Minimum Entropy Rate Simplification of Stochastic Processes: Supplemental Material

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

THIS document contains supplemental material for the IEEE Transactions on Pattern Analysis and Machine Intelligence article "Minimum Entropy Rate Simplification of Stochastic Processes." The supplement is divided into three appendices: the first on MERS for Gaussian processes, and the remaining two on, respectively, the theory and the experimental results of MERS for Markov chains.

A Supplemental Derivations for Gaussian MERS

This section presents detailed derivations of the three solutions to MERS for Gaussian processes presented in Sections 4.1 through 4.3 of the main article, along with arguments for the translation identities in the article Section 4.4.

A.1 Gaussian MERS Solution

We first consider the purely nondeterministic case; the result is easily extended to arbitrary stationary and ergodic Gaussian processes using the Wold decomposition.

Let X and $\widetilde{X}$ be two discrete-time stationary and ergodic purely nondeterministic univariate Gaussian processes, with spectral power density functions $R_X\left(e^{i\omega}\right)$ and $R_{\widetilde{X}}\left(e^{i\omega}\right)$ respectively. These are by necessity positive and defined on $(-\pi,\,\pi]$. The process $\widetilde{X}$ is considered given, and we are trying to find a corresponding X which minimizes the differential entropy rate $h_\infty\left(X\right)$ under a differential KL-divergence rate constraint $d_\infty(X \mid\mid \widetilde{X}) \leq d$ and a constraint $R_X\left(e^{i\omega}\right) \geq r_{\min}$ on the minimum power at any frequency.

Following [1], the differential entropy rate of X can be calculated as

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

Furthermore, the relative differential entropy rate is

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

To minimize the differential entropy rate under our divergence constraint we shall use variational calculus in a computation similar to that of Burg [2], but with an added constraint on the minimum power. Fix a Lagrange multiplier $\lambda > 0$ for the divergence constraint and a set (function) of Lagrange multipliers $(4\pi)^{-1}\,\mu\left(x\right) \geq 0$ for the power constraint, and seek the stationary points of the functional

$$\begin{aligned}\eta_\lambda\left(R_X\left(\cdot\right)\right) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(4\pi^2 e^2 R_X\left(e^{i\omega}\right)\right)\,\mathrm{d}\omega + \lambda\frac{1}{4\pi}\int_{-\pi}^{\pi} \left(\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)} - 1 - \log\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)}\right)\,\mathrm{d}\omega \\ + \int_{-\pi}^{\pi} \frac{1}{4\pi}\mu\left(\omega\right)\left(r_{\min} - R_X\left(e^{i\omega}\right)\right)\,\mathrm{d}\omega \qquad (3)\end{aligned}$$

$$= \frac{1}{4\pi}\int_{-\pi}^{\pi} \left(\log\left(4\pi^2 e^2 R_X\left(e^{i\omega}\right)\right) + \lambda\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)} - \lambda - \lambda\log\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)} + \mu\left(\omega\right)\left(r_{\min} - R_X\left(e^{i\omega}\right)\right)\right)\,\mathrm{d}\omega. \tag{4}$$

- *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.*

The Euler-Lagrange equation is obtained by setting the derivative of the integrand with respect to $R_X\left(e^{i\omega}\right)$ to zero:

$$\frac{\partial}{\partial R_X\left(e^{i\omega}\right)}\left(\log\left(4\pi^2e^2R_X\left(e^{i\omega}\right)\right)+\lambda\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)}-\lambda-\lambda\log\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)}+\mu\left(\omega\right)\left(r_{\min}-R_X\left(e^{i\omega}\right)\right)\right)=0 \tag{5}$$

$$\frac{\partial}{\partial R_X\left(e^{i\omega}\right)}\left((1-\lambda)\log R_X\left(e^{i\omega}\right)+\lambda\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)}+\lambda\log R_{\widetilde{X}}\left(e^{i\omega}\right)-\mu\left(\omega\right)R_X\left(e^{i\omega}\right)\right)=0 \tag{6}$$

$$\frac{1-\lambda}{R_X\left(e^{i\omega}\right)}+\frac{\lambda}{R_{\widetilde{X}}\left(e^{i\omega}\right)}-\mu\left(\omega\right)=0. \tag{7}$$

We can now distinguish two cases, the first being when the minimum-power constraint is not active for a particular ω. In this case complementary slackness requires $\mu\left(\omega\right)=0$, and one obtains

$$\frac{1-\lambda}{R_X\left(e^{i\omega}\right)}+\frac{\lambda}{R_{\widetilde{X}}\left(e^{i\omega}\right)}=0 \tag{8}$$

$$R_X\left(e^{i\omega}\right)=\frac{\lambda-1}{\lambda}R_{\widetilde{X}}\left(e^{i\omega}\right), \tag{9}$$

which is valid as long as

$$\frac{\lambda-1}{\lambda}R_{\widetilde{X}}\left(e^{i\omega}\right)\geq r_{\min}. \tag{10}$$

We see that the MERS simplification of X in this case is a simple scaling (contraction) of the spectral density $R_{\widetilde{X}}\left(e^{i\omega}\right)$ by a factor $\frac{\lambda-1}{\lambda}=\alpha^{-1}$. Obviously, only values $\lambda>1$ are valid in practice; $\lambda\in(0,\,1)$ gives nonsensical, negative spectral powers. In the corresponding interval $\alpha\in(1,\,\infty)$ we recover all MERS solutions between the original $\widetilde{X}$ and fully predictable X, for the case of processes $\mathcal{X}$ with no lower bound on the spectral power ($r_{\min}=0$).

If $r_{\min}>0$, there is a risk that the right-hand side of Equation (9) falls below $r_{\min}$, in which case the minimum-power constraint becomes active, so that $R_X\left(e^{i\omega}\right)=r_{\min}$. For such ω, we obtain (see Equation (7))

$$\frac{1-\lambda}{r_{\min}}+\frac{\lambda}{R_{\widetilde{X}}\left(e^{i\omega}\right)}-\mu\left(\omega\right)=0 \tag{11}$$

$$\mu\left(\omega\right)=\frac{\lambda}{R_{\widetilde{X}}\left(e^{i\omega}\right)}+\frac{1-\lambda}{r_{\min}} \tag{12}$$

$$\mu\left(\omega\right)=\frac{\frac{\alpha}{\alpha-1}}{R_{\widetilde{X}}\left(e^{i\omega}\right)}+\frac{1-\frac{\alpha}{\alpha-1}}{r_{\min}} \tag{13}$$

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

which is strictly positive as long as

$$\alpha^{-1}R_{\widetilde{X}}\left(e^{i\omega}\right)<r_{\min}. \tag{15}$$

Putting the two cases together, we obtain the solution

$$R_X\left(e^{i\omega}\right)=\max\left(r_{\min},\,\alpha^{-1}R_{\widetilde{X}}\left(e^{i\omega}\right)\right) \tag{16}$$

$$\mu\left(\omega\right)=\frac{1}{\alpha-1}\max\left(0,\,\left(\alpha^{-1}R_{\widetilde{X}}\left(e^{i\omega}\right)\right)^{-1}-\left(r_{\min}\right)^{-1}\right) \tag{17}$$

for $\alpha\in(1,\,\infty)$. It is straightforward to verify that this satisfies the requisite optimality criteria. The minimum differential entropy rate achievable is limited by the white noise process $X_{\min}$ defined by

$$R_{X_{\min}}\left(e^{i\omega}\right)\equiv r_{\min}, \tag{18}$$

which has

$$h_\infty\left(X_{\min}\right)=\frac{1}{2}\log\left(4\pi^2e^2r_{\min}\right). \tag{19}$$

If, on the other hand, $r_{\min}\geq\min_\omega R_{\widetilde{X}}\left(e^{i\omega}\right)$, we have $\widetilde{X}\notin\mathcal{X}$ ($\widetilde{X}$ is not in the feasible set), and the minimum achievable distortion is nonzero.

A.2 Weighted Itakura-Saito Divergence MERS

By multiplying the Itakura-Saito criterion by the observed signal spectrum $R_{\widetilde{X}}\left(e^{i\omega}\right)$, we obtain a weighted dissimilarity measure

$$d_{\mathrm{IS}}^{q}(X \mid\mid \widetilde{X}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right) + R_{\widetilde{X}}\left(e^{i\omega}\right)\log\left(\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_X\left(e^{i\omega}\right)}\right)\right)\mathrm{d}\omega \tag{20}$$

which emphasizes correctness of spectral peaks more than spectral valleys. This new measure is nonnegative, since it can be written

$$d_{\mathrm{IS}}^{q}(X \mid\mid \widetilde{X}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} R_{\widetilde{X}}\left(e^{i\omega}\right) f\left(\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_X\left(e^{i\omega}\right)}\right)\mathrm{d}\omega, \tag{21}$$

where $f(t) = t - 1 - \log t$ and $R_{\widetilde{X}}\left(e^{i\omega}\right)$ are both nonnegative. The measure evaluates to zero when $R_X\left(e^{i\omega}\right) = R_{\widetilde{X}}\left(e^{i\omega}\right)$, and is strictly positive whenever $R_X\left(e^{i\omega}\right) \neq R_{\widetilde{X}}\left(e^{i\omega}\right)$ on a set of nonzero measure, assuming $R_{\widetilde{X}}\left(e^{i\omega}\right) > 0$ everywhere (which is required for ergodicity). Expression (20) is superficially similar to the generalized KL-divergence in [3], but with the signs of the $R_X\left(e^{i\omega}\right)$ and $R_{\widetilde{X}}\left(e^{i\omega}\right)$-terms reversed. It is not a Bregman divergence or an f-divergence.

To solve the MERS problem using the weighted Itakura-Saito divergence in (20), we introduce a Lagrange multiplier $\lambda > 0$ for the new dissimilarity constraint, along with $\mu(\omega) \geq 0$ for the lower bound $R_X\left(e^{i\omega}\right) \geq r_{\min}$ as before. This yields the functional

$$\begin{aligned}\eta_\lambda\left(R_X(\cdot)\right) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\left(4\pi^2 e^2 R_X\left(e^{i\omega}\right)\right)\mathrm{d}\omega + \int_{-\pi}^{\pi}\mu(\omega)\left(r_{\min} - R_X\left(e^{i\omega}\right)\right)\mathrm{d}\omega \\ + \lambda\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right) + R_{\widetilde{X}}\left(e^{i\omega}\right)\log\left(\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_X\left(e^{i\omega}\right)}\right)\right)\mathrm{d}\omega.\end{aligned} \tag{22}$$

As usual, the Euler-Lagrange equation is obtained by setting the derivative of the integrand with respect to $R_X\left(e^{i\omega}\right)$ to zero. This produces

$$\frac{\partial}{\partial R_X\left(e^{i\omega}\right)}\left(\log R_X\left(e^{i\omega}\right) + \lambda R_X\left(e^{i\omega}\right) - \lambda R_{\widetilde{X}}\left(e^{i\omega}\right) + \lambda R_{\widetilde{X}}\left(e^{i\omega}\right)\log\left(\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_X\left(e^{i\omega}\right)}\right)\right) = 0 \tag{23}$$

$$\frac{1}{R_X\left(e^{i\omega}\right)} + (\lambda - \mu(\omega)) - \lambda\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_X\left(e^{i\omega}\right)} = 0 \tag{24}$$

$$\frac{1}{\lambda} + \left(1 - \frac{1}{\lambda}\mu(\omega)\right)R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right) = 0, \tag{25}$$

since $R_X\left(e^{i\omega}\right)$ and λ are both nonnegative.

We once again must consider two distinct cases. The first is that the lower-bound constraint is not active for a particular ω, $\mu(\omega) = 0$, and so

$$\frac{1}{\lambda} + R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right) = 0 \tag{26}$$

$$R_X\left(e^{i\omega}\right) = R_{\widetilde{X}}\left(e^{i\omega}\right) - \frac{1}{\lambda}, \tag{27}$$

which is valid whenever

$$R_{\widetilde{X}}\left(e^{i\omega}\right) - \frac{1}{\lambda} \geq r_{\min}. \tag{28}$$

This is a simple subtraction from the power spectral density. For processes with no lower bound on the spectral power (i.e., $r_{\min} = 0$), we see that $R_X\left(e^{i\omega}\right) \to R_{\widetilde{X}}\left(e^{i\omega}\right)$ as $\lambda^{-1} \to 0$ (no simplification), while the entropy rate of X approaches zero as λ^{-1} increases towards $\min_\omega R_{\widetilde{X}}\left(e^{i\omega}\right)$ (complete predictability, under certain continuity assumptions on $R_{\widetilde{X}}\left(e^{i\omega}\right)$).

In the alternative case, the condition in Equation (28) is violated, in which case the lower bound gives $R_X\left(e^{i\omega}\right) = r_{\min}$. Slotting this into Equation (25) and assuming $r_{\min} > 0$, we obtain

$$\frac{1}{\lambda} + \left(1 - \frac{1}{\lambda}\mu(\omega)\right)r_{\min} - R_{\widetilde{X}}\left(e^{i\omega}\right) = 0 \tag{29}$$

$$-r_{\min}\frac{1}{\lambda}\mu(\omega) = R_{\widetilde{X}}\left(e^{i\omega}\right) - \frac{1}{\lambda} - r_{\min} \tag{30}$$

$$\mu(\omega) = \frac{\lambda}{r_{\min}}\left(r_{\min} + \frac{1}{\lambda} - R_{\widetilde{X}}\left(e^{i\omega}\right)\right), \tag{31}$$

which is strictly positive whenever

$$R_{\widetilde{X}}\left(e^{i\omega}\right) - \frac{1}{\lambda} < r_{\min}. \tag{32}$$

Combining the two cases, we arrive at the solution

$$R_X\left(e^{i\omega}\right) = \max\left(r_{\min},\, R_{\widetilde{X}}\left(e^{i\omega}\right) - \lambda^{-1}\right) \tag{33}$$

$$\mu\left(\omega\right) = \frac{\lambda}{r_{\min}} \max\left(0,\, r_{\min} + \lambda^{-1} - R_{\widetilde{X}}\left(e^{i\omega}\right)\right) \tag{34}$$

for $r_{\min} > 0$, valid for all $\lambda \in (0,\, \infty)$. (For $r_{\min} = 0$ we have the previous solution derived in Equation (27).) Exactly like the general solution in Section A.1, there is a nonzero minimum rate $h_\infty\left(X_{\min}\right)$, and the minimum distortion can be nonzero as well.

A.3 Variance-Constrained Gaussian MERS

Another interesting variation of MERS is obtained by constraining the standard relative entropy rate (regular Itakura-Saito divergence), but preventing the X-process from simply shrinking away by only considering processes that satisfy the additional variance constraint

$$\mathrm{Var}(X_t) = \mathrm{Var}(\widetilde{X}_t). \tag{35}$$

For Gaussian processes the variance may be written in terms of the spectral density function as

$$\mathrm{Var}(X_t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R_X\left(e^{i\omega}\right)\, \mathrm{d}\omega, \tag{36}$$

so the constraint (35) may be expressed

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left(R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right)\right)\, \mathrm{d}\omega = 0. \tag{37}$$

Introducing an additional Lagrange multiplier μ for the variance leads to a modified functional

$$\eta_{\lambda,\mu}\left(R_X\left(\cdot\right)\right) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log\left(4\pi^2 e^2 R_X\left(e^{i\omega}\right)\right) + \lambda\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)} - \lambda - \lambda\log\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)}\right)\mathrm{d}\omega$$
$$+ 2\mu\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(R_X\left(e^{i\omega}\right) - R_{\widetilde{X}}\left(e^{i\omega}\right)\right)\mathrm{d}\omega. \tag{38}$$

We again derive the Euler-Lagrange equation as

$$\frac{\partial}{\partial R_X\left(e^{i\omega}\right)}\left((1-\lambda)\log R_X\left(e^{i\omega}\right) + \lambda\frac{R_X\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right)} + 2\mu R_X\left(e^{i\omega}\right)\right) = 0 \tag{39}$$

$$\frac{1-\lambda}{R_X\left(e^{i\omega}\right)} + \frac{\lambda}{R_{\widetilde{X}}\left(e^{i\omega}\right)} + 2\mu = 0 \tag{40}$$

$$\frac{\lambda-1}{R_X\left(e^{i\omega}\right)} = \frac{2\mu R_{\widetilde{X}}\left(e^{i\omega}\right) + \lambda}{R_{\widetilde{X}}\left(e^{i\omega}\right)}, \tag{41}$$

which gives

$$R_X\left(e^{i\omega}\right) = \frac{\left(\lambda-1\right)R_{\widetilde{X}}\left(e^{i\omega}\right)}{2\mu R_{\widetilde{X}}\left(e^{i\omega}\right) + \lambda} \tag{42}$$

$$= \frac{1}{\nu}\frac{R_{\widetilde{X}}\left(e^{i\omega}\right)}{R_{\widetilde{X}}\left(e^{i\omega}\right) + \frac{\alpha}{\nu}}. \tag{43}$$

The new Lagrange multipliers $\nu = \frac{2\mu}{\lambda-1}$ and α here can take on any nonzero real values, though the result only makes sense if $\nu R_{\widetilde{X}}\left(e^{i\omega}\right) + \alpha > 0$ everywhere.

By considering the behavior of the function $f\left(r\right) = \frac{1}{\nu}\frac{r}{r+c}$ for a subset $r \in R = [R_{\min},\, R_{\max}] \subset (0,\, \infty)$ of the positive real line, we see that we that $\nu < 0$, $c < -R_{\max}$ corresponds to positive and increasing concave $f\left(r\right)$ with a derivative $f'\left(r\right) = \frac{1}{\nu}\frac{c}{\left(r+c\right)^2}$ greater than $-\frac{1}{\nu} > 0$. This implies that spectral peaks are emphasized, similar to reverse water-filling. The region $\nu,\, c > 0$, meanwhile, corresponds to a positive but concave f with slope between zero and $\frac{1}{\nu}$. This mapping smooths out spectral peaks and instead maximizes entropy under the constraints.

A.4 Translation Invariance

The fact that $h_\infty(\mu + X) = h_\infty(X)$ when μ is deterministic is theorem 2.4.1 in [1]. For

$$d_\infty(\mu + X \mid\mid \widetilde{\mu} + \widetilde{X}) \geq d_\infty(\widetilde{\mu} + X \mid\mid \widetilde{\mu} + \widetilde{X}), \tag{44}$$

it suffices to prove that

$$d_\infty(\Delta\mu + X \mid\mid \widetilde{X}) \geq d_\infty(X \mid\mid \widetilde{X}), \tag{45}$$

where $\Delta\mu = \mu - \widetilde{\mu}$. Since $\Delta\mu + X$ and X have the same spectral density functions (and covariance matrices), the general theorem 2.4.5 from [1] applies. This establishes

$$\begin{aligned} d_\infty(\Delta\mu + X \mid\mid \widetilde{X}) &= d_\infty(\Delta\mu + X \mid\mid X) + d_\infty(X \mid\mid \widetilde{X}) &(46)\\ &\geq d_\infty(X \mid\mid \widetilde{X}), &(47)\end{aligned}$$

which is the desired result. Equality occurs when $d_\infty(\Delta\mu + X \mid\mid X) = 0 \Leftrightarrow \Delta\mu + X = X \Leftrightarrow \mu = \widetilde{\mu}$.

B Supplemental Derivations for Markov Chain MERS

This section presents an alternative representation of the MERS problem for Markov chains, and subsequently shows how the optimal solution presented in Section 5.1 of the main article can be derived from this representation.

B.1 Bigram Matrix MERS Formulation

The MERS problem for first-order Markov chains is to find a stationary, ergodic Markov chain X that minimizes the entropy rate $H_\infty(X)$ under a divergence constraint $D_\infty(X \mid\mid \widetilde{X}) \leq D$. Here, $\widetilde{X}$ is another, given stationary and ergodic Markov chain over the same state space as X. We may represent X and $\widetilde{X}$ by their transition matrices $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$ with elements $a_{ij} = P(X_{t+1} = j \mid X_t = i)$, and similarly for $\widetilde{\boldsymbol{A}}$. We also have the stationary distribution vector $\boldsymbol{\pi}$ with elements $\pi_i = P(X_t = i) > 0$, as described in the article.

With our notation, the entropy rate and relative entropy rate of a first order Markov chain X can be expressed as

$$H_\infty(X) = -\sum_{ij} \pi_i a_{ij} \log a_{ij} \tag{48}$$

$$D_\infty(X \mid\mid \widetilde{X}) = \sum_{ij} \pi_i a_{ij} \log \frac{a_{ij}}{\widetilde{a}_{ij}}. \tag{49}$$

However, these formulas are cumbersome in practice since they involve $\boldsymbol{\pi}$, which is a normalized version of the leading eigenvector of $\boldsymbol{A}^\intercal$ and thus a complicated function of the matrix elements a_{ij}. A better option is to represent the Markov chain X by its *bigram probability matrix* $\boldsymbol{B}$, having elements $b_{ij} = P(X_t = i \wedge X_{t+1} = j)$. $\boldsymbol{B}$ is related to the transition matrix through the formula $\boldsymbol{B} = \operatorname{diag}(\boldsymbol{\pi})\boldsymbol{A}$, a matrix version of the well-known relationship $P(X_t = i \wedge X_{t+1} = j) = P(X_t = i)\,P(X_{t+1} = j \mid X_t = i)$. Standard requirements for probability distributions give $\boldsymbol{B} \geq \boldsymbol{0}$ and $\boldsymbol{1}^\intercal\boldsymbol{B}\boldsymbol{1} = 1$. The fact that the process is stationary, so that $\pi_i = P(X_t = i) = P(X_{t+1} = i)$, also contributes a constraint $\boldsymbol{B}\boldsymbol{1} = \boldsymbol{B}^\intercal\boldsymbol{1}$ on the marginal distributions. It is easy to see that there is a one-to-one correspondence between $\boldsymbol{A}$ and $\boldsymbol{B}$ in our case.

Using the bigram probability matrix representation, the MERS problem for first-order Markov chains can be written as

$$\min_{\mathbf{B} \in \mathbb{R}^{k \times k}} -\sum_{ij} b_{ij} \log \frac{b_{ij}}{\sum_{j'} b_{ij'}} \tag{50}$$

subject to

$$\begin{aligned} \sum_{ij} b_{ij} \log \frac{b_{ij}}{\widetilde{a}_{ij} \sum_{j'} b_{ij'}} &\leq D &(51)\\ \boldsymbol{1}^\intercal\boldsymbol{B}\boldsymbol{1} &= 1 &(52)\\ \boldsymbol{B}\boldsymbol{1} - \boldsymbol{B}^\intercal\boldsymbol{1} &= \boldsymbol{0} &(53)\\ \boldsymbol{B} &\geq \boldsymbol{0}. &(54)\end{aligned}$$

This formulation is preferable for practical purposes since it does not include any implicit functions of the variables b_{ij}.

B.2 Markov Chain MERS Solution

General nonconvex optimization problems like the above can be demanding to solve. However, it is here possible to solve the problem analytically, using a technique derived from the Blahut-Arimoto algorithm [4], [5] in rate-distortion theory. The first step is to fix a Lagrange multiplier $\lambda \geq 0$ for the dissimilarity constraint (51), and minimize the combined objective function

$$f_\lambda(\boldsymbol{B}) = (\lambda - 1) \sum_{ij} b_{ij} \log \frac{b_{ij}}{\sum_{j'} b_{ij'}} - \lambda \sum_{ij} b_{ij} \log \widetilde{a}_{ij} - \lambda D \tag{55}$$

under the remaining constraints. By considering λ fixed, we define a particular trade-off between simplicity and divergence, corresponding to a point on the rate-distortion curve. Like for the Gaussian case, we shall see that all relevant solutions correspond to $\lambda > 1$, where we can define the factor $\alpha = \frac{\lambda}{\lambda - 1} \in (1, \infty)$ as before. In this region we may divide by $\lambda - 1$, and instead maximize

$$f_\alpha(\boldsymbol{B}) = \sum_{ij} b_{ij} \log b_{ij} - \sum_{ij} b_{ij} \log \sum_{j'} b_{ij'} - \alpha \sum_{ij} b_{ij} \log \widetilde{a}_{ij} - \alpha D. \tag{56}$$

Next, we use the same trick as for the Blahut-Arimoto algorithm, transforming the problem to a minimization over two sets of variables, where each set is straightforward to optimize if the other is fixed. We do this by noting that

$$\max_{\boldsymbol{q} \geq \boldsymbol{0}} \sum_{ij} b_{ij} \log q_i \text{ subject to } \boldsymbol{1}^\intercal \boldsymbol{q} = 1 \tag{57}$$

is solved by $\boldsymbol{q}^\star(\boldsymbol{B}) = \boldsymbol{B}\boldsymbol{1}$ for nonnegative $\boldsymbol{B}$. (Superscripted stars here signify optimal solutions, and are not related to the underlying process $X^\star$ defined within the main article text.) This lets us replace the difficult term $b_{ij} \log \sum_{j'} b_{ij'}$, containing a sum within a logarithm, and instead define an augmented objective function

$$f_\alpha(\boldsymbol{B}, \boldsymbol{q}) = \sum_{ij} b_{ij} \log b_{ij} - \sum_{ij} b_{ij} \log q_i - \alpha \sum_{ij} b_{ij} \log \widetilde{a}_{ij}. \tag{58}$$

Maximizing $f_\alpha(\boldsymbol{B}, \boldsymbol{q})$ subject to constraints (52) through (54), as well as $\boldsymbol{q} \geq \boldsymbol{0}$ and $\boldsymbol{1}^\intercal \boldsymbol{q} = 1$, yields a $\boldsymbol{B}$ that is also optimal for the original problem.

We already know how to optimize $f_\alpha(\boldsymbol{B}, \boldsymbol{q})$ for a given, constant $\boldsymbol{B}$. If we instead fix $\boldsymbol{q}$ and introduce real Lagrange multipliers $(\log \nu - 1)$ (for the sum constraint (52)) and $\log \boldsymbol{\mu}$ (for the stationarity constraints (53)), with ν and all μ_i positive, we can find the optimal $\boldsymbol{B}^\star(\boldsymbol{q})$ from the stationary points of the function

$$\begin{aligned} f_{\alpha,\boldsymbol{\mu},\nu}(\boldsymbol{B}, \boldsymbol{q}) = \sum_{ij} b_{ij} \log b_{ij} - \sum_{ij} b_{ij} \log q_i - \alpha \sum_{ij} b_{ij} \log \widetilde{a}_{ij} \\ + (\log \nu - 1) \sum_{ij} b_{ij} - (\log \nu - 1) + \sum_i \log \mu_i \sum_j (b_{ij} - b_{ji}). \end{aligned} \tag{59}$$

The stationary points satisfy $\nabla_{\boldsymbol{B}} f_{\alpha,\boldsymbol{\mu},\nu}(\boldsymbol{B}^\star, \boldsymbol{q}) = \boldsymbol{0}$, giving

$$\frac{\partial}{\partial b_{ij}} f_{\alpha,\boldsymbol{\mu},\nu}(\boldsymbol{B}^\star, \boldsymbol{q}) = 0 \tag{60}$$

$$\log \frac{b^\star_{ij}}{q_i} + 1 - \alpha \log \widetilde{a}_{ij} + \log \nu - 1 + \log \mu_i - \log \mu_j = 0 \tag{61}$$

$$\frac{b^\star_{ij}}{q_i} (\widetilde{a}_{ij})^{-\alpha} \nu \frac{\mu_i}{\mu_j} = 1 \tag{62}$$

$$b^\star_{ij}(\boldsymbol{q}) = \frac{1}{\nu} \frac{\mu_j}{\mu_i} q_i (\widetilde{a}_{ij})^\alpha. \tag{63}$$

This solution naturally satisfies nonnegativity, with ν being a simple normalization. The vector $\boldsymbol{\mu}$ is more difficult to compute, and will be derived later. For now, we simply assume $\boldsymbol{\mu}$ known.

The formulas for $\boldsymbol{q}^\star(\boldsymbol{B})$ and $\boldsymbol{B}^\star(\boldsymbol{q})$ can, given an initial guess $\boldsymbol{B}^{(0)}$, be used for alternating updates of $\boldsymbol{q}$ and $\boldsymbol{B}$ guaranteed to decrease $f_{\alpha,\boldsymbol{\mu},\nu}$ unless a fixed point or a cycle is reached. At a fixed point of these iterations we have $q^\star_i = \sum_j b^\star_{ij}$. Slotting this into (63) yields

$$a^\star_{ij} = \frac{b^\star_{ij}}{\sum_{j'} b^\star_{ij'}} = \frac{1}{\nu} \frac{\mu_j}{\mu_i} (\widetilde{a}_{ij})^\alpha. \tag{64}$$

This can be written on matrix form as

$$\boldsymbol{A}^\star = \frac{1}{\nu} (\operatorname{diag} \boldsymbol{\mu})^{-1} \widetilde{\boldsymbol{A}}^{(\alpha)} (\operatorname{diag} \boldsymbol{\mu}) \tag{65}$$

where $\widetilde{\boldsymbol{A}}^{(\alpha)}$ denotes *Hadamard power* α of $\widetilde{\boldsymbol{A}}$, the matrix resulting from taking all entries of $\widetilde{\boldsymbol{A}}$ to the power α, i.e., $(\widetilde{\boldsymbol{A}}^{(\alpha)})_{ij} = (\widetilde{a}_{ij})^\alpha$.

To solve for $\boldsymbol{\mu}$ and ν we use the fact that the optimal transition matrix $\boldsymbol{A}$ must satisfy the standard relation $\boldsymbol{A}\mathbf{1} = \mathbf{1}$ in order to properly define a Markov chain X. Hence

$$\frac{1}{\nu} \left(\operatorname{diag} \boldsymbol{\mu}\right)^{-1} \widetilde{\boldsymbol{A}}^{(\alpha)} \left(\operatorname{diag} \boldsymbol{\mu}\right) \mathbf{1} = \mathbf{1} \tag{66}$$

$$\widetilde{\boldsymbol{A}}^{(\alpha)} \boldsymbol{\mu} = \nu \boldsymbol{\mu}, \tag{67}$$

so $\boldsymbol{\mu}$ is a right eigenvector of $\widetilde{\boldsymbol{A}}^{(\alpha)}$ associated with the eigenvalue ν. In addition, we have the requirement that ν and μ_i are all positive, from before; this guarantees that all required inverses exist.

Since the original Markov chain $\widetilde{X}$ is ergodic, $\widetilde{\boldsymbol{A}}$ and hence $\widetilde{\boldsymbol{A}}^{(\alpha)}$ are irreducible. As the matrix elements are nonnegative, aperiodicity of $\widetilde{\boldsymbol{A}}$ also establishes that $\widetilde{\boldsymbol{A}}^{(\alpha)}$ is aperiodic. The Perron-Frobenius theorem then ensures that there is a unique eigenvector of $\widetilde{\boldsymbol{A}}^{(\alpha)}$ with only real, positive elements, and that it is associated with the greatest eigenvalue, which is also real. This establishes that the unique optimal solution to the original MERS problem is the Markov chain X with transition matrix

$$\boldsymbol{A} = \frac{1}{\nu} \left(\operatorname{diag} \boldsymbol{\mu}\right)^{-1} \widetilde{\boldsymbol{A}}^{(\alpha)} \left(\operatorname{diag} \boldsymbol{\mu}\right) \tag{68}$$

where ν is the leading eigenvalue of the Hadamard power $\widetilde{\boldsymbol{A}}^{(\alpha)}$ and $\boldsymbol{\mu}$ is the corresponding right eigenvector with only positive entries. We see that exponent values $\alpha \in (1, \infty)$ recover all solutions between the original $\widetilde{X}$ and complete predictability.

As an aside, Equation (68) shows that $\widetilde{\boldsymbol{A}}^{(\alpha)}$ and $\boldsymbol{A}$ have the same eigenvalues and eigenvectors, except for a scaling by ν and $\operatorname{diag} \boldsymbol{\mu}$.

C Supplemental Material for Markov Chain Experiments

This section presents, in Table 1 on page 8, excerpts from Swedish text data, along with sampled output from word-level Markov models based on this material. Errors, shown in red with wavy underline, are seen to decrease in frequency in the output of models simplified with MERS. Table 2 in Section 6.4 of the main article was derived by word-for-word translation of these Swedish text examples. Please consult the main article for more information.

References

[1] S. Ihara, *Information Theory for Continuous Systems*. World Scientific Publishing Company, 1993.
[2] J. P. Burg, *A New Analysis Technique for Time Series Data*, pp. 42–48. IEEE Press, 1978.
[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.
[4] R. E. Blahut, "Computation of channel capacity and rate distortion function," *IEEE Trans. Inf. Theory*, vol. IT-18, pp. 460–473, 1972.
[5] 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.

ett bad [paus] titta mamman [paus] ta en sko [paus] det är pappan [paus] titta olov [paus] ta ett bad [paus] var är olov nu [paus] ta ett bad [paus] var är olov nu [paus] det är en bil [paus] ta en sko [paus] ta en bil [paus] det är en bil [paus] titta bilen [paus] titta skon [paus] det är en bok [paus] ta ett bad [paus] var är badet nu [paus] titta bilen [paus] titta olov [paus] var är bilen nu [paus] titta skon [paus] ta en bil [paus] hej olov [paus] hej olov [paus] var är olov

(a) Sample from error-free corpus.

pappan nu hej [paus] olov [paus] [paus] hallå olov [paus] [paus] det är pappan [paus] titta ööh skon [paus] [paus] det är ööh olov [paus] [paus] hallå olov [paus] titta ööh boken [paus] [paus] var är olov nu ööh ta en bok [paus] ööh ööh hallå olov [paus] det är olov det är ööh en bil [paus] hallå olov [paus] ööh var är olov nu [paus] ööh var är olov nu [paus] hallå olov [paus] [paus] ta ett bad [paus] titta boken boken hej olov [paus] ööh titta mamman [paus] det är en bil [paus] ööh ta en bil hallå

(b) Sentences disturbed by random speech errors.

en bok [paus] hallå olov [paus] var är olov [paus] var är bilen nu [paus] hallå olov [paus] det är en bok [paus] det är olov [paus] titta pappan [paus] titta badet [paus] det är en bok [paus] var är skon nu [paus] ta en bil [paus] hej olov [paus] ta en sko [paus] det är en bok [paus] var är badet nu [paus] titta bilen [paus] titta bilen [paus] det är olov [paus] det är en bok [paus] ta en bil [paus] titta mamman [paus] hej olov [paus] ta en bil [paus] titta mamman [paus] titta skon [paus] titta

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

badet [paus] hej ööh olov [paus] ta en bok [paus] ta en sko en sko titta boken en bil [paus] det är pappan nu [paus] [paus] var är skon nu [paus] det är olov [paus] det är en sko [paus] hej olov [paus] ööh ta en sko [paus] var är boken nu [paus] det är en bil [paus] ööh ta ett bad ett bad [paus] titta bilen [paus] ööh titta bilen [paus] olov [paus] titta skon [paus] titta pappan var är ööh ett en bil [paus] titta olov det är ett bad [paus] var är är en bil [paus] [paus]

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

hej olov [paus] hallå olov [paus] ööh ta en bil [paus] det är olov nu [paus] ta en bok [paus] det är pappan nu [paus] ta en en bok [paus] ta en bil [paus] det är en sko [paus] var är ööh en bil [paus] hej olov [paus] det är en bil [paus] ta ett en sko [paus] det är en bil [paus] ta en bil [paus] ta ett bad [paus] det är en bok [paus] ööh hej olov [paus] var är olov nu [paus] det är ett bad [paus] ööh ta en bok var är skon nu [paus] det

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

är ett bad [paus] titta mamman [paus] det är en bil [paus] det är [paus] pappan nu [paus] det är olov nu [paus] hej olov [paus] hallå olov [paus] var är badet nu [paus] det är mamman nu [paus] ta en sko [paus] [paus] ta en bok [paus] hej olov [paus] det är mamman [paus] ööh det är ett bad [paus] ta ett bad [paus] hej olov [paus] ööh ta en bil [paus] hallå olov [paus] titta pappan [paus] ta en bok [paus] titta boken [paus] titta badet [paus] det är mamman nu [paus] [paus] hej olov [paus] titta boken

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

var är badet nu [paus] var är skon nu [paus] var är boken nu [paus] det är en bil [paus] det är en bil [paus] var är badet nu [paus] var är skon nu [paus] det är en sko [paus] det är en sko [paus] det är en bok [paus] var är bilen nu [paus] var är badet nu [paus] det är en bok [paus] det är ett bad [paus] var är boken nu [paus] var är badet nu [paus] det är ett bad [paus] det är ett bad [paus] det är ett bad [paus] det är en bok [paus]

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

Table 1

Excerpts from the clean and disturbed data, along with 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).