Statistical Machine Learning (BE4M33SSU)
Lecture 11.

Czech Technical University in Prague

- Why sampling?
- Rejection sampling, importance sampling
- Markov chain models
- Markov chain Monte Carlo sampling

11.1 Why sampling?

- Prediction for graphical models & structured output predictors can be computationally hard

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \Big[\sum_{ij \in E} g_{ij}(y_i, y_j) + \sum_{i \in V} q_i(y_i, x_i) \Big],$$

where $\boldsymbol{y}$ is a labelling of the nodes of a graph (V, E).

Recall: all learning approaches (ERM, EM-algorithm ...) require to solve prediction tasks in each iteration of the learning algorithm.

- Bayesian estimation & risk minimisation

$$p(\theta \mid \mathcal{T}^m) = \frac{p(\mathcal{T}^m \mid \theta) p(\theta)}{\int p(\mathcal{T}^m \mid \theta')\, p(\theta')\, d\theta'}$$

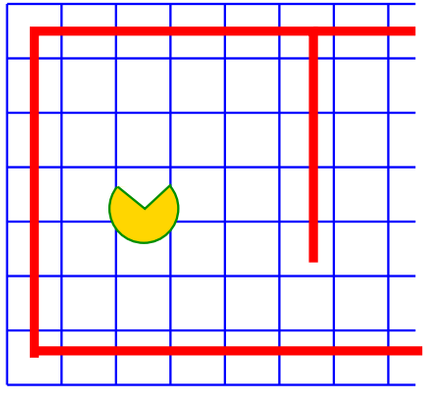
How to compute the normalisation and expected values w.r.t. $p(\theta \mid \mathcal{T}^m)$?

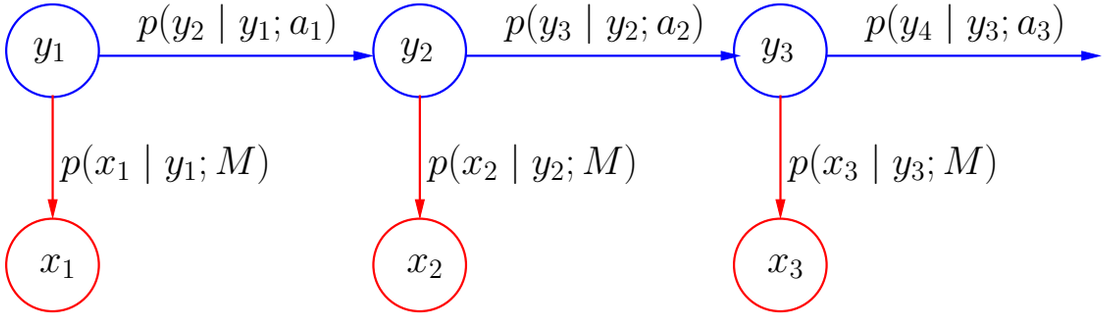
$$\int p(\mathcal{T}^m \mid \theta)\, p(\theta)\, d\theta = ?$$

$$\mathbb{E}(f(\theta)) = \int f(\theta) p(\theta \mid \mathcal{T}^m) d\theta \sim \int f(\theta) p(\mathcal{T}^m \mid \theta) p(\theta) d\theta = ?$$

11.1 Why sampling?

Example 1. Simultaneous localisation and mapping (SLAM) in robotics





- Sequence of poses $y_t \in \mathcal{Y}$, $t = 1, 2 \ldots$, where $\mathcal{Y}$ is a discrete pose space, e.g. $\mathbb{Z}^2 \times S^1$ in the simplest case,
- Sequence of measurements $x_t \in \mathcal{X}$, $t = 1, 2 \ldots$, e.g. laser range-finder, sonar, cameras,
- Sequence of action commands a_t, $t = 1, 2 \ldots$, e.g. mixture of rotation and "move forward" commands,
- Environment map M describes walls, doors, POIs etc. in coordinates of $\mathcal{Y}$.
- Statistical models $p(y_{t+1} \mid y_t; a_t)$ and $p(x_t \mid y_t; M)$.

11.1 Why sampling?

Prediction task: Given y_1, $\boldsymbol{x}_{1:t}$ and $\boldsymbol{a}_{1:t}$ estimate the pose y_t. This requires to solve

$$p(y_t \mid \boldsymbol{x}_{1:t}; \boldsymbol{a}_{1:(t-1)}, M) \sim \sum_{y_2} \dots \sum_{y_{t-1}} \prod_{i=1}^{t} \Big[p(y_i \mid y_{i-1}; a_{i-1})\, p(x_i \mid y_i; M) \Big]$$

$$\sim p(x_t \mid y_t; M) \sum_{y_{t-1}} p(y_t \mid y_{t-1}; a_{t-1})\, p(y_{t-1} \mid \boldsymbol{x}_{1:t-1}; \boldsymbol{a}_{1:t-2}, M)$$

and to predict y_t (depending on the loss function) e.g. by

$$y_t^* = \sum_{y_t \in \mathcal{Y}} y_t\, p(y_t \mid \boldsymbol{x}_{1:t}; \boldsymbol{a}_{1:(t-1)})$$

The pose space $\mathcal{Y}$ is huge and the summation over it is difficult. Moreover, the first formula provides $p(y_t \mid \dots)$ up to an unknown normalisation constant only.

Q: Is it possible to simplify the computations by maintaining a sample of positions y_t^ℓ, $\ell = 1, \dots, N$ as a surrogate of the distribution $p(y_t \mid \dots)$ and by recomputing it recursively from $t \mapsto t+1$?

11.2 Monte Carlo Sampling

The idea of sampling is simple, e.g. computing the expectation of a random variable f

$$\mathbb{E}(f) = \int_{\mathbb{R}^n} f(x)p(x)dx$$

- draw an i.i.d. sample x^ℓ, $\ell = 1, \ldots, N$ from $p(x)$ and
- approximate

$$\mathbb{E}_N(f) = \frac{1}{N}\sum_{\ell}^{N} f(x^\ell) \xrightarrow{N\to\infty} \mathbb{E}(f) = \int_{\mathbb{R}^n} f(x)p(x)dx$$

- recall Hoeffding's inequality: if $a \leqslant f(x) \leqslant b$, then

$$\mathbb{P}\Big(|\mathbb{E}_N(f) - \mathbb{E}(f)| > \epsilon\Big) \leqslant 2\exp\left(-\frac{2N\epsilon^2}{b-a}\right)$$

If $p(x)$ is a standard distribution $\Rightarrow$ it is straightforward to sample from it.

If $p(x)$ is complicated $\Rightarrow$ more sophisticated techniques needed.

11.2 Monte Carlo Sampling

Sample from $p(x)$ by sampling from another, easy-to-sample distribution $q(x)$.

Rejection sampling:

- Assume that $p(x)$ is known up to a proportionality constant,
- the proposal distribution $q(x)$ satisfies $Mq(x) \geqslant p(x)$, $\forall x$ for some M.

Use the following accept/reject procedure

1. sample $x^\ell \sim q(x)$ and $u \sim \mathcal{U}_{(0,1)}$.
2. If $u < \frac{p(x^\ell)}{Mq(x^\ell)}$ then accept x^ℓ and increment the counter ℓ. Otherwise reject.

The accepted x^ℓ are sampled with probability $p(x)$.

Limitations: Especially in high dimensional cases bounding $p(x)$ by $Mq(x)$ can be inefficient, i.e. M too large. The acceptance probability will be too small as a consequence

$$Pr(x \text{ accepted}) = Pr\left(u < \frac{p(x)}{Mq(x)}\right) = \frac{1}{M}$$

11.2 Monte Carlo Sampling

Importance sampling:

Consider an arbitrary proposal distribution $q(x)$ such that its support contains the support of $p(x)$. Rewrite $\mathbb{E}(f)$

$$\mathbb{E}(f) = \int_{\mathbb{R}^n} f(x)\, p(x) dx = \int_{\mathbb{R}^n} f(x) \frac{p(x)}{q(x)} q(x) dx = \int_{\mathbb{R}^n} f(x)\, w(x)\, q(x) dx,$$

where $w(x)$ is known as *importance weight*.

Draw an i.i.d. sample from $q(x)$ and approximate the expectation by

$$\hat{\mathbb{E}}_N(f) = \frac{1}{N} \sum_{\ell=1}^{N} f(x^\ell) w(x^\ell)$$

Importance sampling is still possible if the normalising constant of $p(x)$ is unknown

$$\tilde{\mathbb{E}}_N(f) = \frac{\sum_{\ell=1}^{N} f(x^\ell) w(x^\ell)}{\sum_{\ell=1} w(x^\ell)} = \sum_{\ell=1}^{N} f(x^\ell) \tilde{w}(x^\ell)$$

Limitations: high dimensional cases

11.3 Background: Markov chain models

Example 2 (random walk on a graph)

Given an weighted directed graph (V, E, w) with non-negative weights $w_{ij} \geqslant 0$. Consider a random walk on this graph:

- fix a probability distribution $p(s_1)$ for the first state $s_1 \in V$,
- fix transition probabilities

$$p(s_t = i \mid s_{t-1} = j) = \frac{w_{ij}}{\sum_{k \in \mathcal{N}_j} w_{kj}}$$

- the distribution $p(s_t)$ for the position $s_t \in V$ at time t is given by

$$p(s_t) = \sum_{s_{t-1} \in V} p(s_t \mid s_{t-1})\, p(s_{t-1})$$

Q: How does the distribution $p(s_t)$ behave for $t \to \infty$? (Example: Google's page-rank)

11.3 Background: Markov chain models

Definition A sequence of random variables $\boldsymbol{s} = s_1, s_2, \ldots$, with values $s_i \in K$, is a *Markov Chain Model* if

$$p(s_1, \ldots, s_n) = p(s_1) \prod_{t=2}^{n} p(s_t \mid s_{t-1})$$

holds for any $n > 1$. A Markov chain model is homogeneous if the transition probabilities $p(s_t \mid s_{t-1})$ do not depend on t.

Shorter notations:

- Matrix P with elements $P_{kk'} = p(s_t = k \mid s_{t-1} = k')$ denoting transition probabilities.
- Vector $\boldsymbol{\pi}_t \in \mathbb{R}_+^K$ with components $p(s_t = k)$.

From the definition follows

$$\boldsymbol{\pi}_{t+1} = P\boldsymbol{\pi}_t = P^t \boldsymbol{\pi}_1$$

- A Markov model is *irreducible* if there is a non-zero probability to reach state $k \in K$ beginning in any state $k' \in K$, i.e. for any pair $k, k' \in K$ exists $\tau \geqslant 1$ s.t. $P^\tau_{kk'} > 0$.
- A state $k \in K$ of a Markov model is *a-periodic* if the greatest common denominator of its return times τ such that $P^\tau_{kk} > 0$ is 1.

11.3 Background: Markov chain models

Theorem If a Markov Chain Model is homogeneous and irreducible and all its states are a-periodic, then it has a unique invariant distribution $P\boldsymbol{\pi}^* = \boldsymbol{\pi}^*$ and

$$\boldsymbol{\pi}_{t+1} = P^t \boldsymbol{\pi}_1 \xrightarrow{t \to \infty} \boldsymbol{\pi}^*$$

holds for any distribution $\boldsymbol{\pi}_1$ of the initial states of the model.

Markov Chain Monte Carlo Sampler: MCMC samplers are irreducible and a-periodic Markov chains that have the target distribution as the invariant distribution.

This allows to explore distributions on complex, high dimensional state spaces

11.3 Markov Chain Monte Carlo Sampling

Metropolis-Hastings algorithm Target distribution $p(x)$ and strictly positive conditional proposal distributions $q(x \mid x')$

- Given $x_t = x$, sample a candidate value x^* from $q(x^* \mid x)$
- accept $x_{t+1} = x^*$ with probability $\mathcal{A}(x^*, x) = \min\left(1, \frac{q(x|x^*)p(x^*)}{q(x^*|x)p(x)}\right)$
- otherwise $x_{t+1} = x_t$, i.e. it remains at the previous state.

It can be shown that the invariant distribution of this process is the target distribution $p(x)$.

If x is high dimensional, i.e. $x = \{x_i \mid i \in V\}$ and the conditional distributions

$$p(x_i \mid x_{\overline{V}}) \text{ , where } \overline{V} = V \setminus \{i\}$$

are easy to compute, use a special case of MH sampler:

Gibbs sampler

- Given $x_t = x$, randomly choose a component i, denote $x = (x_i, x_{\overline{V}})$
- set $x_{t+1} = (x_i^*, x_{\overline{V}})$, where x_i^* is sampled from $p(x_i^* \mid x_{\overline{V}})$.

11.3 Markov Chain Monte Carlo Sampling

The acceptance probability for the Gibbs sampler is always 1:

$$\frac{q(x \mid x^*)\, p(x^*)}{q(x^* \mid x)\, p(x)} = \frac{p(x_i \mid x_{\overline{V}})\, p(x_i^*, x_{\overline{V}})}{p(x_i^* \mid x_{\overline{V}})\;\; p(x_i, x_{\overline{V}})} = \frac{p(x_i \mid x_{\overline{V}})\, p(x_i^* \mid x_{\overline{V}})\, p(x_{\overline{V}})}{p(x_i^* \mid x_{\overline{V}})\, p(x_i \mid x_{\overline{V}})\, p(x_{\overline{V}})} = 1$$

The Gibbs sampler is irreducible if the conditional distributions $p(x_i^* \mid x_{\overline{V}})$ are strictly positive for all $i \in V$.