

Noise-Contrastive Estimation for Multivariate Point Processes

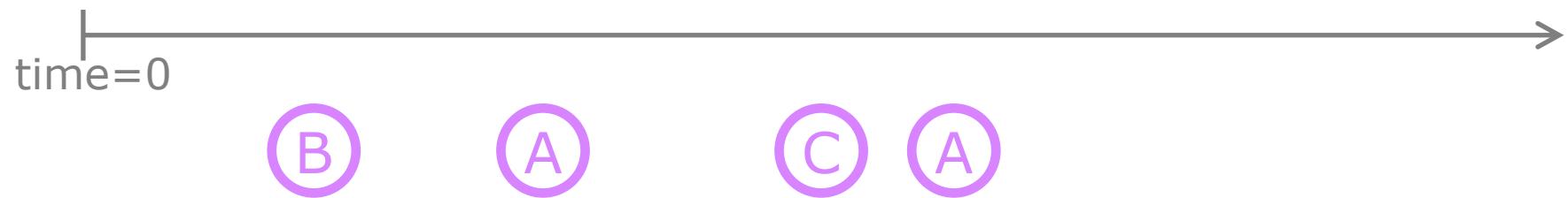
Hongyuan Mei, Tom Wan, Jason Eisner
Johns Hopkins University

MLE: Max log prob of *data*

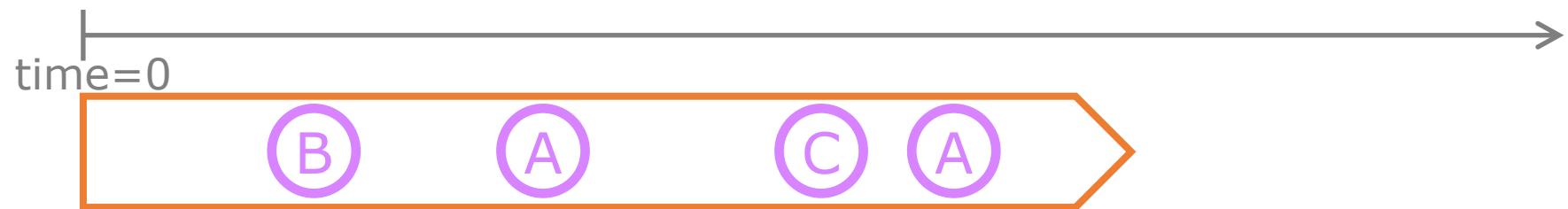
MLE: Max log prob of *data*



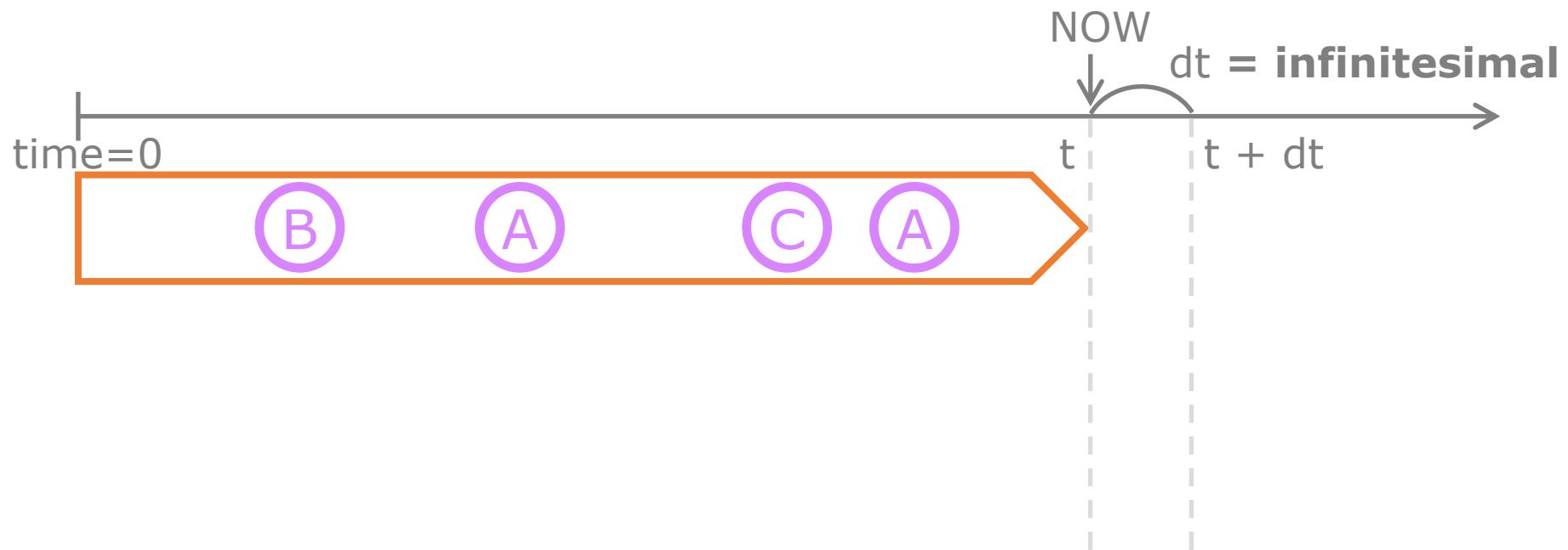
MLE: Max log prob of *data*



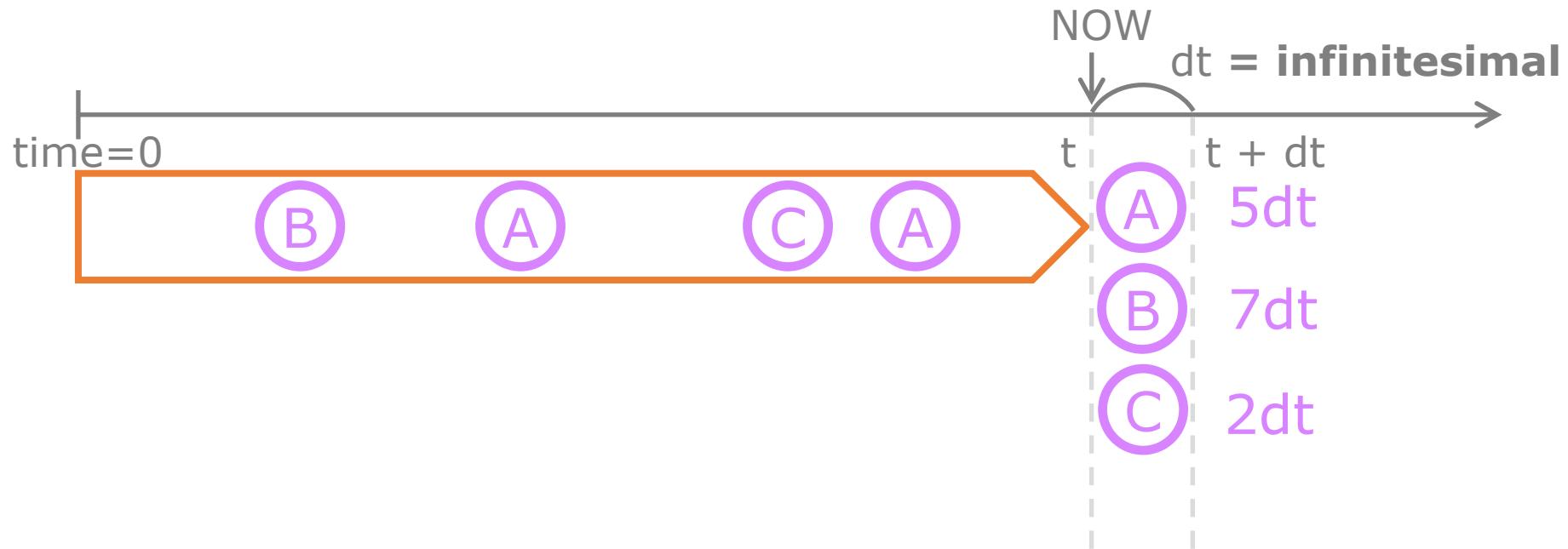
MLE: Max log prob of *data*



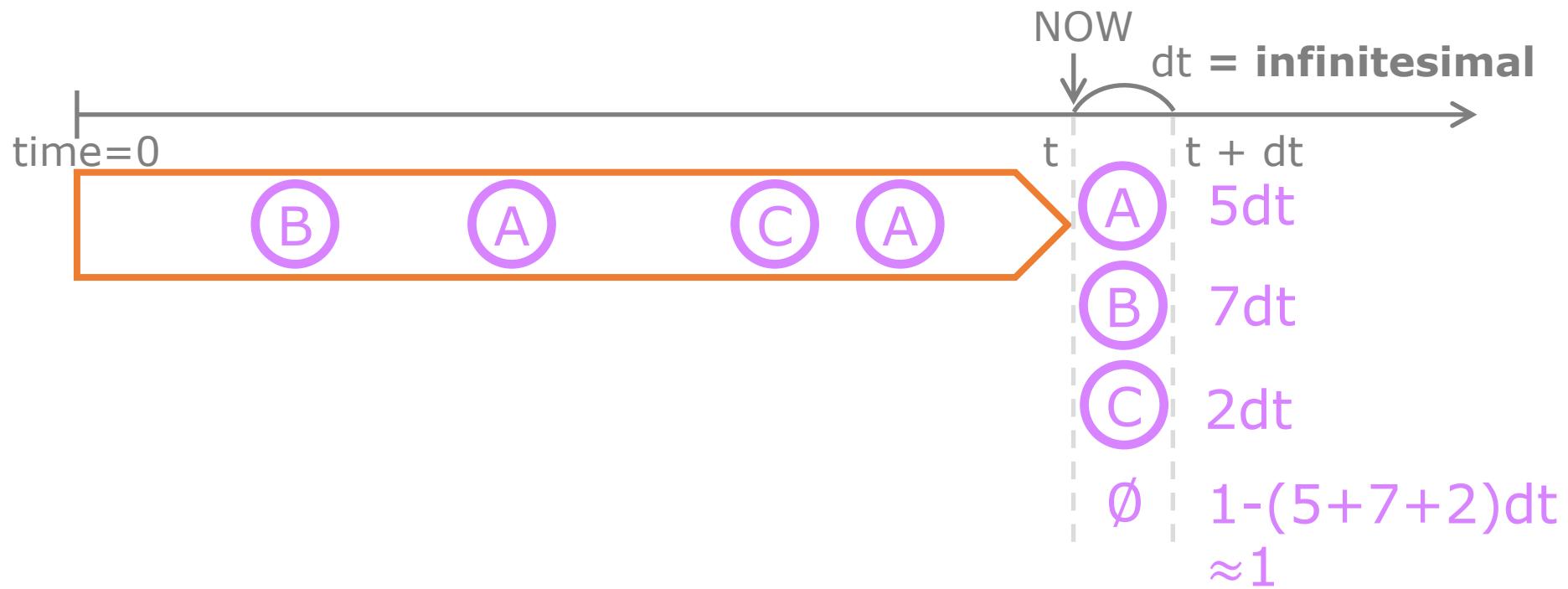
MLE: Max log prob of *data*



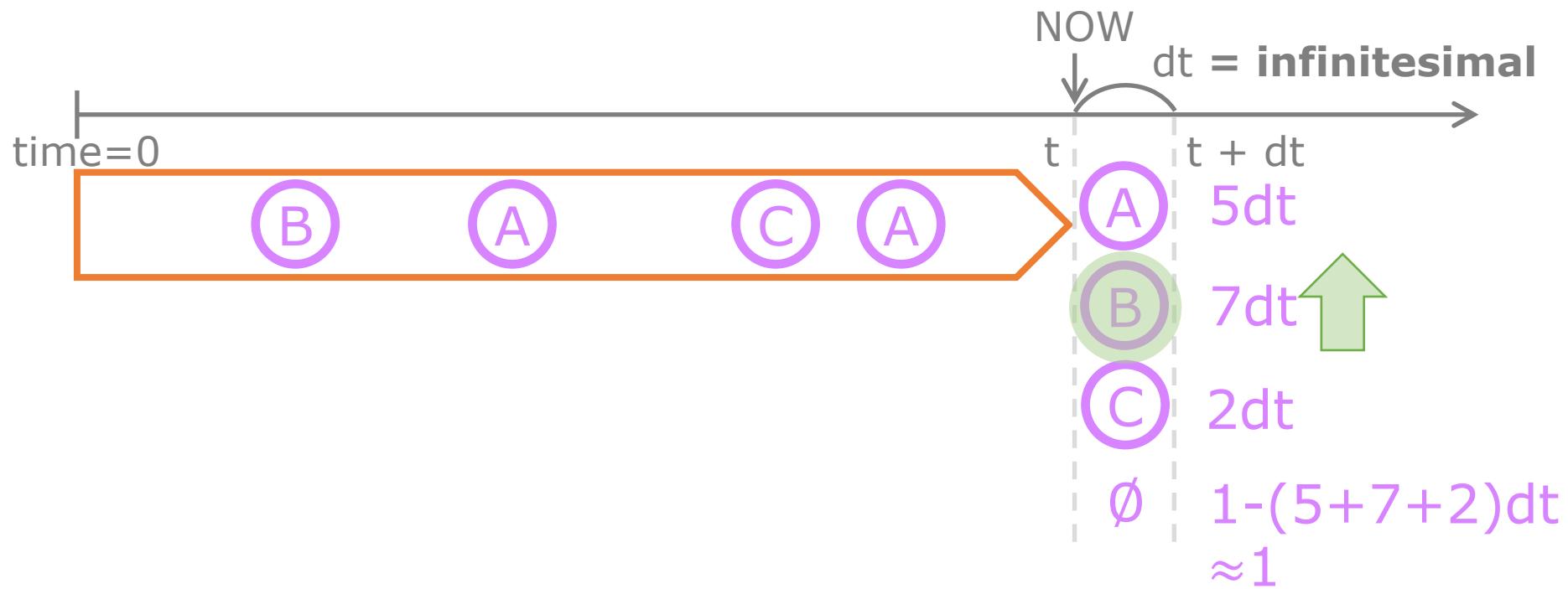
MLE: Max log prob of *data*



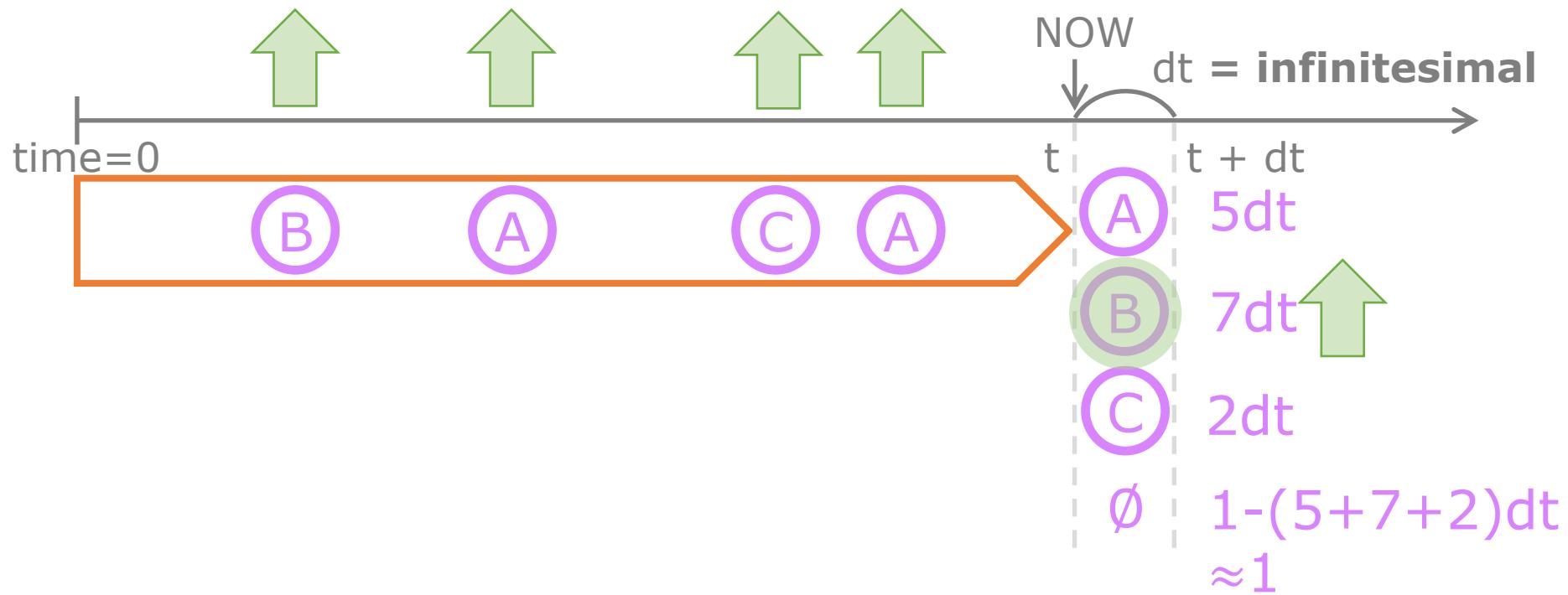
MLE: Max log prob of *data*



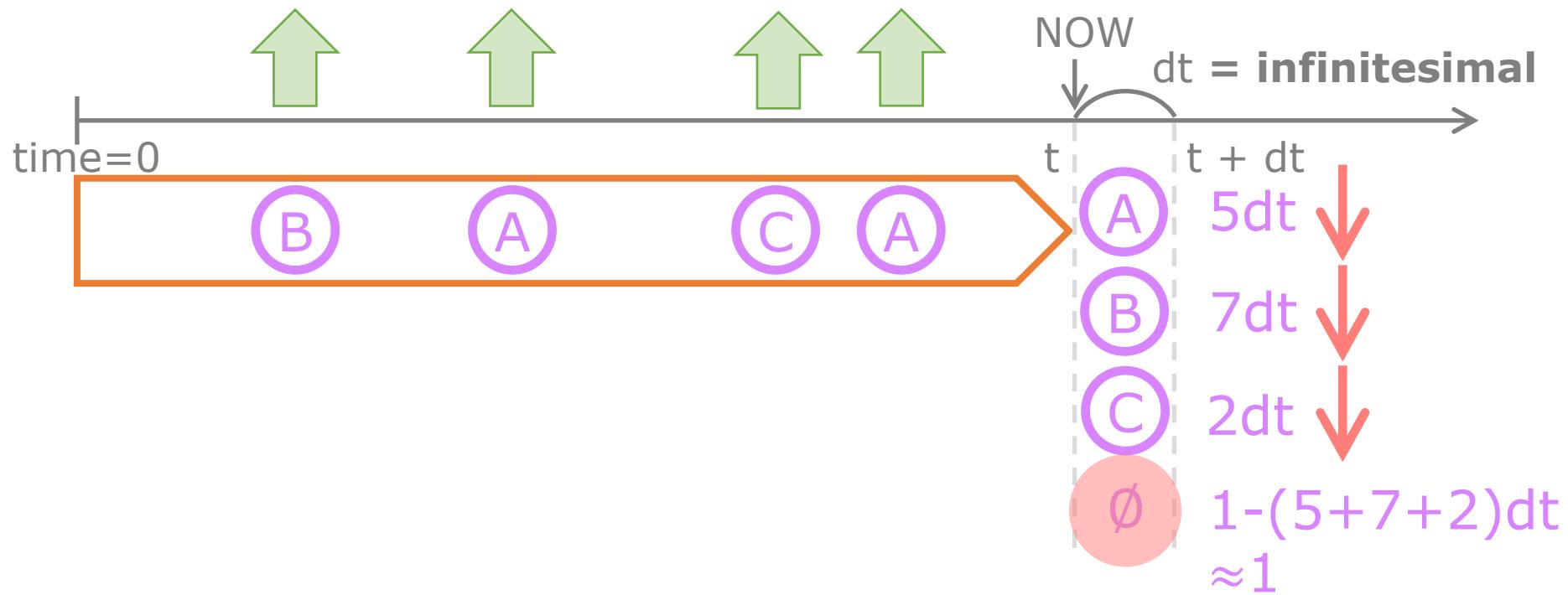
MLE: Max log prob of *data*



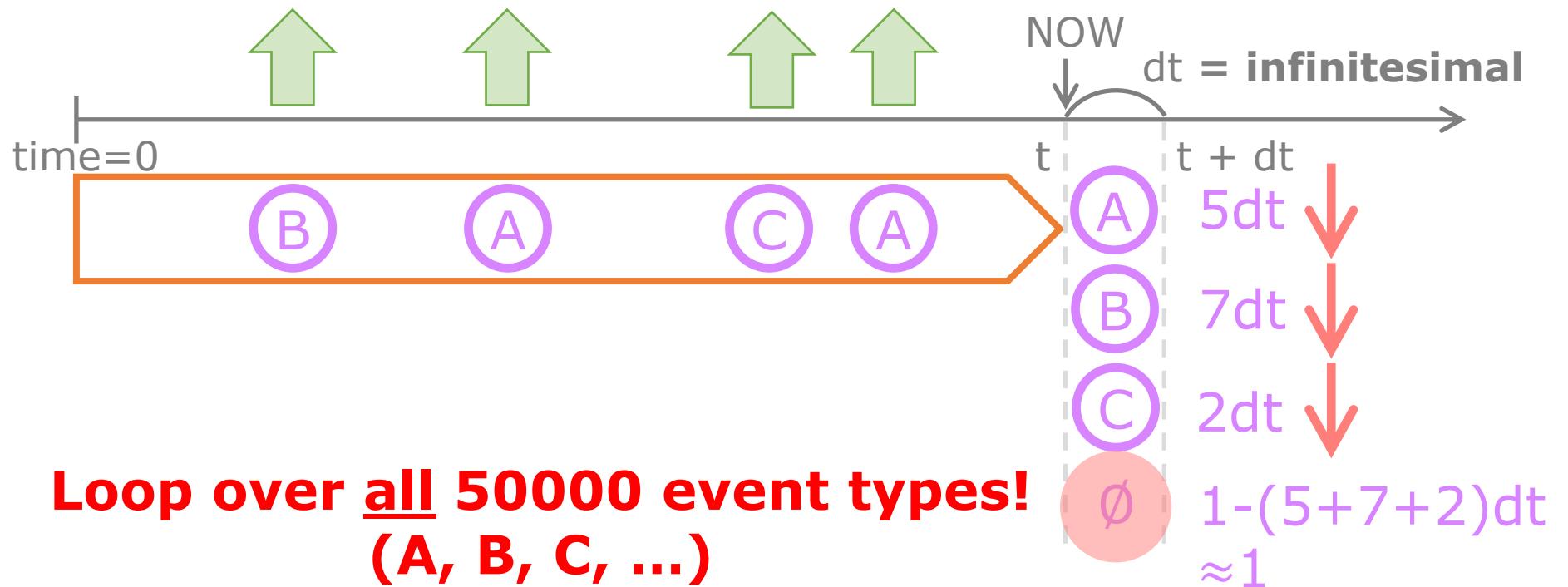
MLE: Max log prob of *data*



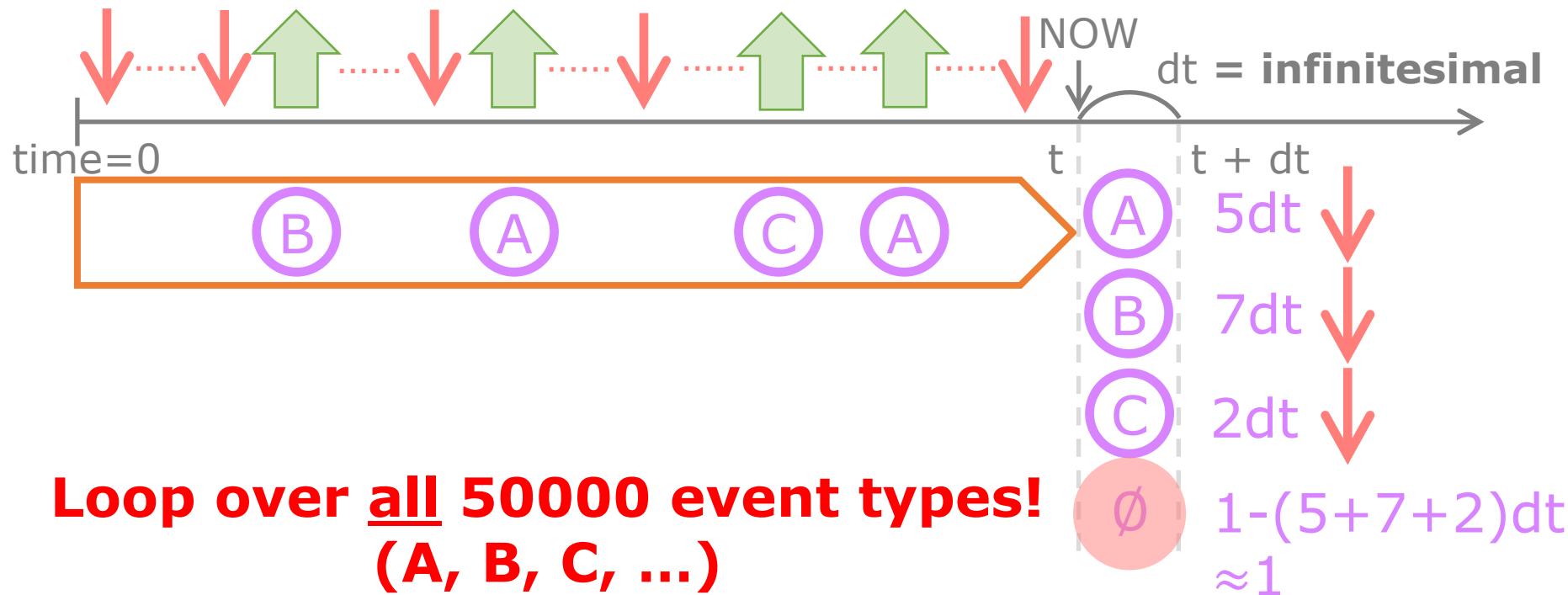
MLE: Max log prob of *data*



MLE: Max log prob of *data*

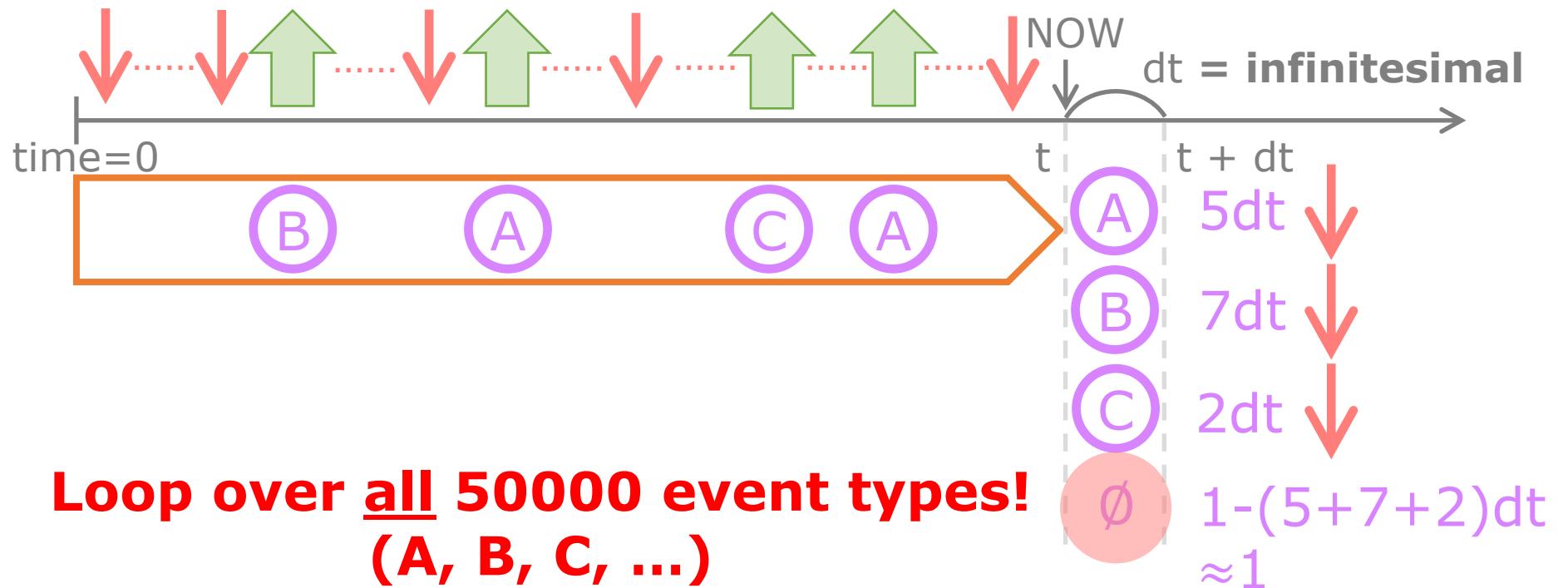


MLE: Max log prob of *data*



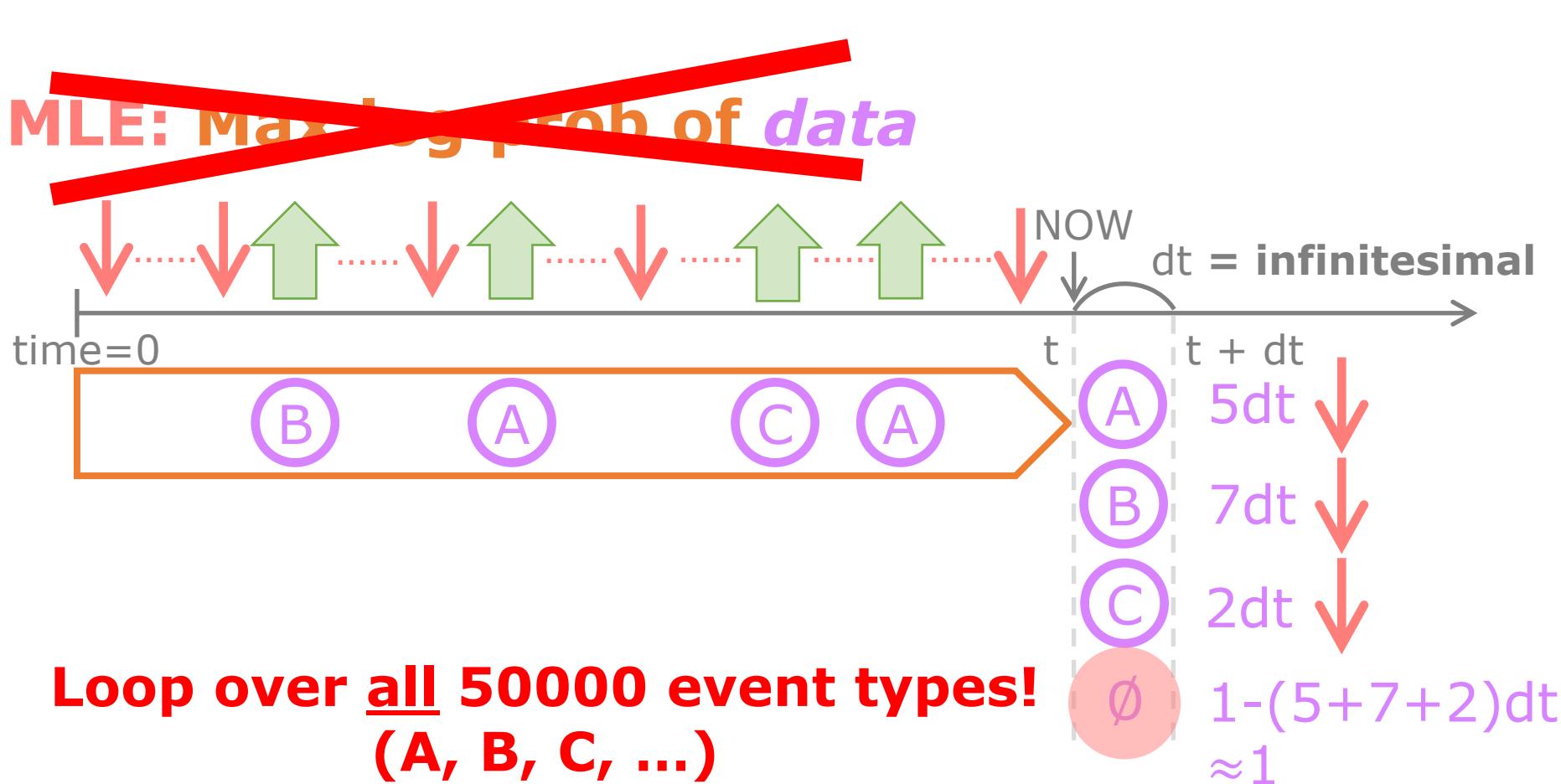
Integrate over infinitely many *non-events*!
(often approx by sampling)

MLE: Max log prob of *data*



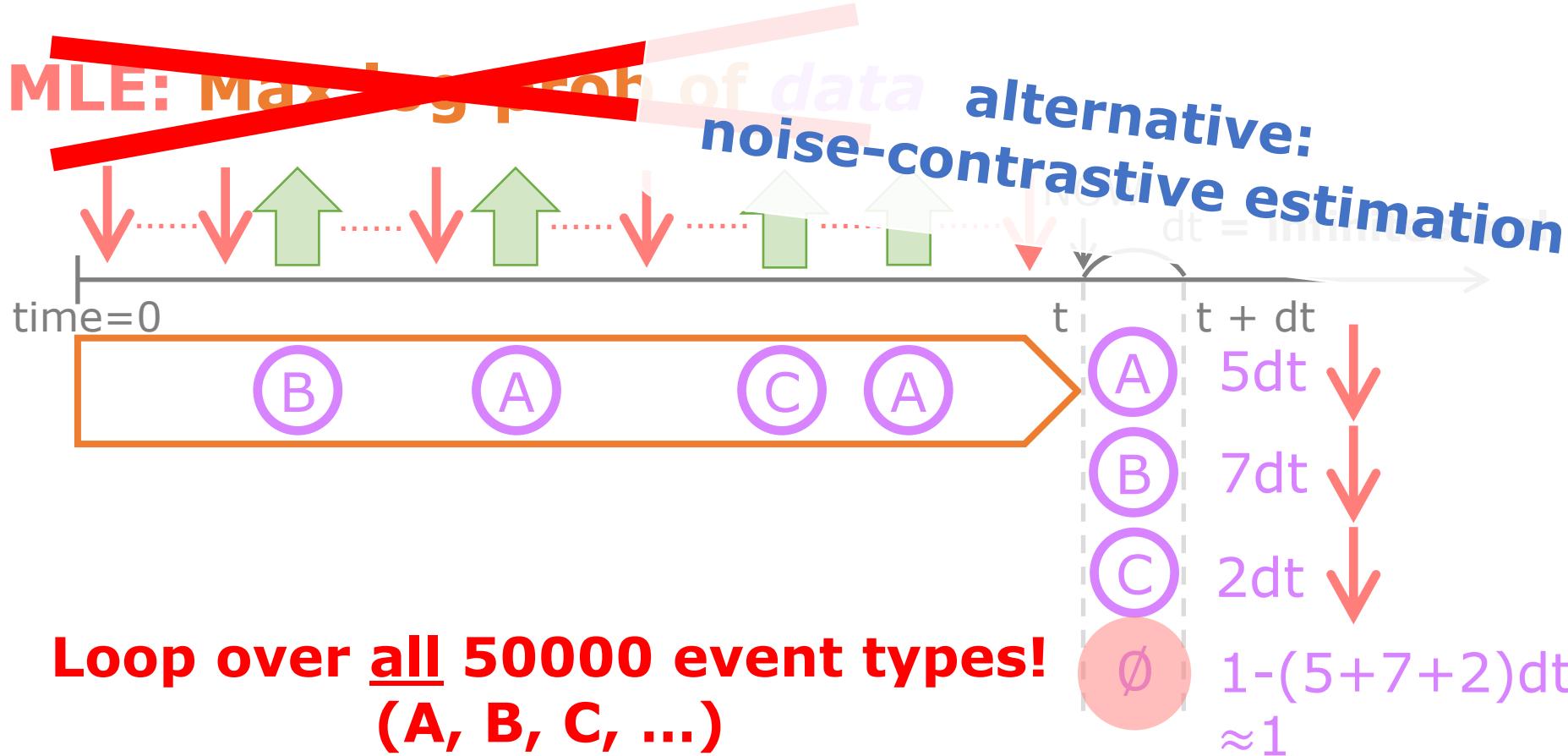
Integrate over infinitely many *non-events*!
(often approx by sampling)

SLOW



Integrate over infinitely many *non-events*!
(often approx by sampling)

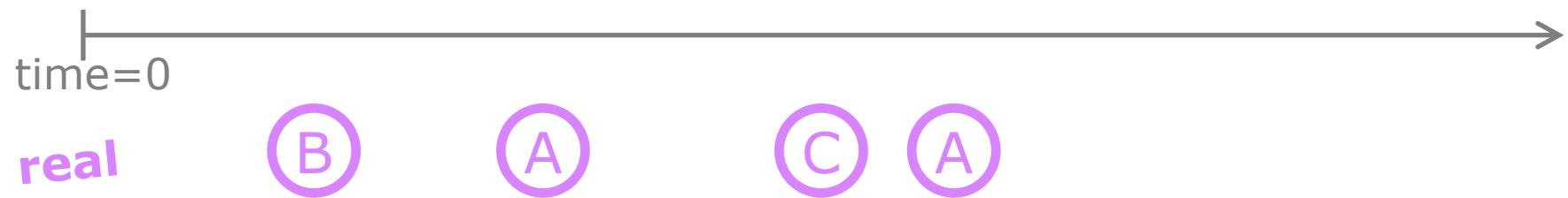
SLOW



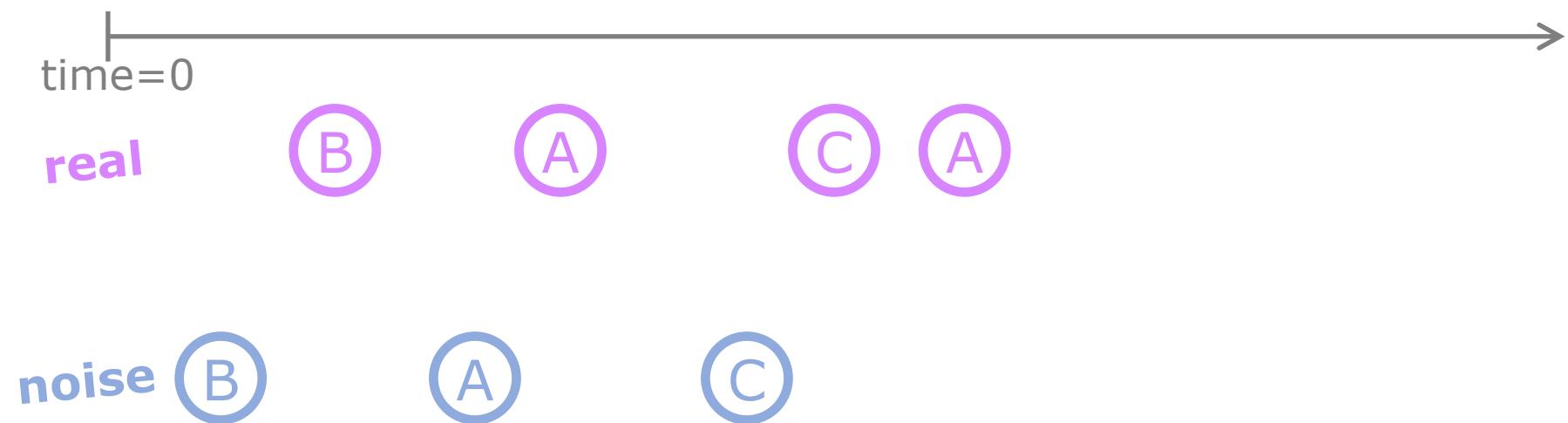
Integrate over infinitely many *non*-events!
(often approx by sampling)

SLOW

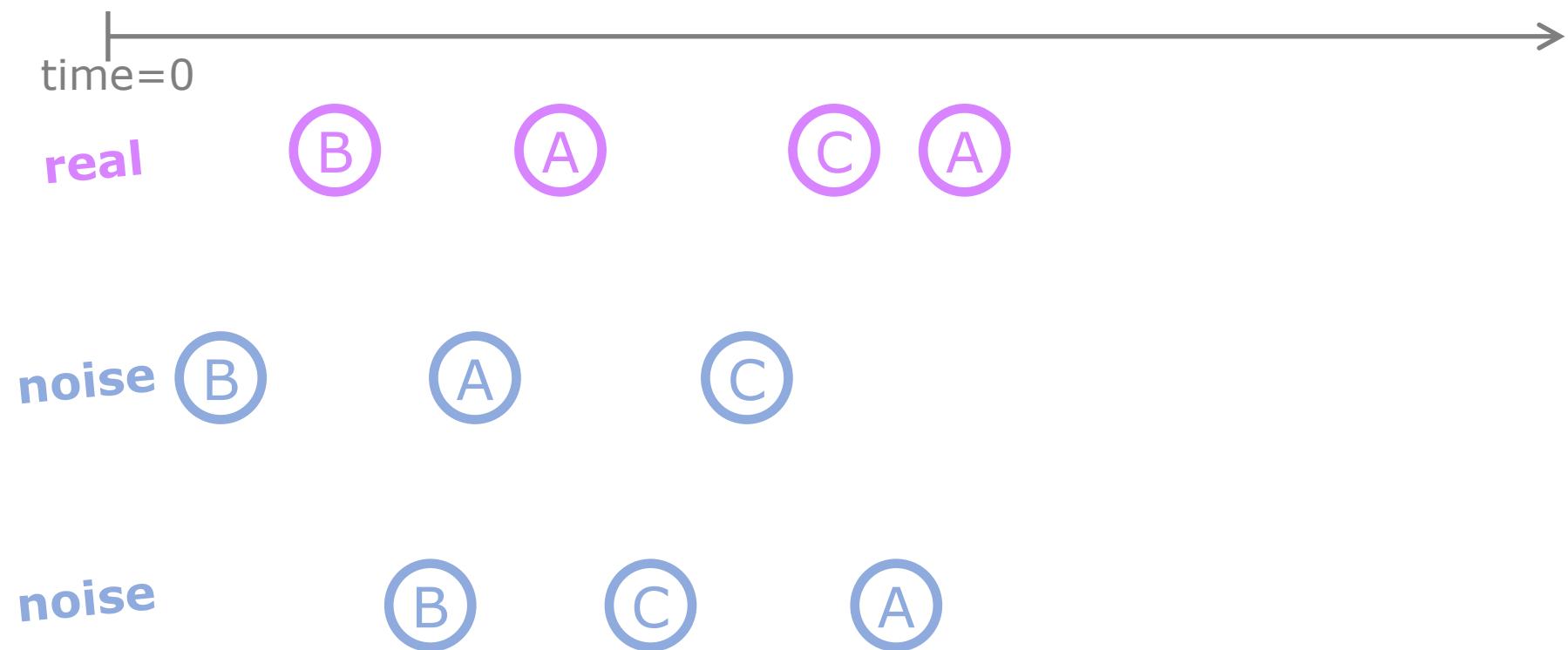
NCE: Max log prob of *correct discrimination*



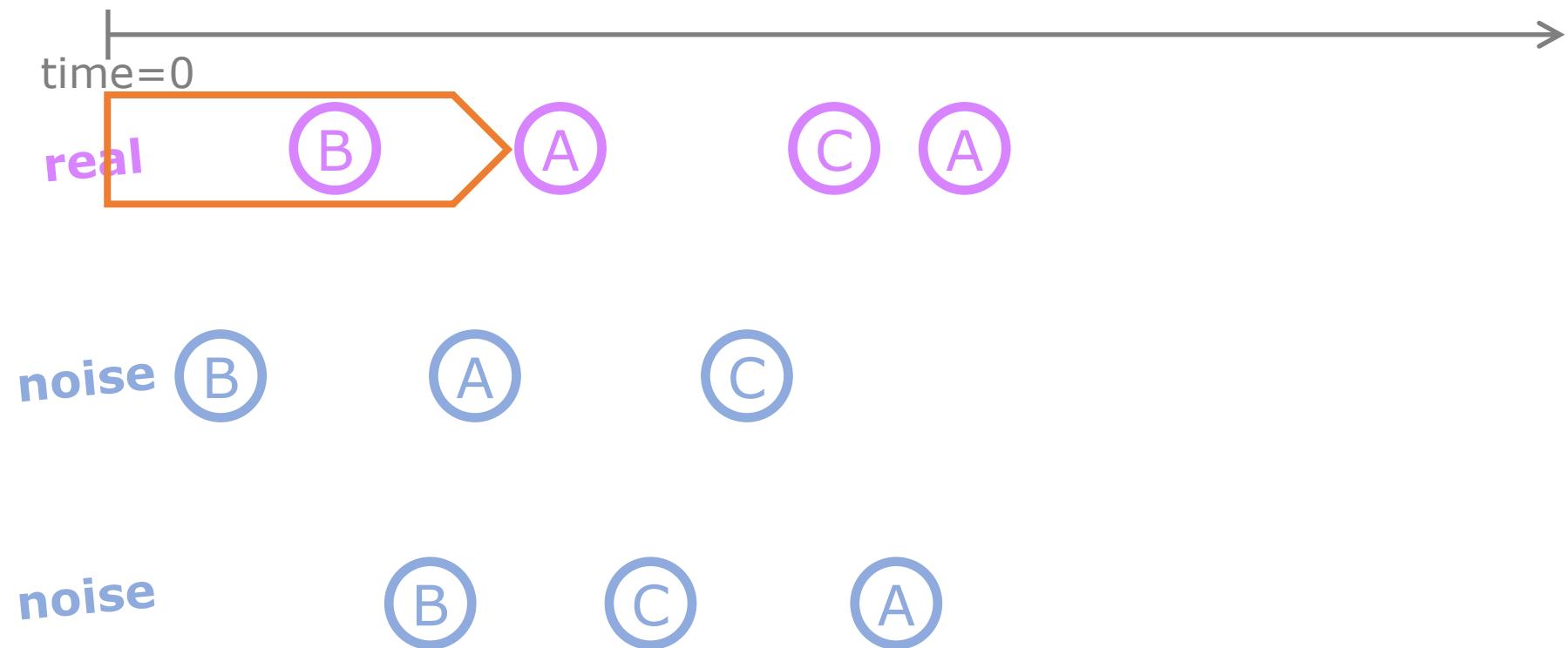
NCE: Max log prob of *correct discrimination*



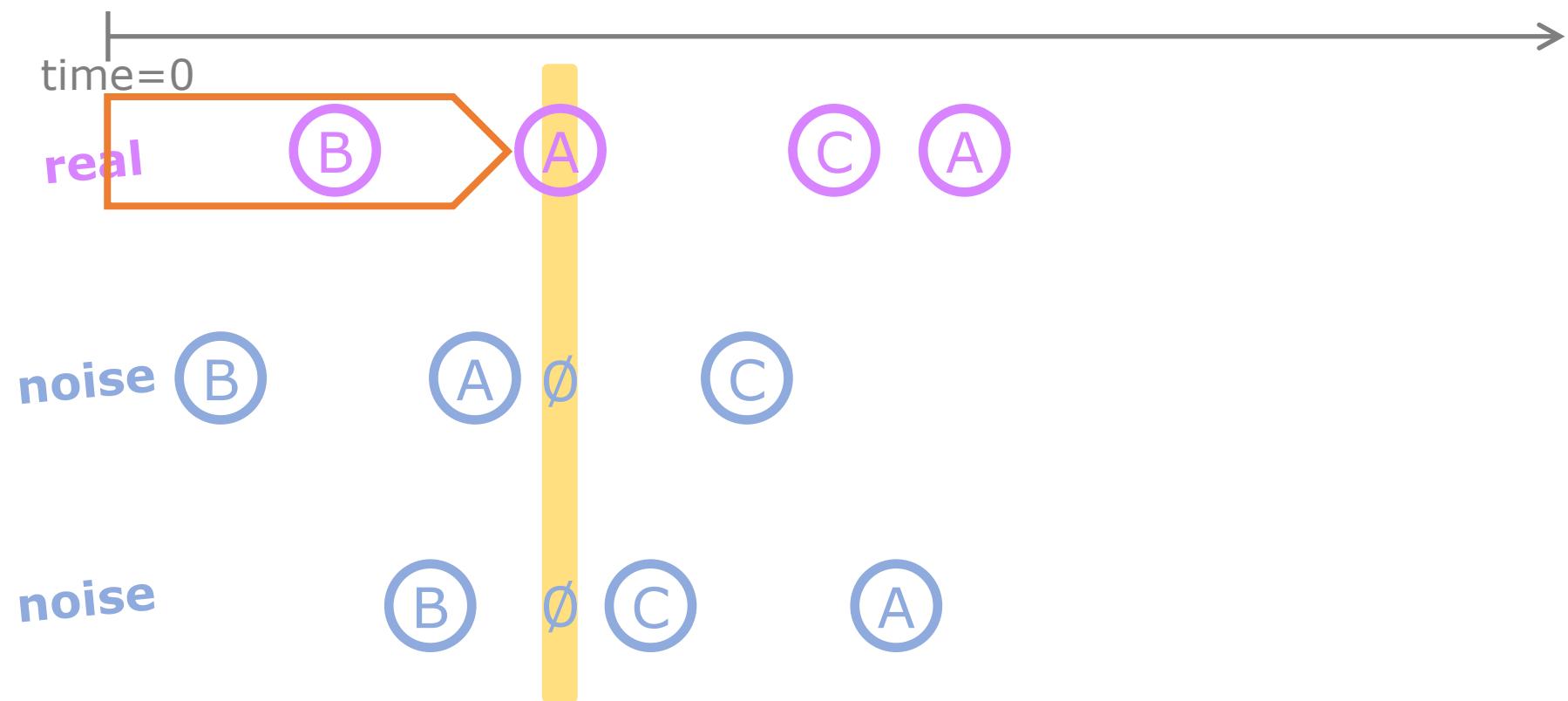
NCE: Max log prob of *correct discrimination*



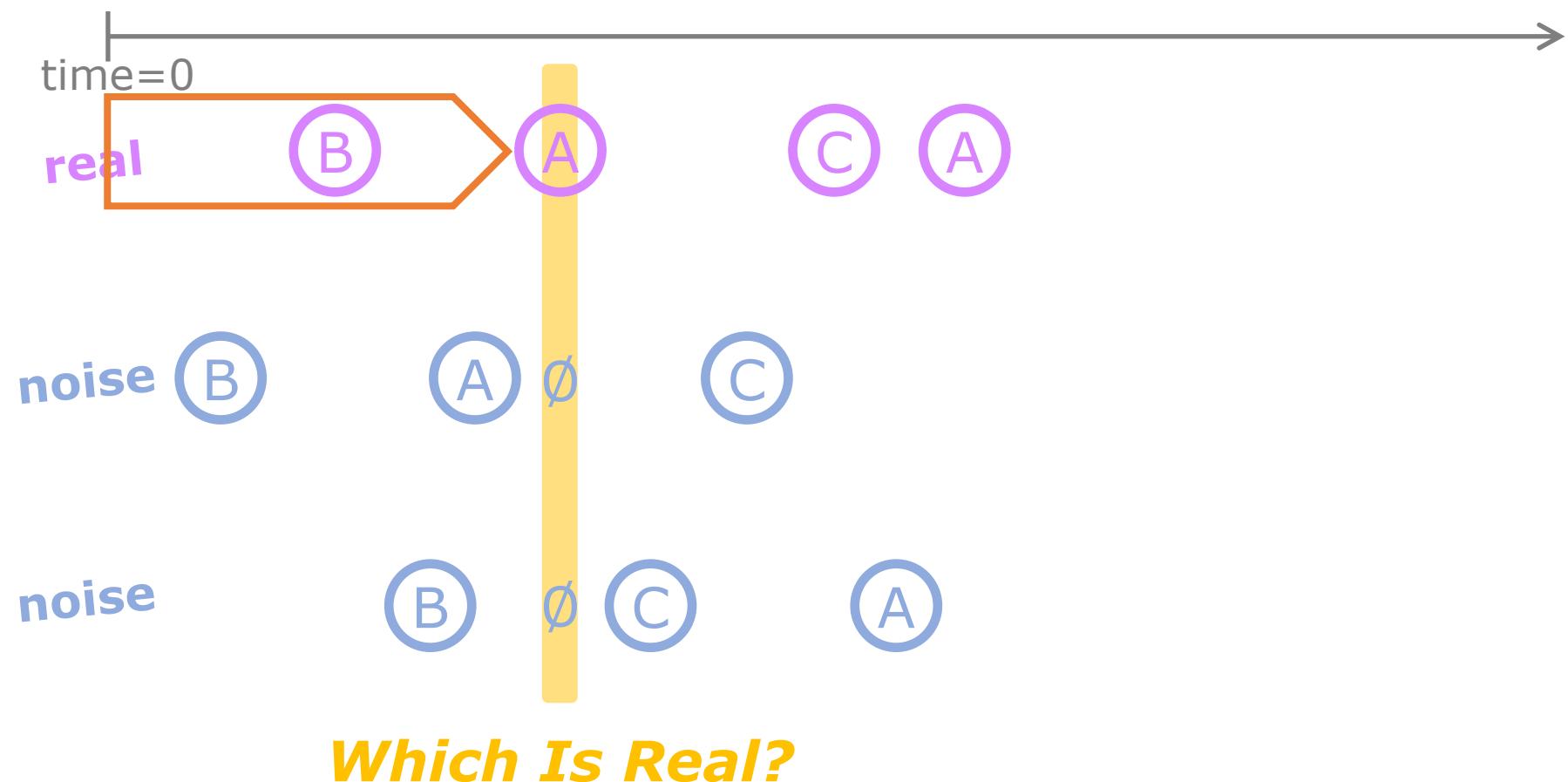
NCE: Max log prob of *correct discrimination*



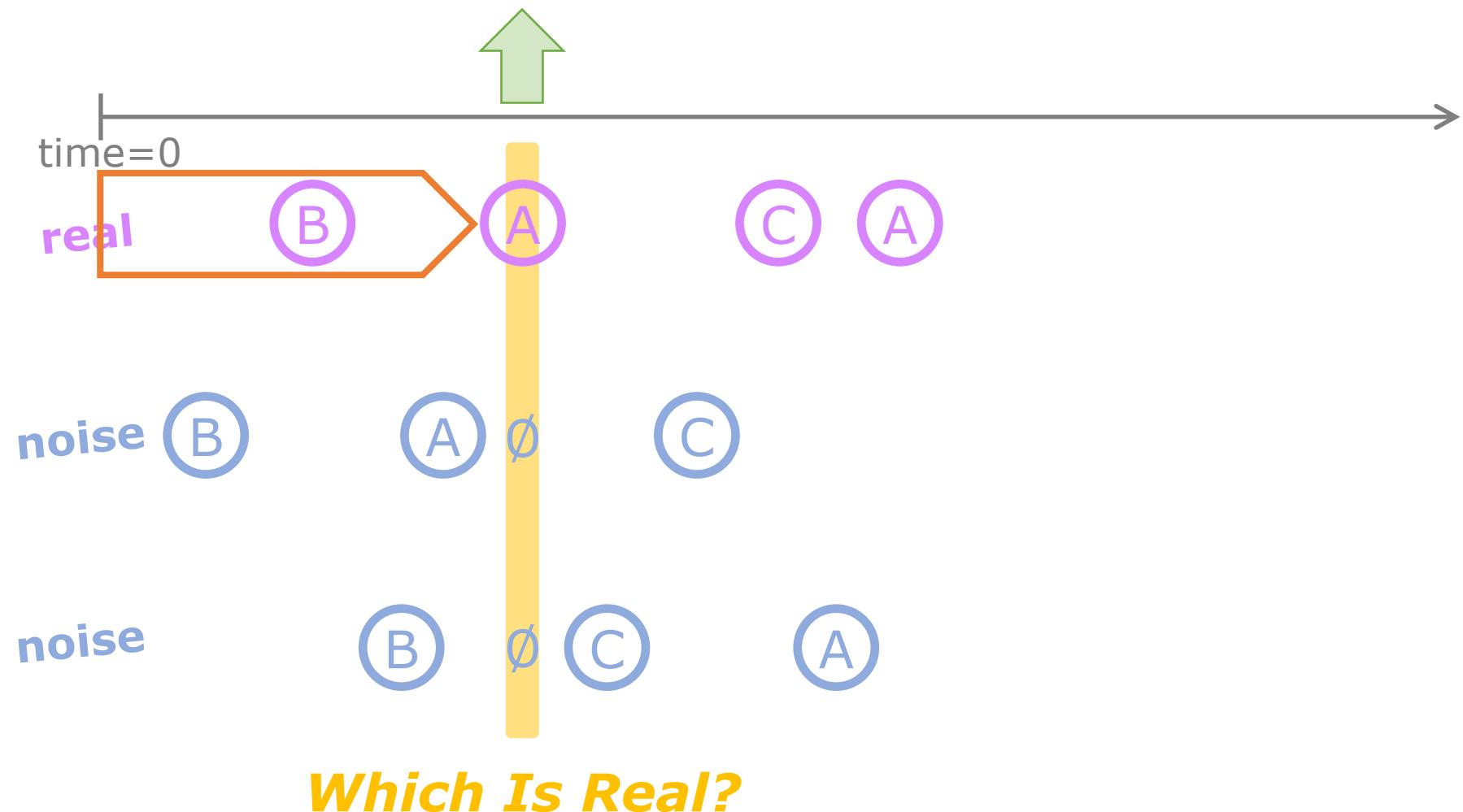
NCE: Max log prob of *correct discrimination*



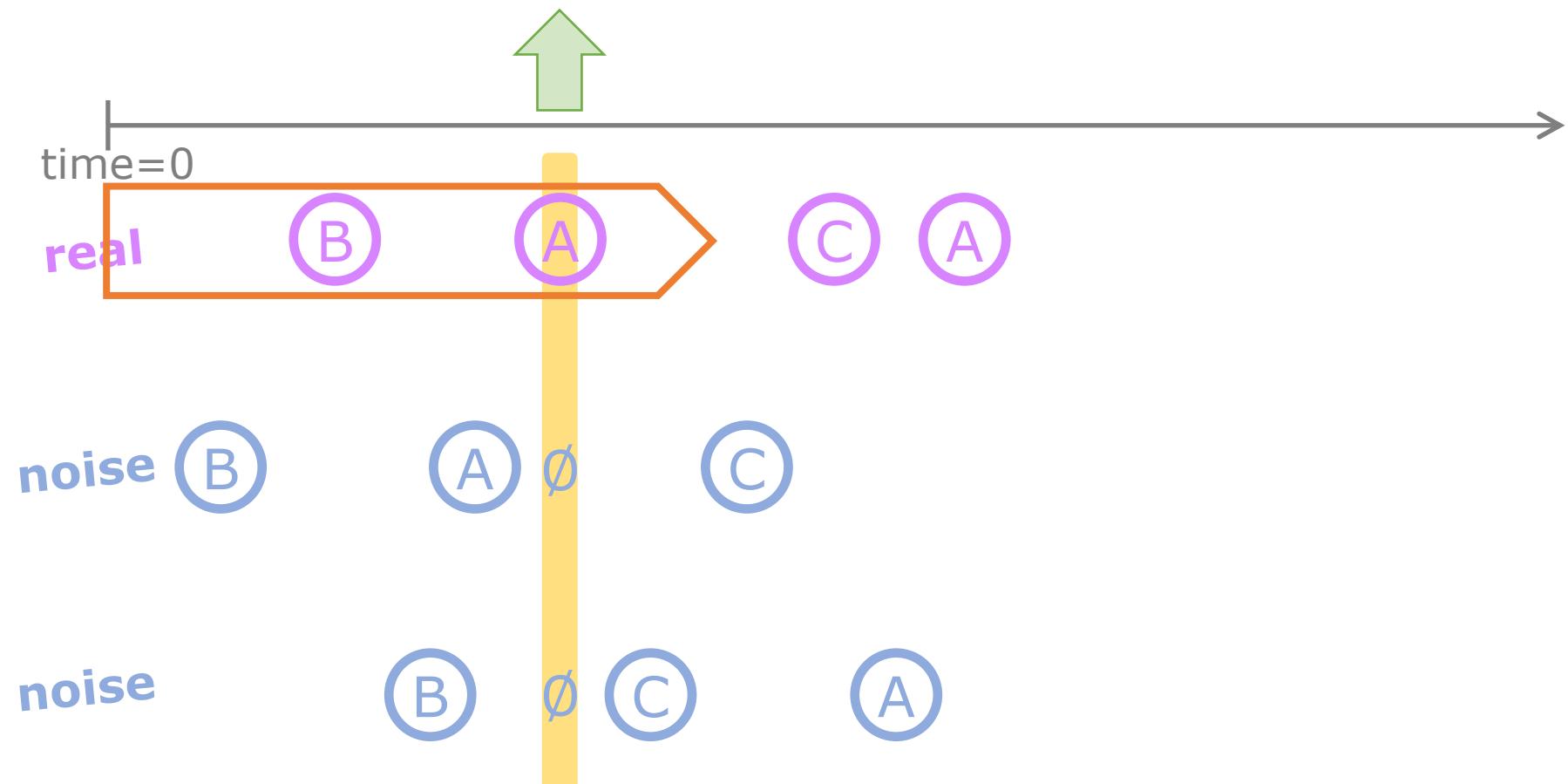
NCE: Max log prob of *correct discrimination*



NCE: Max log prob of *correct discrimination*

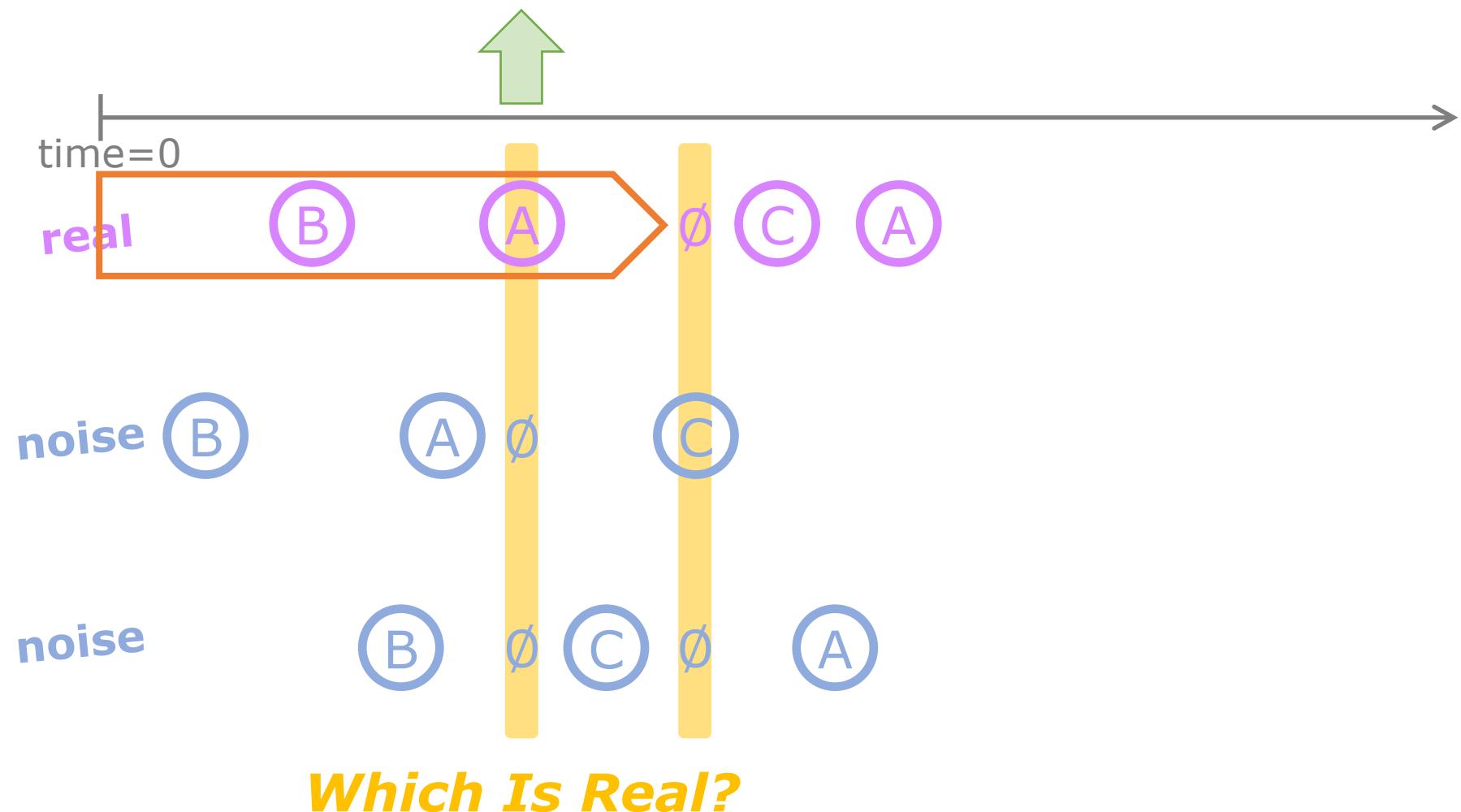


NCE: Max log prob of *correct discrimination*

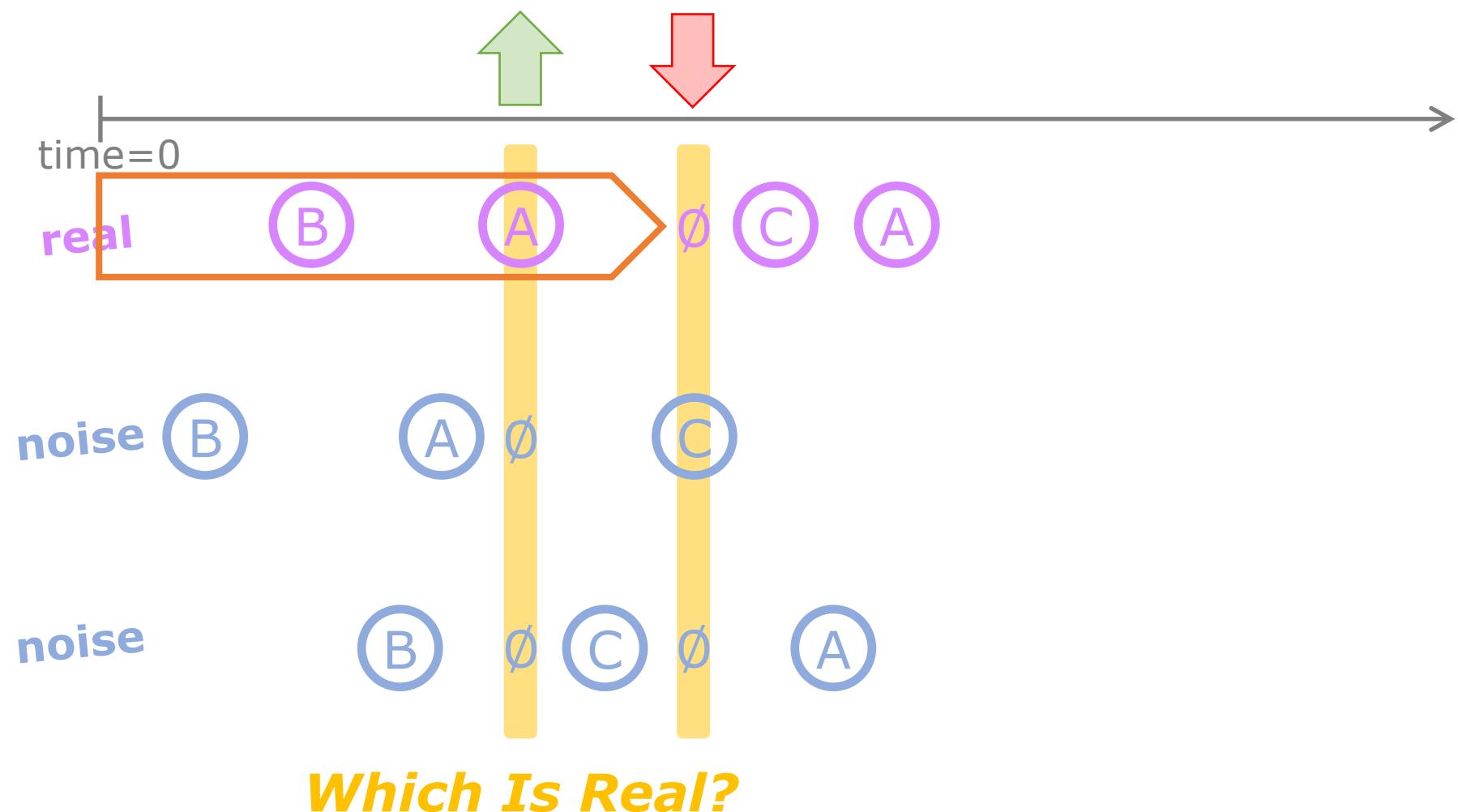


Which Is Real?

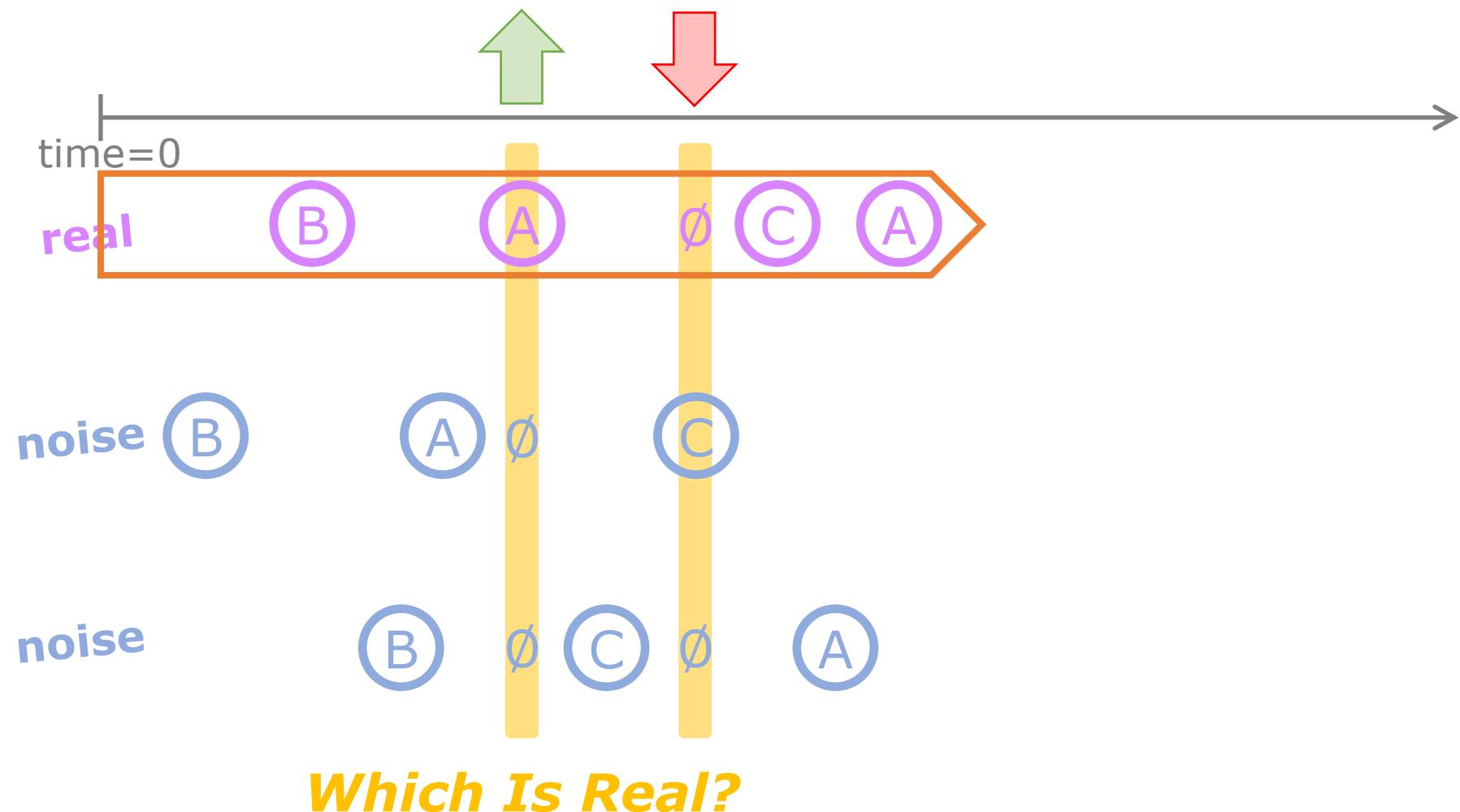
NCE: Max log prob of *correct discrimination*



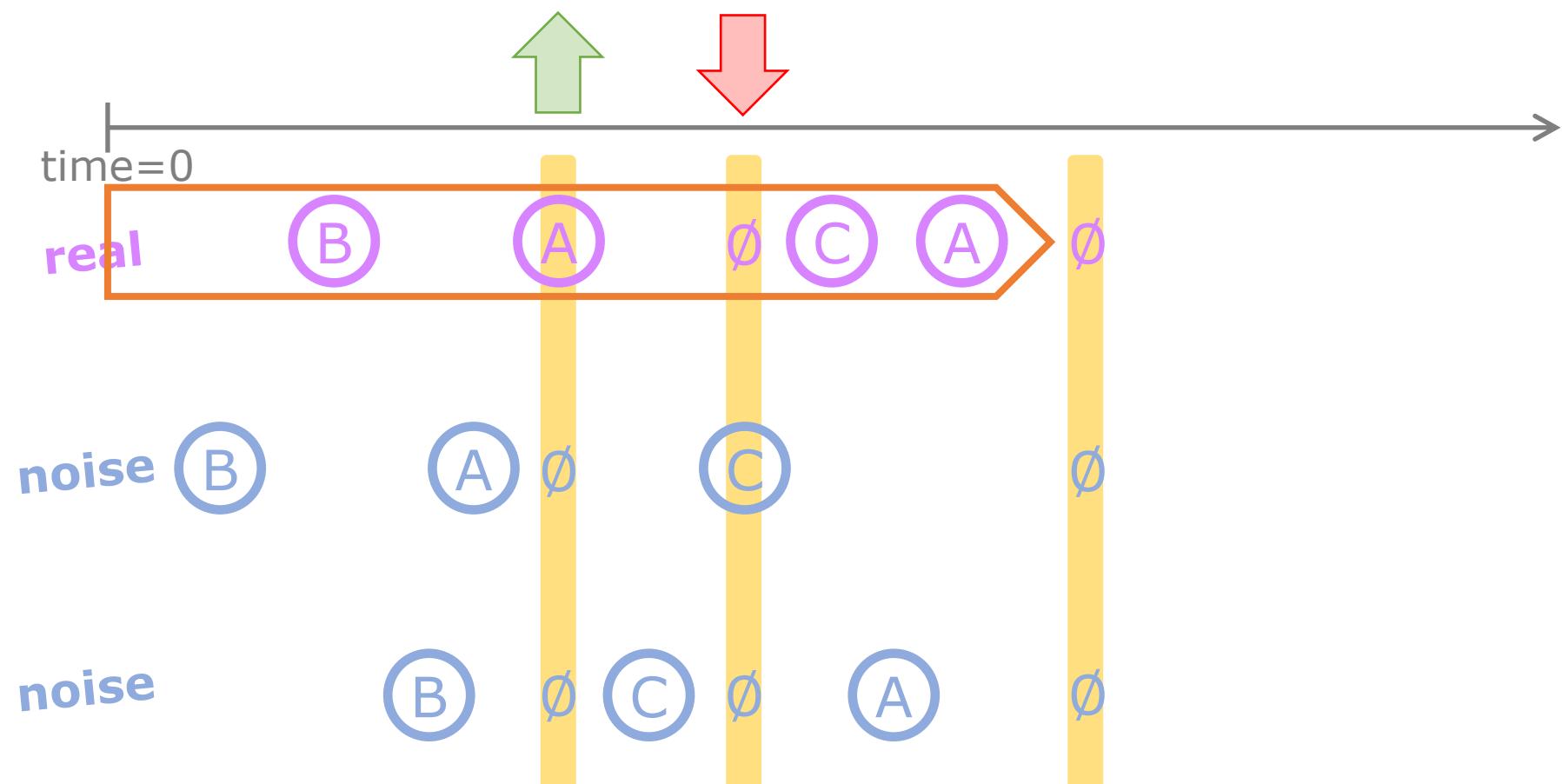
NCE: Max log prob of *correct discrimination*



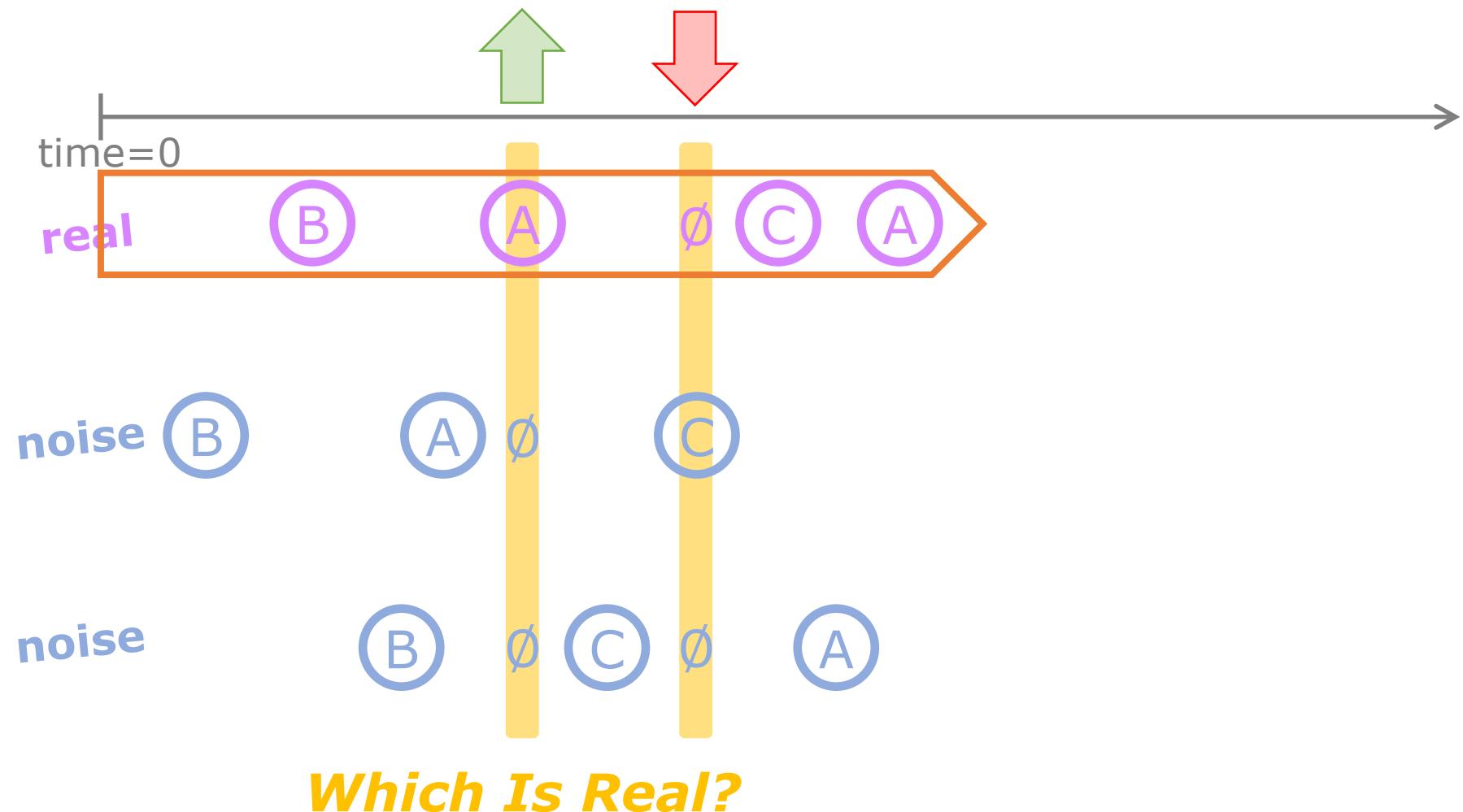
NCE: Max log prob of *correct discrimination*



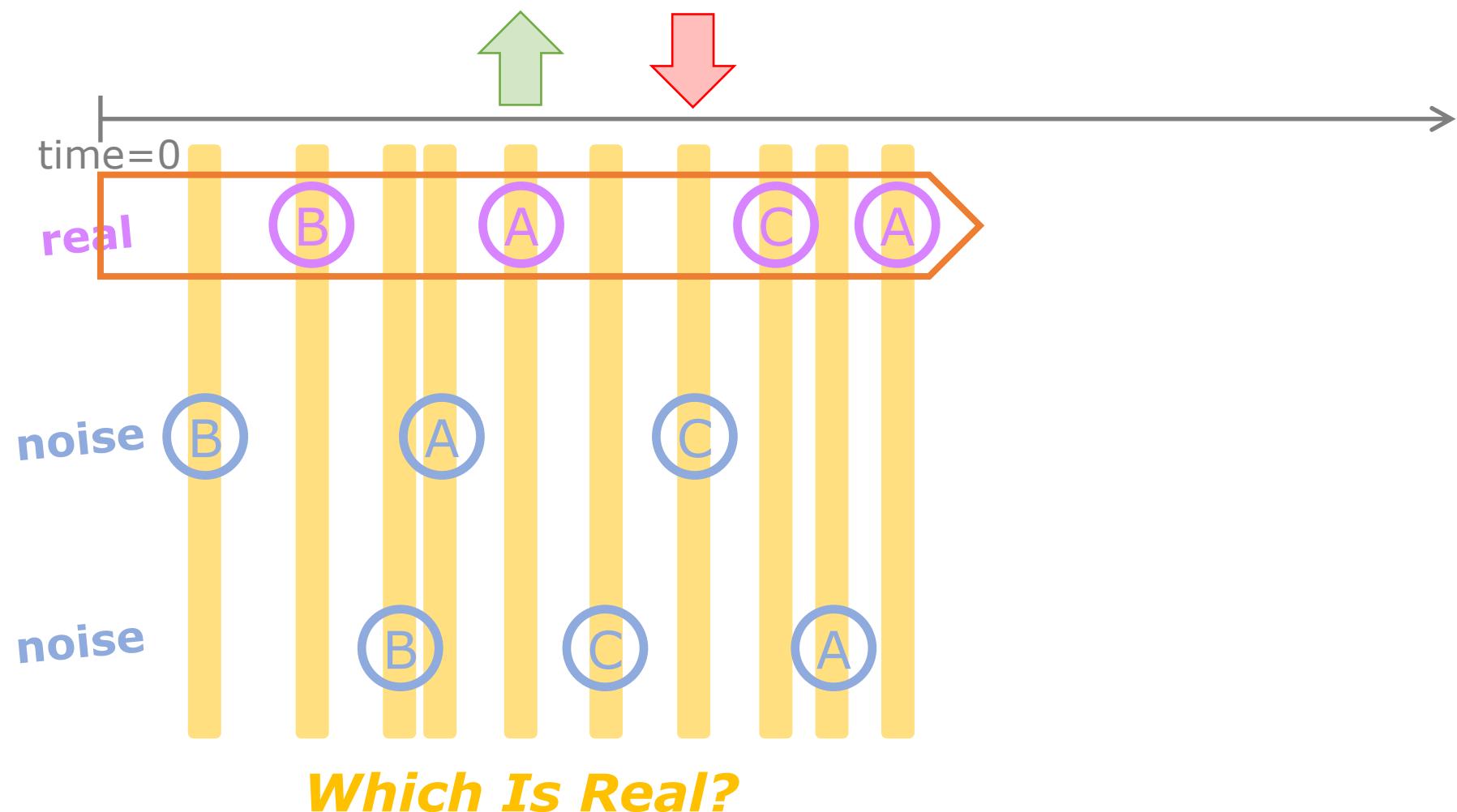
NCE: Max log prob of *correct discrimination*



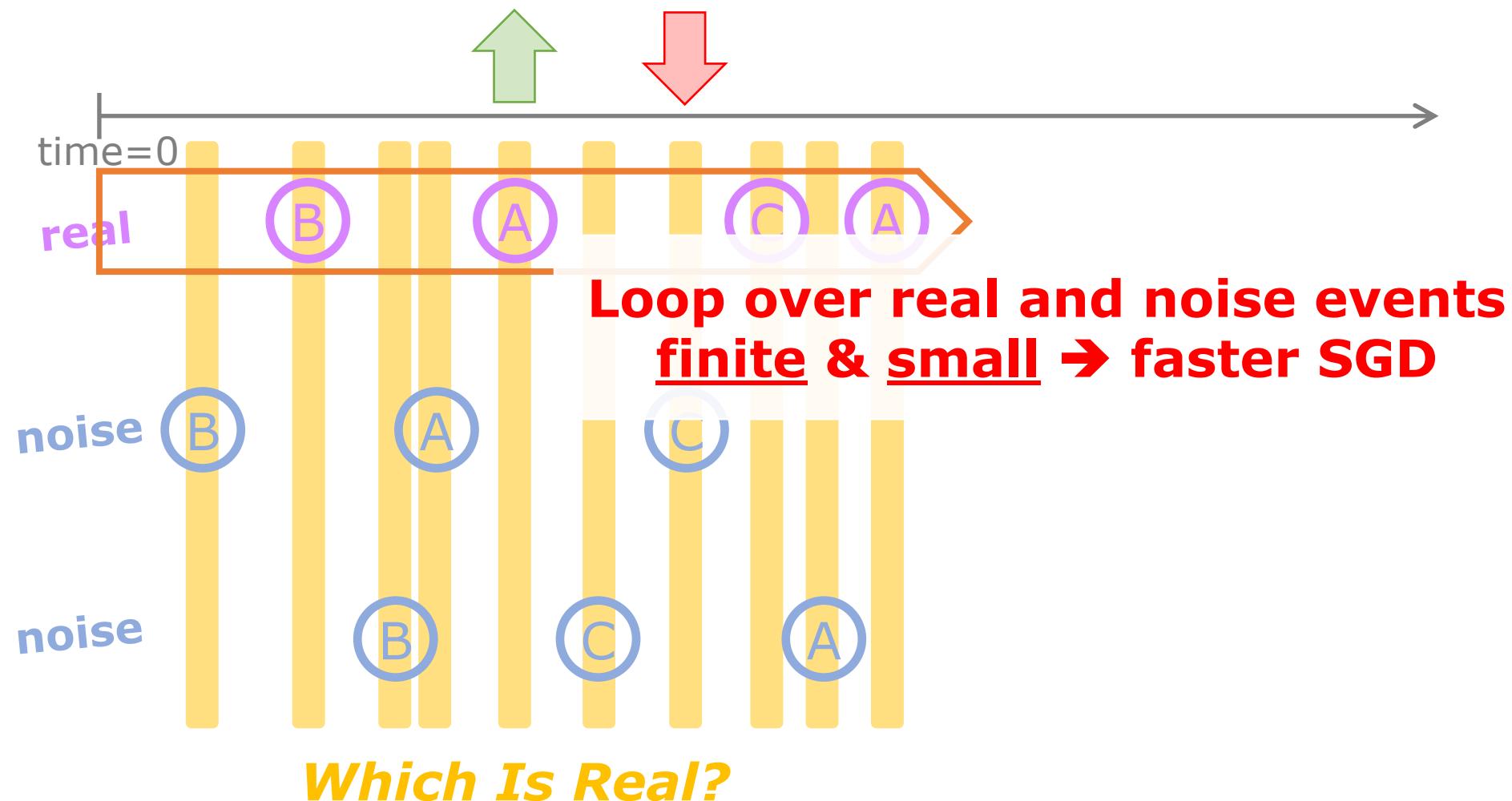
NCE: Max log prob of *correct discrimination*



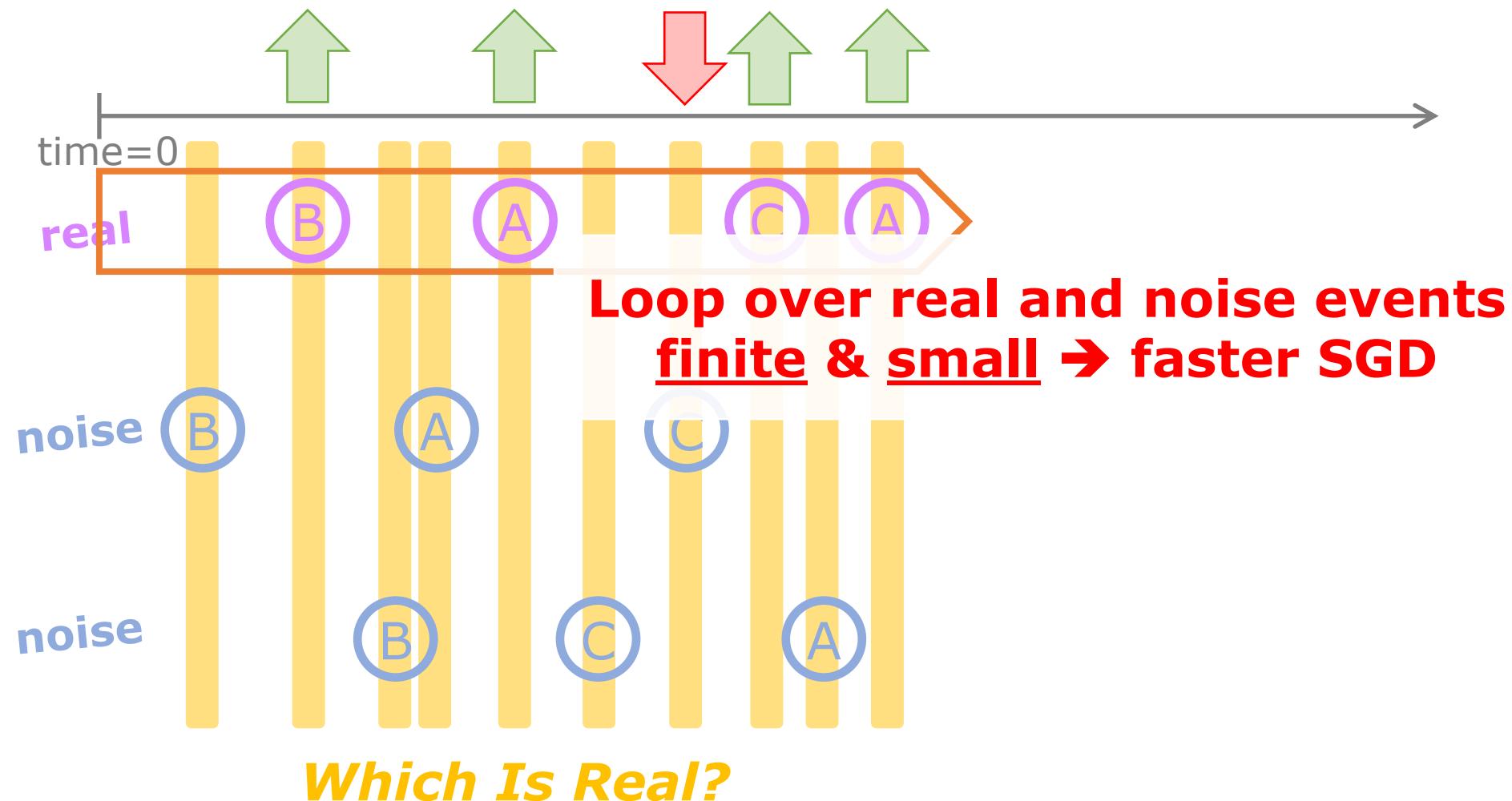
NCE: Max log prob of *correct discrimination*



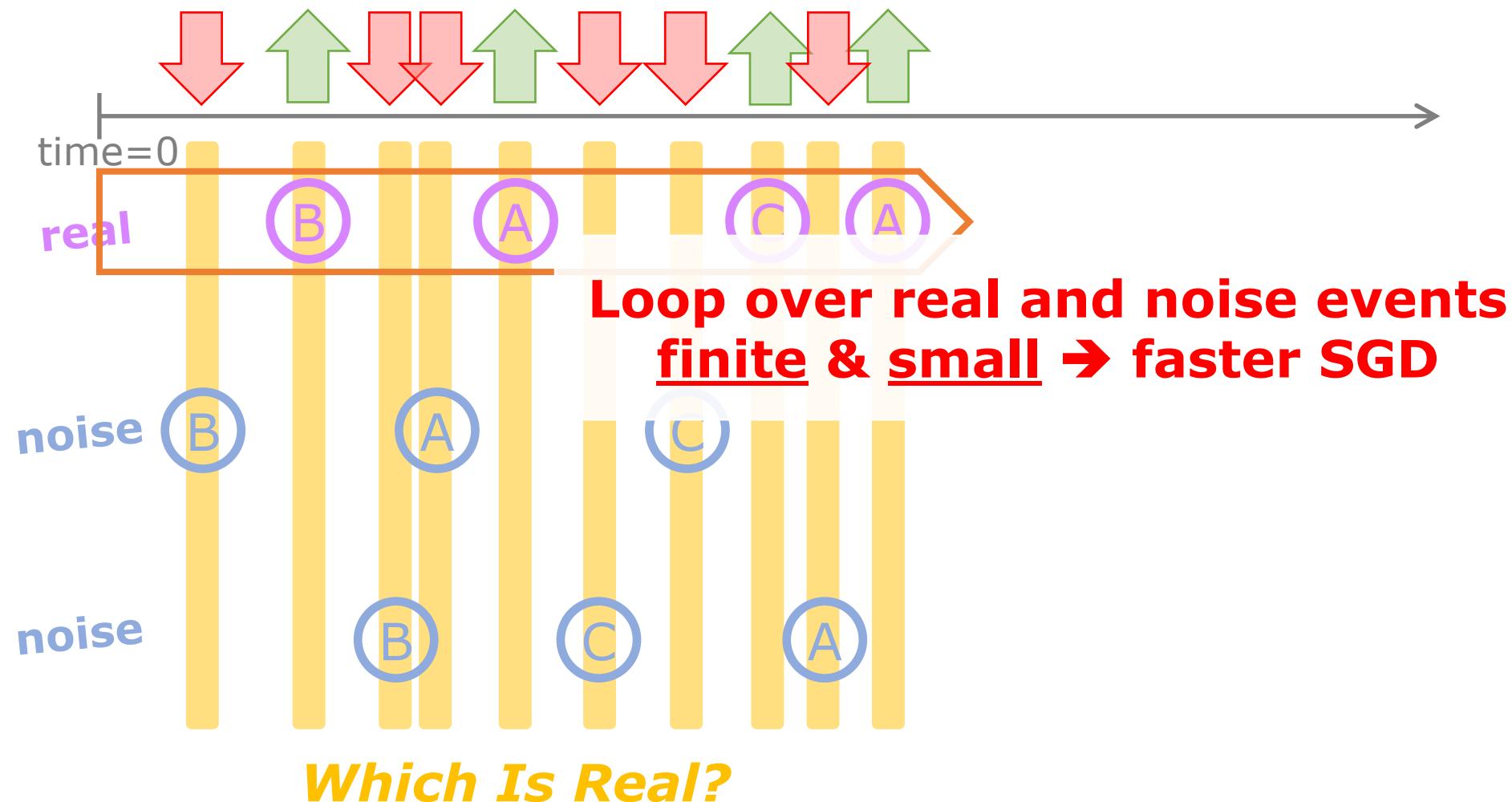
NCE: Max log prob of *correct discrimination*



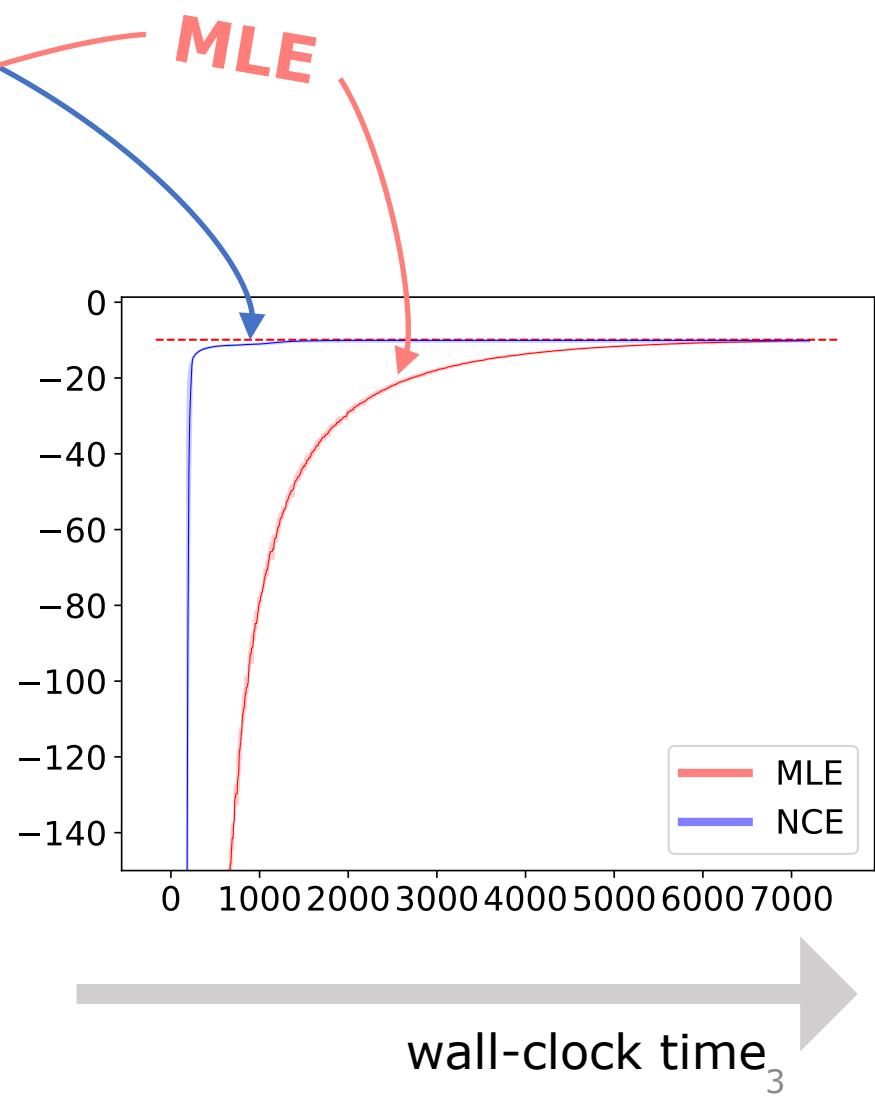
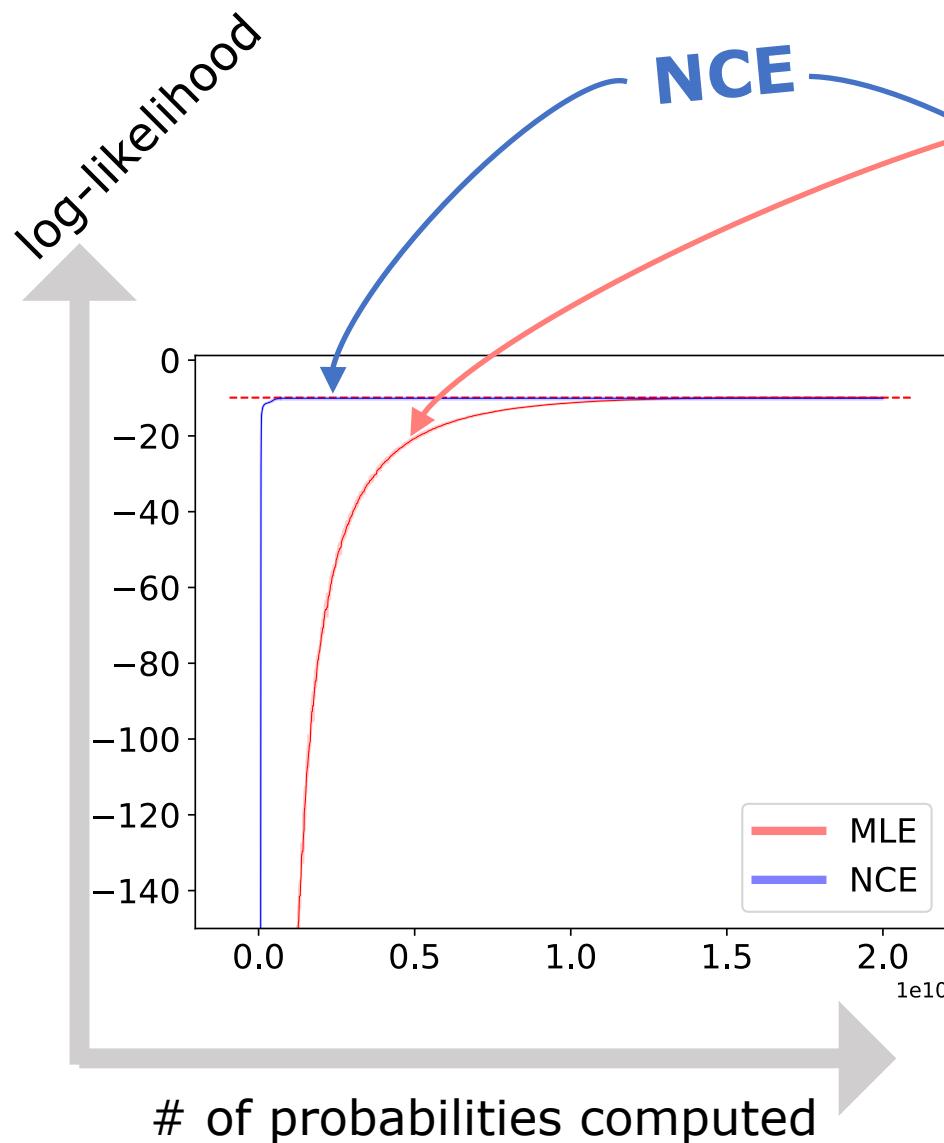
NCE: Max log prob of *correct discrimination*



NCE: Max log prob of *correct discrimination*



NCE vs MLE: what it typically looks like



NCE: More in paper

NCE: More in paper

Theorem 1 (Optimality). *Under assumptions 1 and 2, $\theta \in \operatorname{argmax}_\theta J_{NC}(\theta)$ if and only if $p_\theta = p^*$.*

We first need to highlight the key insight that $H_\theta(k, t, x_{[0,t]}^0)$ in equation (20) is the negative cross-entropy between the following two discrete distributions over $\{\emptyset, 1, \dots, K\}$:

$$\left[\frac{\lambda_k^*(t|x_{[0,t]}^0)}{\Delta_k^*(t|x_{[0,t]}^0)}, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k^*(t|x_{[0,t]}^0)}, \dots, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k^*(t|x_{[0,t]}^0)} \right] \quad (21a)$$

$$\underbrace{\left[\frac{\lambda_k(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)}, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)}, \dots, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)} \right]}_{\text{length } N} \quad (21b)$$

theorems & proofs

Theorem 2 (Consistency). *Under assumptions 1 and 4, for any $\theta \in \operatorname{argmax}_\theta J_{NC}(\theta)$ and $M \geq 1$, with probability 1, we have $J_{NC}^N(\theta) \rightarrow J_{NC}(\theta)$ as $N \rightarrow \infty$. The proof of Theorem 2 is almost identical to the proof of Theorem 1 in Ma & Collins (2018). But we will still spell it out in our notation for completeness.*

The intuition of this theorem is the same as $N \rightarrow \infty$ and they are continuous with respect to θ . Theorem 2 shows that $J_{NC}^N(\theta)$ and $J_{NC}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous with respect to θ . Theorem 2 shows that $J_{NC}^N(\theta)$ and $J_{NC}(\theta)$ has to be the same member of the set $\operatorname{argmax}_\theta J_{NC}(\theta)$. The proof of Theorem 2 is almost identical to the proof of Theorem 1 in Ma & Collins (2018). But we will still spell it out in our notation for completeness.

Theorem 3 (Efficiency). *Under assumptions 2 and 4–7, there exists an integer \bar{M} such that for all $M > \bar{M}$, $\mathbb{E}[J_{NC}^N(\theta) - J_{NC}(\theta)] = 0$ and $\mathbb{E}[J_{NC}^N(\theta) - J_{NC}(\theta)] \geq J_{NC}^N(\theta) - J_{NC}(\theta)$, we have*

The intuition of this theorem is the same as $N \rightarrow \infty$ and they are continuous with respect to θ . Theorem 3 shows that $J_{NC}^N(\theta)$ and $J_{NC}(\theta)$ will become the same member of the set $\operatorname{argmax}_\theta J_{NC}(\theta)$. The proof of Theorem 3 is almost identical to the proof of Theorem 2 in Ma & Collins (2018). But we will still spell it out in our notation for completeness.

Theorem 4 (Large Sample Theory). *Under assumptions 2 and 4–7, by classical large sample theory (Ferguson, 1996), we have*

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{D} \operatorname{Normal}(0, I_M^{-1}) \quad (25)$$

for some non-singular matrix I_M^{-1} . Moreover, there exist a constant $C > 0$ such that for all $M > \bar{M}$,

where $\|I\|$ is the spectral norm of matrix I

$$\|I_M^{-1} - I_*^{-1}\| \leq C/M \quad (39)$$

Proof. We first prove that $\sqrt{N}(\hat{\theta} - \theta^*)$ is asymptotically normal. By the Mean-Value Theorem, we have

$$\nabla_{\theta} J_{NC}^N(\hat{\theta}) = \nabla_{\theta} J_{NC}^N(\theta^*) + (\hat{\theta} - \theta^*) \int_{u=0}^1 \nabla_{\theta}^2 J_{NC}^N(\theta^* + u(\hat{\theta} - \theta^*)) dt \quad (41)$$

NCE: More in paper

Theorem 1 (Optimality). Under assumptions 1 and 2, $\theta \in \operatorname{argmax}_\theta J_{\text{NC}}(\theta)$ if and only if $p_\theta = p^*$.

We first need to highlight the key insight that $H_\theta(k, t, x_{[0,t]}^0)$ in equation (3) is the negative cross-entropy between the following two discrete distributions over $\{\emptyset, 1\}$:

$$\left[\frac{\lambda_k^*(t|x_{[0,t]}^0)}{\Delta_k^*(t|x_{[0,t]}^0)}, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k^*(t|x_{[0,t]}^0)}, \dots, \frac{\lambda_k^q(t|x_t^0)}{\Delta_k^*(t|x_t^0)} \right]$$

$$\left[\frac{\lambda_k(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)}, \underbrace{\frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)}, \dots, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)}}_{\text{length } K} \right]$$

theorems & proofs

how to draw noise fast

Theorem 2 (Consistency). Under assumption 1, for any $\theta \in \operatorname{argmax}_\theta J_{\text{NC}}(\theta)$ and $M \geq 1$, with probability $1 - o(1)$, we have $J_{\text{NC}}(\theta) \rightarrow J_{\text{NC}}(\theta^*)$ and $J_{\text{NC}}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous with respect to θ . The intuition of this theorem is that $J_{\text{NC}}(\theta)$ is some member of the set $\operatorname{argmax}_\theta J_{\text{NC}}(\theta)$. But we will still prove it out in our notation for completeness.

Theorem 3 (Efficiency). Under assumptions 2 and 4–7, there exists an integer \bar{M} such that $\|\mathbf{I}_{\bar{M}}^{-1} - \mathbf{I}_*^{-1}\| \leq C/M$ for some non-singular matrix \mathbf{I}_*^{-1} . Moreover, there exist a constant $C > 0$ such that for all $M > M$, $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, \mathbf{I}_M^{-1})$ as $N \rightarrow \infty$, where $\|\mathbf{I}\|$ is the spectral norm of matrix \mathbf{I} .

Proof. We first prove that $\sqrt{N}(\hat{\theta} - \theta^*)$ is asymptotically normal. By the Mean-Value Theorem, we have

$$\nabla_\theta J_{\text{NC}}^N(\hat{\theta}) = \nabla_\theta J_{\text{NC}}^N(\theta^*) + (\hat{\theta} - \theta^*) \int_{u=0}^1 \nabla_\theta^2 J_{\text{NC}}^N(\theta^* + u(\hat{\theta} - \theta^*)) dt \quad (41)$$

3.1 Efficient Sampling of Noise Events

The thinning algorithm (Lewis & Shedler, 1979; Liniger, 2009) is a rejection sampling method for drawing an event stream over a given observation interval $[0, T]$ from a continuous-time autoregressive process. Suppose we have already drawn the first $i-1$ times, namely t_1, \dots, t_{i-1} . For every future time $t \geq t_{i-1}$, let $\mathcal{H}(t)$ denote the context $x_{[0,t]}$ consisting only of the events at those times, and define $\lambda(t | \mathcal{H}(t)) \stackrel{\text{def}}{=} \sum_{k=1}^K \lambda_k(t | \mathcal{H}(t))$. If $\lambda(t | \mathcal{H}(t))$ were constant at $\bar{\lambda}$, we could draw the next event time as $t_i \sim t_{i-1} + \bar{\lambda}$. But what if $\lambda(t | \mathcal{H}(t))$ is not constant? The thinning algorithm still runs the foregoing method, taking $\bar{\lambda}$ to be any upper bound to λ . This mass is allocated to \emptyset . A draw of $x_{t_i} = \emptyset$ means there was no event at time t_i after all (corresponding to a rejected proposal). Either way, we now continue on to draw t_{i+1} and $x_{t_{i+1}}$ using a version of $\mathcal{H}(t)$ that has been updated to include the event or non-event x_{t_i} . The update to $\mathcal{H}(t)$ affects $\lambda(t | \mathcal{H}(t))$ and the choice of $\bar{\lambda}$. How to sample noise streams. To draw a stream $x_{[0,t]}^0$ of noise events, we run the thinning algorithm, using the noise intensity functions λ_k^q . However, there is a modification: $\mathcal{H}(t)$ is now defined to be $x_{[0,t]}^0$ —the history from the observed event stream, rather than the previously sampled noise events—and is updated accordingly. This is because in equation (6), at each time t , all of $\{x_t^0, x_t^1, \dots, x_t^M\}$ are conditioned on $x_{[0,t]}^0$ (akin to the discrete-time case).⁷ The full pseudocode is given in Algorithm 1 in the supplementary material.

Coarse-to-fine sampling of event types. Although our NCE method has eliminated the need to integrate over t , the thinning algorithm above still sums over k in the definition of $\lambda_k^q(t | \mathcal{H}(t))$. For large K , this sum is expensive if we take the noise distribution on each training minibatch to

NCE: More in paper

theorems & proofs

Theorem 1 (Optimality). Under assumptions 1 and 2, $\theta \in \operatorname{argmax}_{\theta} J_{NC}(\theta)$ if and only if $p_{\theta} = p^*$.

We first need to highlight the key insight that $H_{\theta}(k, t, x_{[0,t]}^0)$ in equation (30) is the negative cross-entropy between the following two discrete distributions over $\{\emptyset, 1\}$:

$$\left[\frac{\lambda_k^*(t|x_{[0,t]}^0)}{\Delta_k^*(t|x_{[0,t]}^0)}, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k^q(t|x_{[0,t]}^0)}, \dots, \frac{\lambda_k^q(t|x_t^0)}{\Delta_k^q(t|x_t^0)} \right]$$

$$\left[\frac{\lambda_k(t|x_{[0,t]}^0)}{\Delta_k(t|x_{[0,t]}^0)}, \frac{\lambda_k^q(t|x_{[0,t]}^0)}{\Delta_k^q(t|x_{[0,t]}^0)}, \dots, \frac{\lambda_k^q(t|x_t^0)}{\Delta_k^q(t|x_t^0)} \right]$$

how to draw noise t_c

Theorem 2 (Consistency). Under assumption 1, $\hat{\theta}_N \rightarrow \theta^*$ as $N \rightarrow \infty$ and $J_{NC}(\hat{\theta}_N) \rightarrow J_{NC}(\theta^*)$ as $N \rightarrow \infty$. Moreover, $\hat{\theta}_N$ is almost surely a member of the set $\operatorname{argmax}_{\theta} J_{NC}(\theta)$.

The intuition of this theorem is that $J_{NC}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous.

Theorem 3 (Efficiency). Under assumptions 2 and 4–7, there exists an integer \bar{M} such that if $M > \bar{M}$, $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, I_M^{-1})$ as $N \rightarrow \infty$.

Moreover, there exist a constant $C > 0$ such that for all $M > \bar{M}$, $\|I_M^{-1} - I_*^{-1}\| \leq C/M$, where $\|I\|$ is the spectral norm of matrix I .

Proof. We first prove that $\sqrt{N}(\hat{\theta} - \theta^*)$ is asymptotically normal. By the Mean-Value Theorem, we have

$$\nabla_{\theta} J_{NC}(\hat{\theta}) = \nabla_{\theta} J_{NC}(\theta^*) + (\hat{\theta} - \theta^*) \int_{u=0}^1 \nabla_{\theta}^2 J_{NC}(\theta^* + u(\hat{\theta} - \theta^*)) du \quad (40)$$

the intuition of this theorem is that $J_{NC}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous.

Theorem 2 (Consistency). Under assumption 1, $\hat{\theta}_N \rightarrow \theta^*$ as $N \rightarrow \infty$ and $J_{NC}(\hat{\theta}_N) \rightarrow J_{NC}(\theta^*)$ as $N \rightarrow \infty$. Moreover, $\hat{\theta}_N$ is almost surely a member of the set $\operatorname{argmax}_{\theta} J_{NC}(\theta)$.

The intuition of this theorem is that $J_{NC}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous.

Theorem 3 (Efficiency). Under assumptions 2 and 4–7, there exists an integer \bar{M} such that if $M > \bar{M}$, $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, I_M^{-1})$ as $N \rightarrow \infty$.

Moreover, there exist a constant $C > 0$ such that for all $M > \bar{M}$, $\|I_M^{-1} - I_*^{-1}\| \leq C/M$, where $\|I\|$ is the spectral norm of matrix I .

Proof. We first prove that $\sqrt{N}(\hat{\theta} - \theta^*)$ is asymptotically normal. By the Mean-Value Theorem, we have

$$\nabla_{\theta} J_{NC}(\hat{\theta}) = \nabla_{\theta} J_{NC}(\theta^*) + (\hat{\theta} - \theta^*) \int_{u=0}^1 \nabla_{\theta}^2 J_{NC}(\theta^* + u(\hat{\theta} - \theta^*)) du \quad (40)$$

the intuition of this theorem is that $J_{NC}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous.

Theorem 2 (Consistency). Under assumption 1, $\hat{\theta}_N \rightarrow \theta^*$ as $N \rightarrow \infty$ and $J_{NC}(\hat{\theta}_N) \rightarrow J_{NC}(\theta^*)$ as $N \rightarrow \infty$. Moreover, $\hat{\theta}_N$ is almost surely a member of the set $\operatorname{argmax}_{\theta} J_{NC}(\theta)$.

The intuition of this theorem is that $J_{NC}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous.

Theorem 3 (Efficiency). Under assumptions 2 and 4–7, there exists an integer \bar{M} such that if $M > \bar{M}$, $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, I_M^{-1})$ as $N \rightarrow \infty$.

Moreover, there exist a constant $C > 0$ such that for all $M > \bar{M}$, $\|I_M^{-1} - I_*^{-1}\| \leq C/M$, where $\|I\|$ is the spectral norm of matrix I .

Proof. We first prove that $\sqrt{N}(\hat{\theta} - \theta^*)$ is asymptotically normal. By the Mean-Value Theorem, we have

$$\nabla_{\theta} J_{NC}(\hat{\theta}) = \nabla_{\theta} J_{NC}(\theta^*) + (\hat{\theta} - \theta^*) \int_{u=0}^1 \nabla_{\theta}^2 J_{NC}(\theta^* + u(\hat{\theta} - \theta^*)) du \quad (40)$$



Figure 4: Learning curves of MLE and NCE on the other real-world social interaction datasets.
 (a) CollegeMsg: neural q
 (b) WikiTalk: Poisson q
 (c) Never redraw new noise samples
 (d) Coarse-to-fine sampling of event types

Figure 5: Ablation Study I. Learning curves of MLE and NCE with $q = p^*$ and different “redraw” strategies.

THANK YOU

**Hongyuan Mei, Tom Wan, Jason Eisner
Johns Hopkins University**