Minimum Entropy Rate Simplification of Stochastic Processes

Gustav Eje Henter, *Member, IEEE*, and W. Bastiaan Kleijn, *Fellow, IEEE*

Abstract—We propose *minimum entropy rate simplification* (MERS), an information-theoretic, parameterization-independent framework for simplifying generative models of stochastic processes. Applications include improving model quality for sampling tasks by concentrating the probability mass on the most characteristic and accurately described behaviors while de-emphasizing the tails, and obtaining clean models from corrupted data (nonparametric denoising). This is the opposite of the smoothing step commonly applied to classification models. Drawing on rate-distortion theory, MERS seeks the minimum entropy-rate process under a constraint on the dissimilarity between the original and simplified processes. We particularly investigate the Kullback-Leibler divergence rate as a dissimilarity measure, where, compatible with our assumption that the starting model is disturbed or inaccurate, the simplification rather than the starting model is used for the reference distribution of the divergence. This leads to analytic solutions for stationary and ergodic Gaussian processes and Markov chains. The same formulas are also valid for maximum-entropy smoothing under the same divergence constraint. In experiments, MERS successfully simplifies and denoises models from audio, text, speech, and meteorology.

Index Terms—G.3.e Markov processes, G.3.p Stochastic processes, H.1.1.b Information theory, H.5.5.c Signal analysis, synthesis, and processing, I.2.7.b Language generation, I.5.1.e Statistical models.

1 Introduction

Real-world observations are frequently corrupted by noise and errors, obscuring simpler processes lying underneath. Examples include field recordings of songbirds or speech recordings in natural environments, where the sources of interference cannot be controlled during data collection. Generative models trained on disturbed data result in complex descriptions that attempt to replicate the errors. Sampling from these models thus produces noisy data. We consider the problem of simplifying these models, so that cleaner and more consistent synthetic output—whether birdsong, speech, or something else—can be produced.

Because the observations are from a disturbed process, different from the actual process to be modeled, the models do not converge on the desired process even in the limit of infinite samples. One response would be to define and train a model with an explicit noise term, but subsequently set this noise to zero when generating new data. This works well if the type of noise is known and easy to describe mathematically, and we are free to choose a model that separates signal and noise. That is not always the case. Given a simple Markov chain, for instance, it is not obvious how to incorporate disturbances without changing the nature of the model significantly.

In this work, we consider the case of model denoising and simplification when no explicit noise model is available, or when the model does not separate signal and noise well. We present a nonparametric framework for simplifying generative models of stochastic processes, based on information theory. The framework can be used for removing disturbances from models of stochastic processes without assumptions about the nature of the errors present, other than that they act to increase the entropy of the data. The need for explicit noise models, as in, e.g., [1], is thus avoided. Neither is there a need to specify a prior, as in Bayesian probability.

Our proposal is also useful for simplifying all-signal models trained on noise-free data, by concentrating on the most characteristic outcomes. Such simplification is relevant in model-based speech synthesis, where random sampling generates unnatural speech due to shortcomings of the acoustic models used [2], and only the most probable outcome is typically generated as output data [3]. For our proposal, the degree of simplification can be adjusted continuously, and unlike ℓ_1-sparsity based model simplification schemes like [4], the results are independent of unitary transformations of the data.

Our simplification procedure acts as a post-processing step applied to already-trained models. As a result, the simplified model is not automatically validated against observed data. On the other hand, this means that the procedure can be applied even when only a model is provided, without any data, e.g., in online learning scenarios where datapoints are not retained. Moreover, the computational effort scales favorably, as it only depends on model size, regardless of the amount of training material used to create that model.

- *G. E. Henter is with the Centre for Speech Technology Research at the University of Edinburgh, United Kingdom. A major portion of this research took place while he was with the Communication Theory laboratory, School of Electrical Engineering at KTH Royal Institute of Technology, Stockholm, Sweden. E-mail: gustav.henter@ee.kth.se.*
- *W. B. Kleijn is with the Communications and Signal Processing Group at Victoria University of Wellington, New Zealand, and the Multimedia Computing Group at TU Delft, The Netherlands.*
  E-mail: bastiaan.kleijn@ecs.vuw.ac.nz.
- *This research was supported by the LISTA (Listening Talker) project. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230.*

Manuscript received Nov. 26 2013; revised Mar. 17 2015; accepted Jan. 23 2016.

The framework leads to straightforward, analytical simplification schemes in a number of important special cases. We show that for an example with a Markov chain grammar learned from synthetic speech transcriptions corrupted by realistic speaker errors, the proposed method is able to remove a significant part of the disturbances in the learned model, as measured objectively by the KL-divergence rate. The simplified model generates sample data of superior subjective quality compared to the original learned model.

The remainder of the article is laid out as follows: Section 2 provides a motivation for probability concentration in data synthesis. Section 3 then introduces the general MERS framework. The sections thereafter provide analytic solutions for two important practical cases. Section 6 presents some experimental results on Markov chains, while Section 7 concludes and suggests further work.

2 Background

In machine learning, a distinction exists between *generative statistical models*, which describe the joint distribution $P(\boldsymbol{X}, Y)$ of labels Y and observations $\boldsymbol{X}$, and *discriminative models* that model the conditional distribution $P(Y \mid \boldsymbol{X})$ only. The former models are more versatile, since they can be used both for classification tasks (estimating Y given $\boldsymbol{X}$), and for sampling from the joint distribution. As a case in point, hidden Markov models lie at the heart of contemporary systems for both speech recognition and model-based speech synthesis, e.g., [5], [6].

One size does not fit all, however. While the same model family may be successful in both generative and discriminative tasks, this does not imply that the same *model* is optimal in all cases; see [7] and [8]. In practice, it is generally necessary to adapt the approach to the context and problem at hand. Specifically, classification and synthesis tasks are often treated differently [7]. For example, training (parameter estimation) for classification problems may employ a discriminative procedure, to emphasize class differences over regular maximum likelihood [9].

2.1 Task-Appropriate Post-Processing

Differences between generative and discriminative tasks go beyond model formulation and training, as it is commonly necessary to post-process the trained models to account for aspects of the problem that may not be represented well by the training data. However, while such post-processing is recognized as a common part of classifier design and has been well studied in that context, the possibility of post-processing for generative tasks is largely unexplored. This article is an initial attempt to fill that gap, and consider generative post-processing in more detail.

For discriminative tasks such as speech recognition, it is common to apply *smoothing* following maximum likelihood parameter estimation [10]. Smoothing increases the amount of randomness and variation in the model, and reduces the impact of the greater variability of real-world data as compared to training data. The practice generally improves the performance of recognizers, which may be trained on clean and grammatically correct speech, but are commonly used in environments with background noise, conversational grammar, and a wide variety of speaker accents.

Mathematically, the issue is that maximum likelihood parameter estimation tends to produce models that assign minimal, often zero, probability to events not observed in the training data. However, as the actual set of possible outcomes may be very large, it is not uncommon for new samples to represent events not previously observed, see [11]. In speech recognition, a small acoustic aberration may then prevent the recognition of a word. Smoothers and Bayesian approaches replace many of the probabilities estimated to be zero with small but nonzero values, thus increasing model variability and improving practical performance. Example techniques include simple additive smoothers such as pseudocount methods, which are related to Bayesian priors, and well-known schemes like those of Jelinek and Mercer [12], Katz [13], and Kneser and Ney [14].

Generative tasks call for the opposite approach. When sampling from a model, it is often preferable to decrease rather than increase its variability. This filters out unlikely and uncharacteristic behavior. For example, it is desirable for a speech synthesizer to use correct grammar, even if the training data is not grammatically perfect. (Compare with human children, who are able to learn excellent grammar from conversational speech alone.) It is thus preferable to focus on the most common and characteristic behaviors, at the expense of less common events. We refer to this as *probability concentration*, as the aim is to concentrate the probability mass or density of a model to select representative outcomes. A simpler, more predictable process is then obtained. The degree of probability concentration that is desired will depend on the particular application. Model-based speech synthesis is an extreme example, where only the most probable outcome is generated [3]. This is the opposite of smoothing for discriminative tasks, where peaked probability distributions are made more uniform, increasing variability.

An alternative view of probability concentration is that we want to reduce or de-emphasize the tails of the process. These may not be well behaved since they can be difficult to estimate from empirical data, and need not have the assumed functional form (recall that the central limit theorem does not apply to extreme values). For the special case of MERS investigated in Sections 4.1 and 5, we obtain a scheme where tails of the distribution function that originally roll off as a power function, $\mathcal{O}(x^{-p})$, after simplification decrease by a greater power $\mathcal{O}(x^{-\alpha p})$, where $\alpha > 1$. Exponentially decreasing tails similarly have their roll-off rate increased by a factor α. Thus, fat tails are made slimmer. We can also perform maximum entropy rate smoothing using the same mathematical solutions, by choosing $\alpha \in (0, 1)$.

2.2 Relations to Sparsity and Denoising

The examples in the introduction assume that errors are inherent in the data acquisition process. This means one cannot rely on just amassing more data in order to converge on a good, low-noise model. Basic Bayesian smoothers like [15] are similarly ineffective, as the impact of the prior there decreases with additional data. Instead, we are compelled to assume the existence of a simpler (less random) underlying structure, and then recover this structure by applying some kind of nonparametric *model denoising*, removing errors from the model rather than from the data. Since errors typically are varied and spread their probability mass over many outcomes, removing uncommon events will concentrate on the least corrupted behaviors of a process. However, the general idea of simplification by decreasing variability is valid even without positing an underlying structure to recover, and applies also outside the domain of denoising.

Probability concentration is related to sparsity, but it is not identical to it. Sparsity is traditionally considered in relation to some basis: a representation is considered sparse if many coefficients are very small or zero, see [16]. One example is speech signals, which become sparse with a unitary transformation to the frequency domain. A major aspect of sparse methods is typically to find the right such sparsifying transformation, for example the Karhunen-Loève transform used in signal compression. Probability concentration, on the other hand, is a property of the model itself rather than a particular representation. Our proposal, in particular, is independent of translations and unitary transformations (or, for discrete-valued variables, any transformation), and thus avoids any need to search for a sparsifying transformation.

The concepts of sparsity and probability concentration partially overlap in models such as Markov chains, which are typically parameterized in terms of probabilities; a Markov chain exhibiting a high degree of probability concentration will have mostly negligible entries in its transition matrix (though the entries need not be identically zero). This connection to sparsity is appealing, since sparse representations tend to compress well [16], may allow fast processing [17], and typically are easier to interpret [15], [18]. Although simple models cannot be motivated from a statistical argument [19], these advantages of sparsity are consistent with Occam's razor.

3 Minimum Entropy Rate Simplification

In this part, we demonstrate how the abstract principle of probability concentration for stochastic processes can be translated into a concrete mathematical framework. We call this *minimum entropy rate simplification*, MERS. We define simplification as a decrease in some quantitative measure of complexity, in our case the entropy rate.

To measure and obtain practical probability concentration, we adopt an approach similar to the well-established rate-distortion framework in lossy source coding. Rate-distortion theory is a compelling starting point for several reasons:

1) The goal is to produce simplified approximations, in the sense that they compress easily and are simple to describe.
2) The results are independent of parameterization, owing to their grounding in information theory.
3) There are well-known solutions exhibiting reverse water-filling, a manifestation of sparsity [20], [21]. We would like to achieve something similar.
4) It has already spawned machine learning spin-offs such as information bottleneck [22] and the semi-supervised method for learning conditional random fields described in [23].
5) By basing our efforts on established theory, we can adapt and reuse its associated tools and techniques.

The central pillar of rate-distortion theory is the trade-off between the chosen degree of simplification (rate decrease) and the distortion it necessarily introduces. We adopt a similar setup, and seek models that are optimally simple in a specific mathematical sense, while not diverging too much from the original model.

3.1 Preliminary Definitions

Let $\widetilde{X} = \{\widetilde{X}_t : t \in \mathbb{Z}\}$ be a given stationary and ergodic stochastic process representing some observed process. Importantly, $\widetilde{X}$ is *not* a set of observations, but a stochastic model that generates them. We shall assume $\widetilde{X}$ to be known, though in practice it generally has to be estimated from observation data first. $\widetilde{X}$ may be either discrete or continuous-valued. We use an underline and indices together to denote contiguous sequences of random variables from a stochastic process, as in $\underline{\widetilde{X}}_t^{t+T} = \{\widetilde{X}_t, \widetilde{X}_{t+1}, \ldots, \widetilde{X}_{t+T}\}$.

Let $\mathcal{X}$ be a given class of stochastic processes on the same sample space Ω as $\widetilde{X}$; typically $\widetilde{X} \in \mathcal{X}$. Probability concentration leads to another stationary and ergodic $X \in \mathcal{X}$ that is similar to the given $\widetilde{X}$, but emphasizes characteristic behavior and suppresses uncommon events. MERS, in particular, maximizes a particular simplicity measure for X, subject to a constraint on the dissimilarity from $\widetilde{X}$, akin to the rate-distortion trade-off in lossy source coding.

In some contexts, we may assume the existence of $X^{\star}$, a clean, stationary, and ergodic underlying stochastic process, for instance a grammar, which is disturbed by an unknown error mechanism to form $\widetilde{X}$, the model process that generates our observations. $X^{\star}$ takes values on the same space as $\widetilde{X}$. The X obtained with MERS may be seen a "cleaned" version of $\widetilde{X}$ and an approximation of $X^{\star}$. Note that we have not made assumptions on the nature of the disturbances, e.g., independence or Markovianity, so the setting is highly general; disturbances can, for example, be omissions, repetitions, as well as noise additions.

3.2 Quantifying Simplicity

The first design choice is how to quantify simplicity. Our choice should capture the degree of probability concentration, and thus the amount of randomness.

The classic measure for quantifying the randomness of a discrete random variable P with pmf $p_P(i)$ is the *information entropy* or *Shannon entropy* [21]

$$H(P) = -\sum_i p_P(i) \log p_P(i) \geq 0. \tag{1}$$

The entropy concept can be generalized to stationary and ergodic discrete-time processes by taking the limit

$$H_\infty(X) = \lim_{T\to\infty} \frac{1}{T} H\left(\left\{\underline{X}_{t+1}^{t+T}\right\}\right), \tag{2}$$

known as the *entropy rate*. This quantifies the unpredictability of the process, measured as the added information (bits, nats, or similar) per time step, and is independent of representation. If the observation space is continuous, one can instead define the *differential entropy* through an integral

$$h(P) = -\int f_P(i) \log f_P(i) \,\mathrm{d}i, \tag{3}$$

with an associated limiting *differential entropy rate* $h_\infty(X)$ for continuous-valued processes. These two quantities are independent of how the model X is parameterized, but not of transformations $y = g(x)$ of the observation space when g is not an isometry (a translation and a unitary linear transformation).

In rate-distortion theory, the entropy rate captures how difficult a signal is to compress: simple signals have concise, efficient descriptions. The entropy rate will fill an analogous role in MERS.

To obtain probability concentration we minimize the entropy rate of X. This should give a more predictable (thus more concentrated) and simpler process. By working with the entropy rate rather than the per-sample entropy $H(X_t)$, we operate our simplification on the space of entire *behaviors* of X, typically involving multiple time steps, as opposed to singular outcomes X_t.

The process X is not observable. In denoising applications we sometimes think of it as a hypothetical underlying, low-noise process. As a matter of fact, the information-theoretic properties of an underlying generating process have been used to quantify process complexity before, for instance in the causal states framework [24].[1]

We note that the entropy of a discrete random variable is a concave function over the unit simplex with minima at the corners. Entropy minimization is therefore not a convex optimization problem, and the concave nature of the objective function could possibly complicate numerical optimization by presenting many local minima. This point is however moot whenever the optimum can be identified analytically, as in the examples later on.

3.3 Preventing Oversimplification

To prevent oversimplification, we maximize the above simplicity under a constraint that we do not stray too far from the original observed process. The latter corresponds to the distortion constraint in rate-distortion theory. Numerous measures of similarity or dissimilarity between distributions exist in the literature (e.g., f-divergences and Bregman divergences [25]), many of which can be extended to stochastic processes in multiple ways. Constraining different measures will lead to different results, and yield different flavors of probability concentration and minimum entropy rate simplification.

In this work, we again look to information theory, in order to define a natural, representation-independent measure of dissimilarity that we may constrain. There, the dissimilarity between two discrete random variables P and Q is commonly quantified by the *relative entropy* (or *Kullback-Leibler divergence*) [21]

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_i p_P(i) \log \frac{p_P(i)}{p_Q(i)} \geq 0. \tag{4}$$

As before, we can take the limit

$$D_\infty(X \,||\, Y) = \lim_{T\to\infty} \frac{1}{T} D_{\mathrm{KL}}\left(\left\{X_{t+1}^{t+T}\right\} \,\middle|\middle|\, \left\{Y_{t+1}^{t+T}\right\}\right) \tag{5}$$

to define the *relative entropy rate* between two stationary stochastic processes X and Y. Like the entropy rate $H_\infty(X)$, this is independent of representation and has compatible units of information per time step. For continuous-valued processes, an analogous concept of differential relative entropy rate, d_∞, is obtained by integrating rather than summing over the observation space. Like the discrete quantity D_∞, this is invariant of parameterization as well as transformation.

In standard situations where the KL-divergence is used, P is the true distribution while Q is an approximation thereof. In MERS, we want to identify a simple candidate underlying model X from a given corrupted, approximate version $\widetilde{X}$. Since $\widetilde{X}$ is the approximate quantity, this suggests a constraint $D_\infty(X \,||\, \widetilde{X}) \leq D$, where D is a user-set maximum tolerable divergence rate which controls the degree of simplification.

Due to the asymmetry between the two arguments of the KL-divergence, this formulation is highly averse to adding behaviors (nonzero-probability outcome sequences) to X that are not in $\widetilde{X}$, but is less sensitive to outcomes being taken away, as is appropriate for simplification. This is in contrast to, e.g., the alternative constraint $D_\infty(\widetilde{X} \,||\, X) \leq D$, which may penalize probability concentration too harshly.

A similar situation to the above, where the second divergence argument and not the first is considered fixed, arises for log-evidence maximization in variational Bayesian methods, cf. [26], also leading to concentrated distributions.

3.4 The General MERS Formulation

Drawing on the above sections, we are now ready to define general minimum entropy rate simplification. Given a stationary and ergodic stochastic process $\widetilde{X}$ and a process dissimilarity measure $\mathrm{Dis}(X, \widetilde{X})$, a general minimum entropy rate simplification of $\widetilde{X}$ in $\mathcal{X}$ for a maximum tolerable dissimilarity D is any process X which solves the optimization problem

$$\min_{X\in\mathcal{X}} H_\infty(X) \tag{6}$$

$$\text{subject to } \mathrm{Dis}(X, \widetilde{X}) \leq D. \tag{7}$$

In this paper, the primary dissimilarity $\mathrm{Dis}(X, \widetilde{X})$ will be the KL-divergence rate $D_\infty(X \,||\, \widetilde{X})$, leading to a formulation

$$\min_{X\in\mathcal{X}} H_\infty(X) \tag{8}$$

$$\text{subject to } D_\infty(X \,||\, \widetilde{X}) \leq D, \tag{9}$$

with D being the maximum tolerable divergence. In case $\widetilde{X}$ is continuous-valued, we replace H_∞ and D_∞ by their differential analogues.

We assume the relevant H_∞ and D_∞ exist; for Markovian processes this is assured [27]. However, the quantities may be difficult to write out explicitly. For instance, the entropy rate of an HMM is a Lyapunov exponent with no known closed-form expression [28]. This echoes rate-distortion theory, where only a few analytic solutions are known [21].

The MERS framework enables a continuum of simplifications, ranging all the way from no modification to

[1] Causal states are however unsuitable for our purposes, as they capture all predictive information in the original process, including any correlations in the noise, and do not actually discard information as would be necessary for strong simplification.

complete predictability, in order to suit a wide array of application scenarios. In practice, different applications call for different trade-offs between simplicity and fidelity, and it is therefore difficult to make general statements on what tuning-parameter values to use. Ultimately, the degree of simplification has to be chosen by the experimenter on a case-by-case basis, for instance via cross-validation if held-out data is available.

4 MERS for Gaussian Processes

Having defined the general MERS framework, we now focus on its concrete implications in a few important special cases. We first discuss the situation where $\widetilde{X}$ is a continuous-valued Gaussian process, where we can draw some parallels to Wiener filtering; discrete Markov chains will be considered in Section 5.

4.1 Purely Nondeterministic Processes

We here present solutions to the MERS problem (8) for two classes of Gaussian processes; derivations are provided in Appendix A. To begin with, let $\mathcal{X}_{\mathrm{nd}}$ be the space of purely nondeterministic stationary and ergodic univariate Gaussian processes, with $X,\ \widetilde{X} \in \mathcal{X}_{\mathrm{nd}}$. Defining the (power) spectral density function of nondeterministic X through

$$R_X\left(e^{i\omega}\right) = \left|\sum_{l=-\infty}^{\infty} \mathbb{E}\left(X_t X_{t+l}\right) e^{i\omega l}\right| \tag{10}$$

for $\omega \in (-\pi,\ \pi]$, with $R_{\widetilde{X}}\left(e^{i\omega}\right)$ defined similarly, the differential entropy rate to minimize becomes

$$h_\infty\left(X\right) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(4\pi^2 e^2 R_X\left(e^{i\omega}\right)\right)\,\mathrm{d}\omega, \tag{11}$$

while the relative entropy rate constraint turns into the Itakura-Saito divergence [29]

$$d_\infty(X \mid\mid \widetilde{X}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)} - 1 - \log\left(\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)}\right)\right)\mathrm{d}\omega, \tag{12}$$

see [30]. We may assume $R_{\widetilde{X}}\left(e^{i\omega}\right) > 0$ almost everywhere, otherwise the process is completely predictable. (Following the same reasoning, (11) means that band-limited processes do not have a meaningful differential entropy rate.)

For generality, we will also consider MERS solutions $X \in \mathcal{X}$ with lower-bounded power spectra, i.e., we let

$$\mathcal{X} = \left\{X \in \mathcal{X}_{\mathrm{nd}} : R_X\left(e^{i\omega}\right) \geq r_{\min}\,\forall\omega\right\} \tag{13}$$

for some given $r_{\min} \geq 0$. This reduces to the unconstrained processes $\mathcal{X}_{\mathrm{nd}}$ when $r_{\min} = 0$.

The problem of minimizing differential entropy rate under our constraints is easily solved through a variational calculus approach similar to the derivation of AR-processes as the maximum-entropy processes under covariance constraints in [31]. We introduce a Lagrange multiplier $\lambda \geq 0$ for the divergence constraint (12), which we can think of as an "exchange rate" between bits of entropy and bits of divergence. Upon seeking stationary points, one obtains from the Euler-Lagrange equation

$$R_X\left(e^{i\omega}\right) = \max\left(r_{\min},\ \frac{\lambda-1}{\lambda}R_{\widetilde{X}}\left(e^{i\omega}\right)\right). \tag{14}$$

This solution is obviously only valid for $\lambda > 1$; for these λ we define the (inverse) scaling factor $\alpha = \frac{\lambda}{\lambda-1} \in (1,\ \infty)$. The MERS solution simply shrinks the spectral magnitude by α^{-1}, until it hits the floor at $r_{\min}$.[2] In an MA(∞)-representation, this uniform spectral scaling corresponds to reducing the variance of the driving Gaussian noise by a factor α^{-1}, or an equivalent scaling of all MA-coefficients. The factor α can be computed as an implicit function of the maximum tolerable dissimilarity d.

It is worth noting that simple spectral scaling indeed produces probability concentration. Specifically, when the minimum-rate bound is inactive the next-step conditional pdf for X given T past samples can be written

$$f_{X_t \mid \underline{X}_{t-T}^{t-1}}(x_t \mid \underline{x}_{t-T}^{t-1}) = \frac{1}{\nu}\left(f_{\widetilde{X}_t \mid \underline{\widetilde{X}}_{t-T}^{t-1}}(x_t \mid \underline{x}_{t-T}^{t-1})\right)^{\alpha}, \tag{15}$$

where $\nu > 0$ is a normalization constant. The ratio of $f_{X_t \mid \underline{X}_{t-T}^{t-1}}$ to $f_{\widetilde{X}_t \mid \underline{\widetilde{X}}_{t-T}^{t-1}}$ is then strictly increasing in $f_{\widetilde{X}_t \mid \underline{\widetilde{X}}_{t-T}^{t-1}}$ for any x_t and $\underline{x}_{t-T}^{t-1}$, meaning that the probability density has become further concentrated to previous high-probability regions.

It is instructive to compare Equation (15) with the Gibbs measure from statistical mechanics [32], which takes the form

$$p_X\left(x\right) = \frac{1}{Z\left(\alpha\right)}\exp\left(-\alpha E\left(x\right)\right), \tag{16}$$

where $E\left(x\right)$ is known as the *energy* of state or configuration x and $Z\left(\alpha\right)$ (the *partition function*) is a normalization constant. The Gibbs measure is a well-known framework that exhibits probability concentration for high values of the inverse-temperature parameter α, in the sense that the system concentrates on the least energetic states as the temperature drops. Adding an artificial temperature parameter is in fact an established method for simplifying a Gibbs model and reducing its thermodynamic entropy; MERS provides an information-theoretic interpretation of this practice.

4.2 Weighted Itakura-Saito Divergence

The Itakura-Saito criterion in (12) is a function of the ratio $R_X\left(e^{i\omega}\right)/R_{\widetilde{X}}\left(e^{i\omega}\right)$, and thus only considers relative spectral differences. In many applications, however, spectral peaks are the most important. One example is speech signal processing, where spectral valleys often are subject to perceptual masking. This suggests weighting the Itakura-Saito divergence by the observed signal power $R_{\widetilde{X}}\left(e^{i\omega}\right)$,

$$\begin{aligned}&d_{\mathrm{IS}}^{q}(X \mid\mid \widetilde{X})\\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right) + R_{\widetilde{X}}\left(e^{i\omega}\right)\log\left(\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_X\left(e^{i\omega}\right)}\right)\right)\mathrm{d}\omega.\end{aligned} \tag{17}$$

As shown in Appendix A.2, minimizing entropy rate while constraining this weighted Itakura-Saito divergence leads to a solution

$$R_X\left(e^{i\omega}\right) = \max\left(r_{\min},\ R_{\widetilde{X}}\left(e^{i\omega}\right) - \lambda^{-1}\right), \tag{18}$$

[2] Interestingly, when $r_{\min} = 0$ the problem of *maximizing* the entropy rate under the divergence constraint (9) is mathematically very similar to MERS, and leads to a solution formula identical to Equation (15), but with exponents $\alpha \in (0,\ 1)$ rather than $\alpha > 1$.

where λ is a Lagrange multiplier associated with the dissimilarity constraint. For $r_{\min} = 0$, the solution is valid for $\lambda^{-1} \in \left(0,\, \min_\omega R_{\widetilde{X}}\left(e^{i\omega}\right)\right]$, with the rate in (11) typically approaching zero at the upper end of this interval, while for $r_{\min} > 0$ the entire range $\lambda^{-1} \in (0,\, \infty)$ can be used, though the rate then has a nonzero minimum value. Interestingly, while (14) corresponds to the same relative spectral reduction everywhere, this solution instead reduces the spectrum everywhere by the same *absolute* amount, until the floor is reached.

4.3 Conserving the Variance

Another variation on MERS is obtained by changing the constraints on the process space $\mathcal{X}$. If we revert back to standard KL-divergence rate based MERS (by constraining d_∞ as given in (12)), but choose $\mathcal{X}$ as the space of stationary and ergodic univariate Gaussian processes having the same variance $\operatorname{Var}(\widetilde{X}_t)$ as the original $\widetilde{X}$, the simplistic variance scaling in (14) is no longer possible. Instead we get a formal solution

$$R_X\left(e^{i\omega}\right) = \frac{1}{\nu} \frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{\beta - R_{\widetilde{X}}\left(e^{i\omega}\right)} \tag{19}$$

(see Appendix A.3). For Lagrange multiplier values $\nu > 0$ and $\beta > \max_\omega R_{\widetilde{X}}\left(e^{i\omega}\right)$ this can be shown to erode away already small values of $R_{\widetilde{X}}\left(e^{i\omega}\right)$, yielding a spectrum where the relative differences between peaks and valleys are increased; other ranges of Lagrange multipliers give maximum entropy rate solutions. Essentially, peaks are conserved since they dominate the energy, while valleys are removed. To achieve a target variance and distortion, one may apply root-finding schemes to solve for appropriate Lagrange multiplier values.

We note that constraining $\operatorname{Var}(X_t)$ prevents X from becoming deterministic in the limit of extreme simplification. Instead, (19) achieves low entropy rates by creating simplifications that are predictable over longer time spans on average.

4.4 General Solution for Gaussian Processes

Solutions (14) and (19) above can easily be extended to general stationary and ergodic univariate Gaussian processes. Let $\widetilde{X}' = \widetilde{\mu} + \widetilde{X}$ and $X' = \mu + X$ be the Wold decompositions [33] of two such processes, with $\widetilde{\mu}$ and μ being the deterministic process components, while X and $\widetilde{X}$ are purely nondeterministic as before. It is assumed that $X \in \mathcal{X} \Rightarrow \mu + X \in \mathcal{X}\,\forall \mu$; $\mathcal{X}$ is closed under deterministic translation. The relation

$$d_\infty(\mu + X \,||\, \widetilde{\mu} + \widetilde{X}) \geq d_\infty(\widetilde{\mu} + X \,||\, \widetilde{\mu} + \widetilde{X}) \tag{20}$$

(see Appendix A.4) then ensures that there always is a MERS optimum of the form $X' = \widetilde{\mu} + X$.

Since (20) only is satisfied with equality when $\mu = \widetilde{\mu}$, the solution is unique. Additionally, the identities $h_\infty\left(X'\right) = h_\infty\left(X\right)$ and

$$d_\infty(\widetilde{\mu} + X \,||\, \widetilde{\mu} + \widetilde{X}) = d_\infty(X \,||\, \widetilde{X}) \tag{21}$$

enable us to reduce the problem to the purely nondeterministic situation above. The general Gaussian MERS solution is therefore $X' = \widetilde{\mu} + X$, which amounts to simplifying the nondeterministic process component as before and keeping the deterministic part of $\widetilde{X}$ unaltered. This is discussed further in Appendix A.4.

4.5 Relation to the Wiener Filter

MERS makes no explicit assumptions about possible disturbances in the input process $\widetilde{X}$. It is however instructive to compare MERS output to that of traditional noise reduction methods containing an explicit noise model, and see when the results agree. A particularly important scenario is that of additive, uncorrelated noise N_t. This satisfies

$$\widetilde{X}_t = X_t^\star + N_t \tag{22}$$

$$R_{\widetilde{X}}\left(e^{i\omega}\right) = R_{X^\star}\left(e^{i\omega}\right) + R_N\left(e^{i\omega}\right). \tag{23}$$

MSE-optimal noise reduction is then performed through Wiener filtering [34].

For the additive noise model, it is easy to see that the unconstrained KL-divergence based MERS solution in (14) can recover the correct signal spectrum, i.e., achieve $R_X\left(e^{i\omega}\right) = R_{X^\star}\left(e^{i\omega}\right)$ for an appropriate choice of α, if the spectrum of the noise is proportional to that of the underlying signal, $R_N\left(e^{i\omega}\right) \propto R_{X^\star}\left(e^{i\omega}\right)$. This reinforces our view that MERS is particularly appropriate when signal and noise are not easily separated.

On the other hand, the solution in (18), obtained by constraining the weighted Itakura-Saito divergence, can be made to coincide with the result of Wiener filtering $R_{\widetilde{X}}\left(e^{i\omega}\right)$ for the important practical case of additive white Gaussian noise. In this scenario, the noise spectral density is constant, $R_N\left(e^{i\omega}\right) = \frac{\sigma^2}{2\pi}$ assuming a noise variance σ^2, and MSE-optimal filtering is performed by choosing $r_{\min} = 0$ and $\lambda = 2\pi\sigma^{-2}$ in MERS.

For variance-constrained MERS, the MERS solution (19) cannot be matched by any traditional Wiener filter. This is because MERS increases the energy of spectral peaks to satisfy the constraint (19), while Wiener filtering can only remove energy. This exemplifies a situation where MERS should primarily be interpreted as a general simplification scheme, rather than a mere denoising procedure.

4.6 Example Application

We illustrate the effects of the various MERS solutions by applying them to an audio signal with compression artifacts (quantization noise). In audio compression, it is standard to assume that the signal waveform follows a Gaussian AR-process, so the previously developed MERS theory is appropriate.

The solid blue graph in Figure 1 shows the power spectrum of a model $X^\star$, estimated from a half-second, Hann-windowed excerpt of the grand piano recording in track 39 (index 1) of the EBU sound quality assessment material [35], downsampled to 8 kHz. The raw spectrum estimate was smoothed with a fifteen-point moving average to reduce the inherent variance of the estimation and to make the peaks wider and easier to resolve. Also shown in the figure (dotted black curve) is the power spectrum of a model $\widetilde{X}$ similarly fitted to a six-bit uniformly quantized version of the signal. As seen in the plot, the effect of this quantization

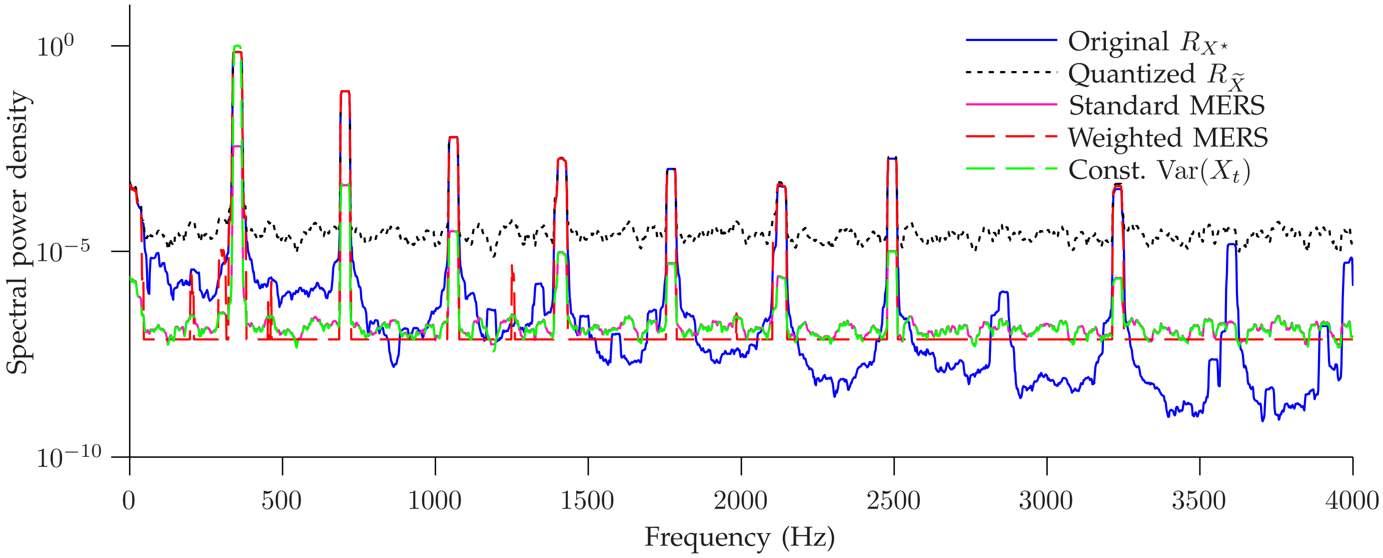


Figure 1. Power spectra of original, disturbed, and minimum entropy rate simplified Gaussian processes.

is to add white noise to the signal. This is most obvious in the spectral valleys, which have become much shallower for the quantized model, while the peaks have hardly changed at all. The figure additionally shows a number of power spectra corresponding to models simplified with the three Gaussian MERS techniques presented in this paper. For each technique, the relevant parameters ($r_{\min}$, α, λ, and/or β) were selected to minimize the root-mean-square (RMS) log-spectral distortion of the simplification X with respect to the unquantized process $X^\star$.

From the figure, we see that the general effect of MERS is to remove energy from the spectral valleys, similar to postfiltering in audio coding. Unweighted and variance-constrained MERS largely overlap (solid pink versus dashed green graphs), but in contrast to the simple spectral scaling of unweighted MERS, which just amounts to a vertical offset in the logarithmic plot, variance-constrained MERS is able to retain, and even emphasize, the most prominent peak in the spectrum. Weighted Itakura-Saito MERS (dashed red graph) goes further, and manages to conserve all spectral peaks above the quantization noise floor, while reducing the spectral valleys to a power level much closer to that of the original signal. This results in models X where the amount of noise has been reduced significantly, and which would exhibit increased subjective audio quality over $\widetilde{X}$ if used in, e.g., audio coding. Quantitatively, the log-spectrum RMS distortion of the best $R_X\left(e^{i\omega}\right)$ is less than 40% of that of the disturbed power spectrum $R_{\widetilde{X}}\left(e^{i\omega}\right)$ we started with.

5 MERS for Markov Chains

For discrete-valued processes there is no natural notion of additive noise, so there is no straightforward analogue of the Wiener filter for removing disturbances. The MERS principle, on the other hand, is equally applicable in continuous and discrete settings, and leads to similar results.

In this part, we apply MERS to discrete processes, namely first-order Markov chains on finite state spaces. Simply stated, these are characterized by the Markovian property $P\left(X_{t+1} \mid \underline{X}_{t-T}^{t}\right) = P\left(X_{t+1} \mid X_t\right)$ for all $T \geq 0$. Note that any Markov chain of finite order p can be can be converted to a first-order process through $Y_{t+1} = \left\{\underline{X}_{t+1}^{t+p}\right\}$, so it is not restrictive to assume a minimum-order process. Markov chains are very common as language models in natural language processing, in addition to being a key building block of hidden Markov models.

5.1 General Solution for Markov Chains

Let X and $\widetilde{X}$ be stationary and ergodic first-order Markov chains with outcomes on a finite-cardinality alphabet $\mathcal{A}$. Without loss of generality we take $\mathcal{A} = \{1, \ldots, k\}$.

Markov chains such as X and $\widetilde{X}$ are usually represented by their *transition matrices*, here written $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$, respectively, the elements of which are the conditional transition probabilities $a_{ij} = P\left(X_{t+1} = j \mid X_t = i\right)$, and similarly for $\widetilde{a}_{ij}$. Specifying the transition matrix completely determines any stationary and ergodic Markov chain, including its stationary distribution, which we write as a vector $\boldsymbol{\pi}$ with elements $\pi_i = P\left(X_t = i\right)$. We assume $\boldsymbol{\pi} > \mathbf{0}$, otherwise the zero-probability states can be removed from consideration and the results again apply.

For the Markov chain the entropy rate (8) minimized by MERS can be expressed as

$$H_\infty\left(X\right) = -\sum_{ij} \pi_i a_{ij} \log a_{ij}, \tag{24}$$

while the relative entropy rate constraint (9) involves

$$D_\infty(X \mid\mid \widetilde{X}) = \sum_{ij} \pi_i a_{ij} \log \frac{a_{ij}}{\widetilde{a}_{ij}}; \tag{25}$$

see [36]. Moreover, the conditions $\boldsymbol{A} \geq \mathbf{0}$ and $\boldsymbol{A}\mathbf{1} = \mathbf{1}$ must be satisfied for $\boldsymbol{A}$ to be a valid Markov chain transition matrix. This problem formulation is cumbersome to optimize since it involves $\boldsymbol{\pi}$, a normalized version of the leading eigenvector of $\boldsymbol{A}^\intercal$, which is a complicated function of the matrix elements a_{ij}.

Despite the complexities of the formulas above, it is possible to derive an analytic solution to the MERS problem for Markov chains. This involves four main steps:

1) Represent X by the *bigram probabilities* $b_{ij} = P\left(X_t = i \cap X_{t+1} = j\right)$, rather than the next-symbol probabilities a_{ij}. This simplifies the entropy and divergence rate expressions to
$$H_{\infty}\left(X\right) = -\sum_{ij} b_{ij} \log \frac{b_{ij}}{\sum_{j'} b_{ij'}} \tag{26}$$
and
$$D_{\infty}(X \mid\mid \widetilde{X}) = \sum_{ij} b_{ij} \log \frac{b_{ij}}{\widetilde{a}_{ij} \sum_{j'} b_{ij'}}, \tag{27}$$
which do not involve any problematic eigenvectors. However, $P\left(X_t = i\right) = P\left(X_{t+1} = i\right)$ (as required for stationarity) contributes an additional constraint $\boldsymbol{B}\mathbf{1} = \boldsymbol{B}^{\intercal}\mathbf{1}$ on the row and column sums of the matrix $\boldsymbol{B}$ of bigram probabilities.
2) Use the Blahut-Arimoto algorithm trick [37], [38] to transform the problem to a minimization over two sets of variables: the bigram probabilities b_{ij} and the variables q_i, the latter representing the single-symbol frequencies $\pi_i = \sum_{j'} b_{ij'}$ in the above expressions.
3) In the resulting formulation, it is possible to solve analytically for the optimal b_{ij} under fixed q_i, and vice versa. This leads to a convergent iterative solution scheme.
4) Interestingly, the fixed point of the Blahut-Arimoto iterations can be identified directly. This leads to an explicit solution formula expressed in $\boldsymbol{A}$, rather than $\boldsymbol{B}$.

A detailed derivation is provided in Appendix B. The result gives the transition matrix of the MERS-optimal X as
$$\boldsymbol{A} = \frac{1}{\nu}\left(\operatorname{diag}\boldsymbol{\mu}\right)^{-1} \widetilde{\boldsymbol{A}}^{(\alpha)} \left(\operatorname{diag}\boldsymbol{\mu}\right), \tag{28}$$
where $\widetilde{\boldsymbol{A}}^{(\alpha)}$ denotes Hadamard power α (elementwise exponentiation) of the original transition matrix $\widetilde{\boldsymbol{A}}$, i.e., $(\widetilde{\boldsymbol{A}}^{(\alpha)})_{ij} = (\widetilde{a}_{ij})^{\alpha}$, while $\boldsymbol{\mu}$ is the unique and positive leading right eigenvector of $\widetilde{\boldsymbol{A}}^{(\alpha)}$, corresponding to the eigenvalue $\nu > 0$, so that $\widetilde{\boldsymbol{A}}^{(\alpha)}\boldsymbol{\mu} = \nu\boldsymbol{\mu}$. The exponent $\alpha = \frac{\lambda}{\lambda - 1} \in (1, \infty)$ is defined in terms of the Lagrange multiplier $\lambda > 1$ for the divergence constraint, similar to the Gaussian case. Fixing α corresponds to a particular simplicity-divergence trade-off.

5.2 Solution Properties

It is easy to see that $a_{ij} = 0$ if and only if $\widetilde{a}_{ij} = 0$, so MERS neither adds nor removes transitions entirely. This ensures that X remains stationary and ergodic. However, the exponent $\alpha > 1$ has the effect of eroding the values of $\widetilde{\boldsymbol{A}}$, which is coupled with a renormalization using $\boldsymbol{\mu}$ and ν. Because the exponentiation decreases small values proportionally more than larger ones, the entropy rate decreases, and many elements may become exceedingly small.

The optimal Markov chain simplification in (28) has several similarities with the Gaussian example in Section 4.1. As before, it is clear that the interval $\alpha \in (1, \infty)$, corresponding to $\lambda > 1$, will yield all MERS solutions between the original $\widetilde{X}$ and a completely predictable process. Exponents $\alpha \in (0, 1)$, meanwhile, yield smoothed, maximum entropy rate solutions, again similar to before. Furthermore, the minimum rate processes can be expressed very similarly between the discrete and continuous cases, using exponentiated conditional pdfs or pmfs normalized by Lagrange multipliers; compare relation (15) with the Markov chain analogue
$$\begin{aligned} &p_{X_t \mid \underline{X}_{t-T}^{t-1}}(x_t \mid \underline{x}_{t-T}^{t-1}) \\ &\qquad = \frac{1}{\nu}\frac{\mu_{x_{t-1}}}{\mu_{x_t}}\left(p_{\widetilde{X}_t \mid \underline{\widetilde{X}}_{t-T}^{t-1}}(x_t \mid \underline{x}_{t-T}^{t-1})\right)^{\alpha}. \end{aligned} \tag{29}$$

The only notable difference between the continuous and discrete solution formulas, (15) and (29), is the unusual renormalization for the Markov chain contributed by the eigenvector $\boldsymbol{\mu}$, which ensures that the solution satisfies $\boldsymbol{A}\mathbf{1} = \mathbf{1}$. We observe that this normalization is similar to the detailed balance condition
$$\boldsymbol{A} = \left(\operatorname{diag}\boldsymbol{\pi}\right)^{-1}\boldsymbol{A}^{\intercal}\left(\operatorname{diag}\boldsymbol{\pi}\right) \tag{30}$$
satisfied by time-reversible Markov chains. In fact, by rewriting (28) as
$$\boldsymbol{A} = \left(\operatorname{diag}\left(\widetilde{\boldsymbol{A}}^{(\alpha)}\left(\operatorname{diag}\boldsymbol{\mu}\right)\mathbf{1}\right)\right)^{-1}\widetilde{\boldsymbol{A}}^{(\alpha)}\left(\operatorname{diag}\boldsymbol{\mu}\right), \tag{31}$$
we recognize it as an instance of the familiar row-sum normalization scheme
$$\boldsymbol{C}_{\text{norm}}\left(\boldsymbol{C}\right) = \left(\operatorname{diag}\left(\boldsymbol{C}\mathbf{1}\right)\right)^{-1}\boldsymbol{C} \tag{32}$$
used, e.g., to convert matrices with raw counts to ML estimates of conditional transition probabilities. Here $\boldsymbol{C} = \widetilde{\boldsymbol{A}}^{(\alpha)}\left(\operatorname{diag}\boldsymbol{\mu}\right)$, so the eigenvector $\boldsymbol{\mu}$ merely acts as a set of column weights for $\widetilde{\boldsymbol{A}}^{(\alpha)}$. The weights are equal and all rows of $\boldsymbol{A}$ are the same if $\widetilde{X}$ is an i.i.d. process.

The most computationally demanding aspect of the solution formula (28) is to find the leading eigenvector of $\widetilde{\boldsymbol{A}}^{(\alpha)}$, but this is a much studied problem for which efficient numerical methods exist [39]. If $\widetilde{\boldsymbol{A}}$ is sparse, as is often the case in practice, extremely large problem sizes can be handled. In computing the seminal PageRank measure of web-page importance [40], which is based on Markov chains, it is common to solve leading-eigenvector problems with billions of columns [41].

6 Markov Chain Experiments

To provide concrete examples of the behavior and performance of Markov chain MERS, we here present its application to three different models. It is interesting to compare this theoretically optimal objective performance against that of relatively straightforward simplification methods. We therefore also apply a simple, thresholding-based scheme for probability concentration to the same examples, and investigate the results both objectively and subjectively.

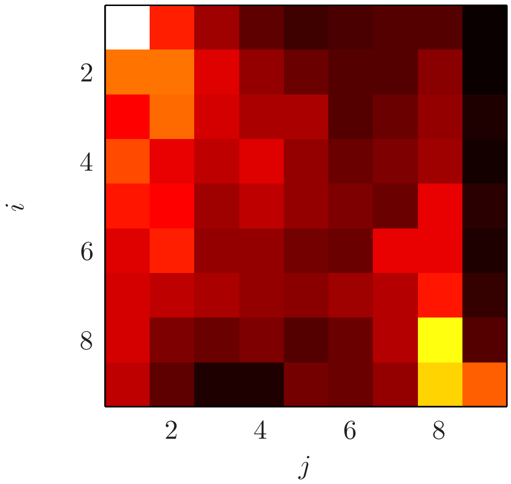


Figure 2. Heat map of cloud-coverage transition probabilities $\widetilde{a}_{ij}$.

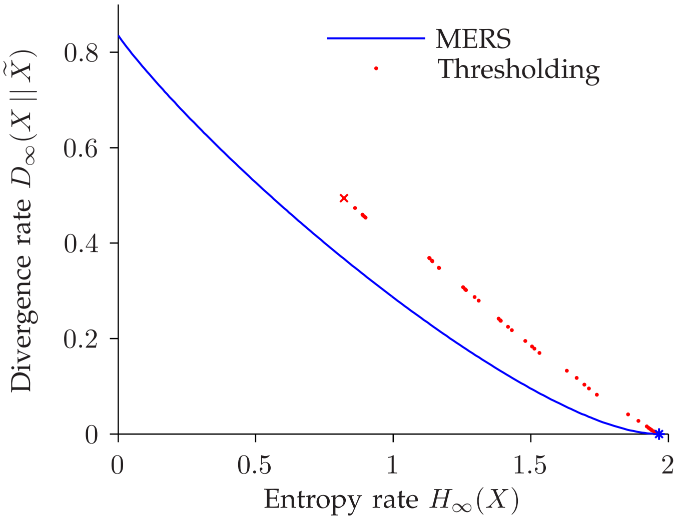


Figure 3. Entropy-divergence trade-offs for the cloud-coverage model.

In the first application, the methods are applied to weather data. Thresholding leads to nonergodic, reducible models, whereas MERS does not. Second, an ML-estimated character-level Markov model of English text is simplified using the two methods. Text synthesized from the resulting models becomes simpler and more consistent, and is at least as reasonable as text from the original Markov chain. In the final example, the methods are applied to denoise a word-level Markov chain of infant-directed speech. Results show that MERS filters out corruptions typical of spontaneous speech, leading to improvements in objective and subjective quality of the grammar in speech generated by the model. Thresholding yields similar subjective quality but inferior objective quality, as it may forbid grammatically legal constructions. First, however, we introduce the thresholding-based baseline simplification.

6.1 Thresholding-Based Simplification

For reference, we will compare the experimental results of MERS to a straightforward, thresholding-based probability concentration scheme, in which a simplified transition matrix is created by removing elements of $\widetilde{\boldsymbol{A}}$ smaller than a threshold $\tau \geq 0$,

$$a'_{ij}(\tau) = \begin{cases} \widetilde{a}_{ij} & \text{if } \widetilde{a}_{ij} \geq \tau \\ 0 & \text{otherwise,} \end{cases} \tag{33}$$

and applying standard row-sum normalization (32) to the result $\boldsymbol{A}'(\tau)$. τ is a free parameter of the method, similar to α from before. For $\boldsymbol{A}'(\tau)$ to be normalizable, we must have $\tau \leq \tau_{\max} = \min_i \max_j (\widetilde{a}_{ij})$.

We note that thresholding-based probability concentration has several qualitative differences from MERS:

1) $\boldsymbol{A}'(\tau)$ is not a smooth function of τ, and the simplification evolves in discrete steps. This means that only a finite number of different entropy-divergence trade-offs are possible. MERS, in contrast, provides a continuum of possible simplifications.
2) Unlike MERS, where small transition probabilities are typically made smaller (apart from the effect of the weighting $\operatorname{diag}\boldsymbol{\mu}$), transitions that are not removed by the thresholding can never have their probability decreased. Uncommon behavior is therefore initially made more likely, not less.
3) Because thresholding removes transitions completely, the resulting Markov chain may be split into several noncommunicating, recurrent parts. (The chain is cut into pieces, in effect.) This is a nonergodic system for which the concepts of entropy and divergence rate lose their meaning. Together with the bound $\tau \leq \tau_{\max}$, this suggests that low-entropy simplifications may not always be attainable.

6.2 Simplifying a Meteorological Model

To illustrate how MERS always yields an ergodic and connected model, whereas other simplifications may not, we consider a small Markov chain based on cloud-coverage data. Specifically, we extracted the percent opaque cloud measurements of the Total Sky Imager fractional sky coverage datastream at the ARM Climate Research Facility Southern Great Plains (SGP) site, Central Facility, in Lamont, OK, (`sgptsiskycoverC1.b1`[3]) recorded between 2000-07-02 and 2012-04-14. The measurement series were converted to daily cloud-cover averages and quantized to a scale of eighths $\{0, {}^1\!/\!_8, {}^2\!/\!_8, \ldots, 1\}$, whereafter a 9×9 first-order Markov chain transition matrix $\widetilde{\boldsymbol{A}}$ was ML-estimated to describe the quantized data series. Due to some missing data, the final matrix was based on 4 021 day pairs.

Figure 2 displays a heat map of the estimated transition matrix $\widetilde{\boldsymbol{A}}$. This shows evidence of bimodal behavior, where either clear or overcast weather is likely to remain largely unchanged, while intermediate states are less common and less predictable. As we shall see, this presents problems for thresholding-based probability concentration.

The Markov process $\widetilde{X}$ defined by the cloud-coverage transition matrix was simplified using the MERS formula (28). As seen in Figure 3, the simplifications for different α trace out a convex curve in (H_∞, D_∞)-space, similar to the rate-distortion function in lossy source coding.

[3] The data can be requested online for free at www.arm.gov.

H_∞	Text sampled from Markov chain
1.05	`e lord is that he judgment. and huldah his city of berothere thren of branch well i command that al`
1.00	`were a comfiture. and it were was made his places an he sleep. take aaron a scaffording with ye str`
0.94	`t all they shall the shall the princenser. selled him. and took king upon the goverthelemish. and s`
0.88	`t word. and ye lord. which the have buried. behold. and he wrother wind of the philistines unto the`
0.82	`he families shall laughteousness abhorring bullock. the people. as the moons. for that israel. and`
0.77	`escaped. and the children. and god. when image. and the land saith the lord of thy strength with hi`
0.71	`them whitherefore this from the lord. whereof. and words of ther children of they servants. and my`
0.65	`eople the counsel answere i will not prophet the son of the oppresert. neighbour own the places of`
0.59	`bring. and them diligent unto the lord of babylon. and moses. and they shall for the lord god. and`
0.54	`egation of thee. for the children of syria. and king. . and unto the tabernacle. and the lord of th`
0.48	` upon the children of the children of the children of the children of the people the children of is`
0.42	`the king the children of israel. because of jerusalem. and said unto the tabernacle of israel. whic`
0.36	`hildren of the children of the people. say unto thee. and they shall the lord. and the children of`
0.30	`dren of the children of the lord. and the tabernacle of thee. and the lord of israel. and the peopl`
0.25	`id unto the lord. and the children of the lord. and the children of the children of the children of`
0.19	`en of the children of the children of the children. and he children of the children of the children`

(a) Minimum entropy rate simplification.

H_∞	Text sampled from Markov chain
1.05	` and the commanded talked from the buried in and thy people. said zecharia. neighbour heave this da`
0.99	`ased to this separable to pastoredom. and he made and it. and spear of the woman of their asa king`
0.93	` the lord said. there bow david. he hath. where. feast. and of heart to the covenant out unto that`
0.87	`e son of aaron the moses. that ther gard the time out and to seed than to death. and israel. the hi`
0.81	` to death of them into him an hand the stranger of thee seen them. and the six hundreds. and he tit`
0.75	`ink the cities together side of ahaz saith moses was it with the city the sons of all be acts of me`
0.69	`d. be and the lord. and said unto the seven them to they with to desolate they were is the shalt th`
0.63	`his hand the profane them. that the lord hath saith the house of the higher is the chief of hosts a`
0.58	`rd god. whole carcases was that the lord hate of her. and said. i will not agains of israel. the lo`
0.52	`of war. and he with the lord saul shalt before the lord god. and the people. and the son of the lor`
0.45	`rd hath side. and the land the saith her the children of the set my people of the son of the city w`
0.39	`conders of the land the lord of the saith the lord shall that was and he son of the land the second`
0.33	`the lord god. and the land. and the lord god. and the shall be as a stonished the soul. the said un`

(b) Thresholding-based simplification. (Note the larger minimum H_∞ compared to 1a, due to τ reaching $\tau_{\max}$.)

Table 1

Random text sampled from simplified models over a range of entropy rates.

Since MERS provides the optimal rate-divergence trade-off, performance below this curve is not possible. Both entropy and KL-divergence rate remain finite everywhere, as expected. At low H_∞, the curve straightens out to a slope near negative unity, and any decrease in entropy immediately translates to an equivalent increase in relative entropy.

Figure 3 additionally shows the entropy-divergence combinations attainable by thresholding $\widetilde{\boldsymbol{A}}$. This simplification evolves in discrete steps in both H_∞ and D_∞ due to the discontinuous nature of the modification, and the achievable trade-offs are therefore drawn as nonconnected dots.

In agreement with theory, the entropy-divergence trade-offs from thresholding are inferior to the optimal curve traced out by MERS. Moreover, even though we have $\tau_{\max} \approx 0.1763$, thresholding produces nonergodic, reducible models for τ exceeding 0.1370. At this point—marked by $\times$ in the figure—the thresholded Markov chain fractures into two distinct, recurrent connected components, comprised of states $\{0,\, ^1/_8,\, ^2/_8\}$ and $\{^7/_8\}$, respectively. The idea of a single, well-defined entropy (or divergence) rate then loses its meaning, as does the rate-divergence graph. It follows that low-rate simplifications are not possible through thresholding for the cloud-cover process.

6.3 Simplifying a Text Model

Next, we investigated a much larger Markov model for character-level text synthesis. Inspired by the application in [42], an initial Markov chain was trained on a four-gram representation of the King James version of the old testament. This is equivalent to a fourth-order character-level Markov chain. The source text was pre-processed by removing verse numbers and converting to lower case, whereafter all contiguous sequences of whitespace, including new lines, were converted to single spaces while all sequences of punctuation, digits, and other non-alphabet characters were converted to single dots. The resulting text contained 3 190 276 samples from 24 309 distinct four-grams, and the trained $\widetilde{\boldsymbol{A}}$-matrix had 77 253 nonzero entries.

The fourth-order Bible text model was simplified using MERS and thresholding. Again, thresholding is incapable of low-entropy simplification, as τ is constrained by $\tau_{\max}$. In this case, $\tau_{\max} \approx 0.1092$, reaching $H_\infty(X'(\tau_{\max})) \approx 0.33$, while MERS can run until $H_\infty(X) \approx 0.19$ ($\alpha \approx 1.95$) before the general-purpose Matlab `eigs`-command begins to experience numerical difficulties with correctly identifying the leading eigenvector.

To investigate the nature and behavior of the two studied probability concentration schemes, it is instructive to generate text from the models they produce. Table 1 shows text strings sampled independently from simplified models over the entire range of entropy rates given above. Samples are quite similar between the two simplification methods, exhibiting classic Markov chain nonsense-text behavior at high rates, but turning increasingly repetitive and predictable as the rate decreases. For both schemes, the simplified samples

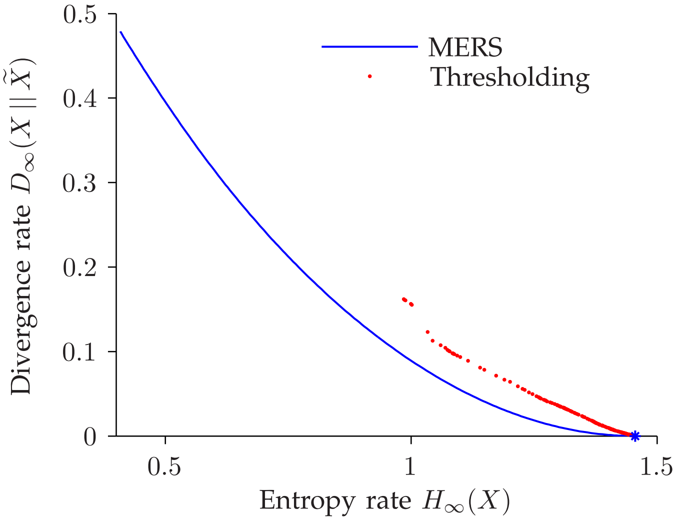


Figure 4. Entropy-divergence trade-offs for the speech grammar.

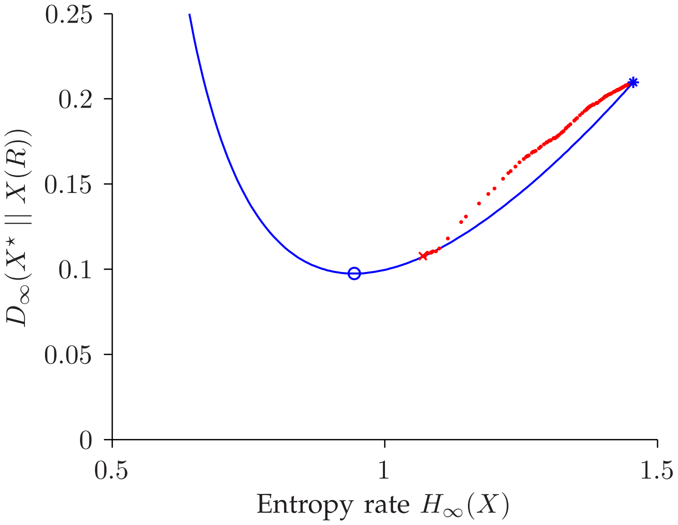


Figure 5. Denoising performance for the speech grammar. A circle marks the optimum.

arguably appear more characteristic of the training text. At low rates, they resemble a cartoon version of the Bible. Simple samples also tend to contain fewer illegal word constructions, suggesting that these models make fewer errors.

6.4 Denoising a Speech Grammar

For our final example, we consider using MERS or thresholding to remove disturbances from a corrupted Markov grammar. The grammar was based on a subset of the matrix-type sentences from the Swedish infant-directed speech corpus in [43], composed of 21 word tokens and a pause marker. Two data sequences of 10^7 symbols were generated by i.i.d. random sampling from this bag of sentences. One sequence was additionally corrupted by a moderate amount of random speech errors such as partial sentences, corrected, repeated, or omitted words, along with disfluencies such as filled pauses (particularly at sentence positions with high branching factors), marked by an additional token. Second-order Markov chains $X^\star$ and $\widetilde{X}$ were then fitted to the clean and the disturbed data, respectively, using maximum likelihood. For $\widetilde{X}$ this gave a 458×458 $\widetilde{\boldsymbol{A}}$-matrix containing 5 550 nonzero elements, about half the maximum possible. (The size of $\widetilde{\boldsymbol{A}}$ was determined by the fact that only 458 out of the $22^2 = 484$ conceivable token bigrams appeared in the sampled data.) $X^\star$, in contrast, was behaviorally very sparse, and had only 135 nonzero transition matrix elements on rows corresponding to positive-probability bigrams.

The Markov chain $\widetilde{X}$, representing the noisy observed process, was subsequently simplified using the MERS formula (28), or by thresholding. The resulting entropy-divergence trade-offs are graphed in Figure 4. We will write $X(R)$ for the optimal solution at a given entropy rate, $H_\infty(X(R)) = R$. As before, limits on τ prevent thresholding from producing low-rate simplifications, while MERS can go further, reaching $\alpha = 46$ before our `eigs`-based implementation becomes unreliable.

As described earlier, we anticipate that simplification will preferably eliminate uncommon and likely erroneous behavior, thus reducing the number of errors (noise) in the model. In our case, we wish to come closer to the clean matrix grammar of the original sentences. Since this is a synthetic example, we can compute this similarity by comparing our denoised model to $X^\star$, the best Markov chain fit to the clean data.

First, we assess denoising performance using an objective measure, by computing the standard KL-divergence rate between the reference process $X^\star$ and the denoised approximations X (MERS) and X' (thresholding). As seen in Figure 5, MERS is capable of recovering an improved model closer to the underlying grammar for a range of divergences. The minimum KL-divergence rate $D_\infty(X^\star \;\|\; X(R)) \approx 0.097$ nats occurs around $R \approx 0.94$, and is less than half the distortion present in the original $\widetilde{X}$. Since MERS does not eliminate transitions entirely, it has finite divergence rate everywhere.

The thresholding scheme initially performs almost as well as MERS for denoising. However, when τ exceeds 0.064 (marked by an × in Figure 5), thresholding eliminates a transition in $\widetilde{\boldsymbol{A}}$ that also occurs in the uncorrupted grammar $\boldsymbol{A}^\star$. This oversimplifies $\widetilde{X}$ to such an extent that constructions which are perfectly legal within the reference grammar are rejected as impossible, and leads to an infinite KL-divergence rate for all thresholded simplifications beyond this point. Thresholding therefore cannot reach the same objective denoising performance as MERS.

Next, we consider the simplified output itself. As illustrated in Table 2,[4] there are clear improvements also in subjective output quality. The table demonstrates that both the underlying grammar and the disturbed conversational speech processes can be well represented by second-order

[4] The table provides a word-by-word English translation of the original Swedish text. Underscores identify particle-based constructions where Swedish instead uses a single word token with a suffix, e.g., “the book” being “boken” in Swedish. The same table in the original Swedish can be found in Appendix C.

```
a bath [pause] look mommy [pause] take a shoe
[pause] it is daddy [pause] look olov [pause] take
a bath [pause] where is olov now [pause] take a
bath [pause] where is olov now [pause] it is a car
[pause] take a shoe [pause] take a car [pause] it
is a car [pause] look the_car [pause] look the_shoe
[pause] it is a book [pause] take a bath [pause]
where is the_bath now [pause] look the_car [pause]
look olov [pause] where is the_car now [pause] look
the_shoe [pause] take a car [pause] hi olov [pause]
hi olov [pause] where is olov
```

(a) Sample from error-free corpus.

```
daddy now hi [pause] olov [pause] [pause] hello olov
[pause] [pause] it is daddy [pause] look um the_shoe
[pause] [pause] it is um olov [pause] [pause] hello
olov [pause] look um the_book [pause] [pause] where
is olov now um take a book [pause] um um hello olov
[pause] it is olov it is um a car [pause] hello olov
[pause] um where is olov now [pause] um where is
olov now [pause] hello olov [pause] [pause] take a
bath [pause] look the_book the_book hi olov [pause]
um look mommy [pause] it is a car [pause] um take a
car hello
```

(b) Sentences disturbed by random speech errors.

```
a book [pause] hello olov [pause] where is olov
[pause] where is the_car now [pause] hello olov
[pause] it is a book [pause] it is olov [pause]
look daddy [pause] look the_bath [pause] it is a
book [pause] where is the_shoe now [pause] take
a car [pause] hi olov [pause] take a shoe [pause]
it is a book [pause] where is the_bath now [pause]
look the_car [pause] look the_car [pause] it is olov
[pause] it is a book [pause] take a car [pause] look
mommy [pause] hi olov [pause] take a car [pause]
look mommy [pause] look the_shoe [pause] look
```

(c) Sample from $X^\star$ fit to error-free data.

```
the_bath [pause] hi um olov [pause] take a book
[pause] take a shoe a shoe look the_book a car
[pause] it is daddy now [pause] [pause] where is
the_shoe now [pause] it is olov [pause] it is a shoe
[pause] hi olov [pause] um take a shoe [pause] where
is the_book now [pause] it is a car [pause] um take
a bath a bath [pause] look the_car [pause] um look
the_car [pause] olov [pause] look the_shoe [pause]
look daddy where is um a a car [pause] look olov it
is a bath [pause] where is is a car [pause] [pause]
```

(d) Sample from $\widetilde{X}$ fit to corrupted data.

```
hi olov [pause] hello olov [pause] um take a car
[pause] it is olov now [pause] take a book [pause]
it is daddy now [pause] take a a book [pause] take
a car [pause] it is a shoe [pause] where is um a
car [pause] hi olov [pause] it is a car [pause]
take a a shoe [pause] it is a car [pause] take a
car [pause] take a bath [pause] it is a book [pause]
um hi olov [pause] where is olov now [pause] it is
a bath [pause] um take a book where is the_shoe now
[pause] it
```

(e) Sample from MERS $X(R)$ at optimum denoising.

```
is a bath [pause] look mommy [pause] it is a car
[pause] it is [pause] daddy now [pause] it is olov
now [pause] hi olov [pause] hello olov [pause] where
is the_bath now [pause] it is mommy now [pause] take
a shoe [pause] [pause] take a book [pause] hi olov
[pause] it is mommy [pause] um it is a bath [pause]
take a bath [pause] hi olov [pause] um take a car
[pause] hello olov [pause] look daddy [pause] take a
book [pause] look the_book [pause] look the_bath
[pause] it is mommy now [pause] [pause] hi olov
[pause] look the_book
```

(f) Sample from thresholded $X'(R)$ at optimum denoising.

```
where is the_bath now [pause] where is the_shoe now [pause] where is the_book now [pause] it is a car
[pause] it is a car [pause] where is the_bath now [pause] where is the_shoe now [pause] it is a shoe
[pause] it is a shoe [pause] it is a book [pause] where is the_car now [pause] where is the_bath now
[pause] it is a book [pause] it is a bath [pause] where is the_book now [pause] where is the_bath now
[pause] it is a bath [pause] it is a bath [pause] it is a bath [pause] it is a book [pause]
```

(g) Sample from MERS $X(R)$ at low rate ($\alpha = 46$, $R \approx 0.41$).

Table 2

Translated excerpts from the clean and disturbed data, along with translations of random samples from the fitted models $X^\star$ and $\widetilde{X}$, the optimally denoised models from MERS and thresholding, and from low entropy-rate MERS (100 symbols each). See Appendix C for the original Swedish.

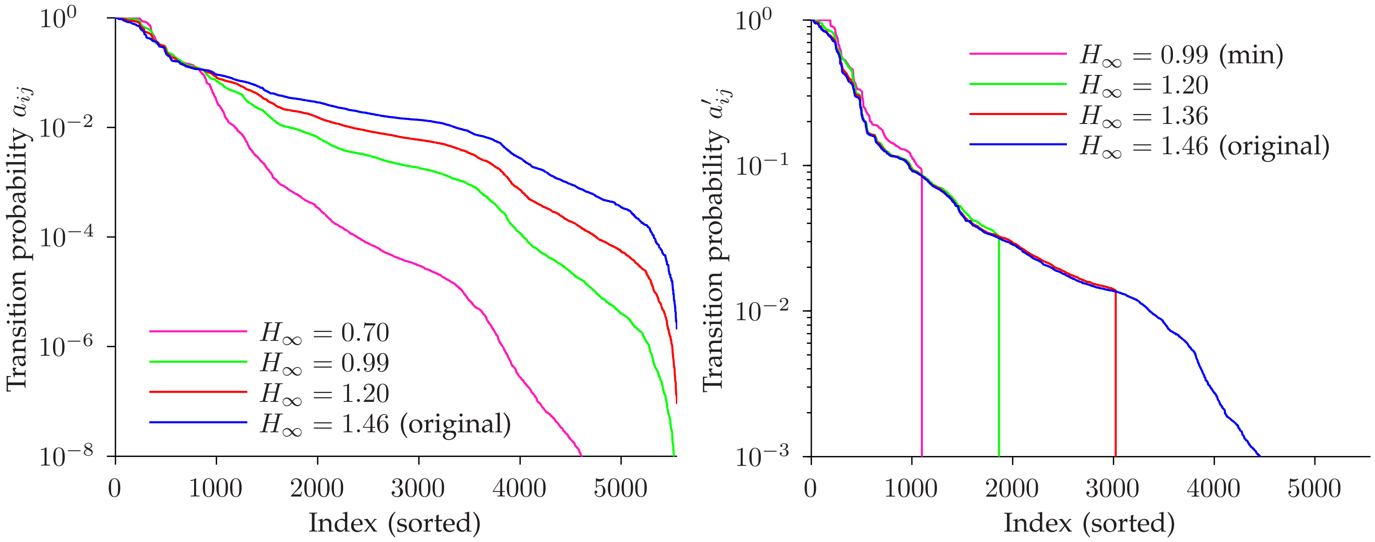
Markov chains. The errors (shown in red with wavy underline) in the conversational data and the model $\widetilde{X}$ derived from it are obvious and pervasive. Samples from denoised models minimizing $D_\infty(X^\star \mid\mid X(R))$ show noticeable improvements, displaying good variety while making significantly fewer mistakes. The simplified process X at low rate contains virtually no errors and is highly consistent, as it only generates sentences from a small subset of the original corpus with any appreciable probability.

It can be noted that the subjective difference between the two optimally denoised samples is small. In general, the fact that thresholding may disallow some legal constructions is seldom apparent in random output (it is easier to notice a presence than an absence), but can be of importance in other uses of the denoised grammar, such as compressing clean text. As always, which approach is preferable is likely to depend on the intended application and its associated constraints.

Finally, we visualize the effects of simplification on the transition matrix $\widetilde{\boldsymbol{A}}$. Figure 6 illustrates how the elements of the transition matrix evolve with increasing degree of simplification, for MERS and thresholding. Under thresholding, matrix entries can only increase, prior to the point where they are removed entirely. Due to how elements are ordered in the plot, this cut-off point, drawn as a vertical drop in element magnitude, moves from right to left in the figure. MERS, in contrast, takes matrix elements to a positive power α, which generally decreases element magnitude and corresponds to a uniform vertical scaling in the logarithmic plot 6a, effectively increasing the slope of the graph. Normalization, in turn, roughly translates the graph upwards a small amount. The largest matrix elements therefore exhibit increased magnitude, while others elements decrease, consistent with probability concentration.

7 Conclusions and Future Work

We have presented MERS, minimum entropy rate simplification for stochastic processes. This is an information-



(a) Minimum entropy rate simplification a_{ij}.

(b) Thresholding a'_{ij}. (Note the different scale from 6a.)

Figure 6. Simplified transition matrix elements at different rates, sorted by magnitude.

theoretic framework for tunable model simplification by concentrating the probability mass on the most representative behaviors. MERS is useful as post-processing for generative models, increasing quality and consistency in synthesis and sampling applications, as demonstrated in the experiments. The independence of unitary transformations of divergence-based MERS sets it apart from sparsity-based model simplification schemes such as [4] (which additionally is limited to i.i.d. processes).

MERS is closely related to a view that data is generated by a low-entropy underlying process $X^\star$, but subsequently influenced by a limited amount of unspecified disturbances to form the observed process $\widetilde{X}$. The framework can thus be considered both as a broad simplification principle, and as nonparametric denoising of stochastic process models.

We see room for future work in both theory and applications. Markov chains and Gaussian processes are highly common practical models, and there may be many situations where an information-theoretic diversity reduction technique is useful. On the theory side, it would be interesting to explore MERS solutions for additional model classes, dissimilarity measures (cf. [25]), and even generalized entropy rate concepts [44]. For the case of HMMs, one may consider performing approximate MERS by minimizing a suitable approximation or bound on the entropy rate, or by applying MERS to the underlying Markov chain of the HMM. Further connections with rate-distortion theory may also be worthy of investigation.

Acknowledgements

The authors wish to thank Associate Professor Richard Heusdens at TU Delft for pointing out some inaccuracies in an earlier version of the article.

References

[1] D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[2] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, “Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech,” in *Proc. Interspeech*, vol. 15, pp. 1504–1508, September 2014.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. ICASSP 2000*, pp. 1315–1318, June 2000.

[4] 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.*, vol. 9, pp. 485–516, Mar. 2008.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 3rd ed., Mar. 2009.

[6] 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*, pp. 294–299, Aug. 2007.

[7] J. Dines, J. Yamagishi, and S. King, “Measuring the gap between HMM-based ASR and TTS,” in *Proc. Interspeech*, pp. 1391–1394, Sept. 2009.

[8] A. Ozerov and W. B. Kleijn, “Asymptotically optimal model estimation for quantization,” *IEEE Trans. Comm.*, vol. 59, no. 4, pp. 1031–1042, 2011.

[9] K. Vertanen, “An overview of discriminative training for speech recognition,” tech. rep., Computer Speech, Text and Internet Technology, University of Cambridge, 15 J. J. Thomson Avenue, Cambridge CB3 0FD, United Kingdom, 2004.

[10] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” in *Proc. ACL 1996*, vol. 34, pp. 310–318, June 1996.

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

[12] F. Jelinek and R. L. Mercer, “Interpolated estimation of Markov source parameters from sparse data,” in *Proc. Workshop Pattern Recognit. Pract.*, vol. 1, May 1980.

[13] S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” in *IEEE Trans. ASSP*, pp. 400–401, 1987.

[14] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in *Proc. ICASSP 1995*, vol. 1, pp. 181–184, 1995.

[15] M. E. Brand, “An entropic estimator for structure discovery,” in *Proc. NIPS 1998*, pp. 723–729, Dec. 1998.

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

[17] S. Arora, E. Hazan, and S. Kale, "A fast random sampling algorithm for sparsifying matrices," in *Proc. RANDOM*, pp. 272–279, Aug. 2006.

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

[19] P. Domingos, "Occam's two razors: The sharp and the blunt," in *Proc. KDD-98*, vol. 4, pp. 37–43, Aug. 1998.

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

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 1991.

[22] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton 1999*, pp. 368–377, Sept. 1999.

[23] Y. Wang, G. Haffari, S. Wang, and G. Mori, "A rate distortion approach for semi-supervised conditional random fields," in *Proc. NIPS*, vol. 22, pp. 2008–2016, Dec. 2009.

[24] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *J. Stat. Phys.*, vol. 104, pp. 817–879, 2001.

[25] W. Stummer and I. Vajda, "On Bregman distances and divergences of probability measures," *IEEE Trans. Inf. Theory*, vol. 58, pp. 1277–1288, Mar. 2012.

[26] H. T. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. UAI*, vol. 15, pp. 21–30, July 1999.

[27] K. Marton and P. C. Shields, "The positive-divergence and blowing-up properties," *Isr. J. Math*, vol. 86, pp. 331–348, Oct. 1994.

[28] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden Markov process," *Theor. Comput. Sci.*, vol. 395, pp. 203–219, Apr. 2008.

[29] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Congr. Acoust.*, vol. 6, pp. C–17–C–20, Aug. 1968.

[30] S. Ihara, *Information Theory for Continuous Systems*. World Scientific Publishing Company, 1993.

[31] J. P. Burg, *A New Analysis Technique for Time Series Data*, pp. 42–48. IEEE Press, 1978.

[32] H.-O. Georgii, *Gibbs Measures and Phase Transitions*. Walter de Gruyter, 2nd ed., 2011.

[33] H. O. Wold, *A Study in the Analysis of Stationary Time Series*. Almqvist & Wiksell, 2nd ed., 1954.

[34] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, 1949.

[35] "Sound Quality Assessment Material recordings for subjective tests: Users' handbook for the EBU SQAM CD," Tech. Rep. 3253, European Broadcasting Union, Geneva, September 2008.

[36] Z. Rached, F. Alajaji, and L. L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, pp. 917–921, May 2004.

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

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

[39] I. Goldhirsch, S. A. Orszag, and B. K. Maulik, "An efficient method for computing leading eigenvalues and eigenvectors of large asymmetric matrices," *J. Sci. Comp.*, vol. 2, no. 1, pp. 33–58, 1987.

[40] L. Page, S. M. Brin, R. Motwani, and T. A. Winograd, "The PageRank citation ranking: Bringing order to the Web," Tech. Rep. 1999-66, Stanford InfoLab, Jan. 1999. Previous number SIDL-WP-1999-0120.

[41] G. M. Del Corso, A. Gullí, and F. Romani, "Comparison of Krylov subspace methods on the PageRank problem," Tech. Rep. TR-05-20, Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy, July 2005.

[42] D. Ron, Y. Singer, and N. Tishby, "The power of amnesia: Learning probabilistic automata with variable memory length," *Mach. Learn.*, vol. 25, no. 2–3, pp. 117–149, 1996.

[43] O. Räsänen, T. Altosaar, and U. K. Laine, "Comparison of prosodic features in Swedish and Finnish IDS/ADS speech," in *Proc. Nordic Prosody X*, vol. 10, Aug. 2008.

[44] C. Beck, "Generalised information and entropy measures in physics," *Contemp. Phys.*, vol. 50, pp. 495–510, May 2009.

Gustav Eje Henter received the Ph.D. degree in electrical engineering (telecommunications) in 2013 and the M.Sc. degree (Civilingenjör) in engineering physics in 2007, both from KTH Royal Institute of Technology in Stockholm, Sweden. In 2011 he was a visiting researcher at Victoria University of Wellington (VUW), New Zealand. He is currently a research fellow at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, United Kingdom, where his research interests include parametric and nonparametric statistical modeling, particularly for speech synthesis.

W. Bastiaan Kleijn received the Ph.D. degree in electrical engineering from Delft University of Technology, The Netherlands (TU Delft); an M.S.E.E. degree from Stanford University; and a Ph.D. degree in soil science and an M.Sc. degree in physics from the University of California, Riverside. He is a professor at Victoria University of Wellington (VUW), New Zealand, and TU Delft, The Netherlands (part-time). He was a professor and head of the Sound and Image Processing Laboratory at KTH Royal Institute of Technology, Stockholm, Sweden, from 1996 until 2010 and a founder of Global IP Solutions, a company that provided the original audio technology to Skype and was later acquired by Google. Before 1996, he was with the Research Division of AT&T Bell Laboratories in Murray Hill, New Jersey. He is an IEEE Fellow.