FAIR RANKING WITH ATTENTION UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Exposure fairness asks that each group of items in a ranked list receive a target share of user attention. Whether a ranking meets this target depends on an attention curve, which specifies how attention is distributed across positions, yet the curve used for evaluation may differ from the one users actually follow. This raises the question of what a fairness score computed under one curve, or a few, guarantees when the true attention curve is unknown. We first characterize the blind spot of a finite set of checked curves, which is the largest exposure disparity that can remain hidden while all checks pass at a given tolerance. For arbitrary targets and multiple, possibly overlapping groups, an exact formula is derived to compute the blind spot from the checked curves within the tolerance. We show that commonly used attention curves can still leave substantial exposure disparities undetected. For exact checks, we also give upper and lower bounds on the number of curves needed to guarantee a desired blind spot. Finally, we design a robust ranking linear program that maximizes utility while controlling exposure disparity under an unknown attention curve. Across Amazon ESCI, TREC Fair Ranking 2022, and German Credit, we show that policies that are fair under a single logarithmic attention curve are still off target on average by 36%–44% of total attention under some monotone attention curve. Under the worst such curve for each query, these policies deliver on average only 42%–64% of the target exposure. They remain off target by 7%–11% of total attention under attention curves estimated from real-world click logs, under which they deliver on average only 80%–91% of the target exposure. By contrast, our robust linear program keeps the mean deviation below 0.06% of total attention, so that it delivers at least 99.8% of the target exposure, with an additional loss of about 1% or less in ranking utility relative to the single-curve policies.

1 INTRODUCTION

Ranking systems decide who gets seen. A *ranking policy* specifies, for each query, a distribution over ranked lists. Exposure-based fairness (Biega et al., 2018; Singh & Joachims, 2018; Diaz et al., 2020) makes this precise by weighting each position with an *attention curve*, the share of user attention that reaches that position, and asking that each group's attention-weighted exposure match a target share. Every such fairness number is conditional on the curve, and the curve is a modeling choice: the logarithmic discount of discounted cumulative gain (DCG), rank-biased precision (RBP) with an assumed probability of reading on, or the discount used by the Text REtrieval Conference (TREC) Fair Ranking track. Real attention depends on the device, the layout, the query, and the user population. Exposure estimates can be unreliable (Heuss et al., 2022), and plausible curves reverse fairness conclusions on the same rankings (Sapiezynski et al., 2019; Raj & Ekstrand, 2022).

How much can one curve hide? Take a list of ten positions and two groups, A and B, that should each receive half of the attention. Under the DCG curve, position one receives 22% of the attention. Consider the policy that always shows an A item first and then, at each of positions 2–10, shows an A item with probability about 0.36. Its DCG-weighted exposure for A is exactly one half, so its disparity under the DCG curve is zero. Nevertheless, a user who reads only the first result sees group A only (Figure 1). The same issue arises for nearly every published curve we surveyed (Section 3.2).

The above example separates two questions: whether a policy is fair under a specified curve, and what that check guarantees under other curves. In this paper, we study the second question and allow the true curve to be any curve whose attention weights are nonnegative, non-increasing, and normalized

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.

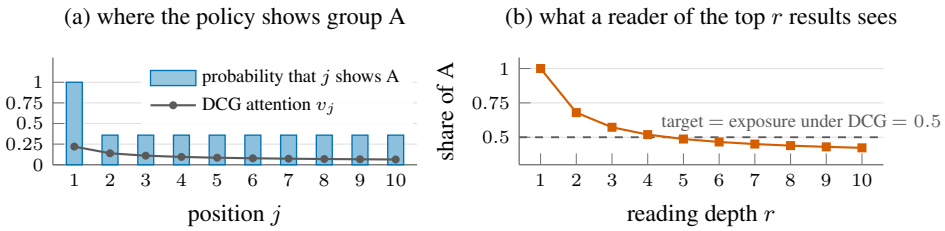


Figure 1: **A policy that is fair under DCG but not for readers of the top results** ($k = 10$, two groups with target 0.5 each). **(a)** The policy always shows group A at position 1 and shows it with probability 0.359 elsewhere. A's exposure under DCG equals its target. **(b)** Share of attention received by group A among the top r positions. Readers of only the top result see A alone, and readers of the whole list see A under-represented.

over the displayed positions. Attention may vary between queries, but is independent of which list the policy draws for a given query. Our main contributions are:

1. **Quantifying what checked curves can miss.** We derive an exact formula for the largest exposure disparity that can remain hidden when every checked curve reports disparity at most δ, which we call the *blind spot* of the checked curves. The formula applies to arbitrary targets and multiple, possibly overlapping groups. We use it to show that commonly used attention curves can leave substantial disparities undetected (Section 3.2). For exact checks ($\delta = 0$), we give upper and lower bounds on the number of curves needed to guarantee a desired blind spot, together with a construction attaining the upper bound. For two groups with equal targets, the bounds match, and our construction uses the fewest possible curves. For a fixed positive blind spot, the number of curves needed grows logarithmically with the length of the ranking (Section 3.3).
2. **Robust ranking under unknown attention.** We design a linear program (LP) that maximizes ranking utility while controlling the largest exposure disparity under any considered attention curve. Thus, robustness to attention uncertainty can be enforced directly at ranking time without changing the relevance scorer (Section 4.1).
3. **Empirical evidence.** Across Amazon ESCI, TREC Fair Ranking 2022, and German Credit, the policies returned by the fair ranking LP of Singh & Joachims (2018), which are fair under a single logarithmic attention curve, still deviate from their target on average by 36%–44% of total attention under some monotone attention curve, and by 7%–11% under attention curves estimated from click logs. Our robust LP keeps the mean deviation below 0.06% of total attention, with an additional loss of about 1% or less in ranking utility relative to that LP (Section 4).

2 Setting

Rankings and policies. A query z comes with n candidate items, of which $k \leq n$ are displayed, at positions 1 (top) to k. A ranking policy is an $n \times k$ matrix P, where P_{ij} is the probability that item i is shown at position j. Every position shows exactly one item, and no item is shown more than once. Hence, the set of feasible policies is

$$\mathcal{P}_{n,k} = \left\{ P \in \mathbb{R}_+^{n \times k} : \sum_{i=1}^{n} P_{ij} = 1 \;\; (j \in [k]), \quad \sum_{j=1}^{k} P_{ij} \leq 1 \;\; (i \in [n]) \right\}. \tag{1}$$

Conversely, every such matrix arises in this way from some random ranking (Proposition A.1).

Groups, occupancy, and targets. Items belong to groups $G_1, \ldots, G_m$, e.g., the brands in a product search or topical categories in content ranking, and groups may overlap. The *occupancy* of group g at position j, $x_{g,j} = \sum_{i \in G_g} P_{ij}$, is the probability that position j shows a member of g. The *target* $p_g \in [0, 1]$ is the share of attention that group g should receive.

Attention curves. An attention curve $v = (v_1, \ldots, v_k)$ specifies the fraction of user attention allocated to each displayed position. We consider *monotone* curves, which are nonnegative, normalized, and non-increasing in position: $\mathcal{V}_k = \{v \in \mathbb{R}_+^k : v_1 \geq \cdots \geq v_k, \; \sum_j v_j = 1\}$. This captures the minimal

assumption that higher-ranked positions receive at least as much attention as lower-ranked ones. The attention curve may vary across queries or user populations, but we assume that it does not depend on which items are displayed.

Exposure and disparity. Under an attention curve v, the *exposure* of group g is defined as $\sum_j v_j x_{g,j}$, i.e., the expected share of total attention received by g. Its *disparity* is the distance of its exposure from the target, $D_g(z; v) = \left|\sum_j v_j x_{g,j}(z) - p_g\right|$. The policy is fair under v when all disparities are zero, and it passes at tolerance δ when they all lie within δ.

Querywise risk. Because the deployed curve is unknown, we evaluate a policy by its *querywise risk*, the expected worst-case disparity over all curves in $\mathcal{V}_k$. Writing $\mathbb{E}_z$ for the average over queries drawn from the query distribution $\mathcal{D}$,

$$\mathcal{R}_{\mathrm{Q}} = \mathbb{E}_{z\sim\mathcal{D}}\left[\sup_{v\in\mathcal{V}_k} \max_g D_g(z; v)\right]. \tag{2}$$

Ranking quality. Each item has a relevance score s_i from a fixed scoring model, and each position j carries a utility weight w_j that decreases with position. The *utility* of a policy is $U(P) = \sum_{i,j} s_i w_j P_{ij}$, the score of what is shown discounted by where it is shown.

3 Measuring Fairness under Unknown Attention

In this section, we study what exposure fairness guarantees remain when the deployed attention curve is unknown. We first show that the largest disparity over all considered curves is attained by one of the k prefix curves. We then quantify what can remain hidden when only finitely many curves are checked, and show that commonly used curves can still leave substantial blind spots. Finally, we give upper and lower bounds on the smallest number of curves needed to guarantee a desired blind spot, together with a construction attaining the upper bound. All proofs are given in the appendix.

3.1 From monotone attention to prefix checks

For $r \in [k]$, let $u^r = (\frac{1}{r}, \ldots, \frac{1}{r}, 0, \ldots, 0)$ be the curve of a user who reads exactly the top r results and pays equal attention to each. We call the top r positions the *prefix* of depth r, and u^r the *prefix curve* of depth r. Under u^r, group g's exposure is its average occupancy over that prefix, and its disparity is the *prefix disparity* at depth r,

$$d_{g,r}(z) = D_g(z; u^r) = \left|\tfrac{1}{r}\textstyle\sum_{j\le r} x_{g,j}(z) - p_g\right|.$$

We use a standard decomposition of monotone weights (Acerbi, 2002; Moffat et al., 2017).

Lemma 1 (Scroll-depth decomposition)**.** *Every $v \in \mathcal{V}_k$ can be written as $v = \sum_{r=1}^{k} \lambda_r u^r$ with weights $\lambda_r = r(v_r - v_{r+1}) \ge 0$ (where $v_{k+1} = 0$) that sum to one.*

The coefficients λ_r describe how attention mass is distributed across prefix depths. By Lemma 1, the signed deviation $\sum_j v_j x_{g,j} - p_g$ is a convex combination of the k signed prefix deviations, with weights $\lambda_1, \ldots, \lambda_k$. Hence the disparity of group g under any $v \in \mathcal{V}_k$ is at most $\max_r d_{g,r}$, and the prefix curve of the maximizing depth attains this bound. The worst-case disparity over $\mathcal{V}_k$ is therefore the largest prefix disparity (Appendix A).

Theorem 2 (Exact querywise certificate)**.** *For every policy and query distribution,*

$$\mathcal{R}_{\mathrm{Q}} = \mathbb{E}_z \max_{g,r} d_{g,r}(z). \tag{3}$$

3.2 Blind spots of finite curve sets

In practice, exposure fairness is typically evaluated under a specified attention model or discount curve (Singh & Joachims, 2018; Diaz et al., 2020; Ekstrand et al., 2023). Prior work shows that fairness measurements can change substantially with the assumed attention model (Sapiezynski et al., 2019; Raj & Ekstrand, 2022). Since the deployed attention curve is unknown, one may also check

several plausible curves rather than rely on a single one. Hence, we ask how much disparity can still remain hidden even when all of these curves report little or no disparity.

Fix a query with groups $G_1, \dots, G_m$ and targets $p_1, \dots, p_m$. Let $x_g = (x_{g,1}, \dots, x_{g,k})$ denote the occupancy vector of group g, and let $\mathcal{X}$ be the set of occupancy profiles realizable by ranking policies. Suppose we check a finite set of curves $\mathcal{S} = \{v^1, \dots, v^q\} \subseteq \mathcal{V}_k$ and require every checked disparity to be at most δ, which we assume at least one policy satisfies. The *blind spot* of $\mathcal{S}$ is the largest disparity that can remain hidden while all these checks are satisfied. By Theorem 2, it is enough to consider prefix curves, so we have

$$B(\mathcal{S};\delta) = \max_{x\in\mathcal{X}} \left\{ \max_{g\in[m],\, r\in[k]} \left|\langle u^r, x_g\rangle - p_g\right| \;\middle|\; \left|\langle v^\ell, x_g\rangle - p_g\right| \le \delta \quad \forall g\in[m],\ \ell\in[q] \right\}. \tag{4}$$

We first consider two groups with equal targets, and prove that the blind spot has a simple form when each group has sufficiently many candidates. Write $\|\cdot\|_1$ for the ℓ_1 norm.

Theorem 3 (Blind spot of two groups with equal targets)**.** *Suppose two groups partition the items, each contains at least k items, and $p_1 = p_2 = \frac{1}{2}$. Then*

$$B(\mathcal{S};\delta) = \max_{r\in[k]} \min_{\alpha\in\mathbb{R}^q} \left\{ \frac{1}{2} \left\| u^r - \sum_{\ell=1}^{q} \alpha_\ell v^\ell \right\|_1 + \delta\|\alpha\|_1 \right\}. \tag{5}$$

The first term measures how accurately the checked curves approximate a prefix curve, and the second measures how their tolerated errors accumulate. Under exact checks ($\delta = 0$), only the approximation error remains: the blind spot is half the largest ℓ_1 distance from a prefix to the span of the checked curves. At positive tolerance, an accurate approximation can still yield a weak guarantee if it requires large coefficients.

The next theorem extends the blind spot formula to arbitrary targets and overlapping groups.

Theorem 4 (Blind spot of general targets)**.**

$$B(\mathcal{S};\delta) = \max_{g,r,\, s\in\{-1,1\}} \min_{\alpha\in\mathbb{R}^{m\times q}} \left\{ \max_{x\in\mathcal{X}} \left[s(\langle u^r, x_g\rangle - p_g) - \sum_{g',\ell} \alpha_{g',\ell}(\langle v^\ell, x_{g'}\rangle - p_{g'}) \right] + \delta\|\alpha\|_1 \right\}. \tag{6}$$

The blind spot is computable using $2mk$ linear programs, one per group, depth, and sign (Appendix C).

We next evaluate Theorem 3 on attention curves from prior work. In this setting, the blind spot is at most $\frac{1}{2}$, and nearly every curve we examine attains this maximum. We consider the DCG curve with weights proportional to $\frac{1}{\log_2(j+1)}$ (Singh & Joachims, 2018), the logarithmic curve with weights proportional to $\frac{1}{\log_2 \max(j,2)}$ used by the TREC Fair Ranking track (Ekstrand et al., 2023), RBP curves with weights proportional to ϕ^{j-1} for continuation probabilities $\phi \in \{0.5, 0.8, 0.95\}$ (Moffat & Zobel, 2008; Diaz et al., 2020), and power-law curves with weights proportional to $j^{-\eta}$

Table 1: **Blind spots of checked curve sets** for $k = 10$, $m = 2$, and $p_1 = p_2 = \frac{1}{2}$. Each entry is the largest disparity that can remain hidden while every checked curve has disparity at most δ. Lower is better, and 0.5 is the largest possible value. For $\delta = 0$, we use the endpoint curves constructed in Section 3.3. For $\delta = 0.01$, we adjust their weights using Proposition D.10. The best single curve is optimized for each tolerance.

		Blind spot	
Checked curves	q	$\delta = 0$	$\delta = 0.01$
best single curve	1	0.409	0.412
five published curves ($\mathcal{S}_{\text{prac}}$)	5	0.125	0.292
our endpoint curves	5	**0.083**	**0.093**
best five prefixes	5	0.125	0.133
all ten prefixes	10	0	0.010

for $\eta \in \{0.5, 1\}$. We denote the latter two families by RBP_ϕ and power_η, respectively, and call these seven curves the *published curves*. Every curve we examine except $\mathrm{RBP}_{0.5}$ has blind spot exactly $\frac{1}{2}$, while the blind spot of $\mathrm{RBP}_{0.5}$ is also within 2^{-k} of it. The maximizing policy is the one from the introduction: it always places group A at position one and places it at every subsequent position with probability $\frac{1-2v_1}{2-2v_1}$. This makes A's exposure under the checked curve exactly $\frac{1}{2}$ whenever $v_1 \leq \frac{1}{2}$. Moreover, even the best possible single attention curve has blind spot $\frac{k-1}{2(k+1)}$ for equal targets (Theorem C.2).

Since no single curve provides a strong guarantee on the blind spot, we next ask whether checking several curves together provides a stronger guarantee. We consider a set $\mathcal{S}_{\mathrm{prac}}$ containing the DCG curve, $\mathrm{RBP}_{0.8}$, $\mathrm{RBP}_{0.95}$, $\mathrm{power}_{0.5}$, and power_1. Even with all five curves checked, the blind spot at $k = 10$ is 0.125 for equal targets (Table 1). A policy can satisfy all five checks because deviations at different positions offset one another under every checked curve, while a substantial disparity remains at a particular reading depth. An example is given in Appendix G. It satisfies all five checks exactly, yet the expected share of one group among the first three positions is 62.5%, while its target share is 50%. By comparison, we show that the set of five curves constructed in Section 3.3, which we call the *endpoint curves*, achieves the smallest possible blind spot among all sets of five curves (0.083; Table 1).

Under exact checks, the five prefix curves at depths $1, 2, 3, 5, 8$ have the same blind spot as $\mathcal{S}_{\mathrm{prac}}$. At tolerance $\delta = 0.01$, however, the blind spot of the prefix curves grows by only 6% to 0.133, while that of $\mathcal{S}_{\mathrm{prac}}$ grows by 133% to 0.292. This is because approximating the sharp cutoff of a prefix curve using smooth curves requires large positive and negative coefficients, which amplify any tolerance in the checked disparities. Appendix J.7 shows the same effect on real policies: on German Credit, the highest-utility policy that passes the checks of $\mathcal{S}_{\mathrm{prac}}$ has querywise risk 0.077 at $\delta = 0$ but 0.167 at $\delta = 0.01$, whereas with our endpoint curves the risk rises only from 0.076 to 0.087.

3.3 The smallest audit for a given blind spot

An *audit* checks a ranking policy using a fixed set of attention curves. Under each curve, it computes every group's exposure and compares it with that group's target. The policy passes the audit at tolerance δ if every resulting disparity is at most δ. The size of an audit is the number of curves it checks. Checking all k prefix curves leaves no blind spot, but fewer curves may suffice when a nonzero blind spot β is acceptable. Under exact checks ($\delta = 0$), let $q^\star$ be the smallest number of curves guaranteeing blind spot at most β on every query with the specified targets.

Our construction groups consecutive reading depths into blocks. For a blind-spot level $0 \leq \beta' < \frac{1}{2}$, set $\gamma = \frac{1+2\beta'}{1-2\beta'}$. Start the first block at $c = 1$, end it at $d = \min(\lfloor \gamma c \rfloor, k)$, and start the next at $d + 1$. For each block, check the *endpoint curve* $v^{c,d} = \frac{1}{2}(u^c + u^d)$. To bound the disparity at depth r, we may approximate u^r with a coefficient $a = \frac{2cd}{r(c+d)}$ chosen for that depth. For $c \leq r \leq d$, we then have $\|u^r - a v^{c,d}\|_1 = \frac{d-c}{d+c} \leq 2\beta'$. For two equal-target groups, the blind spot formula therefore shows that this checked curve bounds every prefix disparity in the block by β'. Let $N(k, \beta')$ be the number of blocks. For example, when $k = 10$ and $\beta' = \frac{1}{12}$, we have $\gamma = 1.4$ and the blocks $[1, 1]$, $[2, 2]$, $[3, 4]$, $[5, 7]$, and $[8, 10]$, so $N(k, \beta') = 5$.

Theorem 5 (Number of curves needed). *For general targets and groups, let* $p_+ = \max_g \max(p_g, 1 - p_g)$ *and* $p_- = \max_g \min(p_g, 1 - p_g)$. *For* $0 \leq \beta < p_-$,

$$N\left(k, \frac{\beta}{2p_-}\right) \leq q^\star \leq N\left(k, \frac{\beta}{2p_+}\right). \tag{7}$$

The above endpoint construction attains the upper bound by setting $\beta' = \frac{\beta}{2p_+}$. For two groups with equal targets, the lower and upper bounds coincide, giving $q^\star = N(k, \beta)$ (Appendix D). Hence, our construction uses the fewest possible curves. For example, $k = 1000$ and $\beta = 0.01$ require exactly 106 curves in this case. For general targets, we have $q^\star \leq N(k, \frac{\beta}{2p_+}) = O(\frac{\log(1+k\beta)}{\beta})$ (Corollary D.7). Thus, for a fixed blind spot, the audit size grows logarithmically with the ranking length.

For two groups with equal targets, the above construction extends to a positive tolerance δ by reweighting the two prefix curves that form each endpoint curve. Let β be the smallest blind spot attainable with q curves under exact checks, and set $\mu = \frac{2(\beta+\delta)}{1+4\beta\delta}$. Keeping the blocks of the above construction with $\beta' = \beta$, we replace the endpoint curve of each block $[c, d]$ by $\frac{c(1+\mu)\,u^c + d(1-\mu)\,u^d}{c(1+\mu)+d(1-\mu)}$. As δ grows, μ increases and the weight shifts toward the shallower prefix curve u^c. At $\delta = 0$, the two weights are equal whenever $d = \gamma c$, and the curve then coincides with $v^{c,d}$. The reweighted endpoint curves have blind spot at most $\frac{\mu}{2} = \frac{\beta+\delta}{1+4\beta\delta}$ (Proposition D.10), whereas no set of at most q curves has blind spot below $\beta + \delta(1 - 2\beta)$ at tolerance δ (Theorem D.9). For $k = 10$, $q = 5$, and $\delta = 0.01$, the reweighted endpoint curves have blind spot 0.093 (Table 1), with a corresponding lower bound of 0.092.

4 Robust Ranking and Experiments

In this section, we turn from measuring fairness to enforcing it. We first give a linear program that maximizes utility while keeping the disparity under every attention curve in $\mathcal{V}_k$ within a tolerance, and then evaluate it on three datasets.

4.1 The robust LP

The single-curve LP of Singh & Joachims (2018) maximizes utility subject to fairness under one attention curve. To control disparity under every curve in $\mathcal{V}_k$, we replace its fairness constraints with prefix bounds. By Theorem 2, these bounds are equivalent to the desired guarantee.

For each query and a chosen tolerance δ, we solve

$$\max_{P \in \mathcal{P}_{n,k}} U(P) \quad \text{s.t.} \quad d_{g,r} \le \delta \quad \text{for all } g \in [m],\ r \in [k], \tag{8}$$

which we call the *robust LP*. The feasible policy set $\mathcal{P}_{n,k}$ is defined in (1). Each disparity bound is the pair of linear inequalities $-\delta \le \frac{1}{r}\sum_{j \le r} x_{g,j} - p_g \le \delta$, so the formulation remains linear for arbitrary targets and overlapping groups.

At $\delta = 0$, the constraints fix each group's probability at every position to its target, $x_{g,j} = p_g$. When groups partition the candidates and these targets are feasible, an optimal policy can be constructed directly by sorting within each group and allocating its required probability mass in score order (Appendix H). The LP provides a common formulation that also handles positive tolerances and overlapping groups.

4.2 Experimental setup

We examine how far policies that are fair under a single attention curve deviate from their targets under other curves, and how much ranking utility the robust LP sacrifices.

Data. *Amazon ESCI* (Reddy et al., 2022) is a product-search dataset with human relevance judgments in four categories: exact, substitute, complement, and irrelevant (ESCI). We use the larger version of the dataset with $k = 10$, containing $60{,}965/10{,}801/22{,}030$ train/validation/test queries, and define groups by the product brands present in each query. All policies use the same multilingual MiniLM relevance scorer, selected on validation data. A calibration map fitted on training data converts its scores into estimated relevance; each brand's target is its share of this estimated relevance. Test judgments are used only for evaluation and do not enter the construction of the policies or their targets (Appendix J).

TREC Fair Ranking 2022 (Ekstrand et al., 2023) is a benchmark for prioritizing Wikipedia articles that need improvement for human editors. It has 50 evaluation topics, and we rank $k = 20$ articles for each. The groups are the four gender categories that the benchmark assigns to articles, and an article may belong to several of them. We use the benchmark's own targets, utility, and attention curve (Appendix I).

To examine how these methods behave beyond information retrieval settings, we also evaluate on *German Credit* (Hofmann, 1994), which is a credit risk dataset containing demographic and

Table 2: **ESCI: ranking quality and disparity of each policy**, as means over the 22,030 test queries (for GDL23 and the prefix-quota MILP, over the 21,955 and 22,017 queries on which they are feasible).

			Disparity under		
Policy	NDCG	Utility	all curves	depths ≥ 3	log-like curves
Utility sort	0.8221	1.000	0.788	0.335	0.264
Single-curve LP	0.7734	0.9538	0.421	0.187	0.075
95%-NDCG point	0.7811	0.9635	0.439	0.195	0.082
Multi-curve LP	0.7692	0.9490	0.039	0.039	0.00028
GDL23	0.7707	0.9512	0.0238	0.0238	0.0238
Prefix-quota MILP	0.8144	0.9963	0.788	0.257	0.229
Robust LP	0.7692	0.9491	**0.00047**	**0.00047**	**0.00040**
95%-NDCG point	0.7835	0.9676	**0.0205**	**0.0205**	**0.0202**

financial information about 1,000 loan applicants. We construct a ranking task in which applicants are prioritized for attention based on predicted creditworthiness. Each query contains ten applicants under 35 and ten aged 35 or older, of whom $k = 10$ are ranked and displayed. We treat the two age groups equally with target $\frac{1}{2}$ each and use relevance scores from a logistic regression model to rank applicants by predicted creditworthiness (Appendix J.6).

Measuring ranking quality. Ground-truth relevance labels are not available at ranking time, so all methods use the same item scores of the relevance model within each dataset. The LPs maximize score-based utility $U(P) = \sum_{i,j} s_i w_j P_{ij}$. We use DCG weights for ESCI and German Credit. For TREC, we use the benchmark weights. We report utility relative to its maximum value, which is attained by sorting the items by decreasing score. To measure the quality that users actually receive, we replace the scores with the ground-truth relevance labels $y_i \geq 0$ and normalize, which gives the normalized discounted cumulative gain (NDCG):

$$\mathrm{NDCG}(P) = \frac{\sum_{i,j} y_i w_j P_{ij}}{\max_{P'} \sum_{i,j} y_i w_j P'_{ij}}.$$

Policies compared. *Utility sort* ranks items by decreasing score and ignores fairness. The *single-curve LP* of Singh & Joachims (2018) uses the DCG curve on ESCI and German Credit, and the benchmark's own logarithmic curve on TREC. The *multi-curve LP* requires fairness under all seven published curves of Section 3.2. *Robust LP* is (8), which maximizes utility while constraining every prefix disparity to be at most δ, thereby guaranteeing the same bound under every monotone attention curve. We run each of the three LPs at its smallest feasible tolerance. *GDL23* is the sampler of Gorantla et al. (2023a), which randomizes over lists whose group counts in the top k stay within a tolerance ω of the target. The *prefix-quota MILP* is a mixed-integer linear program that returns the fixed list with the highest utility whose group counts at every depth r stay between $\lfloor rp_g \rfloor$ and $\lceil rp_g \rceil$ (Celis et al., 2018; Geyik et al., 2019). To compare policies at similar relevance quality, for the single-curve and robust LPs, we choose the smallest tolerance on a validation grid whose NDCG reaches at least 95% of the utility sort's validation NDCG. We call this its 95%-*NDCG point*, and define the 98%-*NDCG* point analogously.

Reported metrics. Disparities are fractions of total attention, averaged over queries. In Table 2, "All curves" is the querywise risk $\mathcal{R}_{\mathrm{Q}}$. "Depths ≥ 3" restricts it to the prefix curves u^r with $r \geq 3$, and "log-like curves" is the largest disparity under the six published curves other than $\mathrm{RBP}_{0.5}$, whose weights decay much faster. We also report the *delivered share* of the target exposure under a curve v, $\mathrm{DS}(v) = \frac{\sum_g \min(\langle v, x_g \rangle, p_g)}{\sum_g p_g}$, which counts each group's exposure up to its target.

4.3 Robustness across attention curves

Table 2 and Figure 2 report the ESCI results. The policy returned by the single-curve LP is fair under the DCG curve, yet its querywise risk is 0.42. On 88% of the queries, the largest disparity is attained by u^1, the curve of users who read only the first result, and a disparity of 0.19 remains under the

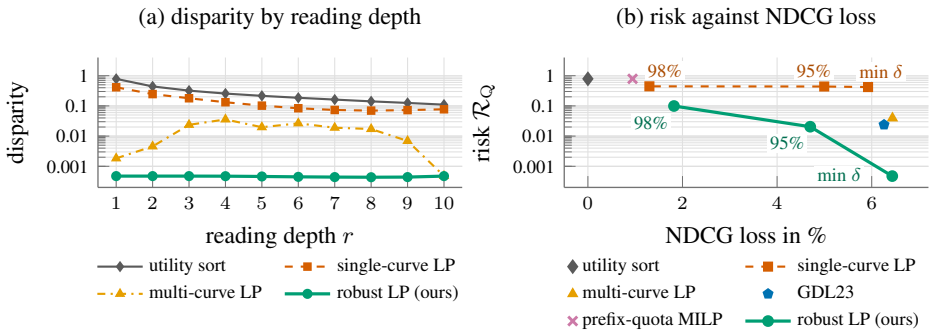


Figure 2: **ESCI: disparity by reading depth, and risk against NDCG loss. (a)** Mean over queries of the largest prefix disparity at each reading depth r. **(b)** Querywise risk $\mathcal{R}_Q$ against the loss of NDCG relative to the utility sort. The single-curve LP and the robust LP are shown at their smallest feasible tolerance, at their 95%-NDCG points, and at their 98%-NDCG points.

Table 3: **Both LPs at their smallest feasible tolerance**, as means over queries. *Click logs*: the curves of Section 4.4 at $k = 10$ (2,000 ESCI queries); *fitted* is the worst curve in the hull of the fitted curves, and *raw* the raw CTR curve. The delivered share takes the worst of these curves on each query.

			Disparity under		Delivered share under	
Dataset	Policy	Utility	all curves	click logs: fitted / raw	all curves	click logs
ESCI	Single-curve LP	0.9538	0.421	0.092 / 0.114	42.1%	79.9%
	Robust LP	0.9491	**0.0005**	**0.0005 / 0.0005**	**99.8%**	**99.8%**
TREC 2022	Single-curve LP	0.9997	0.358	0.074 / 0.081	64.2%	91.3%
	Robust LP	0.9997	**0.0005**	**0.0005 / 0.0005**	**99.9%**	**99.9%**
German Credit	Single-curve LP	0.9754	0.442	0.082 / 0.078	55.8%	90.2%
	Robust LP	0.9655	**0**	**0 / 0**	**100%**	**100%**

prefix curves of depth three or more (Table 2; Figure 2a). The risk of 0.42 depends in part on which of the many optimal solutions of the LP the solver returns: even the most robust optimal solution has risk 0.334, and reducing the risk to 0.016 requires sacrificing 1% of utility and selecting the policy by the prefix criterion of Theorem 2 (Appendix J.10). Under the worst curve of each query, the single-curve LP delivers less than half of the target exposure on 78% of the ESCI queries, and on average only 42% on ESCI, 64% on TREC, and 56% on German Credit, whereas the robust LP delivers on average at least 99.8% on all three datasets (Table 3).

Table 2 also shows that enforcing fairness under more curves reduces the risk but does not eliminate it: the multi-curve LP still has risk 0.039 on ESCI (and 0.071 on TREC; Appendix J.4). Its disparity is concentrated at reading depths three to eight, where none of the seven published curves detects it (Figure 2a). By constraining every prefix, the robust LP reduces the querywise risk to 0.0005. Of the two count-based baselines, the prefix-quota MILP returns a fixed list, which shows a single group at position one (Proposition C.6); on ESCI, its risk equals that of the utility sort, 0.79. GDL23 achieves a risk of 0.0238 at an NDCG of 0.7707, whereas the robust LP at its 95%-NDCG point reaches a lower risk of 0.0205 at a higher NDCG of 0.7835.

The cost of robustness. At their smallest feasible tolerances, the robust LP has 0.5% less utility than the single-curve LP on ESCI (0.9491 against 0.9538; NDCG 0.7692 against 0.7734) and 1.0% less on German Credit. On TREC, the two have the same utility under the benchmark's weights (Appendix I). Both LPs lose about 6% of NDCG relative to the utility sort on ESCI (0.8221), so most of the cost comes from fairness itself rather than from robustness. On ESCI, the robust LP takes 17.9 ms per query on one CPU core, compared with 11.5 ms for the single-curve LP (Appendix J.11).

4.4 Fairness under attention curves from click logs

As a stress test, we estimate real-world attention curves from 325,271 search pages of Baidu-ULTR (Zou et al., 2022; Hager et al., 2024), a public click log for unbiased learning to rank (ULTR), in two ways. The first is a two-tower click model, in which a document shown at position j is clicked with probability $\sigma(\rho(x))\,\sigma(\theta_j)$, where σ is the logistic function. The relevance tower ρ maps the query-document embedding x to a relevance logit, and the bias tower has one logit θ_j per position. The values $\sigma(\theta_j)$ are normalized to sum to one and projected onto monotone attention curves. Estimating the bias tower on different subsets of the pages yields a family of such curves, called the *fitted curves* (Appendix J.9). The second is the *raw click-through-rate (CTR) curve*, the share of clicks at each position. Because it is computed directly from clicks, it does not separate relevance from position bias. Position one receives 40% of the attention under the raw CTR curve, compared with 26%–32% under the fitted curves. Under these curves, the single-curve LP has disparities of 0.07–0.11 on ESCI, German Credit, and TREC (re-ranked with $k = 10$), and delivers on average only 80%, 90%, and 91% of the target exposure, respectively (Table 3). By contrast, the robust LP keeps the mean disparity below 0.0006 under every estimated curve and delivers at least 99.8%.

5 Related Work

Methods for exposure fairness optimize a distribution over rankings under a chosen attention curve (Biega et al., 2018; Singh & Joachims, 2018; Diaz et al., 2020). Other methods constrain group counts in ranking prefixes (Zehlike et al., 2017; Celis et al., 2018; Geyik et al., 2019) or ensure that every sampled list satisfies group quotas (Gorantla et al., 2023a;b). Our robust LP builds on this framework; its prefix constraints yield a guarantee under every monotone attention curve. This guarantee concerns the expected exposure of a randomized policy, whereas quotas on realized lists constrain each sampled ranking.

Prior studies show that fairness measurements are sensitive to the attention model (Sapiezynski et al., 2019; Raj & Ekstrand, 2022). Other work studies fair ranking under uncertainty about the merit of items (Singh et al., 2021), under distributional uncertainty about relevance (Memarrast et al., 2023), and under incomplete exposure estimates caused by dependencies among displayed items (Heuss et al., 2022). We instead hold targets and groups fixed and let the attention curve vary; the blind spot quantifies the disparity that a given set of exposure checks can leave undetected.

The decomposition of monotone weights into mixtures of prefix curves is classical (Kusuoka, 2001; Acerbi, 2002; Moffat et al., 2017). Our blind spot formula specializes optimal recovery from linear measurements (Micchelli & Rivlin, 1977; Traub et al., 1988). For the audit size, we compute the Kolmogorov ℓ_1 width of the prefix curves (Pinkus, 1985) and attain it with monotone attention curves.

6 Limitations and Conclusion

Fairness under one attention curve, or under a few, is a weak guarantee when the true attention curve is unknown. Our blind spot formula gives the largest disparity compatible with a set of passed checks, for arbitrary targets and overlapping groups. When used as the only check, almost every published curve we examined can conceal the largest possible disparity, whereas the set of all k prefix curves conceals none. For two groups with equal targets and zero tolerance, our endpoint curves attain a given blind spot with the fewest possible curves, and, for a fixed positive blind spot, this number grows only logarithmically with the ranking length. We also proposed a robust LP that maximizes ranking utility while keeping the disparity under every monotone attention curve within a chosen tolerance. On three datasets, it lowers the mean worst-case disparity below 0.06% of the total attention, at a utility loss of about 1% or less relative to the single-curve LP.

Our guarantees rest on two modeling choices, and relaxing either one is a natural next step. First, we assume that attention depends on positions and not on the items shown. In exposure models such as the dynamic Bayesian network model used by Kletti et al. (2022), users may stop examining the list once they find a satisfying result, so the attention that a position receives may also depend on the items above it. Second, we take the groups and targets as given. If the targets are known only approximately, the worst-case disparity must also account for the uncertainty in the targets (Theorem E.1).

ETHICS STATEMENT

Targets and attribute governance are normative choices. ESCI catalog brands are weak public proxies for providers, and queries with placeholder brands are excluded. TREC uses track-provided automated page alignments derived from Wikidata, with Unknown retained. These are neither author-inferred nor verified identity labels (Ekstrand et al., 2023). The constructed coarse groups are count-balanced aggregates of brand identifiers, not protected attributes. German Credit groups applicants by age, a protected attribute in many jurisdictions; we use this public benchmark only to evaluate the methods. No experiment affects people. Deployment needs domain justification, privacy and harm review, monitoring, and feedback analysis.

Baidu-ULTR contains hashed queries and documents, embeddings and click positions, and no user identifiers. We use it only to estimate position-attention curves.

REPRODUCIBILITY STATEMENT

The appendix contains the proofs of all results stated in the main text and the robustness checks of the experiments. The supplementary material provides the code and the software versions, the fixed data splits and data hashes, the results for each query, the complete tables, and executable programs that check the theorems. Public datasets are not redistributed but are referenced by their official URLs and checksums. The Baidu-ULTR shards (University of Amsterdam release, CC BY-NC 4.0 license) are listed with their SHA-256 checksums in the manifest of the click experiment.

AI USE STATEMENT

Generative AI assisted with ideation and literature search, with the formulation and stress testing of mathematical claims and proofs, with method and experiment design, with the implementation of methods, experiments, and validation code, with debugging, with the design and audit of the data processing pipeline, with the interpretation of experimental results, and with drafting and editing. We did not use generative AI tools to generate synthetic data or to translate text. The authors independently verified and take responsibility for all text, mathematics, code, data processing, and results.

REFERENCES

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002. doi: 10.1016/S0378-4266(02)00281-9.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 405–414, 2018. doi: 10.1145/3209978.3210063.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán, Revista A*, 5:147–151, 1946.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming*, volume 107 of *Leibniz International Proceedings in Informatics*, pp. 28:1–28:15, 2018. doi: 10.4230/LIPIcs.ICALP.2018.28.

Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 275–284, 2020. doi: 10.1145/3340531.3411962.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the TREC 2022 Fair Ranking track. In *Proceedings of the Thirty-First Text REtrieval Conference (TREC 2022)*, NIST Special Publication 500-338. National Institute of Standards and Technology, 2023. Also available as arXiv:2302.05558.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2221–2231, 2019. doi: 10.1145/3292500.3330691.

Sruthi Gorantla, Amit Deshpande, and Anand Louis. Sampling ex-post group-fair rankings. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 409–417, 2023a. doi: 10.24963/ijcai.2023/46.

Sruthi Gorantla, Anay Mehrotra, Amit Deshpande, and Anand Louis. Sampling individually-fair rankings that are always group fair. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 205–216, 2023b. doi: 10.1145/3600211.3604671.

Philipp Hager, Romain Deffayet, Jean-Michel Renders, Onno Zoeter, and Maarten de Rijke. Unbiased learning to rank meets reality: Lessons from Baidu's large-scale search dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1546–1556, 2024.

Maria Heuss, Fatemeh Sarvi, and Maarten de Rijke. Fairness of exposure in light of incomplete exposure estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 759–769, 2022. doi: 10.1145/3477495.3531977.

Alan J. Hoffman and Joseph B. Kruskal. Integral boundary points of convex polyhedra. In H. W. Kuhn and A. W. Tucker (eds.), *Linear Inequalities and Related Systems*, number 38 in Annals of Mathematics Studies, pp. 223–246. Princeton University Press, 1956.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. Dataset.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 781–789, 2017. doi: 10.1145/3018661.3018699.

Till Kletti, Jean-Michel Renders, and Patrick Loiseau. Pareto-optimal fairness-utility amortizations in rankings with a DBN exposure model. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 748–758, 2022. doi: 10.1145/3477495.3532036.

Shigeo Kusuoka. On law invariant coherent risk measures. In Shigeo Kusuoka and Toru Maruyama (eds.), *Advances in Mathematical Economics*, volume 3, pp. 83–95. Springer, Tokyo, 2001. doi: 10.1007/978-4-431-67891-5_4.

Omid Memarrast, Ashkan Rezaei, Rizal Fathony, and Brian D. Ziebart. Fairness for robust learning to rank. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD 2023, Proceedings, Part I*, volume 13935 of *Lecture Notes in Computer Science*, pp. 544–556, 2023. doi: 10.1007/978-3-031-33374-3_43.

Charles A. Micchelli and Theodore J. Rivlin. A survey of optimal recovery. In Charles A. Micchelli and Theodore J. Rivlin (eds.), *Optimal Estimation in Approximation Theory*, pp. 1–54. Plenum Press, New York, 1977. doi: 10.1007/978-1-4684-2388-4_1.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):2:1–2:27, 2008. doi: 10.1145/1416950.1416952.

Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems*, 35(3):24:1–24:38, 2017. doi: 10.1145/3052768.

Allan Pinkus. *n-Widths in Approximation Theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin, 1985.

Amifa Raj and Michael D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 726–736, 2022. doi: 10.1145/3477495.3532018.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv preprint arXiv:2206.06588*, 2022. doi: 10.48550/arXiv.2206.06588.

Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. Quantifying the impact of user attention on fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 553–562, 2019. doi: 10.1145/3308560.3317595.

Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Chichester, 1986.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2219–2228, 2018. doi: 10.1145/3219819.3220088.

Ashudeep Singh, David Kempe, and Thorsten Joachims. Fairness in ranking under uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11896–11908, 2021.

Joseph F. Traub, Grzegorz W. Wasilkowski, and Henryk Woźniakowski. *Information-Based Complexity*. Academic Press, 1988.

Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988. doi: 10.1109/21.87068.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, pp. 1569–1578, 2017. doi: 10.1145/3132847.3132938.

Lixin Zou, Haitao Mao, Xiaokai Chu, Jiliang Tang, Wenwen Ye, Shuaiqiang Wang, and Dawei Yin. A large scale search dataset for unbiased learning to rank. In *Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks)*, 2022.

APPENDIX ROADMAP

This appendix gives the proofs of the results stated in the main text, a few extensions, and the details of the experiments. Complete tables, additional diagnostics, and executable verification programs are in the supplementary material.

Contents.

Appendix	Contents
A	Proofs of Lemma 1 and Theorem 2: why prefix curves are the worst case; every $P \in \mathcal{P}_{n,k}$ is a distribution over rankings.
B	Restricted classes of attention curves (for example, a cap on the attention of position 1).
C	Proofs of the blind-spot formulas (Theorems 3 and 4); the best single curve; audits that check only prefix curves.
D	Proof of Theorem 5 (the smallest audit) and results for positive tolerances.
E	Uncertainty about the targets in addition to the attention curve.
F	Why the single-curve LP can have many optimal solutions that differ in robustness.
G	Blind spots of the published curves, with explicit policies that pass the checks.
H	The closed-form robust policy at $\delta = 0$, and the smallest attainable risk when no policy is exactly robust.
I, J	Experimental protocol and robustness checks: TREC; ESCI and German Credit; attention curves estimated from click logs.

Notation. The appendix uses the terms and symbols of the main text. Table 4 lists the additional notation that recurs, and Table 5 records the setting in which each main result holds.

A PROOFS FOR SECTION 3.1: PREFIX CURVES ARE THE WORST CASE

Signed deviations and disparities. Fix a policy P, and let $x_{g,j}(z)$ be its occupancy of group g at position j on query z. For the proofs, define the signed deviations

$$a_{g,r}(z) = \frac{1}{r}\sum_{j\le r} x_{g,j}(z) - p_g, \qquad e_g(z;v) = \sum_j v_j x_{g,j}(z) - p_g.$$

The first is the *signed prefix deviation* of group g at depth r. The second is the signed deviation of group g under curve v, which equals its exposure minus its target. The disparities of the main text are the absolute values of these deviations: the prefix disparity of Section 3.1 is $d_{g,r}(z) = |a_{g,r}(z)|$, and the disparity of Section 2 is $D_g(z;v) = |e_g(z;v)|$. We omit z when it is fixed, and write $x_{g,j}(P)$, $a_{g,r}(P)$ or $d_{g,r}(P)$ when the policy matters.

Proof of Lemma 1. Monotonicity gives $\lambda_r \ge 0$. Because $u^r_j = 1/r$ for $j \le r$ and $u^r_j = 0$ for $j > r$, the sum at coordinate j telescopes:

$$\sum_{r=1}^{k} \lambda_r u^r_j = \sum_{r=j}^{k} \frac{r(v_r - v_{r+1})}{r} = \sum_{r=j}^{k} (v_r - v_{r+1}) = v_j.$$

Summing over coordinates gives $\sum_r \lambda_r = \sum_j v_j = 1$. The weights are unique, because this triangular system has the inverse $\lambda_r = r(v_r - v_{r+1})$. □

Proof of Theorem 2. Fix a group g and a query z, and let $v \in \mathcal{V}_k$. Lemma 1 writes $v = \sum_r \lambda_r u^r$ with $\lambda \ge 0$ and $\sum_r \lambda_r = 1$, so linearity gives

$$\sum_j v_j x_{g,j}(z) - p_g = \sum_r \lambda_r \Big(\tfrac{1}{r} \textstyle\sum_{j\le r} x_{g,j}(z) - p_g \Big),$$

and the triangle inequality and $\sum_r \lambda_r = 1$ then give $D_g(z;v) \le \sum_r \lambda_r d_{g,r}(z) \le \max_r d_{g,r}(z)$. Conversely, each prefix curve u^r lies in $\mathcal{V}_k$, which gives the reverse inequality. Maximizing over groups and taking the expectation over z proves the identity. The same argument applies to an empirical average over queries. □

Symbol or term	Meaning
$\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$, $\mathcal{R}_{\mathrm{Q}}^{\mathcal{C}_{\mathrm{dec}}}$	the restricted risk $\mathcal{R}_{\mathrm{Q}}^{C}$ of Appendix B for the reference envelope $C = \mathcal{E}$ and for the bounded-decay class $C = \mathcal{C}_{\mathrm{dec}}$; both classes are defined at the start of Appendix G. An *envelope* is such a restricted class of curves.
$B_k(\mathcal{S}; \delta)$, $B_{k,p}(\mathcal{S}; \delta)$	the equal-target blind spot of Theorem 3, and its analogue for one group with target p (Corollary C.8).
ξ	the ℓ_1 approximation level $\xi = 2\beta$ that corresponds to blind spot β (Appendices C and D); the main text's β' is a blind-spot level, so $\xi = 2\beta'$.
$N_{\mathrm{chain}}(k, \gamma)$, $N_{\mathrm{cover}}(k, \gamma)$	the block counts of Appendix D. The main text's $N(k, \beta')$ equals $N_{\mathrm{cover}}(k, \gamma)$ with $\gamma = \frac{1+2\beta'}{1-2\beta'}$.
floor, minimum risk, excess	the *floor* of an LP on a query is its smallest feasible tolerance; for the robust LP, it is the smallest risk that any policy attains on the query. Rows labeled "minimum risk" run the robust LP at its floor, rows labeled "floor $+0.02$" run it at a tolerance 0.02 above the floor, and an *excess* is a tolerance above the floor.
min-violation extension	when a target cannot be met exactly, the LP first minimizes the violation and then maximizes utility.
TREC curve	the curve of the TREC Fair Ranking track, with weights proportional to $1/\log_2 \max(j, 2)$ (Section 3.2).

Table 4: **Additional notation used in the appendix.**

Result	Setting	Guarantee
Prefix certificate and robust LP (Theorem 2, LP (8))	position-based curves; any groups and targets	exact over all of $\mathcal{V}_k$; no assumption on group sizes or targets
Blind spot of a given set of curves (Theorem 4)	any groups and targets, every tolerance $\delta \geq 0$ at which some policy passes the checks	exact, one LP per group, depth and sign on the query's polytope (Appendix C)
Closed form (Theorem 3 for two groups with equal targets, Corollary C.8 for any target)	disjoint groups with at least k items each	exact, and the largest blind spot over all queries with the same targets (Proposition C.9)
Smallest set of curves (Theorem 5)	any groups and targets, blind spot at most β on every query with these targets, $\delta = 0$	exact for two groups with equal targets, otherwise upper and lower bounds of order $\log(1+k\beta)/\beta$ (Corollaries C.10 and C.11), and the lower bound holds over every linear subspace
Finite tolerance (Appendix D.1)	two groups with equal targets, $\delta > 0$	$q = 1$ exact (Theorem D.11); $q \geq 2$ a universal lower bound and a constructive upper bound, gap open

Table 5: **Setting and guarantee of each result.**

The next proof shows the fact behind the case $\delta = 0$ of Section 4.1: exact robustness, $\mathcal{R}_{\mathrm{Q}} = 0$, forces every position to show each group with probability equal to its target, on almost every query. (The converse is immediate: if $x_{g,j} = p_g$ at every position, then every group's exposure equals p_g under every normalized curve.)

Proof that exact robustness forces the same group mix at every position. By Theorem 2, $\mathcal{R}_{\mathrm{Q}}$ is the expectation of the nonnegative random variable $\max_{g,r} d_{g,r}(z)$. A nonnegative random variable has expectation zero exactly when it vanishes almost surely, so $\mathcal{R}_{\mathrm{Q}} = 0$ forces every prefix disparity to vanish on almost every query. If every prefix deviation vanishes, then $\sum_{j \leq r} x_{g,j} = r p_g$ for every depth r, and subtracting the equations of consecutive depths gives $x_{g,j} = p_g$ at every position. □

Finally, the next standard fact justifies treating a matrix $P \in \mathcal{P}_{n,k}$ as a policy (Section 2).

Proposition A.1 (Every policy matrix is a distribution over rankings). *Let* $\mathcal{P}_{n,k} = \{P \in \mathbb{R}_{+}^{n \times k} : \sum_i P_{ij} = 1 \ (j \in [k]),\ \sum_j P_{ij} \leq 1 \ (i \in [n])\}$ *with* $k \leq n$. *Every vertex of* $\mathcal{P}_{n,k}$ *is a* $0/1$ *matrix,*

which encodes an injective assignment of the k positions to distinct items. Hence every $P \in \mathcal{P}_{n,k}$ is a finite convex combination of at most $nk + 1$ such top-k rankings, computable in polynomial time.

Proof. The constraint matrix of $\mathcal{P}_{n,k}$ is the vertex–edge incidence matrix of the complete bipartite graph $K_{n,k}$, and it is therefore totally unimodular (Schrijver, 1986). The right-hand side is integral, so by the Hoffman–Kruskal theorem (Hoffman & Kruskal, 1956) every vertex of $\mathcal{P}_{n,k}$ is integral. Nonnegativity together with the row and column constraints then forces entries in $\{0, 1\}$, with exactly one 1 per column and at most one per row. Carathéodory's theorem, applied in the affine hull of $\mathcal{P}_{n,k}$, gives the decomposition. Constructively, pad P with $n - k$ columns that carry the residual row slacks to obtain a doubly stochastic $n \times n$ matrix, apply Birkhoff's peeling algorithm (Birkhoff, 1946), and delete the padded columns. □

B Restricted Classes of Attention Curves

An *envelope* is a restricted class of attention curves, a subset of $\mathcal{V}_k$. We consider envelopes defined by finitely many linear inequalities on the curve, for example a cap on the attention of position 1 or a minimum amount of attention beyond position 3. Such a class is useful when more is known about attention than monotonicity. Theorem 2 extends to every such class, with a finite set of worst-case curves in place of the prefix curves (Theorem B.2), and the robust LP extends accordingly. When the envelope is all of $\mathcal{V}_k$, its worst-case curves are the prefix curves.

We work in the mixture weights of Lemma 1. Throughout, $\Delta_k = \{\lambda \in \mathbb{R}^k_+ : \sum_r \lambda_r = 1\}$, and

$$T : \Delta_k \to \mathcal{V}_k, \qquad T\lambda = \sum_{r=1}^{k} \lambda_r u^r.$$

Fix $c^1, \dots, c^q \in \mathbb{R}^k$ and $b \in \mathbb{R}^q$, and put

$$C = \{v \in \mathcal{V}_k : \langle c^i, v\rangle \le b_i,\ i \in [q]\}, \qquad A \in \mathbb{R}^{q\times k}, \quad A_{i,r} := \langle c^i, u^r\rangle, \tag{9}$$

$$\Lambda := \{\lambda \in \Delta_k : A\lambda \le b\}. \tag{10}$$

Here q is the number of inequalities that define the envelope (unlike in Section 3.2, where q is the number of checked curves), $A_{i,r}$ is the value of the i-th constraint function $\langle c^i, \cdot\rangle$ at the prefix curve u^r, and Λ is the set of mixture weights of the curves in C (Lemma B.1). For example, the reference envelope $\mathcal{E} = \{v \in \mathcal{V}_k : v_1 \le 0.45,\ \sum_{j>3} v_j \ge 0.15\}$ of Appendix G has this form with $q = 2$, $c^1 = (1, 0, \dots, 0)$, $b_1 = 0.45$, $c^2 = -(0, 0, 0, 1, \dots, 1)$ and $b_2 = -0.15$; the bounded-decay class $\mathcal{C}_{\mathrm{dec}}$ of that appendix is also an envelope of this form. We assume throughout that $C \neq \emptyset$. The vectors c^i are arbitrary: no monotonicity or sign restriction is imposed on them.

In the mixture weights, the inequalities that define C remain linear, and monotonicity becomes membership in the simplex Δ_k:

Lemma B.1 (Curves correspond one to one to their mixture weights)**.** *T is a linear bijection from Δ_k onto $\mathcal{V}_k$ with inverse $(T^{-1}v)_r = r(v_r - v_{r+1})$, $v_{k+1} := 0$, and*

$$T(\Lambda) = C, \qquad T^{-1}(C) = \Lambda.$$

In particular, C is a (possibly empty) polytope, and $C \neq \emptyset$ if and only if $\Lambda \neq \emptyset$.

Proof. Lemma 1 and its proof give exactly the following: T maps Δ_k into $\mathcal{V}_k$, every $v \in \mathcal{V}_k$ has a preimage, and the preimage is unique and given by the stated triangular inverse. So T is a linear bijection. For $\lambda \in \Delta_k$ and $i \in [q]$, linearity of the inner product gives

$$\langle c^i, T\lambda\rangle = \Big\langle c^i, \sum_r \lambda_r u^r \Big\rangle = \sum_r \lambda_r \langle c^i, u^r\rangle = (A\lambda)_i.$$

Hence, for $v = T\lambda$: $v \in C \iff v \in \mathcal{V}_k$ and $A\lambda \le b \iff \lambda \in \Delta_k$ and $A\lambda \le b \iff \lambda \in \Lambda$. □

For an envelope C, define the restricted risk as the querywise risk over C:

$$\mathcal{R}^C_{\mathrm{Q}} = \mathbb{E}_z\Big[\sup_{v\in C} \max_g D_g(z; v)\Big].$$

For $C = \mathcal{V}_k$, the restricted risk $\mathcal{R}_{\mathrm{Q}}^{C}$ is the querywise risk $\mathcal{R}_{\mathrm{Q}}$ of (2). Write $a_g(z) = (a_{g,1}(z), \ldots, a_{g,k}(z))$ for the vector of signed prefix deviations of group g.

Theorem B.2 (The worst case over a restricted class is attained at its vertices)**.** *For every nonempty polyhedral envelope C in (9),*

$$\mathcal{R}_{\mathrm{Q}}^{C} = \mathbb{E}_z \max_g \max_{\lambda \in \mathrm{ext}(\Lambda)} |\langle \lambda, a_g(z) \rangle| . \tag{11}$$

The identity also holds for an empirical average over queries.

The proof below shows that $\langle \lambda, a_g(z) \rangle = e_g(z; T\lambda)$. The theorem therefore says that the worst case over C is attained at one of the finitely many curves $T\lambda$ with $\lambda \in \mathrm{ext}(\Lambda)$. For $C = \mathcal{V}_k$, these curves are the prefix curves, and the theorem reduces to Theorem 2.

Proof. For $v = T\lambda$, linearity and $\sum_r \lambda_r = 1$ give $e_g(z; v) = \sum_r \lambda_r a_{g,r}(z)$. Taking the absolute value gives the disparity. By Lemma B.1, maximizing over $v \in C$ is equivalent to maximizing over $\lambda \in \Lambda$. The absolute value of this linear function attains its maximum at an extreme point of Λ. Finally, maximize over groups and average over queries. □

The robust LP over an envelope. By Theorem B.2, a policy has disparity at most δ under every curve in C on query z if and only if $|\langle \lambda, a_g(z) \rangle| \le \delta$ for every group g and every $\lambda \in \mathrm{ext}(\Lambda)$. Each of these conditions is a pair of linear inequalities in the policy, because every signed prefix deviation is affine in it. Replacing the prefix constraints of the robust LP (8) by these $2m\,|\,\mathrm{ext}(\Lambda)|$ inequalities therefore gives the highest-utility policy whose disparity under every curve in C is at most δ. For $C = \mathcal{V}_k$, the extreme points of $\Lambda = \Delta_k$ are the standard basis vectors, and the constraints are those of (8). Appendix J.5 uses this form with the 42 extreme curves of $\mathcal{E}$ and the 512 extreme curves of $\mathcal{C}_{\mathrm{dec}}$.

C Blind Spots: Proofs and Supporting Results

This appendix proves Theorems 3 and 4 and the other results on blind spots that the main text cites.

Notation. A set of checked curves $\mathcal{S} = \{v^1, \ldots, v^q\}$ is represented by the $k \times q$ matrix V whose columns are the checked curves, so that $V\alpha = \sum_\ell \alpha_\ell v^\ell$ for $\alpha \in \mathbb{R}^q$. A curve is *normalized* if its entries sum to one, and a curve set is normalized if all its curves are; every curve in $\mathcal{V}_k$ is normalized. The *equal-target blind spot* $B_k(\mathcal{S}; \delta)$ is the blind spot $B(\mathcal{S}; \delta)$ in the setting of Theorem 3: two groups partition the items, each contains at least k items, and both have target $\frac{1}{2}$. By Lemma C.1 below, every $x \in [0,1]^k$ is then the occupancy vector of group 1 under some policy, so $B_k(\mathcal{S}; \delta)$ is the supremum of $\max_r |\langle u^r, x \rangle - \frac{1}{2}|$ over all $x \in [0,1]^k$ with $|\langle v^\ell, x \rangle - \frac{1}{2}| \le \delta$ for every ℓ. (Group 2 has occupancy vector $\mathbf{1} - x$; because all curves are normalized, its deviations are those of group 1 with the sign reversed.) Its single-target analogue $B_{k,p}(\mathcal{S}; \delta)$ is the same supremum with $\frac{1}{2}$ replaced by a target $p \in (0,1)$: the blind spot of one group with target p whose occupancy vector ranges over the cube $[0,1]^k$ (Corollary C.8). Thus $B_{k,1/2} = B_k$. We drop the arguments of B_k when they are clear from the context.

C.1 Equal targets: realizability and a direct proof

We first show when every occupancy vector is realizable, which is the hypothesis behind the formula of Theorem 3, and then prove that formula directly by linear-programming duality.

Lemma C.1 (Which occupancy vectors a policy can realize)**.** *For two groups that partition the items, every $x \in [0,1]^k$ is realizable as the occupancy vector of the first group if and only if both groups contain at least k items.*

Proof. Suppose that each group has at least k items. At each position j, assign mass x_j to an item of the first group and mass $1 - x_j$ to an item of the second group, using fresh items at every position. Conversely, if the first group has fewer than k items, it cannot have occupancy $\mathbf{1}$; if its complement, the second group, has fewer than k items, the second group cannot have occupancy $\mathbf{1}$, so the first group cannot have occupancy $\mathbf{0}$. □

Direct proof of (5) *in Theorem 3.* Let $x \in [0,1]^k$ be the occupancy vector of group 1 and set $y = 2x - \mathbf{1}$, so that $\|y\|_\infty \le 1$. Because every checked curve is normalized, the checks at tolerance δ become $\|V^\top y\|_\infty \le 2\delta$. The largest prefix disparity, which by Theorem 2 is the largest disparity under any curve in $\mathcal{V}_k$, becomes $\frac{1}{2}\max_r |(u^r)^\top y|$. The feasible set of y is centrally symmetric (y is feasible if and only if $-y$ is), so the absolute value can be dropped. For any $a \in \mathbb{R}^k$, linear-programming duality gives

$$\sup_{\|y\|_\infty \le 1,\ \|V^\top y\|_\infty \le 2\delta} a^\top y = \inf_{\alpha \in \mathbb{R}^q} \{\|a - V\alpha\|_1 + 2\delta\|\alpha\|_1\}.$$

Take $a = u^r$, maximize over r, and multiply by $1/2$. At $\delta = 0$, the infimum is the ℓ_1 distance from a to the span of the checked curves. The prefix curves are linearly independent, so a blind spot of zero requires $\operatorname{span} \mathcal{S} = \mathbb{R}^k$.

The dual identity also gives two simple bounds. For $0 \le \delta \le 1/2$, the point $y = 2\delta\mathbf{1}$ is feasible because every checked curve is normalized, and every prefix curve evaluates it to 2δ, which gives $B_k \ge \delta$. If the checked curves include all k prefix curves, then each u^r represents itself with one unit coefficient, and substituting this representation in the dual gives $B_k \le \delta$; for this set, the two bounds coincide. □

C.2 The best single checked curve

Section 3.2 shows that almost every published curve has the largest possible equal-target blind spot, $\frac{1}{2}$, when it is the only checked curve. The next theorem finds the single curve with the smallest equal-target blind spot under exact checks.

Theorem C.2 (The best single curve)**.** *Let $k \ge 2$. For $v \in \mathcal{V}_k$, the equal-target blind spot of the single curve v under exact checks is $B_k(\{v\}; 0) = \frac{1}{2}\max_{r\in[k]} \operatorname{dist}_1\big(u^r, \operatorname{span}\{v\}\big)$ (Theorem 3 at $\delta = 0$), and*

$$\min_{v \in \mathcal{V}_k} B_k(\{v\}; 0) = \frac{k-1}{2(k+1)}, \qquad \textit{attained by} \quad v^\star = \Big(\frac{k+1}{2k},\ \frac{1}{2k},\ \ldots,\ \frac{1}{2k}\Big).$$

Proof. Step 1: $v^\star \in \mathcal{V}_k$. Its entries are nonnegative and nonincreasing ($\frac{k+1}{2k} \ge \frac{1}{2k}$), and they sum to $\frac{k+1}{2k} + \frac{k-1}{2k} = 1$.

Step 2: $B_k(\{v^\star\}; 0) = \frac{k-1}{2(k+1)}$. Parametrize the multiples $\alpha v^\star$ of $v^\star$ by $t = \alpha/(2k)$, so that $\alpha v^\star = (\,(k+1)t,\ t, \ldots, t\,)$. For $r \in [k]$, the ℓ_1 error of approximating u^r by this multiple is

$$h_r(t) := \|u^r - \alpha v^\star\|_1 = \Big|\tfrac{1}{r} - (k+1)t\Big| + (r-1)\Big|\tfrac{1}{r} - t\Big| + (k-r)|t|,$$

a convex piecewise-linear function with breakpoints $t \in \{0,\ \frac{1}{r(k+1)},\ \frac{1}{r}\}$ (note that $0 < \frac{1}{r(k+1)} < \frac{1}{r}$). On $\big(0, \frac{1}{r(k+1)}\big)$, all three arguments are positive, so $h_r(t) = \frac{1}{r} - (k+1)t + (r-1)(\frac{1}{r} - t) + (k-r)t$ and

$$h_r' = -(k+1) - (r-1) + (k-r) = -2r < 0.$$

On $\big(\frac{1}{r(k+1)}, \frac{1}{r}\big)$, the first argument is negative, so $h_r(t) = (k+1)t - \frac{1}{r} + (r-1)(\frac{1}{r} - t) + (k-r)t$ and

$$h_r' = (k+1) - (r-1) + (k-r) = 2(k-r+1) > 0.$$

By convexity, the unique minimizer is $t^\star = \frac{1}{r(k+1)}$, and

$$h_r(t^\star) = 0 + (r-1)\cdot\frac{1}{r}\cdot\frac{k}{k+1} + (k-r)\cdot\frac{1}{r(k+1)} = \frac{(r-1)k + (k-r)}{r(k+1)} = \frac{r(k-1)}{r(k+1)} = \frac{k-1}{k+1},$$

independently of r. Hence $\max_r \operatorname{dist}_1(u^r, \operatorname{span}\{v^\star\}) = \frac{k-1}{k+1}$ and $B_k(\{v^\star\}; 0) = \frac{k-1}{2(k+1)}$.

Step 3: no $v \in \mathcal{V}_k$ does better. We use the elementary direction of ℓ_1/ℓ_∞ duality: for any depth r and any y with $\|y\|_\infty \le 1$ and $\langle v, y\rangle = 0$,

$$\operatorname{dist}_1\big(u^r, \operatorname{span}\{v\}\big) = \min_\alpha \|u^r - \alpha v\|_1 \ \ge\ \min_\alpha \langle u^r - \alpha v, y\rangle = \langle u^r, y\rangle$$

(the inequality is Hölder's inequality, and the last equality uses $\langle v, y\rangle = 0$). So it suffices to exhibit, for each $v \in \mathcal{V}_k$, one depth r and one such y with $\langle u^r, y\rangle \geq \frac{k-1}{k+1}$. Note that $v_1 = \max_j v_j \geq 1/k > 0$.

Case 1: $v_1 \leq \frac{k+1}{2k}$. Since $k \geq 2$, we have $\frac{k+1}{2k} < 1$, so $v_1 < 1$. Put the scalars $a = \min\big(1, \frac{1-v_1}{v_1}\big)$ and $b = \frac{a\,v_1}{1-v_1}$. Then $b \leq 1$: if $\frac{1-v_1}{v_1} \leq 1$, then $a = \frac{1-v_1}{v_1}$ and $b = 1$; otherwise $a = 1$ and $b = \frac{v_1}{1-v_1} < 1$. Set $y = (a, -b, \ldots, -b)$. Then $\|y\|_\infty \leq 1$ and $\langle v, y\rangle = av_1 - b(1-v_1) = 0$, while

$$\langle u^1, y\rangle = a = \min\Big(1, \tfrac{1-v_1}{v_1}\Big) \;\geq\; \min\Big(1, \tfrac{k-1}{k+1}\Big) = \frac{k-1}{k+1},$$

because $s \mapsto \frac{1-s}{s}$ is decreasing, so that $v_1 \leq \frac{k+1}{2k}$ gives $\frac{1-v_1}{v_1} \geq \frac{1-\frac{k+1}{2k}}{\frac{k+1}{2k}} = \frac{k-1}{k+1}$.

Case 2: $v_1 > \frac{k+1}{2k} = \frac{1}{2} + \frac{1}{2k} > \frac{1}{2}$. Set $y = \big(-\frac{1-v_1}{v_1},\, 1, \ldots, 1\big)$. Since $v_1 > \frac{1}{2}$, we have $\frac{1-v_1}{v_1} < 1$, so $\|y\|_\infty \leq 1$, and $\langle v, y\rangle = -\frac{1-v_1}{v_1}v_1 + (1-v_1) = 0$. Moreover,

$$\langle u^k, y\rangle = \frac{1}{k}\Big(k - 1 - \frac{1-v_1}{v_1}\Big) > \frac{1}{k}\Big(k - 1 - \frac{k-1}{k+1}\Big) = \frac{k-1}{k} \cdot \frac{k}{k+1} = \frac{k-1}{k+1},$$

using $v_1 > \frac{k+1}{2k} \Rightarrow \frac{1-v_1}{v_1} < \frac{k-1}{k+1}$.

In both cases $\max_r \operatorname{dist}_1(u^r, \operatorname{span}\{v\}) \geq \frac{k-1}{k+1}$, so $B_k(\{v\}; 0) \geq \frac{k-1}{2(k+1)}$. Together with Step 2, this proves the theorem. □

C.3 Prefix-only audits: span, distance and the best depths

An audit may check prefix curves themselves. This subsection computes the equal-target blind spot of exactly checking the prefix curves of a set of depths, and gives a dynamic program that finds the best set of depths of a given size. Fix $R = \{r_1 < \cdots < r_q\} \subseteq [k]$, put $r_0 = 0$, and let $\mathcal{S}_R = \{u^r : r \in R\}$ be the prefix curve set of the depths in R.

Proposition C.3 (What a few prefix checks can reproduce). *(i)* $\operatorname{span}(\mathcal{S}_R)$ *is the set of vectors* $w \in \mathbb{R}^k$ *that are constant on each block* $(r_{i-1}, r_i]$, $i \in [q]$, *and vanish beyond* r_q*; it is a subspace of dimension exactly* q.

(ii) For $r \in [k]$, *let* $a(r) = \max\{\rho \in R : \rho < r\}$ *(*$:= 0$ *if none) and* $b(r) = \min\{\rho \in R : \rho > r\}$ *(*$:= +\infty$ *if none) be the nearest checked depths below and above* r. *Then*

$$\operatorname{dist}_1\big(u^r, \operatorname{span}(\mathcal{S}_R)\big) = \frac{\min\big(r - a(r),\; b(r) - r\big)}{r} \qquad (\textit{the value is } 0 \textit{ when } r \in R).$$

(iii) Hence the equal-target blind spot of exactly checking the depths in R, $B_k(\mathcal{S}_R; 0)$, *is* $\Psi(R)/2$ *with* $\Psi(R) = \max_{r \in [k]\setminus R} \min\big(r - a(r), b(r) - r\big)/r$.

Proof. (i) On the block $(r_{i-1}, r_i]$, the vector u^{r_j} equals $1/r_j$ if $j \geq i$ and 0 if $j < i$. In either case it is constant there, and it vanishes beyond r_q because $r_j \leq r_q$. So every element of the span lies in the stated set. The span has dimension q because the q vectors $u^{r_1}, \ldots, u^{r_q}$ are linearly independent (their supports are distinct and strictly nested), and the stated set also has dimension q (one free constant per block). Equal dimensions and one inclusion give equality.

(ii) If $r \in R$, then $u^r \in \operatorname{span}(\mathcal{S}_R)$ and the distance is 0. Let $r \notin R$, and let w range over the subspace of (i), with value c_i on block $(r_{i-1}, r_i]$. Since the blocks are disjoint and $c_1, \ldots, c_q$ are free, $\|u^r - w\|_1$ separates into one term per block plus a tail:

$$\|u^r - w\|_1 = \sum_{i=1}^{q} \sum_{j \in (r_{i-1}, r_i]} \big|u_j^r - c_i\big| + \sum_{j > r_q} u_j^r,$$

and each block can be minimized independently. *Blocks up to* $a(r)$ ($r_i \leq a(r) < r$): there $u_j^r = 1/r$ throughout, so $c_i = 1/r$ gives cost 0. *Blocks strictly above* $b(r)$ ($r_{i-1} \geq b(r) > r$): there $u_j^r = 0$, so $c_i = 0$ gives cost 0. *The straddling block* $(a(r), b(r)]$ (present iff $b(r) < \infty$): it contains $r - a(r)$ positions with $u_j^r = 1/r$ and $b(r) - r$ positions with $u_j^r = 0$, so its cost

is $\min_c\big[(r-a(r))|1/r-c|+(b(r)-r)|c|\big]$, a weighted-median problem whose optimum is at $c\in\{0,1/r\}$ with value $\min(r-a(r),b(r)-r)/r$. *Tail*: if $b(r)=\infty$, which means $r>r_q$, the last term contributes $\sum_{j=r_q+1}^{r}1/r=(r-r_q)/r=(r-a(r))/r$, and every block cost is 0. Summing gives (ii) in both cases.

(iii) This is immediate from (ii) and Theorem 3 at $\delta=0$. □

Proposition C.4 (Choosing the best prefix depths by dynamic programming)**.** *Define, for* $0\le a<b\le k$,

$$\mathrm{gap}(a,b)=\max_{a<r<b}\frac{\min(r-a,\ b-r)}{r}\quad(=0\ \text{if}\ b\le a+1),$$

which is computable in $O(1)$*:* $r\mapsto\frac{r-a}{r}$ *is nondecreasing and* $r\mapsto\frac{b-r}{r}$ *is strictly decreasing, so the minimum of the two is unimodal in* r*, and its maximum over the integers is attained at* $\lfloor\frac{a+b}{2}\rfloor$ *or* $\lceil\frac{a+b}{2}\rceil$ *(clipped to* (a,b)*). Then, for* $k\ge 2$ *and* $1\le q\le k$*, every optimal* R *(one that minimizes* $\Psi(R)$ *over* $|R|=q$*) contains* $r_1=1$*, and*

$$\Psi(R)=\max\Big\{\ \mathrm{gap}(r_{i-1},r_i)\ (2\le i\le q),\quad\tfrac{k-r_q}{k}\ \Big\}.$$

Consequently, the recursion $\mathrm{dp}[1][1]=0,\ \mathrm{dp}[t][b]=\min_{t-1\le a<b}\max\big\{\mathrm{dp}[t-1][a],\ \mathrm{gap}(a,b)\big\}$, $\Psi^\star=\min_{q\le b\le k}\max\{\mathrm{dp}[q][b],(k-b)/k\}$ *computes* $\min_{|R|=q}\Psi(R)$ *exactly in* $O(qk^2)$ *time, and the smallest equal-target blind spot of* q *exactly checked prefix curves is* $\Psi^\star/2$.

Proof. $1\in R$. If $r_1\ge 2$, then $r=1$ is unchecked with $a(1)=0$ and $b(1)=r_1$, so $\Psi(R)\ge\min(1-0,\ r_1-1)/1=\min(1,r_1-1)=1$. But $R=\{1\}$ already gives $\Psi=\max_{r\ge 2}(r-1)/r=(k-1)/k<1$. Adding any $q-1$ further depths cannot increase Ψ, so some q-element set that contains 1 has value below 1. Hence, for $k\ge 2$, every R with $r_1\ge 2$ is strictly suboptimal.

Decomposition. Every unchecked r lies in exactly one of the intervals (r_{i-1},r_i), $2\le i\le q$, or $(r_q,k]$; the interval $(0,r_1)=(0,1)$ is empty since $r_1=1$. In the former case, $a(r)=r_{i-1}$ and $b(r)=r_i$, and these depths contribute $\mathrm{gap}(r_{i-1},r_i)$. In the latter case, $b(r)=\infty$, and the contribution is $\max_{r_q<r\le k}(r-r_q)/r=(k-r_q)/k$ because $r\mapsto 1-r_q/r$ is increasing. So $\Psi(R)$ is a maximum of terms that depend only on *consecutive* elements of R, plus a terminal term that depends only on r_q.

Correctness of the recursion. This is the standard bottleneck (min–max) shortest-path DP over the layered graph whose t-th layer is the choice of r_t. By induction on t, $\mathrm{dp}[t][b]$ equals the minimum, over all $1=\rho_1<\cdots<\rho_t=b$, of $\max_{2\le i\le t}\mathrm{gap}(\rho_{i-1},\rho_i)$. The base case $t=1$ is $\mathrm{dp}[1][1]=0$. For the induction step, any optimal chain ending at b has a penultimate element a, and its cost is $\max\{\text{cost of the initial chain ending at } a,\ \mathrm{gap}(a,b)\}$. This is minimized by taking the initial chain that is optimal for $(t-1,a)$, since $\mathrm{gap}(a,b)$ does not depend on that initial chain. Adding the terminal term at $b=r_q$ gives $\Psi^\star$.

Complexity. There are $O(qk)$ states with $O(k)$ transitions each, and each value of gap costs $O(1)$. □

C.4 The number of prefix checks needed for a given blind spot

How many prefix curves must be checked exactly to bring the equal-target blind spot down to β? By Proposition C.3(iii), checking the depths in R gives blind spot at most β exactly when $\Psi(R)\le 2\beta$. The next theorem gives an explicit construction of order $\log(1+k\beta)/\beta$; Corollary D.7 shows that no set of curves, prefix curves or not, can be of smaller order.

Theorem C.5 (Logarithmically many prefix checks suffice)**.** *Let* $0<\beta\le 1/4$ *and put* $\xi=2\beta\in(0,1/2]$*. Let* $q^\star_{\mathrm{prefix}}(k,\beta)=\min\{|R|:R\subseteq[k],\ \Psi(R)\le\xi\}$ *be the smallest number of exactly checked prefix depths whose equal-target blind spot is at most* β*. Then*

$$q^\star_{\mathrm{prefix}}(k,\beta)\ \le\ \frac{6\log(1+k\beta)}{\beta}. \tag{12}$$

Proof. By Proposition C.4, $\Psi(R)\le\xi$ if and only if $\mathrm{gap}(a,b)\le\xi$ for all consecutive $a<b$ in $R\cup\{0\}$ and $r_q\ge(1-\xi)k$.

If $k\xi < 1$, take $R = [k]$: then $q^\star_{\text{prefix}} \le k \le \frac{2\log(1+k\xi)}{\xi}$ because $\log(1+x) \ge x/2$ on $[0,1]$. Otherwise, let $A = \lceil 1/\xi \rceil$ and $\gamma = \frac{1+\xi}{1-\xi}$, and take

$$R = \{1, 2, \dots, A\} \cup \{\rho_1, \rho_2, \dots\}, \qquad \rho_0 = A, \quad \rho_{i+1} = \min\big(k,\ \lfloor \rho_i \gamma \rfloor\big),$$

stopping when $\rho_i = k$. Every consecutive pair (a, b) of R either has $b = a+1$, so that $\text{gap}(a,b) = 0$, or satisfies $b \le a\gamma$, and then $\text{gap}(a,b) \le \max_{a<r<b} \frac{\min(r-a, b-r)}{r} \le \frac{(b-a)/2}{(a+b)/2} = \frac{b-a}{a+b} \le \frac{\gamma-1}{\gamma+1} = \xi$, where the middle inequality bounds the maximum over the integers by the maximum over the reals, which is attained at $r = \frac{a+b}{2}$. All depths $\le A$ are checked and $k \in R$, so the terminal term is 0. Therefore $\Psi(R) \le \xi$. The sequence never stalls: $\rho_i \ge A \ge 1/\xi$ gives $\rho_i \xi \ge 1$, and $\gamma - 1 - \xi = \frac{(1+\xi)\xi}{1-\xi}$, so $\rho_i(\gamma - 1 - \xi) = \rho_i \xi \frac{1+\xi}{1-\xi} \ge \frac{1+\xi}{1-\xi} > 1$. Hence $\rho_{i+1} = \lfloor \rho_i \gamma \rfloor > \rho_i \gamma - 1 > \rho_i(1+\xi)$ whenever $\rho_{i+1} < k$. Thus the number of grid points ρ_i is at most $\frac{\log(k/A)}{\log(1+\xi)} + 1 \le \frac{\log(k\xi)}{\xi \log 2} + 1$, and

$$q^\star_{\text{prefix}} \le |R| \le A + \frac{\log(k\xi)}{0.69\,\xi} + 1 \le \frac{1}{\xi} + \frac{1.45\log(1+k\xi)}{\xi} + 2 \le \frac{6\log(1+k\xi)}{\xi},$$

using $\log(1+k\xi) \ge \log 2$ when $k\xi \ge 1$. Substituting $\xi = 2\beta$ and $\log(1+2k\beta) \le \log 2 + \log(1+k\beta) \le 2\log(1+k\beta)$ (for $k\beta \ge 1$) gives (12). For $k\beta < 1$, use $q^\star_{\text{prefix}} \le k$ as above. □

C.5 The first position: deterministic lists and a single checked curve

When the groups partition the items, every deterministic list shows a single group to all readers of the top result, which bounds its prefix disparity from below (Proposition C.6, cited in Section 4.3). A single checked curve that gives position 1 at most $\min(p, 1-p)$ of the attention has the largest possible blind spot (Proposition C.7, cited in Section 3.2).

Proposition C.6 (A fixed list has a large disparity at the first position)**.** *Suppose that the groups partition the items, and let P be a deterministic top-k list on a query, with g_1 the group of the item at position 1. Then $\max_{g,r} d_{g,r} \ge \max\{1 - p_{g_1},\ \max_{g \ne g_1} p_g\}$, which equals $\max_g \max(p_g, 1-p_g)$ when all targets are equal. No such lower bound holds for overlapping groups.*

Proof. At $r = 1$, the occupancy is $x_{g,1} = \mathbf{1}\{g = g_1\}$, because exactly one item is shown at position 1 and the groups partition the items. Hence $d_{g_1,1} = 1 - p_{g_1}$ and $d_{g,1} = p_g$ for $g \ne g_1$, and taking the largest of these values gives the bound. Under equal targets $p_g = 1/m$, the first term is $1 - 1/m$ and the second $1/m$, whose maximum is $\max_g \max(p_g, 1-p_g)$. For overlapping groups the argument fails: with $G_1 = G_2 =$ all items and $p_1 = p_2 = 1$, every deterministic list has $x_{g,j} = 1$ for all g, j, so $\max_{g,r} d_{g,r} = 0$, while the right-hand side of the bound equals 1. □

Proposition C.7 (One checked curve with an unequal target)**.** *Let two groups partition the items, each with at least k items, let $p \in (0,1)$ be the target share of group 1, and let $v \in \mathcal{V}_k$ be the single checked curve. Then the blind spot of group 1 satisfies $B_{k,p}(\{v\}; 0) \le \max(p, 1-p)$ for every v, with equality whenever $v_1 \le \min(p, 1-p)$.*

Proof. For any occupancy vector $x \in [0,1]^k$ and depth r, $\langle u^r, x\rangle \in [0,1]$, so $|\langle u^r, x\rangle - p| \le \max(p, 1-p)$; this gives the upper bound. For the lower bound, take $x = (1, c, \dots, c)$ with $c = (p - v_1)/(1 - v_1)$. Then $\langle v, x\rangle = v_1 + (1-v_1)c = p$, so the check passes exactly, and $\langle u^1, x\rangle - p = 1 - p$. The point lies in the cube iff $0 \le c \le 1$, which holds iff $v_1 \le p$. Symmetrically, $x = (0, c', \dots, c')$ with $c' = p/(1-v_1)$ passes the check and has $p - \langle u^1, x\rangle = p$, and it lies in the cube iff $v_1 \le 1 - p$. If $p \le 1/2$, the first witness attains $\max(p, 1-p) = 1 - p$ and needs $v_1 \le p = \min(p, 1-p)$. If $p > 1/2$, the second attains p and needs $v_1 \le 1 - p = \min(p, 1-p)$. □

C.6 The blind spot for general targets and groups

We now compute the blind spot for a general target by linear-programming duality (Corollary C.8), prove Theorem 4 for arbitrary groups and targets, and show that the single-target values bound the blind spot of every query and that this bound is attained, by disjoint groups when the targets sum to one and in general by a query with overlapping groups (Proposition C.9).

Corollary C.8 (Blind spot for a general target)**.** *Consider the setting of Proposition C.7: one group with target $p \in (0,1)$ whose occupancy vector ranges over the cube $[0,1]^k$. For every normalized curve set $\mathcal{S}$ with matrix V and every $\delta \geq 0$,*

$$B_{k,p}(\mathcal{S};\delta) = \max_r \max_{s\in\{\pm 1\}} \inf_{\alpha\in\mathbb{R}^q} \Big\{ (1-p)\,\|(su^r - V\alpha)^+\|_1 + p\,\|(su^r - V\alpha)^-\|_1 + \delta\|\alpha\|_1 \Big\},$$

where $w^\pm$ denote the positive and negative parts of a vector w. At $p = \frac{1}{2}$, this is (5). For several groups on one query, the same LP duality applies to the query's occupancy polytope, and the joint blind spot is the maximum, over groups, depths and signs, of the values of one linear program each (proof of Theorem 4 below).

Proof. Write $x = p\mathbf{1} + y$ with $y \in [-p, 1-p]^k$. Because every checked curve and every u^r is normalized, each check reads $|\langle v, y\rangle| \leq \delta$ and the signed prefix deviation at depth r is $\langle u^r, y\rangle$, so $B_{k,p}(\mathcal{S};\delta) = \max_{r,s} \sup\{s\langle u^r, y\rangle : \|V^\top y\|_\infty \leq \delta,\ y \in [-p,1-p]^k\}$. For fixed (r,s), this is a linear program. Its dual, with multipliers α for the slab constraints $|\langle v^\ell, y\rangle| \leq \delta$, is $\inf_\alpha\{h(su^r - V\alpha) + \delta\|\alpha\|_1\}$, where $h(w) = \sup\{\langle w, y\rangle : y \in [-p,1-p]^k\} = (1-p)\|w^+\|_1 + p\|w^-\|_1$ is the support function of the asymmetric box $[-p,1-p]^k$. Strong duality holds because the primal is feasible ($y = 0$) and bounded. The multi-group statement is proved with Theorem 4 below. □

Proof of Theorem 4. Because $|a| = \max(a,-a)$, the blind spot (4) is the largest value, over groups g, depths r and signs, of the linear program

$$\max\Big\{ \pm\big(\langle u^r, x_g\rangle - p_g\big) :\ x \in \mathcal{X},\ \big|\langle v^\ell, x_{g'}\rangle - p_{g'}\big| \leq \delta \text{ for all } g' \in [m],\ \ell \in [q] \Big\},$$

which is feasible by hypothesis and bounded because $\mathcal{X}$ is a polytope. Fix g, r and the sign. Dualize the two inequalities of every check with multipliers $\mu^+_{g',\ell}, \mu^-_{g',\ell} \geq 0$, and write $\alpha = \mu^+ - \mu^-$. The Lagrangian bound is

$$\max_{x\in\mathcal{X}} \Big[\pm\big(\langle u^r, x_g\rangle - p_g\big) - \sum_{g',\ell} \alpha_{g',\ell}\big(\langle v^\ell, x_{g'}\rangle - p_{g'}\big) \Big] + \delta \sum_{g',\ell} \big(\mu^+_{g',\ell} + \mu^-_{g',\ell}\big),$$

and the smallest value of its last term over the decompositions of a fixed α is $\delta\|\alpha\|_1$. Strong duality for linear programs gives (6).

It remains to describe the inner maximum. The set $\mathcal{X}$ is the image of the policy polytope under the linear map $P \mapsto (\sum_{i\in G_{g'}} P_{ij})_{g',j}$, so the bracket is affine in P, and the inner maximum is an assignment problem in which showing item i at position j earns $\sum_{g':\, i\in G_{g'}} w_{g',j}$. Here $w_{g',j} = \pm u^r_j \mathbf{1}\{g' = g\} - \sum_\ell \alpha_{g',\ell} v^\ell_j$ collects the coefficients of x. Because every curve is normalized, the bracket equals $\langle w, x - \bar{x}\rangle$, where $\bar{x}$ is the target occupancy profile, with $\bar{x}_{g,j} = p_g$. The inner maximum is therefore the support function of $\mathcal{X} - \bar{x}$ at w. The argument does not require $\bar{x} \in \mathcal{X}$. □

Proposition C.9 (The single-group formula gives the worst case over queries)**.** *Let $B(\mathcal{S};\delta)$ be the blind spot of a query as defined in Section 3.2, and let $B_{k,p}(\mathcal{S};\delta)$ be the single-target value of Corollary C.8. For every query, $B(\mathcal{S};\delta) \leq \max_g B_{k,p_g}(\mathcal{S};\delta)$. If the groups partition the items, each group has at least k items and the targets sum to one, then equality holds. For arbitrary targets $p_1,\dots,p_m \in (0,1)$, equality holds on the query that contains, for every subset $T \subseteq [m]$, k items that belong to exactly the groups in T. Hence $\max_g B_{k,p_g}(\mathcal{S};\delta)$ is the largest blind spot over all queries with these targets.*

Proof. The inequality. Fix a group g, a depth r and a sign s. If $x \in \mathcal{X}$ passes every check, then x_g lies in $[0,1]^k$ and passes the checks of group g, so $s(\langle u^r, x_g\rangle - p_g)$ is at most the value of the single-group problem on the cube, which is at most $B_{k,p_g}(\mathcal{S};\delta)$ by Corollary C.8. This proves the inequality.

Equality for partitioned groups. Let the groups partition the items, and let each group g' have $n_{g'} \geq k$ items. Then every $x \geq 0$ with $\sum_{g'} x_{g',j} = 1$ for all j is realizable: show each item of group g' at position j with probability $x_{g',j}/n_{g'}$, so that every column of P sums to one and every row sums to at most $k/n_{g'} \leq 1$. Let $y \in [0,1]^k$ attain the single-group value for group g, and define $x_g = y$ and $x_{g'} = \frac{p_{g'}}{1-p_g}(\mathbf{1} - y)$ for $g' \neq g$. Because $\sum_{g'\neq g} p_{g'} = 1 - p_g$, the columns of x sum to one, so

$x \in \mathcal{X}$. For every checked curve v, normalization gives $\langle v, x_{g'} \rangle - p_{g'} = \frac{p_{g'}}{1-p_g}\big(p_g - \langle v, y \rangle\big)$, whose absolute value is at most δ because $p_{g'} \le 1 - p_g$. Hence x passes every check and hides the same disparity as y.

Equality for arbitrary targets. Take the query that contains, for every $T \subseteq [m]$, k items that belong to exactly the groups in T. Let y attain the single-group value for group g; we construct a policy with $x_g = y$ and $x_{g'} = p_{g'}\mathbf{1}$ for $g' \neq g$. At position j, show an item with membership T with probability $\pi_j(T) = \prod_{h \in T} x_{h,j} \prod_{h \notin T}(1 - x_{h,j})$, divided evenly among the k items with that membership. Since $\sum_T \pi_j(T) = 1$, every column of P sums to one, and every row sums to at most $k \cdot \frac{1}{k} = 1$. The occupancy of group h at position j is $\sum_{T \ni h} \pi_j(T) = x_{h,j}$, as intended. Here $x_{g'} = p_{g'}\mathbf{1}$ passes every normalized check exactly, so this policy passes every check and hides the same disparity as y. □

C.7 Audit size for general targets

The next two corollaries bound the audit size for general targets by the audit size $q^\star(k, \cdot)$ for two groups with equal targets, which Appendix D computes exactly and which equals the number $N(k, \cdot)$ of blocks in the main text. Corollary C.11 gives the bounds of Theorem 5.

Corollary C.10 (Upper and lower bounds on the audit size for a general target)**.** *Let $q_p^\star(k, \beta)$ be the smallest size of a normalized curve set whose blind spot for one group with target p under exact checks, $B_{k,p}(\mathcal{S}; 0)$ (Corollary C.8 with $\delta = 0$), is at most β, and write $p_+ = \max(p, 1-p)$ and $p_- = \min(p, 1-p)$. Then*

$$q^\star\Big(k, \tfrac{\beta}{2p_-}\Big) \;\le\; q_p^\star(k, \beta) \;\le\; q^\star\Big(k, \tfrac{\beta}{2p_+}\Big),$$

where $q^\star(k, \cdot)$ is the audit size for two groups with equal targets (Theorem 5). The upper bound is attained by the endpoint curves built for $\beta/(2p_+)$, and the lower bound holds for arbitrary linear subspaces. In particular, $q_p^\star(k, \beta) = \Theta_p(\log(1 + k\beta)/\beta)$ for every fixed $p \in (0, 1)$.

Proof. For every $w \in \mathbb{R}^k$, the support function of the asymmetric box satisfies $p_-\|w\|_1 \le (1-p)\|w^+\|_1 + p\|w^-\|_1 \le p_+\|w\|_1$. Since $\|su^r - V\alpha\|_1 = \|u^r - V(s\alpha)\|_1$ for $s = \pm 1$, taking the infimum over α and the maximum over r and s in Corollary C.8 at $\delta = 0$ gives $2p_- B_k(\mathcal{S}; 0) \le B_{k,p}(\mathcal{S}; 0) \le 2p_+ B_k(\mathcal{S}; 0)$, where $B_k(\mathcal{S}; 0) = \frac{1}{2}\max_r \operatorname{dist}_1(u^r, \operatorname{span}\mathcal{S})$ is the equal-target blind spot. Hence a curve set with $B_k \le \beta/(2p_+)$ has $B_{k,p} \le \beta$; this gives the upper bound, which is attained by the endpoint curves of Section 3.3 at that level. Conversely, any subspace with $B_{k,p} \le \beta$ has $B_k \le \beta/(2p_-)$, so its dimension is at least $q^\star(k, \beta/(2p_-))$ by the subspace lower bound of Corollary D.6. The order statement follows from the logarithmic law for $q^\star(k, \cdot)$ (Corollary D.7, which rests on Corollary D.6), applied at both levels. □

Corollary C.11 (Audit size for several groups)**.** *Let $p_1, \dots, p_m \in (0, 1)$ be arbitrary targets, let $p_+ = \max_g \max(p_g, 1-p_g)$ and $p_- = \max_g \min(p_g, 1-p_g)$, and let $q^\star$ be the smallest size of a normalized curve set whose blind spot $B(\mathcal{S}; 0)$ is at most β on every query with these targets, where $\beta < p_-$. By Proposition C.9, $q^\star$ is also the smallest size of a curve set with $\max_g B_{k,p_g}(\mathcal{S}; 0) \le \beta$. Then $N(k, \beta/(2p_-)) \le q^\star \le N(k, \beta/(2p_+))$, which is Theorem 5 of the main text, because $q^\star(k, \cdot) = N(k, \cdot)$ (Corollary D.6). The upper bound is attained by the endpoint curves built for $\beta/(2p_+)$, and the lower bound holds for arbitrary linear subspaces. Consequently, for fixed targets and $0 < \beta \le p_-/2$, $q^\star$ lies between two multiples of $\log(1 + k\beta)/\beta$ that depend only on the targets, which is the logarithmic growth stated in Section 3.3.*

Proof. The proof of Corollary C.10 gives $2\min(p, 1-p)B_k(\mathcal{S}; 0) \le B_{k,p}(\mathcal{S}; 0) \le 2\max(p, 1-p)B_k(\mathcal{S}; 0)$ for every p. Taking the maximum over groups yields $2p_- B_k(\mathcal{S}; 0) \le \max_g B_{k,p_g}(\mathcal{S}; 0) \le 2p_+ B_k(\mathcal{S}; 0)$, and the argument of that proof applies with these two constants.

For the logarithmic growth, both levels $\beta/(2p_\pm)$ lie in $(0, \frac{1}{4}]$ when $\beta \le p_-/2$, so the logarithmic law of Corollary D.7, which rests on Corollary D.6, applies to them. Moreover, $\beta \le \beta/(2p_-)$ and $\beta/(2p_+) \le \beta$ convert $\log(1 + k\beta/(2p_\pm))$ into $\log(1 + k\beta)$ at the cost of the factors $2p_-$ and $2p_+$. □

D THE SMALLEST AUDIT FOR GENERAL CURVE SETS

This appendix proves Theorem 5 for two groups with equal targets, to which Corollaries C.10 and C.11 (Appendix C) reduce general targets and groups. For exact checks, the smallest audit size $q^\star(k,\beta)$ equals the number of blocks of the endpoint construction of Section 3.3, and no linear subspace of smaller dimension does as well (Corollary D.6). Appendix D.1 treats positive tolerances.

Theorem C.5 bounds the number of checks needed when the checked curves are prefix curves. Because $\mathcal{V}_k$ spans $\mathbb{R}^k$, a set of general monotone curves, each combining many depths, might need only $\Theta(1/\beta)$ checks. The results below determine $q^\star(k,\beta)$ exactly for arbitrary curve sets and show that the factor $\log(1+k\beta)$ remains, so audits of prefix curves are within a small constant factor of optimal (asymptotically, within a factor 1; Corollary D.7).

Relation to classical results. The representation $v = \sum_r \lambda_r u^r$ of Lemma 1 is the standard representation of a monotone weight vector in spectral risk measures, ordered weighted averages and the C/W/L framework (Kusuoka, 2001; Acerbi, 2002; Yager, 1988; Moffat et al., 2017), and it yields Theorem 2 directly. Optimal recovery (Micchelli & Rivlin, 1977; Traub et al., 1988) supplies a general fact: the worst-case error of recovering a linear functional from noisy linear information over a convex balanced set is a radius of information, given by a dual minimization. Theorem 3 is that radius, computed exactly for the functionals $x \mapsto \langle u^r, x\rangle$, the information $V^\top x$ and the set $[-1,1]^k$. It yields the closed-form regularized ℓ_1 expression (5) and, through it, the values in Table 1. The theory of n-widths (Pinkus, 1985) defines $q^\star(k,\beta)$ as the smallest dimension of a subspace that approximates the finite set $\{u^1,\dots,u^k\}$ in ℓ_1 (taken over all subspaces, this smallest dimension is $q^\star_{\mathrm{lin}}(k,\beta)$ below, which equals $q^\star(k,\beta)$ by Corollary D.6), but it gives no value for it. We compute this width of a finite set exactly: Theorem D.5 gives a lower bound through γ-chains, Theorem D.3 matches it with a construction from monotone curves through γ-covers, and the chain–cover duality of Lemma D.2 shows that the two bounds agree. The lower bound holds for every linear subspace, not only for spans of monotone curves. The positive-tolerance results of Appendix D.1 are not covered by these sources.

For a subspace $L \subseteq \mathbb{R}^k$, let $\Psi(L)$ be the largest ℓ_1 distance from a prefix curve to L,

$$\Psi(L) \;:=\; \max_{r\in[k]} \operatorname{dist}_1\big(u^r, L\big),$$

so that, by Theorem 3 at $\delta = 0$, the equal-target blind spot of a curve set $\mathcal{S}$ is $B_k(\mathcal{S};0) = \frac{1}{2}\Psi(\operatorname{span}\mathcal{S})$. Put

$$q^\star(k,\beta) = \min\big\{q :\ \exists\,\mathcal{S}\subset\mathcal{V}_k,\ |\mathcal{S}| = q,\ \Psi(\operatorname{span}\mathcal{S}) \le 2\beta\big\},$$
$$q^\star_{\mathrm{lin}}(k,\beta) = \min\big\{\dim L :\ \Psi(L) \le 2\beta\big\}.$$

The first minimum is over sets of q monotone curves. The second is over *all* linear subspaces $L \subseteq \mathbb{R}^k$, whether or not they are spanned by monotone curves. Trivially $q^\star_{\mathrm{lin}} \le q^\star$, because $\operatorname{span}\mathcal{S}$ has dimension at most $|\mathcal{S}|$.

Several results below are stated for the ℓ_1 level $\xi = 2\beta$ that corresponds to blind spot β (subspaces with $\Psi \le \xi$); the ratio of Section 3.3 is then $\gamma = \frac{1+\xi}{1-\xi}$.

Definition D.1 (γ-chains and γ-covers)**.** *Let $\gamma \ge 1$ and $k \ge 1$.*

*(i) A γ-*chain *is a set of integers $1 \le s_0 < s_1 < \cdots < s_p \le k$ with $s_l > \gamma\, s_{l-1}$ for every $l \in [p]$. Let $N_{\mathrm{chain}}(k,\gamma)$ be the largest cardinality $p+1$ of a γ-chain.*

*(ii) A γ-*cover *is a partition of $[k]$ into consecutive integer intervals $[c_1,d_1],[c_2,d_2],\dots$ (so $c_1 = 1$, $c_{i+1} = d_i + 1$, and the last interval ends at k) with $d_i \le \gamma\, c_i$ for every i. Let $N_{\mathrm{cover}}(k,\gamma)$ be the least number of intervals in a γ-cover.*

Thus a γ-cover splits the depths $1,\dots,k$ into consecutive blocks whose last depth is at most γ times their first, as in Section 3.3; a γ-chain is a set of depths whose consecutive ratios all exceed γ.

Lemma D.2 (Grouping the depths greedily is optimal)**.** *For every $\gamma \ge 1$ and every $k \ge 1$, $N_{\mathrm{chain}}(k,\gamma) = N_{\mathrm{cover}}(k,\gamma)$, and both are attained by the greedy cover*

$$c_1 = 1, \qquad d_i = \min\big(\lfloor \gamma c_i \rfloor,\ k\big), \qquad c_{i+1} = d_i + 1,$$

whose left endpoints $c_1 < c_2 < \dots$ form a longest γ-chain.

Proof. $N_{\text{chain}} \le N_{\text{cover}}$. Fix a γ-cover and a γ-chain. No interval $[c, d]$ of the cover contains two chain elements $s < s'$: that would force $s'/s \le d/c \le \gamma$, contradicting $s' > \gamma s$. As the intervals partition $[k]$, the chain has at most as many elements as the cover has intervals. Taking a longest chain and a cover with the fewest intervals gives the claim.

$N_{\text{cover}} \le N_{\text{chain}}$. The greedy recursion is well posed: $\gamma \ge 1$ gives $\lfloor \gamma c_i \rfloor \ge c_i$, so $c_i \le d_i$ and $c_{i+1} > c_i$, and the recursion terminates with $d_i = k$. By construction $d_i \le \gamma c_i$, so the greedy cover *is* a γ-cover. Let N_{greedy} be its number of intervals; then $N_{\text{cover}} \le N_{\text{greedy}}$. Its left endpoints form a γ-chain: if $i + 1 \le N_{\text{greedy}}$, then $d_i < k$, so $d_i = \lfloor \gamma c_i \rfloor$ and $c_{i+1} = \lfloor \gamma c_i \rfloor + 1 > \gamma c_i$. Hence $N_{\text{chain}} \ge N_{\text{greedy}}$. Combining, $N_{\text{chain}} \ge N_{\text{greedy}} \ge N_{\text{cover}} \ge N_{\text{chain}}$, so all three coincide. □

Upper bound: endpoint curves

For each block $[c_i, d_i]$ of a γ-cover, the construction checks the endpoint curve $v^{c_i,d_i} = \frac{1}{2}(u^{c_i} + u^{d_i})$ of Section 3.3, written v^i below.

Theorem D.3 (Endpoint curves attain the blind spot (upper bound)). *Let $\xi \in [0, 1)$ and $\gamma = \frac{1+\xi}{1-\xi}$. Let $[c_1, d_1], \dots, [c_q, d_q]$ be any γ-cover of $[k]$ and define the* endpoint curves

$$v^i := \tfrac{1}{2}\big(u^{c_i} + u^{d_i}\big) \in \mathcal{V}_k, \qquad i \in [q].$$

Then the v^i are distinct and linearly independent, and

$$\Psi\big(\operatorname{span}\{v^1, \dots, v^q\}\big) \le \max_{i \in [q]} \frac{d_i - c_i}{d_i + c_i} \le \xi, \qquad \textit{hence} \qquad B_k\big(\{v^1, \dots, v^q\}; 0\big) \le \tfrac{\xi}{2}.$$

In particular $q^\star(k, \beta) \le N_{\text{cover}}\big(k, \frac{1+2\beta}{1-2\beta}\big)$.

Proof. Membership and independence. Each v^i is a convex combination of two prefix curves, hence lies in the convex set $\mathcal{V}_k$. Since $c_i \le d_i$, the support of v^i is exactly $[1, d_i]$, and $d_1 < d_2 < \cdots < d_q$. Suppose $\sum_i \alpha_i v^i = 0$. Only v^q is nonzero at coordinate d_q, so reading that coordinate gives $\alpha_q v^q_{d_q} = 0$, hence $\alpha_q = 0$, and downward induction gives $\alpha = 0$. Linear independence implies in particular that the v^i are distinct.

Value. By LP duality, $\operatorname{dist}_1(w, L) = \max\{\langle y, w\rangle : \|y\|_\infty \le 1,\ y \perp L\}$. It therefore suffices to prove the following: every $y \in \mathbb{R}^k$ with $\|y\|_\infty \le 1$ and $\langle v^i, y\rangle = 0$ for all i satisfies $|\langle u^r, y\rangle| \le \theta_i$ for every block $[c_i, d_i]$ and every $r \in [c_i, d_i]$, where $\theta_i := \frac{d_i - c_i}{d_i + c_i}$.

Write $F_r = \sum_{j \le r} y_j$ and $F_0 = 0$, so that $\langle u^r, y\rangle = F_r/r$. Because $|y_j| \le 1$,

$$|F_r - F_s| \le r - s \qquad \text{for all } 0 \le s \le r \le k. \tag{13}$$

Fix a block and abbreviate $a = c_i$, $b = d_i$ and $\theta = \theta_i \in [0, 1)$. Note the identity

$$\frac{b}{a} = \frac{(b + a) + (b - a)}{(b + a) - (b - a)} = \frac{1 + \theta}{1 - \theta}. \tag{14}$$

The constraint $\langle v^i, y\rangle = 0$ reads $\frac{F_a}{a} + \frac{F_b}{b} = 0$, that is,

$$F_b = -\frac{b}{a} F_a. \tag{15}$$

Put $x := F_a$, which may have either sign. We claim that $|F_r| \le \theta r$ for every $r \in [a, b]$. Since the blocks partition $[k]$, this claim proves the theorem.

Upper half. By (13), $F_r \le x + (r - a)$. Using also (15), $F_r \le F_b + (b - r) = -\frac{b}{a}x + (b - r)$. Now

$$x + (r - a) \le \theta r \iff r(1 - \theta) \le a - x \iff r \le \frac{a - x}{1 - \theta},$$

$$-\tfrac{b}{a}x + (b - r) \le \theta r \iff \tfrac{b}{a}(a - x) \le r(1 + \theta) \iff r \ge \frac{b}{a(1 + \theta)}(a - x) \overset{(14)}{=} \frac{a - x}{1 - \theta},$$

where every step divides only by the positive numbers $1-\theta$ and $1+\theta$, so no assumption on the sign of x is used. The two conditions are thus "$r \le T$" and "$r \ge T$" for the *same* threshold $T = \frac{a-x}{1-\theta}$. One of them always holds, so one of the two upper bounds on F_r is at most θr, and hence $F_r \le \theta r$.

Lower half. Symmetrically, $F_r \ge x - (r-a)$ and $F_r \ge -\frac{b}{a}x - (b-r)$, and

$$x - (r-a) \ge -\theta r \iff r \le \frac{a+x}{1-\theta},$$
$$-\tfrac{b}{a}x - (b-r) \ge -\theta r \iff r \ge \frac{b}{a(1+\theta)}(a+x) = \frac{a+x}{1-\theta},$$

again a single threshold, so $F_r \ge -\theta r$. Thus $|F_r| \le \theta r$ on $[a,b]$, and hence $|\langle u^r, y\rangle| \le \theta$ there. Finally, by (14), the cover condition $d_i \le \gamma c_i$ is equivalent to $\theta_i \le \xi$. □

Remark D.4 (Why two endpoints beat one). *In the dual of the proof above, a check of the prefix curve of depth s forces $F_s = 0$. The endpoint curve $\frac{1}{2}(u^a + u^b)$ imposes a weaker constraint: it only ties F_a and F_b together through (15). The weaker constraint is enough because the admissible region $|F_r| \le \theta r$ is a* cone *that widens with r: by (14), the two Lipschitz cones emanating backward from (b, F_b) and forward from (a, F_a) cross exactly on the boundary of that cone, whatever the common value of $|F_a|/a = |F_b|/b$. One curve therefore controls a whole block $[a,b]$ with $b/a \le \gamma$. Prefix curves, in contrast, control only blocks of* additive *half-width at most ξr near depth r.*

Lower bound for every linear subspace

The matching lower bound does not assume that the checked vectors are monotone curves: it holds for every linear subspace of $\mathbb{R}^k$.

Theorem D.5 (No smaller set of curves does better (lower bound)). *Let $\xi \in [0,1)$, $\gamma = \frac{1+\xi}{1-\xi}$, and let $L \subseteq \mathbb{R}^k$ be* any *linear subspace with $\Psi(L) \le \xi$. Then*

$$\dim L \;\ge\; N_{\mathrm{chain}}(k,\gamma).$$

Consequently $q^\star_{\mathrm{lin}}(k,\beta) \ge N_{\mathrm{chain}}\big(k, \frac{1+2\beta}{1-2\beta}\big)$ for every $\beta < \frac{1}{2}$.

Proof. Let $1 \le s_0 < s_1 < \cdots < s_p \le k$ be any γ-chain. Set $s_{-1} := 0$, $u^{s_{-1}} := 0$ and $\rho_l := s_{l-1}/s_l \in [0,1)$ for $0 \le l \le p$ (so $\rho_0 = 0$). We build $p+1$ elements of L that are close enough to vectors with disjoint supports to be linearly independent.

For each l, the distance $\mathrm{dist}_1(u^{s_l}, L)$ is attained, because L is a finite-dimensional subspace and $w \mapsto \|u^{s_l} - w\|_1$ is continuous and coercive on it. Choose $w^l \in L$ with $\|u^{s_l} - w^l\|_1 \le \xi$, and set $w^{-1} := 0$. Define

$$z^l := w^l - \rho_l\, w^{l-1} \in L, \qquad \zeta^l := u^{s_l} - \rho_l\, u^{s_{l-1}} \qquad (0 \le l \le p).$$

Writing $\mathbf{1}_I \in \mathbb{R}^k$ for the indicator vector of a set I of positions, $\rho_l\, u^{s_{l-1}} = \frac{s_{l-1}}{s_l} \cdot \frac{1}{s_{l-1}} \mathbf{1}_{[1,s_{l-1}]} = \frac{1}{s_l}\mathbf{1}_{[1,s_{l-1}]}$, and we get the *annulus* form

$$\zeta^l = \frac{1}{s_l}\,\mathbf{1}_{(s_{l-1},\, s_l]}, \qquad \eta_l := \|\zeta^l\|_1 = \frac{s_l - s_{l-1}}{s_l} = 1 - \rho_l,$$

so $\zeta^0, \dots, \zeta^p$ have *pairwise disjoint supports* $(s_{l-1}, s_l]$. Moreover, by the triangle inequality,

$$\varepsilon_l := \big\|z^l - \zeta^l\big\|_1 = \big\|(w^l - u^{s_l}) - \rho_l(w^{l-1} - u^{s_{l-1}})\big\|_1 \le \xi(1+\rho_l).$$

The chain condition $s_l > \gamma s_{l-1}$ says exactly that $\rho_l < \gamma^{-1} = \frac{1-\xi}{1+\xi}$, which rearranges to

$$\eta_l - \varepsilon_l \;\ge\; (1-\rho_l) - \xi(1+\rho_l) \;=\; (1+\xi)\Big(\frac{1-\xi}{1+\xi} - \rho_l\Big) \;>\; 0 \tag{16}$$

for every $l \in [p]$. For $l = 0$, the positivity $\eta_0 - \varepsilon_0 > 0$ holds because $\eta_0 = 1$ and $\varepsilon_0 \le \xi < 1$.

Now suppose $\sum_{l=0}^{p} \alpha_l z^l = 0$. Using the disjointness of the supports for the first equality, this assumption for the second, and the triangle inequality for the last step,

$$\sum_{l=0}^{p} |\alpha_l|\,\eta_l = \Big\|\sum_l \alpha_l \zeta^l\Big\|_1 = \Big\|\sum_l \alpha_l\big(\zeta^l - z^l\big)\Big\|_1 \le \sum_{l=0}^{p} |\alpha_l|\,\varepsilon_l,$$

so $\sum_l |\alpha_l|(\eta_l - \varepsilon_l) \le 0$, and (16) forces $\alpha = 0$. Hence $z^0, \dots, z^p$ are $p+1$ linearly independent elements of L, and $\dim L \ge p+1$. Maximizing over γ-chains gives $\dim L \ge N_{\mathrm{chain}}(k,\gamma)$. □

THE EXACT AUDIT SIZE

Corollary D.6 below is Theorem 5 of the main text for two groups with equal targets, and Corollary D.7 is the logarithmic growth stated in Section 3.3. The main text's $N(k,\beta')$ equals $N_{\mathrm{cover}}(k,\gamma)$ with $\gamma = \frac{1+2\beta'}{1-2\beta'}$, which corresponds to $\xi = 2\beta'$. In Theorem 5, $q^\star$ is the smallest number of curves whose blind spot is at most β on every query with the given targets. For general targets and groups, Proposition C.9 identifies the worst query, and Corollaries C.10 and C.11 reduce the bounds to the case of equal targets.

Corollary D.6 (The exact audit size, attained by endpoint curves)**.** *For every $k \geq 1$ and $0 \leq \beta < \frac{1}{2}$, with $\gamma = \frac{1+2\beta}{1-2\beta}$,*

$$q^\star(k,\beta) \;=\; q^\star_{\mathrm{lin}}(k,\beta) \;=\; N_{\mathrm{chain}}(k,\gamma) \;=\; N_{\mathrm{cover}}(k,\gamma),$$

and the minimum is attained by the monotone *endpoint curves $\left\{\frac{1}{2}(u^{c_i}+u^{d_i})\right\}_{i\leq N_{\mathrm{cover}}(k,\gamma)}$ of the greedy γ-cover. Equivalently, for $1 \leq q < k$, the Kolmogorov ℓ_1 width of the prefix curves is*

$$\min_{\dim L = q}\ \max_{r\in[k]} \mathrm{dist}_1\big(u^r, L\big) \;=\; \frac{\Gamma_{q+1}(k)-1}{\Gamma_{q+1}(k)+1}, \qquad \Gamma_m(k) := \max_{1\leq s_0<\cdots<s_{m-1}\leq k}\ \min_{1\leq l\leq m-1} \frac{s_l}{s_{l-1}}.$$

For $q \geq k$ the width is 0.

Proof. First, $q^\star \leq N_{\mathrm{cover}}$ by Theorem D.3 applied with $\xi = 2\beta$: the greedy γ-cover has $N_{\mathrm{cover}}(k,\gamma)$ intervals (Lemma D.2), so its endpoint curves are exactly $N_{\mathrm{cover}}(k,\gamma)$ distinct curves in $\mathcal{V}_k$. Next, $q^\star_{\mathrm{lin}} \geq N_{\mathrm{chain}}$ by Theorem D.5, $N_{\mathrm{chain}} = N_{\mathrm{cover}}$ by Lemma D.2, and $q^\star_{\mathrm{lin}} \leq q^\star$; together, these give the equalities. The displayed formula follows by inverting the increasing map $\xi \mapsto \gamma = \frac{1+\xi}{1-\xi}$: a γ-chain of $q+1$ elements exists if and only if $\gamma < \Gamma_{q+1}(k)$, so the least ξ with $N_{\mathrm{chain}}(k,\gamma) \leq q$ solves $\frac{1+\xi}{1-\xi} = \Gamma_{q+1}(k)$. □

Corollary D.7 (The number of curves grows logarithmically)**.** *For $0 < \beta \leq \frac{1}{4}$ and all k, with $\gamma = \frac{1+2\beta}{1-2\beta}$,*

$$\log_\gamma\big(2k\beta\big) \;\leq\; q^\star(k,\beta) \;\leq\; 1+\log_\gamma k, \qquad \textit{and} \qquad \frac{\log(1+k\beta)}{16\,\beta} \;\leq\; q^\star(k,\beta) \;\leq\; 1+\frac{\log k}{4\beta}.$$

In particular $q^\star(k,\beta) = \Theta\big(\log(1+k\beta)/\beta\big)$ for arbitrary sets of normalized monotone curves, the same order as for prefix curves. The upper bound of that order is $q^\star \leq q^\star_{\mathrm{prefix}} \leq 6\log(1+k\beta)/\beta$ (Theorem C.5) together with $q^\star \leq k$; the displayed bound $1+\log k/(4\beta)$ is sharper only when $k\beta \gg 1$. For fixed β, $q^\star(k,\beta) = \frac{\log k}{\log\gamma} + O_\beta(1)$ as $k\to\infty$.

Proof. Write $\xi = 2\beta \in (0, \frac{1}{2}]$, so that $\gamma - 1 = \frac{2\xi}{1-\xi}$ and $\frac{\gamma}{\gamma-1} = \frac{1+\xi}{2\xi}$. By Corollary D.6, it suffices to bound the greedy chain $t_0 = 1$, $t_{i+1} = \lfloor \gamma t_i \rfloor + 1$, whose elements in $[k]$ are the left endpoints $t_i = c_{i+1}$ of the greedy cover and hence form a longest γ-chain (Lemma D.2).

Upper bound. If the greedy cover of Lemma D.2 has one block, the claim is immediate. Otherwise, its left endpoints satisfy $c_i > \gamma^{i-1}$ for every $i \geq 2$: the base case follows from $c_2 = \lfloor \gamma c_1 \rfloor + 1 > \gamma$, and the induction step is identical. By Lemma D.2, the greedy cover has $N_{\mathrm{cover}}(k,\gamma)$ blocks; writing N_{cover} for this number, $\gamma^{N_{\mathrm{cover}}-1} < c_{N_{\mathrm{cover}}} \leq k$. Hence $N_{\mathrm{cover}} < 1+\log_\gamma k$, and the claimed weak inequality follows. Since $\log\gamma = \log\frac{1+\xi}{1-\xi} \geq 2\xi = 4\beta$, this gives $q^\star \leq 1+\frac{\log k}{4\beta}$.

Lower bound. From $t_{i+1} \leq \gamma t_i + 1$ and $t_0 = 1$, $t_i \leq \gamma^i + \sum_{j<i}\gamma^j \leq \gamma^i \frac{\gamma}{\gamma-1} = \gamma^i \frac{1+\xi}{2\xi}$, so $t_i \leq k$ whenever $\gamma^i \leq \frac{2k\xi}{1+\xi}$. As $\xi \leq \frac{1}{2}$, we have $\frac{2\xi}{1+\xi} \geq \xi$, and hence $N_{\mathrm{chain}} \geq 1+\big\lfloor \log_\gamma(k\xi)\big\rfloor \geq \log_\gamma(k\xi) = \log_\gamma(2k\beta)$. Also, if $t_i < \frac{1}{\gamma-1} = \frac{1-\xi}{2\xi}$, then $\gamma t_i < t_i + 1$, so $t_{i+1} = t_i + 1$. The greedy chain therefore contains every integer up to $\min\big(k, \lceil \frac{1-\xi}{2\xi} \rceil\big)$, whence $N_{\mathrm{chain}} \geq \min\big(k, \frac{1}{4\xi}\big) = \min\big(k, \frac{1}{8\beta}\big)$ (using $\xi \leq \frac{1}{2}$).

We now derive the second displayed lower bound from these two bounds. If $k\xi \geq 2$, then $(k\xi)^2 \geq 1+k\xi$, so $\log(k\xi) \geq \frac{1}{2}\log(1+k\xi) \geq \frac{1}{2}\log(1+k\beta)$; with $\log\gamma \leq \frac{2\xi}{1-\xi} \leq 4\xi = 8\beta$, this gives

$q^\star \geq \frac{\log(k\xi)}{\log\gamma} \geq \frac{\log(1+k\beta)}{16\beta}$. If $k\xi < 2$, then $k\beta < 1$, so $\log(1+k\beta) < \log 2 < 1$. If $N_{\text{chain}} \geq \frac{1}{8\beta}$, this already exceeds $\frac{\log(1+k\beta)}{16\beta}$; otherwise $N_{\text{chain}} \geq k \geq \frac{k\beta}{16\beta} \geq \frac{\log(1+k\beta)}{16\beta}$. □

Remark D.8 (Agreement with Theorem C.2 for a single curve). *For $q = 1$, Corollary D.6 gives $\Gamma_2(k) = k$ and hence $\min_{\dim L=1} \Psi(L) = \frac{k-1}{k+1}$, which corresponds to a smallest blind spot of $\frac{k-1}{2(k+1)}$. This is exactly the one-curve optimum proved in Theorem C.2, now shown to be unimprovable even by a single vector that need not be monotone.*

D.1 Positive tolerance

Theorem D.5 and Corollary D.6 concern exact checks ($\delta = 0$). Sets with the same blind spot under exact checks can differ at $\delta > 0$ (Table 1). At a positive tolerance, we give a lower bound on the blind spot of every curve set, a construction with a closed-form Möbius (linear-fractional) upper bound, and the exact optimum for a single curve ($q = 1$). For $q \geq 2$, a matching lower bound remains open. Write $B^\star(k, q, \delta)$ for the smallest blind spot that at most q monotone curves achieve at tolerance δ:

$$B^\star(k, q, \delta) := \inf\big\{B_k(\mathcal{S}; \delta) : \mathcal{S} \subset \mathcal{V}_k,\ |\mathcal{S}| \leq q\big\}.$$

Theorem D.9 (Universal lower bound on the blind spot at positive tolerance). *Let $\mathcal{S}$ be any normalized curve set with matrix V whose columns lie in $\mathcal{V}_k$, and write $\theta = 2B_k(\mathcal{S}; 0)$. Then for every $\delta \in [0, \frac{1}{2}]$,*

$$2B_k(\mathcal{S}; \delta) \ \geq\ \theta + 2\delta(1-\theta).$$

Proof. Fix a depth r attaining θ, so that $\min_\alpha \|u^r - V\alpha\|_1 = \theta$. Let α be arbitrary and put $e = \|u^r - V\alpha\|_1 \geq \theta$. Every column of V and the vector u^r sum to 1, so $1 - \sum_i \alpha_i = \sum_j \big(u^r - V\alpha\big)_j \leq e$, whence $\|\alpha\|_1 \geq \sum_i \alpha_i \geq 1 - e$. Therefore

$$\|u^r - V\alpha\|_1 + 2\delta\|\alpha\|_1 \ \geq\ e + 2\delta(1-e) \ =\ 2\delta + e(1-2\delta) \ \geq\ 2\delta + \theta(1-2\delta),$$

using $\delta \leq \frac{1}{2}$ and $e \geq \theta$. Taking the infimum over α and then the maximum over r gives the claim. □

Thus a tolerance δ raises the blind spot by at least $\delta(1-\theta)$ above its zero-tolerance value $\theta/2$.

The construction for the upper bound keeps the blocks of the zero-tolerance optimum but adjusts the two weights of each endpoint curve to the tolerance; we call the result the *reweighted endpoint curves*.

Proposition D.10 (Reweighted endpoint curves at a positive tolerance). *Let $1 \leq q < k$, $0 \leq \delta < \frac{1}{2}$, and let $\theta^\star = \theta^\star(k, q)$ be the exact $\delta = 0$ width of Corollary D.6. Let $\{[c_i, d_i]\}_{i \leq p}$, $p \leq q$, be the greedy $\Phi(\theta^\star)$-cover attaining it, where $\Phi(t) = (1+t)/(1-t)$. Put*

$$\mu = \Phi^{-1}\big(\Phi(\theta^\star)\Phi(2\delta)\big) = \frac{\theta^\star + 2\delta}{1 + 2\delta\theta^\star}.$$

Define the reweighted endpoint curves $\mathcal{S}^{\mathrm{w}}_\delta = \{v^i : i \leq p\}$ by

$$v^i = \frac{c_i(1+\mu)\,u^{c_i} + d_i(1-\mu)\,u^{d_i}}{c_i(1+\mu) + d_i(1-\mu)}.$$

Then every $v^i \in \mathcal{V}_k$ and

$$2B_k(\mathcal{S}^{\mathrm{w}}_\delta; \delta) \leq \mu, \qquad 2B^\star(k, q, \delta) \leq \frac{\theta^\star + 2\delta}{1 + 2\delta\theta^\star}.$$

Consequently,

$$\theta^\star + 2\delta(1-\theta^\star) \ \leq\ 2B^\star(k, q, \delta) \ \leq\ \frac{\theta^\star + 2\delta}{1 + 2\delta\theta^\star}.$$

The sandwich display extends to $\delta = \frac{1}{2}$, where both sides equal 1.

Proof. Membership follows because each v^i is a convex combination of two members of $\mathcal{V}_k$. For the quantitative claim, the slab dual in the proof of Theorem 3 gives, for the matrix V with columns v^i,

$$2B_k(\mathcal{S}_\delta^{\mathrm{w}};\delta) = \max_{r\in[k]} \max_{\substack{y\in[-1,1]^k \\ \|V^\top y\|_\infty \le 2\delta}} |\langle u^r, y\rangle|.$$

Fix a dual-feasible y and write $F_s = \sum_{j\le s} y_j$, so that $|F_s - F_t| \le |s-t|$. Fix r and let $[c,d] = [c_i, d_i]$ be its cover block. Put

$$D = c(1+\mu) + d(1-\mu), \qquad T = \frac{c(1+\mu) - d(1-\mu)}{D}.$$

The definition of v^i gives

$$\langle v^i, y\rangle = \frac{(1+\mu)F_c + (1-\mu)F_d}{D},$$

so dual feasibility implies

$$|(1+\mu)F_c + (1-\mu)F_d| \le 2\delta D. \tag{17}$$

Moreover, $F_c \ge F_r - (r-c)$ and $F_d \ge F_r - (d-r)$. Substituting these two bounds into the upper half of (17) yields

$$2F_r - 2\mu r + c(1+\mu) - d(1-\mu) \le 2\delta D, \qquad \text{hence} \qquad F_r \le \mu r + \frac{D(2\delta - T)}{2}.$$

Because $[c,d]$ is a block of a $\Phi(\theta^\star)$-cover, $d/c \le \Phi(\theta^\star)$, so

$$\Phi(T) = \frac{c(1+\mu)}{d(1-\mu)} = \frac{\Phi(\mu)}{d/c} \ge \frac{\Phi(\mu)}{\Phi(\theta^\star)} = \Phi(2\delta).$$

The map Φ is increasing on $(-1,1)$, hence $T \ge 2\delta$ and $F_r \le \mu r$. Applying the same argument to $-y$ gives $F_r \ge -\mu r$. Thus $|\langle u^r, y\rangle| = |F_r| \le \mu$ for every depth r. The slab dual now gives $2B_k(\mathcal{S}_\delta^{\mathrm{w}};\delta) \le \mu$, and the definition of $B^\star$ gives the stated constructive upper bound.

For the lower bound, any set $\mathcal{S}$ of at most q curves has $\theta_{\mathcal{S}} := 2B_k(\mathcal{S};0) \ge \theta^\star$: if its span has dimension below q, extending that span to dimension q can only decrease its worst approximation error. Theorem D.9 and the monotonicity of $t \mapsto t + 2\delta(1-t)$ on $[0,1]$ give

$$2B_k(\mathcal{S};\delta) \ge \theta_{\mathcal{S}} + 2\delta(1-\theta_{\mathcal{S}}) \ge \theta^\star + 2\delta(1-\theta^\star).$$

Taking the infimum proves the sandwich. If $\delta = \frac{1}{2}$, the slab constraint of every normalized curve is vacuous on $[-1,1]^k$, so $2B^\star = 1$ directly. □

For $q \ge 2$, we do not know whether this upper bound is tight.

A note on scales. The sandwich is stated for $2B^\star$, and $\theta^\star = 2B^\star(k,q,0)$ is twice the zero-tolerance blind spot, while Table 1 reports blind spots B. For example, at $k=10$ and $q=5$ the greedy cover has blocks $[1,1],[2,2],[3,4],[5,7],[8,10]$ (those of Section 3.3), so $\theta^\star = 1/6$ and $B^\star(10,5,0) = 1/12 = 0.0833$. At $\delta = 10^{-2}$ the sandwich gives

$$\frac{11}{120} = 0.09167 \;\le\; B^\star(10,5,10^{-2}) \;\le\; \frac{4}{43} = 0.09302,$$

which pins the optimum to within 1.5%. The value 0.093 in Table 1 is this constructive upper bound, attained by the reweighted endpoint curves.

At $q=1$, the Möbius formula is exact.

Theorem D.11 (Exact optimum for a single curve at positive tolerance)**.** *Let $k \ge 1$, let $0 \le \delta \le \frac{1}{2}$, and let $\Phi(t) = \frac{1+t}{1-t}$, $\theta^\star = \theta^\star(k,1) = \frac{k-1}{k+1}$ (so $\Phi(\theta^\star) = k$). Put*

$$\mu = \mu(k,\delta) := \frac{k(1+2\delta) - (1-2\delta)}{k(1+2\delta) + (1-2\delta)},$$

$$t^\star = t^\star(k,\delta) := \frac{k(1+2\delta) + (1-2\delta)}{2k} = \frac{1+2\delta}{2} + \frac{1-2\delta}{2k},$$

and

$$v_\delta^\star := \frac{1+2\delta}{2}\,u^1 + \frac{1-2\delta}{2}\,u^k = \Big(t^\star,\ \tfrac{1-2\delta}{2k},\ \ldots,\ \tfrac{1-2\delta}{2k}\Big).$$

Then $v_\delta^\star \in \mathcal{V}_k$, *and*

$$\begin{aligned} &\min_{v\in\mathcal{V}_k} 2B_k(\{v\};\delta) = \mu = \Phi^{-1}\big(\Phi(\theta^\star)\Phi(2\delta)\big), \\ &\textit{i.e.} \qquad \operatorname{artanh}\mu = \operatorname{artanh}\theta^\star + \operatorname{artanh}(2\delta), \end{aligned} \tag{18}$$

the minimum being attained at $v_\delta^\star$ *(read* $\Phi(1) = +\infty$ *and* $\mu = 1$ *at* $\delta = \frac{1}{2}$*).*

Proof. For $v \in \mathcal{V}_k$, write $\bar v_r = \sum_{j\le r} v_j$, so that $\bar v_k = 1$ and $v_1 \ge \frac{1}{k} > 0$. Because $\Phi(\theta^\star) = k$ and $\Phi(2\delta) = \frac{1+2\delta}{1-2\delta}$ for $\delta < \frac{1}{2}$, we have $\Phi(\mu) = \frac{k(1+2\delta)}{1-2\delta} = \Phi(\theta^\star)\Phi(2\delta)$, which is the second equality in (18), and $\log\Phi = 2\operatorname{artanh}$ gives the third. Also $\mu \le 1$, with equality if and only if $\delta = \frac{1}{2}$.

Step 0: $v_\delta^\star \in \mathcal{V}_k$*, and its partial sums.* Its entries are $v_1^\star = t^\star = \frac{1+2\delta}{2} + \frac{1-2\delta}{2k}$ and $v_j^\star = \frac{1-2\delta}{2k}$ for $2 \le j \le k$. The tail is nonnegative exactly when $\delta \le \frac{1}{2}$: this is where the hypothesis is used. Monotonicity holds because $v_1^\star - v_2^\star = \frac{1+2\delta}{2} \ge 0$ and the tail is constant, and $\sum_j v_j^\star = \frac{1+2\delta}{2} + \frac{1-2\delta}{2} = 1$. Hence $v_\delta^\star \in \mathcal{V}_k$, with

$$\bar v_r^\star = \frac{1+2\delta}{2} + \frac{r(1-2\delta)}{2k}, \qquad \text{so} \qquad 2\bar v_r^\star - 1 - 2\delta = \frac{r(1-2\delta)}{k} \quad (1 \le r \le k). \tag{19}$$

Step 1: the single-curve dual. Fix $v \in \mathcal{V}_k$. By (5) with $q = 1$, $2B_k(\{v\};\delta) = \max_{r\in[k]} \inf_{a\in\mathbb{R}}\{\|u^r - av\|_1 + 2\delta|a|\}$. Using $\|w\|_1 = \max\{\langle y, w\rangle : \|y\|_\infty \le 1\}$, the infimum at depth r equals

$$\inf_{a\in\mathbb{R}} \max_{\|y\|_\infty\le 1} \big\{\langle y, u^r\rangle - a\langle y, v\rangle + 2\delta|a|\big\}.$$

The bracket is convex in a and affine in y, and $\{\|y\|_\infty \le 1\}$ is convex and compact, so Sion's minimax theorem permits exchanging $\inf_a$ and $\max_y$. Since $\inf_a\{-a\langle y, v\rangle + 2\delta|a|\}$ equals 0 if $|\langle y, v\rangle| \le 2\delta$ and $-\infty$ otherwise,

$$\begin{aligned} 2B_k(\{v\};\delta) &= \max\big\{|\langle u^r, y\rangle| :\ r \in [k],\ y \in Y(v)\big\}, \\ Y(v) &:= \big\{y \in [-1,1]^k :\ |\langle y, v\rangle| \le 2\delta\big\}; \end{aligned} \tag{20}$$

the absolute value is free because $Y(v)$ is symmetric ($y \in Y(v) \iff -y \in Y(v)$), and the maximum is attained because $Y(v)$ is nonempty and compact. (This is the $q = 1$ case of the slab dual used throughout this appendix.) Taking $a = 0$ bounds the infimum at every depth by $\|u^r\|_1 = 1$, so

$$2B_k(\{v\};\delta) \le 1 \qquad \text{for every } v \in \mathcal{V}_k,\ \delta \ge 0. \tag{21}$$

Step 2: a universal upper bound, valid for every curve. Claim: for all $v \in \mathcal{V}_k$, $y \in Y(v)$ *and* $r \in [k]$,

$$|\langle u^r, y\rangle| \ \le\ 1 - \frac{2\bar v_r - 1 - 2\delta}{r\,v_1}. \tag{22}$$

Indeed, from $\langle y, v\rangle \le 2\delta$, $y_j \ge -1$ and $v_j \ge 0$,

$$\sum_{j\le r} v_j y_j = \langle y, v\rangle - \sum_{j>r} v_j y_j \ \le\ 2\delta + \sum_{j>r} v_j = 2\delta + 1 - \bar v_r,$$

while $1 - y_j \ge 0$ and $0 \le v_j \le v_1$ for $j \le r$ (here the monotonicity of v is used) give

$$\sum_{j\le r} v_j y_j = \bar v_r - \sum_{j\le r} v_j(1 - y_j) \ \ge\ \bar v_r - v_1 \sum_{j\le r}(1 - y_j) = \bar v_r - v_1 r\big(1 - \langle u^r, y\rangle\big).$$

Combining the two displays yields $v_1 r(1 - \langle u^r, y\rangle) \ge 2\bar v_r - 1 - 2\delta$, which is (22) for $\langle u^r, y\rangle$. Applying it to $-y \in Y(v)$ gives it for $-\langle u^r, y\rangle$, hence for $|\langle u^r, y\rangle|$.

Step 3: the upper bound $2B_k(\{v_\delta^\star\};\delta) \le \mu$. Insert (19) into (22) at $v = v_\delta^\star$, where $v_1 = t^\star$:

$$|\langle u^r, y\rangle| \le 1 - \frac{r(1-2\delta)/k}{r\,t^\star} = 1 - \frac{1-2\delta}{k\,t^\star} = 1 - \frac{2(1-2\delta)}{k(1+2\delta) + (1-2\delta)} = \mu,$$
$$r \in [k], \quad y \in Y(v_\delta^\star).$$

The bound is independent of r: $v^\star_\delta$ is precisely the curve whose depth-r obstruction $\frac{2\bar{v}_r-1-2\delta}{r}$ is the same at every depth. By (20), $2B_k(\{v^\star_\delta\};\delta) \le \mu$.

Step 4: the lower bound $2B_k(\{v\};\delta) \ge \mu$ for every $v \in \mathcal{V}_k$. Write $t = v_1 \in [\frac{1}{k}, 1]$. All witnesses below have one of the two shapes $z(c) = (1, c, \dots, c)$ and $y(x) = (x, 1, \dots, 1)$, for which $\sum_j v_j = 1$ gives

$$\langle z(c), v\rangle = t + c(1-t), \quad \langle u^1, z(c)\rangle = 1,$$
$$\langle y(x), v\rangle = xt + (1-t), \quad \langle u^1, y(x)\rangle = x, \quad \langle u^k, y(x)\rangle = \tfrac{x+k-1}{k}.$$

Case A: $t \le \frac{1+2\delta}{2}$. If $t < 1$, take $c = \frac{2\delta - t}{1-t}$. Then $\langle z(c), v\rangle = 2\delta$. Moreover, $c \le 1 \iff 2\delta \le 1$, which holds, while $c \ge -1 \iff 2\delta - t \ge t - 1 \iff t \le \frac{1+2\delta}{2}$, which is the case assumption. So $z(c) \in Y(v)$, and (20) gives $2B_k(\{v\};\delta) \ge \langle u^1, z(c)\rangle = 1 \ge \mu$. If $t = 1$, then $t \le \frac{1+2\delta}{2}$ forces $\delta = \frac{1}{2}$, where $\mu = 1$, and $|\langle y, v\rangle| \le \|y\|_\infty \|v\|_1 \le 1 = 2\delta$ makes the constraint in $Y(v)$ vacuous, so $2B_k(\{v\};\delta) = 1 = \mu$ by (20)–(21).

Case B: $t > \frac{1+2\delta}{2}$. Put

$$x^- = -\frac{1+2\delta-t}{t}, \qquad x^+ = \frac{t-1+2\delta}{t},$$

so that $\langle y(x^-), v\rangle = -(1+2\delta-t) + 1 - t = -2\delta$ and $\langle y(x^+), v\rangle = (t-1+2\delta) + 1 - t = 2\delta$: both witnesses meet the slab exactly. They are feasible: $x^- \le 0 \le 1$ since $t \le 1 \le 1 + 2\delta$, and $x^- \ge -1 \iff t \ge \frac{1+2\delta}{2}$, while $x^+ \le 1 \iff \delta \le \frac{1}{2}$, and $x^+ \ge -1 \iff 2t \ge 1 - 2\delta$, which is implied by $t > \frac{1+2\delta}{2} \ge \frac{1-2\delta}{2}$. Hence $y(x^\pm) \in Y(v)$, and (20) gives

$$2B_k(\{v\};\delta) \;\ge\; \max\Big\{ \underbrace{\tfrac{1+2\delta-t}{t}}_{=|\langle u^1, y(x^-)\rangle|} \;,\; \underbrace{1 - \tfrac{1-2\delta}{kt}}_{=\langle u^k, y(x^+)\rangle} \Big\} =: \max\{f(t), g(t)\}. \tag{23}$$

On $(0, 1]$, f is strictly decreasing and g is nondecreasing (strictly increasing when $\delta < \frac{1}{2}$), and

$$f(t^\star) = \frac{1+2\delta}{t^\star} - 1 = \frac{2k(1+2\delta)}{k(1+2\delta) + (1-2\delta)} - 1 = \mu, \qquad g(t^\star) = 1 - \frac{2(1-2\delta)}{k(1+2\delta) + (1-2\delta)} = \mu,$$

so the two obstructions cross exactly at $t^\star$, with common value μ. Therefore $t \le t^\star$ gives $2B_k(\{v\};\delta) \ge f(t) \ge f(t^\star) = \mu$, and $t \ge t^\star$ gives $2B_k(\{v\};\delta) \ge g(t) \ge g(t^\star) = \mu$. Since $t^\star \ge \frac{1+2\delta}{2}$, Cases A and B exhaust $[\frac{1}{k}, 1]$, and in both $2B_k(\{v\};\delta) \ge \mu$.

Steps 3 and 4 prove (18), with the minimum attained at $v^\star_\delta$.

At $\delta = 0$, the theorem recovers Theorem C.2: then $\mu = \theta^\star = \frac{k-1}{k+1}$ and $v^\star_0 = \frac{1}{2}(u^1 + u^k)$. □

E Uncertainty about Both Attention and Targets

The prefix certificate of Theorem 2 measures disparities against a target p_g, but the target is usually estimated itself. Because the attention curve and the target act on the same scalar, the deviation $\sum_j v_j x_{g,j} - p_g$, the two uncertainties combine in closed form (Theorem E.1). In the simplest case, where each target is known up to $\pm\tau(z)$ and these intervals lie in $[0, 1]$, the worst case over both is the querywise risk at the estimated target plus $\mathbb{E}_z \tau(z)$, so target uncertainty adds a floor to the risk that no policy can remove.

Fix a policy with occupancies $x_{g,j}(z)$, as in Appendix A. Write $\hat{a}_{g,r}(z) = \frac{1}{r}\sum_{j \le r} x_{g,j}(z) - \hat{p}_g(z)$ for the prefix deviations computed at the estimated target $\hat{p}(z)$, and $\mathcal{R}_{\mathrm{Q}}(p)$ for the querywise risk of the policy at target p, so that $\mathcal{R}_{\mathrm{Q}}(\hat{p}) = \mathbb{E}_z \max_{g,r} |\hat{a}_{g,r}(z)|$ (Theorem 2).

Theorem E.1 (Attention and target uncertainty)**.** *Fix, for each query z, an estimated target $\hat{p}(z) \in [0, 1]^m$ and an uncertainty set $T(z)$ for the true target, and define the* joint risk

$$\mathcal{R}_{\mathrm{Q}}^{\mathcal{V},T} = \mathbb{E}_z \sup_{v \in \mathcal{V}_k} \sup_{p \in T(z)} \max_g \Big| \sum_j v_j x_{g,j}(z) - p_g \Big|.$$

For every compact $T(z)$, with $c_{g,r}(z) = \frac{1}{r}\sum_{j\le r} x_{g,j}(z)$, $\underline{p}_g(z) = \min\{p_g : p \in T(z)\}$ and $\overline{p}_g(z) = \max\{p_g : p \in T(z)\}$,

$$\mathcal{R}_{\mathrm{Q}}^{\mathcal{V},T} = \mathbb{E}_z \max_{g,r} \max\big\{c_{g,r}(z) - \underline{p}_g(z),\ \overline{p}_g(z) - c_{g,r}(z)\big\}.$$

Case (i) makes the projections $\underline{p}_g$ and $\overline{p}_g$ explicit for a box of targets, and case (ii) bounds how much an error in the target can change the risk.

(i) (Box.) *If $T(z) = \{p \in [0,1]^m : \|p - \hat{p}(z)\|_\infty \le \tau(z)\}$ (each true target lies within $\tau(z)$ of its estimate), then*

$$\mathcal{R}_{\mathrm{Q}}^{\mathcal{V},T} = \mathbb{E}_z \max_{g,r} \max\big\{\hat{a}_{g,r}(z) + \min\big(\tau(z), \hat{p}_g(z)\big),\\ - \hat{a}_{g,r}(z) + \min\big(\tau(z), 1 - \hat{p}_g(z)\big)\big\}.$$

In particular, $\mathcal{R}_{\mathrm{Q}}^{\mathcal{V},T} = \mathcal{R}_{\mathrm{Q}}(\hat{p}) + \mathbb{E}_z\tau(z)$ whenever $[\hat{p}_g(z) - \tau(z), \hat{p}_g(z) + \tau(z)] \subseteq [0,1]$ for every g.

(ii) (Two-sided sensitivity bound.) *For any two query-dependent targets p and $\hat{p}$, $|\mathcal{R}_{\mathrm{Q}}(p) - \mathcal{R}_{\mathrm{Q}}(\hat{p})| \le \mathbb{E}_z\|p(z) - \hat{p}(z)\|_\infty$ for every policy, with equality attainable.*

Proof. Fix z and omit it from the notation. By Lemma 1 and linearity in v, as in the proof of Theorem 2, $\sup_{v\in\mathcal{V}_k} \max_g |\sum_j v_j x_{g,j} - p_g| = \max_{g,r} |c_{g,r} - p_g|$ for *every* fixed p. Suprema and maxima over independent indices commute, so

$$\sup_{p\in T} \max_{g,r} |c_{g,r} - p_g| = \max_{g,r} \sup_{p\in T} |c_{g,r} - p_g| = \max_{g,r} \max\big(c_{g,r} - \underline{p}_g,\ \overline{p}_g - c_{g,r}\big),$$

where $\underline{p}_g$ and $\overline{p}_g$ are the coordinate projections of T from the statement. They exist because T is compact, and the inner supremum is attained at one of them because $t \mapsto |c_{g,r} - t|$ is convex.

For the box of (i), $\underline{p}_g = \max(0, \hat{p}_g - \tau)$ and $\overline{p}_g = \min(1, \hat{p}_g + \tau)$, so that $\max(c_{g,r} - \underline{p}_g, \overline{p}_g - c_{g,r}) = \max\{(c_{g,r} - \hat{p}_g) + \min(\tau, \hat{p}_g),\ -(c_{g,r} - \hat{p}_g) + \min(\tau, 1 - \hat{p}_g)\}$. Taking $\mathbb{E}_z$ gives the displayed formula.

For (ii), $|\,|a - p_g| - |a - \hat{p}_g|\,| \le |p_g - \hat{p}_g|$ pointwise, for every real a, and a maximum of functions that differ by at most ε differs by at most ε. Equality is attainable, for example, on a degenerate one-query distribution in which one pair (g, r) attains the maximum for both targets, its target difference $|p_g - \hat{p}_g|$ attains the ℓ_∞ norm, and $c_{g,r}$ lies outside $[\min(\hat{p}_g, p_g), \max(\hat{p}_g, p_g)]$. □

F Why the Single-Curve LP Has Many Optimal Solutions

Appendix J.10 measures on ESCI a wide spread of querywise risk over the optimal solutions of the single-curve LP. This section explains the spread, which arises even without tied relevance scores.

The set of optimal solutions of a linear program is a face of its feasible polytope, and here it can be the entire feasible set. Suppose that the constrained curve is the normalized utility weight vector, as in the single-curve LP of our ESCI and German Credit experiments (DCG curve and DCG utility weights; Section 4.2), and that the scores are constant within the constrained groups. Then the fairness constraints *fix the value of the objective*, so the set of optimal solutions is identically the entire feasible set of the single-curve LP (Proposition F.1). Tied scores are one way to make scores constant within groups, and the conclusion also holds with continuous scores.

Setup and notation. Fix one query with n items and $k \le n$ display positions, and let $\mathcal{P}_{n,k}$ be the set of policies (1). Items carry relevance scores $s \in \mathbb{R}^n$, and positions carry utility weights $w \in \mathbb{R}^k_{\ge 0}$ with total weight $W := \sum_{j=1}^k w_j > 0$. Group membership is recorded by a matrix $\Phi \in \{0,1\}^{n\times m}$, with $\Phi_{ig} = 1$ iff item i belongs to group g. The groups are *not* assumed disjoint, and they are *not* assumed to cover the items. The occupancy vector of group g is $x_g(P) \in \mathbb{R}^k$, with $x_g(P)_j = \sum_i \Phi_{ig} P_{ij}$.

The single-curve LP of Singh & Joachims (2018) constrains the exposures under one normalized attention curve $v \in \mathbb{R}^k_{\geq 0}$, $\sum_j v_j = 1$, to lie within δ of the targets $p \in \mathbb{R}^m$, and it maximizes the score-based utility:

$$\max_{P \in \mathcal{F}_\delta} U(P), \qquad U(P) := \sum_{i=1}^{n} \sum_{j=1}^{k} s_i \, w_j \, P_{ij}, \tag{24}$$
$$\mathcal{F}_\delta := \Big\{ P \in \mathcal{P}_{n,k} \; : \; \big|\langle v, x_g(P) \rangle - p_g\big| \leq \delta \;\text{ for } g = 1, \ldots, m \Big\}.$$

Here $\delta = 0$ gives the equality form of the fairness constraint, and $\delta > 0$ gives the tolerance form. Implementations use the tolerance form when the target cannot be met exactly (the min-violation extension of Table 4). Write

$$\mathcal{L} := \operatorname{span}\big\{ \mathbf{1}_n, \; \Phi_{\cdot 1}, \ldots, \Phi_{\cdot m} \big\} \subseteq \mathbb{R}^n$$

for the space of score vectors that are affine in the group indicators. For a vector $r \in \mathbb{R}^n$ (in this section, r is a vector, not a depth), let $r_{(1)} \geq \cdots \geq r_{(n)}$ and $r_{[1]} \leq \cdots \leq r_{[n]}$ denote r sorted in decreasing and in increasing order.

Proposition F.1 (When all fair policies have the same utility)**.** *Assume that the constrained curve is the normalized utility weight vector, $v = w/W$. Then:*

(i) *(Exact invariance.) If $s \in \mathcal{L}$, say $s = a_0 \mathbf{1}_n + \Phi a$ with $a_0 \in \mathbb{R}$ and $a \in \mathbb{R}^m$, then*

$$U(P) \;=\; W\Big(a_0 + \sum_{g=1}^{m} a_g \, p_g\Big) \qquad \textit{for every } P \in \mathcal{F}_0,$$

a constant independent of P. Consequently, U is constant on $\mathcal{F}_0$: every *policy that is feasible for the single-curve LP at $\delta = 0$ is optimal, and the set of optimal solutions of (24) at $\delta = 0$ is all of $\mathcal{F}_0$. This holds for every score vector in $\mathcal{L}$ and every target p.*

(ii) *(Tolerance form gives a band, not a point.) Under the same hypothesis, on $\mathcal{F}_\delta$,*

$$\max_{P \in \mathcal{F}_\delta} U(P) - \min_{P \in \mathcal{F}_\delta} U(P) \;\leq\; 2\, W\, \delta\, \|a\|_1.$$

(iii) *(Quantitative version for scores that vary within groups.) Let s be arbitrary, and take w nonincreasing (as every position discount is). Let $\sigma := \min_{a_0, a} \big\| s - a_0 \mathbf{1}_n - \Phi a \big\|_\infty$ be the Chebyshev (ℓ_∞) distance from s to $\mathcal{L}$, attained at $(a_0^\star, a^\star)$ with residual $r := s - a_0^\star \mathbf{1}_n - \Phi a^\star$, $\|r\|_\infty = \sigma$. Then*

$$\max_{\mathcal{F}_\delta} U - \min_{\mathcal{F}_\delta} U \;\leq\; \underbrace{\sum_{j=1}^{k} w_j \big(r_{(j)} - r_{[j]} \big)}_{\textit{score variation}} \;+\; \underbrace{2\, W\, \delta\, \|a^\star\|_1}_{\textit{tolerance slack}} \;\leq\; 2\, W \big(\sigma + \delta\, \|a^\star\|_1 \big).$$

In particular, at the equality form $\delta = 0$, the whole range of the utility is at most $2W\sigma$. If the groups partition the items, then $\sigma \leq \sigma_{\text{within}} := \max_g \frac{1}{2} \big(\max_{i \in G_g} s_i - \min_{i \in G_g} s_i \big)$, the largest half-range of the scores within a group. Hence if scores vary within groups by at most σ_{within}, the utility varies over the entire feasible set $\mathcal{F}_0$ of the single-curve LP by at most $2W\sigma_{\text{within}}$.

Proof. For $u \in \mathbb{R}^n$, define the linear functional $L_u(P) := \sum_{j=1}^{k} w_j \langle u, P_{\cdot j} \rangle$, where $P_{\cdot j}$ is the j-th column of P. The map $u \mapsto L_u$ is linear, and $U(P) = L_s(P)$. The constraints of (24) fix the values of two families of these functionals.

Position constraints. Every $P \in \mathcal{P}_{n,k}$ satisfies $\sum_i P_{ij} = 1$, hence $L_{\mathbf{1}_n}(P) = \sum_j w_j \cdot 1 = W$ identically.

Fairness constraints. For each group g,

$$L_{\Phi_{\cdot g}}(P) = \sum_j w_j \sum_i \Phi_{ig} P_{ij} = \sum_j w_j \, x_g(P)_j = \langle w, x_g(P) \rangle = W \, \langle v, x_g(P) \rangle,$$

where the last equality holds because $w = Wv$. Therefore $L_{\Phi_{\cdot g}}(P) = Wp_g$ for every $P \in \mathcal{F}_0$, and $L_{\Phi_{\cdot g}}(P) \in W[p_g - \delta, p_g + \delta]$ for every $P \in \mathcal{F}_\delta$.

(i) If $s = a_0\mathbf{1}_n + \Phi a$, then, by linearity of $u \mapsto L_u$,

$$U(P) = L_s(P) = a_0 L_{\mathbf{1}_n}(P) + \sum_{g=1}^{m} a_g L_{\Phi_{\cdot g}}(P) = a_0 W + W \sum_{g=1}^{m} a_g p_g \qquad (P \in \mathcal{F}_0),$$

which is the stated constant.

(ii) For $P, P' \in \mathcal{F}_\delta$, the same expansion gives $U(P) - U(P') = \sum_g a_g \big(L_{\Phi_{\cdot g}}(P) - L_{\Phi_{\cdot g}}(P')\big) = W \sum_g a_g \big(\langle v, x_g(P)\rangle - \langle v, x_g(P')\rangle\big)$. Each bracket lies in $[-2\delta, 2\delta]$, so $|U(P) - U(P')| \le 2W\delta\|a\|_1$.

(iii) Decompose $s = a_0^\star \mathbf{1}_n + \Phi a^\star + r$ and expand once more:

$$U(P) = a_0^\star W + \sum_g a_g^\star L_{\Phi_{\cdot g}}(P) + L_r(P).$$

The first term is a constant. By the computation in (ii), the second term varies by at most $2W\delta\|a^\star\|_1$ over $\mathcal{F}_\delta$. For the third term, note that $\mathcal{F}_\delta \subseteq \mathcal{P}_{n,k}$ and that L_r is linear. Hence the extremes of L_r over $\mathcal{P}_{n,k}$ are attained at vertices, which are exactly the partial permutation matrices (Birkhoff–von Neumann for doubly substochastic matrices with the stated row and column constraints; see also Proposition A.1): such a P selects an ordered k-tuple of distinct items. Maximizing $\sum_j w_j r_{\pi(j)}$ over injections $\pi : [k] \to [n]$ pairs the largest weights with the largest residuals (rearrangement inequality), which gives $\max_{\mathcal{P}_{n,k}} L_r = \sum_{j=1}^{k} w_j r_{(j)}$; likewise, $\min_{\mathcal{P}_{n,k}} L_r = \sum_{j=1}^{k} w_j r_{[j]}$. Subtracting the minimum from the maximum, and adding the tolerance-slack term of the second term, gives the first inequality. The second inequality follows because $r_{(j)} - r_{[j]} \le 2\|r\|_\infty = 2\sigma$ and $\sum_j w_j = W$. Finally, if the groups partition the items, the choice $a_0 = 0$ and $a_g = \frac{1}{2}(\max_{i \in G_g} s_i + \min_{i \in G_g} s_i)$ is feasible for the minimization that defines σ and attains $\|s - \Phi a\|_\infty = \sigma_{\text{within}}$, so $\sigma \le \sigma_{\text{within}}$. □

G Blind Spots of Published Curves

This section gives the details behind the blind spots of published curves in Section 3.2 and Table 1, including explicit occupancy vectors that pass all checks.

Two restricted classes of attention curves (subsets of $\mathcal{V}_k$, which we call envelopes) recur in this appendix and in the experiments. The *reference envelope* $\mathcal{E} = \{v \in \mathcal{V}_k : v_1 \le 0.45,\ \sum_{j>3} v_j \ge 0.15\}$ gives position one at most 45% of the attention and positions four and beyond at least 15%. In the *bounded-decay class* $\mathcal{C}_{\text{dec}} = \{v \in \mathcal{V}_k : 0.5v_j \le v_{j+1} \le 0.97v_j\}$, attention drops by between 3% and 50% from each position to the next. Both classes are polyhedral, so the envelope certificate of Theorem B.2 applies to them: the worst case over the class is attained at one of its finitely many extreme curves. Moreover, $\mathcal{E}$ contains every published curve listed below except $\text{RBP}_{0.5}$.

Table 6 recomputes the exact blind spot (5) for the position discounts used in the literature on fair ranking and on the evaluation of information retrieval. The curves are the following. The DCG curve, $v_j \propto 1/\log_2(j+1)$, is used by Singh & Joachims (2018). The TREC curve of the TREC Fair Ranking track, $v_j \propto 1/\log_2 \max(j, 2)$ (Ekstrand et al., 2023), is *not* the DCG curve and satisfies $v_1 = v_2$. The RBP curves have continuation probabilities $0.5, 0.8, 0.95$ ($\text{RBP}_{0.5}$, $\text{RBP}_{0.8}$ and $\text{RBP}_{0.95}$) (Moffat & Zobel, 2008), and the exposure model of Diaz et al. (2020) with patience parameter 0.5 coincides with $\text{RBP}_{0.5}$ after normalization. The power discounts are $v_j \propto j^{-\eta}$ at $\eta = 0.5, 1$ ($\text{power}_{0.5}$ and power_1); the latter coincides with $1/j$, with reciprocal rank, and with the $\eta = 1$ propensity form of Joachims et al. (2017). After normalization, these ten named curves span only seven distinct directions, the seven published curves of Section 3.2.

Single published curves. Every published curve except $\text{RBP}_{0.5}$ has $v_1 \le \frac{1}{2}$, so its equal-target blind spot is $\frac{1}{2}$, the largest possible value (Proposition C.7). The curve $\text{RBP}_{0.5}$ has $v_1 = \frac{1}{2}(1 - 2^{-k})^{-1} > \frac{1}{2}$ and exact blind spot $\frac{1}{2}(1 - 2^{-(k-1)})$.

Table 6: **What audits built from real, published position-attention curves certify.** For a normalized set of checked curves $\mathcal{S}$ with matrix V, $B_k(\mathcal{S};\delta) = \frac{1}{2}\max_r \min_\alpha\{\|u^r - V\alpha\|_1 + 2\delta\|\alpha\|_1\}$ is the largest group-exposure disparity, at the worst depth $r^\star$, consistent with every checked curve reporting deviation at most δ. All entries are computed by exact linear programs, and $\max_r \mathrm{dist}_1$ is $\max_r \mathrm{dist}_1(u^r, \mathrm{span}\,\mathcal{S}) = 2B_k(\mathcal{S};0)$. $r^\star$ is a depth attaining the maximum at $\delta = 0$ (ties are broken by the solver). The last two columns restrict the hidden curve to the reference envelope $\mathcal{E}$ (42 vertices at $k = 10$, 102 at $k = 20$) and to the bounded-decay class $\mathcal{C}_{\mathrm{dec}}$ (512 vertices at $k = 10$; its 2^{19} vertices at $k = 20$ were not enumerated), evaluated at $\delta = 0$.

Checked curves	q	$\mathrm{rank}\,V$	$\max_r \mathrm{dist}_1$	$B(0)$	$B(10^{-2})$	$r^\star$	$B^{\mathcal{E}}(0)$	$B^{\mathcal{C}_{\mathrm{dec}}}(0)$
$k = 10$								
DCG	1	1	1.0000	0.5000	0.5000	1	0.4275	0.3963
TREC	1	1	1.0000	0.5000	0.5000	1	0.4077	0.3751
RBP 0.5	1	1	0.9980	0.4990	0.5000	1	0.4001	0.3859
RBP 0.8	1	1	1.0000	0.5000	0.5000	1	0.3847	0.3351
RBP 0.95	1	1	1.0000	0.5000	0.5000	1	0.5000	0.4500
power 0.5	1	1	1.0000	0.5000	0.5000	1	0.4315	0.4041
power 1	1	1	1.0000	0.5000	0.5000	1	0.3453	0.2828
$\mathcal{S}_{\mathrm{prac}}$ (5 published)	5	5	0.2505	0.1252	0.2920	3	0.1248	0.0617
all 7 distinct published	7	7	0.1759	0.0880	0.2320	4	0.0880	0.0388
best 5-curve (this paper)	5	5	0.1667	0.0833	0.0930	5	0.0833	0.0619
all k uniform prefixes	10	10	0.0000	0.0000	0.0100	1	0.0000	0.0000
$k = 20$								
DCG	1	1	1.0000	0.5000	0.5000	1	0.5000	—
TREC	1	1	1.0000	0.5000	0.5000	1	0.5000	—
RBP 0.5	1	1	1.0000	0.5000	0.5000	1	0.4500	—
RBP 0.8	1	1	1.0000	0.5000	0.5000	1	0.4413	—
RBP 0.95	1	1	1.0000	0.5000	0.5000	1	0.5000	—
power 0.5	1	1	1.0000	0.5000	0.5000	1	0.5000	—
power 1	1	1	1.0000	0.5000	0.5000	1	0.4289	—
$\mathcal{S}_{\mathrm{prac}}$ (5 published)	5	5	0.4184	0.2092	0.3335	2	0.1979	—
all 7 distinct published	7	7	0.2444	0.1222	0.2657	8	0.1222	—
best 5-curve (this paper)	5	5	0.2500	0.1250	0.1343	13	0.1250	—
all k uniform prefixes	20	20	0.0000	0.0000	0.0100	1	0.0000	—

Restricting the unchecked curve to $\mathcal{E}$. Let $\mathcal{E}$ be the reference envelope defined at the start of this section, and write $B_k^{\mathcal{E}}$ for the blind spot when the unchecked curve is restricted to $\mathcal{E}$ (and B_k for the blind spot over all of $\mathcal{V}_k$). In its mixture representation (Appendix B), $\mathcal{E}$ excludes u^1, u^2, u^3 outright, because a curve supported on the first three positions has zero tail mass. Nevertheless, $\mathcal{S}_{\mathrm{prac}}$ keeps $B_{10}^{\mathcal{E}} = 0.12481$ against $B_{10} = 0.12525$ (99.6%), and the worst case is attained at the extreme point u^4: uniform attention over the top four positions, an ordinary "above the fold" model with $v_1 = 0.25 \le 0.45$ and tail mass $0.25 \ge 0.15$. At $k = 20$, the worst case is attained at u^5, and $B_{20}^{\mathcal{E}} = 0.1979$ against $B_{20} = 0.2092$. The explicit witness at $k = 10$ is the group-A occupancy vector $x = (0.5119, 0.3836, 0.6037, 1, 0, 0, 0.5607, 1, 0.9399, 0)$. Its exposure under u^4 is $0.62481 = \frac{1}{2} + B_{10}^{\mathcal{E}}$, and it passes each of the five checks up to rounding error.

Bounded decay. Take $k = 10$. The upper ratio 0.97 of $\mathcal{C}_{\mathrm{dec}}$ admits the DCG curve, whose largest ratio between adjacent positions is 0.9603, and the interval of ratios also contains every geometric curve RBP_ϕ with continuation probability $\phi \in [0.5, 0.95]$. Positivity and normalization leave nine degrees of freedom. At every vertex, one of the two adjacent-ratio bounds is active at each depth, so direct enumeration gives exactly $2^9 = 512$ vertices. For $\mathcal{S}_{\mathrm{prac}}$, the exact blind spot over $\mathcal{C}_{\mathrm{dec}}$ is 0.06165 at $\delta = 0$ and 0.14216 at $\delta = 0.01$, against 0.12525 and 0.2920 over all monotone curves. The witness over all monotone curves at $\delta = 0$ is the group-A occupancy vector $x = (0.508124, 0.367626, 1, 0, 0.132512, 1, 1, 0.160755, 0, 0.830934)$, a vertex of the checked slice of the cube (the occupancy vectors in $[0,1]^{10}$ that pass all checks). It passes the five checks up to rounding error and gives group A the share $0.62525 = \frac{1}{2} + 0.12525$ (62.5%) of the first three positions; this is the value rounded to 0.125 in Table 1 and Section 3.2. With the narrower upper ratio 0.95, the blind spots are similar: 0.05863 at $\delta = 0$ and 0.13985 at $\delta = 0.01$. Checking the DCG curve alone remains weak even under bounded decay: 0.39632 at $\delta = 0$ and 0.40292 at $\delta = 0.01$.

Hence restrictions motivated by user behavior can substantially shrink the blind spot, but checking the five curves of $\mathcal{S}_{\text{prac}}$ still does not certify fairness over the bounded-decay class. The corresponding witness policy passes all five checks exactly, attains 0.06165 at a vertex of $\mathcal{C}_{\text{dec}}$, and has querywise risk $\mathcal{R}_{\text{Q}} = 0.1078$ over all monotone curves.

H THE ROBUST LP AT ZERO TOLERANCE

This section proves the closed form of the robust LP at $\delta = 0$ stated in Section 4.1 (Theorem H.1) and gives the smallest attainable risk when no policy has zero querywise risk (Theorem H.2).

For Theorem H.1, suppose that the groups form a partition, the target shares sum to one, group g has n_g items, and item i has the learned relevance score s_i (from the fixed relevance model).

Theorem H.1 (At zero tolerance, sorting within each group is optimal)**.** *A policy with zero querywise risk on the query (fair under every monotone curve) exists if and only if $kp_g \le n_g$ for every g. Under this condition, the greedy rule sorts the items within each group by score and assigns mass p_g from group g to every position, using up the items in score order. Its output is feasible, has $\mathcal{R}_{\text{Q}} = 0$, and, among all policies with zero querywise risk, simultaneously maximizes the utility $U(P)$ for every nonnegative nonincreasing utility weight vector w.*

Proof. Existence. Necessity holds because zero querywise risk implies $x_{g,j} = p_g$ at every position (Appendix A), so group g supplies total mass $kp_g \le n_g$, since each item carries total mass at most one. Conversely, under this condition, pouring mass p_g into every column never exhausts the group, and it yields a feasible assignment.

Optimality. Fix a group g and a depth r. Every policy with zero querywise risk places total mass rp_g from group g in the first r columns. The largest total score that group g can place there is the fractional-knapsack value obtained by consuming its items in decreasing score order, and the greedy rule attains this value simultaneously for every group and depth. Let $S_r(P)$ be the total score in the first r positions. Summation by parts gives

$$U(P) = w_k S_k(P) + \sum_{r=1}^{k-1} (w_r - w_{r+1}) S_r(P).$$

All coefficients are nonnegative. Hence, because the greedy rule maximizes every prefix score S_r at once, it maximizes U for every such w. □

Infeasible targets. If the robust LP (8) is infeasible at $\delta = 0$, the experiments of Section 4.2 run it at its smallest feasible tolerance, which is the smallest risk that any policy can attain on the query. For groups that partition the items, the next theorem computes this smallest risk in closed form. It is the ℓ_∞ distance from the target vector p to the set $\mathcal{Q}$ of target vectors under which some policy is fair under every curve in $\mathcal{V}_k$, and it is attained by running the greedy rule of Section 4.1 at the closest such target.

Theorem H.2 (The smallest risk when the targets cannot be met)**.** *Let the groups partition the n items, let group g have n_g items with $\sum_g n_g = n \ge k$, and let $p \in \mathbb{R}^m_{\ge 0}$ be any target vector. By Theorem 2, the largest prefix disparity $\max_{g,r} d_{g,r}(P)$ of a policy P on the query is its largest disparity under any curve in $\mathcal{V}_k$. With*

$$\mathcal{Q} = \left\{ \tilde{p} \in \mathbb{R}^m_{\ge 0} : \ \textstyle\sum_g \tilde{p}_g = 1, \ \tilde{p}_g \le n_g/k \right\},$$

one has

$$\inf_P \max_{g,r} d_{g,r}(P) \ = \ \min_{\tilde{p} \in \mathcal{Q}} \ \|\tilde{p} - p\|_\infty, \tag{25}$$

and the infimum on the left is attained by the greedy rule of Section 4.1 run at any minimizer $p^\star$ of the right-hand side. In particular, an exactly robust policy, one with $\max_{g,r} d_{g,r}(P) = 0$, exists iff $p \in \mathcal{Q}$, which holds iff $\sum_g p_g = 1$ and $kp_g \le n_g$ for every g.

Proof. The set $\mathcal{Q}$ is nonempty (because $\sum_g n_g/k = n/k \ge 1$) and compact, so the right side is attained.

Method (TREC 2022, continuous quality utility)	utility	$\mathcal{R}_{\mathrm{Q}}$	$r \geq 3$	log-like	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$
Utility sort (continuous)	1.000	0.369	0.349	0.302	0.350
Single-curve LP (TREC curve)	0.9856	0.394	0.230	0.073	0.250
Multi-curve LP (all 7)	0.9842	0.064	0.064	0.0005	0.063
Robust LP, minimum risk	0.9840	0.0005	0.0005	0.0005	0.0005

Table 7: **TREC 2022 with the continuous quality estimate as utility** (official target, 50 evaluation topics). Utility is relative to that of the utility sort under the same utility. Besides $\mathcal{R}_{\mathrm{Q}}$ and $\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$, the table reports the largest prefix disparity at depths three and beyond and the largest disparity under the six log-like published curves.

($\geq$). Let P be a policy and set $\tilde{p}_g = \frac{1}{k}\sum_{j\leq k} x_{g,j}(P)$. Then $\tilde{p}_g \geq 0$. Next, $\sum_g \tilde{p}_g = \frac{1}{k}\sum_{j\leq k}\sum_g x_{g,j} = \frac{1}{k}\cdot k = 1$, because the groups partition the items and every column of P (one per position) has unit mass. Moreover, $\tilde{p}_g = \frac{1}{k}\sum_{i\in G_g}\sum_j P_{ij} \leq \frac{n_g}{k}$, because each row sum is at most one. Hence $\tilde{p} \in \mathcal{Q}$. Write $a_{g,r}(P)$ for the signed prefix deviation of Appendix A, so that $d_{g,r} = |a_{g,r}|$. Since $\tilde{p}_g - p_g = a_{g,k}(P)$, $\min_{p'\in\mathcal{Q}} \|p' - p\|_\infty \leq \|\tilde{p} - p\|_\infty = \max_g d_{g,k}(P) \leq \max_{g,r} d_{g,r}(P)$. Taking the infimum over P gives ($\geq$).

($\leq$). Let $p^\star$ attain the right side. Then $p^\star$ is a probability vector with $kp^\star_g \leq n_g$, so, by Theorem H.1, the greedy rule at target $p^\star$ returns a feasible $P^\star$ with $x_{g,j}(P^\star) = p^\star_g$ for every g and j. Hence $a_{g,r}(P^\star) = p^\star_g - p_g$ for every r, and $\max_{g,r} d_{g,r}(P^\star) = \max_g |p^\star_g - p_g| = \|p^\star - p\|_\infty$.

The final claim follows because the right side of (25) vanishes iff $p \in \mathcal{Q}$. □

For arbitrary overlapping groups and any tolerance δ, utility maximization is still a linear program, the robust LP (8), with nk variables and $O(n + k + mk)$ constraints.

I TREC: Scope and Additional Checks

This appendix gives the protocol and the full results of the TREC Fair Ranking 2022 experiment of Section 4 (Table 8).

Data and selection. The track supplies automated alignments of each page (a Wikipedia article) to gender categories, derived from Wikidata, together with an exogenous target for its Task 2 (Ekstrand et al., 2023). We treat the alignments as benchmark metadata: we did not infer them, and they are not verified identity labels. We keep the Unknown category explicitly. The operating point is selected on all 50 training topics, and all 50 evaluation topics are withheld from this selection.

Optimal solutions of the single-curve LP. As on ESCI (Appendix J.10), the single-curve LP can have many optimal solutions, which differ in risk (Appendix F); the single-curve LP of Table 8 is the solution that the solver returned. Over the optimal solutions of a topic, the range of the querywise risk has mean 0.63 under the published six-level scores, which produce large blocks of tied items, 0.07 when these ties are broken in a fixed order that ignores the groups, and 0.16 under the continuous quality estimate released by the track. Ties therefore explain only part of the spread.

Utility cost when the scores have no ties. The official target is derived from the work-needed utility, so each of the three LPs of Table 8 retains 0.9997 of that utility. To measure the cost when the scores are distinct, Table 7 re-solves three LPs with the track's continuous quality estimate as utility: the single-curve LP (TREC curve), the multi-curve LP (all seven published curves) and the robust LP. The robust LP then keeps 0.9840 of the utility of the utility sort, against 0.9856 for the single-curve LP, a loss of 0.2%.

Blind spots on the real occupancy polytopes. The blind spot of Theorem 3 assumes that every occupancy vector in the cube $[0, 1]^k$ can be realized. On the actual occupancy polytopes of the topics, the blind spot of each (topic, group) cell has mean ratio 1.0000 to the corresponding value on the

Policy	Realization	Utility	EE-L	Disparity under all curves	Disparity under depths ≥ 3	Disparity under log-like curves
Utility sort	deterministic	1.000	8.61	0.399	0.343	0.243
Single-curve LP	fractional	0.9997	3.1×10^{-4}	0.358	0.205	0.067
Multi-curve LP	fractional	0.9997	–	0.071	0.071	5.3×10^{-4}
Robust LP	fractional	0.9997	1.8×10^{-4}	$\mathbf{5.3 \times 10^{-4}}$	$\mathbf{5.3 \times 10^{-4}}$	$\mathbf{5.3 \times 10^{-4}}$
GDL23 (21 topics)	fractional	0.9983	0.0552	0.0218	0.0218	0.0218
Prefix-quota MILP (21 topics)	deterministic	1.000	0.1073	0.221	0.126	0.054

Table 8: **TREC 2022: quality and disparity of each policy** (official Task-2 target, 50 evaluation topics). Utility is the benchmark's own utility relative to that of its utility sort, EE-L is the track's expected-exposure loss against its own target, and the disparity columns are as in Table 2. The count-based methods (GDL23 and the prefix-quota MILP) need groups that partition the pages, which 21 topics have.

cube (Corollary C.8), over 177 cells. Checking all groups jointly gives a mean ratio of 0.9761, so the formula on the cube describes the blind spot on the actual polytopes closely.

I.1 Official target and lossless reduction of the candidates

The official target is first computed on the complete set of relevant documents of each topic. The candidates are then grouped by their released alignment signature, the set of categories to which the track aligns a page (its multi-hot vector); pages with identical signatures form one class. The *signature mass* of a policy at a rank is the probability that the rank shows an item of a given signature. Keeping only the best k items of each signature is lossless, even for fractional (randomized) policies:

Lemma I.1 (Only the top items of each membership type matter)**.** *Let I_τ be the items with signature τ. For every feasible assignment P there is a feasible P' supported on the $\min(k, |I_\tau|)$ highest-score items of each I_τ, with the same signature mass at every rank and utility at least that of P. Hence every group occupancy, querywise risk, and optimum at fixed utility is preserved.*

Proof. Fix a signature τ and the masses that P gives to its items at each rank. These rankwise masses sum to at most k. Maximize the score-weighted utility over the corresponding transportation polytope. If an item outside the top k of I_τ carries mass, then some retained item has unused row capacity, and moving mass to it preserves the signature totals at every rank and cannot decrease utility. Repeating this step removes all mass from the excluded items. Combining the assignments of all signatures gives P'. □

On the evaluation topics, this pruning reduces the mean number of candidates from $54{,}752$ to 66.9 without changing any reported fractional optimum. Table 8 reports the results under the official target. It separates the track's metric under its declared attention curve, the expected-exposure loss (EE-L), from the worst-case disparities over classes of curves.

At the same fractional utility, the robust LP's policy has a far smaller worst-case risk than the single-curve LP's policy.

Details of Table 8. The tolerance ω of GDL23 is selected on the training topics. Utility is relative to that of sorting by the work-needed utility. The robust LP's value 5.3×10^{-4} is its mean smallest feasible tolerance (its floor), which is positive where the target cannot be attained; the robust LP is exact on 25 topics. The topic bootstrap gives a half-width of 0.07 for the querywise risk of the single-curve LP. EE-L is the track's squared expected-exposure loss against its own target, computed over the full relevant set.

Finally, the page alignments derived from Wikidata remain a limitation of construct validity, especially the Unknown category and the track's coarsening of the categories.

J Experimental Protocol and Robustness Checks

This appendix gives the protocols of the ESCI and German Credit experiments of Section 4 and the additional analyses that the main text cites. The TREC protocol is in Appendix I.

J.1 Amazon ESCI: data, targets and baselines

Data. We use the Large split of Amazon ESCI with $k = 10$. We keep the queries that have at least ten candidates and at least two nonmissing catalog-brand strings. This leaves $60{,}965/10{,}801/22{,}030$ training/validation/test queries and $432{,}046$ judged test items. Query identifiers are distinct, but an item may appear in several queries. Catalog brands are public product metadata; they are neither seller identities nor protected attributes.

Relevance model, targets and NDCG. All methods use the same fixed relevance model, a multilingual MiniLM model whose configuration and training epoch are chosen on the validation split. An isotonic calibration map, fitted on training items, converts its scores into estimated relevance. Each brand's target on a test query is its share of this estimated relevance, so the targets are computed per query without test relevance judgments. Test judgments are used only for NDCG. NDCG uses the gains of the dataset paper, 1, 0.1, 0.01 and 0 for exact, substitute, complement and irrelevant items, with the discount $1/\log_2(j+1)$. In Table 2, GDL23 is feasible on $21{,}955$ and the prefix-quota MILP on $22{,}017$ of the $22{,}030$ test queries, so their means exclude 75 and 13 queries. The largest query-bootstrap 95% half-width over all policies and reported columns is 0.0031. These intervals resample queries with the relevance model, the targets and the protocol held fixed, so they omit the uncertainty from training the relevance model and from calibrating the targets.

Use of test judgments. Test judgments are read only after a policy has been fixed, to compute NDCG, and enter no other step of the pipeline.

Baselines. Among the baselines of Table 2, the single-curve LP of Singh & Joachims (2018) is implemented faithfully with its equality constraint under the DCG curve, and the GDL23 sampler (Gorantla et al., 2023a), which draws lists whose group counts lie in an interval around the target, is evaluated through its exact marginals. When a method is infeasible on a query, we report this rather than project the target onto a feasible one. The robust LP first minimizes the querywise risk and then maximizes the score-based utility at the selected tolerance δ.

J.2 Comparison at matched NDCG

Section 4.2 also compares the single-curve and robust LPs at similar quality. The operating points of this comparison are selected on the validation split only. For each method, we take the smallest tolerance on a discrete grid whose validation NDCG reaches 95% or 98% of the utility sort's validation NDCG. This tolerance is δ for the robust LP and an excess tolerance on the DCG constraint for the single-curve LP. Validation judgments enter this choice, and test judgments never do. The rule is the same for both LPs, but the grid is discrete, so the two selected points need not have identical utility: at 95%, the robust LP's point retains 0.3% more validation NDCG than the single-curve LP's. The 95%-NDCG rows of Table 2 report these points.

At 95%, the paired differences, single-curve LP minus robust LP, are -0.00243 in NDCG (95% interval $[-0.00291, -0.00195]$), $+0.20481$ in $\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$ and $+0.41821$ in $\mathcal{R}_{\mathrm{Q}}$. The gap in worst-case risk therefore persists at matched utility.

J.3 Two coarse groups with a proportional target

The main ESCI comparison uses fine-grained brand groups and estimated targets. This subsection checks its conclusions with two coarse groups per query and a target that is not estimated.

Table 9 repeats the comparison with two coarse groups per query, formed by assigning whole brands greedily to the lighter of two buckets, and with the exogenous proportional target $p_g = n_g/n$, where n_g is the number of candidates in group g and n the number of candidates of the query. The single-curve LP costs 1.6% of NDCG, and the robust LP a further 0.3% (paired -0.0027

Method (ESCI, two coarse groups, target $p_g = n_g/n$)	NDCG	$\mathcal{R}_{\mathrm{Q}}$	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{C}_{\mathrm{dec}}}$	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$	native error
Utility sort	0.8221	0.474	0.286	0.351	DCG 0.128
Single-curve LP (DCG), exact target	0.8090	0.407	0.178	0.251	DCG 0
GDL23 sampler ($\omega = 0.05$)	0.8051	0.005	0.005	0.005	count tol. 0.05
Prefix-quota MILP (Celis18/Geyik19)	0.8138	0.474	0.159	0.224	quota 0
Robust LP ($\delta = 0$, selected on validation)	0.8063	**0**	**0**	**0**	prefix 0

Table 9: **ESCI with two coarse groups and a proportional target** (22,030 test queries; every method is feasible on every query; δ and ω selected on validation). "Native error" is each method's error under its own criterion: the disparity under the DCG curve, the count tolerance, the quota violation, or the largest prefix disparity.

$[-0.0032, -0.0021]$). The single-curve LP's policy still has $\mathcal{R}_{\mathrm{Q}} = 0.407$, 0.178 over the bounded-decay class ($\mathcal{R}_{\mathrm{Q}}^{\mathcal{C}_{\mathrm{dec}}}$) and 0.251 over the reference envelope ($\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$), whereas the robust LP attains zero over every class. The whole frontier lies within 2% of the utility sort, so the validation rule selects the tightest setting of every method. GDL23 is nearly exact here. The quotas of the prefix-quota MILP are binding, but its risk stays at $\mathcal{R}_{\mathrm{Q}} = 0.474$, as for the utility sort.

Protocol. To form the two coarse groups of a query, we sort its brands by decreasing number of candidates (breaking ties by brand index) and assign each whole brand in turn to the group that currently has fewer candidates; scores, targets and relevance judgments never influence this map. The target is the exogenous proportional target $p_g = n_g/n$. The tolerance δ of the robust LP and the tolerance ω of GDL23 are selected on the validation split by the analogous 95% rule, applied to the validation score-based utility (the main comparison thresholds the validation NDCG instead). The rule selects $\delta = 0$ and $\omega = 0.05$, because $\omega = 0.025$ is infeasible on 10.4% of the queries and 0.05 is the smallest tolerance feasible on every query.

Paired over the 22,030 test queries, the robust LP at $\delta = 0$ minus the single-curve LP has an NDCG difference of -0.0027 $[-0.0032, -0.0021]$ and a querywise-risk difference of -0.4072 $[-0.4091, -0.4053]$. GDL23 minus the robust LP has a risk difference of 0.0050 $[0.0049, 0.0052]$ (intervals from a query bootstrap with 20,000 draws).

J.4 Sources of the worst-case risk

Tables 10 and 11 break down the querywise risk of every policy of Tables 2 and 8 by depth and by curve (Section 4.3). Columns "$r \geq 2$"–"$r \geq 4$" give the largest prefix disparity from the stated depth on, and "$r^\star$=1" is the fraction of queries on which the largest prefix disparity occurs at depth one. Column "pub. hull" is the largest disparity under any of the seven distinct published curves, which equals the risk under their convex hull, and column "u^1" is the disparity under the prefix curve of depth one.

For the policies of the single-curve LP, most of the querywise risk comes from position one: the largest prefix disparity occurs at depth one ($r^\star = 1$) on 86–88% of the ESCI queries and 80% of the TREC topics. The risk beyond depth two (0.19–0.20), however, remains substantial, and the reference envelope $\mathcal{E}$, whose vertices place up to 45% of the attention on position one, does not remove it (Table 12). The multi-curve LP, which constrains all seven published curves at once, has querywise risk 0.039 on ESCI and 0.071 on TREC, and the multi-curve LP of the five published curves $\mathcal{S}_{\mathrm{prac}}$ leaves 0.063 and 0.112.

The query's own blind spot. The equal-target blind spots of the seven curves in Table 6 (0.0880 at $k = 10$ and 0.1222 at $k = 20$) do not bound the risk on these queries, which have several groups and unequal targets. The quantity that does is the query's own blind spot: the largest prefix disparity that any policy passing the same checks can have on the query's occupancy polytope (the blind spot $B(\mathcal{S}; \delta)$ of (4); Theorem 4; one LP per group, depth and sign). We compute it on a random subsample of 1,000 ESCI queries and on all 50 TREC topics, with one protocol for both. The blind spot is evaluated at the tolerance at which the multi-curve LP operates on the query, its floor plus a small numerical slack, so that the policy and the blind spot use the same δ.

Method	NDCG	$\mathcal{R}_{\mathrm{Q}}$	$r \geq 2$	$r \geq 3$	$r \geq 4$	$r^{\star}{=}1$	pub. hull	u^1
Utility sort	0.822	0.788	0.453	0.335	0.271	98%	0.404	0.785
Single-curve LP	0.773	0.421	0.255	0.187	0.141	88%	0.178	0.414
Single-curve LP, 95%-NDCG point	0.781	0.439	0.275	0.195	0.146	86%	0.187	0.430
Single-curve LP under $\mathrm{RBP}_{0.5}$	0.754	0.228	0.211	0.208	0.205	34%	0.121	0.182
Multi-curve LP, $\mathcal{S}_{\mathrm{prac}}$ (5 curves), exact	0.769	0.063	0.063	0.062	0.055	0%	0.003	0.009
Multi-curve LP, $\mathcal{S}_{\mathrm{prac}}$, excess 0.01	0.777	0.155	0.152	0.117	0.099	15%	0.031	0.096
Multi-curve LP, all 7 curves, exact	0.769	0.039	0.039	0.039	0.039	0%	0.0003	0.002
Multi-curve LP, all 7, excess 0.01	0.777	0.115	0.110	0.109	0.097	26%	0.010	0.087
GDL23 ($\omega = 0.075$)	0.771	0.024	0.024	0.024	0.024	31%	0.024	0.024
Prefix-quota MILP	0.814	0.788	0.396	0.257	0.187	98%	0.357	0.785
Robust LP, 95%-NDCG point	0.783	0.020	0.020	0.020	0.020	8%	0.020	0.020
Robust LP, minimum risk	0.769	0.0005	0.0005	0.0005	0.0005	9%	0.0005	0.0005

Table 10: **ESCI: risk by depth and by curve** (brand groups, estimated targets, $k = 10$, 22,030 test queries) for every policy of Table 2, for the single-curve LP under $\mathrm{RBP}_{0.5}$ and for the multi-curve LPs. Means are over the queries on which a method is feasible.

Method	utility	$\mathcal{R}_{\mathrm{Q}}$	$r \geq 2$	$r \geq 3$	$r \geq 4$	$r^{\star}{=}1$	pub. hull	u^1
Utility sort	1.000	0.399	0.360	0.343	0.323	90%	0.311	0.378
Single-curve LP (TREC curve)	1.000	0.358	0.254	0.205	0.163	80%	0.178	0.351
Multi-curve LP, $\mathcal{S}_{\mathrm{prac}}$ (5 curves), exact	1.000	0.112	0.112	0.102	0.088	0%	0.008	0.039
Multi-curve LP, all 7 curves, exact	1.000	0.071	0.071	0.071	0.069	0%	0.001	0.008
GDL23 ($\omega = 0.05$), 21 topics	0.998	0.022	0.022	0.022	0.022	19%	0.022	0.022
Prefix-quota MILP, 21 topics	1.000	0.221	0.145	0.126	0.099	62%	0.110	0.205
Robust LP, minimum risk	1.000	0.001	0.001	0.001	0.001	4%	0.001	0.001

Table 11: **TREC 2022: risk by depth and by curve** (official Task-2 target, $k = 20$, 50 evaluation topics): the same breakdown for the policies of Table 8 and the multi-curve LP of $\mathcal{S}_{\mathrm{prac}}$.

The seven-curve blind spot averages 0.060 on ESCI (median 0.053, maximum 0.330; it exceeds the equal-target value 0.0880 on 18.4% of the queries) and 0.108 on TREC (it exceeds 0.1222 on 30% of the topics). The blind spot of $\mathcal{S}_{\mathrm{prac}}$ averages 0.103 on ESCI and 0.172 on TREC. The DCG curve alone has a mean blind spot of 0.576 on ESCI, and the TREC curve alone 0.632 on TREC: targets far from $1/2$ make a single curve blind beyond $1/2$ (Proposition C.7). The multi-curve LP's realized risk does not exceed its query's blind spot on any sampled ESCI query or TREC topic, up to numerical error, as Theorem 4 requires. On every query where both the multi-curve LP and the robust LP have a zero floor, the multi-curve LP's utility is at least the robust LP's. On the 477 of the 22,030 test queries where the robust LP's floor is positive, the two LPs operate at different floors.

Positive tolerances. When the checks allow a tolerance of 10^{-3}, the seven-curve blind spot averages 0.115 on ESCI and 0.148 on TREC; at 10^{-2}, it averages 0.197 and 0.219. With an excess tolerance of 0.01 on every curve, the multi-curve LP's risk rises to 0.115 on ESCI, which is the magnification of tolerances of Section 3.2 on real polytopes. The multi-curve LPs cost the same NDCG as the robust LP at its floor, and they leave about 80–210 times its risk.

Disparities relative to the target. An absolute disparity is easier to judge against the targets that it misses. This paragraph details the relative measure used in Sections 4.3 and 4.4: the delivered share $\mathrm{DS}(v)$ of the target exposure under a curve v (Section 4.3), which counts each group's exposure only up to its target. It equals one minus the total shortfall of the under-exposed groups, as a fraction of the total target. The total target is one when the groups partition the items, but the targets of overlapping groups need not sum to one. We compute it in three ways. The first way takes the worst monotone curve of each query. The total shortfall is convex in v, so this curve is a prefix curve, and the delivered share is $\min_{r \in [k]} \mathrm{DS}(u^r)$. The second way uses one curve shared by all queries. The mean over queries of the delivered share is concave in v, so its minimum over $\mathcal{V}_k$ is attained at a prefix curve. On all three datasets, this curve is u^1, the curve of users who read only the first result. The third way takes, for each query, the worst of the curves estimated from click logs, namely the 65 vertices of the hull of the fitted curves and the raw CTR curve ($k = 10$, with the TREC instances of Appendix J.9).

Certified class	excess	NDCG	utility	$\mathcal{R}_{\mathrm{Q}}$	$r \geq 3$	log-like	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$
DCG curve only (single-curve LP)	0	0.7660	0.9540	0.414	0.189	0.074	0.215
	0.02	0.7816	0.9727	0.449	0.198	0.088	0.234
DCG curve and u^1	0	0.7638	0.9516	0.239	0.185	0.038	0.194
	0.02	0.7792	0.9704	0.275	0.206	0.052	0.223
prefix depths $\{1, 2, 3, 5, 8\}$	0	0.7712	0.9591	0.068	0.068	0.047	0.068
	0.02	0.7833	0.9734	0.081	0.081	0.057	0.081
all k prefixes (robust LP)	0	0.7623	0.9492	0.0006	0.0006	0.0005	0.0006
	0.02	0.7770	0.9679	0.0206	0.0206	0.0203	0.0206
reference envelope $\mathcal{E}$	0	0.7624	0.9493	0.0019	0.0006	0.0005	0.0006
	0.02	0.7771	0.9681	0.046	0.0208	0.0203	0.0206
bounded decay $\mathcal{C}_{\mathrm{dec}}$	0	0.7623	0.9491	0.0019	0.0009	0.0005	0.0010
	0.02	0.7771	0.9681	0.064	0.0274	0.0202	0.036

Table 12: **Cost of certifying each class of curves** (ESCI, brand groups, 2,000 test queries). Each row constrains the disparity under the named class to its per-query floor plus the excess, and then maximizes utility. Every policy is then scored under all monotone curves, at depths three and beyond, under the log-like curves and over $\mathcal{E}$.

Unlike the ratio $\frac{\langle v, x_g \rangle}{p_g}$ of a single group, the delivered share needs no lower bound on the targets, because a group with a tiny target enters with a tiny weight. A ratio is unstable for tiny targets: the fourth TREC category has a mean target of 0.0025, so the robust LP's absolute disparity of 5×10^{-4} is already a large part of it.

Table 3 reports the resulting means over queries. Under the worst curve of each query, the single-curve LP delivers on average 42% of the target exposure on ESCI, 64% on TREC and 56% on German Credit, and less than half of it on 78% of the ESCI queries. Under the single curve u^1 (not shown in Table 3), it delivers 42%, 65% and 57%, and under the curves estimated from click logs, 80% on ESCI, 91% on TREC and 90% on German Credit. The robust LP delivers 99.8% or more in every column. On ESCI, it delivers at least 99% on 98.1% of the queries, and less only on queries whose targets cannot be realized, where the robust LP runs at a positive floor.

J.5 Classes of curves between one curve and all curves

We measure the cost of certifying classes of curves between a single curve and all monotone curves. For each class, we solve a two-stage LP on a random subsample of 2,000 test queries (Table 12). The first stage minimizes the largest disparity under the class; the minimum is the class's floor on the query. The second stage maximizes the score-based utility under the constraint that the largest disparity under the class is at most this floor plus an excess of 0 or 0.02. The classes are the DCG curve; the DCG curve and u^1; the prefix curves of the five depths $\{1, 2, 3, 5, 8\}$; all k prefix curves; the 42 vertices of $\mathcal{E}$; and the 512 vertices of $\mathcal{C}_{\mathrm{dec}}$. Every resulting policy is then scored under all monotone curves, at depths three and beyond, under the log-like curves and over $\mathcal{E}$. For every class, the floor is positive on at most 2.7% of the queries. A policy that is exact under $\mathcal{E}$ or $\mathcal{C}_{\mathrm{dec}}$ is nearly exact under all prefix curves, so the last three classes (all prefix curves, $\mathcal{E}$ and $\mathcal{C}_{\mathrm{dec}}$) almost coincide at excess 0 and separate only at excess 0.02.

J.6 German Credit: a real protected attribute

The German Credit experiment of Section 4 uses the UCI Statlog German Credit file (1,000 applicants). For each of three seeds, the applicants are split $60/20/20$ into disjoint pools, stratified by age group and label. An ℓ_2-regularized logistic-regression model, fit on the training pool, serves as the relevance model (test accuracies 0.760, 0.775 and 0.750). Each of the 400 test queries of a seed draws ten applicants under 35 and ten aged 35 or older from the test pool, so each group's target of $1/2$ equals its share of the candidates. The utility is the score-based utility with DCG weights, and NDCG uses the binary creditworthiness label. Before the utility is formed, the scores of each query are made positive by min–max scaling. Every method of Table 2 is run with the same code, and Table 13 reports the results. GDL23 at $\omega = 0.05$ is exactly fair here: its bounds on the group counts in the

Method (German Credit, age groups, $k = 10$)	NDCG	utility	$\mathcal{R}_{\mathrm{Q}}$	$r \geq 3$	log-like	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{C}_{\mathrm{dec}}}$
Utility sort	0.8766	1.000	0.500	0.332	0.205	0.381	0.313
Single-curve LP	0.8599	0.9754	0.442	0.208	0.055	0.259	0.177
Multi-curve LP (all 7 curves), exact	0.8578	0.9656	0.0119	0.0119	0	0.0119	0.0048
GDL23 ($\omega = 0.05$)	0.8561	0.9605	0	0	0	0	0
Prefix-quota MILP	0.8630	0.9805	0.500	0.167	0.100	0.222	0.154
Robust LP, minimum risk	0.8578	0.9655	0	0	0	0	0
Robust LP, floor $+0.02$	0.8597	0.9687	0.0200	0.0200	0.0146	0.0200	0.0191

Table 13: **German Credit, age groups,** $k = 10$ (1,200 test queries over three splits, with query-bootstrap 95% half-widths at most 0.01 for every risk column and 0.003 for NDCG). *Log-like* is the largest disparity under the six published curves other than $\mathrm{RBP}_{0.5}$ (the log-like curves of the main text).

	$\delta = 0$		$\delta = 0.01$	
Checked curves	blind spot	risk	blind spot	risk
five published curves ($\mathcal{S}_{\mathrm{prac}}$)	0.125	0.077	0.292	0.167
our endpoint curves	**0.083**	**0.076**	**0.093**	**0.087**
best five prefixes	0.125	0.115	0.133	0.121
all ten prefixes	0	0	0.010	0.010

Table 14: **German Credit: blind spot and risk of the curve sets of Table 1.** "Risk" is the querywise risk of the policy that maximizes utility subject to the checks at tolerance δ.

top k, $\lceil k(p_g - \omega) \rceil$ and $\lfloor k(p_g + \omega) \rfloor$, both equal five, and it places each group's items uniformly over the positions. With the customary age cut at 25 instead of 35, the results are nearly the same: the single-curve LP has querywise risk 0.436 and the robust LP 0, at NDCG 0.8230 and 0.8155 (query-bootstrap 95% half-widths at most 0.006 for $\mathcal{R}_{\mathrm{Q}}$ and for NDCG).

J.7 Checking five curves on real policies

Section 3.2 compares sets of checked curves by their blind spot, which is a worst case over all policies. We test its predictions on actual policies for German Credit. Every query of this dataset has ten candidates per group and targets $\frac{1}{2}$, so Theorem 3 applies exactly to each query. We take the sets of five and ten curves of Table 1: $\mathcal{S}_{\mathrm{prac}}$, our endpoint curves, the best five prefixes (the prefix curves at depths $1, 2, 3, 5, 8$) and all ten prefixes. For each set and tolerance, we maximize utility subject to the checks and report the querywise risk of the resulting policy (Table 14).

The blind spot is tight: for the prefix curves and our endpoint curves, some policy that passes the checks has a risk equal to the blind spot, and for no set does a passing policy exceed the blind spot. The tolerance also matters as Section 3.2 predicts. Under $\mathcal{S}_{\mathrm{prac}}$, the risk rises from 0.077 at $\delta = 0$ to 0.167 at $\delta = 0.01$, whereas under our endpoint curves it rises only from 0.076 to 0.087.

Protocol. Every query of Table 13 has ten applicants per age group, $k = 10$ and target $1/2$ for each group. Hence every vector in $[0, 1]^k$ is the occupancy vector of the first group under some policy, and the equal-target blind spot of Theorem 3 is the query's own blind spot. For each curve set and each $\delta \in \{0, 10^{-3}, 10^{-2}\}$, we solve the multi-curve LP of the set, which maximizes the score-based utility subject to $|\langle v, x_g \rangle - \frac{1}{2}| \leq \delta$ for every curve v of the set, and record the querywise risk $\mathcal{R}_{\mathrm{Q}}$ of the returned policy on the 1,200 test queries. At $\delta = 0.01$, the endpoint curves are the reweighted endpoint curves of Proposition D.10, as in Table 1. Among the policies of Table 13 and the multi-curve LPs of all curve sets, the largest querywise risk of a policy that passes the checks of a prefix curve set or of the reweighted endpoint curves equals the blind spot of the set to four digits at every tolerance, and no passing policy exceeds the blind spot of any set, as Theorem 3 requires. Continuous scores make the optimal solution of each multi-curve LP essentially unique: over its set of optimal solutions, the risk varies by at most 10^{-4}. The differences in Table 14 are therefore properties of the curve sets, and no set dominates the others: the multi-curve LP of the five prefix depths retains more utility (0.976 against 0.970–0.972) at a higher risk.

J.8 PER-QUERY COST OF CERTIFYING EVERY CURVE

We compute the distribution of the paired NDCG loss over the 22,030 ESCI test queries. On each query, the loss of a policy against a comparator is the comparator's NDCG minus the policy's NDCG. We compare the robust LP with the single-curve LP at the two operating points of Table 2 (the smallest feasible tolerance and the 95%-NDCG point), and each of the two LPs with the utility sort.

At the smallest feasible tolerances, the additional cost of the prefix constraints is small on average and zero on three quarters of the queries. It has a tail, however: 18% of the queries lose more than 0.01 NDCG, and 7% lose more than 0.05. On average, the robust LP loses 0.0042 NDCG against the single-curve LP and 0.0529 against the utility sort, which amount to 0.54% and 6.43% of the comparator's mean NDCG. The 95%-NDCG rule that selects the operating points on validation data constrains the mean validation NDCG, not the NDCG of individual queries.

J.9 ATTENTION CURVES ESTIMATED FROM BAIDU-ULTR CLICK LOGS

The click-log experiment of Section 4.4 estimates attention curves from real click logs instead of assuming a restricted class of curves, such as the reference envelope $\mathcal{E}$ (Appendix G). We estimate a curve for each of several deployment contexts that the logs record, such as the height of the top result, and measure how much a check under a single curve hides when the unknown curve is restricted to the estimated curves.

Data and click model. We use the release of Baidu-ULTR (Zou et al., 2022) by the University of Amsterdam (Hager et al., 2024). For each displayed document, it stores a 768-dimensional query–document embedding, the position, the click, the media type, the pixel height and the display time. From six training shards (part-1, splits 0–5), we keep the query pages whose rank-one document is present and that have at least five documents at positions 1–10. This leaves $325{,}271$ query pages (for most queries, the release stores 5–9 of the ten first-page documents; 2.66M document rows in all). The held-out click shard part-0/split-0 ($54{,}263$ pages) is used only for early stopping and for the raw CTR curve below.

We fit a *two-tower* click model, which writes the click probability as the product of a relevance factor and a position factor, $\Pr(\text{click}) = \sigma(\rho(x))\,\sigma(\theta_{c,j})$, where σ is the logistic function. The first factor applies a relevance network ρ to the embedding x, and the second uses a position-bias table $\theta_{c,j}$ with one entry per context c and position j. Contexts are defined in four ways, which we call *context families*: the pooled log, which puts all pages in one context; the pixel height of the top result's card, split at the training terciles (≤ 220 px, middle, > 352 px); the media type of the top result (the four most common types and "other"); and the query length, split at the training quartiles (≤ 7, 8–9, 10–12, > 12 tokens). The four bias tables, one per family, are trained jointly with the shared relevance network (AdamW, two epochs, early stopping on the validation log-loss).

Seed 0 gives the point estimate, and seeds 1–4 refit the model on bootstrap resamples of the queries. The ranges reported below are the minimum and maximum over these four refits and are not confidence intervals. The study is therefore a stress test of the policies under curves estimated from click logs.

From position effects to curves. The curve of context c has entries $v_j \propto \sigma(\theta_{c,j})/\sigma(\theta_{c,1})$, the position factors relative to position 1, normalized to sum to one. An estimated curve need not be non-increasing, so we project it onto the cone of non-increasing vectors (the Euclidean projection, computed by the pool-adjacent-violators algorithm of isotonic regression) and renormalize. The projection moves every curve by an ℓ_1 distance of at most 0.029, except for one media type, which moves by 0.081. Even the pooled curve is changed by the projection, which gives ranks three and four equal weight.

The click-calibrated envelope. We define the *click-calibrated envelope* $\widehat{\mathcal{E}}$, a restricted class of attention curves, as the convex hull of the 65 estimated curves (13 contexts $\times$ 5 seeds). A disparity is the absolute value of a linear function of the curve, so it is convex in the curve, and its worst case over $\widehat{\mathcal{E}}$ is attained at a vertex of the hull, which is one of the estimated curves. If the hull contains only the 13 point estimates, checks under the pooled click curve, the DCG curve and $\text{RBP}_{0.8}$ have blind spots

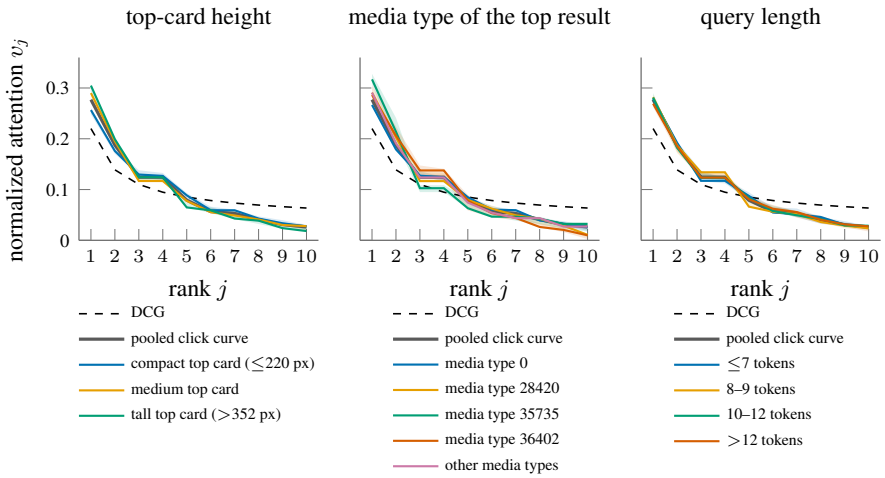


Figure 3: **Attention curves estimated from Baidu-ULTR clicks, by context** ($k = 10$). Lines are the monotone point estimates, and shaded bands the ranges over four refits on bootstrap resamples of the queries. The pooled curve gives rank one a weight of 0.277, against 0.220 for DCG, and puts 0.589 of its mass on the top three positions. Across contexts, v_1 ranges from 0.257 to 0.317.

0.073, 0.188 and 0.123 over the hull (in the setting of Table 15). With the refits, these blind spots are 0.089, 0.221 and 0.139, and the extreme vertices come from the two smallest contexts.

Curves that depend on the displayed list. The top-card height and the media type of the top result are properties of the rank-one result. Under these contexts, the curve therefore depends on the displayed page, which the position-based model of Section 2 excludes. The prefix certificate constrains only the marginal occupancies $x_{g,j}$, and a curve that varies with the displayed list can shift exposure even when every prefix of the marginals is balanced. For example, take two groups A and B with target $\frac{1}{2}$ each, and a policy that shows the lists AB and BA with probability $\frac{1}{2}$ each. All its prefix deviations are zero. Yet if users follow u^1 when AB is shown and u^2 when BA is shown, group A receives exposure $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = 0.75$. The prefix certificate of Theorem 2, which is built from the unconditional position marginals, therefore does not cover curves that depend on the list. (Over the full monotone class, the obstruction is concrete: on every displayed list, the group at rank one fills the whole prefix of depth one, so if users follow u^1 on these lists, the group's exposure is at least its probability of being shown at rank one.) Certifying attention that depends on the list would require a restricted class of curves together with constraints on the occupancies conditional on the rank-one item, which we leave open. Below, we therefore apply each context curve as an exogenous curve to a whole policy, as would an auditor who assumed that context. Query length is the only family that is fixed before ranking, and it barely changes the curve.

Scope of the estimates. The curves are estimated at $k = 10$, the page size of Baidu. This is why the TREC experiments with these curves (Table 3) display ten positions, whereas the main TREC experiment (Section 4.2, Table 8) uses $k = 20$. Two contexts, both media types, have fewer than 20,000 training queries (16,837 and 14,368). The estimates assume that the released subset of documents was not selected on clicks and that the relevance network absorbs the confounding between relevance and position. Hager et al. (2024) report that position-bias models fitted on this log transfer poorly to the expert annotations, so we interpret the curves as descriptions of where clicks fall in each context rather than as validated examination probabilities.

Raw click-through rates. The raw CTR curve of the held-out shard is a naive estimate: the number of clicks at each position, normalized. It has $v_1 = 0.400$, against 0.277 for the two-tower fit. The two curves are at ℓ_1 distance 0.33, and the raw CTR curve lies outside $\widehat{\mathcal{E}}$. The two estimators bracket the share of rank-one clicks that is attributed to relevance rather than to attention.

Audit curve	pooled click	DCG	$\mathrm{RBP}_{0.8}$	top-card contexts	media contexts	length contexts	envelope $\widehat{\mathcal{E}}$
pooled click curve	0	0.117	0.067	0.022–0.041	0.020–0.073	0.010–0.024	0.089
DCG	0.141	0	0.102	0.115–0.180	0.124–0.188	0.130–0.159	0.221
$\mathrm{RBP}_{0.8}$	0.070	0.098	0	0.043–0.108	0.053–0.123	0.060–0.084	0.139

Table 15: **Disparity hidden by a check under one curve** (two groups with target $1/2$ each, as in Theorem 3; $\delta = 0$). Each entry is the largest disparity under the column curve of a policy whose disparity under the row curve, the constrained curve, is zero. The context columns give the range over the context curves of each family, and the last column gives the blind spot over the click-calibrated envelope $\widehat{\mathcal{E}}$.

Dataset	policy	utility	$\mathcal{R}_{\mathrm{Q}}$	pooled click	context curves (max)	$\mathcal{R}_{\mathrm{Q}}^{\widehat{\mathcal{E}}}$	raw CTR curve	$\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$
ESCI (2,000 test queries)	Utility sort	1.000	0.792	0.219	0.255	0.269	0.317	0.409
	Single-curve LP	0.9540	0.414	0.052	0.079	0.092	0.114	0.214
	Single-curve LP (pooled click curve)	0.9482	0.346	0	0.026	0.038	0.062	0.163
	Multi-curve LP (all 7), exact	0.9490	0.039	0.0066	0.0099	0.012	0.0025	0.039
	Prefix-quota MILP	0.9964	0.792	0.176	0.211	0.224	0.276	0.354
	Robust LP, minimum risk	0.9492	0.0006	0.0005	0.0005	0.0005	0.0005	0.0006
	Robust LP, floor $+0.02$	0.9679	0.021	0.020	0.020	0.020	0.020	0.021
German Credit (age 35)	Utility sort	1.000	0.500	0.186	0.213	0.229	0.234	0.381
	Single-curve LP	0.9754	0.442	0.041	0.067	0.082	0.078	0.259
	Multi-curve LP (all 7), exact	0.9656	0.012	0.0005	0.0018	0.0022	0.0002	0.012
	Prefix-quota MILP	0.9805	0.500	0.055	0.064	0.082	0.110	0.222
	Robust LP, minimum risk	0.9655	0	0	0	0	0	0
	Robust LP, floor $+0.02$	0.9687	0.020	0.014	0.015	0.016	0.014	0.020

Table 16: **ESCI and German Credit policies under the click-calibrated envelope** ($k = 10$, as for the estimated curves). Columns: disparity under the pooled click curve, largest disparity under the twelve context curves, risk over $\widehat{\mathcal{E}}$, disparity under the raw CTR curve ($v_1 = 0.400$), and risk over the reference envelope $\mathcal{E}$. ESCI uses a random sample of 2,000 test queries, and German Credit all 1,200 test queries.

Context curves. The context curves differ measurably (Figure 3). Their v_1 runs from 0.257 under a compact top card to 0.317 under one media type, and their top-three mass from 0.562 to 0.634. The refit ranges of the three height terciles do not overlap. Query length barely matters (v_1 from 0.269 to 0.282). As the only checked curve, each of these curves has blind spot $1/2$ over all monotone curves, the largest possible value (Proposition C.7).

Table 15 measures what a check under a single curve hides from these curves. A check under the pooled click curve hides up to 0.073 from a context curve and 0.089 over $\widehat{\mathcal{E}}$. A check under DCG, whose curve puts less weight on rank one (0.220 against 0.277), hides 0.141 from the pooled click curve, 0.115–0.188 from the context curves and 0.221 over $\widehat{\mathcal{E}}$. The largest disparity that one context curve hides from another is 0.103. These values lie between the disparities that remain under the log-like published curves (0.075–0.082, Table 2) and the risk of the single-curve LP over the reference envelope $\mathcal{E}$ (0.21–0.26; Tables 12 and 13).

On TREC, re-ranked with $k = 10$, the policy that is exactly fair under DCG is off target by 0.034–0.054 under the context curves and by 0.072 over $\widehat{\mathcal{E}}$. The policy that is fair under the pooled click curve, the best that an auditor with one estimated curve can deploy, is off by 0.004–0.022 and 0.039. Both have querywise risk 0.28–0.30, whereas the robust LP stays at its floor, 0.0005, under every curve, at essentially the same utility (0.9997–0.9998 of the utility sort's for all three LPs).

ESCI and German Credit under the same curves. Both benchmarks display ten positions, so the Baidu curves apply directly (Table 16). On ESCI, the single-curve LP is off target by 0.052 under the pooled click curve, by 0.079 under the worst context curve and by 0.092 over $\widehat{\mathcal{E}}$. On German Credit, it is off by 0.041, 0.067 and 0.082. By comparison, its risk is 0.214 (ESCI) and 0.259 (German Credit) over the reference envelope $\mathcal{E}$, and 0.414 and 0.442 over all monotone curves. Under the raw CTR curve, the single-curve LP is off by 0.114 on ESCI and 0.078 on German Credit, so the estimator changes the remaining disparity by more than the contexts do.

On ESCI, a policy made exactly fair under the pooled click curve, the best that a single estimated curve can do, is still off by 0.026 under another context curve, by 0.038 over $\widehat{\mathcal{E}}$ and by 0.062 under

Policy	NDCG	Utility	Disparity under all curves	Disparity under depths ≥ 3	Disparity under log-like curves
Single-curve LP	0.7734	0.9538	0.421	0.187	0.075
most robust solution	0.7734	0.9538	0.334	0.160	0.069
most robust at 99% utility	0.7703	0.9507	0.0158	0.0135	0.0051
Robust LP	0.7692	0.9491	**0.00047**	**0.00047**	**0.00040**

Table 17: **ESCI: two selections among the solutions of the single-curve LP**, with the columns of Table 2. The first row is the solution that the solver returned.

the raw CTR curve. It keeps 0.9482 of the utility sort's utility, against 0.9492 for the robust LP, whose disparity is 0.0005 under every estimated curve. Across the three benchmarks, the single-curve LP with the DCG curve therefore leaves 0.04–0.08 under the estimated curves and 0.07–0.09 over their hull, and a single-curve LP under an estimated curve leaves 0.04 over the hull.

J.10 The single-curve LP has many solutions

A linear program can have many optimal solutions, which form a face of its feasible polytope, and a solver returns one of them. For the single-curve LP on ESCI, these solutions differ widely in robustness. On $12{,}044$ of the $22{,}030$ queries, the querywise risks of the most and the least robust solution differ by more than 0.01, and their means are 0.334 and 0.490. The single-curve LP of Table 2 is the solution that the solver returned, and Table 17 adds two selections. The most robust solution has risk 0.334 at the same NDCG. The most robust policy that keeps fairness under the DCG curve and 99% of the optimal utility has risk 0.0158, at an NDCG cost of 0.003.

The solutions are many because the utility weights w_j are proportional to the DCG curve that the LP constrains. When the items of a group have similar scores, the utility is determined by the groups' exposures under this curve, which the fairness constraint fixes (Appendix F). If a single-curve LP is used, the most robust of its optimal solutions should therefore be selected.

For each query, the analysis works at exactly the optimal utility of the single-curve LP. On the $22{,}014$ queries on which the target can be met exactly, the fairness constraints are imposed as equalities; the other 16 queries use the min-violation extension.

We compute the range of the risk over the set of optimal solutions, whose length we call the *risk spread*, for two classes of curves: all monotone curves (risk $\mathcal{R}_{\mathrm{Q}}$) and the reference envelope $\mathcal{E}$ (risk $\mathcal{R}_{\mathrm{Q}}^{\mathcal{E}}$). For each class, the lower endpoint is one LP in epigraph form, which minimizes a bound on all disparities of the class. The upper endpoint is the largest of several LPs, one for every signed objective of a group and a curve: $2mk$ LPs for all monotone curves (the k prefix curves, with both signs, for each of the m groups) and $84m$ LPs for the reference envelope (its 42 vertices, with both signs). A query has on average 12.27 active brands (median 12, maximum 40), and the number of groups is not capped.

Removing the queries with exact score ties leaves the risk spread essentially unchanged (median 0.1003 on the $21{,}115$ remaining queries, against 0.0992 on all $22{,}014$), whereas on the 396 queries without single-item brands its median is zero. This pattern is consistent with Proposition F.1: on average, 72.4% of a query's brands have a single item, so scores are nearly constant within groups, and the constrained DCG curve is proportional to the utility weights. Replacing either the constrained curve by $\mathrm{RBP}_{0.8}$ or the brands by two coarse groups removes almost all of the spread, but the optimal solutions then remain risky (mean querywise risk 0.36–0.45 at optimal utility), whereas the least risky policy within a 1% utility allowance has mean risk 0.008–0.020 in all four combinations. Selecting a low-risk policy among the nearly optimal ones therefore removes most of the risk, whereas changing the constrained curve or the grouping does not.

J.11 Computation time

Per-query solve times on ESCI ($k = 10$, HiGHS LP solver, one CPU core) are 0.06 ms for the utility sort; 11.5 ms for the single-curve LP at exact fairness and 6.3 ms at its 95%-NDCG point; 17.9 ms for the robust LP at its smallest feasible tolerance and 9.6 ms at its 95%-NDCG point; 1.0 ms for the

GDL23 sampler; and 19.5 ms for the prefix-quota MILP. On larger candidate pools, sampled from the relevant set of each TREC topic ($1{,}000$ pages, $k = 20$, five trials, one CPU core), the median solve times are 580 ms for the single-curve LP and $1{,}221$ ms for the robust LP.