Simplified Probability Models for Generative Tasks: a Rate-Distortion Approach

Gustav Eje Henter and W. Bastiaan Kleijn

Sound and Image Processing Lab, School of Electrical Engineering, KTH – Royal Institute of Technology, Stockholm, Sweden

Paper 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.

The Problem

Consider generative models trained on data with interference:

- Speech sounds in realistic environments
- Field recordings of birdsong

The interference is learned with the model even as $N \to \infty$. Sampling from the model also reproduces interference, which is undesirable.

How can disturbances be eliminated from the model without parametric assumptions? (Nonparametric model denoising.)

The Principle

Let $\widetilde{X}_t$ be a given stationary, ergodic stochastic process model learned from disturbed data.

Assume disturbances are generally less common than desirable behaviour.

Removing uncommon, uncharacteristic behaviour (outcomes) from $\widetilde{X}_t$ likely removes more interference than relevant behaviour. This gives an improved and simplified model X_t of the underlying process.

Oversimplification will remove too much relevant behaviour as well, giving poor models. We use the rate-distortion theory framework for optimal trade-off between simplicity (rate) and dissimilarity (distortion).

Simplicity and Dissimilarity

A non-parametric, information theoretic measure of stochastic distribution complexity and variability is information entropy

$$H(P) = -\sum_i P(P=i) \ln P(P=i).$$

For stochastic processes, this is generalized to an entropy rate

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

To obtain simple processes, entropy rate should be minimized.

A simplified model should still be similar to the observations. The classic information theoretic quantifier of the dissimilarity of two distributions is the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_i P(P=i) \ln \frac{P(P=i)}{P(Q=i)}.$$

A KL-divergence rate can be defined similar to the entropy rate as

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

We assume the sought X_t is the "true" model of the desired behaviour, approximated by the known and fixed $\widetilde{X}_t$. Because of asymmetry in the arguments, constraining this divergence rate promotes sparse simplifications.

Differential quantities can be used for continuous-valued processes.

The MERS Proposal

Let Ξ be a class of stationary, ergodic stochastic processes.

The MERS solution X_t for $\widetilde{X}_t$ with max divergence D satisfies:

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

subject to

$$D_\infty(X_t \,||\, \widetilde{X}_t) \leq D.$$

Example: Markov Chains

A (first-order) Markov chain satisfies

$$P(X_{t+1} \,|\, X_t, X_{t-1}, \ldots) = P(X_{t+1} \,|\, X_t).$$

Let $\widetilde{X}_t$ be a given stationary, ergodic Markov chain on $\{1, \ldots, K\}$ described by transition probability matrix $\widetilde{\boldsymbol{A}}$ with elements $\widetilde{a}_{ij} = P(\widetilde{X}_{t+1} = j \,|\, \widetilde{X}_t = i)$.

Let X_t be another Markov chain satisfying the same requirements, described by the bigram probability matrix $\boldsymbol{B}$ with $b_{ij} = P(X_t = i \wedge X_{t+1} = j)$.

To be a minimum entropy rate simplification of $\widetilde{X}_t$, X_t must solve

$$\min_{\boldsymbol{B}} -\sum_{ij} b_{ij} \ln \frac{b_{ij}}{\sum_{j'} b_{ij'}}$$

subject to

$$\sum_{ij} b_{ij} \ln \frac{b_{ij}}{\widetilde{a}_{ij} \sum_{j'} b_{ij'}} \leq D$$
$$(\boldsymbol{B} - \boldsymbol{B}^T)\mathbf{1} = \mathbf{0}$$
$$\mathbf{1}^T \boldsymbol{B} \mathbf{1} = 1$$
$$\boldsymbol{B} \geq \mathbf{0}.$$

An Iterative Solution

A fast algorithm for finding B can be derived, similar to the Blahut-Arimoto procedure for computing points on the rate-distortion curve.

1. Start with a Lagrange multiplier $\alpha > 1$, an initial guess $\boldsymbol{B}^{(0)}$, and $m = 0$.
2. Given B, optimize for q: $\boldsymbol{q}^{(m)} = \boldsymbol{B}^{(m)}\mathbf{1}$.
3. Given q, optimize for B:
   (a) Define $\boldsymbol{B}'^{(m+1)}$ through $b'^{(m+1)}_{ij} = (\widetilde{a}_{ij})^\alpha (\boldsymbol{q})_i^{(m)}$.
   (b) Let $n = 0$ and $\boldsymbol{\mu}^{(n)} = \mathbf{1}$.
   (c) Let $\mu_i^{(n+1)} = \sqrt{\frac{\sum_{j=1,\, j\neq i}^K \mu_j^{(n)} b'^{(m+1)}_{ji}}{\sum_{j=1,\, j\neq i}^K (\mu_j^{(n)})^{-1} b'^{(m+1)}_{ji}}}$.
   (d) Let $n = n + 1$ and repeat from the previous step until convergence.
   (e) Form $\boldsymbol{B}''^{(m+1)} = (\operatorname{diag} \boldsymbol{\mu}^{(n)})^{-1} \boldsymbol{B}'^{(m+1)} (\operatorname{diag} \boldsymbol{\mu}^{(n)})$.
   (f) Normalize to get $\boldsymbol{B}^{(m+1)} = (\mathbf{1}^T \boldsymbol{B}''^{(m+1)} \mathbf{1})^{-1} \boldsymbol{B}''^{(m+1)}$.
4. If not converged, let $m = m + 1$ and repeat from 2.

This converges quickly in practice. The Lagrange multiplier α controls the entropy-divergence trade-off.

A Numerical Example

We take a $K = 12$ Markov chain $\widetilde{X}_t$ trained on Bach chorales using ML. MERS is compared with a reverse water-filling-like scheme where a simplified A is obtained by thresholding each row of $\widetilde{A}$.

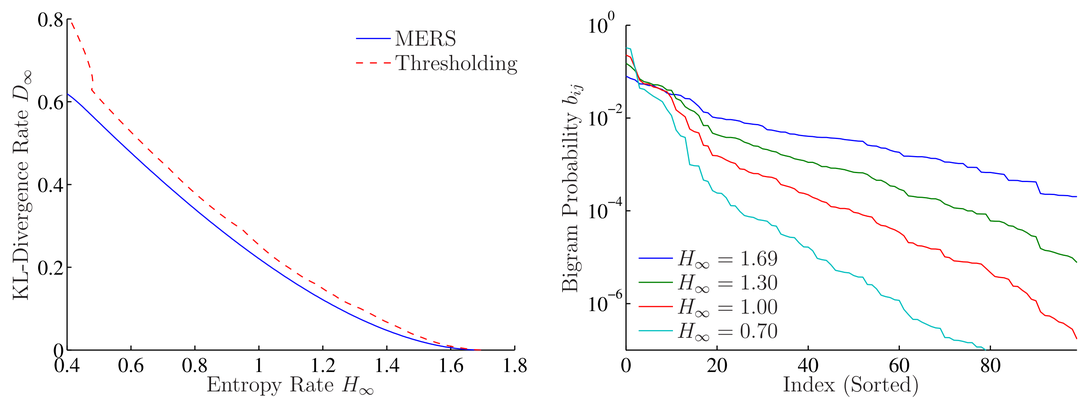

Figure 1: *Entropy-Divergence Curve*

Figure 2: *Bigram Probabilities*

The first graph shows that MERS achieves better trade-offs between simplicity and dissimilarity than thresholding.

The second graph shows the probability concentration properties of MERS. Small matrix elements (unusual transitions) are eroded away. Since many outcomes become very uncommon, a kind of sparsity in the outcome space is produced; only the most characteristic behavior is retained.

Future Work

We are working on MERS solutions for additional classes of stochastic processes. Approximative solutions may be considered for complex cases such as HMMs. We are also developing faster solution formulas.

Other future work includes theory, e.g., properties of the optimal solutions and the simplicity-dissimilarity curve, and practical applications, e.g., improving speech synthesis models.