BEYOND SURROGATES: A QUANTITATIVE ANALYSIS OF RANKING METRIC RELATIONSHIPS

Anonymous authors
Paper under double-blind review

ABSTRACT

The **Consistency** property between surrogate losses and evaluation metrics has been extensively studied to ensure that approaching the optimal surrogate risk leads to metric optimality. However, such guarantees do not ensure that different evaluation metrics improve together during optimization. This gap limits our understanding of the "**Metric Mismatch**" observed when improvements in one metric fail to translate into gains in another. To bridge this gap, this paper proposes a unified theoretical framework designed to quantify the relationships between ranking metrics. For fixed binary-relevance lists, we characterize the optimal linear **Regret Transfer** constants between ranking metrics and show that global regret transfer and local gain sensitivity can exhibit different scaling behavior. We further investigate **Dynamic Gain Transfer** through changes in cumulative relevance across ranking prefixes, deriving exact metric-gain identities and conditions for simultaneous improvement or conflict. To connect these relationships to optimization, we analyze score-margin dynamics and prove that equal-weight population binary cross-entropy gradient flow with independently optimized logits introduces no new relevance-order inversions. Through this framework, we provide a new perspective on how the location and direction of ranking changes govern cross-metric improvements, offering a theoretical basis for understanding metric alignment along optimization trajectories.

1 INTRODUCTION

Modern machine learning heavily relies on the paradigm of metric-driven optimization. In a typical supervised learning scenario, researchers select an evaluation metric $\mathcal{M}$ that aligns with the practical objective, and minimize a differentiable surrogate loss $\mathcal{L}$ that serves as a proxy for $\mathcal{M}$. A common expectation underlying this paradigm is *monotonicity*: a reduction in the risk of $\mathcal{L}$ is expected to translate into an improvement in $\mathcal{M}$. An important theoretical foundation is **Bayes-consistency** (Bartlett et al., 2006; Tewari & Bartlett, 2007), which guarantees that approaching the optimal surrogate risk leads to the Bayes-optimal metric risk. However, this asymptotic guarantee does not imply that every reduction in surrogate risk improves the target metric, nor does it ensure that different evaluation metrics improve together along the optimization trajectory.

A common expectation is that if a surrogate $\mathcal{L}$ is Bayes-consistent with multiple metrics $\{\mathcal{M}_A, \mathcal{M}_B, \cdots\}$, their empirical performance should be mutually aligned during optimization. However, this expectation need not hold. Related discrepancies between offline evaluation and deployment performance have also motivated the study of **"Metric Mismatch"** in recommender systems (Gilotte et al., 2018; Krauth et al., 2020). Understanding such misalignment presents two fundamental hurdles. First, establishing Bayes-consistency is *non-trivial* (Ravikumar et al., 2011; Buffoni et al., 2011; Calauzènes et al., 2012) and can be particularly challenging for complex industrial losses, which incorporate various heuristics and regularization terms. Second, Bayes-consistency alone is an *asymptotic* property: it does not quantify how different metrics respond to individual updates or to changes at different ranking positions. Due to structural differences, metrics may respond differently to the same optimization step. For instance, ranking objectives can place different emphasis on the top of a list (Rudin, 2009); a model may achieve a higher **AUC** by refining the relative ordering of tail items, yet simultaneously degrade the top-ranked results, leading to a decline in a top-heavy metric such as **NDCG**@k. This conflict can arise even on a fixed evaluation set, independently of distribution shift.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

While the Bayes-consistency between surrogate losses and target metrics has been extensively studied (Zhang, 2004a;b; Agarwal, 2014; Yang & Koyejo, 2020; Gao & Zhou, 2015), and quantitative surrogate-to-metric guarantees have been developed (Agarwal, 2014; Ciliberto et al., 2020), the direct relationships between metric changes along optimization trajectories remain less well understood. A central difficulty is that ranking metrics depend on discrete permutations and assign different values to changes at different positions. Consequently, a global bound on the regret of $\mathcal{M}_B$ in terms of the regret of $\mathcal{M}_A$ does not, by itself, determine whether an update that improves $\mathcal{M}_A$ will also improve $\mathcal{M}_B$. Shared optimal rankings describe compatibility at optimality, and regret bounds quantify deviations from that benchmark. Neither determines the direction of finite-step metric changes, which depends on which ranking errors are corrected and which new errors are introduced. In practical applications, performance trade-offs are therefore often assessed through empirical evaluation and A/B testing (Gomez-Uribe & Hunt, 2016). This disconnection presents a critical dilemma for metric-driven optimization: how can we determine whether an improvement in a baseline metric $\mathcal{M}_A$ will translate into a gain or a loss in another metric $\mathcal{M}_B$?

To bridge this gap, this paper proposes a unified theoretical framework based on metric structures to analyze performance transfer among ranking metrics within the same task scenario. We distinguish **global regret transfer**, which compares deviations from optimal metric values, from **dynamic gain transfer**, which characterizes metric changes along an optimization trajectory. Our main analysis considers fixed binary-relevance lists, allowing us to isolate the effects of ranking changes from changes in the evaluation distribution. Specifically, we make the following contributions:

- We characterize the structural relationships among ranking metrics through cumulative relevance across ranking prefixes. This representation provides a common basis for analyzing **AUC**, **NDCG**, and prefix-based metrics such as **Precision**@k and **Recall**@k.
- We derive linear **regret-transfer bounds** between **AUC** and **NDCG** and characterize their dependence on list size and label density. We further distinguish these global bounds from the sensitivity of local ranking changes, showing that global regret transfer and local gain transfer can exhibit different scaling behavior.
- We establish exact **dynamic gain-transfer** identities in terms of changes in cumulative relevance across ranking prefixes. These identities characterize when metrics improve together and when improvements at lower ranks are outweighed by deterioration near the top. We further connect ranking changes to score-margin dynamics and establish a monotonicity result for equal-weight population binary cross-entropy gradient flow with independently optimized logits.

Consequently, we aim to answer two complementary questions for applications: *If a model achieves an ε-regret on metric $\mathcal{M}_A$, what upper bound can be guaranteed for its regret on metric $\mathcal{M}_B$? Moreover, when an optimization step improves $\mathcal{M}_A$, what determines the direction and magnitude of the corresponding change in $\mathcal{M}_B$?* By addressing both questions, we provide an analytical basis for understanding metric trade-offs beyond proximity to a shared optimal ranking. The resulting framework clarifies how the location and direction of ranking changes govern cross-metric improvements, reshaping the understanding of the impact of offline metric growth on online performance.

2 Related Works

2.1 Consistency: From Metrics to Surrogate Optimization

The study of statistical (Bayes-)consistency centers on whether approaching the optimal surrogate risk ensures convergence to the Bayes-optimal target metric risk. Foundational works (Zhang, 2004a;b; Bartlett et al., 2006; Steinwart, 2007) established consistency characterizations for classification surrogates, with Tewari & Bartlett (2007) further clarifying calibration conditions in the multiclass setting. Reid & Williamson (2011) connected surrogate losses, statistical risks, and divergence measures, offering a broader perspective on risk bounds. Task-specific consistency has also been studied for regression (Christmann & Steinwart, 2007), top-k classification (Yang & Koyejo, 2020), and ranking (Cossock & Zhang, 2006; Ravikumar et al., 2011; Gao & Zhou, 2015). In particular, Agarwal (2014) established ranking regret bounds for strongly proper pointwise losses.

Subsequent works examined $\mathcal{H}$-consistency to account for restricted hypothesis classes (Long & Servedio, 2013; Zhang & Agarwal, 2020). Unlike Bayes-consistency, whose benchmark is defined over unrestricted measurable predictors, $\mathcal{H}$-consistency concerns surrogate and target risks within a specified class $\mathcal{H}$. Awasthi et al. (2022) further developed non-asymptotic bounds incorporating hypothesis-class restrictions and minimizability gaps. These results provide guarantees for transferring surrogate-risk control to target-risk control. In general predictor spaces, equality of optimal predictor sets alone does not guarantee vanishing regret transfer along near-optimal sequences. Even when regret-transfer bounds are available, they do not by themselves ensure simultaneous improvement of different metrics along an optimization trajectory.

2.2 Analysis of Evaluation Metrics

Though evaluation metrics are often treated as static measurement tools, their mathematical properties constitute an independent field of study. Amigó et al. (2009) introduced formal constraints for comparing clustering measures, while Gösgens et al. (2021) analyzed desirable properties and their incompatibilities for classification measures. From a statistical perspective, Wang et al. (2013) studied **NDCG**-type measures under different discount functions, and Narasimhan et al. (2014) established consistency results for plug-in classifiers under non-decomposable measures. For broader unifications, Hernández-Orallo et al. (2012) related evaluation metrics through expected classification loss and threshold choice, while Balestra et al. (2024) used a group-theoretic perspective to characterize ranking metrics and their disagreements.

For direct regret relationships, Narasimhan & Agarwal (2013) established weak regret-transfer bounds between classification, ranking, and probability estimation using distribution-dependent transformations, including threshold selection and score calibration. From an optimization perspective, ranking metrics remain constant while the score-induced ordering is unchanged. LambdaRank addresses this structure by incorporating metric changes induced by document swaps into pairwise update weights (Burges, 2010). Building upon these perspectives, our work distinguishes **global regret transfer** from **dynamic gain transfer**. We derive tight linear bounds between **AUC** and **NDCG** on fixed binary-relevance lists, and characterize finite-step metric changes through ranking-prefix changes, identifying when improvements in one metric translate into gains or losses in another.

3 Preliminaries

Let $\mathscr{X} \subseteq \mathbb{R}^d$ denote the input feature space. We consider a fixed list $\mathscr{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with binary relevance labels $y_i \in \{0, 1\}$. Let $m = \sum_{i=1}^n y_i$ and $\ell = n - m$ denote the numbers of positive and negative instances, respectively, and assume $m, \ell \geq 1$. A scoring function $f : \mathscr{X} \to \mathbb{R}$ assigns a real-valued score to each instance and induces a permutation σ_f satisfying

$$f(\boldsymbol{x}_{\sigma_f(1)}) \geq f(\boldsymbol{x}_{\sigma_f(2)}) \geq \cdots \geq f(\boldsymbol{x}_{\sigma_f(n)}). \tag{1}$$

Score ties are resolved by a fixed deterministic rule. We write $\mathfrak{S}_n$ for the set of all permutations of $\{1, \ldots, n\}$. Unless otherwise stated, our metric comparisons are conditional on the fixed list $\mathscr{D}$, whose dependence is suppressed for brevity.

3.1 Surrogate Objective and Ranking Metrics

Consider a scoring function $f_{\boldsymbol{\theta}}$ trained by minimizing a differentiable surrogate $L_{\mathscr{D}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{s}_{\boldsymbol{\theta}}, \boldsymbol{y})$, where $\boldsymbol{s}_{\boldsymbol{\theta}} = (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))_{i=1}^n$ and $\boldsymbol{y} = (y_i)_{i=1}^n$. The objective may aggregate pointwise, pairwise, or listwise loss terms (Liu, 2011), while evaluation depends on the induced ranking.

In our analysis, we mainly consider normalized additive ranking metrics of the form

$$\mathcal{M}(\sigma) = \frac{1}{Z_{\mathcal{M}}} \sum_{r=1}^{n} a_r y_{\sigma(r)}, \quad a_1 \geq \cdots \geq a_n \geq 0, \tag{2}$$

where the weights and positive normalizer $Z_{\mathcal{M}}$ are fixed for the given list, independent of σ, and chosen so that $\mathcal{M} \in [0, 1]$. This family includes **NDCG, Precision@**k**, Recall@**k, and **NDCG@**k, which we study alongside **AUC**. Formal definitions of these metrics are provided in Appendix A.

Throughout the analysis, we use the logarithmic discounts w_r, their adjacent differences d_r, and the ideal discounted gain Z_m defined by

$$w_r = \frac{1}{\log_2(r+1)}, \quad d_r = w_r - w_{r+1}, \quad Z_m = \sum_{i=1}^{m} w_i, \tag{3}$$

where w_r is defined for $1 \le r \le n$ and d_r for $1 \le r < n$.

3.2 Optimal Rankings and Regret

To characterize metric performance independently of a particular model class, we use the best ranking over all permutations in $\mathfrak{S}_n$ as an oracle for the fixed labeled list.

Definition 1 (Optimal Ranking Set)**.** Given an evaluation metric $\mathcal{M}$, define its **optimal value** and **optimal ranking set** as

$$\mathcal{M}^* \coloneqq \max_{\sigma \in \mathfrak{S}_n} \mathcal{M}(\sigma), \quad \mathfrak{S}^*_{\mathcal{M}} \coloneqq \{\sigma \in \mathfrak{S}_n : \mathcal{M}(\sigma) = \mathcal{M}^*\}. \tag{4}$$

Different metrics may admit different sets of optimal rankings. For **AUC** and **NDCG**, the optimal value is 1, attained exactly when every positive instance precedes every negative instance. Prefix-based metrics may impose weaker requirements, depending on their weights and truncation depth.

Definition 2 (Metric Regret)**.** For a permutation $\sigma \in \mathfrak{S}_n$, its **regret** with respect to $\mathcal{M}$ is

$$\mathrm{Reg}_{\mathcal{M}}(\sigma) \coloneqq \mathcal{M}^* - \mathcal{M}(\sigma). \tag{5}$$

Thus, $\mathrm{Reg}_{\mathcal{M}}(\sigma) \ge 0$, with equality iff $\sigma \in \mathfrak{S}^*_{\mathcal{M}}$. Given function f, write $\mathrm{Reg}_{\mathcal{M}}(f) = \mathrm{Reg}_{\mathcal{M}}(\sigma_f)$.

This regret measures deviation from the ideal ranking of the realized labels. It is distinct from **population Bayes-regret** (Bartlett et al., 2006), which compares expected metric risks under a label distribution. Moreover, the oracle ranking need not be attainable by a restricted scoring model. Our global metric bounds apply to all permutations and therefore also apply to rankings produced by any such model, without changing the oracle or requiring the model to attain the ideal ranking.

For subsequent analysis, we define the **cumulative relevance** $H_r(\sigma)$ within the first r positions and the corresponding **oracle deficit** $E_r(\sigma)$ relative to the ideal ranking of the fixed labels as

$$H_r(\sigma) = \sum_{j=1}^{r} y_{\sigma(j)}, \quad E_r(\sigma) = \min(r, m) - H_r(\sigma) \ge 0. \tag{6}$$

The prefix deficits reveal where relevance is missing relative to the ideal ranking, which are not fully captured by a single metric value. E_r is nonnegative because the ideal ranking maximizes relevance within every prefix, while individual position-wise label differences may have either sign.

3.3 Regret Transfer and Dynamic Gain Transfer

In practice, a learnt ranking output is generally suboptimal. To quantify how its regret in one metric constrains in another, we introduce the concept of **regret transfer**.

Definition 3 (Regret Transfer Function)**.** For metrics $\mathcal{M}_A$ and $\mathcal{M}_B$, the **regret transfer function** is

$$\Psi_{A\to B}(\varepsilon) \coloneqq \sup_{\sigma \in \mathfrak{S}_n} \left\{ \mathrm{Reg}_{\mathcal{M}_B}(\sigma) \,\middle|\, \mathrm{Reg}_{\mathcal{M}_A}(\sigma) \le \varepsilon \right\}, \quad \varepsilon \in [0, 1]. \tag{7}$$

We additionally define the **optimal linear transfer constant** as

$$C^*_{A\to B} \coloneqq \inf \left\{ C \ge 0 : \mathrm{Reg}_{\mathcal{M}_B}(\sigma) \le C\, \mathrm{Reg}_{\mathcal{M}_A}(\sigma) \text{ for all } \sigma \in \mathfrak{S}_n \right\}, \tag{8}$$

with $C^*_{A\to B} = +\infty$ if no finite constant exists. A finite constant yields $\Psi_{A\to B}(\varepsilon) \le \min\{1, C^*_{A\to B}\varepsilon\}$. For a fixed finite list, the transfer function is piecewise constant, with possible jumps at attainable source-regret values. Related surrogate-to-target regret bounds quantify how excess surrogate risk controls target risk (Agarwal, 2014; Kotlowski & Dembczynski, 2017), while we instead compare evaluation metrics directly.

To study finite-step changes, let σ_t denote the permutation output at optimization step t. Define the **metric gain** $\Delta_t\mathcal{M}$ and the **prefix change** $h_r(t)$ by

$$\begin{aligned}\Delta_t\mathcal{M} &:= \mathcal{M}(\sigma_{t+1}) - \mathcal{M}(\sigma_t) = \mathrm{Reg}_{\mathcal{M}}(\sigma_t) - \mathrm{Reg}_{\mathcal{M}}(\sigma_{t+1}),\\ h_r(t) &:= H_r(\sigma_{t+1}) - H_r(\sigma_t), \quad 1 \le r < n.\end{aligned} \tag{9}$$

$h_r(t)$ may be positive or negative, indicating improvement or deterioration at a prefix. We use **dynamic gain transfer** to describe the relationship between $\Delta_t\mathcal{M}_A$ and $\Delta_t\mathcal{M}_B$. When $\Delta_t\mathcal{M}_A \neq 0$, their ratio describes the gain transfer of that particular update and is distinct from $\Psi_{A\to B}$ and $C^*_{A\to B}$.

4 Global Regret Transfer

4.1 Prefix Representation

The relationships between ranking metrics depend on how they evaluate errors at different positions. For a fixed binary-relevance list, $E_r(\sigma)$ provide a common representation of these errors.

Proposition 4 (Regret Representation)**.** *Let $\mathcal{M}$ be an additive ranking metric with fixed weights $a_1 \ge \cdots \ge a_n \ge 0$ and positive normalizer $Z_{\mathcal{M}}$. Then*

$$\mathrm{Reg}_{\mathcal{M}}(\sigma) = \frac{1}{Z_{\mathcal{M}}}\sum_{r=1}^{n-1}(a_r - a_{r+1})E_r(\sigma). \tag{10}$$

For prefix-based metrics Precision $@k$ *and* Recall $@k$*,*

$$\mathrm{Reg}_{\mathrm{Precision}\,@k}(\sigma) = \frac{E_k(\sigma)}{k}, \quad \mathrm{Reg}_{\mathrm{Recall}\,@k}(\sigma) = \frac{E_k(\sigma)}{m}. \tag{11}$$

For AUC *and* NDCG*, we have*

$$\mathrm{Reg}_{\mathrm{AUC}}(\sigma) = \frac{1}{m\ell}\sum_{r=1}^{n-1}E_r(\sigma), \quad \mathrm{Reg}_{\mathrm{NDCG}}(\sigma) = \frac{1}{Z_m}\sum_{r=1}^{n-1}d_rE_r(\sigma). \tag{12}$$

The representation follows from discrete summation by parts, while **AUC** follows from counting positive–negative inversions. Full proofs are provided in Appendix B.1. These identities show that **AUC** assigns equal weight to every deficit, whereas **NDCG** weights each deficit by the corresponding adjacent discount difference. For **NDCG**$@k$, define $Z_{m,k} = \sum_{i=1}^{\min(m,k)} w_i$, whose regret is

$$\mathrm{Reg}_{\mathrm{NDCG}\,@k}(\sigma) = \frac{\sum_{r=1}^{k-1} d_rE_r(\sigma) + w_kE_k(\sigma)}{Z_{m,k}}. \tag{13}$$

Unlike **Precision**$@k$ and **Recall**$@k$, **NDCG**$@k$ additionally penalizes deficits within the prefix. Its boundary coefficient is w_k rather than d_k, reflecting the discontinuity introduced by truncation.

Corollary 5 (Relations of Optimal Ranking)**.** *Under the setting of fixed binary-relevance,*

$$\mathfrak{S}^*_{\mathrm{AUC}} = \mathfrak{S}^*_{\mathrm{NDCG}}, \quad \mathfrak{S}^*_{\mathrm{Precision}\,@k} = \mathfrak{S}^*_{\mathrm{Recall}\,@k}. \tag{14}$$

For $1 \le k_1 < k_2 \le n$,

$$\mathfrak{S}^*_{\mathrm{NDCG}} \subseteq \mathfrak{S}^*_{\mathrm{NDCG}\,@k_2} \subseteq \mathfrak{S}^*_{\mathrm{NDCG}\,@k_1}. \tag{15}$$

The proof is also provided in Appendix B.1. We use the common prefix representation in the following section to derive the optimal global regret-transfer constants between ranking metrics.

4.2 Optimal Regret Transfer Constants

With the prefix representation, we first derive optimal regret-transfer constants for **AUC** and **NDCG**, then extend the comparison to truncated metrics.

Theorem 6 (**AUC**-**NDCG** Transfer)**.** *The optimal transfer constants between* **AUC** *and* **NDCG** *are*

$$C^*_{\mathrm{AUC}\to\mathrm{NDCG}} = \frac{(w_1 - w_{m+1})\ell}{Z_m}, \quad C^*_{\mathrm{NDCG}\to\mathrm{AUC}} = \frac{Z_m}{(w_m - w_n)m}. \tag{16}$$

The constants are attained by the ranked label sequences $(0, 1^m, 0^{\ell-1})$ and $(1^{m-1}, 0^\ell, 1)$, respectively, where b^t denotes t consecutive occurrences of label b.

Table 1: Optimal regret-transfer constants for $1 \le k < n$, with conditions appear in parentheses. The quantities s_k, δ_k, F_k, G_k, and T_k are defined in Eqs. 57–59.

Source / Target	**AUC**	**NDCG**	**Precision@k**	**Recall@k**	**NDCG@k**
AUC	1	$\frac{\ell(w_1-w_{m+1})}{Z_m}$	$\frac{m\ell}{k(\lvert k-m\rvert+1)}$	$\frac{\ell}{\lvert k-m\rvert+1}$	$\frac{m\ell F_k}{Z_{m,k}}$
NDCG	$\frac{Z_m}{m(w_m-w_n)}$	1	$\frac{Z_m}{k\delta_k}$	$\frac{Z_m}{m\delta_k}$	$\frac{Z_m w_{s_k}}{Z_{m,k}\delta_k}$
Precision@k	$\frac{n-1}{\ell}$ $(k=m)$	$\frac{m(w_1-w_n)}{Z_m}$ $(k=m)$	1	$\frac{k}{m}$	$\frac{kw_1}{Z_{m,k}}$ $(k\le m)$
Recall@k	$\frac{n-1}{\ell}$ $(k=m)$	$\frac{m(w_1-w_n)}{Z_m}$ $(k=m)$	$\frac{m}{k}$	1	$\frac{mw_1}{Z_{m,k}}$ $(k\le m)$
NDCG@k	$\frac{Z_m G_k}{m\ell}$ $(k\ge m)$	T_k $(k\ge m)$	$\frac{Z_{m,k}}{kw_{s_k}}$	$\frac{Z_{m,k}}{mw_{s_k}}$	1

Table 2: Optimal constants across $1 \le k_1 < k_2 < n$. The coefficient $D_{k_2\to k_1}$ is defined in Eq. 63.

Direction	**Precision**	**Recall**	**NDCG**
$k_2 \to k_1$	$\frac{k_2}{k_1}$ $(k_2 \le m)$	1 $(k_2 \le m)$	$D_{k_2\to k_1}$
$k_1 \to k_2$	$\frac{k_1}{k_2}$ $(k_1 \ge m)$	1 $(k_1 \ge m)$	1 $(k_1 \ge m)$

The proof exploits the constraints $0 \le q_1 \le \cdots \le q_m \le \ell$ on the positive-item displacements $q_i = r_i - i$, where r_i is the rank of the i-th positive item (detailed in Appendix B.2). We next characterize regret transfer among truncated metrics evaluated at the same cutoff k.

Proposition 7 (Transfer at a Fixed Cutoff)**.** *Let $1 \le k < n$, $s_k = \min(k, m)$, and $Z_{m,k} = \sum_{i=1}^{s_k} w_i$.* Precision@$k$ *and* Recall@k *satisfy*

$$\mathrm{Reg}_{\mathrm{Recall}\,@k}(\sigma) = \frac{k}{m}\,\mathrm{Reg}_{\mathrm{Precision}\,@k}(\sigma), \tag{17}$$

with optimal constants $C^*_{\mathrm{Precision}\,@k\to\mathrm{Recall}\,@k} = \frac{k}{m}, C^*_{\mathrm{Recall}\,@k\to\mathrm{Precision}\,@k} = \frac{m}{k}$. Their relation to NDCG@k *is characterized by*

$$C^*_{\mathrm{NDCG}\,@k\to\mathrm{Precision}\,@k} = \frac{Z_{m,k}}{kw_{s_k}}, \quad C^*_{\mathrm{Precision}\,@k\to\mathrm{NDCG}\,@k} = \begin{cases} \dfrac{kw_1}{Z_{m,k}}, & k \le m, \\ +\infty, & k > m. \end{cases} \tag{18}$$

The corresponding constants with Recall@k *follow from Eq. 17.*

The proof is in Appendix B.3. The following corollary characterizes how the cutoff determines transfer between truncated and global metrics.

Corollary 8 (Transfer across Cutoff Depths)**.** *The following relations hold on a fixed binary list.*

(i) **Global and truncated metrics.** *For $1 \le k < n$, let $G \in \{\mathrm{AUC}, \mathrm{NDCG}\}$. Then $C^*_{G\to T} < \infty$ for every $T \in \{\mathrm{Precision}\,@k, \mathrm{Recall}\,@k, \mathrm{NDCG}\,@k\}$. Conversely,*

$$\left.\begin{array}{l} C^*_{\mathrm{Precision}\,@k\to G} < \infty \\ C^*_{\mathrm{Recall}\,@k\to G} < \infty \end{array}\right\} \iff k = m, \quad C^*_{\mathrm{NDCG}\,@k\to G} < \infty \iff k \ge m. \tag{19}$$

(ii) **Different cutoffs.** *For $1 \le k_1 < k_2 \le n$,* NDCG *satisfies $C^*_{\mathrm{NDCG}\,@k_2\to\mathrm{NDCG}\,@k_1} < \infty$ and*

$$C^*_{\mathrm{NDCG}\,@k_1\to\mathrm{NDCG}\,@k_2} < \infty \iff k_1 \ge m. \tag{20}$$

For $\mathcal{M} \in \{\mathrm{Precision}, \mathrm{Recall}\}$ with $1 \le k_1 < k_2 < n$, we have

$$C^*_{\mathcal{M}@k_2\to\mathcal{M}@k_1} < \infty \iff k_2 \le m, \quad C^*_{\mathcal{M}@k_1\to\mathcal{M}@k_2} < \infty \iff k_1 \ge m. \tag{21}$$

Proofs and detailed constants are provided in Appendix B.3 and summarized in Table 1 and 2.

4.3 Asymptotic Scaling

A finite transfer coefficient allows small regret in one metric to control regret in another. As candidate lists grow, the optimal constants in Theorem 6 quantify how this control changes and whether one transfer direction deteriorates faster than the other. Therefore, we examine **AUC** and **NDCG** as representatives under two regimes: a fixed positive count and a fixed positive-to-negative ratio.

Theorem 9 (Scaling of AUC–NDCG Transfer)**.** *Let $n \to \infty$, with $m + \ell = n$ and $m, \ell \geq 1$. The optimal constants between* **AUC** *and* **NDCG** *satisfy:*

1. **Fixed positive count.** *If $m = m_0$ for a fixed positive integer m_0, then*

$$C^*_{\mathrm{AUC}\to\mathrm{NDCG}} = \Theta(n), \quad C^*_{\mathrm{NDCG}\to\mathrm{AUC}} = \Theta(1). \tag{22}$$

2. **Fixed positive proportion.** *If $m/n \to \alpha \in (0, 1)$, then*

$$C^*_{\mathrm{AUC}\to\mathrm{NDCG}} = \Theta(\log n), \quad C^*_{\mathrm{NDCG}\to\mathrm{AUC}} = \Theta(\log n). \tag{23}$$

The proof is in Appendix B.4. The fixed-count regime approximates retrieval with a small, stable relevant set or recommendation evaluation with a fixed set of observed positive interactions, while the proportional regime instead approximates independent sampling from a stable candidate distribution, where a user's interests induce a relevance probability $\alpha \in (0, 1)$. This asymmetry helps explain why strong offline **AUC** may provide limited assurance of relevance among the results displayed to users.

5 Dynamic Metric Transfer

The global regret bounds in Section 4 compare distances from optimality, but do not determine whether an update improves different metrics simultaneously. The following results characterize finite-step metric changes through the prefix changes $h_r(t)$ defined in Eq. 9.

5.1 Gain Identities

Subtracting the prefix-regret representations at σ_t and σ_{t+1} yields

$$\Delta_t \mathcal{M} = \frac{1}{Z_{\mathcal{M}}} \sum_{r=1}^{n-1} (a_r - a_{r+1}) h_r(t). \tag{24}$$

In particular,

$$\Delta_t \,\mathrm{AUC} = \frac{1}{m\ell} \sum_{r=1}^{n-1} h_r(t), \quad \Delta_t \,\mathrm{NDCG} = \frac{1}{Z_m} \sum_{r=1}^{n-1} d_r h_r(t). \tag{25}$$

The corresponding gain identities for **Precision@**k, **Recall@**k, and **NDCG@**k follow from the same prefix representation and are provided in Appendix C.1. Every additive metric with nonnegative, nonincreasing positional weights and a fixed positive normalizer is nondecreasing under the update if and only if $h_r(t) \geq 0$ for all r.

5.2 Transfer Efficiency and Metric Conflict

We still begin with **AUC** and **NDCG**. Suppose $h_r(t) \geq 0$ for all r and $\sum_{r=1}^{n-1} h_r(t) > 0$. Define the distribution of prefix improvements by $p_t(r) = h_r(t) / \sum_{j=1}^{n-1} h_j(t)$. Equation 25 gives

$$\frac{\Delta_t \,\mathrm{NDCG}}{\Delta_t \,\mathrm{AUC}} = \frac{m\ell}{Z_m} \mathbb{E}_{r\sim p_t}[d_r]. \tag{26}$$

Thus, the same **AUC** gain can yield different **NDCG** gains depending on where the prefix improvements occur. Since d_r decreases with r, shifting p_t toward larger ranks in first-order stochastic dominance cannot increase this coefficient. For updates containing both improvements and deteriorations, write $h_r^+ = \max(h_r(t), 0)$ and $h_r^- = \max(-h_r(t), 0)$. Let $P = \sum_r h_r^+$, $Q = \sum_r h_r^-$,

and $\bar{d}_+ = P^{-1}\sum_r d_r h_r^+$, $\bar{d}_- = Q^{-1}\sum_r d_r h_r^-$, with each average set to zero when its denominator vanishes. Then

$$\Delta_t \,\mathrm{AUC} = \frac{P-Q}{m\ell}, \quad \Delta_t \,\mathrm{NDCG} = \frac{P\bar{d}_+ - Q\bar{d}_-}{Z_m}. \tag{27}$$

Hence, **AUC** increases while **NDCG** decreases if and only if $P > Q$ and $P\bar{d}_+ < Q\bar{d}_-$. A larger amount of prefix improvement at lower ranks can therefore coexist with a **NDCG** loss caused by deterioration near the top. For a favorable adjacent swap at boundary r, only $h_r = 1$ is nonzero, so

$$\frac{\Delta \,\mathrm{NDCG}}{\Delta \,\mathrm{AUC}} = \frac{m\ell d_r}{Z_m}. \tag{28}$$

When $m/n \to \alpha \in (0,1)$, the head-swap gain ratio is $\Theta(n \log n)$, versus $\Theta(\log n)$ for the optimal global regret constant: a local update can isolate one boundary, whereas global regret aggregates oracle deficits (Appendix C.3). The decomposition also extends to truncated metrics in Appendix C.2.

5.3 From Optimization to Ranking Changes

The preceding identities determine metric gains once the ranking changes are known. We now examine how optimization generates these changes. Under parameter gradient flow for $L(\boldsymbol{s_\theta})$, let $J = \partial \boldsymbol{s_\theta}/\partial \boldsymbol{\theta}$, $K = JJ^\top$, and $q_{ij} = s_i - s_j$. τ denotes continuous optimization time, which gives

$$\frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}\tau} = -K\nabla_{\boldsymbol{s}} L, \quad \frac{\mathrm{d}q_{ij}}{\mathrm{d}\tau} = -(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top K \nabla_{\boldsymbol{s}} L. \tag{29}$$

A pair changes order when its score margin crosses zero. To isolate the role of coupling, consider independently optimized logits under equal-weight conditional population **BCE**. For fixed candidate features $\boldsymbol{X}$, write $\eta_i = \mathbb{P}(Y_i = 1 \mid \boldsymbol{X})$ and $\rho(s) = (1+e^{-s})^{-1}$. The flow satisfies

$$\frac{\mathrm{d}s_i}{\mathrm{d}\tau} = \eta_i - \rho(s_i), \quad \left.\frac{\mathrm{d}q_{ij}}{\mathrm{d}\tau}\right|_{s_i = s_j} = \eta_i - \eta_j. \tag{30}$$

Thus, if $\eta_i > \eta_j$, the pair can cross only from the incorrect order to the correct order. For deterministic labels $\eta_i = y_i$, every opposite-label crossing moves a positive item forward. All prefix counts are therefore nondecreasing, which guarantees simultaneous nondecrease of AUC and the additive metrics considered above.

With shared parameters, K couples score updates and can reverse a crossing direction even while the loss decreases. The independent-logit result therefore provides a baseline for metric alignment, while the general dynamics explain why the loss alone does not determine whether metrics improve together. Proofs and a shared-parameter counterexample are provided in Appendix C.4.

6 Experiments

Since our theory distinguishes global regret transfer, optimization-induced ranking changes, and finite-step metric gains, we therefore evaluate these results through the following questions:

- **Q1:** Are the global bounds sharp, and do the transfer properties exhibit the predicted scaling?
- **Q2:** Can parameter coupling alter updates despite matched expressivity and decreasing loss?
- **Q3:** Do conflicts persist in real-training, and which positional changes account for them?

Q1. We enumerate all 32,738 binary sequences containing both labels with $n \leq 14$, compute the maximum regret ratios, and compare them with the theoretical coefficients. The derived sharp constants agree within 8×10^{-12}, and the predicted finite-transfer conditions hold throughout. We then scan list sizes up to $n = 262{,}144$, evaluating both extremal rankings and adjacent swaps. The results exhibit the scaling in Theorem 9: linear-versus-bounded growth at fixed positive counts and logarithmic growth in both directions at fixed positive proportions. In the latter regime, the head-swap gain ratio instead grows as $n \log n$, confirming the predicted separation between global regret and local gain coefficients. Details appear in Appendix D.1.

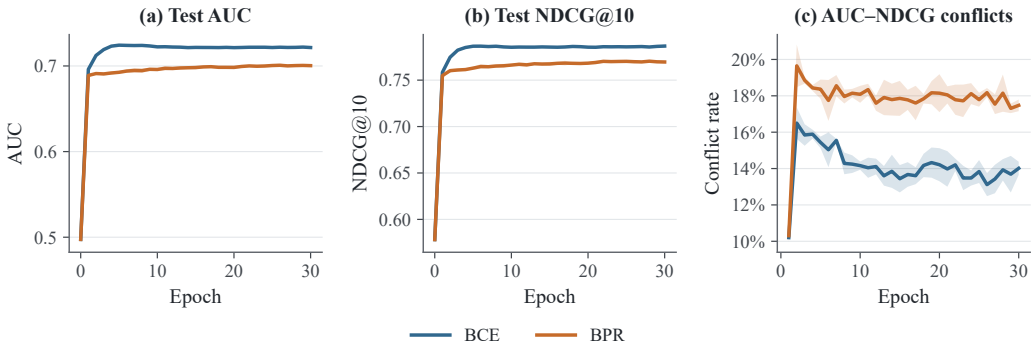


Figure 1: Test metric trajectories and **AUC/NDCG** conflict rates. Curves and shading show means and standard deviations across three seeds. Table 3 includes all moving lists.

Table 3: Conflict and mixed-prefix rates over epochs 2–30, as percentages of moving test-user windows. Values are mean ± standard deviation across seeds.

Loss	AUC/NDCG conflict	AUC/NDCG@10 conflict	Mixed prefixes
BCE	14.19 ± 0.09	14.78 ± 0.12	62.40 ± 0.22
BPR	17.91 ± 0.16	19.65 ± 0.15	69.10 ± 0.16

Q2. To isolate parameter coupling, we optimize 16 logits with **BCE** using independent scores or an invertible linear parameterization. Both represent every score vector and start from identical scores. Across 100 trials per setting, independent optimization introduces no new target-order inversions and, with binary targets, preserves all prefix counts. Coupling introduces inversions in 22/100 binary-target trials despite non-increasing **BCE**. Thus, matched score expressivity and decreasing loss still permit different ranking updates, supporting the mechanism in Section 5.3. Protocols and aggregate results are provided in Appendix D.2.

Q3. We train the same matrix-factorization scorer with **BCE** and **BPR** (Rendle et al., 2009) on **MovieLens-1M** (Harper & Konstan, 2016), using three seeds and 30 epochs. We record scores after every epoch, measure per-user metric conflicts, and decompose their gains through consecutive-checkpoint prefix changes. The data and training protocol are detailed in Appendix D.3.

Figure 1 shows that aggregate performance gains coexist with persistent per-user conflicts. Table 3 shows that mixed prefix changes are substantially more frequent than conflicts. Their positions and magnitudes determine whether metric gains have opposite signs. The signed prefix identities recover observed gains with residual below 8×10^{-16}. Regional decompositions identify windows where tail improvements raise **AUC** while head deterioration dominates **NDCG**, illustrating the theoretical conflict condition in real training (Appendix D.4). Checkpoint-selection comparisons yield small, seed-dependent test-performance differences (Appendix D.5).

7 Conclusion

In this paper, we develop a unified framework that quantifies both **global regret transfer** and **dynamic gain transfer** on fixed binary ranking lists. We derive sharp transfer constants for **AUC, NDCG**, and truncated metrics, identifying how relevance sparsity and evaluation cutoffs govern the strength and direction of metric control. Beyond shared optimal rankings, our prefix-based analysis characterizes how the locations of improvements and deteriorations determine metric agreement or conflict. We establish a separation between global regret control and local swap sensitivity, and connect these ranking changes to optimization dynamics. Together, these results explain metric mismatch through explicit error structures and update mechanisms, providing quantitative criteria for assessing when progress under one metric carries over to another.

AI USE STATEMENT

The research motivation, original theoretical derivations and proofs, initial manuscript were developed entirely by the authors without AI assistance. The authors then use GPT-6 to assist in checking grammar issues, theoretical rigor, experimental implementations and language polishing. The authors carefully reviewed all AI-assisted revisions and take full responsibility for the entire paper, including its theoretical claims, proofs, code, and experimental results.

REPRODUCIBILITY STATEMENT

Proofs of the global regret-transfer results are provided in Appendices B.2, B.3, and B.4; dynamic metric-transfer derivations appear in Appendix C. Appendix D specifies the enumeration ranges, optimization settings, MovieLens preprocessing, data splits, evaluation conventions, and random seeds. Code and configurations for reproducing the experiments is available in the anonymized repository: `https://anonymous.4open.science/r/metric`.

REFERENCES

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *J. Mach. Learn. Res.*, 15(1):1653–1674, 2014. doi: 10.5555/2627435.2638588. URL https://dl.acm.org/doi/10.5555/2627435.2638588.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486, 2009. doi: 10.1007/S10791-008-9066-8. URL https://doi.org/10.1007/s10791-008-9066-8.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1117–1174. PMLR, 2022. URL https://proceedings.mlr.press/v162/awasthi22c.html.

Chiara Balestra, Andreas Mayr, and Emmanuel Müller. Ranking evaluation metrics from a group-theoretic perspective. *CoRR*, abs/2408.16009, 2024. doi: 10.48550/ARXIV.2408.16009. URL https://doi.org/10.48550/arXiv.2408.16009.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

David Buffoni, Clément Calauzènes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 825–832. Omnipress, 2011. URL https://icml.cc/2011/papers/447_icmlpaper.pdf.

Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11 (23-581):81, 2010.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. "on the (non-)existence of convex, calibrated surrogate losses for ranking". In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *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*, pp. 197–205, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/50f3f8c42b998a48057e9d33f4144b8b-Abstract.html.

Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *J. Mach. Learn. Res.*, 21:98:1–98:67, 2020. URL https://jmlr.org/papers/v21/20-097.html.

David Cossock and Tong Zhang. Subset ranking using regression. In Gábor Lugosi and Hans Ulrich Simon (eds.), *Learning Theory, 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006, Proceedings*, volume 4005 of *Lecture Notes in Computer Science*, pp. 605–619. Springer, 2006. doi: 10.1007/11776420_44. URL https://doi.org/10.1007/11776420_44.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In Qiang Yang and Michael J. Wooldridge (eds.), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 939–945. AAAI Press, 2015. URL http://ijcai.org/Abstract/15/137.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline A/B testing for recommender systems. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (eds.), *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pp. 198–206. ACM, 2018. doi: 10.1145/3159652.3159687. URL https://doi.org/10.1145/3159652.3159687.

Carlos Alberto Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manag. Inf. Syst.*, 6(4):13:1–13:19, 2016. doi: 10.1145/2843948. URL https://doi.org/10.1145/2843948.

Martijn Gösgens, Anton Zhiyanov, Aleksey Tikhonov, and Liudmila Prokhorenkova. Good classification measures and how to find them. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17136–17147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8e489b4966fe8f703b5be647f1cbae63-Abstract.html.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, 2016. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

José Hernández-Orallo, Peter A. Flach, and César Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss. *J. Mach. Learn. Res.*, 13:2813–2869, 2012. doi: 10.5555/2503308.2503332. URL https://dl.acm.org/doi/10.5555/2503308.2503332.

Wojciech Kotlowski and Krzysztof Dembczynski. Surrogate regret bounds for generalized classification performance metrics. *Mach. Learn.*, 106(4):549–572, 2017. doi: 10.1007/S10994-016-5591-7. URL https://doi.org/10.1007/s10994-016-5591-7.

Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I. Jordan. Do offline metrics predict online performance in recommender systems? *CoRR*, abs/2011.07931, 2020. URL https://arxiv.org/abs/2011.07931.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011. ISBN 978-3-642-14266-6. doi: 10.1007/978-3-642-14267-3. URL https://doi.org/10.1007/978-3-642-14267-3.

Philip M. Long and Rocco A. Servedio. Consistency versus realizable h-consistency for multiclass classification. 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*, pp. 801–809. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/long13.html.

Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In Christopher J. C. Burges, Léon

Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2913–2921, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 1493–1501, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/32b30a250abd6331e03a2a1f16466346-Abstract.html.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 618–626. JMLR.org, 2011. URL http://proceedings.mlr.press/v15/ravikumar11a/ravikumar11a.pdf.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12:731–817, 2011. doi: 10.5555/1953048.2021029. URL https://dl.acm.org/doi/10.5555/1953048.2021029.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. In Jeff A. Bilmes and Andrew Y. Ng (eds.), *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pp. 452–461. AUAI Press, 2009. URL https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf.

Cynthia Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *J. Mach. Learn. Res.*, 10:2233–2271, 2009. doi: 10.5555/1577069.1755861. URL https://dl.acm.org/doi/10.5555/1577069.1755861.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *J. Mach. Learn. Res.*, 8:1007–1025, 2007. doi: 10.5555/1314498.1390325. URL https://dl.acm.org/doi/10.5555/1314498.1390325.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of NDCG type ranking measures. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *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*, pp. 25–54. JMLR.org, 2013. URL http://proceedings.mlr.press/v30/Wang13.html.

Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. 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*, pp. 10727–10735. PMLR, 2020. URL http://proceedings.mlr.press/v119/yang20f.html.

Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c4c28b367e14df88993ad475dedf6b77-Abstract.html.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004a. ISSN 00905364. URL http://www.jstor.org/stable/3448494.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5:1225–1251, 2004b. URL https://jmlr.org/papers/volume5/zhang04b/zhang04b.pdf.

APPENDIX

A Definitions of Evaluation Metrics **15**
A.1 AUC . . . 15
A.2 Precision and Recall at a Cutoff . . . 15
A.3 NDCG and Truncated NDCG . . . 15
A.4 Fixed Positional Weights . . . 15

B Proofs of Metrics Properties **15**
B.1 Proofs of Proposition 4 and Corollary 5 . . . 15
B.2 Proof of Theorem 6 . . . 16
B.3 Regret Transfer for Truncated Metrics . . . 17
B.3.1 Finite transfer and an exact reduction . . . 17
B.3.2 Proof of Proposition 7 . . . 18
B.3.3 Proof of Corollary 8 . . . 19
B.3.4 Global-truncated transfer constants . . . 20
B.3.5 Transfer across cutoffs . . . 21
B.4 Proof of Theorem 9 . . . 21

C Proofs and Extensions for Dynamic Metric Transfer **22**
C.1 Gain Identities and Simultaneous Improvement . . . 22
C.2 Transfer Efficiency and Metric Conflict . . . 22
C.3 Local Exchanges and Global Regret . . . 23
C.4 Score Dynamics and Population BCE . . . 24

D Experimental Details **24**
D.1 Exact Enumeration and Scaling . . . 24
D.2 Controlled Optimization and Structural Updates . . . 26
D.3 MovieLens Protocol . . . 26
D.4 Trajectory Diagnostics . . . 27
D.5 Checkpoint Selection and Test Performance . . . 28

A Definitions of Evaluation Metrics

We consider a fixed list of $n = m + \ell$ items with binary labels $y_i \in \{0, 1\}$, where $m = \sum_{i=1}^{n} y_i \geq 1$ and $\ell = n - m \geq 1$. Let $\sigma(r)$ denote the item at rank r in descending score order, while score ties are resolved by a fixed, label-independent rule.

A.1 AUC

AUC is the fraction of positive-negative pairs ordered correctly:

$$\mathrm{AUC}(\sigma) = \frac{1}{m\ell} \sum_{r=1}^{n} \sum_{s=r+1}^{n} y_{\sigma(r)}(1 - y_{\sigma(s)}). \tag{31}$$

If $r_1 < \cdots < r_m$ are the ranks of the positive items, then $1 - \mathrm{AUC}(\sigma) = \frac{1}{m\ell} \sum_{i=1}^{m} (r_i - i)$. Each term $r_i - i$ counts the negative items preceding the i-th positive item.

A.2 Precision and Recall at a Cutoff

Define the positive count in the first k positions as $H_k(\sigma) = \sum_{r=1}^{k} y_{\sigma(r)}$. **Precision@**$k$ and **Recall@**k are defined as

$$\mathrm{Precision@}k(\sigma) = \frac{H_k(\sigma)}{k}, \quad \mathrm{Recall@}k(\sigma) = \frac{H_k(\sigma)}{m}. \tag{32}$$

Their oracle values over all permutations are

$$(\mathrm{Precision@}k)^* = \frac{\min(k, m)}{k}, \quad (\mathrm{Recall@}k)^* = \frac{\min(k, m)}{m}. \tag{33}$$

A.3 NDCG and Truncated NDCG

Let $w_r = \frac{1}{\log_2(r+1)}$, $Z_m = \sum_{r=1}^{m} w_r$ and $Z_{m,k} = \sum_{r=1}^{\min(k,m)} w_r$. For binary relevance, full-list and truncated **NDCG** are

$$\mathrm{NDCG}(\sigma) = \frac{\sum_{r=1}^{n} w_r y_{\sigma(r)}}{Z_m}, \quad \mathrm{NDCG@}k(\sigma) = \frac{\sum_{r=1}^{k} w_r y_{\sigma(r)}}{Z_{m,k}}. \tag{34}$$

In following sections, **NDCG** denotes the full-list metric unless a cutoff is explicitly specified.

A.4 Fixed Positional Weights

The additive metric family in Eq.2 has weights and normalizer independent of the permutation. The following choices recover the metrics defined above and is easy to validate:

Metric	**Precision@**k	**Recall@**k	**NDCG**	**NDCG@**k
a_r	$\mathbb{I}(r \leq k)$	$\mathbb{I}(r \leq k)$	w_r	$w_r \mathbb{I}(r \leq k)$
Z_a	k	m	Z_m	$Z_{m,k}$

AUC is connected to these metrics through its prefix regret representation in Proposition 4.

B Proofs of Metrics Properties

B.1 Proofs of Proposition 4 and Corollary 5

Proof of Proposition 4. Let $z_r = y_{\sigma(r)}$, $z_r^* = \mathbf{1}\{r \leq m\}$ denote the ideal binary relevance sequence. Since the weights are non-increasing, this ideal sequence maximizes the additive metric. Hence

$$Z_{\mathcal{M}} \, \mathrm{Reg}_{\mathcal{M}}(\sigma) = \sum_{r=1}^{n} a_r (z_r^* - z_r).$$

By definition,

$$E_r(\sigma) = \sum_{j=1}^{r}(z_j^* - z_j), \quad E_0(\sigma) = E_n(\sigma) = 0.$$

Discrete summation by parts gives

$$\sum_{r=1}^{n} a_r(z_r^* - z_r) = \sum_{r=1}^{n} a_r\big(E_r(\sigma) - E_{r-1}(\sigma)\big) = \sum_{r=1}^{n-1}(a_r - a_{r+1})E_r(\sigma).$$

Dividing by $Z_{\mathcal{M}}$ proves Eq. 10. Taking $a_r = w_r$ and $Z_{\mathcal{M}} = Z_m$ gives the **NDCG** result.

For **AUC**, let $r_1 < \cdots < r_m$ be the ranks of the positive instances. The positive instance at rank r_i has exactly $r_i - i$ negative instances above it. Thus the total number of incorrectly ordered positive–negative pairs is $I(\sigma) = \sum_{i=1}^{m}(r_i - i)$. On the other hand,

$$\sum_{r=1}^{n-1} E_r(\sigma) = \sum_{r=1}^{n-1} \min(r, m) - \sum_{r=1}^{n-1} H_r(\sigma) = \sum_{i=1}^{m}(n - i) - \sum_{i=1}^{m}(n - r_i) = \sum_{i=1}^{m}(r_i - i) = I(\sigma).$$

Since $\mathrm{Reg}_{\mathrm{AUC}}(\sigma) = I(\sigma)/(m\ell)$, the **AUC** result follows. □

Proof of Corollary 5. The coefficients in both **AUC** and **NDCG** representations are strictly positive for $1 \le r < n$. Consequently, both regrets vanish exactly when $E_r(\sigma) = 0$ for all $1 \le r < n$, proving equality of their optimal ranking sets. By Eq. 11, **Precision@**k and **Recall@**k are optimal exactly when $E_k(\sigma) = 0$. They therefore have the same optimal ranking set.

For **NDCG@**k, all coefficients in Eq. 13 are strictly positive. Its regret vanishes exactly when $E_r(\sigma) = 0$ for $1 \le r \le k$. Thus optimality at depth k_2 implies optimality at every smaller depth k_1, which proves Eq. 15. □

B.2 Proof of Theorem 6

The following lemma isolates the displacement constraint responsible for the sharp constants.

Lemma 10 (Discounted regret per inversion)**.** *Let $r_1 < \cdots < r_m$ be the positive ranks and $q_i = r_i - i$. Then $0 \le q_1 \le \cdots \le q_m \le \ell$. Writing $I = \sum_i q_i$ and $D_w = \sum_i (w_i - w_{i+q_i})$, we have*

$$\frac{w_m - w_n}{\ell} I \le D_w \le \frac{w_1 - w_{m+1}}{m} I. \tag{35}$$

Moreover, **AUC** *and* **NDCG** *regrets are $I/(m\ell)$ and D_w/Z_m, respectively.*

Proof. Fix a permutation σ, and let $r_1 < \cdots < r_m$ denote the ranks of its positive instances. Define their displacements from the ideal ranks by $q_i = r_i - i$ with $i = 1, \ldots, m$. Since $r_{i+1} \ge r_i + 1$, these displacements satisfy $0 \le q_1 \le q_2 \le \cdots \le q_m \le \ell$. The upper bound follows because q_i is the number of negative instances preceding the i-th positive instance. Let $I(\sigma) = \sum_{i=1}^{m} q_i, D_w(\sigma) = \sum_{i=1}^{m}\big(w_i - w_{i+q_i}\big)$ denote the inversion count and the unnormalized discounted regret, respectively. Then

$$\mathrm{Reg}_{\mathrm{AUC}}(\sigma) = \frac{I(\sigma)}{m\ell}, \quad \mathrm{Reg}_{\mathrm{NDCG}}(\sigma) = \frac{D_w(\sigma)}{Z_m}. \tag{36}$$

Assume $I(\sigma) > 0$, for $w(x) = \ln 2/\ln(x+1)$, we have

$$w'(x) = -\frac{\ln 2}{(x+1)\ln^2(x+1)} < 0, \quad w''(x) = \frac{\ln 2\,[\ln(x+1) + 2]}{(x+1)^2 \ln^3(x+1)} > 0.$$

Consequently, its adjacent differences $d_r = w_r - w_{r+1}$ are positive and decreasing. Expanding each displacement gives $D_w(\sigma) = \sum_{i=1}^{m}\sum_{j=1}^{q_i} d_{i+j-1}$.

Upper bound. For each $j \in \{1, \dots, \ell\}$, define $S_j = \{i : q_i \geq j\}$. Since every non-empty S_j is a suffix of $\{1, \dots, m\}$ and d_r is decreasing,

$$\frac{1}{|S_j|} \sum_{i \in S_j} d_{i+j-1} \leq \frac{1}{|S_j|} \sum_{i \in S_j} d_i \leq \frac{1}{m} \sum_{i=1}^{m} d_i = \frac{w_1 - w_{m+1}}{m}. \tag{37}$$

The second inequality holds because the average of a suffix of a decreasing sequence cannot exceed the average of the whole sequence. Reordering the sum therefore yields

$$D_w(\sigma) = \sum_{j=1}^{\ell} \sum_{i \in S_j} d_{i+j-1} \leq \frac{w_1 - w_{m+1}}{m} \sum_{j=1}^{\ell} |S_j| = \frac{w_1 - w_{m+1}}{m} I(\sigma). \tag{38}$$

Lower bound. For every i with $q_i > 0$, the inequalities $i \leq m$ and $q_i \leq \ell$, together with the monotonicity of d_r give

$$\frac{1}{q_i} \sum_{j=1}^{q_i} d_{i+j-1} \geq \frac{1}{q_i} \sum_{j=1}^{q_i} d_{m+j-1} \geq \frac{1}{\ell} \sum_{j=1}^{\ell} d_{m+j-1} = \frac{w_m - w_n}{\ell}. \tag{39}$$

Multiplying by q_i and summing over the non-zero displacements gives

$$D_w(\sigma) \geq \frac{w_m - w_n}{\ell} I(\sigma). \tag{40}$$

Combining Eqs. 38 and 40 obtains the final result. □

Proof of Theorem 6. Applying Lemma 10 and substituting Eq. 36 yields

$$\mathrm{Reg}_{\mathrm{NDCG}}(\sigma) \leq \frac{\ell(w_1 - w_{m+1})}{Z_m} \mathrm{Reg}_{\mathrm{AUC}}(\sigma), \quad \mathrm{Reg}_{\mathrm{AUC}}(\sigma) \leq \frac{Z_m}{m(w_m - w_n)} \mathrm{Reg}_{\mathrm{NDCG}}(\sigma).$$

For the ranked label sequence $(0, 1^m, 0^{\ell-1})$, the positive ranks are $r_i = i + 1$, so $q_i = 1$ for every i. Hence $I(\sigma) = m$ and $D_w(\sigma) = \sum_{i=1}^{m} (w_i - w_{i+1}) = w_1 - w_{m+1}$. Therefore,

$$\frac{\mathrm{Reg}_{\mathrm{NDCG}}(\sigma)}{\mathrm{Reg}_{\mathrm{AUC}}(\sigma)} = \frac{\ell(w_1 - w_{m+1})}{Z_m}. \tag{41}$$

For sequence $(1^{m-1}, 0^{\ell}, 1)$, $q_i = 0$ for $i < m$ and $q_m = \ell$. Thus $I(\sigma) = \ell, D_w(\sigma) = w_m - w_n$ and

$$\frac{\mathrm{Reg}_{\mathrm{AUC}}(\sigma)}{\mathrm{Reg}_{\mathrm{NDCG}}(\sigma)} = \frac{Z_m}{m(w_m - w_n)}. \tag{42}$$

Both ratios are attained at rankings with positive regret. No smaller constants can therefore satisfy the respective bounds for all permutations, proving optimality. □

B.3 Regret Transfer for Truncated Metrics

Throughout, the binary list is fixed, with $n = m + \ell$, $m, \ell \geq 1$, and $1 \leq k < n$. Write $Z_{m,k} = \sum_{i=1}^{\min(k,m)} w_i$. The coefficient $C^*_{A \to B}$ is the smallest nonnegative constant satisfying $\mathrm{Reg}_B \leq C \, \mathrm{Reg}_A$ for all rankings, and is $+\infty$ if none exists.

B.3.1 Finite transfer and an exact reduction

Lemma 11 (Finite-domain transfer criterion)**.** *On the finite permutation space,*

$$C^*_{A \to B} < \infty \quad \Longleftrightarrow \quad \mathfrak{S}^*_A \subseteq \mathfrak{S}^*_B. \tag{43}$$

Under this inclusion, $C^*_{A \to B} = \max_{\sigma : \mathrm{Reg}_A(\sigma) > 0} \frac{\mathrm{Reg}_B(\sigma)}{\mathrm{Reg}_A(\sigma)}$ *with an empty maximum defined as zero.*

Proof. A finite bound requires zero source regret to imply zero target regret. Under this condition, the remaining ratios form a finite set, whose maximum is the optimal coefficient. □

Lemma 12 (Interval decomposition and sharp transfer)**.** *Every oracle deficit can be written as*

$$E_r(\sigma) = \sum_{t=1}^{E_m(\sigma)} \mathbf{1}_{\{a_t \le r < c_t\}}, \quad 1 \le a_t \le m < c_t \le n. \tag{44}$$

For $a \le m < c$, let $\sigma^{a,c}$ be the ranking obtained by exchanging positions a and c in the ideal ordering. Then $E_r(\sigma^{a,c}) = \mathbf{1}_{\{a \le r < c\}}$. If $\mathrm{Reg}_M = \sum_r \beta_r^M E_r$ for $M \in \{A, B\}$, with fixed $\beta_r^M \ge 0$, define $g_M(a,c) = \sum_{r=a}^{c-1} \beta_r^M$. Then

$$C^*_{A \to B} = \max_{1 \le a \le m < c \le n} \frac{g_B(a,c)}{g_A(a,c)}. \tag{45}$$

Proof. The deficit E_r is non-decreasing for $r \le m$ and non-increasing for $r \ge m$. Each non-empty integer super-level set $\{r : E_r \ge t\}$ is therefore an interval $\{a_t, \dots, c_t - 1\}$ containing m. Summing its indicators proves the decomposition. Exchanging ideal positions a and c creates exactly the corresponding interval deficit, so every such elementary ranking is feasible. Linearity gives

$$\mathrm{Reg}_M(\sigma) = \sum_{t=1}^{E_m(\sigma)} g_M(a_t, c_t).$$

A zero elementary source regret with positive target regret rules out finite transfer. Otherwise, the maximum elementary ratio bounds every summand and hence every ranking. □

The elementary regrets for the metrics considered here are

$$\begin{aligned} g_{\mathrm{AUC}}(a,c) &= \frac{c-a}{m\ell}, & g_{\mathrm{NDCG}}(a,c) &= \frac{w_a - w_c}{Z_m}, \\ g_{\mathrm{Precision}\,@k}(a,c) &= \frac{\mathbf{1}_{\{a \le k < c\}}}{k}, & g_{\mathrm{Recall}\,@k}(a,c) &= \frac{\mathbf{1}_{\{a \le k < c\}}}{m}, \\ g_{\mathrm{NDCG}\,@k}(a,c) &= \frac{w_a \mathbf{1}_{\{a \le k\}} - w_c \mathbf{1}_{\{c \le k\}}}{Z_{m,k}}. \end{aligned} \tag{46}$$

Substitution into Eq. 45 yields the sharp coefficients below.

B.3.2 Proof of Proposition 7

Proof. The prefix representation gives

$$\mathrm{Reg}_{\mathrm{Precision}\,@k} = \frac{E_k}{k}, \quad \mathrm{Reg}_{\mathrm{Recall}\,@k} = \frac{E_k}{m}. \tag{47}$$

Consequently, $C^*_{\mathrm{Precision}\,@k \to \mathrm{Recall}\,@k} = \frac{k}{m}$ and $C^*_{\mathrm{Recall}\,@k \to \mathrm{Precision}\,@k} = \frac{m}{k}$. Since a positive prefix deficit is attainable for $k < n$, both proportionality constants are optimal.

To compare **NDCG@**k with **Precision@**k, put $s = \min(k, m)$, $Z = Z_{m,k}$, and $t = E_k$. There are exactly $s - t$ positive items in the prefix. Their discounted gain is at most $\sum_{i=1}^{s-t} w_i$, so

$$Z \,\mathrm{Reg}_{\mathrm{NDCG}\,@k} \ge \sum_{i=s-t+1}^{s} w_i \ge t w_s. \tag{48}$$

Consequently,

$$\mathrm{Reg}_{\mathrm{Precision}\,@k} \le \frac{Z}{k w_s} \,\mathrm{Reg}_{\mathrm{NDCG}\,@k}\,. \tag{49}$$

Equality with positive regret is attained by placing positive items at ranks $1, \dots, s-1$, placing no other positives within the cutoff, and assigning the remaining positives beyond k.

Then we suppose $k \le m$. The ideal prefix consists entirely of positives, and $t = E_k$ equals the number of negative items within the prefix. Hence

$$Z \,\mathrm{Reg}_{\mathrm{NDCG}\,@k} = \sum_{\substack{r \le k \\ y_{\sigma(r)} = 0}} w_r \le t w_1, \tag{50}$$

which yields

$$\mathrm{Reg}_{\mathrm{NDCG}\,@k} \le \frac{kw_1}{Z}\,\mathrm{Reg}_{\mathrm{Precision}\,@k}\,. \tag{51}$$

The sequence $(0, 1^m, 0^{\ell-1})$ attains equality as an example.

If $k > m$, the same sequence places all positive items within the prefix. Thus $\mathrm{Reg}_{\mathrm{Precision}\,@k} = 0$, whereas $\mathrm{Reg}_{\mathrm{NDCG}\,@k} > 0$, proving that no finite reverse coefficient exists. The recall constants are

$$C^*_{\mathrm{NDCG}\,@k\to\mathrm{Recall}\,@k} = \frac{Z_{m,k}}{mw_{\min(k,m)}}, \quad C^*_{\mathrm{Recall}\,@k\to\mathrm{NDCG}\,@k} = \begin{cases} \dfrac{mw_1}{Z_{m,k}}, & k \le m, \\ +\infty, & k > m. \end{cases} \tag{52}$$

□

B.3.3 Proof of Corollary 8

Proof. **AUC** and **NDCG** are optimal when all positives precede all negatives. Such a ranking also optimizes every truncated metric considered here. For $k = m$, maximizing **Precision@**k or **Recall@**k requires that the first m items are positive, so their optimal sets coincide with the global optimal set. For $k < m$, the sequence $(1^k, 0^\ell, 1^{m-k})$ optimizes **Precision@**k, **Recall@**k, and **NDCG@**k, but is not globally optimal. For $k > m$, the sequence $(0, 1^m, 0^{\ell-1})$ optimizes **Precision@**k and **Recall@**k but is not globally optimal.

Finally, if $k \ge m$, the strictly decreasing positive discounts imply that **NDCG@**k equals one only when the positive items occupy ranks $1, \ldots, m$. The claims follow from Eq. 43.

□

Proposition 13 (AUC-prefix constants)**.** *The optimal constants from* **AUC** *to the prefix metrics are*

$$C^*_{\mathrm{AUC}\to\mathrm{Precision}\,@k} = \frac{m\ell}{k(|k-m|+1)}, \quad C^*_{\mathrm{AUC}\to\mathrm{Recall}\,@k} = \frac{\ell}{|k-m|+1}. \tag{53}$$

The reverse coefficient is finite only at $k = m$, where both **Precision@**m *and* **Recall@**m *have coefficient* $(n-1)/\ell$*.*

Proof. To prove the forward bounds, let $I = m\ell\,\mathrm{Reg}_{\mathrm{AUC}}$ be the inversion count and $t = E_k$. If $k \le m$, there are t negatives in the prefix and $m - k + t$ positives after it. If $k \ge m$, the numbers of neg and pos are $k - m + t$ and t. In either case, these cross-cutoff pairs give $I \ge t(|k-m|+t)$. For $t > 0$, we have

$$\frac{\mathrm{Reg}_{\mathrm{Precision}\,@k}}{\mathrm{Reg}_{\mathrm{AUC}}} = \frac{m\ell t}{kI} \le \frac{m\ell}{k(|k-m|+1)}. \tag{54}$$

Equality is attained with $t = 1$ by the sequences

$$\begin{cases} (1^{k-1}, 0, 1^{m-k+1}, 0^{\ell-1}), & k \le m, \\ (1^{m-1}, 0^{k-m+1}, 1, 0^{n-k-1}), & k \ge m. \end{cases} \tag{55}$$

The recall constant then follows by proportionality. In the reverse direction, finite constants exist only at $k = m$, where

$$C^*_{\mathrm{Precision}\,@m\to\mathrm{AUC}} = C^*_{\mathrm{Recall}\,@m\to\mathrm{AUC}} = \frac{n-1}{\ell}.$$

Indeed, if $t = E_m > 0$, the prefix contains t negatives and the suffix contains t positives. The inversion count satisfies

$$I \le t(m-t) + t^2 + t(\ell - t) = t(n-t).$$

Therefore,

$$\frac{\mathrm{Reg}_{\mathrm{AUC}}}{\mathrm{Reg}_{\mathrm{Precision}\,@m}} = \frac{I}{\ell t} \le \frac{n-1}{\ell}, \tag{56}$$

in which equality is attained by $(0, 1^{m-1}, 0^{\ell-1}, 1)$.

□

B.3.4 GLOBAL-TRUNCATED TRANSFER CONSTANTS

For $1 \le k < n$, define the quantities used in Table 1:

$$s_k = \min(k, m), \quad b_k = \max(k, m), \quad \delta_k = w_{s_k} - w_{b_k+1}, \quad Z_{m,k} = \sum_{i=1}^{s_k} w_i. \tag{57}$$

Let

$$F_k = \max\left\{ \frac{w_1}{b_k}, \frac{w_{s_k}}{b_k - s_k + 1}, \mathbf{1}_{\{k>m\}} \frac{w_1 - w_{m+1}}{m} \right\}, \quad \Lambda_{n,m} = \max_{1 \le a \le m} \frac{n-a}{w_a}. \tag{58}$$

For $k \ge m$, define

$$G_k = \begin{cases} \Lambda_{n,m}, & k = m, \\ \max\left\{ \Lambda_{n,m}, \frac{k-m}{w_m - w_k} \right\}, & k > m, \end{cases} \quad T_k = \begin{cases} 1 - w_n/w_1, & k = m, \\ 1, & k > m. \end{cases} \tag{59}$$

Proposition 14 (Sharp global–truncated constants)**.** *The optimal coefficients satisfy*

$$\begin{aligned} C^*_{\mathrm{NDCG}\to\mathrm{Precision}\,@k} &= \frac{Z_m}{k\delta_k}, \quad C^*_{\mathrm{NDCG}\to\mathrm{Recall}\,@k} = \frac{Z_m}{m\delta_k}, \\ C^*_{\mathrm{Precision}\,@m\to\mathrm{NDCG}} &= C^*_{\mathrm{Recall}\,@m\to\mathrm{NDCG}} = \frac{m(w_1 - w_n)}{Z_m}. \end{aligned} \tag{60}$$

Both **P/R**@k*-to-*$\mathbf{NDCG}$ *coefficients are infinite for $k \ne m$. Moreover,*

$$C^*_{\mathrm{AUC}\to\mathrm{NDCG}\,@k} = \frac{m\ell F_k}{Z_{m,k}}, \quad C^*_{\mathrm{NDCG}\to\mathrm{NDCG}\,@k} = \frac{Z_m w_{s_k}}{Z_{m,k}\delta_k}. \tag{61}$$

In reverse,

$$C^*_{\mathrm{NDCG}\,@k\to\mathrm{AUC}} = \begin{cases} +\infty, k < m, \\ Z_m G_k/(m\ell), & k \ge m, \end{cases} \quad C^*_{\mathrm{NDCG}\,@k\to\mathrm{NDCG}} = \begin{cases} +\infty, & k < m, \\ T_k, & k \ge m. \end{cases} \tag{62}$$

Proof. By Lemma 12, it suffices to maximize elementary regret ratios over $a \le m < c$. For count-based targets, positive regret requires $a \le k < c$, and $\min_{\substack{a \le s_k \\ c \ge b_k+1}} (w_a - w_c) = \delta_k$, attained at $(s_k, b_k + 1)$. At $k = m$, both elementary count regrets equal $1/m$, while $\max_{a \le m < c}(w_a - w_c) = w_1 - w_n$, attained at $(1, n)$. This proves Eq. 60. For $k < m$ and $k > m$, respectively, the pairs $(k+1, m+1)$ and $(1, m+1)$ have zero count regret but positive full-list regret.

For **AUC**-to-$\mathbf{NDCG}@k$ transfer, the elementary ratio is

$$\frac{m\ell}{Z_{m,k}} \times \begin{cases} 0, & a > k, \\ (w_a - w_c)/(c-a), & c \le k, \\ w_a/(c-a), & a \le k < c. \end{cases}$$

Since d_r decreases, $\max_{a \le m < c \le k} \frac{w_a - w_c}{c-a} = \frac{w_1 - w_{m+1}}{m}$ when $k > m$, which attained at $(1, m+1)$. In the crossing region,

$$\max_{a \le k < c} \frac{w_a}{c-a} = \max_{1 \le a \le s_k} \frac{w_a}{b_k + 1 - a} = \max\left\{ \frac{w_1}{b_k}, \frac{w_{s_k}}{b_k - s_k + 1} \right\},$$

where $a \le m < c$ remains implicit. The last equality follows because $f(x) = (b_k + 1 - x)\ln(x+1)$ satisfies $f''(x) = -\frac{b_k + x + 3}{(x+1)^2} < 0$, so its minimum occurs at an endpoint.

For $\mathbf{NDCG}$-to-$\mathbf{NDCG}@k$ transfer, the elementary ratio divided by $Z_m/Z_{m,k}$ is zero for $a > k$, one for $c \le k$, and $w_a/(w_a - w_c)$ for $a \le k < c$. The last expression increases with a and decreases with c, giving $\max_{a \le k < c} \frac{w_a}{w_a - w_c} = \frac{w_{s_k}}{\delta_k} > 1$, which attained at $(s_k, b_k + 1)$.

For either reverse direction, $k < m$ admits the zero-source, positive-target pair $(k+1, m+1)$. For $k \ge m$, $Z_{m,k} = Z_m$, and the elementary $\mathbf{NDCG}@k$-to-**AUC** ratio is

$$\frac{Z_m}{m\ell} \times \begin{cases} (c-a)/(w_a - w_c), & c \le k, \\ (c-a)/w_a, & c > k. \end{cases}$$

Monotonicity of the averages of d_r gives $\max_{a\le m<c\le k}\frac{c-a}{w_a-w_c}=\frac{k-m}{w_m-w_k}$ when $k>m$, which attained at (m,k), whereas

$$\max_{a\le m,\ c>k}\frac{c-a}{w_a}=\max_{1\le a\le m}\frac{n-a}{w_a}=\Lambda_{n,m}.$$

The first region is empty at $k=m$, yielding G_k.

Finally, the **NDCG@**k-to-**NDCG** ratio equals one for $c\le k$ and $1-w_c/w_a$ for $c>k$. The former region exists exactly when $k>m$, while the latter has maximum $1-w_n/w_1$, attained at $(1,n)$. □

B.3.5 Transfer across cutoffs

Proposition 15 (P/R@k at different cutoffs)**.** *For* $1\le k_1<k_2<n$*,*

$$C^*_{\text{Precision @}k_2\to\text{Precision @}k_1}=\begin{cases}k_2/k_1, & k_2\le m,\\ +\infty, & k_2>m,\end{cases}\quad C^*_{\text{Precision @}k_1\to\text{Precision @}k_2}=\begin{cases}k_1/k_2, & k_1\ge m,\\ +\infty, & k_1<m.\end{cases}$$

For Recall, each finite coefficient is replaced by one.

Proof. Monotonicity of the deficits gives $E_{k_1}\le E_{k_2}$ when $k_2\le m$ and $E_{k_2}\le E_{k_1}$ when $k_1\ge m$. Substituting $\text{Reg}_{\text{Precision @}k}=E_k/k$ and $\text{Reg}_{\text{Recall @}k}=E_k/m$ proves the finite bounds, attained by $\sigma^{1,n}$ with $E_{k_1}=E_{k_2}=1$. If $k_2>m$, the elementary pair $(a,c)=(\min(k_1,m),\max(k_1,m)+1)$ gives $E_{k_2}=0<E_{k_1}$. If $k_1<m$, the pair $(a,c)=(k_1+1,\max(k_2,m)+1)$ gives $E_{k_1}=0<E_{k_2}$. These prove the respective infinite cases. □

Proposition 16 (NDCG at different cutoffs)**.** *For* $1\le k_1<k_2\le n$*, let* $s_1=\min(k_1,m)$ *and* $b_1=\max(k_1,m)$*. The coefficient in Table 2 is*

$$D_{k_2\to k_1}:=C^*_{\text{NDCG @}k_2\to\text{NDCG @}k_1}=\begin{cases}\dfrac{Z_{m,k_2}}{Z_{m,k_1}}, & k_2\le m,\\ \dfrac{Z_m w_{s_1}}{Z_{m,k_1}(w_{s_1}-w_{b_1+1})}, & k_2>m.\end{cases}\tag{63}$$

In reverse,

$$C^*_{\text{NDCG @}k_1\to\text{NDCG @}k_2}=\begin{cases}+\infty, & k_1<m,\\ 1-w_n/w_1, & k_1=m,\ k_2=n,\\ 1, & k_1\ge m,\ (k_1,k_2)\ne(m,n).\end{cases}\tag{64}$$

Proof. For deeper-to-shallower transfer, elementary rankings with $a>k_1$ have zero target regret. For $a\le k_1$, the regret ratio is

$$\frac{Z_{m,k_2}}{Z_{m,k_1}}\times\begin{cases}1, & c\le k_1 \text{ or } c>k_2,\\ \dfrac{w_a}{w_a-w_c}, & k_1<c\le k_2.\end{cases}$$

The second region exists exactly when $k_2>m$, in which its factor exceeds one and is maximized at $(a,c)=(s_1,b_1+1)$. When $k_2\le m$, the baseline ratio is attained at $(1,m+1)$. This proves Eq. 63.

For reverse transfer, $k_1<m$ admits the pair $(k_1+1,m+1)$ with zero source regret and positive target regret. If $k_1\ge m$, both normalizers equal Z_m, and extending the cutoff cannot increase regret. The resulting bound of one is attained at $(1,n)$ when $k_2<n$, or at $(m,m+1)$ when $k_1>m$. In the remaining case $k_1=m$, $k_2=n$, the elementary ratio is $1-w_c/w_a$, whose maximum $1-w_n/w_1$ is attained at $(1,n)$. □

B.4 Proof of Theorem 9

Proof. Theorem 6 gives

$$C^*_{\text{AUC}\to\text{NDCG}}=\frac{\ell(w_1-w_{m+1})}{Z_m},\quad C^*_{\text{NDCG}\to\text{AUC}}=\frac{Z_m}{m(w_m-w_n)}.\tag{65}$$

With $w_r = \ln 2/\ln(r+1)$ and $b = \lfloor\sqrt{m}\rfloor$, monotonicity yields

$$\frac{m \ln 2}{\ln(m+1)} \le Z_m \le b + \frac{(m-b)\ln 2}{\ln(b+2)}.$$

Hence, as $m \to \infty$, $Z_m = \Theta\big(\frac{m}{\ln m}\big)$. If $m/n \to \alpha \in (0,1)$, then $m, \ell = \Theta(n)$, $Z_m = \Theta(n/\ln n)$, and $w_1 - w_{m+1} \to 1$. Moreover,

$$w_m - w_n = \frac{\ln 2 \, \ln((n+1)/(m+1))}{\ln(m+1)\ln(n+1)} = \Theta\left(\frac{1}{(\ln n)^2}\right), \tag{66}$$

since $\ln((n+1)/(m+1)) \to \ln(1/\alpha) > 0$. Substitution into Eq. 65 gives

$$C^*_{\mathrm{AUC}\to\mathrm{NDCG}} = \Theta(\ln n), \quad C^*_{\mathrm{NDCG}\to\mathrm{AUC}} = \Theta(\ln n).$$

If $m = m_0 \ge 1$ is fixed, then $\ell/n \to 1$ and $w_n \to 0$, so

$$\frac{C^*_{\mathrm{AUC}\to\mathrm{NDCG}}}{n} \longrightarrow \frac{w_1 - w_{m_0+1}}{Z_{m_0}} > 0, \quad C^*_{\mathrm{NDCG}\to\mathrm{AUC}} \longrightarrow \frac{Z_{m_0}}{m_0 w_{m_0}} > 0.$$

Thus $C^*_{\mathrm{AUC}\to\mathrm{NDCG}} = \Theta(n)$ and $C^*_{\mathrm{NDCG}\to\mathrm{AUC}} = \Theta(1)$. □

C Proofs and Extensions for Dynamic Metric Transfer

C.1 Gain Identities and Simultaneous Improvement

For a fixed evaluation list,

$$\Delta_t \mathcal{M} = \mathrm{Reg}_{\mathcal{M}}(\sigma_t) - \mathrm{Reg}_{\mathcal{M}}(\sigma_{t+1}), \quad E_r(\sigma_t) - E_r(\sigma_{t+1}) = h_r(t).$$

Subtracting the prefix-regret representations proves Eqs. 24 and 25.

If every $h_r(t) \ge 0$, non-negative prefix weights imply non-decrease of every additive metric considered. Conversely, choosing $a_j = \mathbf{1}_{\{j \le r\}}$ and $Z_{\mathcal{M}} = r$ gives **Precision**@r, whose gain is $h_r(t)/r$. Thus simultaneous non-decrease of the entire family requires $h_r(t) \ge 0$ for every $r < n$.

For $1 \le k \le n$, substitution of the truncated prefix weights gives

$$\begin{aligned} \Delta_t \,\mathrm{Precision}\,@k &= \frac{h_k(t)}{k}, \quad \Delta_t \,\mathrm{Recall}\,@k = \frac{h_k(t)}{m}, \\ \Delta_t \,\mathrm{NDCG}\,@k &= \frac{\sum_{r<k} d_r h_r(t) + w_k h_k(t)}{Z_{m,k}}, \end{aligned} \tag{67}$$

where $h_n(t) = 0$.

C.2 Transfer Efficiency and Metric Conflict

Suppose $h_r(t) \ge 0$ and $S_t = \sum_r h_r(t) > 0$. With $p_t(r) = h_r(t)/S_t$, Eq. 25 gives

$$\frac{\Delta_t \,\mathrm{NDCG}}{\Delta_t \,\mathrm{AUC}} = \frac{m\ell}{Z_m} \sum_r d_r p_t(r).$$

If p_t is supported on $\{a, \dots, b\}$, $\frac{m\ell d_b}{Z_m} \le \frac{\Delta_t \,\mathrm{NDCG}}{\Delta_t \,\mathrm{AUC}} \le \frac{m\ell d_a}{Z_m}$.

For two such updates on the same list, suppose

$$F_v(k) := \sum_{r \le k} p_v(r) \le \sum_{r \le k} p_u(r) =: F_u(k), \quad k < n. \tag{68}$$

Summation by parts yields

$$\mathbb{E}_p[d_r] = d_{n-1} + \sum_{k=1}^{n-2} (d_k - d_{k+1}) F_p(k).$$

Since $d_k - d_{k+1} \geq 0$, this proves

$$\frac{\Delta_v \,\mathrm{NDCG}}{\Delta_v \,\mathrm{AUC}} \leq \frac{\Delta_u \,\mathrm{NDCG}}{\Delta_u \,\mathrm{AUC}}.$$

For signed changes, substituting $h_r = h_r^+ - h_r^-$ gives Eq. 27, and hence

$$\Delta_t \,\mathrm{AUC} > 0, \quad \Delta_t \,\mathrm{NDCG} < 0 \quad \Longleftrightarrow \quad P > Q, \quad P\bar{d}_+ < Q\bar{d}_-.$$

For example, with $n = 8$ and $m = 3$, moving positive ranks from $(1, 6, 8)$ to $(2, 5, 7)$ gives $h_1 = -1$, $h_5 = h_7 = 1$, and $\Delta \,\mathrm{AUC} = \frac{1}{15} > 0, \Delta \,\mathrm{NDCG} = \frac{-d_1 + d_5 + d_7}{Z_3} < 0$.

For an additive metric, let $b_r^{\mathcal{M}} = a_r - a_{r+1}$. Under nonnegative prefix changes with $S_t > 0$,

$$\frac{\Delta_t \mathcal{M}}{\Delta_t \,\mathrm{AUC}} = \frac{m\ell}{Z_{\mathcal{M}}} \mathbb{E}_{p_t}[b_r^{\mathcal{M}}].$$

For signed changes, define

$$\bar{b}_{\mathcal{M},+} = \frac{\sum_r b_r^{\mathcal{M}} h_r^+}{P}, \quad \bar{b}_{\mathcal{M},-} = \frac{\sum_r b_r^{\mathcal{M}} h_r^-}{Q},$$

setting each average to zero when its denominator vanishes. Then

$$\Delta_t \mathcal{M} = \frac{P\bar{b}_{\mathcal{M},+} - Q\bar{b}_{\mathcal{M},-}}{Z_{\mathcal{M}}}.$$

Thus $\Delta_t \mathcal{A} > 0$ and $\Delta_t \mathcal{B} < 0$ exactly when $P\bar{b}_{\mathcal{A},+} > Q\bar{b}_{\mathcal{A},-}, P\bar{b}_{\mathcal{B},+} < Q\bar{b}_{\mathcal{B},-}$.

Precision@k and **Recall@k** use $b_r = \mathbf{1}_{\{r=k\}}$, while **NDCG@$k$** uses

$$b_r = d_r \mathbf{1}_{\{r<k\}} + w_k \mathbf{1}_{\{r=k\}}, \quad r < n.$$

These weights need not decrease with r, so the stochastic-order comparison above does not generally extend to truncated metrics. At a common cutoff,

$$\Delta_t \,\mathrm{Recall}\,@k = \frac{k}{m} \Delta_t \,\mathrm{Precision}\,@k,$$

but both may conflict with **NDCG@k**. For $m = 2$, $\ell = 3$, and $k = 4$, the update

$$(1, 0, 0, 0, 1) \longrightarrow (0, 0, 1, 1, 0)$$

gives

$$\Delta \,\mathrm{Precision}\,@4 = \frac{1}{4}, \quad \Delta \,\mathrm{Recall}\,@4 = \frac{1}{2}, \quad \Delta \,\mathrm{NDCG}\,@4 = \frac{w_3 + w_4 - w_1}{Z_{2,4}} < 0.$$

C.3 Local Exchanges and Global Regret

A favorable adjacent swap at boundary r gives $h_j = \mathbf{1}_{\{j=r\}}$, so $\Delta_r \,\mathrm{AUC} = \frac{1}{m\ell}, \Delta_r \,\mathrm{NDCG} = \frac{d_r}{Z_m}$. Since every boundary is realizable across rankings and d_r decreases,

$$\max_{1 \leq r < n} \frac{\Delta_r \,\mathrm{NDCG}}{\Delta_r \,\mathrm{AUC}} = \frac{m\ell d_1}{Z_m}, \quad \max_{1 \leq r < n} \frac{\Delta_r \,\mathrm{AUC}}{\Delta_r \,\mathrm{NDCG}} = \frac{Z_m}{m\ell d_{n-1}}. \tag{69}$$

If $m/n \to \alpha \in (0, 1)$, then $m\ell = \Theta(n^2)$ and $Z_m = \Theta(n/\ln n)$. Therefore,

$$\frac{m\ell d_1}{Z_m} = \Theta(n \ln n), \quad C^*_{\mathrm{AUC} \to \mathrm{NDCG}} = \Theta(\ln n),$$

where the second relation follows from Theorem 9.

C.4 Score Dynamics and Population BCE

For $L = L(\boldsymbol{s}_{\boldsymbol{\theta}})$ and $J = \partial \boldsymbol{s}_{\boldsymbol{\theta}} / \partial \boldsymbol{\theta}$, the chain rule gives $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -J^\top \nabla_{\boldsymbol{s}} L$, $\frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}\tau} = -JJ^\top \nabla_{\boldsymbol{s}} L$. Subtracting the j-th coordinate from the i-th proves Eq. 29. For independent logits, the summed conditional **BCE** is

$$L_{\mathrm{BCE}}(\boldsymbol{s}) = \sum_i \big[\ln(1 + e^{s_i}) - \eta_i s_i\big].$$

Writing $\rho(s) = (1 + e^{-s})^{-1}$ gives $\frac{\mathrm{d}s_i}{\mathrm{d}\tau} = \eta_i - \rho(s_i)$, $\frac{\mathrm{d}q_{ij}}{\mathrm{d}\tau} + c_{ij}(\tau) q_{ij} = \eta_i - \eta_j$, where

$$c_{ij}(\tau) = \int_0^1 \rho'\big(s_j(\tau) + u q_{ij}(\tau)\big)\,\mathrm{d}u \in (0, 1/4].$$

Let $A_{ij}(\tau) = \int_0^\tau c_{ij}(u)\,\mathrm{d}u$. The solution is

$$q_{ij}(\tau) = e^{-A_{ij}(\tau)} \left[q_{ij}(0) + (\eta_i - \eta_j) \int_0^\tau e^{A_{ij}(u)}\,\mathrm{d}u \right].$$

For $\eta_i > \eta_j$, the bracket is strictly increasing, so a margin crosses zero at most once, from negative to positive. For $\eta_i = \eta_j$, its sign is preserved. When $\eta_i = y_i$, every opposite-label crossing promotes the positive item, making all prefix counts non-decreasing. The gain identities then imply simultaneous metric non-decrease.

For gradient descent, $s_i^+ = s_i + \gamma(\eta_i - \rho(s_i))$. If $\eta_i > \eta_j$ and $q_{ij} \geq 0$,

$$q_{ij}^+ \geq (1 - \gamma/4) q_{ij} + \gamma(\eta_i - \eta_j) > 0 \quad (0 < \gamma \leq 4),$$

by the $1/4$-Lipschitz property of ρ. Thus this sufficient step-size condition preserves correct orderings. For mean **BCE**, the effective step size is γ/n.

Shared parameters can reverse the boundary direction. Consider

$$s_1 = \theta_1, \quad s_2 = 2\theta_1 + \theta_2, \quad \eta_1 = 0.9, \quad \eta_2 = 0.8.$$

At $\boldsymbol{\theta} = 0$,

$$JJ^\top = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}, \quad \nabla_{\boldsymbol{s}} L = \begin{pmatrix} -0.4 \\ -0.3 \end{pmatrix}, \quad \frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}\tau} = \begin{pmatrix} 1 \\ 2.3 \end{pmatrix}.$$

Hence $\mathrm{d}q_{12}/\mathrm{d}\tau = -1.3 < 0$ despite $\eta_1 > \eta_2$. By continuity, sufficiently close correctly ordered initializations can cross into the incorrect ordering, while

$$\frac{\mathrm{d}L}{\mathrm{d}\tau} = -\|\nabla_{\boldsymbol{\theta}} L\|^2 \leq 0.$$

D Experimental Details

D.1 Exact Enumeration and Scaling

Enumeration protocol. For every $n \in \{2, \ldots, 14\}$ and $m \in \{1, \ldots, n-1\}$, we enumerate the $\binom{n}{m}$ binary label sequences. Permuting items with the same label leaves all metrics considered here unchanged, so this exhausts the relevant ranking states without enumerating redundant item permutations. We evaluate **AUC**, **NDCG**, **Precision**@k, **Recall**@k and **NDCG**@k for every $1 \leq k \leq n$. **Regret** is the gap to the fixed-list oracle.

For each ordered metric pair, we first search for a ranking with zero source regret and positive target regret. Such a witness rules out every finite linear coefficient. Otherwise, we maximize the target-to-source regret ratio over rankings with positive source regret, assigning coefficient zero when the target regret vanishes identically. We compare this value with a closed-form sharp constant where one is established. For remaining pairs, we check only the predicted finiteness condition.

The enumeration covers 32,738 distinct binary sequences, 22,750 ordered-pair checks at a common cutoff, and 6,930 additional cross-cutoff checks for $n \leq 10$. All checks pass, and the maximum absolute discrepancy for a closed-form constant is 7.78×10^{-12}. Each checked pair is archived with a maximizing ranking or a zero-source-regret witness. These finite checks corroborate the formulas; the uniform conclusions follow from the proofs.

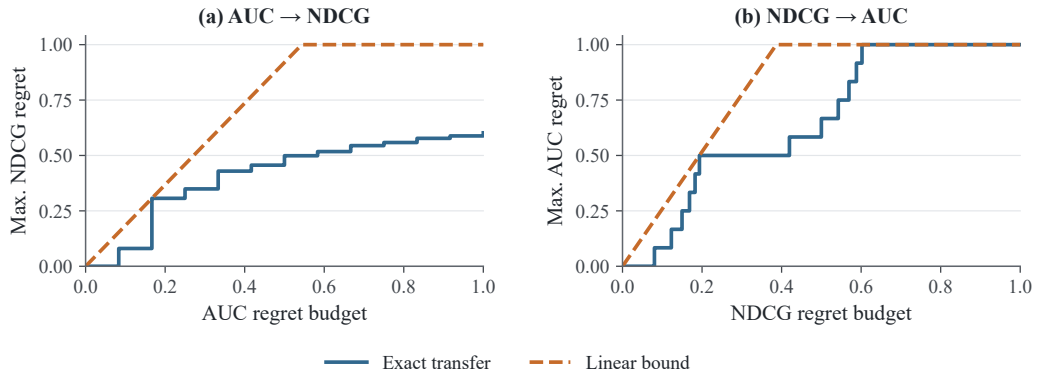


Figure 2: Exact transfer functions and optimal linear bounds for $n = 8, m = 2$. The sharp coefficients define linear envelopes of the staircase transfer functions.

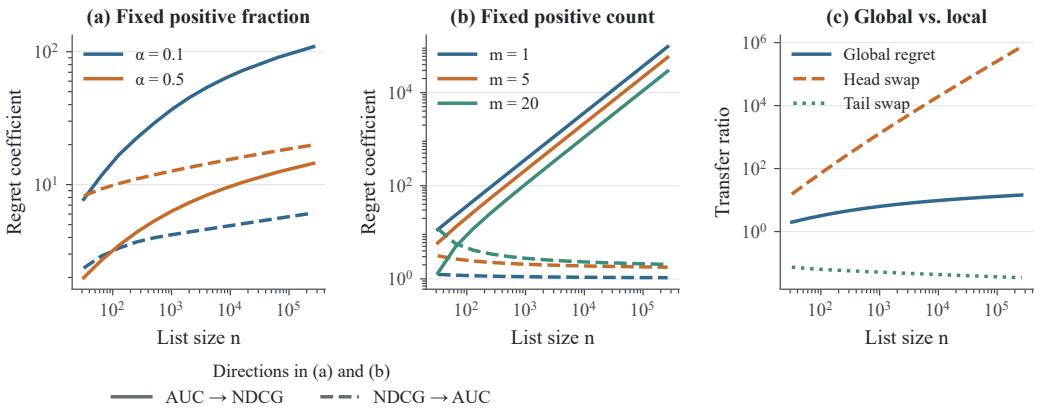


Figure 3: Global and local transfer scaling. (a) Fixed positive fractions. (b) Fixed positive counts. (c) Global **AUC**-to-**NDCG** coefficients and adjacent-swap gain ratios at $\alpha = 0.5$. Colors and line styles identify the parameters and transfer directions in (a)–(b).

Transfer functions versus linear envelopes. For $n = 8$ and $m = 2$, we explicitly evaluate the transfer function at every attainable source-regret value in both **AUC-NDCG** directions. Figure 2 displays the resulting staircase and its optimal clipped linear envelope.

Scaling protocol. We use $n \in \{2^5, 2^6, \ldots, 2^{14}, 2^{16}, 2^{18}\}$, with $m \in \{1, 5, 20\}$ for fixed-count sequences and $m = \lfloor \alpha n \rfloor$, $\alpha \in \{0.1, 0.5\}$, for proportional sequences. This gives 60 settings. For each setting, we calculate the theoretical constants and independently evaluate their attaining label sequences, $(0, 1^m, 0^{\ell-1})$ and $(1^{m-1}, 0^{\ell}, 1)$. We also construct feasible rankings containing a negative–positive adjacent pair at the first or last boundary, exchange that pair, and compute both metric increments directly.

Figure 3 illustrates Theorem 9 and the local/global distinction in Section 5.2. Both global coefficients have logarithmic order in the proportional regime, while the fixed-count regime exhibits the linear-versus-bounded directional separation. The head-swap gain ratio in the proportional regime has the larger order $n \log n$. We additionally archive constants divided by their predicted rates, rather than estimating a power-law exponent for logarithmic growth. These curves concern fixed-list coefficients, not rates of loss convergence.

Table 4: Number of trajectories exhibiting at least one event out of 100 per setting. The target order is the order of $\boldsymbol{y}$ or $\boldsymbol{\eta}$, respectively. No run exhibits a recorded **BCE** increase.

BCE target	Scores	New inversion	Prefix decrease	Metric conflict
Binary	Independent	0	0	0
Binary	Coupled	22	18	0
Fractional	Independent	0	61	16
Fractional	Coupled	100	91	7

D.2 Controlled Optimization and Structural Updates

Matched score expressivity. We conduct 100 base trials with 16 scores. Each trial draws $\eta_i \sim \mathrm{Uniform}(0.05, 0.95)$, initial scores $s_i^0 \sim \mathcal{N}(0, 1.5^2)$, and a matrix $B = I + E$ with independent $E_{ij} \sim \mathcal{N}(0, 0.25^2)$. Every sampled B is verified to be invertible. Independent scores use $K = I$, while coupled scores use $K = BB^\top/\lambda_{\max}(BB^\top)$, corresponding to a rescaled invertible linear score parameterization. Both settings start from the same score vector and can represent every vector in $\mathbb{R}^{16}$. For the **BCE** objective, updates are

$$\boldsymbol{s}^{t+1} = \boldsymbol{s}^t - \gamma K\big(\rho(\boldsymbol{s}^t) - \boldsymbol{z}\big), \quad \gamma = 0.5, \tag{70}$$

for 1,500 steps. We test both $\boldsymbol{z} = \boldsymbol{\eta}$ and a deterministic binary target $\boldsymbol{z} = \boldsymbol{y}$. The diagnostic binary vector is formed by thresholding η_i at 0.5, with its first and last entries set to zero and one to ensure both classes. It is held fixed throughout each trial. In the fractional-target setting, this vector is used only to illustrate realized-label metric behavior.

The independent-score step size satisfies the sufficient finite-step ordering condition $\gamma \le 4$ for the summed objective. At every step we inspect all score pairs for newly introduced target-order inversions, using tolerance 10^{-10}. We also record prefix decreases and strict **AUC/NDCG** conflicts for the fixed diagnostic labels. Counts describe endpoint comparisons at the recorded steps.

Table 4 shows that independent optimization introduces no target-order inversions, with binary targets, prefix counts also remain non-decreasing. Coupled scores introduce inversions despite decreasing **BCE**. With fractional targets, independent scores preserve the target order while still producing conflicts on the fixed binary evaluation labels. Metric conflict therefore depends on the resulting prefix changes, beyond whether target-order inversions occur.

Structural updates. Separately, for $n = 128, m = 32$, we construct 100 feasible rankings at each of boundaries 1, 65, and 127 containing an adjacent negative–positive pair, then perform the favorable swap. A fixed mixed update, with positive ranks $(1, 6, 8) \to (2, 5, 7)$ on a length-eight list, illustrates a direction conflict. These experiments check the gain formulas on actual rankings rather than unconstrained synthetic prefix vectors. The maximum identity residual is below 2.4×10^{-16}.

D.3 MovieLens Protocol

We use **MovieLens-1M** (Harper & Konstan, 2016), containing 1,000,209 ratings from 6,040 users over 3,706 observed items. Ratings of at least four define positive labels, while lower observed ratings define negatives. Unrated items are not assigned negative labels. For each user, ratings are sorted by timestamp. We allocate $\lfloor 0.70 n_u \rfloor$ interactions to training, $\lfloor 0.15 n_u \rfloor$ to validation, and the remainder to testing. The resulting interaction counts are 697,378, 147,221, and 155,610, respectively.

We evaluate the common cohort of users whose validation and test lists each contain at least ten candidates and both label classes. This retains 3,518 users. Their test lists have lengths between 11 and 348, with median 27. These dense, observed-rating candidate pools support an analysis of metric changes on real training trajectories.

The score model is $s_{ui} = \boldsymbol{p}_u^\top \boldsymbol{q}_i + b_u + b_i + b_0$, with 16-dimensional embeddings initialized independently from $\mathcal{N}(0, 0.1^2)$ and zero biases. For seeds 27, 28, and 29, **BCE** and **BPR** start from identical parameters. Both use **Adam** with learning rate 0.005, batch size 1,024, default moments $(0.9, 0.999)$, and $\epsilon = 10^{-8}$, for 30 epochs. The regularizer is $\lambda/2$ times the sum of squared embeddings and user/item biases, with $\lambda = 10^{-5}$.

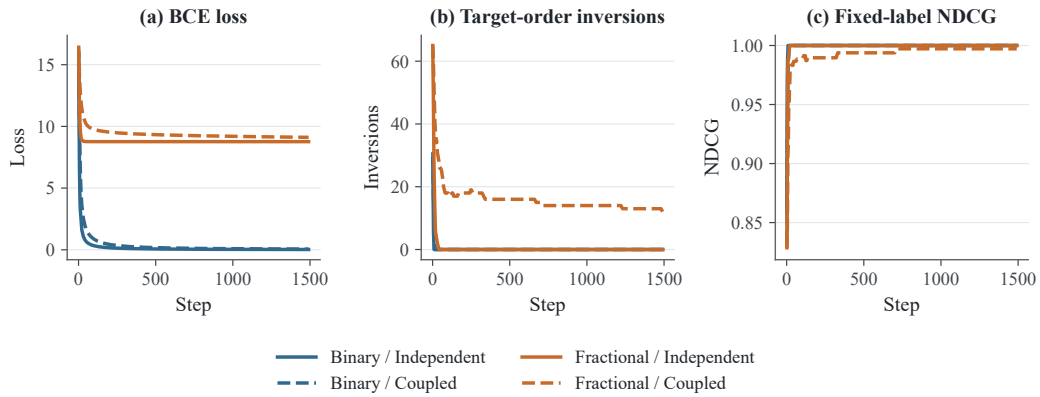


Figure 4: Controlled **BCE** optimization with independent and coupled scores. Each curve shows the first trial of its setting. Inversions use the training target; **NDCG** uses fixed binary labels. Aggregate results appear in Table 4.

Table 5: Conflict and mixed-prefix rates across reporting periods. Values are percentages of moving test-user windows, reported as mean ± standard deviation across three seeds.

Loss	Epochs	AUC/NDCG conflict	AUC/NDCG@10 conflict	Mixed prefixes
BCE	1–30	14.03 ± 0.09	14.58 ± 0.12	61.45 ± 0.23
BCE	2–30	14.19 ± 0.09	14.78 ± 0.12	62.40 ± 0.22
BCE	21–30	13.65 ± 0.21	13.87 ± 0.16	61.25 ± 0.20
BPR	1–30	17.64 ± 0.15	19.29 ± 0.14	68.04 ± 0.17
BPR	2–30	17.91 ± 0.16	19.65 ± 0.15	69.10 ± 0.16
BPR	21–30	17.69 ± 0.33	19.55 ± 0.27	68.84 ± 0.36

BCE averages the logistic loss over training ratings, while **BPR** (Rendle et al., 2009) uses the pairwise logistic objective $\log(1 + \exp(-(s_{ui^+} - s_{ui^-})))$. The objectives shares architecture, initialization, epoch count, batch size, and optimizer settings. We save initial scores and scores after every epoch. Our reported analysis focuses on **AUC**, **NDCG**, and **NDCG@10**. Performance is averaged uniformly across evaluation users.

D.4 Trajectory Diagnostics

For each test user and consecutive epoch pair, we compute metric increments and net prefix changes h_r. A moving window has $P + Q > 0$, and mixed prefix changes have $P, Q > 0$. Metric conflicts satisfy $\Delta\,\mathrm{AUC}\ \Delta\,\mathrm{NDCG} < -10^{-24}$; truncated conflicts replace **NDCG** with **NDCG@10**. The main results use epochs 2-30. Table 5 also reports all epochs and the final ten epochs. Rates are calculated over moving windows within each seed, then summarized by their mean and sample standard deviation across three seeds.

Conflicts occur under both objectives and persist through the final ten epochs. Mixed prefix changes are more frequent than conflicts: their positions and magnitudes determine whether the two metrics move in opposite directions. No single-positive test list exhibits an **AUC/NDCG** conflict. The prefix identities recover metric increments with maximum absolute residual 7.65×10^{-16}.

To locate the contributions, define $r_h = \min(10, n-1)$ and $r_m = \min(n-1, \max(10, \lceil n/2 \rceil))$. Head, middle, and tail boundaries are $1 \le r \le r_h$, $r_h < r \le r_m$, and $r_m < r < n$. For each region $\mathcal{R}$, the signed NDCG contributions are

$$\frac{1}{Z_m} \sum_{r \in \mathcal{R}} d_r h_r^+, \quad -\frac{1}{Z_m} \sum_{r \in \mathcal{R}} d_r h_r^-.$$

Figure 5 shows a uniform reservoir sample of up to 5,000 moving windows and mean regional contributions over all moving windows in each conflict category, pooled across seeds and objectives.

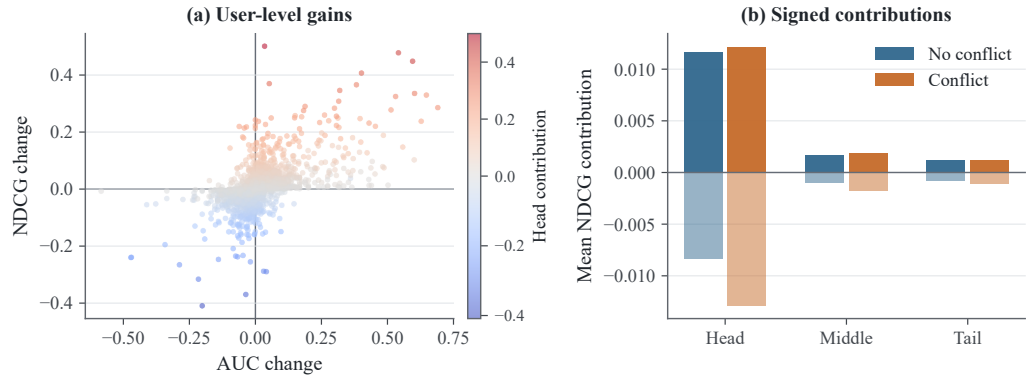


Figure 5: Metric gains and regional **NDCG** contributions. The scatter shows sampled moving user windows; bars show mean positive and negative contributions for windows with and without metric conflict.

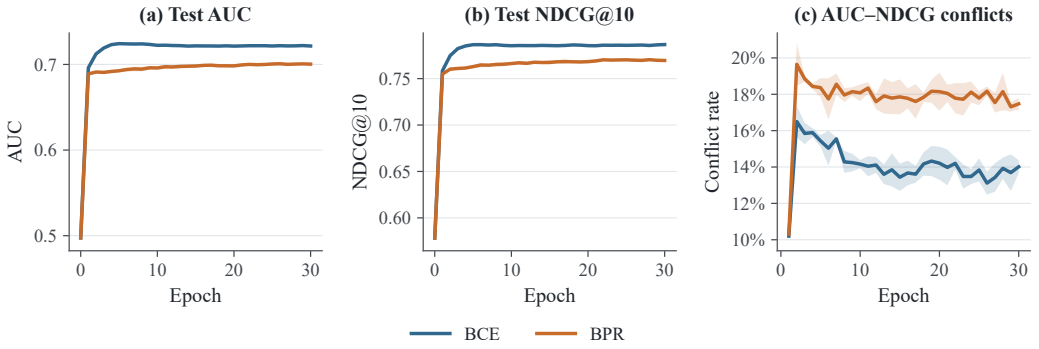


Figure 6: Test performance and **AUC-NDCG** conflicts across training epochs. Curves and shaded regions show means and standard deviations over three seeds. Conflict rates use moving multi-positive user lists.

Figure 6 summarizes performance and conflict rates throughout training.

D.5 Checkpoint Selection and Test Performance

We compare checkpoints selected by validation **AUC** and validation **NDCG**@10, choosing the earliest epoch in case of ties. Table 6 reports their test performance across three seeds.

Both selection criteria yield similar test performance. Selecting by NDCG@10 increases mean test NDCG@10 from 0.7863 to 0.7867 for BCE and from 0.7697 to 0.7698 for BPR. Using 1,000 paired user-bootstrap resamples per run, all six 95% percentile intervals for the NDCG@10 difference contain zero (Table 7). The differences are small and vary in sign across seeds.

Table 6: Test performance after validation-based checkpoint selection; mean $\pm$ sample standard deviation across three seeds.

Loss	Validation criterion	AUC	NDCG	NDCG@10
BCE	AUC	0.7242 ± 0.0005	0.8896 ± 0.0005	0.7863 ± 0.0014
BCE	NDCG@10	0.7240 ± 0.0007	0.8895 ± 0.0005	0.7867 ± 0.0006
BPR	AUC	0.7004 ± 0.0013	0.8815 ± 0.0004	0.7697 ± 0.0012
BPR	NDCG@10	0.6994 ± 0.0012	0.8815 ± 0.0009	0.7698 ± 0.0009

Table 7: Effect of selecting by validation NDCG@10 instead of validation AUC. Differences are in test NDCG@10; intervals are paired user-bootstrap intervals within each run, not uncertainty intervals across seeds.

Loss/seed	AUC epoch	NDCG epoch	Difference	95% interval
BCE/27	6	5	$+0.00102$	$[-0.00061, 0.00251]$
BCE/28	6	5	-0.00045	$[-0.00195, 0.00101]$
BCE/29	5	7	$+0.00064$	$[-0.00111, 0.00241]$
BPR/27	26	24	-0.00005	$[-0.00251, 0.00225]$
BPR/28	24	28	-0.00046	$[-0.00303, 0.00206]$
BPR/29	29	23	$+0.00083$	$[-0.00174, 0.00359]$