

Particle Filtering

a brief introductory tutorial

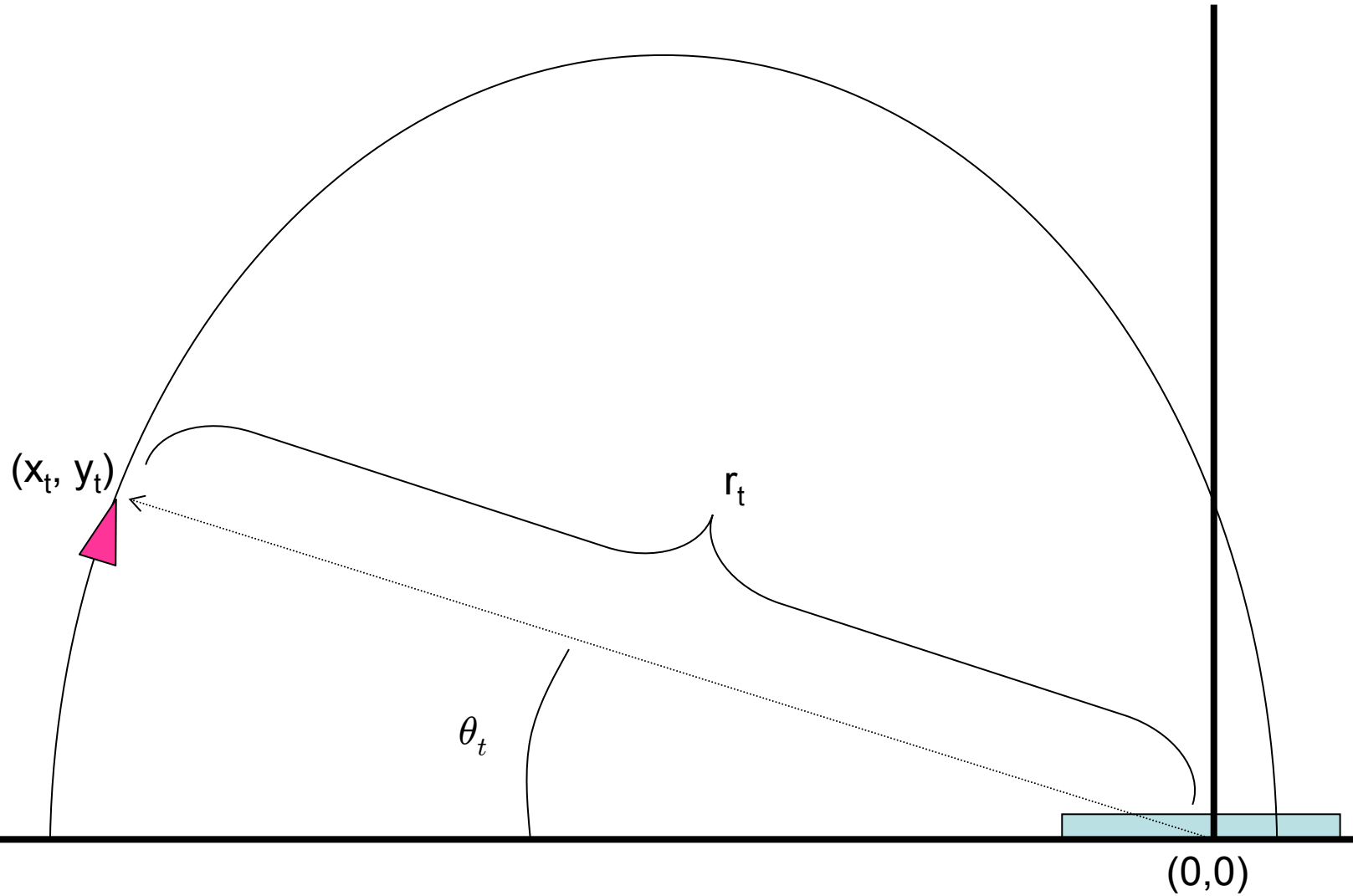
Frank Wood

Gatsby, August 2007

Problem: Target Tracking

- A ballistic projectile has been launched in our direction and may or may not land near enough to kill us
- We have orders to launch an interceptor only if there is a $>5\%$ chance it will land near enough to kill us (the projectile's kill radius is 100 meters) – the interceptor requires 5 seconds to destroy incoming projectiles
- We live under an evil dictator who will kill us himself if we unnecessarily launch an interceptor (they're expensive after all)

Problem Schematic



Problem problems

- Only noisy observations are available
- True trajectory is unknown and must be inferred from the noisy observations
- Full details will be given at the end of the lecture

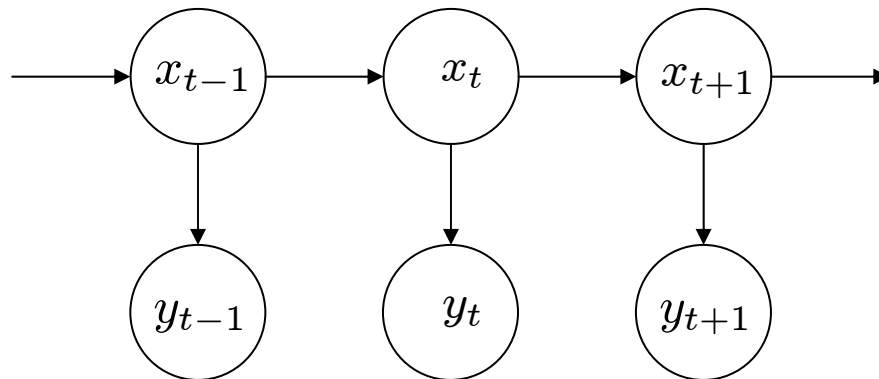
Probabilistic approach

- Treat true trajectory as a sequence of latent variables
- Specify a model and do inference to recover the distribution over the latent variables (trajectories)

Linear State Space Model (LSSM)

- Discrete time
- First-order Markov chain

$$x_{t+1} = ax_t + \epsilon, \epsilon \sim N(\mu_\epsilon, \sigma_\epsilon^2)$$
$$y_t = bx_t + \eta, \eta \sim N(\mu_\eta, \sigma_\eta^2)$$



LSSM

- Joint distribution

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

- For prediction we want the posterior predictive

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

- and posterior (filtering) distributions

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

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...

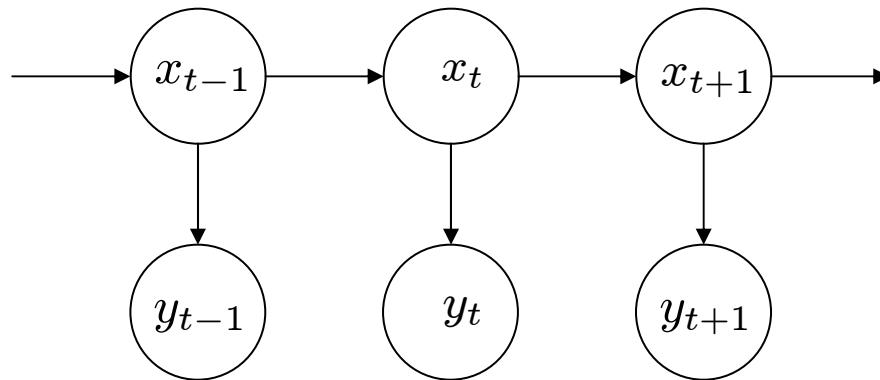
$$p(x_i | \mathbf{y}_{1:i})$$

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

Exploiting LSSM model structure...

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

Use Bayes' rule and the conditional independence structure dictated by the LSSM graphical model



\propto

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

for sequential inference

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

Particle filtering steps

- 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

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

As applied in this stage of the particle filter

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

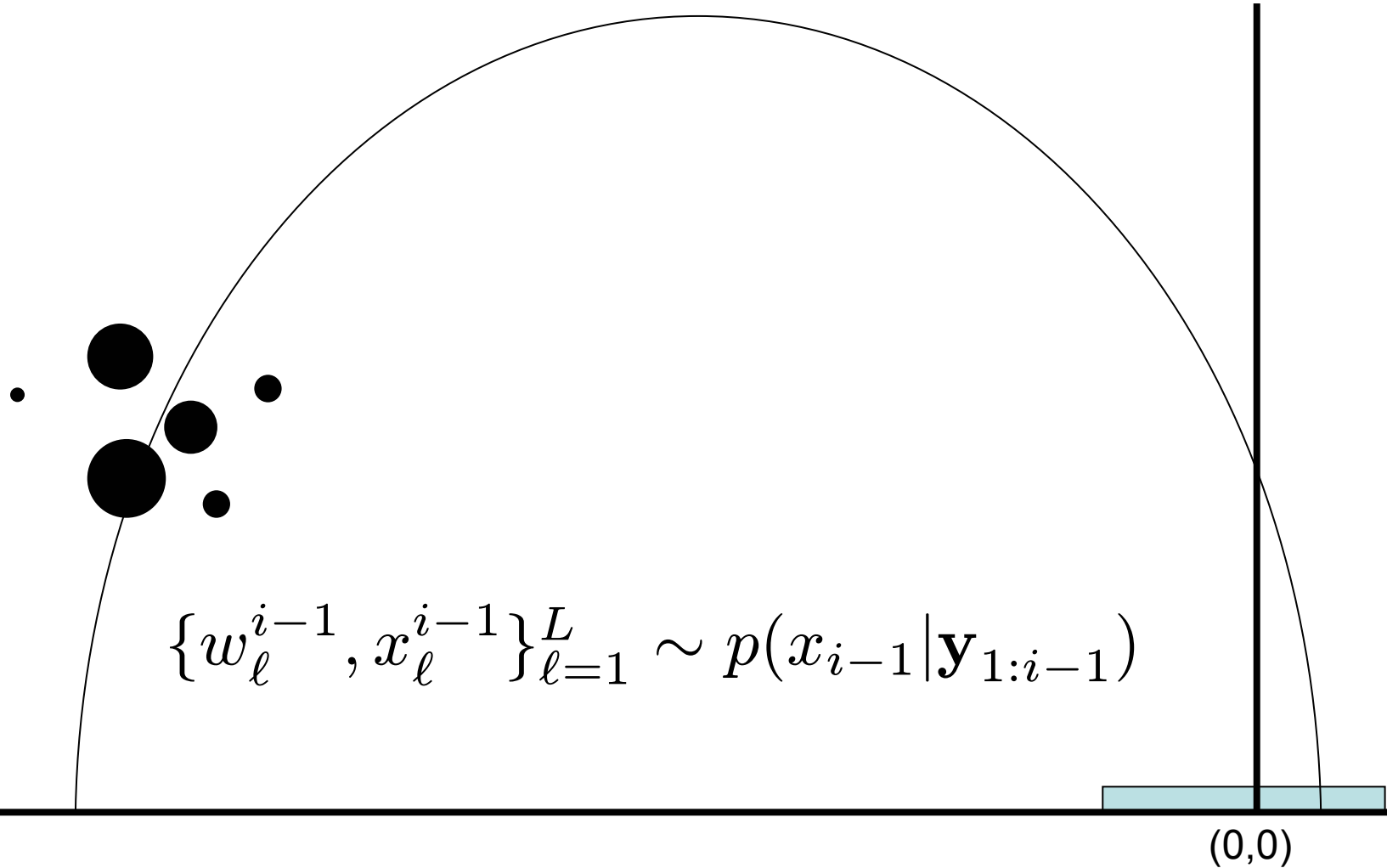
$$\approx \sum_{\ell=1}^L w_{\ell}^{i-1} p(x_i | x_{\ell}^{i-1})$$

Samples from
Finite mixture model

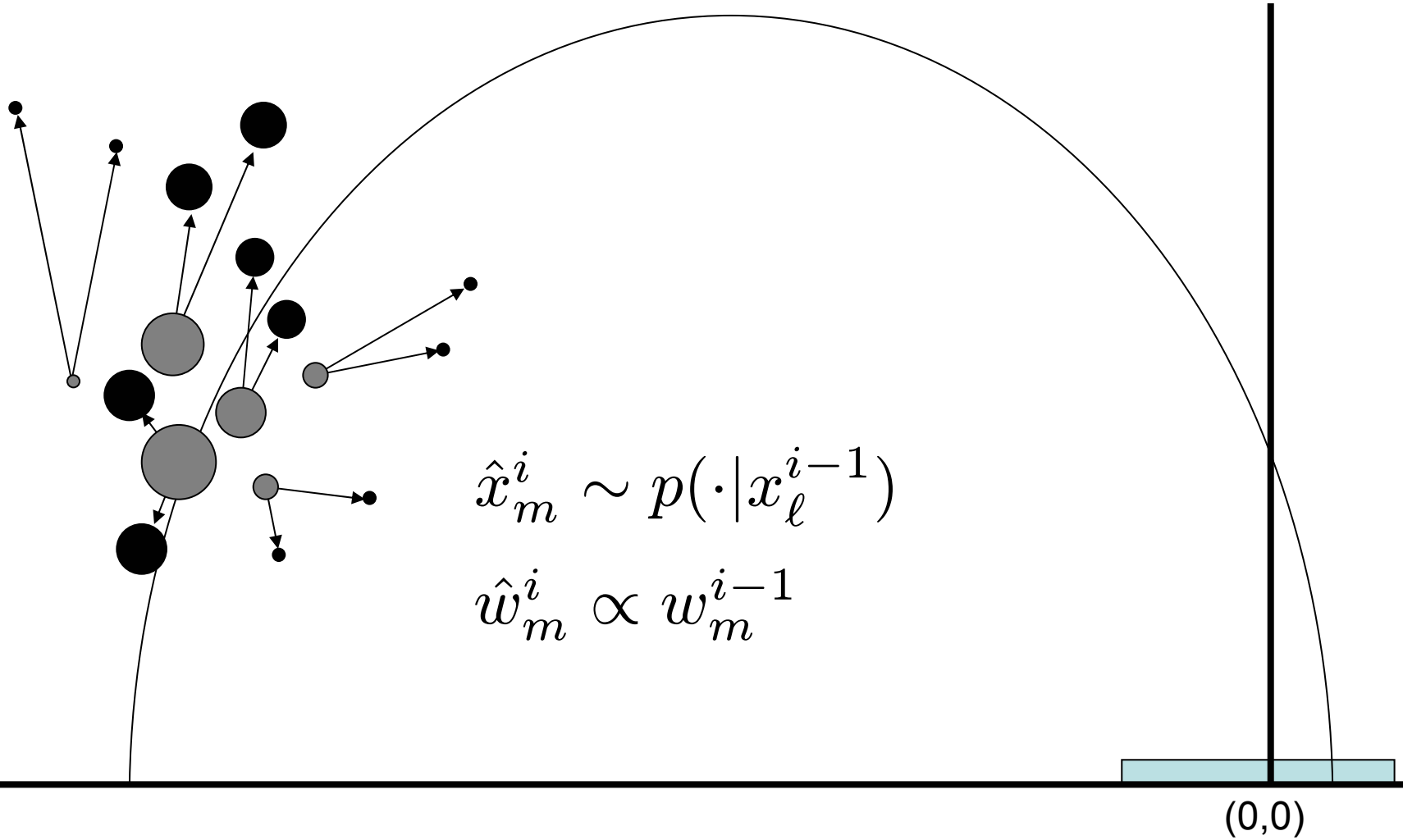
Intuitions

- Particle filter name comes from physical interpretation of samples and the fact that the filtering distribution is often the distribution of interest

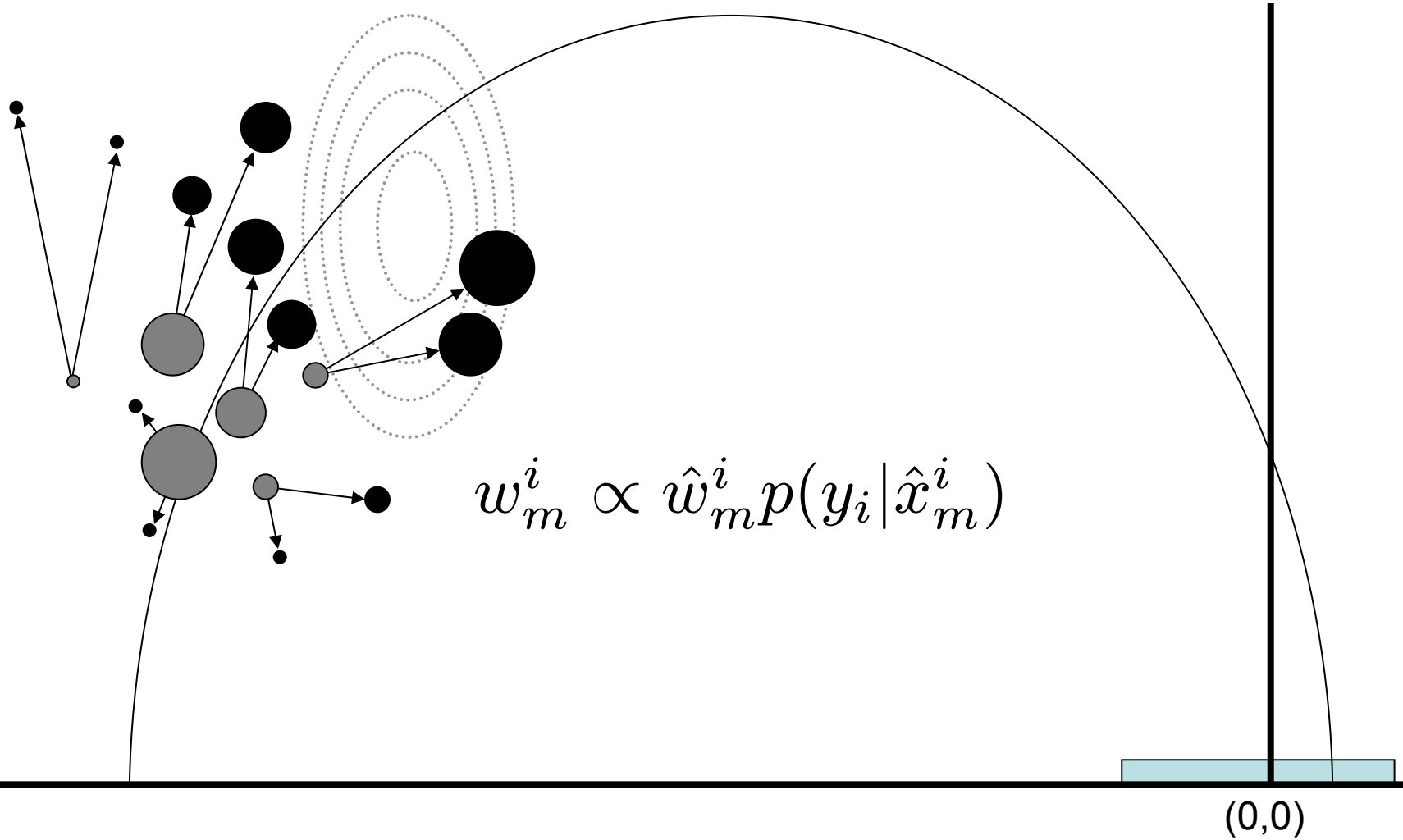
Start with samples representing the hidden state



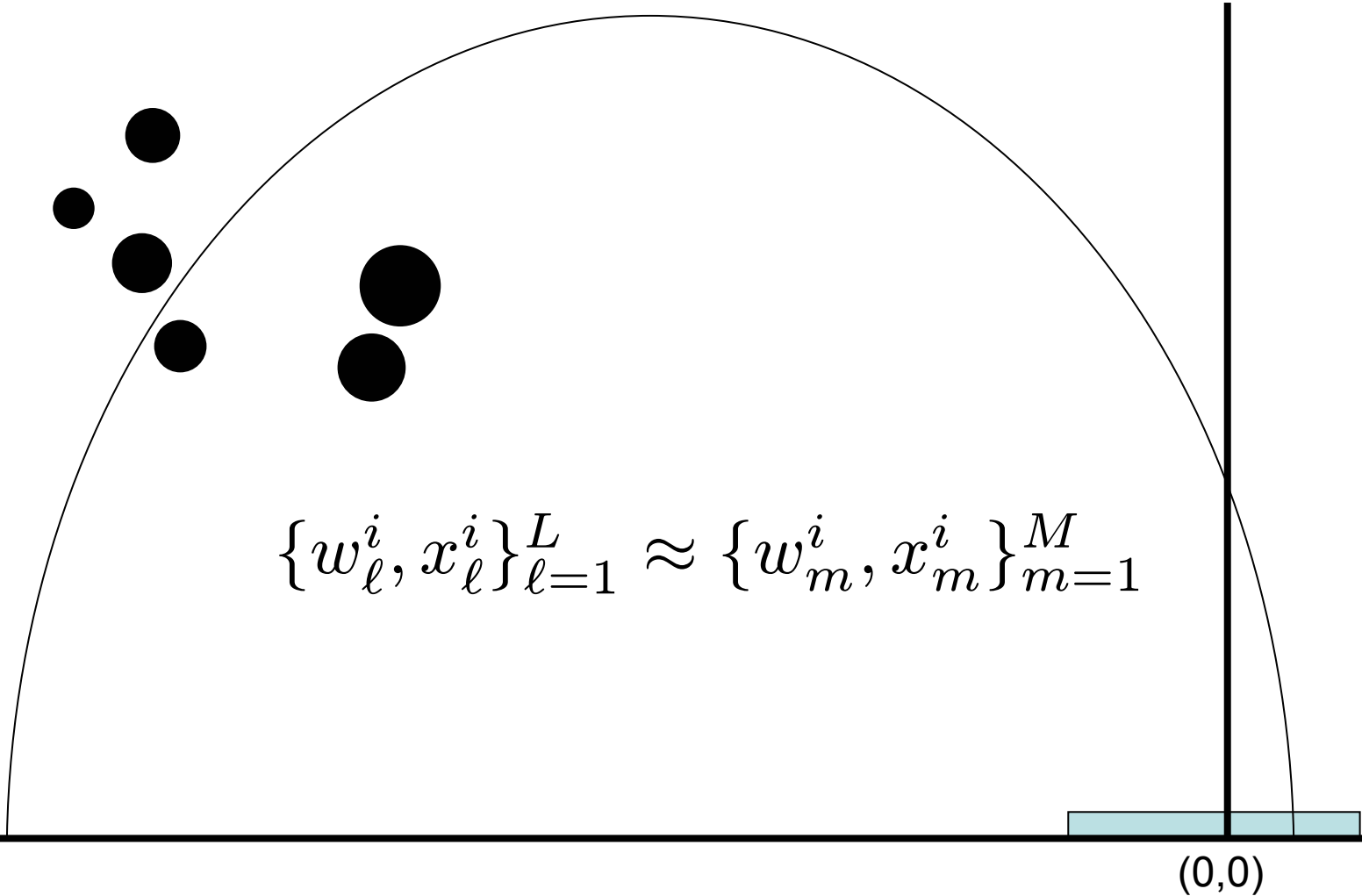
Evolve them according to the state model



Re-weight them by the likelihood



Down-sample / resample



Particle filtering

- Consists of two basic elements:
 - Monte Carlo integration

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

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

- Importance sampling

Importance sampling

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

Original
distribution:
hard to
sample from,
easy to
evaluate

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

Proposal
distribution:
easy to
sample from

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

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

Importance
weights

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

Importance sampling

un-normalized distributions

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

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

$$q(x) = \frac{\tilde{q}(x)}{Z_q}$$

Un-normalized proposal distribution: still easy to sample from

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

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

New term:
ratio of
normalizing
constants

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

Normalizing the importance weights

Un-normalized importance weights

Takes a little algebra

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

Normalized importance weights

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

PF'ing: Forming the posterior predictive

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

Posterior up to observation $i - 1$

$$p(x_i | y_{1:i-1}) = \int p(x_i | x_{i-1}) p(x_{i-1} | y_{1:i-1}) dx_{i-1}$$

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

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

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

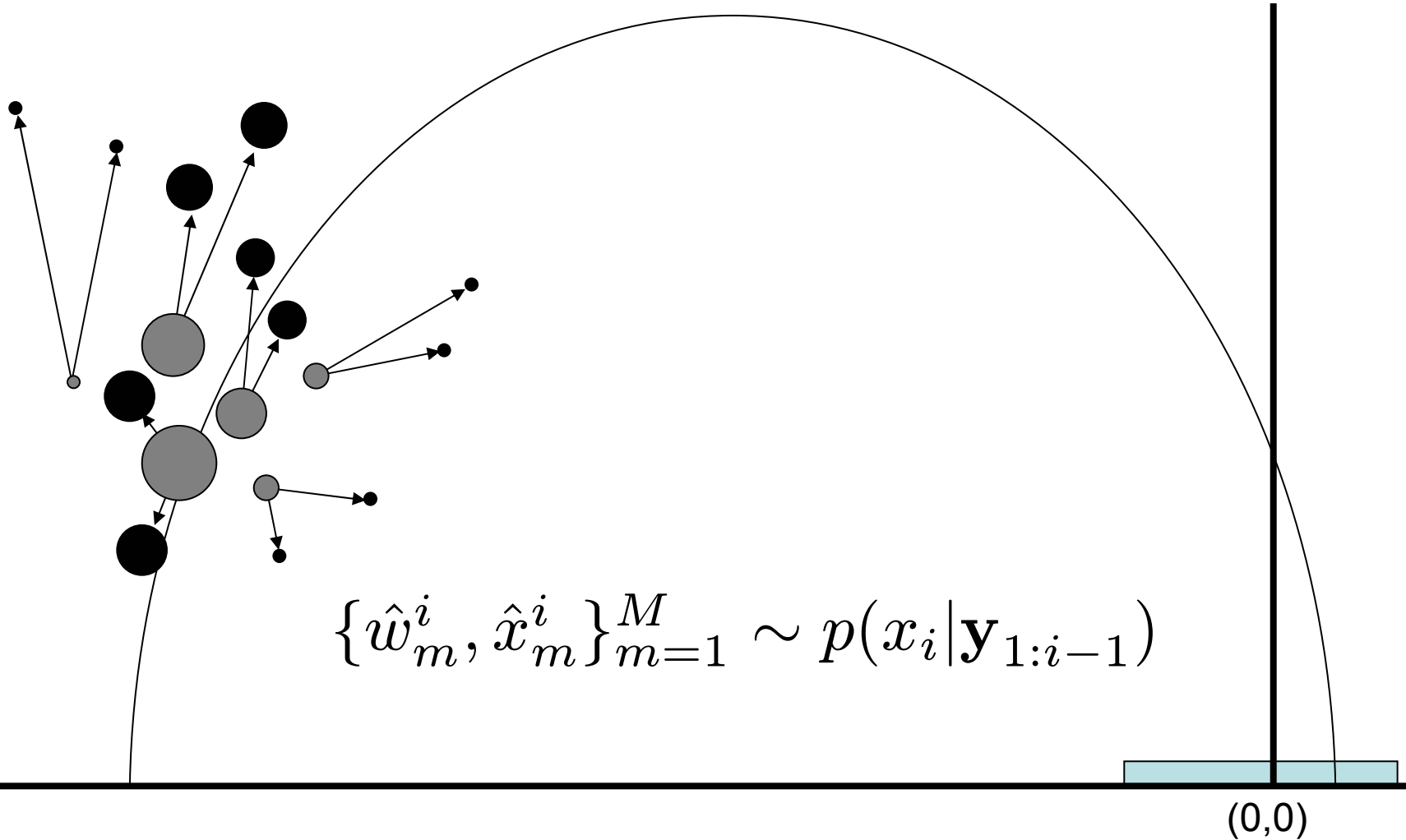
- For instance sample from each mixture component several times such that M , the number of samples drawn, is two times L , the number of densities in the mixture model, and assign weights

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

Not the best

- Most efficient Monte Carlo estimator of a function $I(x)$
 - 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 I over the mixture density
- Take home: smarter sampling possible

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

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

Distribution from which we want to sample

$$q(x_i)$$

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

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

Proposal distribution

Sequential importance sampling

- Weighted posterior samples arise as

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

- Normalization of the weights takes place as before

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

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

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

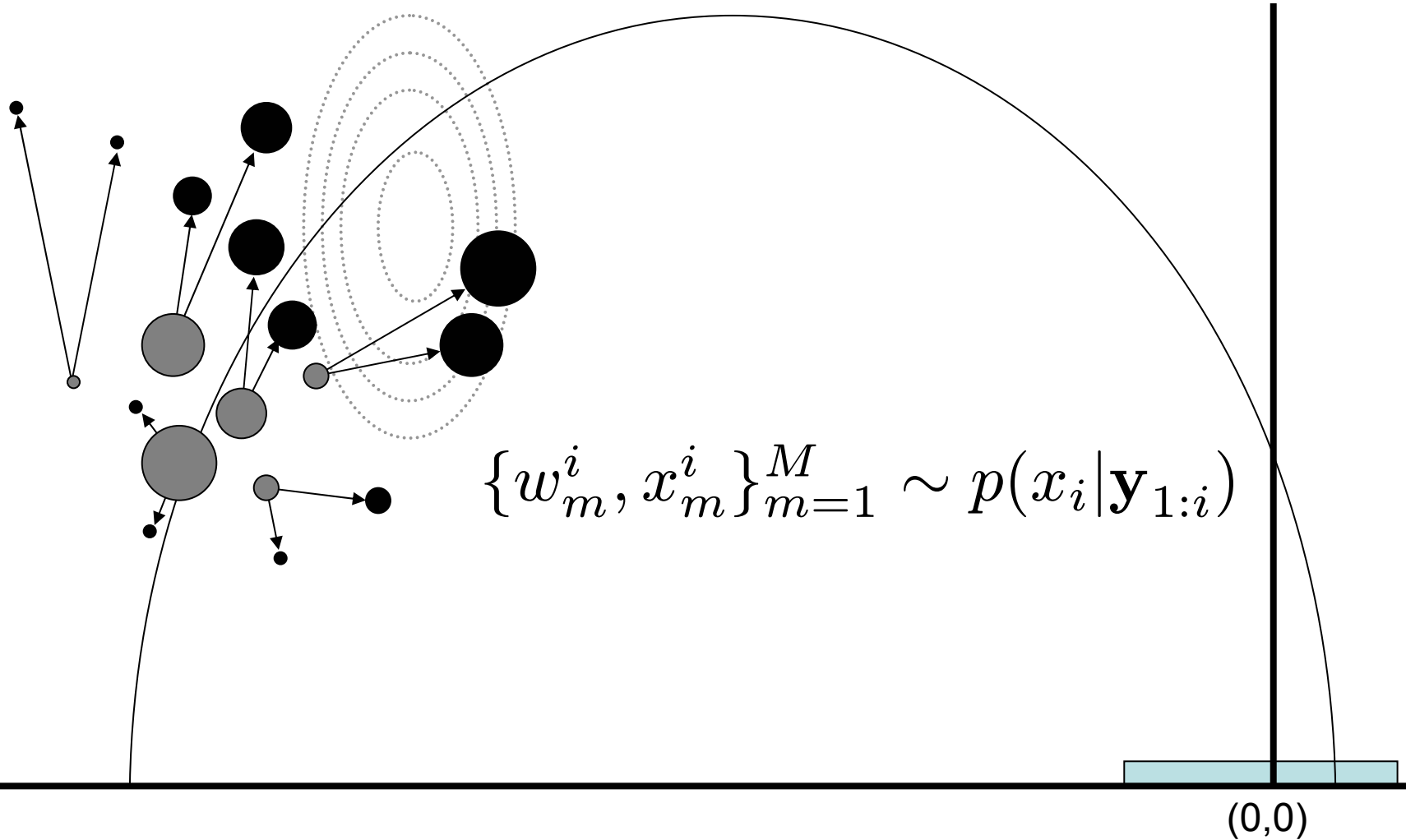
An alternative view

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

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

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

Importance sampling from the posterior distribution



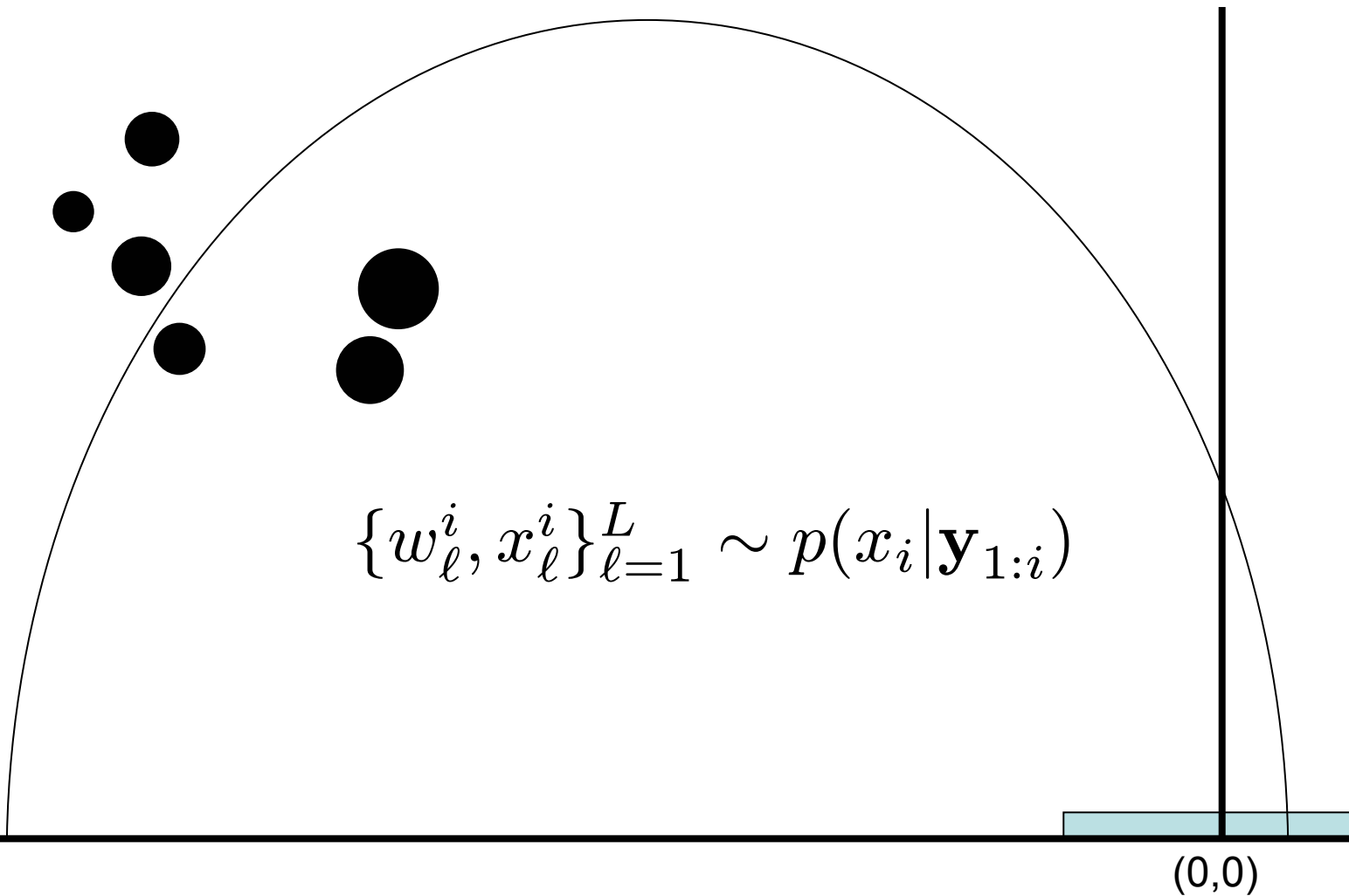
Sequential importance re-sampling

- 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

Down-sampling the particle set



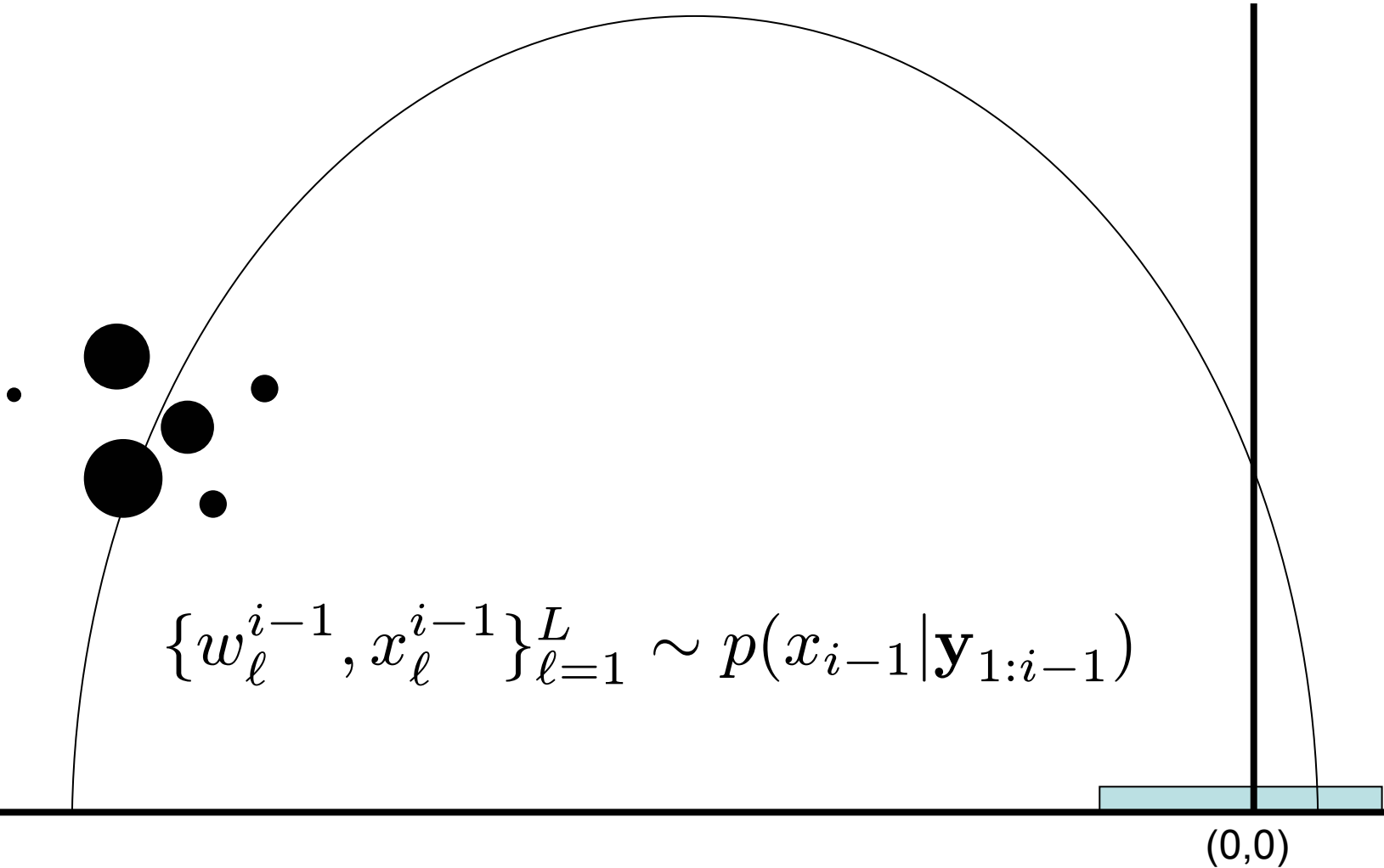
Recap

- 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

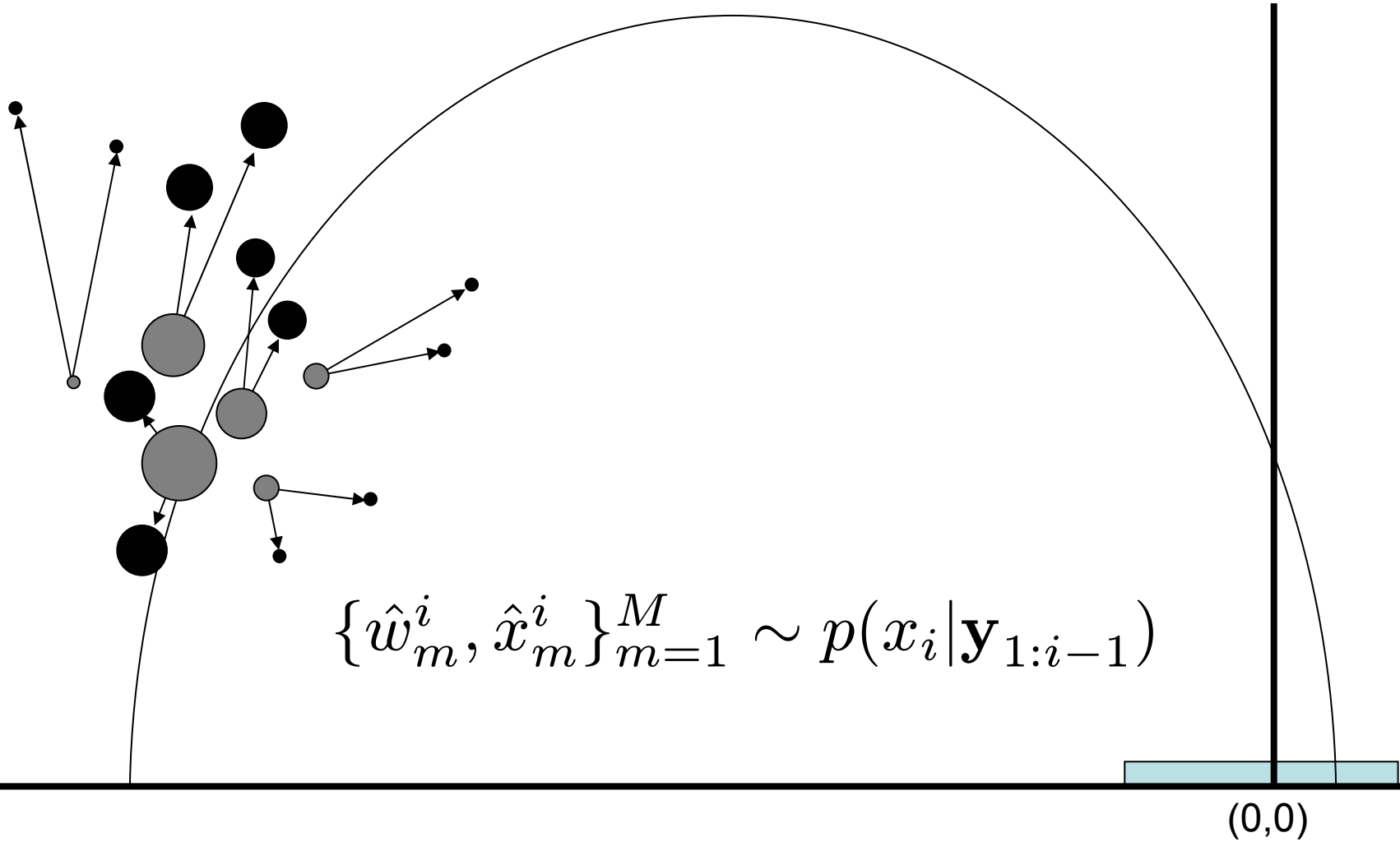
SIR Particle Filter

- The Sequential Importance Sampler (SIS) particle filter multiplicatively updates weights at every iteration and never resamples particles (skipped)
 - Concentrated weight / particle degeneracy
- The Sequential Importance Resampler (SIR) particle filter avoids many of the problems associated with SIS particle filtering but always uses representations with L particles and does not maintain a weighted particle set

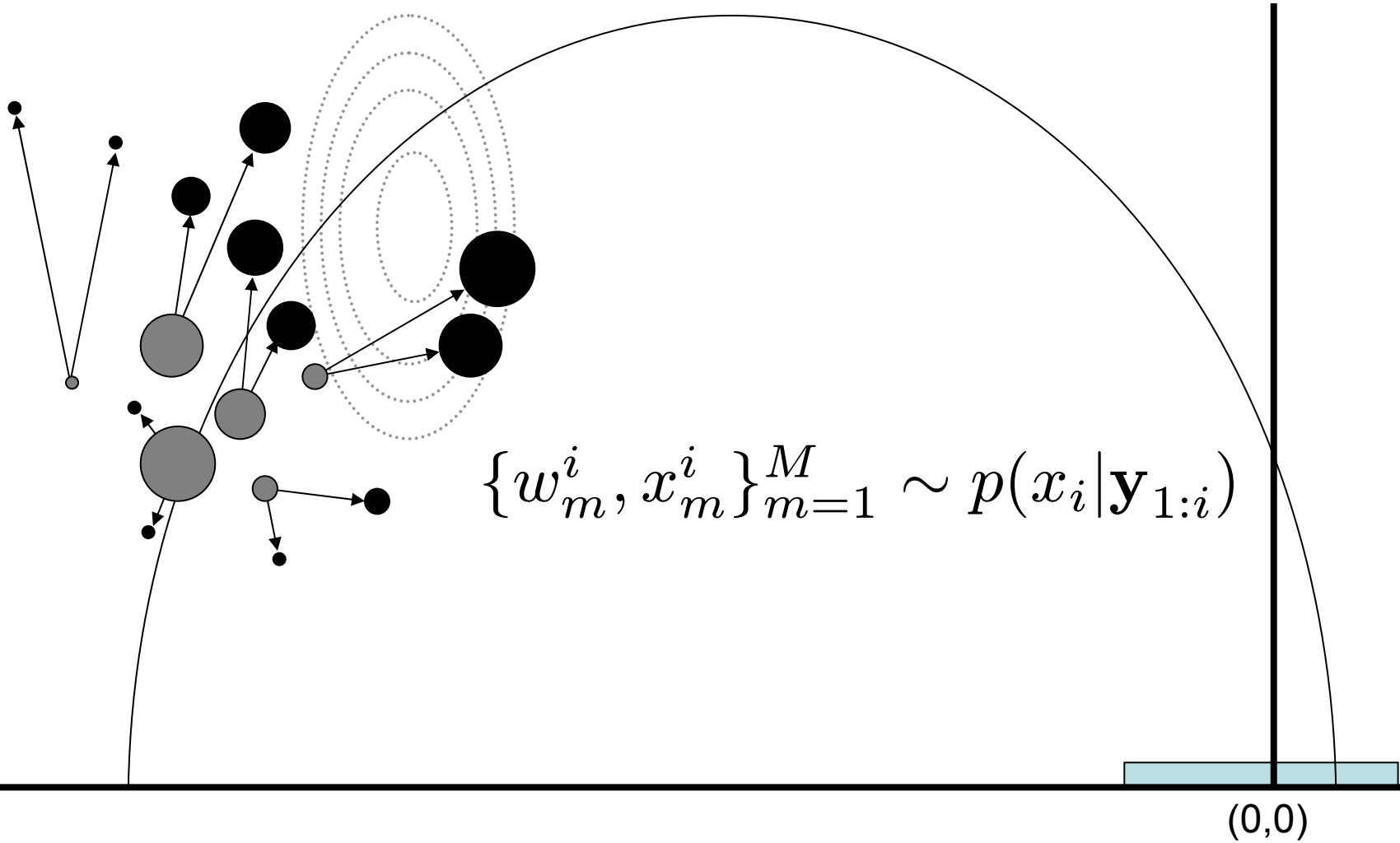
Particle evolution



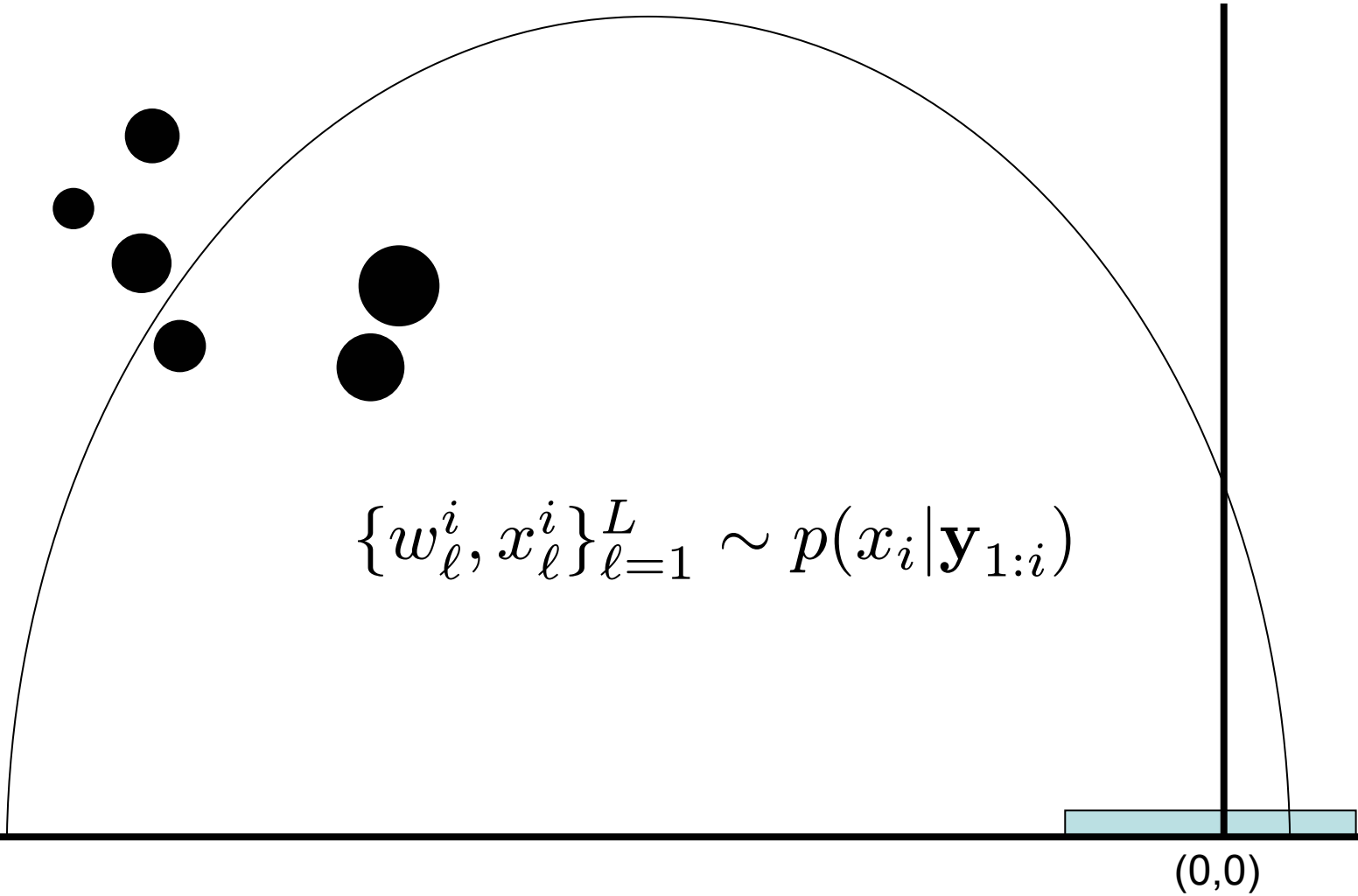
Particle evolution



Particle evolution



Particle evolution



LSSM Not alone

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

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

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

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

- for $n=1:K$

- choose k according to π_k
- sample x_n from f_k
- set w_n equal to $1/N$

- for $n=1:K$

- 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 $1/c$

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.

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

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 $\text{Gamma}(1,100)$ 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_0 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

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

Sampling

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

$$\{w_\ell, x_\ell\}_{\ell=1}^L \sim p(\cdot)$$
$$\lim_{L \rightarrow \infty} \sum_{\ell=1}^L w_\ell \delta_{x_\ell} \rightarrow p(\cdot)$$

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

The diagram illustrates Bayes' theorem using color-coded components:

- Posterior** ($p(\theta|x)$) is shown in a pink box on the left.
- Likelihood** ($p(x|\theta)$) is shown in a yellow box in the numerator of the first fraction.
- Prior** ($p(\theta)$) is shown in a blue box in the numerator of the first fraction.
- Evidence** ($p(x)$) is shown in a green box in the denominator of the first fraction.
- The **Marginalized Prior** ($\int p(x|\theta)p(\theta)d\theta$) is shown in yellow and blue boxes in the denominator of the second fraction.

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

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

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

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

$$p(x|\theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}$$

x successes in N trials, θ probability of success

Conjugacy Continued

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

$$\theta|x \sim \text{Beta}(\alpha + x, \beta + N - x)$$

$$Z(x) = \left(\frac{\Gamma(\alpha + \beta + N)}{\Gamma(\alpha + x)\Gamma(\beta + N - x)} \right)^{-1}$$

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

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

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

Relating the Posterior to the Posterior Predictive

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

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

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

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