Sampling

Prof. Dr. H. H. Takada

Quantitative Research – Itaú Asset Management
Institute of Mathematics and Statistics – University of São Paulo

Sampling

Sampling concerns drawing realisations $\mathcal{X} = \{x^1, \ldots, x^L\}$ from a distribution $p(x)$. For a discrete variable x, in the limit of a large number of samples, the fraction of samples in state x tends to $p(x = \mathsf{x})$. That is,

$$\lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \mathbb{I}\left[x^l = \mathsf{x}\right] = p(x = \mathsf{x})$$

In the continuous case, one can consider a small region Δ such that the probability that the samples occupy Δ tends to the integral of $p(x)$ over Δ.

Using sampling to approximate averages

Given a finite set of samples, one can then approximate expectations using

$$\langle f(x) \rangle_{p(x)} \approx \frac{1}{L} \sum_{l=1}^{L} f(x^l) \equiv \hat{f}_{\mathcal{X}}$$

The subscript in $\hat{f}_{\mathcal{X}}$ emphasises that the approximation is dependent on the set of samples drawn.

Sampling as an approximation techniques

- If we have a procedure that in some sense 'faithfully' draws samples from $p(x)$, then we can use this to approximate averages.
- This suggests a general class of approximation methods for computing averages with respect to otherwise computationally intractable distributions.
- The art therefore is in finding sampling procedures that can indeed draw samples from distributions that are both analytically and computationally intractable.

The marginal requirement

A sampling procedure produces realisations of the set $\mathcal{X}$ and can be considered as a distribution $\tilde{p}(\mathcal{X})$. Provided the marginals of the sampling distribution are equal to the marginals of the target distribution, $\tilde{p}(x^l) = p(x^l)$, then the average of the approximation with respect to draws of the sample set $\mathcal{X}$ is

$$\left\langle \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})} = \frac{1}{L} \sum_{l=1}^{L} \left\langle f(x^l) \right\rangle_{\tilde{p}(x^l)} = \left\langle f(x) \right\rangle_{p(x)}$$

Hence the mean of the sample approximation is the exact mean of f provided only that the marginals of $\tilde{p}(\mathcal{X})$ correspond to the required marginals $p(x)$.

Dependent samples?

Even if the individual samples $x^1, \ldots, x^L$ are dependent, that is $\tilde{p}(\mathcal{X})$ does not factorise into $\prod_l \tilde{p}(x^l)$, provided that the marginal $\tilde{p}(x^l) = p(x)$, then the sample average is unbiased.

The variance desideratum

For any sampling method, an important issue is the variance of the sample estimate. If this is low only a small number of samples are required since the sample mean must be close to the true mean (since it is unbiased). Defining

$$\Delta\hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{X}} - \left\langle \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})}, \qquad \Delta f(x) = f(x) - \langle f(x) \rangle_{p(x)}$$

the variance of the approximation is (assuming $\tilde{p}(x^l) = p(x)$, for all l)

$$\begin{aligned}
\left\langle \Delta^2 \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})} &= \frac{1}{L^2} \sum_{l,l'} \left\langle \Delta f(x^l) \Delta f(x^{l'}) \right\rangle_{\tilde{p}(x^l, x^{l'})} \\
&= \frac{1}{L^2} \left(L \left\langle \Delta^2 f(x) \right\rangle_{\tilde{p}(x)} + \sum_{l \neq l'} \left\langle \Delta f(x^l) \Delta f(x^{l'}) \right\rangle_{\tilde{p}(x^l, x^{l'})} \right)
\end{aligned}$$

The property of independent samples

Provided the samples are independent

$$\tilde{p}(\mathcal{X}) = \prod_{l=1}^{L} \tilde{p}(x^l)$$

and $\tilde{p}(x) = p(x)$, then $\tilde{p}(x^l, x^{l'}) = p(x^l)p(x^{l'})$. The term above $\left\langle \Delta f(x^l) \right\rangle \left\langle \Delta f(x^{l'}) \right\rangle$ is zero since $\langle \Delta f(x) \rangle = 0$. Hence

$$\left\langle \Delta^2 \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})} = \frac{1}{L} \left\langle \Delta^2 f(x) \right\rangle_{p(x)}$$

and the variance of the approximation scales inversely with the number of samples. In principle, therefore, provided the samples are independently drawn from $p(x)$, only a small number of samples is required to accurately estimate the expectation. Importantly, this result is independent of the dimension of x.

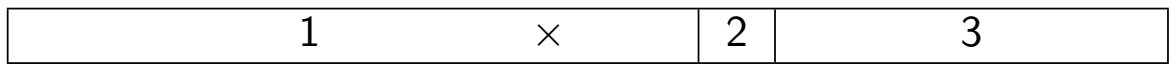
Drawing independent samples

- The critical difficulty is in actually generating independent samples from $p(x)$.
- Drawing samples from high-dimensional distributions is generally difficult and few guarantees exist to ensure that in a practical timeframe the samples produced are independent.
- Whilst a dependent sampling scheme may be unbiased, the variance of the estimate can be high such that a large number of samples may be required for expectations to be approximated accurately.
- There are many different sampling algorithms, all of which work in principle, but each working in practice only when the distribution satisfies particular properties; for example many schemes such as Markov chain Monte Carlo methods do not produce independent samples and a large number of samples may be necessary to produce a satisfactory approximation.

Univariate sampling: discrete distributions

Consider the one dimensional discrete distribution $p(x)$ where $\mathrm{dom}(x) = \{1, 2, 3\}$, with

$$p(x) = \begin{cases} 0.6 & x = 1 \\ 0.1 & x = 2 \\ 0.3 & x = 3 \end{cases}$$



We then draw a sample uniformly from $[0, 1]$, say $u = 0.66$. Then the sampled state would be state 2, since this is in the interval $(c_1, c_2]$.

Continuous case

- First we calculate the cumulant density function

$$C(y) = \int_{-\infty}^{y} p(x)dx$$

- Then we sample u uniformly from $[0, 1]$, and obtain the corresponding sample x by solving $C(x) = u \Rightarrow x = C^{-1}(u)$.

Special cases

For certain distributions, such as Gaussians, numerically efficient alternative procedures exist, usually based on co-ordinate transformations.

Rejection Sampling

Assume we have an efficient sampling procedure for $q(x)$. Can we use this to help us sample from another distribution $p(x)$?
We assume $p(x)$ is known only up to a normalisation constant Z,

$$p(x) = p^*(x)/Z$$

Using an auxiliary variable

Consider a binary auxiliary variable $y \in \{0, 1\}$ and define $q(x, y) = q(x)q(y|x)$ with

$$q(x, y = 1) = q(x)q(y = 1|x)$$

We can use the term $q(y = 1|x)$ to our advantage if we set

$$q(y = 1|x) \propto p(x)/q(x)$$

since then

$$q(x, y = 1) \propto p(x)$$

and sampling from $q(x, y)$ gives us a procedure for sampling from $p(x)$.

Rejection Sampling

We assume we can find a positive M such that

$$q(y = 1|x) = \frac{p^*(x)}{Mq(x)} \leq 1 \quad \forall x$$

Drawing from $p(x)$

- We first draw a candidate x^{cand} from $q(x)$
- Then draw y from $q(y|x^{cand})$.
- To do this we draw a value u uniformly between 0 and 1. If this value is less then $q(y = 1|x)$, we set $y = 1$, otherwise we set $y = 0$.
- If $y = 1$ we take x^{cand} as an independent sample from $p(x)$ – otherwise no sample is made (it is 'rejected').

Expected acceptance rate

The expected rate that we accept a sample is

$$q(y=1) = \int_x q(y=1|x)q(x) = \frac{Z}{M}$$

so that, to increase the acceptance rate, we seek the minimal M subject to $p^*(x) \leq Mq(x)$.

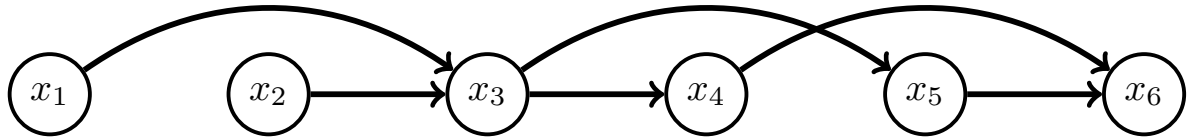
High dimensions

In high dimensions for vector $\mathbf{x}$, $q(y=1|\mathbf{x})$ will often decrease exponentially with the number of dimensions of $\mathbf{x}$. Rejection sampling is therefore a potentially useful method of drawing independent samples in very low dimensions, but is likely to be impractical in higher dimensions.

Multi-variate sampling

- One way to generalise the one dimensional discrete case to a higher dimensional distribution $p(x_1, \ldots, x_n)$ is to translate this into an equivalent one-dimensional distribution.
- We enumerate all the possible joint states $(x_1, \ldots, x_n)$, giving each a unique integer i from 1 to the total number of states, and construct a univariate distribution with probability $p(i) = p(\mathrm{x})$ for i corresponding to the multi-variate state x.
- In general, this procedure is impractical since the number of states grows exponentially with the number of variables $x_1, \ldots, x_n$.

Ancestral Sampling for Belief Nets



- We first rename the variable indices so that parent variables always come before their children

$$p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_3)p(x_6|x_4, x_5)$$

- One can sample first from those nodes that do not have any parents (here, x_1 and x_2). Given these values, one can then sample x_3, and then x_4 and x_5 and finally x_6.
- Such a procedure is straightforward. This procedure holds for both discrete and continuous variables.
- Ancestral or 'forward' sampling is a case of perfect sampling since each sample is indeed independently drawn from the required distribution.

Gibbs Sampling

- We consider a particular variable, x_i, for which we wish to draw a sample. Using Bayes' rule we may write

$$p(x) = p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

- Given a joint initial state x^1, from which we can read off the 'parental' state $x_1^1, \ldots, x_{i-1}^1, x_{i+1}^1, \ldots, x_n^1$, we then draw a sample x_i^2 from

$$p(x_i|x_1^1, \ldots, x_{i-1}^1, x_{i+1}^1, \ldots, x_n^1) \equiv p(x_i|x_{\setminus i})$$

  We assume this distribution is easy to sample from since it is univariate. We call this new joint sample (in which only x_i has been updated) $x^2 = \left(x_1^1, \ldots, x_{i-1}^1, x_i^2, x_{i+1}^1, \ldots, x_n^1\right)$.
- One then selects another variable x_j to sample and, by continuing this procedure, generates a set $x^1, \ldots, x^L$ of samples in which each x^{l+1} differs from x^l in only a single component.

Gibbs Sampling

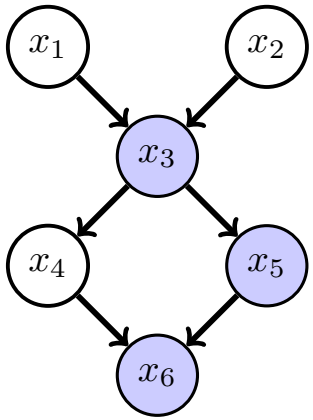


Figure: The shaded nodes are the Markov blanket of x_4 for the belief network. To draw a sample from $p(x_4|x_{\setminus 4})$ we clamp x_3, x_5, x_6 into their evidential states and draw a sample from $p(x_4|x_3)p(x_6|x_4, x_5)/Z$ where Z is a normalisation constant.

- $p(x_i|x_{\setminus i})$ is defined by the Markov blanket of x_i:

$$p(x_i|x_{\setminus i}) \propto p(x_i|\mathrm{pa}\,(x_i)) \prod_{j \in \mathrm{ch}(i)} p(x_j|\mathrm{pa}\,(x_j))$$

- For continuous variable x_i the summation is replaced with integration.
- Evidence is readily dealt with by clamping for all samples the evidential variables into their evidential states. There is also no need to sample for these variables, since their states are known.
- Whilst Gibbs sampling is particularly straightforward to implement, a drawback is that the samples are strongly dependent.

Markov Chain Monte Carlo (MCMC)

- We assume we have a multi-variate distribution in the form

$$p(x) = \frac{1}{Z} p^*(x)$$

  where $p^*(x)$ is the unnormalised distribution and $Z = \int_x p^*(x)$ is the normalisation constant.
- We assume we are able to evaluate $p^*(x = \mathrm{x})$, for any state x, but not $p(x = \mathrm{x})$ since Z is intractable.
- The idea in MCMC sampling is to sample, not directly from $p(x)$, but from a different distribution such that, in the limit of a large number of samples, effectively the samples will be from $p(x)$.
- To achieve this we forward sample from a Markov transition whose stationary distribution is equal to $p(x)$.

Markov chains

- Consider the conditional distribution $q(x^{l+1}|x^l)$. After a long time $L \gg 1$, the samples are from $q_\infty(x)$ which is defined as

$$q_\infty(x') = \int_x q(x'|x) q_\infty(x)$$

- The idea is, for a given distribution $p(x)$, to find a transition $q(x'|x)$ which has $p(x)$ as its stationary distribution. If we can do so, then we can draw samples from the Markov chain by forward sampling and take these as samples from $p(x)$ as the chain converges towards its stationary distribution.
- Note that for every distribution $p(x)$ there will be more than one transition $q(x'|x)$ with $p(x)$ as its stationary distribution. This is why there are very many different MCMC sampling methods, each with different characteristics and varying suitability for the particular distribution at hand.

Metropolis-Hastings sampling

$\tilde{q}(x'|x)$ is a so-called proposal distribution and $0 < f(x', x) \leq 1$ a positive function. Our interest is to set $f(x, x')$ such that the stationary distribution of $q(x'|x)$ is equal to $p(x)$ for any proposal $\tilde{q}(x'|x)$. That is

$$\begin{aligned} p(x') &= \int_x q(x'|x)p(x) \\ &= \int_x \tilde{q}(x'|x)f(x', x)p(x) + p(x')\left(1 - \int_{x''} \tilde{q}(x''|x')f(x'', x')\right) \end{aligned}$$

In order that this holds, we require (changing the integral variable from x'' to x)

$$\int_x \tilde{q}(x'|x)f(x', x)p(x) = \int_x \tilde{q}(x|x')f(x, x')p(x')$$

This can be achieved by using the acceptance function

$$f(x', x) = \min\left(1, \frac{\tilde{q}(x|x')p(x')}{\tilde{q}(x'|x)p(x)}\right) = \min\left(1, \frac{\tilde{q}(x|x')p^*(x')}{\tilde{q}(x'|x)p^*(x)}\right)$$

Metropolis-Hastings sampling

1: Choose a starting point x^1.
2: **for** $i = 2$ to L **do**
3: Draw a candidate sample x^{cand} from the proposal $\tilde{q}(x'|x^{l-1})$.
4: Let $a = \frac{\tilde{q}(x^{l-1}|x^{cand})p(x^{cand})}{\tilde{q}(x^{cand}|x^{l-1})p(x^{l-1})}$
5: **if** $a \geq 1$ **then** $x^l = x^{cand}$ ▷ Accept the candidate
6: **else**
7: draw a random value u uniformly from the unit interval $[0, 1]$.
8: **if** $u < a$ **then** $x^l = x^{cand}$ ▷ Accept the candidate
9: **else**
10: $x^l = x^{l-1}$ ▷ Reject the candidate
11: **end if**
12: **end if**
13: **end for**

Gaussian proposal distribution

A common proposal distribution for vector $\mathbf{x}$ is

$$\tilde{q}(\mathbf{x}'|\mathbf{x}) = \mathcal{N}\left(\mathbf{x}'|\mathbf{x}, \sigma^2\mathbf{I}\right) \propto e^{-\frac{1}{2\sigma^2}(\mathbf{x}'-\mathbf{x})^2}$$

for which $\tilde{q}(\mathbf{x}'|\mathbf{x}) = \tilde{q}(\mathbf{x}|\mathbf{x}')$ and the acceptance criterion becomes

$$f(\mathbf{x}', \mathbf{x}) = \min\left(1, \frac{p^*(\mathbf{x}')}{p^*(\mathbf{x})}\right)$$

If the unnormalised probability $p^*(x')$ of the candidate state is higher than the current state, $p^*(x)$, we therefore accept the candidate. Otherwise, we accept the candidate only with probability $p^*(\mathbf{x}')/p^*(\mathbf{x})$. If the candidate is rejected, the new sample is taken to be a copy of the previous sample $\mathbf{x}$.

Gaussian proposal distribution in high dimensions

- In high dimensions it is unlikely that a random candidate sampled from a Gaussian will result in a candidate probability higher than the current value.
- Because of this, only very small jumps (σ^2 small) are likely to be accepted. This limits the speed at which we explore the space $\mathbf{x}$ and increases the dependency between samples.

Optimisation versus Sampling

The acceptance function highlights that sampling is different from finding the optimum. Provided $\mathbf{x}'$ has a higher probability than $\mathbf{x}$, we accept $\mathbf{x}'$. However, we may also accept candidates that have a *lower* probability than the current sample.

Importance Sampling

Consider $p(x) = \frac{p^*(x)}{Z}$ where $p^*(x)$ can be evaluated but $Z = \int_x p^*(x)$ is an intractable normalisation constant. The average of $f(x)$ with respect to $p(x)$ is given by

$$\int_x f(x)p(x) = \frac{\int_x f(x)p^*(x)}{\int_x p^*(x)} = \frac{\int_x f(x)\frac{p^*(x)}{q(x)}q(x)}{\int_x \frac{p^*(x)}{q(x)}q(x)}$$

Let $x^1, \ldots, x^L$ be samples from $q(x)$, then we can approximate the average by

$$\int_x f(x)p(x) \approx \frac{\sum_{l=1}^L f(x^l)\frac{p^*(x^l)}{q(x^l)}}{\sum_{l=1}^L \frac{p^*(x^l)}{q(x^l)}} = \sum_{l=1}^L f(x^l)w^l$$

where we define the normalised importance weights

$$w^l = \frac{p^*(x^l)/q(x^l)}{\sum_{l=1}^L p^*(x^l)/q(x^l)}, \qquad \text{with} \sum_{l=1}^L w^l = 1$$

Dynamic Bayes networks

$$p(v_{1:t}, h_{1:t}) = p(v_1|h_1)p(h_1)\prod_{t=2}^{t} p(v_t|h_t)p(h_t|h_{t-1})$$

where $v_{1:t}$ are observations and $h_{1:t}$ are the random variables. Our interest is to draw sample paths $h_{1:t}$ given the observations $v_{1:t}$. In some models, such as the HMM, this is straightforward. However, in other cases, for example when the observation distribution $p(v_t|h_t)$ is intractable to normalise, we can approximate sampling instead.

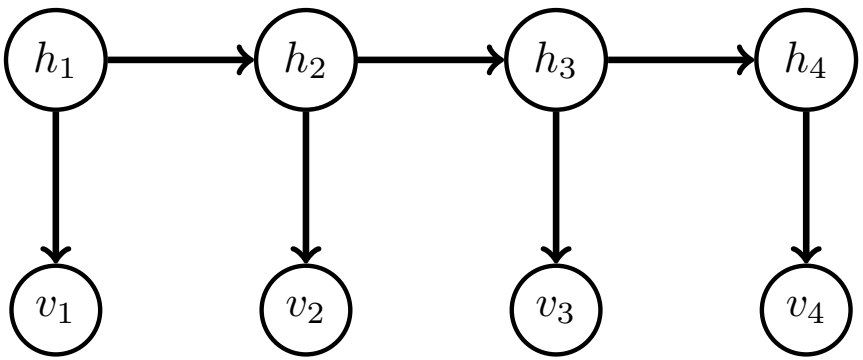


Figure: A Dynamic Bayesian Network. In many applications of interest, the emission distribution $p(v_t|h_t)$ is non-Gaussian, leading to the formal intractability of filtering/smoothing.

Particle filtering

Using ρ to represent the filtered distribution,

$$\rho(h_t) \propto p(h_t|v_{1:t})$$

the exact filtering recursion is

$$\rho(h_t) \propto p(v_t|h_t) \int_{h_{t-1}} p(h_t|h_{t-1})\rho(h_{t-1})$$

A PF can be viewed as an approximation of the above in which $\rho(h_{t-1})$ is approximated by a sum of δ-peaks:

$$\rho(h_{t-1}) \approx \sum_{l=1}^{L} w_{t-1}^l \delta\left(h_{t-1}, h_{t-1}^l\right)$$

where w_{t-1}^l are the normalised importance weights $\sum_{l=1}^{L} w_{t-1}^l = 1$, and h_{t-1}^l are the particles.

Particle filtering

The ρ message is represented as a weighted mixture of delta-spikes where the weight and position of the spikes are the parameters of the distribution. Using the PF approximation, we have

$$\rho(h_t) = \frac{1}{Z} p(v_t|h_t) \sum_{l=1}^{L} p(h_t|h_{t-1}^l) w_{t-1}^l$$

The constant Z is used to normalise the distribution $\rho(h_t)$. Although $\rho(h_{t-1})$ was a simple sum of delta peaks, in general $\rho(h_t)$ will not be – the delta-peaks get 'broadened' by the hidden-to-hidden and hidden-to-observation factors. Our task is then to approximate $\rho(h_t)$ as a new sum of delta-peaks. Below we discuss a method to achieve this for which explicit knowledge of the normalisation Z is not required. This is useful since in many tracking applications the normalisation of the emission $p(v_t|h_t)$ is unknown.

A Monte-Carlo sampling approximation

A simple approach to forming an approximate mixture-of-delta functions representation of $\rho(h_t)$ is to generate a set of sample points using importance sampling and use these as the new particles. That is we generate a set of samples $h_t^1, \ldots, h_t^L$ from some importance distribution $q(h_t)$ which gives the unnormalised importance weights

$$\tilde{w}_t^l = \frac{p(v_t|h_t^l) \sum_{l'=1}^{L} p(h_t^l|h_{t-1}^{l'}) w_{t-1}^{l'}}{q(h_t^l)}$$

Defining the normalised weights:

$$w_t^l = \frac{\tilde{w}_t^l}{\sum_{l'} \tilde{w}_t^{l'}}$$

we obtain an approximation

$$\rho(h_t) \approx \sum_{l=1}^{L} w_t^l \delta\left(h_t, h_t^l\right)$$

A Monte-Carlo sampling approximation

Ideally one would use the importance distribution that makes the importance weights unity, namely

$$q(h_t) \propto p(v_t|h_t) \sum_{l=1}^{L} p(h_t|h_{t-1}^l) w_{t-1}^l$$

However, this is often difficult to sample from directly due to the unknown normalisation of the emission $p(v_t|h_t)$. A simpler alternative is to sample from the transition mixture:

$$q(h_t) = \sum_{l=1}^{L} p(h_t|h_{t-1}^l) w_{t-1}^l$$

To do so, one first samples a component l^* from the histogram with weights $w_{t-1}^1, \ldots, w_{t-1}^L$. Given this sample index, say l^*, one then draws a sample from $p(h_t|h_{t-1}^{l^*})$. In this case the unnormalised weights become simply

$$\tilde{w}_t^l = p(v_t|h_t^l)$$

A toy face-tracking example

At time t a binary face template is in a two-dimensional location $\mathbf{h}_t$, which describes the upper-left corner of the template. The face moves randomly according to

$$\mathbf{h}_t = \mathbf{h}_{t-1} + \sigma \boldsymbol{\eta}_t$$

where $\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\eta}_t | \mathbf{0}, \mathbf{I})$. In addition, a fraction of the binary pixels in the whole image are selected at random and their states flipped. The aim is to try to track the upper-left corner of the face through time.

A toy face-tracking example

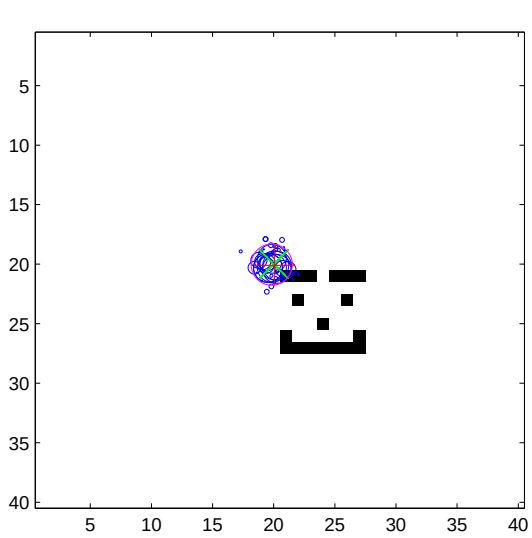
We need to define the emission distribution $p(\mathbf{v}_t|\mathbf{h}_t)$ on the binary pixels with $v_i \in \{0, 1\}$. Consider the following compatibility function

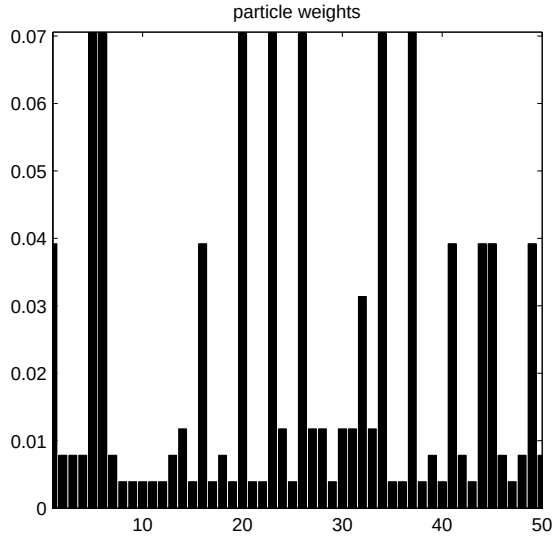
$$\phi(\mathbf{v}_t, \mathbf{h}_t) = \mathbf{v}_t^{\mathsf{T}} \tilde{\mathbf{v}}(\mathbf{h}_t)$$

where $\tilde{\mathbf{v}}(\mathbf{h}_t)$ is the vector representing the whole image with a clean face placed at position $\mathbf{h}_t$ and zeros outside of the face template. Then $\phi(\mathbf{v}_t, \mathbf{h}_t)$ measures the overlap between the face template at a specific location and the noisy image restricted to the template pixels. The compatibility function is maximal when the observed image $\mathbf{v}_t$ has the face placed at position $\mathbf{h}_t$. We therefore define

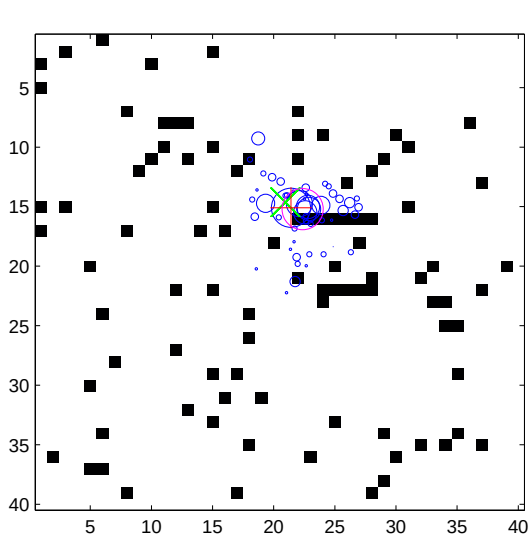
$$p(\mathbf{v}_t|\mathbf{h}_t) \propto \phi(\mathbf{v}_t, \mathbf{h}_t)$$

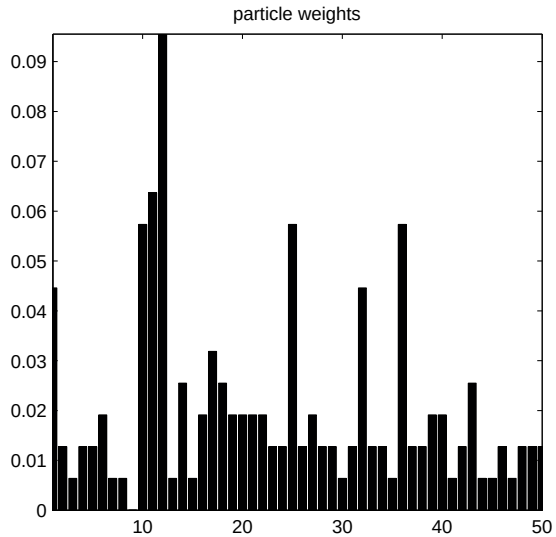
A Monte-Carlo sampling approximation





(a)





(b)

Figure: Tracking an object with a particle filter containing 50 particles. The small circles are the particles, scaled by their weights. The correct corner position of the face is given by the '×', the filtered average by the large circle 'o', and the most likely particle by '+'. **(a)**: Initial position of the face without noise and corresponding weights of the particles. **(b)**: Face with noisy background and the tracked corner position after 20 timesteps. The Forward-Sampling-Resampling PF method is used to maintain a healthy proportion of non-zero weights.