INTERIOR-POINT METHODS AS AN ALTERNATIVE TO ESTIMATE THE CONTEXT-FREE GRAMMAR PARAMETERS

Autor: Esther Sofía Mamián López
Orientador: Dr. Aurelio Ribeiro Leite de Oliveira
Co-orientador: Dr. Fredy Angel Amaya Robayo

27 de abril de 2015

Conteúdo

Goal

Compute the probability of a sentence or sequence of words.

Application

- Machine Translation
  - P(high winds tonight) $>$ P(large winds tonight)
- Speech Recognition
  - P(I saw a van) $>>$ P(eyes awe of an)

Language model

- Let $\chi_i \in \Sigma^*$ a sentence or a word.
  A model that computes either of these $P(\chi)$ or $P(\chi_1, \chi_2, ..., \chi_n,)$, is called a language model.

Language model

- Let $\chi_i \in \Sigma^*$ a sentence or a word.
  A model that computes either of these $P(\chi)$ or $P(\chi_1, \chi_2, ..., \chi_n,)$, is called a language model.
- Better call **Grammar**

Context-Free Grammar

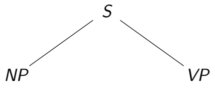
Context-free grammar were first used in the study of human language. One way understanding the relationship of terms such as *noun*, *verb* and *preposition* and their respective phrase leads to a natural recursion because noun phrases may appear inside verbs phrases and vice versa. Context-free grammar can capture important aspects of these relationships.

Grammar

1) $S \rightarrow NP\ VP$	7) $NP \rightarrow NP\ PP$
2) $PP \rightarrow P\ NP$	8) $NP \rightarrow astronomers$
3) $VP \rightarrow V\ NP$	9) $NP \rightarrow eyes$
4) $VP \rightarrow VP\ PP$	12) $NP \rightarrow saw$
5) $P \rightarrow with$	10) $NP \rightarrow starts$
6) $V \rightarrow saw$	11) $NP \rightarrow telescopes$

Grammar

1) $S \rightarrow NP\ VP$
2) $PP \rightarrow P\ NP$
3) $VP \rightarrow V\ NP$
4) $VP \rightarrow VP\ PP$
5) $P \rightarrow with$
6) $V \rightarrow saw$
7) $NP \rightarrow NP\ PP$
8) $NP \rightarrow astronomers$
9) $NP \rightarrow eyes$
12) $NP \rightarrow saw$
10) $NP \rightarrow starts$
11) $NP \rightarrow telescopes$



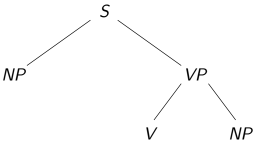
Language and Grammar

Context-free Grammar and Probabilistic Context-Free Grammar

Grammar

1) $S \rightarrow NP\ VP$
2) $PP \rightarrow P\ NP$
3) $VP \rightarrow V\ NP$
4) $VP \rightarrow VP\ PP$
5) $P \rightarrow with$
6) $V \rightarrow saw$

7) $NP \rightarrow NP\ PP$
8) $NP \rightarrow astronomers$
9) $NP \rightarrow eyes$
12) $NP \rightarrow saw$
10) $NP \rightarrow starts$
11) $NP \rightarrow telescopes$



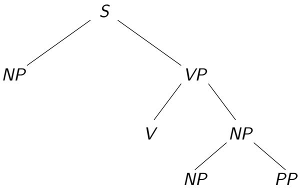
Language and Grammar

Context-free Grammar and Probabilistic Context-Free Grammar

Grammar

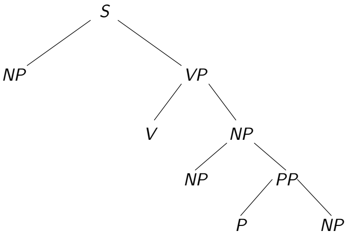
1) $S \rightarrow NP\ VP$
2) $PP \rightarrow P\ NP$
3) $VP \rightarrow V\ NP$
4) $VP \rightarrow VP\ PP$
5) $P \rightarrow with$
6) $V \rightarrow saw$

7) $NP \rightarrow NP\ PP$
8) $NP \rightarrow astronomers$
9) $NP \rightarrow eyes$
12) $NP \rightarrow saw$
10) $NP \rightarrow starts$
11) $NP \rightarrow telescopes$



Grammar

1) $S \rightarrow NP\ VP$	7) $NP \rightarrow NP\ PP$
2) $PP \rightarrow P\ NP$	8) $NP \rightarrow astronomers$
3) $VP \rightarrow V\ NP$	9) $NP \rightarrow eyes$
4) $VP \rightarrow VP\ PP$	12) $NP \rightarrow saw$
5) $P \rightarrow with$	10) $NP \rightarrow starts$
6) $V \rightarrow saw$	11) $NP \rightarrow telescopes$



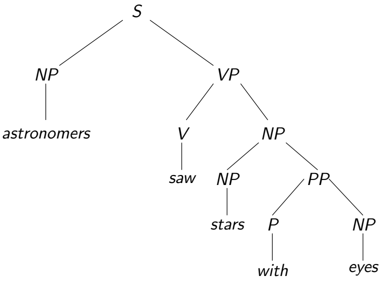
Language and Grammar

Context-free Grammar and Probabilistic Context-Free Grammar

Grammar

1) $S \rightarrow NP\ VP$
2) $PP \rightarrow P\ NP$
3) $VP \rightarrow V\ NP$
4) $VP \rightarrow VP\ PP$
5) $P \rightarrow with$
6) $V \rightarrow saw$

7) $NP \rightarrow NP\ PP$
8) $NP \rightarrow astronomers$
9) $NP \rightarrow eyes$
12) $NP \rightarrow saw$
10) $NP \rightarrow starts$
11) $NP \rightarrow telescopes$

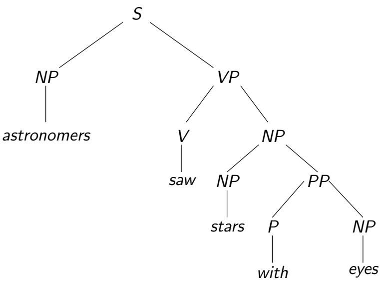


Grammar

1) $S \rightarrow NP\ VP$	7) $NP \rightarrow NP\ PP$
2) $PP \rightarrow P\ NP$	8) $NP \rightarrow astronomers$
3) $VP \rightarrow V\ NP$	9) $NP \rightarrow eyes$
4) $VP \rightarrow VP\ PP$	12) $NP \rightarrow saw$
5) $P \rightarrow with$	10) $NP \rightarrow starts$
6) $V \rightarrow saw$	11) $NP \rightarrow telescopes$

$\chi =$ astronomers saw stars with eyes
$\eta =$ saw saw saw with saw

Parser tree



Grammar

1) $S \rightarrow NP\ VP$	7) $NP \rightarrow NP\ PP$
2) $PP \rightarrow P\ NP$	8) $NP \rightarrow astronomers$
3) $VP \rightarrow V\ NP$	9) $NP \rightarrow eyes$
4) $VP \rightarrow VP\ PP$	12) $NP \rightarrow saw$
5) $P \rightarrow with$	10) $NP \rightarrow starts$
6) $V \rightarrow saw$	11) $NP \rightarrow telescopes$

A grammar describes a Language

Definition

A context-free grammar is a 4-tuple $G = (\Sigma, N, S, P)$ *where:*

- Σ *is a finite set of terminal symbols.*
- N *is a finite set of nonterminal symbols, with* $N \cap \Sigma = \emptyset$.
- $S \in N$ *is the start symbol.*
- P *is a finite set of rules/production of the form* $A \longrightarrow \gamma$, *where* $A \in N$ *and* γ *is a sentential form, i.e.* $\gamma \in (N \cup \Sigma)^*$.

Definition

A context-free grammar is in Chomsky Normal Form (CNF) if every rule is of the form $A \longrightarrow BC$ *e* $A \longrightarrow a$*, where* $A, B, C \in N$ *and* $a \in \Sigma$*.*

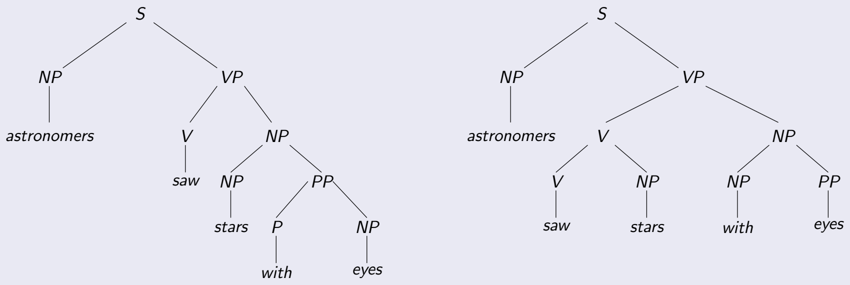
Definition

The language $L(G)$ *generate from* G *is the set of terminal strings that have derivation from start symbol. That is,*

$$L(G) = \{\chi \in \Sigma^{+} : S \overset{*}{\Rightarrow} \chi\},$$

where Σ^{+} *is a set of string/sentence in* Σ*.*

Ambiguity

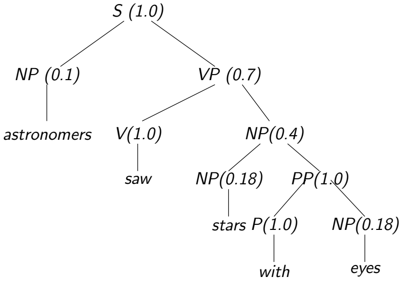

S
NP
VP
astronomers
V
NP
saw
NP
PP
stars
P
NP
with
eyes
S
NP
VP
astronomers
V
NP
V
NP
NP
PP
saw
stars
with
eyes

Definition

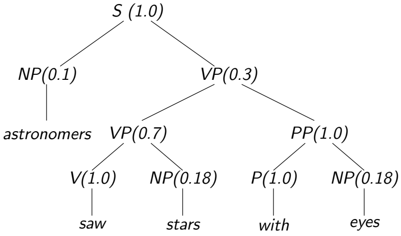
A Probabilistic Context-Free Grammar is a 5-tuple $G = (\Sigma, N, S, P, p)$ *where:*

- Σ *is a finite set called terminal symbols.*
- N *is a finite set called nonterminal symbols. Where* $N \cap \Sigma = \emptyset$*.*
- $S \in N$ *is the start symbol.*
- P *is a finite set of rules/production of the form* $A \longrightarrow \gamma$*, where* $A \in N$ *and* γ *is a sentential form .* $\gamma \in (N \cup \Sigma)^*$*.*
- p *is a probability function*
  - $p : P \rightarrow [0, 1]$
  - $\forall A \in N, \sum_{(A \longrightarrow \alpha) \in \Gamma_A} p(A \longrightarrow \alpha) = 1$

Parse tree t1



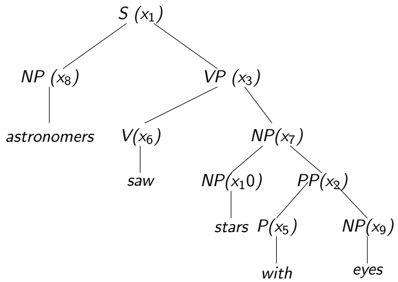
Parse tree t2



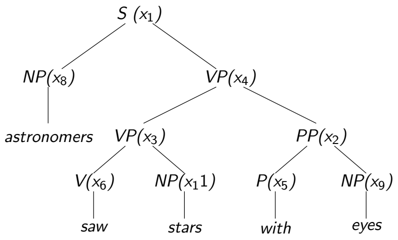
$$p(t1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$$
$$p(t2) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 1.0 \times 0.18 \times 1.0 \times 0.18 = 0.0006804$$

$$p(\chi) = p(t1) + p(t2) = 0.0015876$$

Parse tree t1



Parse tree t2



$$p(t1) = x_1 x_8 x_3 x_6 x_7 x_{10} x_2 x_5 x_9$$
$$p(t2) = x_1 x_8 x_4 x_3 x_2 x_6 x_{11} x_5 x_9$$

$$p(\chi) = p(t1) + p(t2) = x_1 x_8 x_3 x_6 x_7 x_{10} x_2 x_5 x_9 + x_1 x_8 x_4 x_3 x_2 x_6 x_{11} x_5 x_9$$

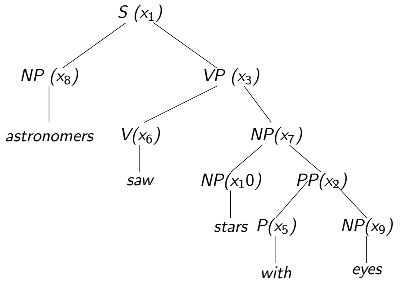
Learning Grammar/Training Grammar

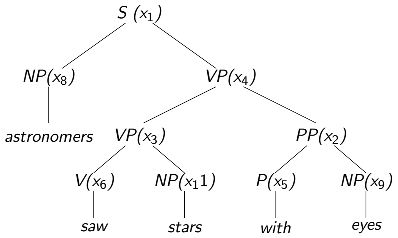
- A corpus Ω is a big finite set of sentences and words of a language.
- We train parameters of our model on a training set. A training set is a subset the corpus.
- We test the model's performance on data of corpus that we haven't seen.

Definition

In the context-free grammar the likelihood function of a language corpus Ω *is*

$$Pr(\Omega|G_p) = \prod_{\chi \in \Omega} Pr(\chi|G_p).$$





$p(t1) = x_1 x_8 x_3 x_6 x_7 x_{10} x_2 x_5 x_9$
$p(t2) = x_1 x_8 x_4 x_3 x_2 x_6 x_{11} x_5 x_9$

$p(\chi) = p(t1) + p(t2) = x_1 x_8 x_3 x_6 x_7 x_{10} x_2 x_5 x_9 + x_1 x_8 x_4 x_3 x_2 x_6 x_{11} x_5 x_9$
$p(\eta) = p(t1) + p(t2) + p(t3) = x_1 x_8 x_3 x_6 x_7 x_{10} x_2 + x_1 x_4 x_3 x_2 x_6 x_{11} + x_1^2 x_2 x_8 x_4 x_3^3$

likelihood function

If $\Omega = \{\chi, \eta\}$ we have that likelihood function is $Pr(\Omega | G_p) = p(\chi) \times p(\eta)$

Problema

$$\begin{aligned} \text{maximize} \quad & Pr(\Omega | G_p) \\ \text{such as} \quad & \sum_{x_i \in \Psi_A} x_i = 1 \\ & 0 \leq x_i \leq 1, \quad i = 1, ..., |P|, \end{aligned}$$

where Ψ_A is finite set of rules probabilities with the same antecedent.

$$\left\{\begin{array}{ll} \text{Maximize} & f(x) \\ \text{Such as} & Ax = b \\ & 0 \leq x \leq u \end{array}\right. \overset{\mu\to 0}{\Longleftrightarrow} \left\{\begin{array}{ll} \text{Minimize} & f(x) - \mu\left(\sum_{i=1}^{n} ln(x_i) + \sum_{i=1}^{n} ln(v_i)\right) \\ \text{Such as} & Ax = b \\ & x + v = u, \end{array}\right.$$

com $\mu > 0$.

where $f : \Re^n \longrightarrow \Re$, $x \in \Re^n$, $A \in \Re^{m\times n}$, $b \in \Re^m$ e $u \in \Re^n$.

For each fixed value μ and using the associated Lagrangiano

$$\mathcal{L}(x, v, y, w) = f(x) - \mu\left(\sum_{i=1}^{n} ln(x_i) + \sum_{i=1}^{n} ln(v_i)\right) + y^T(b - Ax) + w^T(u - v - x).$$

$$\begin{pmatrix} \nabla f(x) - \mu X^{-1}e + A^T y - w \\ w - \mu V^{-1}e \\ Ax - b \\ u - v - x \end{pmatrix} = 0.$$

$$\begin{pmatrix} H(x) + \mu X^{-2} & 0 & A^T & -I \\ 0 & \mu V^{-2} & 0 & I \\ A & 0 & 0 & 0 \\ I & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta v \\ \Delta y \\ \Delta w \end{pmatrix} = \begin{pmatrix} r_d \\ r_b \\ r_p \\ r_u \end{pmatrix},$$

$$\begin{pmatrix} H(x)+\mu X^{-2} & 0 & A^T & -I \\ 0 & \mu V^{-2} & 0 & I \\ A & 0 & 0 & 0 \\ I & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta v \\ \Delta y \\ \Delta w \end{pmatrix} = \begin{pmatrix} r_d \\ r_b \\ r_p \\ r_u \end{pmatrix},$$

onde $H(x) = \nabla^2 f(x)$, $r_d = \mu X^{-1} - A^T y - w - \nabla f(x)$, $r_b = \mu V^{-1} e - w$, $r_p = b - Ax$ e $r_u = u - x - v$

Algorithm 1.1: Barreira Logarítmica Canalizado

Dados: $(x^0, v^0) > 0,\ y^0,\ w^0,\ \sigma = \frac{1}{\sqrt{n}}\ k = 1$

inicio

Calcular o valor para $\mu = \sigma\frac{\gamma}{n}$, onde $\gamma = (v^k)^t w^k$;

Calcular as direções $\Delta x, \Delta y$ resolvendo o sistema linear:

$$\begin{pmatrix} \hat{H} & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_d + V^{-1}W(r_u - W^{-1}r_b) \\ r_p \end{pmatrix} \qquad \text{onde } \hat{H} = H(x) + \mu X^{-2} + V^{-1}W\ ;$$

Determinar α usando o teste da razão;

Atualizar o passo:

$$(x^{k+1}, v^{k+1}, y^{k+1}, w^{k+1}) = (x^k, v^k, y^k, w^k) + \alpha(\Delta x, \Delta v, \Delta y, \Delta w);$$

se *o critério de parada é satisfeito* **então**

  parar;

senão

  faça $k = k + 1$ e volte ao Passo 2;

fim se

fim

$$\begin{array}{cl} \text{minimize} & f(x) \\ \text{such as} & Ax = b \\ & 0 \leq x \leq u \end{array}$$

onde $f : \Re^n \longrightarrow \Re$, $x \in \Re^n$, $A \in \Re^{m\times n}$, $b \in \Re^m$ e $u \in \Re^n$.

$$\begin{pmatrix} \nabla f(x) - \mu X^{-1}e + A^T y - w \\ w - \mu V^{-1}e \\ Ax - b \\ x + v - u \end{pmatrix} = 0.$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{such as} & Ax = b \\ & 0 \leq x \leq u \end{array}$$

onde $f : \Re^n \longrightarrow \Re$, $x \in \Re^n$, $A \in \Re^{m \times n}$, $b \in \Re^m$ e $u \in \Re^n$.

z

$$XZe = \mu e \qquad \begin{pmatrix} \nabla f(x) - \mu X^{-1}e + A^T y - w \\ w - \mu V^{-1}e \\ Ax - b \\ x + v - u \end{pmatrix} = 0.$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{such as} & Ax = b \\ & 0 \leq x \leq u \end{array}$$

onde $f : \Re^n \longrightarrow \Re$, $x \in \Re^n$, $A \in \Re^{m \times n}$, $b \in \Re^m$ e $u \in \Re^n$.

$VWe = \mu e$

$$\begin{pmatrix} \nabla f(x) - z + A^T y - w \\ \cancel{w - \mu V^{-1} e} \\ XZe - \mu e \\ Ax - b \\ x + v - u \end{pmatrix} = 0.$$

$$\begin{array}{cl} \text{minimize} & f(x) \\ \text{such as} & Ax = b \\ & 0 \leq x \leq u \end{array}$$

onde $f : \Re^n \longrightarrow \Re$, $x \in \Re^n$, $A \in \Re^{m\times n}$, $b \in \Re^m$ e $u \in \Re^n$.

$$\begin{pmatrix} \nabla f(x) - z + A^T y - w \\ VWe - \mu e \\ XZe - \mu e \\ Ax - b \\ x + v - u \end{pmatrix} = 0, \quad (x, v, w, z) \geq 0, \quad \mu > 0.$$

$$\begin{pmatrix} H(x) & 0 & -I & A^T & -I \\ 0 & W & 0 & 0 & V \\ Z & 0 & X & 0 & 0 \\ A & 0 & 0 & 0 & 0 \\ I & I & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta v \\ \Delta z \\ \Delta y \\ \Delta w \end{pmatrix} = \begin{pmatrix} r_d \\ r_b \\ r_c \\ r_p \\ r_u \end{pmatrix}$$

$r_d = z - \nabla f(x) - A^t y - w$, $r_c = \mu e - XZe$ $r_b = \mu V^{-1} e - w$, $r_p = b - Ax$ e $r_u = u - x - v$

Algorithm 1.2: Primal-Dual Canalizado

Dados: y^0, $(x^0, v^0, w^0, z^0) > 0$, $\sigma = \frac{1}{2n}$, $k = 1$.

inicio

Escolher $\mu = (\frac{\gamma}{2n})$, onde $\gamma = (x)^t z + (v)^t w$;

Resolver o sistema linear

$$\begin{pmatrix} \hat{H} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_d - X^{-1} r_c + V^{-1}(r_b - w r_u) \\ r_p \end{pmatrix},$$

onde $\hat{H} = H(x) + X^{-1}Z + V^{-1}W$.

Determinar α usando o teste da razão *backtracking*;

Atualizar o passo $(x^{k+1}, v^{k+1}, z^{k+1}, y^{k+1}, w^{k+1}) = (x^k, v^k, z^k, y^k, w^k) + \alpha(\Delta x, \Delta v, \Delta z, \Delta y, \Delta w)$;

se *o critério de parada é satisfeito* **então**

parar;

senão

faça $k = k + 1$ e volte ao início;

fim se

fin

Numerical Result

Tamanho amostra	Method	Iteration	Time (sec)	Perplexity
20	Logarithmic Barrier	23	14,6692	3,3422
	Primal dual Logarithmic Barrier	9	8,2426	3,5171
500	Logarithmic Barrier	19	670, 4786	3,4382
	Primal dual Logarithmic Barrier	1	64, 7130	3,4422
4000	Logarithmic Barrier	20	12448, 4759	3,4382
	Primal dual Logarithmic Barrier	1	837, 2390	3,4382

Tabela : Numerical Result of the Grammar 1 (5,12,25)

Tamanho amostra	Method	Iteration	Time (sec)	Perplexity
20	Logarithmic Barrier	8	27,6678	19,4277
	Primal dual Logarithmic Barrier	8	28,4952	22,9811
500	Logarithmic Barrier	8	379, 3173	19,5804
	Primal dual Logarithmic Barrier	8	314, 8892	22,9811
4000	Logarithmic Barrier	17	28498, 3234	19,3310
	Primal dual Logarithmic Barrier	8	9031, 0355	23,0644

Tabela : Numerical Result of the Grammar 2 (14,13,47)

Numerical Result

Tamanho amostra	Method	Iteration	Time (sec)	Perplexity
20	Logarithmic Barrier	23	14,6692	3,3422
	Primal dual Logarithmic Barrier	9	8,2426	3,5171
500	Logarithmic Barrier	19	670, 4786	3,4382
	Primal dual Logarithmic Barrier	1	64, 7130	3,4422
4000	Logarithmic Barrier	20	12448, 4759	3,4382
	Primal dual Logarithmic Barrier	1	837, 2390	3,4382
		Perplexidade por palavra Inside-Outside		25, 424

Tabela : Numerical Result of the Grammar 1 (5,12,25)

Tamanho amostra	Method	Iteration	Time (sec)	Perplexity
20	Logarithmic Barrier	8	27,6678	19,4277
	Primal dual Logarithmic Barrier	8	28,4952	22,9811
500	Logarithmic Barrier	8	379, 3173	19,5804
	Primal dual Logarithmic Barrier	8	314, 8892	22,9811
4000	Logarithmic Barrier	17	28498, 3234	19,3310
	Primal dual Logarithmic Barrier	8	9031, 0355	23,0644

Tabela : Numerical Result of the Grammar 2 (14,13,47)

Conclusion

- The approach of interior point methods is very robust, converging when the value of stop criterion ϵ is extremely low.
- The times obtained for convergence, iteration and perplexity shows that further studies testing theses methods in big grammars are very promising, and eventually, it could be used as an alternative in natural language.

Future Propose

- To use C++ for implementing the methods.
- To implement a variant predictor-corrector Logarithmic Barrier Method and predictor-corrector Primal-Dual Logarithmic Barrier Method. The number of iterations required is expected to be lower.
- To search for another approaches to calculate the Hessian.
- To implement an hybrid approach between the Inside-Outside and Interior-Point methods.

Bibliografía

José-Miguel Benedí and Joan-Andreu Sánchez.
Estimation of stochastic context-free grammars and their use as language models.
Computer Speech and Language, 19(3):249–274, 2005.

AS El-Bakry, Richard A Tapia, T Tsuchiya, and Yin Zhang.
On the formulation and theory of the newton interior-point method for nonlinear programming.
Journal of Optimization Theory and Applications, 89(3):507–541, 1996.

Jhon Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman.
Introduction to Automata Theory, Languages, And Computation.
Addison-Wesley, segunda edition, 2001.

Johnny Ferney Ibarra and Astrid Yineeth Alvarez.
Uso del algoritmo inside-outside en la estimación de parámetros de una gramática incontextual probabilística, Universidad del Cauca. Departamento de Matemáticas, 2009.

Sanjay Mehrotra.
On the implementation of a primal-dual interior point method.
Siam Jornual on Optimization, 2(4):575–601, 1992.

Joan Andreu Sánchez Peiró.
Estimación de Gramáticas Incontextuales Probabilísticas y su Aplicación en Modelación del Lenguaje.
PhD thesis, Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y computación, 1999.

Stuart Russell and Peter Norvig.
Inteligência Artificial.
Elsevier, 2004.

Stephen J. Wright.
Primal-Dual Interior-Point Methods.
SIAM, 1997.