

Noise-Contrastive Estimation for Multivariate Point Processes

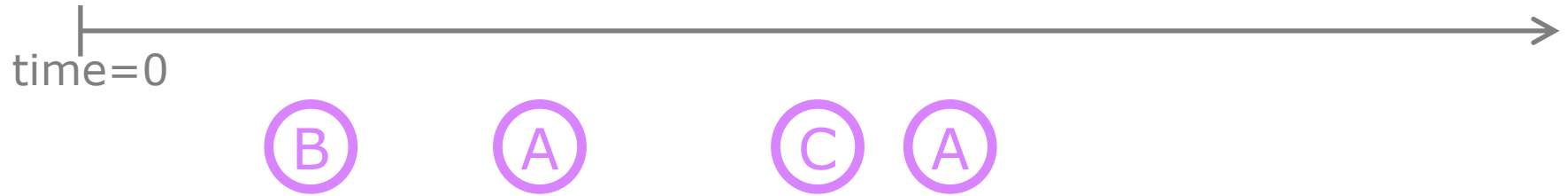
**Hongyuan Mei, Tom Wan, Jason Eisner
Johns Hopkins University**

MLE: Max log prob of *data*

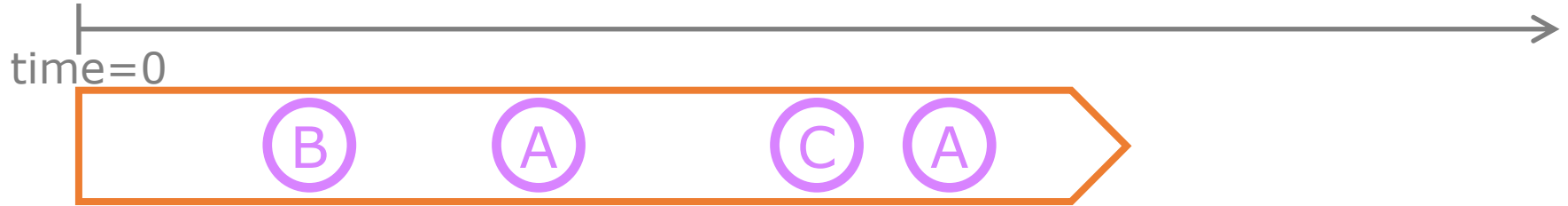
MLE: Max log prob of *data*



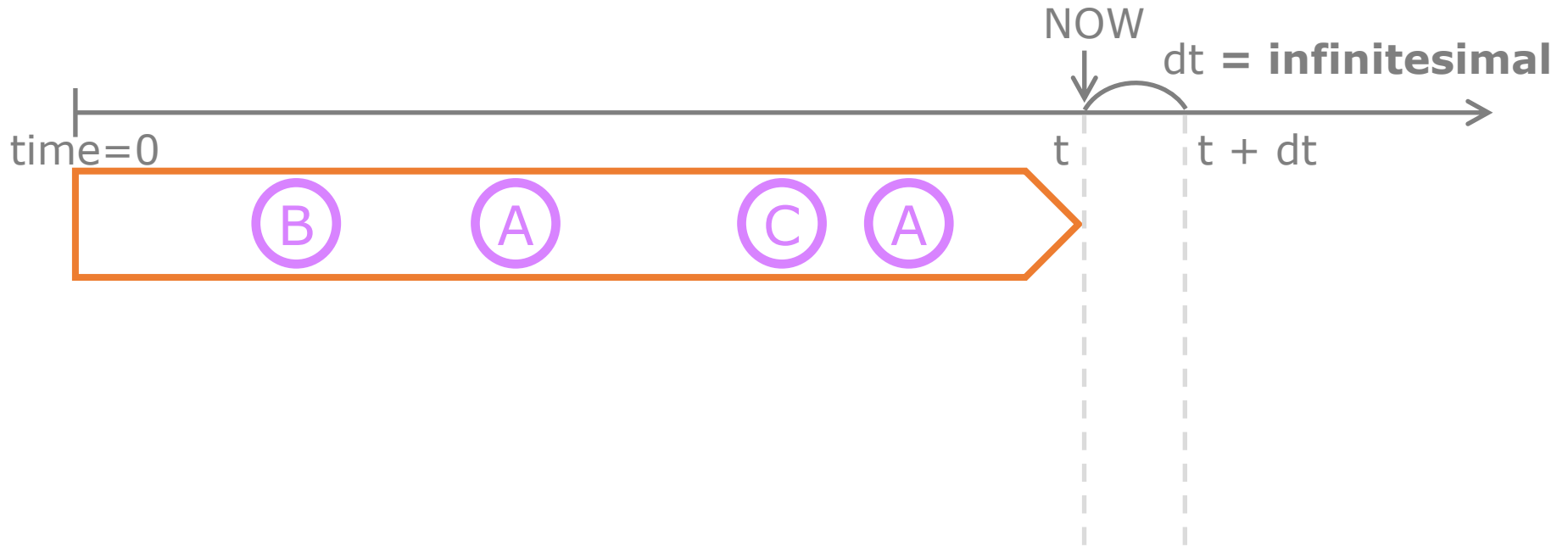
MLE: Max log prob of *data*



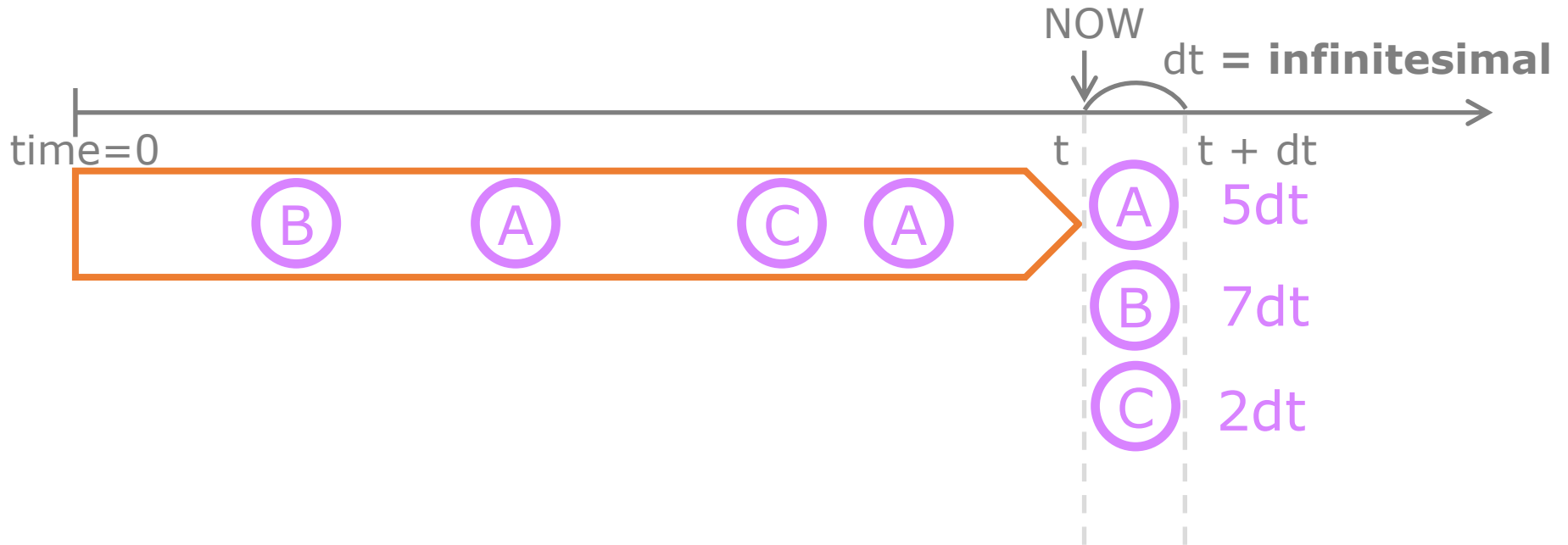
MLE: Max log prob of *data*



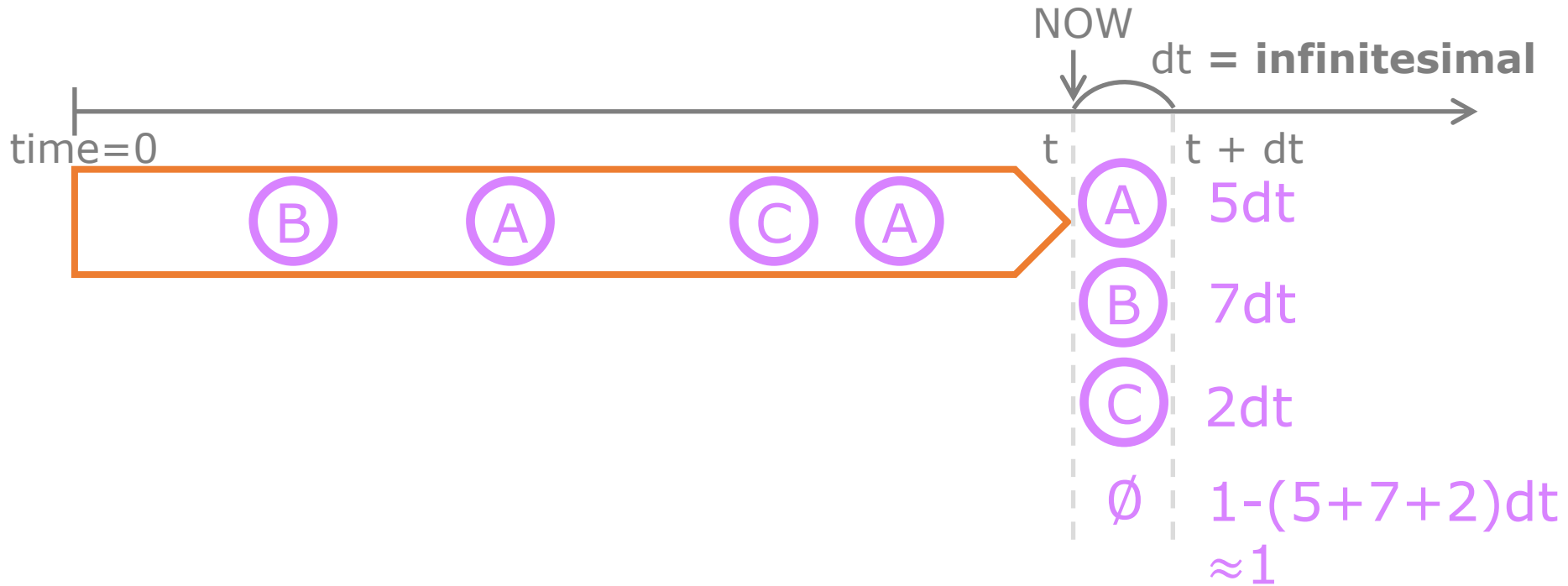
MLE: Max log prob of *data*



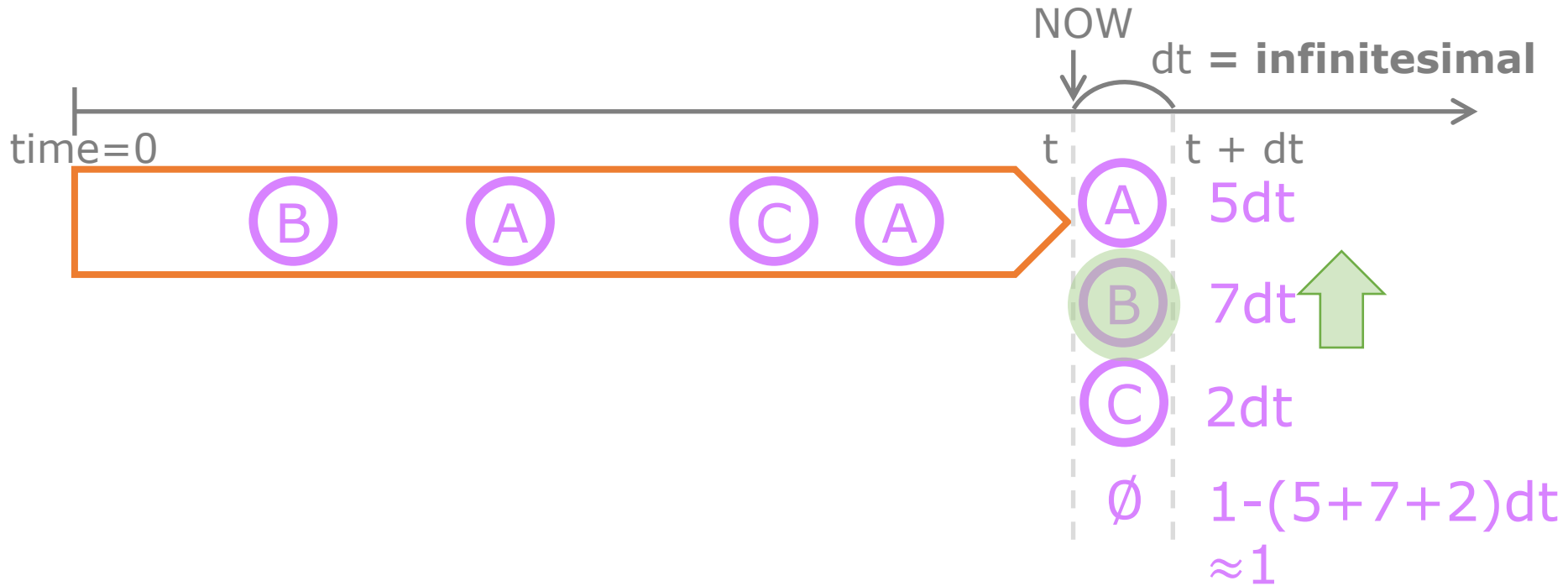
MLE: Max log prob of *data*



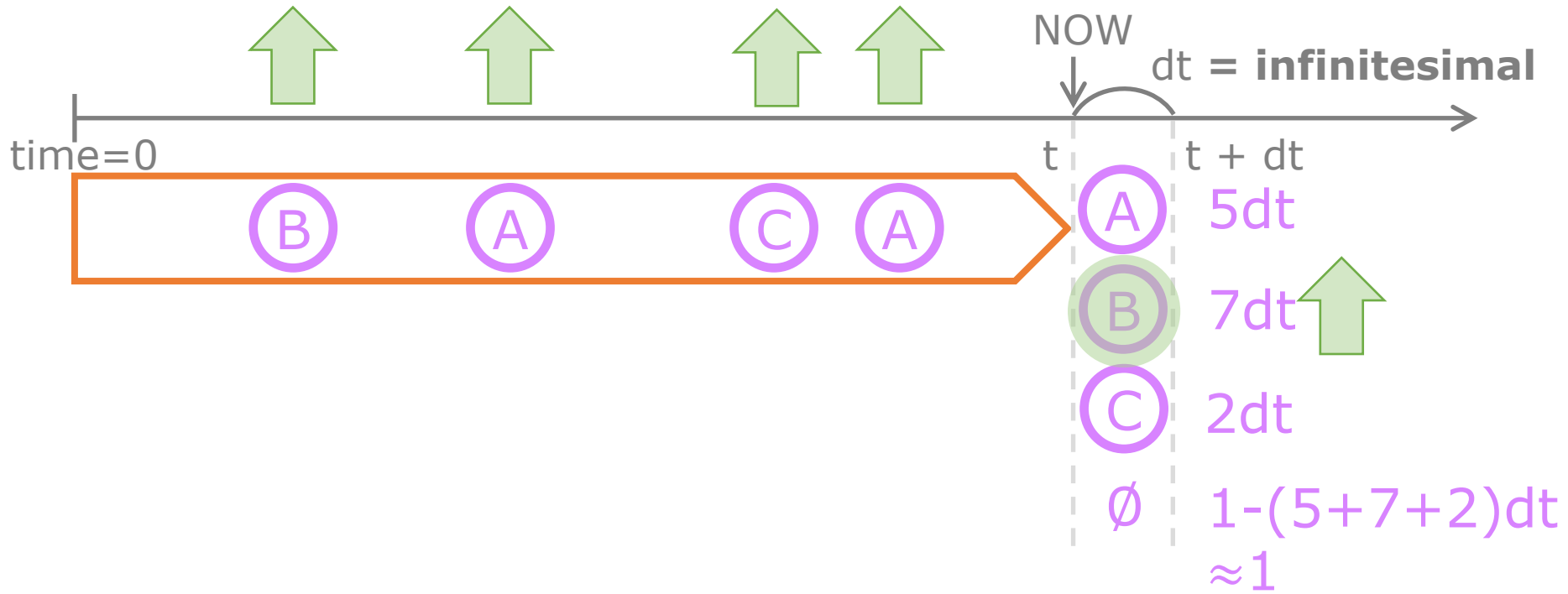
MLE: Max log prob of *data*



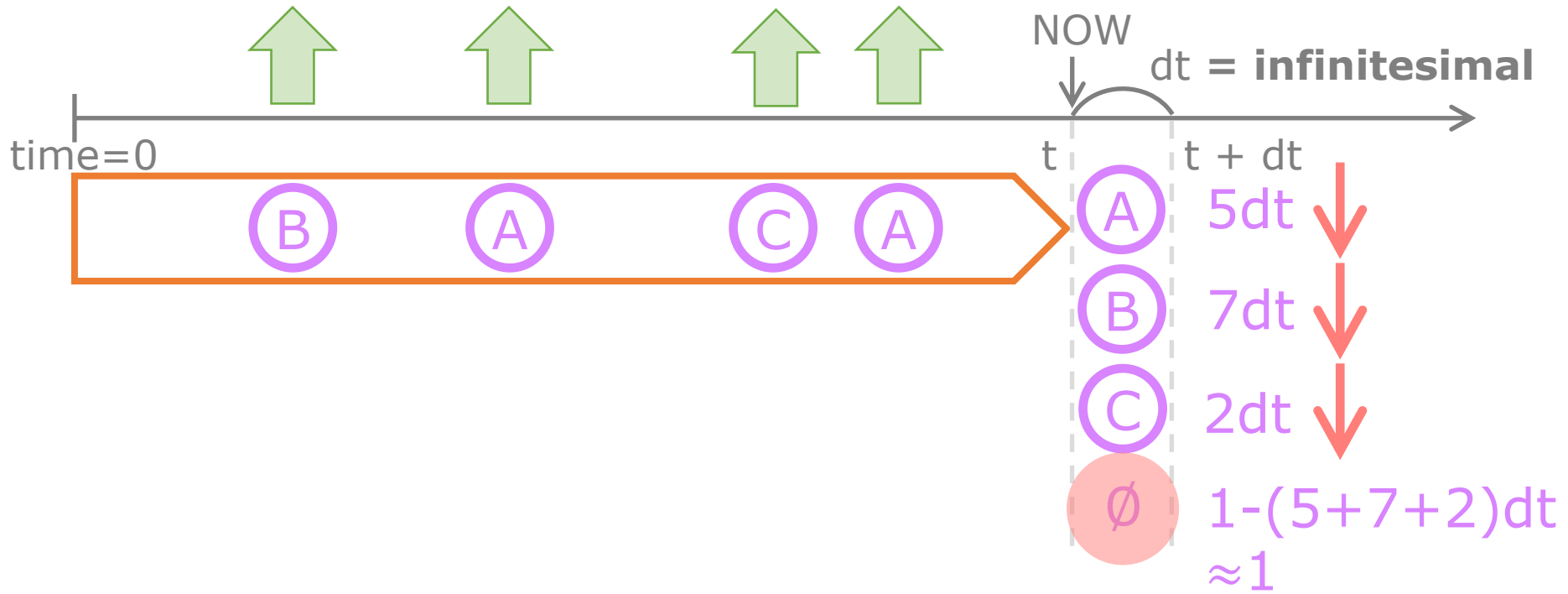
MLE: Max log prob of *data*



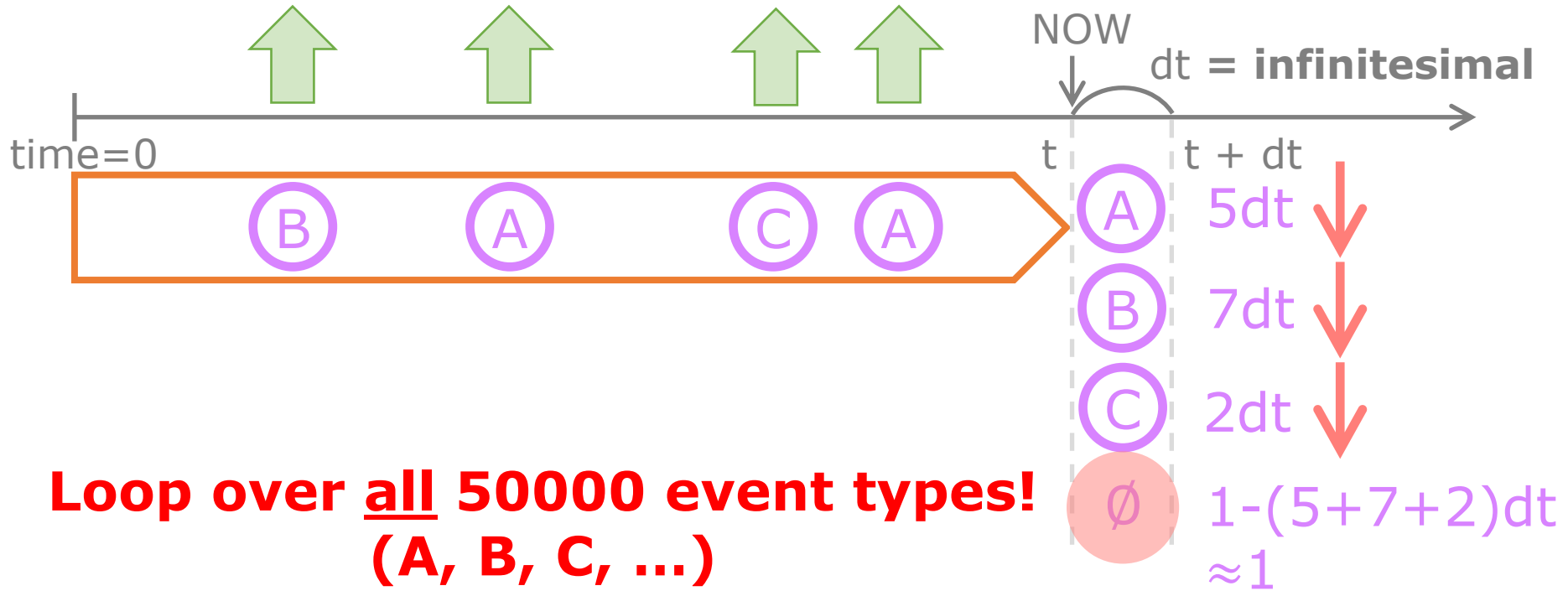
MLE: Max log prob of *data*



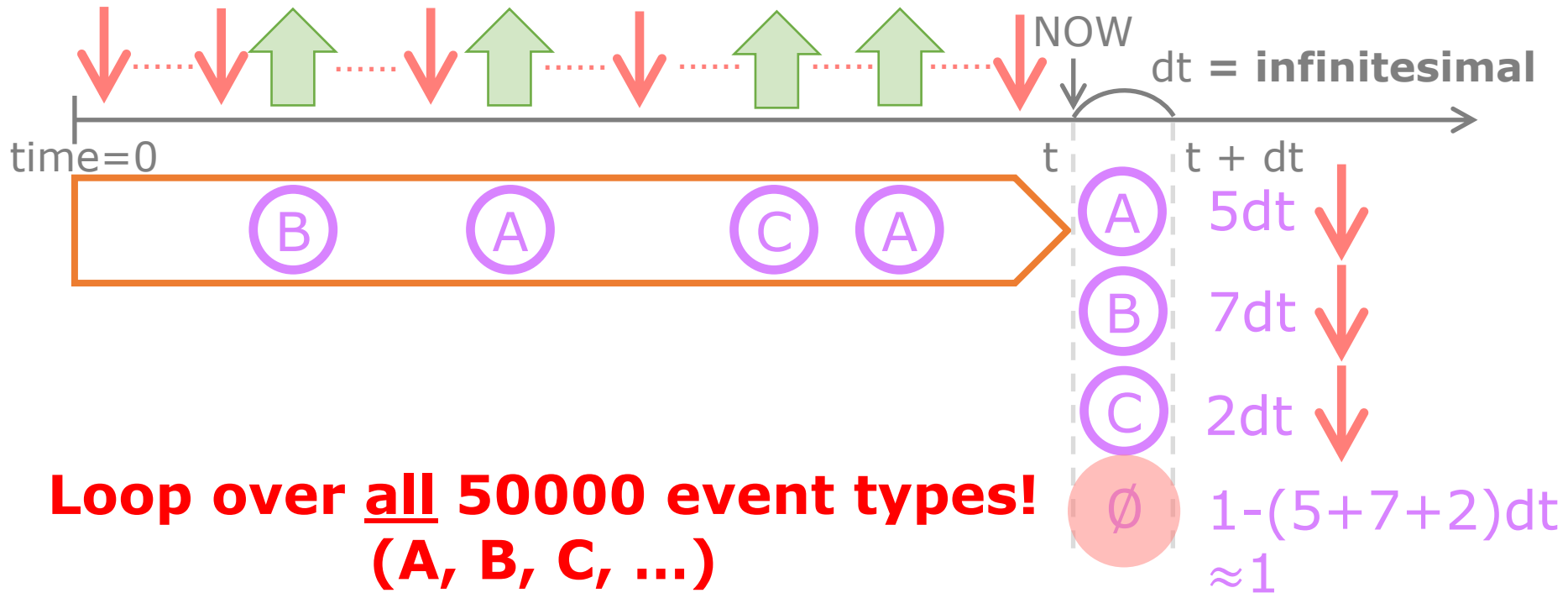
MLE: Max log prob of *data*



MLE: Max log prob of *data*

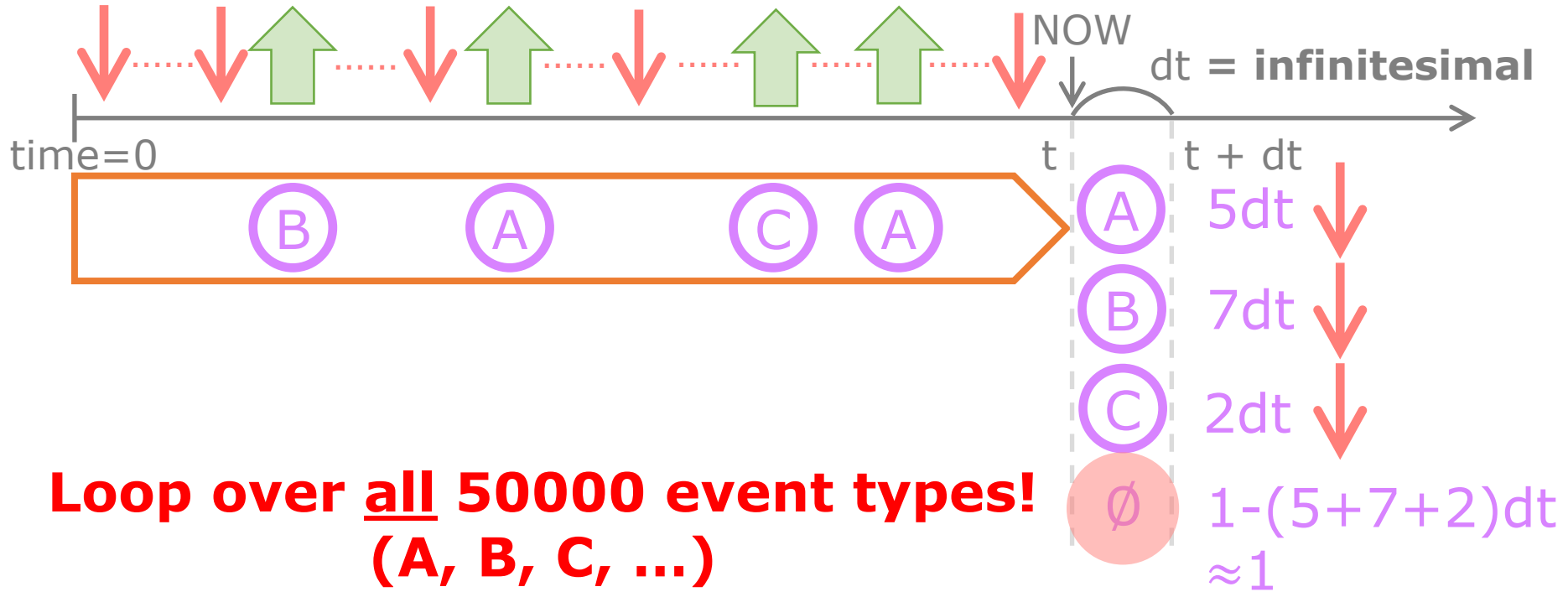


MLE: Max log prob of *data*



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

MLE: Max log prob of *data*

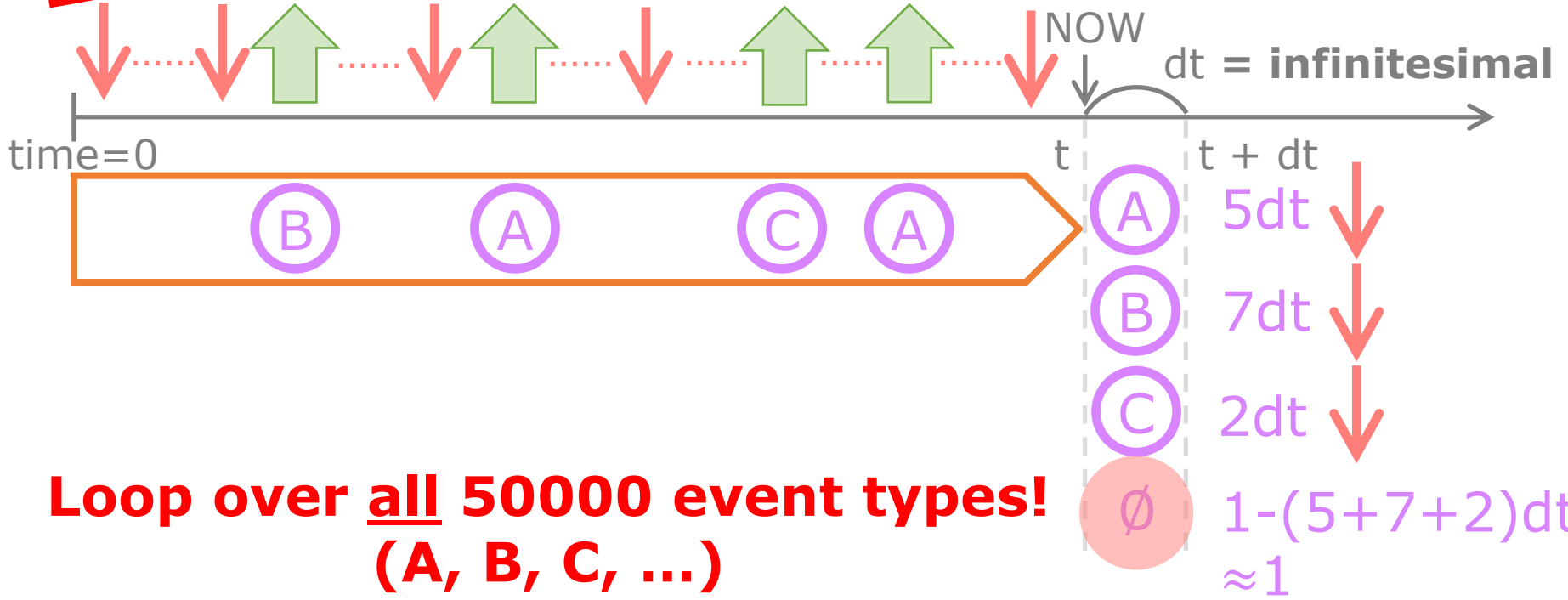


**Loop over all 50000 event types!
(A, B, C, ...)**

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

SLOW

~~MLE: Max log prob of data~~

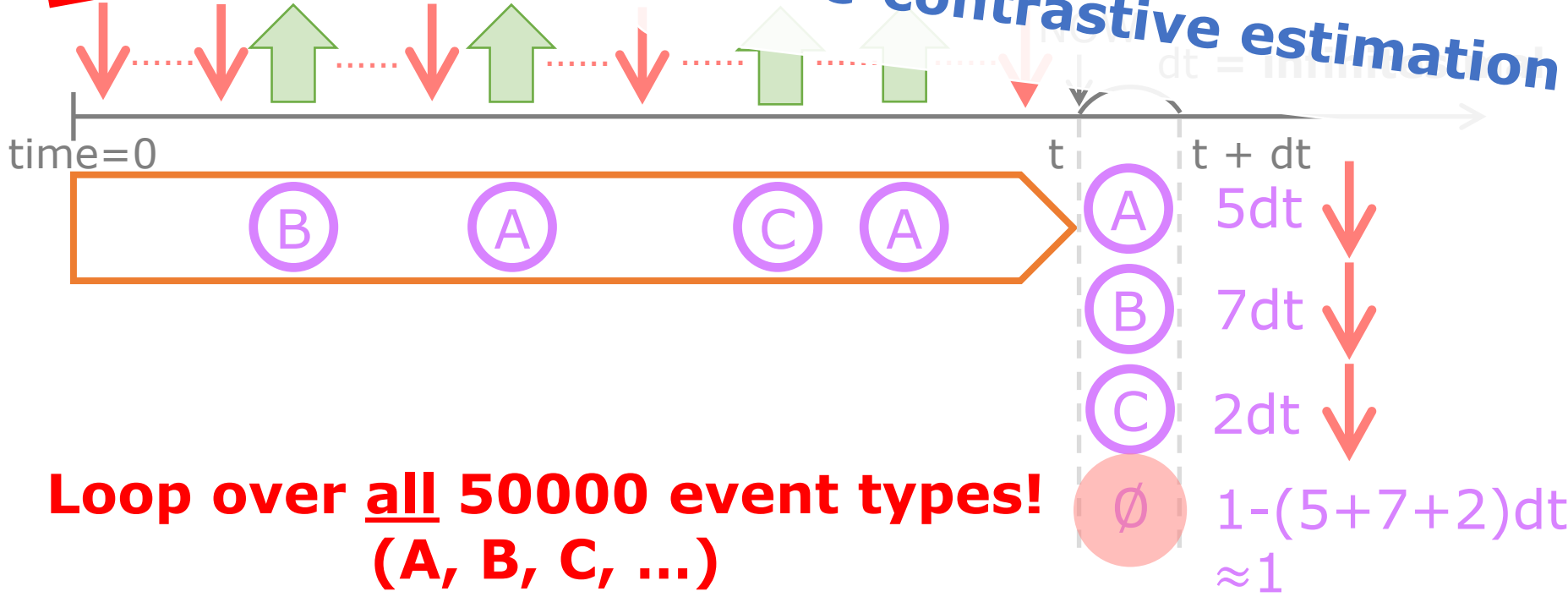


**Loop over all 50000 event types!
(A, B, C, ...)**

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

SLOW

~~MLE: Max log prob of data~~ alternative: noise-contrastive estimation

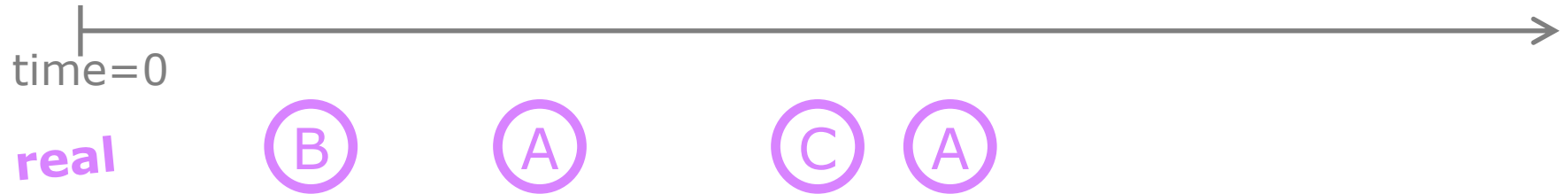


**Loop over all 50000 event types!
(A, B, C, ...)**

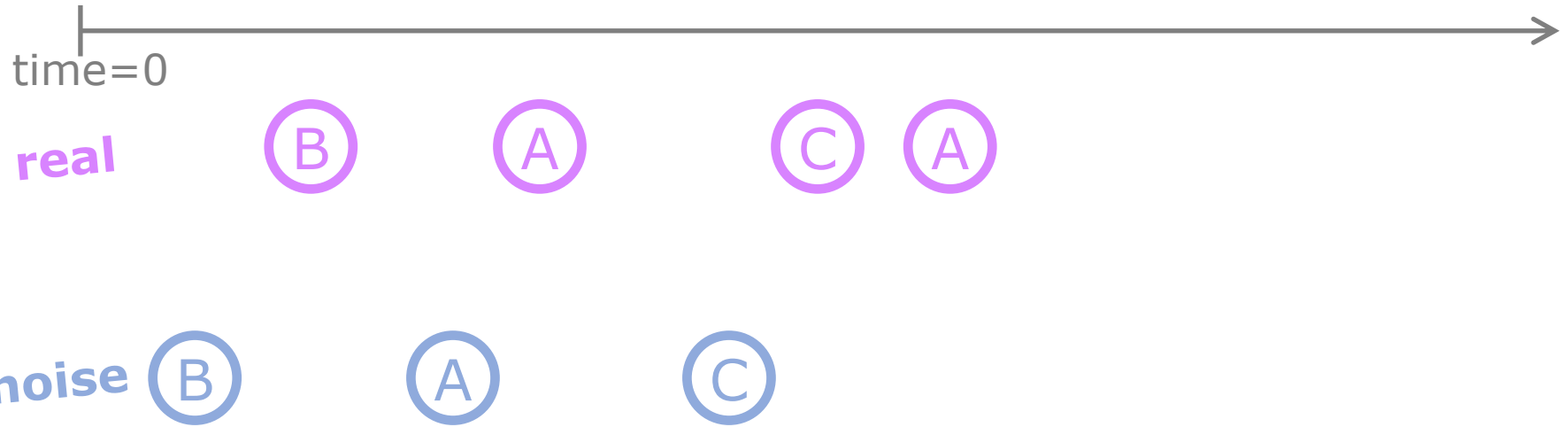
**Integrate over infinitely many *non-events*!
(often approx by sampling)**

SLOW

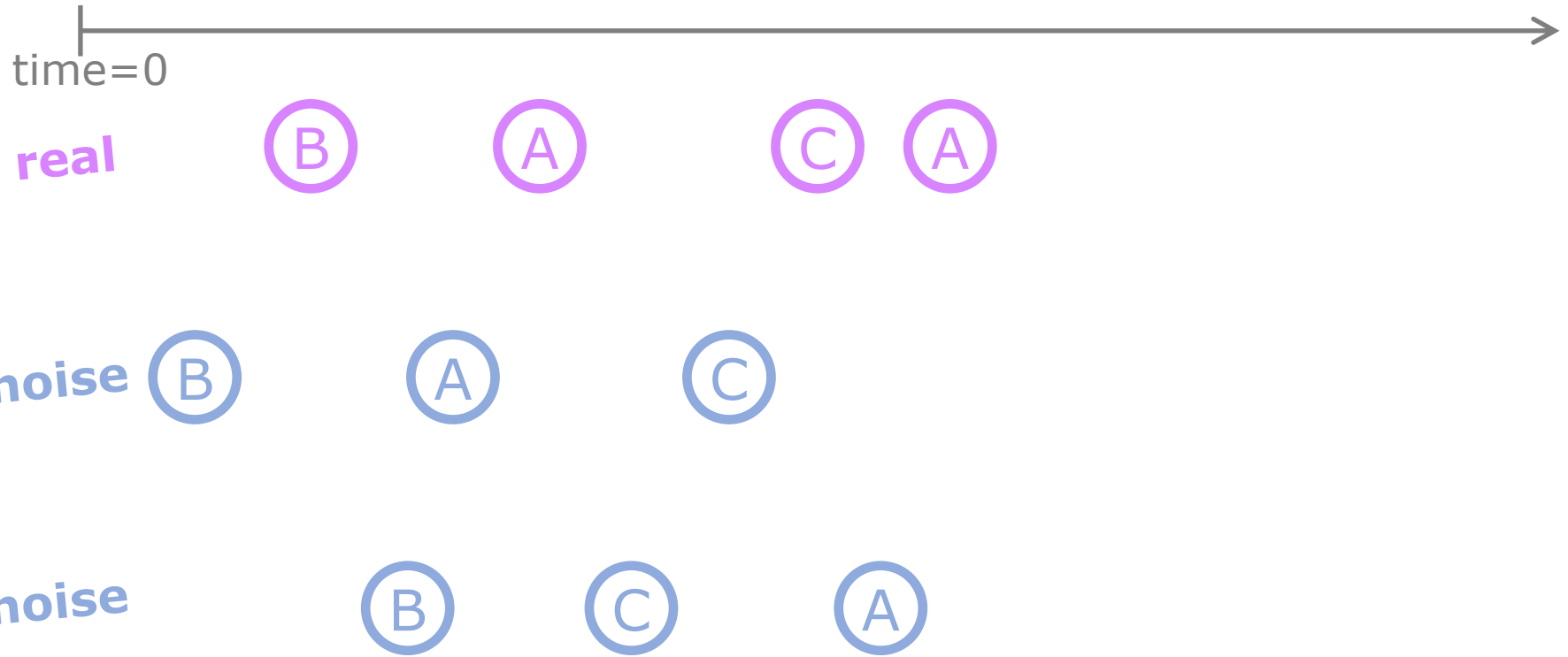
NCE: Max log prob of *correct discrimination*



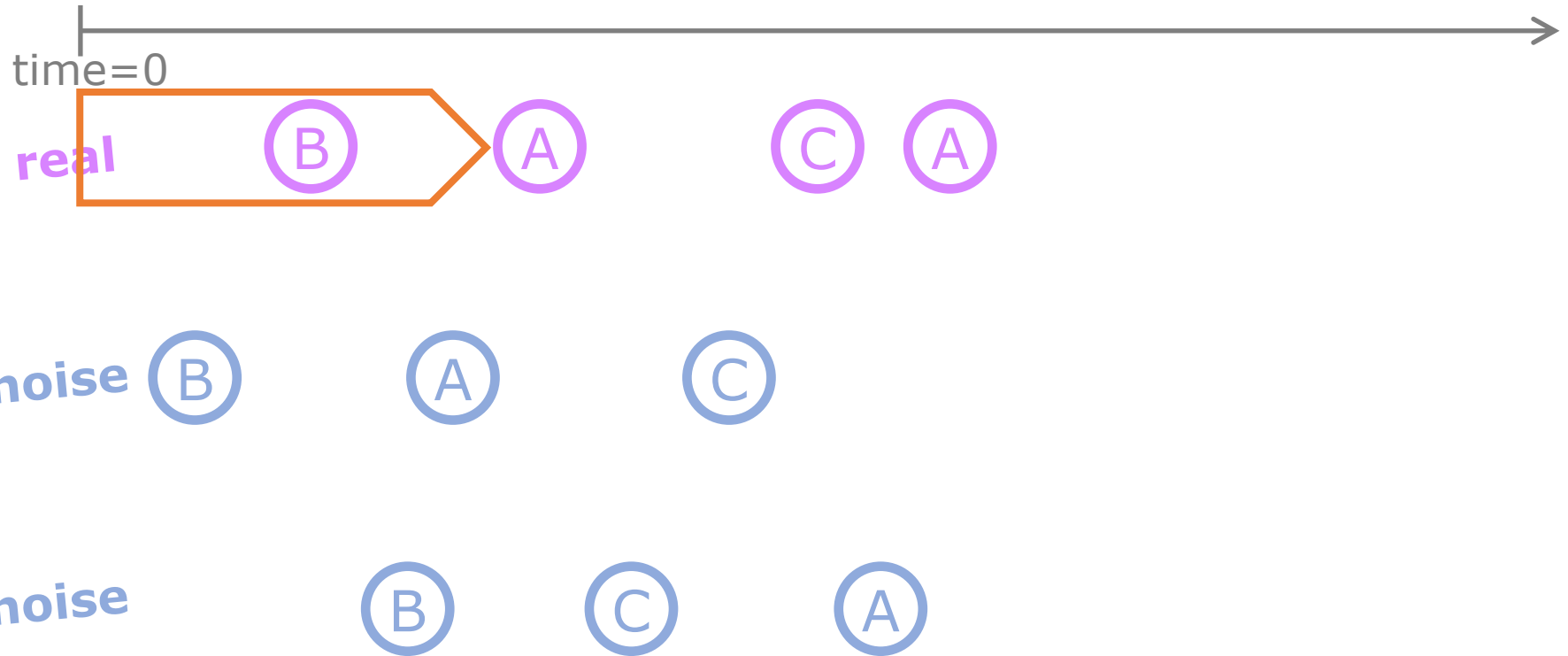
NCE: Max log prob of *correct discrimination*



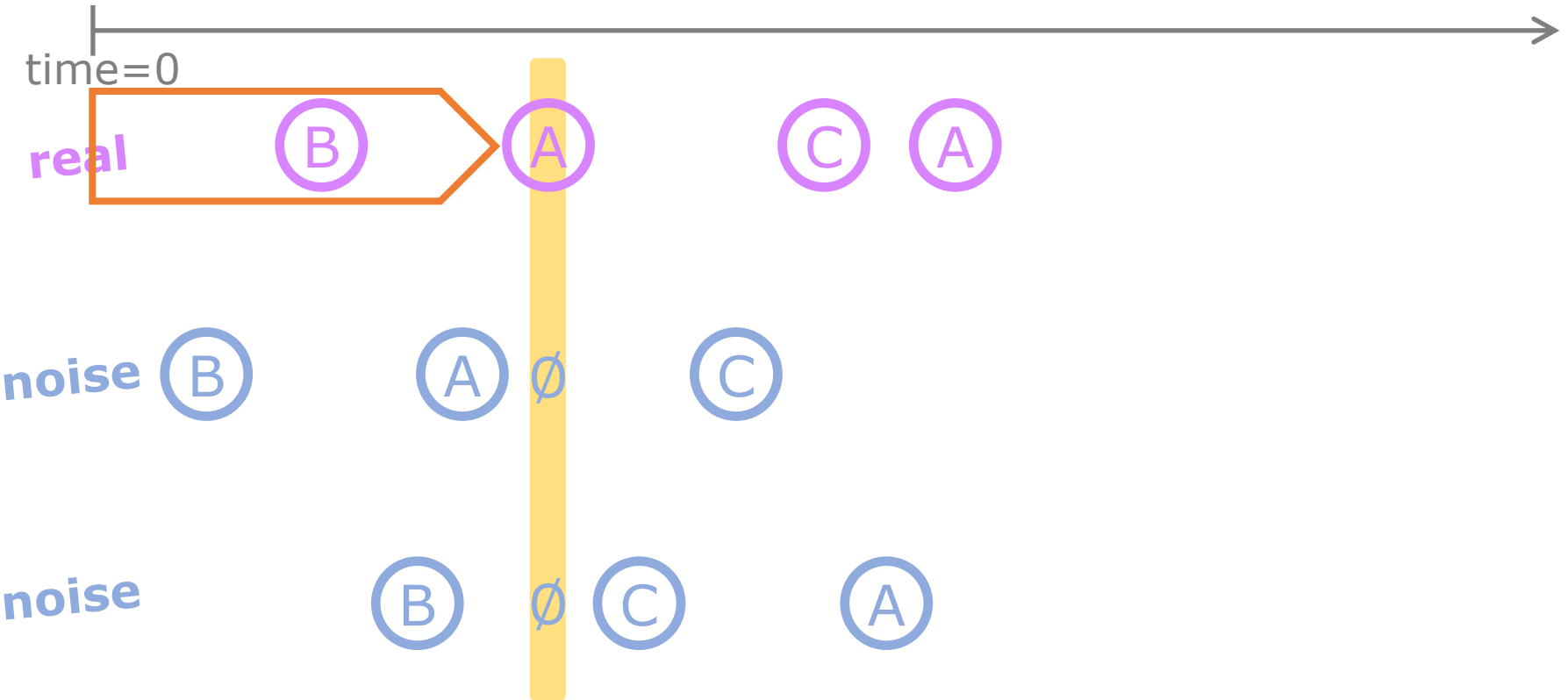
NCE: Max log prob of *correct discrimination*



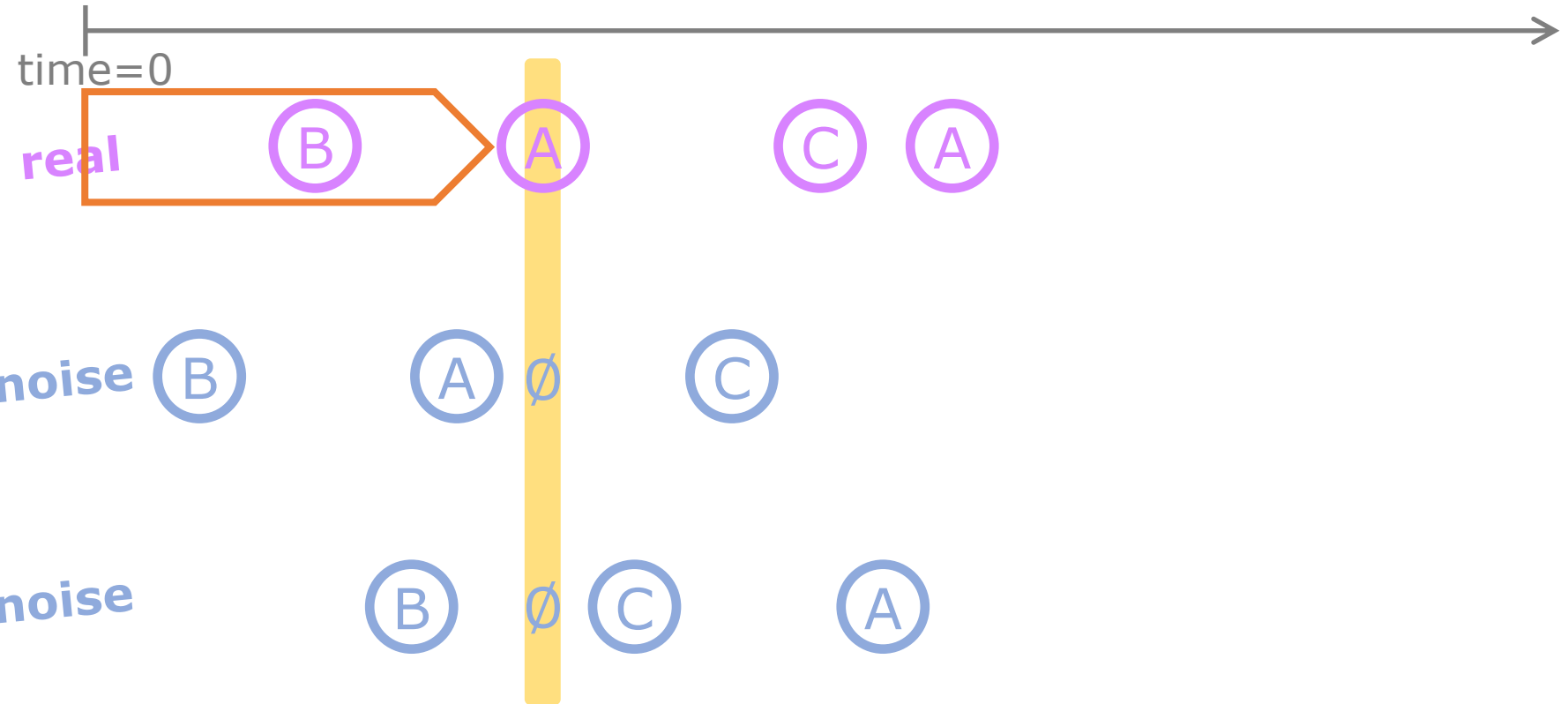
NCE: Max log prob of *correct discrimination*



NCE: Max log prob of *correct discrimination*

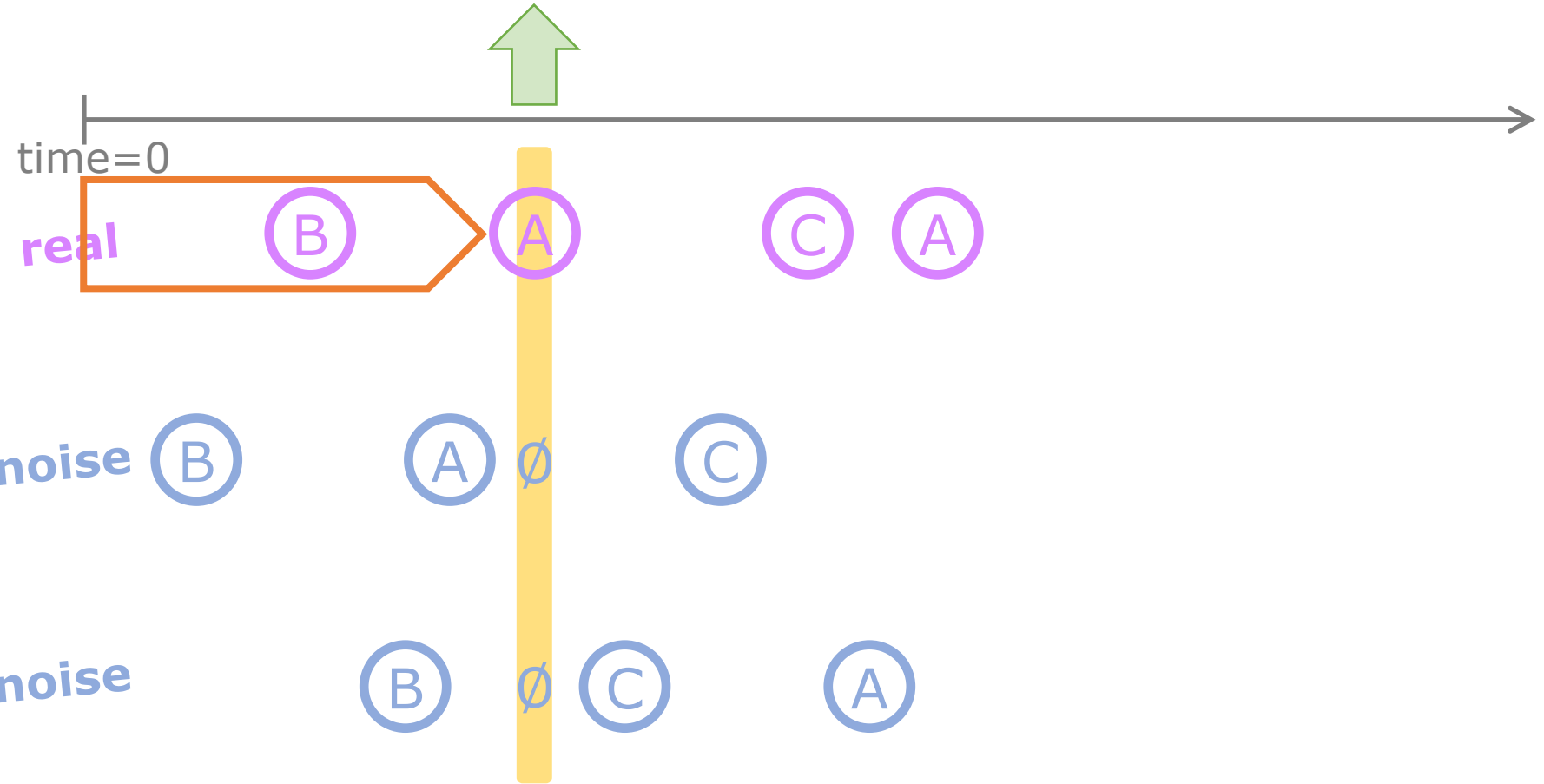


NCE: Max log prob of *correct discrimination*



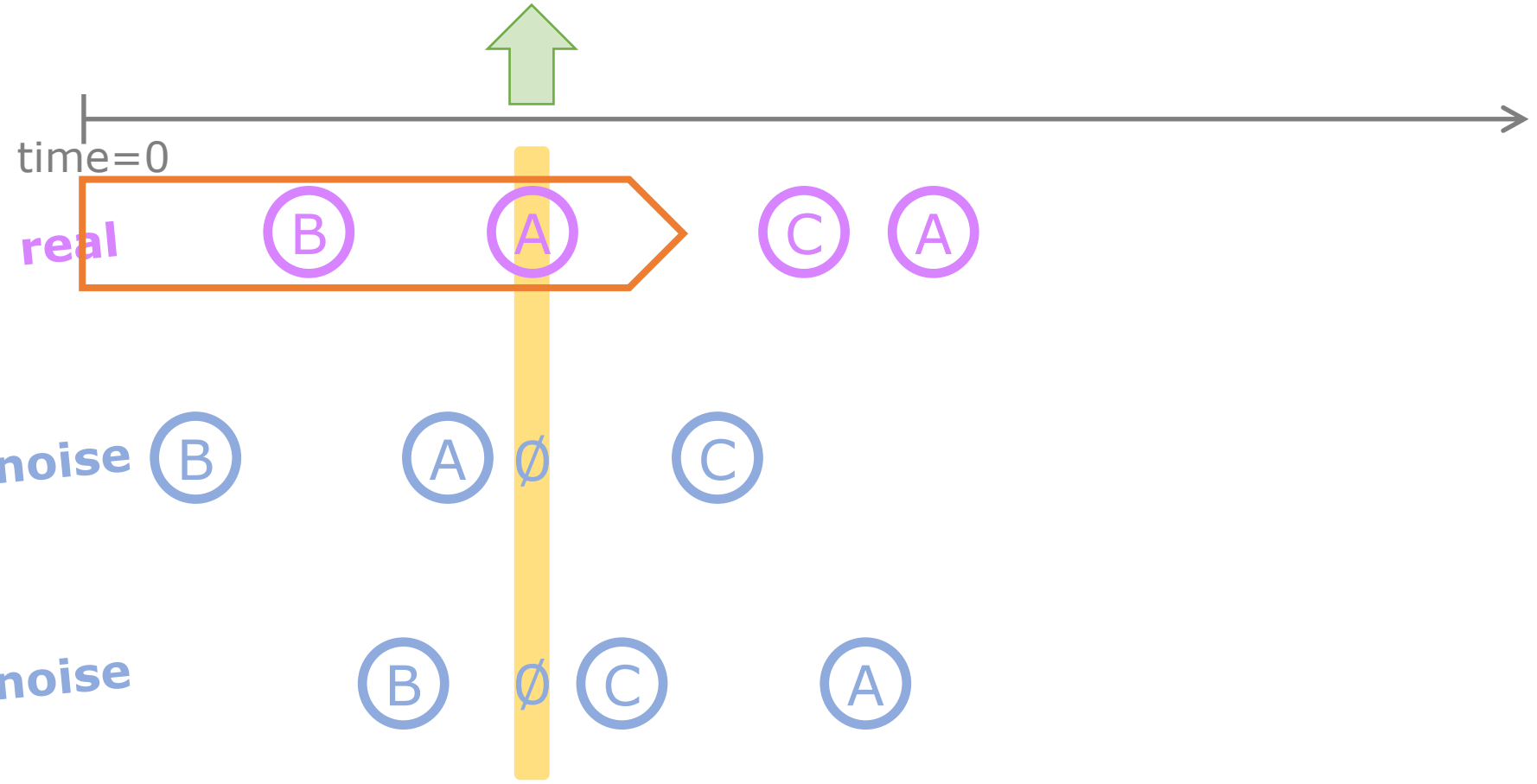
Which Is Real?

NCE: Max log prob of *correct discrimination*



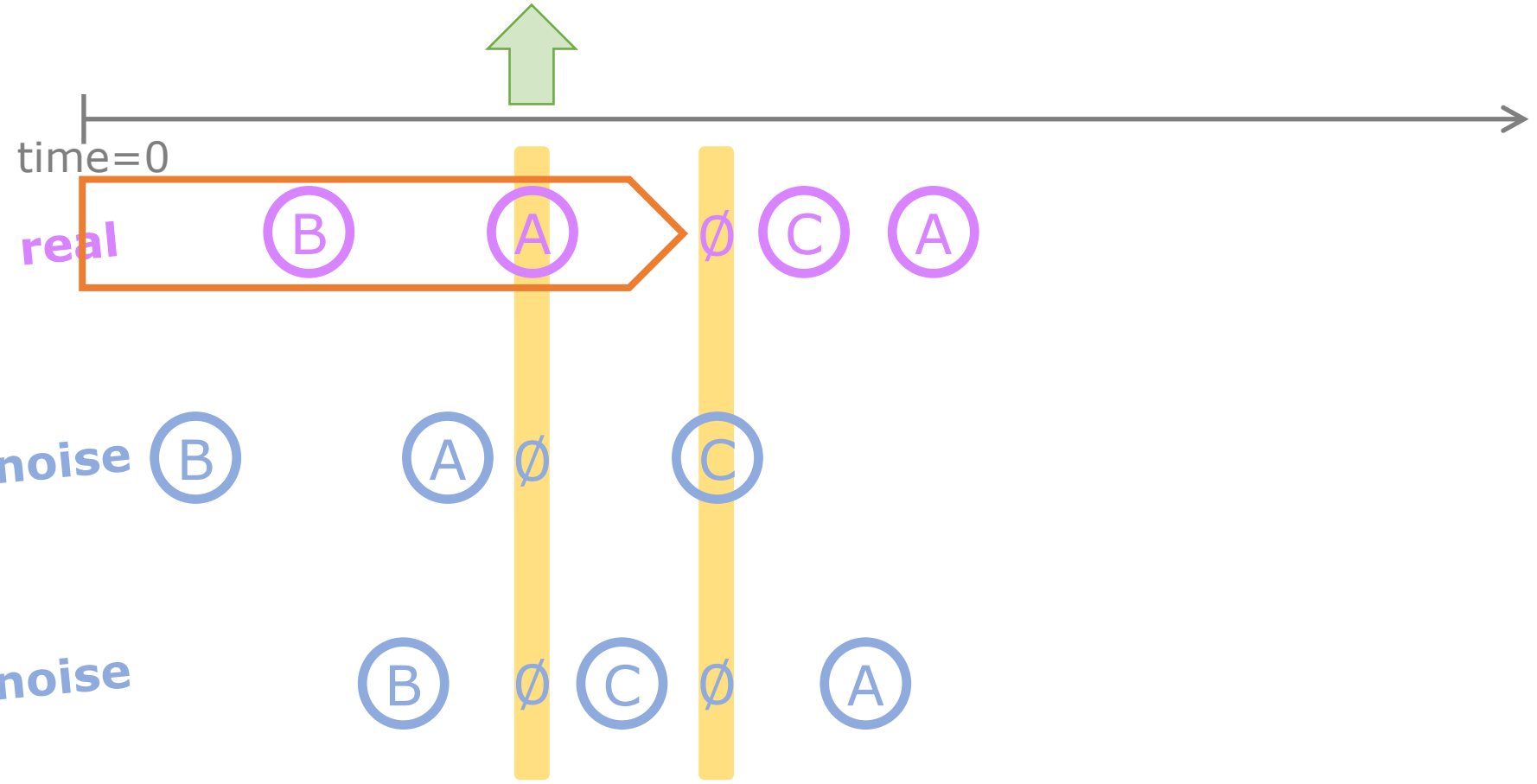
Which Is Real?

NCE: Max log prob of *correct discrimination*



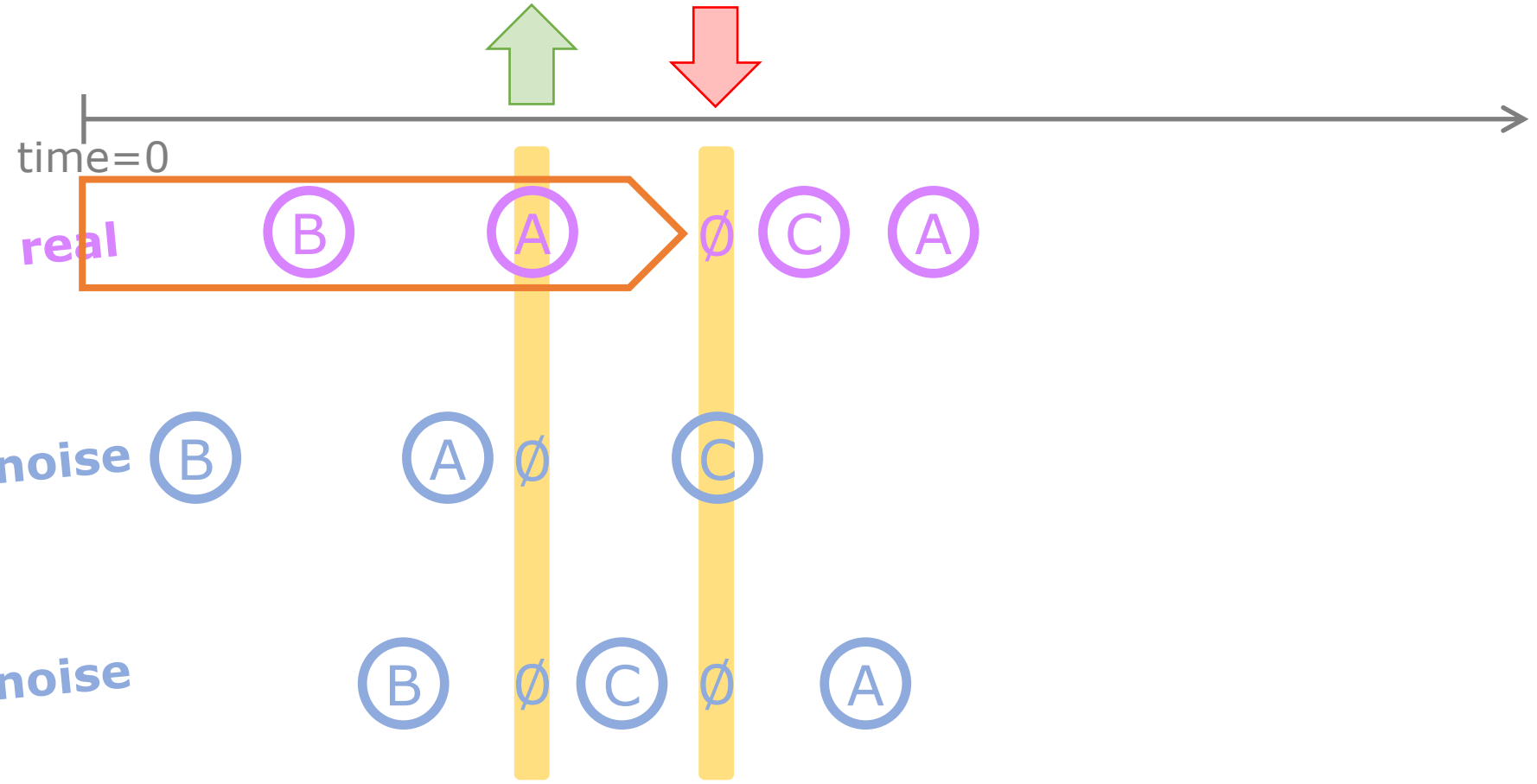
Which Is Real?

NCE: Max log prob of *correct discrimination*



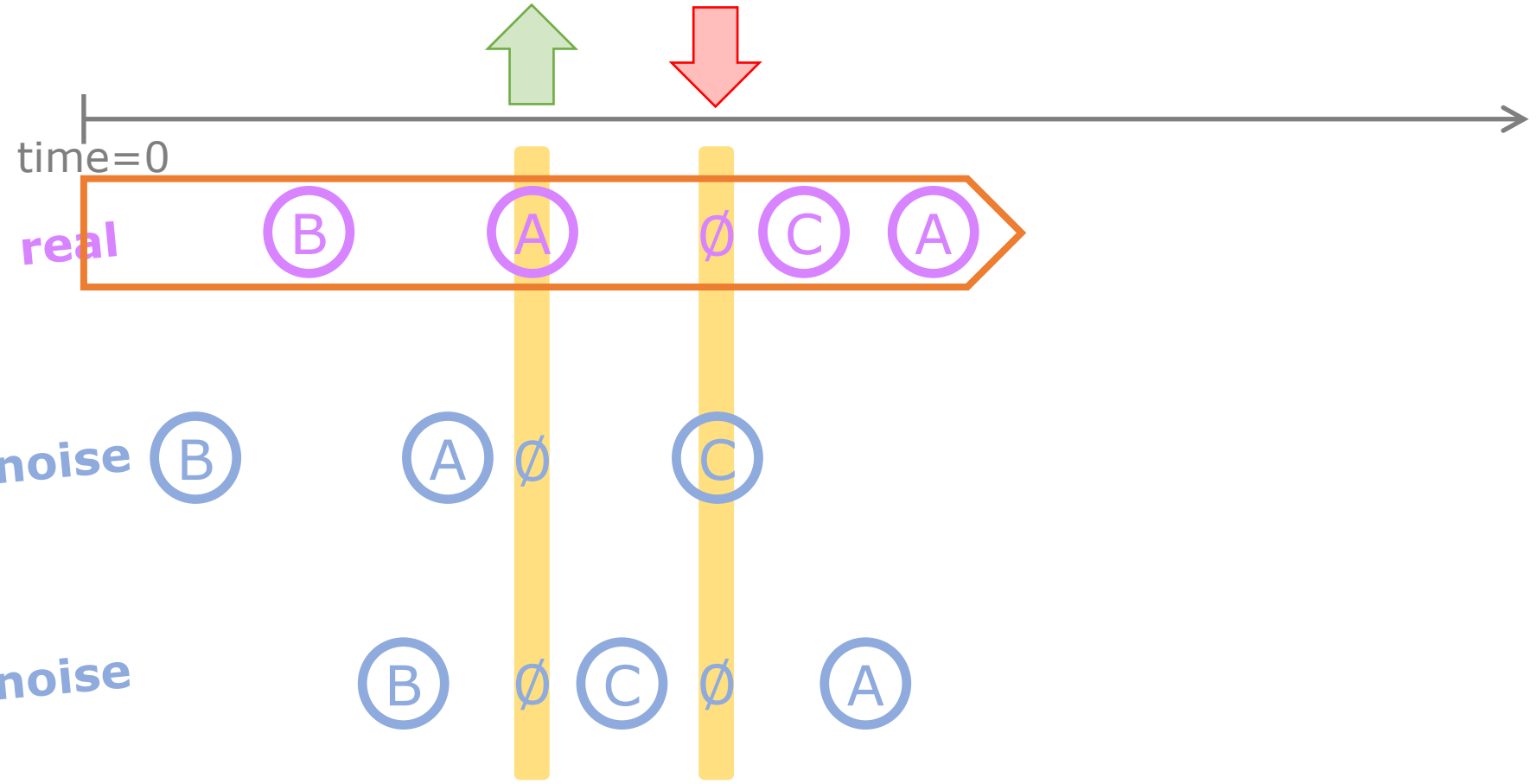
Which Is Real?

NCE: Max log prob of *correct discrimination*



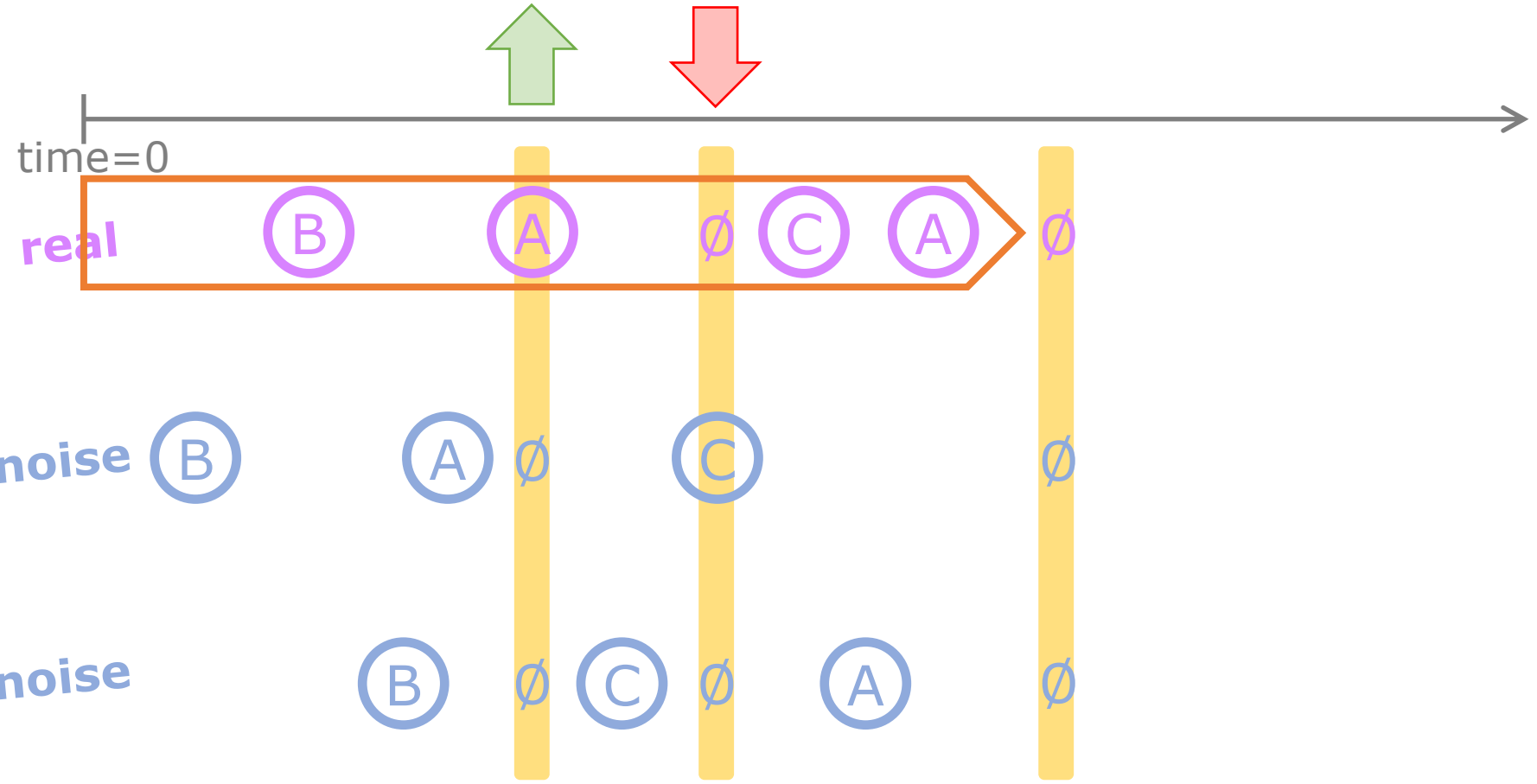
Which Is Real?

NCE: Max log prob of *correct discrimination*



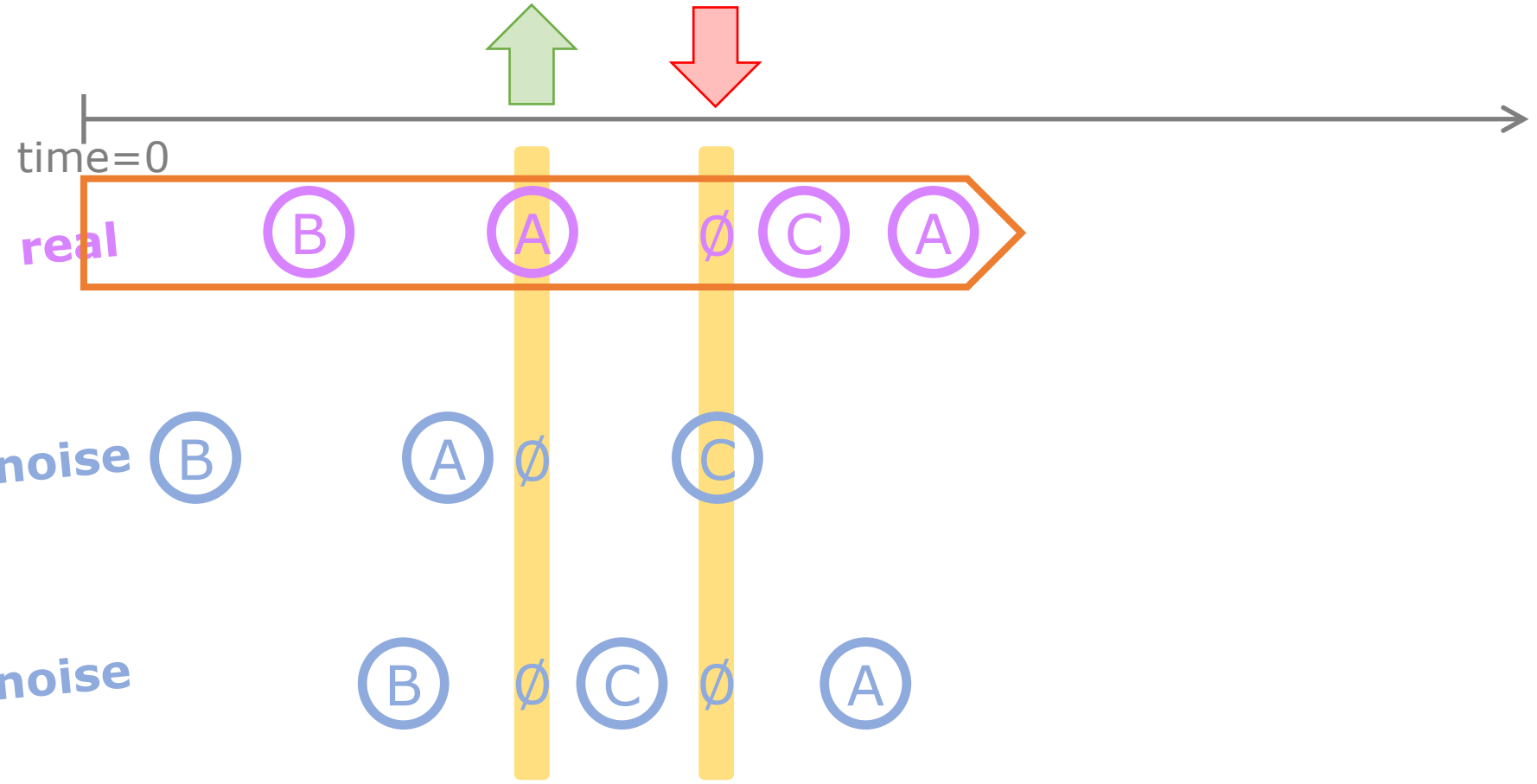
Which Is Real?

NCE: Max log prob of *correct discrimination*



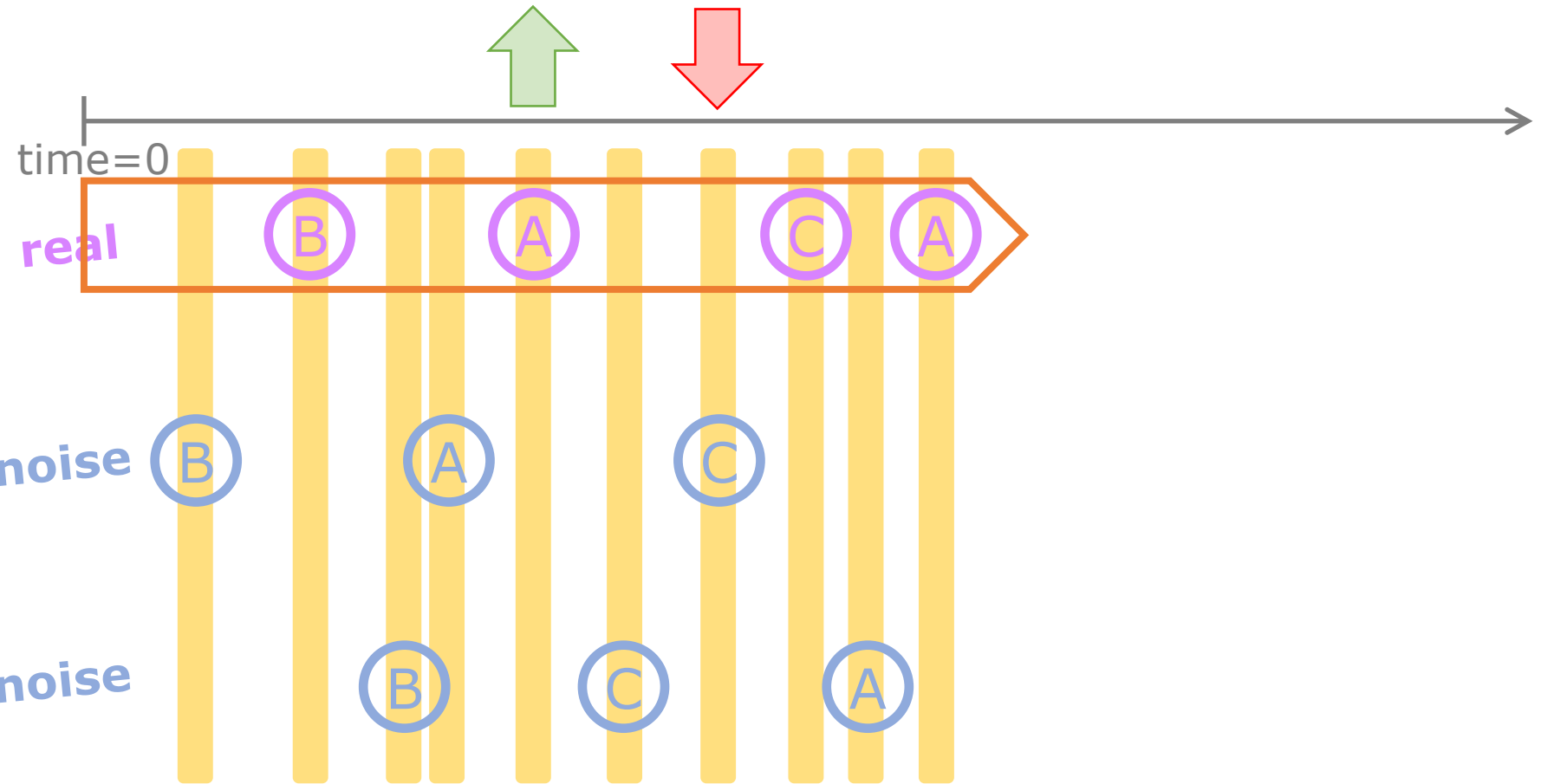
Which Is Real?

NCE: Max log prob of *correct discrimination*



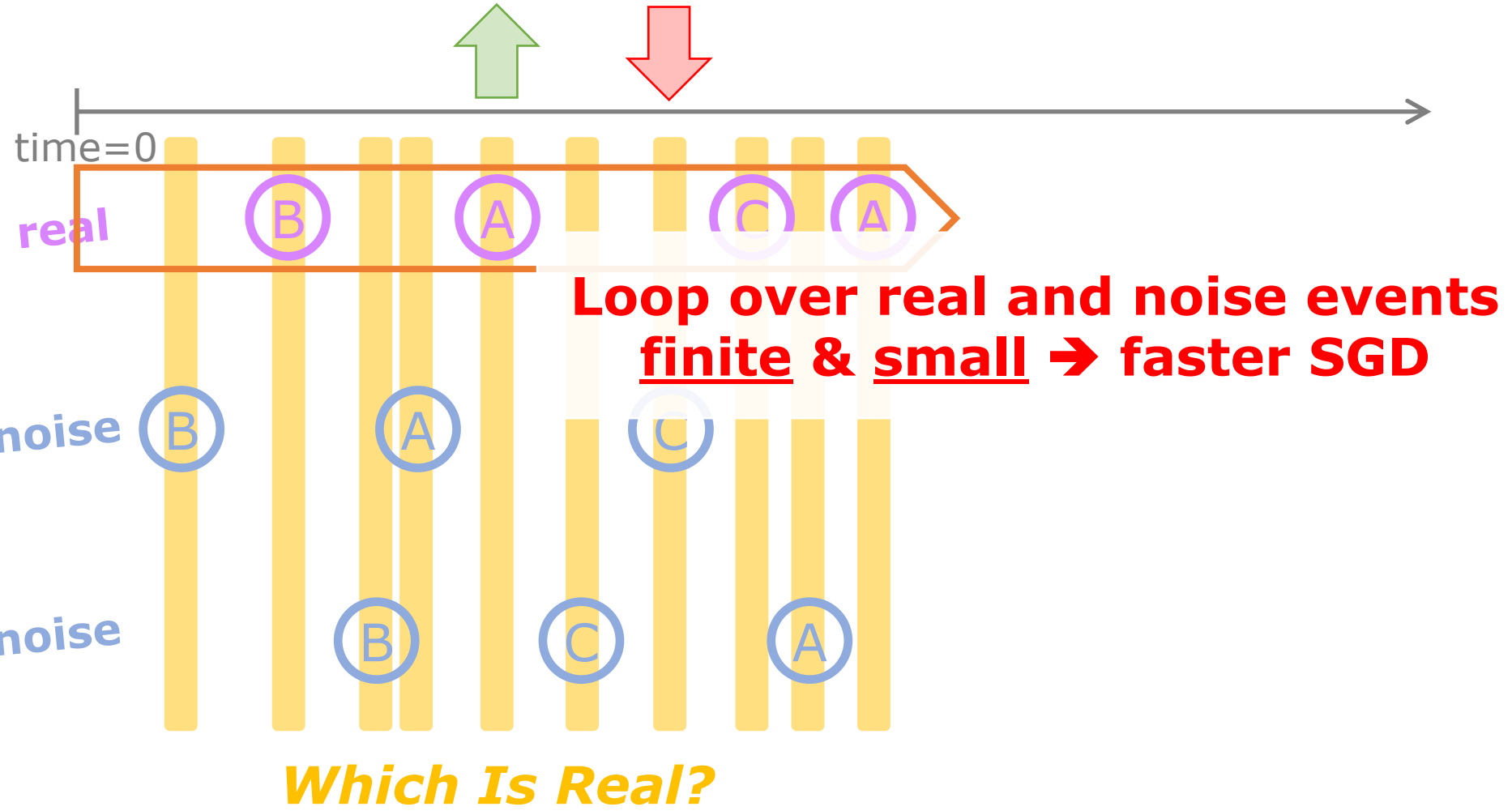
Which Is Real?

NCE: Max log prob of *correct discrimination*

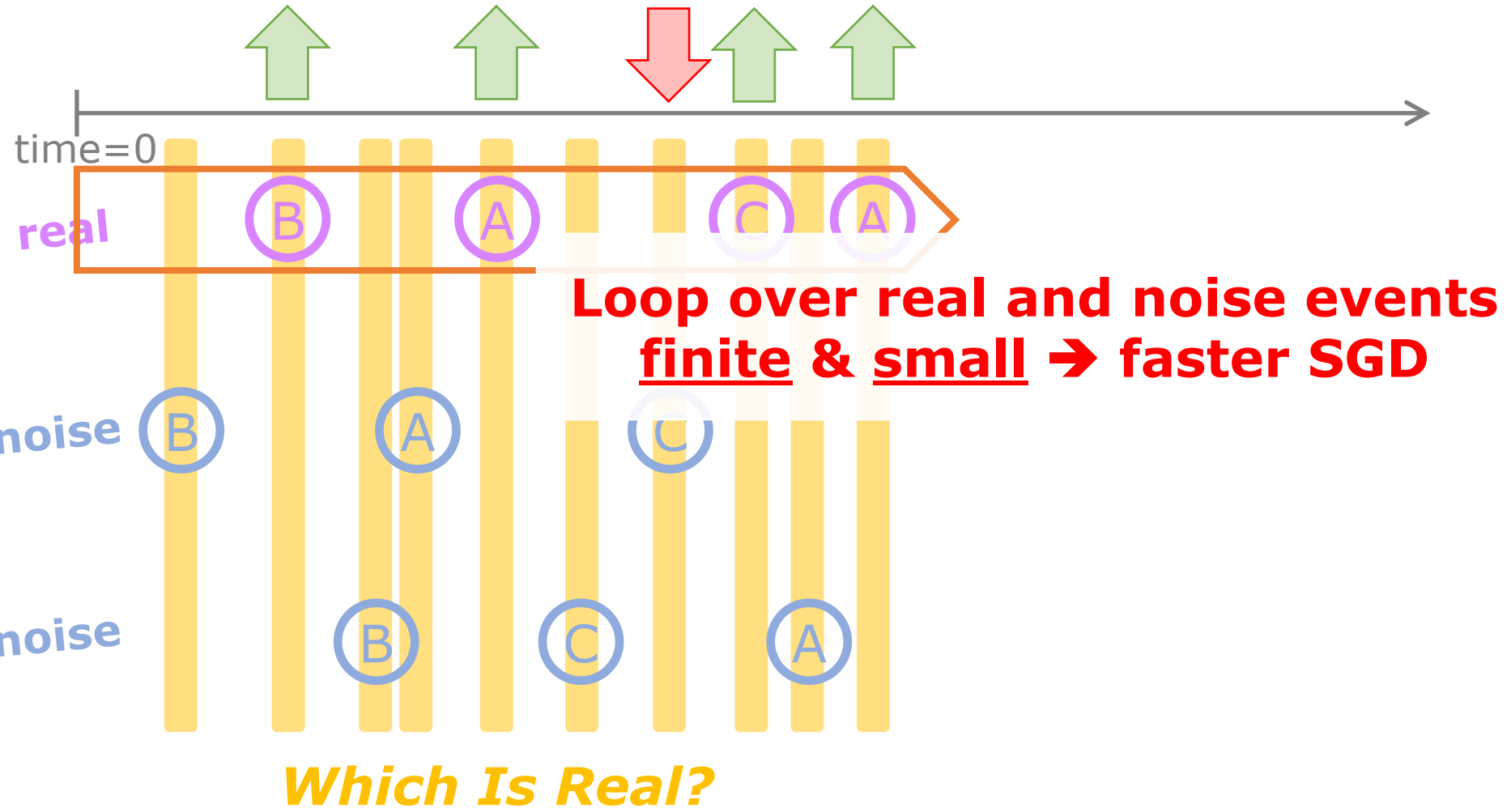


Which Is Real?

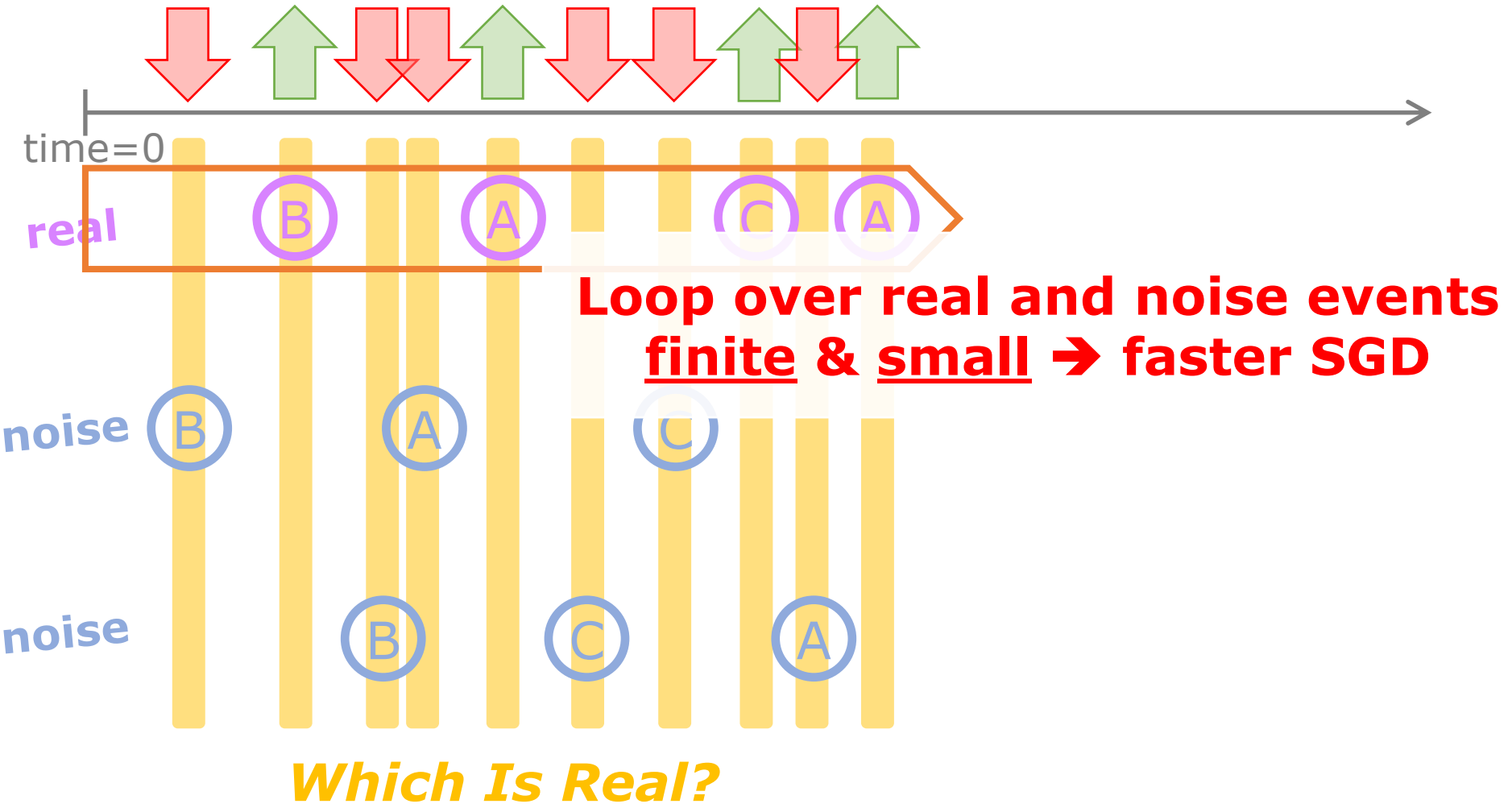
NCE: Max log prob of *correct discrimination*



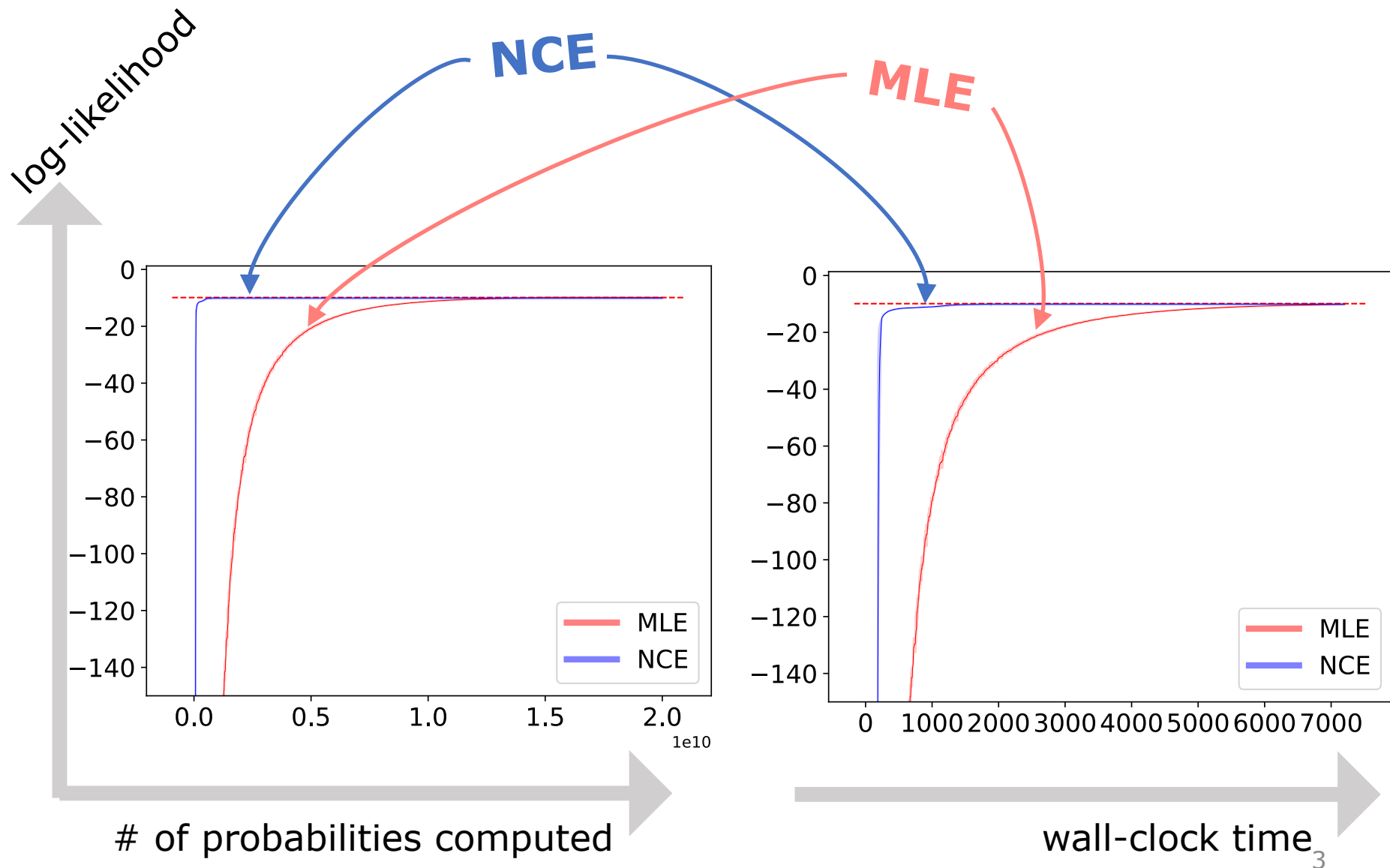
NCE: Max log prob of *correct discrimination*



NCE: Max log prob of *correct discrimination*



NCE vs MLE: what it typically looks like



NCE: More in paper

NCE: More in paper

Theorem 1 (Optimality). Under assumptions 1 and 2, $\theta \in \operatorname{argmax}_{\theta} J_{\text{NC}}(\theta)$ 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 (20) is the negative cross-entropy between the following two discrete distributions over $\{\emptyset, 1, \dots, K\}$:

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

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

theorems & proofs

Theorem 2 (Consistency). Under assumptions 1 and 2, for any $\theta \in \Theta_{\text{NC}}$, $J_{\text{NC}}^N(\hat{\theta}) \rightarrow J_{\text{NC}}(\theta)$ and $J_{\text{NC}}^N(\hat{\theta}) \rightarrow J_{\text{NC}}(\theta)$ as $N \rightarrow \infty$. Here $\|\cdot\|$ is the L_2 norm.

The intuition of this theorem is that as $N \rightarrow \infty$, the empirical distribution $p_{\hat{\theta}}$ converges to the true distribution p_{θ} . In our case, $J_{\text{NC}}^N(\hat{\theta})$ and $J_{\text{NC}}(\theta)$ will become the same as $N \rightarrow \infty$ and they are continuous functions of p_{θ} . The set Θ_{NC} is compact, so $J_{\text{NC}}^N(\hat{\theta})$ has to be the same as $J_{\text{NC}}(\theta)$ for any $\theta \in \operatorname{argmax}_{\theta} J_{\text{NC}}(\theta)$. This is almost identical to the proof of Theorem 2 in Ma & Collins (2018). But we will state 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 $M > \bar{M}$, $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, \mathbf{I}_M^{-1})$ as $N \rightarrow \infty$. Moreover, there exist a constant $C > 0$ such that for all $M > \bar{M}$, $\|\mathbf{I}_M^{-1} - \mathbf{I}_M^{-*}\| \leq C/M$. (39)

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

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

NCE: More in paper

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

We first need to highlight the key insight that $H_{\theta}(k, t, x_{[0,t]}^0)$ in equation (6) is the negative cross-entropy between the following two discrete distributions over $\{\emptyset, 1\}$:

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

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

how to draw noise fast

theorems & proofs

Theorem 2 (Convexity). Under assumptions 1 and 2, $J_{\text{NCE}}(\theta)$ is almost everywhere concave and $J_{\text{NCE}}(\theta)$ is almost everywhere convex. $J_{\text{NCE}}(\theta)$ is almost everywhere concave and $J_{\text{NCE}}(\theta)$ is almost everywhere convex.

Theorem 3 (Efficiency). Under assumptions 2 and 4-7, there exists an integer M such that for any compact set Θ' of θ , $J_{\text{NCE}}(\theta) - J_{\text{NCE}}(\theta^*) \geq J_{\text{NCE}}^M(\theta) - J_{\text{NCE}}^M(\theta^*)$.

B.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 $x_{[0,t]}^0$ 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 $t_i \sim t_{i-1} + \text{Exp}(\bar{\lambda})$. We would then set $x_{t_i} = \emptyset$ for all of the intermediate times $t \in (t_{i-1}, t_i)$, and finally draw the type x_{t_i} of the event at time t_i , choosing k with probability $\lambda_k(t_i | \mathcal{H}(t_i)) / \bar{\lambda}$. 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: $\bar{\lambda} \geq \lambda(t | \mathcal{H}(t))$ for all $t \geq t_{i-1}$. In this case, there may be "leftover" probability mass not allocated to any k . This mass is allocated to a rejected proposal. If $x_{t_i} = \emptyset$ means there was no event at time t_i , 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 $x_{[0,t]}^m$ of noise events, we run the thinning algorithm, using the noise intensity functions λ_k^{θ} . 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 $\{x_{t_1}^0, x_{t_2}^0, \dots, x_{t_M}^0\}$ are conditioned on $x_{[0,t]}^0$ (akin to the discrete-time case). 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 $\lambda^{\theta}(t | \mathcal{H}(t))$. For large K , this sum is expensive if we take the noise distribution on each training minibatch to

$$\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, \mathbf{I}_M^{-1}) \text{ as } N \rightarrow \infty$$

$$\|\mathbf{I}_M^{-1} - \mathbf{I}_*^{-1}\| \leq C/M$$

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

$$\nabla_{\theta} J_{\text{NCE}}^M(\hat{\theta}) = \nabla_{\theta} J_{\text{NCE}}^M(\theta^*) + (\hat{\theta} - \theta^*) \int_{\theta^*}^{\hat{\theta}} \nabla_{\theta}^2 J_{\text{NCE}}^M(\theta^* + u(\hat{\theta} - \theta^*)) dt \quad (41)$$

NCE: More in paper

Theorem 1 (Optimality). Under assumptions 1 and 2, $\theta \in \operatorname{argmax}_{\theta} J_{\text{NCE}}(\theta)$ 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 (6) is the negative cross-entropy between the following two discrete distributions over $\{\emptyset, 1\}$

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

theorems & proofs

how to draw noise for

more results & analysis

Theorem 2 (Conv and $M \geq 1$, with norm. The intuition of this theorem is the same as $N \rightarrow \infty$ and they are continuous to some member of the set $\operatorname{argmax}_{\theta} J_{\text{NCE}}(\theta)$ in Ma & Collins (2018). But we will study the assumption in Theorem 2, by classical large sample theory (Ferguson & Theorem 3 (Efficiency). Under assumptions 2 and 4-7, there exists an integer M for some non-singular matrix \mathbf{I}_M^* . Moreover, there exist a constant $C > 0$ such that for all $m > M$ where $\|\mathbf{I}\|$ is the spectral norm of matrix \mathbf{I} .

$$\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow \text{Normal}(0, \mathbf{I}_M^*) \text{ as } N \rightarrow \infty$$

$$\|\mathbf{I}_M^* - \mathbf{I}_m^*\| \leq C/M$$

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

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

B.1 Efficient Sampling of Noise Events

The thinning algorithm (Lewis & Shedler, 1975) drawing an event stream over a given observation time process. Suppose we have already drawn the future time $t \geq t_{i-1}$, let $\mathcal{H}(t)$ denote the context and define $\lambda(t | \mathcal{H}(t)) \triangleq \sum_{k=1}^K \lambda_k(t | \mathcal{H}(t))$. If the next event time as $t_i \sim t_{i-1} + \text{Exp}(\lambda)$. We will draw $\lambda_k(t_i | \mathcal{H}(t_i)) / \lambda$, and finally draw the type x_{t_i} of the event. If the foregoing method, taking $\bar{\lambda}$ to be any upper bound: $\bar{\lambda} \geq \lambda(t | \mathcal{H}(t))$ for all $t \geq t_{i-1}$. In this case there may be "leftover" probability mass not allocated to any k . 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 $\{x_{t_1}^0, x_{t_2}^0, \dots, x_{t_M}^0\}$ are conditioned on $x_{[0,t]}^0$ (akin to the discrete-time case). 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 $\lambda^q(t | \mathcal{H}(t))$. For large K , this sum is expensive if we take the noise distribution on each training minibatch to



THANK YOU

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