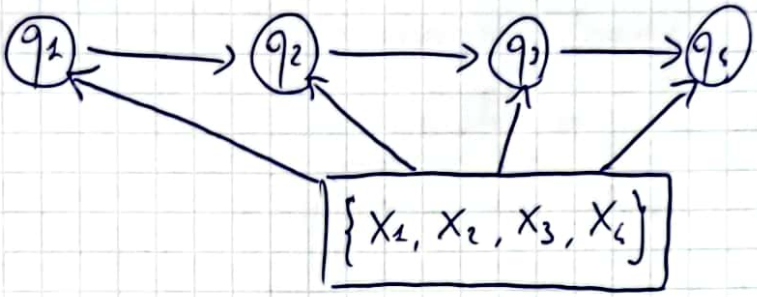
CONDITIONAL RANDOM FIELDS

HIDDEN MARKOV MODELS ⟶ GENERATIVE (or UNSUPERVISED)
as they estimate $\mathbb{P}(q, x)$

CRFs ⟶ DISCRIMINATIVE (or SUPERVISED)
as they estimate $\mathbb{P}(q \mid x)$

How is a CRF made?



LINEAR-CHAIN CRF: obeys Markov property.

GLOBAL CONDITIONING: all X_t influence all q_t.

q_t is the HIDDEN STATE at time t, in our love letter example it could be Q, R, D, or G.

$\{X_t\}_t$ is the OBSERVABLE SEQUENCE, the messages in the example.

CRF estimate:

$$\mathbb{P}(q_t \mid \{X_t\}) = \frac{1}{Z} \exp\left(\sum_i \lambda_i t_i(q_{t-1}, q_t) + \sum_i \mu_i s_i(q_t, X_t) \right)$$

Elements of the formula:

Z = normalization factor (to have a probability at the end).

λ_i, μ_i = parameters to be estimated from learning.

$t_i(q_{t-1}, q_t)$ = TRANSITION FEATURES

↳ for a training data instance in which $q_{t-1} = Q$ and $q_t = R$, we will have, for instance:

$t_{(Q,R)}(q_{t-1}, q_t) = 1$

$t_{(Q,D)}(q_{t-1}, q_t) = \emptyset$

$t_{(D,G)}(q_{t-1}, q_t) = \emptyset$

} all pairs that do not correspond to the transitions observed in data are $\emptyset$

$s_i(X_t, q_t)$ = STATE FEATURES

↳ these are the features of documents X_t, such as the ones we saw for love letters: length, count("?"), count("love"), etc.

exp * ⟶ The exponential function arises from the MAXIMUM ENTROPY PRINCIPLE

Intuitively: you shall impose no regularity on estimated distribution beyond what data tell you.

Formally: be as close as possible to the uniform distribution, i.e. maximize entropy.

Given functions $t_i(\cdot)$ and $s_i(\cdot)$, each data point can be expressed as a concatenation of those features *.

Ex:

count("?") = 1

count("love") = 1

length = 6

<< Do you really love me ? >>

$q_t = Q$, $q_{t-1} = D$ } $t_{(D,Q)} = 1$, $t_{(k,l)} = \emptyset$ $\forall (k,l) \neq (D,Q)$

Any iterative parameter estimation algorithm can be used to learn λ_i and μ_i. E.g.: LBFGS, SGD, ...