# Maximum Entropy Model for Melodic Patterns

**Jason Sakellariou** JASON.SAKELLARIOU@LIP6.FR
LIP6, UPMC Université Paris VI, France & Sony CSL, Paris, France

**Francesca Tria**
Sony CSL, Paris, France & Complex Systems Lagrange Lab, Institute for Scientific Interchange (ISI), Torino, Italy

**Vittorio Loreto**
Sony CSL, Paris, France & Department of Physics, Sapienza Universita di Roma, Italy & Institute for Scientific Interchange (ISI), Torino, Italy

**François Pachet**
Sony CSL, Paris, France & LIP6, UPMC Université Paris VI, France

## Abstract

We introduce a model for music generation where melodies are seen as a network of interacting notes. Starting from the principle of maximum entropy we assign to this network a probability distribution, which is learned from an existing musical corpus. We use this model to generate novel musical sequences that mimic the style of the corpus. Our main result is that this model can reproduce high-order patterns despite having a polynomial sample complexity. This is in contrast with the more traditionally used Markov models that have an exponential sample complexity.

## 1. Introduction

Many complex systems exhibit a highly non-trivial structure that is difficult to capture with simple models. Several biological systems form networks of interacting components (neurons, proteins, genes, whole organisms) and their collective behaviour is characterized by a complex mosaic of correlations between its various parts. Arguably, the ultimate biological origin of purely intellectual constructs such as language or music, should allow us to look at them from a similar point of view, i.e. as complex networks of interacting components. In both cases one would suspect that essential features of their complexity arise from high-order, combinatorial interactions. However, a number of works in recent years have shown that models based on *pairwise* interactions alone capture most of the correlation structure of some biological systems (Schneidman et al., 2006; Lezon et al., 2006; Bialek & Ranganathan, 2007; Mora et al., 2010; Bialek et al., 2012) and even English words (Stephens & Bialek, 2010).

One of the most popular strategies for algorithmic music composition is Markov chains (see for example (Roads, 1996)). In order to capture the long-range structure of musical phrases, high-order Markov models must be used. This leads to serious overfitting issues: the number of $k$-grams in the musical corpus is orders of magnitude smaller than their total potential number, which is exponential in $k$. Typical musical corpora contain a few hundred notes when the total number of different pitches is a few tens.

In this paper we propose a model consistent with pairwise correlations between notes across many different time-distances. We show that, for musical data, combining pairwise constraints restricts the space of solutions enough for higher-order patterns to emerge. That way, we capture long range musical patterns while avoiding the overfitting issues of high-order models. This approach cannot be implemented as an extension to Markov models, however, and a different framework is needed. This framework is provided by the *maximum entropy principle* (Jaynes, 1957). Maximum entropy models consistent with pairwise correlations are variations of the Ising or Potts models of statistical mechanics (see for instance (Baxter, 2007)), which have a long and rich history as theoretical models for statistical order and phase transitions. These models belong to the large family of *Probabilistic Graphical Models* which offer a very general framework for modeling statistical dependencies. Our model can be used for generating sequences

*Proceedings of the $32^{nd}$ International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. 

that mimic some aspects of the musical style of a given corpus.

## 2. The Model

Music has many aspects (melody, harmony, rhythm, form, sound, etc) which renders realistic models extremely complicated. In this paper we focus on monophonic pitch sequences, for simplicity. A pitch sequence is a sequence of integers $\{s_1, \ldots, s_N\}$ encoding note pitches ordered as they appear in a real melody but disregarding any other information about duration, onset, velocity etc. The variables $s_i$ take values from some finite alphabet representing the different pitches appearing in the corpus. In our model we are interested in reproducing the correct frequencies of single notes and of pairs of notes at distance $k$

$$f(\sigma) \equiv \frac{1}{N}\sum_{i=1}^{N}\delta(\sigma, s_i) \quad (1)$$

$$f_k(\sigma, \sigma') \equiv \frac{1}{N-k}\sum_{|i-j|=k}\delta(\sigma, s_i)\delta(\sigma', s_j) \quad (2)$$

with $k = 1, \ldots, K_{\text{max}}$. In the above formulas $\delta(\cdot,\cdot)$ is the Kronecker delta symbol. The sums run over the whole corpus (i.e. the original pitch sequence). The *maximum entropy* recipe states that we have to look for the distribution that maximizes the entropy while being consistent with the frequency counts in eqs. (1) and (2). This is the least biased estimate of the true distribution, given the available information. A simple calculation leads to the following Boltzmann-Gibbs distribution

$$P(s_1, \ldots, s_N) = \frac{1}{Z}\exp\left(-\mathcal{H}(s_1, \ldots, s_N)\right) \quad (3)$$

where $\mathcal{H}$ is called the Hamiltonian and is a sum of "potentials" each corresponding to one of the above constraints

$$\mathcal{H}(s_1, \ldots, s_N) = -\sum_{i=1}^{N} h(s_i) - \sum_{k=1}^{K_{\text{max}}}\sum_{\substack{i,j \\ |i-j|=k}} J_k(s_i, s_j) \quad . \quad (4)$$

The partition function

$$Z \equiv \sum_{s_1}\sum_{s_2}\cdots\sum_{s_N}\exp\left(-\mathcal{H}(s_1, \ldots, s_N)\right) \quad (5)$$

guaranties that the distribution is normalized. We will refer to the $h$'s as the local fields and to the $J_k$'s as the interaction potentials. Inspired by statistical physics, these quantities can be thought as external fields acting on the variables on one hand and interactions between variables on the other. The Hamiltonian then gives the energy of the system by summing the contribution of all the above terms. According to distribution (3), sequences with low energy have an exponentially larger probability. Therefore, the effect of the above "potentials" is to bias the individual and pair probabilities so as to make the model consistent with the corpus. A graphical representation of our model can be seen in Figure 1.

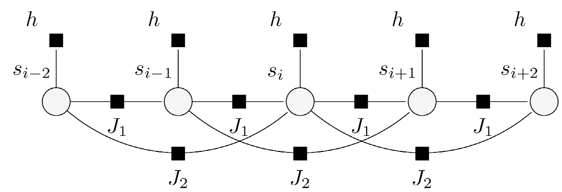


*Figure 1.* Factor graph representing the factorisation of the distribution (3) for $K_{\text{max}} = 2$. Factors (squares) are connected to the variables (circles) according to eq. (4)

## 3. Results

We first need to train our model. Training the model means finding the set of local fields and interactions that can best reproduce the individual and pair frequencies of the corpus. In other words we want to maximize the likelihood of the $h$'s and $J$'s given the data. This procedure is difficult computationally since it relies on exact inference, a NP-hard problem in general(Cooper, 1990). Among the different existing approximate solutions we chose to use *pseudo-likelihood maximization* first introduced in (Ravikumar et al., 2010).

### 3.1. Pair Frequencies

Once the potentials have been found one can generate new pitch sequences by sampling from distribution (3). This can be simply done by the Metropolis Algorithm (Metropolis et al., 1953). Figure 2 shows a scatter plot for pair frequencies of the corpus VS the ones generated by our model. For this particular example we used as a corpus the content of (WeimarJazzDatabase) consisting of 257 transcriptions of famous Jazz improvisations. There is very good agreement for the more frequent pairs and, as expected, small pair probabilities are reproduced less accurately.

To better appreciate what the model does it is informative to look at pair frequencies for different distances separately. Figure 3 shows colormaps of matrices given by eq. (2) for $k = 1$ and 5 for three cases: the original sequence, here Partita No. 1 in B minor BWV 1002 by Johann Sebastian Bach (part II double), a first-order Markov model and our maximum entropy model. The Markov model, by construction, reproduces perfectly the frequencies of neighbouring notes ($k = 1$). However it completely fails at greater dis-

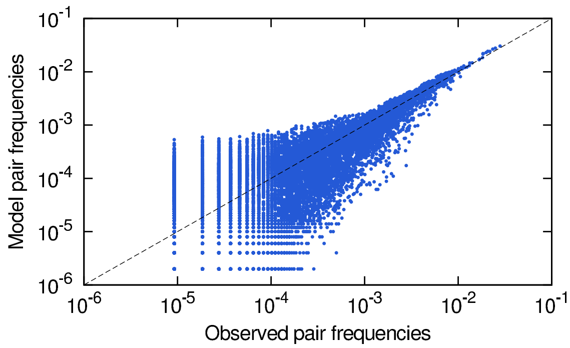


*Figure 2.* Model VS observed pair frequencies. The observed ones are from the corpus (WeimarJazzDatabase). The model frequencies come from a $K_{\max} = 10$ model.

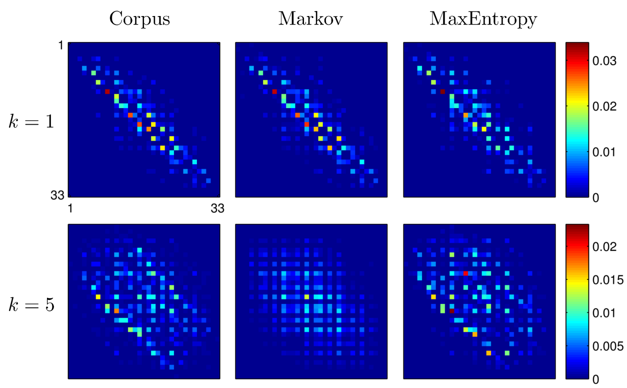


*Figure 3.* Matrices of note-pairs frequencies. Colormaps representing matrices obtained by counting note-pair occurences (see eq. (2)). From left to right: the original sequence (J.S. Bach's first violin partita), a first-order Markov model and our maximum entropy model. Top row: $k = 1$, bottom: $k = 5$

tances. In the bottom row we see that the particular information contained in the $k = 5$ matrix of the corpus is almost completely lost for the Markov case. The maximum entropy model, however, performs equally well on both cases. Here we used a model with $K_{\max} = 10$ so the training will make sure to select a set of potentials that better reproduce the pair frequencies for all distances up to 10. The reason for using a first-order Markov model for comparison with our model is that they have comparable sample complexities. First-order Markov models have $O(|\mathcal{X}|^2)$ parameters while our model has $O(K_{\max}|\mathcal{X}|^2)$, where $|\mathcal{X}|$ is the alphabet size. In contrast, a $K_{\max}$-order Markov model has $O(|\mathcal{X}|^{(K_{\max}+1)})$. A high-order Markov model would reproduce correctly the pair frequencies at all distances by trivially copying the whole corpus. Examples of the generated sequences can be heard in the Flow Machines webpage (MaxEntropyExamples).

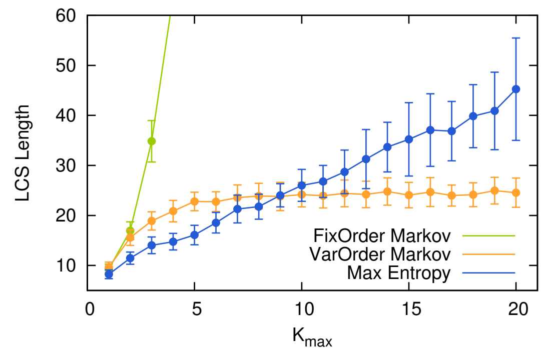


*Figure 4.* Longest common string lengths between the corpus (J.S. Bach's first violin partita) and sequences of size $N = 5000$ generated from various models. We average over 100 sequences for each model.

### 3.2. Style Imitation

Musical style imitation is a concept which is difficult to formalize. However, most musicians would agree that it involves "creatively" rearranging existing material. By that we don't mean a mere reshuffling of melodic patterns. In the new sequence, these patterns must succeed "naturally" one another just as in the original one. This is a requirement from a model that aims at imitating a musical style. Additionally, we would want to be able to control the characteristic size of the patterns that are used as building blocks of the new sequence.

Markov models guarantee this "naturalness" by ensuring that every $k$-gram is continued according to conditional probabilities estimated from the corpus. However, it is very difficult to have a good control on the size of patterns. There is a sharp transition between an under-fitted regime, where small chunks are copied, and an over-fitted regime where very large ones are reproduced exaclty.

In contrast, in our model the size of chunks copied by the corpus scale linearly with $K_{\max}$. In particular we look for the *longest common string* (LCS) between generate sequences from the different models and the corpus. The results can be seen in Figure 4, where we also included results from a *variable order Markov model* (Pachet, 2003) for completeness. Here $K_{\max} = 1, \ldots, 20$ represents the order for the fixed order Markov models (FOM), the maximal order for the variable length Markov models (VOM) and the maximal interaction range for our maximum entropy model (MaxEnt). The LCS length for FOM explodes at a very small value of $k$ while, for the VOM, it reaches a plateau and longer strings are not reachable. In our model the LCS length seems to have a linear relation with $K_{\max}$. As we discussed earlier, this property is desirable since it allows to fine-tune the characteristic size of the chunks involved.

## 4. Conclusion

We presented a maximum entropy model that captures pairwise correlations between notes in a musical sequence, at various distances. The model is used to generate original sequences that mimic a given musical style. The particular topology of this model (see Figure 1) leads to the emergence of high-order patterns, despite the pairwise nature of the information used, which in turn has the benefits of a quadratic, in the alphabet size, sample complexity. This allows us to maintain control over the sizes of the common strings between the corpus and the generated sequences in a way that outperforms the more commonly used Markov models (of both fixed and variable order).

Beyond this technical advantage, graphical models like this one offer a very general setting for modeling statistical depedencies. Work in progress will extend this model in order to account for other aspects of music, such as rhythm or polyphony. The general idea is the same: additional information (e.g. note durations) is captured by additional variables which are coupled with *pairwise* interactions. In that way, we keep the quadratic sample complexity.

## Acknowledgments

This research is conducted within the Flow Machines project which received funding from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 291156

## References

Baxter, Rodney J. *Exactly solved models in statistical mechanics*. Courier Corporation, 2007.

Bialek, William and Ranganathan, Rama. Rediscovering the power of pairwise interactions. *arXiv preprint arXiv:0712.4397*, 2007.

Bialek, William, Cavagna, Andrea, Giardina, Irene, Mora, Thierry, Silvestri, Edmondo, Viale, Massimiliano, and Walczak, Aleksandra M. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

Cooper, Gregory F. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, 42(2):393–405, 1990.

Jaynes, Edwin T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Lezon, Timothy R, Banavar, Jayanth R, Cieplak, Marek, Maritan, Amos, and Fedoroff, Nina V. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences*, 103(50):19033–19038, 2006.

MaxEntropyExamples. http://www.flow-machines.com/MaximumEntropy/.

Metropolis, Nicholas, Rosenbluth, Arianna W, Rosenbluth, Marshall N, Teller, Augusta H, and Teller, Edward. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Mora, Thierry, Walczak, Aleksandra M, Bialek, William, and Callan, Curtis G. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010.

Pachet, Francois. The continuator: Musical interaction with style. *Journal of New Music Research*, 32(3):333–341, 2003.

Ravikumar, Pradeep, Wainwright, Martin J, Lafferty, John D, et al. High-dimensional ising model selection using $\ell_1$-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Roads, Curtis. *The computer music tutorial*. MIT press, 1996.

Schneidman, Elad, Berry, Michael J, Segev, Ronen, and Bialek, William. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

Stephens, Greg J and Bialek, William. Statistical mechanics of letters in words. *Physical Review E*, 81(6):066119, 2010.

WeimarJazzDatabase. http://jazzomat.hfm-weimar.de/dbformat/dboverview.html. The Weimar Jazz Database contains detailed transcriptions of famous jazz improvisations. As of March 2015 the database contains 257 songs.