SIMPLIFIED PROBABILITY MODELS FOR GENERATIVE TASKS: A RATE-DISTORTION APPROACH

Gustav Eje Henter and W. Bastiaan Kleijn

Sound and Image Processing Laboratory, KTH (Royal Institute of Technology)
Osquldas väg 10, SE-100 44, Stockholm, Sweden
phone: +46 8 790 7420, email: {gustav.henter,bastiaan.kleijn}@ee.kth.se

ABSTRACT

We consider using sparse simplifications to denoise probabilistic sequence models for generative tasks such as speech synthesis. Our proposal is to find the least random model that remains close to the original one according to a KL-divergence constraint, a technique we call *minimum entropy rate simplification* (MERS). This produces a representation-independent framework for trading off simplicity and divergence, similar to rate-distortion theory. Importantly, MERS uses the cleaned model rather than the original one for the underlying probabilities in the KL-divergence, effectively reversing the conventional argument order. This promotes rather than penalizes sparsity, suppressing uncommon outcomes likely to be errors. We write down the MERS equations for Markov chains, and present an iterative solution procedure based on the Blahut-Arimoto algorithm and a bigram matrix Markov chain representation. We apply the procedure to a music-based Markov grammar, and compare the results to a simplistic thresholding scheme.

1. INTRODUCTION

In machine learning, an interesting duality exists between *discriminative tasks* such as speech recognition, and *generative tasks* such as speech synthesis. Both can be seen as mappings between observation space and model space, but in opposite directions. Generative and discriminative tasks alike can be addressed using so-called *generative models*, of which Markov chains and hidden Markov models (HMMs) [1] are common examples. Hidden Markov models, in particular, are used in modern systems for speech recognition as well as speech synthesis [2, 3].

However, just because the same model family can be applied for generative and discriminative problems, it does not follow that the exact same *model* will be optimal in both cases. On the contrary, the practical requirements for a good model typically differ between the two applications, so what is optimal in one case need not be the best strategy in the other; see [4]. While the problem of adapting models to increase recognition performance has been widely studied [5, 6, 7, 8], we will consider the converse task of improving models for purposes of sampling and synthesis. To this end, we propose *minimum entropy rate simplification* (MERS), a rate-distortion like framework for simplifying and sparsifying estimated probability models for stochastic sequences, removing noise and errors inherited from the training data.

The paper is laid out as follows: section 2 describes the benefits of sparse, simplified generative models. We then introduce and discuss the general MERS framework in section 3. Thereafter, in section 4, we describe the concrete optimization problem that arises in the special case of Markov chains, present a solution algorithm, and apply it to a simple music grammar. Section 5 then rounds off with conclusions and suggestions for future work.

2. BACKGROUND

As an example of the different requirements in discriminative versus generative settings, we shall consider the dual topics of speech recognition and speech synthesis. In both cases, the typical approach to learning revolves around one or more generative models trained on human speech data, often using a maximum likelihood estimation technique such as Baum-Welch training for hidden Markov models [9], a special case of the EM-algorithm [10]. However, after the training stage, paths diverge [4].

2.1 Discriminative Tasks and Smoothing

For recognition tasks, it is common practice to apply some kind of *smoothing* to ML-estimated models [7], which increases the amount of randomness and number of possible outcomes. This is to allow for the vast variety of different behaviours present in real, spontaneous speech, which is typically much greater than the training data can represent [11]. *Additive smoothers*, including pseudocount methods such as Laplace's rule, are a common choice, though many alternatives exist, e.g., [6, 8]. Pseudocounts are often motivated and interpreted as a Bayesian prior.

The net result of smoothing is to increase the probability of rare events, and assign small, nonzero probabilities to events previously deemed impossible by the model. Were this not done, many erroneous or simply unusual constructs that tend to occur in the real world may have probability zero under the unsmoothed maximum likelihood model [11]. These zeroes are known to be problematic and degrade practical performance: for example, if only grammatically correct interpretations of speech have nonzero probabilities, any grammatical mistake by the speaker might make recognition impossible.

2.2 Generative Tasks

In the case of speech synthesis and other generative tasks, the situation is the opposite to the above: one would like to decrease rather than increase the room for errors in samples from the model. This would reduce the importance of occasional idiosyncrasies in the training data and generally filter out unlikely and uncharacteristic behaviour, such as grammatically incorrect speech, that the initially estimated model might still allow. These mistakes and unpredictable behaviours are generally undesirable from a communication standpoint in a practical speech synthesis system, even if removing them makes the model in some sense less realistic.

As another example, we may consider synthesizing species-specific birdsong. A training dataset of field recordings may not always be clean, but could include background sounds and occasional interference from other singing birds. By reducing the range of behaviours that can be expressed by the model, more consistent output may be obtained, where disturbances from the training material are eliminated or suppressed.

In both examples above we presume errors to be inherent to the data process rather than a finite sampling effect. This is a situation where Bayesian approaches such as [12]—which in terms of objective is quite similar to our MERS proposal—are not directly applicable, since the impact of the Bayesian prior decreases with an increasing amount of data.

2.3 Sparsity

We have described the need for simplifying stochastic processes so that uncommon, uncharacteristic behaviour is removed or reduced.

The result could be seen as a sort of caricature of the original process, exaggerating the most prominent features, or we may consider it a "sharpening," in the sense that it has the opposite effect of smoothing. However, it can also be considered a kind of sparsification of the model output—we want to decrease the number of possible outcomes and reduce the size of the typical set. In many model classes where parameters can be interpreted as probabilities, including Markov chains and HMMs, this directly carries over to imply sparsity in model parameter space, though not all models may admit such an interpretation.

Sparse representations, in general, is a topic area that has seen much recent interest. Well-known techniques such as least angle regression (LARS/Lasso) [13, 14], support vector machines [15], and the wavelet paradigm in signal processing [16] can all be considered examples of sparse approaches. Methods for obtaining sparse probability models have also been addressed before, e.g., [17], but this does not apply to stochastic processes that are not i.i.d., as considered here.

General advantages of sparse representations include that they compress easily [16], may be faster to process [18], and tend to be more amenable to human interpretation [13, 12]. In general, sparse representations echo the principle of Occam's razor that "plurality should not be posited without necessity."

Though explicit constraints provide one route to sparsity, e.g., [19], sparsity in several of the methods above emerges as a natural by-product of their construction. A classic example of emergent sparsity is reverse water-filling in source coding, where, in optimal coding of stationary stochastic Gaussian processes, certain frequencies (or variables in the discrete case) are omitted completely from the compressed description [20, 21]. This will serve as the model for our efforts for identifying sparse simplifications of stochastic processes.

3. RATE-DISTORTION SIMPLIFICATION

We shall now adapt the rate-distortion framework from source coding to the task of simplifying stochastic processes, which yields the MERS framework. This will involve a brief discussion of traditional rate-distortion theory and how it applies to our problem, after which we describe how to select suitable analogues of rate and distortion for stochastic processes. In section 4, we shall then consider the important special case of a Markov chain.

3.1 Rate-Distortion Theory

Let X be a stochastic variable with known, fixed distribution $F_X(x)$, and let $\widehat{X}$ be some approximation of X reconstructed from partial information about X. In our particular application, these variables will be stochastic processes, and we aim to find a simple underlying X using the incomplete information available through the disturbed observations $\widehat{X}$.

Rate-distortion theory in lossy source coding concerns the trade-off between the rate $R = I(X, \widehat{X})$ (the average number of bits transmitted) and the expected distortion of the signal given the transmitted information. The latter is quantified through a *distortion measure* $D(x, \widehat{x}) \geq 0$, with equality if $\widehat{x} = x$. The goal of source coding is to choose a distribution $F_{\widehat{X}|X}(\widehat{x} \mid x)$ for $\widehat{X}$ that strikes an optimal balance between the contradictory objectives of low average rate (simplicity) and low mean distortion (dissimilarity). Coding can thus be seen as a simplification scheme, similar to what we want to derive.

By constraining either rate or distortion, the other variable can be minimized. It does not matter which is fixed; using Lagrange multipliers, both approaches can be recast as unconstrained minimization of a weighted functional, as in

$$\min_{F_{\widehat{X}|X}(\widehat{x}|x)} I(X, \widehat{X}) - \beta \mathbb{E}_{F_X(x)} D(X, \widehat{X}). \tag{1}$$

The minima over the range of Lagrange multipliers $\beta \geq 0$ define a convex, nonincreasing *rate-distortion function* $D(R)$ which lower bounds compression performance for a given distribution $F_X(x)$; only rate-distortion pairs (R, D) on or above the curve are achievable.

The balance between rate and distortion above can be adjusted continuously through the variable β. Similar information-theoretic trade-offs recur in, for instance, the opposing forces of relevance and compression within the information bottleneck framework [22], and the semi-supervised CRF learning framework in [23]. We shall let a trade-off of the same form as above define our model simplification scheme, by studying the rate and distortion components in turn, arguing for natural generalizations that produce sparse simplifications.

3.2 Rate Minimization

Let $\widetilde{X}_t$ and X_t for $t \in \mathbb{Z}$ be strictly stationary and ergodic stochastic processes over a space $\mathscr{X}$. $\mathscr{X}$ may be either discrete or continuous. We will take the properties of $\widetilde{X}_t$ to be known and fixed—typically, this is a model with parameters estimated from possibly impure training data—and seek a suitable X_t, a cleaned version of the $\widetilde{X}_t$-process that strikes an optimal balance between simplicity and similarity.

We want X_t to be a simplification of $\widetilde{X}_t$ where rare events are removed or generally de-emphasized. Intuitively, the fewer outcomes that are possible, the less random the output becomes, a notion formalized by the classic entropy concept. To obtain an optimally simple stochastic process X_t, it thus appears sensible to minimize the *entropy rate* of the model as

$$H_\infty(X_t) = \lim_{T\to\infty} \frac{1}{T} H(X_t, X_{t+1}, \ldots, X_{t+T-1}), \tag{2}$$

or the analogous *differential entropy rate*

$$h_\infty(X_t) = \lim_{T\to\infty} \frac{1}{T} h(X_t, X_{t+1}, \ldots, X_{t+T-1}) \tag{3}$$

if $\mathscr{X}$ is not discrete. These are parameter-independent information-theoretic measures of disorder, with units of bits, nats or similar, depending on the logarithm used to define the entropies.

The lower the entropy rate, the more predictable a process becomes, and on average only a few outcomes will have any appreciable probability. The extreme points where the rate is identically zero correspond to processes that, with probability one, are completely deterministic once a single sample is known (and are thus not necessarily ergodic).

We note that the entropy of a general stochastic variable is a concave function over the unit simplex with minima at the corners. Thus algorithms minimizing the entropy rate, as we wish to do for X_t here, might converge on points that are not globally optimal. This is not necessarily a grave concern—methods such as hidden Markov models work well in practice despite the fact that the training algorithms are not certain to find global optima.

3.3 Distortion Constraint

In rate-distortion theory, a low rate is balanced against the undesirable distortion it induces in the reconstructed variable $\widehat{X}$. To prevent oversimplification, we similarly want to ensure that the difference between X_t and the measurements $\widetilde{X}_t$ is not too great, according to some appropriate measure of distortion. Selecting this distortion measure is not as straightforward as minimizing the rate.

The classic case of reverse water-filling with Gaussian variables occurs for the squared error distortion function $d(x, \widehat{x}) = \|x - \widehat{x}\|_2^2$, but this measure is parameterization dependent and not even defined for categorical variables. Instead, we describe an approach involving the *Kullback-Leibler divergence*, or *relative entropy*, which in the discrete case has the form

$$D_{\mathrm{KL}}(P||Q) = \sum_i p_P(i) \log \frac{p_P(i)}{p_Q(i)}, \tag{4}$$

given random variables P and Q.

This is an information-theoretic error measure that also is parameterization-independent, and has been used in other frameworks inspired by rate-distortion theory such as the information bottleneck approach.

To properly adapt the KL-divergence for our purposes, we must keep in mind the different roles of the two arguments P and Q. In many applications, the divergence $D_{\mathrm{KL}}(P||Q)$ is interpreted as the mean reduction in per-symbol log-likelihood that occurs when using the probability model Q, compared to the true sample distribution P. In source coding, this equals the average excess number of bits, nats, or similar, consumed when coding the variable P using a code optimal for Q. Commonly, then, the argument P represents the fixed, actual distribution of the data (known or inferred from observations), while Q is some approximation thereof. This strongly discourages sparsity in Q—in source coding, for instance, it is vital that nonzero probability symbols all have finite length codewords—so care must be taken not to rule out sparse output when using this divergence.

In our case, a different understanding of what constitutes P and Q is appropriate, compared to many other applications. We consider the observations from $\widetilde{X}_t$ as a corruption, a noisy approximation of some clean underlying process X_t, for example birdsong or grammatically correct speech. Hence it makes sense to reverse the conventional ordering and constrain the limiting *relative entropy rate* $D^{\infty}_{\mathrm{KL}}(X_t||\widetilde{X}_t)$ defined through

$$\lim_{T\to\infty}\frac{1}{T}D_{\mathrm{KL}}((X_t,\ldots,X_{t+T-1})\,||(\widetilde{X}_t,\ldots,\widetilde{X}_{t+T-1})), \tag{5}$$

or the analogous differential entropy rate for continuous-valued processes. We shall assume these quantities exist, which is assured if the processes are Markovian [24]. Another notable example where the second argument in the KL-divergence is considered fixed instead of the first one is variational Bayesian inference [25].

Constraining $D^{\infty}_{\mathrm{KL}}(X_t||\widetilde{X}_t)$ represents a belief that short sequences from the underlying X_t are not too unlikely to be observed in $\widetilde{X}_t$ unaltered. This harshly punishes needless non-sparsity; any X_t-process which has additional nonzero probability outcome sequences compared to the observations $\widetilde{X}_t$ will incur an infinite penalty. Excess sparsity in X_t, so that $\Omega(X_t)\subset\Omega(\widetilde{X}_t)$, leads to a more modest, finite divergence.

3.4 Method Overview

Summing up the reasoning above, we propose to find a simplified, sparse model X_t of a process $\widetilde{X}_t$ by solving the optimization problem

$$\min_{X_t\in\Xi} H_{\infty}(X_t) \tag{6}$$

subject to

$$D^{\infty}_{\mathrm{KL}}(X_t||\widetilde{X}_t)\le D, \tag{7}$$

where D is a free parameter and Ξ is a class of stationary, ergodic discrete-time models. We take appropriate differential entropies and divergences if the outcome set $\mathscr{X}$ is not discrete. Note that this reduces to a problem from regular rate-distortion theory if the processes considered are i.i.d.

Just as in rate-distortion theory and the information bottleneck framework, the opposing goals of low rate and low distortion in the problem enable a continuous trade-off between the original estimated process $\widetilde{X}_t$ and complete determinism at zero rate, controlled by the tolerable distortion D. The different possible optima trace out a nonincreasing rate-distortion function $R(D)$. From a variational, information-theoretic perspective, the optimal X_t is determined by trading bits of increased order for bits of divergence at the exchange rate specified by the Lagrange multiplier β corresponding to the constraint (7).

A practical issue with rate-distortion theory is that few closed-form solutions have been found. Many of these are available in [26, 21]. Sometimes, fundamental quantities such the entropy rate (6) may be difficult to write down explicitly, for instance if X_t is a hidden Markov process [27]. Nevertheless, there are many important cases where this is not a problem. A particularly useful example is the class of stationary and ergodic Markov chains, which we will consider next.

4. SPARSE MARKOV CHAINS

We have presented a general, abstract rate-distortion problem for simplifying probability models, with the intent of eliminating noise and disturbances for sampling applications. To get a more concrete impression of how the MERS framework operates, we shall now address how it applies to a simple example, namely that of ordinary Markov chains. We write down the explicit optimization problem that results for this particular case, present an iterative solution algorithm, and demonstrate sparsity in an application.

4.1 Optimization Problem Formulation

Let $\widetilde{M}_t$ be a given stationary, ergodic first-order Markov chain on an alphabet $\mathscr{A}$ of cardinality $N<\infty$, defined by the transition probability matrix $\widetilde{\mathbf{A}}\in[0,1]^{N\times N}$ such that

$$(\widetilde{\mathbf{a}})_{ij}=P(\widetilde{M}_{t+1}=j\mid\widetilde{M}_t=i) \tag{8}$$

for all $i,j\in\mathscr{A}$. Given $\widetilde{\mathbf{A}}$, the process has a unique stationary distribution $\widetilde{\pi}\in[0,1]^N$ such that $(\widetilde{\pi})_i=P(\widetilde{M}_t=i)$, which solves the eigenvector equation $\widetilde{\mathbf{A}}^T\widetilde{\pi}=\widetilde{\pi}$. We shall require $\widetilde{\pi}>\mathbf{0}$ (meaning that all elements are greater than zero), else some symbols in $\mathscr{A}$ are not emitted and can be removed from consideration.

Now let M_t be another stationary, ergodic first-order Markov chain on $\mathscr{A}$. Instead of the transition matrix $\mathbf{A}$, we let M_t be defined by its *bigram probability matrix* $\mathbf{B}$ with elements

$$(\mathbf{b})_{ij}=P(M_t=i\wedge M_{t+1}=j). \tag{9}$$

The stationary distribution vector π is again required to have all positive elements and satisfies $\pi=\mathbf{B}\mathbf{1}=\mathbf{B}^T\mathbf{1}$ due to stationarity, where $\mathbf{1}$ is a column vector of all ones. It is possible to transform this back to a regular transition-matrix representation using the relation

$$\mathbf{B}=(\operatorname{diag}\pi)\,\mathbf{A}. \tag{10}$$

With the above definitions, the minimum-rate simplification M_t of $\widetilde{M}_t$ for a given Kullback-Leibler distortion D is the solution to the optimization problem

$$\min_{\mathbf{B}\in[0,1]^{N\times N}}-\sum_{i,j}(\mathbf{b})_{ij}\log\frac{(\mathbf{b})_{ij}}{\sum_{j'}(\mathbf{b})_{ij'}} \tag{11}$$

subject to

$$\sum_{i,j}(\mathbf{b})_{ij}\log\frac{(\mathbf{b})_{ij}}{(\widetilde{\mathbf{a}})_{ij}\sum_{j'}(\mathbf{b})_{ij'}}\le D \tag{12}$$

$$\left(\mathbf{B}-\mathbf{B}^T\right)\mathbf{1}=\mathbf{0} \tag{13}$$

$$\mathbf{1}^T\mathbf{B}\mathbf{1}=1 \tag{14}$$

$$\mathbf{B}\ge\mathbf{0}. \tag{15}$$

Equations (11) and (12) are Markov chain versions of (6) and (7), respectively, derived using the formulas in [28]. The two final constraints are required for $\mathbf{B}$ to describe a proper probability distribution, while the relation (13) between the marginals of $\mathbf{B}$ is necessary to obtain a stationary process.

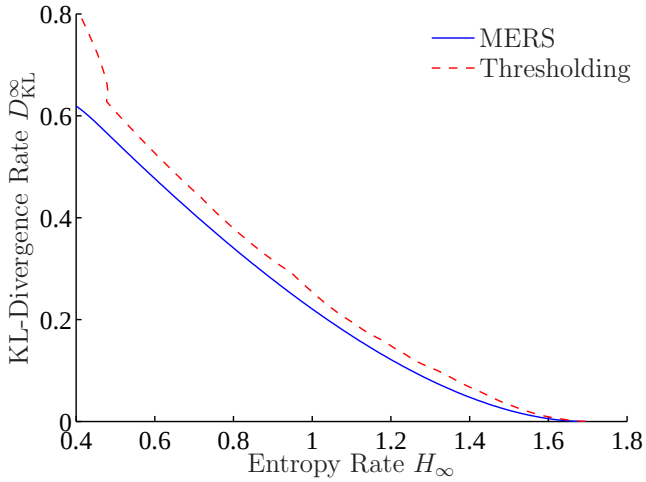

Figure 1: Simplicity-Divergence Curve

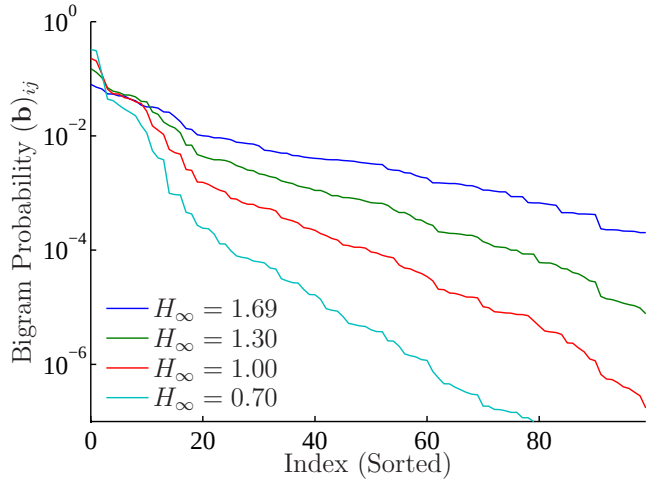

Figure 2: Rate Dependence of Matrix Elements

4.2 Iterative Solution

The MERS problem above is nonconvex and typically demanding to solve using brute force numerical minimization techniques. However, it is possible to derive fast iterative solution procedures, analogous to the Blahut-Arimoto [29, 30] for calculating points on the rate-distortion curve in rate-distortion theory. We introduce an auxiliary variable $\mathbf{q} = \mathbf{B}\mathbf{1}$ and interlace iterative optimization of $\mathbf{B}$ and $\mathbf{q}$. Given a parameter $\alpha > 1$, a counter $m = 0$, and an initial guess $\mathbf{B}^{(0)}$ satisfying the above constraints for $\mathbf{B}$, this yields an algorithm:

1. Given $\mathbf{B}$, optimize for $\mathbf{q}$: $\mathbf{q}^{(m)} = \mathbf{B}^{(m)}\mathbf{1}$.
2. Given $\mathbf{q}$, optimize for $\mathbf{B}$:
   (a) Define $\mathbf{B}'^{(m+1)}$ through $(\mathbf{b}')_{ij}^{(m+1)} = (\widetilde{\mathbf{a}})_{ij}^{\alpha}(\mathbf{q})_i^{(m)}$.
   (b) Let $n = 0$ and $\mu^{(n)} = \mathbf{1}$.
   (c) Let $(\mu)_i^{(n+1)} = \sqrt{\frac{\sum_{j=1,\, j\neq i}^{N} (\mu)_j^{(n)} (\mathbf{b}')_{ji}^{(m+1)}}{\sum_{j=1,\, j\neq i}^{N} \left((\mu)_j^{(n)}\right)^{-1} (\mathbf{b}')_{ij}^{(m+1)}}}$.
   (d) Let $n = n+1$ and repeat from 2c until convergence.
   (e) Form $\mathbf{B}''^{(m+1)} = \left(\operatorname{diag}\mu^{(n)}\right)^{-1} \mathbf{B}'^{(m+1)} \left(\operatorname{diag}\mu^{(n)}\right)$.
   (f) Normalize to get $\mathbf{B}^{(m+1)} = \left(\mathbf{1}^T \mathbf{B}''^{(m+1)} \mathbf{1}\right)^{-1} \mathbf{B}''^{(m+1)}$.
3. If not converged, let $m = m+1$ and repeat from 1.

This algorithm can be derived by introducing a Lagrange multiplier for the divergence constraint (12), and then splitting the problem into a minimization problem over two sets of variables, $\mathbf{B}$ and $\mathbf{q}$ (representing π) using the same trick as for the Blahut-Arimoto algorithm. Minimizing over one parameter set is straightforward if the other set is fixed, leading to an alternating minimization scheme as above. The inner loop at 2c provides an iterative solution to an equation of the form

$$(\operatorname{diag}\mu)^{-1}\mathbf{B}'(\operatorname{diag}\mu)\mathbf{1} = (\operatorname{diag}\mu)\mathbf{B}'^{T}(\operatorname{diag}\mu)^{-1}\mathbf{1}, \qquad (16)$$

which is necessary to find Lagrange multipliers μ so that constraint (13) is satisfied. While we have no formal convergence guarantees, the algorithm converges quickly in practice. Because of nonconvexity, the obtained solutions are not necessarily globally optimal.

We note that only $\mathbf{B}$-matrix entries where the corresponding value in $\widetilde{\mathbf{A}}$ is nonzero need to be considered for the computations; all other entries are zero at the optimum. The parameter α in the algorithm adjusts the trade-off between simplicity and divergence. To achieve a target entropy or divergence rate, it may be necessary solve the problem for several different α, and use a root-finding procedure to converge on the appropriate value. This is a common trait of the Blahut-Arimoto algorithm and its descendants such as the bottleneck equations in the information bottleneck method [22].

4.3 Numerical Example

To illustrate the behaviour of MERS in practice we present a brief numerical example from the domain of music. A process $\widetilde{M}_t$ was created by first extracting the pitch sequences from the Bach chorale data described in [31]. Taking differences between consecutive elements in each chorale, sequences of pitch changes were obtained; these were wrapped onto a single octave using modulo 12. An eleven-state Markov chain $\widetilde{M}_t$ was then fitted to the resulting data using maximum likelihood. (Only eleven states were necessary since there were no instances of six-semitone pitch increases.) This music model was then simplified using MERS for a number of different α-values, yielding the results shown.

Figure 1 graphs the high and medium rate sections of the simplicity-divergence curve, the MERS analogue of the rate-distortion curve, for this example. Simplicity and divergence are here defined by equations (11) and (12), respectively. (The low entropy region of the curve is omitted since the calculation of μ is slow there.) The curve is smooth and reflects the law of diminishing returns: near full rate, we can take away some variation with little effect on divergence, but as entropy rate is decreased further the removed bits (or fractions thereof) carry progressively greater importance.

For comparison, the figure also includes the performance of a simplistic thresholding scheme, where all the probability mass in $\widetilde{\mathbf{A}}$ below a certain threshold is removed. The threshold is different for every row, such that that an equal mass p is removed from each row; the matrix is then renormalized by dividing by $1-p$. Evidently, this reverse water-filling-like scheme achieves inferior simplicity-divergence trade-offs as the parameter p is varied, compared to the MERS curve.

Figure 2 illustrates progressively increasing sparsity as the rate is decreased. The lines correspond to the entries of the matrix $\mathbf{B}$ ordered by decreasing magnitude for a number of different entropy rates. As $H_\infty(M_t)$ goes down, only a few bigrams have increased probability, whereas most matrix entries (typically corresponding to less common two-note sequences) approach zero at an increasing pace and rapidly become insignificant. This sparse $\mathbf{B}$-matrix representation corresponds to sparsity and simplicity in outcome space for M_t, as desired; at low rate, only relatively few, highly typical sequences tend to be observed in practice.

5. DISCUSSION AND CONCLUSIONS

We have presented MERS, a rate-distortion based framework for simplifying stationary, ergodic stochastic processes, including the special case of Markov chains. The framework revolves around minimizing entropy rate under a Kullback-Leibler divergence constraint designed to promote sparsity in outcome space, so that only a few outcomes have any appreciable probability. Numerical experiments confirm this behaviour.

MERS simplifications are similar to reverse water-filling in source coding in the sense that less prominent aspects of the original distribution are filtered out. This is particularly appropriate for improving models for generative tasks, and may recover an approximation of the underlying sparsity structure from models disturbed by imperfections and occasional erratic outcomes.

A distinguishing advantage of MERS is that the information-theoretic nature of the framework ensures wholly parameterization independent results. This contrasts with many typical approaches to sparsity such as thresholding schemes or Lasso-influenced methods, e.g., [17], that rely on constraining or minimizing the ℓ_1-norm of the model parameters.

We see room for future work in both theory and applications. On the theory side, we intend to explore fundamental aspects of the proposed framework in greater depth, such as the properties of the MERS rate-distortion function and the emergence of sparsity. For applications, we are pursuing analytic solutions to the MERS problem for some important special cases. This should open the door to apply MERS to potentially large problems in a number of different contexts.

In the case of more general processes such as hidden Markov models, merely computing entropy rates can be a difficult problem [27], and analytic solutions may not be possible. It would be interesting to consider minimizing suitable upper bounds on the entropy and divergence rates, which could produce approximate rate-distortion simplifications while being computationally feasible.

REFERENCES

[1] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, Feb. 1989.

[2] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Department, 3rd edition, Dec. 2006.

[3] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda. The HMM-based speech synthesis system version 2.0. In *Proc. ISCA SSW6*, pages 294–299, Bonn, Germany, Aug. 2007.

[4] J. Dines, J. Yamagishi, and S. King. Measuring the gap between HMM-based ASR and TTS. In *Proc. Interspeech*, pages 1391–1394, Brighton, UK, Sept. 2009.

[5] A. Nádas, D. Nahamoo, and M. A. Picheny. On a model-robust training method for speech recognition. *IEEE Trans. Acoust. Speech Signal Process.*, 36(9):1432–1436, Sept. 1988.

[6] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Trans. ASSP*, pages 400–401, 1987.

[7] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. ACL 1996*, volume 34, pages 310–318, Santa Cruz, CA, USA, June 1996.

[8] L. K. Saul and F. C. N. Pereira. Aggregate and mixed-order Markov models for statistical language processing. In *Proc. EMNLP-2*, volume 2, pages 81–89, Providence, RI, USA, Aug. 1997.

[9] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41(1):164–171, 1970.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B*, 39(1):1–38, 1977.

[11] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-5(2):179–190, Mar. 1983.

[12] M. E. Brand. An entropic estimator for structure discovery. In *Proc. NIPS 1998*, pages 723–729, Denver, CO, USA, Dec. 1998.

[13] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B*, 58:267–288, 1996.

[14] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407–499, Apr. 2004.

[15] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273 –297, Sept. 1995.

[16] S. Mallat. *A Wavelet Tour of Signal Processing, the Sparse Way*. Academic Press, 3rd edition, Dec. 2008.

[17] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, Mar. 2008.

[18] S. Arora, E. Hazan, and S. Kale. A fast random sampling algorithm for sparsifying matrices. In *Proc. RANDOM*, pages 272–279, Barcelona, Spain, Aug. 2006.

[19] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, Nov. 2004.

[20] R. A. McDonald and P. M. Schultheiss. Information rates of Gaussian signals under criteria constraining the error spectrum. *Proc. IEEE*, 52:415–416, 1964.

[21] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 2nd edition, 1991.

[22] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. Allerton 1999*, pages 368–377, Allerton House, IL, USA, Sept. 1999.

[23] Y. Wang, G. Haffari, S. Wang, and G. Mori. A rate distortion approach for semi-supervised conditional random fields. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 2008–2016, Vancouver, Canada, Dec. 2009.

[24] K. Marton and P. C. Shields. The positive-divergence and blowing-up properties. *Isr. J. Math*, 86(1):331 –348, Oct. 1994.

[25] H. T. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th UAI*, volume 15, pages 21–30, Stockholm, Sweden, July 1999.

[26] T. Berger. *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[27] G. Seroussi, P. Jacquet, and W. Szpankowski. On the entropy of a hidden Markov process. In *Proc. DCC*, pages 362–371, Snowbird, UT, USA, Mar. 2004.

[28] Z. Rached, F. Alajaji, and L. L. Campbell. The Kullback-Leibler divergence rate between Markov sources. *IEEE Trans. Inf. Theory*, 50(5):917–921, May 2004.

[29] R. E. Blahut. Computation of channel capacity and rate distortion function. *IEEE Trans. Inf. Theory*, IT-18:460–473, 1972.

[30] S. Arimoto. An algorithm for calculating the capacity of an arbitrary discrete memoryless channel. *IEEE Trans. Inf. Theory*, IT-18:14–20, 1972.

[31] D. Conklin and I. H. Witten. Multiple viewpoint systems for music prediction. *J. New Music Res.*, 24(1):51–73, 1995.