Triggering, Hill-Climbing and

Can a stochastic trigger-based learner

William G. Sakas
City University
Graduate

the Conservative Learner:
afford Greediness as a constraint?

and Janet Dean Fodor
of New York
Center

Goals of Language Learning Theory: (0)

- a learning system that is guaranteed to converge on the target grammar
- …. and do so in polynomial time (= number of input sentences)

Background (1)

Theory of grammars:

- Universal principles and (binary) parameters
- Noiseless input (no ungrammatical sentences)
- No memory for past inputs or grammars (no batch processing)

Mathematical perspective:

- the learning algorithm may be viewed as a Markov process, in which each state represents a language licensed by a grammar (see, for example, Berwick & Niyogi, 1996)

The Greediness Constraint (2)

The learner shifts to a new grammar only if the new grammar licenses the current input (see, for example, Gibson & Wexler – 1994)

Unconstrained Error Driven Learner (UED Learner):

a stochastic learner that shifts to a new grammar (randomly selected) if and only if the current grammar does not license the current input

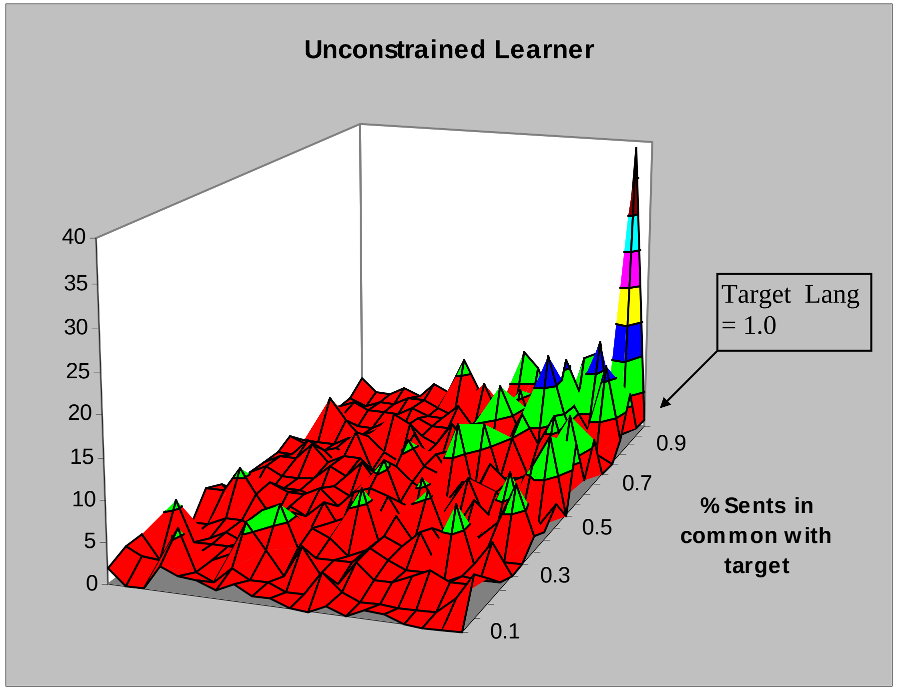
Our Claims (3)

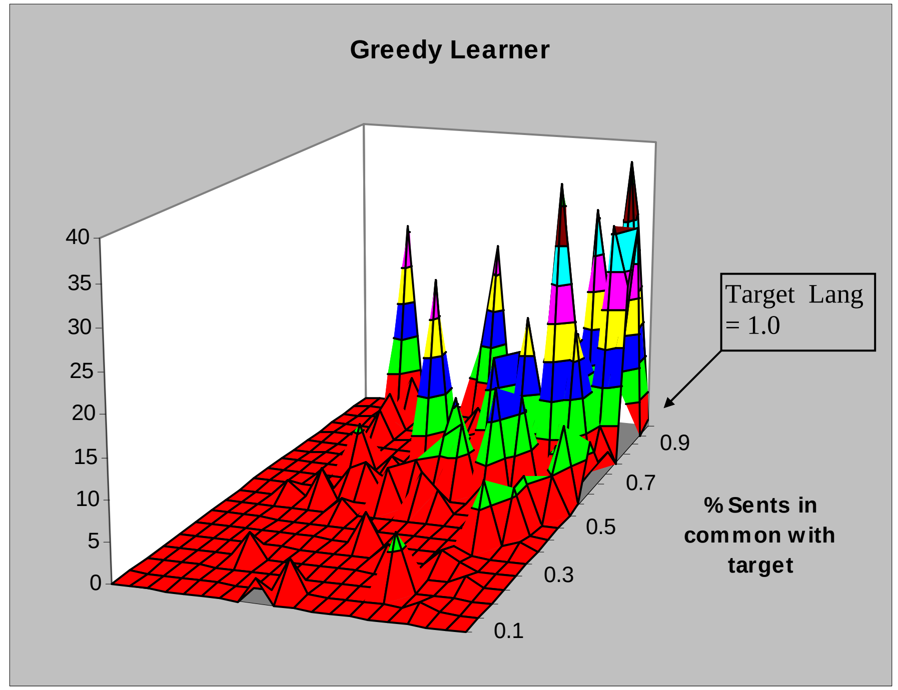
1) Adding the Greediness Constraint to an Unconstrained Error Driven Learner can only increase the time to convergence – regardless of the language space.

2) The UED learner requires a number of inputs that is exponential in the number of parameters, and is therefore implausible as a model for human learning.

3) Therefore, the UED with the Greediness constraint is exponential and implausible.

Greediness biases the learner's search (4) towards the area around the target.





The X-Y plane depicts language states of increasing similarity with the target language.
The vertical Z axis depicts the number of inputs the learner consumes while in state (x,y).
The graphs reflect data from one representative simulation trial.

The Paradox of Greediness (5)

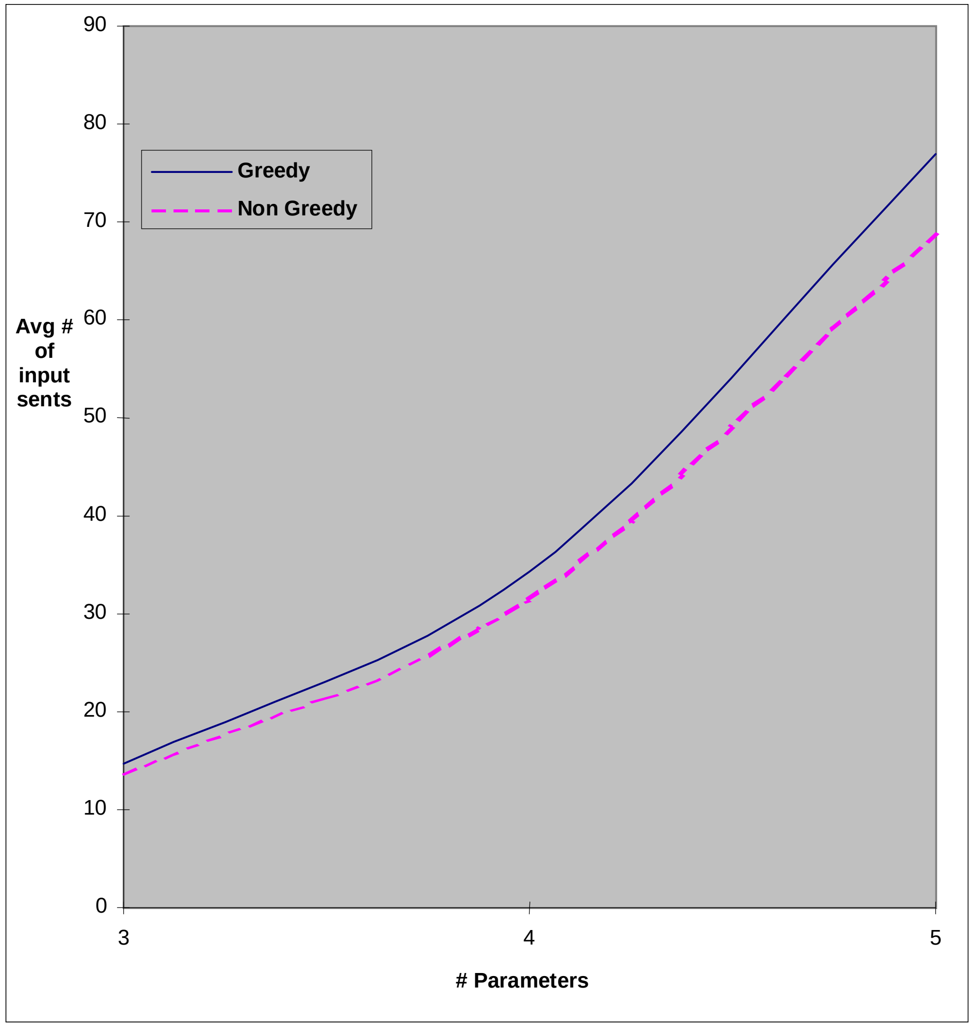
- **Perception:** Over time, Greediness will increase the *probability* that the current grammar *is* the target grammar
- **Reality:** Over time, Greediness increases the *similarity* of the current grammar* *to* the target grammar

But (perhaps counter-intuitively) -

As the similarity between the current grammar and the target grammar increases, the learner is less likely to encounter an input trigger that will shift it to the target.

*If there is not a smoothness relationship between grammars and languages, then technically Greediness favors similarity of languages

Simulation of performance with and without Greediness: (6)



Experiment:

1K trials on each of 1K randomly generated language spaces
3, 4, or 5 parameters in each space
12 sentences in each target language
1-11 sentences in each non-target language

The non-greedy learner consumes less sentences than the greedy one - at least for up to 5 parameters.

BUT – only small spaces can be explored practicably in this way.

Informal Summary of Argument (7)

1) Start with a non-greedy learner that, on average, attains the target with N inputs.
2) Add Greediness. The effect is to decrease the frequency of shifting from one grammar to another.
3) This conservatism directs the search, but does so at the cost of shifting less frequently.
4) The benefit gained by Greediness does not overcome the cost of less frequent shifting.

The learner with Greediness attains the target in N + X steps where X depends on the cost of NOT shifting

Outline of Proof: (8)

$\boldsymbol{\pi}$ = probability that the learner picks a particular grammar G_i (here π is constant)
$\boldsymbol{\alpha_i}$ = probability that the current input can be parsed by G_i

Let U = the transient sub-matrix of the transition matrix that describes the UED Learner.
Probability of a shift from G_i to $G_{j,}$ for the UED = P (G_i~G_j) = $\pi(1-\alpha_i)$

Let K = a matrix, which when added to U, describes Greediness applied to the UED learner.
k_{ij} = probability that the current input can be not be parsed by either G_i or G_j. times π
(note that $k_{ij} = k_{ji}$)

Probability of a shift from G_i to $G_{j,}$ for the Greedy learner = probability that the current input *s* can be parsed by G_j given that *s* cannot be parsed by G_i = P(G_i~G_j) = $\pi(1-\alpha_i) - k_{ij}$

U	G_0	G_1	G_2
G_0	$\boldsymbol{\alpha_0}$	$\boldsymbol{\pi(1-\alpha_0)}$	$\boldsymbol{\pi(1-\alpha_0)}$
G_1	$\boldsymbol{\pi(1-\alpha_1)}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\pi(1-\alpha_1)}$
G_2	$\boldsymbol{\pi(1-\alpha_2)}$	$\boldsymbol{\pi(1-\alpha_2)}$	$\boldsymbol{\alpha_2}$

K	G_0	G_1	G_2
G_0	$\mathbf{k_{01}+k_{02}}$	$\mathbf{-k_{01}}$	$\mathbf{-k_{02}}$
G_1	$\mathbf{-k_{01}}$	$\mathbf{k_{01}+k_{12}}$	$\mathbf{-k_{12}}$
G_2	$\mathbf{-k_{02}}$	$\mathbf{-k_{12}}$	$\mathbf{k_{02}+k_{12}}$

U+K	G_0	G_1	G_2
G_0	$\boldsymbol{\alpha_0+k_{01}+k_{02}}$	$\boldsymbol{\pi(1-\alpha_0)-k_{01}}$	$\boldsymbol{\pi(1-\alpha_0)-k_{02}}$
G_1	$\boldsymbol{\pi(1-\alpha_1)-k_{01}}$	$\boldsymbol{\alpha_1+k_{01}+k_{12}}$	$\boldsymbol{\pi(1-\alpha_1)-k_{12}}$
G_2	$\boldsymbol{\pi(1-\alpha_2)-k_{02}}$	$\boldsymbol{\pi(1-\alpha_2)-k_{12}}$	$\boldsymbol{\alpha_2+k_{02}+k_{12}}$

Define $|\mathbf{X}|_\Sigma$ as the sum of all the elements of matrix X.

If the UED takes a shorter time to converge on average than the Greedy Learner, then:
| fundamental matrix of UED $|_\Sigma \leq$ | fundamental matrix of UED+Greediness $|_\Sigma$. or,

$| \mathbf{(I-U)'} = I+U+U^2+U^3+U^4 \text{ .... } |_\Sigma \leq | \mathbf{(I-(U+K))'} = I+(U+K)+(U+K)^2+(U+K)^3 \text{ .... } |_\Sigma$

expanding the right hand side, and rearranging the terms we have:
$| I+U+U^2+U^3+U^4+ \text{ .... } |_\Sigma \leq | I+U+U^2+\ldots+K+UK+KU+K^2+UUK+UKU+ \ldots\ldots |_\Sigma$

applying the fact that $|X+Y|_\Sigma = |X|_\Sigma+|Y|_\Sigma$ we're left with:
$|I|_\Sigma+|U|_\Sigma+|U^2|_\Sigma+\ldots. \leq |I|_\Sigma+|U|_\Sigma+|U^2|_\Sigma+\ldots.+|UK|_\Sigma+|KU|_\Sigma+|UUK|_\Sigma+|UKU|_\Sigma+|K^2|_\Sigma\ldots.$

this is obviously true if the $||_\Sigma$ **of each of the terms that involves a K is >= 0**.

We show that $|KX|_\Sigma = |XK|_\Sigma = |K^i|_\Sigma = 0$, and that $|UKU|_\Sigma$ is the sum of terms of the form $k_i(r_u-r_v)(c_u-c_v)$, where k_i is positive and r_x = sum of row *x* of U, and c_x = sum of column *x* of U. Since $r_u-r_v \leq 0 \Leftrightarrow c_u-c_v \leq 0$, for any row sum and column sum of U - each term is positive. And finally by induction, that the $||_\Sigma$ of the all terms bracketed by U on the left and right are positive.

Performance of the UED Learner without Greediness is Exponential (10)

- Assume that all languages have a certain percentage of sentences in common with the target language call this percentage α
- Assume n parameters; 2^n languages. From any non-target state the probability of attaining the target is: the probability that the current input is not licensed by the current grammar times the probability of picking the target state: $\mathbf{(1-\alpha) \bullet 1/(2^n - 1)}$
- Thus, on average, the number of inputs required is $\mathbf{(2^n - 1)/(1-\alpha)}$
- Note that the number of inputs required is exponential in the number of parameters.

Conclusions: (11)

- Greediness carries a processing cost: the learner must parse each novel sentence twice

- Greediness can only increase the number of sentences consumed by the UED Learner before convergence

- Greediness does not mitigate the inefficiency of error driven random walk learning

Future Research (12)

- Are there language learning systems for which greediness is beneficial? For example:

  Genetic Algorithms (Clark)
  Neural Networks (Elman)
  Cue-Based Learners (Lightfoot, Bertolo et al)
  Structural Trigger Learners (Fodor)

- Do the consequences of Greediness depend on the content of what is learned or the mechanism of learning?