Deep Learning Techniques for Music Generation
3. Generation by Feedforward Architectures

Jean-Pierre Briot

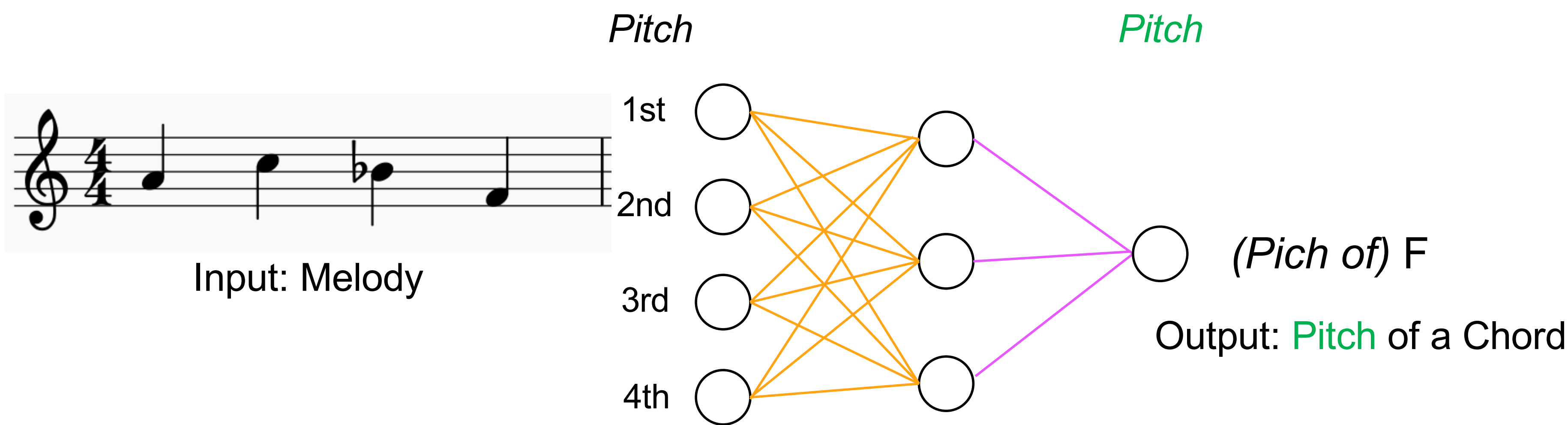
Jean-Pierre.Briot@lip6.fr

Laboratoire d'Informatique de Paris 6 (LIP6)
Sorbonne Université – CNRS

Programa de Pós-Graduação em Informática (PPGI)
UNIRIO

Direct Use – Feedforward – Ex 1

- Feedforward Architecture
- Prediction Task
- Ex1: Predicting a chord associated to a melody segment
  - scale/mode -> tonality

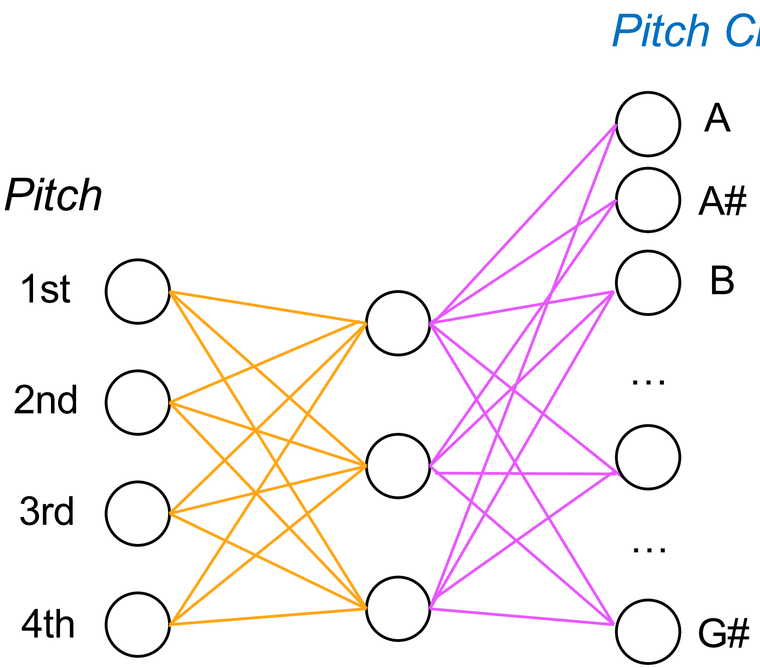


  - Training on a corpus/dataset <melody, chord>
  - Production (Prediction)

Direct Use – Feedforward – Ex 1

- Feedforward Architecture
- Classification Task
- Ex1: Predicting a chord associated to a melody segment
  - scale/mode -> tonality

Input: Melody

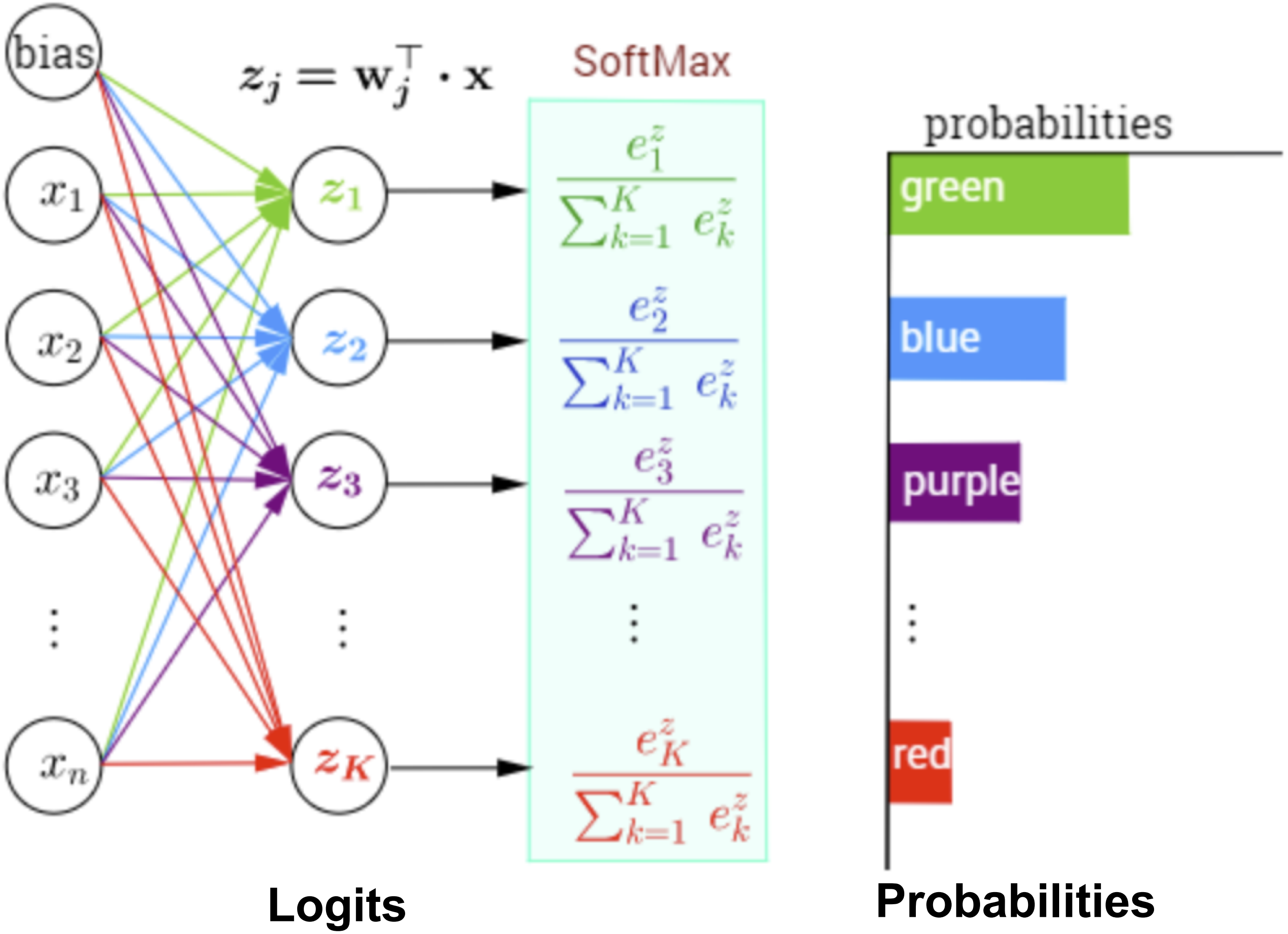


F

Output: Chord (Pitch Class)

  - Training on a corpus/dataset <melody, chord>
  - Production (Classification)

Softmax

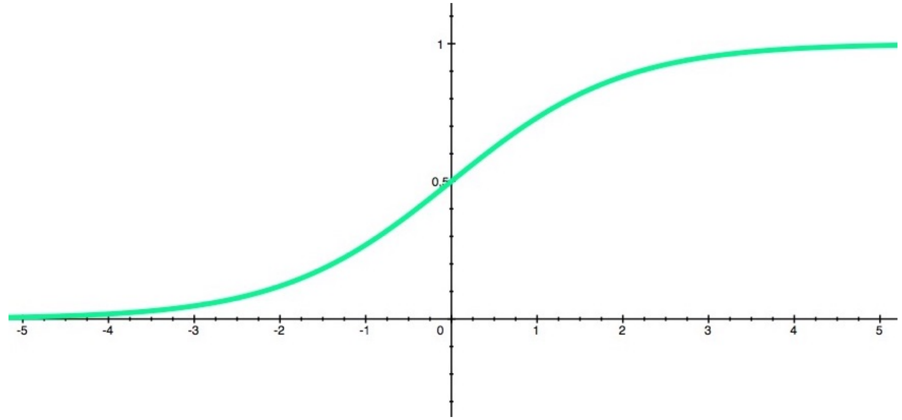


bias
x_1
x_2
x_3
x_n
$z_j = \mathbf{w}_j^\top \cdot \mathbf{x}$
z_1
z_2
z_3
z_K
SoftMax
$\frac{e_1^z}{\sum_{k=1}^K e_k^z}$
$\frac{e_2^z}{\sum_{k=1}^K e_k^z}$
$\frac{e_3^z}{\sum_{k=1}^K e_k^z}$
$\frac{e_K^z}{\sum_{k=1}^K e_k^z}$
probabilities
green
blue
purple
red
Logits
Probabilities

Softmax and Sigmoid

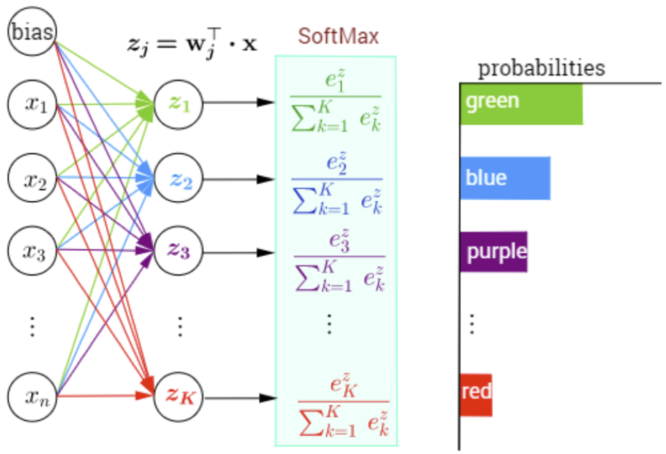
- Softmax is the generalization of Sigmoid
- From Binary classification to Categorical (Multiclass) classification

Sigmoid



Probability € [0, 1]

Softmax



$\sum$ Probabilities = 1

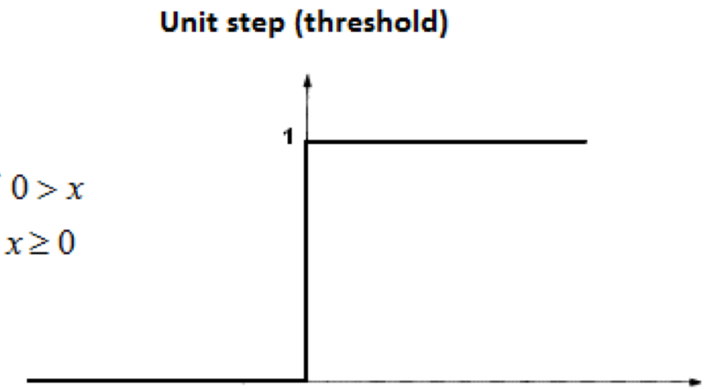
Logits **Probabilities**

Softmax and Sigmoid

- Step function and Argmax are NOT differentiable
- No gradient -> No possibility of back propagation

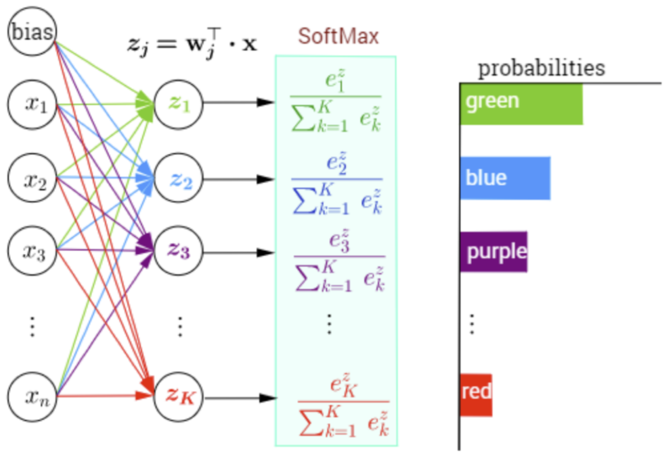
Step function (Perceptron)

$$f(x) = \begin{cases} 0 \text{ if } 0 > x \\ 1 \text{ if } x \geq 0 \end{cases}$$



Probability € {0, 1}

Argmax



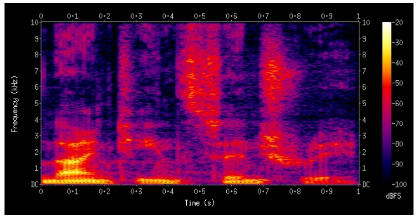
Logits **Probabilities**

Probability(Argmax) = 1

Representation

- Audio
  - Waveform

  - Spectrogram (Fourier Transform)



  - Other (ex: MFCC)

- Symbolic
  - Note

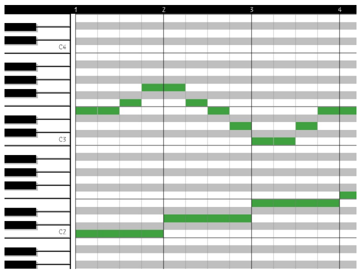
  - Rest

  - Note hold

  - Duration
  - Chord

  - Rhythm

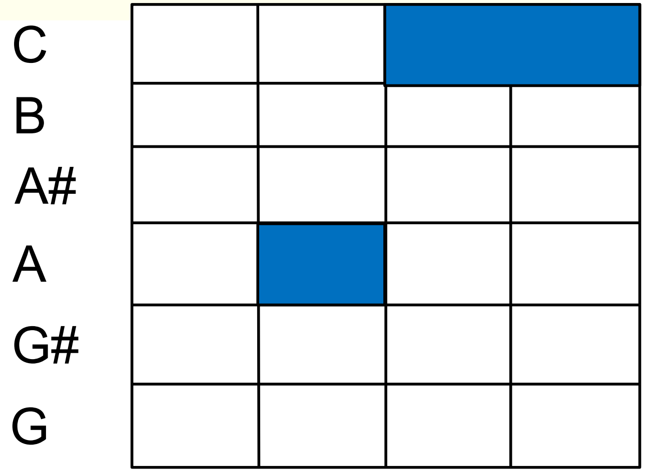
  - Piano Roll



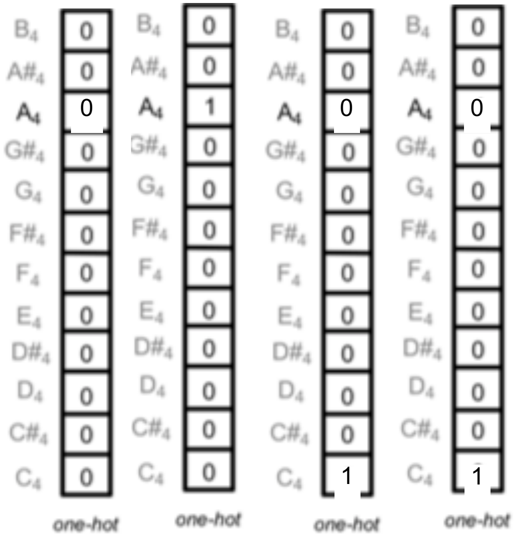
  - MIDI
  - ABC, XML…

Representation

Score



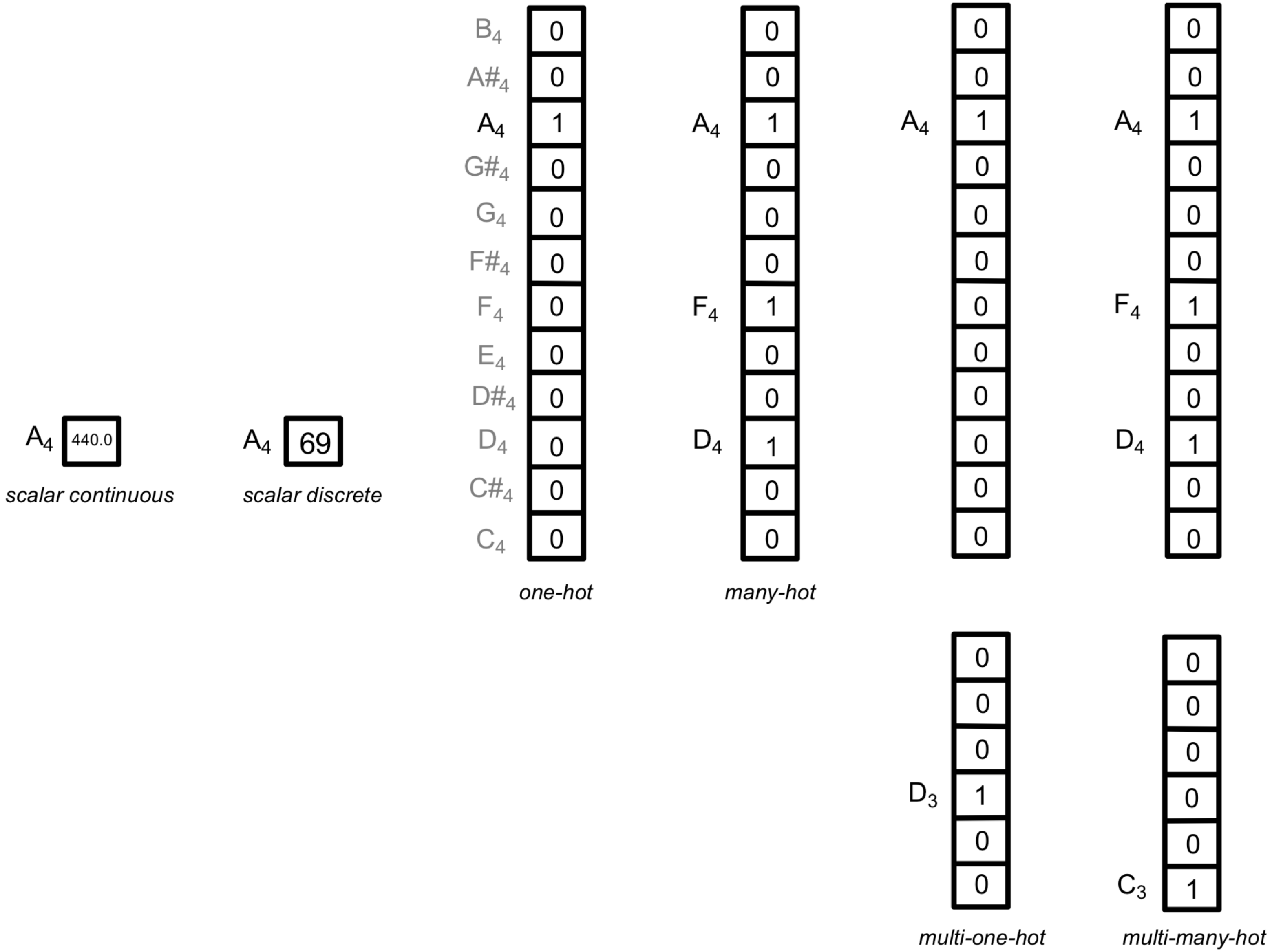
Piano Roll



One hot Encoding

Encoding of Features (ex : Note Pitch)

- Value
  - *Analogic*
- One-Hot
  - *Digital*
- Embedding
  - *Constructed*

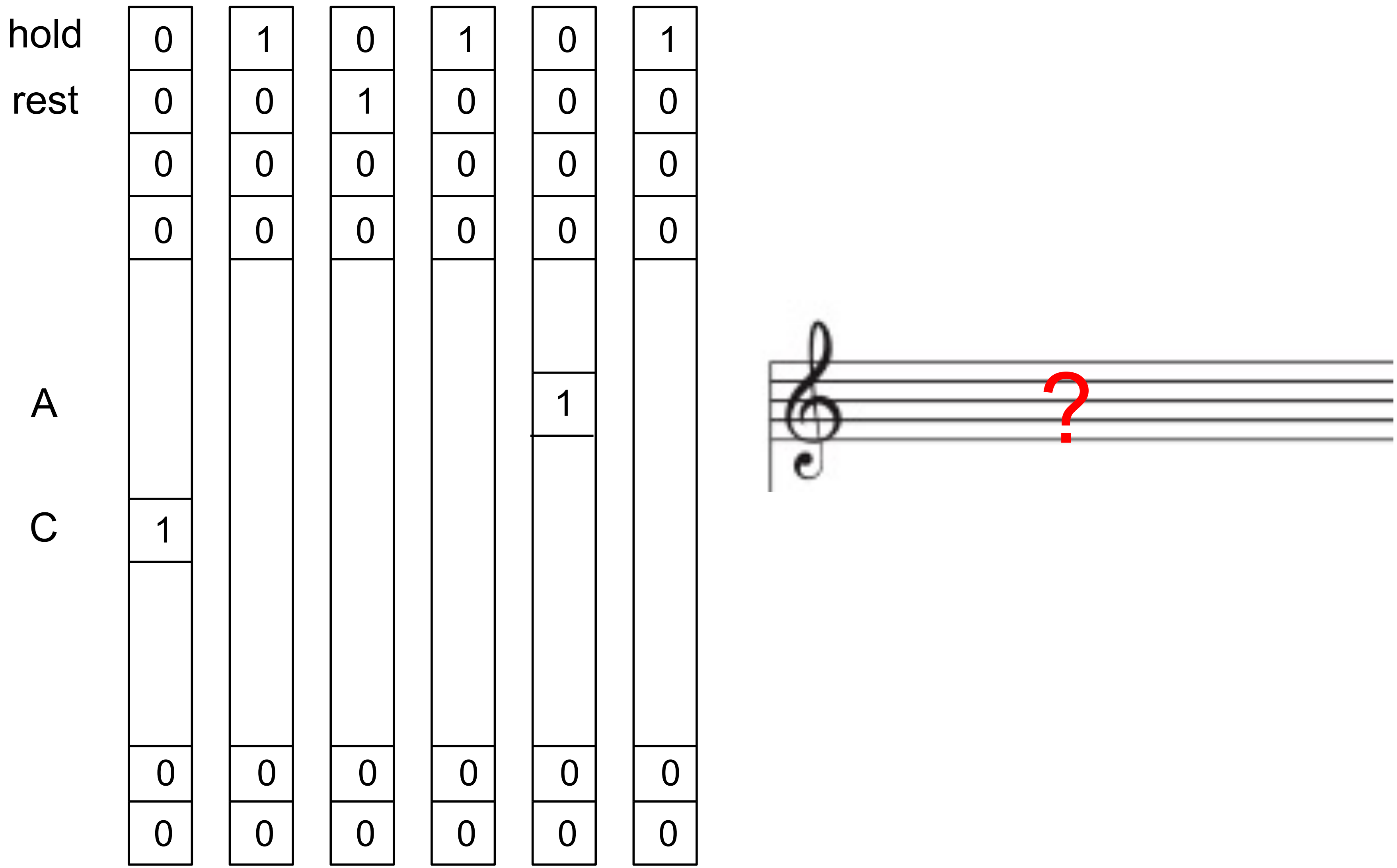


Encoding

- Rest
  - Zero-hot
    - » But ambiguity with low probability notes
  - One more one-hot element
  - …

- Hold
  - One more one-hot element
    - » But only for monophonic melodies
  - Replay matrix
  - …

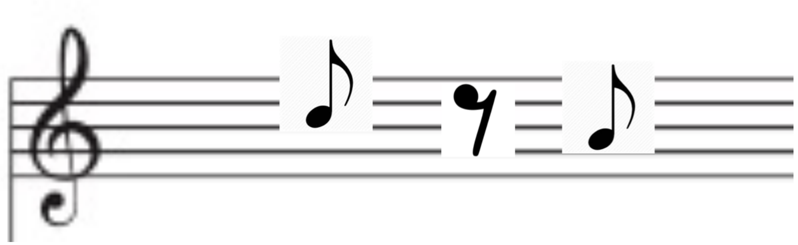
Representation



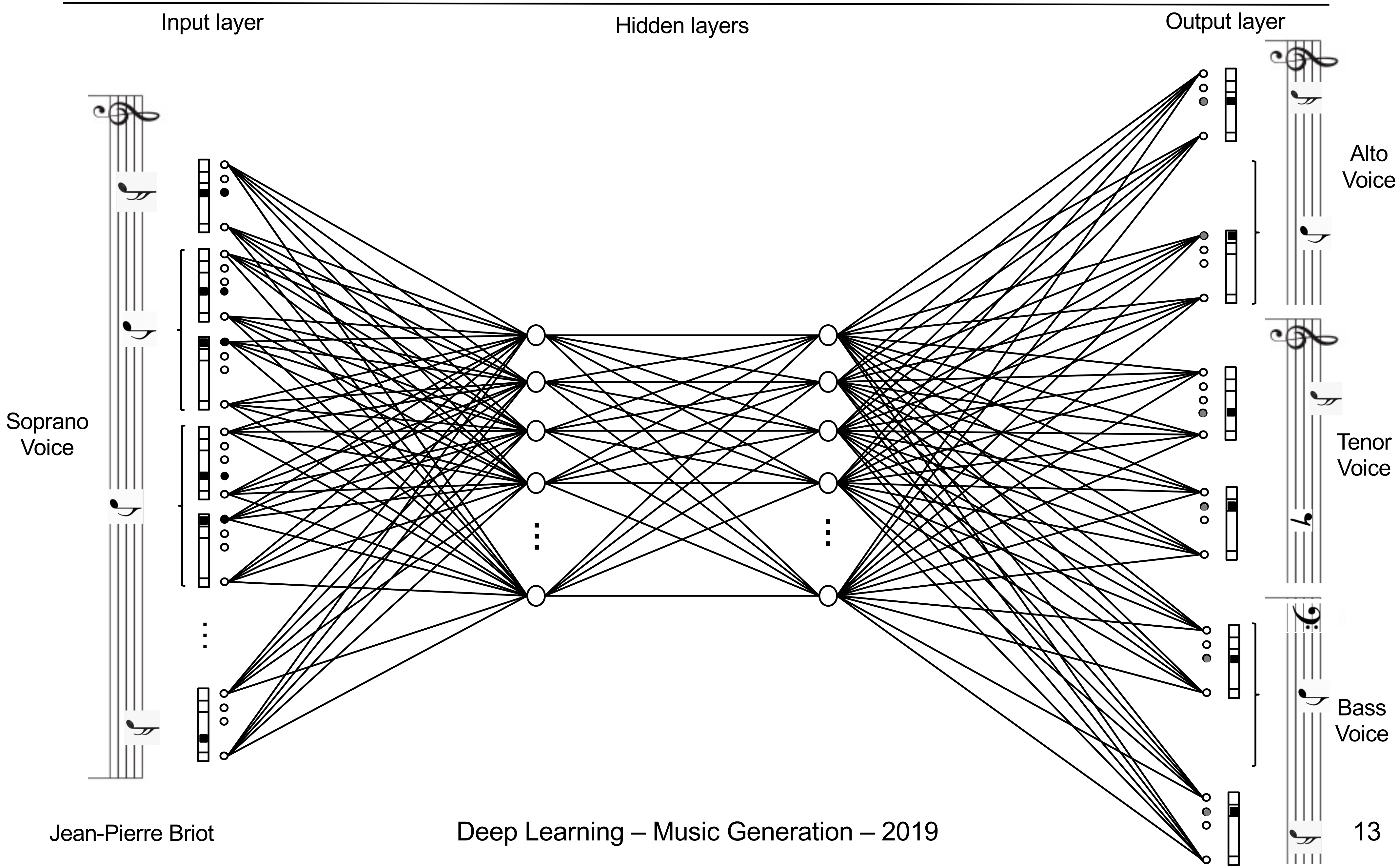
Representation

hold	0	1	0	1	0	1
rest	0	0	1	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
A					1	
C	1					
	0	0	0	0	0	0
	0	0	0	0	0	0

If time slice = sixteenth 𝅘𝅥𝅯



Music / Representation / Network



Code

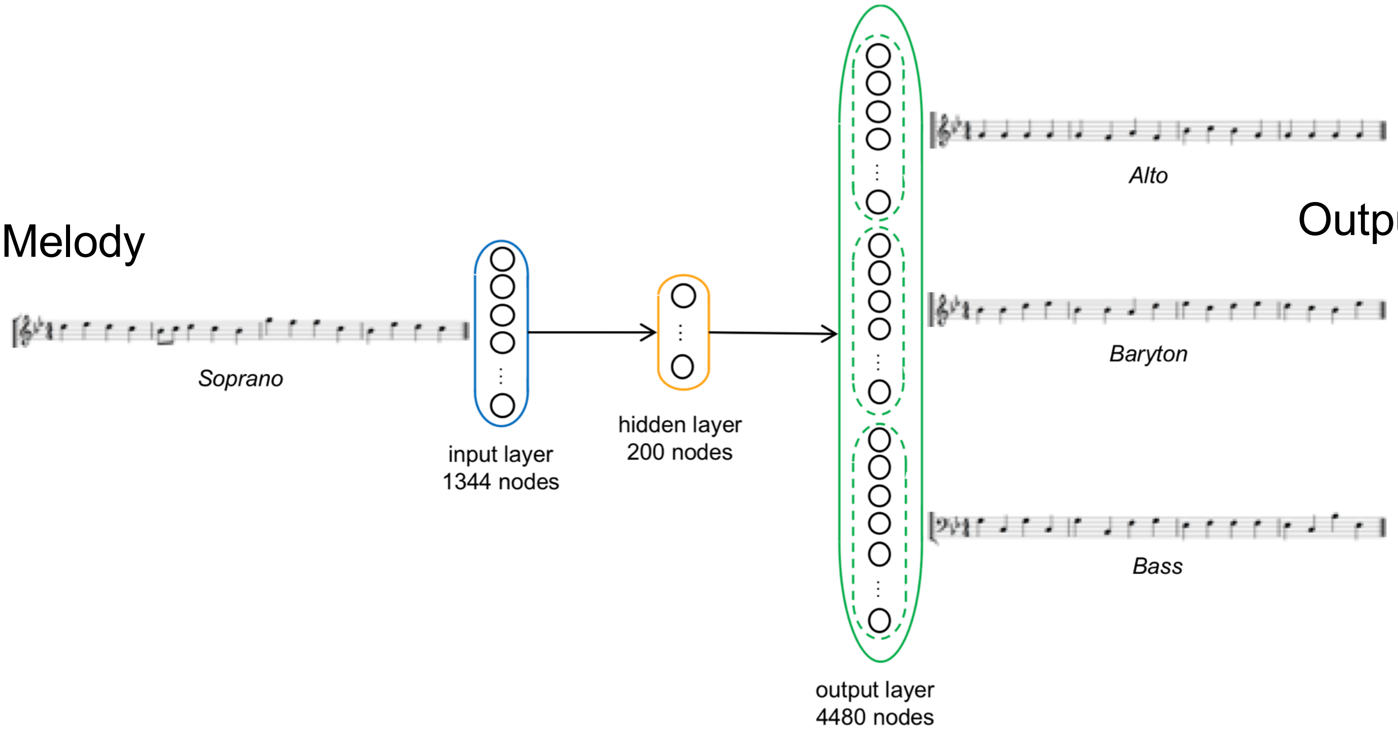
- Python (3)
- Keras
- Theano or TensorFlow
- Music21

Direct Use – Feedforward – Ex 2: ForwardBach

- Feedforward Architecture
- Prediction Task
- Ex2: Counterpoint (Chorale) generation

- Training on the set of (389) J.S. Bach Chorales (Choral Gesang)

Input: Melody



Output: 3 Melodies

ForwardBach

Bach BWV 344 Chorale
(Training Example)

Original

Regenerated

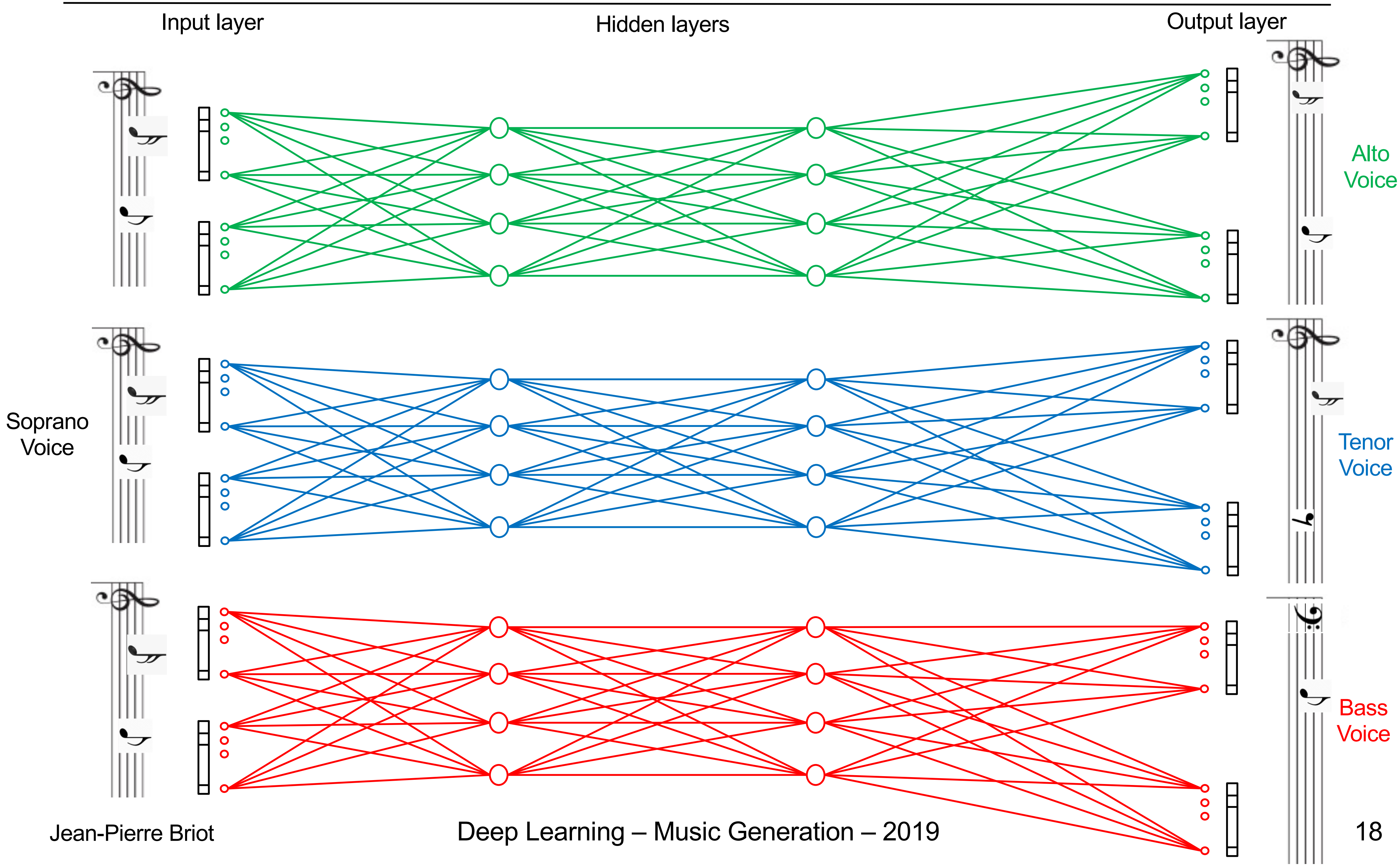
ForwardBach

Bach BWV 423 Chorale
(Test Example)

Original

Regenerated

Music / Representation / Network
Alternative 3 Models Architecture [Cotrim & Briot, 2019]



Forward3Bach [Cotrim & Briot, 2019]

Bach BWV 423 Chorale
(Test Example)

Original

Single Architecture Regenerated

Triple Architecture Regenerated

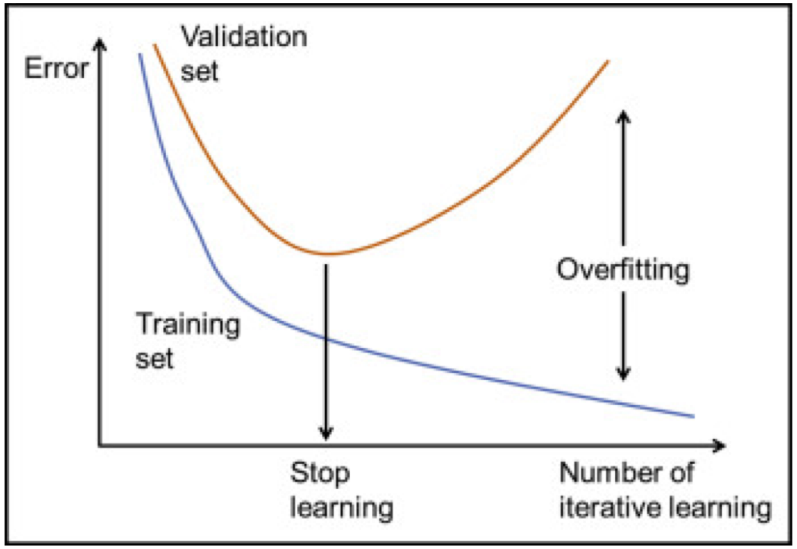
Comparison ? What happened ?

Overfitness Limitations

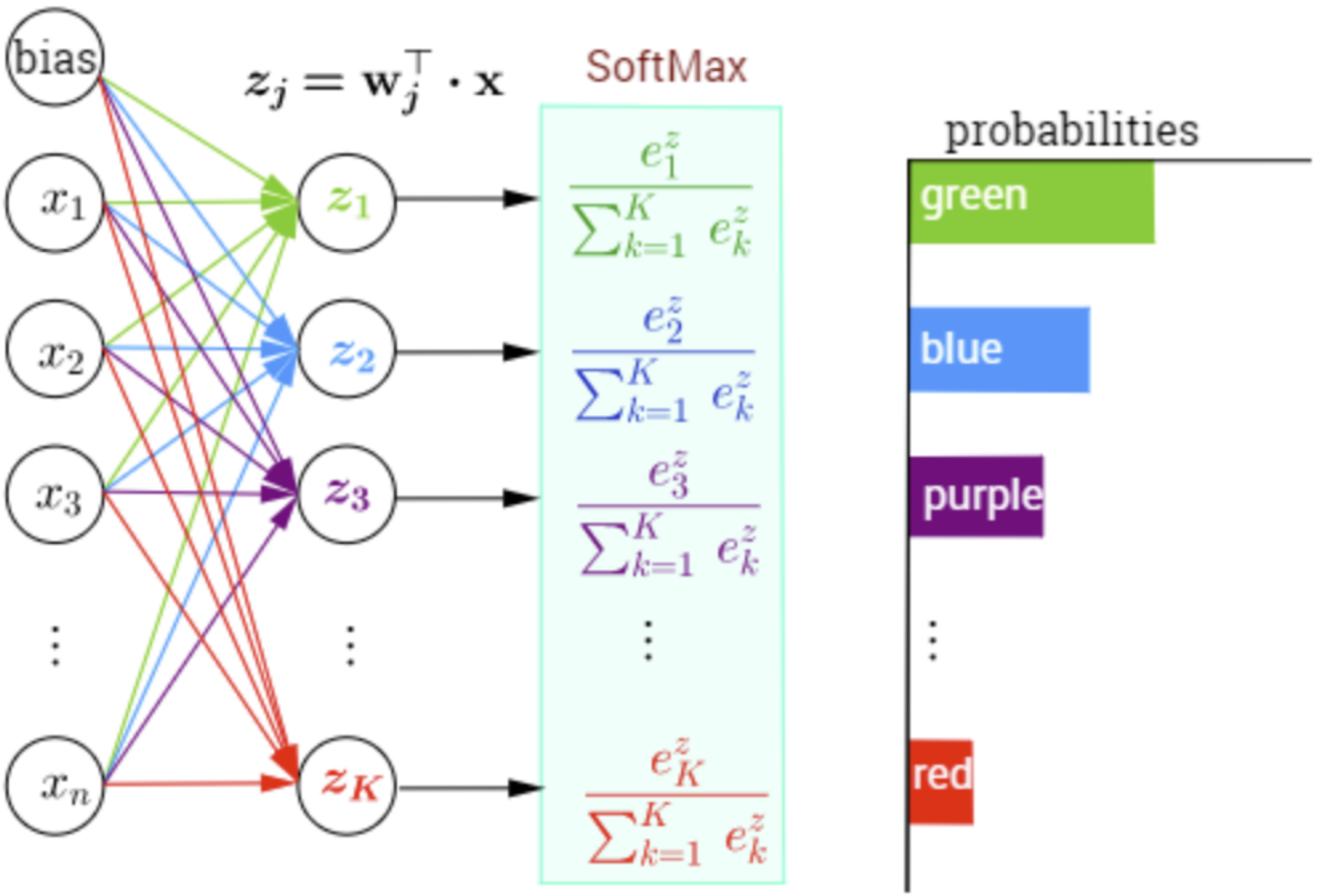
- Musical accuracy is not that good (yet)
- Regeneration of training example is better than Regeneration of test/validation example
- Case of Overfitness

Techniques

- Limit Accuraccy and Control Overfitness
- More Examples (Augment the Corpus)
  - Keeping a Good Style Representation, Coverture and Consistency
  - More Consistency and Coverture
  - Transpose (Align) All Chorales to Only One Key (ex: C)
- More Synthetic Examples
  - More Coverture
  - Transpose All Chorales in All Keys (12)
- Regularization
  - Weight-based
    - L1, L2
  - Connexion-based
    - Dropout
  - Epochs-based
    - Early-Stop
  - Analysis of Learning Curves



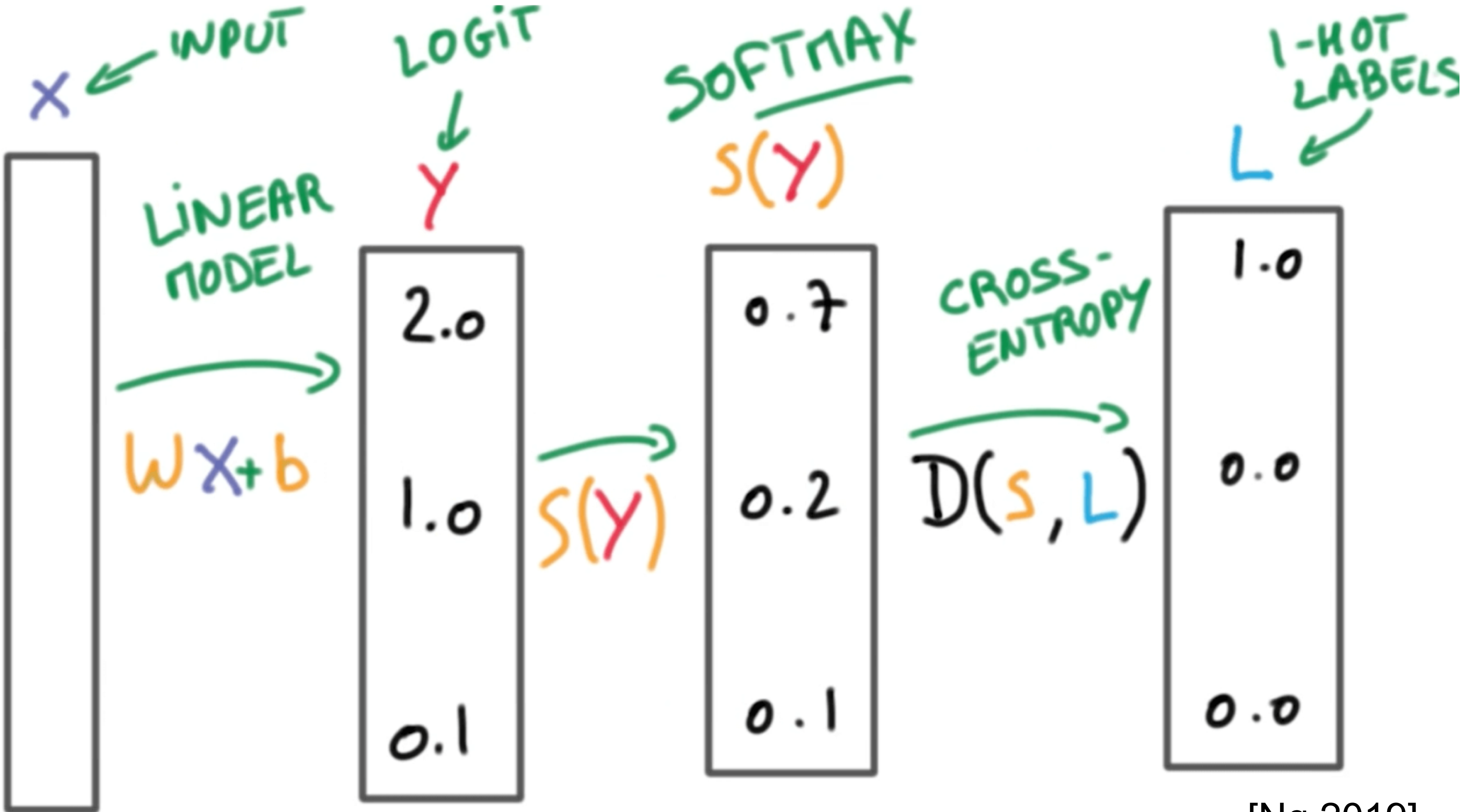
Softmax



Logits **Probabilities**

Softmax + Cross-Entropy

- Cross-Entropy measures dissimilarity between two probability distributions (prediction and target/true value)



[Ng 2019]

Output Activation and Cost/Loss Functions

Task	*Type of the output* ($\hat{y}$)	*Encoding of the target* (y)	*Output activation function*	*Cost (loss)*	*Application*
Regression	Real	$\mathbb{R}$	Identity (Linear)	Mean squared error	
Classification	Binary	$\{0, 1\}$	Sigmoid	Binary cross-entropy	
Classification	Multiclass single label	One-hot	Softmax	Categorical cross-entropy	Monophony
Classification	Multiclass multilabel	Many-hot	Sigmoid	Binary cross-entropy	Polyphony
Multiple Classification	Multi Multiclass single label	Multi One-hot	Sigmoid	Binary cross-entropy	Multivoice
			Multi Softmax	Multi Categorical cross-entropy	

Ex. multiclass single label: Classification among a set of possible notes for a monophonic melody, with only one single possible note choice (single label)

Ex. multiclass multilabel: Classification among a set of possible notes for a single-voice polyphonic melody, therefore with several possible note choices (several labels)

Ex. multi multiclass single label: Multiple classification among a set of possible notes for multivoice monophonic melodies, therefore with only one single possible note choice for each voice; Multiple classification among a set of possible notes for a set of time slices for a monophonic melody, therefore for each time slice with only one single possible note choice

Output Activation and Cost/Loss Functions

Task	*Type of the output* ($\hat{y}$)	*Encoding of the target* (y)	*Output activation function*	*Cost (loss)*	*Interpretation*
Regression	Real	$\mathbb{R}$	Identity (Linear)	Mean squared error	none
Classification	Binary	$\{0, 1\}$	Sigmoid	Binary cross-entropy	none
Classification	Multiclass single label	One-hot	Softmax	Categorical cross-entropy	argmax or sampling
Classification	Multiclass multilabel	Many-hot	Sigmoid	Binary cross-entropy	argsort and > threshold & max-notes
Multiple Classification	Multi Multiclass single label	Multi One-hot	Sigmoid	Binary cross-entropy	*p* argmax or sampling
			Multi Softmax	Multi Categorical cross-entropy	

Other cost functions:

Mean absolute error, Kullback-Leibler (KL) divergence…

Output Activation and Cost/Loss Functions

Output type	*Output activation function*	*Output value ($\hat{y}$)*	*Cost*	*Target (true) value (y)*	*Interpretation*	*Meaning*
Real	Identity	439.7	Mean squared error	440		A_4
Binary	Sigmoid	0.96	Categorical cross-entropy	1	> 0.5	True

Output Activation and Cost/Loss Functions

Output type	*Output activation function*	*Output value* ($\hat{y}$)	*Cost*	*Target (true) value* (y)	*Interpretation*	*Meaning*
Multiclass single label	Softmax	$\begin{bmatrix} 0.03 \\ 0.02 \\ 0.52 \\ 0.03 \\ 0.04 \\ 0.13 \\ 0.05 \\ 0.02 \\ 0.03 \\ 0.07 \\ 0.04 \\ 0.02 \end{bmatrix}$	$\longleftrightarrow$ Categorical cross-entropy	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	Argmax	A_4

Output Activation and Cost/Loss Functions

Output type	*Output activation function*	*Output value ($\hat{y}$)*	*Cost*	*Target (true) value (y)*	*Interpretation*	*Meaning*
		0.07	↔	0		
		0.03	↔	0		
		0.54	↔	1		A_4
		0.03	↔	0		
		0.07	↔	0		
Multiclass multilabel	Sigmoid	0.02	↔	0	Argsort & > *threshold* & < *max number notes*	
		0.45	↔	1		F_4
		0.02	↔	0		
		0.09	↔	0		
		0.63	↔	1		D_4
		0.12	↔	0		
		0.04	↔	0		
			Binary cross-entropy			

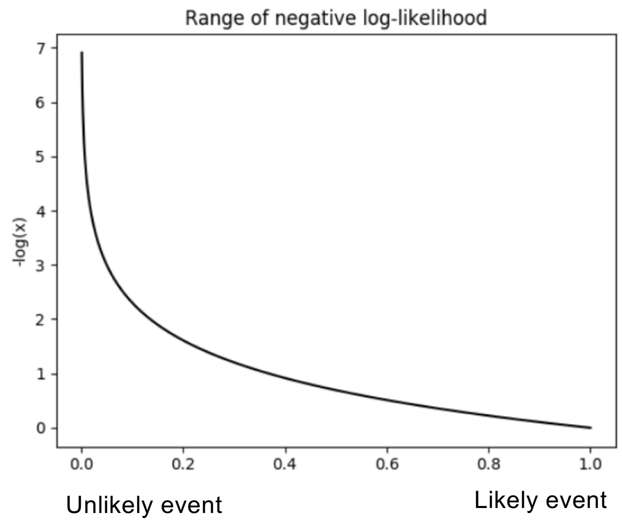
Output Activation and Cost/Loss Functions

Output type	*Output activation function*	*Output value ($\hat{y}$)*	*Cost*	*Target (true) value (y)*	*Interpretation*	*Meaning*
		0.03		0		
		0.02		0		
		0.52		1		A_4
		0.03		0		
		0.04		0		
		0.13		0	Argmax	
		0.05		0		
		0.02		0		
		0.03		0		
		0.07		0		
		0.04		0		
		0.02	Binary cross-entropy	0		
Multi multiclass single label	Sigmoid					
		0.05		0		
		0.02		0		
		0.04		0		
		0.03		0		
		0.02		0		
		0.05		0	Argmax	
		0.53		1		F_4
		0.06		0		
		0.03		0		
		0.09		0		
		0.03		0		
		0.05	Binary cross-entropy	0		

(Summary of) Principles of Loss Functions

- Probability theory + Information theory

See also Maximum likelihood principle



- Intuition:
  - Information content (Likely Event) : Low
  - Information content (Unlikely Event) : High

- Self-information: I(x) = log (1/P(x)) = - log P(x)
- Ex: I(note=B) = - log P(note=B)

- Entropy of Probability distribution : Σ_i I(note=$Note_i$), weighted by P(note=$Note_i$)
- H(note) = Σ_i P(note=$Note_i$) I(note=$Note_i$)

- Expectation-based alternative definition:
- Expectation: Mean value of f(x) when x~P: $E_{x\sim P}$ [f(x)] = Σ_x P(x) f(x)
- H(note) = $E_{note\sim P}$ I(x) = $E_{note\sim P}$ [- log P(note)] = - $E_{note\sim P}$ [log P(note)]

KL-Divergence and Cross-Entropy

- Measures of Differences between Distributions (over a same variable: note)
  - Assymetric $D_{KB}(P||Q) =/= D_{KB}(Q||P)$ $H(P,Q) =/= H(Q,P)$

- Kullback-Leibler Divergence (KL- Divergence):
- $D_{KB}(P||Q) = E_{note \sim P}$ [log P(note)/Q(note)] = $E_{note \sim P}$ [log P(note) - log Q(note)]

- Categorical Cross-Entropy:
- $H(P,Q) = E_{note \sim P}$ [- log Q(note)] = - $E_{note \sim P}$ [log Q(note)]

- Difference with KL-Divergence: log P(note) term, constant with respect to Q

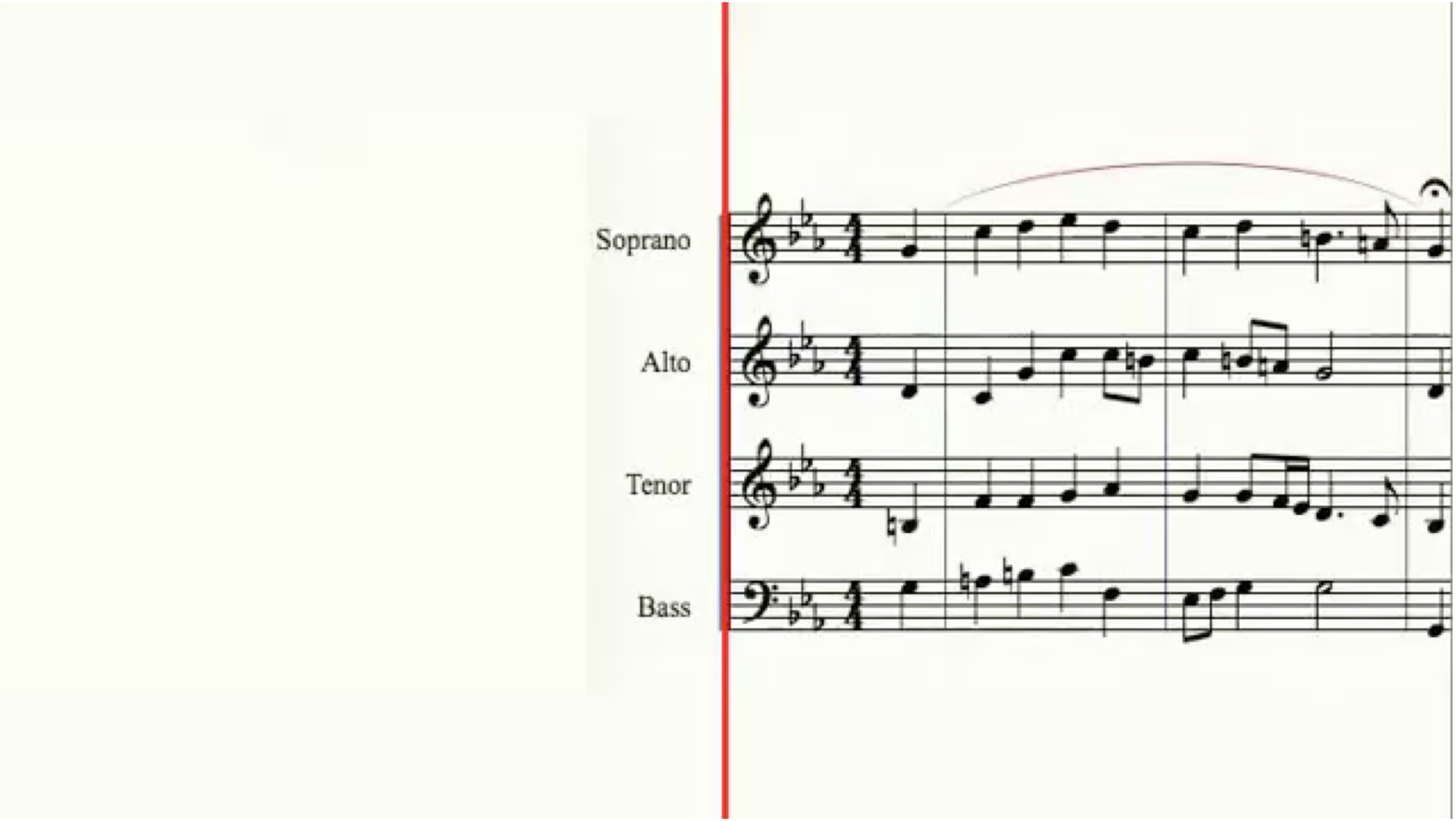
- $D_{KB}(y||\hat{y}) = E_{note \sim P} [\log y - \log \hat{y}] = \sum_i y_i (\log y_i - \log \hat{y}_i)$
- $H(y, \hat{y}) = - E_{note \sim P} [\log \hat{y}] = - \sum_i y_i \log \hat{y}_i$

- Binary Cross-Entropy:
- $H_B(y, \hat{y}) = - (y_0 \log \hat{y}_0 + y_1 \log \hat{y}_1) = - (y \log \hat{y} + (1-y) \log (1-\hat{y}))$

ForwardBach Brazilian Hymn Counterpoint

ForwardBach Brazilian Hymn Counterpoint (2 times slower and removing the intro)

DeepBach – Demo [Hadjeres, 2017]



https://www.youtube.com/watch?v=QiBM7-5hA6o

Reorchestration of God Save the Queen by DeepBach [Hadjeres, 2018]

https://www.youtube.com/watch?time_continue=1&v=x-W0ixD9Cpg