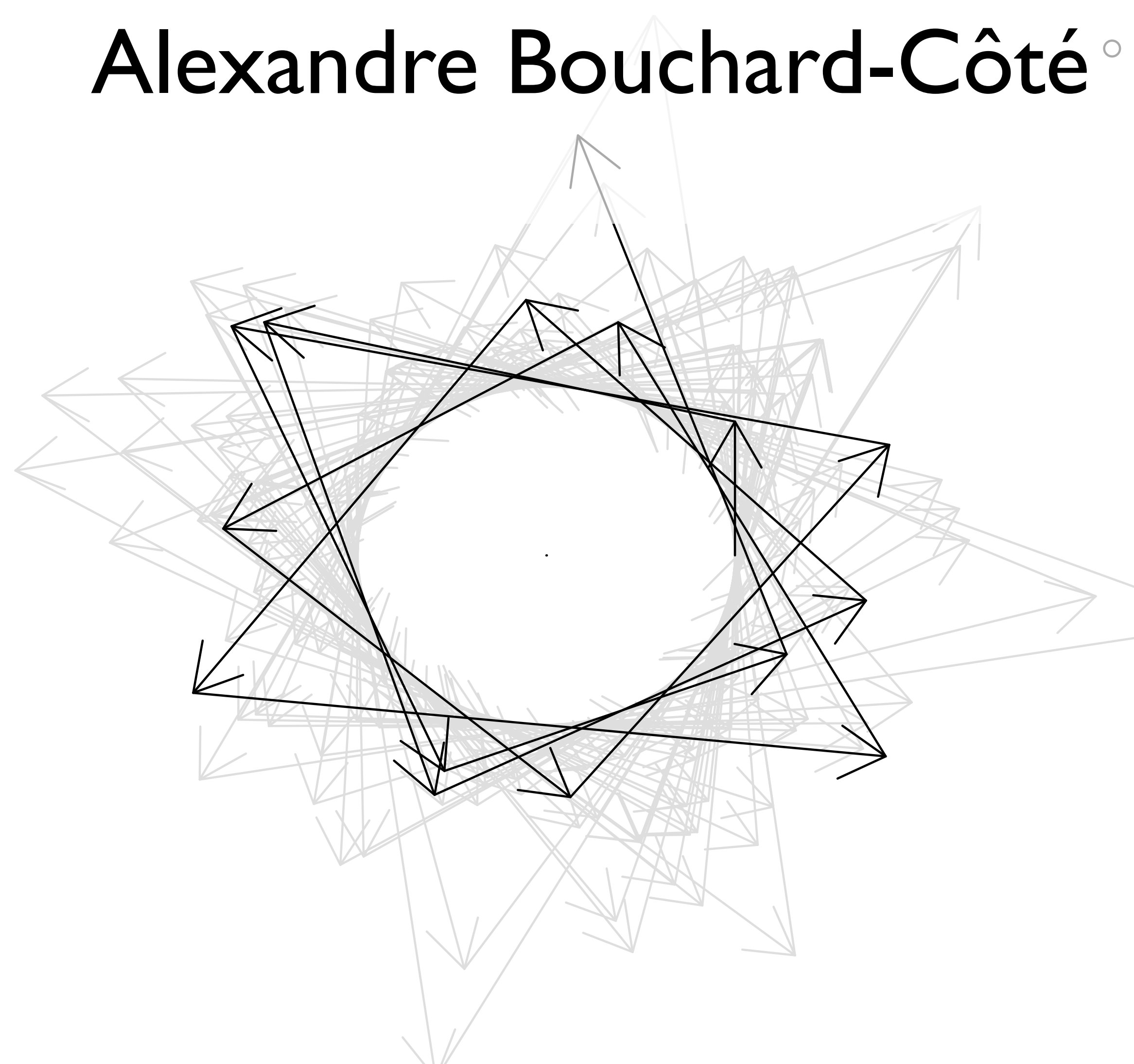
Monte Carlo methods

Alexandre Bouchard-Côté



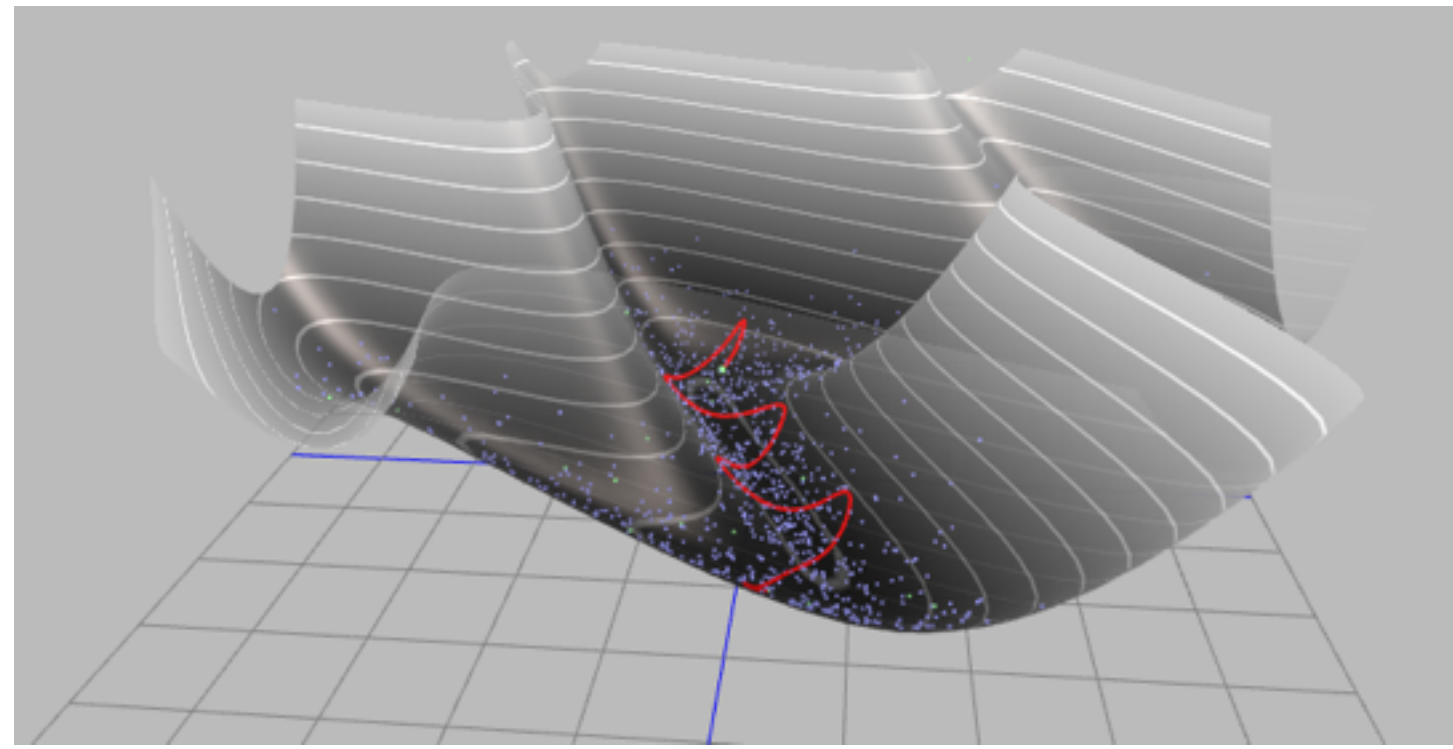
Announcements

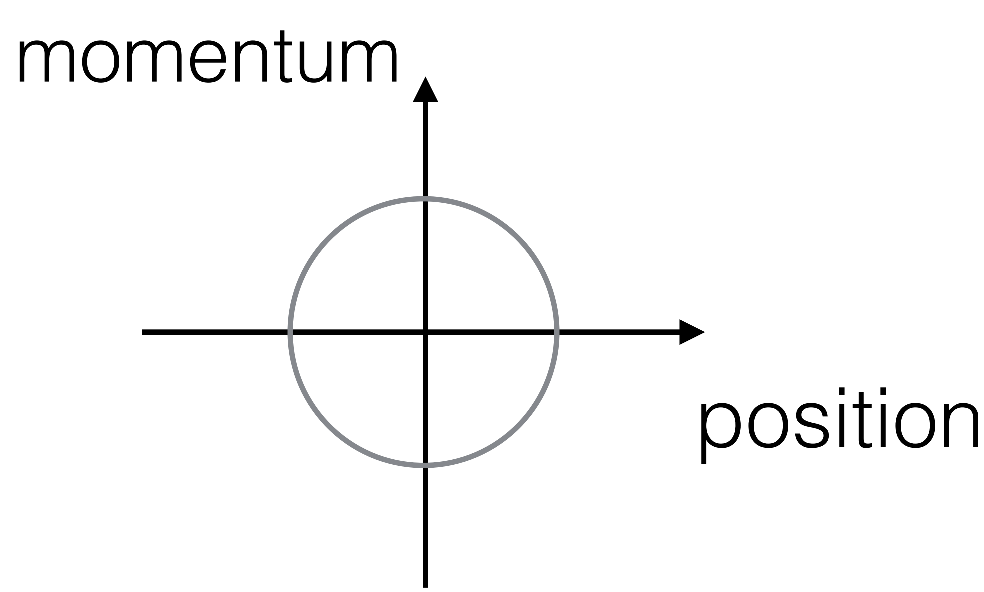
- Assignment due tomorrow morning
- Project
  - Due: April 26 (send code + pdf by email)
  - Send informal plan for project by email by tomorrow morning
- One final lab next Wednesday (participation points) during office hours
- Extra optional lecture Friday DMP 110, 4:00-4:45

Gradient-based methods

Hamiltonian Monte Carlo: intuition

- Physical ball rolling on the *energy*
  - U(x) = -log(p(x))
- Motion described by the *Hamiltonian flow*
- Phase space on a Gaussian target:





HMC: auxiliary variables

- Physics' notation: z = (q, p)
  - position q
  - Augment the state with a momentum random variable p
- Put an auxiliary distribution on p, with f(p) = exp(-K(p)) and s.t. K(p) = K(-p), e.g. normal.

$$H(q,p) = U(q) + K(p), \quad U(q) = q^2/2, \quad K(p) = p^2/2$$

- Can think of p as a velocity (when the mass matrix, i.e. covariance of f(x) is identity).
- Statistical notation would be then z = (x, v)

Exact HMC

- MCMC kernel is a non-reversible
- Given by a Dirac delta: k(z, dz') = $\delta_{\Phi(z)}$(dz')
- Φ is the Hamiltonian flow, i.e. solutions of the differential equations

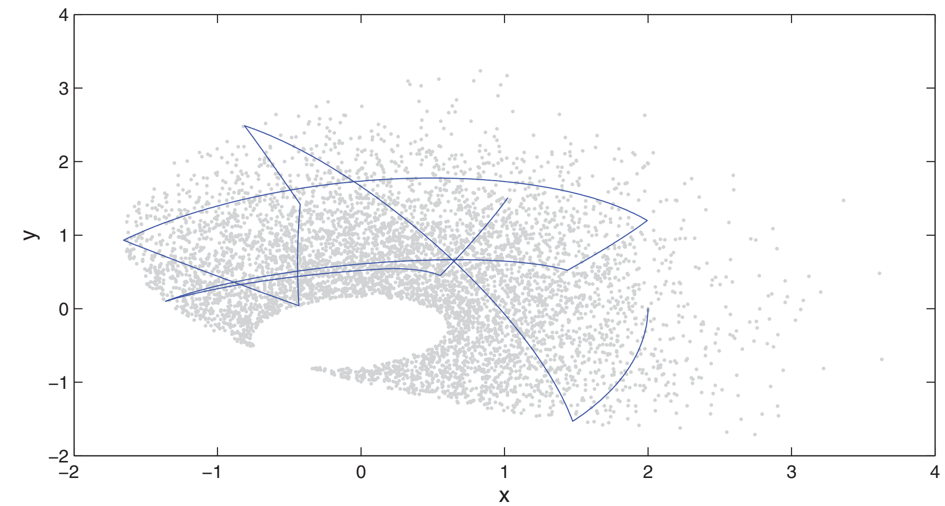
$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \frac{dq_i}{dt} &= [M^{-1}p]_i \\ \frac{dp_i}{dt} &= -\frac{\partial U}{\partial q_i} \end{aligned}$$

- Exact HMC: Analytic solution only in special cases, e.g. for (truncated) normal target we get:

$$q(t) = r\cos(a+t), \quad p(t) = -r\sin(a+t)$$

Application: truncated normal distributions

- See Pakman and Paninski (2014)
- Truncated normal arise in many practical contexts:
  - Probit and tobit models
  - Bayesian splines for positive functions
  - Bayesian lasso



$$y_i = \text{sign}(w_i)$$

$$w_i = -\mathbf{z}_i \cdot \boldsymbol{\beta} + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, 1)$$

Exact HMC: invariance

- MCMC kernel is non-reversible
- Given by a Dirac delta: $k(z, dz') = \delta_{\Phi(z)}(dz')$
- Invariance equivalent to:
  - given $Z \sim$ extended target π'
    $\pi'(x, v) = \pi(x) \times \text{normal}(v)$
  - Define $Y = \Phi(Z)$
  - Do we have $Y \sim \pi'$?

Exact HMC: invariance

- By change of variable formula, break into two factors:

$$f_Y(y) = f_Z(\Phi^{-1}(y))\,|\det J_{\Phi^{-1}}(y)|$$

hence ingredient to show Y ~ π' are:

- Φ invertible (yes, for inverse set v ⟵ -v)
- *Conservation of Hamiltonian*: first factor is constant
- *Volume preservation*: second factor is constant

Conservation of Hamiltonian

- Want f(z) = f(Φ(z))
- Enough: no infinitesimal Hamiltonian changes, H' = 0 [prime notation: derivative w.r.t *t*]
- Use total derivative identity

$$\frac{dH}{dt} = \sum_{i=1}^{d}\left[\frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial H}{\partial p_i}\right]$$

- Then substitute our choice of the differential equation:

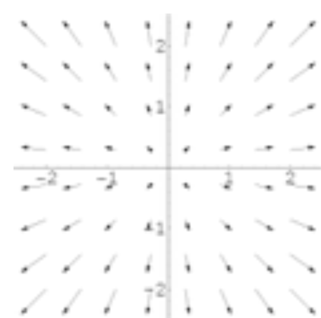
$$\begin{aligned}\frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i}\\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}\end{aligned} \implies \sum_{i=1}^{d}\left[\frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial H}{\partial p_i}\right] = \sum_{i=1}^{d}\left[\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i}\right] = 0$$

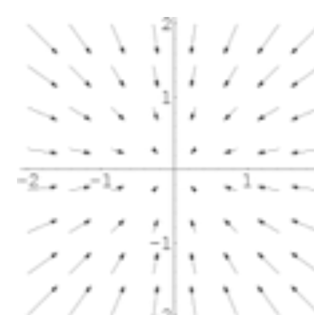
Volume preservation

The preservation of volume by Hamiltonian dynamics can be proved in several ways. One is to note that the divergence of the vector field defined by equations (2.1) and (2.2) is zero, which can be seen as follows:

$$\sum_{i=1}^{d}\left[\frac{\partial}{\partial q_i}\frac{dq_i}{dt}+\frac{\partial}{\partial p_i}\frac{dp_i}{dt}\right] = \sum_{i=1}^{d}\left[\frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i}-\frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i}\right] = \sum_{i=1}^{d}\left[\frac{\partial^2 H}{\partial q_i \partial p_i}-\frac{\partial^2 H}{\partial p_i \partial q_i}\right] = 0 \quad (2.13)$$

A vector field with zero divergence can be shown to preserve volume (Arnold, 1989).

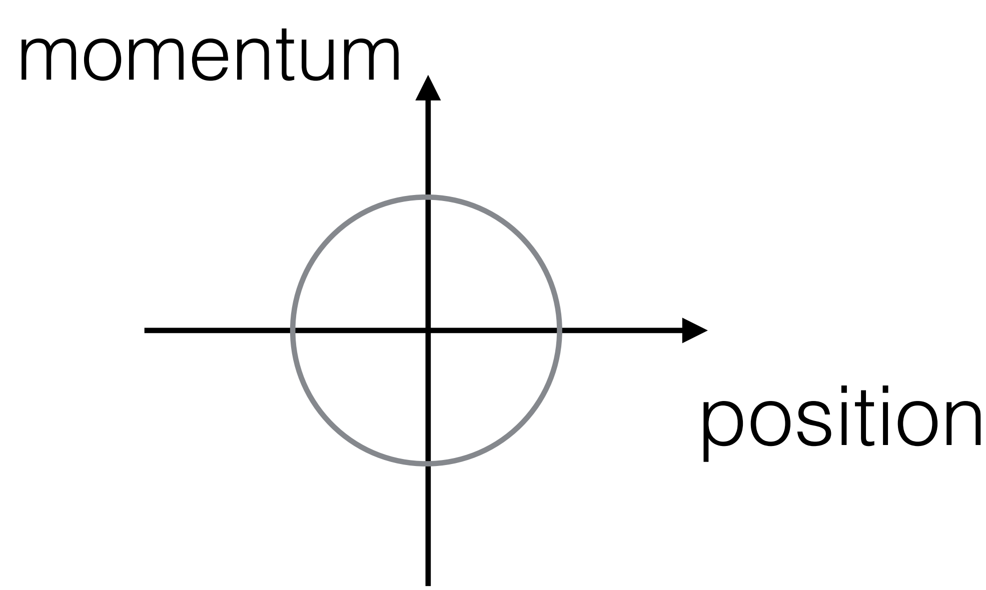




- See Neal (2012). MCMC using Hamiltonian dynamics for another, more direct argument

Exact HMC: irreducibility

- Easy to see non irreducible in phase space

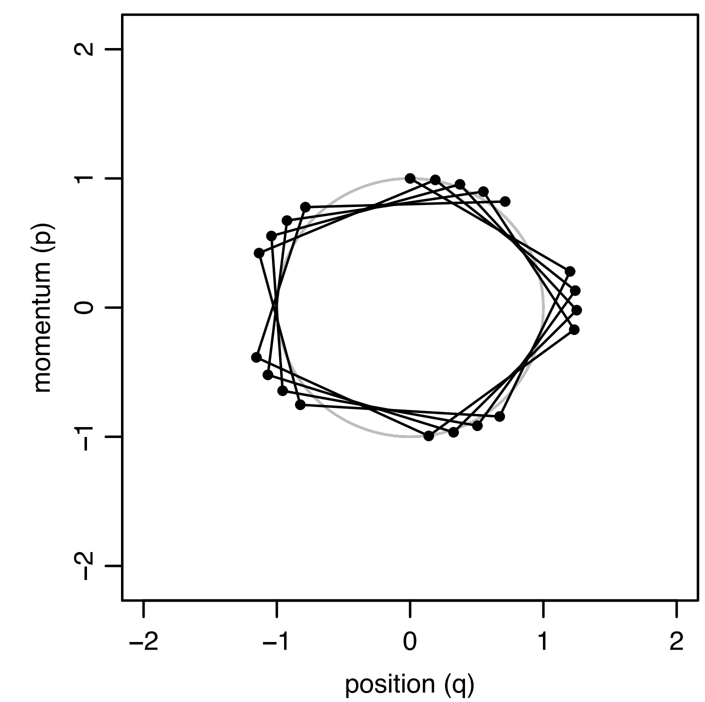


- Solution: refresh momentum

Symplectic HMC

- We can’t simulate the exact Hamiltonian flow for most targets of interest.
- Idea:
  - solve the differential equation using numerical methods and initial condition given by current point
    - can be done so that volume still preserved (e.g. with leap-frog integrator)
  - Hamiltonian no longer exactly preserved, so use MH to accept-reject

Symplectic HMC



- Numerical solution example:
- Algorithm: numerically follow the evolution of diff. equation
  - Simplest version: Euler method

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}$$

$$\frac{dq_i}{dt} = [M^{-1}p]_i$$

$$p_i(t+\varepsilon) = p_i(t) + \varepsilon \frac{dp_i}{dt}(t) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t))$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{dq_i}{dt}(t) = q_i(t) + \varepsilon \frac{p_i(t)}{m_i}$$

- Need something better: leap-frog integrator (will see why soon when going over invariance)

Rough idea

- Use accept-reject
- Proposal: deterministic, given by numerical solution of DE followed for a fixed number of steps
- Accept-reject to take into account numerical error
- Why is this not quite correct?

Important, overlooked condition on proposal *q*

- Mutual absolute continuity condition:

$$\int_A \pi(\mathrm{d}x) q(x, B) > 0 \Leftrightarrow \int_B \pi(\mathrm{d}x) q(x, A) > 0$$

- For example, in a discrete state space where the target has full support, this means:

$$q(x, y) > 0 \Leftrightarrow q(y, x) > 0$$

- This can be tricky in combinatorial spaces (more on that soon)

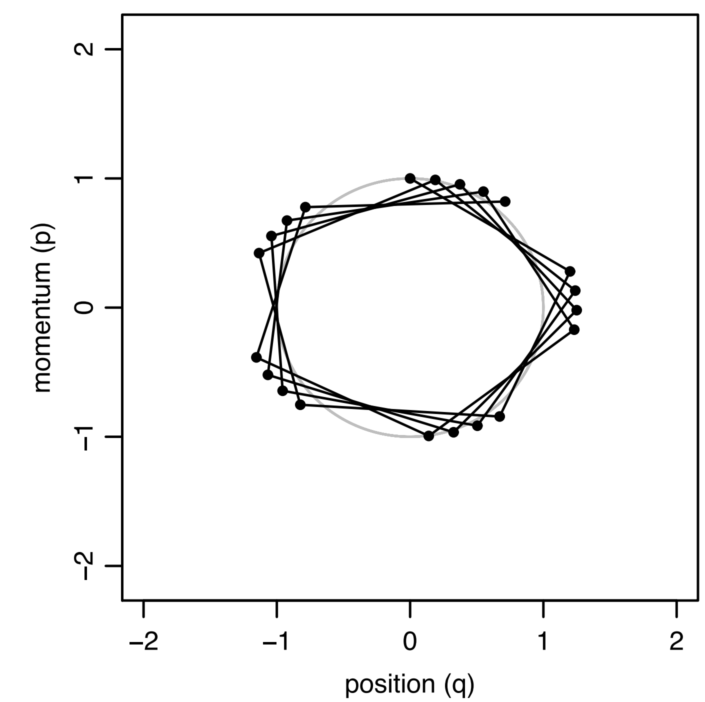
Symplectic HMC

- 2 moves, which have to be deterministically cycled

  1. Φ: an MH move with proposal given by:
     - follow the ~~exact~~ discretized trajectory
     - flip the momentum, R(q,p) = R(q, -p)
  2. Momentum refreshment

- What properties do we need for invariance?

Symplectic HMC



- Numerical solution example:
- Algorithm: numerically follow the evolution of diff. equation
    - Replace Euler by leaf-frog

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}$$

$$\frac{dq_i}{dt} = [M^{-1}p]_i$$

$$p_i(t+\varepsilon/2) = p_i(t) - (\varepsilon/2)\frac{\partial U}{\partial q_i}(q(t))$$

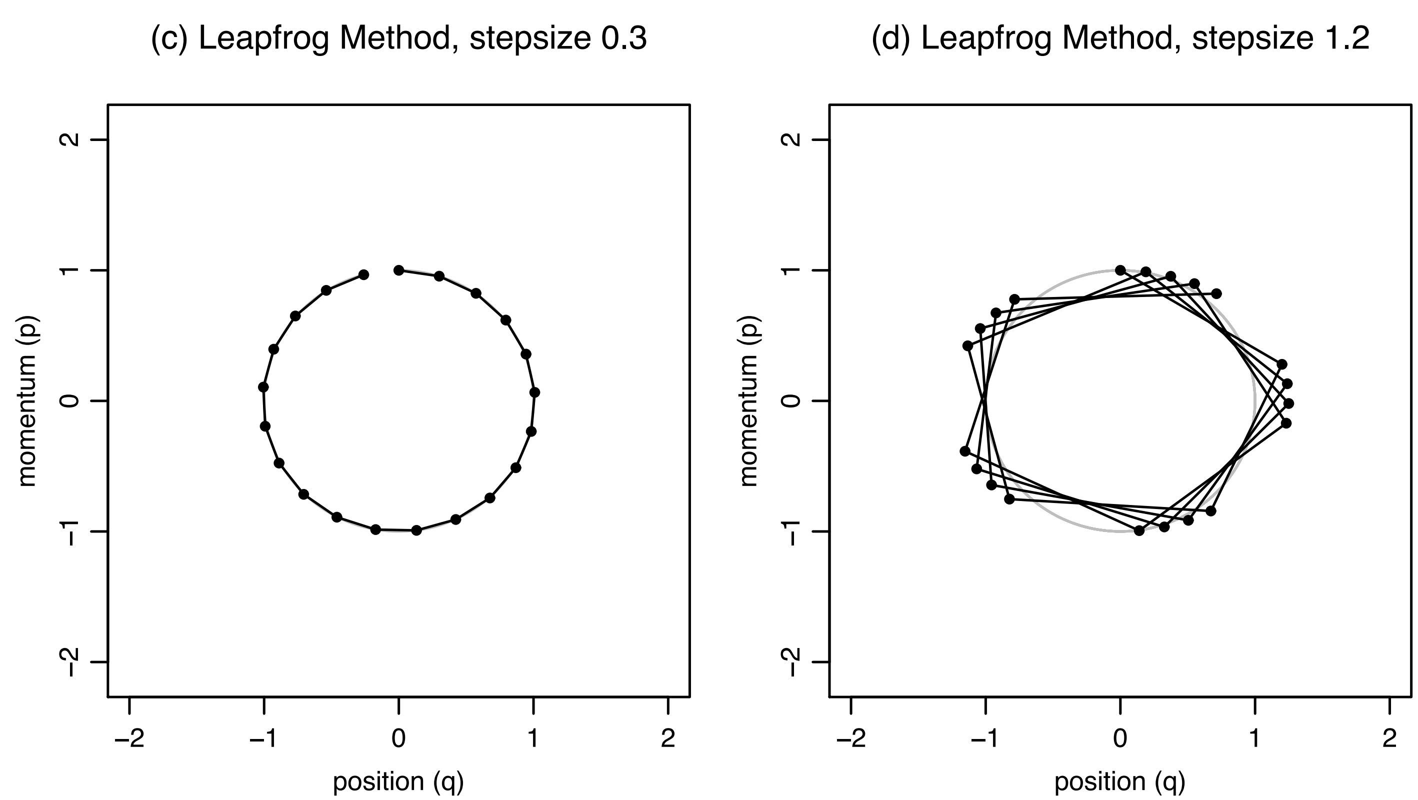
$$q_i(t+\varepsilon) = q_i(t) + \varepsilon\frac{p_i(t+\varepsilon/2)}{m_i}$$

$$p_i(t+\varepsilon) = p_i(t+\varepsilon/2) - (\varepsilon/2)\frac{\partial U}{\partial q_i}(q(t+\varepsilon))$$

- Properties: let R(q,p) = (q,-p) (flip)
    - involution: R(Φ(R(Φ(z)))) = z
    - hence, volume preservation

Practical considerations

- Two <u>critical</u> parameters to tune:
  1. L: number of leap-frog steps
  2. epsilon: step size
- For L: Hoffman 2011, Sohl-Dickstein 2016
- For epsilon: mostly heuristics/adaptation



Special case: Metropolis-Adjusted Langevin (MALA)

- Use one leap frog step, and use the following order for the kernels
  - Refresh velocity first
  - Then do one leap frog, which simplifies into:

$$
\begin{aligned}
q_i^* &= q_i - \frac{\varepsilon^2}{2} \frac{\partial U}{\partial q_i}(q) + \varepsilon p_i \\
p_i^* &= p_i - \frac{\varepsilon}{2} \frac{\partial U}{\partial q_i}(q) - \frac{\varepsilon}{2} \frac{\partial U}{\partial q_i}(q^*)
\end{aligned}
$$

Dimensionality scaling

running time = $\boxed{\text{number of samples needed to get a tolerance (with probability 95\%)}}$ x $\boxed{\text{compute cost per sample}}$

HMC	$d^{1/4}$	d^{1}
MALA	$d^{1/3}$	d^{1}
Random walk MH	d^{1}	d^{1}

SMC

Organization

- SMC on product spaces
- Transforming other problems into product spaces (sequential change of measure)

Motivations for SMC on product spaces

- Sequential predictions / streaming data / HMM / state space models
  - latent state from noisy observation
  - change point
- Time series where 'time' is not time
  - genomics: 'time' = position on genome
    - observations: SNP
    - latent: haploblock (chunk shared by several individuals)

Common feature: the latent space is a product space $F_t = E_1 \times E_2 \times \cdots \times E_t$ indexed by the integers $t \in \{1, \ldots, n\}$.

Sequence of targets

- As in PT we now have a sequence of targets
  - but: with different dimensionality now vs. fixed dimensionality for PT
- In the product space context, sometimes we care about all targets (real time predictions), sometimes, we care only about the last one
- Typical problems:
  - integrating test functions
  - + computing normalization Z (e.g. for model selection, where Z = P(data)

Building block: sequential importance sampling

- Rewrite self-normalized importance sampling so that it can be done with a sequence of targets
- Use the following identities:

$$\gamma(x_{1:n}) = \frac{\gamma(x_{1:n})}{\gamma(x_{1:n-1})}\frac{\gamma(x_{1:n-1})}{\gamma(x_{1:n-2})}\cdots\frac{\gamma(x_{1:1})}{\gamma(x_{\emptyset})}, \qquad q(x_{1:n}) = q(x_1|x_{\emptyset})q(x_2|x_{1:1})q(x_3|x_{1:2})\ldots q(x_n|x_{1:n-1}),$$

- Yields the recursions

$$\begin{aligned} x_t^i &\sim q(\cdot|x_{1:t-1}^i) \\ x_{1:t} &= (x_{1:t-1}^i, x_t^i), \end{aligned} \qquad w_t^i = w_{t-1}^i \frac{\gamma(x_{1:t})}{\gamma(x_{1:t-1})}\frac{1}{q(x_t|x_{1:t-1})}.$$

- Does not work! (Why?) But forms basis of SMC

Fix: resampling

- Intuition: prune particles with low normalized weights
- Constraints: we still want consistency
- Idea: resample N times according to the normalized weights
  - *multinomial resampling*

Notation for our goals

Given a model (joint)...: $\gamma_t(\mathbf{x}_t) = p(\mathbf{x}_t, \mathbf{y}_t)$

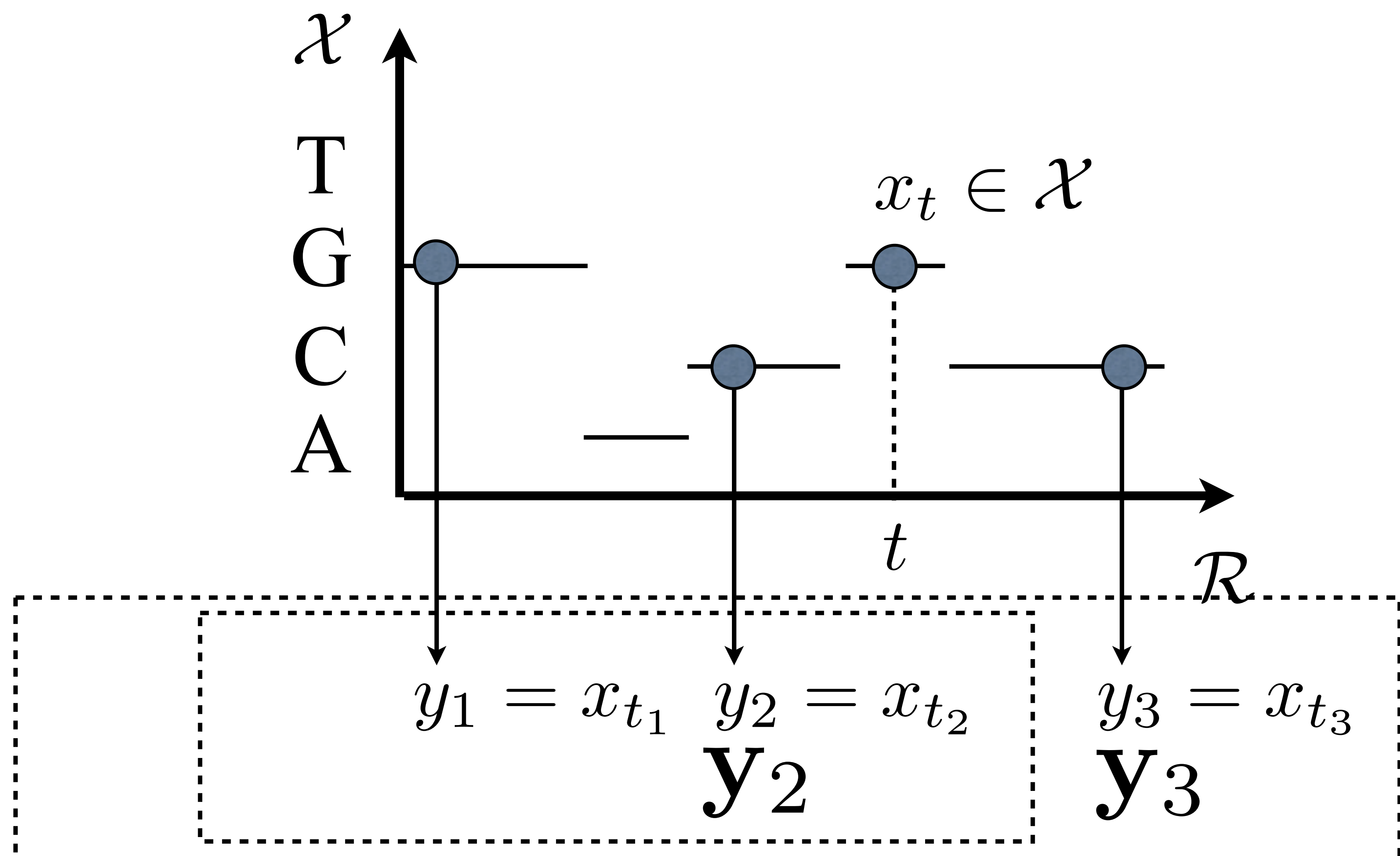
Sample from a *target distribution*: $\pi_t(\mathbf{x}_t) = p(\mathbf{x}_t|\mathbf{y}_t)$

$$\pi_t(\mathbf{x}_t) = \frac{\gamma_t(\mathbf{x}_t)}{Z_t}$$

.. and/or evaluate the normalization: $Z = p(\mathbf{y}_t)$

Notation

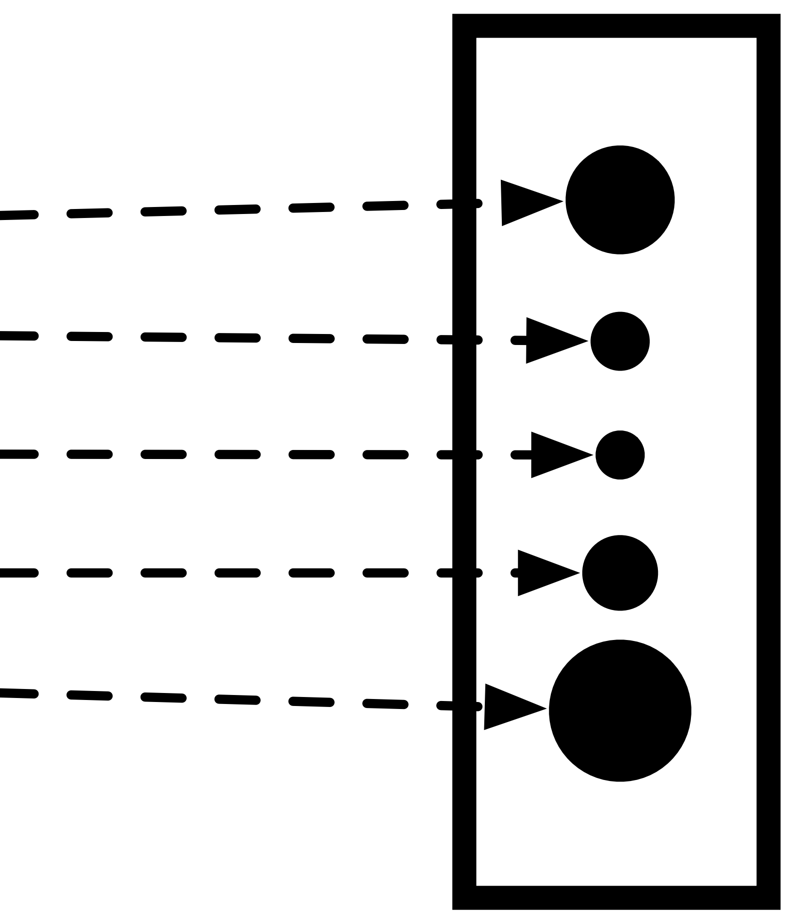
$\mathcal{X}$	State space
$x_t \in \mathcal{X}$	Point in that space
	Subscript: process index
$\mathbf{x}_t$	Many points in the state space
$\mathbf{y}_t$	Many observations



Standard SMC

Output: competing ‘hypotheses’ $\mathbf{x}_t^i$

$t =$ last time observed



Hypothesis
$i = 1$

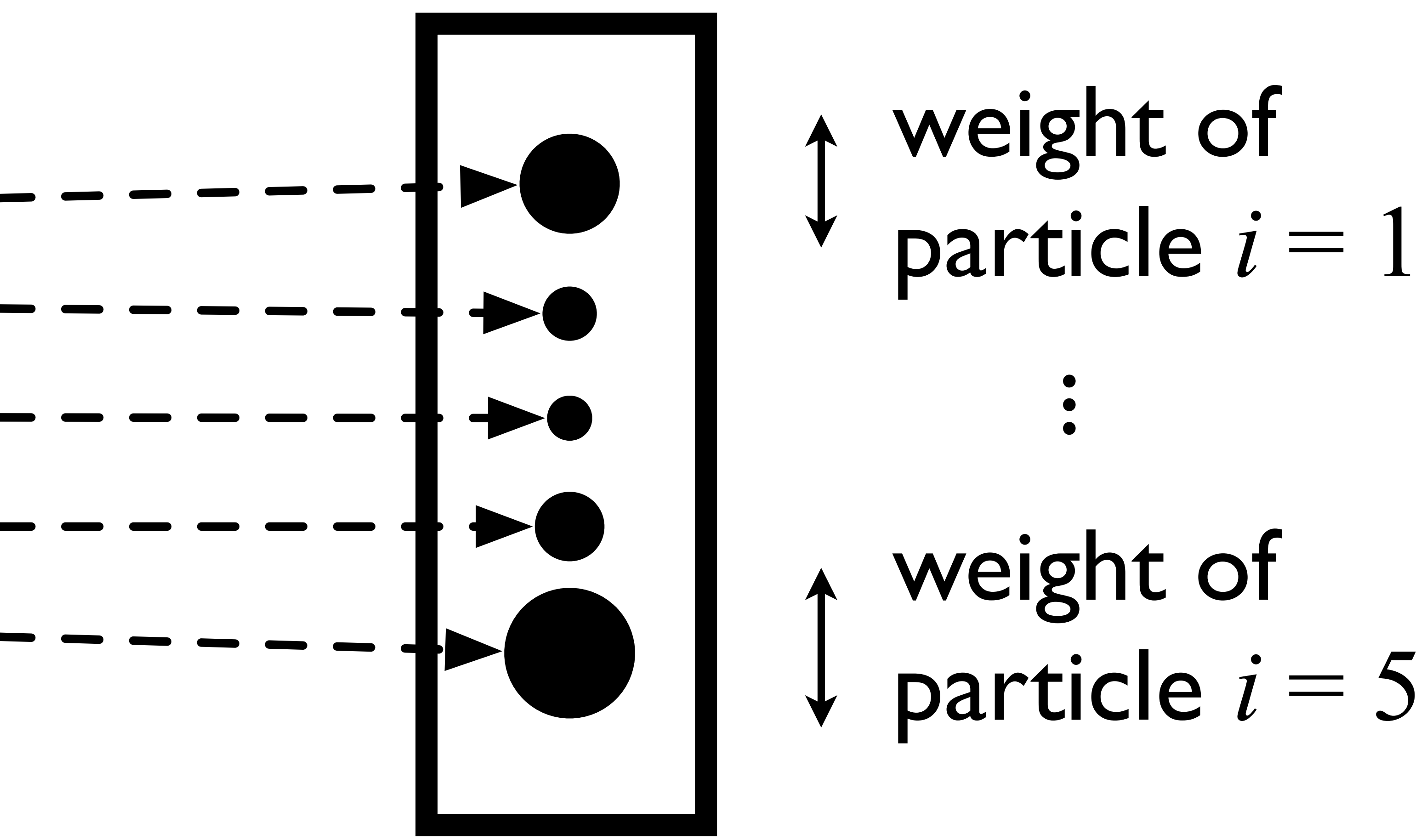
$\vdots$

Hypothesis
$i = 5$

Standard SMC

Output: competing 'hypotheses' $\mathbf{x}_t^i$

weight for each of these w_t^i

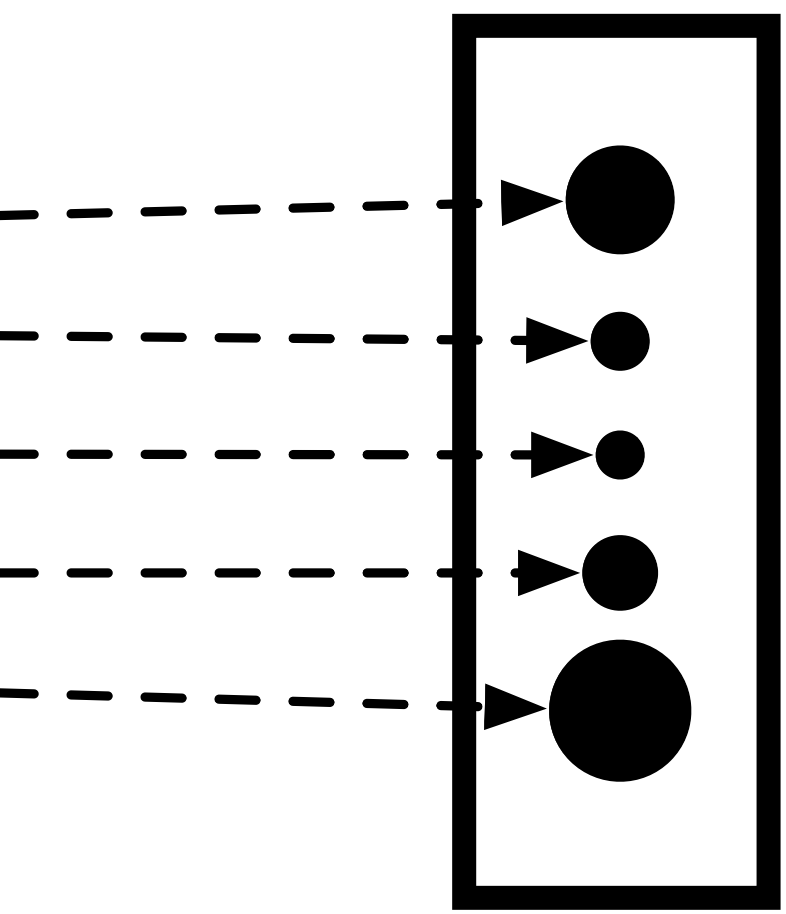


Standard SMC

Output: competing 'hypotheses' $\mathbf{x}_t^i$

weight for each of these w_t^i

Can view these as a (random) distribution

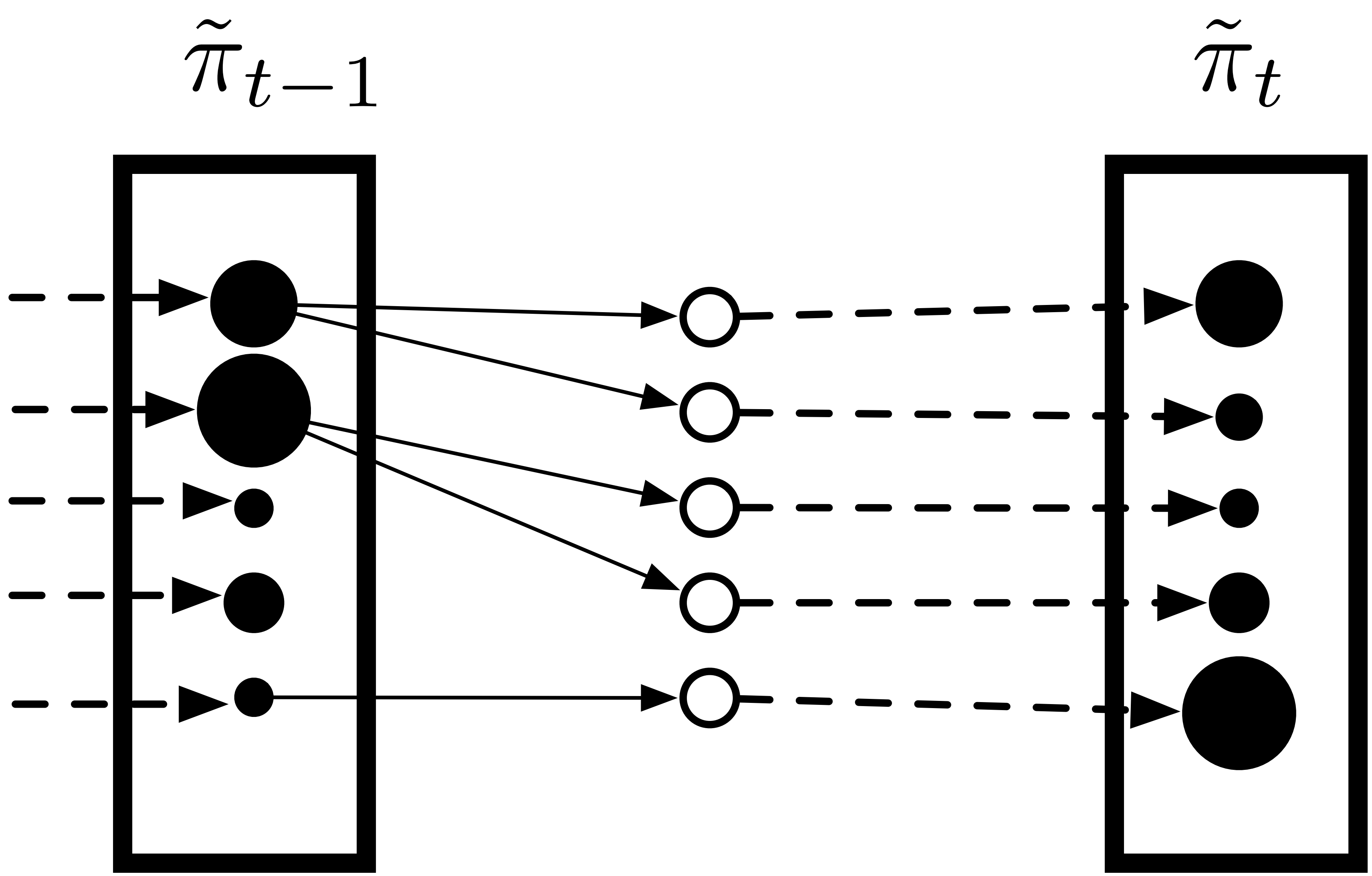


$$\tilde{w}_t^i = \frac{w_t^i}{\sum_j w_t^j}$$

$$\tilde{\pi}_t(\cdot) = \sum_i \tilde{w}_t^i \delta_{\mathbf{x}_t^i}(\cdot)$$

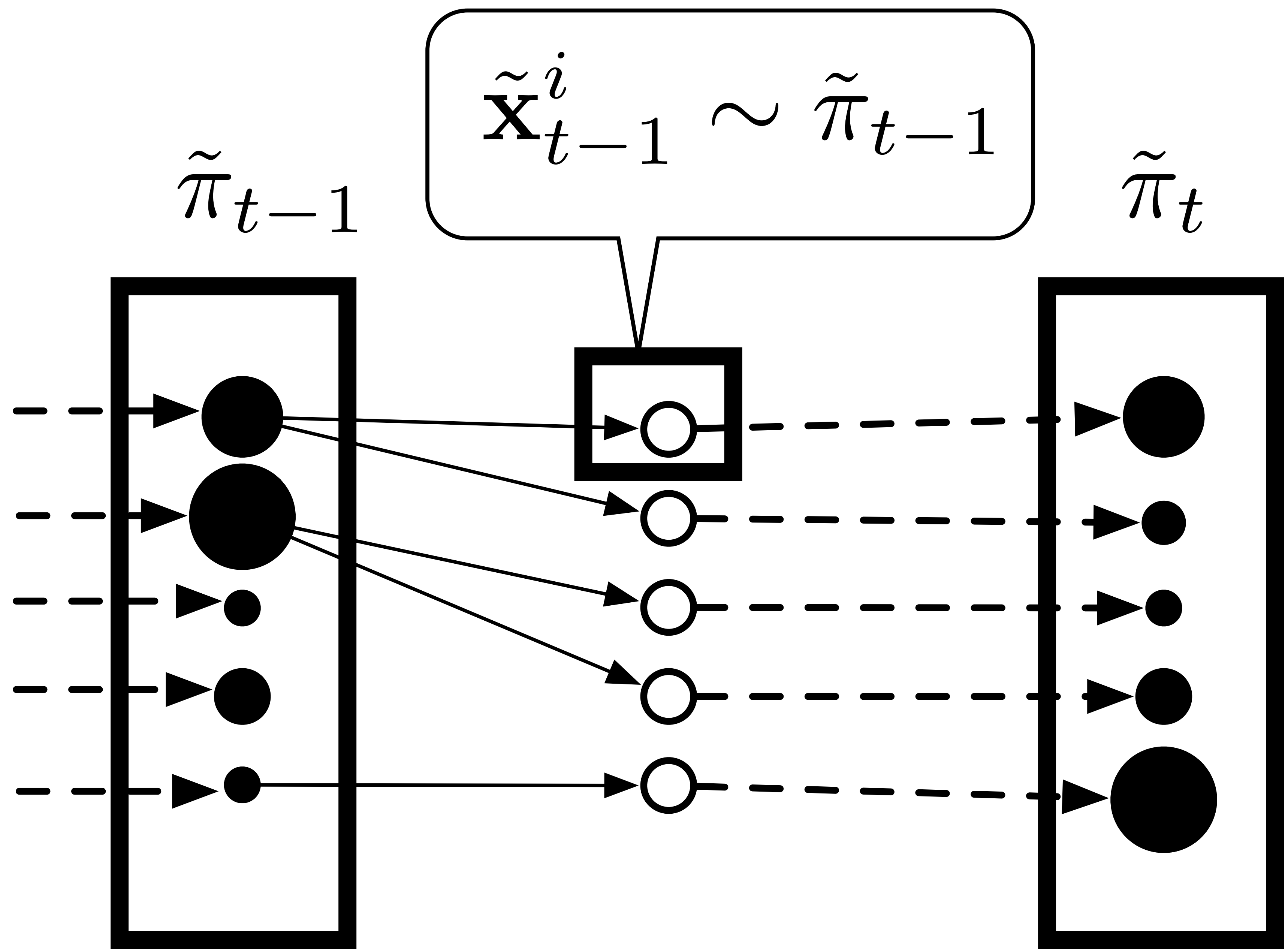
Standard SMC inner working: 1. Assume inductively that we have computed approximation for:

$$\pi_{t-1}(\mathbf{x}_{t-1}) = p(\mathbf{x}_{t-1}|\mathbf{y}_{t-1})$$



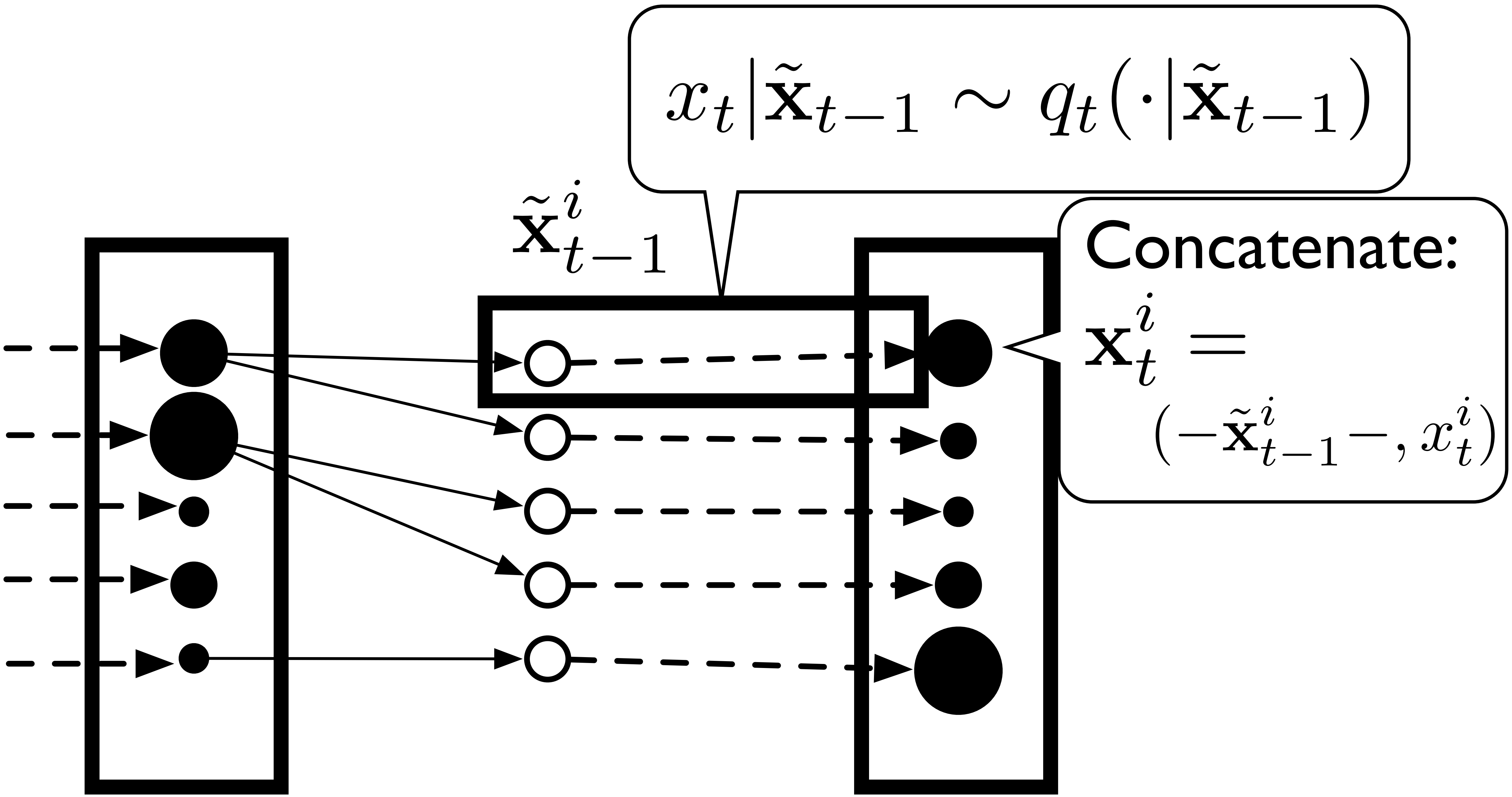
Standard SMC inner working: 1. Assume inductively...

2. Sample from $\tilde{\pi}_{t-1}$



Standard SMC inner working:

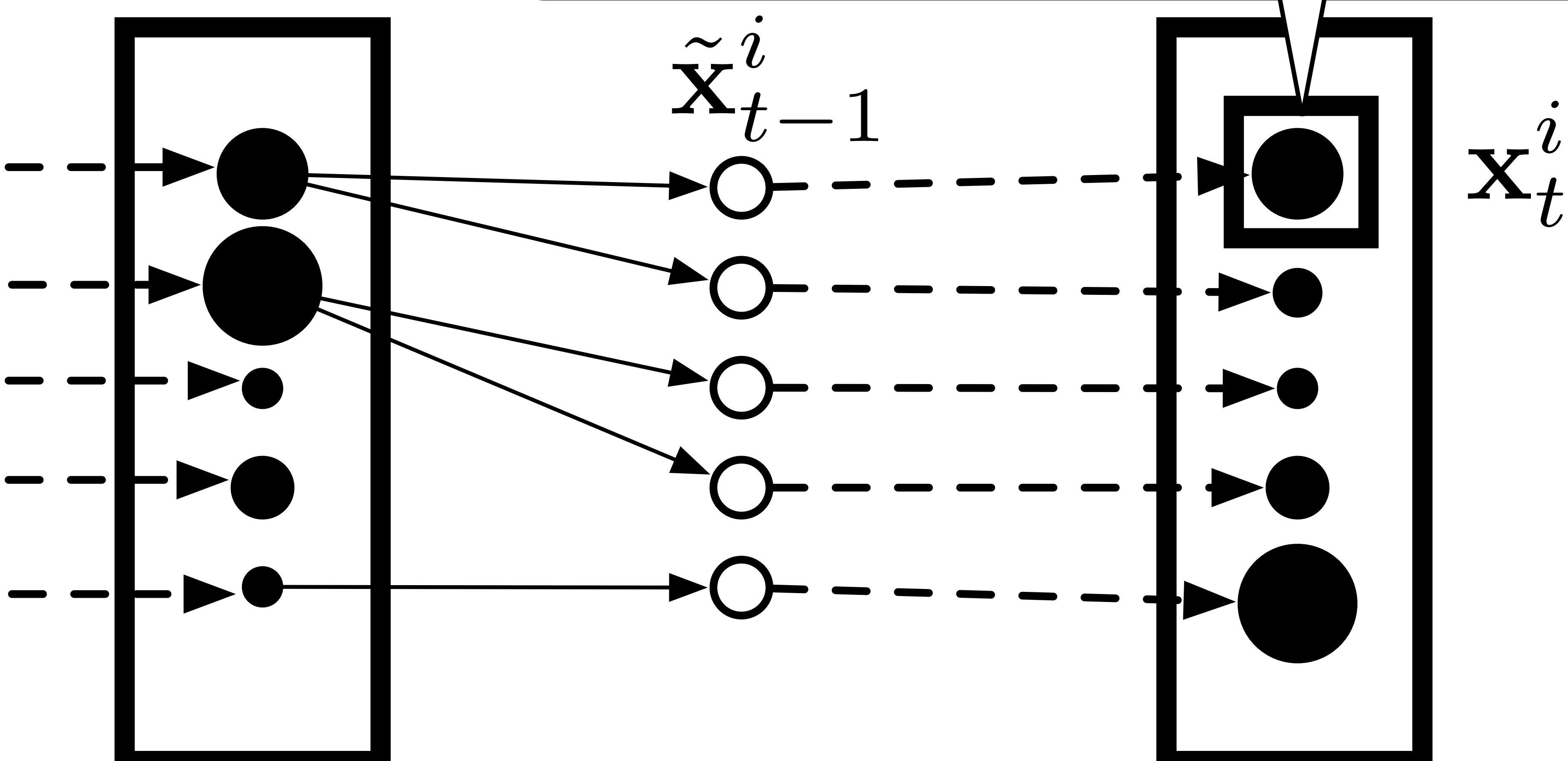
1. Assume inductively...
2. Sample from $\tilde{\pi}_{t-1}$
3. Propose (extend):



Standard SMC inner working:

1. Assume inductively...
2. Sample from $\tilde{\pi}_{t-1}$
3. Propose (extend)
4. Reweigh:

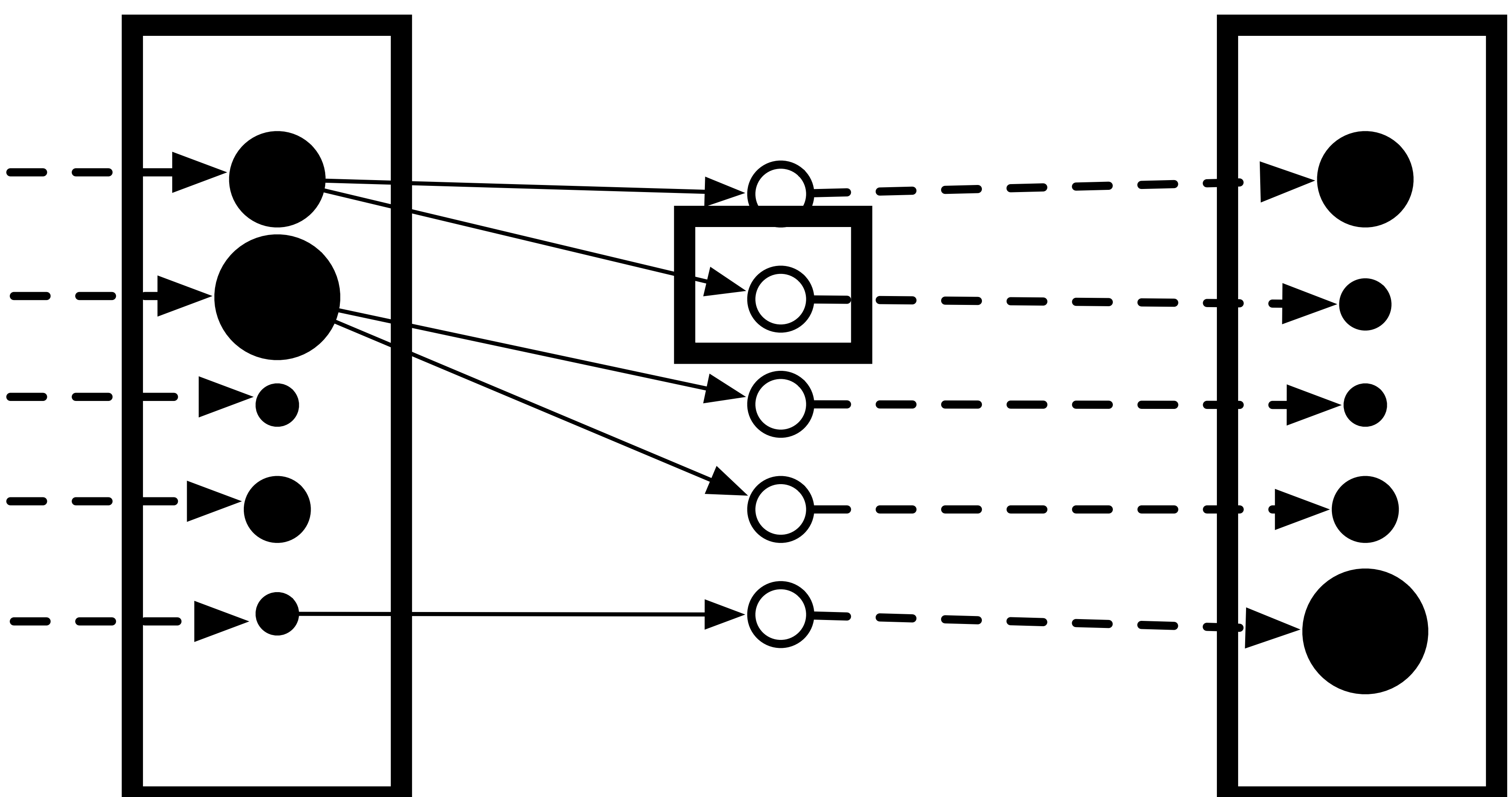
$$w_t^i = \frac{\pi_t(\mathbf{x}_t^i)}{\pi_{t-1}(\tilde{\mathbf{x}}_{t-1}^i)} \frac{1}{q_t(x_t^i|\tilde{\mathbf{x}}_{t-1}^i)}$$



Standard SMC inner working:

1. Assume inductively...
2. Sample from $\tilde{\pi}_{t-1}$
3. Propose (extend)
4. Reweigh

Repeat for each particle (5 times)



Some pointers

- Theory: see Del Moral, 2013 for LLN, CLT
- How to build MC intervals: see J. Olsson, R. Douc (2018)
- Proposals:
  - sometimes, forced to pick dynamics
  - else, various options, e.g. *lookahead proposal*

Resampling

- Efficient implementation
  - Poisson process trick, see March 7
- Often important not perform resampling at every step
  - Monitor relative ESS (March 12) after each proposal round
    $$\frac{(E_q[\tilde{W}])^2}{E_q[\tilde{W}^2]} \approx \frac{(\frac{1}{n}\sum \tilde{W}^{(i)})^2}{\frac{1}{n}\sum(\tilde{W}^{(i)})^2}$$
  - Resample when it drops under a threshold (0.5) typically
- Finally, alternatives to multinomial resampling exist, see Mathieu Gerber, Nicolas Chopin, Nick Whiteley, 2017 for recent analysis of those

Organization

- SMC on product spaces
- **Transforming other problems into product spaces (sequential change of measure)**

AIS / Jarzynski's trick

- Target spaces F_t, not product spaces,
    - important e.g. $F_t = S$ (change of measure)
- Auxiliary spaces:

$$S_{1:n} = S \times S \times \ldots S$$

- Distribution on those? Use a *backward kernel B*

$$\pi_{1:n}(x_{1:n}) = \pi_n(x_n) \prod_{m<n} B_m(x_m|x_{m+1})$$

- Get weight update:

$$\tilde{w}(x_{1:n-1}, x_{1:n}) = \frac{\gamma_n(x_n)}{\gamma_{n-1}(x_{n-1})} \frac{B_{n-1}(x_{n-1}|x_n)}{K_n(x_n|x_{n-1})}$$

Example

- Setup: change of measure on annealed distributions
- K_n: π_n invariant kernel (from MH)
- Problem: cannot compute weight in general

$$\tilde{w}(x_{1:n-1}, x_{1:n}) = \frac{\gamma_n(x_n)}{\gamma_{n-1}(x_{n-1})} \frac{B_{n-1}(x_{n-1}|x_n)}{K_n(x_n|x_{n-1})}$$

- Idea: use fact we are free to pick B as we wish; use

$$B_{n-1}(x_{n-1}|x_n) = \frac{\pi_n(x_{n-1})K_n(x_n|x_{n-1})}{\pi_n(x_n)}$$

- Weight update simplifies (check)

$$\tilde{w} = \frac{\gamma_n(x_{n-1})}{\gamma_{n-1}(x_{n-1})}$$