6.438 Recitation 11 Notes 11/21/2014

Today:
- Markov Chain Monte Carlo Methods (MCMC)
  - Metropolis - Hastings
  - Gibbs Sampling
- Particle Filters

1°. MCMC

* Recap: Metropolis - Hastings

- Setup & Goal: Target distribution $P_x(\underline{x}) = \dfrac{P^*_x(\underline{x})}{Z}$ ← known (numerator) ← not known (Z)

  Want: samples $\underline{x}^{(1)}, \underline{x}^{(2)}, \cdots$ from $P_x(\underline{x})$.

  · MCMC does work for continuous variables → Assume: $\underline{x} \in \{1, 2, \cdots |\mathcal{X}|\}^N$

  · We'll derive for discrete case for simplicity of presentation.

- Basic Idea:
  - Construct a Markov Chain P, s.t P has unique Stationary distribution $\underline{\pi}$ & $\underline{\pi}(\underline{x}) = P_{\underline{x}}(\underline{x})$ $\forall \underline{x} \in |\mathcal{X}|^N$
  - The states of P are clearly all possible $\underline{x} \in \mathcal{X}^N$

    ⇒ We need to specify the transition probabilities. $[P_{ij}]$ ← This is a matrix of size $|\mathcal{X}|^N \times |\mathcal{X}|^N$

- Metropolis - Hastings Algorithm:

(1). Find some transition probability matrix $[K_{ij}]$

- $[K_{ij}]$ can be chosen as you like. e.g. uniform.
- Different choices of $[K_{ij}]$ don't affect correctness. but can affect mixing time
- $K_{ii} > 0$ $\forall i \in \{1, 2, \cdots N\}$. and any state $\underline{X}$ can go to any other state $\underline{X}'$ in finite # of steps.

(2) Define
$$P_{ij} = \begin{cases} K_{ij} \cdot \underbrace{\min\left\{1, \dfrac{P^*_x(j) \cdot K_{ji}}{P^*_x(i)\, K_{ij}}\right\}}_{R(i,j)} & (\text{if } i \neq j) \\ 1 - \sum_{k \neq i} P_{ik} & (\text{if } i = j) \end{cases}$$

(3) Pick any initial state $\underline{x}^{(0)}$

```
For $t = 1$ : Max-Nb-Iter
    - Given the current state $i \triangleq \underline{x}^{(t-1)}$, propose a new state $j$
      according to $[k_{ij}]$
    - With probability $R(i,j)$, set $\underline{x}^{(t)} = j$
      With probability $1 - R(i,j)$, set $\underline{x}^{(t)} = i$
End
```

- Comments. (1). The algorithm will spit out a sequence of samples

$$\underbrace{\underline{x}^{(0)}, \underline{x}^{(1)}, \underline{x}^{(2)}, \cdots \underline{x}^{(t)},}_{\text{"burn-in" period}} \boxed{\underline{x}^{(t+1)}}, \cdots \underline{x}^{(t+s)}, \boxed{\underline{x}^{(t+s+1)}}, \cdots \boxed{\underline{x}^{(t+2s)}}, \cdots\cdots$$

choose one sample from every s successive samples to reduce correlation

(2) Intuition about $R(i,j)$

$P_x(j)\, k_{ji}$: "flow from $j \to i$"

$P_x(i)\, k_{ij}$: "flow from $i \to j$"

- if "flow $j \to i$" $\geq$ "flow $i \to j$": add flow $i \to j$, i.e. make the transition from i to j
- if "flow $j \to i$" $<$ "flow $i \to j$": control flow $i \to j$, i.e. only make the transition $i \to j$ with some probability < 1

(3) In practice, don't compute/store $\underbrace{[P_{ij}]}_{|\mathcal{X}|^N \times |\mathcal{X}|^N, \text{ too big}}$

* Gibbs Sampling

- can be viewed as a special case of Metropolis-Hastings, with $[k_{ij}]$ chosen using the following process: current state $= \underline{x}$

(1). uniformly pick a coordinate $k \in \{1, 2, \dots N\}$.

(2). $\forall \ell \neq k$, $\ell \in \{1, 2, \cdots N\}$. set $x_\ell' = x_\ell$

x_k' is sampled from $P_{x_k | \underline{x}_{\setminus k}}(\cdot \mid \underline{x}_{\setminus k})$ ← $\underline{x}_{\setminus k}$ denotes all other coordinates of $\underline{x}$

- Claim: Let us denote the new state as $\underline{x}'$. If $R(\underline{x}, \underline{x}')$ is defined as before (i.e. $R(\underline{x}, \underline{x}') = \min\left\{1, \dfrac{P_{\underline{x}}(\underline{x}') \cdot k_{\underline{x}'\underline{x}}}{P_{\underline{x}}(\underline{x})\, k_{\underline{x}\underline{x}'}}\right\}$) is always 1.

Proof: $P_{\underline{x}}(\underline{x}) \cdot K_{\underline{x},\underline{x}'} = \underline{P_{\underline{x}}(\underline{x})} \cdot \frac{1}{N} P(x_k' \mid \underline{x}_{\setminus k}')$

$= \frac{1}{N} \underline{\boxed{P(x_k \mid \underline{x}_{\setminus k})} \cdot P(\underline{x}_{\setminus k})} \quad P(x_k' \mid \underline{x}_{\setminus k}')$

$= \frac{1}{N} \boxed{P(x_k \mid \underline{x}_{\setminus k}')} \cdot \underline{\underline{P(\underline{x}_{\setminus k}') \, P(x_k' \mid \underline{x}_{\setminus k}')}}$

$= \frac{1}{N} P(x_k \mid \underline{x}_{\setminus k}') \quad \underline{\underline{P_{\underline{x}}(\underline{x}')}}$

$= K_{\underline{x}',\underline{x}} \quad P_{\underline{x}}(\underline{x}')$.

$\Rightarrow R(\underline{x}, \underline{x}') = \min\left\{1, \frac{P_{\underline{x}}(\underline{x}') \, K_{\underline{x}',\underline{x}}}{P_{\underline{x}}(\underline{x}) \, K_{\underline{x},\underline{x}'}}\right\} = 1 \Rightarrow$ always accept.

- Notice the correlation is even higher in Gibbs sampling, because only 1 coordinate is changed at one time.

Variants exist: e.g. check homework problem 8.2 for block gibbs sampling

* Failure Modes of MCMC

- Islands of high-probability states, with low-probability states in between
- All states have very small probability except for one state.

  e.g. $P(\underline{x}_0) = 1/2 \qquad P(\underline{x}) = \frac{1}{2(2^{100}-1)} \quad \forall \underline{x} \neq \underline{x}_0 \quad \& \quad \underline{x}, \underline{x}_0 \in \{0,1\}^{100}$

  will have long sequences of $\underline{x}_0$ & long sequences of $\underline{x} \neq \underline{x}_0$

2° Particle Filters

* Importance Sampling

- Importance Sampling produces <
  - estimate of expectation of a given function ✓ $E_p[f(x)]$
  - samples from $P_{\underline{x}}(\cdot)$ ✗
- Algorithm:

  (1) Propose some distribution $q_{\underline{x}}(\cdot)$ "proposal distribution" that is easy to sample from e.g. uniform distribution

  (2) Let $\underline{x}^{(1)}, \underline{x}^{(2)}, \dots \underline{x}^{(K)}$ are iid samples from $q_{\underline{x}}(\cdot)$

  (3) Compute weights: $w^k = w(\underline{x}^k) = \frac{P^*(\underline{x}^k)}{q(\underline{x}^k)} \quad k = 1, 2, \cdots K$

  (4) Compute expected value of given function f:

  $$E_p[f(\underline{x})] \approx \frac{\frac{1}{K}\sum_{k=1}^{K} w^k \cdot f(\underline{x}^k)}{\frac{1}{K}\sum_{k=1}^{K} w^k}$$

- Comments: (1). Can prove: $$\lim_{k\to\infty} \frac{\frac{1}{K}\sum_{k=1}^{K} w^k \cdot f(\underline{x}^k)}{\frac{1}{K}\sum_{k=1}^{K} w^k} = E_{p_{\underline{x}}}[f(\underline{x})] \quad (*)$$

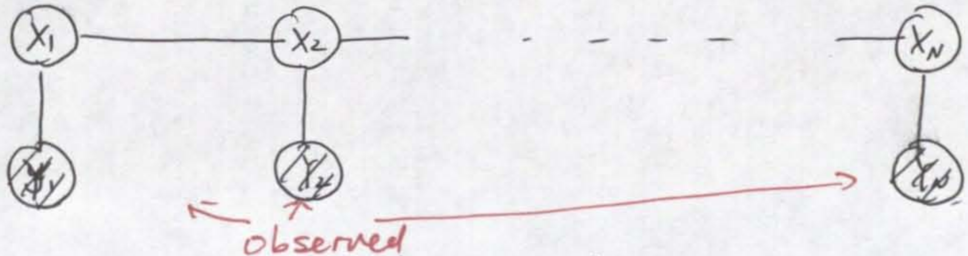
(2). $q_{\underline{x}}(\cdot)$ can be any distribution (that is easy to sample from). But choice of $q_{\underline{x}}(\cdot)$ will affect the speed of convergence in (*)

(3) Undesirable situation: $q_{\underline{x}}(\cdot)$ has low probability where $P_{\underline{x}}(\cdot)$ has high probability.

(4) One way to look at importance sampling is that we're approximating $P_{\underline{x}}(\cdot)$ using a bunch of weighted particles. $\{w^k, \underline{x}^k\}_{k=1}^{K}$. Particle filter is built based on this intuition.

* Particle Filters

- Introduced on HMM ← can generalize to trees
- Can deal with distributions over any continuous variables ← not just Gaussians
-



- View 1: (as introduced in lecture)
  - Target Distribution: $P_{x_1^N | y_1^N}(x_1^N | y_1^N) = \frac{P_{x_1^N, y_1^N}(x_1^N, y_1^N)}{P_{y_1^N}(y_1^N)}$ ← p^* known; z not known
  - use $P_{x_1^N}(x_1^N)$ as the proposal distribution $q(\cdot)$
    ($P_{x_1^n}(x_1^n) = P_{x_1}(x_1) \underset{x_2|x_1}{P}(x_2|x_1) \cdots P_{x_n|x_{n-1}}(x_n|x_{n-1})$ can be sampled easily)
  - Let $\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots \underline{x}^{(K)}$ be iid samples from $P_{x_1^N}(\cdot)$
  - $w^{(k)} = \frac{p^*(\underline{x}^{(k)})}{q(\underline{x}^{(k)})} = \frac{P_{x_1^N, y_1^N}(x_1^{N(k)}, y_1^N)}{P_{x_1^N}(x_1^{N(k)})} = P_{y_1^N|x_1^N}(y_1^N | x_1^{N(k)}) = \prod_{i=1}^{N} P_{y_i|x_i}(y_i | x_i^{(k)})$
  - $E_{P_{x_1^n|y_1^n}}[f(\underline{x})] \approx \frac{\frac{1}{K}\sum_{i=1}^{K} w^{(k)} f(\underline{x}^{(k)})}{\frac{1}{K}\sum_{i=1}^{K} w^{(k)}}$

- View 2.

- Target distribution: $P_{X_n|Y_1^n}(\cdot \mid y_1^n)$ $\quad n = 1, 2, \cdots N$
- initialization: $P_{X_1}(\cdot)$: $\{w_1^{(k)}, x_1^{(k)}\}_{k=1}^{K}$ $\quad w_1^{(k)} = \frac{1}{K}$.
- For $n = 1 : N-1$

$$x_{n+1}^{(k)} \sim P_{X_{n+1}|X_n}(\cdot \mid x_n^{(k)}) \quad k = 1, 2 \dots K$$

$$\tilde{w}_{n+1}^{(k)} = \tilde{w}_n^{(k)} \quad k = 1, 2, \cdots K.$$

"prediction step." $\{w_n^{(k)}, x_n^{(k)}\}_{k=1}^{K} \Rightarrow \{\tilde{w}_{n+1}^{(k)}, x_{n+1}^{(k)}\}_{k=1}^{K}$

$$w_{n+1}^{(k)} = \tilde{w}_{n+1}^{(k)} \cdot P_{Y_{n+1}|X_{n+1}}(y_{n+1} \mid x_{n+1}^{(k)}) \quad k = 1, 2 \dots K.$$

$\leftarrow$ "update step" $\{\tilde{w}_{n+1}^{(k)}, x_{n+1}^{(k)}\}_{k=1}^{K} \Rightarrow \{w_{n+1}^{(k)}, x_{n+1}^{(k)}\}_{k=1}^{K}$

End

- resampling: (heuristic) if $\hat{N}_{eff} = \frac{1}{\sum_{k=1}^{K}(w_n^{(k)})^2}$ too small.

- Related to forward pass of sum-product:

$$P_{n+1|n}(x_{n+1} \mid y_1^n) = \int_{x_n} P_{X_{n+1}|X_n}(x_{n+1} \mid x_n)\, P_{n|n}(x_n \mid y_1^n)\, dx_n$$

$$P_{n+1|n+1}(x_{n+1} \mid y_1^{n+1}) = \frac{1}{Z} P_{Y_{n+1}|X_{n+1}}(y_{n+1} \mid x_{n+1})\, P_{n+1|n}(x_{n+1} \mid y_1^n)$$

* Beyond HMM: Particle Filters on Trees

- BP equations on trees:

$$m_{i \to j}(x_j) = \int_{x_i} \psi_{ij}(x_i, x_j)\, \phi_i(x_i) \prod_{\ell \in N(i) \setminus j} m_{\ell \to i}(x_i)\, dx_i$$

separate into 2 steps:

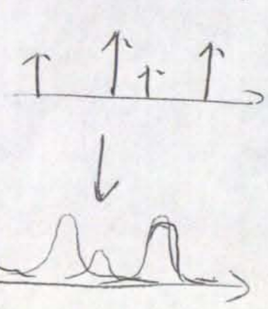
$$(1): \quad \phi_{ij}(x_i) \triangleq \phi_i(x_i) \prod_{\ell \in N(i) \setminus j} m_{\ell \to i}(x_i)$$

$$(2): \quad m_{i \to j}(x_j) = \int_{x_i} \psi_{ij}(x_i, x_j)\, \phi_{ij}(x_i)\, dx_i$$

- step (1): $m_{\ell \to i}(x_i)$ is now represented by a weighted particle set $\{W_{\ell \to i}^{(k)}, x_{\ell \to i}^{(k)}\}_{k=1}^{K}$ i.e. $m_{\ell \to i}(x_i) \approx \sum_{k=1}^{K} W_{\ell \to i}^{(k)}\, \delta(x_i - x_{\ell \to i}^{(k)})$

But we have problems multiplying $M_{\ell_1\to i}(x_i)$ & $M_{\ell_2\to i}(x_i)$ in this form

(continuous x_i $\Rightarrow \mathbb{P}(x^{(k)}_{\ell_1\to i} = x^{(k)}_{\ell_2\to i}) = 0$)



Instead, use small Gaussians:

$$M_{\ell\to i}(x_i) \approx \underbrace{\sum_{k=1}^{K} W^{(k)}_{\ell\to i}\, \mathcal{N}(x_i;\, x^{(k)}_{\ell\to i},\, J^{-1}_{\ell\to i})}_{\text{mixture of Gaussians}}$$

$\Rightarrow M_{\ell_1\to i}(x_i)\cdot M_{\ell_2\to i}(x_i)$ also a mixture of Gaussians.

For further details: "Efficient Multiscale Sampling from Products of Gaussian Mixtures" by Ihler et al 2003.

- Step (2). Just a summation over the samples.

Rosencrantz & Guildenstern Are Dead

Film Discussion Questions

1. What do you make of the opening credits? As the play opens in a place "without visible character," the film begins with credits on a black screen. However, the film introduces "audible character" in the form of a western-themed soundtrack. How does this contrast with your expectation of "two Elizabethans"?

2. Unlike a play, film can cut from scene to scene instantaneously. It can completely change the setting and mise-en-scene without limitation. Does the film take advantage of this fact, or does it try to respect the story's original medium by forcing characters around in circles so that that they repeatedly end up in the same room of the castle?

3. How are sound effects used in the film? Extradiegetic chimes are added at crucial moments – first after Guildenstern brings up the question of suspense, then again when the coin is finally tails. The film inserts dramatic echoes to some of Guildenstern's most excited moments that momentarily halt the flow of dialogue and narrative. Ambient animal noises can be heard throughout the film and are at one point revealed to be coming from Rosencrantz. Do these elements add anything to the narrative other than perhaps to emphasize moments of significance to the less sophisticated audience of cinema?

4. As discussed in Thursday's class, two individuals on a stage or screen are far more distinguishable from one another than words on a page. It was mentioned that being able to visually tell them apart would make it easier to tell which character was which. After watching the film, was it clear who was Rosencrantz and who was Guildenstern? Or did the fact that you could differentiate them as Gary Oldman and Tim Roth (or the long-haired one and short-haired one) eliminate the necessity of thinking of each one specifically as Rosencrantz or Guildenstern?

5. What were the effects of having the question game played on an actual tennis court?

6. What is the significance of wind in the film? Hamlet is said to be "at the mercy of the elements," but Rosencrantz and Guildenstern are also plagued by their quest to find out the direction of the wind. As an example, the windmill toy spins after Rosencrantz insists that "there isn't any wind."

7. Let's discuss the puppet show. The tragedians use puppets in addition to live action to enact the deaths from Hamlet. Why does the film version include this story-telling mode when the original play does not? Is it simply a way to visually emphasize the complex literary layers of *Rosencrantz and Guildenstern are Dead*?

8. The film ends with the Tragedians driving away in their caravan. What is the effect of ending with life rather than ending on a stage filled with death (as both Hamlet and the play of R&G do)?

6.438 Algorithms for Inference
Fall 2014