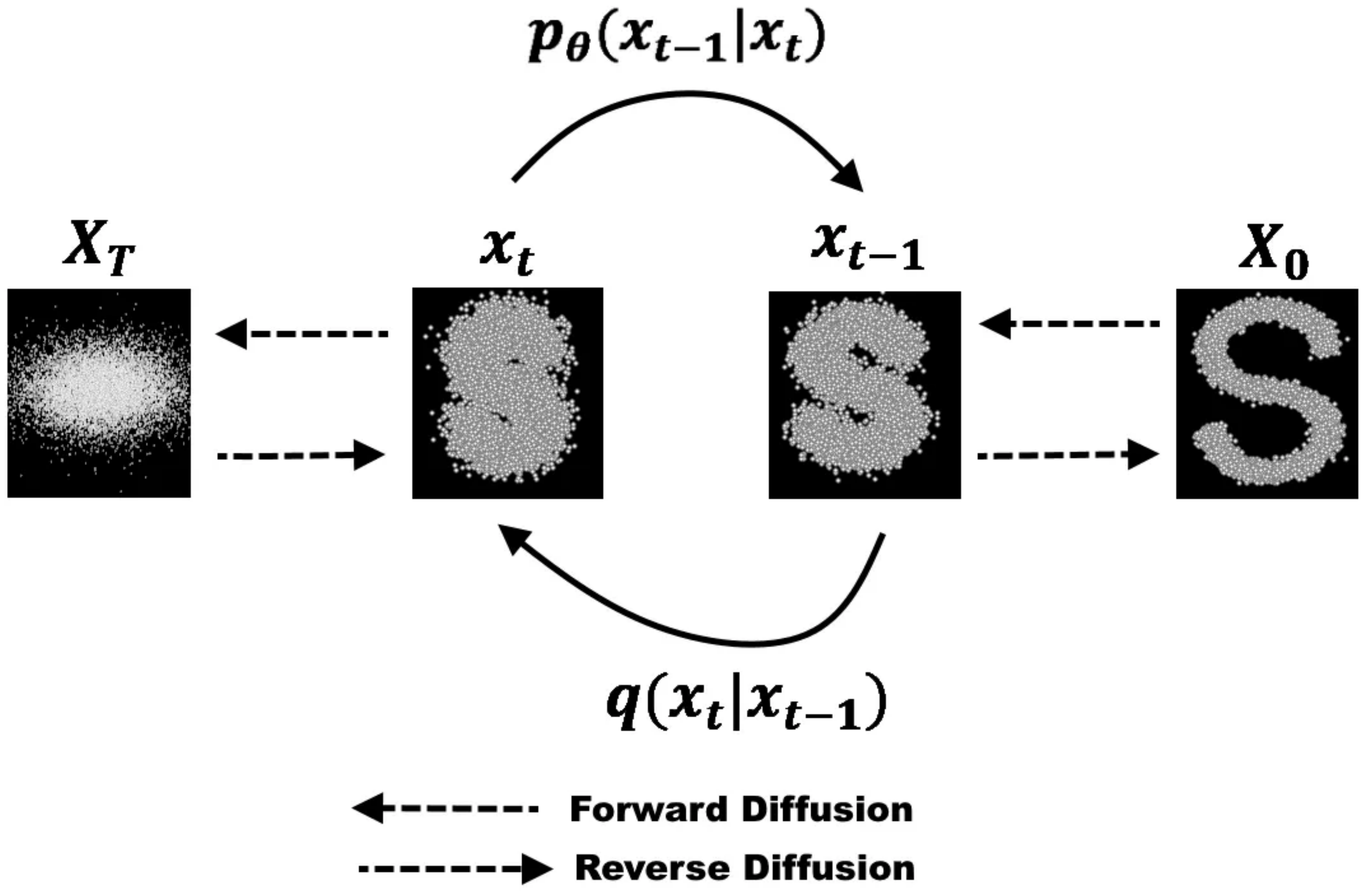
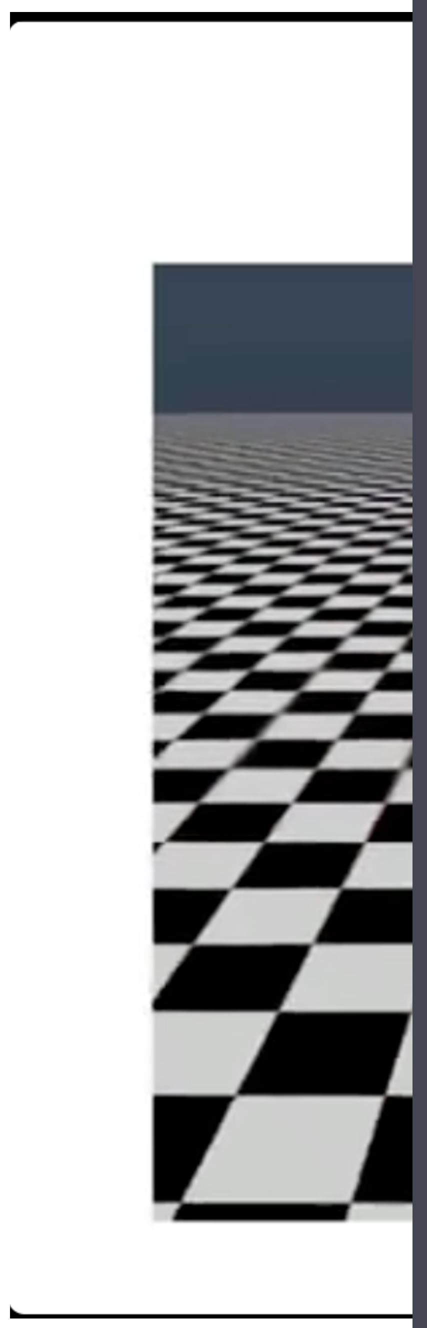
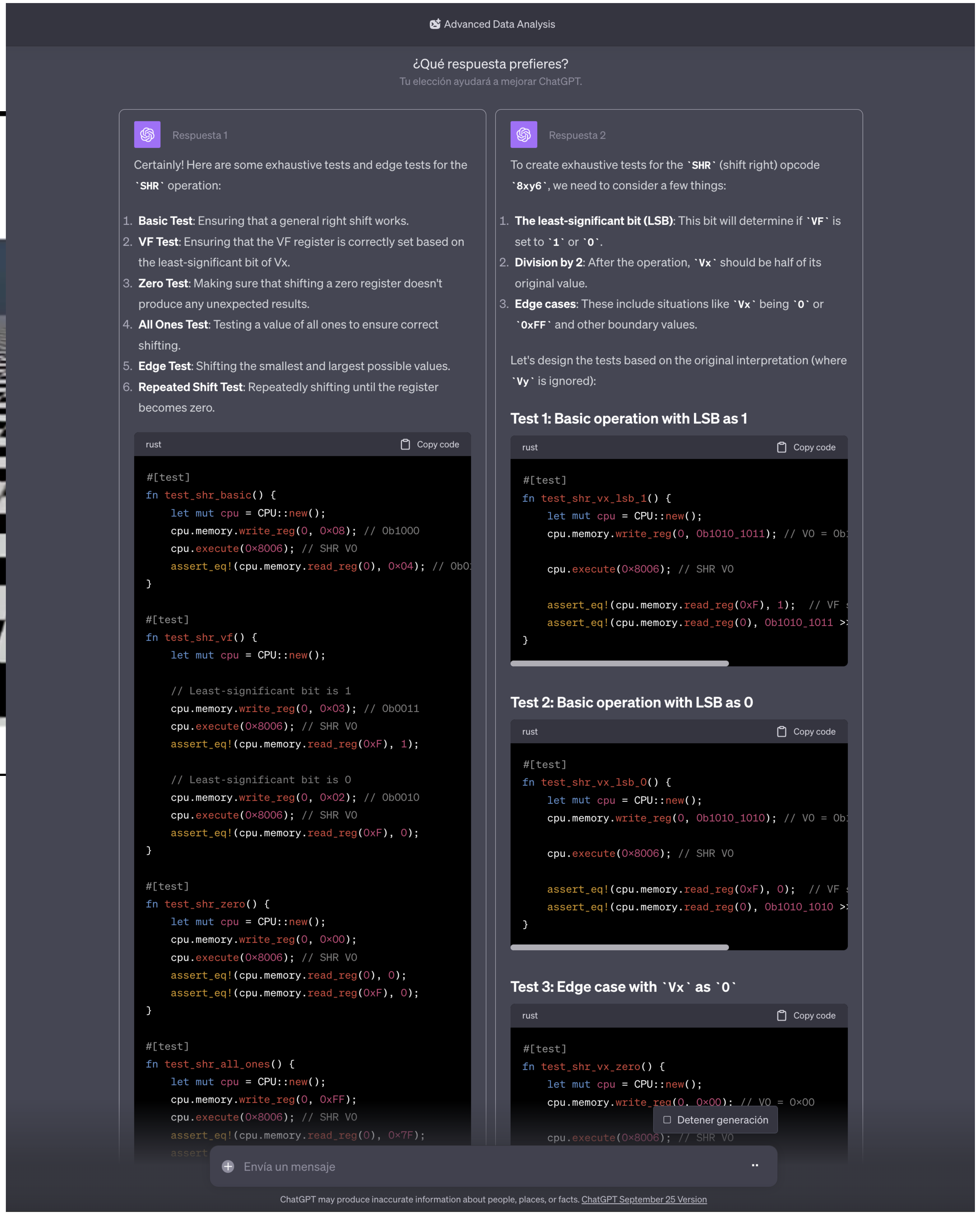
Preference Learning with Response Time: Robust Losses and Guarantees

Ayush Sawarni, Vasilis Syrgkanis

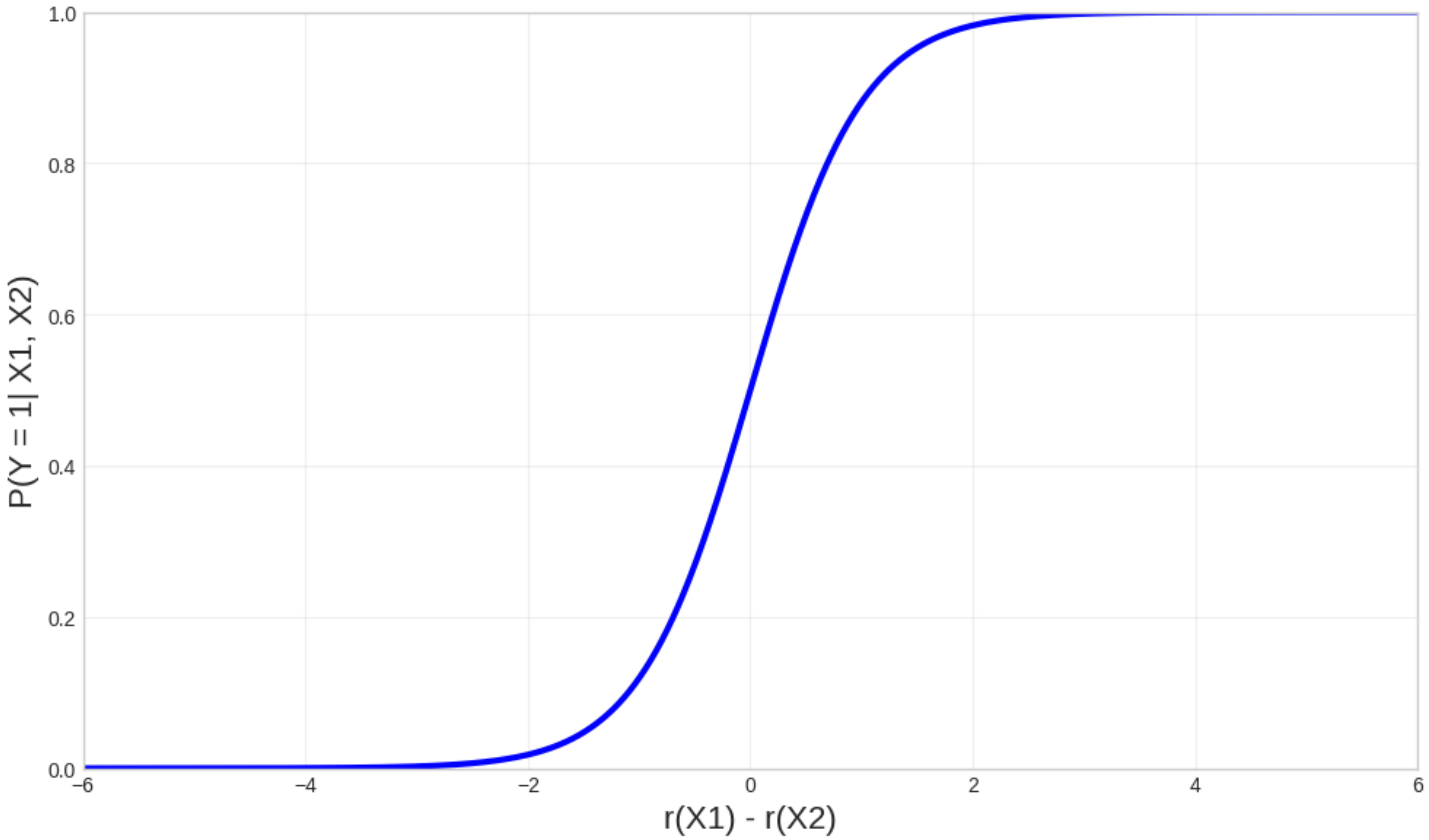
Learning from Human Preferences







Preference Learning



Observe: Signal strength reduces as r(X1) - r(X2) grows. **Vanishing** gradient problem. Learning reward model gets harder.

Using response time

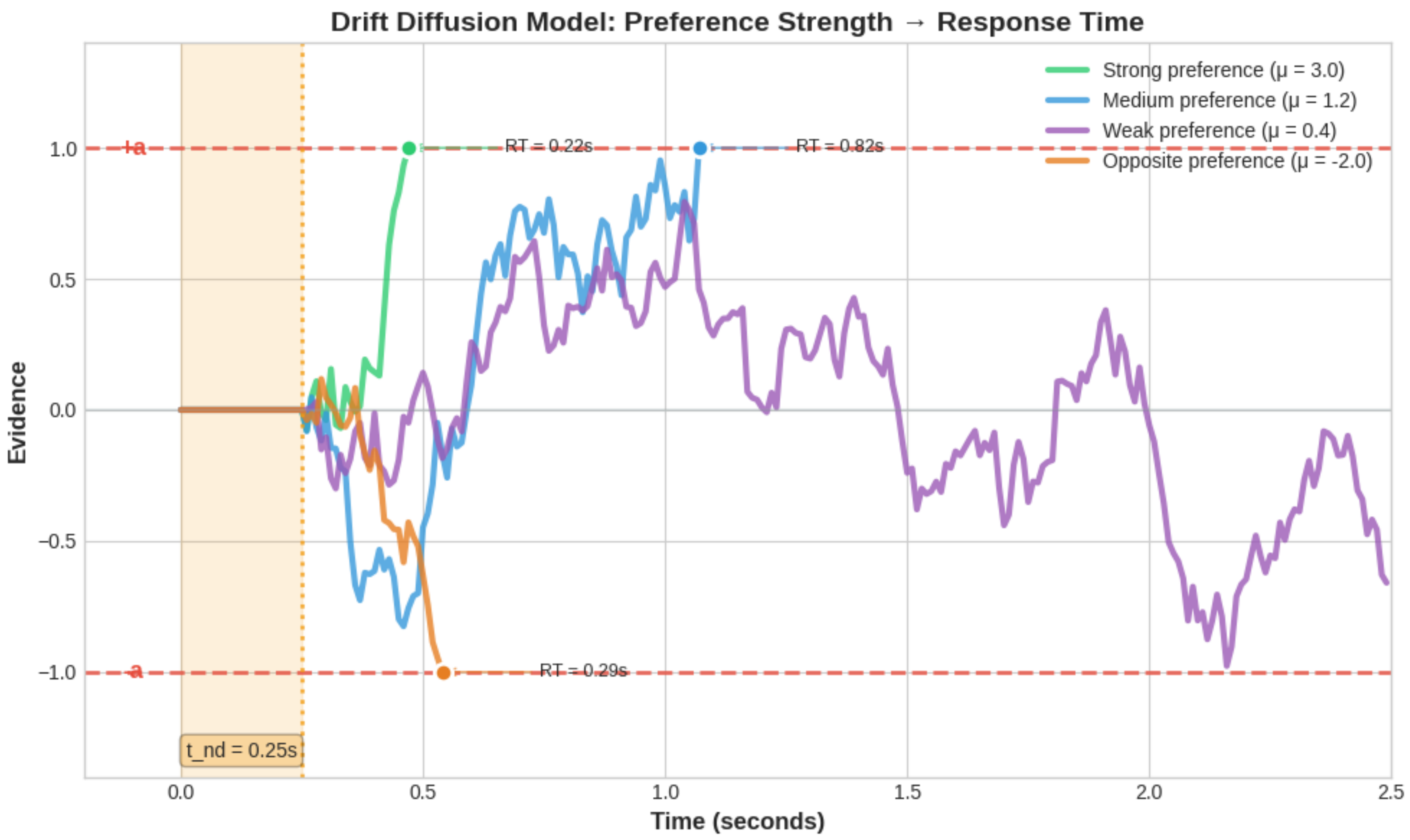
- Intuitively: stronger preference would lead to a shorter response time.
- Revealed strength of preferences. *(Chabris et al., 2008), (Konovalov, Krajbich 2019), (Bavard et al., 2024).*
- Cognitive Psychology, Neuroscience, Economics - Drift Diffusion Models.

  EZ-Diffusion Model (Wagenmakers et. al. 2007) - computationally efficient parameter estimation.
- Based the reward difference $r(X)$ the decision making happens via the following process.

  A. User spends t_{nondec} time encoding the choices.

  B. Makes a decision Y after and "evidence accumulation" phase of duration T

  C. Evidence accumulation is modelled as a diffusion process with drift $r(X)$



Drift Diffusion Model: Preference Strength → Response Time
Strong preference (μ = 3.0)
Medium preference (μ = 1.2)
Weak preference (μ = 0.4)
Opposite preference (μ = -2.0)
+a
-a
RT = 0.22s
RT = 0.82s
RT = 0.29s
t_nd = 0.25s
Evidence
Time (seconds)
1.0
0.5
0.0
−0.5
−1.0
0.0
0.5
1.0
1.5
2.0
2.5

Learning r

- MLE of response time?
- Likelihood is **not** tractable. :(

$$f_T(t) = \frac{a}{\sqrt{2\pi t^3}} \sum_{k=-\infty}^{\infty} (2k+1) \exp\left[-\frac{\big((2k+1)a - t \cdot r(X)\big)^2}{2t}\right], \qquad t > 0.$$

- Maybe use the moment expressions. Can show the following relation.

$$r(X) = \frac{\mathbb{E}[Y \mid X]}{\mathbb{E}[T \mid X]}$$

- Loss function using the above relation.

Goal: Construct a loss function to learn r

- **Objective**: given $Z_i = (X_i^1, X_i^2, Y_i, T_i)$ $i = 1, \cdots, n$ - learn the reward function r
- With **only** preferences, the following loss function naturally follows from BT model.
  - $\mathscr{L}^{\text{logloss}}(r) = \mathbb{E}\left[\log\left(1 + \exp(-2\, Y\, r(X))\right)\right].$
  - **Naive loss (with response time)**: We may start with the following loss function
    - $\mathscr{L}^{\text{non-ortho}} = \mathbb{E}\left[\left(Y - r(X)\mathbb{E}[T \mid X]\right)^2\right]$

$\mathbb{E}[T \mid X]$ is unknown, learn $t_o(X) = \mathbb{E}[T \mid X]$ Loss function is sensitive to the estimate of t_o.

- Learning t_0 can be hard.
- t_o may be highly non linear. **Example** - linear $r(X) = \langle \theta, x \rangle$, $t(X) = \dfrac{\tanh(r(X))}{r(X)}$ (non-linear, MSE non-convex in θ).

Learning with Nuisances - Orthogonal Statistics

Neyman orthogonality Loss functions that are less sensitive to errors in Nuisance estimate.

Directional derivative (Gateaux derivative) of a functiona $F: \mathcal{F} \rightarrow \mathbb{R}$ at f direction h is defined by

$$D_f F(f)[h] = \frac{d}{dt} F(f + th)\Big|_{t=0}$$

Goal: Orthogonal loss function with respect to response time function $t(X)$.

Orthogonal Loss

$$\mathscr{L}(r; \mathfrak{r}, t) = \mathbb{E}\left[(Y - (T - t(X))\mathfrak{r}(X) - r(X)t(X))^2\right]$$

- t : an estimate of $t_o(X)$
- $\mathfrak{r}$: a crude estimate of reward function (minimizing log-loss for example)
- Orthogonal wrt to nuisance pair $g = (t, \mathfrak{r})$

First stage:

- Minimize the log-loss function to estimate reward model $\mathfrak{r}$.
- Can fit a neural network to estimate **t(X)**.

Second stage:

- Use $t(.)$ and $\mathfrak{r}(.)$ in second stage with ortho-loss function

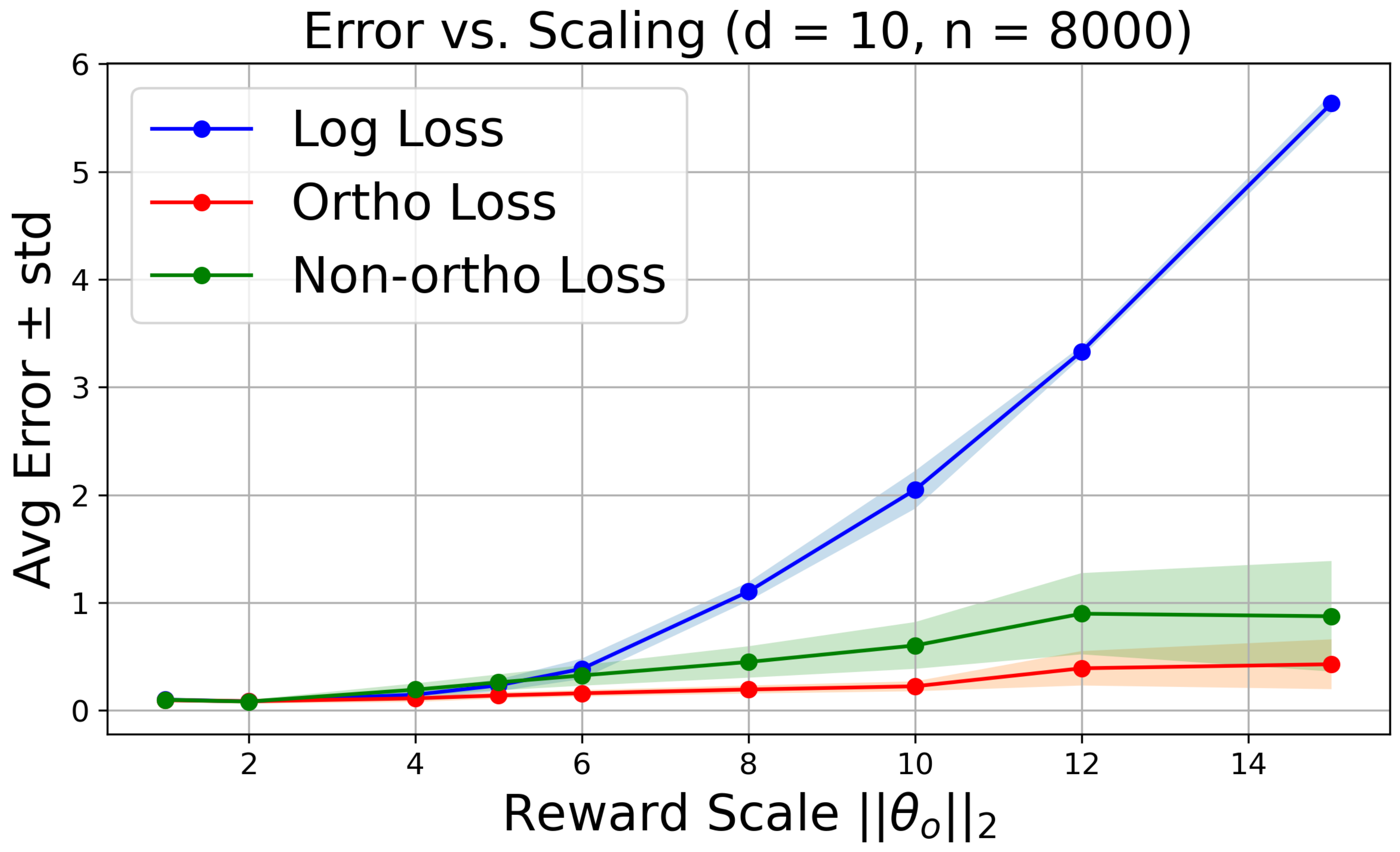
Theoretical results:

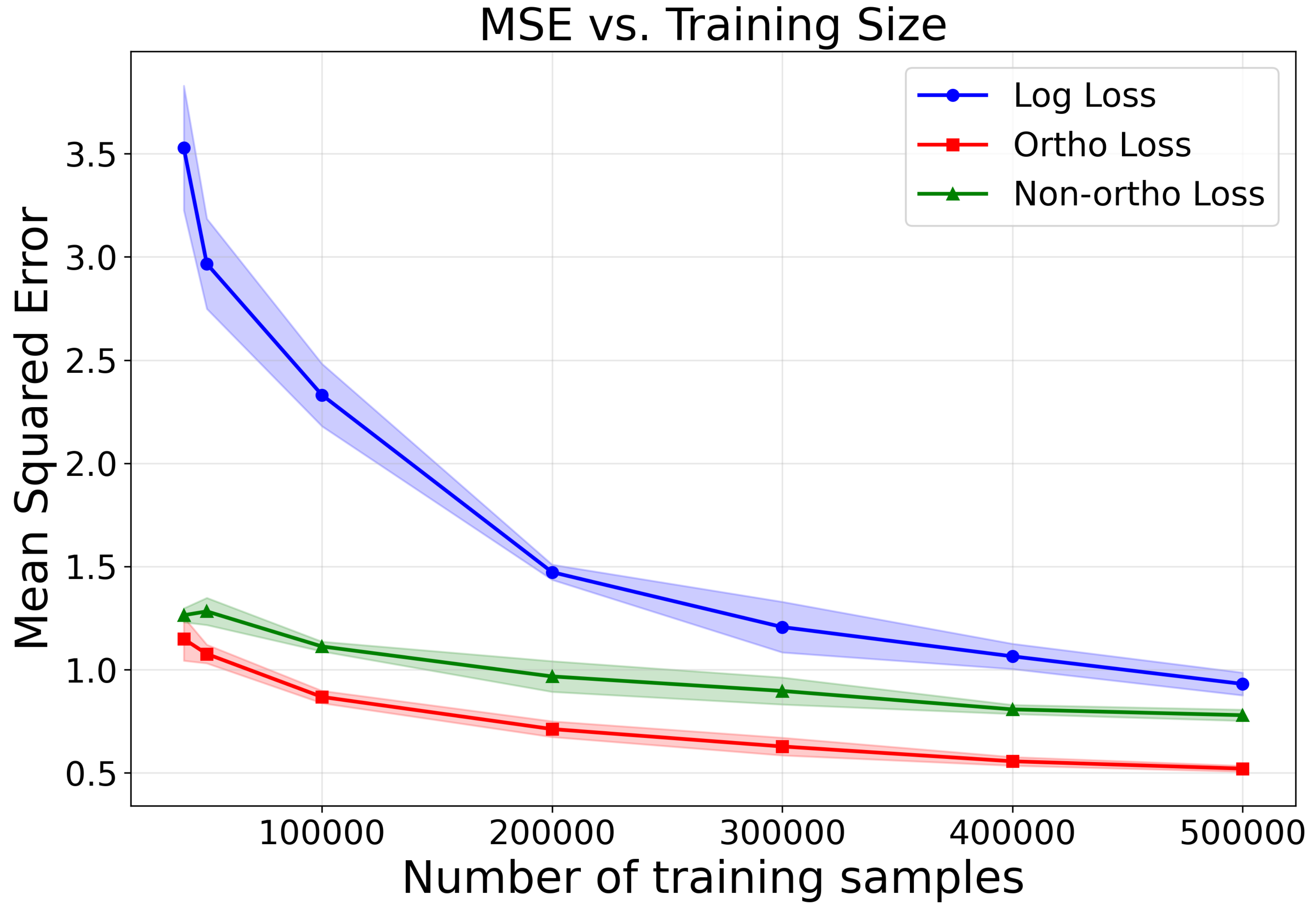
- Linear reward $r_o(X) := \langle \theta_o, X \rangle$ **Goal:** Estimate θ_o. Assume $|\theta_o| \leq S$ and X to be uniformly bounded by 1.

Key result:

- Variance of the **choice-only estimator** grows *exponentially* with S.
- Variance of the **ortho-loss estimator** grows only *polynomially* with S.

Experiments





Future work - Implicit reward learning

DPO style objective: two stage learning

First stage (learn policy $\boldsymbol{\phi}$ standard DPO loss) - $\mathfrak{r}(X) = c\left(\log\frac{\phi(X^1)}{\pi_{\text{ref}}(X^1)} - \log\frac{\phi(X^2)}{\pi_{\text{ref}}(X^2)}\right)$

- Second stage:

$$\mathscr{L}^{\text{ortho}}(r;\mathfrak{r}) = \mathbb{E}\left[\left(Y - T\,\mathfrak{r}(X) + \tanh\big(\mathfrak{r}(X)\big) - r(X)\frac{\tanh\big(\mathfrak{r}(X)\big)}{\mathfrak{r}(X)}\right)^2\right]$$

Thank you!

https://arxiv.org/abs/2505.22820