INFORMATION COMPLEXITY AND ESTIMATION

Dror Baron[1]

[1]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA, barondror@ncsu.edu

ABSTRACT

We consider an input x generated by an unknown stationary ergodic source X that enters a signal processing system J, resulting in $w = J(x)$. We observe w through a noisy channel, $y = z(w)$; our goal is to estimate x from y, J, and knowledge of $f_{Y|W}$. This is *universal* estimation, because f_X is unknown. We provide a formulation that describes a trade-off between information complexity and noise. Initial theoretical, algorithmic, and experimental evidence is presented in support of our approach.

1. INTRODUCTION

Universal algorithms [1–5] achieve the best possible performance asymptotically – without knowing the input statistics. These algorithms have had tremendous impact in lossless compression, which is crucial for data backups and transmissions. In sharp contrast, universal algorithms have made much less impact in other areas.

Estimation algorithms attempt to recover an input from noisy measurements (Figure 1). Numerous estimation problems have received great attention including the additive noise scalar channel, $y = x + z$ [6]; linear matrix multiplication with additive noise, $y = Jx+z$ with applications including compressed sensing [7–9], finance, medical and seismic imaging; universal lossy compression [4, 5, 10], where the goal is to find compressible x that is sufficiently close to y; nonlinear regression, where $J(x)$ is nonlinear; and distributed signal processing.

In these estimation problems, the common goal is to estimate the input x from knowledge of the noisy measurements y and measurement system J. To do so, we must exploit all statistical structure in x. A particularly challenging type of statistical structure is the appearance of spatial or temporal dependencies in data. In images, such dependencies can be captured by dictionary learning or employing energy compacting transforms. In other problems, the statistical dependencies might be more subtle. Following the lead of universal lossless compression, we assume that the input x was generated by an unknown stationary ergodic source X. It is well known that stationary ergodic models capture the statistics of text files well, and hence the success of universal lossless compressors. Stationary ergodic models have also been incorporated in speech denoising and enhancement, and appear prominently in hidden Markov models.

One approach to *universal estimation* relies on Kolmogorov complexity [11]. For a prospective $\widehat{x}$, the Kolmogorov complexity $K(\widehat{x})$ is the length of the shortest computer program that can compute $\widehat{x}$. Donoho [12] proposed a Kolmogorov-based estimator for the white scalar channel, $y = x + z$. Despite related extensions to compressed sensing [8, 9], *what is missing in the literature is a universal approach in arbitrary measurement systems that would support noise and unknown stationary ergodic input distributions.*

We propose to perform universal estimation in (potentially nonlinear) signal processing systems from noisy measurements. The algorithmic component of our work features a harmonious marriage of scalar quantization, universal lossless compression, and Markov chain Monte Carlo. We evaluate the estimated input $\widehat{x}$ over a quantized grid and optimize for the trade-off between information complexity (lossless coding length) of $\widehat{x}$ and how well $\widehat{x}$ explains the measurements y. We report promising preliminary theoretical and numerical results.

2. INFORMATION COMPLEXITY FORMULATION

We focus on the setting where the lengths M of the output y and N of the input x both grow to infinity, $M, N \to \infty$. We further assume that their ratio is finite and positive, $\lim_{N\to\infty} \frac{M}{N} = \delta > 0$. Similar settings have been discussed in the literature, e.g., [13]. Since x was generated by an unknown source, we must search for an estimation mechanism that is agnostic to the specific distribution f_X.

Kolmogorov complexity: For $x \in \mathbb{R}^N$, the Kolmogorov complexity [11] of x, denoted by $K(x)$, is the length of the shortest computer program that can compute x. To be more precise, $K(x)$ is the length of the shortest input to a Turing machine [14] that generates x and then halts. We limit our discussion to Turing machines whose "input tapes" consist of bits. Consider the shortest program $\mathcal{P}(x)$ that generates x. From the perspective of a source encoder [6], we say that $\mathcal{P}(x)$ is a code for x.

Having linked Turing machines [14] and data compression [6], let us *temporarily* limit the discussion to *discrete valued* x generated by a stationary ergodic source X. Each such x is generated with probability $p_X(x)$, and it is easily shown that the per-symbol Kolmogorov coding length $K(x)$ converges to the entropy rate H almost surely, $\lim_{N\to\infty} \frac{1}{N} K(x) = H$ [6]. Noting that universal lossless compressors [1, 2] achieve H asymptotically

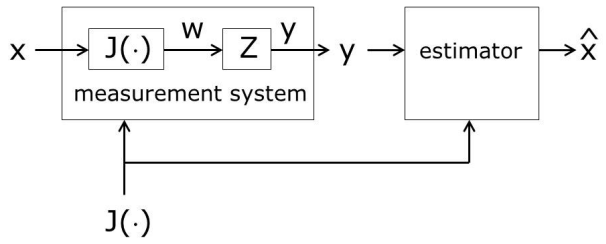

Figure 1. **Measurement and estimation system***: An* input $x \in \mathbb{R}^N$ *generated by an unknown stationary ergodic source* X *is processed by a known (potentially nonlinear) operator* J *to produce* $w = J(x) \in \mathbb{R}^L$*. A probabilistic* noise *operator* z *that implies a known probability density* $f_{Y|W}(y|w = J(x))$ *is applied to* w*, the measurements are* $y = z(J(x))$*. Our goal is to estimate* x *using* $y \in \mathbb{R}^M$ *and* J*, resulting in* $\widehat{x} \in \mathbb{R}^N$*. Although our emphasis is on real-valued* w, x, y*, discrete-valued signals and operators are allowed.*

for discrete valued stationary ergodic sources [6], we see that these algorithms achieve the per-symbol Kolmogorov complexity almost surely.

Kolmogorov sampler: For additive white Gaussian noise, $y = x + z$, Donoho [12] proposed the *Kolmogorov sampler*,

$$\widehat{x}_{KS} = \arg\min_{\hat{x}}\{K(\hat{x}) - \log(f_Z(z = y - \hat{x}))\}.$$

For stationary ergodic X, $\widehat{x}_{KS}$ is sampled from the posterior $f_{X|Y}(y|x)$, where the mean square error, $E[(\widehat{x}_{KS} - x)^2]$, is twice larger than the Bayesian minimum mean square error (MMSE) [12].

In a later paper, Donoho et al. discussed a Kolmogorov estimator for compressed sensing $y = Jx$ [8]; their estimator ignores noise, and is of limited practical interest. For the noisy version of this problem, $y = Jx + z$, Haupt and Nowak [9] described a complexity measure that, when optimized, produces the LASSO algorithm [15]. To the best of our knowledge, Haupt and Nowak did not pursue complexity based regularization beyond iid signals and additive white Gaussian noise (AWGN).

Quantization and estimation: The overwhelming majority of real numbers have infinite Kolmogorov complexity. Nonetheless, some scalars $x \in \mathbb{R}^N$ can be represented by a finite length $\mathcal{P}(x)$. In practice, it is impossible to compute $K(x)$ even for discrete alphabets. At the same time, we have seen that universal lossless source codes [1, 2] achieve per-symbol Kolmogorov coding length almost surely [6]. To represent continuous valued $\widehat{x}$, we apply a scalar quantizer, $Q : \widehat{x} \in \mathbb{R}^N \longrightarrow x' \in Q^N$, and then compress $x' = Q(\widehat{x})$ with a universal lossless compressor U with coding length $U(x')$, where quantization levels $Q \subset \mathbb{R}$ consist of a *finite* subset of $\mathbb{R}$, and performing an optimization over $\widehat{x} \in Q^N$ reduces the complexity of the estimation problem from infinite to combinatorial. Note that we generate x' by independently quantizing each entry of x with Q. This *encoder* first describes the quantizer Q and then compresses $Q(x)$. The coding length, which we desire to minimize, is denoted by $U(Q(x))$ or $U(x)$.

It would seem that we must search for a good scalar quantizer Q (Section 3), but *data-independent* reproduction levels are of theoretical interest,

$$\mathcal{R} \triangleq \left\{-\frac{\gamma^2}{\gamma}, -\frac{\gamma^2 - 1}{\gamma}, \ldots, \frac{\gamma^2}{\gamma}\right\}, \quad \gamma = \lceil \log(N) \rceil.$$

As N increases, $\mathcal{R}$ will quantize a broader range of values of x to a greater resolution. An encoder based on $\mathcal{R}$ need not describe the structure of the data-independent quantizer, because N is known. That is, $U(\mathcal{R}(x))$ only accounts for the length of the universal code U.

Universal MAP estimation: We perform maximum *a posteriori* (MAP) estimation over possible sequences $\widehat{x} \in \mathcal{R}^N$, where the prior $p_X(x) = 2^{-U(\widehat{x})}$ utilizes the coding length $U(\widehat{x})$ of some universal lossless compressor [1, 2],

$$\widehat{x}_{MAP} = \arg\min_{\hat{x} \in \mathcal{R}^N}\{U(\hat{x}) - \log(f_{Y|W}(y|w = J(\hat{x})))\}, \tag{1}$$

where we note that $\mathcal{R}(\hat{x}) = \hat{x}$ for $\hat{x} \in \mathcal{R}^N$. Our MAP estimator is applicable to *any* signal processing system J and supports *any* probabilistic noise operators, it is closely related to universal prediction [2, 3].

Estimation performance: We have promising preliminary theoretical results using the data-independent quantizer $\mathcal{R}$. In universal lossy source coding of analog (continuous valued) sources [4], we have shown with Weissman that $\widehat{x}_{MAP}$ (1) achieves the rate distortion function for finite variance stationary ergodic sources in an appropriate asymptotic sense. That is, $U(\mathcal{R}(\widehat{x}))$ offers a sufficiently good approximation to $K(\widehat{x})$ in universal lossy compression, where we chose $U(\widehat{x})$ to be empirical entropy of blocks of $q = O(\log(N))$ symbols in $\widehat{x}$. In universal compressed sensing [16], we have shown with Duarte that under minor technical conditions on f_X, performing MAP estimation over the discrete alphabet $\mathcal{R}$ converges to the MAP estimate over the continuous distribution f_X asymptotically, where we used i.i.d. zero-mean Gaussian noise $z \in \mathbb{R}^M$ with known variance. It remains to be seen whether $\mathcal{R}$ or other data-independent quantizers are useful for *arbitrary* nonlinear measurement systems.

In terms of the mean square error, we would expect $\widehat{x}_{MAP}$ to perform well in Donoho's scalar channel setting, $y = x + z$. With Duarte [16], we have promising results for the compressed sensing (linear matrix multiplication) channel, $y = Jx + z$, where we approximated $\widehat{x}_{MAP}$ (1) by a Markov chain Monte Carlo (MCMC) [17] algorithm (Section 3). Figure 2 illustrates recovery results from Gaussian measurement matrices for a source with i.i.d. Bernoulli entries with nonzero probability of 3%. Our MCMC algorithm outperforms ℓ_1-norm minimization, which is a well-known compressed sensing reconstruction (estimation) algorithm [7], except when the number of measurements M is low. Comparing MCMC to the minimum mean square error (MMSE) achievable in the Bayesian regime with known statistics [13], the square error achieved by MCMC is *three* times larger. One is left to wonder whether the mean square error performance of our algorithm might also be double the MMSE, particularly in the limit of infinite computation (Section 3).

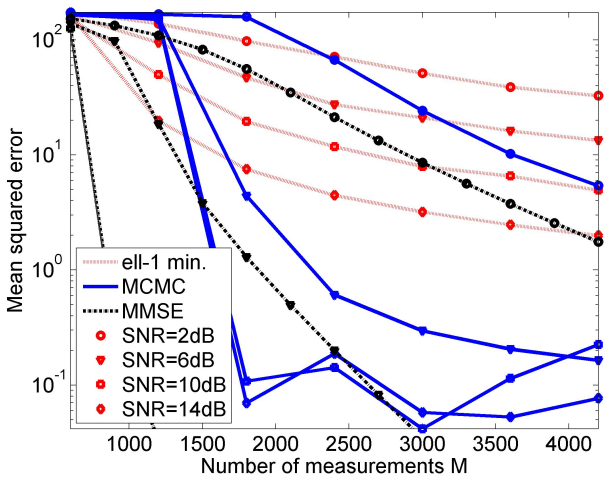

Figure 2. *Universal Markov chain Monte Carlo (MCMC) [16] and ℓ_1-norm minimization [7] recovery results for a source with i.i.d. Bernoulli entries with nonzero probability of 3% as a function of the number of Gaussian random measurements M for different signal to noise ratio (SNR) values.*

Taking Kolmogorov beyond MAP: The Kolmogorov sampler $\widehat{x}_{KS}$ samples from the posterior [12]; it throws away all the statistical information it has on signals $\widehat{x}$ that differ from $\widehat{x}_{KS}$. Seeing that the mean square error obtained by $\widehat{x}_{KS}$ is double the MMSE, there is great potential to reduce estimation error over our Kolmogorov-based MAP estimator $\widehat{x}_{MAP}$ (1). We therefore propose Kolmogorov-based conditional expectation,

$$\begin{aligned}\widehat{x}_{MSE} &= E[x|J,y] \\ &= \frac{\sum_{\hat{x}\in\mathcal{R}^N} \hat{x}\cdot 2^{-U(\hat{x})} f_{Y|W}(y|w=J(\hat{x}))}{\sum_{\hat{x}\in\mathcal{R}^N} 2^{-U(\hat{x})} f_{Y|W}(y|w=J(\hat{x}))},\end{aligned}$$

where we employ the universal prior, $p_X(\hat{x}) = 2^{-U(\mathcal{R}(\hat{x}))}$. It is well known that conditional expectation achieves the MMSE of the Bayesian regime, and this estimator should perform well. Interestingly, when the signal to noise ratio (SNR) is low, the Bayesian MMSE is sizable, and achieving double the MMSE is unimpressive. In these low SNR settings, $\widehat{x}_{MSE}$ should estimate *much* better than $\widehat{x}_{MAP}$.

In some signal processing systems, one wants to minimize some other (not necessarily quadratic) distortion metric $D(x,\widehat{x})$. The universal prior is readily invoked by defining the Kolmogorov conditional probability,

$$p_{X|Y}(x|y) = \frac{p_{Y|X}p_X}{p_Y} \propto p_{Y|X}p_X,$$

and taking the minimizing expression gives the Kolmogorov-based estimator for $D(\cdot)$,

$$\widehat{x}_D = \arg\min_w \left\{ \sum_{\widehat{x}\in\mathcal{R}^N} D(\widehat{x},w) f_{Y|W}(y|w=J(\hat{x}))2^{-U(\hat{x})} \right\}.$$

For scalar channels and iid noise, Sivaramakrishnan and Weissman [18] described a universal denoising algorithm that estimates x by $\widehat{x}_{SW}$, its expected error $E[D(x,\widehat{x}_{SW})]$ converges to the Bayesian risk asymptotically in an appropriate stochastic setting. For scalar channels and iid noise, our expected estimation error $E[D(x,\widehat{x}_D)]$ should also be asymptotically optimal. The performance in arbitrary signal processing systems J is an open question.

3. ALGORITHMS

In principle, $\widehat{x}_{MAP}$ can be computed by evaluating the Kolmogorov-based posteriors of $|\mathcal{R}|^N$ possible sequences $\mathcal{R}(x)$. This is better than continuous estimation, but still computationally intractable. Instead, we perform this optimization using Markov chain Monte Carlo (MCMC) [5, 17], where $U(\widehat{x}) = H_q(\widehat{x})$ is the empirical entropy of blocks of $q = O(\log(N))$ symbols of $\widehat{x}$.

Markov chain Monte Carlo: We use MCMC [17] to approximate $\widehat{x}_{MAP}$, which is the globally optimal MAP minimizer. To keep things simple, assume that $\widehat{x} \in \mathcal{R}^N$ is a candidate estimate. Define the Boltzmann PDF,

$$f_s(\widehat{x}) \triangleq \frac{1}{\zeta_s}\exp(-s[H_q(\widehat{x}) - \log(f_{Y|W}(y|w=J(\widehat{x})))]), \quad (2)$$

where $H_q(x)$ is the empirical entropy of blocks of q symbols in x [2, 4, 5, 16], $q = O(\log(N))$ to ensure convergence of the empirical entropy to the entropy rate [6], $s > 0$ is inversely related to temperature in an analogous statistical physics heat-bath setting [17], and ζ_s is a normalization constant. To sample from the Boltzmann PDF (2), we use a *Gibbs sampler*: in each iteration, a single element $\widehat{x}_n$ is generated by resampling from the PDF, while the rest of $\widehat{x}$ remains unchanged. The key idea is to reduce temperatures slowly enough for the randomness of Gibbs sampling to eventually drive MCMC out of any local minimum toward the globally optimal $\widehat{x}_{MAP}$.

Adaptive quantizer: Jalali and Weissman [5] have used MCMC to approach the fundamental rate distortion (RD) limits [6] in lossy compression of binary inputs. For continuous valued (analog) sources [4], using the data-independent quantizer $\mathcal{R}$ in MCMC asymptotically achieves the RD function universally for stationary ergodic continuous amplitude sources. However, $\mathcal{R}$ grows with the input length, slowing down the convergence to the RD function, and is thus an impediment in practice.

To address this issue, we next propose an MCMC-based algorithm that uses an *adaptive quantizer* Q. The ground-breaking work by Rose on the discrete nature of the Shannon codeboook for iid sources when the Shannon lower bound is not tight [19] suggests that, for most sources of practical interest, restriction of the quantizer Q to a smaller number of levels does not stand in the way of attaining the fundamental compression limits. When employed on such sources, our latter algorithm zeroes in on the finite quantizer, and thus enjoys rates of convergence commensurate with the small-quantizer setting.

Numerical results: In universal lossy compression of analog sources [4], we have developed an algorithm that optimizes the quantizer for square error, and have promising preliminary results. Figure 3 compares results for an

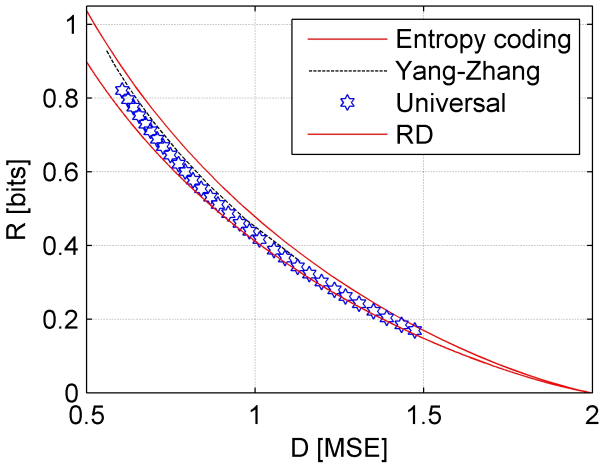

Figure 3. **Universal lossy compression**: *Rate R vs. distortion D of entropy coding [20], results by Yang and Zhang [10], average rate and distortion of our universal lossy compression algorithm [4], and the RD function [6] for length-15000 iid Laplace inputs, $f_X(x) = \frac{1}{2}e^{-|x|}$.*

iid Laplace input, $f_X(x) = \frac{1}{2}e^{-|x|}$, achieved by entropy coding [20], a deterministic approach by Yang and Zhang [10], and our universal MCMC algorithm [4].

In our universal compressed sensing work with Duarte [16], we focused on development of a fast routine for optimizing the quantizer; this routine greatly accelerates the algorithm. We have seen in Figure 2 for a source with i.i.d. Bernoulli entries with nonzero probability of 3% that MCMC outperforms ℓ_1-norm minimization, except when the number of measurements M is low. We have additional results, but omit these for brevity; MCMC generally estimates the input signal x well, but much work remains to be done.

4. ACKNOWLEDGMENTS

I wish to thank Marco Duarte and Tsachy Weissman for permission to discuss our joint work [4, 16]. Thanks also to Neri Merhav, Deanna Needell, and Phil Schniter for enlightening discussions.

5. REFERENCES

[1] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[2] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, Jul. 1984.

[3] A. C. Singer and M. Feder, "Twice universal linear prediction of individual sequences," in *Proc. Int. Symp. Inf. Theory (ISIT1998)*, Aug. 1998.

[4] D. Baron and T. Weissman, "An MCMC approach to universal lossy compression of analog sources," submitted for publication and Arxiv preprint arXiv:1107.2972, 2011.

[5] S. Jalali and T. Weissman, "Rate-distortion via Markov chain Monte Carlo," in *Proc. Int. Symp. Inf. Theory (ISIT2008)*, Jul. 2008, pp. 852–856.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2006.

[7] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[8] D. Donoho, H. Kakavand, and J. Mammen, "The simplest solution to an underdetermined system of linear equations," in *Int. Symp. Inf. Theory (ISIT)*, Jul. 2006, pp. 1924–1928.

[9] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4036–4048, Sep. 2006.

[10] E. Yang and Z. Zhang, "Variable-rate trellis source encoding," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 586–608, Mar. 1999.

[11] A.N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems Inf. Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[12] D.L. Donoho, "The Kolmogorov sampler," 2002.

[13] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983, 2005.

[14] A.M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[15] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Stat. Soc. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[16] D. Baron and M. F. Duarte, "Universal MAP estimation in compressed sensing," submitted to Allerton Conf. on Comm., Control, and Computing, 2011.

[17] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, pp. 721–741, Nov. 1984.

[18] K. Sivaramakrishnan and T. Weissman, "Universal denoising of discrete-time continuous-amplitude signals," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5632–5660, 2008.

[19] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1939–1952, Nov. 1994.

[20] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, Kluwer, 1993.