Stochastic Music: From Markov Chains to Transformers

Nicholas G. Polson
Booth School of Business
University of Chicago
ngp@chicagobooth.edu

August 10, 2026

Abstract

In November 1949 Bell Telephone Laboratories circulated Technical Memorandum MM-49-150-29, *Composing Music by a Stochastic Process*, by John R. Pierce and Mary E. (Betty) Shannon. The memorandum generates tonal music from a finite state Markov chain with a hand elicited transition table and hard rules imposed on top. It appears one year after Claude Shannon's Markov approximations to English and seven years before the Illiac Suite. We review what the historical record supports, develop the mathematics of the construction, and then argue that every structural element of the 1949 scheme survives in an autoregressive transformer: the chain becomes an adaptive order kernel smoother, the rule set becomes exponential tilting through temperature, truncation and preference optimisation, and the elicited table becomes an estimated conditional with parameter sharing. The change is one of function class and estimation, not of generative principle. We give a worked three chord example, a Bayesian treatment of the transition table, a free energy identity that calibrates the strength of a rule set, the Diaconis and Freedman theorem as the representation result behind in context learning, a reading of Bowie's Verbasizer as the zeroth order case in which the artist supplies the tilt by hand, a survey of machine learning in music, and an honest account of what the comparison does not deliver.

Keywords: Algorithmic composition; Attention; Bayesian predictive; De Finetti; Entropy rate; Exponential tilting; Generative models; In context learning; Markov chains; Transformers.

1 Introduction

Generative modelling of sequences has a clean origin. Shannon (1948) fitted approximations to English of order zero, one and two by drawing letters and words from frequency tables, and printed the sample paths. The demonstration was not about language. It was that a source carrying the right redundancy produces output a reader takes for almost the real thing, and that the resemblance is a number.

Within a year the device reached music. Pierce and Shannon (1949) put a stochastic process on a musical alphabet and generated pieces from it. No computer was involved. The transition probabilities were written down by hand, the randomness came from tables, and the output was notated and played. It is among the earliest documented experiments in algorithmic composition, and among the earliest research reports at Bell Labs to carry a woman's name.

We have three aims. The first is to separate what the record establishes about the memorandum and its author from what has accumulated around it. The second is to work out the mathematics, since the memorandum is usually disposed of in a sentence and the sentence hides the structure worth having. The third, from Section

12 onward, is to make the correspondence with modern autoregressive models exact where it can be made exact and to say plainly where it cannot.

We do not claim that a transformer is a Markov chain. We claim something narrower and more useful. The 1949 memorandum makes three decisions, choosing a state space, choosing a conditional law, and imposing constraints on the output, and those are the three decisions still being made. Stating the modern machinery in the old vocabulary shows how much of it is substitution and locates the one place where the substitution buys something a table never could.

1.1 Betty Shannon and the memorandum

Mary Elizabeth Moore (1922 to 2017) read mathematics at the New Jersey College for Women, graduated Phi Beta Kappa, and joined Bell Laboratories as a human computer, later a Technical Assistant in the mathematics department. She worked on microwave and radar problems. She met Claude Shannon at the Labs in 1948, married him in 1949, and left Bell in 1951.

The standard accounts make her Claude Shannon's closest technical collaborator. She assisted in building Theseus, the maze solving mouse, and the wearable roulette computer built with Ed Thorp. The item of interest here is different, because it carries her own name. The IEEE Information Theory Society obituary and the Soni and Goodman profile both single out the memorandum as an unusual credit for a woman on a research report in 1949.

The bibliographic record is: J. R. Pierce and M. E. Shannon, *Composing Music by a Stochastic Process*, Bell Telephone Laboratories Technical Memorandum MM-49-150-29, 15 November 1949. Bell Labs memoranda were internal documents. This one has not been widely digitised, and the substantive public discussion is Yu and Varshney (2017).

Pierce popularised the exercise the following year in *Science for Art's Sake*, written under the pseudonym J. J. Coupling for *Astounding Science Fiction*, November 1950. He returned to the theme for four decades, in *Symbols, Signals and Noise* (1961), in *The Computer as a Musical Instrument* (1960), and through the Bell Labs computer music program with Max Mathews. The 1949 memorandum is the head of that line.

2 The 1949 construction

Let $\mathcal{S} = \{s_1, \dots, s_K\}$ be a finite musical alphabet. In the harmonic reading the states are chords, possibly with inversion and voicing attached. Let $P = (p_{ij})$ be a stochastic matrix,

$$p_{ij} \;=\; \Pr(X_{t+1} = s_j \mid X_t = s_i), \qquad p_{ij} \geq 0, \quad \sum_j p_{ij} = 1.$$

A composition of length n is a sample path $X_{1:n} = (X_1, \dots, X_n)$ with law

$$q(x_{1:n}) \;=\; \mu(x_1) \prod_{t=2}^{n} p_{x_{t-1} x_t},$$

where μ is the initial distribution. Sampling requires only a table and a source of uniform variates, which in 1949 meant random number tables and, according to secondary accounts we cannot verify, dice.

Three features of the design are worth isolating, since everything below returns to them.

(D1) The state space is a modelling choice, not a given. Music is not symbolic. Choosing chords as states fixes a resolution: harmonic motion is representable, voice leading within a chord is not, and rhythm

has to be smuggled in by attaching duration to the state or by fixing it externally. Every subsequent system, including every modern one, makes this choice first, and most of the apparent disagreement between systems is disagreement about $\mathcal{S}$.

(D2) The transition table was elicited, not estimated. The entries encode functional harmony: dominant resolves to tonic with high probability, tonic moves freely, and so on. The chain is therefore a prior over pieces. No corpus is fitted. It is the opposite pole from current practice, and the cleanest surviving example of a generative model built entirely from theory.

(D3) Hard constraints sit on top of the chain. Voice leading rules such as prohibition of parallel fifths cannot be expressed in a first order table over chord symbols, so they are enforced by rejecting or repairing paths. Generation is therefore not sampling from q but from q conditioned on an admissible set. Section 6 treats this properly.

3 Chain theory: stationarity, mixing, and what a first order table can hold

3.1 Stationary law

If P is irreducible and aperiodic on the finite set $\mathcal{S}$ there is a unique π with $\pi^\top P = \pi^\top$, and

$$\|\mu^\top P^n - \pi^\top\|_{TV} \;\le\; C\,|\lambda_2|^{\,n},$$

where λ_2 is the eigenvalue of P of second largest modulus. Setting P chooses two things at once: the long run harmonic profile π, which is the left Perron eigenvector, and the local motion, which is the rows. Empirical chord frequencies in a long generated piece converge to π almost surely by the ergodic theorem.

3.2 A worked example

Take $\mathcal{S} = \{\mathrm{I}, \mathrm{IV}, \mathrm{V}\}$ and

$$P \;=\; \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.2 & 0.5 \\ 0.7 & 0.1 & 0.2 \end{pmatrix},$$

with rows and columns ordered I, IV, V. The dominant resolves to the tonic with probability 0.7 and the subdominant prefers the dominant, which is the elicitation in (D2) in miniature. Solving $\pi^\top P = \pi^\top$ gives

$$\pi \;=\; (0.472,\; 0.216,\; 0.312).$$

The row entropies are

$$H_{\mathrm{I}} = 1.571, \qquad H_{\mathrm{IV}} = 1.485, \qquad H_{\mathrm{V}} = 1.157 \quad \text{bits},$$

so the dominant is the most predictable state, as intended. The eigenvalues other than one are $\lambda_{2,3} = -0.1 \pm 0.2i$, with $|\lambda_2| = 0.224$. The chain forgets its initial condition in under one step of relaxation time. This is the central weakness of the model stated numerically: a first order chain with a substantial spectral gap has no memory, so it cannot produce a piece that returns to material stated thirty bars earlier. Complex eigenvalues give a faint oscillation in the decay, which is the closest a chain of this size comes to periodic structure.

4 Entropy rate, redundancy, and the size of a style

4.1 Entropy rate and typical sets

The entropy rate of the stationary chain is

$$H(P) \;=\; \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \;=\; -\sum_{i=1}^{K} \pi_i \sum_{j=1}^{K} p_{ij} \log_2 p_{ij} \quad \text{bits per symbol.}$$

In the example above $H(P) = 1.423$ bits against a maximum of $\log_2 3 = 1.585$, so the redundancy is

$$R \;=\; 1 - \frac{H(P)}{\log_2 K} \;=\; 0.102.$$

By the Shannon, McMillan and Breiman theorem, $-n^{-1} \log_2 q(X_{1:n}) \to H(P)$ almost surely, so the number of typical paths of length n is $2^{nH(P)}$ to first order in the exponent. A style is a set with a measurable size. For the example, a phrase of thirty two chords has roughly $2^{45.5} \approx 5 \times 10^{13}$ typical realisations. The model is not memorising and it is not enumerable. It is a probability measure with a computable log cardinality, and that is the object the 1949 experiment put on the table.

4.2 The redundancy dial

The design claim implicit in the memorandum, and stated explicitly by Pierce later, is that listenable output requires an interior value of R. At $R = 1$ the piece is a fixed loop. At $R = 0$ it is uniform noise on the alphabet. Tonality is redundancy. The interest of the construction is that redundancy became a dial rather than a description.

Brooks, Hopkins, Neumann and Wright (1957) exhibited the same tradeoff empirically by fitting chains of order one through eight to a corpus of hymn tunes. Low order gives incoherence. High order gives plagiarism, because the conditional collapses onto the single continuation observed in the corpus. This is the tradeoff between bias and variance for sequence models, discovered in music before it was named, and Section 8 shows that it is the problem transformers actually solve.

4.3 Perplexity

Modern practice reports perplexity, $\mathrm{PPL} = 2^{\widehat{H}}$ where $\widehat{H}$ is the empirical cross entropy rate of the model on held out data. It is the same quantity in exponentiated form. Cross entropy exceeds entropy by the Kullback Leibler divergence,

$$\widehat{H} \;=\; H(p_{\mathrm{true}}) + \mathrm{KL}(p_{\mathrm{true}} \| p_\theta),$$

so minimising perplexity is minimising divergence from the source, and the minimum attainable value is not zero but $H(p_{\mathrm{true}})$. Reported perplexities are therefore uninterpretable in isolation: the number confounds the irreducible entropy of the source with the deficiency of the model, and the two can only be separated by comparison on a fixed corpus and a fixed alphabet.

Log loss is the natural scale for a further reason. It is a strictly proper scoring rule, so the forecaster's expected score is optimised uniquely at the true predictive, and no hedging strategy improves it. It is also the only proper local rule on a discrete alphabet, in the sense that it depends on the forecast through the probability assigned to the outcome that occurred and nothing else. Whatever one believes about music, the quantity being minimised in training has an unambiguous decision theoretic meaning.

What it does not have is any connection to musical quality. A model that attains the entropy rate of the source is indistinguishable from it in the compression sense and may still be unlistenable, since compression rewards hedging in exactly the places where a composer must commit. That gap is the subject of Section 16.

5 Prediction, gambling, and what redundancy is worth

Redundancy is not merely an aesthetic quantity. It is the exploitable part of a source, and the exchange rate is exact.

5.1 Guessing, betting, and the doubling rate

Shannon (1951) estimated the entropy of English by having a subject guess the next letter, recording the number of guesses required, and bounding the entropy rate from the distribution of that count. The device turns an unobservable rate into a prediction experiment. It is the ancestor of held out perplexity. Cover and King (1978) refined it into a gambling estimate: a subject sequentially bets a fraction of capital on each candidate next symbol, and the wealth attained bounds the entropy.

Consider a fair odds market on the next symbol of the chain, where a stake on s_j returns K times the stake if $X_{t+1} = s_j$ and nothing otherwise. A gambler who reinvests all capital, placing fraction b_j on s_j, has log optimal proportions $b_j = p_{ij}$ given the current state s_i. This is the Kelly (1956) rule, and the resulting asymptotic growth rate of capital is

$$W \;=\; \lim_{n\to\infty} \frac{1}{n} \log_2 \frac{C_n}{C_0} \;=\; \log_2 K - H(P) \;=\; R \log_2 K, \tag{1}$$

almost surely, where R is the redundancy of Section 4. The growth rate available to a forecaster equals the redundancy of the source, measured in the same bits.

In the worked example of Section 2, $R = 0.102$ and $\log_2 K = 1.585$, so a gambler who knows P compounds at 0.162 bits per chord, roughly twelve percent growth per symbol. A listener is running the same computation without the stake. Expectation, surprise and resolution, the vocabulary in which music is usually discussed, are the phenomenology of a predictor with an edge, and (1) prices the edge.

5.2 Side information

If the gambler observes Y correlated with the next symbol, the increase in growth rate is exactly the mutual information,

$$\Delta W \;=\; I(X_{t+1}; Y \mid X_t),$$

which is the cleanest statement of what a conditioning signal is worth. Key signatures, metre, lyrics and a text prompt are all instances of Y, and the identity says that a conditioning variable earns its place only to the extent that it is informative about the continuation *given what is already known*. The conditioning is on X_t, so a signal that merely restates the current state is worth nothing however strongly it correlates with the output. Much conditioning in practice is redundant in this technical sense.

Two corollaries are worth stating. First, $\Delta W \le H(Y)$, so a coarse conditioning variable caps the achievable gain no matter how well the model uses it, which bounds what any amount of architecture can extract from a short text prompt. Second, the gain is additive over a chain of conditioning variables only when they are conditionally independent given the state, and otherwise $I(X_{t+1}; Y_1, Y_2 \mid X_t) \le I(X_{t+1}; Y_1 \mid X_t) + I(X_{t+1}; Y_2 \mid X_t)$ may fail in either direction. Synergy between conditioning signals is possible, and so is complete redundancy. Ablation studies that remove one signal at a time measure neither.

6 Constraints as exponential tilting

Let $A \subset \mathcal{S}^n$ be the admissible set defined by the rules in (D3). Rejection sampling produces the conditional law $q(\cdot \mid A)$. Softening the rules to a penalty V gives the Gibbs form

$$p(x_{1:n}) \;=\; \frac{1}{Z(\lambda)}\, q(x_{1:n})\, \exp\{-\lambda V(x_{1:n})\}, \qquad Z(\lambda) = \sum_{x_{1:n}} q(x_{1:n}) e^{-\lambda V(x_{1:n})}, \tag{2}$$

with $\lambda \to \infty$ recovering the hard constraint. Two facts make (2) the right way to read the 1949 recipe.

Proposition 1. *If $V(x_{1:n}) = \sum_{t=2}^{n} v(x_{t-1}, x_t)$ is additive along the path, the tilted law is again a Markov chain. Let $M = (p_{ij} e^{-\lambda v(i,j)})$, let ρ be its Perron root and r the corresponding right eigenvector. The tilted transition matrix is*

$$\tilde{p}_{ij} \;=\; \frac{p_{ij} e^{-\lambda v(i,j)} r_j}{\rho\, r_i}.$$

The proof is the standard Perron Frobenius change of measure. The consequence is that soft voice leading rules cost nothing: they are absorbed into a new table. Hard rules that are not additive along the path, which is the interesting case, are not absorbable, and generate and test is then the honest algorithm. Hiller and Isaacson (1957) built the Illiac Suite on exactly this principle, which is rejection sampling against counterpoint rules.

Proposition 2. *The tilted law* (2) *solves*

$$p \;=\; \arg\min_{p} \Big\{ \lambda\, \mathbb{E}_p[V] + \mathrm{KL}(p \,\|\, q) \Big\}.$$

The two propositions combine into a statement that fixes the calibration of λ. Write $\Lambda(\lambda) = \log \rho(\lambda)$ for the log Perron root of $M(\lambda) = (p_{ij} e^{-\lambda v(i,j)})$, and let $\tilde{\pi}$ be the stationary law of the tilted chain.

Proposition 3. *For additive V the per symbol Kullback Leibler rate between the tilted and the original chain is*

$$\mathcal{D}(\lambda) \;=\; \lim_{n\to\infty} \frac{1}{n} \mathrm{KL}(\tilde{p}_{1:n} \,\|\, q_{1:n}) \;=\; -\lambda\, \mathbb{E}_{\tilde{\pi}}[v] - \Lambda(\lambda),$$

so that

$$\lambda\, \mathbb{E}_{\tilde{\pi}}[v] + \mathcal{D}(\lambda) \;=\; -\Lambda(\lambda), \qquad \Lambda'(\lambda) \;=\; -\,\mathbb{E}_{\tilde{\pi}}[v].$$

Proof. From the tilted kernel, $\log(\tilde{p}_{ij}/p_{ij}) = -\lambda v(i,j) + \log r_j - \log r_i - \Lambda(\lambda)$. Taking expectations under $\tilde{\pi}$ and the tilted kernel, the telescoping term $\mathbb{E}[\log r_{X_{t+1}} - \log r_{X_t}]$ vanishes by stationarity, which gives the first display. The second is the variational identity of Proposition 2 evaluated at the optimum. The third follows from differentiating Λ and using the normalisation of the Perron eigenvector. □

Three consequences. The optimal value of the constrained problem is the free energy $-\Lambda(\lambda)$, so the whole family is indexed by a single convex function. Convexity of Λ makes $\mathbb{E}_{\tilde{\pi}}[v]$ monotone decreasing in λ, so the rule strength is a genuine dial with no reversals. And by Gartner and Ellis the empirical violation rate $n^{-1} V(X_{1:n})$ satisfies a large deviation principle under q with rate function

$$I(v) \;=\; \sup_{\lambda} \{-\lambda v - \Lambda(\lambda)\},$$

so a target violation rate v is achieved by $\lambda = -I'(v)$. A composer who wants parallel fifths in one bar per thousand can solve for λ rather than guess it, and the same calculation prices the tilt in Section 12.

This variational identity is the bridge to Section 12. Preference optimisation of a language model maximises expected reward subject to a KL penalty against a reference model, and the optimum is $p^\star(x) \propto$

$p_{\mathrm{ref}}(x)\exp\{r(x)/\beta\}$. That is (2) with q the pretrained model and $-\lambda V$ the reward. The voice leading rule set of 1949 and the reward model of today occupy the same slot in the same formula.

7 The Bayesian version the memorandum did not use

7.1 Dirichlet multinomial

Given a corpus with transition counts n_{ij}, place independent priors $P_{i\cdot} \sim \mathrm{Dir}(\alpha_{i1},\dots,\alpha_{iK})$ on the rows. Rows are a posteriori independent with $P_{i\cdot} \mid \text{data} \sim \mathrm{Dir}(\alpha_{ij}+n_{ij})$, and the one step predictive is

$$\Pr(X_{t+1}=s_j \mid X_t=s_i,\text{data}) \;=\; \frac{n_{ij}+\alpha_{ij}}{n_{i\cdot}+\alpha_{i\cdot}}, \tag{3}$$

which is additive smoothing. The 1949 table is the limit $n_{ij}=0$, so the generated music is a draw from the prior predictive. Every subsequent system is a point on the line between (3) with α dominant and (3) with n dominant.

7.2 Sparsity and the horseshoe

Functional harmony is sparse. Most chord pairs never occur, and the sparsity pattern carries more of the style than the surviving numerical values do. Write $p_{ij} \propto \exp(\eta_{ij})$ and put a global and local prior on the logits,

$$\eta_{ij} \mid \lambda_{ij},\tau \sim N(0,\tau^2\lambda_{ij}^2), \qquad \lambda_{ij}\sim C^+(0,1), \qquad \tau\sim C^+(0,1),$$

the horseshoe. The half Cauchy tail on λ_{ij} leaves genuinely large logits essentially unshrunk, while the infinite spike at the origin pulls the rest to zero. In the Gaussian mean case the posterior mean is $\mathbb{E}[(1-\kappa_{ij}) \mid \text{data}]\,\hat\eta_{ij}$ with shrinkage weight $\kappa_{ij}=1/(1+\tau^2\lambda_{ij}^2)$, and the horseshoe places a $\mathrm{Beta}(1/2,1/2)$ prior on κ, mass at both endpoints and little in between. The prior is deciding which transitions exist, not merely regularising their sizes, and τ sets the expected number that survive.

That is exactly what the 1949 elicitation did by hand. A harmony textbook is a statement about the support of P, delivered with infinite confidence and no data. The horseshoe is the same statement made from a corpus and made revisable. Seen this way the two traditions of Section 10 differ in the strength of a prior, not in kind, and the choice between them is a question about how much one trusts the textbook relative to the sample.

7.3 Hierarchical smoothing and backoff

For higher order chains the counts $n_{i_1\dots i_k,j}$ are sparse for even moderate k. The classical fix is backoff to lower orders. The Bayesian version is a hierarchical Pitman Yor process on the context tree, in which the predictive of order k is centred on the predictive of order $k-1$,

$$G_{\mathbf{u}} \mid G_{\sigma(\mathbf{u})} \sim \mathrm{PY}(d_{|\mathbf{u}|},\theta_{|\mathbf{u}|},G_{\sigma(\mathbf{u})}),$$

where $\sigma(\mathbf{u})$ drops the oldest symbol of the context $\mathbf{u}$. The resulting predictive is interpolated Kneser Ney smoothing, which was the strongest method based on n grams for two decades. The point for Section 8 is that this is a hierarchical shrinkage solution to the curse of order, and a transformer is a different solution to the same problem.

8 The curse of order and how it is broken

A chain of order k on an alphabet of size K has $K^k(K-1)$ free parameters. For a modest symbolic music alphabet, $K = 10^2$ and $k = 8$ gives 10^{16} cells and no corpus will ever fill them. The Brooks et al. finding follows immediately: at high order almost every observed context has count one, the predictive is a point mass on the single observed continuation, and the model copies.

There are three ways out, and they are the history of the field.

1. *Shrinkage across orders.* Backoff and hierarchical Pitman Yor, Section 7. The context tree is pruned adaptively; variable order chains and context tree weighting are the same idea with an explicit model average.
2. *Structural restriction.* Impose that the conditional depends on the context only through a low dimensional summary. Hidden Markov models take the summary to be a latent state with its own dynamics, and the filtering recursion $\alpha_t(z) \propto p(x_t \mid z) \sum_{z'} \alpha_{t-1}(z') p(z \mid z')$ carries all the history that matters.
3. *Parameter sharing in a smooth function class.* Represent the context by a vector $h_t \in \mathbb{R}^d$ and set $p_\theta(x_t \mid x_{1:t-1}) = \mathrm{softmax}(W h_t)$ with $h_t = f_\theta(x_{1:t-1})$. The number of parameters is now governed by d and the architecture, not by K^k. Similar contexts share statistical strength through the geometry of the embedding rather than through an explicit tree.

Route three is the transformer. The essential statistical content is that a table lookup indexed by an exact context has been replaced by a smooth function of a learned representation of the context, which is a regularisation device.

8.1 Choosing the order

The order is a model index and can be treated as one. With independent Dirichlet priors on the rows of an chain of order k, the marginal likelihood is available in closed form,

$$p(x_{1:n} \mid k) \;=\; \prod_{\mathbf{u} \in \mathcal{S}^k} \frac{B(\boldsymbol{\alpha}_\mathbf{u} + \mathbf{n}_\mathbf{u})}{B(\boldsymbol{\alpha}_\mathbf{u})}, \qquad B(\mathbf{a}) = \frac{\prod_j \Gamma(a_j)}{\Gamma(\sum_j a_j)},$$

where $\mathbf{n}_\mathbf{u}$ counts continuations of the context $\mathbf{u}$. Posterior model probabilities follow immediately, and the expansion for large n is

$$-2 \log p(x_{1:n} \mid k) \;=\; -2 \log p(x_{1:n} \mid \hat{P}_k) + K^k(K-1) \log n + O(1),$$

the Schwarz penalty with the dimension count of Section 8. The penalty grows geometrically in k while the fit improves at best linearly in n, so the posterior concentrates on small k for any corpus a musicologist will ever assemble. This is the Brooks et al. finding derived rather than observed.

The lesson generalises. A model class whose dimension grows geometrically with the resolution of the conditioning cannot be selected into by data. Either the dimension must be controlled by shrinkage, as in the hierarchical construction above, or the conditioning must pass through a low dimensional learned summary, which is route three. Selection between fixed tables is not a live option, and no amount of data makes it one.

9 Rhythm, or what the chord alphabet leaves out

Design choice (D1) buys harmonic motion at the cost of time. A chain over chord symbols has no clock. The 1949 scheme handles this by fixing durations externally or by folding duration into the state, which multiplies K and worsens the sparsity of Section 8.

The alternative is to model onsets as a point process on $[0, T]$ and marks as the pitches attached to them. Let $N(dt \times dy)$ be a marked point process with conditional intensity $\lambda(t \mid \mathcal{F}_{t-})$ relative to the history. A homogeneous Poisson intensity gives no metre. A self exciting intensity

$$\lambda(t \mid \mathcal{F}_{t-}) \;=\; \lambda_0 \;+\; \sum_{t_i < t} g(t - t_i)$$

with g concentrated near the beat produces clustering at metrical positions, and a shot noise intensity $\lambda(t) = \sum_i a_i g(t - \tau_i)$ driven by a latent subordinator gives bar level and phrase level bursts. The stationary autocovariance of a shot noise process is $\mathrm{Cov}(\lambda(0), \lambda(h)) \propto \int g(u) g(u + h)\, du$, so the kernel g is the object that encodes metrical structure, and the correlation decays at whatever rate g dictates rather than exponentially.

This matters for the comparison to follow. Symbolic transformers discretise time into a token grid and recover metre by learning positional regularities, which works but puts a representational choice where a modelling choice belongs. The continuous time formulation is the honest one for rhythm, and it is not what either 1949 or current practice does.

10 The line of descent

The memorandum sits at the head of a short and clearly marked lineage.

- Olson and Belar at RCA carried out statistical analysis of melody around 1950 and built an electronic composing machine driven by transition frequencies estimated from a small corpus. This is the estimated counterpart to the elicited table of (D2).
- Brooks, Hopkins, Neumann and Wright (1957) fitted chains of order one through eight to hymn tunes and reported the order tradeoff directly. It is the first clean statement of the tradeoff between bias and variance for sequence models.
- Hiller and Isaacson produced the Illiac Suite in 1956 to 57 on the ILLIAC, the first substantial score generated by a computer, using generate and test against counterpoint rules. That is rejection sampling from a constrained set, Section 6 with $\lambda = \infty$.
- Xenakis formalised stochastic composition in *Formalized Music* (1963), using Markov chains in Analogique A and B and Poisson and Gaussian mechanisms elsewhere, and treating the sound mass rather than the note as the object of distributional assumption.
- Pierce and Mathews built the Bell Labs computer music program through the 1960s, producing the MUSIC languages and eventually, at Stanford, the Bohlen Pierce scale. The memorandum is the direct ancestor of that program.

Two features of the lineage are worth noting. Estimation from a corpus appears immediately, within a year, so the theory driven and data driven approaches are contemporaries rather than successive eras. And the constraint mechanism is present from the start, which is why the alignment problem of Section 12 has an unbroken history rather than a recent one.

11 Cut ups, and the Verbasizer

A parallel tradition ran outside the laboratory and reached the same structure by a different route. Burroughs and Gysin cut printed text into fragments and reassembled them at random, and David Bowie used the technique on lyrics from the mid 1970s, physically cutting up newspaper and manuscript, drawing fragments, and discarding what did not work. In 1995, for the album *Outside*, Bowie and the programmer Ty Roberts built a Macintosh application called the Verbasizer that automated the procedure: words and sentences were entered in columns, columns could be restricted by part of speech and weighted, and the program returned randomised recombinations.

The Verbasizer deserves a paragraph of mathematics, because it is the degenerate case of everything above and the degeneracy is the instructive part. Let a line have m slots with a syntactic template fixing the part of speech at each. The generator draws independently,

$$p(x_{1:m}) \;=\; \prod_{j=1}^{m} w_j(x_j),$$

a product measure. There is no transition matrix, and no dependence at all: the mutual information $I(X_i; X_j)$ is zero for every pair, and the entropy rate is $m^{-1}\sum_j H(w_j)$, the largest it can be given the marginals. This is the zeroth order model of Shannon (1948), with the columns supplying the part of speech constraint that a first order chain would otherwise have to learn.

The step that matters comes next. Bowie kept some lines and crossed out others. Selection is conditioning on an acceptance set A, and the accepted law is $p(\cdot \mid A)$, which is not a product measure. Conditioning on a function of several coordinates induces dependence among coordinates that were independent, the collider mechanism. The apparatus supplies maximum entropy and no structure, and the artist supplies the tilt of (2) by hand, at a value of λ that is never written down. The generative model contributes variance; the human contributes the constraint. Bowie was explicit that the output was raw material rather than product, and that the point was to be given something unexpected to respond to.

Two remarks follow. First, the division of labour between an unconstrained sampler and a selecting human is the oldest interaction pattern in generative art, and it is exactly what a modern system internalises when the reward model of Section 12 replaces the artist's pencil. Second, a browser reimplementation of the Verbasizer now exposes a temperature control, which is Proposition 5 arriving in a lyric tool by a route with no contact whatever with Bell Labs in 1949.

12 Transformers as adaptive order stochastic composers

12.1 The factorisation is unchanged

An autoregressive transformer defines

$$p_\theta(x_{1:n}) \;=\; \prod_{t=1}^{n} p_\theta(x_t \mid x_{1:t-1}), \qquad p_\theta(\cdot \mid x_{1:t-1}) = \mathrm{softmax}(W h_t).$$

This is the chain rule, not a new modelling idea. The 1949 model is the special case $h_t = e_{x_{t-1}}$, a one hot encoding of the previous symbol, and W the log of the transition table. Sampling is done the same way in both cases, symbol by symbol from a categorical law.

12.2 Attention is a kernel smoother over the past

Within a head, with queries $q_t = W_Q h_t$, keys $k_s = W_K h_s$ and values $v_s = W_V h_s$,

$$a_{ts} \;=\; \frac{\exp(q_t^\top k_s/\sqrt{d})}{\sum_{u\le t}\exp(q_t^\top k_u/\sqrt{d})}, \qquad \mathrm{Attn}(h)_t \;=\; \sum_{s\le t} a_{ts} v_s.$$

This is a Nadaraya Watson estimator with kernel $\mathcal{K}(q,k) = \exp(q^\top k/\sqrt{d})$ and bandwidth controlled by $\sqrt{d}$ and by the scale of W_Q, W_K. The weights are data dependent, so the effective neighbourhood is chosen at run time.

Equivalently, $a_{t\cdot}$ is a Gibbs measure over past positions with energy $-q_t^\top k_s$ and inverse temperature $1/\sqrt{d}$. Attention is therefore itself an instance of (2), applied to positions rather than to paths, with the uniform measure on $\{1,\dots,t\}$ as the base. Every stage of the modern pipeline is an exponential tilt of something: attention tilts over positions, the output softmax tilts over the alphabet, temperature retilts the output, and preference optimisation tilts the whole path measure.

The bandwidth is not a nuisance parameter. Define the effective context length at position t as the perplexity of the attention distribution,

$$L_t^{\mathrm{eff}} \;=\; \exp\Big\{-\sum_{s\le t} a_{ts}\log a_{ts}\Big\} \;\in\; [1,t]. \tag{4}$$

The two endpoints are exactly the two models under comparison. At $L_t^{\mathrm{eff}} = 1$ with the mass on $s = t-1$ the head is a first order chain and the 1949 memorandum is recovered as a special case. At $L_t^{\mathrm{eff}} = t$ with uniform weights the head is a bag of history summary with no order at all. A trained model sits between, at a value that varies with t and with the head.

Here is the sharpest single statement of the relationship to 1949. A fixed order chain uses a fixed window of the past with fixed weight. A transformer uses a weighted average over the whole visible past with weights that depend on the content of the past. It is an adaptive order, adaptive bandwidth chain. The effective order at position t can be read off from the entropy of the attention distribution $a_{t\cdot}$: concentrated attention on $s = t-1$ is a bigram model, diffuse attention is a long context model, and the model interpolates without being told which regime it is in. The hierarchical Pitman Yor construction of Section 7 chooses the order by shrinkage across a fixed tree. Attention chooses it by learned similarity in an embedded space.

12.3 Induction heads are in context estimators

A documented mechanism in trained transformers is the induction head: locate an earlier occurrence of the current suffix and copy what followed it. Written out, the contribution to the predictive is

$$\Pr(x_t = j \mid x_{1:t-1}) \;\propto\; \sum_{s<t} \mathbf{1}\{x_{s-1} = x_{t-1}\}\,\mathbf{1}\{x_s = j\},$$

which is the empirical transition frequency computed on the current context window. That is (3) with counts drawn from the prompt rather than from the training corpus, and with the pretrained weights supplying the α term. The model estimates a chain in context and mixes it with what it learned in training. The 1949 table has not vanished. It is being estimated afresh inside every forward pass.

12.4 De Finetti, Diaconis and Freedman

The Bayesian representation behind this is exact, and worth stating carefully. A sequence is *Markov exchangeable* if its law is invariant under permutations preserving transition counts and the initial state.

Theorem 4 (Diaconis and Freedman, 1980)**.** *A recurrent, Markov exchangeable process on a countable state space is a mixture of Markov chains: there is a measure ν on stochastic matrices with*

$$\Pr(X_{1:n} = x_{1:n}) \;=\; \int \mu(x_1) \prod_{t=2}^{n} p_{x_{t-1}x_t} \, \nu(dP).$$

This is the Markov analogue of de Finetti's theorem and it is the representation result behind in context learning. A model whose predictive is $\Pr(x_{t+1} \mid x_{1:t}) = \int p_{x_t,x_{t+1}} \, \nu(dP \mid x_{1:t})$ updates a posterior over transition tables as the sequence proceeds. Empirically, transformers behave this way on synthetic Markov data: the in context predictive tracks the Bayesian posterior predictive for the chain generating the prompt. The subjectivist reading is the useful one. There is no true P in the model. There is a predictive rule, and the mixture representation is a theorem about the rule, not a metaphysical claim about the source. Pierce and Shannon chose a single P by fiat. A transformer carries a posterior over P implicitly and updates it from the prompt.

12.5 Temperature and truncation are the tilting of Section 6

Sampling from a trained model is rarely sampling from p_θ. Temperature scaling uses

$$p_\tau(x_t \mid x_{1:t-1}) \;\propto\; p_\theta(x_t \mid x_{1:t-1})^{1/\tau},$$

which is (2) with $V = -\log p_\theta$ and $\lambda = 1/\tau - 1$. That this is the redundancy dial of Section 4 is a one line fact rather than an analogy.

Proposition 5. *Let $p_\beta(x) \propto p(x)^\beta$ with $\beta = 1/\tau$, and write $E(x) = -\log p(x)$. Then*

$$\frac{\partial}{\partial \beta} H(p_\beta) \;=\; -\,\beta \operatorname{Var}_{p_\beta}(E) \;\le\; 0,$$

so the entropy of the sampling distribution is nondecreasing in τ, strictly unless E is degenerate.

Proof. p_β is a one parameter exponential family with natural parameter $-\beta$ and sufficient statistic E. Its log partition function $\psi(\beta) = \log \sum_x e^{-\beta E(x)}$ satisfies $\psi'(\beta) = -\mathbb{E}_\beta[E]$ and $\psi''(\beta) = \operatorname{Var}_\beta(E)$. Since $H(p_\beta) = \psi(\beta) + \beta\, \mathbb{E}_\beta[E] = \psi(\beta) - \beta\psi'(\beta)$, differentiating gives $-\beta\psi''(\beta)$. □

The variance of the log probability is thus the sensitivity of style to the knob. A model whose next symbol log probabilities are nearly deterministic is unresponsive to temperature, which is the numerical signature of the collapse that Section 8 attributes to high order chains fitted on sparse counts. Top k and nucleus sampling restrict the support to a high probability set and renormalise, which is the hard constraint of (D3): admissible continuations only, everything else rejected. The practitioner's observation that low temperature gives repetitive output and high temperature gives incoherence is the Brooks et al. tradeoff of 1957, restated.

12.6 Alignment is a rule set

As noted after Proposition 2, KL regularised preference optimisation produces

$$p^\star(x) \;\propto\; p_{\mathrm{ref}}(x) \exp\{r(x)/\beta\}.$$

The reward r plays the role that prohibition of parallel fifths played in 1949. It is a constraint on admissible output that the base generative process does not express, imposed by tilting rather than by redesigning the process. The 1949 memorandum had a rule set written by music theorists. Modern systems have one learned from preference data. The algebra is identical, and so is the failure mode: tilt too hard and the tilted law collapses onto a small set, which in music is a loop and in language is mode collapse.

12.7 What is genuinely new

Three things do not appear in the 1949 picture at all.

First, the representation is learned. The alphabet in (D1) is fixed by hand, but the map from symbol sequences to h_t is estimated, so states that behave alike become neighbours in $\mathbb{R}^d$ without anyone declaring them similar. This is what breaks the curse of order in practice.

Second, the model is trained on a scale at which the distinction between prior and data disappears. The elicited table of (D2) is replaced by an estimate whose effective prior is architectural.

Third, long range structure is representable. A first order chain has a spectral gap and therefore an exponentially decaying memory, as the worked example showed. Direct attention to position s from position t gives a path of length one between them regardless of $|t - s|$, so correlation need not decay. Whether it should decay, and at what rate for music, is an open empirical question. Relative position encodings, introduced for music precisely because absolute position is the wrong invariance for metre and repetition, are an attempt to build the right decay into the kernel.

12.8 Symbolic and audio models

Two representation regimes descend from (D1). Symbolic models keep a discrete alphabet of note and timing events and are the direct heirs of the memorandum. Audio models replace the alphabet with a learned discrete codebook from a vector quantised autoencoder, then run the same autoregressive machinery over codes, or abandon the left to right factorisation entirely for a diffusion or flow model on a continuous latent. The last of these is the one genuine break with 1949: it is not a stochastic process running forward in musical time but a transport of noise to data in an artificial time. Even there the tilting structure of Section 6 returns as classifier free guidance, which is again an exponential reweighting of a base measure.

12.9 Summary of the correspondence

Table 1 states the mapping compactly. The left column is the 1949 memorandum, the right column is a modern autoregressive system, and the middle column is the mathematical object common to both.

The row that carries the real content is the third. Everything else is substitution of one implementation for another. Adaptivity of order is the property that a fixed table cannot have at any size, and it is what the smooth function class of Section 8 buys.

13 Form, and what neither model represents

Both the 1949 chain and a trained transformer are flat models of a sequence. Music is not flat. A sonata exposition returns transposed, a rondo alternates a refrain with episodes, a fugue states a subject and answers it at the fifth. These are hierarchical relations between segments, not statistical regularities between adjacent symbols, and the difference is measurable.

Pierce and Shannon, 1949	Common object	Transformer
Chord alphabet $\mathcal{S}$, fixed by hand	State space, choice (D1)	Learned tokenisation or VQ codebook
Transition table P, $K \times K$	Conditional $p(x_t \mid x_{<t})$	$\mathrm{softmax}(Wh_t)$, $h_t = f_\theta(x_{<t})$
Fixed order one	Effective order	Adaptive, set by attention entropy
Elicited from harmony texts	Source of parameters	Estimated from corpus; prior is architectural
Voice leading rules, rejection	Tilt $q\, e^{-\lambda V}/Z$	Reward model, guidance, nucleus truncation
Random number tables	Sampling	Ancestral sampling with temperature τ
Redundancy R chosen by design	Entropy rate H	τ, and the training objective
Exponential memory, $\lvert\lambda_2\rvert < 1$	Correlation decay	Direct paths, decay need not be exponential
Listening	Evaluation	Listening, plus perplexity

Table 1: The three design decisions of Section 2 and their modern counterparts.

13.1 Correlation decay

For a stationary chain with spectral gap $1 - |\lambda_2| > 0$ and any square integrable f, g,

$$|\mathrm{Cov}(f(X_0), g(X_h))| \;\le\; \|f\|_{2,\pi}\, \|g\|_{2,\pi}\, |\lambda_2|^{h},$$

so dependence dies geometrically. In the worked example $|\lambda_2| = 0.224$ and the correlation at lag ten is below 10^{-6}. No choice of the nine free entries of P produces a recapitulation. To slow the decay one must push $|\lambda_2|$ toward one, which by the same eigenvalue is to make the chain nearly reducible, and a nearly reducible chain does not modulate: it stays where it started. Long memory and harmonic mobility are in direct conflict within the model class. This is the precise sense in which a first order chain cannot represent form, and enlarging K does not help.

Attention removes the constraint by construction, since position t reaches position s in one step whatever $|t - s|$. That guarantees representability, not correctness. Whether trained models use the capacity is an empirical question, and the honest summary is that generated music holds local coherence over far longer spans than a chain but that long range returns are often approximate.

13.2 Grammar

The alternative tradition models form directly. Generative theories of tonal music posit a tree over the piece with reductions at each level, and probabilistic context free grammars make that tree a random object with rule probabilities in place of transition probabilities. The marginal law of the leaf sequence then has long range dependence by construction, because two distant leaves may be siblings in the tree at a high level.

The comparison is instructive. A grammar buys hierarchy and pays in tractability and in the difficulty of eliciting or estimating rule probabilities. A transformer buys hierarchy at no representational cost, pays nothing at inference, and offers no account of what the hierarchy is. Neither has displaced the other, and the honest position is that hierarchical structure in music remains a modelling problem rather than a solved one.

13.3 Segment level latent structure

An intermediate construction is a hierarchical model with a slow latent chain over sections and a fast conditional over symbols,

$$Z_m \sim \text{chain on sections}, \qquad X_{1:n_m} \mid Z_m = z \sim p_z(\cdot),$$

which is a hidden semi Markov model. The slow chain supplies form, the fast conditional supplies surface, and the two timescales are explicit rather than emergent. The stationary autocovariance is then a mixture of geometric terms with rates set by the two spectra, so the model can hold both fast harmonic motion and slow sectional return. That so simple a device is rarely used now says more about fashion than about its merits.

14 Machine learning in music

The field that grew out of the memorandum now divides into three problems, and it is useful to keep them apart because the statistical difficulty is different in each.

14.1 Discriminative tasks

Transcription, beat and downbeat tracking, key and chord recognition, instrument identification and source separation are supervised problems with a loss function and a ground truth. They are the easy case in the sense that Section 16 does not bite: one can hold out data and count errors. Deep networks displaced hand built spectral features in all of them over roughly a decade.

What remains difficult is not architectural. Annotated audio is expensive, so labels are scarce and the effective sample size is far smaller than the number of frames suggests, since frames within a recording are strongly dependent and the independent unit is the track. Standard splits frequently violate this, placing different excerpts of one performance on both sides of the partition, and the resulting optimism is large. Album and artist effects compound it: a model can score well by recognising the recording rather than the music, which is confounding of the plainest kind and is why artist conditional splits are now standard practice in careful work.

Note also that ground truth here is a human annotation, not a physical fact. Chord labels for the same passage disagree between annotators at rates that are a substantial fraction of the reported error rates of the best systems, so the achievable ceiling is set by annotator agreement rather than by anything intrinsic. Reported accuracy above that ceiling is a sign of fitting an annotator, not of solving the task.

14.2 Generative models

The sequence of function classes for $p_\theta(x_t \mid x_{<t})$ runs: transition table, hidden Markov model, recurrent network, gated recurrent network, transformer. Eck and Schmidhuber showed in 2002 that an LSTM could hold blues structure over timescales where an n gram could not, which is the first demonstration that the memory problem of Section 8 was tractable. Chorale harmonisation became the standard benchmark, since Bach supplies a clean corpus with explicit rules, and systems in the DeepBach line combine a learned conditional with Gibbs sampling over voices, which is the 1949 constraint mechanism run as an MCMC sweep rather than as rejection.

The Music Transformer introduced relative position attention for exactly the reason given in Section 12: metre and repetition are invariances in elapsed time, not in absolute position, so the kernel should depend on $t - s$. Autoregressive models over quantised audio codes, in the Jukebox line, extended the same machinery from symbols to waveforms by learning the alphabet with a vector quantised autoencoder. Text conditioned

audio models and diffusion systems followed. The break with 1949 in the diffusion case is real, since the sampler no longer runs forward in musical time, but the conditioning mechanism is again an exponential reweighting of a base measure.

14.3 The representation question is still open

Choice (D1) has not been settled and shows no sign of being settled. A piano roll grid makes simultaneity trivial and makes expressive timing inexpressible. A MIDI like event stream makes timing exact and makes simultaneity a matter of adjacency in a list, so chords have an arbitrary order and the model must learn to ignore it. Beat relative encodings build metre into the alphabet and thereby prevent the model from ever being wrong about it, which is an advantage until the music has no stable beat. A learned audio codebook defers the choice to an autoencoder and then inherits whatever that autoencoder discarded.

Each choice fixes a group of transformations under which the model is automatically invariant and a second group under which it cannot be. Transposition invariance is free in a pitch interval encoding and must be learned in an absolute pitch one; tempo invariance is free in a beat relative encoding and must be learned in a milliseconds one. The representation is a prior on invariances, and it is by far the strongest prior in any of these systems.

This matters for reading the literature. Reported improvements routinely confound the representation with the model, and there is no accepted protocol that varies one while holding the other fixed. Two systems reported at the same perplexity on different tokenisations are not comparable at all, since the alphabet sets the units. The 1949 authors chose an alphabet in a paragraph and got on with it. Seventy five years later the choice is made by convention rather than by argument, which is not progress.

15 The generative view

The memorandum is a generative model in the strict sense. No likelihood is ever evaluated. The only interface with the world is simulation: one draws a path and listens. This posture is now standard. In generative Bayesian computation one learns a map $x = f_\theta(u, z)$ from base randomness u and conditioning information z to the object of interest, and inference proceeds by pushing samples through the map. Density evaluation is optional and often impossible.

Two consequences follow, and both were visible in 1949.

The first is that model criticism has to be simulation based. If one cannot evaluate the likelihood one cannot compute a likelihood ratio, and comparison of generative models falls back on discrepancy between simulated and observed distributions, which is the posterior predictive check, or on a learned discriminator, which is the adversarial version of the same thing.

The second is that the interesting parameters are not the ones being estimated. In the 1949 model the entries of P matter far less than the choice of $\mathcal{S}$ and the rule set. In a modern system the weights matter far less than the tokenisation and the tilting. Effort follows estimability rather than importance, which is a general hazard and not specific to music.

16 Evaluation

There is no loss function for music. Entropy rate measures the size of the style, not the quality of the sample, and perplexity rewards a model for hedging in exactly the places where a composer should commit. Assessment therefore falls back on listening, which was true in 1949 and remains true.

Two cautions apply to listening studies of generated music and are routinely ignored. Comparisons are made across many models, many prompts, many excerpt lengths and many musical dimensions, and the reported effect is the largest of these. Without accounting for the search, the significance attached to it is meaningless, which is the look elsewhere effect in its usual form. Second, human raters are comparing samples, not distributions. A model that produces excellent samples one time in ten will win a curated listening test and lose on any measure of the generated measure as a whole. The 1949 authors, generating a handful of pieces by hand, had no way to distinguish these and did not claim to.

17 Limitations

We have not read MM-49-150-29. The primary document does not appear to be publicly available, and Section 2 is a reconstruction consistent with the secondary literature rather than a report of its contents. Claims that circulate online about the memorandum, in particular that dice were used and that the output took rondo form, trace to unreliable sources and should not be relied on without the original. The state space may have been notes rather than chords, or both, and the order may have been higher than one. Nothing in the mathematics above depends on which, but the historical claim would.

The division of labour between Pierce and Shannon is not documented. Pierce was the senior figure and the one who wrote on the topic for four decades, so attributing the conception to Betty Shannon would overstate the evidence. What the record supports is joint authorship on a research memorandum.

The transformer material of Section 12 is a set of structural correspondences, not theorems about trained models. The kernel smoother reading of attention is exact as algebra and only suggestive as a description of what a trained network does. The induction head account is empirical and partial. The Diaconis and Freedman theorem is a statement about Markov exchangeable laws, and no trained transformer is known to satisfy Markov exchangeability, so the theorem organises the intuition without licensing the conclusion. Claims that in context learning *is* Bayesian inference should be read as claims about behaviour on synthetic tasks.

Finally, entropy rate figures for music depend entirely on the representation, the alphabet size and the corpus. Comparisons across studies are rarely meaningful, and we have quoted none. The numbers in the worked example are illustrative and are not estimates of anything.

18 Conclusion

The 1949 memorandum makes three choices: a discrete state space, a conditional law given the recent past, and a rule set enforced by rejection. Seventy five years later the same three choices are made, with a learned tokenisation in place of the chord alphabet, a kernel smoother over the whole visible past in place of the transition table, and a reward model in place of the counterpoint rules. The entropy rate still fixes the size of the style. The redundancy is still the dial between a loop and noise, and it is still worth exactly the growth rate of a gambler who knows the source. The constraint is still an exponential tilt of a base measure, priced by a free energy.

Three things did change. The conditional is now a smooth function of a learned representation rather than a lookup, so the order need not be fixed in advance. The parameters come from a corpus rather than from a textbook, which moves the prior from the transition table into the architecture and the tokenisation, where it is harder to see and no less strong. And memory need not decay geometrically, so form is at last representable, if not yet well understood.

What did not change is the part that matters for judging any of it. There is still no loss function for music. The quantity we minimise measures the size of a style and says nothing about whether a particular draw is worth hearing, and the gap between those two things is as wide now as it was when two researchers at Bell Labs rolled for a chord progression and played the result to see.

References

[1] Brooks, F. P., Hopkins, A. L., Neumann, P. G. and Wright, W. V. (1957). An experiment in musical composition. *IRE Transactions on Electronic Computers*, EC 6, 175 to 182.

[2] Braga, M. (2016). The Verbasizer was David Bowie's 1995 lyric writing Mac app. *Vice*, 11 January 2016.

[3] Burroughs, W. S. and Gysin, B. (1978). *The Third Mind*. Viking Press, New York.

[4] Cover, T. M. and King, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, 24, 413 to 421.

[5] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken.

[6] Coupling, J. J. (John R. Pierce) (1950). Science for art's sake. *Astounding Science Fiction*, November 1950, 83 to 91.

[7] Diaconis, P. and Freedman, D. (1980). de Finetti's theorem for Markov chains. *Annals of Probability*, 8, 115 to 130.

[8] Eck, D. and Schmidhuber, J. (2002). Finding temporal structure in music: blues improvisation with LSTM recurrent networks. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 747 to 756.

[9] Hadjeres, G., Pachet, F. and Nielsen, F. (2017). DeepBach: a steerable model for Bach chorales generation. In *Proceedings of ICML*, 1362 to 1371.

[10] Hiller, L. A. and Isaacson, L. M. (1959). *Experimental Music: Composition with an Electronic Computer*. McGraw Hill, New York.

[11] Huang, C. Z. A. et al. (2019). Music transformer: generating music with long term structure. In *International Conference on Learning Representations*.

[12] IEEE Information Theory Society (2017). Mary Elizabeth Moore Shannon, 95. Obituary.

[13] Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35, 917 to 926.

[14] Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press, Cambridge.

[15] Olson, H. F. and Belar, H. (1961). Aid to music composition employing a random probability system. *Journal of the Acoustical Society of America*, 33, 1163 to 1170.

[16] Pierce, J. R. (1960). The computer as a musical instrument. *Journal of the Audio Engineering Society*, 8, 139 to 142.

[17] Pierce, J. R. (1961). *Symbols, Signals and Noise: The Nature and Process of Communication*. Harper, New York.

[18] Pierce, J. R. and Shannon, M. E. (1949). *Composing Music by a Stochastic Process*. Bell Telephone Laboratories Technical Memorandum MM-49-150-29, 15 November 1949.

[19] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379 to 423 and 623 to 656.

[20] Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30, 50 to 64.

[21] Soni, J. and Goodman, R. (2017). Betty Shannon, unsung mathematical genius. *Scientific American*, 22 July 2017.

[22] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461 to 464.

[23] Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman Yor processes. In *Proceedings of ACL*, 985 to 992.

[24] Vaswani, A. et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30.

[25] Xenakis, I. (1963). *Formalized Music: Thought and Mathematics in Composition*. Indiana University Press, Bloomington.

[26] Yu, H. and Varshney, L. R. (2017). On "Composing Music by a Stochastic Process": from computers that are human to composers that are not human. *IEEE Information Theory Society Newsletter*, December 2017.

A Algebra for the worked example

With $\mathcal{S} = \{\mathrm{I}, \mathrm{IV}, \mathrm{V}\}$ and P as in Section 2, the balance equations $\pi^\top P = \pi^\top$ read

$$\pi_1 = 0.4\pi_1 + 0.3\pi_2 + 0.7\pi_3, \qquad \pi_2 = 0.3\pi_1 + 0.2\pi_2 + 0.1\pi_3, \qquad \pi_3 = 0.3\pi_1 + 0.5\pi_2 + 0.2\pi_3.$$

The second gives $\pi_2 = 0.375\pi_1 + 0.125\pi_3$. Substituting into the third gives $0.7375\pi_3 = 0.4875\pi_1$, so $\pi_3 = 0.6610\pi_1$ and $\pi_2 = 0.4576\pi_1$. Normalising, $\pi_1(1 + 0.4576 + 0.6610) = 1$ and

$$\pi \;=\; (0.4722,\; 0.2161,\; 0.3121).$$

For the spectrum, $\operatorname{tr} P = 0.8 = 1 + \lambda_2 + \lambda_3$ and $\det P = 0.05 = \lambda_2\lambda_3$, so λ_2, λ_3 solve $x^2 + 0.2x + 0.05 = 0$, giving $\lambda_{2,3} = -0.1 \pm 0.2i$ and $|\lambda_2| = \sqrt{0.05} = 0.2236$. The relaxation time is $1/\log(1/|\lambda_2|) = 0.667$ steps. The eigenvalues are complex, so the decay carries a slow oscillation, with argument $\arg\lambda_2 = 2.034$ radians and hence a pseudo period of $2\pi/2.034 = 3.09$ chords. That is the only periodic structure in the model, and it is not metre.

The row entropies are

$$H_{\mathrm{I}} = 2(0.3\log_2\tfrac{1}{0.3}) + 0.4\log_2\tfrac{1}{0.4} = 1.5710,$$
$$H_{\mathrm{IV}} = 1.4855, \qquad H_{\mathrm{V}} = 1.1568,$$

so

$$H(P) = 0.4722(1.5710) + 0.2161(1.4855) + 0.3121(1.1568) = 1.4233 \text{ bits},$$

against $\log_2 3 = 1.5850$, giving $R = 0.1021$ and a Kelly growth rate of $R\log_2 3 = 0.1618$ bits per chord by (1). A thirty two chord phrase has $2^{32H(P)} = 2^{45.5} \approx 5.1 \times 10^{13}$ typical realisations.

All figures are illustrative. They describe a table written down to make the algebra transparent, and they are not estimates from any corpus.