**Lý thuyết về Conditional Random Fields**

Trường ngẫu nhiên có điều kiện (CRF) là một lớp thuộc mô hình phân biệt thích hợp để dự đoán liên quan đến ngữ cảnh hoặc các trạng thái lân cận có ảnh hưởng tới kết quả dự đoán. CRFs được ứng dụng trong nhận dạng thực thể, dán nhãn POS, dự đoán gen, giảm tiếng ồn và nhận dạng vật thể.

**Mô hình phân biệt (Discriminative classifier)**

Các mô hình học máy có hai mảng chính là mô hình sinh mẫu (generative) và mô hình phân biệt. CRF là mô hình phân biệt. Mô hình tổng quát được các ranh giới giữa các class khác nhau. Còn mô hình sinh mẫu tổng quát hóa cách dữ liệu được sinh ra. Naive Bayes là một mô hình sinh mẫu, logistic regression là một mô hình phân biệt dựa trên Maximum Liklihood.

CRF ứng dụng Logistic Regression vào chuỗi đầu vào. CRFs có những điểm tương đồng với mô hình Hidden Markov, nhưng HMM là mô hình sinh mẫu.

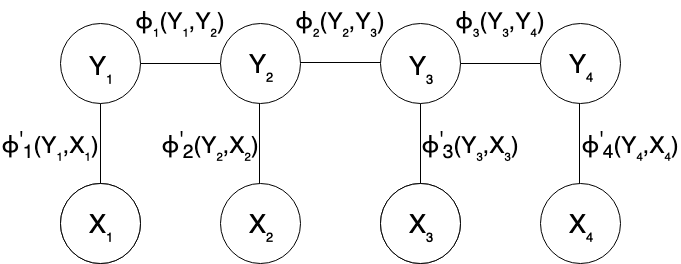
**Trường sinh ngẫu nhiên có điều kiện**

Giả sử ta có Markov Random Field và chia nó làm hai tập ngẫu nhiên là Y và X.

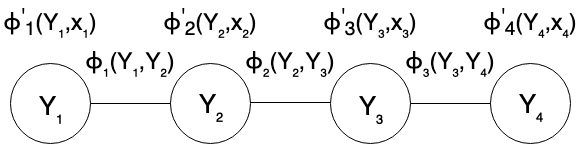
CRFs là trường hợp đặc biệt của Markov Random Field. Trong đồ đồ thị phải thỏa mãn:

* Khi ta xác định đồ thị dựa vào X. Ví dụ khi giá trị của biến ngẫu nhiên trong X không thay đổi, tất cả biến ngẫu nhiên trong tập Y được xác định dựa trên Markov p(Yᵤ/X,Yᵥ, u≠v) = p(Yᵤ/X,Yₓ, Yᵤ~Yₓ). Yᵤ~Yₓ tức Yᵤ and Yₓ are là lân cận trong đồ thị.

Ví dụ đồ thị thỏa mãn luật trên được biểu diễn qua cấu trúc chuỗi như sau:



Mô hình CRF tổng quát hóa xác suất P(Y/X) (X luôn được cho trước). Vì vậy đồ thị cũng có thể biểu diễn như sau:



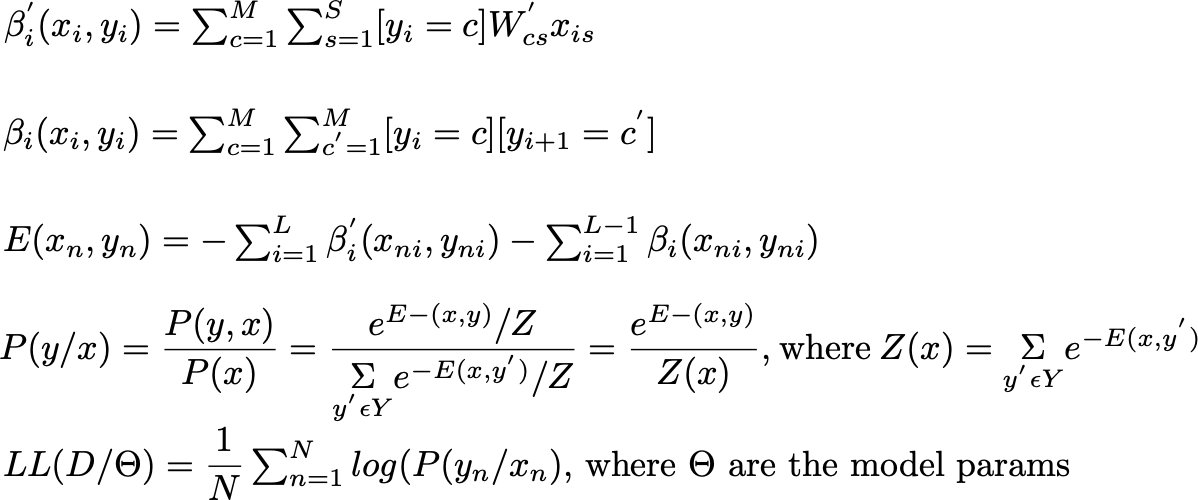
Vì ta đang tổng quát hóa dựa trên X và tìm kết quả tương ứng cho Yi với mỗi Xi, X và Y còn được gọi là biến dấu hiệu và biến nhãn.

**Lý thuyết CRF và tối ưu hóa lân cận**

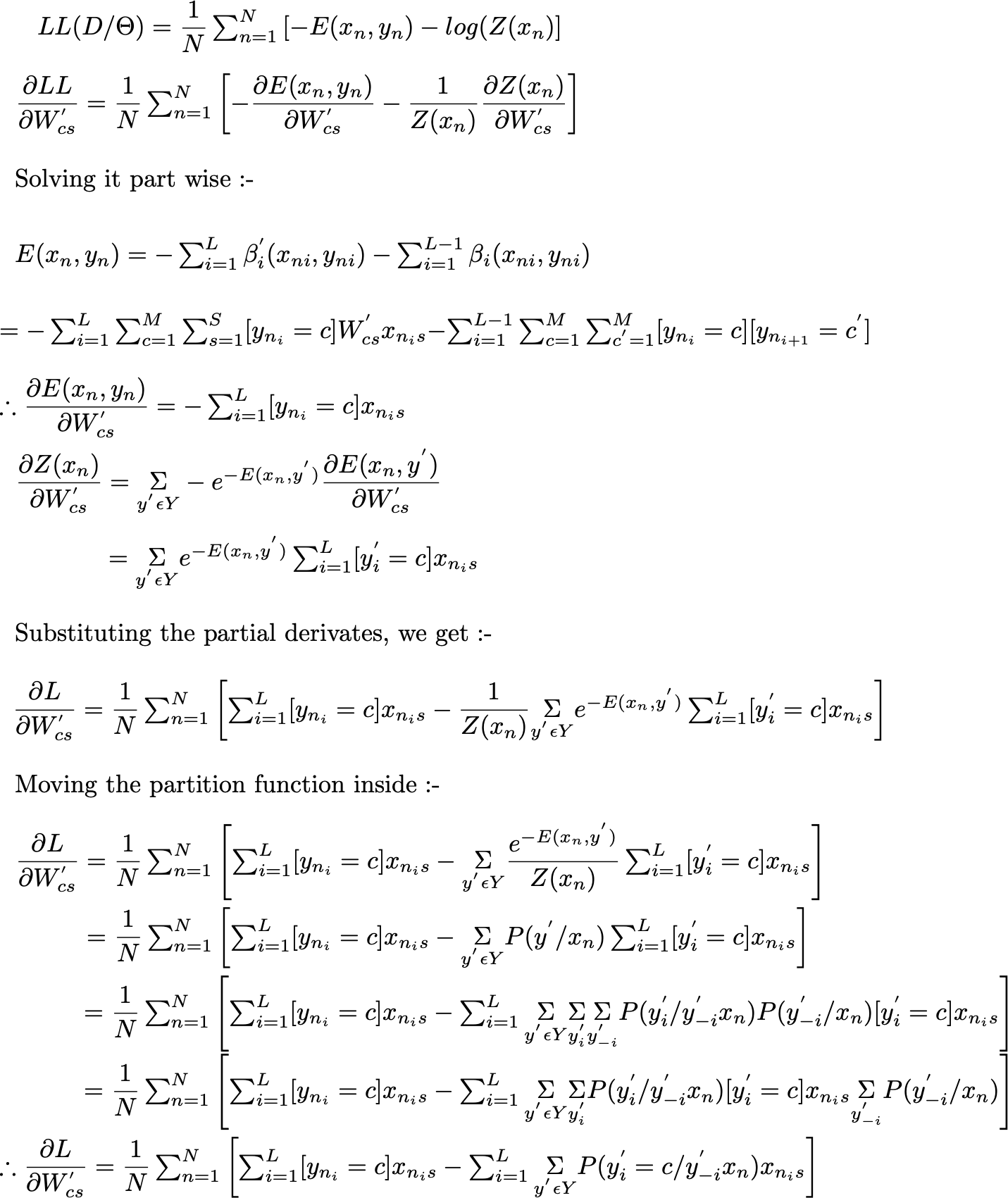
Đầu tiên ta cần định nghĩa parameter và phương trình cho các xác suất bằng cách sử dụng phân phối Gibbs:

1. Miền nhãn: giả sử những biến ngẫu nhiên trong tập Y có miền như sau: {m ϵ ℕ | 1≤m ≤M}. M là số tự nhiên.
2. Miền và cấu trúc của biến dấu hiệu: giả sử biến ngẫu nhiên thuộc tập X là vector số thực có kích thước là ∀ Xᵢ ϵ X, Xᵢ ϵ Rˢ.
3. L là chiều dài của chuỗi CRF. Ví dụ L nhãn và L biến dấu hiệu.
4. Đặt βᵢ(Yᵢ, Yⱼ) = Wcc’ nếu Yᵢ = c, Yⱼ = c’và j = i+1, 0 với những trường hợp còn lại.
5. Đặt β’ᵢ(Yᵢ, Xᵢ) = W’c . Xᵢ, nếu Yᵢ = c và 0 những trường hợp còn lại
6. Tổng số parameter là M x M + M x S. Ví dụ có một parameter cho mỗi phép chuyển nhãn (M x M phép chuyển nhãn có thể có) và có S parameter cho mỗi nhãn (M nhãn có thể có). Ta sẽ nhân với những biến đã có (vector có kích thước S) của nhãn đó.
7. Đặt D = {(xn, yn)} for n=1 to N là dữ liệu huấn luyện bao gồm N mẫu.

Theo đó, các lân cận có thể biểu diễn như sau:

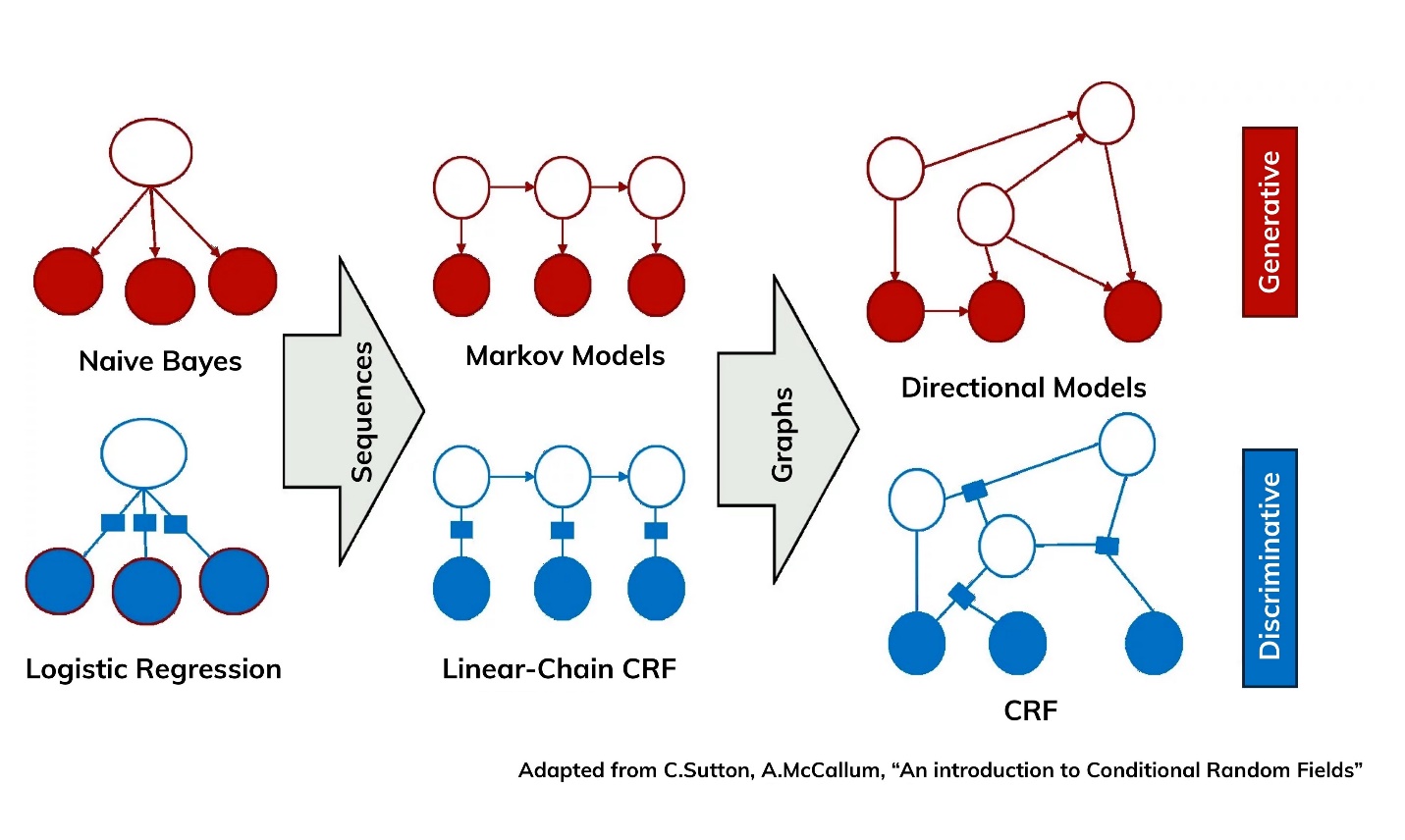


Độ dốc của log lân cận với W’cs được tính như sau:



**Khác biệt giữa CRFs và HMM**

Mặc dù đều sử dụng mô hình dữ liệu tuần tự nhưng CRFs và HMM là hai mô hình khác nhau. HMM là mô hình sinh mẫu, mô hình này đưa ra kết quả bằng cách mô hình hóa các phân phối xác suất. Trong khi CRFs là mô hình phân biệt và nó mô hình hóa các phân phối xác suất có điều kiện. CRFs không dựa vào các biến độc lập (các nhãn độc lập với nhau). HMM dựa trên Naive Bayes dựa trên Logistic Regression và CRFs cũng dựa trên Logistic Regression.



**Ứng dụng của CRFs**

CRFs thường được sử dụng trong xử lý ngôn ngữ tự nhiên trong đó có dán nhãn POS. Nhãn POS của một câu dựa trên những từ trước đó bằng cách sử dụng những hàm đặc trưng. Ta có thể sử dụng CRFs để phân biệt từ nào trong câu có POS nào tương ứng. Một ứng dụng khác tương tự là nhận dạng thực thể hoặc rút trích danh từ trong câu. CRF cũng được sử dụng để nhận dạng thành phần trong ảnh và dự đoán gene.