Lecture 6: Logical foundations

1 Introduction: different formal approaches

2 Penalty logic

3 Penalty logic and Bayesian networks

4 Penalty logic and Dempster-Shafer theory

5 Penalty logic and neural nets

6 Learning

1 Introduction: different formal approaches

- Brewka (1994); Besnard, Mercer & Schaub (2002) [for a copy go to http://www.cs.uni-potsdam.de/wv/pdfformat/bemesc02a.pdf]: Optimality Theory through Default Logic with priorities. The priorities are handled by a total ordering defined on the system of defaults. See also Nicolas Rescher's (1964) book "Hypothetical reasoning" which clearly expresses the very same idea.
- Dick de Jongh & Fenrong Liu (2006). They take an approach in terms of priority sequences of logical expressions, an idea that comes close to Brewka (1994).
- Pinkas (1992) introduced penalty logic and used it to model high-level (logical) properties of neural networks (see also Pinkas, 1995)
- Lima et al. (Lima, Morveli-Espinoza, & Franca, 2007) improve on it.
- Prince (2002) and Pater et al. (2007; 2007) compare OT hierarchies and systems with weighted constraints.

2 Penalty logic

The presentations follows Darwiche & Marquis (2004) and Blutner (2004). Let's consider the language $\mathcal{L}_{\mathrm{At}}$ of propositional logic (referring to the alphabet At of atomic symbols).

Definition 1: A triple $<\mathrm{At}, \Delta, k>$ is called a *penalty knowledge base* (PK) iff (i) Δ is a set of consistent sentences built on the basis of At (the possible hypotheses); (ii) $k: \Delta \Rightarrow (0, \infty)$ (the penalty function).

Intuitively, the penalty of an expression δ represents what we should pay in order to get rid of δ. If we pay the requested price we no longer have to satisfy δ. Hence, the larger $k(\delta)$ is, the more important δ is.

From some PK we can extract the system $W = \{[\alpha, k(\alpha)]: \alpha \in \Delta\}$ which is called the *weighted base* of the system PK (see Darwiche & Marquis)

Definition 2: Let α be a formula of our propositional language $\mathcal{L}_{At}$. A *scenario of* α *in PK*(*W*) is a subset Δ' of Δ such that $\Delta' \cup \{\alpha\}$ is consistent. The cost $K_{PK}(\Delta')$ of a scenario Δ' in PK is the sum of the penalties of the formulas of PK that are not in Δ':

$$K_{PK}(\Delta') = \sum_{\delta \in (\Delta - \Delta')} k(\delta)$$

Definition 3: An *optimal scenario of* α *in PK* is a scenario the cost of which is not exceeded by any other scenario (of α in PK), so it is a penalty minimizing scenario. With regard to a penalty knowledge base PK, the following cumulative consequence relation can be defined:

$\alpha \mid\sim_{PK} \beta$ iff β is an ordinary consequence of each optimal scenario of α in PK.

Hence, penalties may be used as a criterion for selecting preferred consistent subsets in an inconsistent knowledge base, thus inducing a non-monotonic inference relation.

Example 1

Weighted base W: $\{\langle a \wedge b, 2\rangle, \langle \neg b, 1\rangle\}$

Optimal scenario for a in W:
$\Delta_1 = \{a \wedge b\}$ $K_{PK}(\Delta_1) = 1$

Optimal scenario for ¬a in W: (violating $a \wedge b$ or b, respectively)
$\Delta_2 = \{\neg b\}$ $K_{PK}(\Delta_2) = 2$

$a \mid\!\sim_W b$

$\neg a \mid\!\sim_W \neg b$

Example 2

First Law: A robot may not injure a human being.
Second Law: A robot must follow (obey) the orders given it by human beings, except where such orders would conflict with the First Law.
Third Law: A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Weighted base W

$\neg I$	5	(first law)
F	2	(second law)
P	1	(third law)
$(S \wedge F) \rightarrow K$	1000	(S: giving the order to kill her)
$K \rightarrow I$	1000	(K: the robot kills her)

Two scenarios for S in W (violating F and $\neg$I, respectively)

$\Delta_1 = \{\neg I, P, (S \wedge F) \rightarrow K, K \rightarrow I\}$ $K_{PK}(\Delta_1) = 2$

$\Delta_2 = \{F, P, (S \wedge F) \rightarrow K, K \rightarrow I\}$ $K_{PK}(\Delta_2) = 5$

$S \mid\!\sim_W \neg I$

Semantics

Consider a *penalty knowledge base* PK = <At, Δ, k>. Let ν denote an ordinary (total) interpretation for the language $\mathcal{L}_{At}$ (ν: At$\rightarrow$\{0,1\}). The usual clauses apply for the evaluation $[\![.]\!]_\nu$ of the formulas of $\mathcal{L}_{At}$ relative to ν. The following function indicates how strongly an interpretation ν conflicts with the space of hypotheses Δ of a penalty knowledge base PK:

Definition 4 (system energy of an interpretation)

$$\mathcal{E}_{PK}(\nu) =_{def} \sum_{\delta \in \Delta} k(\delta) [\![\neg\delta]\!]_\nu$$

$\mathcal{E}_{PK}(\nu)$ is also called *violation rank* (Pinkas), *cost* (deSaint-Cyr et al.), *weight* (Darwiche & Marquis) of the interpretation.

Example 1 again

Weighted base W: $\{\langle a \wedge b, 2\rangle, \langle \neg b, 1\rangle\}$.
Let us consider the following four interpretations over the variables appearing in *W*, Var(*W*):

- $\nu 1 = (a, b)$ $\mathcal{E}_{PK}(\nu 1) = 1$
- $\nu 2 = (a, \neg b)$ $\mathcal{E}_{PK}(\nu 2) = 2$
- $\nu 3 = (\neg a, b)$ $\mathcal{E}_{PK}(\nu 3) = 3$
- $\nu 4 = (\neg a, \neg b)$ $\mathcal{E}_{PK}(\nu 4) = 2$

Hence, the interpretation with minimum energy is $\nu 1$.

Preferred models

Let α be a wff of the language $\mathcal{L}_{At.}$ An interpretation ν is called a *model* of α just in case $[\![\alpha]\!]_\nu = 1$.

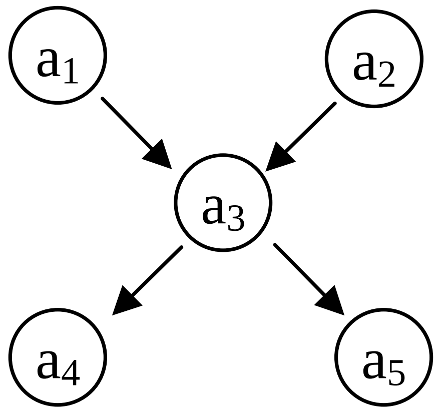
Definition 4

A *preferred model* of α is a model of α with minimal energy $\mathcal{E}$ (with regard to the other models of α). As the semantic counterpart to the syntactic notion $\alpha \mid\!\sim_{PK} \beta$ given in Definition 3 we can define the following relation:
$\alpha \mid\!\approx_{PK} \beta$ iff each preferred model of α is a model of β.

As a matter of fact, the syntactic notion (Definition 3) and the present semantic notion (21) coincide. Hence, the logic is sound and complete. A proof can be found in Pinkas (1995).

Example 1, continued: $a \mid\!\approx b$; $\neg a \mid\!\approx \neg b$.

3 Penalty logic and Bayesian networks

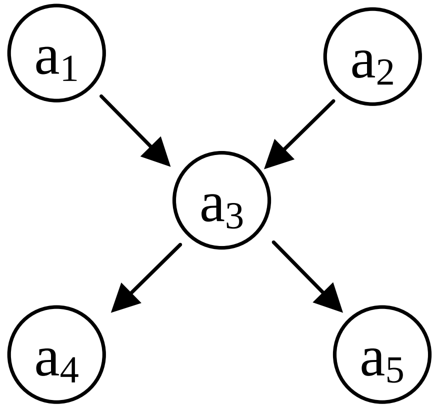


Consider a Bayesian network with binary random variables $a_1, a_2, \ldots, a_n$. Consider a partial specification of these random variables described by a set of "interpretations" V. Let α be a conjunction of literals (atoms or their negation) that describes this set V, i.e. $V = \{\nu: \nu(\alpha) = 1\}$.

Finding a most probable world model: find the specification of the random variables that maximizes the probability $\mu(\nu)$ of the joint distribution; in other words, find argmax $_{\nu \in V}[\mu(\nu)]$.

Example: $\alpha = a_1 \wedge \neg a_2$, find an optimal specification of the random variables $\{a_3, a_4, a_5\}$ maximizing the joint probability $\mu(a_1 = 1, a_2 = 0, a_3 = 0/1, a_4 = 0/1, a_5 = 0/1)$. Of course, the concrete solution depends on the details of the conditioned probability tables.

Global semantics and finding a most probable world model (Kooij, 2006)



$$\mu(a_1, \ldots, a_n) = \prod_{i=1}^{n} \mu(a_i / \text{Parents}(a_i))$$

In the example:

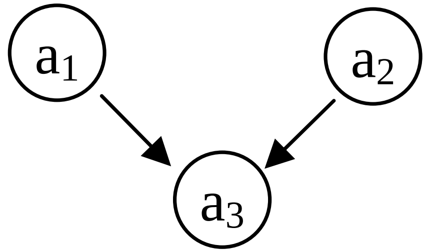
$\mu(a_1, \ldots, a_5) = \mu(a_1) \cdot \mu(a_2) \cdot \mu(a_3/a_1,a_2) \cdot \mu(a_4/a_3) \cdot \mu(a_5/a_3)$

$$\begin{aligned} &\text{argmax}_{\nu \in V}\ \mu(a_1 = \nu(a_1), \ldots, a_n = \nu(a_n)) \\ &= \text{argmax}_{\nu \in V}\ \mu(\nu) \\ &= \text{argmin}_{\nu \in V}\ -\log \mu(\nu) \\ &= \text{argmin}_{\nu \in V}\ \sum_{i=1}^{n} -\log \mu(a_i = \nu(a_i) / \text{Parents}(a_i) = \nu(\ldots)) \end{aligned}$$

The log-terms will be interpreted as penalties of corresponding rules:

$$\langle (\wedge_{x \in Parents(a_i)} x = \nu(x)) \rightarrow a_i = \nu(a_i),\ -\log \mu(a_i = -\nu(a_i) / \text{Parents}(a_i) = \nu(\ldots)) \rangle$$

Example



Consider the weighted rules connected with the a_3-part of the CPTs:

a_1	a_2	$\mu(a_3 = T / a_1, a_2)$	weighted rule for $a_3 = T$
F	F	0.8	$\langle \neg a_1 \wedge \neg a_2 \rightarrow a_3, -\log 0.2 \rangle$
F	T	0.4	$\langle \neg a_1 \wedge a_2 \rightarrow a_3, -\log 0.6 \rangle$
T	F	0.5	$\langle a_1 \wedge \neg a_2 \rightarrow a_3, -\log 0.5 \rangle$
T	T	0.3	$\langle a_1 \wedge a_2 \rightarrow a_3, -\log 0.7 \rangle$

a_1	a_2	$\mu(a_3 = F / a_1, a_2)$	weighted rule for $a_3 = F$
F	F	0.2	$\langle \neg a_1 \wedge \neg a_2 \rightarrow \neg a_3, -\log 0.8 \rangle$
F	T	0.6	$\langle \neg a_1 \wedge a_2 \rightarrow \neg a_3, -\log 0.4 \rangle$
T	F	0.5	$\langle a_1 \wedge \neg a_2 \rightarrow \neg a_3, -\log 0.5 \rangle$
T	T	0.7	$\langle a_1 \wedge a_2 \rightarrow \neg a_3, -\log 0.3 \rangle$

The mapping theorem

Assume a Bayesian network is mapped into a penalty knowledge base in the indicated way. Then finding a most probable world model of a conjunction of literals α and finding a *preferred model* (minimal energy) of α with regard to the penalty knowledge base are equivalent tasks (leading to the same optimal interpretation)

Comment

Looking for preferred models in penalty logic can be interpreted as a kind of qualitative reasoning in Bayesian networks. Which values of a set of random variables give a maximal probability for a given specification α of a proper subset of these random variables? The concrete probability value for the specification α doesn't matter. What counts is the optimality of the assignment.

4 Penalty logic and Dempster-Shafer theory

Dempster-Shafer theory is a theory of *evidence*. There are different pieces φ_i of evidence that give rise to a certain belief function and a (dual) plausibility function. Different pieces of evidence can be combined by means of Dempster's rule of combination.

A standard application is in medical diagnostics where some positive test result X can give a positive evidence for some disease Y but a negative test result gives absolutely no evidence for or against the disease.

Definition (mass function)

A mass function on a domain Ω of possible worlds (for a given piece of information) is a function m: $2^W \rightarrow [0, 1]$ such that the following two conditions hold:

$$\mathrm{m}(\emptyset) = 0.$$

$$\Sigma_{V \subseteq \Omega}\, \mathrm{m}(V) = 1$$

Definition (belief/plausibility function based on m)

Let m be a mass function on Ω. Then for every $U \subseteq \Omega$:

$$\mathrm{Bel}(U) =_{\mathrm{def}} \Sigma_{V \subseteq U}\, \mathrm{m}(V)$$
$$\mathrm{Pl}(U) =_{\mathrm{def}} \Sigma_{V \cap U \neq \emptyset}\, \mathrm{m}(V)$$

Dempster's rule of combination

Suppose m_1 and m_2 are basic mass functions over W. Then $m_1 \oplus m_2$ is given by Dempster's combination rule without renormalization:

$$m_1 \oplus m_2\,(U) = \Sigma_{V_i \cap V_j = U}\, m_1(V_i) \cdot m_2(V_j)$$

Facts:

Assume $m(U) = \oplus_{i=1}^{n} m_i\,(U)$; Pl plausibility function based on m; Pl_i plausibility function based on m_i. Then we have:

W

1. $Pl(\{\nu\}) = \sum_{\substack{V \\ \nu \in V}} m(V)$; $Pl_i(\{\nu\}) = \sum_{\substack{V \\ \nu \in V}} m_i(V)$
2. $Pl(\{\nu\}) = \prod_{i=1}^{n} Pl_i(\{\nu\})$ ["contour function"]

Relating penalties to Dempster-Shafer theory

Let be $W = \{[\alpha_i, k(\alpha_i)]: \alpha_i \in \Delta\}$ a *weighted base* of a system PK in our language $\mathcal{L}_{At}$.

Each formula α_i represents a piece of evidence for $V_i = \{\nu: \nu \models \alpha_i\}$. Formally, this is expressed by the following mass function m_i:

$$m_i(V_i) = 1 - e^{-k(\alpha i)}\ ;\ m_i(\Omega) = e^{-k(\alpha i)}$$

Using facts 1 and 2 it can be shown that[1]

$$Pl(\{\nu\}) = e^{-\mathcal{E}_{PK}(\nu)}$$

This brings to light a relation between penalties and evidence where each formula of the knowledge base is considered to be given by a distinct source, this source having a certain probability to be faulty, and all sources being independent.

[1] For a proof see deSaint-Cyr, Lang, & Schiex (1994).

5 Penalty logic and neural nets

Main thesis: Certain activities of connectionist networks can be interpreted as nonmonotonic inferences. In particular, there is a strict correspondence between Hopfield networks and penalty/reward nonmonotonic inferential systems. There is a direct mapping between the information stored in such (localist) neural networks and penalty/reward knowledge bases.

- Certain logical systems are singled out by giving them a "deeper justification".
- Understanding Optimality Theory: Which assumptions have a deeper foundation and which ones are pure stipulations?
- New methods for performing nonmonotonic inferences: Connectionist methods (simulated annealing etc.)

Hopfield network - fast dynamics

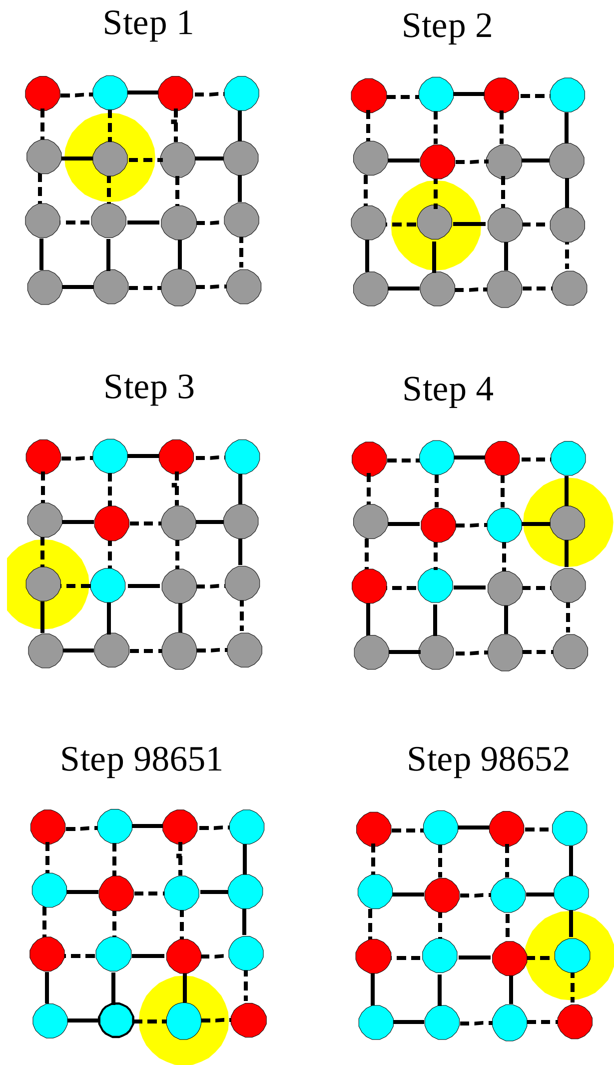
Let the interval [-1,+1] be the *working range* of each neuron

+1: maximal firing rate
0: resting
-1 : minimal firing rate)

$S = [-1, 1]^n$
$w_{ij} = w_{ji}$, $w_{ii} = 0$

ASYNCHRONOUS UPDATING:

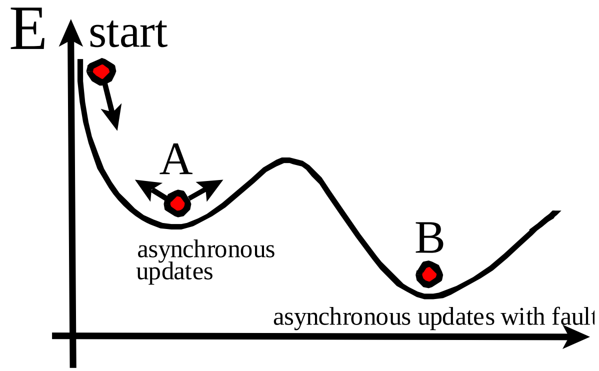
$$s_i(t+1) = \begin{cases} \theta\left(\Sigma_j w_{ij} \cdot s_j(t)\right), & \text{if } i = \text{rand}(1,n) \\ s_i(t), & \text{otherwise} \end{cases}$$



Summarizing the main results

Theorem 1 (Cohen & Großberg 1983)
Hopfield networks are resonance systems (i.e. $\lim_{n\to\infty} f^n(s)$ exists and is a resonance for each $s \in S$ and $f \in F$)

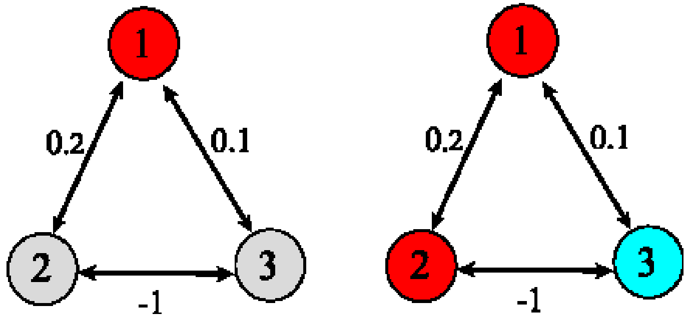
Theorem 2 (Hopfield 1982)
$E(s) = -\frac{1}{2} \sum_{i,j} w_{ij}\, s_i\, s_j$ is a *Ljapunov-function* of the system in the case of asynchronous updates. The output states $\lim_{n\to\infty} f^n(s)$ can be characterized as *the local minima* of E



Theorem 3 (Hopfield 1982)
The output states $\lim_{n\to\infty} f^n(s)$ can be characterized as *the global minima* of E if certain stochastic update functions f are considered (faults!).

Example

$$w = \begin{pmatrix} 0 & 0.2 & 0.1 \\ 0.2 & 0 & -1 \\ 0.1 & -1 & 0 \end{pmatrix}$$



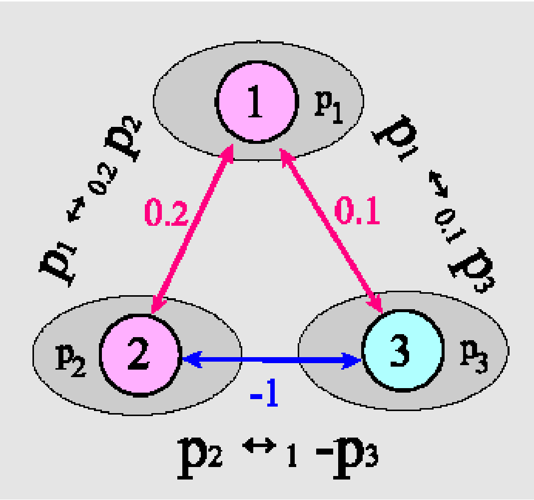
Input Output

$E(s) = -0.2s_1s_2 - 0.1s_1s_3 + s_2s_3$

			E
<1 0 0>	≤	<1 0 0>	0
		<1 0 1>	-0.1
		<1 1 0>	-0.2
		<1 1 1>	0.7
		<1 1-1>	-1.1

$ASUP_w(<1\ 0\ 0>) = \min_E(s) = <1\ 1\ {-1}>$

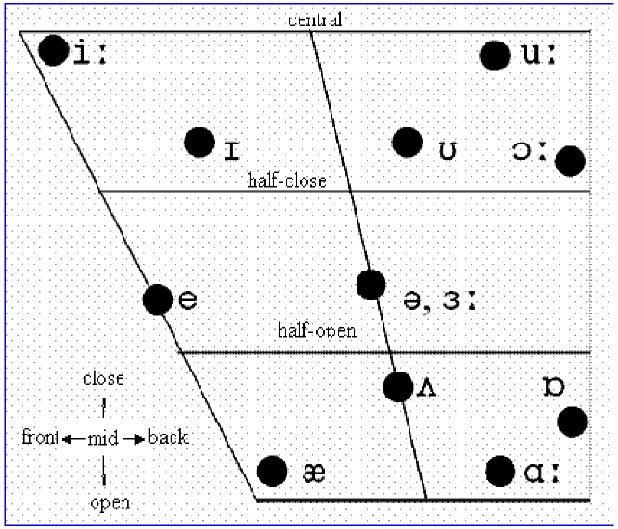
The correspondence between symmetric networks and penalty knowledge bases



1. relate the nodes of the networks to atomic symbols a_i of $\mathcal{L}_{At.}$ $At = \{p_1,p_2,p_3\}$
2. translate the network in a corresponding weighted base $W = \{\langle p_1 \leftrightarrow p_2, 0.2\rangle, \langle p_1 \leftrightarrow p_3, 0.1\rangle, \langle p_2 \leftrightarrow \neg p_3, 1\rangle\}$
3. relate states and interpretations: $s \cong \nu$ iff $s_i = \nu(a_i)$
4. observe that the energy of a network state is equivalent to the energy of an interpretation: E(s) $= \mathcal{E}_{PK}(\nu) =_{def} \sum_{\delta\in\Delta} k(\delta)\, [\![\neg\delta]\!]_\nu$ E.g.:

E(<1 1 1>) = 0.7 = –0.2–0.1+1
E(<1 1-1>) = –1.1 = –0.2+0.1–1
…

Example from phonology

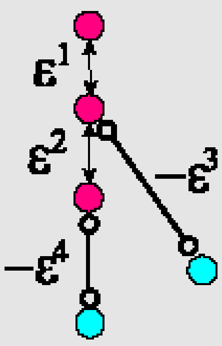


–back	+back	
/i/	/u/	+high
/e/	/o/	–high/–low
/æ/	/ɔ/ /a/	+low

The phonological features may be represented as by the atomic symbols BACK, LOW, HIGH, ROUND. The generic knowledge of the phonological agent concerning this fragment may be represented as a Hopfield network using *exponential weights* with basis $0 < \varepsilon \leq 0.5$.

Exponential weights and strict constraint ranking

Strong Constraints: LOW → ¬HIGH; ROUND → BACK



VOC	/a/	/i/	/o/	/u/	/ɔ/	/e/	/æ/
BACK	+	−	+	+	+	−	−
LOW	+	−	−	−	+	−	+
HIGH	−	+	−	+	−	−	−
ROUND	−	−	+	+	+	−	−

Assigned Poole-system

VOC ↔ε^1 BACK; BACK ↔ε^2 LOW

LOW ↔ε^4 ¬ROUND; BACK ↔ε^3 ¬HIGH

Keane's markedness conventions

Conclusion

- As with weighted logical system, OT looks for an optimal satisfaction of a system of conflicting constraints
- The exponential weights of the constraints realize a strict ranking of the constraints:
- Violations of many lower ranked constraints count less than one violation of a higher ranked constraint.
- The grammar doesn't count!

6 Learning

Translating connectionist and standard statistic methods of learning into an update mechanism of a penalty logical system.

Boersma & Hayes (2001): gradual learning algorithm (stochastic OT)
Goldwater & Johnson (2003): maximum entropy model
Jäger (2003): Comparison between these two models
Pater, Bhatt & Potts (2007)

These papers are also a starting point for understanding iterated learning and the modelling of (cultural) language evolution.

References

Blutner, R. (2004). *Neural Networks, Penalty Logic and Optimality Theory*. Amsterdam: ILLC.

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry, 32*, 45-86.

Darwiche, A., & Marquis, P. (2004). Compiling propositional weighted bases. *Artificial Intelligence, 157*, 81-113.

de Jongh, D., & Liu, F. (2006). *Optimality, Belief and Preference*: Institute for Logic, Language and Computation (ILLC), University of Amsterdam.

deSaint-Cyr, F. D., Lang, J., & Schiex, T. (1994). Penalty logic and its link with Dempster-Shafer theory, *Proceedings of the 10th Int. Conf. on Uncertainty in Artificial Intelligence (UAI'94)* (pp. 204-211).

Goldwater, S., & Johnson, M. (2003). *Learning OT constraint rankings using a maximum entropy model*. Paper presented at the Stockholm Workshop on Variation within Optimality Theory, Stockholm.

Jäger, G. (2003). Maximum entropy models and Stochastic Optimality Theory. Potsdam: University of Potsdam.

Kooij, J. F. P. (2006). Bayesian Inference and Connectionism. Penalty Logic as The Missing Link, *Essay written for the course "Neural Networks and Symbolic Reasoning"*. Amsterdam.

Lima, P. M. V., Morveli-Espinoza, M. M. M., & Franca, F. M. G. (2007). *Logic as Energy: A SAT-Based Approach* (Vol. 4729). Berlin, Heidelberg: Springer.

Pater, J., Bhatt, R., & Potts, C. (2007). Linguistic optimization. *Ms., University of Massachusetts, Amherst*.

Pater, J., Potts, C., & Bhatt, R. (2007). Harmonic Grammar with linear programming. *Ms., University of Massachusetts, Amherst.[ROA-827]*.

Pinkas, G. (1992). *Logical inference in symmetric connectionist networks*. Unpublished Doctoral thesis, Washington University, St Louis, Missouri.

Pinkas, G. (1995). Reasoning, connectionist nonmonotonicity and learning in networks that capture propositional knowledge. *Artificial Intelligence*, *77*, 203-247.

Prince, A. (2002). Anything goes, *A new century of phonology and phonological theory* (pp. 66–90). New Brunswick: Rutgers University.

Rescher, N. (1964). *Hypothetical Reasoning*: North-Holland Pub. Co.