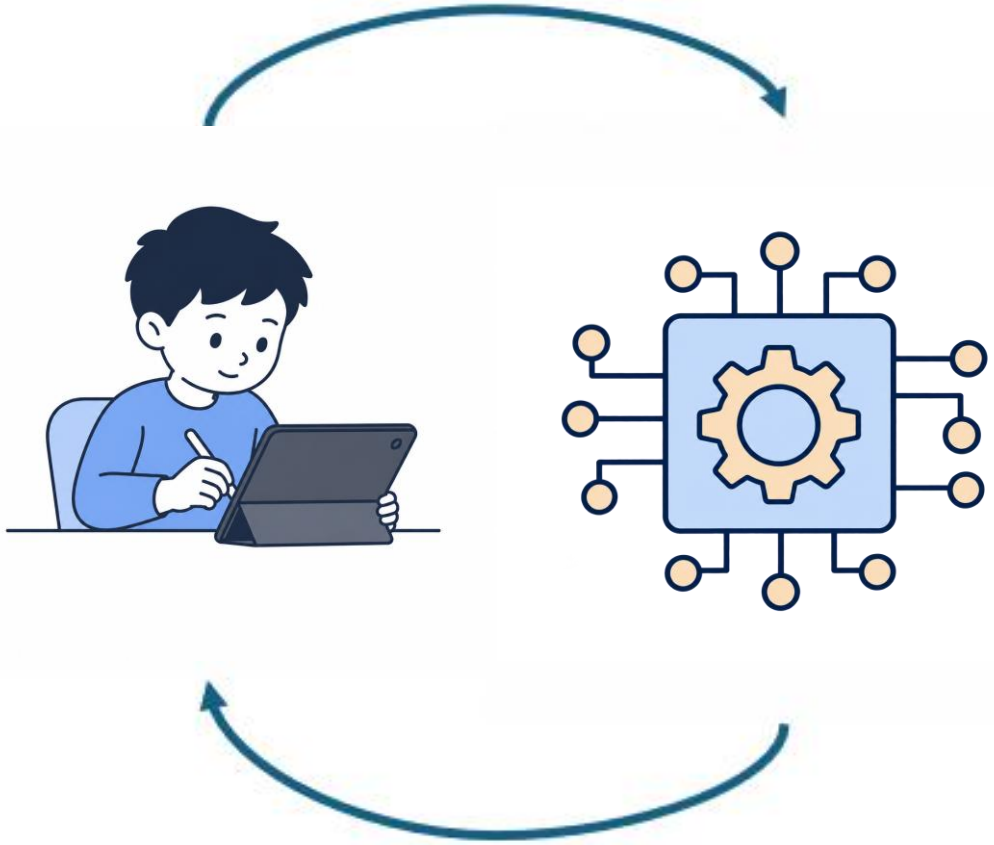
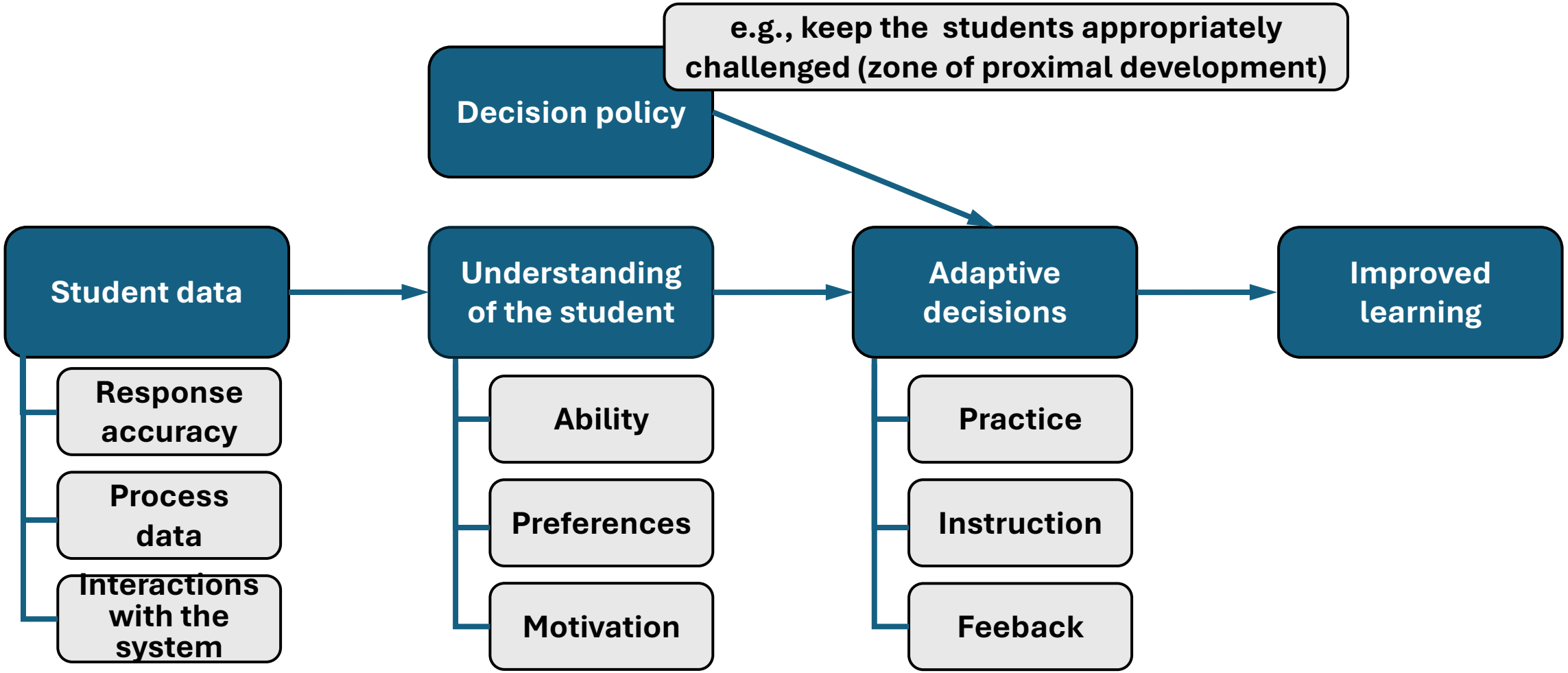
Psychometric algorithms for measurement in adaptive learning environments

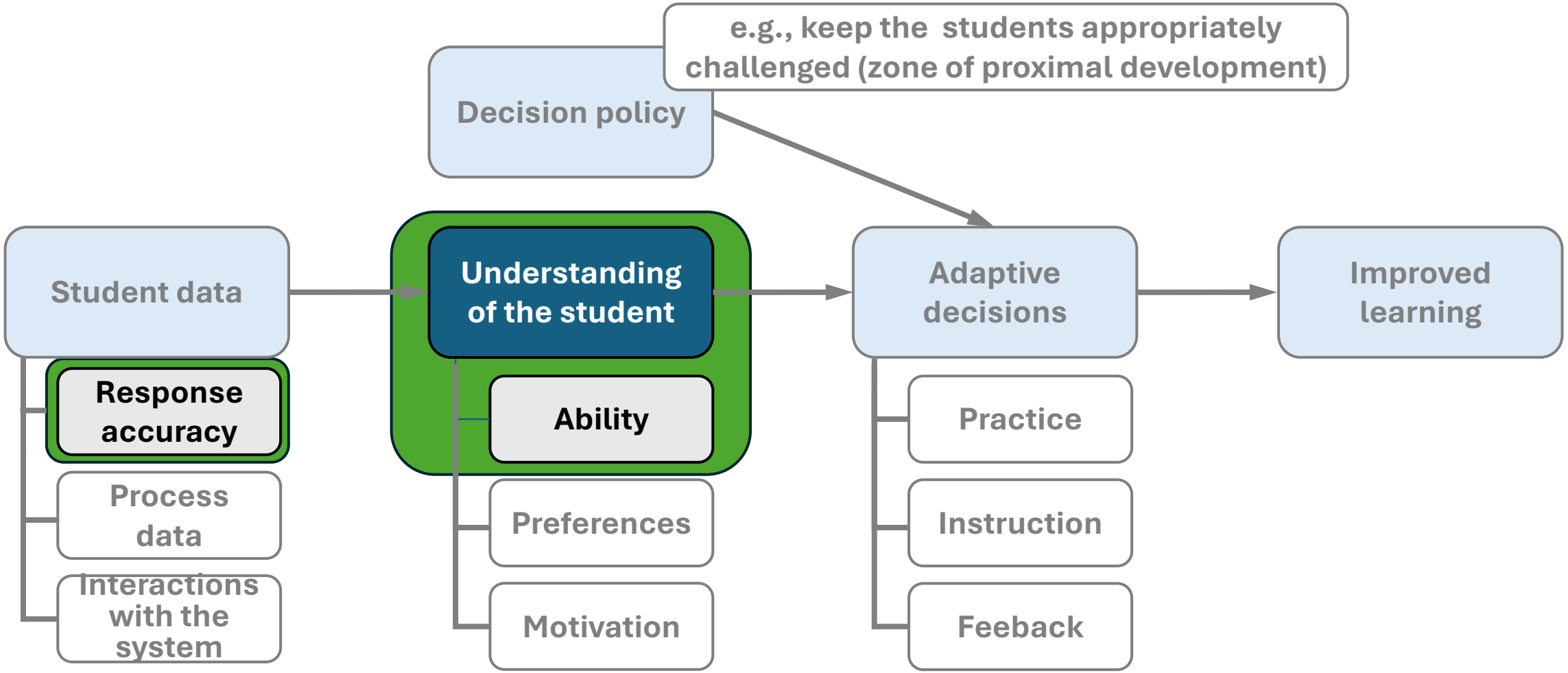
Maria Bolsinova



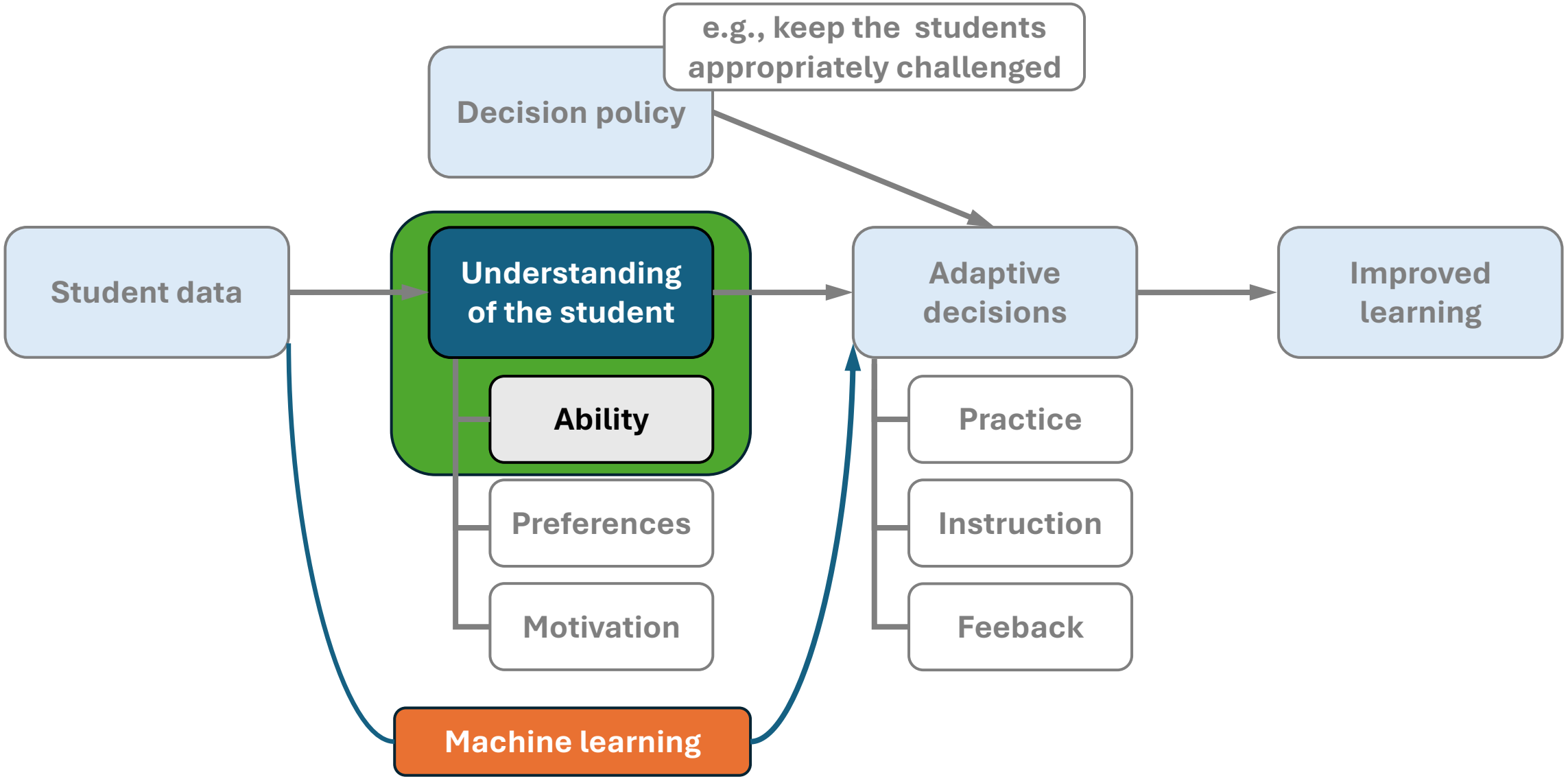
Adaptive learning pipeline



Focus of this presentation



Machine-learning-based approaches



Machine-learning-based approaches

Transparent

Logistic knowledge tracing

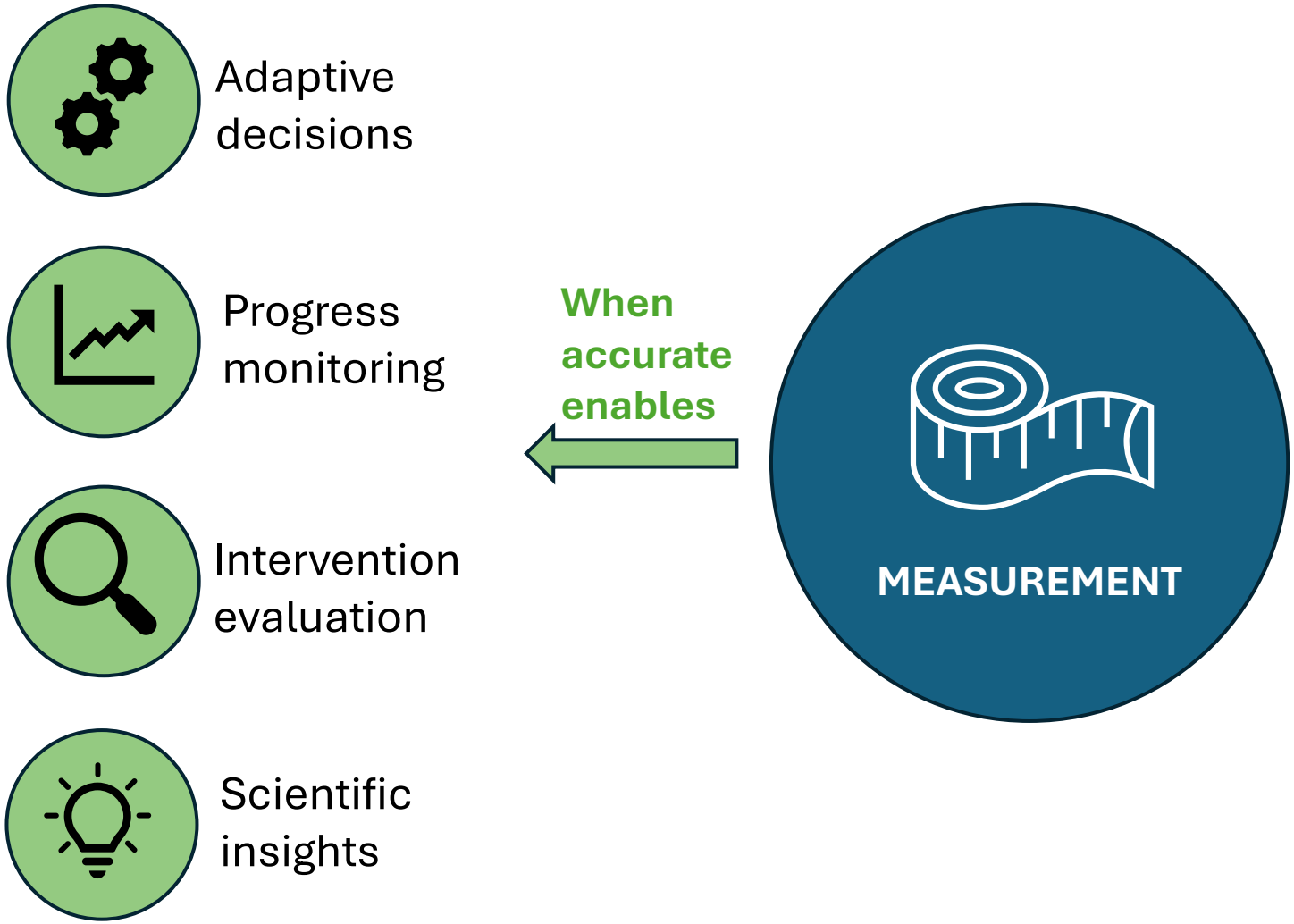
- Summaries of observed data are used to predict the next response (e.g., # previous correct responses related to a skill)
- Interpretable but do not provide measures of changing ability
- Prediction quality often close to black-box models

Black-box

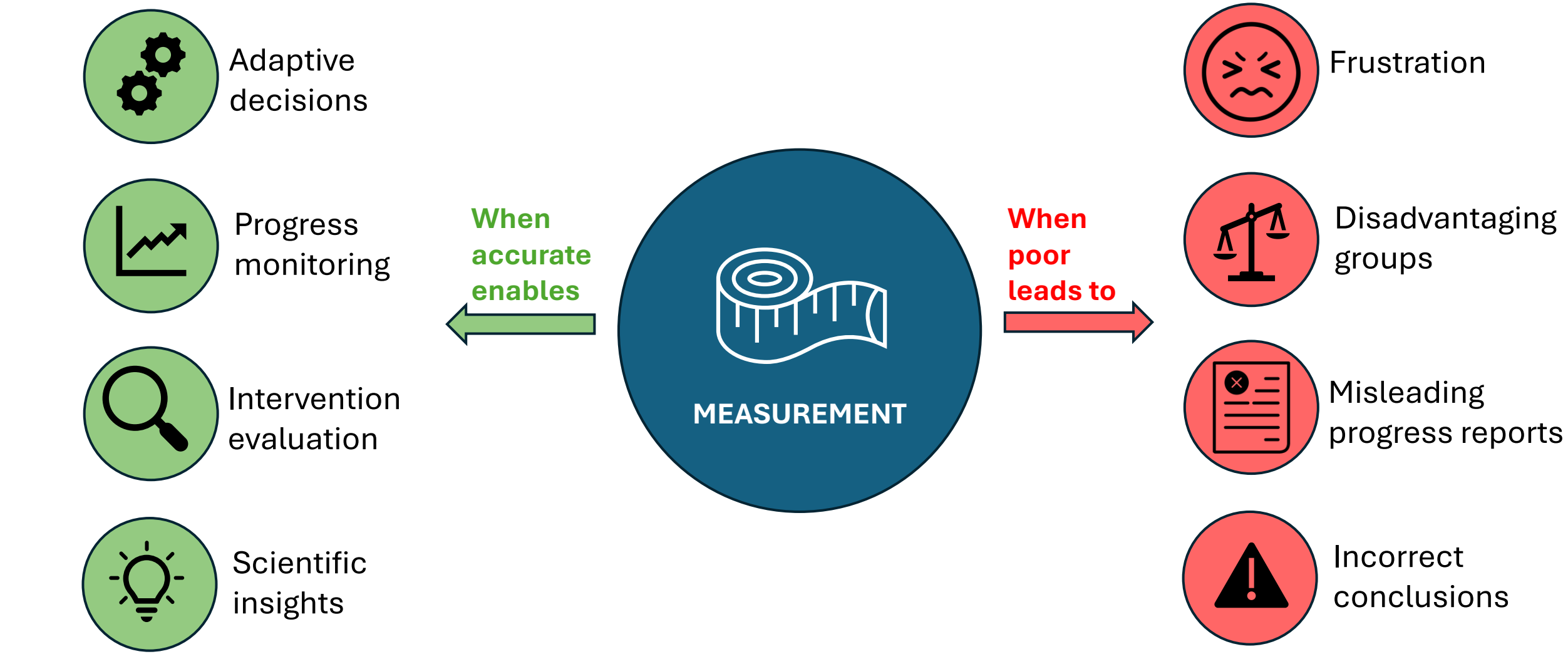
Deep knowledge tracing

- Neural-network based
- High prediction accuracy
- Low interpretability
- “Measures” of ability are sometimes used to make the predictions explainable
  - ? Reliability
  - ? Validity
  - ? Fairness

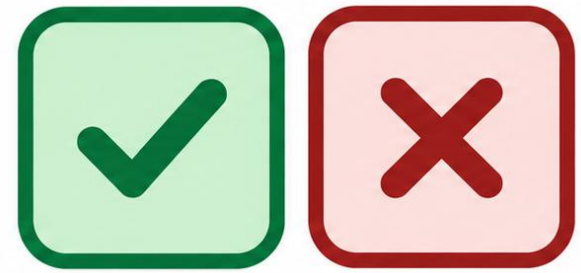
Measurement as the foundation of adaptive learning



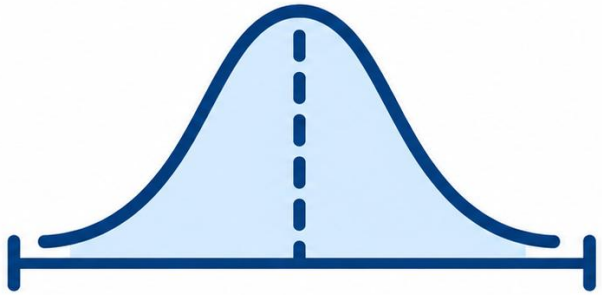
Measurement as the foundation of adaptive learning



Measurement-based approaches



Binary skills



Continuous abilities

Educational data mining tradition → Bayesian knowledge tracing

Psychometric tradition → Dynamic cognitive diagnostic models

(Dynamic) item response theory

Tracing development of binary skills

$\theta_t = \{0,1\}$, binary skill (knowledge component) at timepoint t

Dynamic model is a Markov chain characterized by:

$$p_{initial} = Pr(\theta_1 = 1)$$

$$p_{forgetting} = Pr(\theta_{t+1} = 0 | \theta_t = 1)$$

$$p_{learning} = Pr(\theta_{t+1} = 1 | \theta_t = 0)$$

In BKT $p_{forgetting}$ is often fixed to 0.

Measurement model:

$$Pr(X_{it} = 1 | \theta_t) = (1 - \theta_t) p_{i,guessing} + \theta_t p_{i,slipping}$$

In BKT parameters are often shared by the items

BKT algorithm

Parameters shared by students are estimated in training data, usually with expectation-maximization algorithm

When tracing a student's skill:

- **Step 1:** Update the probability of mastery using the dynamic model (prior probability at timepoint t)
- **Step 2:** Update the probability of mastery given the observed response x={0,1} using the Bayes theorem (posterior at timepoint t):

$$\Pr(\theta_t = 1 | X_t = x) = \frac{\Pr(X_t=x|\theta_t=1)\Pr(\theta_t=1)}{\Pr(X_t=x|\theta_t=1)\Pr(\theta_t=1)+\Pr(X_t=x|\theta_t=0)\Pr(\theta_t=0)}$$

Differences between BKT and dynCDM

BKT: different skills are usually considered **in isolation**

- If an item relates to multiple knowledge components, it is used to update each of these skills separately each time assuming a simple single-skill measurement model

CDM: the skill profile is typically considered **simultaneously**

- When an item measures multiple skills, they all enter a measurement model together via different functions depending on a model (DINA, DINO, etc.)

Tracking development of continuous abilities

Continuous ability θ_t changes over time and observations depend on it through an item response theory (IRT) model

$$\Pr(X_{it} = 1) = \frac{\exp(\theta_t - \delta_i)}{1 + \exp(\theta_t - \delta_i)}$$

δ_i - item difficulty (potentially also time-varying).

Option 1: no explicit dynamic model:

- Elo rating system
- Urnings rating system

Option 2: Explicit dynamic model:

- Dynamic IRT
- Different updating algorithms

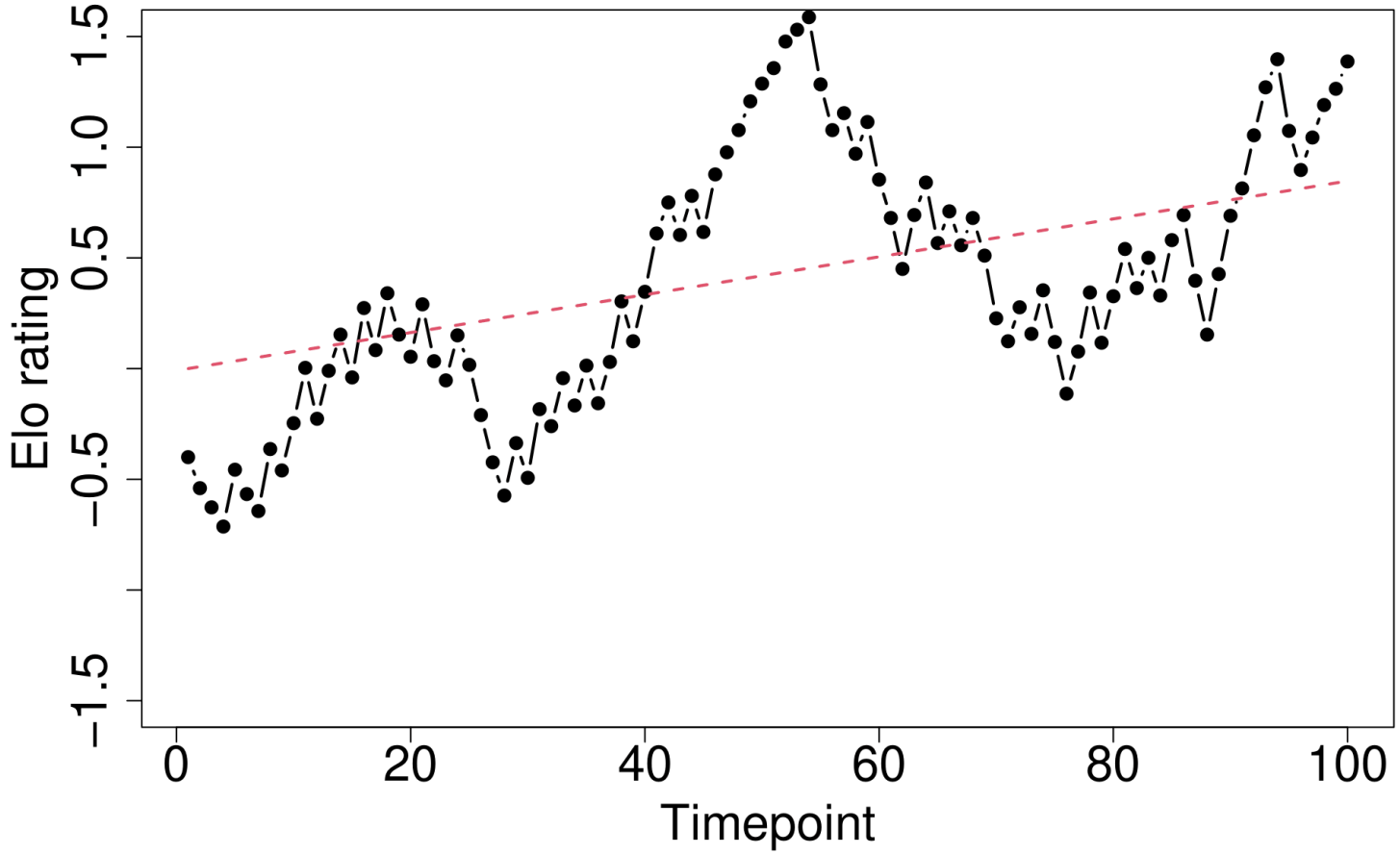
Elo rating system for measurement

After student j ‘competes’ with item i, their ‘ratings’ are updated:

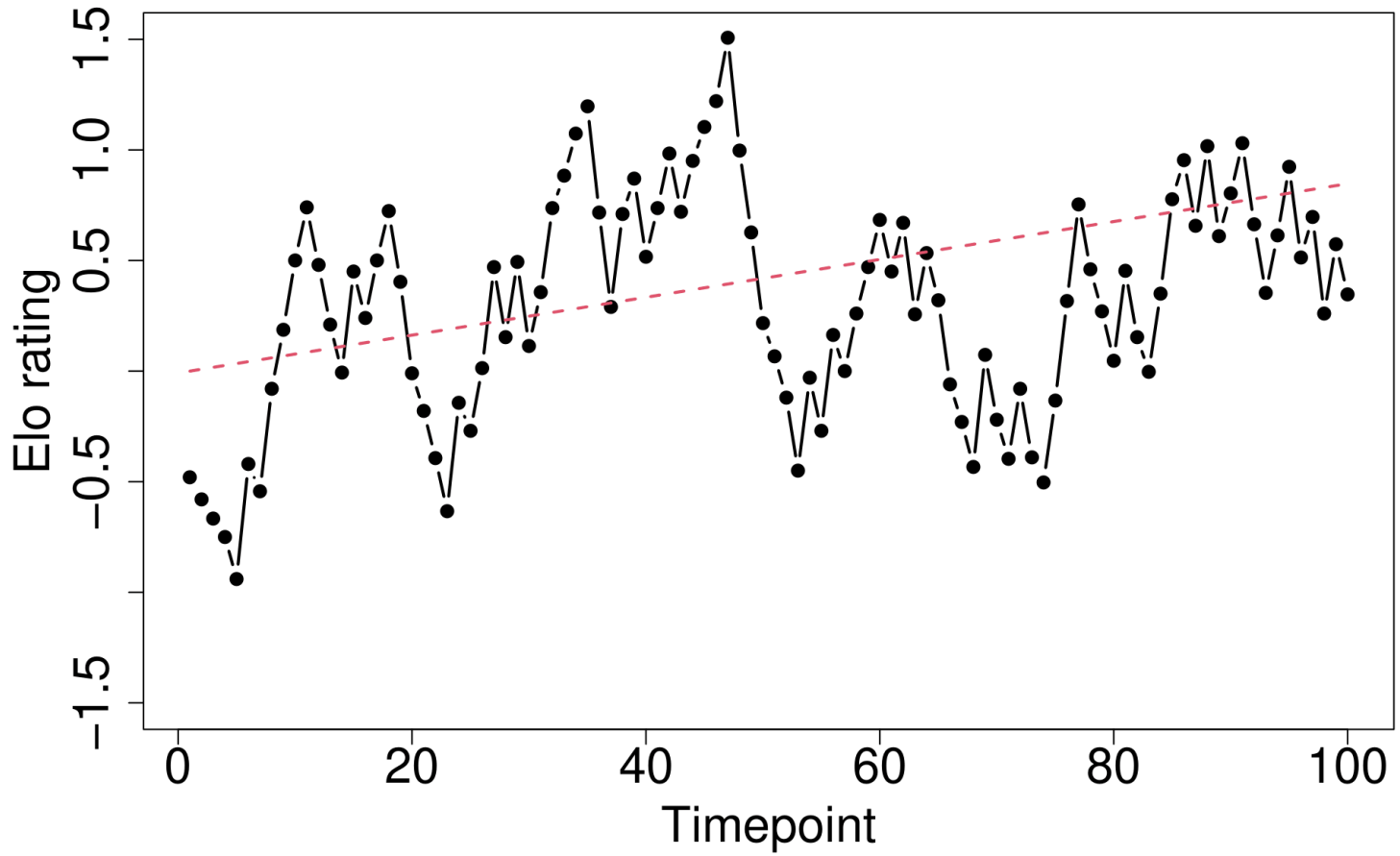
$$\theta_{\mathrm{j,t}} = \theta_{j,(t-1)} + K(\mathrm{X_{ji}} - E(X_{\mathrm{ji}}))$$
$$\delta_{\mathrm{i,t}} = \delta_{\mathrm{i,(t-1)}} - \mathrm{K}(\mathrm{X_{ji}} - \mathrm{E}(\mathrm{X_{ji}}))$$

θ – ability, δ – difficulty, $\mathrm{X_{ji}}$ – observed accuracy, $E(X_{\mathrm{ji}})$ – expected accuracy (Rasch model)

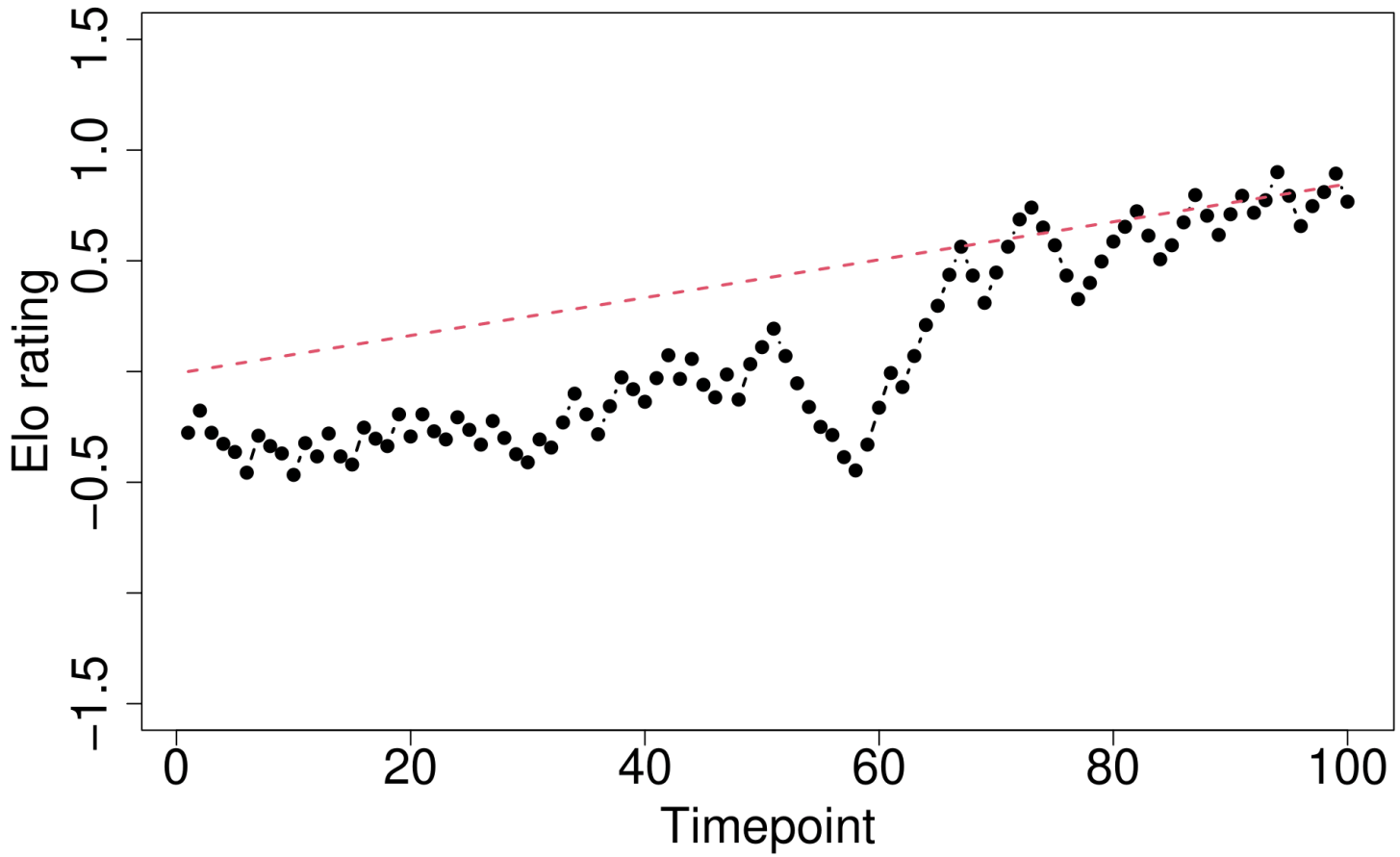
Tracking with Elo: K=0.2



Tracking with Elo: K=0.4



Tracking with Elo: K=0.1



Measurement properties of Elo

How reliable are they?

Are they unbiased?

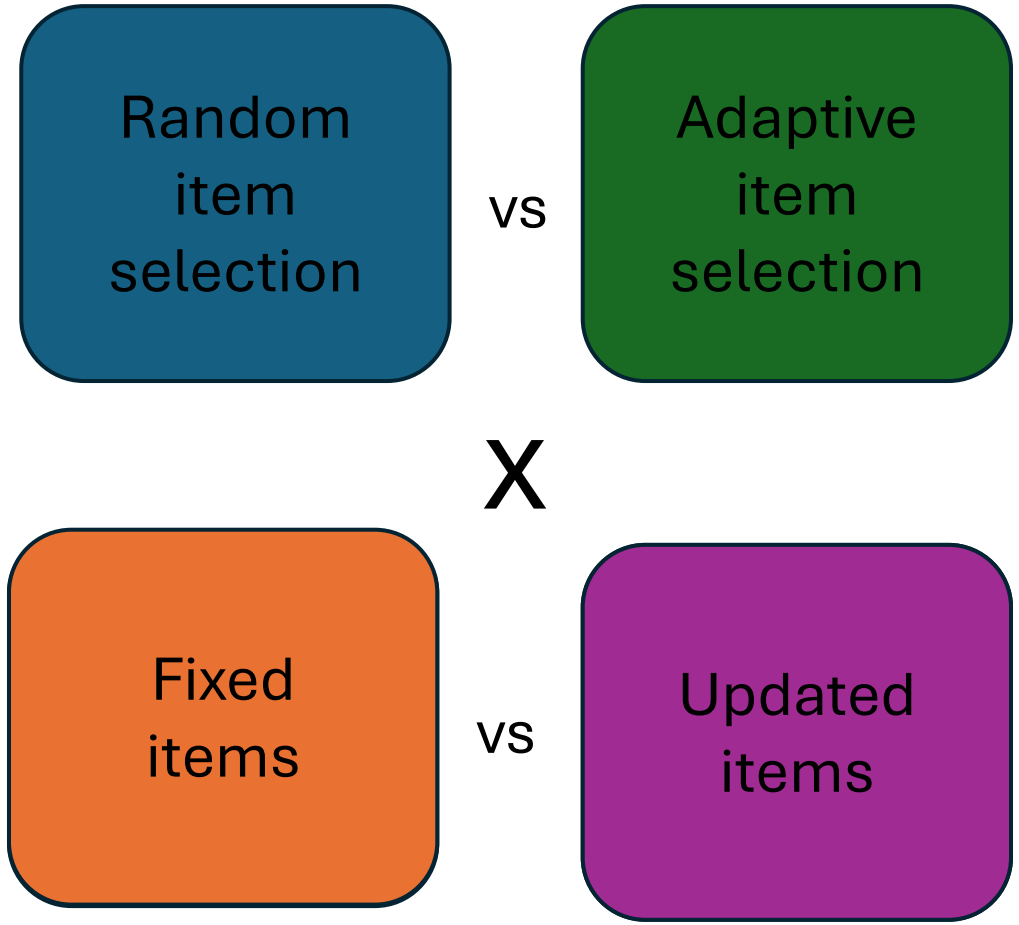
Simulation setup

Stable abilities and difficulties

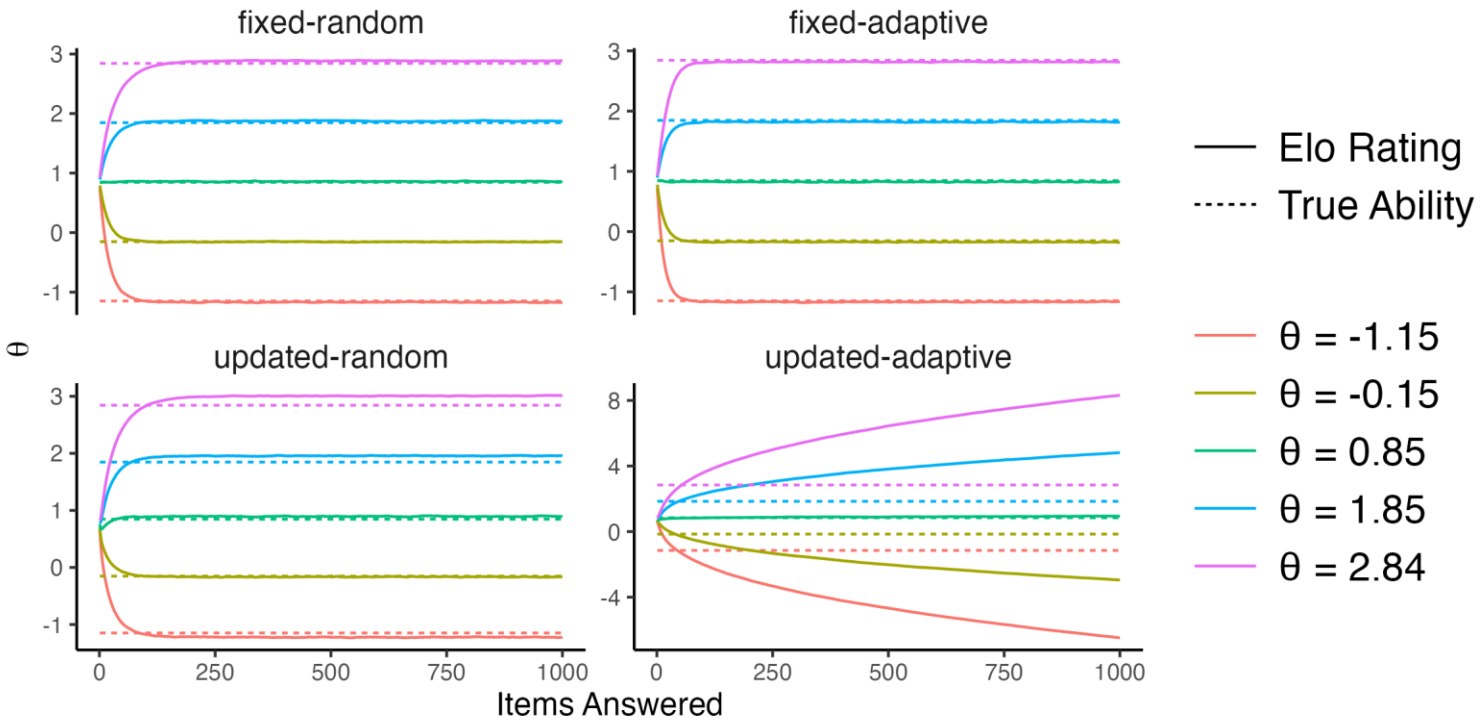
Obtain distributions of the ratings at each timepoint

Do they stabilize? If yes, what is the mean and variance?

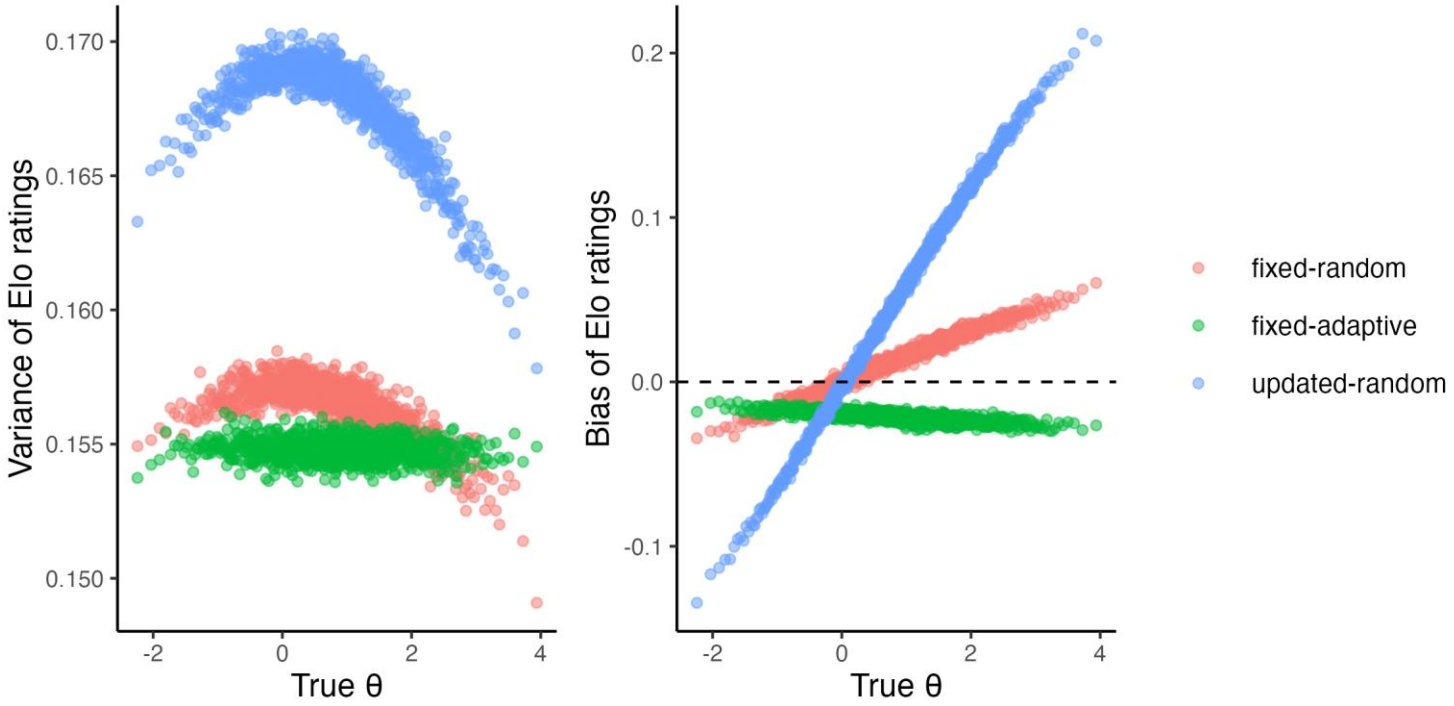
Simulation scenarios



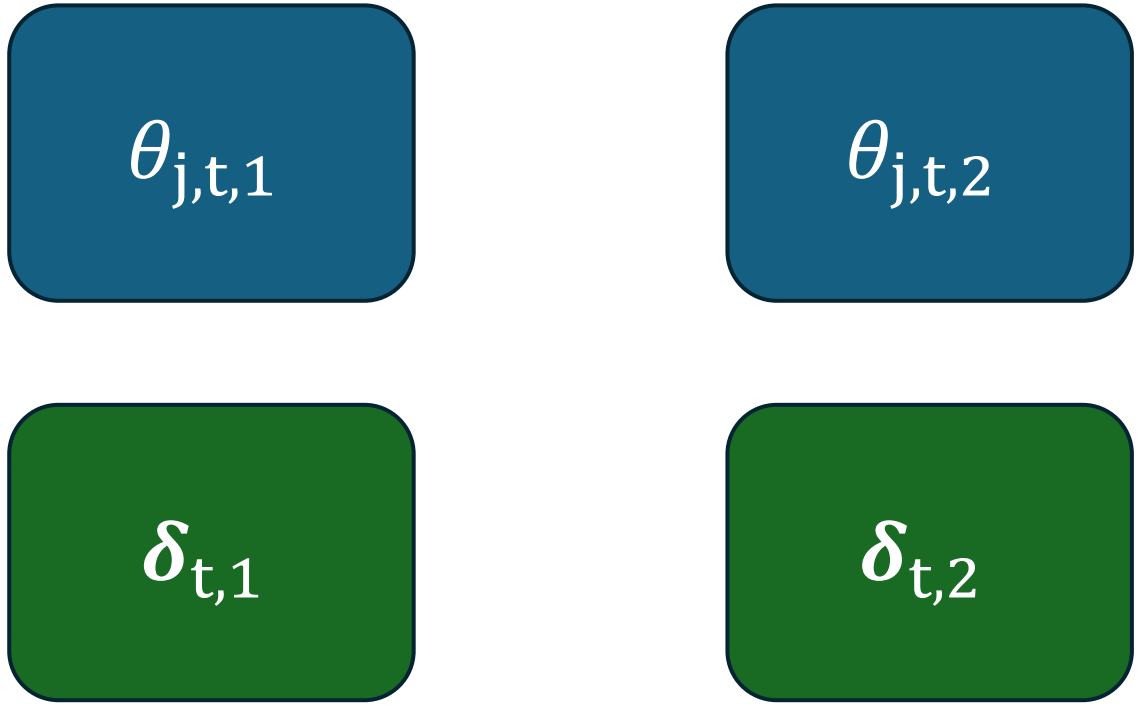
Do Elo ratings converge?



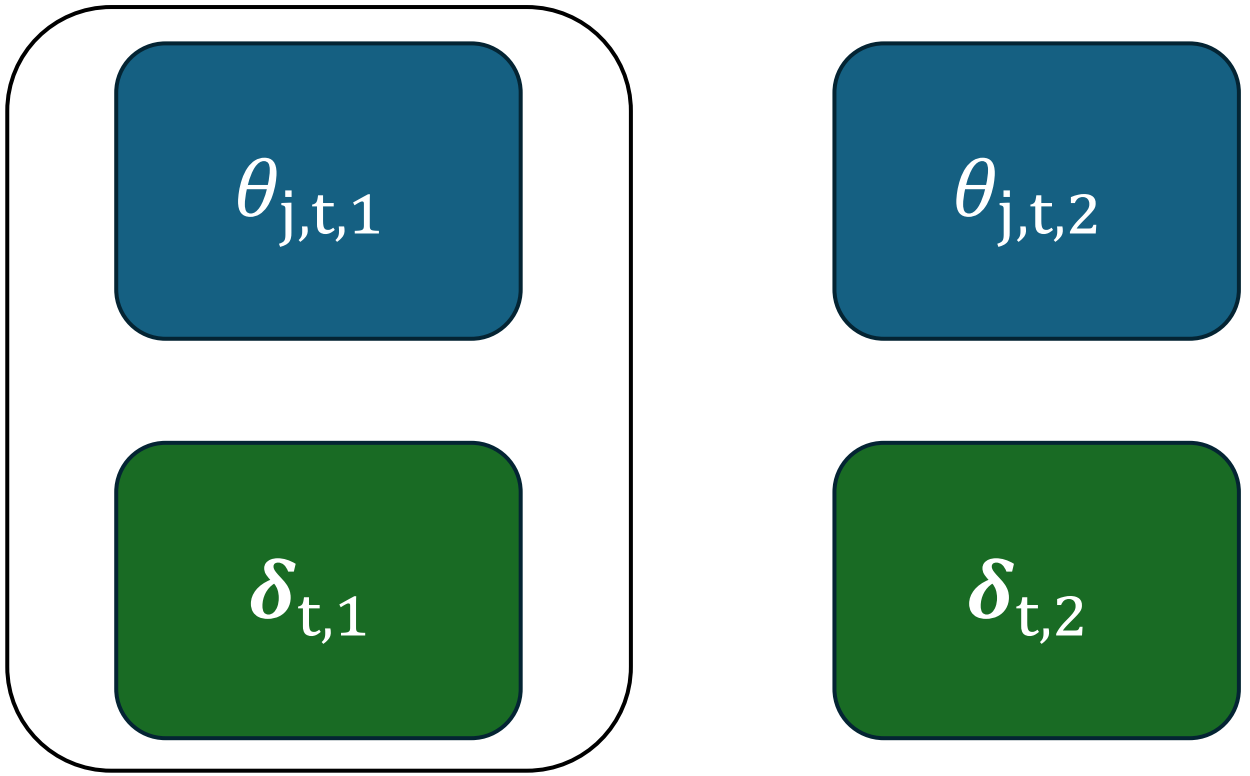
Measurement error and bias



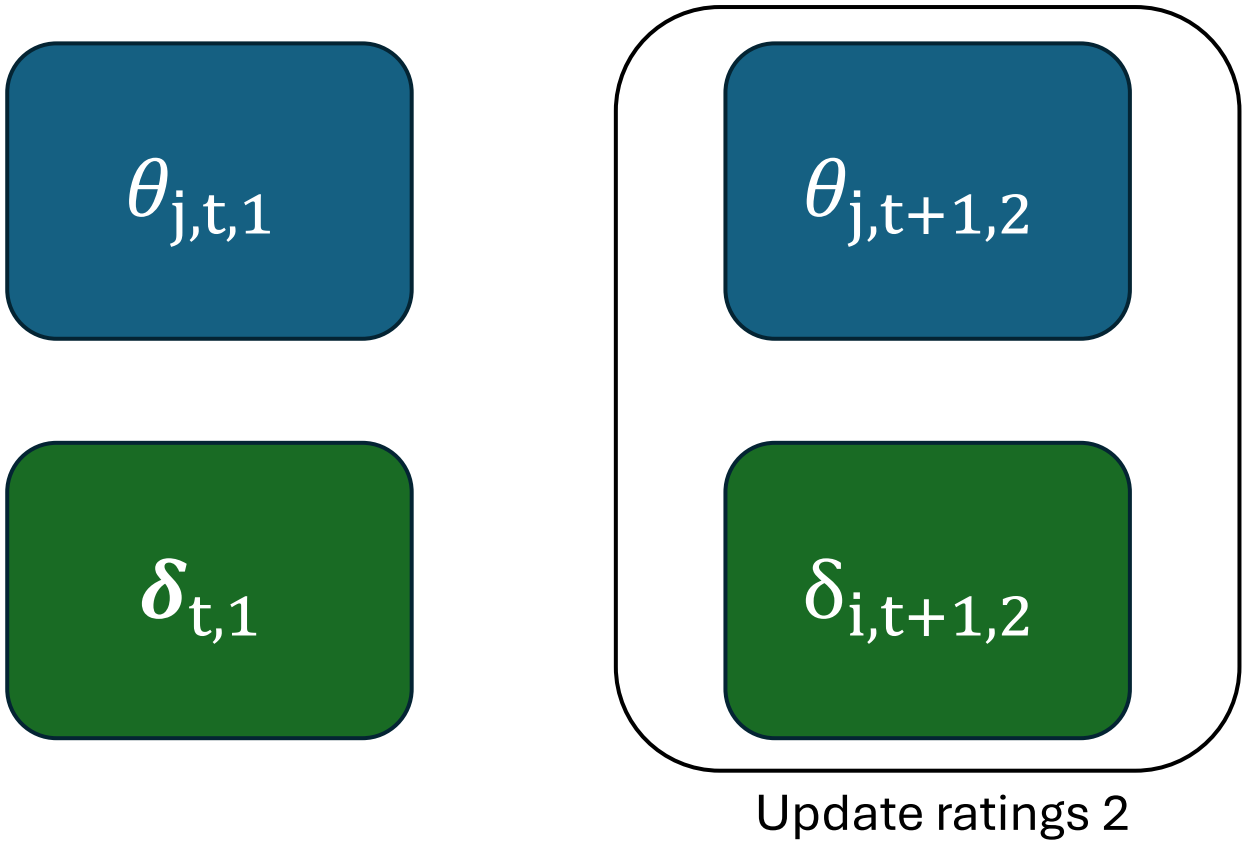
Solution to variance inflation: Parallel Elo



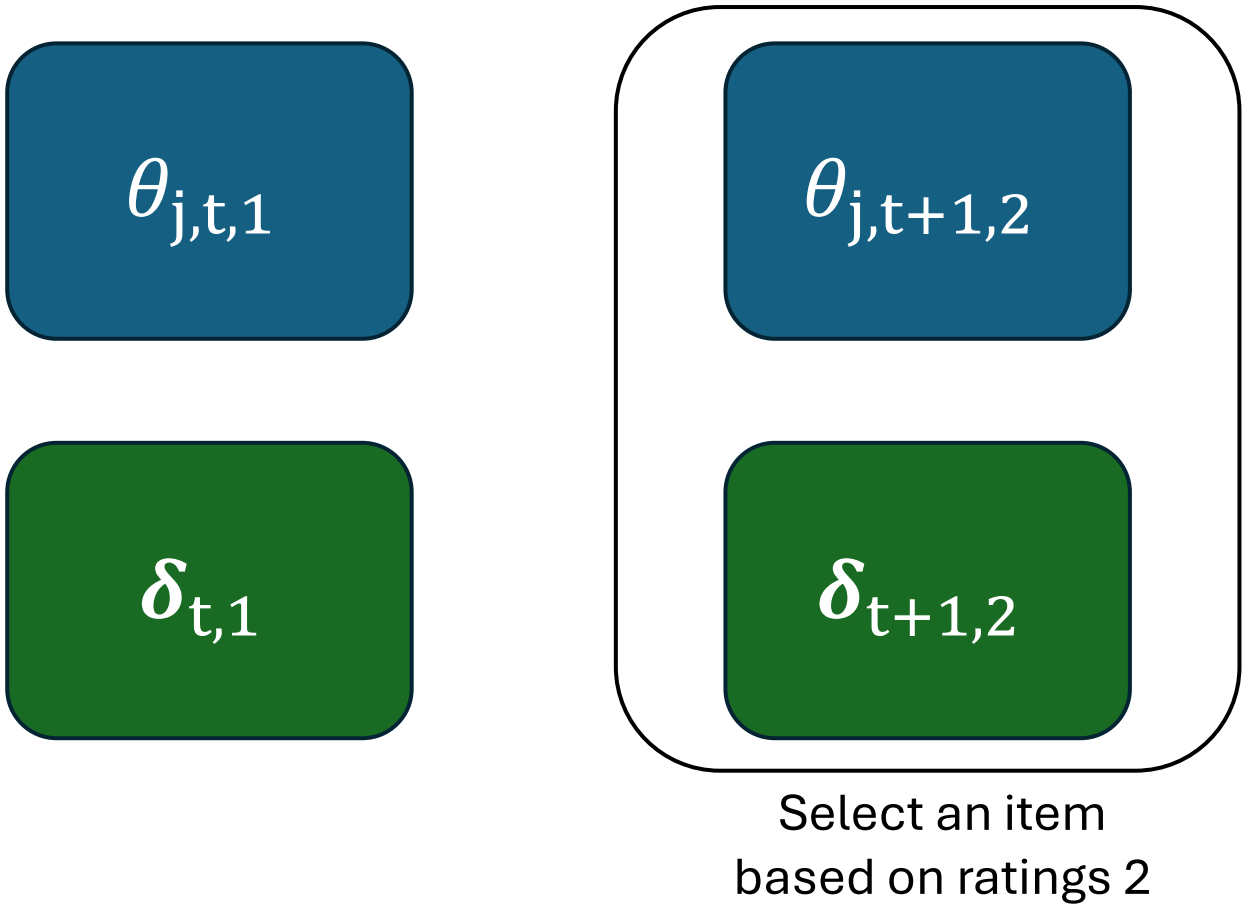
Solution to variance inflation: Parallel Elo



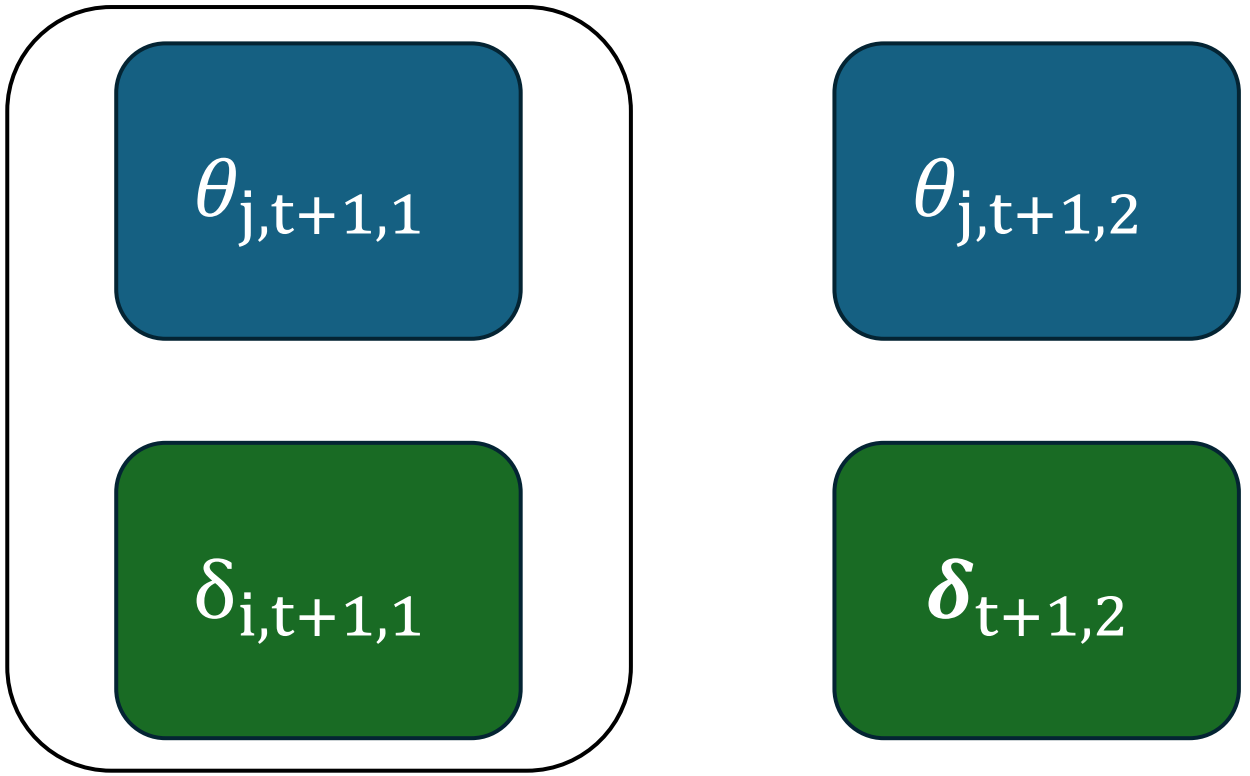
Solution to variance inflation: Parallel Elo



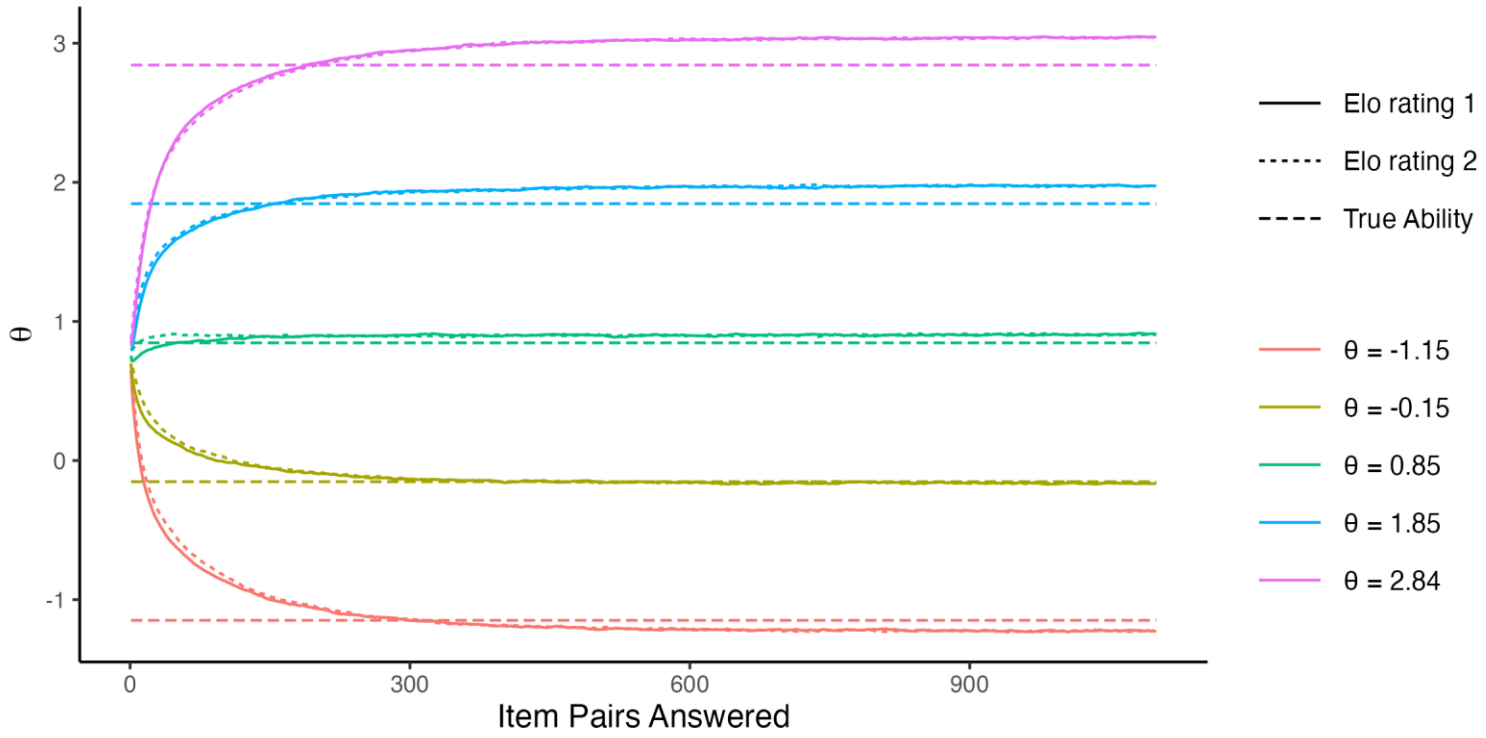
Solution to variance inflation: Parallel Elo



Solution to variance inflation: Parallel Elo



Parallel Elo: Convergence



Alternative to Elo: Urnings

1) Transform the parameters to the [0,1] scale:

$$\pi_j = \frac{\exp(\theta_j)}{1 + \exp(\theta_j)}$$

2) Track the transformed parameters on a discrete grid:

$$\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n\text{-}1}{n}, \frac{n}{n}$$

Urnings rating system for measurement

After student j ‘competes’ with item i, their ‘ratings’ are updated:

$$\frac{u_{j,t}}{n} = \frac{u_{j,t-1}}{n} + \frac{1}{n}(X_{ji} - X^*_{ji})$$
$$\frac{v_{i,t}}{n} = \frac{v_{i,t-1}}{n} - \frac{1}{n}(X_{ji} - X^*_{ji})$$

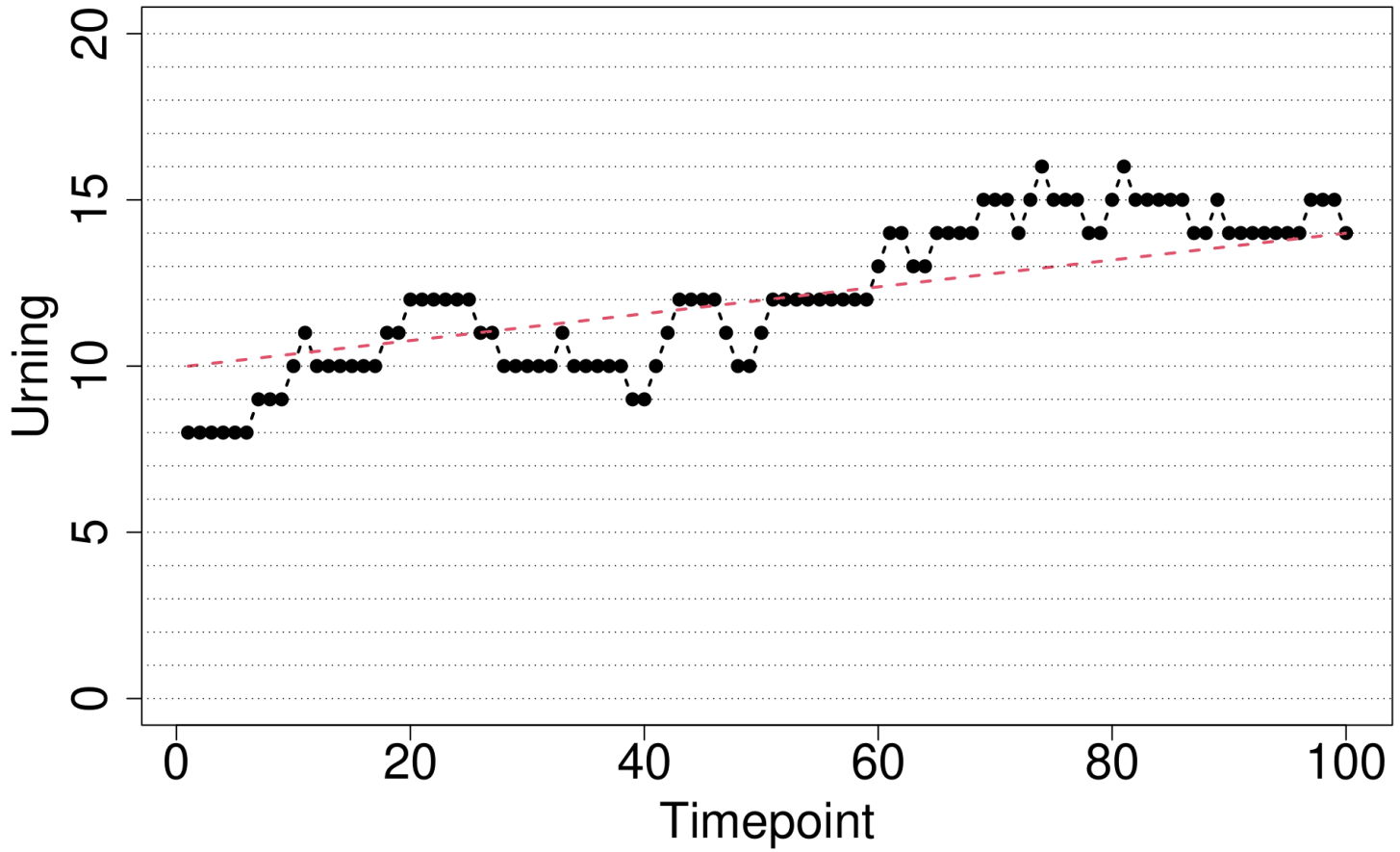
$\frac{u_{j,t}}{n}$ – transformed ability rating,

$\frac{v_{i,t}}{n}$ – transformed difficulty rating, X_{ji} – obseverd response,

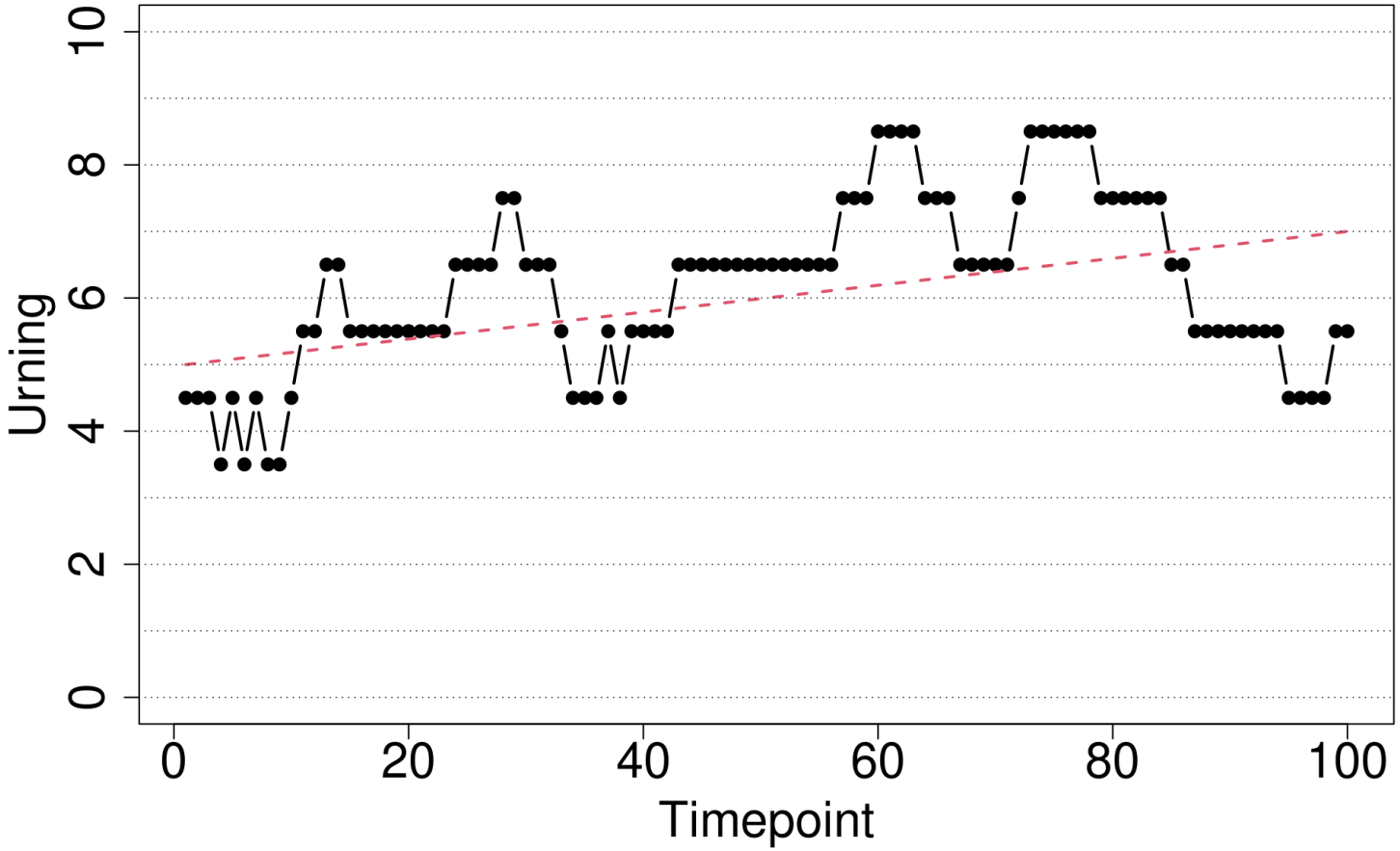
X^*_{ji} – response simulated with augmented ratings $\frac{u_{j,t} + X_{ji}}{n+1}$ and $\frac{v_{i,t} + 1\text{-}X_{ji}}{n+1}$

Adapted item selection is corrected for in a Metropolis-Hastings step

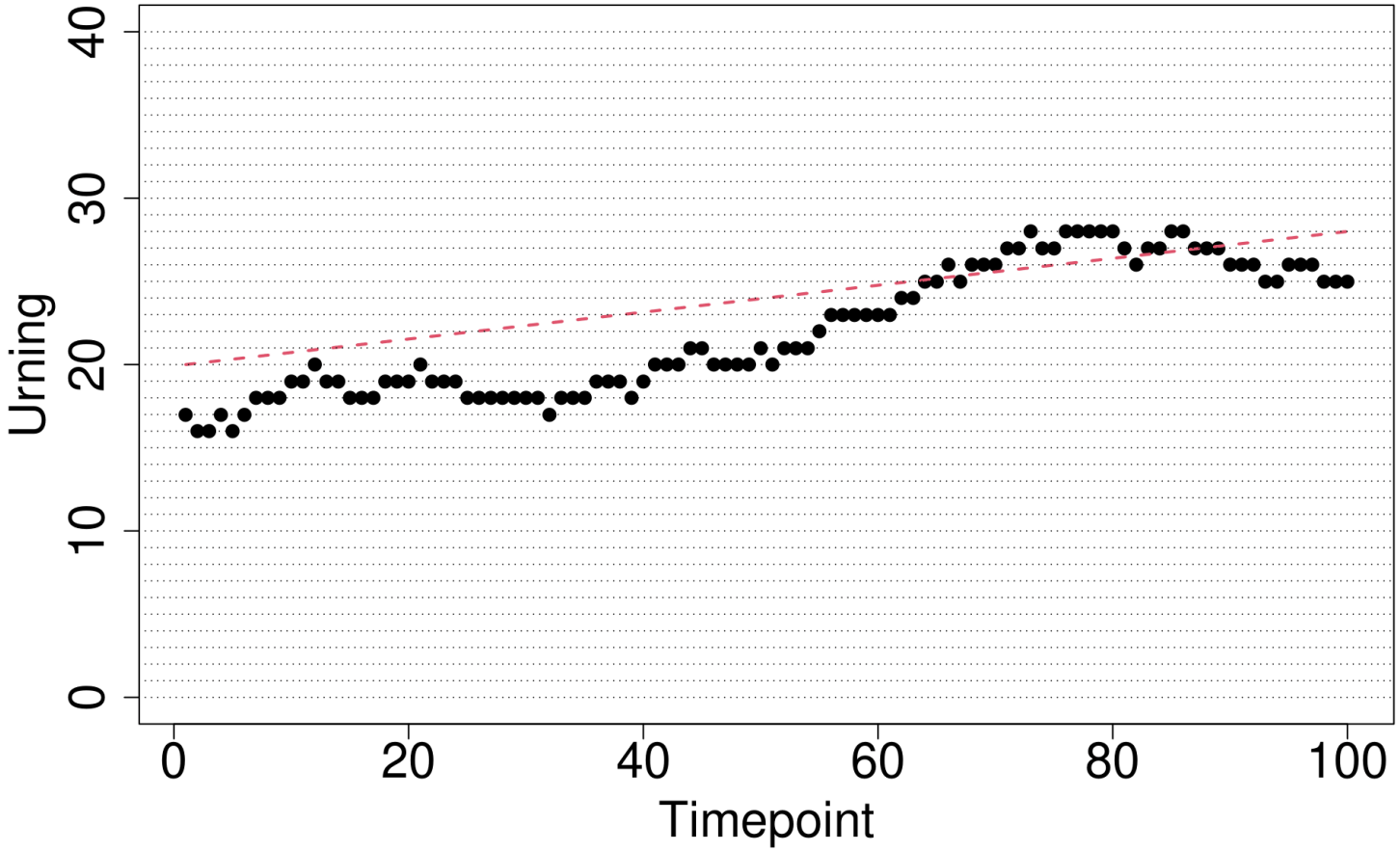
Tracking with urnings: n=20



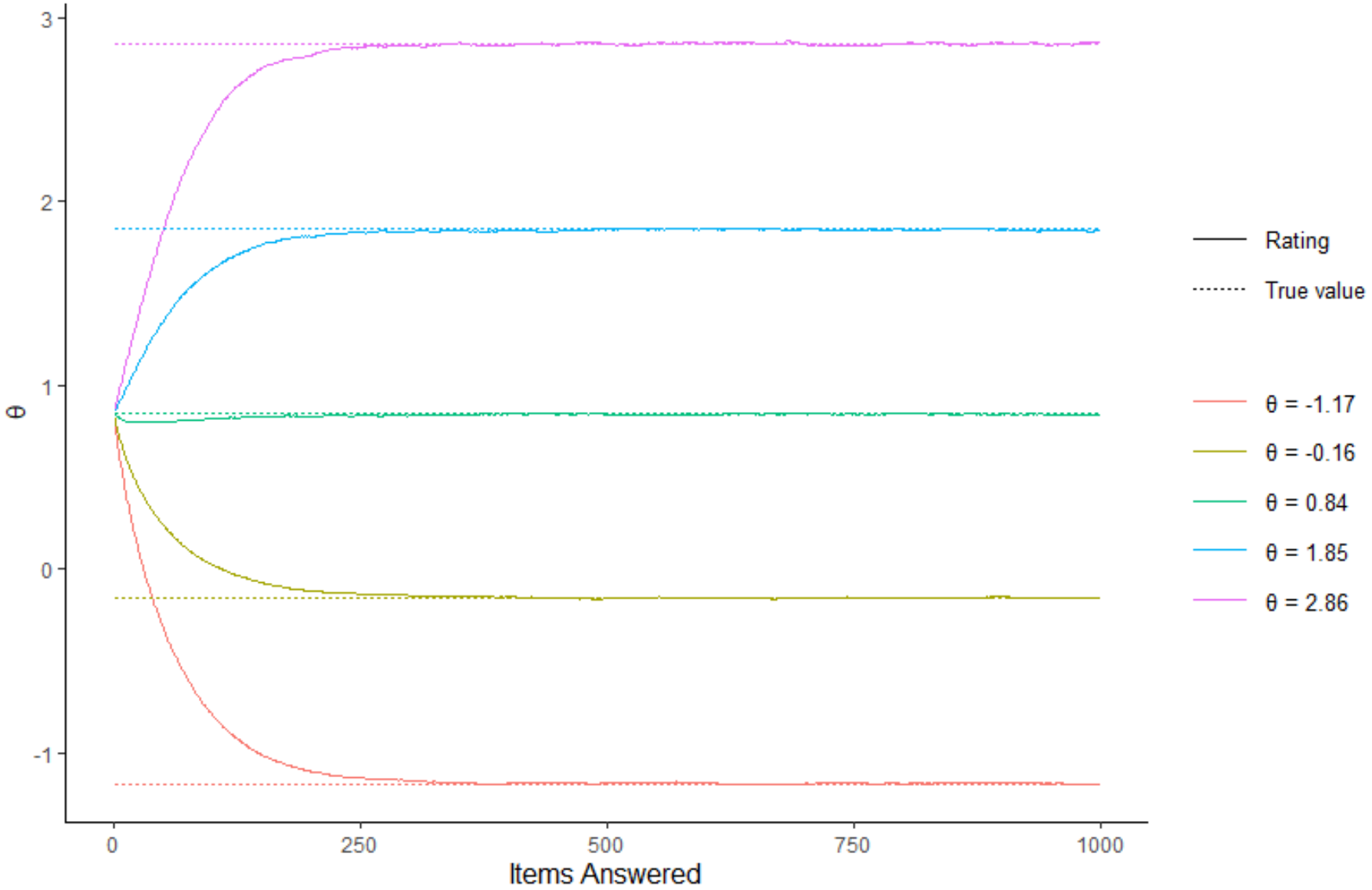
Tracking with urnings: n=10



Tracking with urnings: n=40



Urnings: convergence

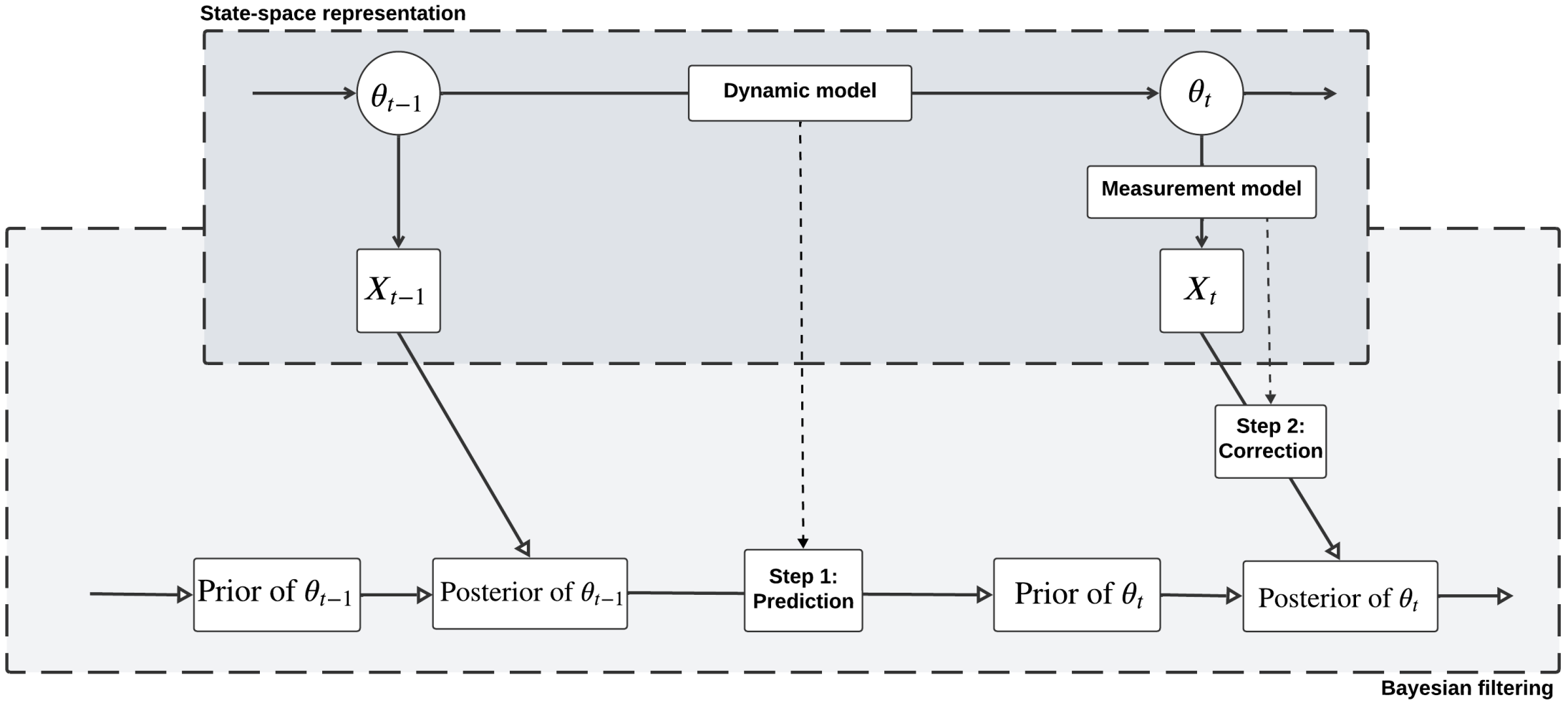


No dynamic model: Strength or missed opportunities?

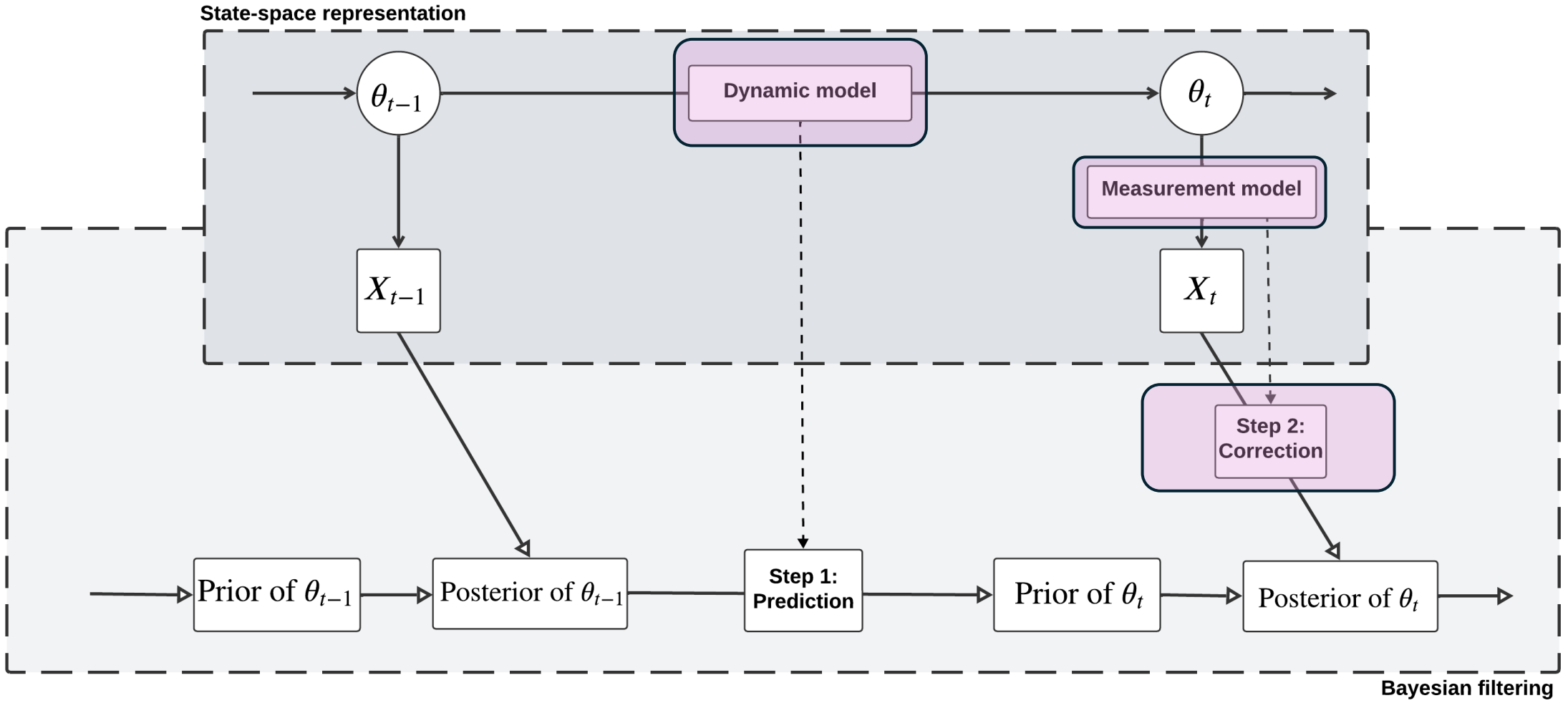
- Specifying a growth model suitable for all student is challenging
- Having no explicit model can provide robustness
- Fixed step size (K, n) or heuristics for changing it play a role of an implicit growth model

- Having an explicit dynamic model allows to adjust step size in a principled way
- Multidimensionality can be leveraged if transfer between skills is modeled
- Additional sources of information can be incorporated in the dynamic model (learning trajectories, seasonal effects, student characteristics)

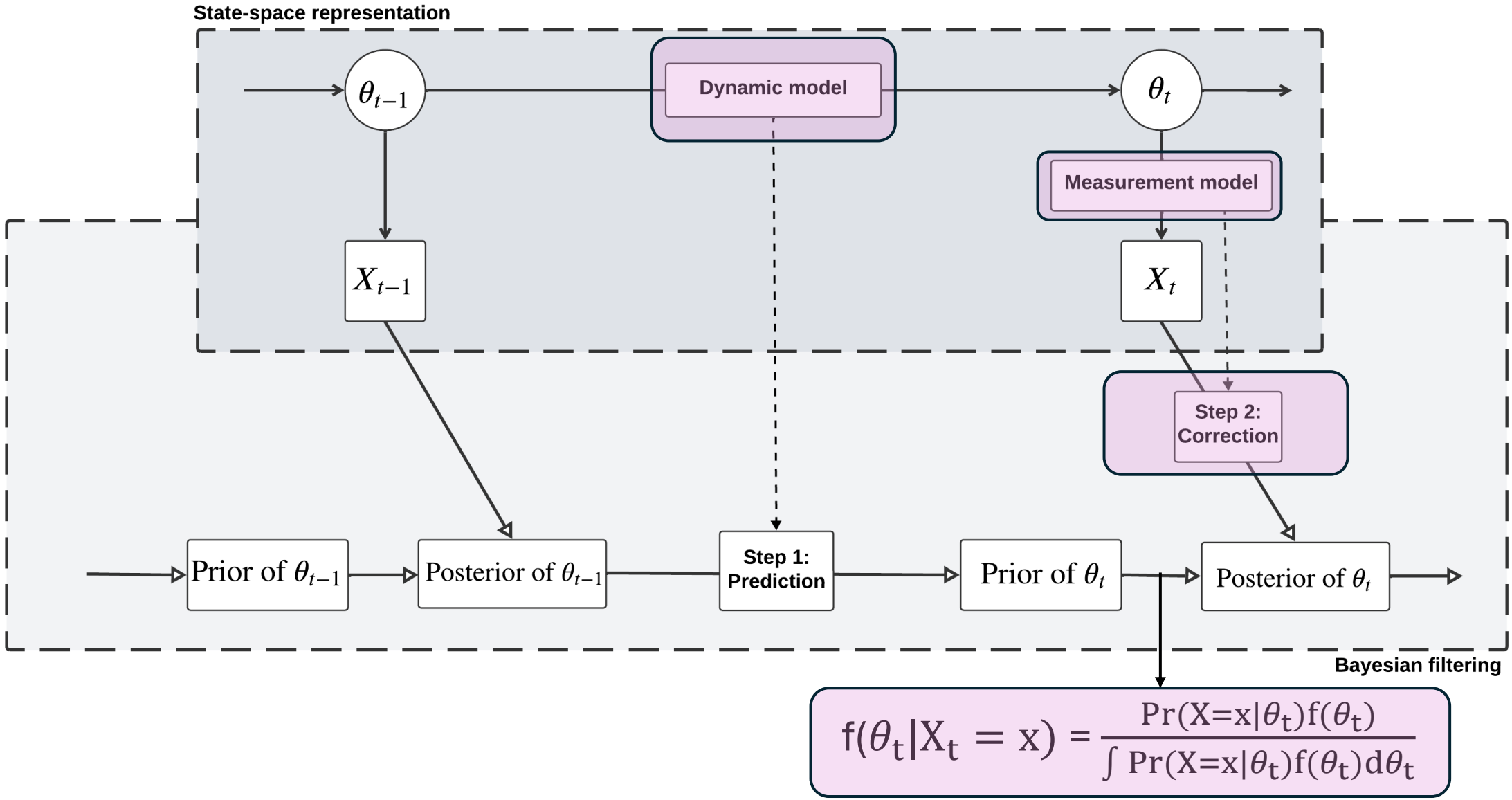
State-space representation of dynamic IRT



Different algorithms: Choices



Different algorithms: Choices



Dynamic models

How does θ_{t+1} depend on θ_t?

Random walk (σ_ϵ^2 - process variance)

$$\theta_{t+1} = \theta_t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

Wiener process (unequally spaced timepoints, Δt - time between t and t+1)

$$\theta_{t+1} = \theta_t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2 \Delta t)$$

Wiener process with drift (μ - expected change per unit of time)

$$\theta_{t+1} = \theta_t + \mu\Delta t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2 \Delta t)$$

Diminishing returns ($\rho\theta_t$ - correction factor for growth slowing down with ability)

$$\theta_{t+1} = \theta_t + \mu(1\text{-}\rho\theta_t)\Delta t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2 \Delta t)$$

How is the posterior approximated?

Only mean and variance are tracked

1. **Extended Kalman filter**: linearization, posterior is approximated with a normal distribution most closely matching at the current mean
2. **Moment-matching/Trueskill**: posterior is approximated with a normal with the same mean and variance as the actual posterior
3. **Particle filters**: posterior is approximated by a sample of points (particles) which preserve the full shape of the distribution

Extended Kalman filter

For simplicity, assume that item parameter are known

Prior at timepoint 1: $N(\mu_1, \sigma_1^2)$

Prediction at timepoint t (via dynamic model, here Wiener process)

$$\mu_t \leftarrow \mu_{t-1}$$

$$\sigma_t^2 \leftarrow \sigma_t^2 + \sigma_\epsilon^2 \Delta t$$

Correction after every response (via measurement model)

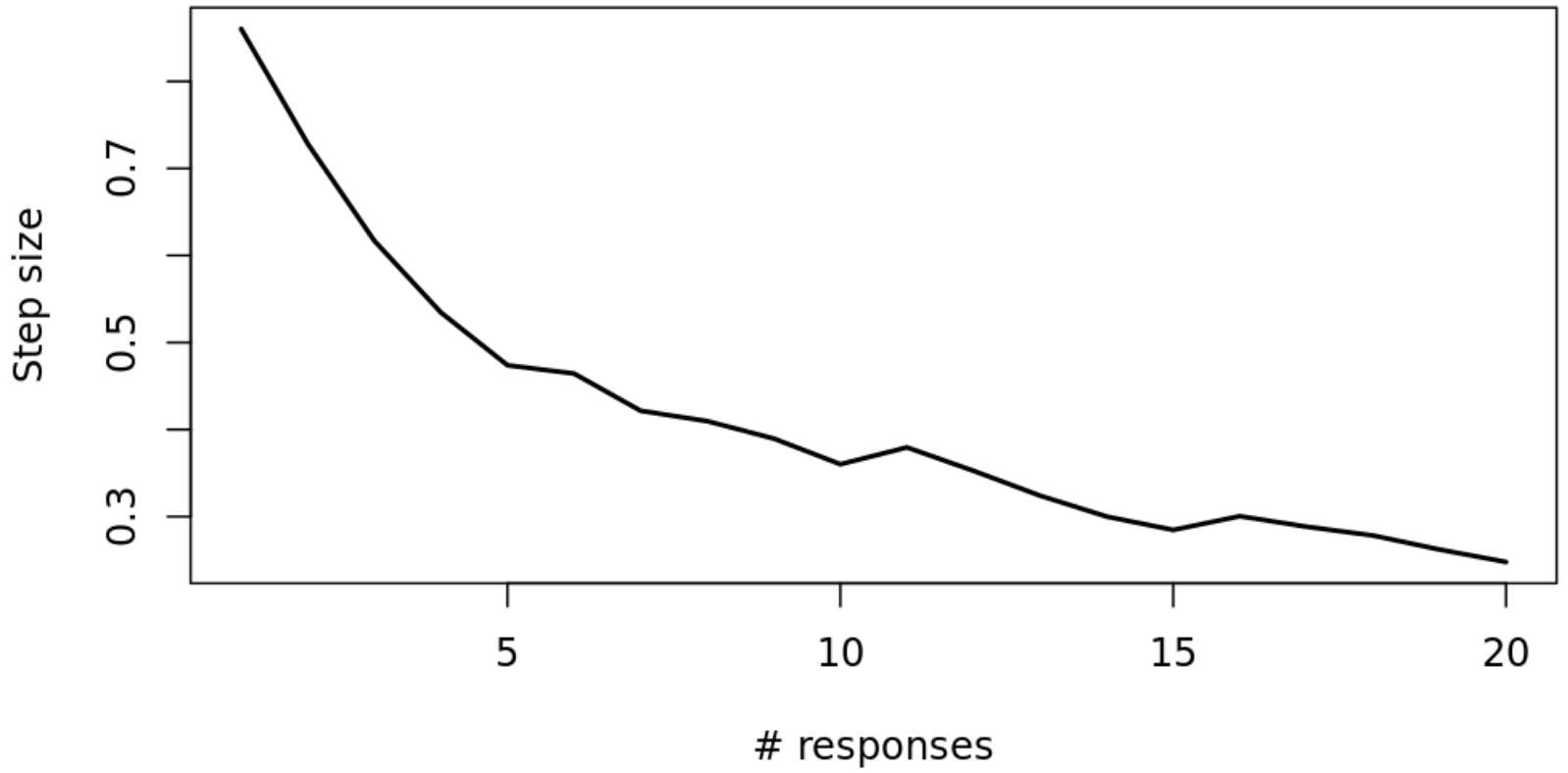
$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(X_t - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

$\mathrm{E}(X_t)$ and $\mathrm{Var}(X_t)$ are computed with ability equal to μ_t

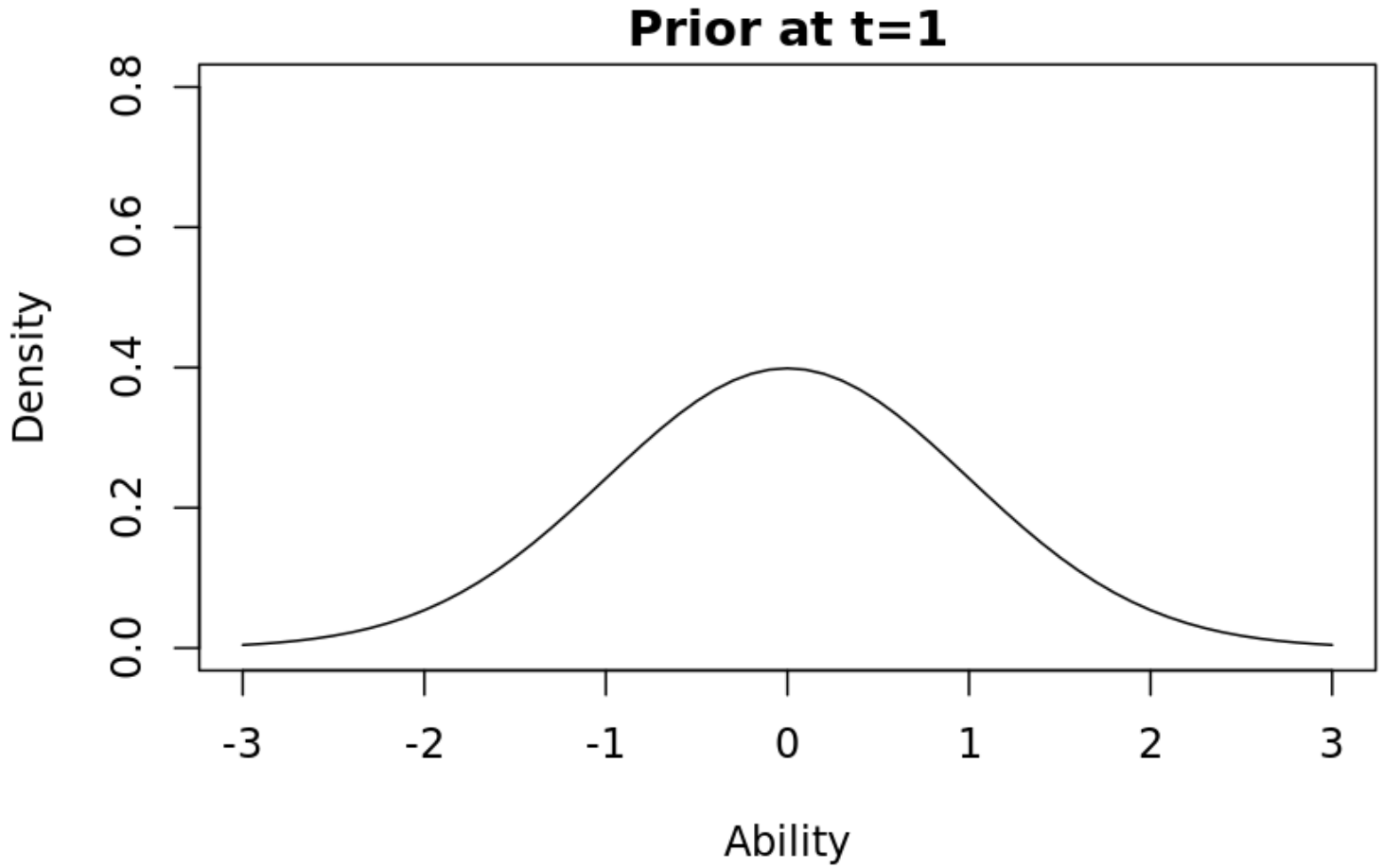
EKF = Elo with dynamic K

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)} (X_t - E(X_t))$$



Extended Kalman filter

Set μ_1, σ_1^2

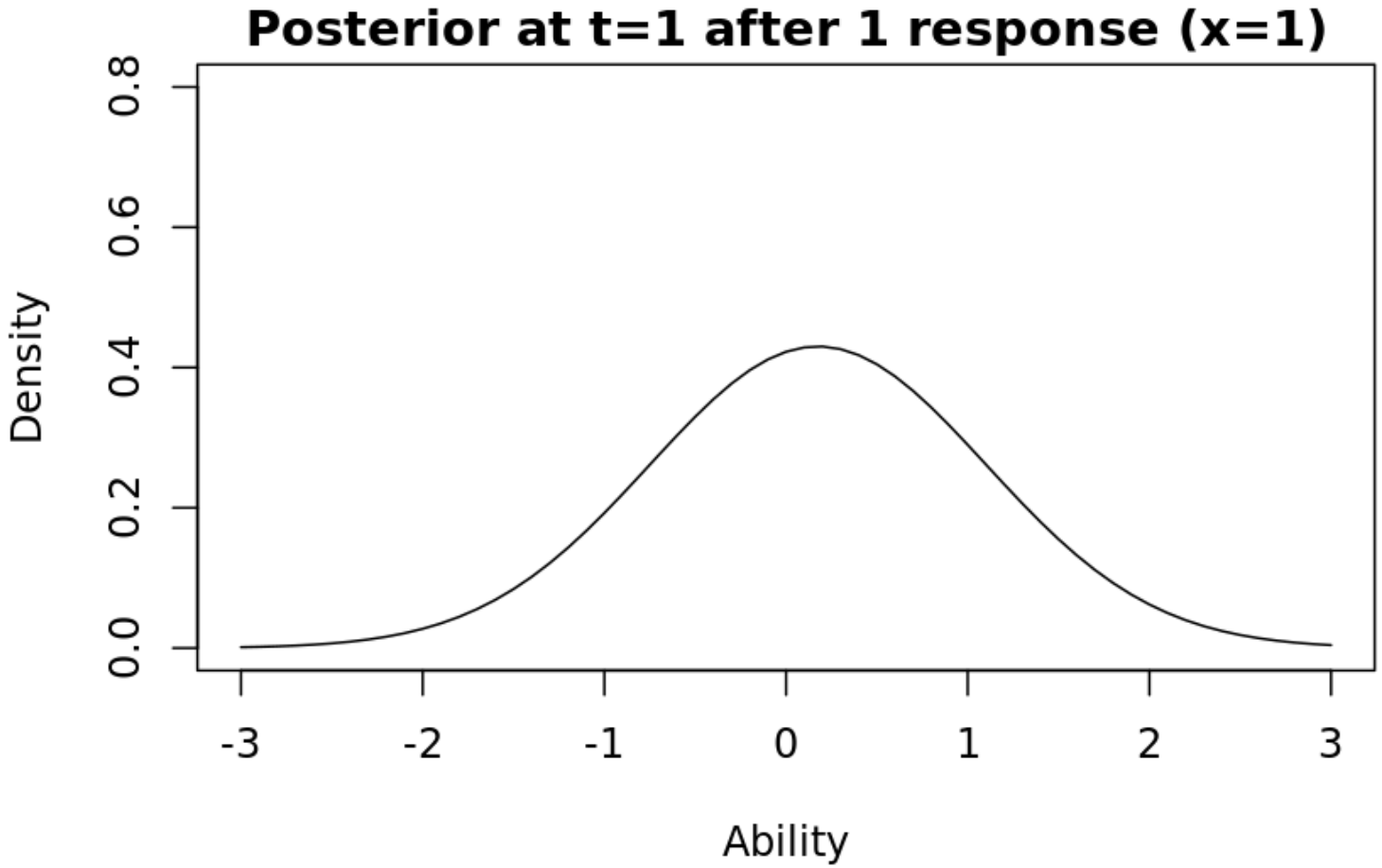


Extended Kalman filter

Observe $X_{1,1}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

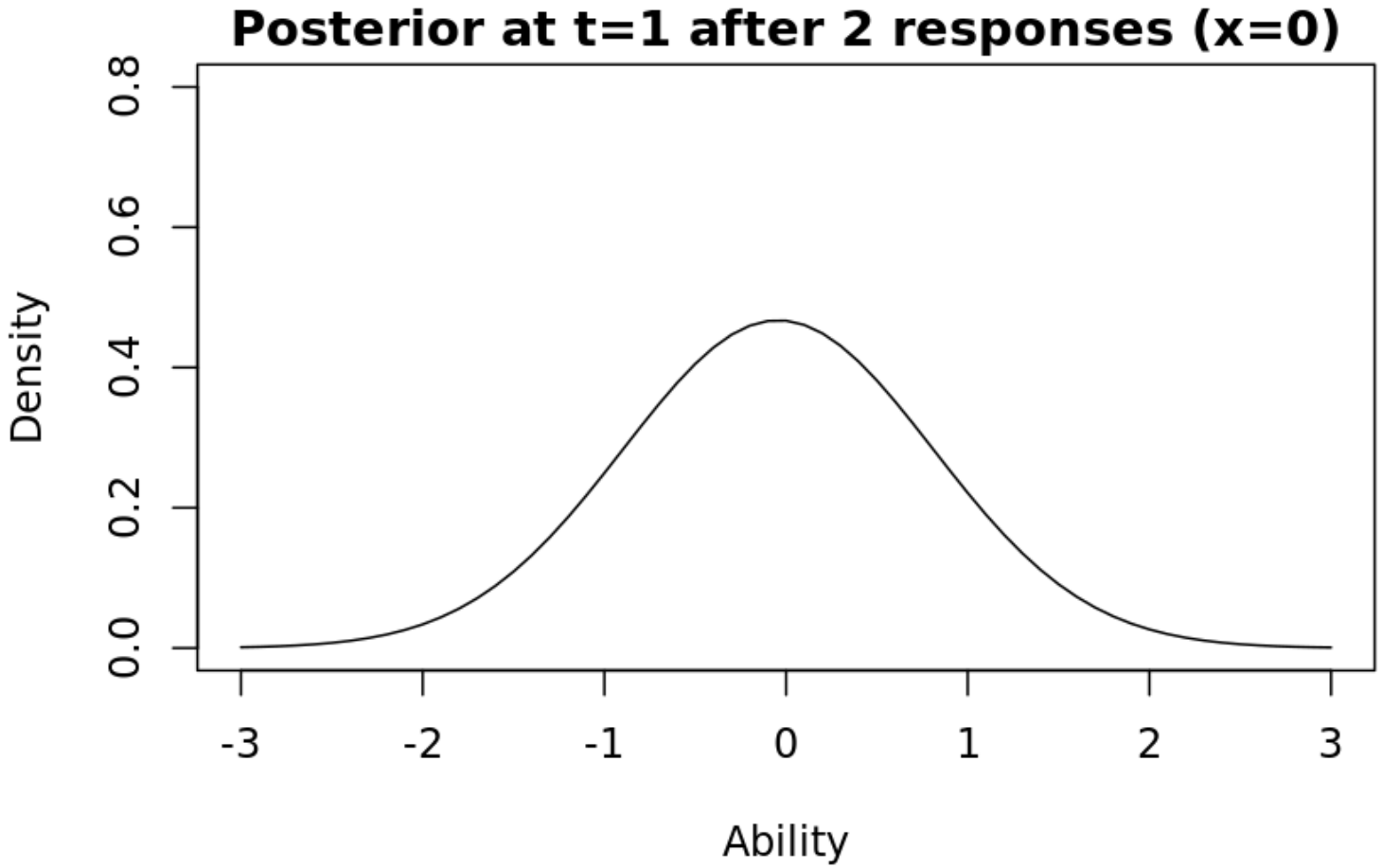


Extended Kalman filter

Observe $X_{1,2}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

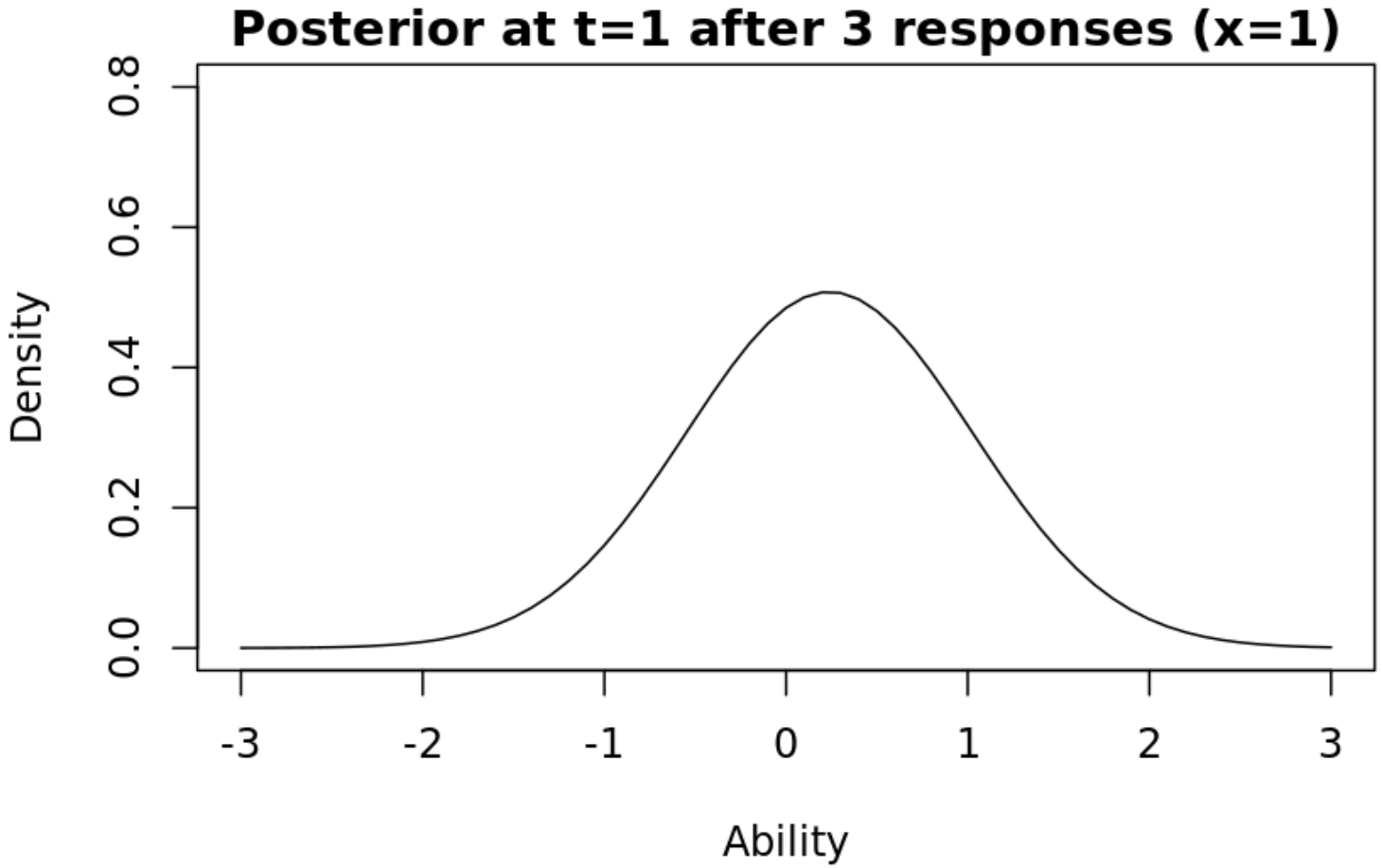


Extended Kalman filter

Observe $X_{1,3}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

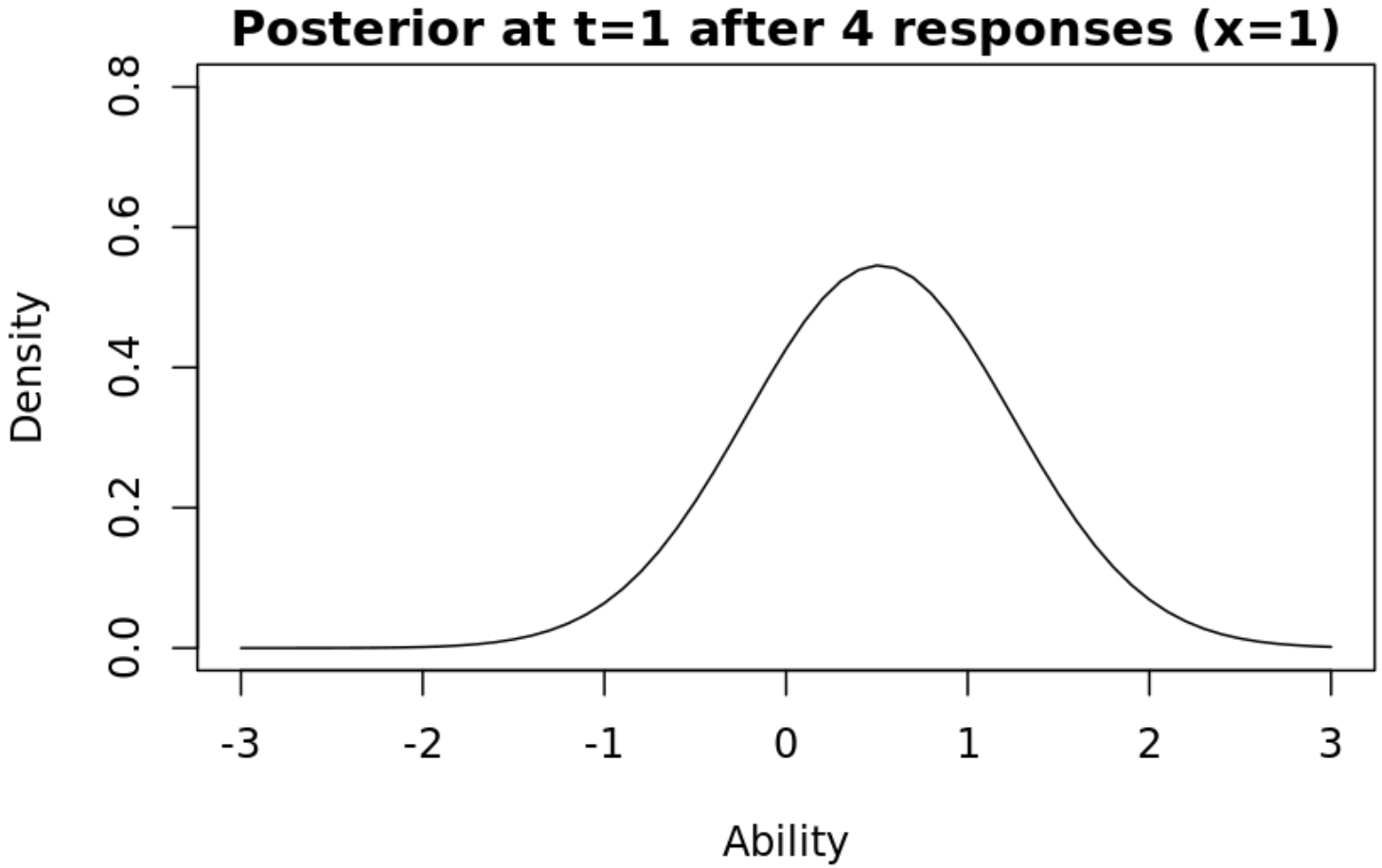


Extended Kalman filter

Observe $X_{1,4}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

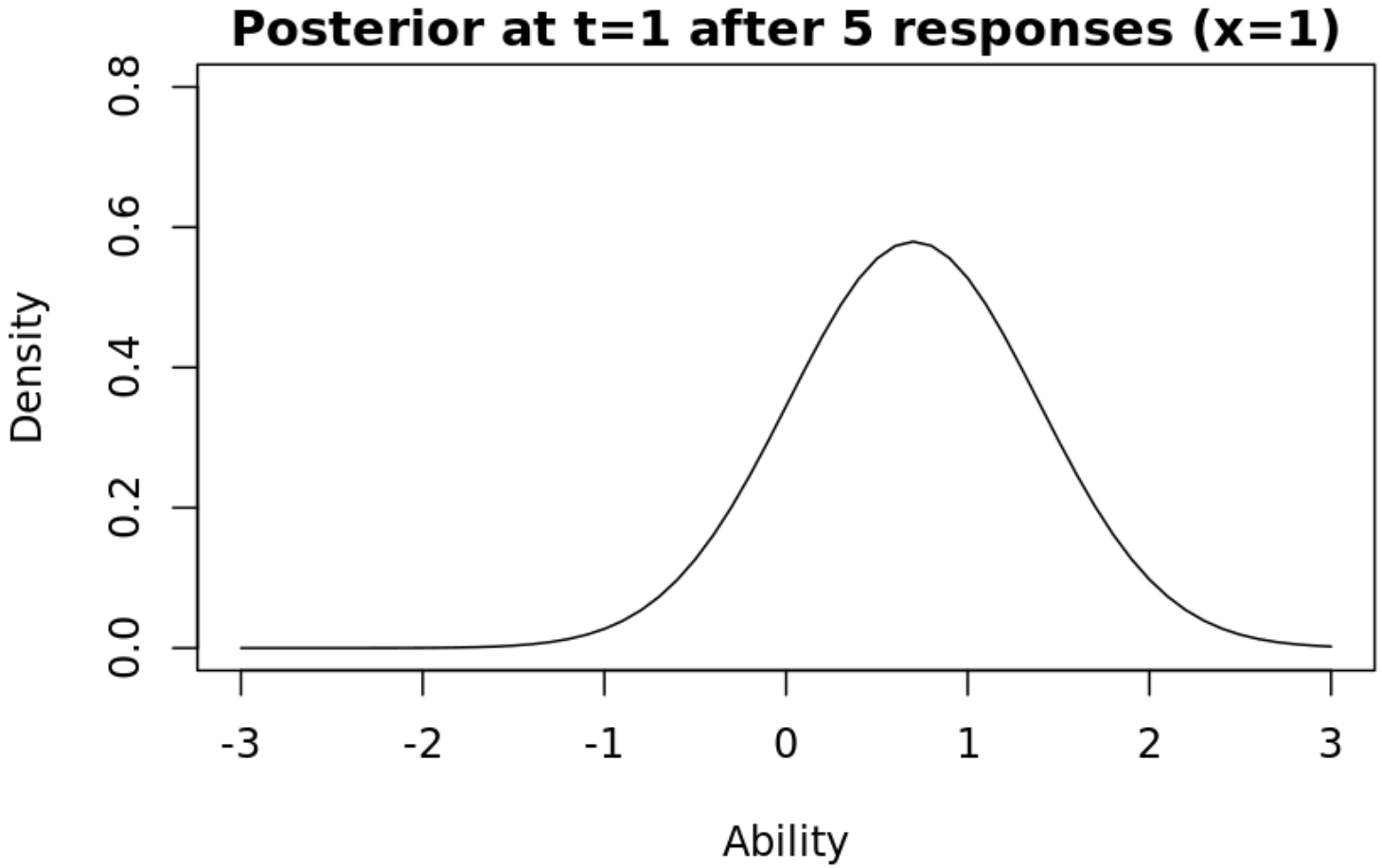


Extended Kalman filter

Observe $X_{1,5}=x$

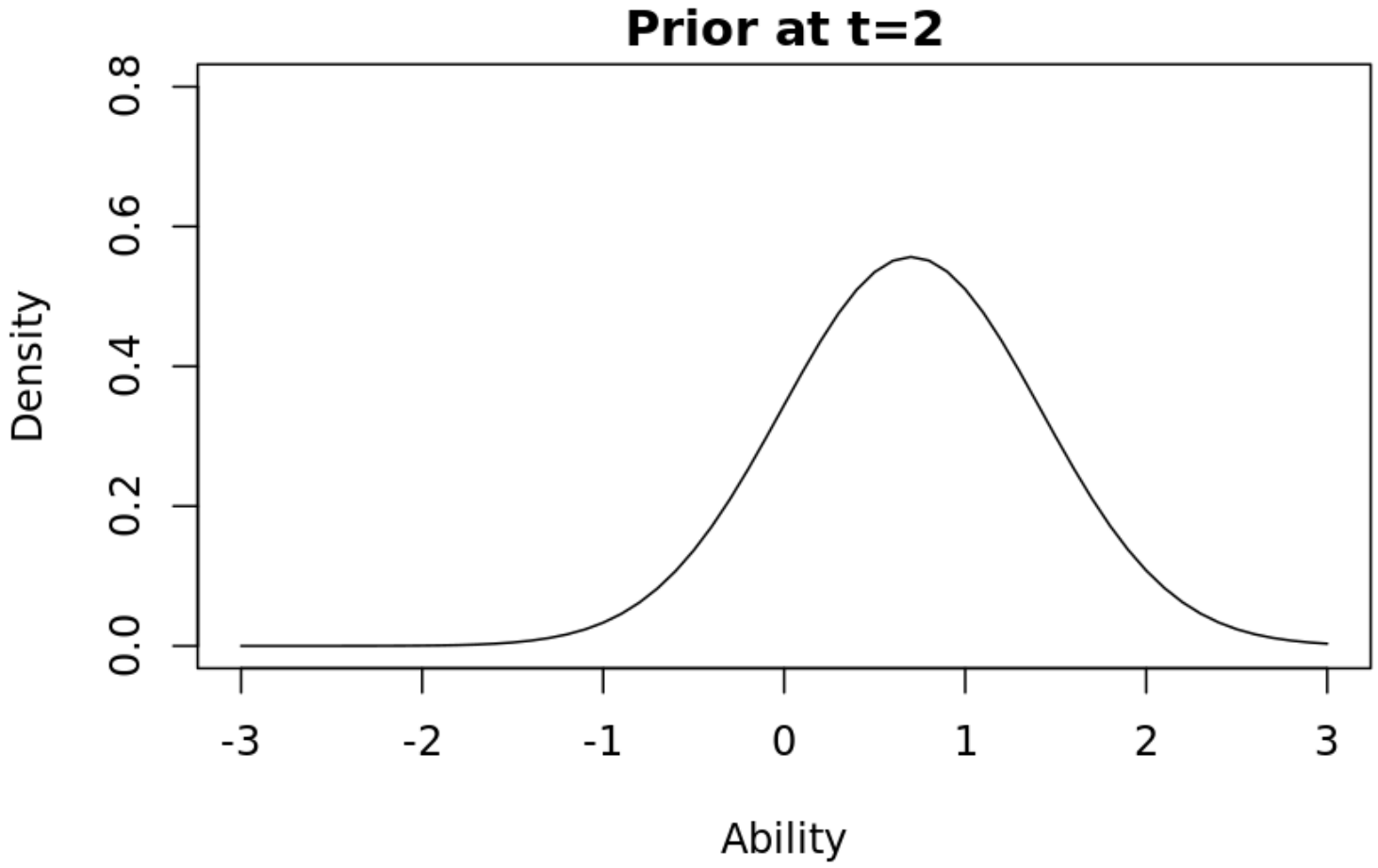
$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$



Extended Kalman filter

Predict: $\mu_t = \mu_{t-1}, \sigma_t^2 = \sigma_{t-1}^2 + \sigma_\epsilon^2 \Delta t$

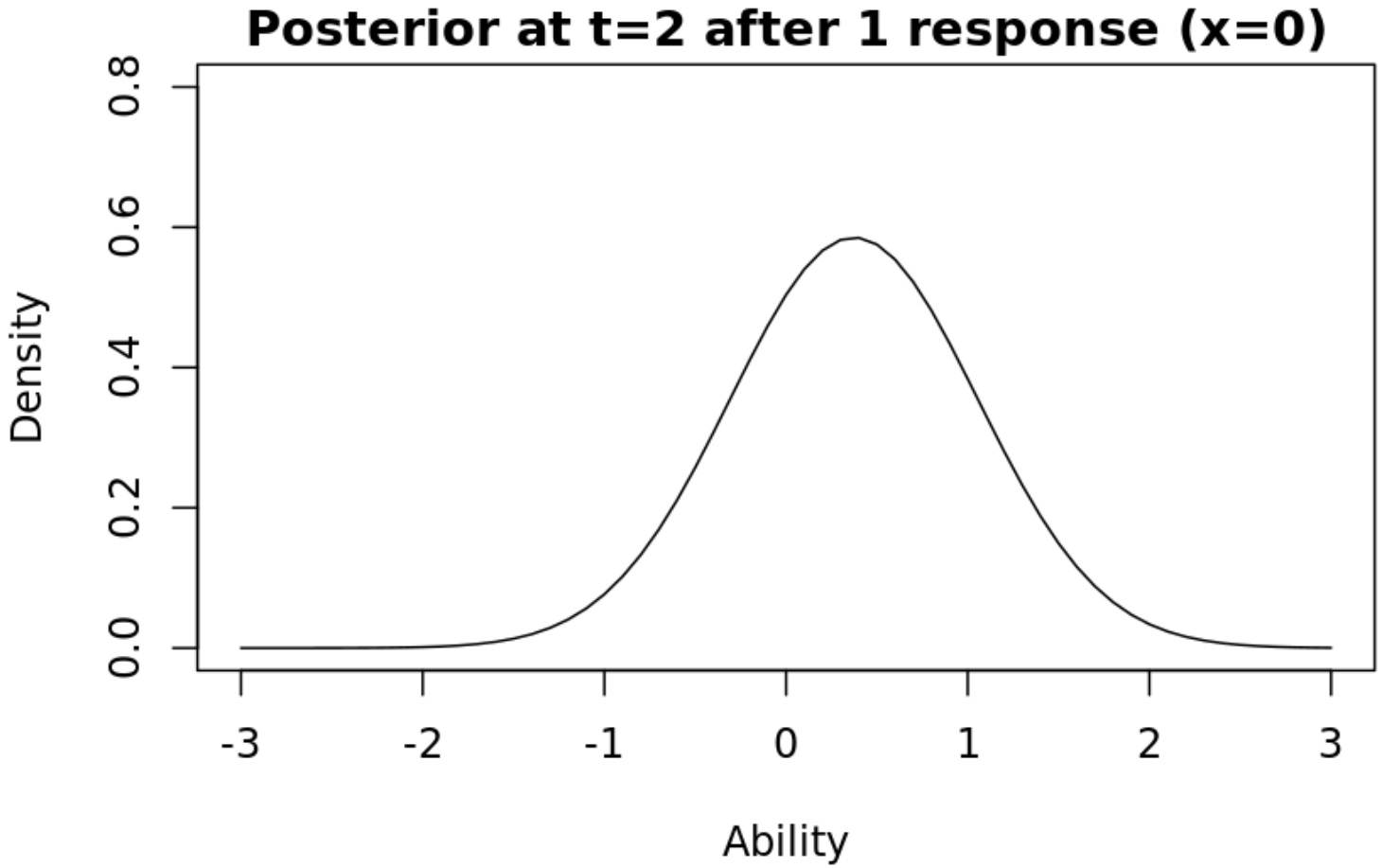


Extended Kalman filter

Observe $X_{2,1}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

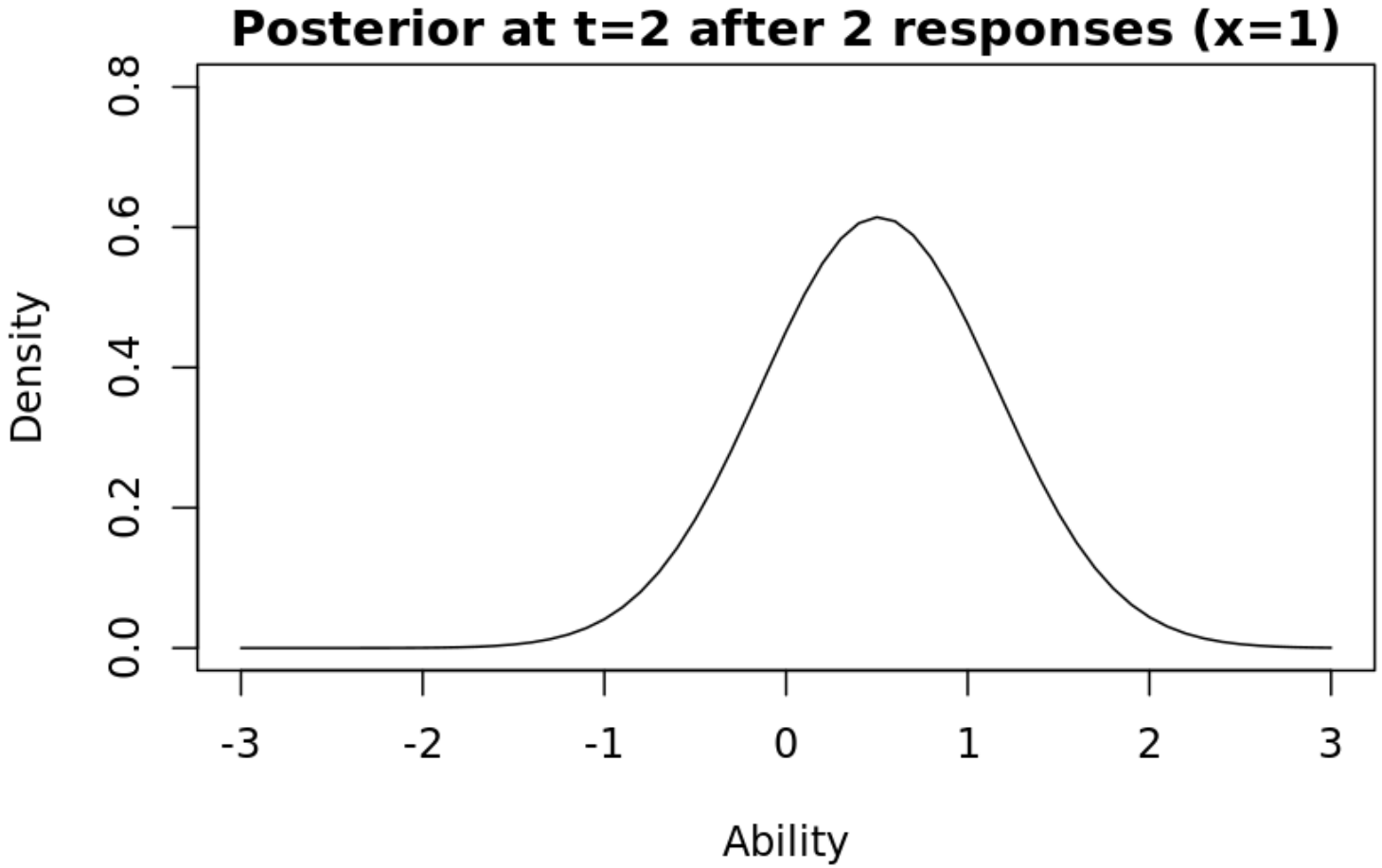


Extended Kalman filter

Observe $X_{2,2}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

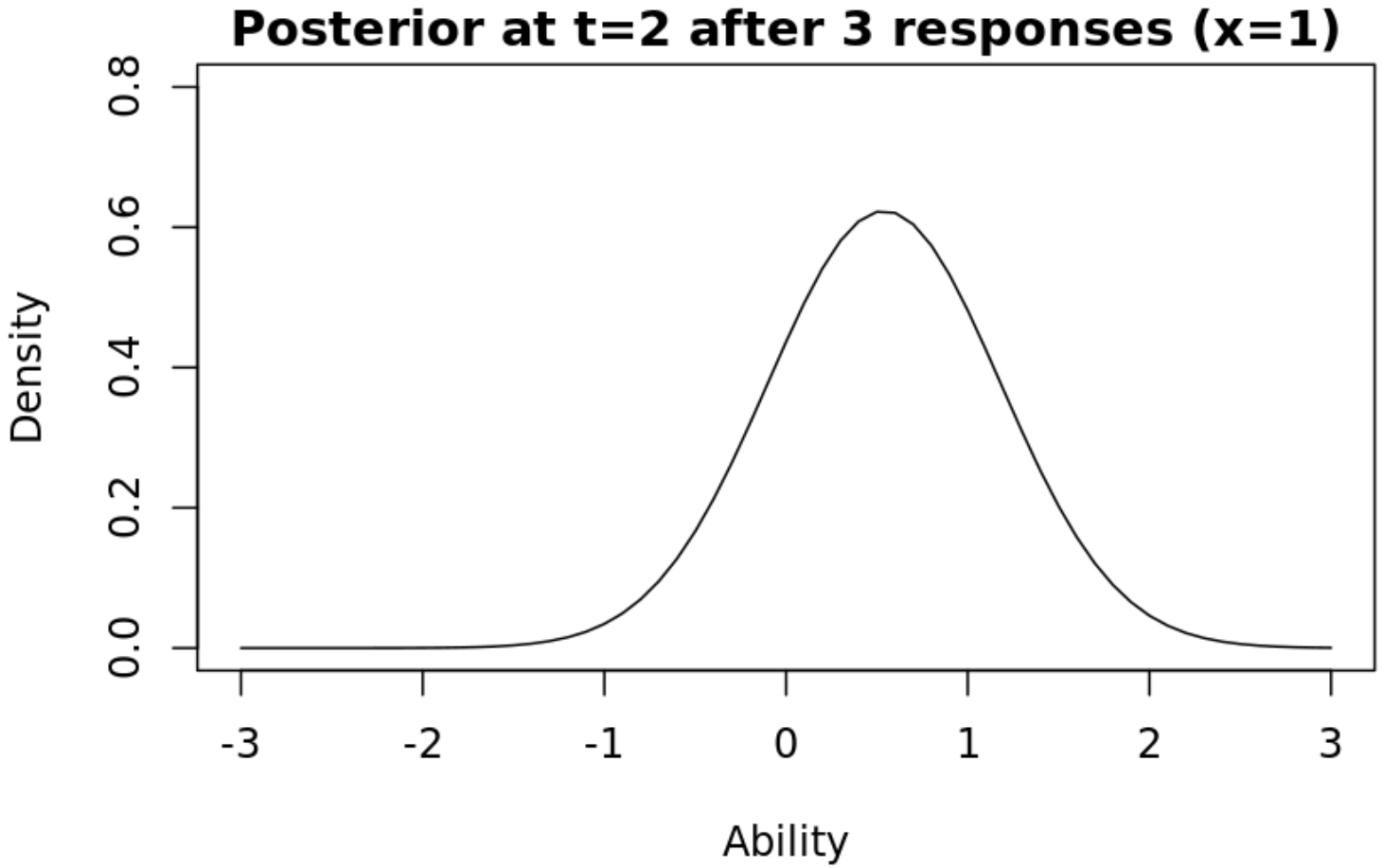


Extended Kalman filter

Observe $X_{2,3}$=x

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

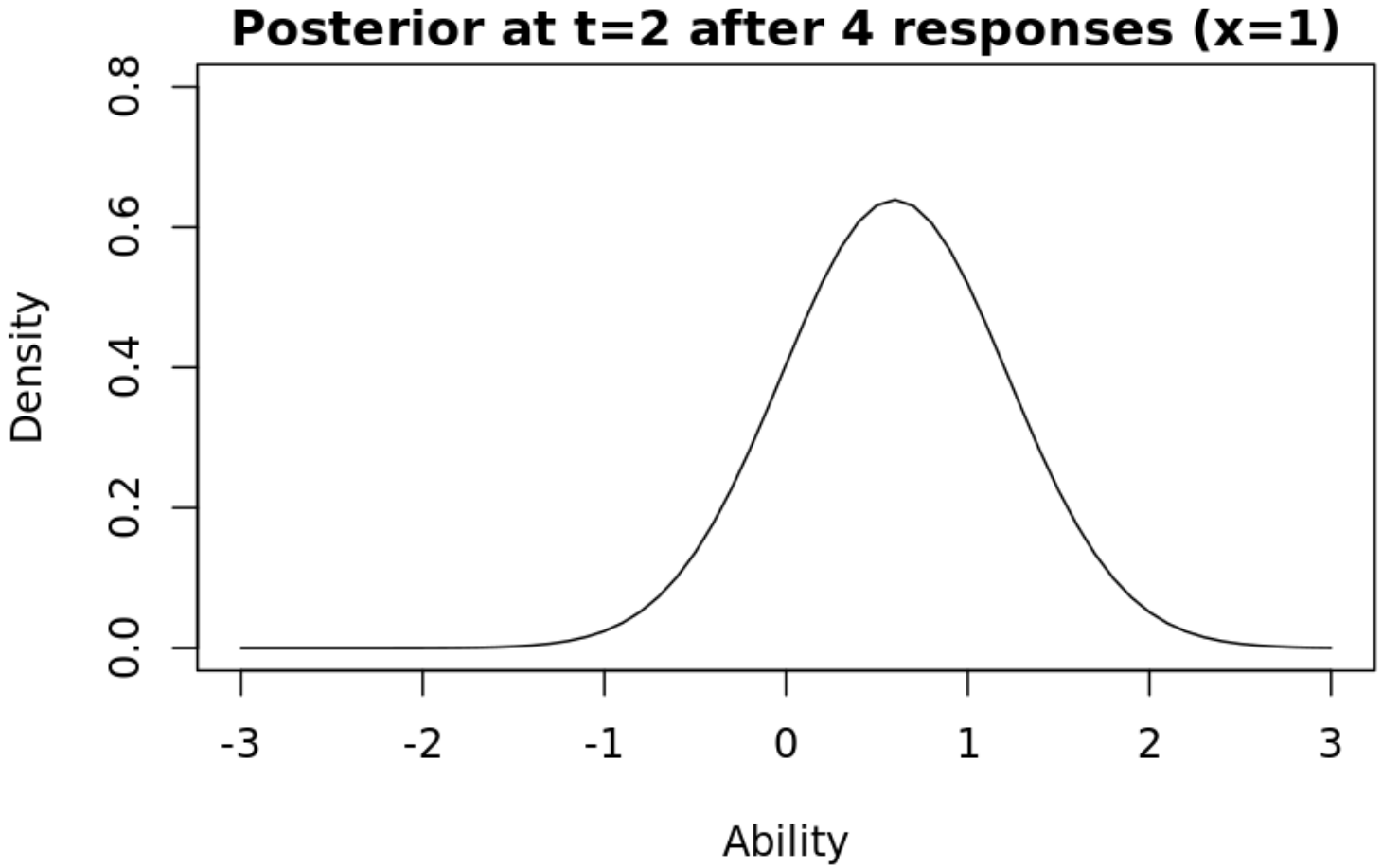


Extended Kalman filter

Observe $X_{2.4}$=x

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$

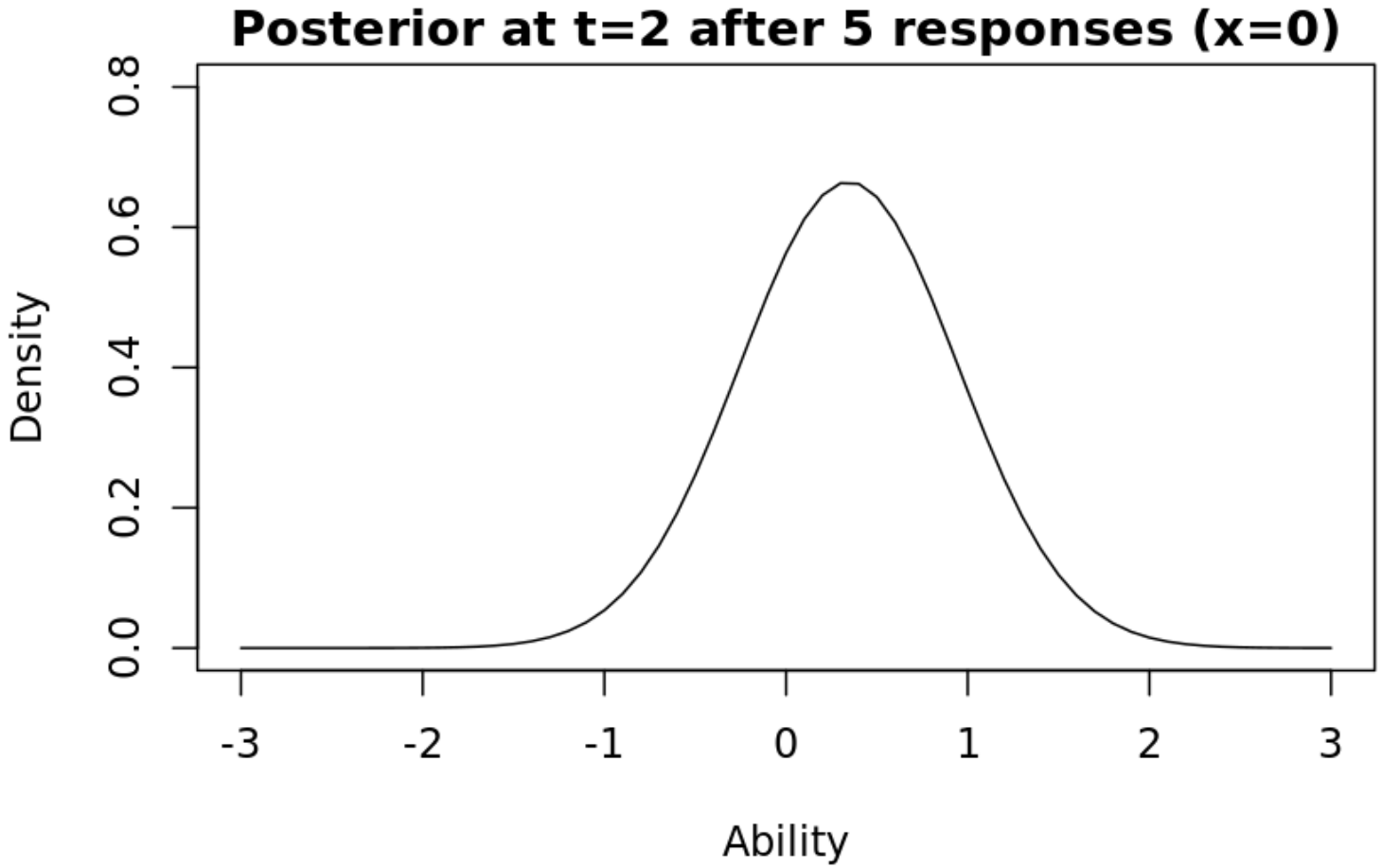


Extended Kalman filter

Observe $X_{2,5}=x$

$$\mu_t \leftarrow \mu_t + \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}(x - \mathrm{E}(X_t))$$

$$\sigma_t^2 \leftarrow \frac{1}{1/\sigma_t^2 + \mathrm{Var}(X_t)}$$



Moment-matching a.k.a Trueskill a.k.a Assumed density filter

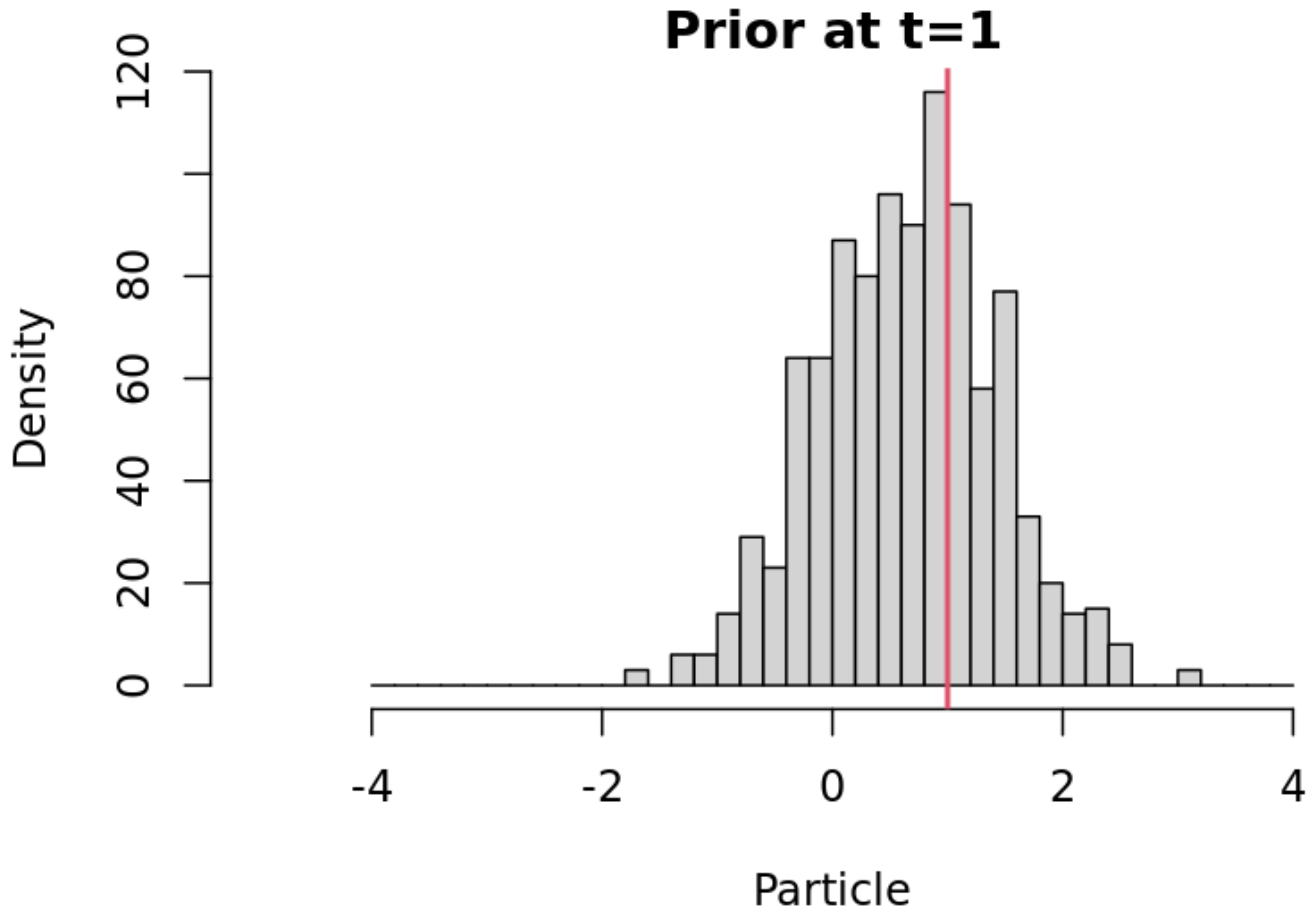
- If instead of the Rasch model we use a normal ogive model:

$$\Pr(X_{it} = 1) = \Phi(\theta_t - \delta_i)$$

- Then the mean and variance of the posterior have a closed form
- Prediction step is the same as in EKF
- Correction step sets the updated moments to the moments of the actual posterior
- Similar dynamics for the mean and variance as in EKF

Particle filter

Sample $\theta^{(1)}, \ldots, \theta^{(P)}$ from f(θ_1)

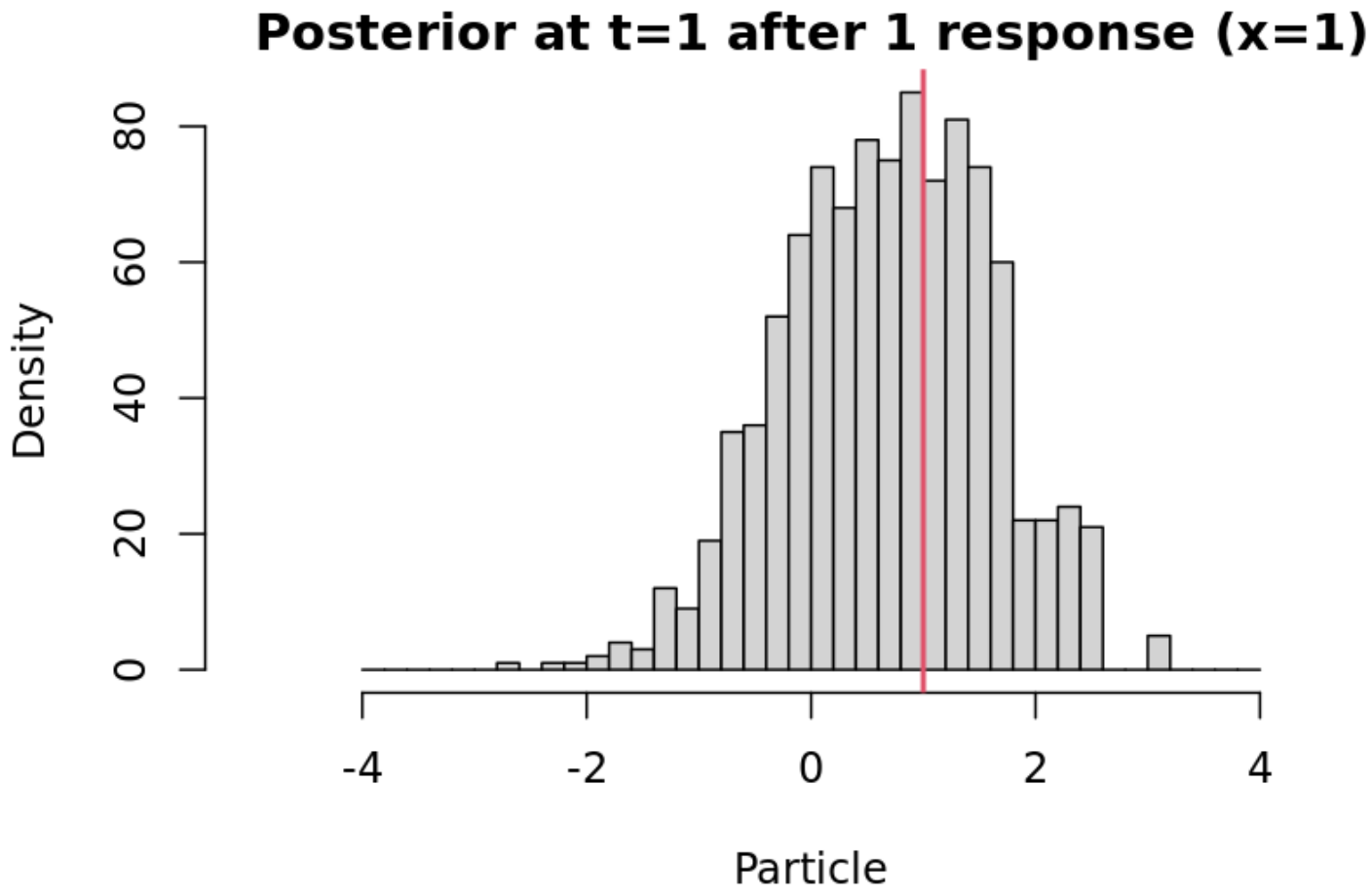


Particle filter

Observe $X_{1,1}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles using the weights

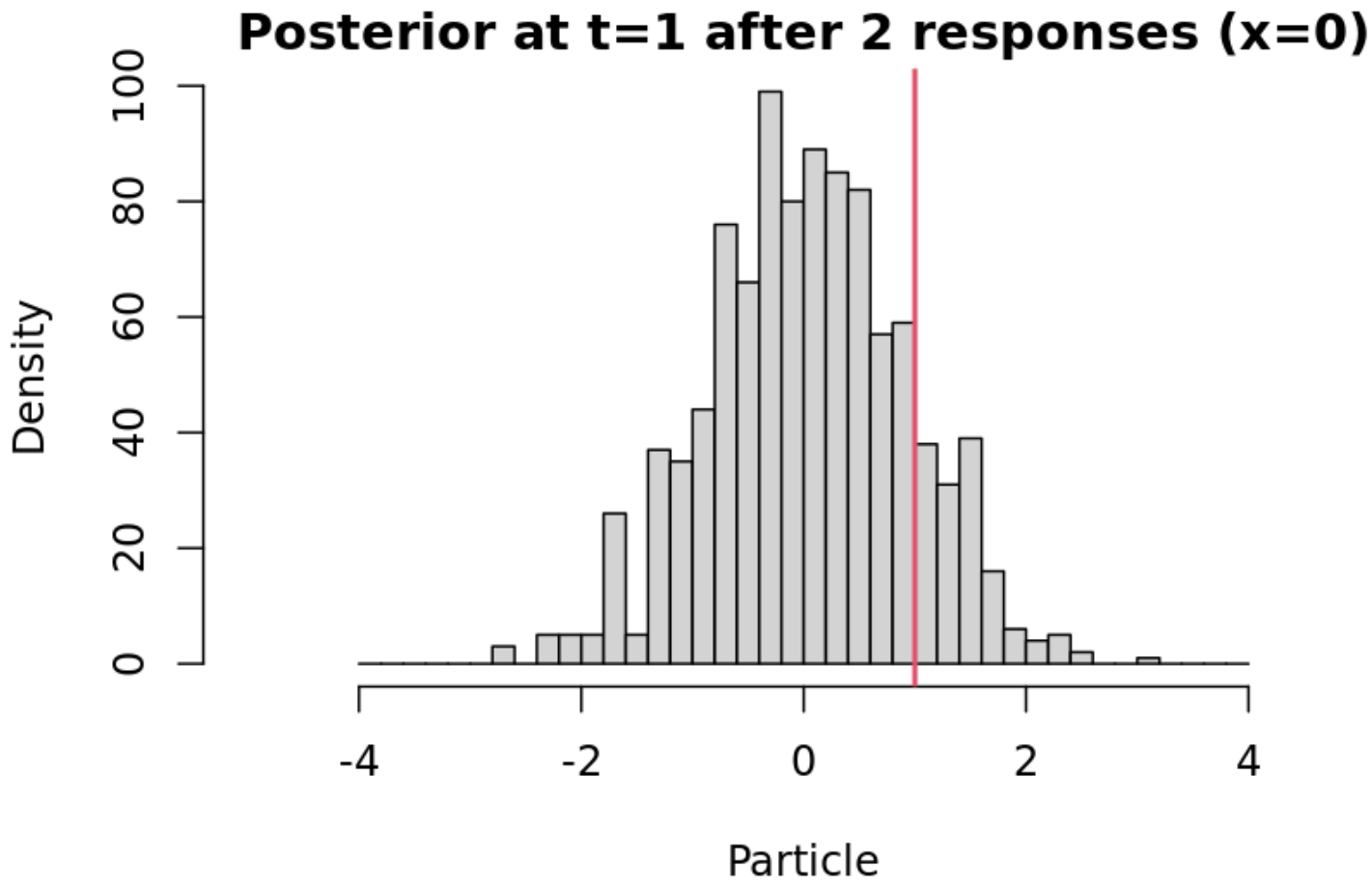


Particle filter

Observe $X_{1,2}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights

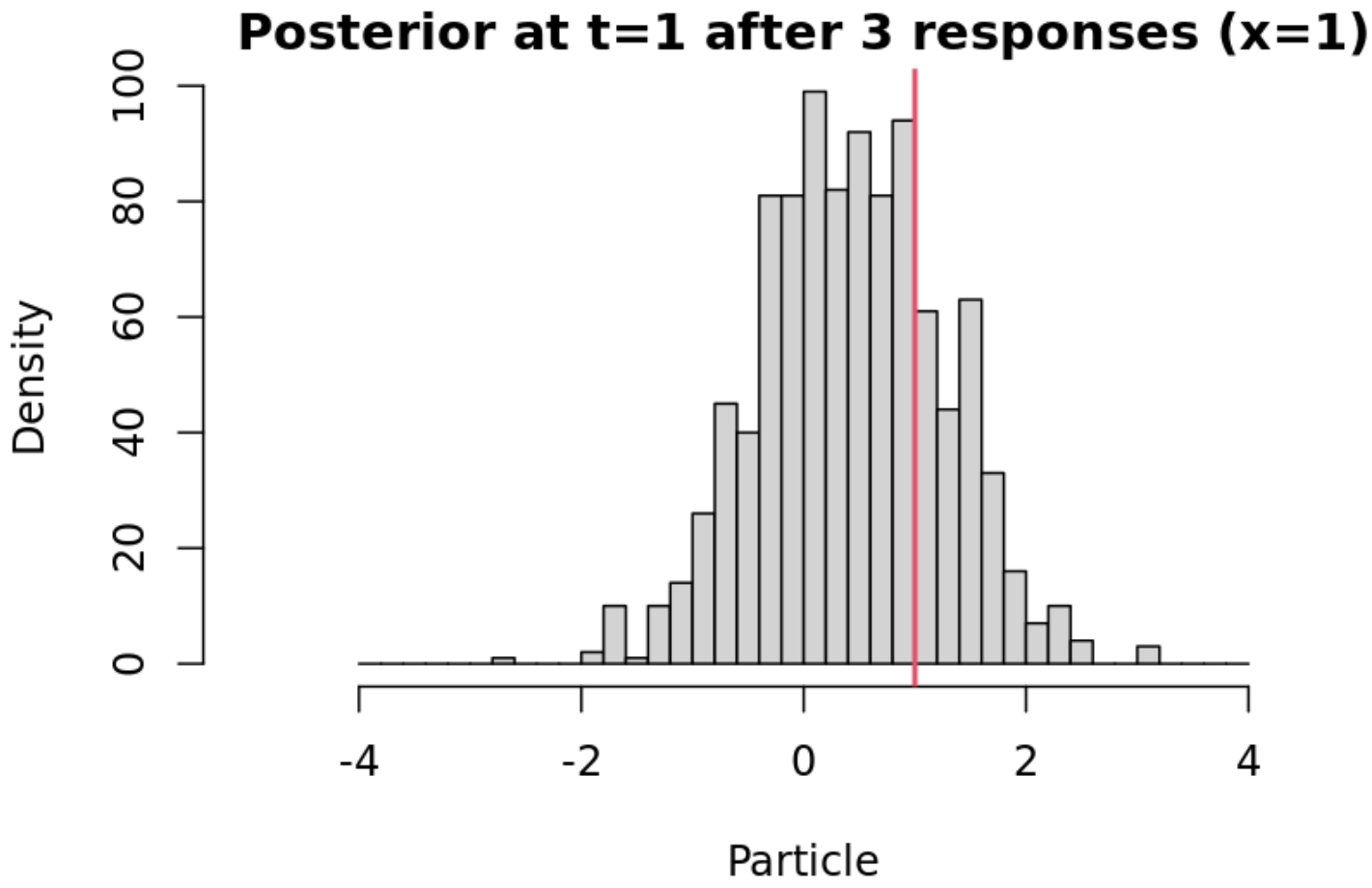


Particle filter

Observe $X_{1,3}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights

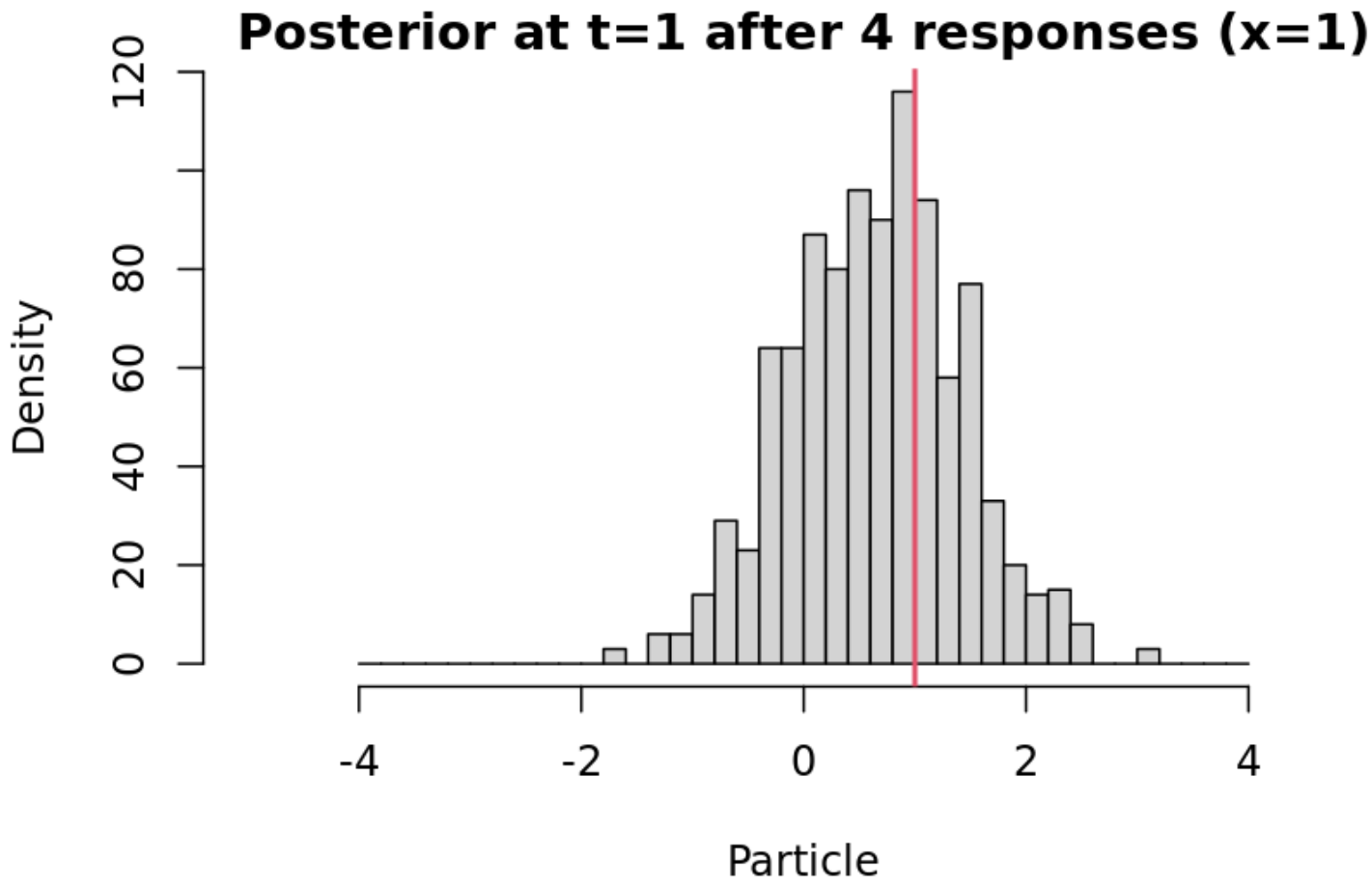


Particle filter

Observe $X_{1,4}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights

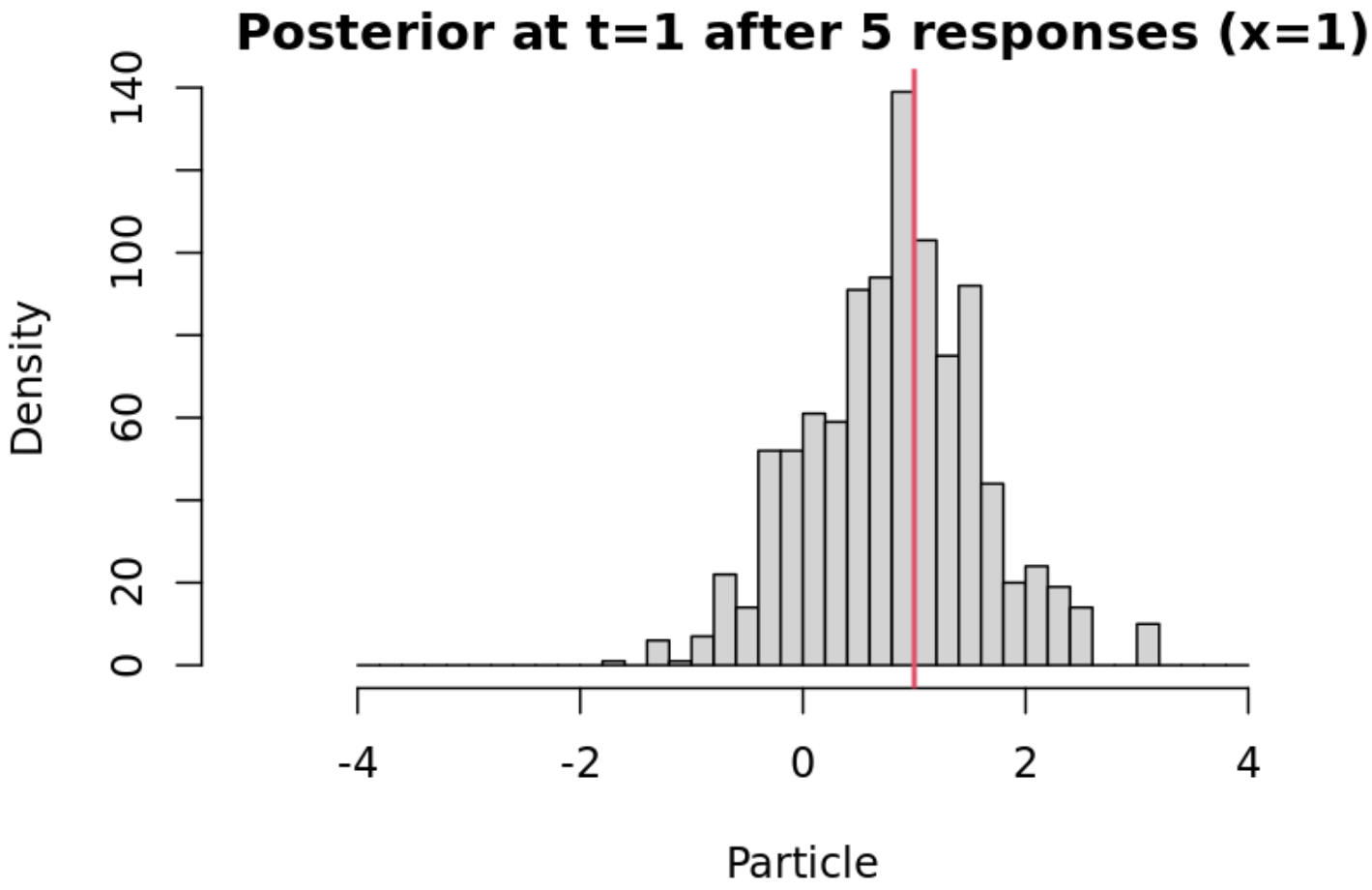


Particle filter

Observe $X_{1,5}$=x

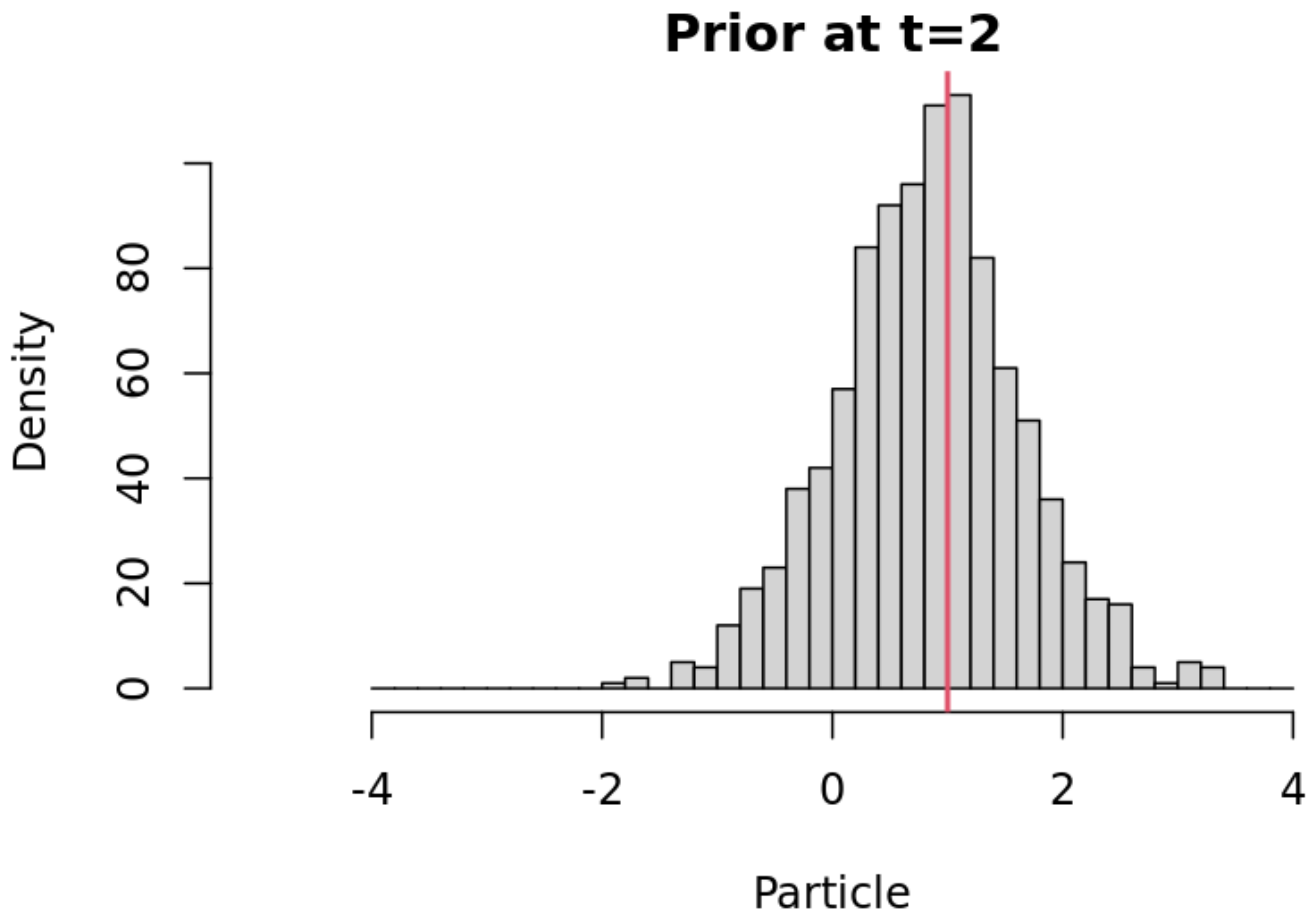
Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights



Particle filter

Sample $\theta^{(1)}, \ldots, \theta^{(P)}$ from $f(\theta_2|\theta_1)$

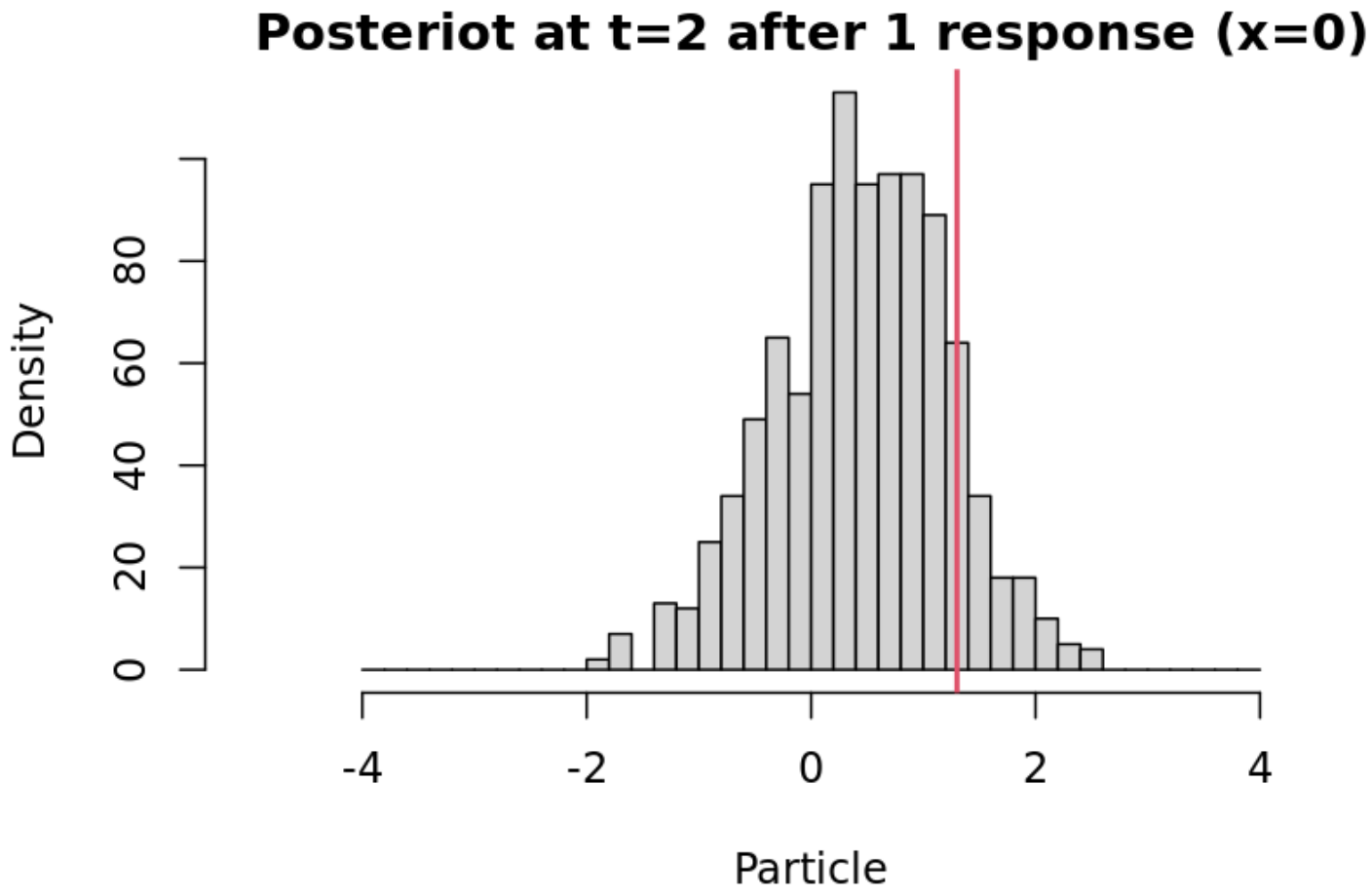


Particle filter

Observe $X_{2,1}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles using the weights

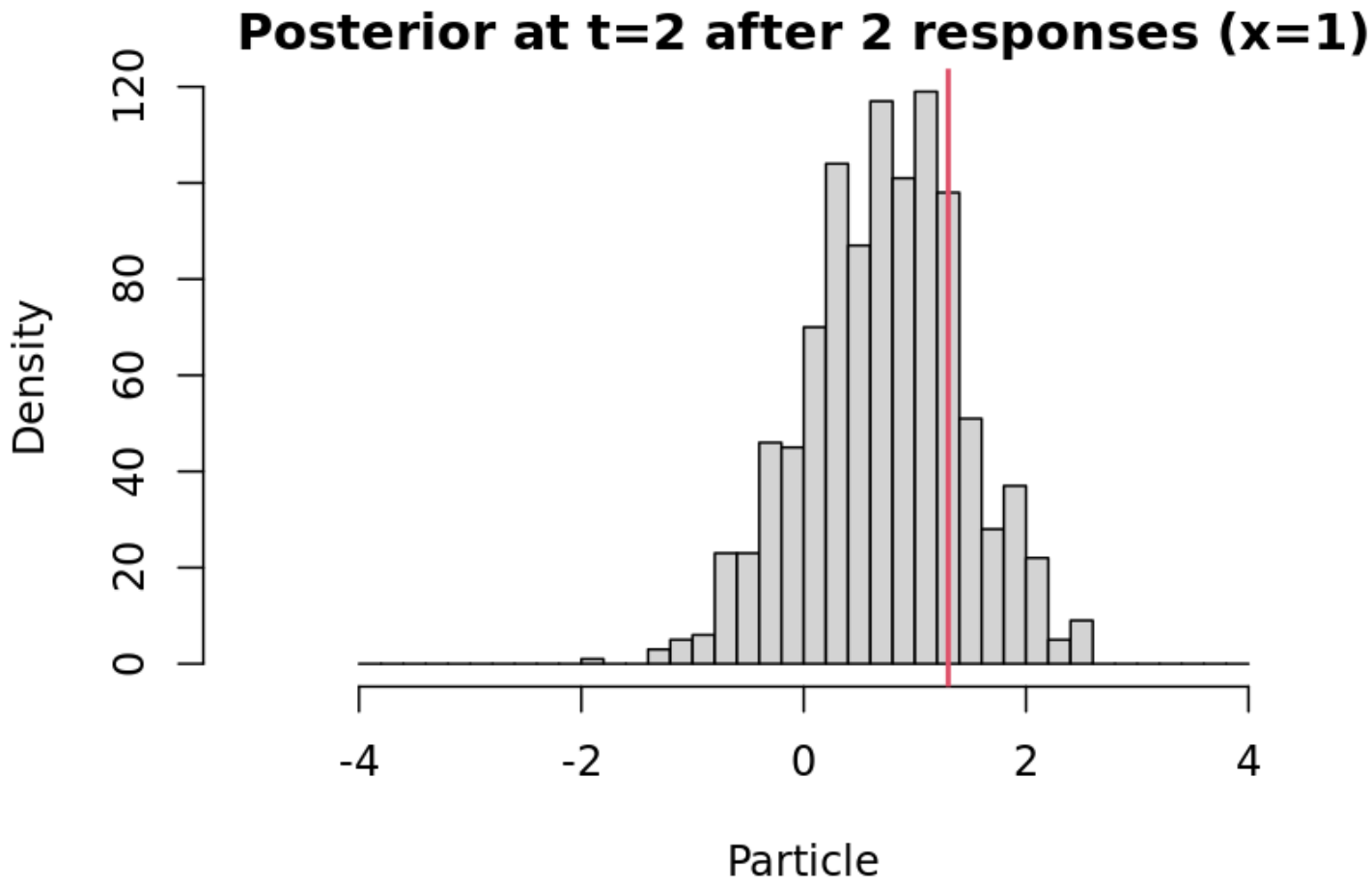


Particle filter

Observe $X_{2,2}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights

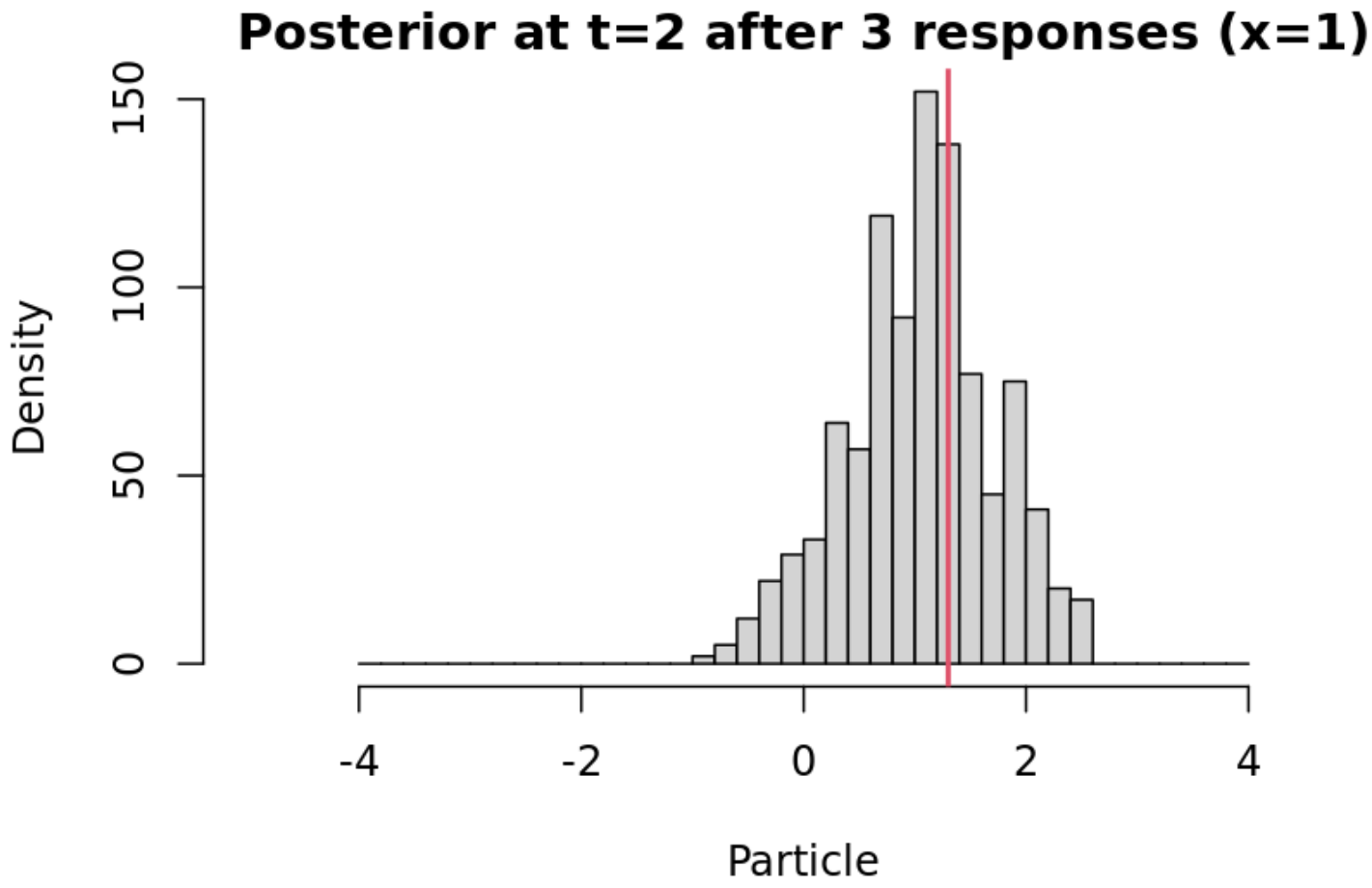


Particle filter

Observe $X_{2,3}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights

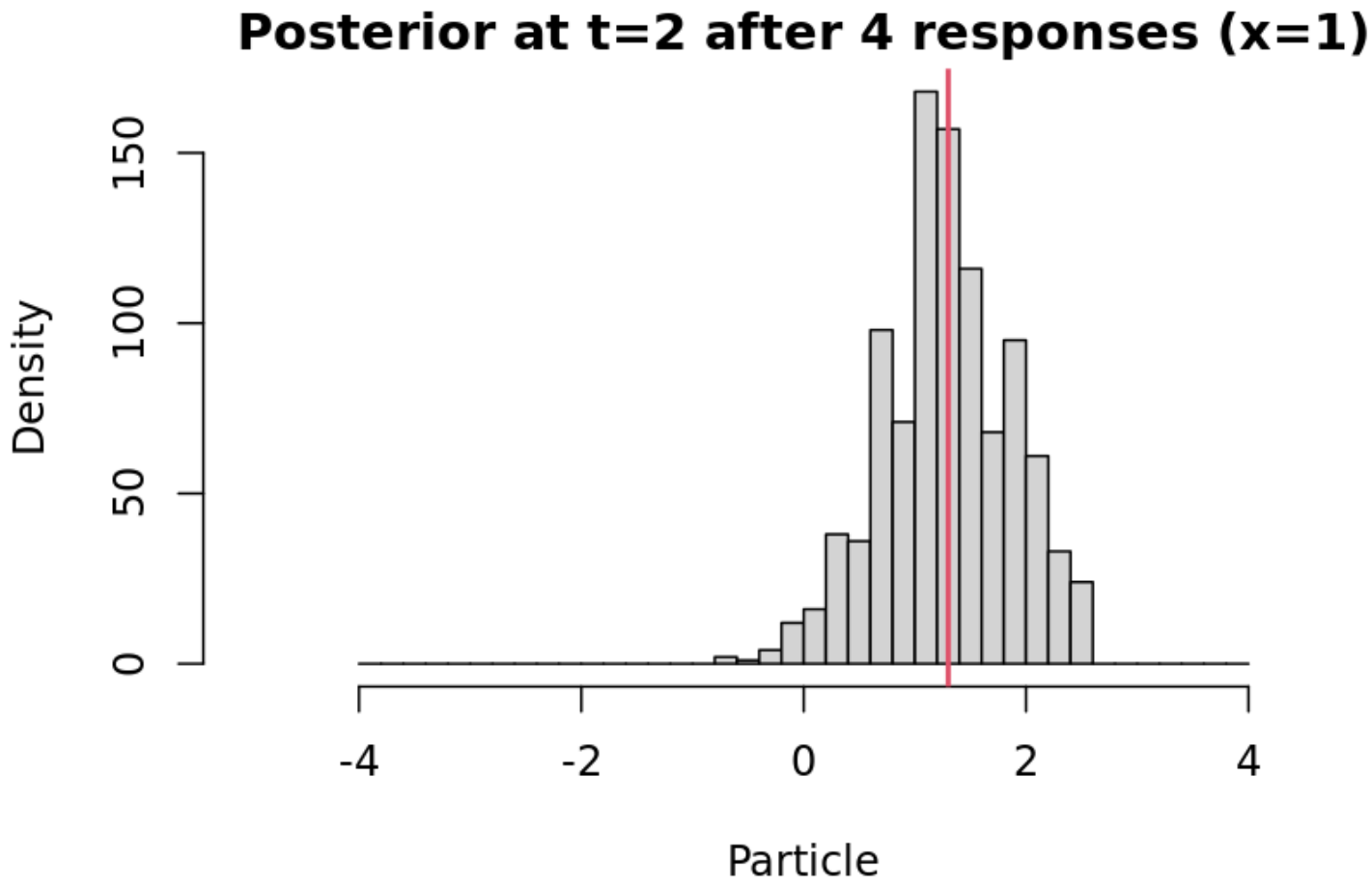


Particle filter

Observe $X_{2,4}$=x

Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights

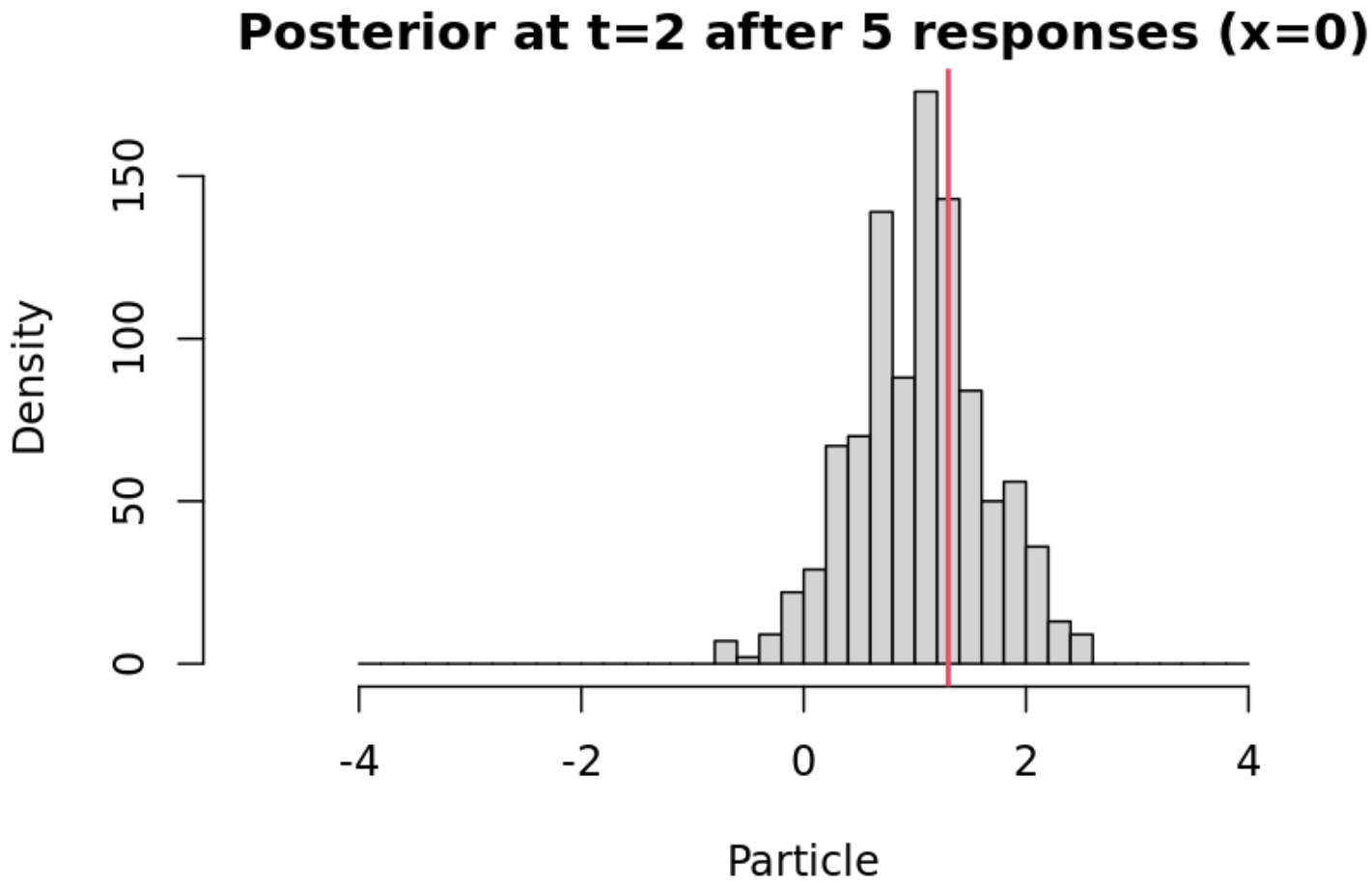


Particle filter

Observe $X_{2,5}$=x

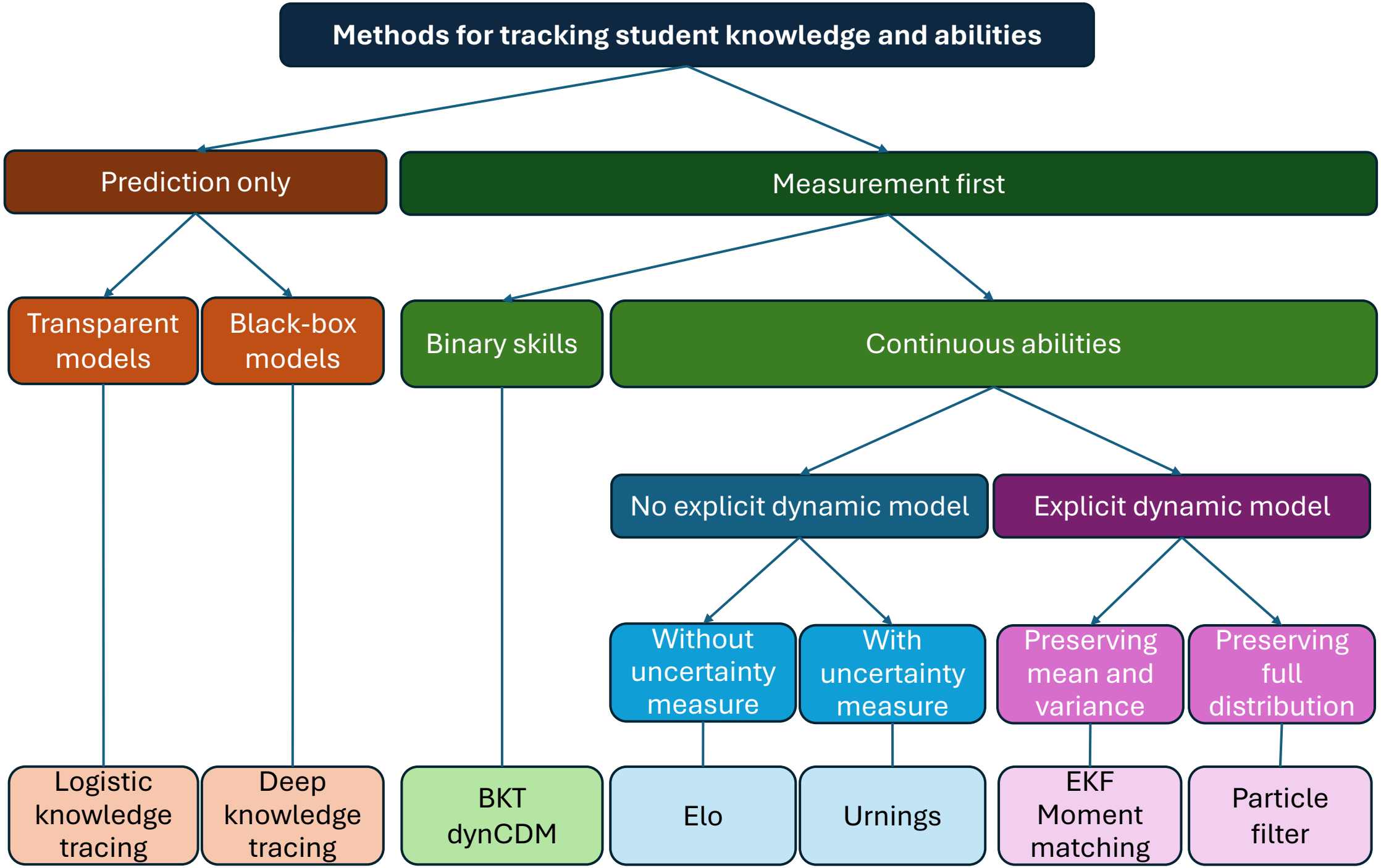
Compute weights for $\theta^{(1)}, \ldots, \theta^{(P)}$ proportional to the posterior: $w^{(p)} \propto Pr(X_{1,1} = x)$

Resample the particles Using the weights



Particle filter: Remarks

- Preserves the *shape of the distribution*
- *Easy to adapt* to new measurement and dynamic model
- Accuracy of approximation depends on the *number of particles* (P)
- Caution with *particle degeneracy*
- High *computational and memory* costs


Methods for tracking student knowledge and abilities
Prediction only
Measurement first
Transparent models
Black-box models
Binary skills
Continuous abilities
No explicit dynamic model
Explicit dynamic model
Without uncertainty measure
With uncertainty measure
Preserving mean and variance
Preserving full distribution
Logistic knowledge tracing
Deep knowledge tracing
BKT dynCDM
Elo
Urnings
EKF Moment matching
Particle filter

Open questions

Different methods are rarely compared on the same datasets

- When is which method the most suitable and would perform best?
- Is BKT better than IRT when working with fine-grained skills?
- How do “measures” from DKT compare to IRT measures in terms of measurement properties?
- When is prediction enough and when not?
- How to evaluate measurement quality beyond prediction accuracy?

Achieving high measurement precision is challenging

- Can we gain additional information from process data without sacrificing validity and fairness?
- Can we improve measurement through extending the dynamic model?
- Can we improve measurement by taking the hierarchical structure of the skills into account?