrained Structure Function (CSF), a relaxation of the KSF that allows for any increasing sequence of predictive families. The core idea is to replace the minimization in Definition 1 with an optimization over a restricted family of models.

We consider datasets $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ that consist of inputs $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$. We are interested in measuring the complexity of the entire dataset $\mathcal{D}$, where there is often redundant information among the many input-label pairs.

We adapt the notion of predictive families from Xu et al. (2020) as follows:

Definition 2 (Predictive family, $\mathcal{V}$)**.** *Let $\Omega = \{f : \mathcal{X} \to \mathcal{P}(\mathcal{Y})\}$, where $\mathcal{P}(\mathcal{Y})$ is the set of probability measures over the Borel algebra on $\mathcal{Y}$. We say that $\mathcal{V} \subseteq \Omega$ is a predictive family if it satisfies*

$$\forall f \in \mathcal{V}, \forall P \in \text{range}(f), \quad \exists f' \in \mathcal{V}, \quad \textit{s.t.} \quad \forall x \in \mathcal{X}, f'[x] = P. \tag{2}$$

Given a predictive family and a dataset, we define the $\mathcal{V}$-complexity as

Definition 3 ($\mathcal{V}$-complexity)**.** *Let $\mathcal{D}$ be a dataset, and $\mathcal{V}$ a predictive family. The $\mathcal{V}$-complexity of $\mathcal{D}$ is the lowest achievable negative log likelihood of $\mathcal{D}$ under $\mathcal{V}$:*

$$h(\mathcal{D}; \mathcal{V}) = \inf_{f \in \mathcal{V}} \left(-\sum_i \log f(y_i; x_i) \right). \tag{3}$$

Taking motivation from this overall structure, we consider arbitrary nested sequences of predictive families. This predictive family is much more amenable to the minimization in (3), and has the property that $\mathcal{V}_i \subset \mathcal{V}_{i+1}$ for each i. We first show that such nested sequences satisfy the same decreasing property as the Kolmogorov structure function.

Definition 4 (Constrained Structure Function, CSF)**.** *Let $\mathcal{D}$ be a dataset and $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \ldots \subset \mathcal{V}_N$ a nested sequence of predictive families.*

3.2 Properties of the CSF

Basic properties. We show that based on only the constraint that $\mathcal{V}_i \subset \mathcal{V}_{i+1}$ for each i, the Constrained Structure Function (CSF) shares many desirable properties with the Kolmogorov Structure Function (KSF).

Proposition 1. *Let $\mathcal{D}$ be a dataset and $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \ldots \subset \mathcal{V}_N$ be defined as in Definition 4. We have*

1. ***Decreasing in** N: $h(\mathcal{D}; \mathcal{V}_1) \geq h(\mathcal{D}; \mathcal{V}_2) \geq \ldots \geq h(\mathcal{D}; \mathcal{V}_N)$.*
2. ***Nonnegative:** $h(\mathcal{D}; \mathcal{V}) \geq 0$ for all $\mathcal{V}$.*
3. ***Zero limit:** For any $\epsilon > 0$, there exists $\mathcal{V}$ such that $h(\mathcal{D}; \mathcal{V}) < \epsilon$.*
4. ***Monotonicity in** $\mathcal{D}$: If $\mathcal{D} \subseteq \mathcal{D}^+$, then $h(\mathcal{D}; \mathcal{V}) \leq h(\mathcal{D}^+; \mathcal{V})$ for all $\mathcal{V}$.*

These properties demonstrate the validity of the CSF as a measure of the complexity of a dataset.

Additive decomposition. Recall that we consider *datasets* rather than general binary strings. As datasets consists of unordered pairs of inputs and labels, a standard choice of model class for datasets is a factorized model. In such a model, the joint distribution over the dataset is the product of the conditional distributions over each input-label pair

$$\log p(y_1, \ldots, y_n | x_1, \ldots, x_n) = \sum_i \log p(y_i | x_i). \tag{4}$$

This mild factorization assumption is satisfied by standard discriminative models including neural networks. For such $\mathcal{V}_1 \subset \ldots, \mathcal{V}_N$, we can compute a datapoint-wise analogue of the CSF, which shows the learnability of the datapoint (x, y) in the context of the entire dataset $\mathcal{D}$. We formally define the datapoint-specific CSF as follows:

Definition 5. *Let $\mathcal{D}$ be a dataset, $\mathcal{V}$ a predictive family, and $(x, y) \in \mathcal{D}$. We define the datapoint-specific CSF as*

$$h(x, y, \mathcal{D}; \mathcal{V}) = -\log f(y; x), \quad \textit{where} \quad f = \arg\min_{f \in \mathcal{V}} \left(-\sum_{(x,y) \in \mathcal{D}} \log f(y; x) \right). \tag{5}$$

Intuitively, datapoints that are more "typical" will have a more rapidly decreasing datapoint-specific CSF, since even the predictive models with lower capacity will fit them. An important property is that the sum of all datapoint-wise CSFs is equal to the CSF of the full dataset $\mathcal{D}$.

Proposition 2. *The CSF of a dataset $\mathcal{D}$ is the sum of the datapoint-specific CSFs of its elements:*

$$h(\mathcal{D}; \mathcal{V}) = \sum_{(x,y) \in \mathcal{D}} h(x, y, \mathcal{D}; \mathcal{V}). \tag{6}$$

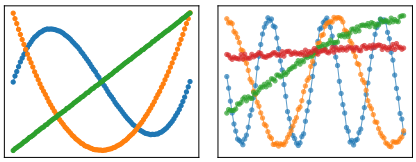

Figure 2: Visualizations of regression tasks: polynomial functions (left) and sine functions (right).

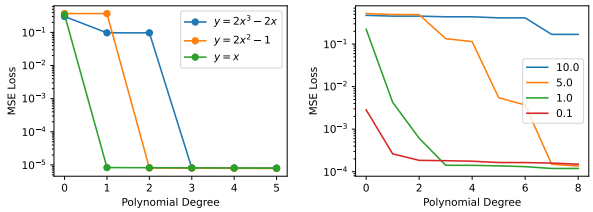

Figure 3: CSF of polynomial regression on functions in Figure 2. The CSF decreases most at the smallest degree that matches each function's overall shape.

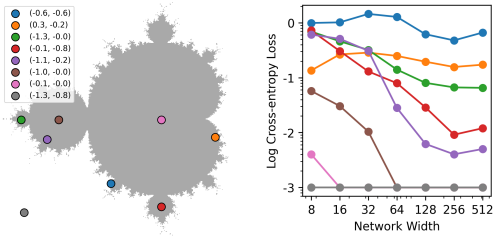

Figure 4: Graphs of the CSF at representative points in the Mandelbrot set.

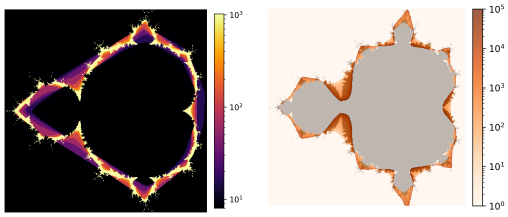

Figure 5: A two-dimensional map of properties of the datapoint-wise CSF on the Mandelbrot set: (left) smallest width that achieves loss 0.3 and (right) x-intercept of the best linear fit.

4 Experiments

4.1 Didactic Example: The CSF in Polynomial Regression

We demonstrate the CSF on a simple regression setting to build intuition for what properties of a dataset are captured by this function. We consider two types of regression tasks with varying complexity: polynomials of degree $(1, 2, 3)$ and sinewaves with frequencies $(0.1, 1.0, 5.0, 10.0)$. We construct the training dataset by uniformly sampling $n = 100$ points from $[0, 1]$. For the sinewaves only, we add Gaussian noise with standard deviation $\sigma = 0.01$ to the targets. These functions are visualized in Figure 2. As predictive families for the CSF, we consider the class of polynomials of degree $d = 0$ up to $d = 8$. We show the CSFs for these tasks in Figure 3; recall that the CSF will always converge to zero as $d \to \infty$, and the CSFs of different datasets differ only in lower degrees ($d < n$). First, note that for the datasets generated from polynomial functions, the CSF sharply decreases at the degree of the underlying function. For example, the CSF of the cubic function decreases most at degree 3, indicating that most of the structure of this function is captured by a cubic polynomial. The CSF decreases more smoothly for the datasets generated from sinewaves because these functions do not precisely fit any low-degree polynomial. Still, the CSF decreases at rates consistent with the functions' shapes, with higher frequencies fitting at a slower rate. This experiment demonstrates that the overall shape of the CSF can be a good indicator of the complexity of a dataset.

4.2 Analyzing the Mandelbrot Set with Pointwise CSF

We consider a binary classification problem where the input is a two-dimensional point (x, y) and the target is 1 if the point belongs to the Mandelbrot set (Brooks & Matelski, 1981). The boundary of the Mandelbrot set is known to have a fractal structure with broad patterns at each scale but with exceptions to these patterns at smaller scales. As predictive families, we consider two-layer ReLU networks with width $\in \{8, 16, 32, 64, 128, 256, 512\}$. The pointwise CSF for these functions at some representative points in Figure 4 shows that the CSF reflects the extent to which each point is typical in the overall shape of the function. The gray and pink points are densely surrounded by points with label 0 and 1, respectively, and this fact is reflected in their rapidly decreasing CSF. The red point is part of a smaller "pocket" of points with label 1, and its CSF decreases at a slow but steady rate. Finally, the blue point is quite close to the decision boundary with many fluctuations, and its CSF stays at a high value due to the difficulty in predicting its label. In Figure 5, we further visualize pointwise CSFs through a heat map of (1) the smallest function that achieves a particular loss and (2) the x-intercept of linear regression to each CSF. Both graphs reflect how much each point's output is typical within the dataset, with slightly differing tendencies: the first graph has highest values on the less common label in each region, while the second graph "overshoots" and has the highest value on the more common label in each region. We note that both are properties of the CSF, and we can extract different information by considering different statistics of each pointwise CSF.

References

Alessandro Achille, Glen Mbeng, and Stefano Soatto. Dynamics and reachability of learning tasks. *arXiv preprint arXiv:1810.02440*, 2018.

Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. The information complexity of learning tasks, their structure and their distance. *Information and Inference: A Journal of the IMA*, 10(1):51–72, 2021.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Robert Brooks and J Peter Matelski. The dynamics of 2-generator subgroups of psl (2, c). In *Riemann surfaces and related topics: Proceedings of the 1978 Stony Brook Conference, Ann. of Math. Stud*, volume 97, pp. 65–71, 1981. 4

Yann Dubois, Douwe Kiela, David J Schwab, and Ramakrishna Vedantam. Learning optimal representations with the decodable information bottleneck. *Advances in Neural Information Processing Systems*, 33:18674–18690, 2020.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html. 1, 6

Yansong Gao and Pratik Chaudhari. An information-geometric distance on the space of tasks. In *International Conference on Machine Learning*, pp. 3553–3563. PMLR, 2021.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D Manning. Conditional probing: measuring usable information beyond a baseline. *arXiv preprint arXiv:2109.09234*, 2021.

AN Kolmogorov. Complexity of algorithms and objective definition of randomness. *Uspekhi Mat. Nauk*, 29(4):155, 1974. 1, 2

Andrei N Kolmogorov. Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1):1–7, 1965. 1

Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will DNNs choose? a study from the parameter-space perspective. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=qRDQi3ocgR3.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 1

Alexander Kh Shen. The concept of (α, β)-stochasticity in the kolmogorov sense, and its properties. In *Soviet Math. Dokl*, volume 28, pp. 295–299, 1983. 6

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Nikolai K Vereshchagin and Paul MB Vitányi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004. 2

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*, 2020.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH. 1, 3, 6

Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KB5onONJIAU.

A Appendix

A.1 Important Special Cases of the CSF

We now review the relationship between the CSF and a few important existing measures of complexity. Many existing notions of complexity can be viewed as special cases of the CSF with a particular choice of predictive family $\mathcal{V}$ The comparative advantage of the CSF is that by using predictive families such as neural networks, we can capture the complexity of datasets that are not amenable to existing measures.

Kolmogorov strucure function. We now show that the Kolmogorov structure function (Definition 1) itself is a special case of $\mathcal{V}$-complexity for a specific choice of predictive family $\mathcal{V}$. Recall that for integer α, the value of the structure function $h_x(\alpha)$ is defined as the maximum log likelihood of x among all distributions with Kolmogorov complexity at most α. Therefore, $h_x(\alpha)$ is equal to the $\mathcal{V}$-complexity with respect to the predictive family $\mathcal{V}_\alpha$, which consists of all functions that map to uniform distributions over sets with description length at most α. Note that such predictive families have the property $\mathcal{V}_{\alpha=1} \subset \mathcal{V}_{\alpha=2} \subset \ldots \subseteq \Omega$ and $h(\mathcal{D}; \mathcal{V}_{\alpha=1}) \geq h(\mathcal{D}; \mathcal{V}_{\alpha=2}) \geq \ldots \geq 0$. While this is an appealing connection, the difficulty in direct evaluation remains: we cannot efficiently search over the predictive family $\mathcal{V}_\alpha$ for any value of α.

Predictive $\mathcal{V}$-entropy and $\mathcal{V}$-information. The $\mathcal{V}$-complexity is closely related to predictive $\mathcal{V}$-entropy (Xu et al., 2020), a measure of usable information under computational constraints. For $\mathcal{V} = \mathcal{V}$, the predictive $\mathcal{V}$-entropy is defined as

$$H_{\mathcal{V}}(Y \mid X) = \inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim p(X,Y)} \left[-\log f(y; x) \right], \tag{7}$$

where $p(X, Y)$ is a data distribution. Equation (3) replaces the expectation with respect to $p(X, Y)$ with an average over the given dataset $\mathcal{D}$. The two notions differ in that $\mathcal{V}$-complexity is a property of the dataset $\mathcal{D}$ itself, whereas $\mathcal{V}$-entropy is a property of an underlying data distribution, which we may not have access to. The $\mathcal{V}$-information of two random variables X, Y is an analogue of the Shannon mutual information $I(X; Y)$, but with respect to a predictive family $\mathcal{V}$. It is formally defined as

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y \mid X). \tag{8}$$

Pointwise $\mathcal{V}$-information Our notion of the datapoint-specific CSF is related to the pointwise $\mathcal{V}$-information (Ethayarajh et al., 2022), a measure of usable information within a single datapoint in the context of the entire dataset. For individual datapoint (x, y), PVI is defined as

$$\mathrm{PVI}(x \to y) = -\log_2 g[\varnothing](y) + \log_2 g'[x](y), \tag{9}$$

where $g, g' \in \mathcal{V}$ are two functions that achieve maximum likelihood conditioned on $\varnothing$ and x, respectively. A key property is that the predictive $\mathcal{V}$-information is the expected value of the PVI under the data distribution:

$$I_{\mathcal{V}}(X \to Y) = \mathbb{E}_{x,y \sim P(X,Y)}[\mathrm{PVI}(x \to y)]. \tag{10}$$

This is analogous to how the CSF is the sum of the datapoint-specific CSFs.

A.2 Equivalance between set-based and distribution-based definitions of KSF

Definition 6 (Kolmogorov structure function, set-based)**.** *Let $K(\cdot)$ be the Kolmogorov complexity: the length of the shortest program that outputs the given object and halts. Let x be data and $\alpha \in \mathbb{N}^+$ an integer bounding the complexity of sets to consider. The structure function of x maps the complexity α to a minimal cardinality of x-containing sets as:*

$$h_x : \alpha \mapsto \min_S \{\log |S| : S \ni x, K(S) \leq \alpha\}. \tag{11}$$

The use of sets S can be seen as an instance of a probabilistic model which puts weight $\frac{1}{|S|}$ on each member of S. This model class is known to be equivalent, up to a logarthmic additive term, to the model class of computable probaility distributions, as shown in the following lemma.

Lemma 1. (Shen, 1983) *Let S be the set of strings of length n, and let P be a computable probability distribution on S. We have*

$$-\log P(x) = \log |S| + O(\log n). \tag{12}$$