

---

# Active preference learning for ordering items in- and out-of-sample

---

**Herman Bergström\***

Chalmers University of Technology  
and University of Gothenburg  
hermanb@chalmers.se

**Emil Carlsson\*†**

Sleep Cycle AB  
Chalmers University of Technology  
and University of Gothenburg

**Devdatt Dubhashi**

Chalmers University of Technology  
and University of Gothenburg

**Fredrik D. Johansson**

Chalmers University of Technology  
and University of Gothenburg

## Abstract

Learning an ordering of items based on pairwise comparisons is useful when items are difficult to rate consistently on an absolute scale, for example, when annotators have to make subjective assessments. When exhaustive comparison is infeasible, actively sampling item pairs can reduce the number of annotations necessary for learning an accurate ordering. However, many algorithms ignore shared structure between items, limiting their sample efficiency and precluding generalization to new items. It is also common to disregard how noise in comparisons varies between item pairs, despite it being informative of item similarity. In this work, we study active preference learning for ordering items with contextual attributes, both in- and out-of-sample. We give an upper bound on the expected ordering error of a logistic preference model as a function of which items have been compared. Next, we propose an active learning strategy that samples items to minimize this bound by accounting for aleatoric and epistemic uncertainty in comparisons. We evaluate the resulting algorithm, and a variant aimed at reducing model misspecification, in multiple realistic ordering tasks with comparisons made by human annotators. Our results demonstrate superior sample efficiency and generalization compared to non-contextual ranking approaches and active preference learning baselines.

## 1 Introduction

The success of supervised learning is built on annotating items at great volumes with small error. For subjective assessments, however, assigning a value from an arbitrary rating scale can be difficult and prone to inconsistencies, causing many to favor *preference feedback* from pairwise comparisons (Yannakakis and Martínez, 2015; Christiano et al., 2017; Ouyang et al., 2022; Zhu et al., 2023). Preference feedback is sufficient to learn an *ordering* of items (Fürnkranz and Hüllermeier, 2003), but for  $n$  items, there are  $\mathcal{O}(n^2)$  possible pairs of items to compare. A common solution is to use crowd-sourcing (Chen et al., 2013; Yang et al., 2021; Larkin et al., 2022), but many tasks require domain *expertise*, making annotations *expensive* to collect. This is the case in the field of medical imaging, where annotations require trained radiologists (Phelps et al., 2015; Jang et al., 2022; Lidén et al., 2024; Tärnåsen and Bergström, 2023). So, how can we learn the best ordering possible from a limited number of comparisons?

---

\*Equal contribution. †Work was mainly performed while the author was a PhD student at Chalmers University of Technology.

Classically, this problem is solved by active learning, sampling comparisons based on preference feedback and estimated item scores (Herbrich et al., 2006; Maystre and Grossglauser, 2017; Heckel et al., 2018). However, consider a radiologist who wants to quantify the expression of a disease in a collection of X-ray images. Purely preference-based algorithms utilize only the outcomes of comparisons but ignore the contents of the X-rays, which can reveal similarities between items and inform an ordering strategy. Moreover, the set we want to order is often larger than the set of items observed during training—we may want to rank new X-rays in relation to previous ones. This cannot be solved by learning per-item scores alone. As an alternative, active learning for classification can be used to fit a map from pairs of item contexts  $x_i, x_j$  (e.g., the contents of images) to the comparison  $i \succ_{\gamma} j$ , that can be applied to old and new items alike (Houlsby et al., 2011; Qian et al., 2015). However, as we show in Section 4, learning this map to recover a *complete ordering* is distinct from the tasks preference learning is commonly used for, and existing algorithms lack theoretical justification for this application. Moreover, formal results for related problems, such as contextual bandits or reinforcement learning (Das et al., 2024; Filippi et al., 2010; Zhu et al., 2023; Bengs et al., 2022), do not translate directly to effective active sampling criteria for ordering. There is a small body of work on learning a contextual model to recover the complete ordering (Jamieson and Nowak, 2011; Ailon, 2011) but these either assume noiseless preference feedback or that the noise is unrelated to the similarity of items, which is unrealistic for subjective assessments.

**Contributions.** We propose using a contextual logistic preference model to support efficient in-sample ordering and generalization to new items. Our analysis yields the first bound on the expected ordering error achievable given a collected set of comparisons (Section 4). This result justifies an active sampling principle that accounts for both epistemic and aleatoric uncertainty which we implement in a greedy deterministic algorithm called GURO (Section 5). We further propose a hybrid variant of the contextual preference model, compatible with GURO as well as existing sampling strategies, that overcomes model misspecification by adding per-item parameters (Section 5.1). We evaluate GURO and baseline algorithms in four diverse ordering tasks, three of which utilize comparisons performed by human annotators (Section 6). Our sampling strategy compares favorably to active preference learning baselines, and our hybrid model benefits both GURO and other sampling criteria, achieving the low variance of contextual models and the low bias of fitting per-item parameters. This results in faster convergence in-sample, better generalization to new items, and efficient continual learning when new items are added.

## 2 Ordering items with active preference learning

Our goal is to learn an ordering of items  $\mathcal{I}$  according to an unobserved score  $y_i \in \mathbb{R}$ , defined for each item  $i \in \mathcal{I}$ . The ground-truth ordering of  $\mathcal{I}$  is determined by a comparison function  $\pi_{ij} := \mathbb{1}[y_i > y_j]$ , where  $\pi_{ij} = 1$  indicates that  $i$  ranks higher than  $j$ . We assume there are no ties.

We define the *ordering error*  $R_{\mathcal{I}}(h)$  of a learned comparison function  $h : \mathcal{I} \times \mathcal{I} \rightarrow \{0, 1\}$  as the frequency of pairwise inversions under a uniform distribution of item pairs,

$$R_{\mathcal{I}}(h) = \frac{2}{n(n-1)} \sum_{i \neq j \in \mathcal{I}} \mathbb{1}[h(i, j) \neq \pi_{ij}], \quad (1)$$

where  $n = |\mathcal{I}|$ . This error is equivalent to the normalized Kendall’s Tau distance (Kendall, 1948).

Hypotheses  $h$  are learned from *preference feedback*—noisy pairwise comparisons  $C_{ij} \in \{0, 1\}$  for items  $(i, j)$  related to their score, for example, provided by human annotators.  $C_{ij} = 1$  indicates that an annotator perceived that item  $i$  has a higher score than  $j$ , i.e., that they prefer  $i$  over  $j$ . *Our goal is to minimize the ordering error  $R_{\mathcal{I}}(h)$  for a fixed budget  $T \geq 1$  of adaptively chosen comparisons.*

We are interested in contextual problems, where each item  $i \in \mathcal{I}$  is endowed with item-specific attributes  $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ . As we will see, this permits more sample-efficient ordering and learning algorithms that can order items out-of-sample, trained on comparisons of a subset of items  $\mathcal{I}_D \subseteq \mathcal{I}$  and generalizing to  $\mathcal{I} \setminus \mathcal{I}_D$ . Ordering algorithms based *only* on preference feedback cannot solve this problem since observed comparisons are uninformative of new items.

Our *active preference learning* scenario proceeds as follows: 1) A learner is given an annotation budget  $T$ , a pool of items  $\mathcal{I}_D \subseteq \mathcal{I}$  and item attributes  $x_i$  for  $i \in \mathcal{I}_D$ . 2) Over rounds  $t = 1, \dots, T$ , the learner requests a comparison of two items  $i_t, j_t \in \mathcal{I}_D$  according to a sampling criterion and

receives noisy binary preference feedback  $c_t \sim p(C_{ij})$ , independently of previous comparisons. 3) After  $T$  rounds, the learner returns a comparison function  $h : \mathcal{I} \times \mathcal{I} \rightarrow \{0, 1\}$ . We denote the history of accumulated observations until and including time  $t$  by  $D_t = ((i_1, j_1, c_1), \dots, (i_t, j_t, c_t))$ .

We assume that comparisons  $C_{ij}$  follow a logistic model applied to the difference between item scores,  $p(C_{ij} = 1) = \sigma(y_i - y_j)$ , the so-called Bradley-Terry model (Bradley and Terry, 1952), which assumes linear stochastic transitivity (Oliveira et al., 2018). Throughout,  $\sigma(z) = 1/(1 + e^{-z})$  and  $\dot{\sigma}(z)$  its derivative at  $z$ . Specifically, we study the case where  $y_i$  is a linear function of item attributes,  $y_i = \theta_*^\top x_i$ , with  $\theta_* \in \mathbb{R}^d$  the ground-truth coefficients. Thus, comparisons are determined by a logistic regression model applied to the attribute difference vector  $z_{ij} := x_i - x_j$ ,

$$p(C_{ij} = 1) = \sigma(\theta_*^\top z_{ij}). \quad (2)$$

We face two kinds of uncertainty when actively learning the model in (2): *epistemic* and *aleatoric*. Epistemic uncertainty, or model uncertainty, is the uncertainty about the true parameter  $\theta_*$ , while aleatoric uncertainty is the irreducible uncertainty about labels due to noisy annotation.

### 3 Related work

**Active Preference Learning:** *Preference learning* (Fürnkranz and Hüllermeier, 2003; Chu and Ghahramani, 2005) is related to the problem of *learning to rank* (Burges et al., 2005; Busse et al., 2012). When using adaptively chosen comparisons it may be posed as an *active learning* or *bandit* problem (Brinker, 2004; Long et al., 2010; Silva et al., 2014; Ling et al., 2020). Non-contextual active learners, such as TrueSkill (Herbrich et al., 2006; Minka et al., 2018), Hamming-LUCB (Heckel et al., 2018), and Probe-Rank (Lou et al., 2022) produce in-sample preference orderings, but must be updated if new items are to be ranked. Contextual algorithms, such as BALD (Houlsby et al., 2011), mitigate this by exploiting item structure and Kirsch and Gal (2022) show that many recently proposed contextual active learning strategies may be unified in a framework based on Fisher information. Similarly, methods have been proposed to recover a linear preference model by adaptively sampling paired comparisons (Qian et al., 2015; Massimino and Davenport, 2021; Canal et al., 2019). Still, this setting differs from ours in that we emphasize recovering the full ordering, not perfectly estimating the parameters. While it is true that knowing the parameters is sufficient to order the list, reducing uncertainty for all parameters equally will likely be wasteful (see Section 4). Ailon (2011) offer guarantees for ordering using contextual features in the noiseless setting, while Jamieson and Nowak (2011) analyze the setting where noise is unrelated to item similarity.

**Bandits:** Bandit algorithms with *relative* or *dueling* feedback (Yue and Joachims, 2009; Bengs et al., 2021; Yan et al., 2022) also learn from pairwise comparisons, and have been proposed both in contextual (Dudík et al., 2015) and non-contextual settings (Yue et al., 2012) to minimize regret or identify top- $k$  items. Bengs et al. (2022) proposed CoLSTIM, a contextual dueling bandit for regret minimization under linear stochastic transitivity, matching (2), and Di et al. (2023) gave variance-aware regret bounds for this setting. However, algorithms that find the top- $k$  items, such as pure exploration bandits (Fang, 2022; Jun et al., 2021), can be arbitrarily bad at learning a full ordering (see Appendix D). Related are also George and Dimitrakakis (2023) who learn Kemeny rankings in non-contextual dueling bandits, and Wu et al. (2023b) who minimize Borda regret. Zhu et al. (2023) studies the problem of estimating a preference model from offline data. Our analysis uses techniques from logistic bandits (Filippi et al., 2010; Li et al., 2017; Fauray et al., 2020; Kveton et al., 2020).

**RLHF:** Preference learning is commonly used when training large language models through reinforcement learning with human feedback (RLHF) (Christiano et al., 2017; Bai et al., 2022; Ouyang et al., 2022; Wu et al., 2023a). In this line of work, Zhu et al. (2023) provide guarantees on the sample complexity of learning a preference model from offline data. They leverage similar tools from statistical learning and bandits as we do. In contrast to their work, we provide sampling strategies for the online setting. Mehta et al. (2023) consider active learning for RLHF in a dueling bandit framework where the goal is to optimize a contextual version of the Borda regret. Concurrent work by Mukherjee et al. (2024) and Das et al. (2024) studies a similar problem, as we do here, in the RLHF setting but with the objective to identify an optimal policy in a contextual bandit with dueling feedback. In contrast to their objective, we are interested in recovering the ordering of items. Das et al. (2024) use similar bandit techniques as we do, and their selection criterion, when adapted for ordering, corresponds to our NormMin baseline (see Section 6).

## 4 Which comparisons result in a good ordering?

We give an upper bound on the ordering error  $R_{\mathcal{I}}(h)$  for a hypothesis  $h$  fit using the feedback from a given set of  $T$  comparison queries  $\mathcal{Q}_T = ((i_1, j_1), \dots, (i_t, j_t))$ . In other words, the bound attempts to answer the question “if we make queries  $\mathcal{Q}_T$ , how good can we expect our resulting model to be at ordering the items in  $\mathcal{I}$ ?”. That is, we condition on the queries themselves and reason about the uncertainty due to the stochastic feedback  $c_t$ . In Section 5, we use insights from the result to design an active learning algorithm.

We restrict our analysis to the logistic model in (2) and denote by  $R(\theta) \equiv R_{\mathcal{I}}(h_\theta)$  the risk of the hypothesis defined by  $h_\theta(i, j) = \mathbb{1}[\theta^\top z_{ij} > 0]$ . Recall that  $z_{ij} = x_i - x_j$  for  $i, j \in \mathcal{I}$ , and define  $z_t \equiv z_{i_t j_t}$  as the difference between attributes for the pair of items selected at round  $t = 1, \dots, T$ . Let  $\theta_t$  be the maximum-likelihood estimate (MLE) fit to  $t$  rounds of feedback,  $D_t$

$$\theta_t = \arg \max_{\theta} \sum_{s=1}^t (c_s \log \sigma(\theta^\top z_s) + (1 - c_s)(1 - \sigma(\theta^\top z_s))) . \quad (3)$$

Let  $\Delta_{ij} > 0$  lower bound the margin of comparison,  $|\sigma(z_{ij}^\top \theta_*) - 1/2| > \Delta_{ij}$  for all  $i, j \in \mathcal{I}$  and define  $\Delta_* = \min_{i \neq j} \Delta_{ij}/|i - j|$ . Next, let  $\mathbf{H}_t(\theta) := \sum_{s=1}^t \dot{\sigma}(z_s^\top \theta) z_s z_s^\top$  be the Hessian of the negative log-likelihood of observations at time  $t$  under (2), given the parameter  $\theta$ , also known as *observed Fisher information*. We define  $\tilde{\mathbf{H}}_t(\theta) := \frac{1}{t} \mathbf{H}_t(\theta)$ . For a square matrix  $V$ , we define  $\|x\|_V = \sqrt{x^\top V x}$ . We make the following assumptions for our analysis:

**Assumption 1.**  $\theta_*$  satisfies  $\|\theta_*\|_2 \leq S$  for some  $S > 0$ .

**Assumption 2.**  $\forall i \in \mathcal{I}$ , we have  $\|x_i\|_2 \leq Q$  for  $Q > 0$ .

**Assumption 3.**  $\mathbf{H}_T(\theta_T)$  and  $\mathbf{H}_T(\theta_*)$  have full rank and minimum eigenvalues larger than  $\lambda_0 > 0$ .

Assumption 1 implies that  $\theta_*$  lies in some ball with radius  $S$  and cannot have unbounded coefficients. Assumption 2 states that there exists an upper bound on the norm of the feature vectors. This assumption is trivially satisfied whenever we have a finite set of data points. Both assumptions 1 and 2 are standard in the bandit literature and only required for our analysis. Assumption 3 is naturally satisfied for sufficiently large  $T$  by any sampling strategy with support on  $d$  linearly independent vectors, or can be ensured by allowing for a burn-in phase of  $d$  samples. Assumption 3 ensures the uniqueness of  $\theta_t$ .

We start by stating the following concentration result for the deviation of  $\sigma(z_{ij}^\top \theta_T)$  from the true probability  $\sigma(z_{ij}^\top \theta_*)$ . Recall that, while queries  $\mathcal{Q}_T$  (and therefore  $\{z_t\}_{t=1}^T$ ) are fixed, the stochasticity in the feedback  $c_t$  implies that  $\theta_T$  and consequently  $\tilde{\mathbf{H}}_T^{-1}(\theta_T)$  are random. The proof of Lemma 1 is found in Appendix C and builds on results for optimistic algorithms in logistic multi-armed bandits (Filippi et al., 2010; Faury et al., 2020).

**Lemma 1** (Concentration Lemma). *Define, for all pairs of items  $i, j \in \mathcal{I}$ , and any  $\Delta > 0$ ,*

$$\alpha_{ij}(\Delta) := \exp\left(\frac{-\Delta^2 T}{8dC_1(\dot{\sigma}(z_{ij}^\top \theta_T)\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)})^2}\right), \quad \beta_{ij}(\Delta) := \exp\left(\frac{-\Delta T}{dC_1\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}^2}\right).$$

*Then, if  $\alpha := \alpha_{ij}(\Delta)$ ,  $\beta := \beta_{ij}(\Delta)$  and  $\alpha, \beta \leq \frac{1}{4dT}$ ,*

$$P(|\sigma(z_{ij}^\top \theta_*) - \sigma(z_{ij}^\top \theta_T)| > \Delta) \leq 2dT(\alpha + \beta).$$

*$C_1$  depends on  $S, \lambda_0, Q$  from Assumptions 1–3 (see Appendix C for definition and proof).*

The concentration result in Lemma 1 is *verifiable* (given by observables) since the upper bound depends only on the maximum likelihood estimate  $\theta_T$  at time  $T$ , not on  $\theta_*$ . We present a sharper, *unverifiable* bound in Appendix C which instead depends on  $\theta_*$  but does not suffer from the explicit scaling with  $d$  in the definitions of  $\alpha$  and  $\beta$ . The bound in Lemma 1 can also be expressed in terms of  $\mathbf{H}_T^{-1}(\theta_T)$  by using the equality  $\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}^2 = \frac{1}{T}\|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}^2$ . As long as our sampling strategy ensures that the minimum eigenvalue of  $\tilde{\mathbf{H}}_T(\theta_T)$  does not tend to zero, i.e., the strategy is *strongly consistent* (Chen et al., 1999), we have  $\alpha_{ij}(\Delta_{ij}) \sim \exp[-\Delta_{ij}^2 T / (\dot{\sigma}(z_{ij}^\top \theta_T)^2 \|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}^2)]$  and

$\beta_{ij}(\Delta_{ij}) \sim \exp[-\Delta_{ij}T/\|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}^2]$ . Since  $\Delta_{ij}^2 < \Delta_{ij} < 1/2$  by definition, we can view  $\alpha$  as the *first-order* term and  $\beta$  as the *second-order* term of our bound.

Lemma 1 formally captures the intuition that it should be easier to sort when annotations contain little noise, i.e.,  $\dot{\sigma}(z_{ij}^\top \theta_T)$  is small. Especially, we observe  $\dot{\sigma}(z_{ij}^\top \theta_T) \approx 0$  for pairs where  $\Delta_{ij}$  is sufficiently large, causing the first-order term to vanish, leaving us with the faster decaying second-order term  $\beta$ . Lemma 1 also tells us that the hardest pairs to guarantee a correct ordering for are the ones with both high *aleatoric* uncertainty under the MLE model, e.g., where annotators disagree or labels are noisy, captured by  $\dot{\sigma}(z_{ij}^\top \theta_T)$ , as well as high *epistemic* uncertainty captured by  $\|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}$ .

A direct consequence of Lemma 1 is the following bound on the ordering error of  $h_{\theta_T}$  over  $\mathcal{I}$ ,

$$\mathbb{E}[R(\theta_T)] \leq \sum_{i \neq j} \frac{2 \min\{2dT(\alpha_{ij}(\Delta_{ij}) + \beta_{ij}(\Delta_{ij})), 1\}}{n(n-1)}.$$

The right-hand side in the above inequality can be bounded further by utilizing that  $\Delta_{ij} \geq |i-j|\Delta_*$ . Together with Markov's inequality, this yields the following bound on  $P(R(\theta_T) \geq \epsilon)$ .

**Theorem 1** (Upper bound on the ordering error).

Let  $\alpha_* := \max_{i \neq j} \alpha_{ij}(\Delta_*)$  and  $\beta_* := \max_{i \neq j} \beta_{ij}(\Delta_*)$ , with  $\alpha, \beta$  from Lemma 1. Then, for  $\alpha_*, \beta_* \leq \frac{1}{4dT}$  and any  $\epsilon \in (0, 1)$ , the ordering error  $R(\theta_T)$  satisfies

$$P(R(\theta_T) \geq \epsilon) \leq \frac{4dT}{\epsilon n} \left( (\alpha_*^{-1} - 1)^{-1} + (\beta_*^{-1} - 1)^{-1} \right) \approx \frac{4dT}{\epsilon n} (\alpha_* + \beta_*),$$

where  $\alpha_*$  and  $\beta_*$  decay exponentially with  $T$ .

Theorem 1 suggests that the probability of  $R(\theta_T) \geq \epsilon$  decays exponentially with a rate that depends on the quantities  $\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T) \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}$  and  $\max_{i,j} \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}^2$ . Both quantities are random variables that depend on the particular sampling strategy that yields  $\mathbf{H}_T$ . Focusing on the leading term,  $\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T) \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}$ , Theorem 1 suggests that an active learner should gather data to minimize this quantity and obtain the smallest possible bound. The factor  $\|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}^2$  is the weighted norm of  $z_{ij}$  w.r.t. the inverse of the observed Fisher information (cf. Kirsch and Gal (2022)). It controls the shape of the confidence ellipsoid around  $\theta_T$  and the width of the confidence interval around  $\theta_T^\top z_{ij}$ . The leading term in Theorem 1 re-scales this quantity with aleatoric noise under the MLE estimate  $\theta_T$ . This suggests that higher epistemic (model) certainty is needed in directions with high aleatoric uncertainty—where item similarity increases noise in comparisons.

In Appendix C.3, we comment on i) generalizations to regularized preference models, ii) applications to generalized linear models with other link functions, iii) lower bounds on the ordering error, and iv) an algorithm-specific upper bound.

## 5 Greedy uncertainty reduction for ordering (GURO)

We present an active preference learning algorithm based on greedy minimization of the bound in Theorem 1, called GURO. We begin with fully contextual preference models of the form  $\sigma(\theta^\top z_{ij})$  and return in Section 5.1 to parameterization variants to reduce the effects of model misspecification.

The main component of the bound in Theorem 1 to be controlled by an active learner is the term

$$\max_{i,j \in \mathcal{I}} \dot{\sigma}(z_{ij}^\top \theta_T) \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)}, \quad (4)$$

which represents the highest uncertainty in the comparison of any items  $i, j \in \mathcal{I}$  under the model  $\theta_T$ . A smaller value of (4) yields a smaller bound and a stronger guarantee. Recall that, for any  $t = 1, \dots, T$ ,  $\theta_t$  is the MLE estimate of the ground-truth parameter  $\theta_*$  with respect to the observed history  $D_t$ . Both factors in (4) are determined by the sampling strategy that yielded the item pairs  $(i_t, j_t)$  in  $D_T$  and, therefore,  $\mathbf{H}_T$  and  $\theta_T$  (the results of comparisons  $c_{ij}$  are outside the control of the algorithm, but  $z_{ij}$  are known).

Direct minimization of (4), for a subset  $\mathcal{I}_D$ , is not feasible without access to comparisons  $c_{ij}$  and their likelihood under  $\theta_T$ . Instead, we adopt a greedy, alternating approach: In each round, a)

---

**Algorithm 1** Greedy Uncertainty Reduction for Ordering (GURO), [BayesGURO]

---

**Require:** Training items  $\mathcal{I}_D$ , attributes  $\mathbf{X} = \{x_i\}_{i \in \mathcal{I}_d}$

- 1: Initialize  $\theta_0$
  - 2: **for**  $t = 1, \dots, T$  **do**
  - 3:     Draw  $(i_t, j_t)$  based on  $\theta_t$  according to (5) [or (7)]
  - 4:     Observe  $c_t$  from noisy comparison (annotator)
  - 5:      $D_t = D_{t-1} \cup \{i_t, j_t, c_t\}$
  - 6:      $\theta_t = \text{MLE}(D_t)$  according to (3) [or  $\theta_t = \text{MAP}(D_t)$  as in (11) in the Appendix]
  - 7: **end for**
  - 8: Return  $h_T$
- 

single pair is sampled for comparison by maximizing (4) under the current model estimate, and b)  $\theta_t$  is recomputed based on  $D_t$ . Specifically, at  $t = 1, \dots, T$ , we sample,

$$i_t, j_t = \arg \max_{i, j \in \mathcal{I}_D, i \neq j} \dot{\sigma}(z_{ij}^\top \theta_{t-1}) \|z_{ij}\|_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})}. \quad (5)$$

We refer to this sampling criterion as Greedy Uncertainty Reduction for Ordering (GURO), since it reduces the uncertainty of  $\theta_t$  in the direction of  $z_{ij}$ . To see this, consider the change of  $\mathbf{H}_t(\theta_t)$  after a single play of  $i_t, j_t$ . The Sherman-Morrison formula (Sherman and Morrison, 1950) yields,

$$\mathbf{H}_t^{-1}(\theta_{t-1}) = \mathbf{H}_{t-1}^{-1}(\theta_{t-1}) - \dot{\sigma}(z_t^\top \theta_{t-1}) \frac{\mathbf{H}_{t-1}^{-1}(\theta_{t-1}) z_t z_t^\top \mathbf{H}_{t-1}^{-1}(\theta_{t-1})}{1 + \dot{\sigma}(z_t^\top \theta_{t-1}) \|z_t\|_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})}^2}, \quad (6)$$

where  $z_t := z_{i_t j_t}$ . With  $\xi$  as the second term in (6), it holds for all  $i < j \in \mathcal{I}$ , with  $\mathbf{H}_{t-1} = \mathbf{H}_{t-1}(\theta_{t-1})$ , that  $\|z_{ij}\|_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})}^2 = \|z_{ij}\|_{\mathbf{H}_{t-1}^{-1}}^2 - \|z_{ij}\|_{\xi}^2 \leq \|z_{ij}\|_{\mathbf{H}_{t-1}^{-1}}^2$ . The inequality is strict for the pair  $i_t, j_t$  in (5). As  $\theta_t$  converges to  $\theta_*$ , this pair becomes representative of the maximizer of (4) provided there is no major systematic discrepancy between  $\mathcal{I}_D$  and  $\mathcal{I}$ .

Surprisingly, GURO can also be justified from a Bayesian analysis. Consider a Bayesian model of the parameter  $\theta$  with  $p(\theta)$  the prior belief and  $p(\theta | D_t)$  the posterior after observing the preference feedback in  $D_t$ . A natural active learning strategy is to sample items  $i_t, j_t$  for which the model preference is highly uncertain under the posterior distribution,

$$i_t, j_t = \arg \max_{i, j \in \mathcal{I}_D, i < j} \hat{\mathbb{V}}_{\theta | D_{t-1}}[\sigma(\theta^\top z_{ij})], \quad (7)$$

where  $\hat{\mathbb{V}}_{\theta | D_{t-1}}[\sigma(\theta^\top z_{ij})]$  is a finite-sample estimate of the variance in predictions, computed by sampling from the posterior. In Appendix B.3, we show that the first-order Taylor expansion of the true variance is equal to the GURO criterion. Hence, we refer to sampling according to (7) as BayesGURO. Unlike GURO, BayesGURO can incorporate prior knowledge through  $p(\theta)$  and benefits from controlled stochasticity through the empirical estimate  $\hat{\mathbb{V}}$ , which makes it appropriate for batched algorithms—a deterministic criterion would construct batches of a single item pair. Both GURO and BayesGURO are presented in Algorithm 1.

**Computational Complexity:** Running the algorithms requires  $O(n^2)$  operations each iteration to evaluate the sampling criteria (Equation 5 or 7) on all possible pairs, a problem shared by many active preference learning algorithms (Qian et al., 2015; Canal et al., 2019; Houlsby et al., 2011). A way of mitigating this computational complexity is to, at each time step, sample a fixed number of comparisons and only evaluate on these, similar to the approach taken in Canal et al. (2019). When only looking at a sample of  $m \ll n^2$  pairs, the complexity is reduced to  $O(m)$ . While making  $m$  too small can hurt the sample complexity, we describe in Appendix E how we implemented this sub-sampling strategy to speed up computations in one of our experiments and observed no noticeable change in performance. Lastly, we want to highlight that in many realistic scenarios, the computational burden pales in comparison to the time it takes to query an annotator.

### 5.1 Preference models for in- and out-of-sample ordering

Our default preference model  $h(i, j) = \mathbb{1}[f(i, j) > 0]$  is based on a *fully contextual* scoring function

$$f_\theta(x_i, x_j) = \theta^\top (x_i - x_j), \quad (8)$$

fit with a logistic likelihood  $\sigma(f(i, j)) \approx p(C_{ij} = 1)$ . The model’s strength is that the variance in its estimates grows with  $d$ , but not with  $n = |\mathcal{I}|$ , often resulting in quicker convergence than non-contextual methods for moderate dimension  $d$  (see, e.g., Figure 2c). The fully contextual model also generalizes to unseen items as long as the attributes for  $\mathcal{I}_D$  span attributes observed for  $\mathcal{I}$ .

The limitations of a fully contextual model are model misspecification (error due to the functional form), and noise (error due to  $C$  not being fully determined by  $X$ ). The former can be mitigated by applying the linear model to a representation function  $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ ,  $f_\theta(x_i, x_j) = \theta^\top (\phi(x_i) - \phi(x_j))$ . A good representation  $\phi$ , e.g., from a foundation model, can mitigate misspecification and admit different input modalities. As demonstrated in Figure 5 in the Appendix, even a representation pre-trained for a different task can perform much better than a random initialization.<sup>2</sup>

Noise due to insufficiencies in  $X$  cannot be mitigated by a representation  $\phi(x)$ ; If annotators consistently compare items based on features  $U$  not included in  $X$ , no function  $h(X_i, X_j)$  can perfectly order the items. However, for in-sample ordering of  $\mathcal{I}_D$ , adding per-item parameters  $\zeta_i \in \mathbb{R}$  to the scoring function, one for each item  $i \in \mathcal{I}_D$ , can mitigate both misspecification and noise,

$$f_{\theta, \zeta}(x_i, x_j) = \theta^\top (\phi(x_i) - \phi(x_j)) + (\zeta_i - \zeta_j). \quad (9)$$

We call this a *hybrid* model and apply it in “GURO Hybrid” and baselines in experiments. The term  $\zeta_i - \zeta_j$  can correct the residual of the fully contextual model, which is small if a) the context captures the most relevant information about the ordering, and b) the functional form  $\theta^\top \phi(x_i)$  is nearly well-specified. Using  $\zeta_i - \zeta_j$  alone is sufficient in-sample, but has high variance (the dimension is  $n$  instead of  $d$ ) and poor generalization ( $\zeta_i$  are unknown for items  $i \notin \mathcal{I}_D$ ). In practice, we use L2 regularization to prevent the model from learning an arbitrary  $\theta$  by using the full expressivity of  $\zeta_i$  (see Appendix E for details). Empirically, our hybrid models exhibit the best of both worlds: When  $\phi$  is poor, the model recovers and competes with non-contextual models (Figure 5); when  $\phi$  is good, convergence matches fully contextual models (Figure 2).

## 6 Experiments

We evaluate GURO (Algorithm 1) and GURO Hybrid (see Section 5.1) in four image ordering tasks, one with logistic (synthetic) preference feedback, and three tasks based on real-world feedback from human annotators<sup>3</sup>. We provide a synthetic experiment in Appendix E.2 that includes empirical estimates of the bound in Theorem 1. The experiments include five diverse baseline algorithms, described next. BALD (Houlsby et al., 2011) is *a priori* the strongest baseline since it is a contextual active learning algorithm, unlike the others. Its selection criterion greedily maximizes the decrease in posterior entropy, which amounts to reducing the epistemic uncertainty and includes a term to downplay the influence of aleatoric uncertainty. This is not always beneficial, as suggested by our analysis in Section 4, since learners may require several comparisons of high-uncertainty pairs to get the order right. CoLSTIM (Bengs et al., 2022) is a contextual bandit algorithm, developed for regret minimization and is not expected to perform well here. It is included to illustrate the mismatch between regret minimization and our setting.

TrueSkill (Herbrich et al., 2006; Graepel, 2012) is a non-contextual skill-rating system that models the score of each item as a Gaussian distribution, disregarding item attributes, and has been adopted in various works to score items based on subjective pairwise comparisons (Larkin et al., 2022; Naik et al., 2014; Sartori et al., 2015). We use the sampling rule from Hees et al. (2016), designed for ordering. Finally, we include Uniform sampling, and to illustrate the importance of accounting for aleatoric uncertainty, we use a version of GURO called NormMin that ignores the  $\dot{\sigma}(z_{ij}^\top \theta_t)$  term and plays the pair maximizing  $\|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)}$ , i.e., it minimizes the *second-order* term in Lemma 1. NormMin corresponds to the selection criterion in the concurrent work Das et al. (2024), adapted to our problem of finding the correct ordering. We refer the reader to Appendix E.2 for a detailed comparison where NormMin performs significantly worse than Uniform on certain problem instances, and Appendix E for details regarding the implementation and the choice of hyperparameters for GURO, BayesGURO, and baselines.

<sup>2</sup>It is feasible to update representations during exploration (Xu et al., 2022; Singh and Chakraborty, 2021), but we do not consider that here.

<sup>3</sup>Our code is available at: <https://github.com/Healthy-AI/GURO>

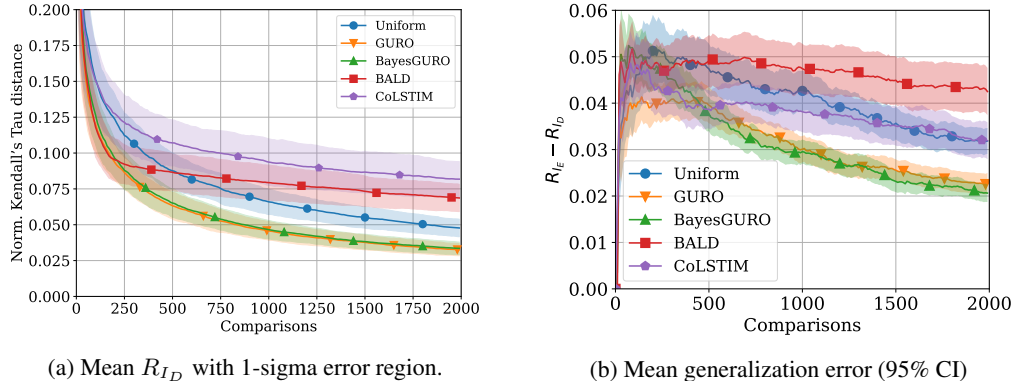


Figure 1: **X-RayAge**. Performance of active sampling strategies when comparisons are simulated using a logistic model according to (2). In-sample Kendall’s Tau distance  $R_{I_D}$  on 200 images (left) and generalization error  $R_{I_E} - R_{I_D}$  for models trained on 150 images and evaluated on 150 images from a different distribution (right). All results are averaged over 100 different random seeds.

## 6.1 Ordering X-ray images under the logistic model

Our first task (X-RayAge) is to order X-ray images based on perceived age (Ieki et al., 2022) where the preference feedback follows a (well-specified) logistic model. We base this experiment on the data from the Kaggle competition “X-ray Age Prediction Challenge” (Felipe Kitamura, 2023) which contains more than 10 000 de-identified chest X-rays, along with the person’s true age. Features were extracted using the 121-layer DenseNet in the TorchXrayVision package (Cohen et al., 2022) followed by PCA projection, resulting in 35 features. A ridge regression model,  $\theta_*$ , was fit to the true age ( $R^2 \approx 0.67$ ). During active learning, feedback is drawn from  $p(C_{ij} = 1) = \sigma(\theta_*^\top z_{i,j} \cdot \lambda)$ , where  $\lambda$  (set to 0.1) controls the noise level. We only include the fully contextual models here since they are well-specified by design, meaning  $\mathcal{I}$  can be ordered using only contextual features.

In the first setting, we sub-sample 200 X-ray images uniformly at random from the full set. A ground-truth ordering of these elements is derived using the learned linear model. Figure 1a shows the ordering error over 2 000 iterations. GURO and BayesGURO perform similarly, both better than the baselines. BALD starts off converging about as fast as GURO, but plateaus, most likely as a result of actively avoiding comparisons with high aleatoric uncertainty—pairs where annotators disagree in their preferences. The poor performance of CoLSTIM highlights the discrepancy between regret minimization and recovering a complete ordering.

In the second setting, we evaluate how well the algorithms generalize to new items. First, we sample 300 X-ray images from the full dataset. Next, we split these into two sets, with one ( $I_D$ ) containing the youngest 50% and the other ( $I_E$ ) the oldest 50%. The algorithms were then trained to order the list containing the younger subjects, but were simultaneously evaluated on how well they could sort the list containing the older subjects. The continuously measured difference in ordering error evaluated on  $I_E$  and  $I_D$  are presented in Figure 1b. While all algorithms are worse at ordering items in  $I_E$ , GURO and BayesGURO achieve the lowest average difference. Together with Figure 1a, this means that our proposed algorithms achieved the best in-sample and out-of-sample orderings. For completeness, the in-sample performance of algorithms in the generalization experiment in Figure 1b are included in Appendix E.2.

## 6.2 Ordering items with human preference data

Next, we evaluate our algorithm on three publicly available datasets to study the algorithms’ performance when preference feedback comes from human annotators (see Table 1 for an overview, detailed information of datasets in Appendix E.1). The datasets are IMDB-WIKI-SbS (Pavlichenko and Ustalov, 2021), where annotators have stated which of two people appear older, ImageClarity (Zhang et al., 2016), where modified versions of the same image have been compared according to the level of distortion, as well as the extended WiscAds dataset (Carlson and Montgomery, 2017), where labels correspond to which political advertisement is perceived as more negative toward an



Table 1: Datasets with preference feedback from annotators. Pretrained models are ResNet34 (He et al., 2016), all-mpnet-base-v2 (Reimers and Gurevych, 2019), and FaceNet (Schroff et al., 2015).

Dataset	$n$	$d$	#comparisons	Data type	Embedding Model
<b>ImageClarity</b>	100	63	27 730	Image	ResNet34 (Imagenet)
<b>WiscAds</b>	935	162	9 528	Text	all-mpnet-base-v2
<b>IMDB-WIKI-SbS</b>	6072	75	110 349	Image	FaceNet (CASIA-Webface)

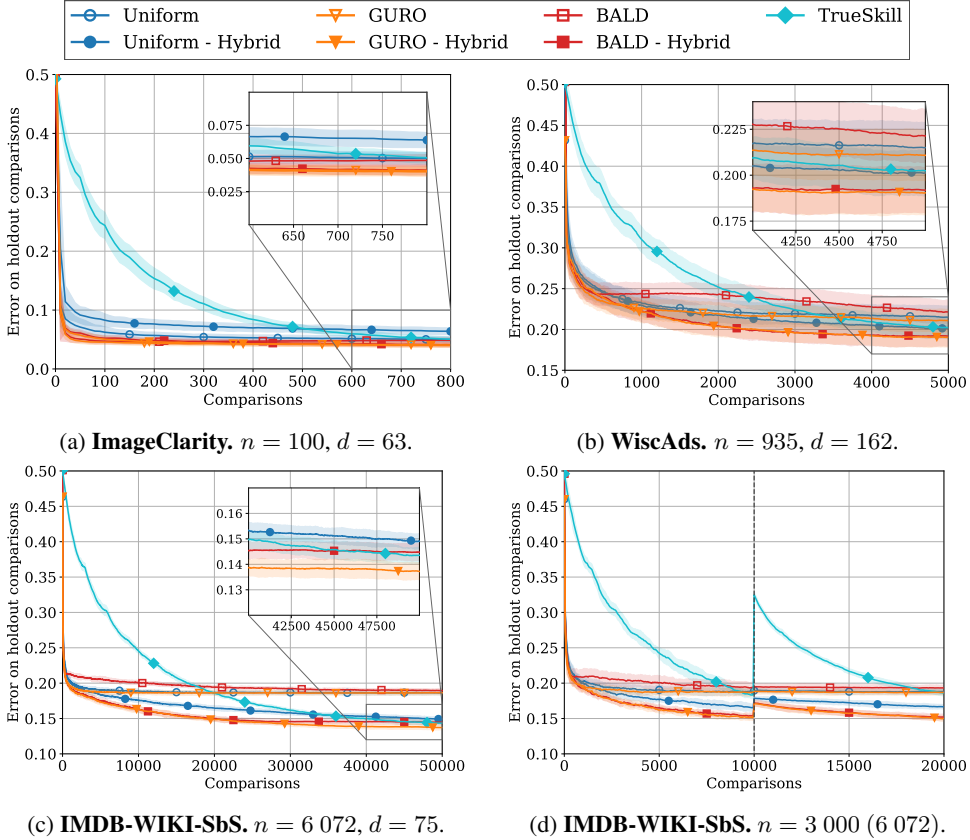


Figure 2: The empirical error  $\hat{R}_{D'}(h)$  on a holdout comparison set  $D'$  when comparisons are made by human annotators. The plots are averaged over 100 (a,b) or 10 (c,d) seeds, and the shaded area represents one standard deviation above and below the mean. For every seed, 10% of comparisons were used for the holdout set. In (d) we initially order a list  $\mathcal{I}_D$  of 3 000 images. After 10 000 comparisons the remaining 3 072 images,  $\mathcal{I} \setminus \mathcal{I}_D$ , are added.

opponent. In all datasets, pairs of items were sampled uniformly for annotation. For each experiment, we construct a feature vector  $\phi(x_i) \in \mathbb{R}^d$  for all  $n$  items using a pre-trained embedding model followed by PCA, applied to reduce computational complexity. We restrict algorithms to only query pairs for which an annotation exists and remove the annotation from the pool once queried. In cases where multiple annotations exist for the same pair, the feedback is chosen randomly among these.

The images in the ImageClarity dataset have been constructed to have an objective ground truth ordering but this is not the case for WiscAds or IMDB-WIKI-SbS. As the ground-truth ordering is generally unknown also in real-world applications, we evaluate methods by the error on a held-out set of comparisons  $D'$ ,  $\hat{R}_{D'}(h) = \frac{1}{|D'|} \sum_{(i,j,c) \in D'} \mathbb{1}[h(i,j) \neq c]$ . This serves as an empirical analog of Kendall’s Tau distance and a minimizer of  $\hat{R}_{D'}(h)$  will minimize  $R_{\mathcal{I}}(h)$  for sufficiently large  $D'$ , but will not converge toward 0 since there is inherent noise in annotations. This metric makes no assumptions on the ground truth ordering unlike the alternative approach of fitting an ordering to all available comparisons, see e.g., Maystre and Grossglauser (2017). In Appendix E.2,

we show results for the latter that highlight the limitations of estimating a “ground-truth” ordering, as well as the similar results when measuring the distance to the objective ground-truth ordering of the ImageClarity dataset. The longest trajectory (single seed) for any algorithm took less than 35hrs to complete on one core of an Intel Xeon Gold 6130 CPU and required at most 10 GB of memory.

In all experiments, we compare fully contextual (8) and hybrid (9) versions of GURO, BALD, and Uniform, as well as TrueSkill. The results of each experiment can be seen in Figure 2. Figure 2a shows that the ImageClarity dataset is the easiest to order using contextual (non-hybrid) features. This is expected, as features relevant to the level of distortion are low-level. In this case, the choice of adaptive strategy has a modest impact on the ordering error. Figures 2b and 2c highlight the differences between modeling strategies. The fully contextual algorithms initially improve rapidly, achieving a rough ordering of the items, before plateauing and not making any real improvements. This indicates that the features are informative enough to roughly order the list, but insufficient for retrieving a more granular ordering. The non-contextual TrueSkill converges at a much slower pace but keeps improving steadily throughout. Perhaps most interesting are the hybrid algorithms, which seemingly reap the benefits of both methods, improving as quickly as the contextual methods, but avoiding the plateau. In fact, in Figure 5 in the Appendix we show that the hybrid models perform comparably to TrueSkill even when features are completely uninformative.

The limitations of BALD are most noticeable in the fully contextual case, where it plateaus at a higher error compared to GURO and Uniform. This is however not as prominent when we use BALD in conjunction with our hybrid model, likely a result of the increased dimensionality of the model causing BALD Hybrid to attribute more of the observed errors to model uncertainty. While this initially causes the algorithm to avoid fewer comparisons that are subject to aleatoric uncertainty, the final iterations in Figure 2c suggest that BALD Hybrid can still run into this issue given enough samples. In all experiments, GURO and GURO Hybrid perform better than or similar to our baselines, never worse. Additionally, Figures 2b and 2c showcase how our hybrid model can increase performance when used with existing sampling strategies, such as BALD or Uniform.

The final experiment, visible in Figure 2d, is a few-shot scenario where after some time, additional images are added to the pool of items. IMDB-WIKI-SbS was used as it contained the highest number of both images and comparisons. The initial pool consists of 3 000 images sampled from the dataset. After 10 000 steps, the remaining 3 072 images were added to the pool. The results again emphasize the differences between our three types of models; the increase in error of the fully contextual model is very slight, likely a result of added samples being drawn from the same distribution. For TrueSkill, the error increases drastically as a result of the algorithm not having seen these items before and having no way of generalizing the results of previous comparisons to them. Lastly, the hybrid algorithms seem to be moderately affected. The error increases as the model has not yet tuned any of the added per-item parameters, but the extent is much smaller than for TrueSkill as the model can provide a rough ranking of the out-of-sample elements using the contextual features.

## 7 Conclusion

We have demonstrated the benefits of utilizing contextual features in active preference learning to efficiently order a list of items. Empirically, this leads to quicker convergence, compared to non-contextual methods, and allows algorithms to generalize out-of-sample. We derived an upper bound on the ordering error and used it to design an active sampling strategy that outperforms or matches baselines on realistic image and text ordering tasks. Both theoretical and empirical results highlight the benefit of accounting for noise in comparisons when learning from human annotators.

The optimality of our sampling strategy remains an open question. A future direction is to derive a lower bound on the ordering error, and prove an—ideally matching—algorithm-specific upper bound. However, constructing upper bounds for related fixed-budget tasks is an open problem (Qin, 2022). Moreover, motivated by the annotation setting, our focus has been on reducing sample complexity and we leave it to future work to explore potential linear approximations of the sampling criteria and other trade-offs between sample complexity and computational complexity. Further, our approach can potentially be improved by performing representation learning throughout the learning process. Finally, our experiments are constrained to a limited amount of already-collected (offline) human preference data, causing different algorithms to select disproportionately similar comparisons. Future work should evaluate the strategies in an online setting.

## Acknowledgements

FDJ and HB are supported by Swedish Research Council Grant 2022-04748. FDJ is also supported in part by the Wallenberg AI, Autonomous Systems and Software Program founded by the Knut and ALice Wallenberg Foundation. EC and DD are supported by Chalmers AI Research Centre (CHAIR). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

## References

- Nir Ailon. Active learning ranking from pairwise preferences with almost optimal query complexity. *Advances in Neural Information Processing Systems*, 24, 2011.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.
- Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *The Journal of Machine Learning Research*, 22 (1):278–385, 2021.
- Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pages 1764–1786. PMLR, 2022.
- Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.
- Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Klaus Brinker. Active learning of label ranking functions. In *Proceedings of the twenty-first international conference on Machine learning*, page 17, 2004.
- Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hultender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.
- Ludwig M. Busse, Morteza Haghiri Chehreghani, and Joachim M. Buhmann. The information content in sorting algorithms. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 2746–2750, 2012. doi: 10.1109/ISIT.2012.6284021.
- Gregory Canal, Andy Massimino, Mark Davenport, and Christopher Rozell. Active embedding search via noisy paired comparisons. In *International Conference on Machine Learning*, pages 902–911. PMLR, 2019.
- David Carlson and Jacob M. Montgomery. A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts. *American Political Science Review*, 111(4):835–843, November 2017. ISSN 0003-0554, 1537-5943.
- Kani Chen, Inchi Hu, and Zhiliang Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4): 1155–1163, 1999.

- Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, pages 193–202, New York, NY, USA, February 2013. Association for Computing Machinery. ISBN 978-1-4503-1869-3. doi: 10.1145/2433396.2433420.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144, 2005.
- Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarnera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRyVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022.
- Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient rlhf via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024.
- Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902 – 1933, 2004. doi: 10.1214/009117904000000397.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. *arXiv preprint arXiv:2310.00968*, 2023.
- Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.
- Boli Fang. Fixed-budget pure exploration in multinomial logit bandits. In *International Joint Conference on Artificial Intelligence*, 2022.
- Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.
- Paulo Kuriki Felipe Kitamura, Lilian Mallagoli. Spr x-ray age prediction challenge, 2023.
- Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.
- Johannes Fürnkranz and Eyke Hüllermeier. Pairwise preference learning and ranking. In *European conference on machine learning*, pages 145–156. Springer, 2003.
- Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.
- Anne-Marie George and Christos Dimitrakakis. Eliciting kemeny rankings, 2023.
- Thore Graepel. Score-based bayesian skill learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD-12)*, January 2012.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

- Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pages 1057–1066. PMLR, 2018.
- Jörn Hees, Benjamin Adrian, Ralf Biedert, Thomas Roth-Berghofer, and Andreas Dengel. Tssort: Probabilistic noise resistant sorting. *arXiv preprint arXiv:1606.05289*, 2016.
- Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning, December 2011. arXiv:1112.5745 [cs, stat].
- Hiroataka Ieki, Kaoru Ito, Mike Saji, Rei Kawakami, Yuji Nagatomo, Kaori Takada, Toshiya Kariyasu, Haruhiko Machida, Satoshi Koyama, Hiroki Yoshida, Ryo Kurosawa, Hiroshi Matsunaga, Kazuo Miyazawa, Kouichi Ozaki, Yoshihiro Onouchi, Susumu Katsushika, Ryo Mat-suoka, Hiroki Shinohara, Toshihiro Yamaguchi, Satoshi Kodera, Yasutomi Higashikuni, Kat-suhito Fujiu, Hiroshi Akazawa, Nobuo Iguchi, Mitsuaki Isobe, Tsutomu Yoshikawa, and Is-sei Komuro. Deep learning-based age estimation from chest X-rays indicates cardiovascular prognosis. *Communications Medicine*, 2(1):1–12, December 2022. ISSN 2730-664X. doi: 10.1038/s43856-022-00220-6. Number: 1 Publisher: Nature Publishing Group.
- Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.
- Ikbeom Jang, Garrison Danley, Ken Chang, and Jayashree Kalpathy-Cramer. Decreasing annotation burden of pairwise comparisons with human-in-the-loop sorting: Application in medical image artifact rating. *arXiv preprint arXiv:2202.04823*, 2022.
- Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5148–5157. PMLR, 18–24 Jul 2021.
- Maurice George Kendall. *Rank correlation methods*. Griffin, 1948.
- Andreas Kirsch and Yarin Gal. Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. Expert Certification.
- Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.
- Andrew Larkin, Ajay Krishna, Lizhong Chen, Ofer Amram, Ally R. Avery, Glen E. Duncan, and Perry Hystad. Measuring and modelling perceptions of the built environment for epidemiological research using crowd-sourcing and image-based deep learning models. *Journal of Exposure Science & Environmental Epidemiology*, 32(6):892–899, November 2022. ISSN 1559-064X. doi: 10.1038/s41370-022-00489-8. Number: 6 Publisher: Nature Publishing Group.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.
- Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.
- Mats Lidén, Antoine Spahr, Ola Hjelmgren, Simone Bendazzoli, Josefin Sundh, Magnus Sköld, Göran Bergström, Chunliang Wang, and Per Thunberg. Machine learning slice-wise whole-lung CT emphysema score correlates with airway obstruction. *European Radiology*, 34(1):39–49, January 2024. ISSN 1432-1084. doi: 10.1007/s00330-023-09985-3.

- Suiyi Ling, Jing Li, Anne Flore Perrin, Zhi Li, Lukáš Krasula, and Patrick Le Callet. Strategy for boosting pair comparison and improving quality assessment accuracy. *arXiv preprint arXiv:2010.00370*, 2020.
- Bo Long, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng. Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274, 2010.
- Hao Lou, Tao Jin, Yue Wu, Pan Xu, Quanquan Gu, and Farzad Farnoud. Active ranking without strong stochastic transitivity. *Advances in neural information processing systems*, 35:297–309, 2022.
- Andrew K Massimino and Mark A Davenport. As you like it: Localization via paired comparisons. *Journal of Machine Learning Research*, 22(186):1–39, 2021.
- Lucas Maystre and Matthias Grossglauser. Just sort it! a simple and effective approach to active preference learning. In *International Conference on Machine Learning*, pages 2344–2353. PMLR, 2017.
- Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267*, 2023.
- Tom Minka, Ryan Clevon, and Yordan Zaykov. Trueskill 2: An improved bayesian skill rating system. *Technical Report*, 2018.
- Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human feedback, 2024.
- Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 779–785, 2014.
- IFD Oliveira, S Zehavi, and O Davidov. Stochastic transitivity: Axioms and models. *Journal of Mathematical Psychology*, 85:25–35, 2018.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
- Nikita Pavlichenko and Dmitry Ustalov. Imdb-wiki-sbs: An evaluation dataset for crowdsourced pairwise comparisons. *CoRR*, abs/2110.14990, 2021.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Andrew S. Phelps, David M. Naeger, Jesse L. Courtier, Jack W. Lambert, Peter A. Marcovici, Javier E. Villanueva-Meyer, and John D. MacKenzie. Pairwise comparison versus Likert scale for biomedical image assessment. *AJR. American journal of roentgenology*, 204(1):8–14, 2015. ISSN 0361-803X. doi: 10.2214/ajr.14.13022.
- Li Qian, Jinyang Gao, and HV Jagadish. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment*, 8(11):1322–1333, 2015.
- Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, pages 5650–5654. PMLR, 2022.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.

- Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.
- Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pages 968–994. PMLR, 2022.
- Andreza Sartori, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM Transactions on Interactive Intelligent Systems (TiS)*, 5(2):1–27, 2015.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. doi: 10.1109/CVPR.2015.7298682. arXiv:1503.03832 [cs].
- Jack Sherman and Winifred J. Morrison. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1):124 – 127, 1950. doi: 10.1214/aoms/1177729893.
- Rodrigo M Silva, Marcos A Gonçalves, and Adriano Veloso. A two-stage active learning method for learning to rank. *Journal of the Association for Information Science and Technology*, 65(1): 109–128, 2014.
- Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1794–1834. PMLR, 07–10 Jul 2017. URL <https://proceedings.mlr.press/v65/simchowitz17a.html>.
- Ankita Singh and Shayok Chakraborty. Deep active learning with relative label feedback: An application to facial age estimation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.
- Hanna Tärnåsen and Herman Bergström. Rank based annotation system for supervised learning in medical imaging. Master’s thesis, Chalmers University of Technology, 2023.
- Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96):6227, 2024. ISSN 2475-9066. doi: 10.21105/joss.06227.
- Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment, 2023a.
- Yue Wu, Tao Jin, Hao Lou, Farzad Farnoud, and Quanquan Gu. Borda regret minimization for generalized linear dueling bandits. *arXiv preprint arXiv:2303.08816*, 2023b.
- Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration. In *International Conference on Learning Representations*, 2022.
- Xinyi Yan, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 567–577, 2022.
- Miao Yang, Ge Yin, Yixiang Du, and Zhiqiang Wei. Pair comparison based progressive subjective quality ranking for underwater images. *Signal Processing: Image Communication*, 99:116444, 11 2021. ISSN 09235965. doi: 10.1016/j.image.2021.116444.
- Georgios N. Yannakakis and Héctor P. Martínez. Ratings are overrated! *Frontiers in ICT*, 2, July 2015. doi: 10.3389/fict.2015.00013.

- Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208, 2009.
- Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.
- Xiaohang Zhang, Guoliang Li, and Jianhua Feng. Crowdsourced top-k algorithms: an experimental evaluation. *Proceedings of the VLDB Endowment*, 9(8):612–623, April 2016. ISSN 2150-8097. doi: 10.14778/2921558.2921559. URL <https://dl.acm.org/doi/10.14778/2921558.2921559>.
- Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 23–29 Jul 2023.



## A Notation

Table 2: Notation

$\mathcal{I}$	Collection of items $\mathcal{I} = \{1, \dots, n\}$
$n$	Number of items
$d$	Dimension of item attributes
$x_i \in \mathbb{R}^d$	Context attributes for item $i \in \mathcal{I}$
$z_{ij} \in \mathbb{R}^d$	$z_{ij} = x_i - x_j$ for $i, j \in \mathcal{I}$
$y_i$	Score for item $i \in \mathcal{I}$
$c_t \in \{0, 1\}$	The outcome of the comparison at time $t$ , 1 if $i_t$ was preferred to $j_t$
$D_t$	$D_t = ((i_1, j_1, c_1), \dots, (i_t, j_t, c_t))$
$\theta \in \mathbb{R}^d$	Model parameter
$\theta_* \in \mathbb{R}^d$	Model parameter of the environment
$\theta_t \in \mathbb{R}^d$	Estimated parameter at time $t$
$\sigma(\cdot)$	Sigmoid (logistic) function
$\dot{\sigma}(\cdot)$	derivative of $\sigma(\cdot)$
$\mathbf{H}_t(\theta)$	Hessian of the negative log-likelihood $\mathbf{H}_t(\theta) := \sum_{s=1}^t \dot{\sigma}(z_s^\top \theta) z_s z_s^\top$
$\tilde{\mathbf{H}}_t(\theta)$	Hessian normalized by number of plays $\tilde{\mathbf{H}}_t(\theta) := \frac{1}{t} \mathbf{H}_t(\theta)$
$\theta_{B,t} \in \mathbb{R}^d$	The MAP estimate of $\theta$ at time $t$
$\mathbf{H}_{B,t}$	The Hessian in the Bayesian setting, adjusted by the prior covariance $\mathbf{H}_{B,0}^{-1}$
$\ z_{ij}\ _{\mathbf{H}_t^{-1}(\theta)}$	$\ z_{ij}\ _{\mathbf{H}_t^{-1}(\theta)} = \sqrt{z_{ij}^\top \mathbf{H}_t^{-1}(\theta) z_{ij}}$
$S_t$	$S_t := \sum_{s=1}^t \epsilon_s z_s$ , where $\epsilon_s = c_s - \sigma(z_s^\top \theta_*)$
$h$	Comparison model (binary output)
$f$	Comparison logit (typically linear), e.g., $f_\theta(i, j) = \theta^\top (x_i - x_j)$
$S$	An upper bound on the norm of $\theta_*$ , that is $\ \theta_*\ _2 \leq S$

## B Algorithms

### B.1 MLE estimator for logistic regression

The log-likelihood  $L_t(\theta)$  of data  $D_t = \{(i_s, j_s, c_s)\}_{s=1}^t$ , with  $z_s = x_{i_s} - x_{j_s}$ , under a logistic regression model with parameters  $\theta$  is defined by

$$L_t(\theta) = \sum_{s=1}^t \left( c_s \log \sigma(\theta^\top z_s) + (1 - c_s)(1 - \sigma(\theta^\top z_s)) \right).$$

The maximum likelihood estimator (MLE) at time  $t$  is the parameters

$$\theta_t = \arg \max_{\theta} L_t(\theta). \quad (10)$$

The regularized estimator with ridge/ $\ell_2$  penalty with parameter  $\lambda$  is

$$\theta_t^R = \arg \min_{\theta} -L_t(\theta) + \lambda \|\theta\|_2^2.$$

### B.2 Bayesian estimator for logistic regression

$\theta_{B,t}$  is the MAP estimate of  $\theta$  at time  $t$  according to the log likelihood

$$\theta_{B,t} = \arg \max_{\theta} \ln p(\theta | D_t), \quad (11)$$

where

$$\begin{aligned} \ln p(\theta | D_t) = & -\frac{1}{2}(\theta - \theta_{B,0})^\top \mathbf{H}_{B,0}^{-1}(\theta - \theta_{B,0}) \\ & + \sum_t c_t \ln(\sigma(z_{i_t, j_t}^\top \theta)) + (1 - c_t) \ln(1 - \sigma(z_{i_t, j_t}^\top \theta)) + \text{const.} \end{aligned}$$

The hessian at time  $t$  is defined as

$$\mathbf{H}_{B,t} = \mathbf{H}_{B,0} + \sum_{(i,j) \in D_t} \dot{\sigma}(z_{i,j}^\top \theta_{B,t}) z_{i,j} z_{i,j}^\top = \mathbf{H}_{B,0} + \mathbf{H}_t.$$

Moreover, if priors  $\theta_{B,0} = \mathbf{0}$  and  $\mathbf{H}_{B,0}^{-1} = I_d$  are used, the log likelihood boils down to:

$$\ln p(\theta | D_t) = -\frac{1}{2} \|\theta\|_2^2 + \sum_t c_t \ln(\sigma(z_{i_t, j_t}^\top \theta)) + (1 - c_t) \ln(1 - \sigma(z_{i_t, j_t}^\top \theta)) + \text{const}$$

which implies that the MAP estimate will be the same as the MLE estimate with ridge regularisation in the frequentist setting. Similarly, the Hessian becomes:

$$\mathbf{H}_{B,t} = \mathbf{H}_t + I_d$$

Sequential updates are also possible in the Bayesian setting by using your current estimates as the new priors. Note that this will give slightly different results, as the calculation of  $\mathbf{H}_{B,t}$  depends on the current estimate of  $\theta_{B,t}$ .

### B.3 Stochastic Bayesian uncertainty reduction (BayesGURO)

We describe BayesGURO, a Bayesian sampling criterion, closely related to GURO. Consider a Bayesian model of the parameter  $\theta$  with  $p(\theta)$  the prior belief and  $p(\theta | D_t)$  the posterior after observing the preference feedback in  $D_t$ . A natural strategy for learning more about the ordering of  $\mathcal{I}$  is to sample items  $i_t, j_t$  based on an estimate of the posterior variance of predictions for their comparison,

$$i_t, j_t = \arg \max_{i,j \in \mathcal{I}_D, i < j} \hat{\mathbb{V}}_{\theta | D_{t-1}}[\sigma(\theta^\top z_{ij})]. \quad (12)$$

Here,  $\hat{\mathbb{V}}_{\theta | D_t}[\sigma(\theta^\top z_{ij})]$  is an estimate of the variance of probabilities  $\sigma(\theta^\top z_{ij})$ , computed from finite samples drawn from the posterior of  $\theta$ . Estimating the variance in this way both i) allows for tractable implementation, and ii) induces controlled stochasticity in the selection of item pairs. This can be useful in batched learning settings so that multiple pairs can be sampled within the same batch. A deterministic criterion would return the same item pair every time until  $\theta$  is updated. We refer to the sampling criterion in (7) as BayesGURO.

For the logistic model considered in Section 4, using Laplace approximation with a Normal prior  $\mathcal{N}(0, \mathbf{H}_{B,0}^{-1})$  on  $\theta$ , the Bayesian criterion in (7) is related to the GURO sampling criterion in (5) through the first-order Taylor expansion of the variance:

$$\mathbb{V}_{\theta|D_t}(\sigma^\top z_{ij}) \approx (\dot{\sigma}(\mathbb{E}_{\theta|D_t}[\theta^\top z_{ij}]))^2 \mathbb{V}_{\theta|D_t}[\theta^\top z_{ij}] = (\dot{\sigma}(\theta_{B,t}^\top z_{ij}) \|z_{ij}\|_{\mathbf{H}_{B,t}^{-1}(\theta_{B,t})})^2,$$

where  $\theta_{B,t}$  is the MAP estimate of  $\theta$  at time  $t$  and  $\mathbf{H}_{B,t}$  is the Hessian adjusted by the prior covariance  $\mathbf{H}_{B,0}^{-1}$  (further described in Appendix B.2). Thus, to a first-order approximation, for a large number of posterior samples, the GURO and BayesGURO active learning criteria are equivalent, save for the influence of the prior. In practice, we find that the Bayesian variant lends itself well to sequential updates of the posterior. The choice of prior  $p(\theta)$ , which could be useful under strong domain knowledge, and the stochasticity of using few posterior samples to approximate  $\mathbb{V}$  make the two criteria distinct.

## B.4 Uniform sampling

The uniform sampling algorithm is given in Algorithm 2. The corresponding Bayesian version replaces line 5 with the MAP estimate.

---

### Algorithm 2 Uniform sampling algorithm

---

**Require:** Training items  $\mathcal{I}_D$ , attributes  $\mathbf{X} = \{x_i\}_{i \in \mathcal{I}_d}$

- 1: **for**  $t = 1, \dots, T$  **do**
- 2:   Sample  $(i_t, j_t)$  uniformly
- 3:   Observe  $c_t$  from noisy comparison (annotator)
- 4:    $D_t = D_{t-1} \cup \{i_t, j_t, c_t\}$
- 5:   Let  $\theta_t = \text{MLE}(D_t)$
- 6: **end for**
- 7: Return  $h_T$

---

## B.5 BALD

---

### Algorithm 3 BALD bandit

---

**Require:** Training items  $\mathcal{I}_D$ , attributes  $\mathbf{X} = \{x_i\}_{i \in \mathcal{I}_d}$

- 1: Initialize  $\theta_{B,0} = \mathbf{0}$ ,  $\mathbf{H}_{B,0} = \lambda^{-1} I$
- 2: **for**  $t = 1, \dots, T$  **do**
- 3:   Draw  $(i_t, j_t) = \arg \max_{i,j} H[y | z_{i,j}, D_{t-1}] - \mathbb{E}_{\theta \sim p(\theta|D_{t-1})}[H[y | z_{i,j}, \theta]]$
- 4:   Observe  $c_t$  from noisy comparison (annotator)
- 5:    $D_t = D_{t-1} \cup \{i_t, j_t, c_t\}$
- 6:   Let  $\theta_t = \text{MAP}(D_t)$
- 7:   Update  $\mathbf{H}_{B,t} \leftarrow \mathbf{H}_{B,0} + \sum_{(i,j) \in D_t} \dot{\sigma}(z_{i,j}^\top \theta_t) z_{i,j} z_{i,j}^\top$
- 8: **end for**
- 9: Return  $h_T$

---

Where the posterior is calculated as in Appendix B.2 and  $H[y | z_{i,j}, D_{t-1}] - \mathbb{E}_{\theta \sim p(\theta|D_{t-1})}[H[y | z_{i,j}, \theta]]$  is approximated as in Appendix B.5.1.

### B.5.1 Deriving the BALD sampling criterion

The BALD criteria formalized using our notation becomes

$$\arg \max_{i,j} H[y | z_{i,j}, D_t] - \mathbb{E}_{\theta \sim p(\theta|D_t)}[H[y | z_{i,j}, \theta]],$$

where  $H$  represents Shannon's entropy

$$h(p) = -p \log_2(p) - (1-p) \log_2(1-p).$$

The first term of the equation becomes

$$H[y | z_{ij}, D_t] = h(\Pr(y | z_{ij}, D_t)) = h\left(\int \Pr(y | z_{ij}, \theta) \Pr(\theta | D_t) d\theta\right).$$

Here  $\Pr(y | z_{ij}, D_t)$  is the predictive distribution for our Bayesian logistic regression model. As covered in Bishop and Nasrabadi (2006, Chapter 4), this expectation cannot be evaluated analytically but can be approximated using the probit function  $\Phi$ :

$$\Pr(y | z_{ij}, D_t) \approx \Phi\left(\frac{\theta_t^\top z_{i,j}}{\sqrt{\lambda^{-2} + \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2}}\right) \approx \sigma\left(\frac{\theta_t^\top z_{i,j}}{\sqrt{1 + \frac{\pi \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2}{8}}}\right).$$

Next, the term  $\mathbb{E}_{\theta \sim p(\theta | D_t)}[H[y | z_{ij}, \theta]]$  must be calculated. The true definition is

$$\mathbb{E}_{\theta \sim p(\theta | D_t)}[H[y | z_{ij}, \theta]] = \int h(\sigma(\theta^\top z_{i,j})) \mathcal{N}(\theta | \theta_t, \mathbf{H}_t^{-1}) d\theta.$$

To make this a one variable integral, let  $X = \theta^\top z_{i,j}$  define a new random variable. Since  $\theta \sim \mathcal{N}(\theta_t, \mathbf{H}_t^{-1})$ , and  $z_{i,j}$  is just a constant vector, we know that  $X$  will follow a univariate normal distribution  $X \sim \mathcal{N}(\theta_t^\top z_{i,j}, \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2)$ . This allows us to rewrite the integral as

$$\int h(\sigma(\theta^\top \mathbf{z})) \mathcal{N}(\theta | \theta_t, \mathbf{H}_t^{-1}) d\theta = \int h(\sigma(x)) \mathcal{N}(\theta_t^\top z_{i,j}, \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2) dx.$$

However, this integral has no closed form solution. Instead we perform the same strategy as in Houlby et al. (2011) and do a Taylor expansion of  $\ln h(\sigma(\theta^\top \mathbf{z}))$ . The third-order Taylor expansion gives us

$$h(\sigma(x)) \approx \exp\left(-\frac{x^2}{8 \ln 2}\right).$$

Inserting this, the term can be approximated as

$$\begin{aligned} \int h(\sigma(x)) \mathcal{N}(x | \theta_t^\top z_{i,j}, \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2) dx &\approx \int \exp\left(-\frac{x^2}{8 \ln 2}\right) \mathcal{N}(x | \theta_t^\top z_{i,j}, \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2) dx \\ &= \frac{C}{\sqrt{\|z_{ij}\|_{\mathbf{H}_t^{-1}}^2 + C^2}} \exp\left(-\frac{(\theta_t^\top z_{i,j})^2}{2(\|z_{ij}\|_{\mathbf{H}_t^{-1}}^2 + C^2)}\right), \end{aligned}$$

where  $C = \sqrt{4 \ln 2}$ . Finally, we arrive at an estimation of the objective function we wish to maximize:

$$\begin{aligned} H[y | z_{ij}, D_t] - \mathbb{E}_{\theta \sim p(\theta | D_t)}[H[y | z_{ij}, \theta]] &\approx h\left(\sigma\left(\frac{\theta_t^\top z_{i,j}}{\sqrt{1 + \frac{\pi}{8} \|z_{ij}\|_{\mathbf{H}_t^{-1}}^2}}\right)\right) \\ &\quad - \frac{C}{\sqrt{\|z_{ij}\|_{\mathbf{H}_t^{-1}}^2 + C^2}} \exp\left(-\frac{(\theta_t^\top z_{i,j})^2}{2(\|z_{ij}\|_{\mathbf{H}_t^{-1}}^2 + C^2)}\right) \end{aligned}$$

## C Proofs of Lemma 1 and Theorem 1

### C.1 Proof of Lemma 1

*Proof.* We now proceed to bound, under Assumptions 1–3 w.r.t. time  $t$ ,

$$P\left(|\sigma(z_{ij}^\top \theta_t) - \sigma(z_{ij}^\top \theta_*)| > \Delta\right).$$

From the self-concordant property of logistic regression we have (Faury et al., 2020)

$$|\sigma(z_{ij}^\top \theta_t) - \sigma(z_{ij}^\top \theta_*)| \leq \dot{\sigma}(z_{ij}^\top \theta_t) |z_{ij}^\top (\theta_t - \theta_*)| + \frac{1}{4} |z_{ij}^\top (\theta_t - \theta_*)|^2. \quad (13)$$

We will prove a high probability bound on the event

$$\dot{\sigma}(z_{ij}^\top \theta_t) |z_{ij}^\top (\theta_t - \theta_*)| + \frac{1}{4} |z_{ij}^\top (\theta_t - \theta_*)|^2 \leq \Delta. \quad (14)$$

Directly trying to bound the LHS in Equation 14 will result in a rather messy expression. Instead, we define the events

$$\begin{aligned} \mathcal{E}_1 &:= \left\{ \dot{\sigma}(z_{ij}^\top \theta_t) |z_{ij}^\top (\theta_t - \theta_*)| \leq \frac{\Delta}{2} \right\} \\ \mathcal{E}_2 &:= \left\{ \frac{1}{4} |z_{ij}^\top (\theta_t - \theta_*)|^2 \leq \frac{\Delta}{2} \right\}. \end{aligned}$$

Clearly  $\mathcal{E}_1 \cup \mathcal{E}_2$  implies the expression in Equation 14. Assume we have bounds on the complement of these events,  $P(\mathcal{E}_1^c) \leq \alpha$  and  $P(\mathcal{E}_2^c) \leq \beta$ . Then

$$\begin{aligned} P\left(|\sigma(z_{ij}^\top \theta_t) - \sigma(z_{ij}^\top \theta_*)| > \Delta\right) &\leq \alpha + \beta + \alpha\beta \\ &\leq 2\alpha + 2\beta. \end{aligned}$$

We now proceed to bound the probability of these complements separately.

**Step 1. Relating  $\theta_t$  to  $\theta_*$ :** The first challenge in our analysis is to relate  $\theta_*$  and  $\theta_t$ . In contrast to linear regression, where we have a closed-form expression for  $\theta_t$ , there is no analytical solution for  $\theta_t$  given a set of observation. However, we know that  $\theta_t$  is the MLE, corresponding to

$$\theta_t = \arg \max_{\theta} L_t(\theta)$$

where

$$L_t(\theta) = \sum_{s=1}^t c_s \log \sigma(z_s^\top \theta) + (1 - c_s) \log (1 - \sigma(z_s^\top \theta)).$$

We have

$$\nabla_{\theta} L_t(\theta) = \sum_{s=1}^t c_s z_s - \underbrace{\sum_{s=1}^t \sigma(z_s^\top \theta) z_s}_{g_t(\theta)}$$

and hence  $g_t(\theta_t) = \sum_{s=1}^t c_s z_s$ .

A standard trick in logistic bandits (Filippi et al., 2010; Faury et al., 2020; Jun et al., 2021) is to relate  $\theta_* - \theta_t$  to  $g_t(\theta_*) - g_t(\theta_t)$ . Especially, the following equality is due to the mean-value theorem (see Filippi et al. (2010))

$$g_t(\theta_*) - g_t(\theta_t) = \mathbf{H}_t(\theta') (\theta_* - \theta_t) \quad (15)$$

where  $\theta'$  is some convex combination of  $\theta_*$ ,  $\theta_t$ . Note that  $\mathbf{H}_t(\theta')$  has full rank.

Using Equation 15 yields

$$\left| z_{ij}^\top (\theta_* - \theta_t) \right| = \left| z_{ij}^\top \mathbf{H}_t^{-1}(\theta') (g_t(\theta_*) - g_t(\theta_t)) \right|$$

Furthermore, since  $g_t(\theta_t) = \sum_{s=1}^t c_s z_s$ , due to  $\nabla_{\theta} L_t(\theta_t) = 0$ , we have

$$g_t(\theta_t) - g_t(\theta_*) = \sum_{s=1}^t \underbrace{\left( c_s - \sigma(z_s^\top \theta_*) \right)}_{\epsilon_s} z_s$$

where  $\epsilon_s$  is a sub-Gaussian random variable with mean 0 and variance  $\nu_s^2 := \dot{\sigma} (z_s^\top \theta_*)$ . We define

$$S_t := \sum_{s=1}^t \epsilon_s z_s.$$

We now have

$$\left| z_{ij}^\top (\theta_* - \theta_t) \right| = \left| z_{ij}^\top \mathbf{H}_t^{-1}(\theta') S_t \right|$$

and Lemma 10 in Faury et al. (2020) states that  $\mathbf{H}_t^{-1}(\theta') \preceq (1 + 2S)\mathbf{H}_t^{-1}(\theta_*)$  where  $\|\theta_*\|_2 \leq S$ . Hence,

$$\left| z_{ij}^\top (\theta_* - \theta_t) \right| \leq (1 + 2S) \left| z_{ij}^\top \mathbf{H}_t^{-1}(\theta_*) S_t \right|$$

## Step 2. Tail bound for vector-valued martingales:

We will now prove an upper bound on the probability that  $|z_{ij}^\top \mathbf{H}_t^{-1}(\theta_*) S_t|$  deviates much from a certain threshold. This step is based on the proof of Lemma 1 in Filippi et al. (2010) which itself is based on a derivation of a concentration inequality in Rusmevichientong and Tsitsiklis (2010). The difference compared to Filippi et al. (2010) is that we work with the Hessian  $\mathbf{H}_t(\theta_*)$  instead of the design matrix for linear regression  $V_t = \sum_s x_s x_s^\top$ . This requires us to construct a slightly different martingale.

Let  $A$  and  $B$  are two random variables such that

$$\mathbb{E} \left[ \exp \left\{ \gamma A - \frac{\gamma^2}{2} B^2 \right\} \right] \leq 1, \forall \gamma \in \mathbb{R} \quad (16)$$

then due to Corollary 2.2 in de la Peña et al. (2004) it holds that  $\forall a \geq \sqrt{2}$  and  $b > 0$

$$P \left( |A| \geq a \sqrt{(B^2 + b) \left( 1 + \frac{1}{2} \log \left( \frac{B^2}{b} + 1 \right) \right)} \right) \leq \exp \left\{ \frac{-a^2}{2} \right\}. \quad (17)$$

Let  $\eta \in \mathbb{R}^d$  and consider the process

$$M_t^\gamma(\theta_*, \eta) := \exp \left\{ \gamma \eta^\top S_t - \gamma^2 \|\eta\|_{\mathbf{H}_t(\theta_*)}^2 \right\}. \quad (18)$$

We will now proceed to prove that  $M_t^\gamma(\theta, \eta)$  is a non-negative super martingale satisfying Equation 16. Note that

$$\gamma \eta^\top S_t - \gamma^2 \|\eta\|_{\mathbf{H}_t(\theta_*)}^2 = \sum_{s=1}^t \underbrace{\left( \gamma \eta^\top z_s \epsilon_s - \dot{\sigma}(\theta^\top z_s) \gamma^2 (\eta^\top z_s)^2 \right)}_{F_s} = \sum_{s=1}^t F_s.$$

Further we use the fact that  $\epsilon_s$  is sub-Gaussian with parameter  $\nu_s$ , i.e.,

$$\mathbb{E} [\exp\{\lambda \epsilon_s\}] \leq \exp\{\nu_s^2 \lambda^2\}, \forall \lambda > 0.$$

Let  $D_{s-1}$  denote the observations up until time  $s$ , then

$$\begin{aligned} \mathbb{E} [\exp\{F_s\} \mid D_{s-1}] &= \mathbb{E} \left[ \exp \left\{ \underbrace{\gamma \eta^\top z_s \epsilon_s}_{\lambda} \right\} \right] \exp \left\{ - \underbrace{\dot{\sigma}(\theta^\top z_s) \gamma^2 (\eta^\top z_s)^2}_{\nu_s^2} \right\} \\ &\leq \exp\{\nu_s^2 \lambda^2\} \exp\{-\nu_s^2 \lambda^2\} = 1. \end{aligned}$$

This also implies

$$\mathbb{E} [M_t^\gamma(\theta_*, \eta) \mid D_{t-1}] \leq M_{t-1}^\gamma(\theta_*, \eta)$$

and  $M_t^\gamma(\theta_*, \eta)$  is a super-martingale satisfying

$$\mathbb{E} \left[ \exp \left\{ \gamma \eta^\top S_t - \gamma^2 \|\eta\|_{\mathbf{H}_t(\theta_*)}^2 \right\} \right] \leq 1, \forall \gamma \geq 0$$

and we can apply the results of de la Peña et al. (2004).

We now follow the last step of the proof of Lemma 1 in Filippi et al. (2010). We let  $a = \sqrt{2 \log \frac{1}{\delta}}$  for some  $\delta \in (0, 1/e)$  and let  $b = \lambda_0 \|\eta\|_2^2$ . We have with probability at least  $1 - \delta$

$$|\eta^\top S_t| \leq \sqrt{2 \log \frac{1}{\delta}} \sqrt{\|\eta\|_{\mathbf{H}_t(\theta_*) + \lambda_0 \|\eta\|_2^2}^2 \left( 1 + \frac{1}{2} \log \left( 1 + \frac{\|\eta\|_{\mathbf{H}_t(\theta_*)}^2}{\lambda_0 \|\eta\|_2^2} \right) \right)}.$$

Rearranging and using the fact that  $\lambda_0 \|\eta\|_2^2 \leq \|\eta\|_{\mathbf{H}_t(\theta_*)}^2 \leq t \|\eta\|_2$  yields

$$|\eta^\top S_t| \leq \rho(\lambda_0) \|\eta\|_{\mathbf{H}_t(\theta_*)} \sqrt{2 \log \frac{t}{\delta}}. \quad (19)$$

where  $\rho$  is defined as

$$\rho(\lambda_0) = \sqrt{3 + 2 \log \left( 1 + \frac{4Q^2}{\lambda_0} \right)}.$$

We take  $M_t$  to be a matrix such that  $M_t^2 = \mathbf{H}_t(\theta_*)$  and note that for any  $\tau > 0$

$$P \left( \|S_t\|_{\mathbf{H}_t^{-1}(\theta_*)}^2 \geq d\tau^2 \right) \leq \sum_{i=1}^d P \left( |S_t^\top M_t^{-1} e_i| \geq \tau \right)$$

where  $e_i$  is the  $i$ :th unit vector. Equation 19 with  $\eta = M_t^{-1} e_i$  together with  $\|M_t^{-1} e_i\|_{\mathbf{H}_t(\theta_*)} = 1$  yield that the following holds with probability at least  $1 - \delta$

$$\|S_t\|_{\mathbf{H}_t^{-1}(\theta_*)} \leq \rho(\lambda_0) \sqrt{2d \log t} \sqrt{\log \frac{d}{\delta}}. \quad (20)$$

### Step 3. (Unverifiable) High-probability bounds on $\mathcal{E}_1$ and $\mathcal{E}_2$ .

We now have enough machinery to state high-probability bounds for our two events. These bounds will be *unverifiable* in the sense that they depend on the true parameter  $\theta_*$  which is not known to us during runtime. We derive verifiable bounds in the next step of the proof.

Recall that  $\mathbf{H}_t^{-1}(\theta_*)$  is symmetric. We apply Equation 19 with  $\eta = \mathbf{H}_t^{-1}(\theta_*) z_{ij}$  and  $\alpha > 0$  in place of  $\delta$ . First, we note that  $\|\mathbf{H}_t^{-1}(\theta_*) z_{ij}\|_{\mathbf{H}_t(\theta_*)} = \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)}$  which implies with probability at least  $1 - \alpha$

$$\left| z_{ij}^\top \mathbf{H}_t^{-1}(\theta_*) S_t \right| = \left| S_t^\top \mathbf{H}_t^{-1}(\theta_*) z_{ij} \right| \leq \rho(\lambda_0) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)} \sqrt{2 \log \frac{t}{\alpha}}. \quad (21)$$

We solve for smallest possible  $\alpha \in (0, 1/e)$  such that

$$(1 + 2S) \rho(\lambda_0) \dot{\sigma}(z_{ij}^\top \theta_*) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)} \sqrt{2 \log \frac{t}{\alpha}} \leq \frac{\Delta}{2}$$

Rearranging yields

$$\alpha \leq \exp \left\{ \frac{-\Delta^2}{8\rho^2(\lambda_0)(1+2S)^2 \left( \dot{\sigma}(z_{ij}^\top \theta_*) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)} \right)^2 + \log t} \right\}. \quad (22)$$

For  $\mathcal{E}_2$  and the bound on its probability,  $\beta > 0$  we have

$$\frac{1}{4} |z_{ij}^\top (\theta_t - \theta_*)|^2 \leq \frac{1}{2} (1 + 2S)^2 \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)}^2 \rho^2(\lambda_0) \log \frac{t}{\beta} \leq \frac{\Delta}{2}$$

and

$$\beta \leq \exp \left\{ \frac{-\Delta}{\rho^2(\lambda_0)(1+2S)^2 \left( \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)} \right)^2 + \log t} \right\}. \quad (23)$$

Note that both Equation 22 and Equation 23 are under the assumption that the RHS satisfy  $< 1/e$  since this is required in order to apply the results of de la Peña et al. (2004). As we discuss in the main text, these quantities are approaching zero as  $O(te^{-t})$ , ignoring various constants, for reasonable sampling strategies and will satisfy this condition eventually.

### Step 4. (Verifiable) High-probability bounds on $\mathcal{E}_1$ and $\mathcal{E}_2$ .

The bounds in the previous step depend on the true parameter  $\theta_*$  which we do not have access to in practice. We again use Lemma 10 of Fauray et al. (2020) together with Cauchy-Schwartz

$$\begin{aligned} |z_{ij}^\top (\theta_* - \theta_t)| &= \left| z_{ij}^\top \mathbf{H}_t^{-1/2}(\theta_t) \mathbf{H}_t^{1/2}(\theta_t) S_t \right| \\ &\leq (1 + 2S) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \|S_t\|_{\mathbf{H}_t^{-1}(\theta_*)}. \end{aligned}$$

Connecting to Equation (13), we have

$$\begin{aligned} |\sigma(z_{ij}^\top \theta_t) - \sigma(z_{ij}^\top \theta_*)| &\leq \dot{\sigma}(z_{ij}^\top \theta_t) |z_{ij}^\top (\theta_t - \theta_*)| + \frac{1}{4} |z_{ij}^\top (\theta_t - \theta_*)|^2 \\ &\leq \dot{\sigma}(z_{ij}^\top \theta_t) (1 + 2S) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \|S_t\|_{\mathbf{H}_t^{-1}(\theta_*)} \\ &\quad + \left( \frac{1}{4} (1 + 2S) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \|S_t\|_{\mathbf{H}_t^{-1}(\theta_*)} \right)^2. \end{aligned}$$

Using Equation 20, and the fact that all terms are strictly greater than 0, we have with probability at least  $1 - \alpha$

$$(1 + 2S) \dot{\sigma}(z_{ij}^\top \theta_t) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \|S_t\|_{\mathbf{H}_t^{-1}(\theta_*)} \leq (1 + 2S) \dot{\sigma}(z_{ij}^\top \theta_t) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \rho(\lambda_0) \sqrt{2d \log t} \sqrt{\log \frac{d}{\alpha}}. \quad (24)$$

We solve for smallest  $\alpha \in (1/e)$  such that Equation 24 is smaller than  $\Delta_{ij}/2$ . This yields

$$\alpha \leq \exp \left\{ \frac{-\Delta^2}{8d\rho^2(\lambda_0)(1 + 2S)^2 \left( \dot{\sigma}(z_{ij}^\top \theta_t) \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \right)^2} + \log dt \right\}.$$

Same steps for  $\beta$  yields

$$\beta \leq \exp \left\{ \frac{-\Delta}{d\rho^2(\lambda_0)(1 + 2S)^2 \left( \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)} \right)^2} + \log dt \right\}.$$

For brevity, define  $C_1 = \rho^2(\lambda_0)(1 + 2S)^2$ .

Using the definition of  $\tilde{\mathbf{H}}_t$  yields the statement of Lemma 1.  $\square$

## C.2 Proof of Theorem 1

*Proof.* We let  $i \succ j$  denote that  $i$  is preferred to  $j$ . W.l.o.g assume  $1 \succ 2 \succ \dots \succ n$ . The key observation is that for any  $i$  and  $j$  such that  $i < j$  it holds that

$$\Delta_{i,j} > (j - i) \Delta_*.$$

If we get the wrong relation between  $i, j$  then  $\sigma(z_{ij}^\top \theta_*) - \sigma(z_{ij}^\top \theta_T) > (j - i) \Delta_*$ . Lemma 1 implies

$$\begin{aligned} P(\sigma(z_{ij}^\top \theta_*) - \sigma(z_{ij}^\top \theta_T) > (j - i) \Delta) &\leq dT \left( \exp \left\{ \frac{-(j - i) \Delta^2}{8d\rho^2(\lambda_0)(1 + 2S)^2 \left( \dot{\sigma}(z_{ij}^\top \theta_T) \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)} \right)^2} \right\} \right. \\ &\quad \left. + \exp \left\{ \frac{-(j - i) \Delta}{d\rho(\lambda_0)(1 + 2S)^2 \left( \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)} \right)^2} \right\} \right). \end{aligned}$$

$\underbrace{\hspace{15em}}_{\alpha_{ij}^{j-i}} \quad \underbrace{\hspace{15em}}_{\beta_{ij}^{j-i}}$

Let  $R(\theta_T)$  be the ordering error of the  $n$  items. Then, under a uniform distribution over items we have

$$\mathbb{E}[R(\theta_T)] \leq \frac{4dT}{n(n-1)} \left( \underbrace{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \alpha_{ij}^{j-i}}_A + \underbrace{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \beta_{ij}^{j-i}}_B \right) \quad (25)$$

$A$  and  $B$  will be upper bounded using the same argument. We now upper bound sum  $A$



Let  $\alpha_* := \exp \left\{ \frac{-\Delta_*^2}{8dC_1 \max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T) \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_*)}^2} \right\}$  then

$$\begin{aligned} A &\leq \sum_{i=1}^{n-1} \sum_{j=i+1}^n \alpha_*^{j-i} \leq (n-1) \left( \sum_{j=0}^n \alpha_*^j - 1 \right) \\ &\leq (n-1) \left( \frac{1}{1-\alpha_*} - 1 \right). \end{aligned}$$

This follows from the definition of  $\delta_{1,*}$  and properties of the geometric sum. It is easy to see that  $\frac{1}{1-e^{-x}} - 1 = \frac{1}{e^x - 1}$ . Hence,

$$\frac{4dT}{n(n-1)} A \leq \frac{4dT}{n} (\alpha_*^{-1} - 1)^{-1}.$$

For  $B$  we perform the same steps with  $\beta_* := \exp \left\{ \frac{-\Delta_*}{dC_1 \max_{i,j} \|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_*)}^2} \right\}$  to get

$$\frac{4dT}{n(n-1)} A \leq \frac{4dT}{n} (\beta_*^{-1} - 1)^{-1}.$$

Combing yields and

$$\mathbb{E}[R(\theta_T)] \leq \frac{4dT}{n} \left( (\alpha_*^{-1} - 1)^{-1} + (\beta_*^{-1} - 1)^{-1} \right)$$

By Markov's inequality we have

$$P(R(\theta_T) \geq \epsilon) \leq \frac{4dT}{\epsilon n} \left( (\alpha_*^{-1} - 1)^{-1} + (\beta_*^{-1} - 1)^{-1} \right). \quad (26)$$

□

### C.3 Extensions of current theory

**Regularized estimators.** In our analysis in Section 4, we have assumed that  $\theta_T$  is the maximum likelihood estimate and that  $\mathbf{H}(\theta_T)$  has full rank. This can be relaxed by considering  $\ell_2$  (Ridge) regularization where  $\theta_{\lambda_0, T}$  is the optimum of the regularized log-likelihood with regularization  $\lambda_0 \mathbf{I}$  and  $\mathbf{H}_{\lambda_0}(\theta_{\lambda_0, T}) = \sum_{s=1}^T \dot{\sigma}(z_s^\top \theta_{\lambda_0, T}) z_s z_s^\top + \lambda_0 \mathbf{I}$ . The same machinery used to prove Lemma 1 (Filippi et al., 2010; Fauray et al., 2020) can be applied to this regularized version with small changes to the final bound.

**Generalized linear models.** It is also possible to derive similar results for generalized linear models with other link functions,  $\mu(z_{ij}^\top \theta_*)$ , by using the general inequality  $\mathbf{H}(\theta) \geq \kappa^{-1} \mathbf{V}$  with  $\mathbf{V} = \sum_{s=1}^T z_s z_s^\top$  and  $\kappa \geq 1 / \min_{z_{ij}} \dot{\mu}(z_{ij}^\top \theta_*)$ . We conjecture that this will yield a scaling of  $\sim \exp(-\Delta^2 T / \kappa)$  where, unfortunately,  $\kappa$  might be very large. For a more thorough discussion on the dependence on  $\kappa$  in generalized linear bandits, see Lattimore and Szepesvári (2020, Chapter 19).

**Lower and algorithm-specific upper bounds on the ordering error.** A worst-case lower bound on the ordering error can be constructed in the fixed-confidence setting, where the goal is to minimize the number of comparisons until a correct ordering is found with a given confidence, by following Garivier and Kaufmann (2016). This involves defining the set of *alternative* models  $\text{Alt}(\theta_*)$  which differs from  $\theta_*$  in their induced ordering of  $\mathcal{I}$ . The bound is then constructed by optimizing the frequency of comparisons of each pair of items so that such alternative models are distinguished as much as possible from the true parameter. We have left this result out of the paper as we find it uninformative in the regime when the number of comparisons is small, (see Simchowitz et al. (2017) for a discussion on the limitations of these asymptotic results in the standard bandit setting). Constructing a lower bound for our fixed-budget setting, of learning as good an ordering as possible with a fixed number of comparisons, is much more challenging. The fixed-confidence result yields a bound for the fixed-budget case (Garivier and Kaufmann, 2016), but constructing either a tight lower bound or a tight algorithm-specific upper bound is an open problem (Fang, 2022).

## D Comparison with regret minimization

Bengs et al. (2022) considered a problem formulation where the goal is to learn a parameter  $\theta$  which determines the utility  $Y_{i,t}$  for a set of arms  $i = 1, \dots, n$  as a function of observed context vectors  $x_{i,t}$  in a sequence of rounds  $t = 1, \dots, T$ ,

$$Y_{i,t} = \theta^\top X_{i,t}.$$

The probability that item  $i$  is preferred over  $j$  (denoted  $i \succ j$ ) in round  $t$  is decided through a comparison function  $F$ ,

$$\Pr(i \succ j \mid X_{i,t}, X_{j,t}) = F(Y_{i,t} - Y_{j,t}).$$

The goal in their setting is to, in each round, select two items  $(i_t, j_t)$  so that their maximum (or average) utility is as close as possible to the utility of the best item. The expected regret in their average-utility setting is

$$\mathfrak{R}_{BSH} = \mathbb{E}\left[\sum_{t=1}^T 2Y_{i_t^*,t} - Y_{i_t,t} - Y_{j_t,t}\right].$$

**Proposition 1** (Informal). *An algorithm which achieves minimal regret in the setting of Bengs et al. (2022) can perform arbitrarily poorly in our setting.*

*Proof.* The optimal choice of arm pair in the BSH setting is the optimal and next-optimal arm  $(i_t^*, i_t')$  such that  $i_t^* \succ i_t' \succ j$  for any other arms  $j$ . Assume that the ordering of all other arms  $j$  is determined by a feature  $X_{j,t}(k)$  but that  $X_{i_t^*,t}(k) = X_{i_t',t}(k)$ . Then, no knowledge will be gained about arms other than the top 2 choices under the BSH regret. As the number of arms grows larger, the error in our setting grows as well.  $\square$

Saha (2021) study the same average-utility regret setting and give a lower bound under Gumbel noise. Saha and Krishnamurthy (2022) investigated where there is a computationally efficient algorithm that achieves the derived optimality guarantee.

## E Experiment details

For BayesGURO and BALD, the posterior  $p(\theta | D_t)$  is estimated using the Laplace approximation as described in Bishop and Nasrabadi (2006, Chapter 4). With this approximation, the covariance matrix is the same as the inverse of the Hessian of the log-likelihood. For both methods, the priors  $\theta_{B,0} = \mathbf{0}^d$  and  $\mathbf{H}_{B,0}^{-1} = I_d$  were used, and sequential updates were performed every iteration. The sample criterion for BALD under a logistic model is given in Appendix B.5.1. For BayesGURO, 50 posterior samples were used to estimate  $\hat{\mathbb{V}}_{\theta|D_t}[\sigma(\theta^T z_{ij})]$  for every  $z_{ij}$ . The hybrid algorithms follow the same structure with the added constraint that each per-item parameter  $\zeta_i$  is independent of other parameters. This allows for efficient updates of  $\mathbf{H}_{B,t}^{-1}$  by using sparsity in the covariance.

GURO, CoLSTIM, and Uniform use LogisticRegression from Scikit-learn (Pedregosa et al., 2011) with default Ridge regularization ( $C = 1$ ) and the lbfgs optimizer. The former two updates  $\theta_t$  every iteration using the full history,  $D_t$  in all experiments except for IMDB-WIKI-SbS, where GURO updates  $\theta_t$  every 25th iteration. This caused no noticeable change in performance as GURO still updates  $\mathbf{H}_t^{-1}$  every iteration using the Sherman-Morrison formula. Note that when using the Sherman-Morrison formula in practice, you only get an estimate of  $\mathbf{H}_t^{-1}(\theta_t)$  since previous versions have been calculated using older estimates of  $\theta$ . This method for approximating the inverse hessian is covered in Bishop and Nasrabadi (2006, Chapter 5) and when we compared it to calculating  $\mathbf{H}_t^{-1}(\theta_t)$  from scratch every iteration we observed that the methods performed equally. The design matrix for CoLSTIM is updated as in Bengs et al. (2022): the confidence width  $c_1$  was chosen to be  $\sqrt{d \log(T)}$ , and the perturbed values were generated using the standard Gumbel distribution.

To increase computational efficiency for the large IMDB-WIKI-SbS dataset, the hybrid algorithms did not evaluate all  $\sim 100\,000$  comparisons at every time step. Instead, a subset of 5 000 comparisons was first sampled, and the highest-scoring pair in this set was chosen. This resulted in a large speed-up and no noticeable change in performance during evaluation.

### E.1 Datasets

**ImageClarity** Data available at <https://dbgroup.cs.tsinghua.edu.cn/ligl/crowdtopk>. This dataset contained differently distorted versions of the same image. To extract relevant features, we used a ResNet34 model (He et al., 2016) that had been pre-trained on Imagenet (Deng et al., 2009). After PCA projection feature dimensionality was reduced to  $d = 63$ . The dataset consisted of 100 images and 27 730 comparisons. Since the type of distortion is the same for all images, the dataset has a true ordering with regards to the strength of the distortion applied.

**WiscAdds** Data available at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OZRGEE> (license: CC 0 1.0). The WiscAdds dataset, containing 935 political texts, has been extended with 9 528 pairwise comparisons by Carlson and Montgomery (2017). In comparisons, annotators have stated which of two texts has a more negative tone toward a political opponent. To extract general features from the text, sentences were embedded using the pre-trained all-mpnet-base-v2 model from the Sentence-Transformers library (Reimers and Gurevych, 2019). After applying PCA to the sentence embeddings, each embedding had a dimensionality of  $d = 162$ .

**IMDB-WIKI-SbS** Data available at <https://github.com/Toloka/IMDB-WIKI-SbS> (license: CC BY). IMDB-WIKI-SbS consists of close-up images of actors of different ages. For each comparison, the label corresponds to which of two people appears older. The complete dataset consists of 9 150 images and 250 249 comparisons, but images that were grayscale or had a resolution lower than  $160 \times 160$  were removed, resulting in 6 072 images and 110 349 comparisons. We extract features from each image using the Inception-ResNet implemented in FaceNet (Schroff et al., 2015) followed by PCA, resulting in  $d = 75$  features per image.

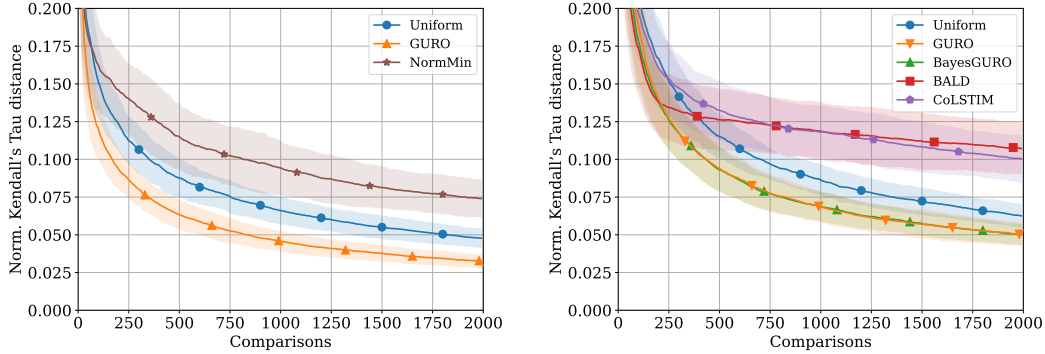
### E.2 Additional figures

#### X-RayAge

To highlight the importance of the first-order term in Lemma 1, we evaluated NormMin on the same X-ray ordering task as in Figure 1a. The results, shown in Figure 3a, indicate that not only does the algorithm perform worse than GURO, but is seemingly also outperformed by a uniform sampling strategy. Furthermore, for completeness, we include Figure 3b which shows the in-sample error,  $R_{I_D}$ , during the generalization experiment.

#### Synthetic Example and Illustration of Upper Bound

In this setting, 100 synthetic data points were generated. Each data point consisted of 10 features, where the feature values were sampled according to a standard normal distribution. The true model,  $\theta_*$ , was generated by sampling each value uniformly between  $-3$  and  $3$ . The pairwise comparison feedback was simulated the same

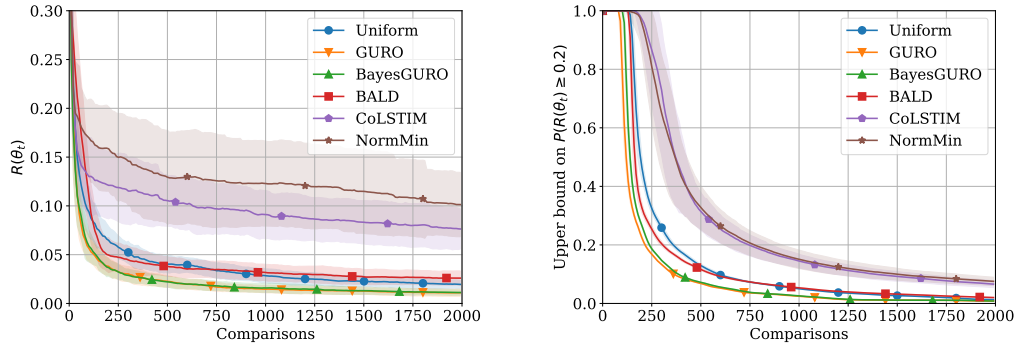


(a) **X-RayAge**. NormMin included in the experiment shown in Figure 1a.

(b) **X-RayAge**. The in-sample error  $R_{ID}$  for the generalization experiment performed in Figure 1b.

Figure 3: Additional figures from the **X-RayAge** experiment.

way as in Section 6.1, with  $\lambda = 0.5$ . The upper bound of the probability that  $R(\theta_t) \geq 0.2$  was calculated every iteration according to Theorem 1. Each algorithm was run for 2000 comparisons, updating every 10th, the results of which can be seen in Figure 4. We observe in Figure 4b that our greedy algorithms are seemingly the fastest at minimizing the upper bound. The order of performance follows the same trend as in the experiments of Section 6.



(a) The risk  $R(\theta_t)$ , defined as the normalized Kendall's tau distance between estimated and true orderings.

(b) The probability that the frequency of pairwise inversions is  $\geq 20\%$  after every comparison, according to (1).

Figure 4: The loss (left) along with the upper bound (right) when ordering a list of size 100 in a synthetic environment. The results have been averaged over 50 seeds.

### Randomly initialized representation

As discussed in Section 6.2, the performance of our contextual approach will depend on the quality of the representations. To underscore the practical usefulness of our algorithms, we have performed the same experiment as in Figure 2c, but this time the model used to extract image features was untrained (i.e., the weights were random). As to be expected, the results, shown in Figure 5, demonstrate that the fully contextual algorithms have no real way of ordering the items according to these uninformative features. However, GURO Hybrid performs similarly to TrueSkill, despite model misspecification. This is promising, since you may not know in advance how informative the extracted features will be for the target ordering task.

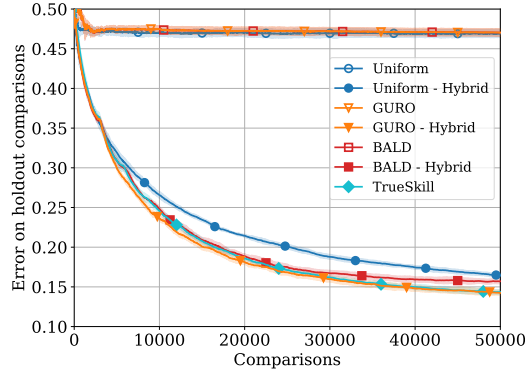


Figure 5: **IMDB-WIKI-SbS**. The same experiment as presented in Figure. 2c, but the model used for feature extraction is untrained.

### ImageClarity ground truth

The ImageClarity dataset consists of multiple versions of the *same image*, with the *same distortion* applied to it to varying degrees. Due to this artificial construction, the pairwise comparisons should, given enough samples, reflect the magnitudes of the applied distortions. In Figure 6 we perform the same experiment as in Figure 2a, but instead of evaluating on a holdout comparison set, we measure the distance to the ground-truth ordering. The overall results are very similar, although we do see a slight increase in the performance of contextual algorithms compared to the non-contextual TrueSkill.

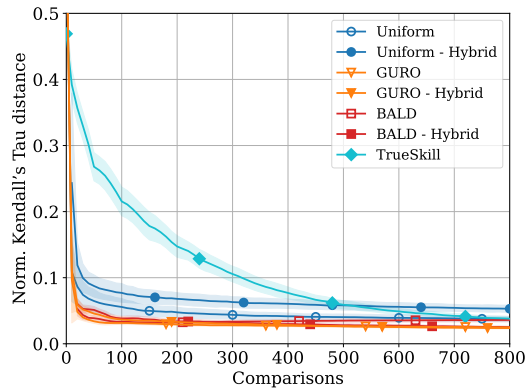
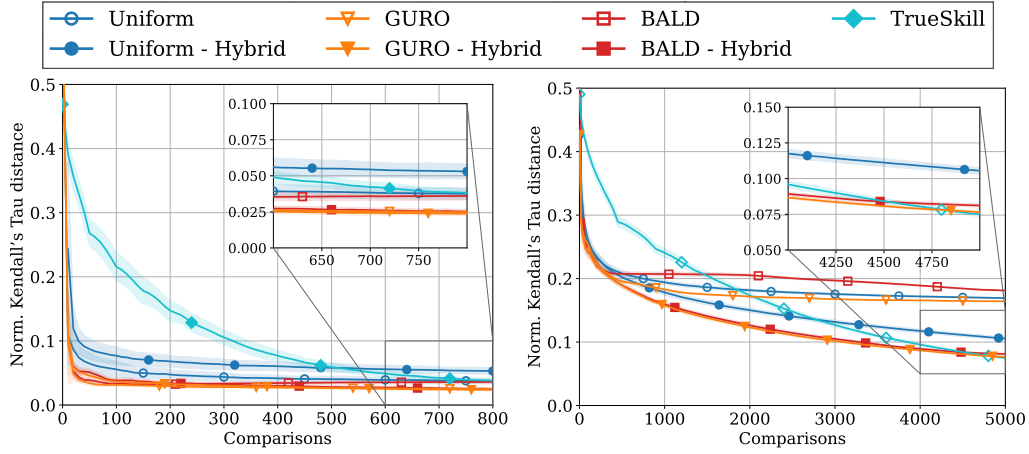


Figure 6: **ImageClarity**. Same experiment as in Figure 2a, but now measuring the distance to the ground-truth ordering. Averaged over 25 seeds along with the 1-sigma error region.

### Ground truth ordering using the Bradley-Terry model

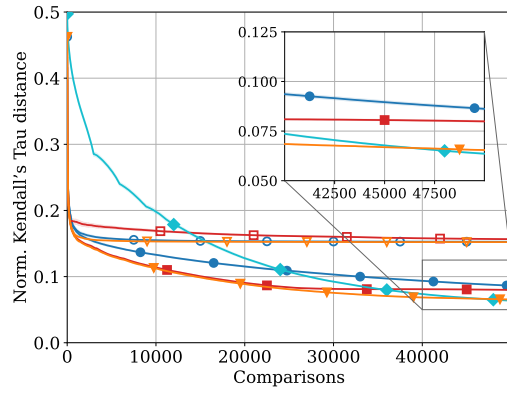
An alternate approach to evaluate ordering quality is to estimate a "ground-truth" ordering by applying the popular Bradley-Terry (BT) model (Bradley and Terry, 1952) to all available comparisons. We used the CrowdKit library (Ustalov et al., 2024) to find the MLE scores for each item and ordered the elements accordingly. In Figure 7 we run the same experiments as in Figure 2, but instead measure the distance to the constructed BT ordering. The overall trends remain, but for (b) and (c) there is a slight shift for the later iterations. More specifically we see non-contextual TrueSkill eventually overtaking the contextual algorithms.

The issue is that algorithms with orderings closer to the maximum likelihood estimate of the BT model will be favored. To exemplify this we use the ImageClarity dataset since it contains the largest number of comparisons relative to the number of items. We sample 1 000 comparisons and let this be the collection that is available to the algorithms. We further construct two target orderings, one from the BT estimate using the sampled subset of comparisons, and a second, more probable ordering, from the BT estimate using all 27 730 available comparisons. Figure 8 shows the distance between the GURO and TS algorithms and the different target orderings, where dashed lines indicate the distance to the ordering generated using the full list of comparisons. If we only look at the distance to the ordering produced using our subset of comparisons, TrueSkill seemingly outperforms GURO after about 350 comparisons. However, if we instead measure the distance to the more probable order-



(a) **ImageClarity**.  $n = 100$ ,  $d = 63$ .

(b) **WiscAds**.  $n = 935$ ,  $d = 162$ .



(c) **IMDB-WIKI-SbS**.  $n = 6\ 072$ ,  $d = 75$ .

Figure 7: The same experiment as presented in Figure. 2c, but we instead measure the distance to a ground-truth estimated using all available comparisons.

ing, we see that GURO converges toward a lower distance. Note that these are the same orderings, evaluated against different targets. This is likely the effect we observe in Figure 7b and c, but not in Figure 7a as a result of the high amount of comparisons available to us.

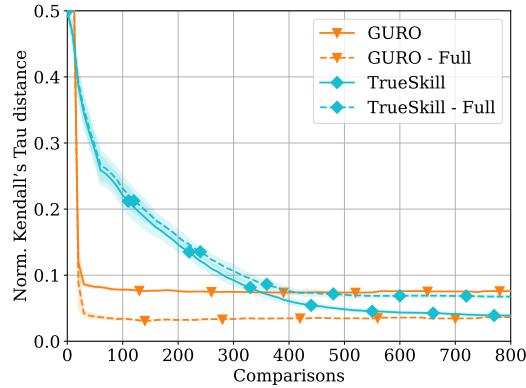


Figure 8: **ImageClarity**. The same experiment as presented in Figure 2a, but we instead measure the distance to target orderings that correspond to the maximum likelihood estimate of the BT model using different numbers of comparisons. The dashed lines show the distance to the BT estimate using all 27 730 comparisons.

# NeurIPS Paper Checklist

## 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: In the introduction, we clearly state where in the paper each contribution can be found.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

## 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In Section 7 we highlight limitations such as real data experiments not being performed in an online setting and the lack of optimality guarantees for our algorithm.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

## 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We clearly define our problem setting in Section 2, including relevant assumptions. Additionally, in Section 4 we list the three assumptions that our analysis depends on, and include the complete proofs in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Among our main contributions are the GURO and BayesGURO sampling strategies, which we not only offer algorithmic descriptions of but also provide the code for. We also describe our experimental setup in Section 6 and Appendix E including hyperparameters and pre-trained models used for feature extraction.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code in a publicly available repository. We have also provided links from which the datasets can be downloaded.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.



- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental details, including splits and hyperparameters, in Section 6.2 as well as Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In Section 6 we either provide 1-sigma error bars or the 95% confidence interval, explicitly stating which in the figure captions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Section 6.2 we state the compute resources used for the most demanding experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We have read and ensured that our work conforms to the ethical guidelines stipulated.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In the introduction, we describe our motivating example of annotating medical images. Enabling quantitative analysis of these images could be highly beneficial for medical research. We do however not see a direct path from our active sampling criterion to a negative application.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The algorithms we provide do not pose any particular risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the authors of all datasets we use according to their wishes in Section 6. We further include URL's in Appendix E.1 along with licenses for all available datasets except ImageClarity, for which we could not find the relevant information.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not introduce any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We have not performed any crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We did not perform any experiments with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.