Probabilistic Graphical Models
Christoph Lampert <chl@ist.ac.at>
TA: all of us <pgm_2016@lists.ist.ac.at>
Exercise Sheet 4 v1.0

1 Variational Inequality

Show: the difference between the left and the right hand side in the variational lower bound

$$\log p(x;\theta) \geq \mathbb{E}_{z\sim q} \log p(x,z;\theta) - \mathbb{E}_{z\sim q} \log q(z) \tag{1}$$

is exactly the Kullback-Leibler divergence (in which direction?) between $q(z)$ and $p(h|x,\theta)$.

Note: this shows as well that the inequality is an equality for $q(z) = p(h|x,\theta)$.

2 Stochastic Matrices

A matrix $M = (m_{ij})_{ij}$ with $m_{ij} \geq 0$ is called *stochastic* if $\sum_i m_{ij} = 1$ for every j. Let λ be an eigenvalue of M with eigenvector v. Show: if $\sum_i v_i > 0$, then $\lambda = 1$.

3 Maximum Entropy Distributions (if we got there in the lecture)

Derive at the same level of formality as in the lecture:

a) the maximum entropy distribution on $\mathbb{R}$ with mean 0 and variance 1

b) the maximum entropy distribution on $[0,\infty) \subset \mathbb{R}$ with mean 1

The maximum entropy distribution might not exist for some conditions. Try to find

c) the maximum entropy distribution on $\mathbb{R}$ with mean 1. What does wrong?

4 Inference in Hidden Markov Models

4.1 Analysis

a) Imagine an HMM with continuous emissions according to a Gaussian distribution. Derive the EM update formulae for the means and covariances.

b) Derive an algorithm to efficiently compute the gradient of the log likelihood for an HMM with discrete transition and emission matrices.

4.2 Implementation

Implement an HMM in the *Burglar* situation:

- There are 25 latent states (corresponding to a 5×5 floor grid).
- For each grid position the probability of a creak floor or bumping into furniture is given by these maps:

'creaks' map

'bumps' map

Dark squares means probability 0.9, light squares means probability 0.1. Creaks and bumps occur independently.

- In each time step, the burglar will move to any possible neighboring square *top, down, left* or *right* with equal probability.
- In the first step, the burglar might be anywhere with equal probability.
- You make the following 10 observations:

t	1	2	3	4	5	6	7	8	9	10
creaks	n	y	n	y	n	y	y	y	y	y
bumps	y	n	n	y	n	y	n	n	y	y

a) Compute the filtered probabilities $p(h_t|v_{1:t})$ for every $t = 1, \ldots, T$, and visualize the result

b) Compute the smoothed probabilities $p(h_t|v_{1:t})$ for every $t = 1, \ldots, T$, and visualize the result

c) Assume that the burglar also has the option to stay where he is with the same probability as each possible movement. Rerun a) and b) accordingly.

d) Assume that you are certain that the burglar is not in the house before $t = 1$. Which probabilities change? Rerun a) and b) accordingly.

e) Assume that you observe either only the cracks, or only the bumps. How do the filtered and smoothed probabilities change?

5 Learning Hidden Markov Models

5.1 Analysis

a) In the implementation part below, the goal is to estimate the emission and transition probability tables a, A and B, from observation sequences. Before running the code: write down an estimate how many sequences will be needed to get estimate with ±10% of the true values?

b) The EM algorithm requires an initialization. Show that EM does not work (well), if you initialize all unknown values to uniform constants (that fulfill the conditions, e.g. $a_i = \frac{1}{H}$ where H is the number of hidden states, etc).

c) There are situations in which the EM algorithm does not work well. Can you design a setting for which is EM is useless? (if not: there's a hint at the end of the sheet)

c) A few years ago, a new method for HMM learning was suggested: [Daniel Hsu, Sham M. Kakade, and Tong Zhang. *A spectral algorithm for learning hidden Markov models*. Journal of Computer and System Sciences, 78(5):1460-1480, 2012.], available at `http://www.cs.columbia.edu/~djhsu/`.

Read the paper and answer the following questions: what's the main claim? How does the proposed algorithm differ from EM? Does the new algorithm perform maximum likelihood estimation? If not, what else does it optimize?

Hint for 5.1c): Try making v_t a deterministic function of h_t and θ.

5.2 Implementation

We want to learn the probabilities of the above burglar situation from observed sequences, $v^1, \ldots, v^N$.

a) Adopt the above HMM with starting state at one of the boundary locations. Write a routine that can generate random trajectories for the burglar and corresponding observation sequences of arbitrary length. Create $N = 10, 100, 1000, 10000$ observation sequences of length 10.

b) Implement the HMM-EM algorithm and run it on the $N = 10, 100, 1000, 10000$ sequences. Observe how the probability estimates change?

c) Make plots of the Kullback-Leibler divergences and the L^2 distance (w.r.t. to the parameter entries) of the true a, A, B and p and their respective estimates. Discuss the results.

d) For $N = 10, 100, 1000, 10000$, repeat Exercise 4a) and 4b) with the estimated probabilities. How do the results differ?

e)* Assume that you could influence the length of the observation sequences while keeping the total number of observations. What would you prefer? More shorter sequences, or fewer longer sequences?

f)* Repeat d) with sequences of different lengths.