

# The Particle Filter

- **Non-parametric** implementation of Bayes filter
- Represents the belief (posterior)  $\text{Bel}(x_t)$  by a set of **random state samples**.
- This representation is **approximate**.
- Can represent distributions that are **not Gaussian**.
- Can model **non-linear** transformations.

## Basic principle:

- Set of state hypotheses (“particles”)
- Survival-of-the-fittest



# The Bayes Filter Algorithm (Rep.)

$$\text{Bel}(x_t) = \eta p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) \text{Bel}(x_{t-1}) dx_{t-1}$$

Algorithm Bayes\_filter ( $\text{Bel}(x)$ ,  $d$ )

1. if  $d$  is a sensor measurement  $z$  then
2.      $\eta = 0$
3.     for all  $x$  do
4.          $\text{Bel}'(x) \leftarrow p(z \mid x) \text{Bel}(x)$
5.          $\eta \leftarrow \eta + \text{Bel}'(x)$
6.     for all  $x$  do  $\text{Bel}'(x) \leftarrow \eta^{-1} \text{Bel}'(x)$
7. else if  $d$  is an action  $u$  then
8.     for all  $x$  do  $\text{Bel}'(x) \leftarrow \int p(x \mid u, x') \text{Bel}(x') dx'$
9. return  $\text{Bel}'(x)$



# Mathematical Description

Set of weighted samples:

$$\mathcal{X}_t := \{ \langle x_t^{[1]}, w_t^{[1]} \rangle, \langle x_t^{[2]}, w_t^{[2]} \rangle, \dots, \langle x_t^{[M]}, w_t^{[M]} \rangle \}$$

State hypotheses

Importance weights

The samples represent the probability distribution:

$$p(x) = \sum_{i=1}^M w_t^{[i]} \cdot \delta_{x_t^{[i]}}(x)$$

Point mass  
distribution  
("Dirac")



# The Particle Filter Algorithm

Algorithm *Particle\_filter*( $\mathcal{X}_{t-1}, u_t, z_t$ ) :

1.  $\bar{\mathcal{X}}_t = \mathcal{X}_t = \emptyset$
2. **for**  $m = 1$  **to**  $M$  **do**
3. sample  $x_t^{[m]} \sim p(x_t \mid u_t, x_{t-1}^{[m]})$
4.  $w_t^{[m]} \leftarrow p(z_t \mid x_t^{[m]})$
5.  $\bar{\mathcal{X}}_t \leftarrow \bar{\mathcal{X}}_t \cup \langle x_t^{[m]}, w_t^{[m]} \rangle$
6. **for**  $m = 1$  **to**  $M$  **do**  
draw  $i$  with prob.  $\propto w_t^{[i]}$   
 $\mathcal{X}_t \leftarrow \mathcal{X}_t \cup \langle x_t^{[i]}, 1/M \rangle$
7. **return**  $\mathcal{X}_t$

Sample from  
proposal

Compute sample  
weights

Resampling



# Localization with Particle Filters

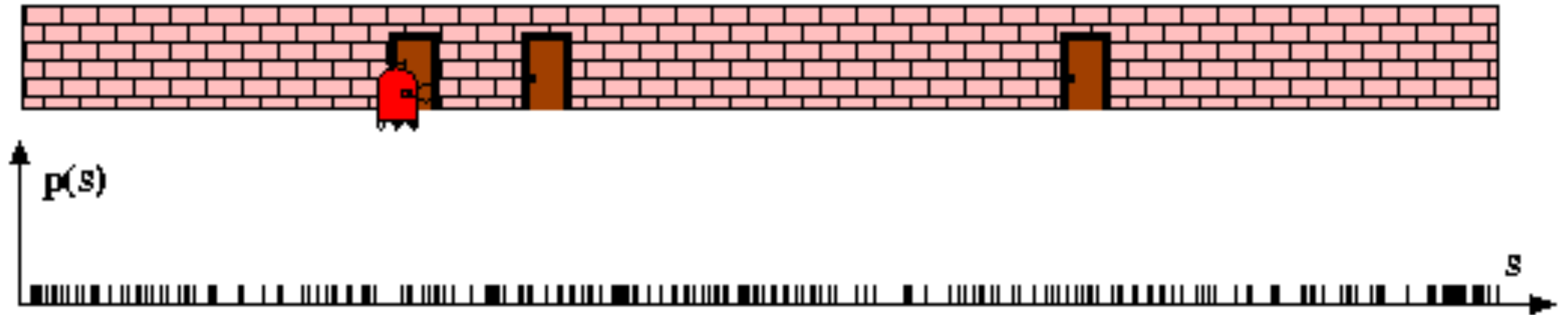
- Each particle is a potential **pose** of the robot
- Proposal distribution is the motion model of the robot (**prediction step**)
- The observation model is used to compute the importance weight (**correction step**)

Randomized algorithms are usually called Monte Carlo algorithms, therefore we call this:

**Monte-Carlo Localization**



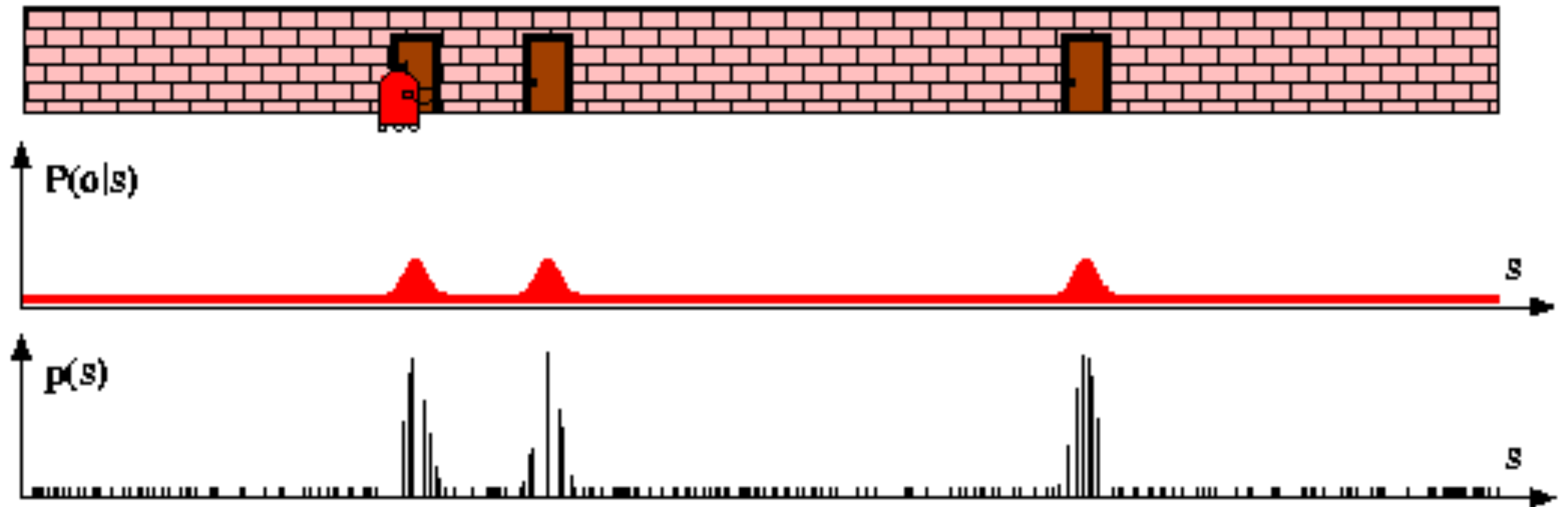
# A Simple Example



- The initial belief is a uniform distribution (global localization).
- This is represented by an (approximately) uniform sampling of initial particles.



# Sensor Information

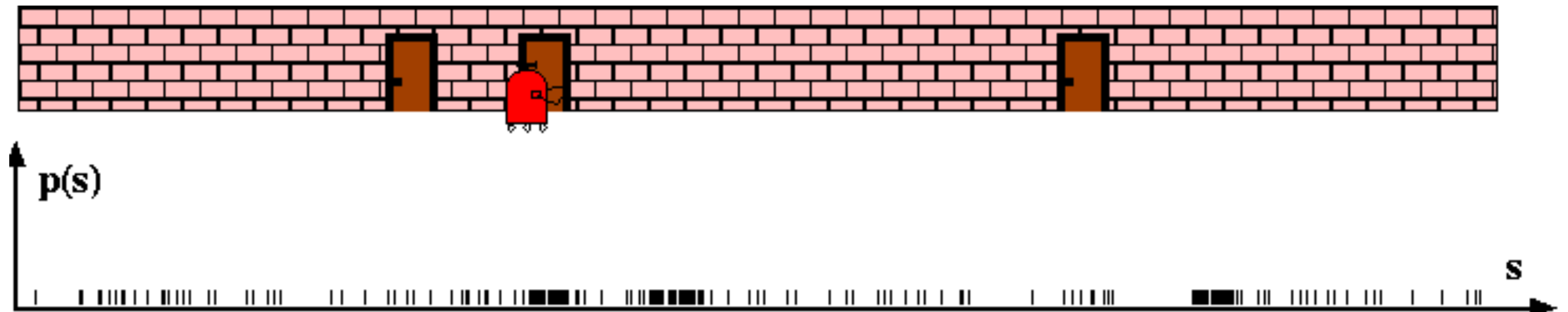


The sensor model  $p(z_t | x_t^{[m]})$  is used to compute the new importance weights:

$$w_t^{[m]} \leftarrow p(z_t | x_t^{[m]})$$



# Robot Motion



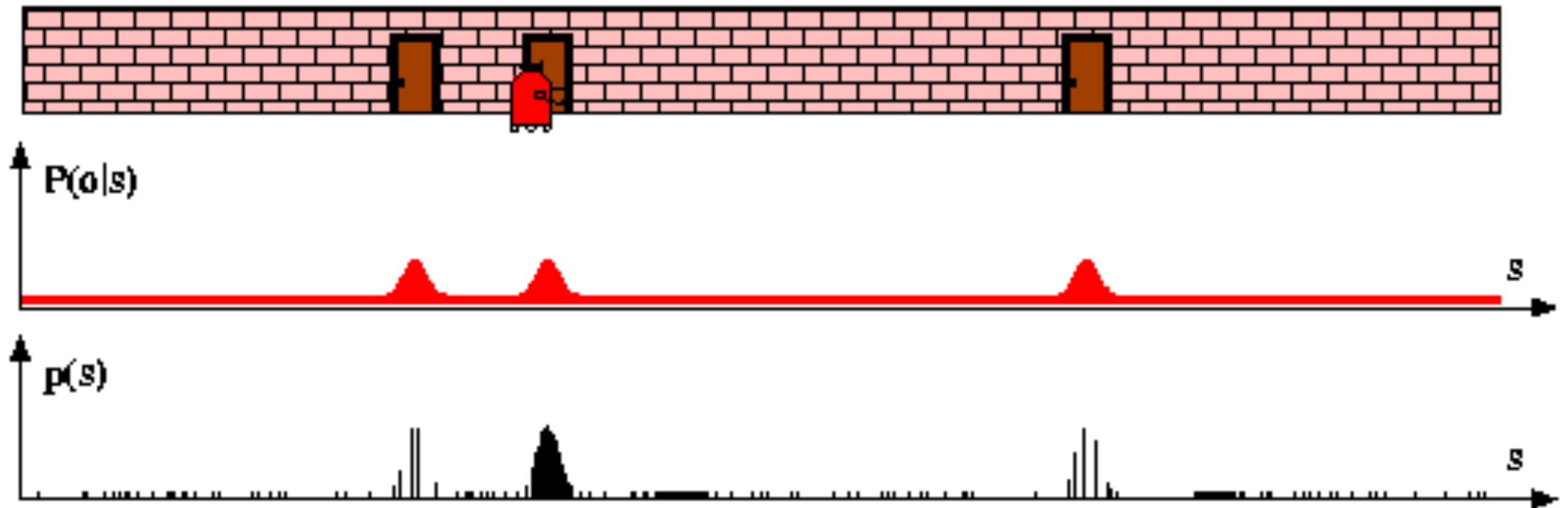
After resampling and applying the motion model

$p(x_t \mid u_t, x_{t-1}^{[m]})$  the particles are distributed more densely at three locations.





# Sensor Information

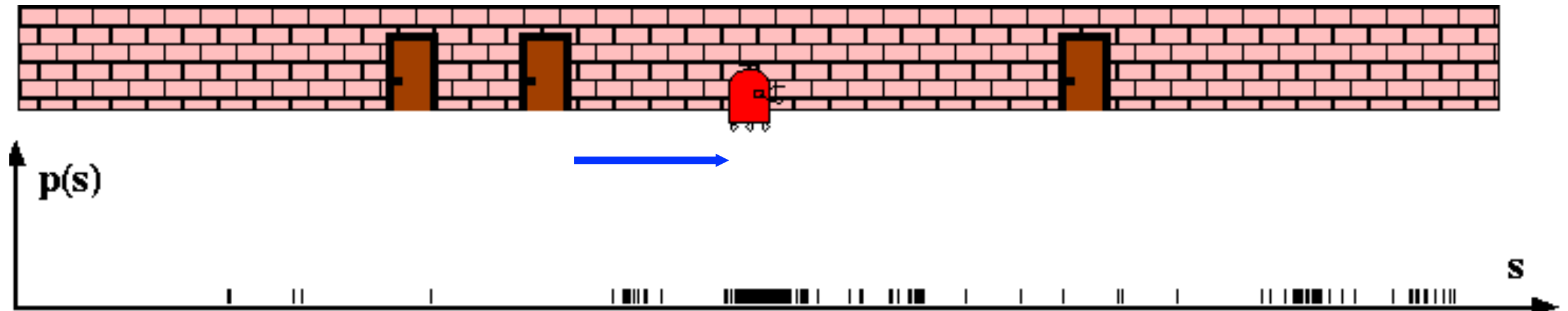


Again, we set the new importance weights equal to the sensor model.

$$w_t^{[m]} \leftarrow p(z_t \mid x_t^{[m]})$$



# Robot Motion



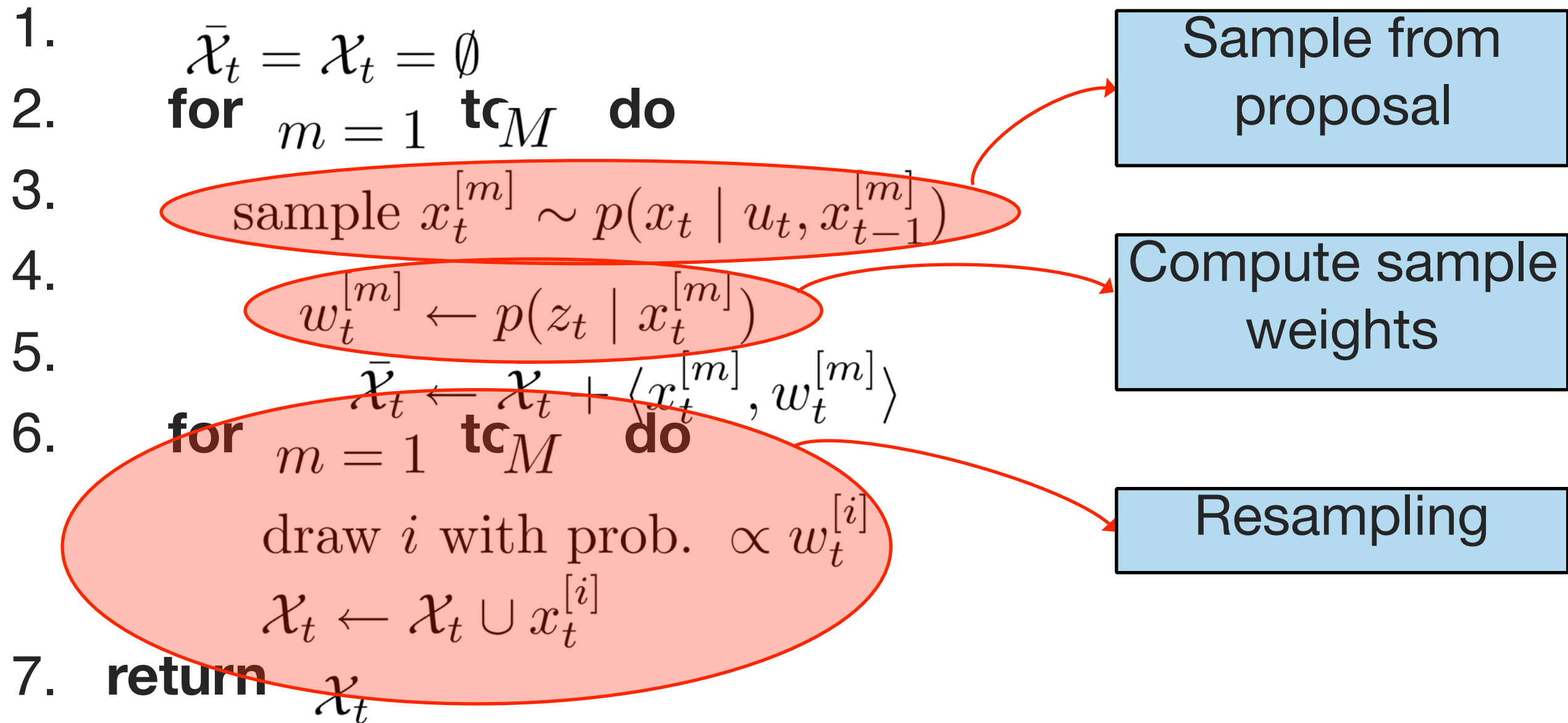
Resampling and application of the motion model:  
One location of dense particles is left.

**The robot is localized.**



# A Closer Look at the Algorithm...

Algorithm *Particle\_filter* ( $\mathcal{X}_t, u_t, z_t$ ) :



# Sampling from Proposal

This can be done in the following ways:

- Adding the motion vector to each particle directly  
(this assumes perfect motion)
- Sampling from the motion model  
e.g. for a 2D motion with translation velocity  $v$  and rotation velocity  $w$  we have:

$$\text{sample } x_t^{[m]} \sim p(x_t | u_t, x_{t-1}^{[m]})$$

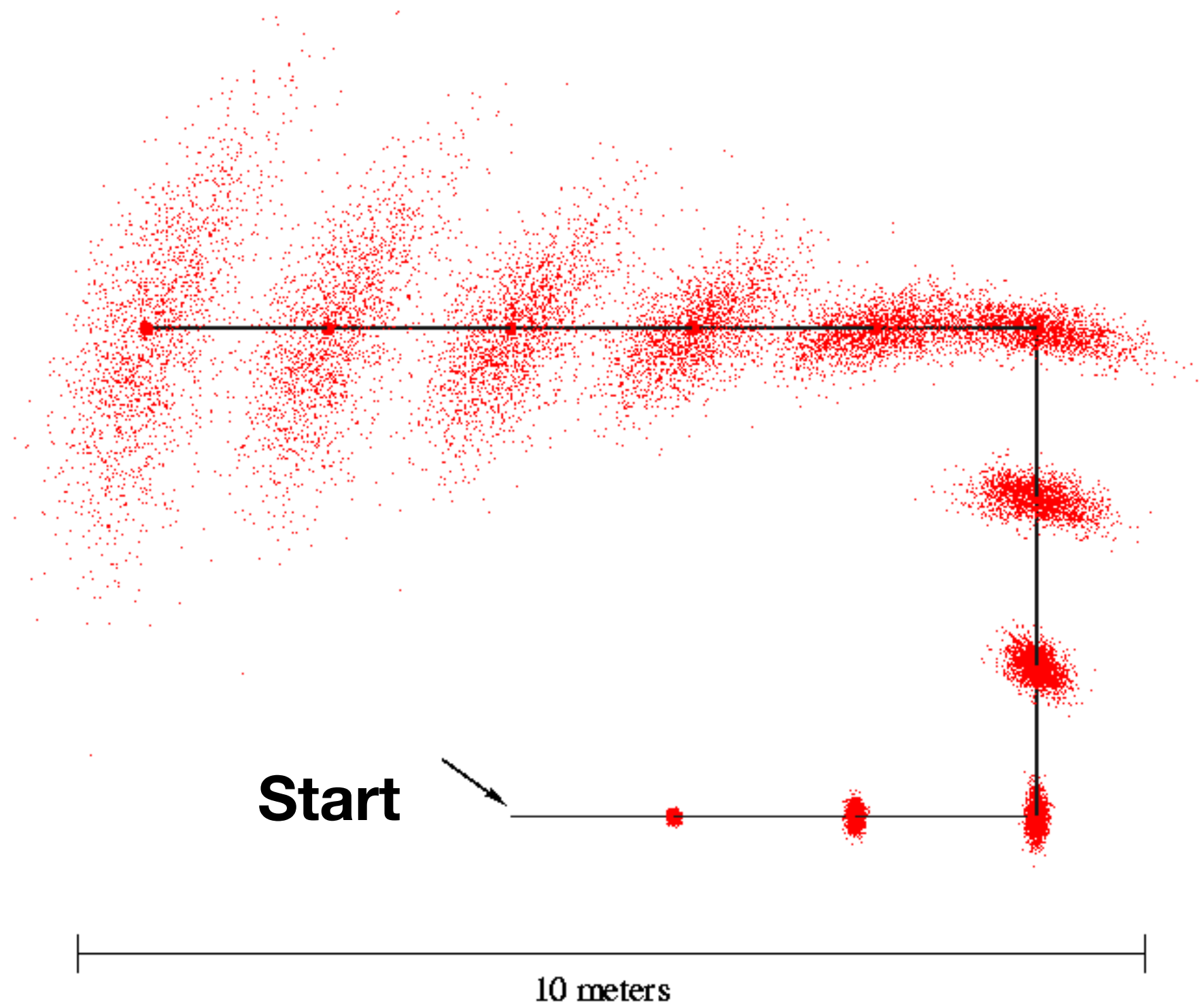
$$\mathbf{u}_t = \begin{pmatrix} v_t \\ w_t \end{pmatrix} \quad \mathbf{x}_t = \begin{pmatrix} x_t \\ y_t \\ \theta_t \end{pmatrix}$$

Diagram illustrating the components of the motion vector  $\mathbf{u}_t$  and the state vector  $\mathbf{x}_t$ :

- $\mathbf{u}_t$  is a 2D vector with components  $v_t$  (translation velocity) and  $w_t$  (rotation velocity).
- $\mathbf{x}_t$  is a 3D vector with components  $x_t$  (position),  $y_t$  (position), and  $\theta_t$  (orientation).
- The components  $x_t$  and  $y_t$  are grouped under the label "Position".
- The component  $\theta_t$  is grouped under the label "Orientation".



# Motion Model Sampling (Example)



# Computation of Importance Weights

Computation of the sample weights:

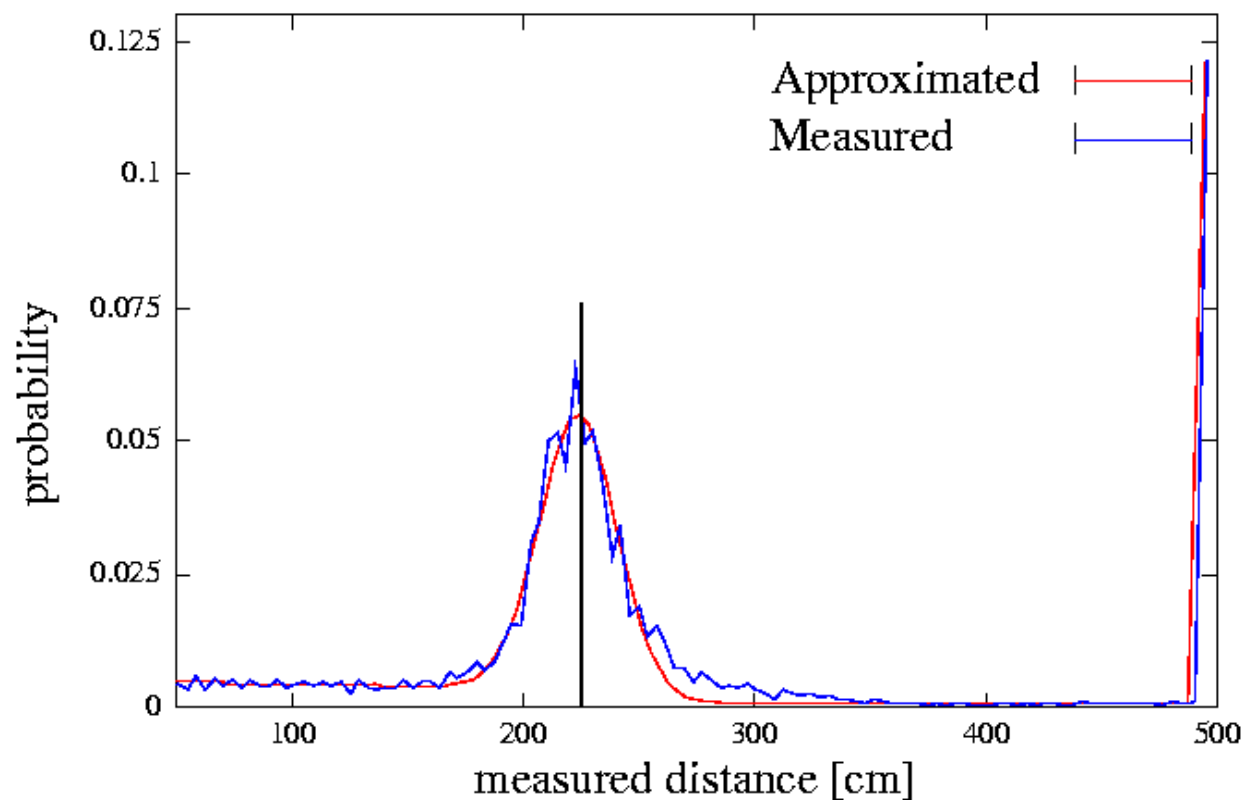
- Proposal distribution:  $g(x_t^{[m]}) = p(x_t^{[m]} | u_t, x_{t-1}^{[m]})\text{Bel}(x_{t-1}^{[m]})$   
(we sample from that using the motion model)
- Target distribution (new belief):  $f(x_t^{[m]}) = \text{Bel}(x_t^{[m]})$   
(we can not directly sample from that  $\rightarrow$  importance sampling)  
 $w_t^{[m]} \leftarrow p(z_t | x_t^{[m]})$
- Computation of importance weights:

$$w_t^{[m]} = \frac{f(x_t^{[m]})}{g(x_t^{[m]})} \propto \frac{p(z_t | x_t^{[m]})p(x_t^{[m]} | u_t, x_{t-1}^{[m]})\text{Bel}(x_{t-1}^{[m]})}{p(x_t^{[m]} | u_t, x_{t-1}^{[m]})\text{Bel}(x_{t-1}^{[m]})} = p(z_t | x_t^{[m]})$$

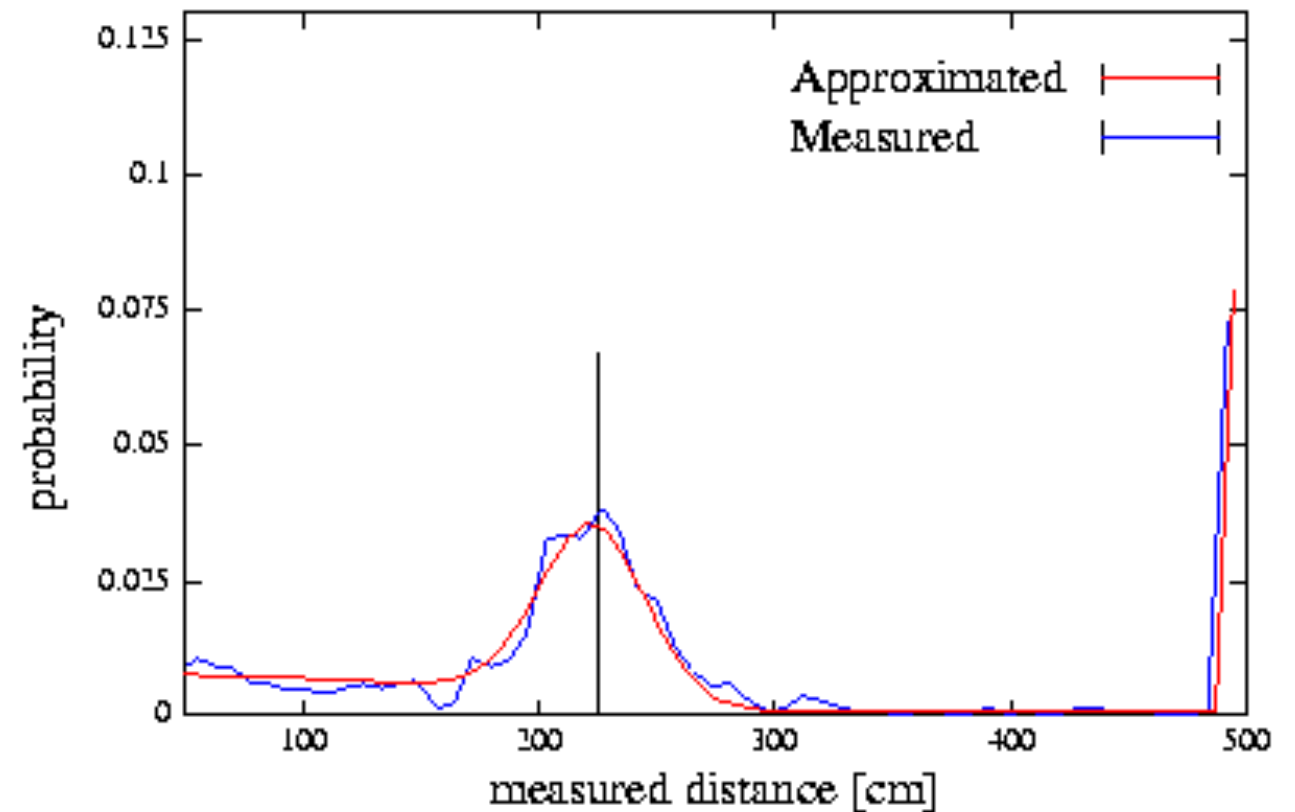


# Proximity Sensor Models

- How can we obtain the sensor model  $p(z_t | x_t^{[m]})$  ?
- Sensor Calibration:



Laser sensor



Sonar sensor



# Resampling

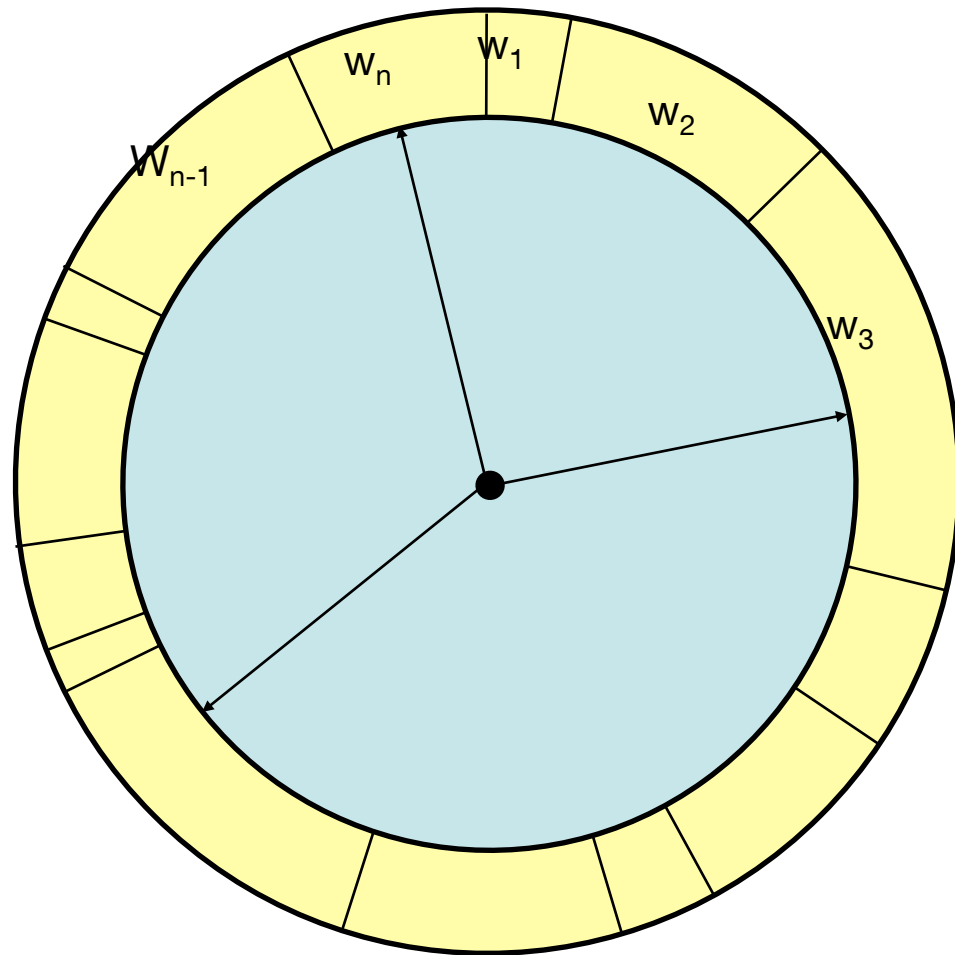
- Given: Set  $\bar{\mathcal{X}}_t$  of weighted samples.
- Wanted : Random sample, where the probability of drawing  $x_i$  is equal to  $w_i$ .
- Typically done  $M$  times with replacement to generate new samples. **for  $m = 1$  to  $M$  do**

draw  $i$  with prob.  $\propto w_t^{[i]}$   
 $\mathcal{X}_t \leftarrow \mathcal{X}_t \cup x_t^{[i]}$

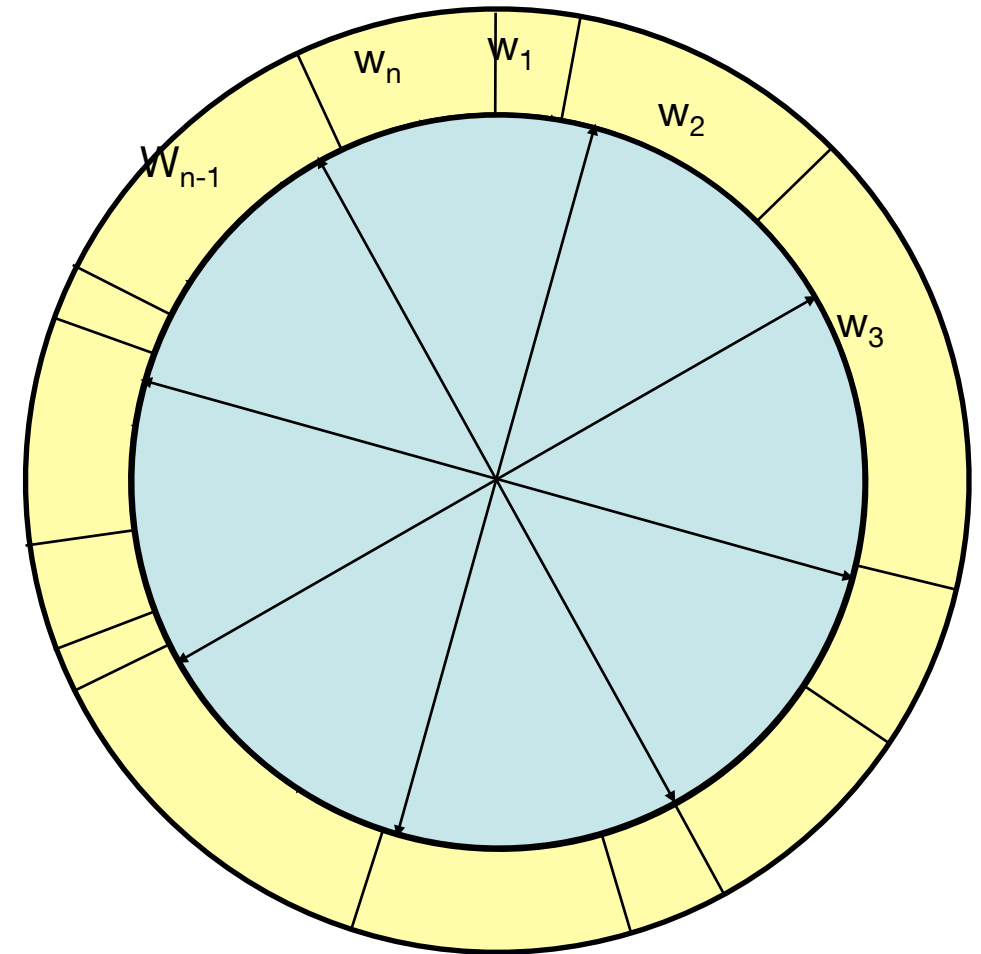




# Resampling



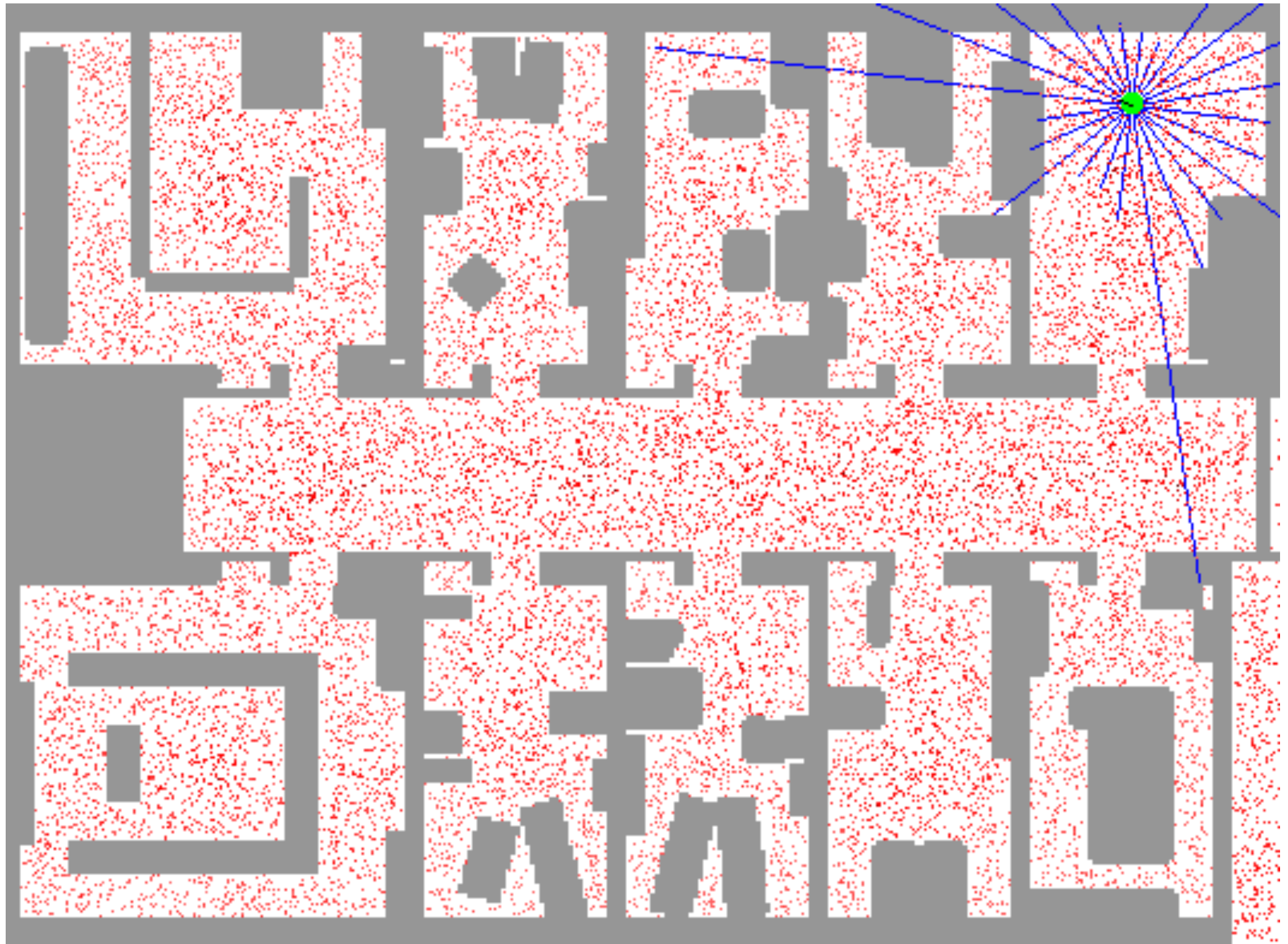
- Standard  $n$ -times sampling results in high variance
- This requires more particles
- $O(n \log n)$  complexity



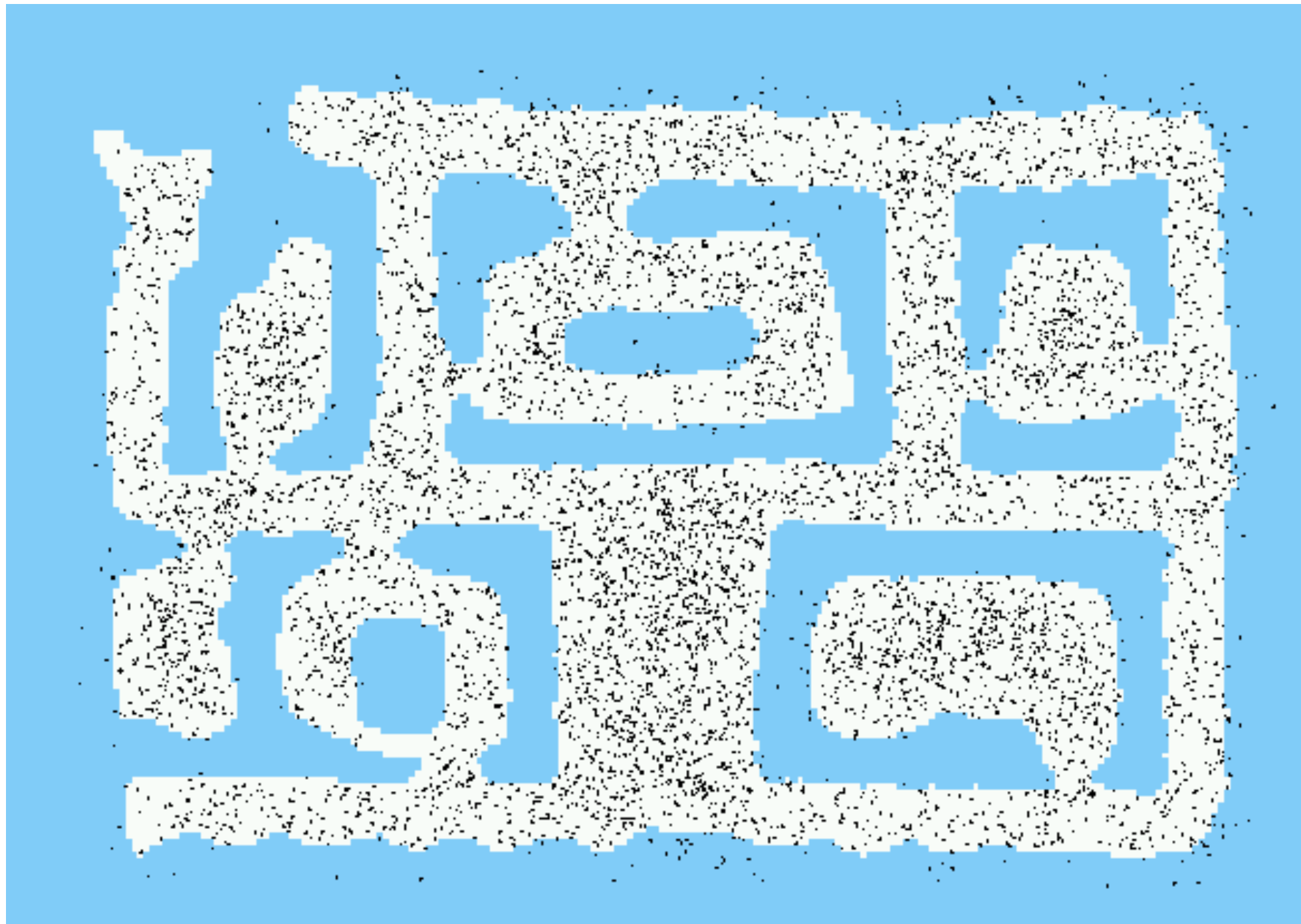
- Instead: low variance sampling only samples once
- Linear time complexity
- Easy to implement



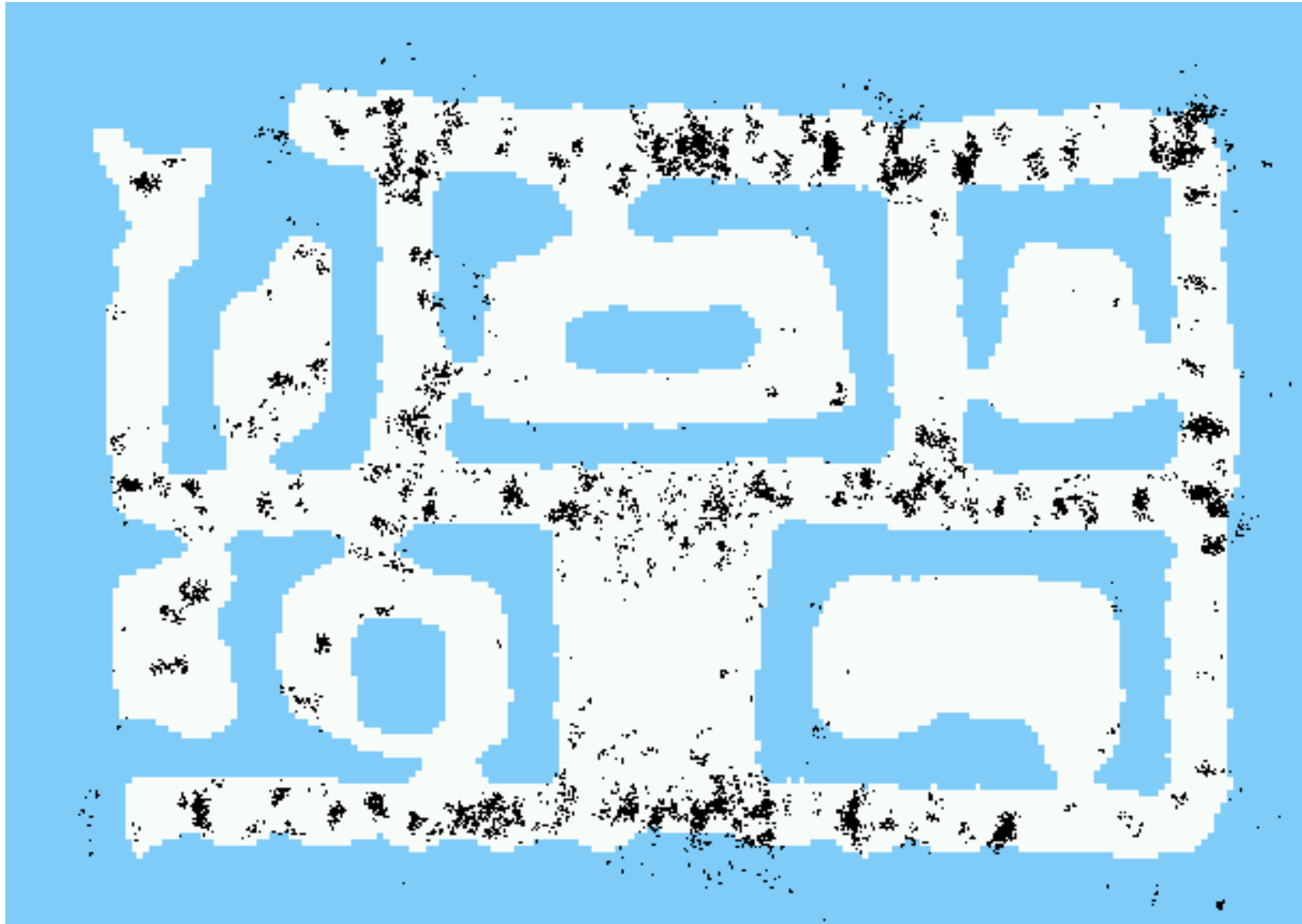
# Sample-based Localization (sonar)



# Initial Distribution



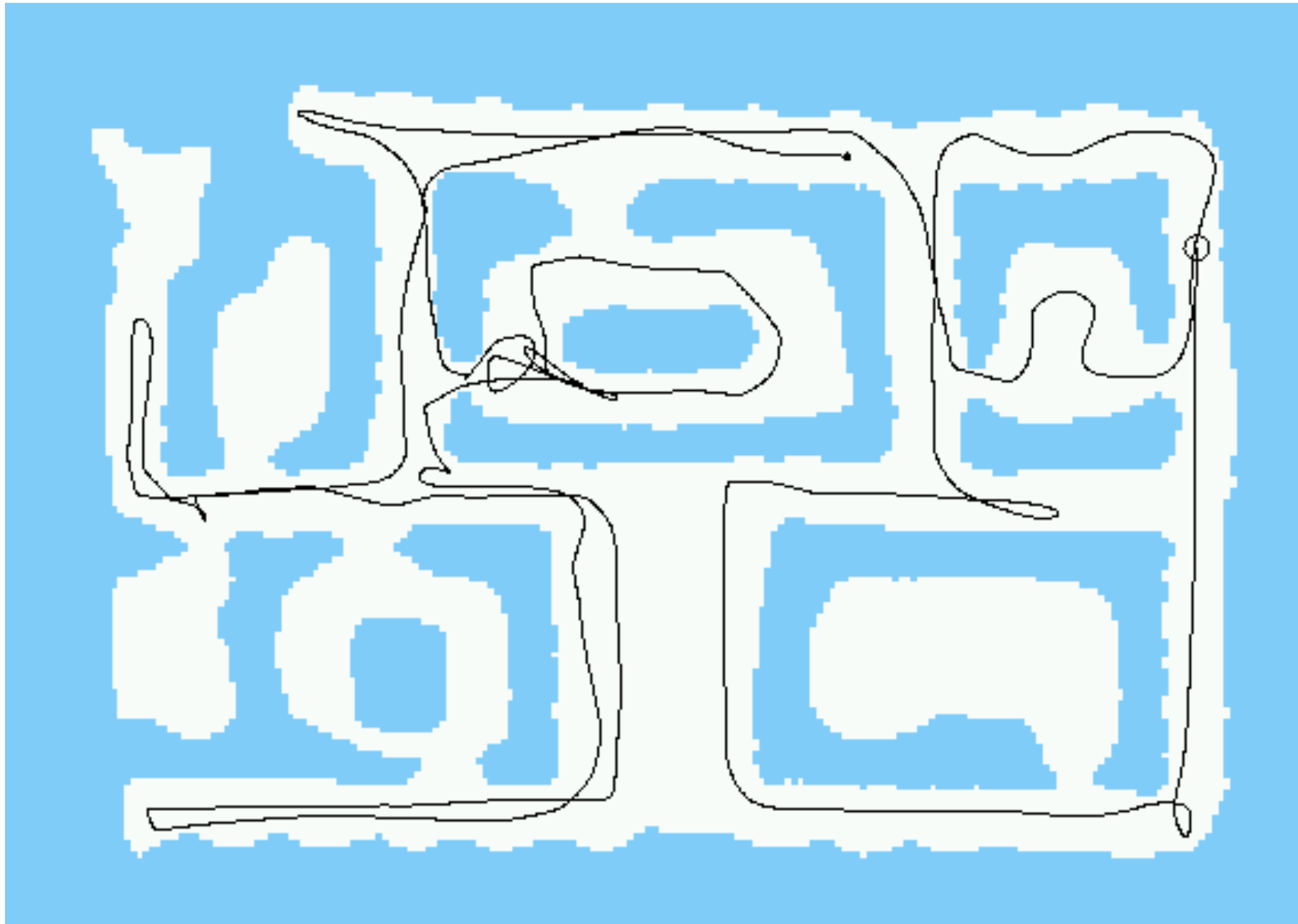
# After Ten Ultrasound Scans



# After 65 Ultrasound Scans



# Estimated Path





# Kidnapped Robot Problem

The approach described so far is able to

- track the pose of a mobile robot and to
- globally localize the robot.
- How can we deal with localization errors (i.e., the kidnapped robot problem)?

**Idea:** Introduce uniform samples at every resampling step

- This adds new hypotheses



# Summary

- There are mainly 4 different types of sampling methods: Transformation method, rejection sampling, importance sampling and MCMC
- Transformation only rarely applicable
- Rejection sampling is often very inefficient
- Importance sampling is used in the particle filter which can be used for robot localization
- An efficient implementation of the resampling step is the low variance sampling







# Markov Chain Monte Carlo

# Markov Chain Monte Carlo

- In high-dimensional spaces, rejection sampling and importance sampling are very inefficient
- An alternative is Markov Chain Monte Carlo (MCMC)
- It keeps a record of the current state and the proposal depends on that state
- Most common algorithms are the Metropolis-Hastings algorithm and Gibbs Sampling

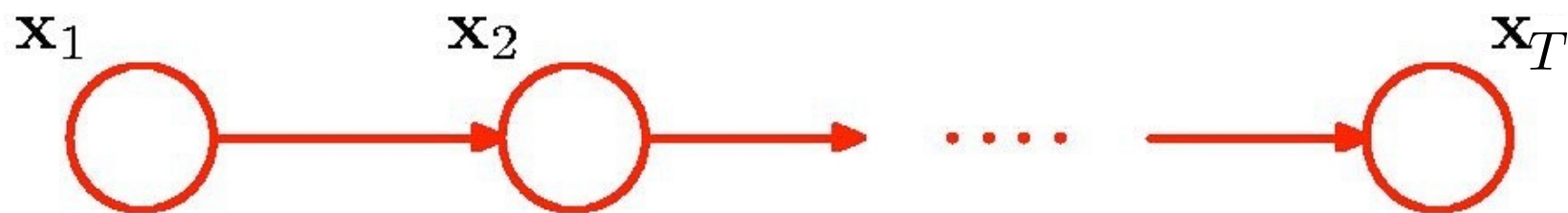


# Markov Chains Revisited

A Markov Chain is a distribution over discrete-state random variables  $\mathbf{x}_1, \dots, \mathbf{x}_M$  so that

$$p(\mathbf{x}_1, \dots, \mathbf{x}_T) = p(\mathbf{x}_1)p(\mathbf{x}_2 \mid \mathbf{x}_1) \cdots = p(\mathbf{x}_1) \prod_{t=2}^T p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$

The graphical model of a Markov chain is this:

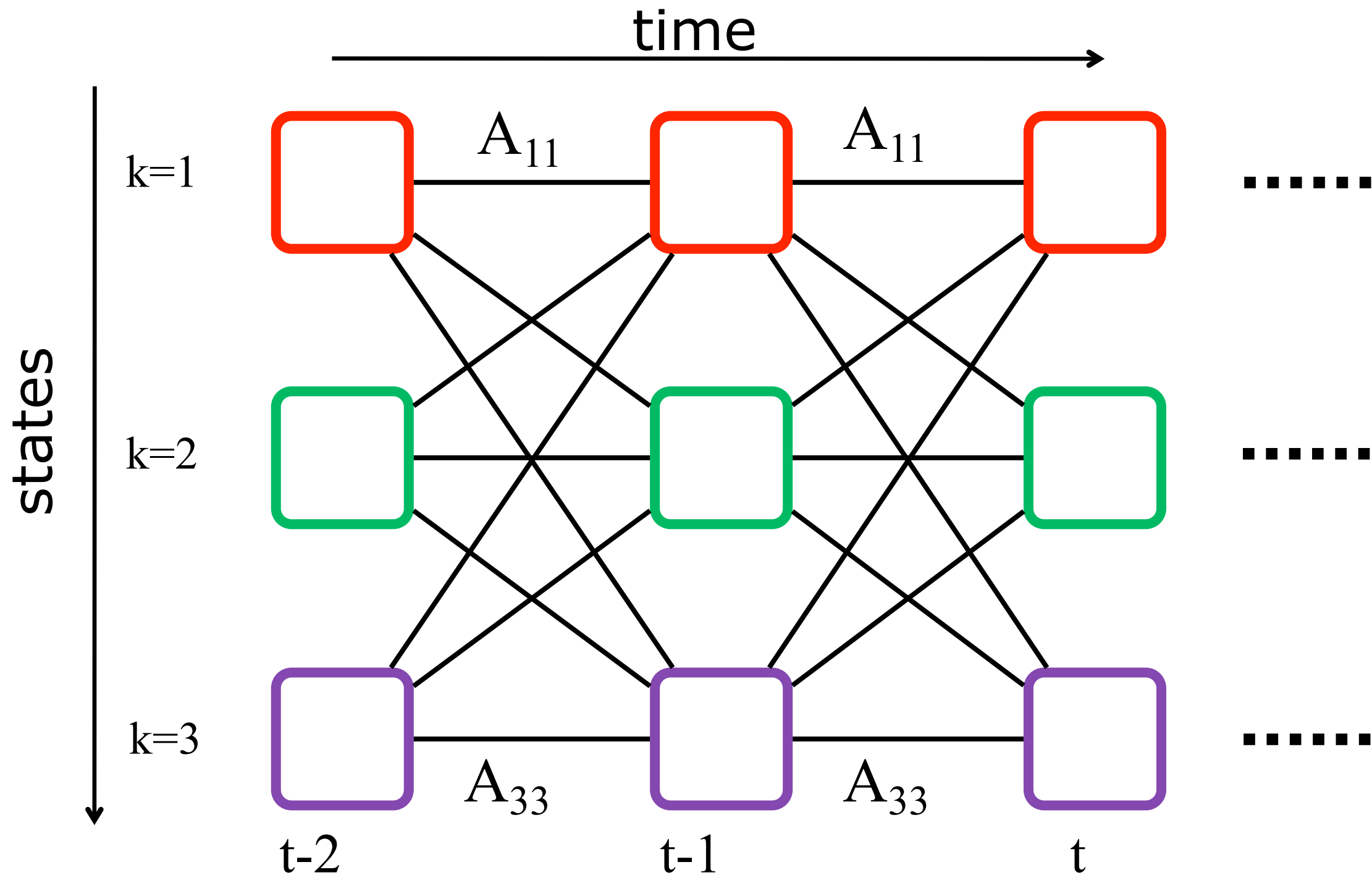


We will denote  $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$  as a row vector  $\pi_t$

A Markov chain can also be visualized as a **state transition diagram**.



# The State Transition Diagram



# Some Notions

- The Markov chain is said to be **homogeneous** if the transitions probabilities are all the same at every time step  $t$  (here we only consider homogeneous Markov chains)
- The transition matrix is **row-stochastic**, i.e. all entries are between 0 and 1 and all rows sum up to 1
- Observation: the probabilities of reaching the states can be computed using a vector-matrix multiplication



# The Stationary Distribution

The probability to reach state  $k$  is  $\pi_{k,t} = \sum_{i=1}^K \pi_{i,t-1} A_{ik}$

Or, in matrix notation:  $\pi_t = \pi_{t-1} A$

We say that  $\pi_t$  is **stationary** if  $\pi_t = \pi_{t-1}$

## Questions:

- How can we know that a stationary distributions exists?
- And if it exists, how do we know that it is unique?



# The Stationary Distribution (Existence)

To find a stationary distribution we need to solve the eigenvector problem  $A^T \mathbf{v} = \mathbf{v}$

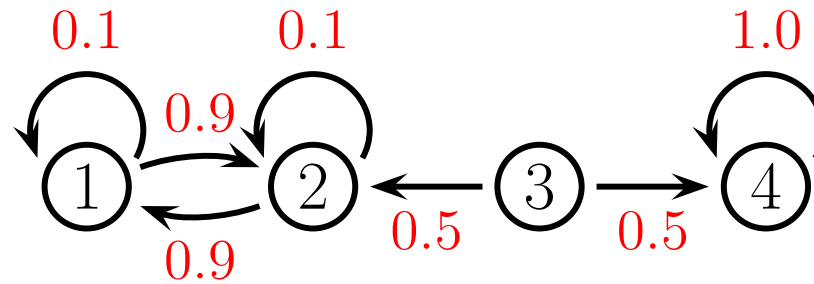
The stationary distribution is then  $\pi = \mathbf{v}^T$  where  $\mathbf{v}$  is the eigenvector for which the eigenvalue is 1.

This eigenvector needs to be normalized so that it is a valid distribution.

**Theorem (Perron-Frobenius):** Every row-stochastic matrix has such an eigen vector, but this vector may not be unique.



# Stationary Distribution (Uniqueness)



- A Markov chain can have many stationary distributions
- Sufficient for a unique stationary distribution: we can reach every state from any other state in finite steps at non-zero probability (i.e. the chain is **ergodic**)
- This is equivalent to the property that the transition matrix is **irreducible**:

$$\forall i, j \exists m \quad (A^m)_{ij} > 0$$





# Main Idea of MCMC

- So far, we specified the transition probabilities and analysed the resulting distribution
- This was used, e.g. in HMMs

Now:

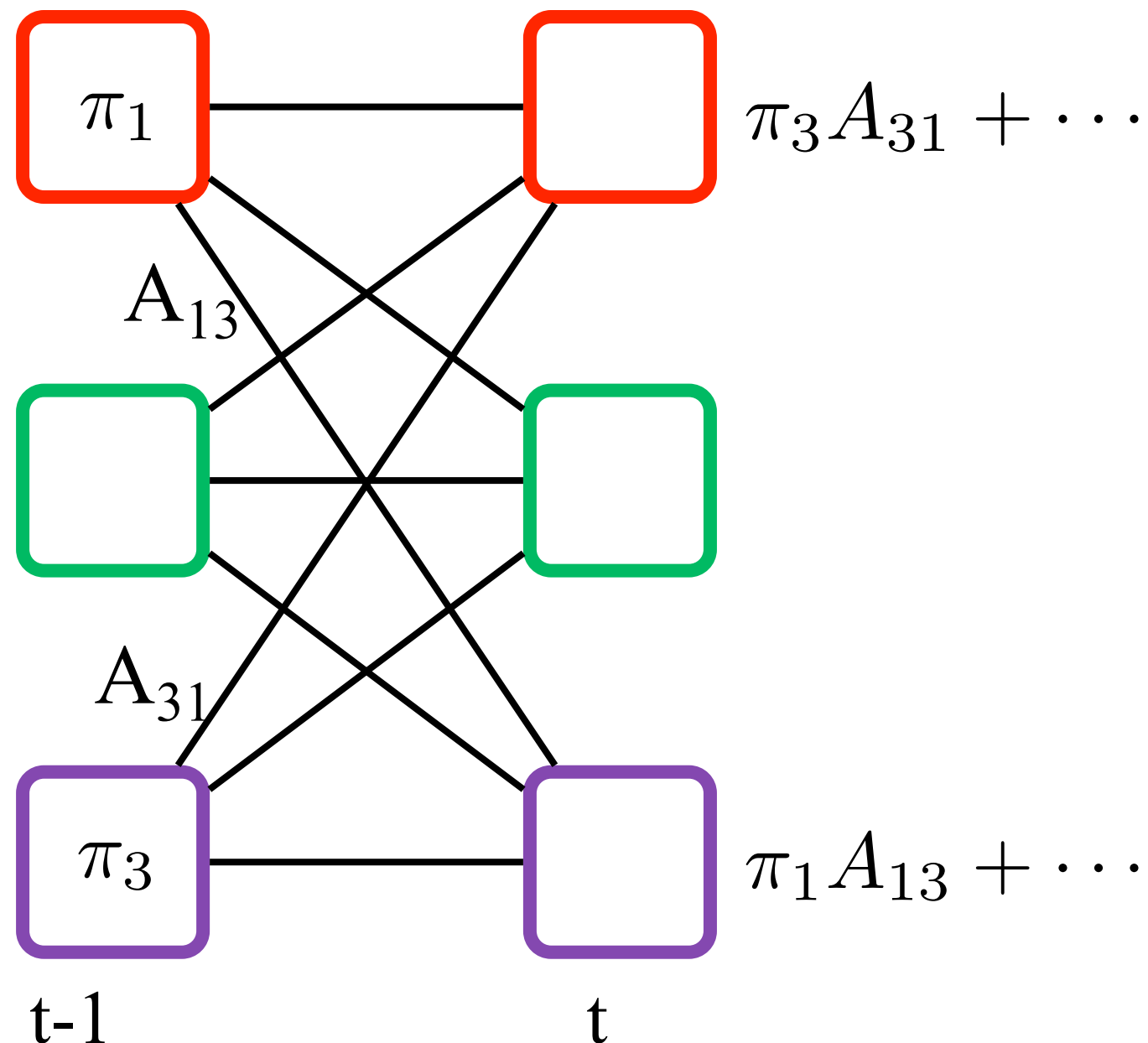
- We want to **sample** from an arbitrary distribution
- To do that, we design the transition probabilities so that the resulting stationary distribution is our desired (target) distribution!



# Detailed Balance

**Definition:** A transition distribution  $\pi_t$  satisfies the property of **detailed balance** if  $\pi_i A_{ij} = \pi_j A_{ji}$

The chain is then said to be **reversible**.



# Making a Distribution Stationary

**Theorem:** If a Markov chain with transition matrix  $A$  is irreducible and satisfies detailed balance wrt. the distribution  $\pi$ , then  $\pi$  is a stationary distribution of the chain.

**Proof:**

$$\sum_{i=1}^K \pi_i A_{ij} = \sum_{i=1}^K \pi_j A_{ji} = \pi_j \sum_{i=1}^K A_{ji} = \pi_j \quad \forall j$$

it follows  $\pi = \pi A$ .

This is a sufficient, but not necessary condition.



# Sampling with a Markov Chain

The idea of MCMC is to sample state transitions based on a **proposal distribution**  $q$ .

The most widely used algorithm is the Metropolis-Hastings (MH) algorithm.

In MH, the decision whether to stay in a given state is based on a given probability.

If the proposal distribution is  $q(\mathbf{x}' | \mathbf{x})$ , then we stay in state  $\mathbf{x}'$  with probability

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

Unnormalized target distribution  $\rightarrow$



# The Metropolis-Hastings Algorithm

- Initialize  $x^0$
- for  $s = 0, 1, 2, \dots$ 
  - define  $x = x^s$
  - sample  $x' \sim q(x' \mid x)$ 
    - compute acceptance probability
$$\alpha = \frac{\tilde{p}(x')q(x \mid x')}{\tilde{p}(x)q(x' \mid x)}$$
    - compute  $r = \min(1, \alpha)$
    - sample  $u \sim U(0, 1)$
    - set new sample to

$$x^{s+1} = \begin{cases} x' & \text{if } u < r \\ x^s & \text{if } u \geq r \end{cases}$$



# Why Does This Work?

We have to prove that the transition probability of the MH algorithm satisfies detailed balance wrt the target distribution.

**Theorem:** If  $p_{MH}(\mathbf{x}' | \mathbf{x})$  is the transition probability of the MH algorithm, then

$$p(\mathbf{x})p_{MH}(\mathbf{x}' | \mathbf{x}) = p(\mathbf{x}')p_{MH}(\mathbf{x} | \mathbf{x}')$$

**Proof:**



# Why Does This Work?

We have to prove that the transition probability of the MH algorithm satisfies detailed balance wrt the target distribution.

**Theorem:** If  $p_{MH}(\mathbf{x}' | \mathbf{x})$  is the transition probability of the MH algorithm, then

$$p(\mathbf{x})p_{MH}(\mathbf{x}' | \mathbf{x}) = p(\mathbf{x}')p_{MH}(\mathbf{x} | \mathbf{x}')$$

**Note: All formulations are valid for discrete and for continuous variables!**



# Choosing the Proposal

- A proposal distribution is valid if it gives a non-zero probability of moving to the states that have a non-zero probability in the target.
- A good proposal is the Gaussian, because it has a non-zero probability for all states.
- **However:** the variance of the Gaussian is important!
  - with low variance, the sampler does not explore sufficiently, e.g. it is fixed to a particular mode
  - with too high variance, the proposal is rejected too often, the samples are a bad approximation



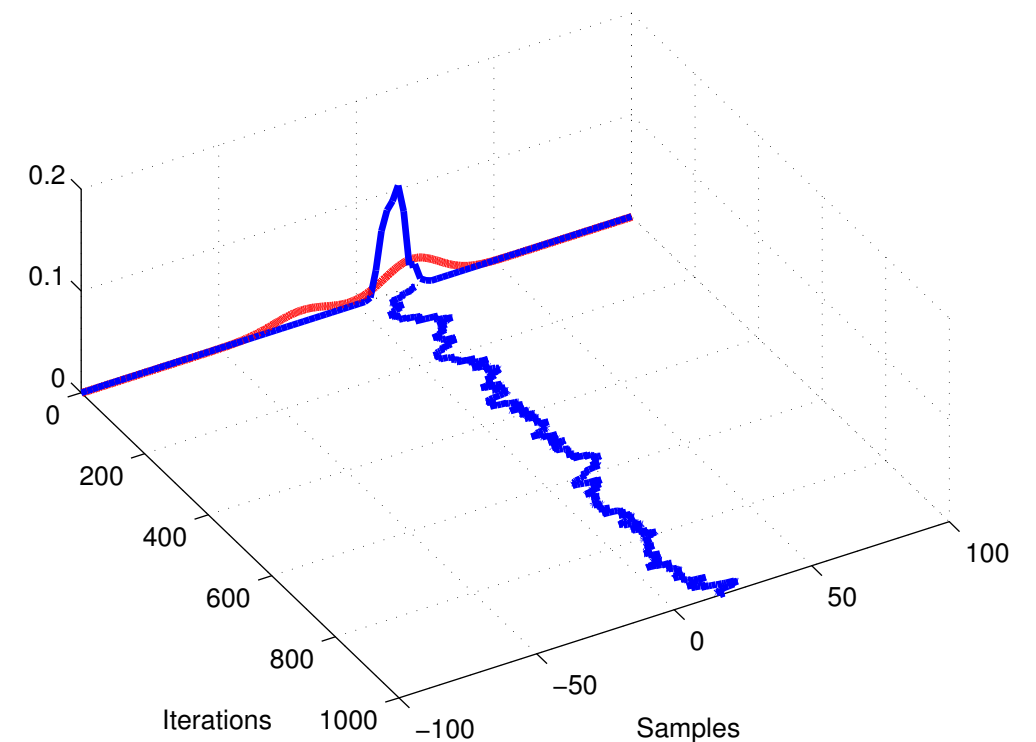


# Example

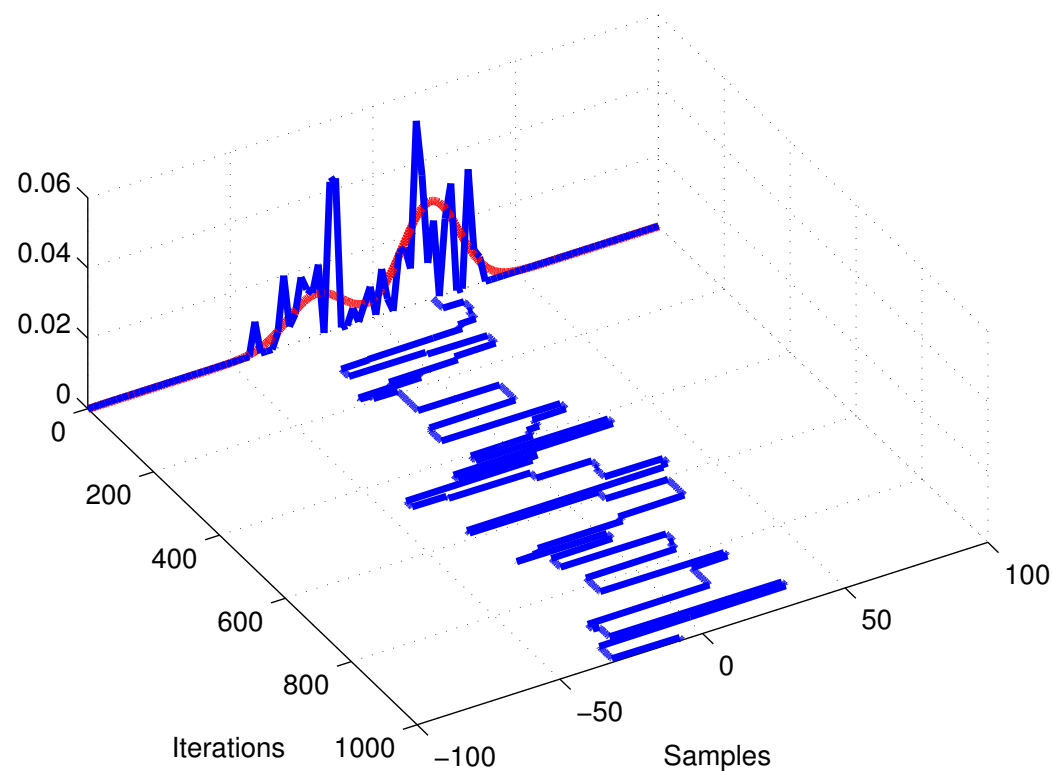
Target is a mixture of 2  
1D Gaussians.

Proposal is a Gaussian  
with different variances.

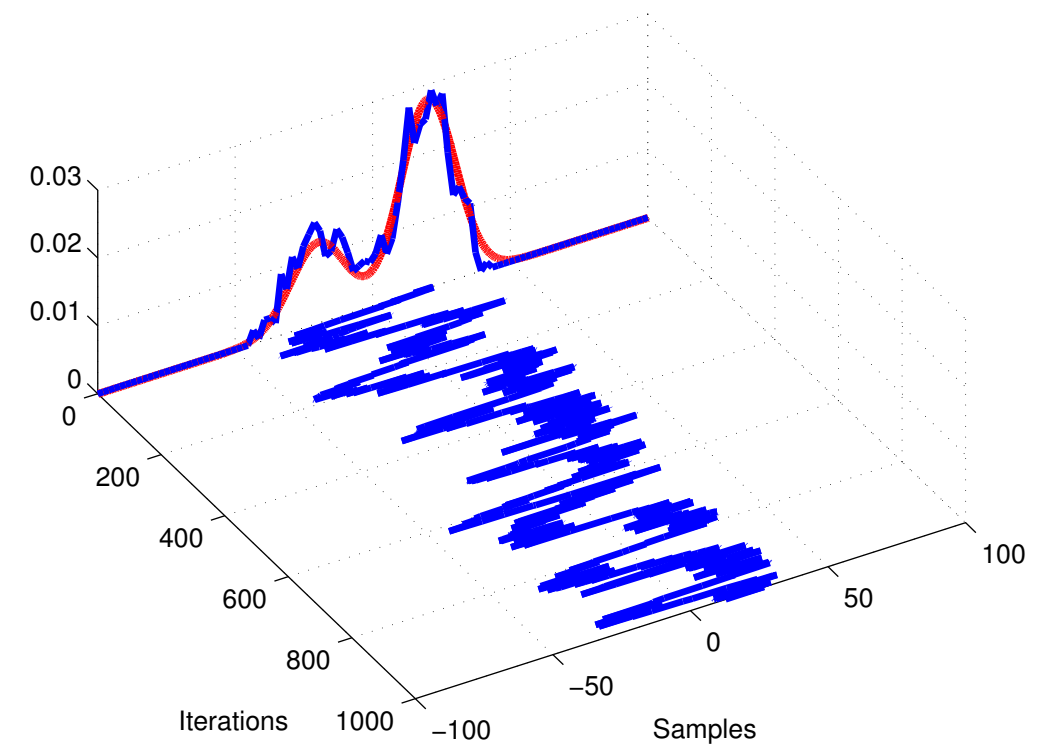
MH with  $N(0, 1.000^2)$  proposal



MH with  $N(0, 500.000^2)$  proposal



MH with  $N(0, 8.000^2)$  proposal



# Gibbs Sampling

- Initialize  $\{z_i : i = 1, \dots, M\}$
- For  $\tau = 1, \dots, T$ 
  - Sample  $z_1^{(\tau+1)} \sim p(z_1 \mid z_2^{(\tau)}, \dots, z_M^{(\tau)})$
  - Sample  $z_2^{(\tau+1)} \sim p(z_2 \mid z_1^{(\tau+1)}, \dots, z_M^{(\tau)})$
  - ...
  - Sample  $z_M^{(\tau+1)} \sim p(z_M \mid z_1^{(\tau+1)}, \dots, z_{M-1}^{(\tau+1)})$

**Idea:** sample from the full conditional

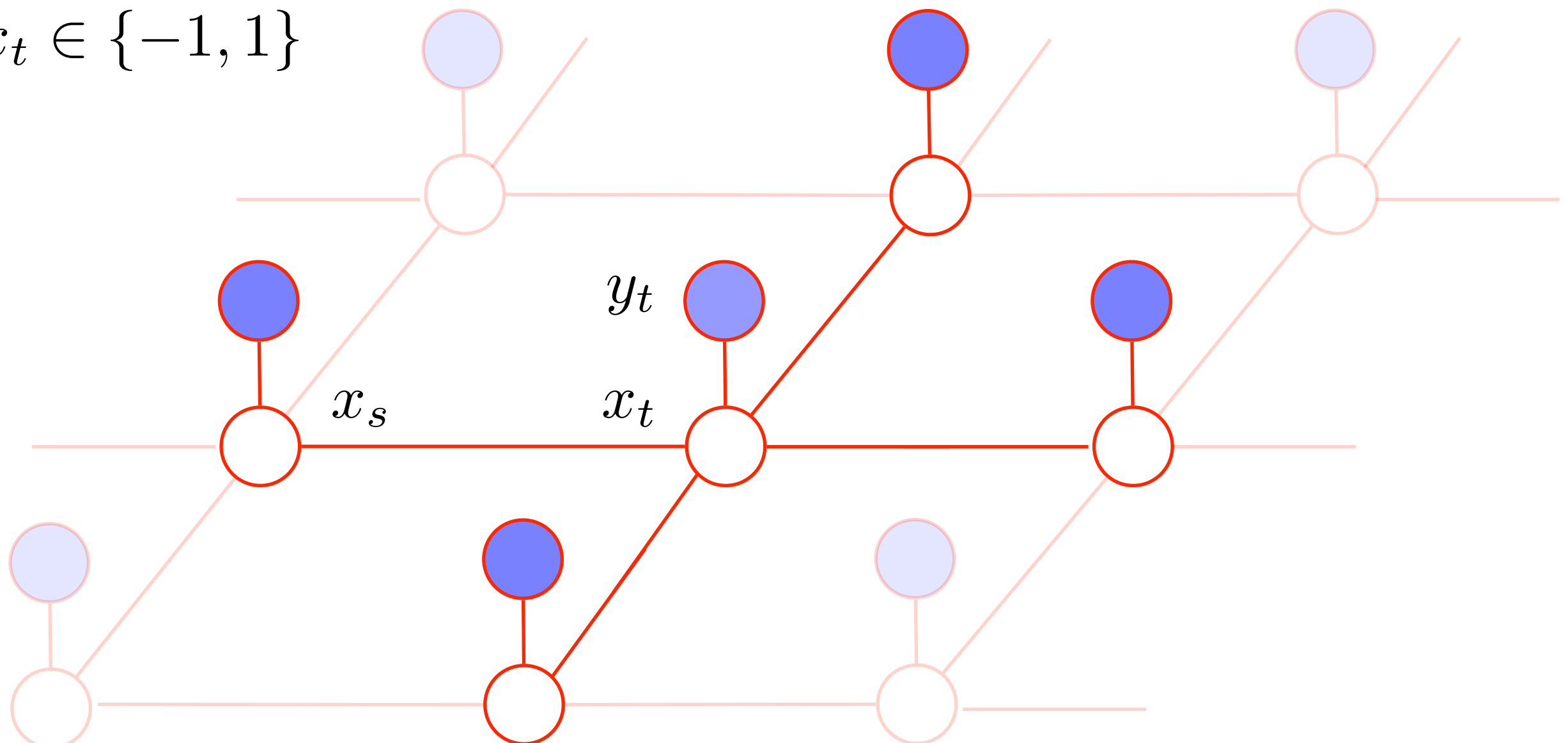
This can be obtained, e.g. from the Markov blanket in graphical models.



# Gibbs Sampling: Example

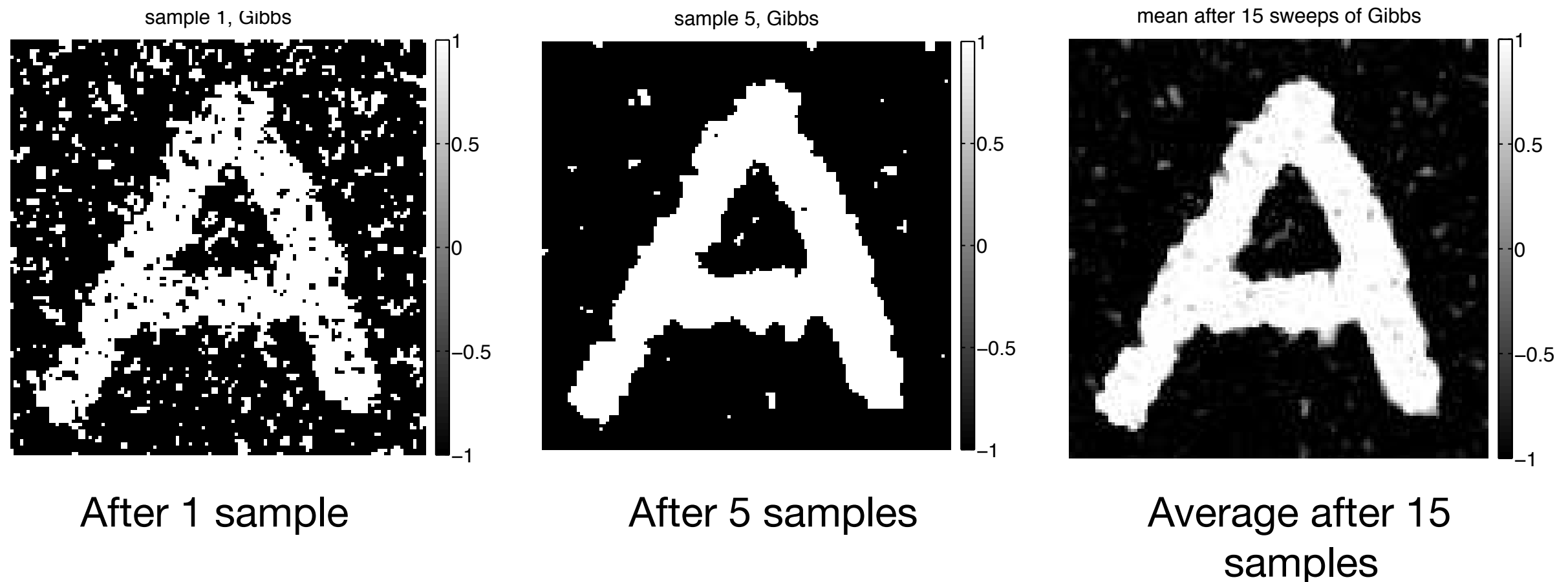
- Use an MRF on a binary image with edge potentials  $\psi(x_s, x_t) = \exp(J x_s x_t)$  (“Ising model”) and node potentials  $\psi(x_t) = \mathcal{N}(y_t | x_t, \sigma^2)$

$$x_t \in \{-1, 1\}$$



# Gibbs Sampling: Example

- Use an MRF on a binary image with edge potentials  $\psi(x_s, x_t) = \exp(J x_s x_t)$  (“Ising model”) and node potentials  $\psi(x_t) = \mathcal{N}(y_t | x_t, \sigma^2)$
- Sample each pixel in turn



# Gibbs Sampling is a Special Case of MH

- The proposal distribution in Gibbs sampling is

$$q(\mathbf{x}' \mid \mathbf{x}) = p(x'_i \mid \mathbf{x}_{-i}) \mathbb{I}(\mathbf{x}'_{-i} = \mathbf{x}_{-i})$$

- This leads to an acceptance rate of:

$$\alpha = \frac{p(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')}{p(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x})} = \frac{p(x'_i \mid \mathbf{x}'_{-i})p(\mathbf{x}'_{-i})p(x_i \mid \mathbf{x}'_{-i})}{p(x_i \mid \mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x'_i \mid \mathbf{x}_{-i})} = 1$$

- Although the acceptance is 100%, Gibbs sampling does not converge faster, as it only updates one variable at a time.



# Summary

- Markov Chain Monte Carlo is a family of sampling algorithms that can sample from arbitrary distributions by moving in state space
- Most used methods are the Metropolis-Hastings (MH) and the Gibbs sampling method
- MH uses a proposal distribution and accepts a proposed state randomly
- Gibbs sampling does not use a proposal distribution, but samples from the full conditionals





# **11. Variational Inference**

# Motivation

- A major task in probabilistic reasoning is to evaluate the **posterior** distribution  $p(Z | X)$  of a set of latent variables  $Z$  given data  $X$  (**inference**)

**However:** This is often not tractable, e.g. because the latent space is high-dimensional

- Two different solutions are possible: sampling methods and variational methods.
- In variational optimization, we seek a tractable distribution  $q$  that **approximates** the posterior.
- Optimization is done using **functionals**.





# Motivation

- A major task in probabilistic reasoning is to evaluate the **posterior** distribution  $p(Z | X)$  of a set of latent variables  $Z$  given data  $X$  (**inference**)
- **Careful: Different notation!**  
**In Bishop (and in the following slides)**  
 $Z$  are hidden states  
and  $X$  are observations
- In variational optimization, we seek a tractable distribution  $q$  that **approximates** the posterior.
- Optimization is done using **functionals**.



# Variational Inference

In general, variational methods are concerned with mappings that take **functions** as input.

Example: the entropy of a distribution  $p$

$$\mathbb{H}[p] = \int p(x) \log p(x) dx \quad \text{“Functional”}$$

Variational optimization aims at finding **functions** that minimize (or maximize) a given functional.

This is mainly used to find approximations to a given function by choosing from a family.

The aim is mostly tractability and simplification.



# The KL-Divergence

**Aim:** define a functional that resembles a “difference” between distributions  $p$  and  $q$

**Idea:** use the average additional amount of information:

$$-\int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} - \left( -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \right) = -\int p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = \text{KL}(p||q)$$

This is known as the **Kullback-Leibler** divergence

It has the properties:  $\text{KL}(q||p) \neq \text{KL}(p||q)$

$$\text{KL}(p||q) \geq 0$$

$$\text{KL}(p||q) = 0 \Leftrightarrow p \equiv q$$

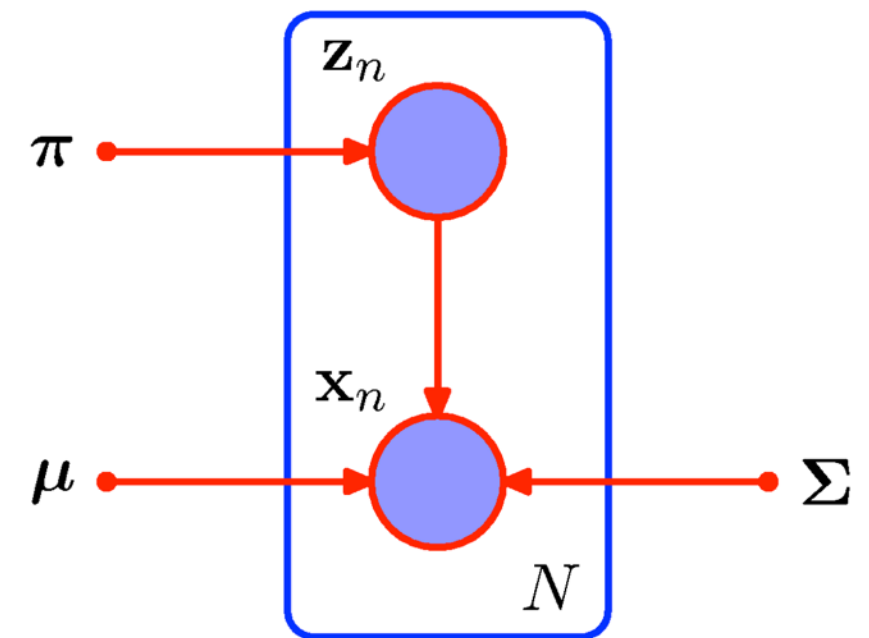
This follows from Jensen's inequality



# Example: A Variational Formulation of EM

Assume for a moment that we observe  $X$  and the binary latent variables  $Z$ . The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^N p(\mathbf{z}_n \mid \pi) p(\mathbf{x}_n \mid \mathbf{z}_n, \mu, \Sigma)$$



# Example: A Variational Formulation of EM

Assume for a moment that we observe  $X$  and the binary latent variables  $Z$ . The likelihood is then:

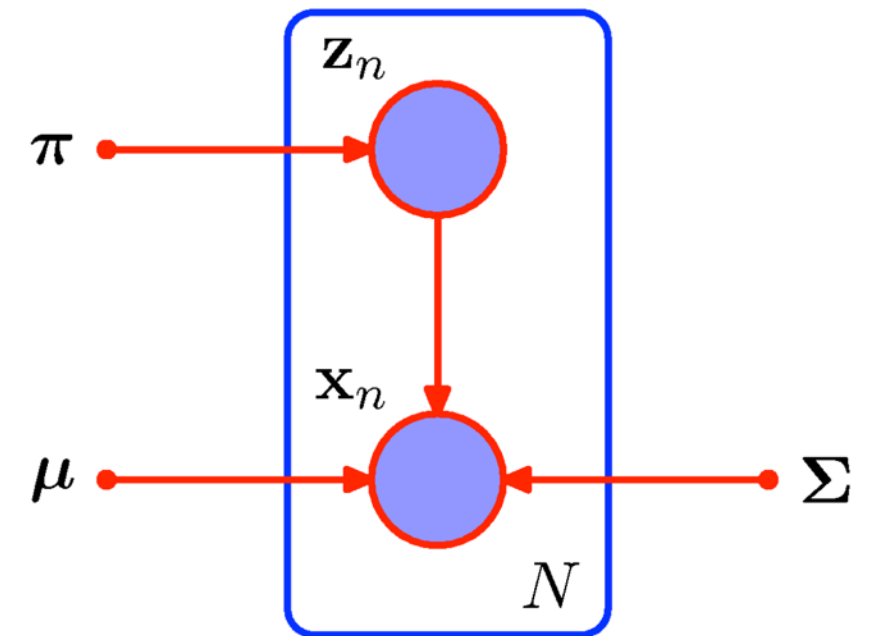
$$p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \prod_{n=1}^N p(\mathbf{z}_n \mid \boldsymbol{\pi}) p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\mu}, \Sigma)$$

**Remember:**

$$z_{nk} \in \{0, 1\}, \quad \sum_{k=1}^K z_{nk} = 1$$

where  $p(\mathbf{z}_n \mid \boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{z_{nk}}$  and

$$p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\mu}, \Sigma) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)^{z_{nk}}$$



# Example: A Variational Formulation of EM

Assume for a moment that we observe  $X$  and the binary latent variables  $Z$ . The likelihood is then:

$$p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^N p(\mathbf{z}_n \mid \boldsymbol{\pi}) p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

**Remember:**

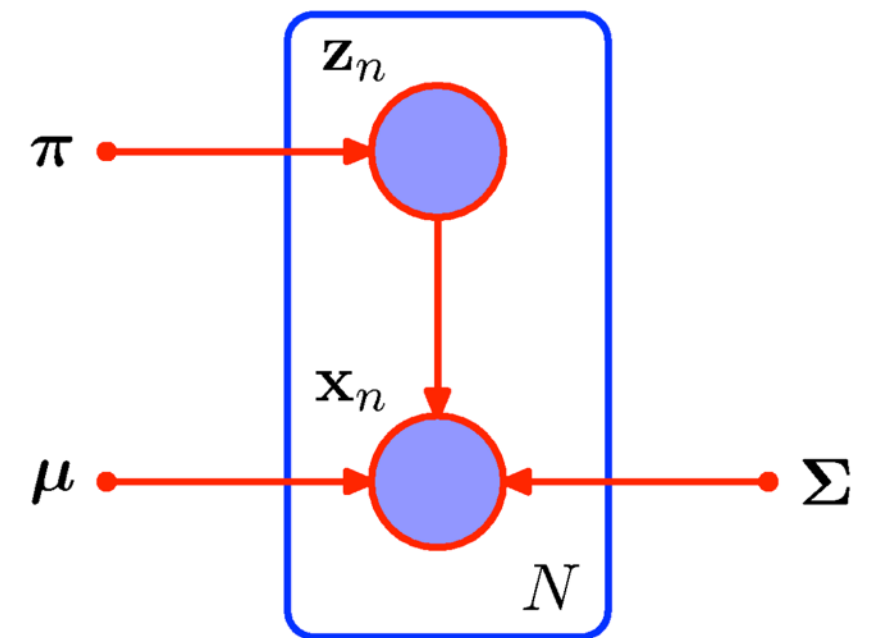
$$z_{nk} \in \{0, 1\}, \quad \sum_{k=1}^K z_{nk} = 1$$

where  $p(\mathbf{z}_n \mid \boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{z_{nk}}$  and

$$p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

which leads to the log-formulation:

$$\log p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$



# The Complete-Data Log-Likelihood

$$\log p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

- This is called the **complete-data log-likelihood**
- Advantage: solving for the parameters  $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$  is much simpler, as the log is inside the sum!
- We could switch the sums and then for every mixture component  $k$  only look at the points that are associated with that component.
- This leads to simple closed-form solutions for the parameters
- However: the latent variables  $Z$  are not observed!

