Sequential Monte Carlo and Particle Filtering

Frank Wood

Gatsby, November 2007

Importance Sampling

- Recall:
  - Let's say that we want to compute some expectation (integral)

$$E_p[f] = \int p(x) f(x) dx$$

and we remember from Monte Carlo integration theory that with samples from $p()$ we can approximate this integral thusly

$$E_p[f] \approx \frac{1}{L} \sum_{\ell=1}^{L} f(x_\ell)$$

What if $p()$ is hard to sample from?

- One solution: use importance sampling
  - Choose another easier to sample distribution $q()$ that is similar to $p()$ and do the following:

$$
\begin{aligned}
E_p[f] &= \int p(x) f(x) dx \\
&= \int \frac{p(x)}{q(x)} f(x) q(x) dx \\
&\approx \frac{1}{L} \sum_{\ell=1}^{L} \frac{p(x_\ell)}{q(x_\ell)} f(x_\ell)
\end{aligned}
$$

$$\boxed{x_\ell \sim q(\cdot)}$$

I.S. with distributions known only to a proportionality

- Importance sampling using distributions only known up to a proportionality is easy and common, algebra yields

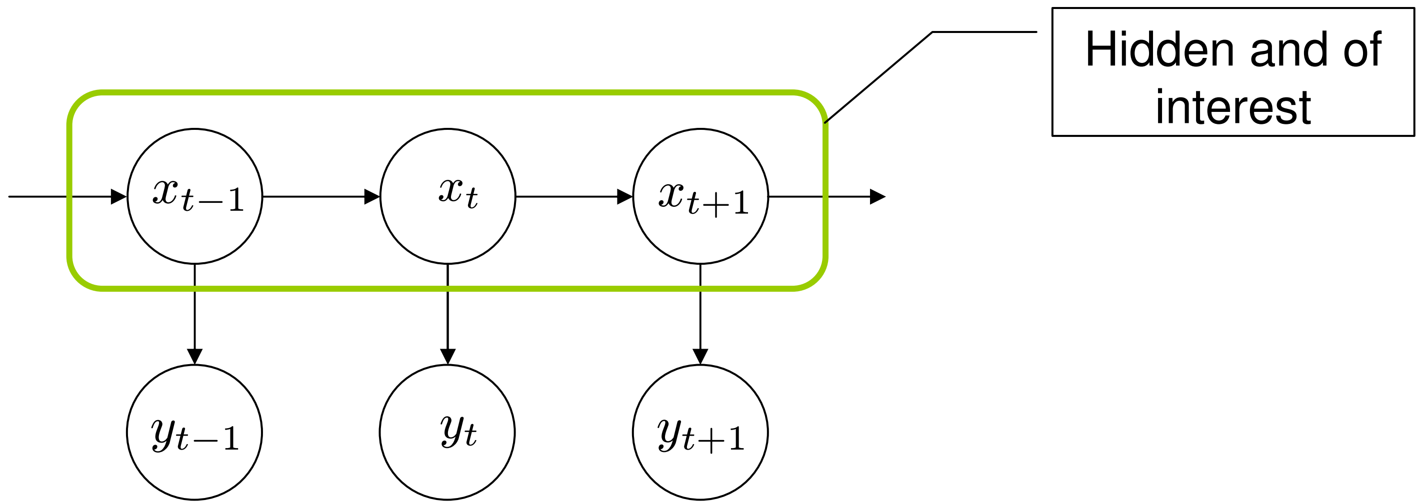
$$
\begin{aligned}
E_p[f] &\approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L}\frac{\tilde{p}(x_\ell)}{\tilde{q}(x_\ell)}f(x_\ell) \\
&\approx \sum_{\ell=1}^{L} w_\ell f(x_\ell)
\end{aligned}
$$

$$\tilde{r}_\ell = \frac{\tilde{p}(x_\ell)}{\tilde{q}(x_\ell)}$$

$$w_\ell = \frac{\tilde{r}_\ell}{\sum_{\ell=1}^{L}\tilde{r}_\ell}$$

A model requiring sampling techniques

- Non-linear non-Gaussian first order Markov model



$$p(\mathbf{x}_{1:i}, \mathbf{y}_{1:i}) = \prod_{i=1}^{N} p(y_i|x_i)p(x_i|x_{i-1})$$

Filtering distribution hard to obtain

- Often the filtering distribution is of interest

$$p(x_i|\mathbf{y}_{1:i-1}) \propto \int \ldots \int p(\mathbf{x}_{1:i}, \mathbf{y}_{1:i}) d_{x_1} \ldots d_{x_{i-1}}$$

- It may not be possible to compute these integrals analytically, be easy to sample from this directly, nor even to design a good proposal distribution for importance sampling.

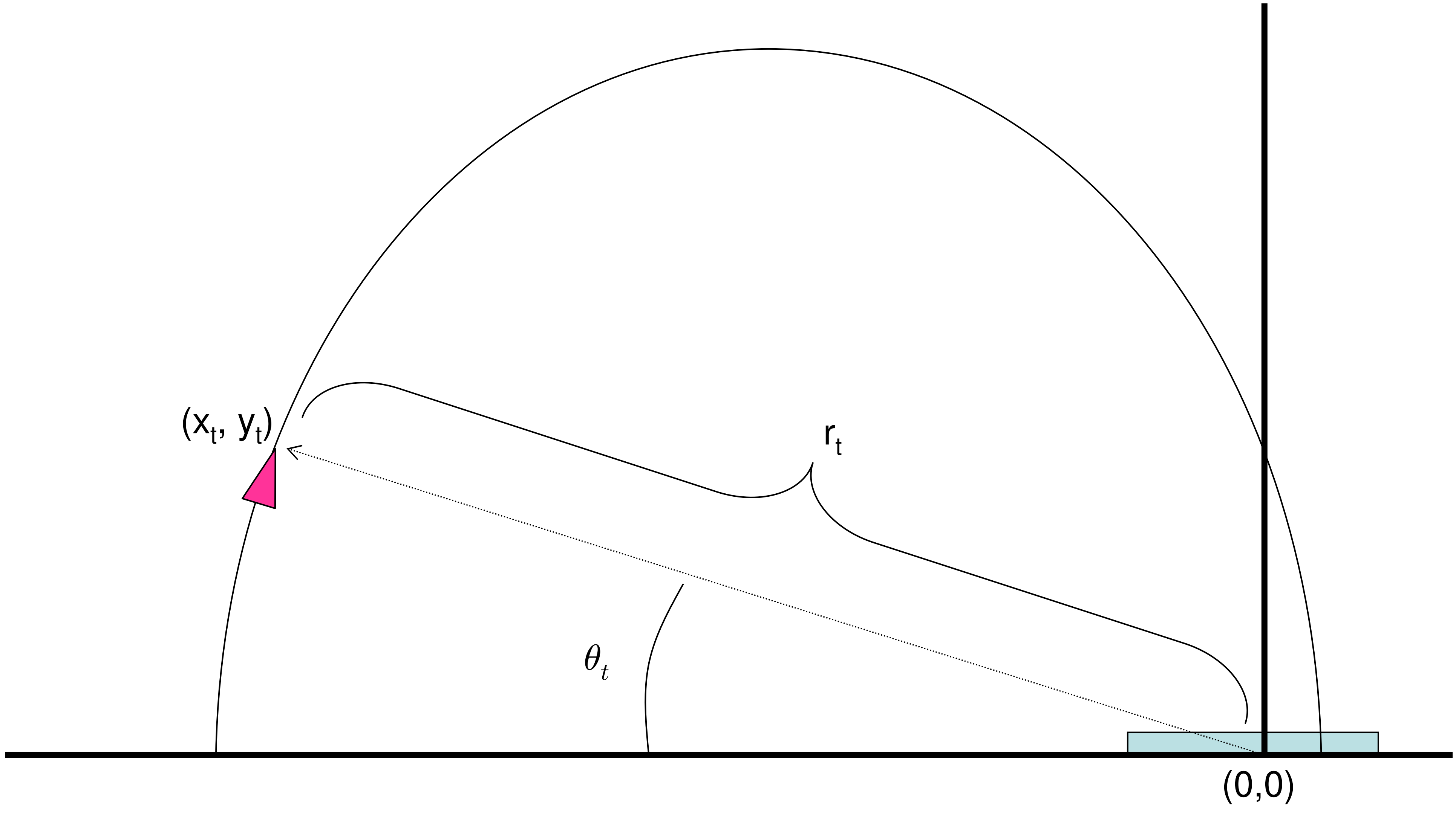
A solution: sequential Monte Carlo

- Sample from sequence of distributions that “converge” to the distribution of interest
- This is a very general technique that can be applied to a very large number of models and in a wide variety of settings.
- Today: particle filtering for a first order Markov model

Concrete example: target tracking

- A ballistic projectile has been launched in our direction.
- We have orders to intercept the projectile with a missile and thus need to infer the projectiles current position given noisy measurements.

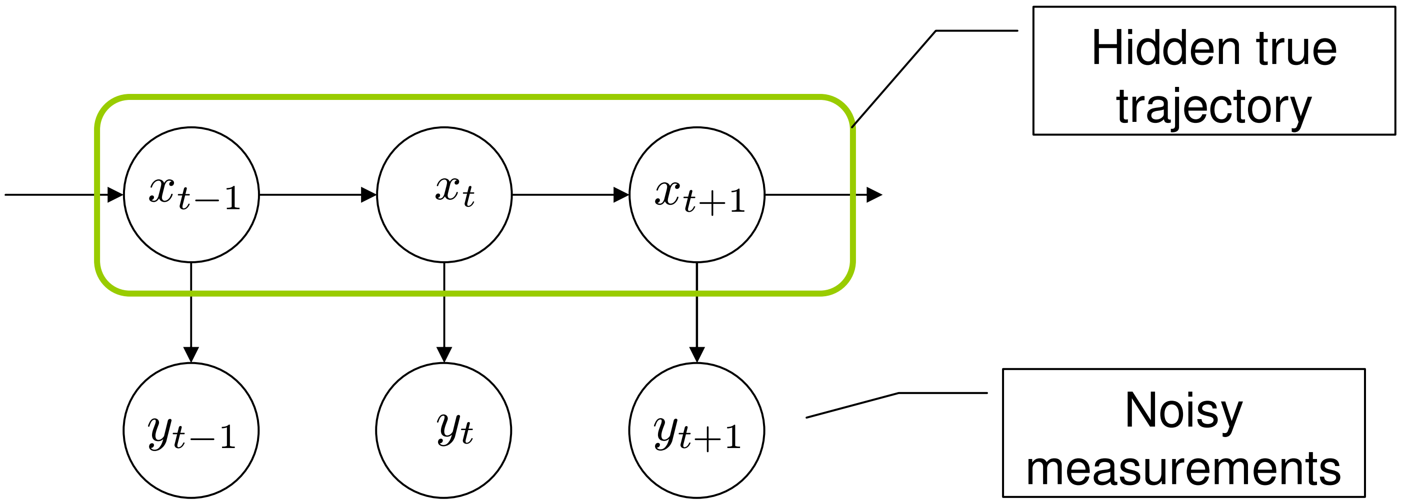
Problem Schematic



Probabilistic approach

- Treat true trajectory as a sequence of latent random variables
- Specify a model and do inference to recover the position of the projectile at time *t*

First order Markov model



$$
\begin{aligned}
p(x_{t+1}|x_t) &= a(\cdot;\theta_a, x_t) \\
p(y_t|x_t) &= h(\cdot;\theta_h, x_t)
\end{aligned}
$$

Posterior expectations

- We want filtering distribution samples

$$\boxed{p(x_i|\mathbf{y}_{1:i}) \propto p(y_i|x_i)\int p(x_i|x_{i-1})p(x_{i-1}|\mathbf{y}_{1:i-1})dx_{i-1}}$$

$$\propto p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1})$$

Write this down!

so that we can compute expectations

$$E_p[f] = \int f(x_i)p(x_i|\mathbf{y}_{1:i})dx_i$$

Importance sampling

$$\boxed{\tilde{p}(x_i)} \propto p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1}) \qquad \boxed{q(x_i)} = p(x_i|\mathbf{y}_{1:i-1})$$

Distribution from which we want samples

Proposal distribution

- By identity the posterior predictive distribution can be written as

$$q(x_i) = p(x_i|\mathbf{y}_{1:i-1}) = \int p(x_i|x_{i-1})p(x_{i-1}|\mathbf{y}_{1:i-1})dx_{i-1}$$

Basis of sequential recursion

- If we start with samples from

$$\{w_\ell^{i-1}, x_\ell^{i-1}\}_{\ell=1}^L \sim p(x_{i-1}|\mathbf{y}_{1:i-1})$$

then we can write the proposal distribution as a finite mixture model

$$q(x_i) = p(x_i|\mathbf{y}_{1:i-1}) \approx \sum_{\ell=1}^L w_\ell^{i-1} p(x_i|x_\ell^{i-1})$$

and draw samples accordingly

$$\{\hat{w}_m^i, \hat{x}_m^i\}_{m=1}^M \sim q(\cdot)$$

Samples from the proposal distribution

- We now have samples from the proposal

$$\{\hat{w}_m^i, \hat{x}_m^i\} \sim q(x_i)$$

- And if we recall

$$\boxed{\tilde{p}(x_i)} = p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1}) \qquad \boxed{q(x_i)} = p(x_i|\mathbf{y}_{1:i-1})$$

Distribution from which we want samples

Proposal distribution

Updating the weights completes importance sampling

$$\hat{r}^i_m = \frac{\tilde{p}(\hat{x}^i_m)}{q(\hat{x}^i_m)} = p(y_i|\hat{x}^i_m)\hat{w}^i_m$$

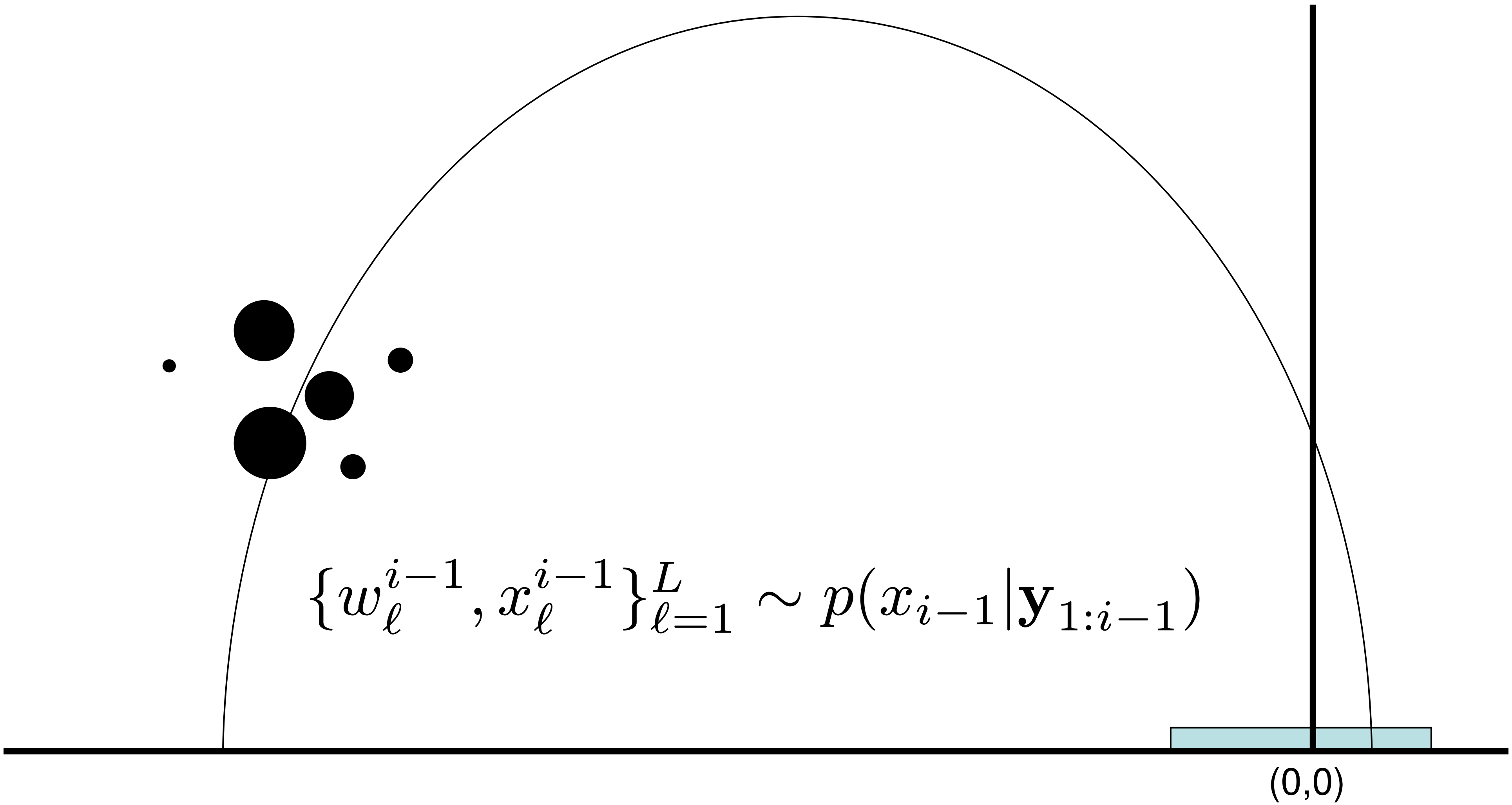
- We are left with M weighted samples from the posterior up to observation i

$$\boxed{w^i_m} = \frac{\hat{r}^i_m}{\sum_{m=1}^{M} \hat{r}^i_m} \qquad \boxed{\{w^i_m, x^i_m\}_{m=1}^{M}} \sim p(x_i|\mathbf{y}_{1:i})$$

Intuition

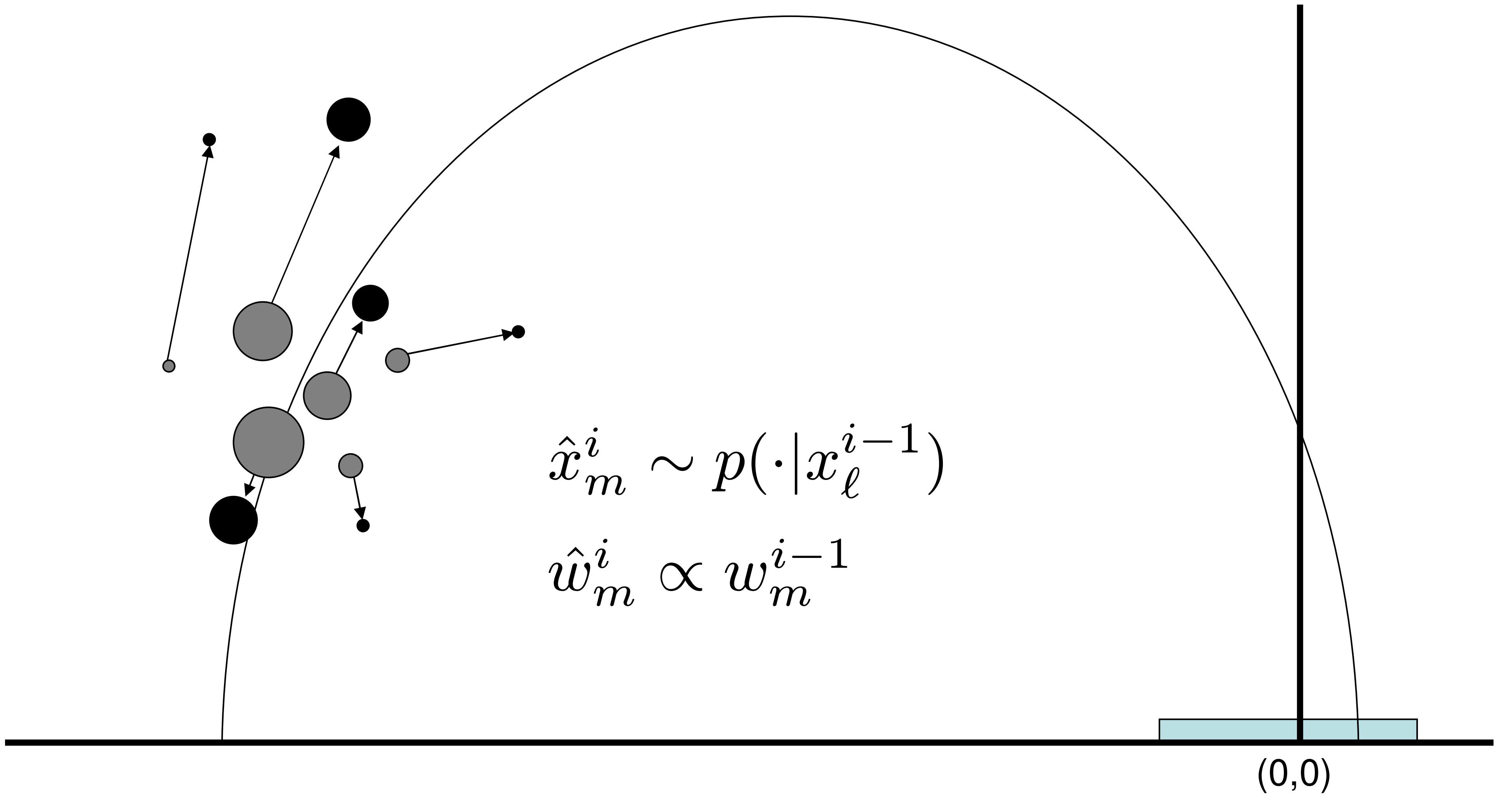
- Particle filter name comes from physical interpretation of samples

Start with samples representing the hidden state



$$\{w_\ell^{i-1}, x_\ell^{i-1}\}_{\ell=1}^{L} \sim p(x_{i-1}|\mathbf{y}_{1:i-1})$$

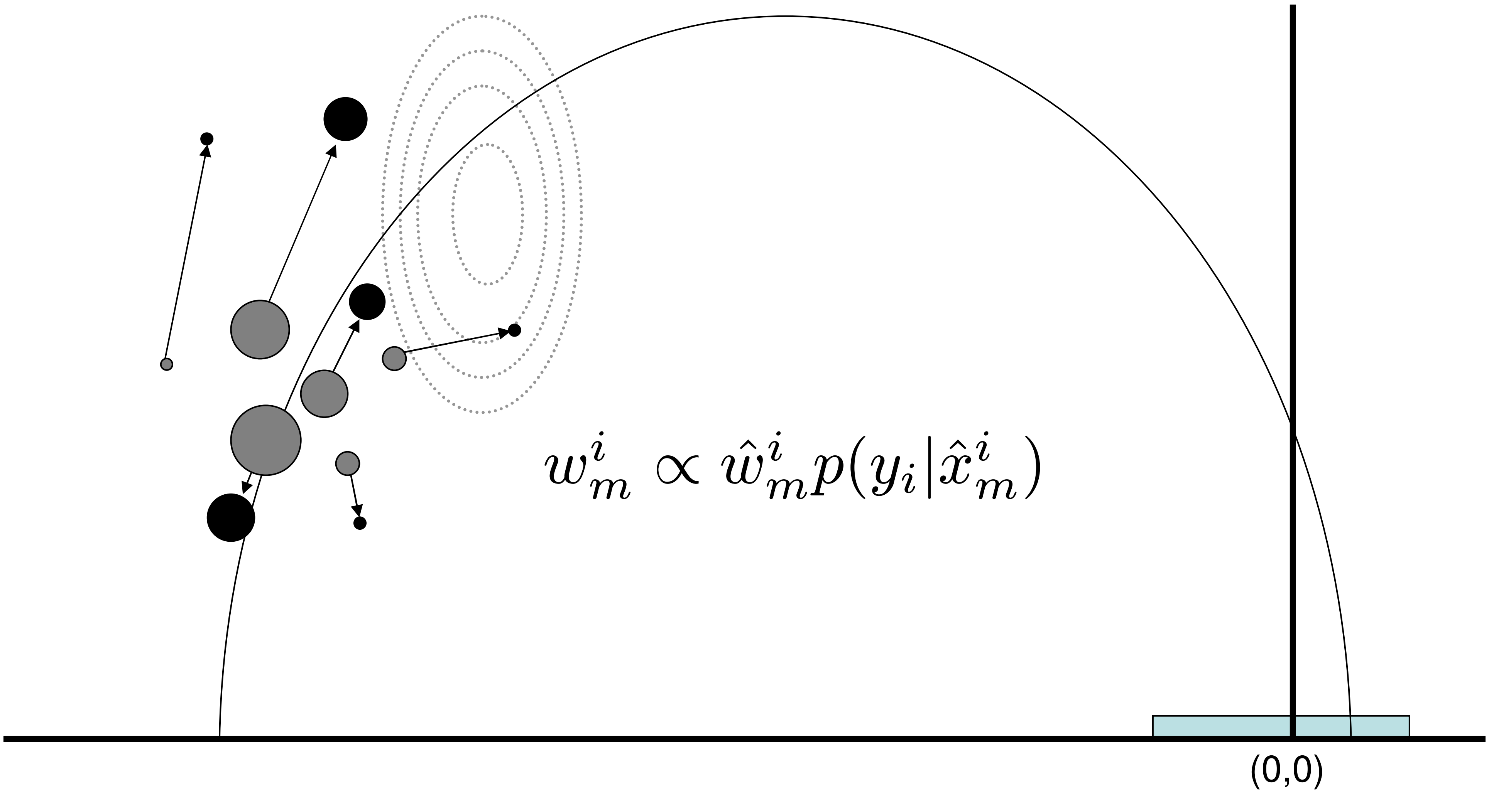
Evolve them according to the state model



$$\hat{x}_m^i \sim p(\cdot|x_\ell^{i-1})$$

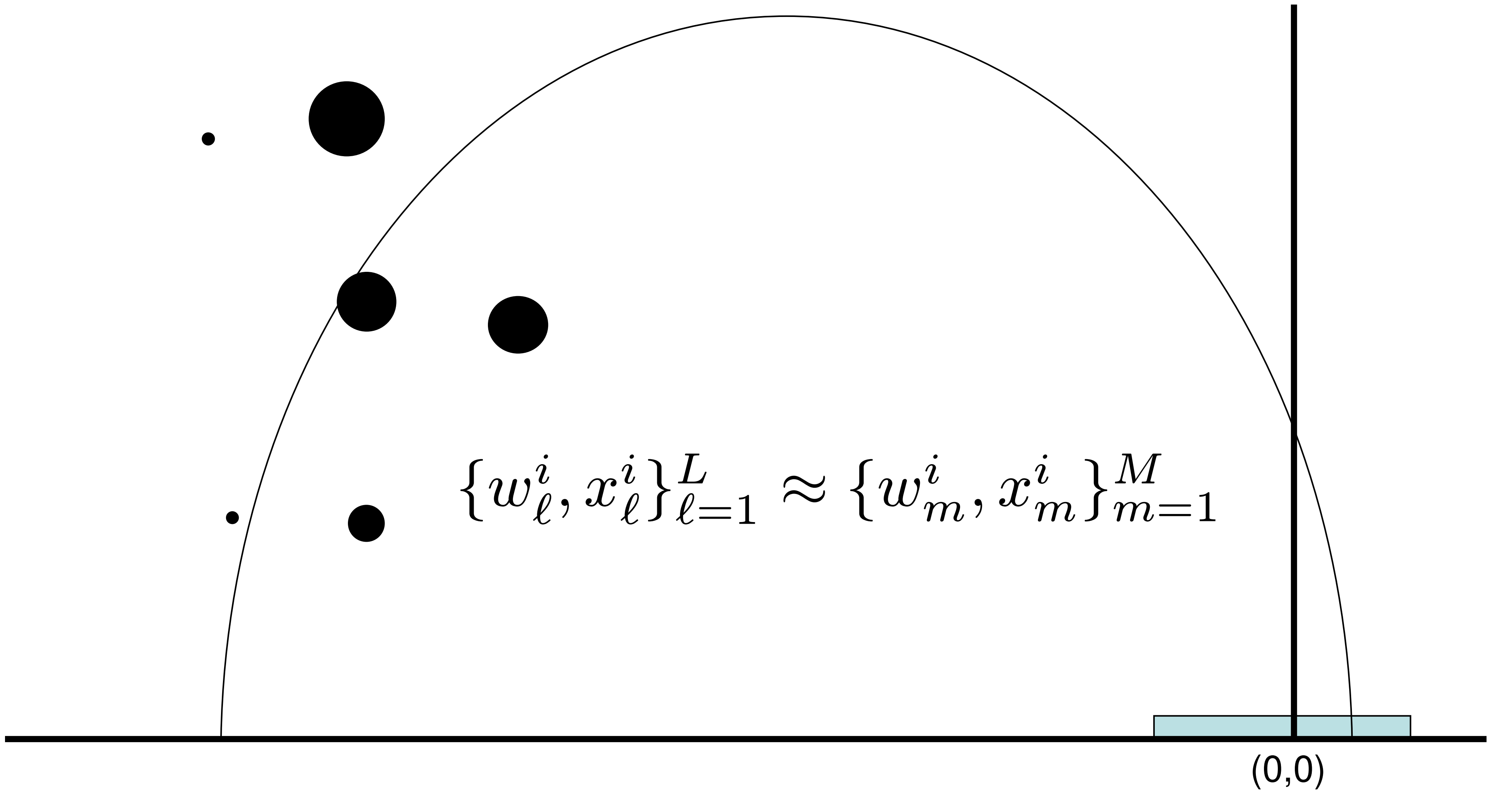
$$\hat{w}_m^i \propto w_m^{i-1}$$

Re-weight them by the likelihood



$$w_m^i \propto \hat{w}_m^i p(y_i|\hat{x}_m^i)$$

Results in samples one step forward



SIS Particle Filter

- The Sequential Importance Sampler (SIS) particle filter multiplicatively updates weights at every iteration and thus often most weights get very small
- Computationally this makes little sense as eventually low-weighted particles do not contribute to any expectations.
- A measure of particle degeneracy is “effective sample size”

$$\hat{N}_{eff} = \frac{1}{\sum_{\ell=1}^{L} (w_\ell^i)^2}$$

- When this quantity is small, particle degeneracy is severe.

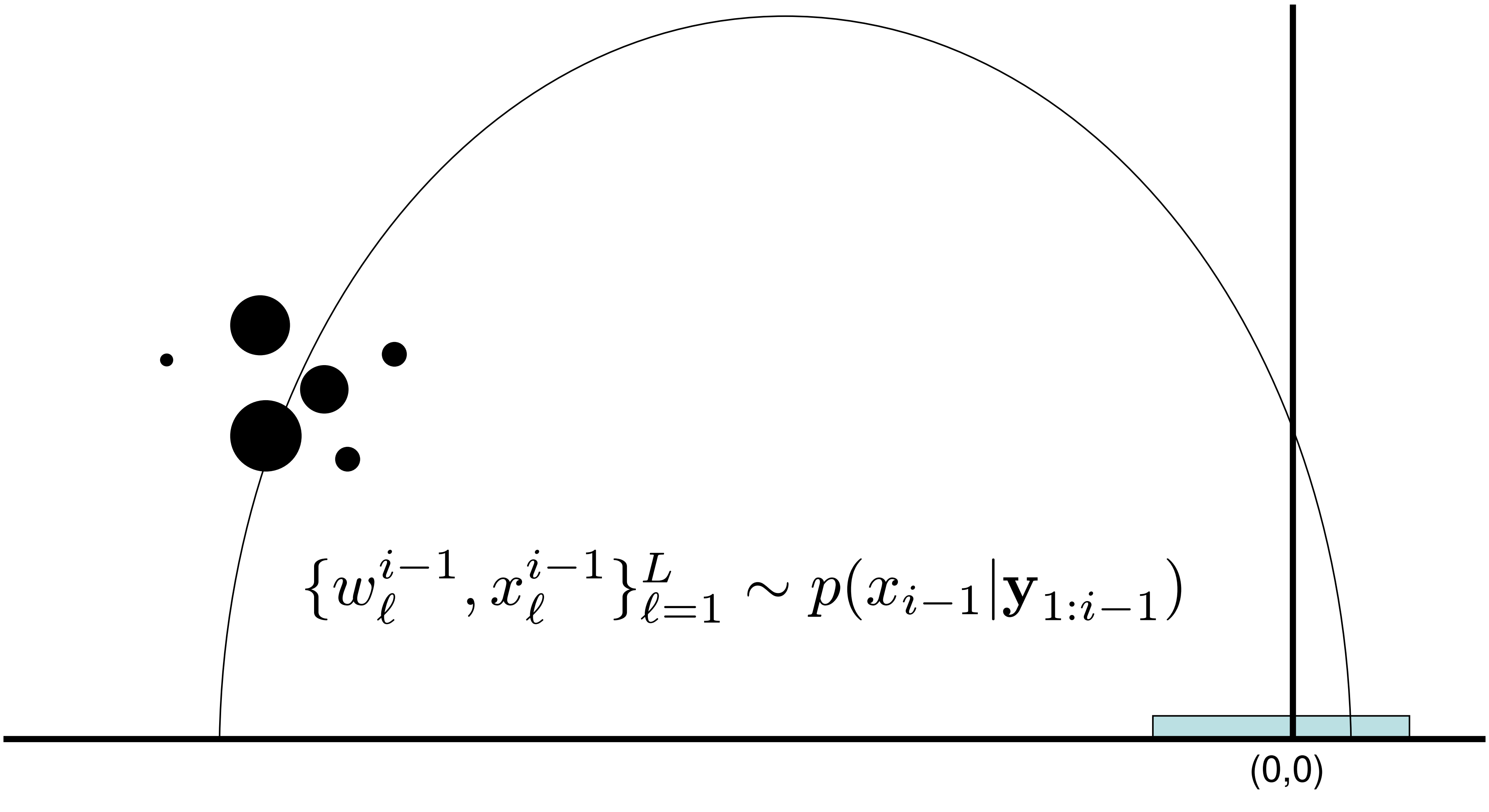
Solutions

- Sequential Importance *Re*-sampling (SIR) particle filter avoids many of the problems associated with SIS pf'ing by re-sampling the posterior particle set to increase the effective sample size.
- Choosing the best possible importance density is also important because it minimizes the variance of the weights which maximizes $\hat{N}_{eff}$

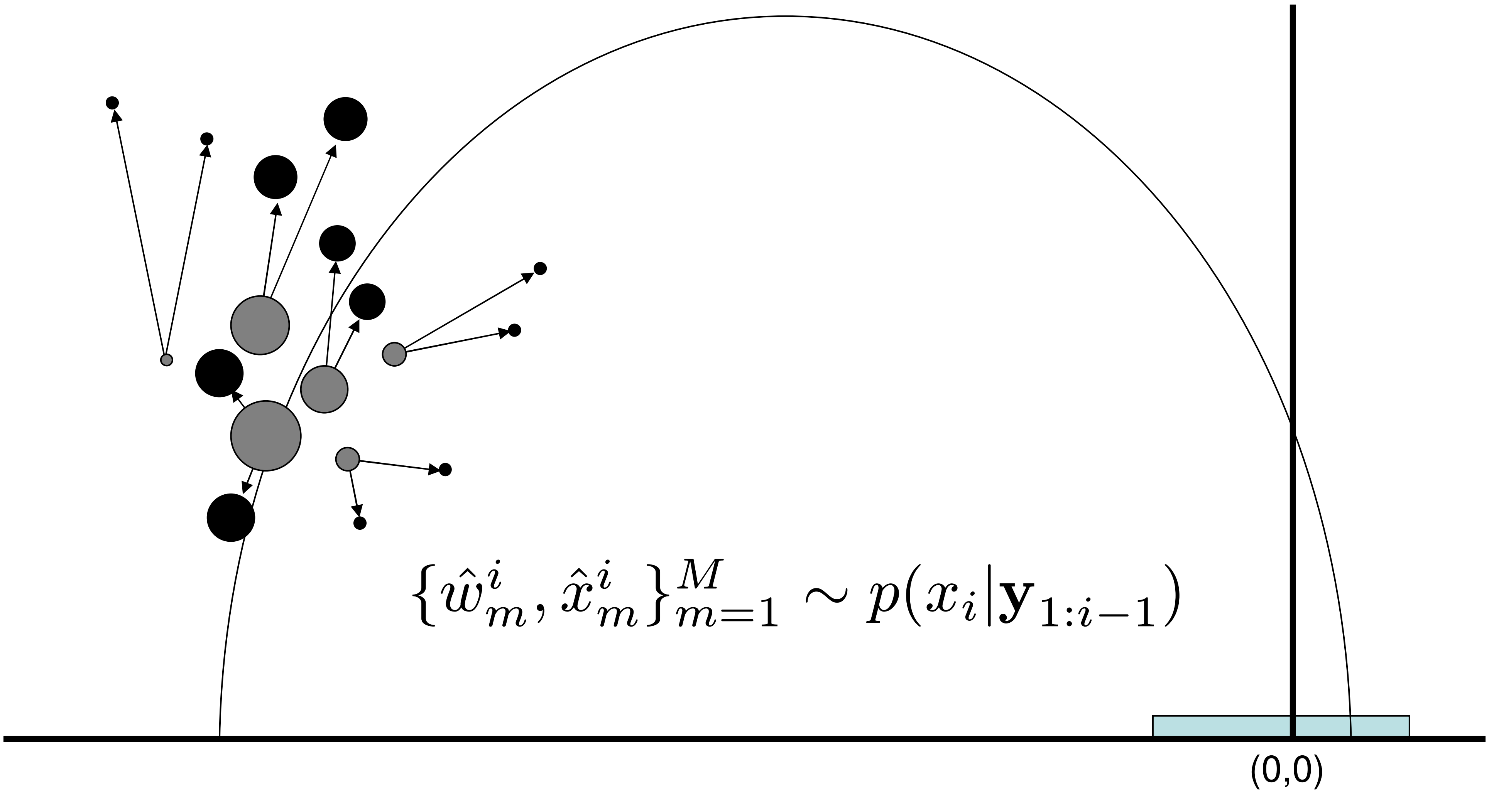
Other tricks to improve pf'ing

- Integrate out everything you can
- Replicate each particle some number of times
- In discrete systems, enumerate instead of sample
- Use fancy re-sampling schemes like stratified sampling, etc.

Initial particles

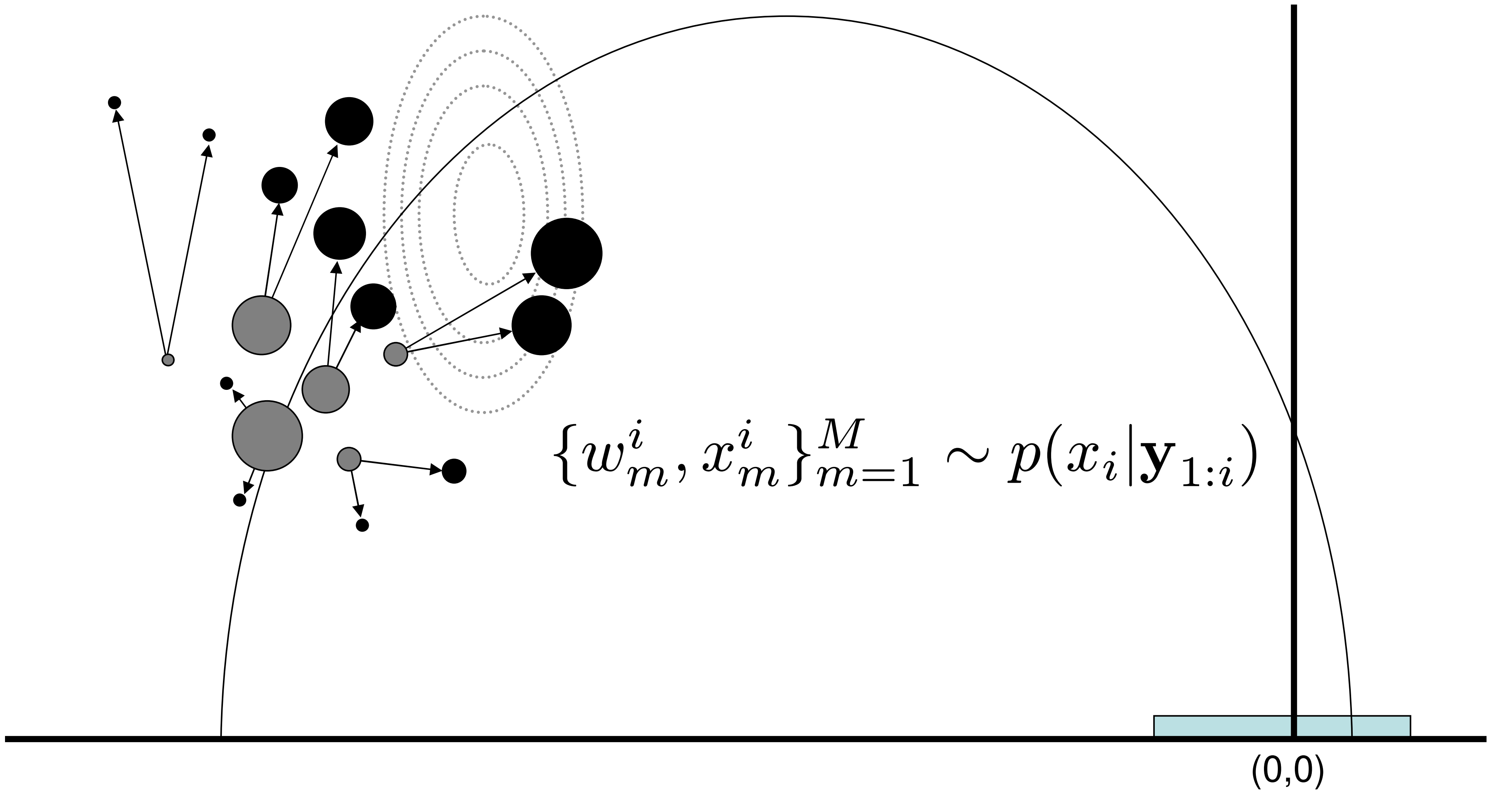


Particle evolution step



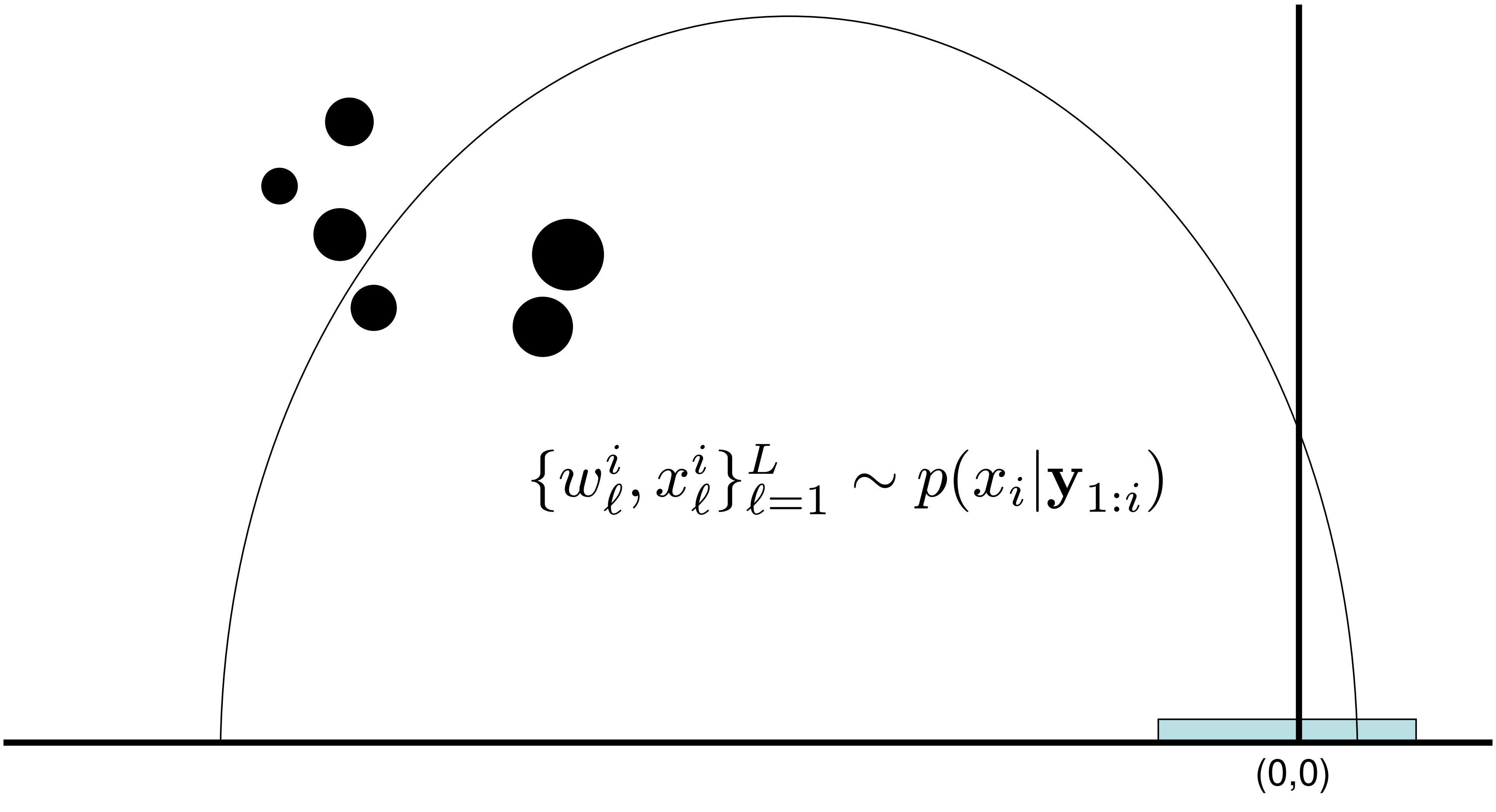
$$\{\hat{w}_m^i, \hat{x}_m^i\}_{m=1}^M \sim p(x_i|\mathbf{y}_{1:i-1})$$

Weighting step



$$\{w_m^i, x_m^i\}_{m=1}^M \sim p(x_i|\mathbf{y}_{1:i})$$

Resampling step



Wrap-up: Pros vs. Cons

- Pros:
  - Sometimes it is easier to build a “good” particle filter sampler than an MCMC sampler
  - No need to specify a convergence measure
- Cons:
  - Really *filtering* not smoothing
- Issues
  - Computational trade-off with MCMC

Thank You

Tricks and Variants

- Reduce the dimensionality of the integrand through analytic integration
  - Rao-Blackwellization
- Reduce the variance of the Monte Carlo estimator through
  - Maintaining a weighted particle set
  - Stratified sampling
  - Over-sampling
  - Optimal re-sampling

Particle filtering

- Consists of two basic elements:
  - Monte Carlo integration

$$p(x) \approx \sum_{\ell=1}^{L} w_\ell \delta_{x_\ell}$$

$$\lim_{L \to \infty} \sum_{\ell=1}^{L} w_\ell f(x_\ell) = \int f(x) p(x) dx$$

  - Importance sampling

PF'ing: Forming the posterior predictive

$$\{w_\ell^{i-1}, x_\ell^{i-1}\}_{\ell=1}^{L} \sim \boxed{p(x_{i-1}|y_{1:i-1})}$$

Posterior up to observation $i-1$

$$\begin{aligned} p(x_i|y_{1:i-1}) &= \int p(x_i|x_{i-1})p(x_{i-1}|y_{1:i-1})dx_{i-1} \\ &\approx \boxed{\sum_{\ell=1}^{L} w_\ell^{i-1} p(x_i|x_\ell^{i-1})} \end{aligned}$$

The proposal distribution for importance sampling of the posterior up to observation i is this approximate posterior predictive distribution

Sampling the posterior predictive

- Generating samples from the posterior predictive distribution is the first place where we can introduce variance reduction techniques

$$\{\hat{w}_m^i, \hat{x}_m^i\}_{m=1}^M \sim p(x_i|\mathbf{y}_{1:i-1}), \qquad p(x_i|\mathbf{y}_{1:i-1}) \approx \sum_{\ell=1}^{L} w_\ell^{i-1} p(x_i|x_\ell^{i-1})$$

- For instance sample from each mixture component several twice such that M, the number of samples drawn, is two times L, the number of densities in the mixture model, and assign weights

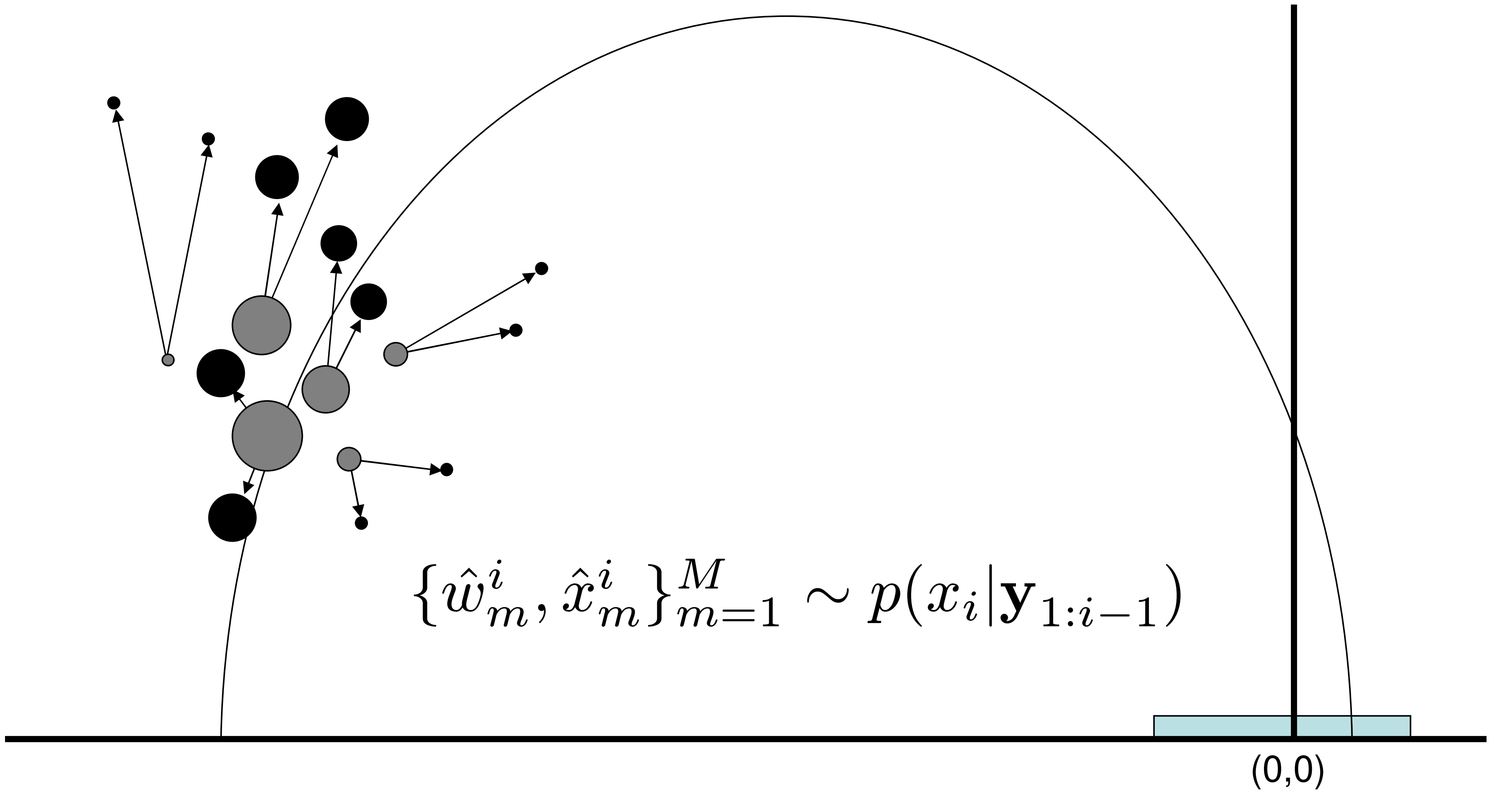
$$\hat{w}_m^i = \frac{w_\ell^{i-1}}{2}$$

Not the best

- Most efficient Monte Carlo estimator of a function $\mathit{\Gamma}$(x)
  - From survey sampling theory: Neyman allocation
  - Number drawn from each mixture density is proportional to the weight of the mixture density times the std. dev. of the function $\mathit{\Gamma}$ over the mixture density
- Take home: smarter sampling possible

[Cochran 1963]

Over-sampling from the posterior predictive distribution



Importance sampling the posterior

- Recall that we want samples from

$$\begin{aligned} p(x_i|\mathbf{y}_{1:i}) &\propto p(y_i|x_i)\int p(x_i|x_{i-1})p(x_{i-1}|\mathbf{y}_{1:i-1})dx_{i-1} \\ &\propto p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1}) \end{aligned}$$

- and make the following importance sampling identifications

Proposal distribution

$$\boxed{\tilde{p}(x_i)} = p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1})$$

Distribution from which we want to sample

$$\begin{aligned} \boxed{q(x_i)} &= p(x_i|\mathbf{y}_{1:i-1}) \\ &\approx \sum_{\ell=1}^{L} w_\ell^{i-1} p(x_i|x_\ell^{i-1}) \end{aligned}$$

Sequential importance sampling

- Weighted posterior samples arise as

$$\boxed{\{\hat{w}_m^i, \hat{x}_m^i\}} \sim q(\cdot)$$

- Normalization of the weights takes place as before

$$\boxed{w_m^i} = \frac{\hat{r}_\ell^i}{\sum_{\ell=1}^{L} \hat{r}_\ell^i} \qquad \hat{r}_m^i = \frac{\tilde{p}(\hat{x}_m^i)}{q(\hat{x}_m^i)} = p(y_i|\hat{x}_m^i)\hat{w}_m^i$$

- We are left with M weighted samples from the posterior up to observation i

$$\boxed{\{w_m^i, x_m^i\}_{m=1}^M} \sim p(x_i|\mathbf{y}_{1:i})$$

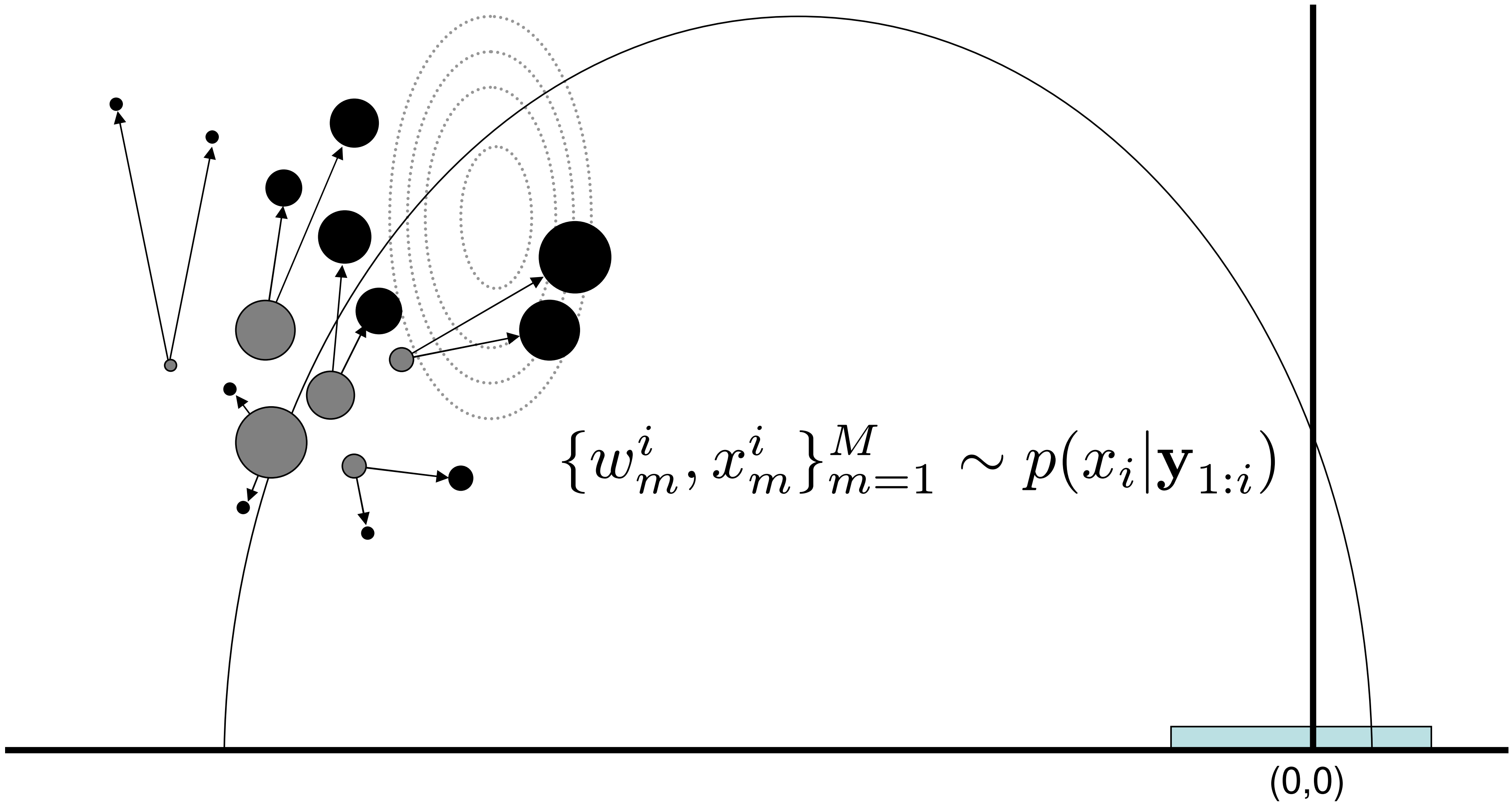
An alternative view

$$\{\hat{w}_m^i, \hat{x}_m^i\} \quad \sim \quad \sum_{\ell=1}^{L} w_\ell^{i-1} p(x_i | x_\ell^{i-1})$$

$$p(x_i|\mathbf{y}_{1:i-1}) \approx \textstyle\sum_{m=1}^{M} \hat{w}_m^i \delta_{\hat{x}_m^i}$$

$$p(x_i|\mathbf{y}_{1:i}) \approx \textstyle\sum_{m=1}^{M} p(y_i|\hat{x}_m^i)\hat{w}_m^i \delta_{\hat{x}_m^i}$$

Importance sampling from the posterior distribution



Sequential importance re-sampling

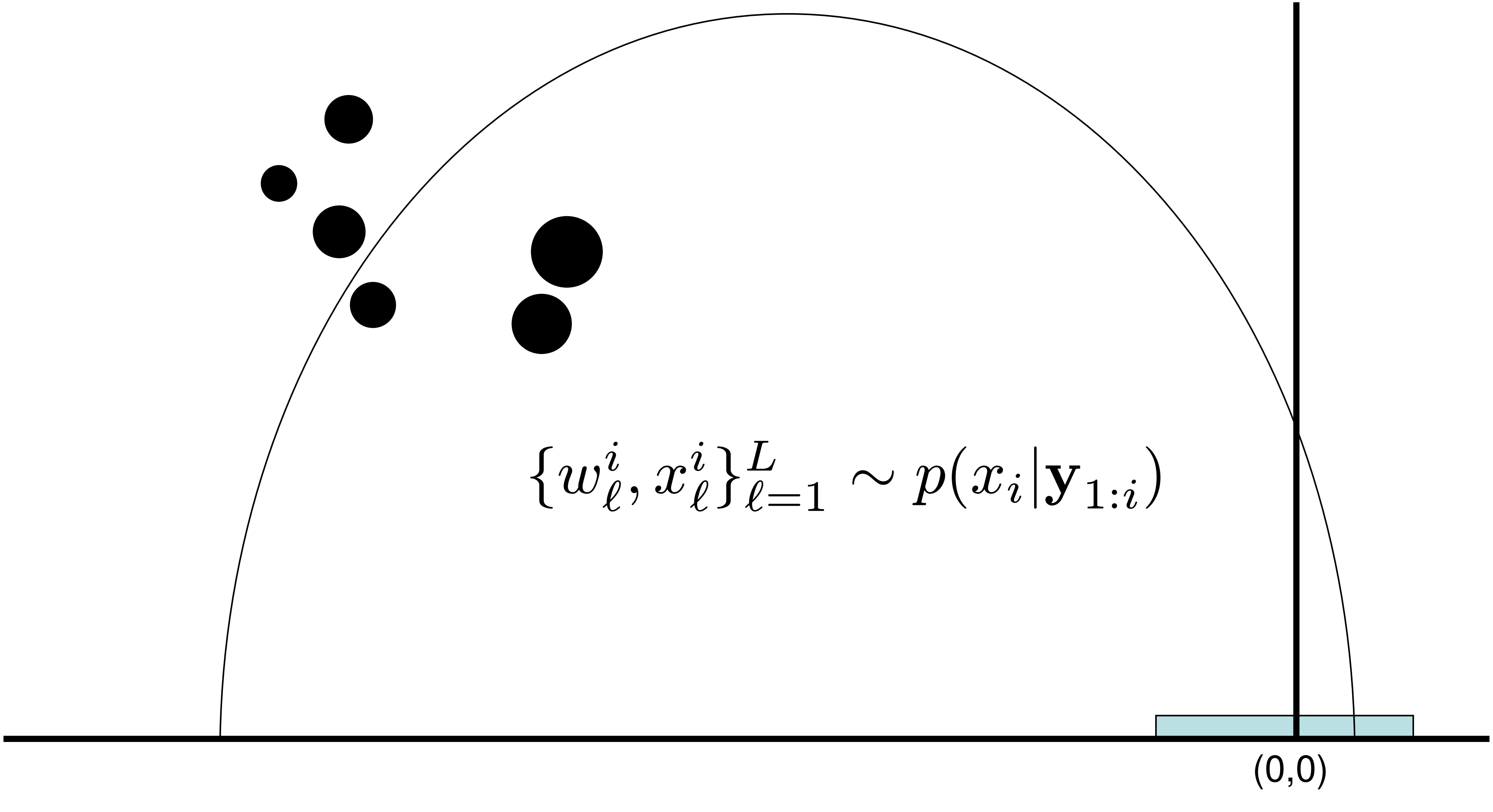
- Down-sample L particles and weights from the collection of M particles and weights

$$\{w_\ell^i, x_\ell^i\}_{\ell=1}^L \approx \{w_m^i, x_m^i\}_{m=1}^M$$

this can be done via multinomial sampling or in a way that provably minimizes estimator variance

[Fearnhead 04]

Down-sampling the particle set



$$\{w_\ell^i, x_\ell^i\}_{\ell=1}^L \sim p(x_i|\mathbf{y}_{1:i})$$

Recap

- Starting with (weighted) samples from the posterior up to observation i-1
- Monte Carlo integration was used to form a mixture model representation of the posterior predictive distribution
- The posterior predictive distribution was used as a proposal distribution for importance sampling of the posterior up to observation i
- $M > L$ samples were drawn and re-weighted according to the likelihood (the importance weight), then the collection of particles was down-sampled to L weighted samples

LSSM Not alone

- Various other models are amenable to sequential inference, Dirichlet process mixture modelling is another example, dynamic Bayes' nets are another

Rao-Blackwellization

- In models where parameters can be analytically marginalized out, or the particle state space can otherwise be collapsed, the efficiency of the particle filter can be improved by doing so

Stratified Sampling

- Sampling from a mixture density using the algorithm on the right produces a more efficient Monte Carlo estimator

$$\{w_n, x_n\}_{n=1}^{K} \quad \sim \quad \sum_k \pi_k f_k(\cdot)$$

• for $\mathrm{n=1:K}$ – choose k according to π_k – sample x_n from f_k – set w_n equal to $1/\mathrm{N}$	• for $\mathrm{n=1:K}$ – choose f_n – sample x_n from f_n – set w_n equal to π_n

Intuition: weighted particle set

- What is the difference between these two discrete distributions over the set {a,b,c}?
  - (a), (a), (b), (b), (c)
  - (.4, a), (.4, b), (.2, c)

- Weighted particle representations are equally or more efficient for the same number of particles

Optimal particle re-weighting

- Next step: when down-sampling, pass all particles above a threshold c through without modifying their weights where c is the unique solution to

$$\sum_{m=1}^{M} \min\{cw_m^i, 1\} = L$$

- Resample all particles with weights below c using stratified sampling and give the selected particles weight $1/c$

Fearnhead 2004

Result is provably optimal

- In the down-sampling step

$$\{w_\ell^i, x_\ell^i\}_{\ell=1}^L \approx \{w_m^i, x_m^i\}_{m=1}^M$$

- Imagine instead a "sparse" set of weights of which some are zero

$$\{\tilde{w}_\ell^i, x_\ell^i\}_{\ell=1}^M \approx \{w_m^i, x_m^i\}_{m=1}^M$$

- Then this down-sampling algorithm is optimal w.r.t.

$$\sum_{m=1}^M E_w[(\tilde{w}_m^i - w_m^i)^2]$$

Fearnhead 2004

Problem Details

- Time and position are given in seconds and meters respectively
- Initial launch velocity and position are both unknown
- The maximum muzzle velocity of the projectile is 1000m/s
- The measurement error in the Cartesian coordinate system is N(0,10000) and N(0,500) for x and y position respectively
- The measurement error in the polar coordinate system is N(0,.001) for θ and Gamma(1,100) for r
- The kill radius of the projectile is 100m

Data and Support Code

http://www.gatsby.ucl.ac.uk/~fwood/pf_tutorial/

Laws of Motion

- In case you've forgotten:

$$\mathbf{r} = (v_0 \cos(\alpha))t\mathbf{i} + ((v_0 \sin(\alpha)t - \frac{1}{2}gt^2)\mathbf{j}$$

- where v_o is the initial speed and α is the initial angle

Good luck!

Monte Carlo Integration

- Compute integrals for which analytical solutions are unknown

$$\int f(x)p(x)dx$$

$$p(x) \approx \sum_{\ell=1}^{L} w_\ell \delta_{x_\ell}$$

Monte Carlo Integration

- Integral approximated as the weighted sum of function evaluations at L points

$$\int f(x)p(x)dx \approx \int f(x) \sum_{\ell=1}^{L} w_\ell \delta_{x_\ell} dx$$

$$p(x) \approx \sum_{\ell=1}^{L} w_\ell \delta_{x_\ell} \quad = \quad \sum_{\ell=1}^{L} w_\ell f(x_\ell)$$

Sampling

- To use MC integration one must be able to sample from p(x)

$$
\begin{aligned}
\{w_\ell, x_\ell\}_{\ell=1}^{L} &\sim p(\cdot) \\
\lim_{L\to\infty} \sum_{\ell=1}^{L} w_\ell \delta_{x_\ell} &\to p(\cdot)
\end{aligned}
$$

Theory (Convergence)

- Quality of the approximation independent of the dimensionality of the integrand
- Convergence of integral result to the "truth" is $O(1/n^{1/2})$ from L.L.N.'s.
- Error is independent of dimensionality of x

Bayesian Modelling

- Formal framework for the expression of modelling assumptions
- Two common definitions:
  - using Bayes' rule
  - marginalizing over models

Posterior

Prior

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

Likelihood

Evidence

Posterior Estimation

- Often the distribution over latent random variables (parameters) is of interest
- Sometimes this is easy (conjugacy)
- Usually it is hard because computing the evidence is intractable

Conjugacy Example

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ x|\theta &\sim \text{Binomial}(N, \theta) \end{aligned}$$

$$\begin{aligned} p(\theta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ p(x|\theta) &= \binom{N}{x} \theta^{x}(1-\theta)^{N-x} \end{aligned}$$

x successes in N trials, θ probability of success

Conjugacy Continued

$$
\begin{aligned}
p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \\
&= \frac{1}{Z(x)}p(x|\theta)p(\theta) \\
&= \frac{1}{Z(x)}\theta^{\alpha-1+x}(1-\theta)^{\beta-1+N-x}
\end{aligned}
$$

$$
\begin{aligned}
\theta|x &\sim \mathrm{Beta}(\alpha+x, \beta+N-x) \\
Z(x) &= \left(\frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+x)\Gamma(\beta+N-x)}\right)^{-1}
\end{aligned}
$$

Non-Conjugate Models

- Easy to come up with examples

$$\begin{aligned} \sigma^2 &\sim N(0, \alpha) \\ x|\sigma^2 &\sim N(0, \sigma^2) \end{aligned}$$

Posterior Inference

- Posterior averages are frequently important in problem domains
  - posterior predictive distribution

$$p(x_{i+1}|\mathbf{x}_{1:i}) = \int p(x_{i+1}|\theta, \mathbf{x}_{1:i})p(\theta|\mathbf{x}_{1:i})d\theta$$

  - evidence (as seen) for model comparison, etc.

Relating the Posterior to the Posterior Predictive

$$p(x_i|\mathbf{y}_{1:i}) \quad = \quad \int p(x_i, \mathbf{x}_{1:i-1}|\mathbf{y}_{1:i})d\mathbf{x}_{1:i-1}$$

$$\propto \quad p(x_i|\mathbf{y}_{1:i-1})$$

Relating the Posterior to the Posterior Predictive

$$
\begin{aligned}
p(x_i|\mathbf{y}_{1:i}) &= \int p(x_i, \mathbf{x}_{1:i-1}|\mathbf{y}_{1:i}) d\mathbf{x}_{1:i-1} \\
&\propto \int p(y_i|x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1}) p(x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1}) d\mathbf{x}_{1:i-1}
\end{aligned}
$$

$$
\propto \quad p(x_i|\mathbf{y}_{1:i-1})
$$

Relating the Posterior to the Posterior Predictive

$$
\begin{aligned}
p(x_i|\mathbf{y}_{1:i}) &= \int p(x_i, \mathbf{x}_{1:i-1}|\mathbf{y}_{1:i})d\mathbf{x}_{1:i-1} \\
&\propto \int p(y_i|x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1})p(x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1})d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i) \int p(x_i|\mathbf{x}_{1:i-1}, \mathbf{y}_{1:i})p(\mathbf{x}_{1:i-1}|\mathbf{y}_{1:i})d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1})
\end{aligned}
$$

Relating the Posterior to the Posterior Predictive

$$
\begin{aligned}
p(x_i|\mathbf{y}_{1:i}) &= \int p(x_i, \mathbf{x}_{1:i-1}|\mathbf{y}_{1:i})d\mathbf{x}_{1:i-1} \\
&\propto \int p(y_i|x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1})p(x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1})d\mathbf{x}_{1:i-} \\
&\propto p(y_i|x_i) \int p(x_i|\mathbf{x}_{1:i-1}, \mathbf{y}_{1:i})p(\mathbf{x}_{1:i-1}|\mathbf{y}_{1:i})d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i) \int p(x_i|x_{i-1})p(\mathbf{x}_{1:i-1}|\mathbf{y}_{1:i-1})d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1})
\end{aligned}
$$

Importance sampling

$$E_x[f(x)] = \int p(x) f(x) dx$$

$$= \int \frac{\boxed{p(x)}}{q(x)} f(x) \boxed{q(x)} dx$$

$$\approx \frac{1}{L} \sum_{\ell=1}^{L} \frac{p(x_\ell)}{q(x_\ell)} f(x_\ell)$$

Original distribution: hard to sample from, easy to evaluate

Proposal distribution: easy to sample from

$$x_\ell \sim q(\cdot)$$

Importance weights

$$\boxed{r_\ell} = \frac{p(x_\ell)}{q(x_\ell)}$$

Importance sampling un-normalized distributions

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

Un-normalized distribution to sample from, still hard to sample from and easy to evaluate

$$q(x) = \frac{\boxed{\tilde{q}(x)}}{Z_q} \qquad x_\ell \sim \boxed{\tilde{q}(\cdot)}$$

Un-normalized proposal distribution: still easy to sample from

$$E_x\left[f(x)\right] \approx \frac{1}{L}\sum_{\ell=1}^{L} \frac{p(x_\ell)}{q(x_\ell)} f(x_\ell)$$

$$\approx \boxed{\frac{Z_q}{Z_p}} \frac{1}{L}\sum_{\ell=1}^{L} \frac{\tilde{p}(x_\ell)}{\tilde{q}(x_\ell)} f(x_\ell)$$

New term: ratio of normalizing constants

Normalizing the importance weights

Un-normalized importance weights

Takes a little algebra

$$\boxed{\tilde{r}_\ell} = \frac{\tilde{p}(x_\ell)}{\tilde{q}(x_\ell)} \qquad \boxed{\frac{Z_q}{Z_p} \approx \frac{L}{\sum_{\ell=1}^{L} \tilde{r}_\ell}} \qquad \boxed{w_\ell} = \frac{\tilde{r}_\ell}{\sum_{\ell=1}^{L} \tilde{r}_\ell}$$

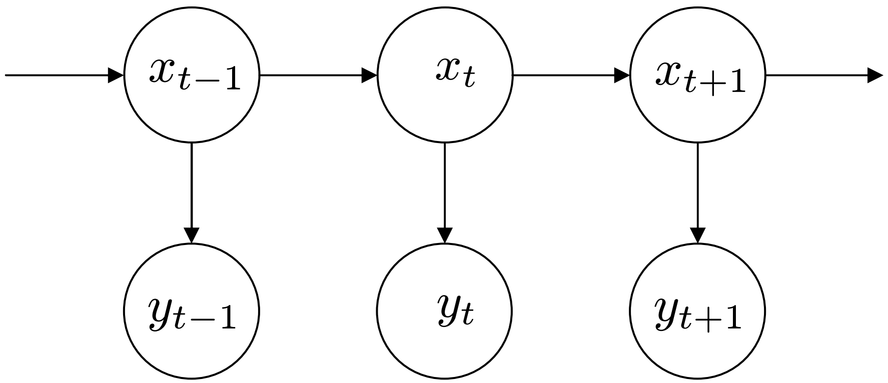
Normalized importance weights

$$\begin{aligned} E_x[f(x)] &\approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{\ell=1}^{L} \frac{\tilde{p}(x_\ell)}{\tilde{q}(x_\ell)} f(x_\ell) \\ &\approx \sum_{\ell=1}^{L} w_\ell f(x_\ell) \end{aligned}$$

Linear State Space Model (LSSM)

- Discrete time
- First-order Markov chain

$$
\begin{aligned}
x_{t+1} &= a x_t + \epsilon, \epsilon \sim N(\mu_\epsilon, \sigma_\epsilon^2) \\
y_t &= b x_t + \eta, \eta \sim N(\mu_\eta, \sigma_\eta^2)
\end{aligned}
$$



Inferring the distributions of interest

- Many methods exist to infer these distributions
  - Markov Chain Monte Carlo (MCMC)
  - Variational inference
  - Belief propagation
  - etc.
- In this setting sequential inference is possible because of characteristics of the model structure and preferable due to the problem requirements

Exploiting LSSM model structure…

$p(x_i|\mathbf{y}_{1:i})$

$\propto \qquad p(x_i|\mathbf{y}_{1:i-1})$

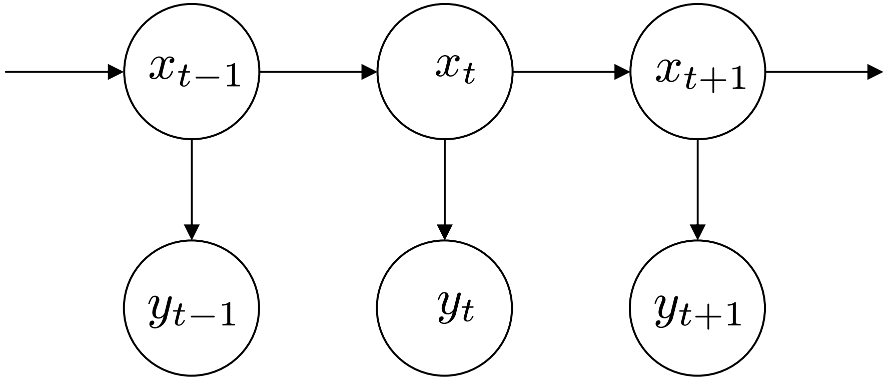
Particle filtering

$$\boxed{p(x_i|\mathbf{y}_{1:i})} \propto p(y_i|x_i) \int p(x_i|x_{i-1}) \boxed{p(x_{i-1}|\mathbf{y}_{1:i-1})} dx_{i-1}$$

Exploiting Markov structure…

$p(x_i|\mathbf{y}_{1:i})$

Use Bayes' rule and the conditional independence structure dictated by the first order Markov hidden variable model



$\propto \qquad p(x_i|\mathbf{y}_{1:i-1})$

for sequential inference

$$
\begin{aligned}
p(x_i|\mathbf{y}_{1:i}) &= \int p(x_i, \mathbf{x}_{1:i-1}|\mathbf{y}_{1:i}) d\mathbf{x}_{1:i-1} \\
&\propto \int p(y_i|x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1}) p(x_i, \mathbf{x}_{1:i-1}, \mathbf{y}_{1:i-1}) d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i) \int p(x_i|\mathbf{x}_{1:i-1}, \mathbf{y}_{1:i}) p(\mathbf{x}_{1:i-1}|\mathbf{y}_{1:i}) d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i) \int p(x_i|x_{i-1}) p(\mathbf{x}_{1:i-1}|\mathbf{y}_{1:i-1}) d\mathbf{x}_{1:i-1} \\
&\propto p(y_i|x_i) \int p(x_i|x_{i-1}) p(x_{i-1}|\mathbf{y}_{1:i-1}) dx_{i-1} \\
&\propto p(y_i|x_i) p(x_i|\mathbf{y}_{1:i-1})
\end{aligned}
$$

An alternative view

$$\{\hat{w}_m^i, \hat{x}_m^i\} \quad \sim \quad \sum_{\ell=1}^{L} w_\ell^{i-1} p(x_i | x_\ell^{i-1})$$

$$p(x_i|\mathbf{y}_{1:i-1}) \approx \textstyle\sum_{m=1}^{M} \hat{w}_m^i \delta_{\hat{x}_m^i}$$

$$p(x_i|\mathbf{y}_{1:i}) \approx \textstyle\sum_{m=1}^{M} p(y_i|\hat{x}_m^i) \hat{w}_m^i \delta_{\hat{x}_m^i}$$

Sequential importance sampling inference

- Start with a discrete representation of the posterior up to observation i-1
- Use Monte Carlo integration to represent the posterior predictive distribution as a finite mixture model
- Use importance sampling with the posterior predictive distribution as the proposal distribution to sample the posterior distribution up to observation i

What?

Posterior predictive distribution

State model

Likelihood

$$p(x_i|\mathbf{y}_{1:i}) \propto p(y_i|x_i)\int p(x_i|x_{i-1})p(x_{i-1}|\mathbf{y}_{1:i-1})dx_{i-1}$$

$$\propto p(y_i|x_i)p(x_i|\mathbf{y}_{1:i-1})$$

Start with a discrete
representation of this
distribution

Monte Carlo integration

General setup

$$p(x) \approx \sum_{\ell=1}^{L} w_\ell \delta_{x_\ell} \qquad \lim_{L\to\infty} \sum_{\ell=1}^{L} w_\ell f(x_\ell) = \int f(x) p(x) dx$$

As applied in this stage of the particle filter

$$\begin{aligned} p(x_i|\mathbf{y}_{1:i-1}) &= \int p(x_i|x_{i-1}) p(x_{i-1}|\mathbf{y}_{1:i-1}) dx_{i-1} \\ &\approx \boxed{\sum_{\ell=1}^{L} w_\ell^{i-1} p(x_i|x_\ell^{i-1})} \end{aligned}$$

Samples from

Finite mixture model