

Data Efficient Sample Selection for In-Context Learning

Venktesh V 

Stockholm University
venktesh.viswanathan
@dsv.su.se

Orcid: 0000-0001-5885-2175

Cem Levi 

SKIM Group B.V.
c.levi
@skimgroup.com

Orcid: 0009-0003-8367-101X

Avishek Anand 

TU Delft
avishek.anand
@tudelft.nl

Orcid: 0000-0002-0163-0739

Abstract

The In-context learning (ICL) paradigm aids large language models (LLMs) to adapt to new tasks without need for fine-tuning. However, selecting an optimal combination of demonstration examples from a large pool of example subsets is a challenging problem. Existing approaches for selection do not model the complex relationship between ICL samples and downstream LLM performance. They typically perform static task-level selection, choosing subsets once offline, which can fail to generalize to unseen queries. We introduce DEARICL (Data Efficient Algorithm for Ranking) ICL samples, a new framework that models demonstration example selection as a subset ranking problem. DEARICL employs a non-linear surrogate employing a differentiable sorting objective within a gap-index bandit algorithm. The gap-index based approach enables fine-grained separation of good arms and borderline arms which is used as an auxiliary objective to train the non-linear surrogate through sufficient sampling of borderline arms, supporting instance-level subset ranking. On exemplar selection benchmarks with open-source LLMs, DEARICL achieves **8.08–15.9%** accuracy gains over strong linear bandit baselines, with low sample complexity. Code and data: ¹

1 Introduction

Selecting representative instances from a large pool is a recurring need across applications such as training models, domain adaptation, and few-shot learning. In this work, we focus on *in-context learning* (ICL) with large language models (LLMs), where LLMs can solve new tasks when given a small sequence of demonstrations (v, w) or $(v, \text{rationale}, w)$ in context (Brown et al., 2020). However, naively choosing demonstrations (randomly or heuristically) performs poorly (Purohit et al., 2024; Li and Qiu, 2023). One principled

way to formulate this is as a *subset selection* problem: choosing a small set of examples that captures the task’s salient structure, according to a task-specific reward signal. The class of approaches that aim to select such near-optimal example subsets that achieves the best average performance across all test / validation queries from a target task are broadly categorized as static (task-level) selection methods (Li and Qiu, 2023; Zhang et al., 2025).

Since, ICL performance depends on *subset interactions*, the joint effect of examples provided together, renders a large search space of example subsets. For sample-efficiency, some of the existing task-level selection methods formulates subset selection as *top- m arm identification* in multi-armed bandits (MABs), where each subset of examples is an arm and rewards are sampled based on LLM performance on the task (Réda et al., 2021). Casting ICL demonstration choice as top- m arm identification provides a principled route to select representative subsets using the task-level rewards, while reducing number of LLM evaluations. Recent works (Purohit et al., 2025, 2024), builds upon gap-index bandit frameworks like GIFA (Réda et al., 2021), introducing challenger sampling mechanisms to handle the large search space. While effective, these approaches have two major limitations. First, they rely on a *linear surrogate* to map arm features to rewards which cannot capture the complex dependencies between subsets and task performance.

Secondly, a fundamental limitation of task-level (static) selection approaches is that they can fail when new queries require skills absent from the fixed set of examples. Dynamic selection chooses examples per test query, which improves flexibility but involves searching through a large space of combinations and applying MAB algorithm per instance is computationally expensive. Existing approaches only employ heuristic semantic matching (Ye et al., 2023b; Rubin et al., 2022) based on embeddings that are not adapted for the subset selec-

¹<https://github.com/VenkteshV/DearICL>

tion setting. However, training ranking models for subset ranking that can clearly distinguish between optimal example subsets from sub-optimal ones for each test query is expensive. It requires sampling of sufficient number of borderline example subsets that are hard to distinguish for the subset ranker.

Hence, it is critical to devise a sample-efficient mechanism that can help learning of subset ranking at instance-level which can also model the non-linear reward landscape of LLM based feedback. To overcome these limitations, we propose DEARICL (**D**ata **E**fficient **A**lgorithm for **R**anker training for demonstration samples selection in ICL), a new bandit framework for instance-level subset selection. DEARICL is a gap-index MAB framework which focuses on top- m selection as auxiliary objective in the outer loop, while fitting a non-linear surrogate in the inner loop for instance-level subset ranking. The top- m selection aims to implicitly induce a separation between near-optimal subsets (arms) and sub-optimal ones as discussed in literature (Katariya et al., 2019) and hence the gap-index framework ensures sufficient number of borderline arms are sampled to induce this separation. Hence, by using this as auxiliary objective, our proposed approach, ensures that borderline subsets with rewards sampled at instance-level (each validation query) are used to fit the ranker in the inner loop. This enables the surrogate to clearly distinguish borderline example subsets from useful ones for each instance during inference for ranking example subsets for each test query. We provide **theoretical guarantees**, including bounds on pairwise gap error and sample complexity. Our approach is sample-efficient while improving performance (**8.08–15.9%**) on **smaller LLMs**.

2 Related Works

Exemplar Selection for ICL. The rise of LLMs has transformed them into general-purpose answering engines through emergent capabilities like ICL (Brown et al., 2020; Wei et al., 2022a,b; Wang et al., 2023; Kojima et al., 2022; Chen et al., 2023) where a few examples are provided to LLMs to demonstrate the task. To eliminate manual selection, several automated methods have emerged, such as reinforcement learning (Zhang et al., 2022; Lu et al., 2023), trained retrievers (Xiong et al., 2024), Determinantal Point Processes (Ye et al., 2023a) and constrained optimization (Tonglet et al., 2023). Additionally, instance-level selection meth-

ods that are learning-free, such as similarity-based (Rubin et al., 2022), complexity-based (Fu et al., 2023), and MMR (Ye et al., 2023b), have been explored. However, instance-level methods increase inference-time computational costs. To address this, a pre-selected, representative set of exemplars is chosen for ICL, akin to coreset selection methods (Guo et al., 2022), though the key difference is that ICL does not involve parameter updates. While existing methods (Purohit et al., 2025, 2024; Wan et al., 2025) propose bandit-based or bayesian optimization based task-level example selection algorithm, they focus on static selection, which does not transfer well to new test questions.

Learning to Rank approaches: The existing works from online learning to rank literature are related to our surrogate (Zoghi et al., 2017; Grotov and de Rijke, 2016; Li et al., 2019), which learn the ranking models from user interaction or click data. However, our approach differs fundamentally, as we do not rely on direct user feedback but from LLM feedback, obviating challenges like de-biasing clicks.

Top- m identification in stochastic bandits: The objective of top- m arm identification is to identify those arms with highest means. While fixed-confidence (Kalyanakrishnan et al., 2012) and fixed-budget settings (Bubeck et al., 2013) exist, our focus is the fixed-confidence setting, where the error probability to estimate the top- m arms should be smaller than a predefined parameter $\delta \in (0, 1)$. Adaptive sampling algorithms like UGapE (Gabilon et al., 2012) and LUCB (Kalyanakrishnan et al., 2012), along with uniform sampling methods (Kaufmann and Kalyanakrishnan, 2013; Chen et al., 2017), have been introduced for the fixed confidence setup, but they lack efficiency in terms of sample complexity. While efficient adaptive sampling methods for linear bandits, such as (Fiez et al., 2019), RAGE (Zhang et al., 2024), LTS (Jedra and Proutière, 2020), PEPS (Li et al., 2024), LinGapE (Xu et al., 2018) and LinGame (Degenne et al., 2020), have been proposed, they primarily address best-arm identification ($m = 1$). GIFA (Réda et al., 2021) was the first unified framework for efficient top- m arm identification, but requires a significant number of gap-index computations and comparisons, leading to high sample complexity.

3 Methods

3.1 Problem Definition

In ICL, the model processes a sequence of input–output demonstrations followed by a new test input, and is expected to generate the corresponding output. We frame the problem of ICL as choosing representative example subset from a large dataset of existing examples (Li and Qiu, 2023) or training instances (Ho et al., 2020). While static selection of these subsets can be modeled formally as a *top- m arm identification problem* in the multi-armed bandit framework due to its ability to identify near optimal example subsets (Li and Qiu, 2023; Zhang et al., 2025), we discuss the limitations of this static selection and demonstrate how a **new formulation** can be used as auxiliary objective for training non-linear surrogates as rankers dynamic selection. Our framework primarily focuses on sample-efficient fitting of non-linear surrogate utilizing classical top- m selection in bandit literature as auxiliary objective. This surrogate is then employed for instance-level (dynamic) selection through subset ranking.

3.1.1 Subset Selection for ICL - Preliminaries and Limitations of Existing Formulation

In ICL, a large language model processes a sequence of *demonstrations* (i.e., input–output pairs) followed by a test input, and is expected to generate the corresponding test output. We denote each input by v (e.g., a natural language question, a math problem, or a sentence to translate) and each output by w (e.g., an answer, solution, or translation). In some tasks, v may also contain additional reasoning such as rationales or chain-of-thought annotations; for example, solving math word problems may require showing intermediate steps, while translation tasks typically do not.

Let $\mathcal{Q} = \{(v_i, w_i)\}_{i=1}^n$ be a pool of n candidate demonstrations, and let (v_{test}, w_{test}) denote a test instance. To perform inference on v_{test} , we select a subset $S \subseteq \mathcal{Q}$ of k demonstrations and concatenate them with the test input to form the context $E = [(v_{i_1}, w_{i_1}), \dots, (v_{i_m}, w_{i_m}), v_{test}]$, $\hat{w}_{test} = \mathbb{P}_{LLM}(\cdot | E)$

The quality of the chosen subset S has a direct impact on LLM performance. However, the candidate pool might comprise of good near-optimal subsets and sub-optimal subsets which could impact performance. This motivates the identification of the *top- m subsets* of demonstrations that are

most useful for ICL to prune the sub-optimal subsets from the search space. Formally, let $\mathcal{S}(\mathcal{Q})$ denote the set of all k -sized subsets of $\mathcal{Q}$, and let $\mathcal{V} = \{(v_{val}(j), w_{val}(j))\}_{j=1}^{n'}$ be a validation set. The objective is to identify $\{a_1, \dots, a_m\} \subseteq \mathcal{S}(\mathcal{Q})$ that maximize expected performance on $\mathcal{V}$.

The inference procedure with top- m subsets can then be written as: $E = \phi(\{a_1, \dots, a_m\}, v_{test})$, $\hat{w}_{test} = \mathbb{P}_{LLM}(\cdot | E)$, where ϕ constructs the context from selected subsets. For task-level selection, ϕ may pick the subset with the lowest validation loss on $\mathcal{V}$.

Nevertheless, static subsets selected offline as described above may fail to generalize to unseen queries, while fully online selection is computationally prohibitive since running bandit algorithms to convergence per test instance incurs high latency. This motivates the need for a principled formulation that is both *expressive*, capturing complex dependencies in LLM rewards, and *efficient*, supporting practical inference-time selection. We cast it as an auxiliary objective due to its ability to sample borderline arms to adaptively train a non-linear surrogate in the internal loop.

3.1.2 Top- m Arm Selection as Auxiliary objective

We present the *top- m arm identification task* in a multi-armed bandit (MAB) setting. Each candidate subset $a_i \in \mathcal{S}(\mathcal{Q})$ is treated as an arm, with reward defined by LLM performance on $\mathcal{V}$. Since $|\mathcal{S}(\mathcal{Q})|$ can be large, the search space $\mathcal{S}(\mathcal{Q})^m$ of all possible m -subsets is extremely challenging.

Let the true reward for an arm a be $\rho(a; \theta^*) = \mathcal{F}_{\theta^*}(x_a)$, where $x_a \in \mathbb{R}^n$ are arm features and θ^* are the true parameters. Each evaluation of an arm yields a noisy observation:

$$\hat{\rho}(a; \theta) = \rho(a; \theta) + \eta, \quad \mathbb{E}[e^{\lambda\eta}] \leq \exp\left(\frac{\lambda^2 \chi^2}{2}\right),$$

where η is sub-Gaussian with variance χ^2 . The top- m identification objective is to output $\hat{S}_m$ such that $\mathbb{P}(\hat{S}_m \neq S_m^*) \leq \delta$, $S_m^* = \{1, 2, \dots, m\}$,

Gap-index methods and their limitations. Gap-index bandit algorithms such as GIFA (Réda et al., 2021) address top- m arm identification by iteratively estimating arm parameters and comparing the most *ambiguous arms* using gap indices. However, GIFA relies on a *linear surrogate* $\mathcal{F}_\theta$, which is restrictive when modeling rewards induced by LLMs. Linear surrogates cannot capture complex, non-linear dependencies between subsets and

task performance. Additionally, **simply substituting a non-linear surrogate in existing gap-index frameworks is insufficient** for our problem setup. In DEARICL, we propose a novel gap-index bandits framework with a non-linear ranking surrogate, with theoretical bounds on pairwise gap error and sample complexity which applies to the auxiliary top-m selection objective and also number of arm samplings required to fit the non-linear surrogate. After convergence on top-m selection objective, the non-linear surrogate can be employed for dynamic subset selection at inference time. New confidence widths and index computations are devised in the *gap-index framework* to incorporate the **non-linear surrogate** within the **auxiliary objective** of separating optimal arms from sub-optimal ones (top-m selection) and to ensure convergence.

3.2 DEARICL: A sample-efficient algorithm fitting of non-linear Ranking surrogate

Based on above discussion of the problem setup we propose a gap-index based bandit algorithm DEARICL with a non-linear ranking surrogate. The top-m identification algorithm implicitly induces a clustering to separate top-m near-optimal arms from other borderline or sub-optimal arms (Katariya et al., 2019). Hence, the top-m identification is used as *auxiliary objective* to fit the non-linear surrogate, so that it samples sufficient number of borderline arms apart from just repeatedly sampling good arms. Since, the task can also be viewed as learning to rank (LTR) the example subsets (arms), we adopt a *differentiable sorting* objective (Swezey et al., 2021) to train the non-linear surrogate. This surrogate learns to approximate the rewards of arms during the offline run of the bandit algorithm and hence can be used during runtime to rank example subsets for dynamic selection. At each step, an exemplar subset (arm) is regarded as a document whose current empirical mean is estimated by $\mathcal{F}_\theta$. $\mathcal{F}_\theta$ is a multi-layer connected network architecture with RELU activations.

$$\hat{\rho}(a_i) = \frac{1}{n'} \sum_{i=1}^{n'} (\mathcal{F}_\theta(x_a(v_{val}(i)), v_{val}(i))), \quad (1)$$

$$x_a(v_{val}(i)) = \left[\mathcal{H}(v_{val}(i)), \frac{1}{k} \sum_{l=1}^k \mathcal{H}(v_{i_l}; w_{i_l}) \right]$$

where $v_{val}(i)$ from $\mathcal{V}$ is treated as a query and the example subset (arm) is treated as the document. We average the sentence embeddings of examples

obtained using an encoder $\mathcal{H}$ in the arm to provide a single feature representation for the arm x_a of dimension d with respect to query $v_{val}(i)$. The representation with query embedding is input to the surrogate $\mathcal{F}_\theta$ to obtain empirical mean estimate.

Algorithm 1 DEARICL

Require: $\mathcal{Q}$ (training exemplars); k (prompt size); $\mathcal{S}$ (all k -subsets of $\mathcal{Q}$); m (top-m : target); m' (challenger size); $\mathcal{G}$ (data generator); N (query size); $\mathcal{F}_\theta$ (surrogate model)

Ensure: $\mathcal{F}_\theta$ (trained surrogate) and estimated top- m set

- 1: **Initialize**
- 2: $U_0 \leftarrow$ random m arms from $\mathcal{S}$ $\triangleright$ Current Top- m
- 3: $C_0 \leftarrow$ the next best m' arms where $m' < m$ (is resampled every iteration) $\triangleright$ Challengers
- 4: $\mathcal{D} \leftarrow \emptyset$ or random M subsets from $\mathcal{G}(\mathcal{Q})$ to solve cold-start problem
- 5: $t \leftarrow 1$
- 6: **while** $\neg (B_t(ch_t, b_t) \leq \epsilon)$ **do**
- 7: **(A) IDENTIFY ARMS TO SWAP**
- 8: $m_t \leftarrow \arg \min_{a \in U_{t-1}} \hat{\rho}_t(a)$ $\triangleright$ Weakest in U_t
- 9: $c_t \leftarrow \arg \max_{a \in C_{t-1}} \hat{\rho}_t(a)$ $\triangleright$ Strongest challenger
- 10: **(b) Border Update (Swap if Needed)**
- 11: **if** $\hat{\rho}_t(c_t) \geq \hat{\rho}_t(m_t)$ **then**
- 12: Exchange m_t and c_t between U_{t-1} and C_{t-1}
- 13: **end if**
- 14: $U_t, C_t \leftarrow$ updated sets
- 15: **(c) Expand Candidate Pool**
- 16: $M_t \leftarrow$ random m' arms from $(U_t \cup C_{t-1})^c$
- 17: $C_t \leftarrow \text{top}_{m'}(M_t \cup C_{t-1}; \hat{\rho}_t)$
- 18: **(D) RECOMPUTE AMBIGUITY FRONTIER**
- 19: $b_{t+1} \leftarrow \arg \max_{b \in U_t} \max_{ch \in C_t} B_t(ch, b)$
- 20: $ch_{t+1} \leftarrow \arg \max_{ch \in C_t} B_t(ch, b_{t+1})$
- 21: **(E) ACQUIRE FEEDBACK & RETRAIN**
- 22: $a_{t+1} \leftarrow \text{selection_rule}(U_t, C_t)$
- 23: $r_{t+1}(a_{t+1}) \leftarrow R(\psi(a_{t+1}), \mathcal{V})$ $\triangleright$ LLM calls
- 24: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{G}(r_{t+1}(a_{t+1}), a_{t+1})$
- 25: Retrain $\mathcal{F}_\theta$ for one epoch on $\mathcal{D}$
- 26: $t \leftarrow t + 1$
- 27: **end while**
- 28: **Return** $\mathcal{F}_\theta$

Then the arm (example subset) being played provides the score based on LLM output on multiple validation samples from $\mathcal{V}$ as rewards. For a single val. sample, reward is

$$\mathcal{R}(\psi(a_i), v_{val}(i)) = \gamma \left(\mathbb{P}_{LLM}(\cdot | \psi(a_i, v_{val}(i))) \right) \quad (2)$$

Here γ could indicate accuracy or other relevance measures like BertScore which compares the generated output from LLM $\hat{w}_{val}$ with ground truth w_{val} and outputs a relevance score. And ψ denotes the context / prompt generator function based on given subset of examples and the query to be answered. Hence reward for an arm

can be obtained as: $r(a_i) = \mathcal{R}(\psi(a_i), \mathcal{V}) = [\mathcal{R}(\psi(a_i), v_{val}(i)) \cdot \mathcal{R}(\psi(a_i), v_{val}(n'))]$ which is then added as relevance labels of subset (arm) with respect to validation samples (queries) to the dataset D , used to fit the non-linear surrogate to correct estimation errors with ranking loss (Step 24, 25 Alg. 1) where $\mathcal{G}$ is the data formatter/loader.

An overview of the gap-index based bandit algorithm with differentiable sorting surrogate is shown in Algorithm 1. First a shortlist of good arms U_0 is initialized to random m arms in Steps 1-3. The dataset $\mathcal{D}$ which is used to fit the non-linear ranking surrogate is initialized to empty set or random data to solve the cold start problem (Line 4). The updated U_t is computed by selecting the worst-arm in U_t with lowest empirical mean in current step and swap it with the best challenger arm c_t in the challenger shortlist C_t (Lines 8-14). The empirical means for above steps are computed using the formulation in Equation 1. In Lines 15-17, we uniformly sample m' arms from $(U_t \cup C_{t-1})^c$, to generate the set M_t , and then select the top- m' arms from $M_t \cup C_{t-1}$ to generate the updated C_t . The most ambiguous arms b_t (guess for m-best arm) and ch_t (a potentially misassessed arm m-best arm) which determine the stopping criterion are computed with help of gap-indices as shown in Steps 18-20. The gap-index between any two arms i, j is computed as: $B_t(i, j) = \hat{\rho}_t(i) - \hat{\rho}_t(j) + W_t(i, j)$. Here in gap-index computation, $W_t(i, j)$ is computed as per **Equation 5** (the RHS of the inequality from **Theorem 3.2**) accounting for the non-linear ranking based surrogate in DEARICL. $C_t \cup U_t$ bounds the amount of comparisons required for gap-index computations unlike GIFA. The intuition here is that once the gap-index between most ambiguous (borderline) arms approaches ϵ (Line 6), there is no confusion between the empirically estimated top- m arms and it's closest competitor in the challenger shortlist. Then we employ a greedy selection rule (Réda et al., 2021), where the arm that minimizes the variance between b_t and ch_t is selected. The error in the empirical mean estimates with respect to rewards at instance-level (validation sample) for the non-linear surrogate is computed by the version of loss $\mathcal{L}(\vec{\rho}, \vec{r}) = -\widehat{\text{NDCG}}(\vec{\rho}, \vec{r})$ employing a relaxed version of NDCG metric (Swezey et al., 2021). Here $\vec{r}$ and $\vec{\rho}$ are rewards and empirical mean vectors (across arms equivalent to documents) for a query (validation sample) from D . Then through a Stochastic Gradient Descent (SGD)

step the surrogate is updated to better estimate the empirical mean ((Lines 22-25)). Hence, using the top- m subsets (arms) selection formulation, which aims to clearly separate near-optimal arms from sub-optimal ones, as an auxiliary objective, our approach provides an **efficient mechanism** to learn a differentiable sorting model for instance-level subset ranking objective based on LLM feedback.

3.3 Sample Complexity Theoretical Bounds

Following (Réda et al., 2021), we obtain a high probability $(1 - \delta)$ upper bound on sample complexity of DEARICL which is non-trivial and different from linear MAB variants.

Definition 3.1. (Good Gap indices) $\mathcal{E} \triangleq \bigcap_{t>0} \bigcap_{i,j \in [K]} \left(\rho_i - \rho_j \in [-B_t(j, i), B_t(i, j)] \right)$,

with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ which denotes that a good choice of gap indices $B_t(i, j)$ satisfies event $\mathcal{E}$ with probability greater than or equal to $1 - \delta$. For the above event to hold, it is essential to prove the following bound on pairwise-gap error

Theorem 3.2. In a fixed-confidence setting, $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all pairs $i, j \in \mathcal{A}$:

$$\begin{aligned} & \left| (\hat{\rho}_t(i) - \hat{\rho}_t(j)) - (\rho(i) - \rho(j)) \right| \leq \\ & c_t \sqrt{\frac{2\hat{V}_{ij,t} \log(2K^2/\delta)}{N}} + \epsilon_t^{\text{stab}} + b_i + b_j \\ & \quad + \frac{4M \log(2K^2/\delta)}{3N}. \end{aligned} \quad (3)$$

where $\hat{V}_{ij,t}$ is the empirical variance of MC-dropout differences, for N stochastic predictions (MC dropout forward passes) using set $\mathcal{V}$ $\bar{d}_{ij} := \frac{1}{N} \sum_{k=1}^N (y_i^{(k)} - y_j^{(k)})$, $\hat{V}_{ij,t} := \frac{1}{N} \sum_{k=1}^N (y_i^{(k)} - y_j^{(k)} - \bar{d}_{ij})^2$, ϵ_t^{stab} is the SGD stability error, and b_i, b_j are surrogate approximation biases.

Theorem 3.2 Proof structure: Appendix I.

Theorem 3.3. Sample complexity: For DEARICL, on event $\mathcal{E}$ on which the algorithm is (ϵ, K, δ) -PAC, stopping time τ_δ satisfies $\tau_\delta \leq \inf\{u \in \rho^{*+} : u > 1 + H^\epsilon(\rho) \frac{\log(2K^2/\delta)}{N} + \mathcal{O}(K)\}$, where, for algorithm with the largest variance selection rule²: $H^\epsilon(\rho) \triangleq 18c_t^2 \sum_{a \in [K]} \sigma_{a,t}^2 \cdot \max\left\{\epsilon^{-2}, \left(\frac{\epsilon + \Delta(a)}{3}\right)^{-2}\right\}$.

Theorem 3.3 Proof: On event $\mathcal{E}$, we first demonstrate that the Lemma 3.4 below holds. Then using stopping criterion and Lemma 3.4 we derive the upper bound on sample complexity. The detailed

²or pulling both arms in $\{b_t, c_t\}$ at time t

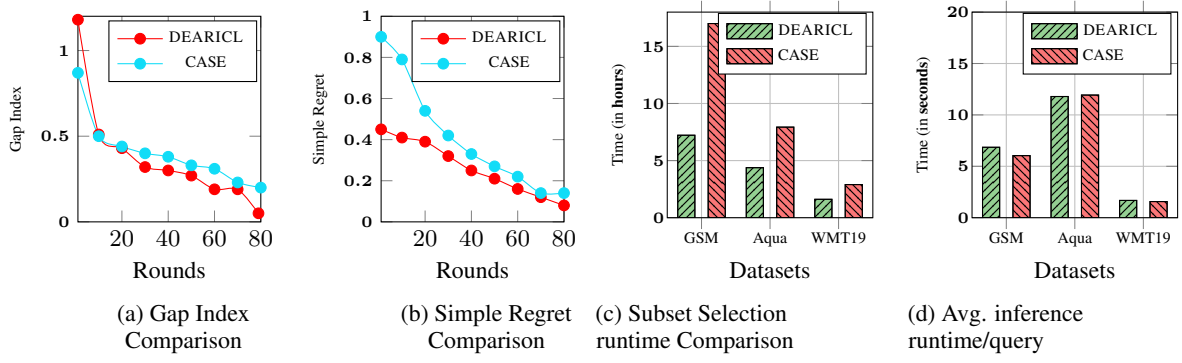


Figure 1: Top- m arm identification by DEARICL, CASE for AquaRAT. (a) Gap Index ($B_t(ch_t, b_t)$) comparison and (b) Simple regret comparison, (c) subset selection time, (d) Average inference time/query.

Method	GSM8K	AquaRat	WMT19
Task level			
Zero-shot-COT (Kojima et al., 2022)	37.37	36.61	51.22
Few-Shot COT (Wei et al., 2022b)	63.22	40.94	61.64
LENS (Li and Qiu, 2023)	64.97	44.88	64.09
EXPLORA (Purohit et al., 2024)	69.92	47.24	66.26
BRIDGE (Wan et al., 2025)	64.52	48.81	64.96
Static CASE (Purohit et al., 2025)	67.00	46.06	65.40
DEARICL_{static} (ours)	69.90	50.00	68.83
Instance Level			
KNN (Rubin et al., 2022)	61.07	41.31	68.36
MMR (Ye et al., 2023b)	66.48	45.84	68.86
PiRank (Swezey et al., 2021)	69.30	37.00	71.30
NeuralNDCG (Pobrotyn and Bialobrzewski, 2021)	72.23	44.09	71.13
ListNet (Cao et al., 2007; Pobrotyn et al., 2020)	70.30	43.30	71.26
UDR (Li et al., 2023)	67.30	44.09	71.22
NDCGLoss 2++ (Wang et al., 2018)	70.00	43.70	72.56
NeuralNDCG with Normalized Data	71.33	45.66	71.80
Instance level (Bandit Approaches)			
<i>CASE_{dynamic}</i> (Purohit et al., 2025)	70.00	47.51	71.70
Instance level (Bandit + LTR)			
DEARICL_{dynamic} (ours)	75.66 ($\uparrow 8.08\%$) [†]	55.11 ($\uparrow 15.99\%$) [†]	78.89 ($\uparrow 10.03\%$) [†]
DEARICL_{dynamic} (-exploration) (ablation)	70.30	37.00	71.30
DEARICL_{dynamic} (random) (ablation)	69.06	49.21	70.38

Table 1: Results using 11ama3.2:3b[†] indicates statistical significance (t-test) over *CASE_{dynamic}* at 0.05 level. We report performance using the official metrics: Exact Match (EM) (AquaRAT, GSM8K) and BertScore (Zhang et al., 2020) (WMT19) for the respective datasets.

proof is available in Appendix J.2. The bound holds for arms in $U_T \cup C_T$. It implies that the top- m arms from $U_T \cup C_T$ are present in U_T with prob. $1 - \delta$, if $T > \tau_\delta$, and K is the size of $U_T \cup C_T$.

Lemma 3.4. *On the event $\mathcal{E}$, for all $t > 0$,*

$$B_t(ch_t, b_t)(t) \leq \min(-(\Delta(b_t) \vee \Delta(ch_t)) + 2W_t(b_t, ch_t), 0) + W_t(b_t, ch_t)$$

, where $a \vee b = \max(a, b)$. In summary, Theorem 3.2 helps support Definition 3.1. Theorem 3.2 establishes a uniform high-probability bound on pairwise gap estimation error, ensuring that the empirical gap between any two arms concentrates around the true gap. This directly justifies Definition 3.1, which defines the “good-gap” event

$\mathcal{E}$ under which all pairwise comparisons are well-behaved. Since event $\mathcal{E}$ follows from Theorem 3.2, conditioned on event $\mathcal{E}$, Lemma 3.4 characterizes how the algorithm’s adaptive confidence radii shrink over time, and Theorem 3.3 then converts these shrinkage properties into a high-probability sample-complexity upper bound for DEARICL which represents the expected number of LLM calls for convergence of auxiliary objective in DEARICL.

4 Experimental Setup

We aim to answer the following research questions:

RQ1: Does DEARICL with auxiliary objective converge faster than state-of-the-art approaches for

Method	GSM8K	AquaRat	WMT19
$CASE_{dynamic}$	1554.19	2378.04	106.54
Neural NDCG	10.59	18.51	5.20
PiRank	6.10	11.59	1.80
DEARICL $_{dynamic}$	6.04	11.59	1.68

Table 2: Inference Time Comparison (in seconds)

top-m ICL sample subsets identification?

RQ2: Does instance-level selection using surrogate from DEARICL with auxiliary objective lead to improved downstream task performance?

RQ3: Can DEARICL lead to improved task performance without sacrificing efficiency ?

4.1 Experimental Setup

Datasets and Metrics: For mathematical reasoning, we use GSM8K and AquaRAT. For demonstrating generalization, we also evaluate on a translation task WMT 2019. Detailed descriptions of the datasets are provided in Appendix B. We report performance using the official metrics: Exact Match (EM) (AquaRAT,GSM8K) and BertScore (Zhang et al., 2020) (WMT19) for the respective datasets. For **reward (LLM feedback)**, we compute BertScore (Zhang et al., 2020) between generated and ground truth rationales, answers for GSM8K and AquaRAT, between translations for WMT19 following (Jauregi Unanue et al., 2021).

LLMs and hyperparameters: We primarily evaluate on relatively stable open-source LLMs like Llama3.2-3b. We also report performance on closed source models like gpt-4o-mini in Appendix E and results using open models like Deepseek-R1:7B (DeepSeek-R1-Distill-Qwen-7B) are shown in Table 7. For all approaches we set max_tokens to predict to 1000 with temperature of 0.25.

Baselines: We compare with existing state-of-the-art example selection algorithms like CASE (Purohit et al., 2025), BRIDGE (Wan et al., 2025), EXPLORA (Purohit et al., 2024) and since our surrogate is based on LTR philosophy we compare with LTR baselines. The number of top arms to be identified was set as $m = |U_t| = 10$ and $|C_t| = 5$. The confidence parameter was fixed at $\delta = 0.05$, controlling the probability of incorrectly identifying the top-m arms. The stopping criterion which is the gap between U_t and N_t was also kept at $\varepsilon = 0.1$. The example subsets ($\mathcal{S}$) are formed, by sampling *with replacement* one example from each of the 5 clusters formed from training set. We use the same hyperparameters in MAB setup for DEARICL for

fair comparison. **Dynamic CASE** ($CASE_{dynamic}$) method is obtained by applying CASE (Purohit et al., 2025), for each test instance instead of single offline run. For instance-level selection using the surrogate (DEARICL $_{dynamic}$) the search space is union of all U_t and C_t shortlists from all rounds of the bandit run. **Learning to Rank baselines** - We compare to diverse LTR approaches including PiRank (Swezey et al., 2021) and learned rankers for ICL like UDR (Li et al., 2023). The same hyperparameter values were also used for all LTR baselines and for DEARICL. Optimization was performed using Adam with learning rate of $1e-4$. The training ran for 100 epochs with a batch size of 16. LTR baselines hyperparameters are in Table 4.

5 Results

5.1 Empirical Verification of Convergence in DEARICL

To answer **RQ1**, we record, compare and analyze the gap-index and simple regret across rounds. Since the stopping criterion is directly dependent on gap-index, it should decrease across rounds with minor fluctuations for convergence. The gap-index across rounds is compared across different algorithms as shown in Figure 1a. We observe that for DEARICL, gap-index decreases gradually and approaches 0 demonstrating that our proposed non-linear ranking surrogate based MAB algorithm converges demonstrating it’s correctness empirically. We also observe that it converges earlier than CASE. While CASE initially shows a monotonically decreasing trend in the gap, it stagnates after round 70 struggling to approach ϵ . We observe that this is primarily due to the model struggling to distinguish between truly good arms and borderline challenger arms that appear to be good. We also observe simple regret as shown in Figure 1b to analyze if the estimate of good arms improves over time as the surrogate better learns to estimate the means (utility) of the arms. For DEARICL, simple regret is calculated as loss of the ranked subsets in U_t with respect to their true ranking based on LLM feedback based rewards. For CASE, it is RMSE between optimal LLM reward and predicted empirical means owing to it’s linear modeling of rewards. We observe that the simple regret of the set U_t - the current estimate of top-m arms decreases gradually. The gap and simple regret for other datasets are reported in Appendix D

5.2 Performance Comparison for Example Subsets Selection

To answer **RQ2**, we compare DEARICL with static (task-level) and learning to rank based dynamic example selection approaches as shown in Table 1. Since DEARICL outputs top- m exemplar subsets as part of auxiliary objective, we compare DEARICL_{static} with other task-level selection based inference approaches. We observe that DEARICL_{static} outperforms existing approaches including CASE across datasets. We hypothesize that this is primarily due to the parameterized non-linear surrogate that models the arm feature-rewards relationship better than CASE and existing approaches. We also employ the non-linear ranking surrogate trained during the selection to rank example subsets dynamically for each test instance with results indicated by DEARICL_{dynamic} in Table 1. Firstly we observe that DEARICL_{dynamic} outperforms static example selection approaches demonstrating need for instance-level selection as static set of examples may not generalize to unseen queries. Also from the table, we observe that DEARICL_{dynamic} significantly outperforms existing approaches like KNN and MMR which aim to retrieve examples based on similarity and diversity to the test example respectively. This is primarily because, DEARICL_{dynamic} learns to rank subsets by taking into consideration the impact of a particular combination of examples on downstream LLM performance through the training process in the bandit optimization step. Whereas KNN and MMR retrieve examples independently without considering how they may interact together and affect LLM performance. DEARICL_{dynamic} also outperforms dynamic version of the CASE, which employs a bandit-based selection algorithm per instance and scores subsets as a whole. For instance, DEARICL_{dynamic} achieves upto **15.99%** over dynamic CASE on AquaRAT. We observe that the improvements are from better modeling of reward structure and auxiliary objective leading to clear separation of borderline and top- m arms.

Comparing to LTR approaches, from Table 1, we observe that DEARICL_{dynamic} outperforms existing LTR approaches trained with different objectives. The LTR approaches are also trained to rank example subsets as a whole for fair comparison and demonstrate gains over static selection approaches. However, we observe that the mechanism to iteratively fit the non-linear surrogate in our approach

helps the model clearly distinguish between borderline arms and top- m arms through sampling of most ambiguous arms and reduction of the gap between them. However, in classical LTR training approaches, there is no principled mechanism to sample ambiguous borderline arms and only adopt heuristic negative sampling without considering downstream task performance, leading to subpar performance: ablation in Table 1. We also perform an ablation where we replace the proposed gap-index based exploration with random exploration (DEARICL_{dynamic} (random)). The performance comparison demonstrates that our proposed gap-index exploration leads to better performance compared to random exploration. A qualitative analysis of selected ICL demonstrations is discussed in Appendix H.

Total # LLM calls	GSM8K	AquaRat	WMT19
CASE	10200	4360	87
DEARICL _{static}	2600	1580	48

Table 3: Results comparing the number of LLM calls.

5.3 Efficiency Comparison

To answer **RQ3**, we compare the convergence time of CASE and DEARICL (Figure 1c), inference time across the three datasets. We observe that convergence time of DEARICL is less than CASE (providing approximately **2x** speedup on AquaRAT and **2.35 x** speedup on GSM8K) for identifying top- m arms. We observe that this is primarily due to **faster convergence** of DEARICL based on stopping criterion. This is because the non-linear ranking-based surrogate in DEARICL quickly learns to distinguish between top- m arms and borderline challenger arms by sampling better ambiguous arms through auxiliary objective. This is also evident from comparison of gap-index across rounds between CASE and DEARICL in Figure 1a. For instance, in AquaRAT, DEARICL converges in **79 rounds**, whereas CASE requires **238 rounds** and for GSM8k, DEARICL requires **130 rounds** but CASE requires **510 rounds**. We also observe the sample-efficiency of DEARICL from Table 3, where DEARICL significantly reduces the number of LLM calls (by **63-75%** approx.) during the top- m selection using an auxiliary objective (offline) compared to prior approaches like CASE.

We also plot the average inference times per query on test set. Particularly, we compare

DEARICL_{dynamic} with static CASE to measure the latency overhead of surrogate ranker at inference time in DEARICL_{dynamic}. We observe from Figure 1d, that the latency overhead is negligible with DEARICL_{dynamic}, only a few milliseconds over static CASE. Additionally, we also compare the average runtime/query of DEARICL_{dynamic} with other approaches (Table 2). CASE_{dynamic} applies bandit selection at test time, resulting in high runtime. In contrast, DEARICL_{dynamic} is more efficient due to its lightweight ranker, avoiding extra inference overhead unlike NeuralNDCG. PiRank has similar runtime (similar architecture) but performs worse. All static methods (EXPLORA, etc.) have same runtime as CASE (Figure 1d)

6 Conclusion

In this work, we propose a sample-efficient gap-index based MAB framework (DEARICL) as auxiliary objective to fit a non-linear surrogate that models example subset scores to enhance ICL. DEARICL provides a mechanism to learn ranking surrogates from LLM feedback for instance-level example subsets selection. DEARICL converges faster than existing MAB algorithms. It adds only negligible overhead during inference for instance-level setting. The proposed algorithm can also be extended to other ranking tasks in the future with appropriate loss function changes.

7 Acknowledgments

The experiments 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 (proposal NAISS 2025/5-631) and Zeus cluster provided by the Department of Computer and Systems Sciences at Stockholm University.

8 Limitations

While our approach proposes a novel bandit framework for enhancing In-context learning, accounting for complex non-linear reward relationships, it has a few limitations. Firstly, the task relies on supervised question-answer pairs which may not be available for all domains. However, the feature design in our current algorithm can be easily modified to first select a subset of questions and augment them with answers using LLMs or other data augmentation mechanisms. Since this is not in the scope

of the current work, we defer this for future work. Additionally, while we provide sample-complexity bounds as our core objective is to reduce the number of LLM calls for efficiency, regret bounds could also be useful. While we do currently demonstrate empirically convergence of regret and gap-index, theoretical regret bounds are hard to derive for gap-index frameworks. We would refer to existing gap-index frameworks for linear bandits (Réda et al., 2021), which also do not focus on regret bounds and only on sample-complexity. Since we propose a new gap-index framework for non-linear stochastic settings, this required the development of a new theoretical framework for sample-complexity, as shown in the paper, and hence our current work does not focus on regret bounds. However, we believe that our algorithm is useful for settings beyond ICL and can be extended by the community with regret bounds and new advances. Additionally, our framework could be extended to other relaxed versions of learning-to-rank losses, which is also an avenue for further research.

9 Ethical Considerations and Risks

We use only publicly available math reasoning/machine translation datasets that do not contain private or harmful information. While we employ LLMs for QA systems in our experiments, we do not prompt them in a format that would elicit harmful or biased information.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. 2013. [Multiple identifications in multi-armed bandits](#). In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 258–265. JMLR.org.
- Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. [Learning to rank: from pairwise approach to listwise approach](#). In *Machine Learning*,

- Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 129–136. ACM.
- Lijie Chen, Jian Li, and Mingda Qiao. 2017. [Nearly instance optimal sample complexity bounds for top-k arm selection](#). In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 101–110. PMLR.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#). *Trans. Mach. Learn. Res.*, 2023.
- Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. 2020. [Gamification of pure exploration for linear bandits](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2432–2442. PMLR.
- Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian J. Ratliff. 2019. [Sequential experimental design for transductive linear bandits](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10666–10676.
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. [Complexity-based prompting for multi-step reasoning](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. 2012. [Best arm identification: A unified approach to fixed budget and fixed confidence](#). In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 3221–3229.
- Artem Grotov and Maarten de Rijke. 2016. [Online learning to rank for information retrieval: SIGIR 2016 tutorial](#). In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1215–1218. ACM.
- Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. [Deepcore: A comprehensive library for coreset selection in deep learning](#). In *Database and Expert Systems Applications - 33rd International Conference, DEXA 2022, Vienna, Austria, August 22-24, 2022, Proceedings, Part I*, volume 13426 of *Lecture Notes in Computer Science*, pages 181–195. Springer.
- Moritz Hardt, Ben Recht, and Yoram Singer. 2016. [Train faster, generalize better: Stability of stochastic gradient descent](#). In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1225–1234. JMLR.org.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. [BERTTune: Fine-tuning neural machine translation with BERTScore](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 915–924, Online. Association for Computational Linguistics.
- Yassir Jedra and Alexandre Proutière. 2020. [Optimal best-arm identification in linear bandits](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. 2012. [PAC subset selection in stochastic multi-armed bandits](#). In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.
- Sumeet Katariya, Ardhendu Tripathy, and Robert D. Nowak. 2019. [Maxgap bandit: Adaptive algorithms for approximate ranking](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11045–11055.
- Emilie Kaufmann and Shivaram Kalyanakrishnan. 2013. [Information complexity in bandit subset selection](#). In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 228–251. JMLR.org.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Tor Lattimore and Csaba Szepesvári. 2020. [Bandit Algorithms](#). Cambridge University Press.

- Shuai Li, Tor Lattimore, and Csaba Szepesvári. 2019. [Online learning to rank with features](#). In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3856–3865. PMLR.
- Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. [Unified demonstration retriever for in-context learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.
- Xiaonan Li and Xipeng Qiu. 2023. [Finding support examples for in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pages 6219–6235. Association for Computational Linguistics.
- Zhaoqi Li, Kevin Jamieson, and Lalit Jain. 2024. [Optimal exploration is no harder than thompson sampling](#). In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 1684–1692. PMLR.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. [Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Andreas Maurer and Massimiliano Pontil. 2009. [Empirical bernstein bounds and sample-variance penalization](#). In *Annual Conference Computational Learning Theory*.
- Przemyslaw Pobrotyn, Tomasz Bartczak, Mikolaj Synowiec, Radoslaw Bialobrzeski, and Jaroslaw Bojar. 2020. [Context-aware learning to rank with self-attention](#). *CoRR*, abs/2005.10084.
- Przemyslaw Pobrotyn and Radoslaw Bialobrzeski. 2021. [NeuralIndcg: Direct optimisation of a ranking metric via differentiable relaxation of sorting](#). *CoRR*, abs/2102.07831.
- Kiran Purohit, Venkatesh V, Sourangshu Bhattacharya, and Avishek Anand. 2025. [Sample efficient demonstration selection for in-context learning](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.
- Kiran Purohit, Venkatesh V, Raghuram Devalla, Krishna Mohan Yerragorla, Sourangshu Bhattacharya, and Avishek Anand. 2024. [EXPLORA: Efficient exemplar subset selection for complex reasoning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5388, Miami, Florida, USA. Association for Computational Linguistics.
- Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. 2021. [Top-m identification for linear bandits](#). In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1108–1116. PMLR.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.
- Robin M. E. Swezey, Aditya Grover, Bruno Charron, and Stefano Ermon. 2021. [Pirank: Scalable learning to rank via differentiable sorting](#). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21644–21654.
- Jonathan Tonglet, Manon Reusens, Philipp Borchert, and Bart Baesens. 2023. [SEER : A knapsack approach to exemplar selection for in-context hybridqa](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13569–13583. Association for Computational Linguistics.
- Xingchen Wan, Han Zhou, Ruoxi Sun, and Sercan Ö. Arik. 2025. [From few to many: Self-improving many-shot reasoners through iterative optimization and generation](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2018. [The lambdaloss framework for ranking metric optimization](#). In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 1313–1322, New York, NY, USA. Association for Computing Machinery.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy

- Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *Trans. Mach. Learn. Res.*, 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng Yang, Qingxin Cao, Haiming Wang, Xiongwei Han, Jing Tang, Chengming Li, and Xiaodan Liang. 2024. [Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Liyuan Xu, Junya Honda, and Masashi Sugiyama. 2018. [A fully adaptive algorithm for pure exploration in linear bandits](#). In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 843–851. PMLR.
- Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. [Compositional exemplars for in-context learning](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.
- Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. [Complementary explanations for effective in-context learning](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pages 4469–4484. Association for Computational Linguistics.
- Jianfei Zhang, Bei Li, Jun Bai, Rumei Li, Yanmeng Wang, Chenghua Lin, and Wenge Rong. 2025. [Selecting demonstrations for many-shot in-context learning via gradient matching](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11686–11704, Vienna, Austria. Association for Computational Linguistics.
- Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2024. [Trained transformers learn linear models in-context](#). *J. Mach. Learn. Res.*, 25:49:1–49:55.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. [Active example selection for in-context learning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Masrour Zoghi, Tomás Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvári, and Zheng Wen. 2017. [Online learning to rank in stochastic click models](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 4199–4208. PMLR.

A Appendix

B Dataset Description

AquaRAT: It comprises 100,000 algebraic word problems in the train set with dev and test set each comprising 254 problems. The problems are provided along with answers and rationales providing the step-by-step solution to the problem.

GSM8K: This dataset consists of linguistically diverse math problems that require multi-step reasoning. The dataset consists of 8.5K problems and we evaluate on the test set of 1319 questions.

WMT 19: We focus on en-zh (english-chinese) translation split. Test sets are a few thousand sentences (for example, 3,981 in WMT18 for zh-en direction for test). Train set has 173k english-chinese sentence pairs.

C LTR Hyperparameters

The hyperparameters for learning to rank baselines are detailed in Table 4.

D Gap and simple regret analysis for GSM8K and WMT19

Similar to the gap index and simple regret analysis on AquaRAT in Section 5.1, we also plot the gap indices and simple regret across rounds for GSM8K and WMT 2019 as shown in Figure 2. We observe a similar trend where gap index between ambiguous arms approaches ϵ , the stopping criterion and the regret also minimizes across rounds demonstrating the empirical convergence of DEARICL.

E Results on alternative LLMs

We also evaluate static and dynamic versions of our proposed approach DEARICL using alternative LLMs like gpt-4o-mini and Deepseek-R1:7B (DeepSeek-R1-Distill-Qwen-7B). We choose these LLMs as it has shown relatively stable performance across benchmarks. We extract the most competitive baselines and method from Table 1 and evaluate them using gpt-4o-mini and Deepseek-R1:7B (DeepSeek-R1-Distill-Qwen-7B). All our experiments are carried out in a transfer setting where exemplars selected using Llama3.2:3b in the optimization loop are employed directly for inference on test set using gpt-4o-mini or Deepseek. The results using gpt-4o-mini are as shown in Table 6 and results for and Deepseek-R1:7B (DeepSeek-R1-Distill-Qwen-7B) are shown in Table 7.

F Data ablation and extension to other LTR loss functions

We also perform data ablations where we train the LTR approaches on training data collected across rounds by DEARICL to isolate the effect of our auxiliary objective tailored to our setup. The results are shown in Table 5. We observe that when compared to Table 1, the performance of some baselines show improvement on some benchmarks. But they still fall behind DEARICL_{dynamic}.

Though training on quality subsets in logged data from DEARICL_{dynamic} helps improve performance, they miss the benefit of the online adaptive MAB outer loop, optimizing for auxiliary objective. Note that our proposed learning process uses utility from the surrogate, which is also continuously updated from feedback and based on a specific ranking loss. Hence, the borderline arms are adaptively sampled based on the gap-index, which depends on the estimated utility from surrogate.

Our framework is agnostic to the surrogate architecture: any non-linear scorer can be substituted in the inner loop. We use a lightweight architecture because the ranker is invoked per test instance at inference, where it delivers a runtime advantage over per-instance bandit search. Our choice of lightweight architecture additionally keeps the network’s Lipschitz constant controlled. The one component that interacts with our theory is the ranking loss. Our sample-complexity analysis is derived via uniform stability, which requires the inner-loop surrogate loss to be L -Lipschitz ; Lemma I.1 establishes this property for our relaxed-NDCG formulation. A stability-controlled loss enables reliable learning from the small, adaptively collected sample that the gap-index acquisition produces. Other LTR objectives can be substituted in the inner loop empirically; however, carrying the sample-complexity related guarantees over requires establishing the analogous constants for each loss (beyond scope of this work and is reserved for future work), which is why our analysis is scoped to relaxed-NDCG, and extending Lemma I.1’s analysis to further losses is a natural follow-up work.

G Dataset prompts

The prompts are given in Figures 3, 4. The prompts for WMT19 are in the github repo.

Framework	Loss function	Hyperparameters	Network architecture
PiRank	PiRank surrogate loss	Presented in Section 4	(256, 256, 128, 64)
Other LTR baselines	Neural NDCG	$N = 2$, $d_{ff} = 384$, $h = 1$, dropout = 0.1	(768, 96)
	ListNet	$N = 4$, $d_{ff} = 512$, $h = 2$, dropout = 0.3	(768, 128)
	Neural NDCG	$N = 4$, $d_{ff} = 512$, $h = 4$, dropout = 0.3	(768, 96)
	Neural NDCG With Normalized data	$N = 2$, $d_{ff} = 384$, $h = 1$, dropout = 0.1	(768, 96)
	NDCGLoss 2++	-	(256, 512, 1024, 512, 256)

Table 4: Details of hyperparameters used in different LTR model configurations. Categorized by loss function and framework.

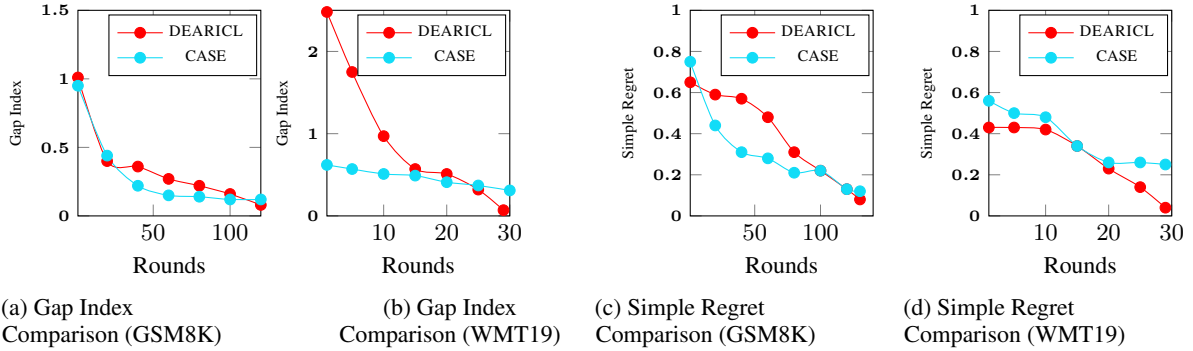


Figure 2: Top- m arm identification by DEARICL, CASE for GSM8k, WMT19. (a,b) Gap Index ($B_t(ch_t, b_t)$) comparison and (c,d) Simple regret comparison

Task performance	GSM8K	AquaRat	WMT19
NDCGLoss 2++	72.11	49.00	74.06
NeuralNDCG	69.83	48.30	72.75
NeuralNDCG (norm. data)	72.06	49.21	72.97
UDR	71.63	48.03	72.65
ListNet	71.79	46.06	73.64
DEARICL_{dynamic}	75.66	55.11	78.89

Table 5: Ablation results: Training LTR approaches on training data collected across rounds by DEARICL leveraging llama3.2:3b

such example is shown in Table 8. The remaining selected ICL samples are provided in the GitHub repo. The test instance is a Profit, Loss & Interest problem that requires algebra. Subset selected by DEARICL_{dynamic} contains three ICL samples (1, 2, 4) that contain the required skills to solve the test instance, whereas NeuralNDCG contains none.

H Qualitative Analysis

We also analyzed the quality of a few demonstration samples selected for the test instances. One

I Proof for Theorem 3.2

Notation and setup. Let $\mathcal{A}$ be the arm set, $|\mathcal{A}| = K$. At round t the PiRank surrogate (with dropout) outputs N stochastic predictions (N MC-dropout

Method	GSM8K	AquaRat	WMT
Task level			
EXPLORA (Purohit et al., 2024)	93.63	69.29	84.55
LENS (Li and Qiu, 2023)	76.19	64.56	83.57
Static CASE	91.13	73.23	83.49
DEARICL _{static}	94.84	77.16	86.25
Instance level			
CASE _{dynamic} (Purohit et al., 2025)	92.19	76.77	86.36
DEARICL _{dynamic}	94.66	81.88 †	92.41 †

Table 6: Results across datasets (we use 5-shot for all methods) using gpt-4o-mini. † indicates statistical significance (t-test) over CASE_{dynamic} at 0.05 level

Method	GSM8K	AquaRat	WMT
Task level			
EXPLORA (Purohit et al., 2024)	82.63	68.10	78.59
LENS (Li and Qiu, 2023)	77.33	57.87	76.98
Static CASE	83.09	69.29	78.26
DEARICL _{static}	86.12	70.47	79.27
Instance level			
CASE _{dynamic} (Purohit et al., 2025)	85.98	69.68	79.78
DEARICL _{dynamic}	88.40 †	74.41 †	81.50

Table 7: Results across datasets using Deepseek-R1:7b (we use 5-shot for all methods). † indicates statistical significance (t-test) over CASE_{dynamic} at 0.05 level

Question	A travel company wants to charter a plane to the Bahamas. Chartering the plane costs \$5,000. So far, 12 people have signed up for the trip. If the company charges \$200 per ticket, how many more passengers must sign up for the trip before the company can make any profit on the charter?
Method	Exemplars
NeuralNDCG	Question: A group consists of 4 couples in which each of the 4 boys has one girlfriend. In how many ways can they be arranged in a straight line such that boys and girls occupy alternate positions? Question: A man cycling along the road noticed that every 12 minutes a bus overtakes him and every 4 minutes he meets an oncoming bus. If all buses and the cyclist move at a constant speed, what is the time interval between consecutive buses? Question: Find large number from below question. The difference of two numbers is 1365. On dividing the larger number by the smaller, we get 6 as quotient and the 15 as remainder? Question: 64, 81, 100, 144, 196, ?, Find the missing number(?).
DEARICL_{dynamic}	Question: A certain sum is invested at simple interest at 18% p.a. for two years instead of investing at 12% p.a. for the same time period. Therefore the interest received is more by Rs. 840. Find the sum? Question: An investor can sell her MicroTron stock for 36\$ per share and her Dynaco stock for 60\$ per share. If she sells 300 shares altogether, some of each stock, at an average price per share of 40\$, how many shares of Dynaco stock has she sold? Question: Anna left for city A from city B at 5.20 a.m. She traveled at the speed of 80 km/hr for 2 hrs 15 min. After that the speed was reduced to 60 km/hr. If the distance between two cities is 350 kms, at what time did Anna reach city? Question: On a sum of money, the S.I. for 2 years is Rs. 660, while the C.I. is Rs. 696.30, the rate of interest being the same in both the cases. The rate of interest is?

Rationale is not shown to conserve space. However, in our experiments, all exemplars for AquaRat include rationales. The test instance is a Profit, Loss & Interest problem that requires algebra. Subset selected by DEARICL_{dynamic} contains three similar questions(1, 2,.4.), whereas NeuralNDCG contains none.

Table 8: Exemplar subsets selected by NeuralNDCG and DEARICL_{dynamic} using LLama3.2:3b for a particular AquaRat test instance where DEARICL_{dynamic} had a correct answer and the NeuralNDCG model did not.

AQUA Prompt

Instruction: You are a helpful, respectful, and honest assistant helping to solve math word problems or tasks requiring reasoning or math. Follow given examples and solve the problems in step by step manner.

Exemplars :

[Question]: *The average age of three boys is 45 years and their ages are in proportion 3:5:7. What is the age in years of the youngest boy?*

[Options]: A) 9, B) 10, C) 11, D) 12, E) 13

[Explanation]: $3x + 5x + 7x = 45$,

$x = 3$,

$3x = 9$

[Answer]: **The option is A**

...

...

Test Input : Question: Options:

Explanation: [INS] Answer: [INS]

Figure 3: Prompt for Aqua

GSM8K Prompt

Instruction: You are a helpful, respectful and honest assistant helping to solve math word problems or tasks requiring reasoning or math. Follow given examples and solve the problems in step by step manner.

Exemplars :

[Question]: *Samir just turned half the age Hania was 10 years ago. If in five years Hania will be 45 years old, what will Samir's age be five years from now?*

[Explanation]: If in five years, Hania will be 45 years old, currently she is $45 - 5 = 40$ years old. Samir just turned half the age Hania was 10 years ago, which means she is $30/2 = 15$ years old. In five years, Samir will be $15 + 5 = 20$ years old.

[Answer]: **20 years old**

...

...

Test Input : Question:

Explanation: [INS] Answer: [INS]

Figure 4: Prompt for GSM8K

forward passes) per arm a :

$$\{y_a^{(k)}\}_{k=1}^N, \quad \hat{\rho}_t(a) := \frac{1}{N} \sum_{k=1}^N y_a^{(k)}.$$

Define the sample mean difference and sample variance for pair (i, j) :

$$\bar{d}_{ij} := \frac{1}{N} \sum_{k=1}^N (y_i^{(k)} - y_j^{(k)}),$$

$$\hat{V}_{ij,t} := \frac{1}{N} \sum_{k=1}^N (y_i^{(k)} - y_j^{(k)} - \bar{d}_{ij})^2.$$

Let c_t be an annealed multiplier (user-specified) and define

$$B_t(ch_t, b_t) := \hat{\rho}_t(ch_t) - \hat{\rho}_t(b_t) + W_t(b_t, ch_t).$$

In order to prove Lemma 3.4, we first need to demonstrate that event $\mathcal{E}$ hold with probability $\geq 1 - \delta$ for DEARICL. However, the estimated means $\hat{\rho}_t$ and $W_t(i, j)$ for any two arms i, j is computed in a different manner for DEARICL than existing linear stochastic bandit frameworks like CASE and GIFA. Hence, we establish that the confidence event $\mathcal{E}$ holds here by deriving an upper bound on pairwise-gap error as explained below. This is one of our main **contributions** which further helps in deriving a high complexity upper bound on sample complexity.

To recap the event is defined as, $\mathcal{E} \triangleq \bigcap_{t>0} \bigcap_{i,j \in [K]} (\rho_i - \rho_j \in [-B_t(j, i), B_t(i, j)])$, Expanding the event $\mathcal{E}$

$$\rho_i - \rho_j \geq (\hat{\rho}_i(i) - \hat{\rho}_i(j)) - W_t(i, j)$$

and symmetrically,

$$\rho_i - \rho_j \leq (\hat{\rho}_i(i) - \hat{\rho}_i(j)) + W_t(i, j)$$

Hence it follows,

$$(\hat{\rho}_i(i) - \hat{\rho}_i(j)) - (\rho_i - \rho_j) \leq W_t(i, j)$$

Let

$$\mathbb{E}_{ij}(t) := (\hat{\rho}_t(i) - \hat{\rho}_t(j)) - (\rho(i) - \rho(j))$$

denote the pairwise-gap error. This error can be further decomposed as follows based on the source of randomness / errors

$$\begin{aligned}
\mathbb{E}_{ij}(t) = & \underbrace{\frac{[(\hat{\rho}_t(i) - \hat{\rho}_t(j)) - (\mathbb{E}_{\text{drop}}[y_i] - \mathbb{E}_{\text{drop}}[y_j])]}{2}}_{\text{(Term 1) MC Dropout noise}} \\
& + \underbrace{\frac{[(\mathbb{E}_{\text{drop}}[y_i] - \mathbb{E}_{\text{weights,drop}}[y_i]) - (\mathbb{E}_{\text{drop}}[y_j] - \mathbb{E}_{\text{weights,drop}}[y_j])]}{2}}_{\text{(Term 2) Weight randomness / SGD-induced drift}} \\
& + \underbrace{\frac{[(\mathbb{E}_{\text{weights,drop}}[y_i] - \rho(i)) - (\mathbb{E}_{\text{weights,drop}}[y_j] - \rho(j))]}{2}}_{\text{(Term 3) Model bias}}
\end{aligned}$$

Each of the above terms can be bounded individually to prove the following theorem. Restating Theorem 3.2,

Theorem (MC-Dropout gap concentration). In a fixed-confidence setting, $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all pairs $i, j \in \mathcal{A}$:

$$\begin{aligned}
& |(\hat{\rho}_t(i) - \hat{\rho}_t(j)) - (\rho(i) - \rho(j))| \\
& \leq c_t \sqrt{\frac{2\hat{V}_{ij,t} \log(2K^2/\delta)}{N}} + \varepsilon_t^{\text{stab}} + b_i + b_j \\
& \quad + \frac{4M \log(2K^2/\delta)}{3N}.
\end{aligned}$$

where $\hat{V}_{ij,t}$ is the empirical variance of MC-dropout differences, $\varepsilon_t^{\text{stab}}$ is the SGD stability error, and b_i, b_j are surrogate approximation biases.

Term 1 - Monte-Carlo Dropout Noise

For arms i, j we define Monte Carlo Dropout samples of their differences as:

$$Z_k = (y_i(k) - y_j(k)) - (\mathbb{E}_{\text{drop}}[y_i] - \mathbb{E}_{\text{drop}}[y_j])$$

Our goal is to bound the empirical mean $\hat{Z} = \frac{1}{N} \sum_{k=1}^N Z_k$ which is equivalent to Term 1.

Assuming Z_k 's are independent (as dropout masks are independent) and bounded, we apply Bernstein's inequality as it states that,

For $\{Z_k\}_{k=1}^N$ be independent, mean-zero random variables with $|Z_k| \leq b$ almost surely, and let

$$\hat{Z}_N := \frac{1}{N} \sum_{k=1}^N Z_k, \quad \sigma^2 := \text{Var}(Z_k).$$

Then, for any $\epsilon > 0$, the (scalar) Bernstein inequality states: if $Z_1, \dots, Z_N$ are independent, mean-zero, and satisfy $|Z_k| \leq b$, then for any $\epsilon > 0$,

$$\Pr(|\hat{Z}| \geq \epsilon) \leq 2 \exp\left(-\frac{N\epsilon^2}{2\sigma^2 + \frac{2}{3}b\epsilon}\right).$$

Here $b = 2M$ and $\sigma^2 = \text{Var}(y_i - y_j)$.

To get a confidence radius ϵ such that the event holds with probability at least $1 - \delta$ we set

$$2 \exp\left(-\frac{N\epsilon^2}{2\sigma^2 + \frac{2}{3}b\epsilon}\right) = \delta.$$

This yields the inequality:

$$2 \left(-\frac{N\epsilon^2}{2\sigma^2 + \frac{2}{3}b\epsilon}\right) \geq \log \delta.$$

hence,

$$\left(\frac{N\epsilon^2}{2\sigma^2 + \frac{2}{3}b\epsilon}\right) \geq \log(2/\delta).$$

Following (Maurer and Pontil, 2009; Lattimore and Szepesvári, 2020) we aim to solve the inequality:

$$\left(\frac{N\epsilon^2}{2\sigma^2 + \frac{2}{3}b\epsilon}\right) \geq \log(2/\delta).$$

This can be expressed as a quadratic in ϵ ,

$$N\epsilon^2 - \frac{2}{3}b\epsilon(\log(2/\delta)) - 2\sigma^2 \log(2/\delta) \geq 0$$

We are interested in the smallest positive ϵ (denoted ϵ^*) that makes the Bernstein exponent achieve the desired log level. Now, solving the quadratic equation and using this principle we get,

$$\epsilon^* = \frac{\frac{2}{3}b \log(2/\delta) + \sqrt{(\frac{2}{3}b \log(2/\delta))^2 + 8N\sigma^2 \log(2/\delta)}}{2N}. \quad (4)$$

To obtain an upper bound for square root term we use the inequality,

$$\sqrt{a^2 + x} \leq a + \sqrt{x}, \text{ where } a = (\frac{2}{3}b \log(2/\delta)) \text{ and } x = 8N\sigma^2 \log(2/\delta)$$

hence,

$$\begin{aligned}
& \sqrt{\left(\frac{2}{3}b \log(2/\delta)\right)^2 + 8N\sigma^2 \log(2/\delta)} \leq \\
& \left(\frac{2}{3}b \log(2/\delta)\right) + \sqrt{8N\sigma^2 \log(2/\delta)}
\end{aligned}$$

Using above in Equation 4

$$\epsilon^* \leq \frac{\frac{4}{3}b \log(2/\delta) + \sqrt{8N\sigma^2 \log(2/\delta)}}{2N}$$

$$\epsilon^* \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{N}} + \frac{2b \log(2/\delta)}{3N}.$$

Thus, with probability at least $1 - \delta$, (recall that generally $\Pr(|\bar{Z}| \leq \epsilon) \geq 1 - \delta$)

$$|\bar{Z}| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{N}} + \frac{4M \log(2/\delta)}{3N}.$$

Replacing σ^2 by the empirical variance:
Define the sample variance estimator

$$\begin{aligned} \hat{V}_{ij,t} &:= \frac{1}{N} \sum_{k=1}^N \left[(y_i^{(k)} - y_j^{(k)}) - \bar{d}_{ij} \right]^2, \\ \bar{d}_{ij} &:= \frac{1}{N} \sum_{k=1}^N (y_i^{(k)} - y_j^{(k)}). \end{aligned}$$

Then by concentration of empirical variance (via Bernstein or Bennett bounds), $\hat{V}_{ij,t}$ is close to σ^2 with high probability, so we may plug $\hat{V}_{ij,t}$ into the bound:

$$|\bar{Z}| \leq \sqrt{\frac{2\hat{V}_{ij,t} \log(2/\delta)}{N}} + \frac{4M \log(2/\delta)}{3N}.$$

It is to be noted that we interpret the MC-dropout variance as solely as a variance estimator and we derive concentration bounds as standard in bandit analysis.

Union bound over all pairs. We require the inequality to hold for all pairs $(i, j) \in \mathcal{A}$ simultaneously. Since there are at most K^2 ordered pairs, set

$$\delta = \frac{\delta}{K^2}.$$

By a union bound, with probability at least $1 - \delta$,

$$\begin{aligned} \forall i, j \in \mathcal{A} : & \left| (\hat{\rho}_t(i) - \hat{\rho}_t(j)) - (\mathbb{E}_{\text{drop}}[y_i] - \mathbb{E}_{\text{drop}}[y_j]) \right| \\ & \leq W_t(i, j). \end{aligned}$$

This defines the desired pairwise confidence width $W_t(i, j)$ under Monte Carlo dropout.

Term 2 — Weights Randomness / SGD Stability

We need to bound

$$\left| \mathbb{E}_{\text{drop}}[y_a] - \mathbb{E}_{\text{weights, drop}}[y_a] \right|,$$

that is, the gap between the conditional dropout mean (given current weights trained on the dataset)

and the expectation over randomness in the training set and weights.

To do this it is first essential that the predictions of the differentiable sorting surrogate does not deviate a lot in each round of arm sampling. This translates to proving that for one-epoch SGD the *uniform stability* criterion holds.

This criterion guarantees that a randomized algorithm is uniformly stable, if for all data sets differing in only one element, the learned models produce nearly the same predictions. This is applicable to our setup, as in each round after sampling reward from a arm, this new sample (arm+reward) is added to the training set to update the non-linear surrogate with SGD simulating a single epoch of NN training.

For one-epoch SGD, (Hardt et al., 2016) show uniform stability bounds of the form

$$\sup_z \left| \ell(\text{SGD}(S), z) - \ell(\text{SGD}(S^{(i)}), z) \right| \leq \epsilon_t^{\text{stab}},$$

, where S and $S^{(i)}$ differ in atmost one data sample and $\ell(\text{SGD}(S^{(i)}), z)$ denotes the loss of the randomized algorithm (Diffsort neural network), which can be translated to a bound on predictions. Under our conditions (bounded gradients G and **Lipschitz loss L - as shown in Lemma I.1**), one-epoch SGD has stability that decays with the dataset size and step size; we encapsulate this as ϵ_t^{stab} .

Concretely, there exist constants (depending on G, L, η_t) such that $\forall a$:

$$\left| \mathbb{E}_{\text{dropout}}[y_a \mid \text{weights}] - \mathbb{E}_{\text{weights, dropout}}[y_a] \right| \leq \epsilon_t^{\text{stab}}.$$

Hence for the pair (i, j) , the contribution is at most $2\epsilon_t^{\text{stab}}$; we absorb a factor of 2 into the constant and state the theorem with one ϵ_t^{stab} representing the pairwise bound (or keep $+\epsilon_t^{\text{stab}}$ per side — we used one in the statement for brevity).

Intuition: Because we train only one epoch per new sample, the model is only mildly unstable: removing or adding one sample cannot arbitrarily change predictions. That bounded change becomes a bias term in the final gap bound.

Term 3 - Model Bias :

The per arm bias can be defined as:

$$b_a := \left| \mathbb{E}_{\text{weights, drop}}[y_a] - \rho(a) \right|.$$

By the triangle inequality,

$$\begin{aligned} & |\mathbb{E}_{\text{weights, drop}}[y_i] - \rho(i)| - \\ & |\mathbb{E}_{\text{weights, drop}}[y_j] - \rho(j)| \\ & \leq |\mathbb{E}_{\text{weights, drop}}[y_i] - \rho(i)| + \\ & |\mathbb{E}_{\text{weights, drop}}[y_j] - \rho(j)| = b_i + b_j \end{aligned}$$

Hence,

$$\begin{aligned} & |\mathbb{E}_{\text{weights, drop}}[y_i] - \rho(i)| - \\ & |\mathbb{E}_{\text{weights, drop}}[y_j] - \rho(j)| \\ & \leq b_i + b_j \end{aligned}$$

This implies that the model-bias contribution to the pairwise error is at most the sum of the two per-arm biases. If the surrogate is well-specified $b_a = 0$.

Where empirically, b_a can be computed as the deviation in empirical mean of the arm over the rounds with respect to a moving average of estimated empirical means over past rounds.

Combining bounds for Term 1, 2 and 3 yields the expression for W_t in Theorem 3.2. The event $\mathcal{E}$ holds for this W_t which is one of our **main theoretical contributions**. Then the proof for Lemma 3.4 follows from (Réda et al., 2021) which yields a high probability upper bound on sample complexity as stated in Theorem 3.3. We include proof of Lemma 3.4 in Appendix J.1 for completion. We then derive the **high probability upper bound on sample complexity** In Appendix J.2.

Lemma I.1. Let $\text{NDCG}_\tau(\hat{\rho}, \mathbf{r})$ denote the differentiable (temperature- τ) relaxation of NDCG obtained via a soft permutation matrix $P_\tau(\hat{\rho})$. Assume: (i) $\|\mathbf{r}\|_\infty \leq R_{\max}$, (ii) the discount function $D(\cdot)$ is bounded by $D_{\max}$, (iii) P_τ is L_P -Lipschitz with $L_P = O(1/\tau)$. Then the loss $L(\hat{\rho}, \mathbf{r}) = -\widehat{\text{NDCG}}_\tau(\hat{\rho}, \mathbf{r})$ is L -Lipschitz in $\hat{\rho}$ with

$$L \leq \frac{R_{\max} D_{\max}}{Z(\mathbf{r})} L_P = O\left(\frac{1}{\tau}\right).$$

Proof. The relaxed NDCG can be written as

$$\widehat{\text{NDCG}}_\tau(\hat{\rho}, \mathbf{r}) = \frac{1}{Z(\mathbf{r})} \sum_{i,j} P_\tau(i, j; \hat{\rho}) r_j D(i),$$

which is linear in P_τ . Let $\hat{\rho}_1, \hat{\rho}_2$ be arbitrary. Then

$$\begin{aligned} & |\widehat{\text{NDCG}}_\tau(\hat{\rho}_1, \mathbf{r}) - \widehat{\text{NDCG}}_\tau(\hat{\rho}_2, \mathbf{r})| \leq \\ & \frac{1}{Z(\mathbf{r})} \sum_{i,j} |P_\tau(i, j; \hat{\rho}_1) - P_\tau(i, j; \hat{\rho}_2)| |r_j| |D(i)| \\ & \leq \frac{R_{\max} D_{\max}}{Z(\mathbf{r})} \sum_{i,j} |P_\tau(i, j; \hat{\rho}_1) - P_\tau(i, j; \hat{\rho}_2)| \\ & \leq \frac{R_{\max} D_{\max}}{Z(\mathbf{r})} \|P_\tau(\hat{\rho}_1) - P_\tau(\hat{\rho}_2)\|_1 \\ & \leq \frac{R_{\max} D_{\max}}{Z(\mathbf{r})} L_P \|\hat{\rho}_1 - \hat{\rho}_2\|. \end{aligned}$$

where L_P denotes the Lipschitz constant of the soft permutation operator P_τ . Thus $L(\hat{\rho}, \mathbf{r}) = -\widehat{\text{NDCG}}_\tau(\hat{\rho}, \mathbf{r})$ is L -Lipschitz.

J Sample Complexity Upper Bound

J.1 Proof of Lemma 3.4

Proof. We primarily follow the proof structure of GIFA framework (Réda et al., 2021) and (Purohit et al., 2025).

Preliminaries Recap: Let $\mathcal{S}_m^*$ be the true set of top- m arms and $(\mathcal{S}_m^*)^c$ denote the true set remaining worst arms. The gap-index between any two arms i, j is computed as: $B_t(i, j) = \hat{\rho}_t(i) - \hat{\rho}_t(j) + W_t(i, j)$.

, where

$$W_t = c_t \sqrt{\frac{2\hat{V}_{ij,t} \log(2K^2/\delta)}{N}} + \varepsilon_t^{\text{stab}} + b_i + b_j + \frac{4M \log(2K^2/\delta)}{3N}. \quad (5)$$

as derived in the proof for Theorem 3.2 To prove Lemma 3.4, we introduce the following property,

Property 1: For $b_t \in U_t$ and $ch_t \in C_t$ it holds that $\hat{\rho}_t(b_t) \geq \hat{\rho}_t(ch_t)$. Hence, it follows that $B_t(ch_t, b_t) = \hat{\Delta}_t(ch_t, b_t) + W_t(b_t, ch_t) \leq W_t(b_t, ch_t)$ as $\hat{\Delta}_t(ch_t, b_t) < 0$ From property 1, we can establish that $B_t(ch_t, b_t) \leq W_t(b_t, ch_t)$. Hence, to show that

$$B_t(ch_t, b_t) \leq -(\Delta(b_t) \vee \Delta(ch_t)) + 3W_t(b_t, ch_t)$$

we consider the following scenarios:

(i) $b_t \in \mathcal{S}_m^*$ and $ch_t \notin \mathcal{S}_m^*$: In that case,

$$\Delta(b_t) = \rho(b_t) - \rho(m+1); \Delta(ch_t) = \rho(m) - \rho(ch_t)$$

is the true gap of the arms.

As event $\mathcal{E}$ holds from Theorem 3.2 and Appendix I,

$$\begin{aligned} B_t(ch_t, b_t) &= -B_t(b_t, ch_t) + 2W_t(b_t, ch_t) \\ &\leq \Delta(ch_t, b_t) + 2W_t(b_t, ch_t) \end{aligned}$$

As $ch_t \notin \mathcal{S}_m^*$,

$$\rho(ch_t) \leq \rho(m+1)$$

$$\Delta(ch_t, b_t) \leq \rho(m+1) - \rho(b_t) = -\Delta(b_t)$$

But as $b_t \in \mathcal{S}_m^*$, it also holds that $\rho(b_t) \geq \rho(m)$, and $\Delta(ch_t, b_t) \leq \rho(ch_t) - \rho(m) = -\Delta(ch_t)$. Hence,

$$\begin{aligned} B_t(ch_t, b_t) &\leq -(\Delta(b_t) \vee \Delta(ch_t)) + 2W_t(b_t, c_t) \\ &\leq -(\Delta(b_t) \vee \Delta(ch_t)) + 3W_t(b_t, c_t). \end{aligned}$$

(ii) $b_t \notin \mathcal{S}_m^*$ and $ch_t \in \mathcal{S}_m^*$:

$$\Delta(ch_t) = \rho(ch_t) - \rho(m+1);$$

$$\Delta(b_t) = \rho(m) - \rho(b_t)$$

By Property 1,

$$\begin{aligned} B_t(ch_t, b_t) &\leq W_t(b_t, ch_t) \\ &\leq \hat{\Delta}_t(b_t, ch_t) + W_t(b_t, ch_t) = B_t(b_t, ch_t) \end{aligned}$$

as $\hat{\rho}_t(b_t) \geq \hat{\rho}_t(ch_t)$. Further, as $\mathcal{E}$ holds,

$$\begin{aligned} B_t(b_t, ch_t) &= -B_t(ch_t, b_t) + 2W_t(b_t, ch_t) \\ &\leq \Delta(b_t, ch_t) + 2W_t(b_t, ch_t) \end{aligned}$$

As $b_t \notin \mathcal{S}_m^*$, $\rho(b_t) \leq \rho(m+1)$ and hence $\Delta(b_t, ch_t) \leq \rho(m+1) - \rho(ch_t) = -\Delta(ch_t)$. As $ch_t \in \mathcal{S}_m^*$, $\rho(ch_t) \geq \rho(m)$ and hence $\Delta(b_t, ch_t) \leq \rho(b_t) - \rho(m) = -\Delta(b_t)$. Hence,

$$\begin{aligned} B_t(ch_t, b_t) &\leq -(\Delta(b_t) \vee \Delta(ch_t)) + 2W_t(b_t, c_t) \\ &\leq -(\Delta(b_t) \vee \Delta(ch_t)) + 3W_t(b_t, c_t). \end{aligned}$$

(iii) $b_t \notin \mathcal{S}_m^*$ and $ch_t \notin \mathcal{S}_m^*$: We state that there exists a $b \in \mathcal{S}_m^*$ that belongs to C_t . At any time t ,

$$M_t \leftarrow \text{random } m' \text{ arms from } (U_t \cup C_{t-1})^c$$

$$C_t \leftarrow \text{top}_{m'}(M_t \cup C_{t-1}; \hat{\rho}_{(t-1)})$$

Due to the above sampling approach adopted for C_t which captures the next m' arms with the highest means, it follows by definition of estimated means

that C_t captures at least one arm in $\mathcal{S}_m^*$. Given that $\mathcal{E}$ holds and $b \in \mathcal{S}_m^*$,

$$W_t(b_t, ch_t) \geq B_t(ch_t, b_t) \geq B_t(b, b_t)$$

ch_t by the definition is one of the most ambiguous arms posing largest threat to b_t as it has the largest gap with respect to b_t $B_t(ch_t, b_t) \geq B_t(b, b_t)$. Hence, $B_t(ch_t, b_t) \geq B_t(b, b_t)$. From this and event $\mathcal{E}$ it follows

$$\begin{aligned} B_t(ch_t, b_t) &\geq B_t(b, b_t) \geq \rho(b) - \rho(b_t) \\ &\geq \rho(m) - \rho(b_t) \end{aligned}$$

. Hence $W_t(b_t, ch_t) \geq B_t(ch_t, b_t) \geq \Delta(b_t)$. Using event $\mathcal{E}$,

$$\begin{aligned} B_t(ch_t, b_t) &\leq \Delta(ch_t, b_t) + 2W_t(b_t, ch_t) \\ &= (\rho(ch_t) - \rho(m)) + \\ &\quad (\rho(m) - \rho(b_t)) + 2W_t(b_t, ch_t) \end{aligned}$$

From above Eq and since $B_t(ch_t, b_t) \geq \Delta(b_t)$,

$$\begin{aligned} B_t(ch_t, b_t) &\leq -\Delta(ch_t) + \Delta(b_t) + 2W_t(b_t, ch_t) \\ &\leq -\Delta(ch_t) + 3W_t(b_t, ch_t) \end{aligned}$$

Also from Property 1 and $W_t(b_t, ch_t) \geq \Delta(b_t)$, it holds that

$$\begin{aligned} B_t(ch_t, b_t) &\leq W_t(b_t, ch_t) = -W_t(b_t, ch_t) \\ &\quad + 2W_t(b_t, ch_t) \leq -\Delta(b_t) + 2W_t(b_t, ch_t) \\ &\leq -\Delta(b_t) + 3W_t(b_t, ch_t) \end{aligned}$$

Hence $B_t(ch_t, b_t) \leq -(\Delta(b_t) \vee \Delta(ch_t)) + 3W_t(b_t, c_t)$.

(iv) $b_t \in \mathcal{S}_m^*$ and $ch_t \in \mathcal{S}_m^*$: Then there exists a $s \notin \mathcal{S}_m^*$ and $s \in U_t$. In that case,

$$\begin{aligned} \Delta(b_t) &= \rho(b_t) - \rho(m+1); \Delta(ch_t) \\ &= \rho(ch_t) - \rho(m+1) \end{aligned}$$

Also by definition of b_t and ch_t , it holds that $B_t(ch_t, b_t) = \max_{i \in U_t} \max_{j \in C_t} [B_t(j, i)]$. Since there exists $s \in U_t$ and $ch_t \in C_t$,

$$\begin{aligned} B_t(ch_t, b_t) &= \max_{i \in U_t} \max_{j \in C_t} [B_t(j, i)] \\ &\geq \max_{j \in C_t} B_t(j, s) \geq B_t(ch_t, s) \\ &\geq \rho(ch_t) - \rho(s) \geq \rho(ch_t) - \rho(m+1) \end{aligned}$$

As $\rho(ch_t) - \rho(m+1) = \Delta(ch_t)$, $B_t(ch_t, b_t) \geq \Delta(ch_t)$. By property 1, $B_t(ch_t, b_t) \leq W_t(b_t, ch_t)$. Hence,

$$\Delta(ch_t) \leq B_t(ch_t, b_t) \leq W_t(b_t, ch_t)$$

On event $\mathcal{E}$ it follows that $B_t(ch_t, b_t) \leq \rho(ch_t) - \rho(b_t) + 2W_t(b_t, ch_t)$ as $(B(ch_t, b_t) \leq W_t(b_t, ch_t))$. Then $\rho(ch_t) - \rho(b_t)$ can be expressed as $\rho(ch_t) - \rho(m+1) + \rho(m+1) - \rho(b_t)$. hence,

$$\begin{aligned} B_t(ch_t, b_t) &\leq \rho(ch_t) - \rho(m+1) + \rho(m+1) \\ &\quad - \rho(b_t) + 2W_t(b_t, ch_t) \\ &\leq \Delta(ch_t) - \Delta(b_t) + 2W_t(b_t, ch_t) \end{aligned}$$

We already know that $B_t(ch_t, b_t) \geq \Delta(ch_t)$ resulting in,

$$(a) \quad B_t(ch_t, b_t) \leq -\Delta(b_t) + 3W_t(b_t, ch_t)$$

Now to prove $B_t(ch_t, b_t) \leq -\Delta(ch_t) + 3W_t(b_t, ch_t)$, we rely on property 1,

$$\begin{aligned} B(ch_t, b_t) &\leq W_t(b_t, ch_t) \\ &\leq -W_t(b_t, ch_t) + 2W_t(b_t, ch_t) \end{aligned}$$

As $W_t(b_t, ch_t) \geq \Delta(ch_t)$, $-W_t(b_t, ch_t) \leq -\Delta(ch_t)$. Hence,

$$\begin{aligned} (b) \quad B(ch_t, b_t) &\leq W_t(b_t, ch_t) \leq -W_t(b_t, ch_t) \\ &\quad + 2W_t(b_t, ch_t) \leq -\Delta(ch_t) + W_t(b_t, ch_t) \\ &\leq -\Delta(ch_t) + 3W_t(b_t, ch_t) \end{aligned}$$

From (a) and (b)

$$B_t(ch_t, b_t) \leq -(\Delta(b_t) \vee \Delta(ch_t)) + 3W_t(b_t, ch_t) \quad (6)$$

□

J.2 Proof Blueprint for Theorem 3.3

Proof. We now convert (6) into sampling bounds by using the stopping rule and the explicit form of W_t . The intuition is similar to Lemma 8 in GIFA (Réda et al., 2021), where once the stopping rule $B_t(ch_t, b_t) \leq \varepsilon$ triggers, arms with non-zero gap must have been sampled enough times so that the width is small relative to the gap. We invert this relation to obtain a per-arm bound.

Stopping rule. Assume the algorithm stops when

$$B_t(ch_t, b_t) \leq \varepsilon.$$

On the event $\mathcal{E}$, by Lemma 3.4 at time $t < \text{stopping time}$ we have

$$\varepsilon \leq B_t(ch_t, b_t) \leq -(\Delta(b_t) \vee \Delta(ch_t)) + 3W_t(b_t, ch_t).$$

Rearrange to get

$$3W_t(b_t, ch_t) \geq \varepsilon + (\Delta(b_t) \vee \Delta(ch_t)).$$

$$W_t(b_t, ch_t) \geq \frac{\varepsilon + \Delta_a}{3}$$

Hence, for any arm a that remains active (i.e. is sampled further until elimination), when it is sampled at time t its associated width at that time must satisfy the above inequality (with a playing the role of b_t or ch_t in the identity). Substituting leading term of $W_t(i, j)$ (first term in Equation 5): $c_t \sqrt{\frac{2\hat{V}_{b_t, ch_t, t} \log(2K^2/\delta)}{N}}$ (we ignore stability and bias terms for clarity and also because they are negligible in sample complexity upper bound derivation)

$$c_t \sqrt{\frac{2\hat{V}_{b_t, ch_t, t} \log(2K^2/\delta)}{N}} \geq \frac{\varepsilon + \Delta_a}{3}$$

Since the predictive variance decreases with the number of arm samples $n_t(a)$

$$\hat{V}_{b_t, ch_t, t} \leq \frac{\sigma_{a,t}^2}{n_t(a)}$$

Where $\sigma_{a,t}^2$ is the effective variance of arm a

Substituting this in the inequality from earlier we get,

$$c_t \sigma_{a,t} \sqrt{\frac{2 \log(2K^2/\delta)}{N \cdot n_t(a)}} \geq \frac{\varepsilon + \Delta_a}{3}$$

Taking square on both sides

$$c_t^2 \sigma_{a,t}^2 \frac{2 \log(2K^2/\delta)}{N \cdot n_t(a)} \geq \frac{(\varepsilon + \Delta_a)^2}{9}$$

$$n_t(a) \leq 18 c_t^2 \sigma_{a,t}^2 \frac{\log(2K^2/\delta)}{N(\varepsilon + \Delta_a)^2}$$

To account for case when Δ_a is tiny we replace $(\varepsilon + \Delta_a)^{-2}$ yielding

$$n_t(a) \leq 18c_t^2 \sigma_{a,t}^2 \frac{\log(2K^2/\delta)}{N}.$$

$$\max\left\{\varepsilon^{-2}, \left(\frac{\varepsilon+\Delta_a}{3}\right)^{-2}\right\}$$

Per-arm and total bounds. Formally, on event $\mathcal{E}$, for every arm a ,

$$\mathcal{N}_T(a) \leq 18c_t^2 \sigma_{a,t}^2 \frac{\log(2K^2/\delta)}{N}.$$

$$\max\left\{\varepsilon^{-2}, \left(\frac{\varepsilon+\Delta(a)}{3}\right)^{-2}\right\}. \quad (7)$$

Summing over arms yields the total-sample upper bound

$$T \leq 18c_t^2 \frac{\log(2K^2/\delta)}{N} \sum_{a \in \mathcal{A}} \sigma_{a,t}^2.$$

$$\max\left\{\varepsilon^{-2}, \left(\frac{\varepsilon+\Delta(a)}{3}\right)^{-2}\right\}. \quad (8)$$

□

The above equation leads to the upper bound on sample complexity as stated in Theorem 3.3.

□