Learning in graphical models & MCMC methods **Fall 2014**

Cours 8 — November 19

Enseignant: Guillaume Obozinski *Scribe: Khalife Sammy, Maryan Morel*

The web page of the course: `http://www.di.ens.fr/~fbach/courses/fall2014/`

8.1 HMM (cntd.)

8.2 Learning on graphical models

8.3 Approximate inference

8.3.1 Sampling methods

We often need to comptue the expectancy of a function f under some distribution p that cannot be computed. Let X be a random variable following the distribution p, we want to compute $\mu = \mathbb{E}[f(X)]$.

Example 8.3.1 $X = (X_1, ..., X_n)$,

$$f(X) = \delta(X = x_A)$$

$$\mathbb{E}[f(X)] = \mathbb{P}(X = x_A)$$

If we know how to sample from p, we can use the following method :

Algorithm 1 Monte Carlo Estimation

1: Draw $X_1, ..., X_n \overset{i.i.d.}{\sim} p$
2: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$

This method relies on the two following propositions :

Proposition 8.1 (Law of Large Numbers (LLN))

$$\hat{\mu} \xrightarrow{a.s.} \mu \text{ } if \text{ } ||\mu|| < \infty$$

Proposition 8.2 (Central Limit Theorem (CLT)) *For X a scalar random variable, if $\mathbb{V}ar(f(X)) = \sigma^2 < \infty$, then*

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

thus $\mathbb{E}(||\hat{\mu} - \mu||_2^2) = \frac{\sigma^2}{n}$

How to sample from a specific distribution ?

1. Uniform distribution on [0, 1] : use `rand`
2. Bernoulli distribution of parameter p : $X = \mathbf{1}_{\{U<p\}}$ with $U \sim \mathcal{U}([0,1])$
3. Using inverse transform sampling :

$$\forall x \in \mathbb{R} \qquad F(x) = \int_{-\infty}^{x} p(t)dt = \mathbb{P}(X \in [-\infty, x])$$

$$X = F^{-1}(U) \text{ avec } U \sim \mathcal{U}([0,1])$$

Proof $\mathbb{P}(X \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y)$ ■

Example 8.3.2 *Exponential distribution (one of the rare cases admitting an explicit inverse CDF[1])*

$$p(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}_+}(x)$$

$$X = -\frac{1}{\lambda} \ln(U)$$

8.3.2 Rejection sampling

Assume that $p(x)$ is known up to a constant

$$p(x) = \frac{\tilde{p}(x)}{Z_p}$$

Assume that we can construct and compute q_k such that

$$\tilde{p}(x) < kq_k(x)$$

with q_k a probability distribution. Assume we can sample from q We define the rejection sampling (R.S.) algorithm as :

Algorithm 2 Rejection Sampling Algorithm

1: Draw X from q
2: Accept X with probability $\frac{\tilde{p}(x)}{kq_k(x)} \in [0,1]$, otherwise, reject the sample

[1] Cumulative Distribution Function

Proof

$$\begin{aligned}
\mathbb{P}(X = x, X \text{ is accepted}) &= \mathbb{P}(X = x, X \text{ is accepted}) \\
&= \mathbb{P}(X \text{ is accepted}|X = x)\mathbb{P}(X = x) \\
&= \frac{\tilde{p}(x)}{kq(x)}q(x) \\
&= \frac{\tilde{p}(x)}{k}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(X \text{ is accepted}) &= \int \frac{\tilde{p}(x)}{k}\mathrm{d}x \\
&= \frac{Z_p}{k}
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{P}(X = x|X \text{ is accepted}) &= \frac{\tilde{p}(x)}{k}\frac{k}{Z_p} \\
&= p(x)
\end{aligned}$$

■

Remark 8.3.1 *In practice, finding q and k such that acceptance has a reasonably large probability is hard.*

8.3.3 Importance Sampling

Assume $X \sim p$. We aim to compute the expectancy of a function f :

$$\begin{aligned}
\mathbb{E}_p(f(X)) &= \int f(x)p(x)\mathrm{d}x \\
&= \int \frac{f(x)p(x)}{q(x)}q(x)\mathrm{d}x \\
&= \mathbb{E}_q\left(f(Y)\frac{p(Y)}{q(Y)}\right) \qquad \text{with } Y \sim q \\
&= \mathbb{E}_q(g(Y)) \\
&\approx \frac{1}{n}\sum_{j=1}^{n} g(Y_j) \qquad \text{with } Y_j \overset{iid}{\sim} q \\
&= \frac{1}{n}\sum_{j=1}^{n} f(Y_j)\frac{p(Y_j)}{q(Y_j)}
\end{aligned}$$

$w(Y_i) = \frac{p(Y_j)}{q(Y_j)}$ are called *importance weights*. Remind that

$$\mu = \mathbb{E}_p(f(X)) \approx \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Thus we get :

$$\begin{aligned}
\mathbb{E}(\hat{\mu}) &= \frac{1}{n}\sum \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx \\
Var(\hat{\mu}) &= \frac{1}{n}Var_{q(x)}\left(\frac{f(x)p(x)}{q(x)}\right)
\end{aligned}$$

Lemme 8.3 *If* $\forall x,\ |f(x)| \leq M$,

$$Var(\hat{\mu}) \leq \frac{M^2}{n}\int \frac{p(x)^2}{q(x)}dx.$$

Proof

$$\begin{aligned}
Var(\hat{\mu}) = & \ \tfrac{1}{n}Var_{q(x)}\left(\tfrac{f(x)p(x)}{q(x)}\right) \\
\leq & \ \tfrac{1}{n}\int \tfrac{f(x)^2 p(x)^2}{q(x)^2}q(x)dx \\
\leq & \ \tfrac{M^2}{n}\int \tfrac{p(x)^2}{q(x)}dx.
\end{aligned}$$

■

Remark 8.3.2

$$\begin{aligned}
\int \frac{p(x)^2}{q(x)}dx &= \int \frac{p^2(x) - 2p(x)q(x) + q^2(x)}{q(x)}dx + \int \frac{2p(x)q(x) - q^2(x)}{q(x)}dx \\
&= \underbrace{\int \frac{(p(x) - q(x))^2}{q(x)}dx}_{\chi^2 \text{ divergence between } p \text{ and } q.} +1
\end{aligned}$$

Hence, importance sampling will give good results if q has mass where p has. Indeed, if for some y, $q(y) << p(y)$, importance weights $Var(\hat{\mu})$ may be very large.

Extension of Importance Sampling Assume we only know p and q up to a constant : $p(x) = \frac{\tilde{p}(x)}{Z_p}$ and $q(x) = \frac{\tilde{q}(x)}{Z_p}$, and only $\tilde{p}(x)$ and $\tilde{q}(x)$ are known.

$$\begin{aligned}
\mathbb{E}\left(f(Y)\frac{\tilde{p}(Y)}{\tilde{q}(Y)}\right) &= \mathbb{E}\left(f(Y)\frac{p(Y)}{q(Y)}\frac{Z_p}{Z_q}\right) = \mu\frac{Z_p}{Z_q} \\
\hat{\tilde{\mu}} &= \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)} \xrightarrow{a.s.} \mu\frac{Z_p}{Z_q}
\end{aligned}$$

Take f to be a constant, we get

$$\begin{aligned}
\hat{Z}_{p/q} &= \frac{1}{n}\sum_{i=1}^{n} \frac{p(Y_i)}{q(Y_i)} \xrightarrow{a.s.} \frac{Z_p}{Z_q} \\
\hat{\mu} &= \frac{\hat{\tilde{\mu}}}{\hat{Z}_{p/q}} \xrightarrow{a.s.} \mu
\end{aligned}$$

Remark 8.3.3 *Even if $Z_p = Z_q = 1$, renormalizing by $\hat{Z}_{p/q}$ often improves the estimation.*

8.4 Markov Chain Monte Carlo (MCMC)

Context $x \in \mathcal{X}$, $\mathcal{X}$ finite. We aim to build a Markov chain $X_0, X_1, \ldots$ such that its density $q_t(x) = p(X_t = x)$ converges to a target distribution $p(x)$.

8.4.1 Reminder on Markov chains

Consider order 1 homogenous Markov chains, i.e.

$$\mathbb{P}(X_t = y | X_{t-1} = x) = \mathbb{P}(X_{t-1} = y | X_{t-2} = x)$$

Definition 8.4 (Time Homogenous Markov chain)

$$\begin{aligned}
\forall t \geq 0\ \forall (x,y) \in \mathcal{X} \qquad & p(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0) \\
&= p(X_{t+1} = y \mid X_t = x) \\
&= p(X_1 = y \mid X_0 = x) \\
&= S(x,y)
\end{aligned}$$

Definition 8.5 (Transition matrix) *Let $k = card(\mathcal{X}) < \infty$. We define the matrix $S \in \mathbb{R}^{k \times k}$ such that $\forall x, y \in \mathcal{X}, S(x,y) = \mathbb{P}(X_t = y | X_{t-1} = x)$. S is called* transition matrix *of the Markov chain $(X_k)_k$.*

Properties 8.4.1 *If $k = card(\mathcal{X}) < \infty$, then:*

- $S \succeq 0$
- $S\mathbf{1} = \mathbf{1}$ *(i.e. column sum is equal to 1)*

S is a stochastic matrix

Definition 8.6 (Stationary Distribution) *The distribution π on $\mathcal{X}$ is stationary if $S^T\Pi = \Pi$ where*

$$\Pi = \pi(x)_{x\in\mathcal{X}}$$

Equivalently,

$$i.e.\ \forall x, y\ \ \pi(y) = \sum_x \pi(x)S(x,y)$$

If $\mathbb{P}(X_n = x) = \pi(x)$ with π a stationary distribution of S, then we have $\mathbb{P}(X_{n+1} = y) = \sum_x \mathbb{P}(X_{n+1} = y|X_n = x)\mathbb{P}(X_n = x) = \sum_x S(x,y)\pi(x) = \pi(y)$

Theorem 8.7 (Perron-Frobenius) *Every stochastic matrix S has* at least *one stationary distribution π*

Definition 8.8 (Regular Markov Chain) *A markov chain is regular (or equivalently aperiodic irreductible) if $\forall x, y \in \mathcal{X}, S(x,y) > 0$*

Proposition 8.9 *If a Markov chain is regular, then its transition matrix has a unique stationary distribution π and for any initial distribution q_0 on X_0, if $q_t(\cdot) = \mathbb{P}(X_t = \cdot)$, then $q_t \underset{t\to+\infty}{\longrightarrow} \pi$ Let q_n be the distribution of X_n, then for all distribution q_0 we get*

$$q_n \to \pi$$

Goal We want to find

$$\pi(x) = \frac{1}{Z}\prod_{c\in\mathcal{C}} \psi_c(x_c)$$

We try to reverse engineer this distribution by finding a Markov chain converging to π

Definition 8.10 (Detailed Balance) *A Markov chain is* reversible *if for the transition matrix S,*

$$\exists\pi, \forall x, y \in \mathcal{X}, \pi(x)S(x,y) = \pi(y)S(y,x)$$

This equation is called detailed balance equation. *It can be reformulated*

$$\mathbb{P}(X_{t+1} = y, X_t = x) = \mathbb{P}(X_{t+1} = x, X_t = y)$$

Proposition 8.11 *If π satisfies detailed balance, then π is a stationary distribution and* $\sum_x S(x,y)p(x) = \sum_x p(y)S(y,x) = p(y)\sum_x S(y,x) = p(y)$

8.4.2 Metropolis-Hastings Algorithm

Proposal transition $T(x, z) = \mathbb{P}(Z = z|X = x)$

Acceptance probability $\alpha(x, t) = \mathbb{P}(\text{Accept z } |X = x, Z = z)$

α is not a transition matrix.

Algorithm 3 Metropolis Hastings

Initialize x_0 from $X_0 \sim q$
for $t = 1, \ldots, T$ **do**
  Draw z_t from $\mathbb{P}(Z = \cdot|X_{t-1} = x_{t-1}) = T(x_{t-1}, \cdot)$
  With probability $\alpha(Z_t, x_{t-1})$, set $x_t = z_t$, otherwise, set $x_t = x_{t-1}$
end for

Proposition 8.12 *With that choice of $\alpha(x, z)$, if $T(\cdot, \cdot)$ is regular, then the Metropolis-Hastings algorithm defines a Markov chain that converges to π.*

Explanation $\mathbb{P}(X_t = x_t|X_{t-1} = x_{t-1}) = S(x_{t-1}, x_t)$

$$\begin{aligned} \forall z \neq x, S(x, z) &= T(x, z)\alpha(x, z) \\ S(x, x) &= T(x, x) + \sum_{z \neq x} T(x, z)(1 - \alpha(x, z)) \end{aligned}$$

Let π be given : we want to choose S such that we have *detailed balance* :

$$\begin{aligned} \pi(x)S(x, z) &= \pi(z)S(z, x) \\ \pi(x)T(x, z)\alpha(x, z) &= \pi(z)T(z, x)\alpha(z, x) \end{aligned}$$

Then

$$\frac{\alpha(x, z)}{\alpha(z, x)} = \frac{\pi(z)T(z, x)}{\pi(x)T(x, z)} \; (*)$$

If

$$\alpha(x, z) = \min\left(1, \frac{\pi(z)T(z, x)}{\pi(x)T(x, z)}\right)$$

then

$$\begin{cases} \alpha(x, z) \in [0, 1] \\ (*) \text{ is satisfied } \implies \text{ detailed balance} \end{cases}$$