Revenue, Relevance, Arbitrage and More: Joint Optimization Framework for Search Experiences in Two-Sided Marketplaces

Andrew Stanton
Etsy, Inc
astanton@etsy.com

Akhila Ananthram
Etsy, Inc
aananthram@etsy.com

Congzhe Su
Etsy, Inc
csu@etsy.com

Liangjie Hong
LinkedIn Inc.
hongliangjie@gmail.com

ABSTRACT

Two-sided marketplaces, such as eBay, Etsy and Taobao, provide services to satisfy the demand from both buyers and sellers; buyers seek the relevant and interesting item to purchase, while sellers reach out to their audience and grow their business. Concurrently, platforms work to realize their business objectives, ranging from growing user bases to maximizing revenue. It is challenging to obtain a globally favorable outcome for buyer, seller and platform and often results with the Search experience, one of the most important entry point in E-commerce, attempting to satisfy conflicting needs from multipe parties. To address this issue, we formulate market-level metrics as constraints and demonstrate tuning conflicting metrics for business needs. We explore using Evolutionary Strategies to optimize policies, improving both group-level and market-level metrics for all parties simultaneously. We evaluate the proposed method offline on the top 5,000 queries in Etsy search logs and present results, validating our framework is successful in improving the joint objectives without sacrificing buyer's experience, seller's demand, and platform's business interest.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; *Information retrieval diversity*.

KEYWORDS

neural networks, learning to rank, evolutionary strategies, e-commerce

ACM Reference Format:
Andrew Stanton, Akhila Ananthram, Congzhe Su, and Liangjie Hong. 2020. Revenue, Relevance, Arbitrage and More: Joint Optimization Framework for Search Experiences in Two-Sided Marketplaces. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

As online shopping becomes a dominant avenue for global buyers, E-commerce companies strive to meet a wide range of often conflicting goals when showing products for their communities. While optimizing for buyer *Conversion Rate*, and therefore higher general revenue or Gross-Merchandise-Value (GMV), is commonly considered the top objective, it is usually far from driving a healthy and growing business. In fact, many E-commerce companies, especially those with two-sided marketplaces, face a number of challenges due to over-shackling to GMV optimization alone. Without appropriate tempering, such a platform is usually unable to satisfy the short term needs of buyers and sellers as well as the long term needs of the business.

A typical two-sided marketplace such as eBay, Etsy and Taobao has two distinct groups of customers where buyers use the platform to seek the most relevant and interesting item to purchase and sellers view the same platform as a tool to reach out to their audience and grow their business. On top of that, the platform normally would have its own objectives ranging from growing both buyer and seller user bases to GMV maximization. It is not difficult to see that it would be challenging to obtain a globally favorable outcome for all parties.

Take showing relevant products to a buyer through the search experience as an example. For a particular purchase intent, or sometimes with a specific item in mind, a buyer would likely discover a spectrum of product listings from multiple sellers in a typical two-sided marketplace. Some seem to be more relevant than others and some even might look the same. A buyer has to decide among these items with a positive experience such that he/she would return to the marketplace next time. For sellers, however, they view search result pages as prominent real estate to gain customers' attention and therefore potentially increase their market share. Maximizing their success in search benefits both their brand and take home pay, regardless of buyers' overall experience on the site or ramifications to other sellers. On the mission of growing a marketplace, the platform generally needs to step in and sometimes artificially advantage under represented segments of sellers to give them more exposure, creating a reasonably fair competition. However, it might be equally risky to put established sellers in disadvantaged situations, who originally rank well on their own merit, and provide a sub-standard experience to buyers as potentially fewer relevant and lower quality goods are exposed higher through search results.

This scenario exemplifies a microcosm of the marketplace: any interventions are likely to impact either buyers or sellers unfairly to course correct for a greater perceived need. On another hand, a two-sided platform also needs to be cautious about the situation where the rich get richer and the poor get poorer, namely the *Matthew Effect*, a factor of constant battle in most marketplaces [2].

SIGIR eCom'20, July 30, 2020, Virtual Event, China

While power sellers often perform well due to better commercial strategies and product fit than their peers, a healthy marketplace cannot be solely dominated by a small segment of power sellers and continue to grow. To make things even more complicated, platforms, often operating as modern corporations, subsequently attempt to compensate for these inefficiencies with organizations and teams devoted to their respective customer: for example, buyer, seller, and core market. Hence, as each team attempts to solve a particular problem set, competing needs are demanded of the search experience with each team expecting tuning for their particular business focus. As often these asks are ill-defined and heuristically measured, a grand challenge for building a search ranking algorithm to satisfy all fronts is presented.

In this paper, we address how a company-aligned search experience can be provided with competing business metrics that E-commerce companies typically tackle. As far as we know, this is a pioneering work to consider multiple different aspects of business metrics in two-sided marketplaces to optimize a search experience. We demonstrate that many problems are difficult or impossible to decompose down to credit assigned scores on individual documents, rendering standard point-wise approaches to multi-objective [30] or standard diversity-based[26] learning to rank algorithms inadequate. Instead, we express market-level metrics as constraints and discuss to what degree multiple potentially conflicting objectives can be tuned to business needs. In addition, we propose a policy learner in the form of *Evolutionary Strategies* to jointly optimize both group-level and market-level metrics simultaneously, sidestepping traditional cascading methods and manual interventions.

The paper is organized as follows. In §2, we discuss related work in several different directions. We then formulate and define a wide range of metrics including relevancy metrics, diversity metrics and a number of newly proposed market-level metrics relevant to E-commerce interests in Section §3. We follow up with a proposed set of policies to optimize the above metrics in §4. Finally in §5, we empirically evaluate the effectiveness of proposed method on Etsy search logs data, showing how different weightings influence the ultimately delivered rankings.

2 RELATED WORK

There are number of different facets of work that need to be considered when ranking across a variety of soft constraints.

2.1 Diversity

Diversity in Learning to Rank has long and storied past as it relates to Web Search. The simplest solution typically falls under a heuristic based approach. Carbonell and Goldstein[5] formulated the problem as a selection: Documents are chosen greedily based on a linear combination of query-document relevance and maximal margin relevance (MMR), with each step picking the document which the highest combined score. After selection, the MMR scores are updated to reflect the newly picked document. MMR is rooted in the idea of *novelty*, the idea that maximizing the differences in similarity between the set of already selected documents and the remaining set results in a more diverse outcome. Subsequent work attempts to define better heuristics for selection [26]. Dang and Croft propose PM-2 [13], a diversification method based on proportionality; they argue diversification should be biased toward the overall subtopic proportionality of the entire query-set rather than attempting to balance it uniformly. xQuAD [27] attempts to understand diversity as a combination of an originating query and derived sub-queries.

In the learning space, the closest related work comes from Xia et al. who describe PAMM [33], a method for optimizing diversity and relevancy via a Perceptron. Novel to the paper is the idea of direct optimization of the evaluation metrics rather than utilizing heuristics or optimizing surrogate functions. PAMM works generally by sampling both positively and negatively ranked lists and attempts to maximize the margin between them.

2.2 Policy Learning

Policy optimization in LTR space has come in a few flavors over the years, with most of its history focused in online LTR. Radlinski et al. [22] first discussed diversity-based online optimization in the form of using multi-armed bandits to minimize page abandonment. They condition the expected reward on previous documents selected, considering each remaining document an "arm", allowing for suitable exploration/exploitation trade-off against the expected reward. More recent work proposes modeling user behavior as an MDP[15] with the goal of learning how browser sessions can be utilized in re-ranking. They proceed to describe a policy gradient method to learn optimal ranking policies given the learned SS-MDP.

Singh et al.[28] explores incorporating both buyer and seller level utility into learning ranking policies via policy gradient methods, selecting candidates using the Plackett-Luce model. While similar in concept to our work, optimization scales less well and is constrained to differentiable models - limitations our approach does not suffer from. Additionally, Plackett-Luce is unable to incorporate conditional features for selection, limiting the types of objectives it is able to optimize.

Most applicable to our proposed method is utilizing black box optimization in learning to rank. Salimans et al. recently showed that Evolutionary Strategies (ES) [25] were well suited for learning reinforcement problems, applying a variation of ES known as Natural Evolutionary Strategies[32] to Atari game learning. ES is an optimization approach that estimates a gradient step in policy space by perturbing a policy several times with a noise distribution, typically a Normal distribution, evaluating them according to an unknown fitness function, and combining them to from a directional step. While intuitively similar to Finite Differences gradient estimation, they exhibit properties more conducive to fitness functions which are non-differentiable or discontinuous[19], properties which naturally arise in sorting problems such as ranking. They proposed a scalable algorithm and demonstrated the optimizer's tolerance to stochastic environments. [9] showed competitive results to Salimans via *Canonical ES* - a simpler version of the $(\lambda, 1)$ variants of Evolutionary Strategies. Concurrently, Ibrahim et al.[16] applied perhaps the simplest type of (1+1)-Evolutionary Strategies to learn policies on linear models to directly optimize the average nDCG across all query sets.

2.3 Popularity Bias and the Matthew Effect

Measuring market level performance within search is fairly under researched in the space of E-Commerce. Perhaps closest to our particular use case is from Hentenryck et al. [31] who analyzed the impacts of social influence on a trial-offer market, showing that ranking on conditional purchase rate lead to natural monopolies by the highest quality products. Follow up work attempts to compensate for the Matthew Effect [2] by intervening with a stochastic policy to randomize products of similar quality. They show how segmenting products into different "worlds" and conditioning popularity on which world a user observes results in a stable market where products of similar quality obtain equivalent market share.

Blake et al. [3] estimated the actual costs associated with running searches at eBay, showing that the cost of searching was substantially lower than previously measured. More relevant to this work, Moshary et al. [21] showed how factors such as price saliency, that is how easy it is to determine the true price of a good, can significantly impact the cost of items purchased - highlighting how not all buyer beneficial changes result in an improved bottom line.

2.4 Multi-Objective Optimization

A number of other works have started incorporating multiple objectives within ranking models, taking the form of either ensembles of expert models focused on individual objectives or single models incorporating some weighted combination of target goals. Most related to our work is from Momma et al. Momma et al. [20] which incorporates multiple-objectives via inclusion of Augmented Lagrangian constraints within the LambdaMART gradient computation.

Recent work from Carmel et al. [6] introduced a stochastic label aggregation approach to transform multiple sets of ranked documents into a single objective optimization problem. They further prove their approach is superior to deterministic, linearly interpolated combinations of ranked labels.

3 METRICS FOR OPTIMIZATION

In this section, we outline metrics for optimization from a typical two-sided marketplace. We review classic relevancy metrics in §3.1 as well as diversity metrics in §3.2, serving the foundation of metrics for modern search ranking. We introduce a new class of market-level metrics in §3.3. We list all notations used through out the paper in Table 1.

3.1 Relevancy Metrics

First and foremost is the concept of relevancy, rooted originally in the well-known *Probability Ranking Principle* (PRP) framework [11, 24] which states that documents should be ordered independently in decreasing presentation of relevance. That is, the most relevant document for a query should be placed first. To measure that principal, industry has standardized around two core metrics for evaluating the efficacy of their ranking systems: NDCG [18] and ERR [8]. We now give a brief overview of their formulation.

Normalized Discounted Cumulative Gain (NDCG): NDCG is an ordered relevance metric measuring the agreement between a goldset list of documents and the permutation return by the ranking policy. It is typically evaluated to some position K, indicating only the first K documents should be considered for evaluation. Usually values for K are small, emphasizing the importance of getting the first few documents correct.

Table 1: Notation

Notation	Description
$Q = \{q_1, ..., q_n\}$	Unique query set
N_i	Number of documents in query set q_i
$D = \{d_1, ..., d_j\}$	Document set
$Feats(d_i)$	Feature vector for document d_i
$Y \in \{1, 2, 3, 4, 5\}$	Relevance grades
$y_{i,j} \in Y$	Relevance score for q_i and d_j
π	Ranking policy
$R(\pi, q_i) =< d_1, \ldots) >$	Ranked documents for query q_i
$V(\pi, q_i, d_j)$	Evaluator for query-doc set
$S(\pi, Q, D) \in [0, 1]$	Market-level objective function

Expected Reciprocal Rank (ERR): ERR [8] is proposed as an adjustment to NDCG, attempting to factor in a prior to how users actually consider documents for engagement. While graded labels are still assigned to documents independently, ERR is grounded on the idea of the cascade user model [12]: that previous evaluations of documents influences the likelihood that a buyer will continue browsing. While a full discussion on ERR is out of the scope of this paper, we provide the following formulation used in evaluation. Given some mapping R of relevance grades to probability of relevance, we can define ERR as:

$$R(g) = \frac{2^g - 1}{2^{\max(Y)}}$$

$$p_0 = 1, p_j = p_{j-1}(1 - R(y_{i,j}))$$

$$\text{ERR}_0 = 0, \text{ERR}_j = \text{ERR}_{j-1} + p_{j-1}\frac{R(y_{i,j})}{j}$$

3.2 Diversity Metrics

Evaluation of sub-topic diversity is a rich field with many contributions [1, 7, 10, 34]. While there are many diversity questions related to ordered result sets, such under-specification, we focus on metrics revolving around the idea that there exist ambiguity in the topicality behind a query. For example, while there is likely a strong relationship between the query “lace bridal veil” and the `/clothing/wedding/accessory/veils` taxonomy, for other queries there is less implicit understanding. On Etsy, we serve a large number of inspirational (cheerful, happy, beautiful), stylistic (geometric, upcycled, animal print), and occasion (gifts for him, bridesmaid presents, stocking stuffers) queries which have high taxonomic (interchangeably used with topicality) diversity. In cases where there is low certainty of strong topicality, ranking benefits from increasing coverage of different sub topics early on in the presented result set. Indeed, empirical results have shown increasing diversity improves user engagement metrics [23].

ERR-IA: ERR-IA [8] is an extension to ERR that incorporates the notion of diversity. It incorporates topicality by computing the ERR for each subtopic independently, then weighing their importance by the ratio of each subtopic with respect to entire result set. Let

$\Pr(t \mid q)$ be the probability of a topic, t, for a given query, q. We can define ERR-IA as:

$$R(y_i, t) = \begin{cases} R(y_i) & i \in t \\ 0 & else \end{cases}$$

$$\text{ERRIA@K} = \sum_{j=1}^{K} \frac{1}{j} \sum_{t} \Pr(t \mid q) \prod_{i=1}^{j-1} (1 - R(y_i, t)) R(y_j, t)$$

Or, using the ERR formula from before, we can define it as:

$$\begin{aligned} \text{scores} &= [y_i * \mathbb{1}[i \in cat_t], \forall i \in [1, K]] \\ \text{ERRIA@K} &= \sum_{t} \Pr(t \mid q) * \text{ERR@K}(\text{scores}) \end{aligned} \tag{1}$$

3.3 Market-Level Metrics

Two sided marketplaces, unlike traditional web search, suffer from multiple challenges typically framed in the form of *inequality*; the realization that there exists some skew in the marketplace we wish to correct. In this sub section we introduce a wide range of different types of market corrections and present potential metrics for optimization. Before we introduce the first proposed metric, we discuss the notion of query-set dependence below.

Query-Set Dependence: Let us consider the case of balancing between highly successful power sellers and new sellers on the site. While majority of the sales will come from a fraction of the entire seller base, an over saturation of a small proportion of sellers in search has the potential to squeeze out new shops on the site. As there is strong evidence that faster time to first sale increases a seller's *Life Time Value* (LTV), we would ideally like to improve their exposure. Applying a diversity constraint between power sellers and new sellers seems reasonable: for each query set, increase coverage between the two distributions in the top K spots. However, this ignores a fundamental problem: different queries have different amounts of traffic associated with them, preventing us from properly balancing the market. Much like the PRP models before, diversity metrics assume query-level independence: we only consider how the impact of diversification adjusts the metrics within each query set, not its contribution to the overall marketplace.

Below we discuss different methods for addressing marketplace challenges.

3.3.1 Weighted Importance Ranking. Often times there are subsets of the traffic we deem more critical for the ranking policy to consider. For example, it benefits the business ensure the highest revenue generating queries are correct, even if it is at the expense of queries deemed less important. To account for this, we incorporate a weight for each query; given a scoring function S_{sub} and a set of importance weights W_i, we can describe our score function S as:

$$S(\pi) = \frac{\sum_{i=1}^{N} W_i * S_{sub}(\pi, q_i, d_i)}{\sum_{i=1}^{N} W_i} \tag{2}$$

This provides a useful feedback mechanism to the policy learner: much like the intuition that rankers optimizing for NDCG or ERR should focus their energy on improving the rank of documents higher in the page, providing importance feedback provides guidance to the policy on which queries it should focus time on optimizing.

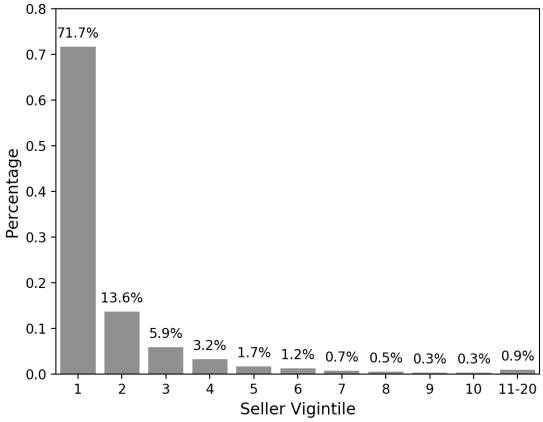

Figure 1: Seller Distribution of shop revenue

3.3.2 Outlier Skew. When dealing with implicit feedback data, chronic cold starts, or otherwise uncertainty in the relevance set, optimizing for the expected NDCG across all queries can lead to an over-sensitivity to outliers: query sets that are either trivial or impossible to successfully rank given the features.

To address the influence, we propose a simple change to the optimization function to maximize the scores at given percentiles instead of the mean. Given the set of scores, M, a set of percentiles to evaluate (e.g. 25th, 75th, etc.), Percentiles:

$$\begin{aligned} M &= \{V(\pi, q_1, d_q), ..., V(\pi, q_N, d_N)\} \\ S(\pi) &= \frac{\prod_{p \in \text{Percentiles}} M_{(p)}}{|\text{Percentiles}|} \end{aligned} \tag{3}$$

Empirically, we have found that optimizing quantiles can provide a smoother distribution.

3.3.3 Incentives. Many cases of market bias boil down to minimizing arbitrage. For example, it might be observed that Buyers find items with low list prices attractive and yet are surprised when confronted with costly shipping. Sellers will often discover these user behaviors can yield a higher collection of clicks, a standard ranking signal, and will list items with artificially low prices to improve their ranking in Search, benefiting themselves at the expense of the marketplace.

To counteract this undesired behavior, we propose a simple maximization approach across the top ranked documents for each query. Given some user-provided query i, document j indicator, $B_{i,j} \in \{0, 1\}$, indicating that a document for a given query exhibits a quality we wish to incentivize, a ranking of documents for query q_i, $r_i = R(\pi, q_i)$, and the number of positions, K, to consider:

$$S(\pi) = \frac{\sum_{i=1}^{|Q|} \sum_{p=1}^{K} B_{i,r_{i,p}}}{K|Q|} \tag{4}$$

Combined with Weighted Importance Ranking in Equation 2, we can influence desired behaviors conditioned on relevance and group-level diversity.

3.3.4 Inequality. The final class of market level metrics fall under the guise of *inequality*: some imbalance in "wealth" distribution across tiers of sellers that we wish to correct. Without interventions, marketplaces often fall under the phenomena of the Matthew Effect; wealth accumulates in a small segment of the population, squeezing out other Sellers. Etsy is no different: Figure 1 shows how visibility accumulates in the wealthy few: the top few% of sellers account for the bulk of sales. Indeed, it can be shown that models that maximizing conditional purchase rate will result in natural monopolies [31] due to social signals present in the marketplace (reviews, best sellers, number of sales, etc.). While beneficial to maximizing sales, it carries inherent business risk: any loss of sellers occupying monopolistic positions for popular queries will have outsize effect on bottom line KPIs.

While work has been done examining how different rankings can impact a market's health (e.g. maximizing number of purchases) [2], there is little discussion on improving proportional representation across tiers of sellers jointly.

Gini Index: Given we can estimate or measure the wealth distribution apriori, we propose minimizing the Gini Index[14], also known as the Gini Coefficient, across the marketplace. Based on the Lorenz Curve, it estimates the difference between full income equality and the actual observed wealth distribution of different sub populations. In a completely equal world, the wealth distribution matches the proportion of the population. For example, if new sellers comprised 10% of the total sellers on Etsy, they should account for 10% of total sales. Conversely, as one sub population starts to disproportionately earn more than other groups, the Gini subsequently measures the market as more unequal. Given a cumulative proportion of the population, X_i and a cumulative proportion of wealth W_i, ordered such that $w_i/x_i \leq w_{i+1}/x_{i+1}$ we can define the Gini Index indicator:

$$\begin{aligned} \text{Gini} &= 1 - \sum_{i=2}^{|\mathcal{Q}|} (X_i - X_{i-1})(\mathrm{W}_i + \mathrm{W}_{i-1}) \\ S(\pi) &= 1 - \text{Gini} \end{aligned} \tag{5}$$

There are a number of different ways to compute the population: total count of different sub populations across all inventory, such as seller country origin, the subset represented in the train dataset, etc. Similarly, there are many ways we can measure wealth; for E-commerce businesses, purchase count per query can be a reasonable approximation for wealth. Given some function IsSubPop:

$$w_i = \sum_{j=1}^{|\mathcal{Q}|} \text{IsSubPop}(\pi, q_j, i) \cdot Purchases(q_j)$$

In the case where we are more interested in the traffic distribution, W can be set to the query volume. For simplicity, in our experiments, we compute Gini for rank=1. However, computing the Gini over multiple rank positions is similarly straight forward when we have access to the observation probability O of a document at position p (such as from a click model):

$$\begin{aligned} O(\text{Pos}) &\in [0, 1] \\ S_{\text{pos}}(\pi) &= O(p) * \text{Gini} \end{aligned} \tag{6}$$

4 ALGORITHM OVERVIEW

Learning a ranking policy requires multiple levels of information: individual scores in the case of relevancy, group-level metrics for diversity, and market-level measurements to account for a variety of skews. Further difficulty arises from the large number of rank orderings: many metrics are neither continuous or differentiable.

To optimize the policies, we compose a linear combination of all metrics (e.g. NDCG, incentives, Gini, etc) into a final *Fitness Function*, which our proposed optimizer tries to maximize.

$$F(\pi) = \frac{\sum_{i=1}^{|S|} W_i \cdot S_i(\pi)}{\sum_{i=1}^{|S|} W_i} \tag{7}$$

Below we describe a greedy, group-level policy optimizing either a static or stochastic value function followed by a proposed optimizer to learn the policies.

4.1 Greedy Algorithm

As has been shown through numerous previous works[26], the assumption of independence during prediction is violated when considering group-level diversity. Consequently, most work on diversity utilize a heuristic, second pass algorithm to select subsequent documents during the ranking process. Zhai et al. showed that a simple greedy algorithm performed well in maximizing MMR[34] (Algorithm 1) which selects documents one at a time such that they maximize the desired diversity function:

Algorithm 1: Greedy Algorithm

```
Input: parameters σ, value function v, documents D
Output: ⟨d_1, d_2, ..., d_k⟩
for i = 1, 2, 3, ..., K do
    d_i = arg max_{d∈D} v(σ, d_i, ⟨d_1, d_2, ..., d_{i−1}⟩)
    D = D − {d_i}
end
```

4.1.1 Static Value Functions. While the greedy policy classically focuses on a heuristic document similarity as the selection criteria for the value function, we instead learn a parameterization over the greedy algorithm and utilize a simple average over the features of previously selected documents to represent aggregate state.

We define a static value function as:

$$\begin{aligned} s_1 = 0, s_i &= \frac{\sum_{j=1}^{i-1} Feats(d_i)}{i-1} \\ v(\sigma, d_i, \langle d_1, .., d_{i-1}\rangle) &= \Phi(\sigma, s_{i-1} - Feats(d_i)) \end{aligned} \tag{8}$$

where Φ is a fully connected neural network.

4.1.2 Stochastic Value Functions. In this section we introduce a stochastic value function, SVF for short. Queries are historically assumed independent much the same way that the PRP assumes documents can be arranged independently of each other. However, it's easy to see how this assumption is violated.

Revisiting the previous example of new sellers and power sellers, we present a set of rankings from two policies π_1 and π_2:

$$Q1_{\pi_1} = \{P, N, P, N\}, Q2_{\pi_1} = \{P, N, P, N\}$$
$$Q1_{\pi_2} = \{N, P, N, P\}, Q2_{\pi_2} = \{N, P, N, P\}$$

It is clear that the group level diversity metrics for each policy are optimal given these two seller categories, but also equally clear that the diversity of sellers occupying the first position is poor. In expectation, one can also see how blending the two policies would result in the highest overall reward:

$$\pi = \arg\max_{p \in \{\pi_1, \pi_2\}} \text{Uniform}(0, 1)$$

Inspired by the observation that neural networks can be viewed as an exponential set of sub-networks [29] and the above observation, we introduce a simple stochastic feature into the network to allow for the blending of learned sub-policies:

$$f \sim \text{Uniform}(0, 1)$$
$$v(\sigma, d_i, \langle d_1, .., d_{i-1} \rangle) = \Phi(\sigma, (s_{i-1} - Feats(d_i)) \oplus f) \quad (9)$$

Rather than break apart θ into explicit policy sets, we rely on feature masking in 2 to produce thinned, sub-networks for optimization, deferring to the optimizer to learn how best to incorporate the noise.

4.2 Evolutionary Strategies

Learning a policy that devolves into various forms of sorting is difficult; given slight variations to the underlying parameters can result in large swings in scores. To solve this challenge, we reach for recent work using ES from the LTR and Reinforcement Learning space to maximize our desired objectives. Intuitively, you can view ES as a crowd sourcing algorithm, as we explore an area in the weight space by looking at different gradients from the current location, then combine those gradients based on their fitness score to determine our next location. Below we provide a high level description of the $(1 + \lambda)$ variety of ES.

Given a parameter set θ (henceforth known as the parent), a fitness function f, and shaping function H, we sample λ search gradients from the Normal distribution: $i \in \{1, 2, .., \lambda\}, \epsilon_i \sim \mathcal{N}(\mu, I)$ where $\mu = 0$ and $I = 1$ are typical parameters for the noise distribution; while both Salimans and Chrabaszcz explored adjusting I, neither found it significantly changed the results. We further augment the algorithm by masking parameters with probability (p), reducing the effective search space per pass [16]. For each search gradient, we compute its fitness with respect to the parent $\text{Fitness}_i = f(\theta_{\text{parent}} + \epsilon_i)$. We proceed to run all scores through a shaping function which scales each gradient by some rank function to smooth out the impact of outlier fitness scores: $\epsilon'_i = \epsilon_i * H(\text{Fitness}_i, \text{Fitness}_*)$. We compute our candidate parent as the sum of gradients scaled by σ and compare it to the previous parent, replacing the parent if the candidate improves.

$$P = \{\theta_{\text{parent}}, \theta_{\text{parent}} + \sigma * \sum_{i=1}^{\lambda} \epsilon'_i\}$$
$$\theta = \arg\max_{p \in P} F(p)$$

There are a few variations which are commonly used. The first is always updating $\theta_{\text{parent}} = \theta_{\text{candidate}}$ regardless of improvement of fitness. The second is with respect to the shaping function: Wierstra et al. explored the impact of fitness shaping functions in Natural Evolutionary Strategies [32] and found that so long as they were monotonic with respect to utility rank, they improved the robustness. The final one is the number of Search gradients used during candidate construction, which also correspond to the theoretical underpinnings: Salimans et al. used all gradients as part of its computation due to assumptions made in NES whereas [9] implement a canonical variant which only uses the best μ children.

We consolidate all of these variants into the following generalized $(1 + \lambda) - ES$ algorithm in 2.

Algorithm 2: Generalized $(1 + \lambda)$-ES

Input: θ_0 - parameters, F - fitness function, H - shaping function, (p) - mask probability, update $\in \{True, False\}$, $\lambda \in \mathcal{I}^+$, $\mu \in \mathcal{I}^+$, iters $\in \mathcal{I}^+$
for *i = {1, 2, .., iters}* **do**
  for *c = {1..λ}* **do**
    $\epsilon_c \sim \mathcal{N}(0, 1) \cdot \text{Bern}(p)$
    $s_c = F(\theta_{i-1} + \epsilon_c)$
  end
  $Sort(\epsilon_*, s_*)$ in non-increasing order $\{s_1 \geq s_2 \geq ..s_\lambda\}$
  $\theta_{candidate} = \theta_{i-1} + \sum_{j=1}^{\mu} \epsilon_j \cdot H(s_j, \{s_1, s_2, .., s_\mu\})$
  if *update* **then**
    $\theta_i \leftarrow \theta_{candidate}$
  else
    $P = \{\theta_{i-1}, \theta_{candidate}\}$
    $\theta_i \leftarrow P_{\arg\max_{p \in P} F(p)}$
  end
end

We add a few additions on top of the Generalized ES algorithm. First we observe that the Greedy policy complexity is $O(N^2)$, leading to slowdown on large document sets during training. To mitigate this, we uniformly sub-sample the document set each pass, for each query. While left as a hyper parameter, we found setting it to twice the K value used to compute NDCG sufficient for convergence. Secondly, we used mini-batching instead of optimizing the entire dataset for each search gradient, significantly speeding up the algorithm. Finally, we utilize the the masking strategy as in ES-Rank[16] except that we sample from the Bernoulli distribution with some probability (p).

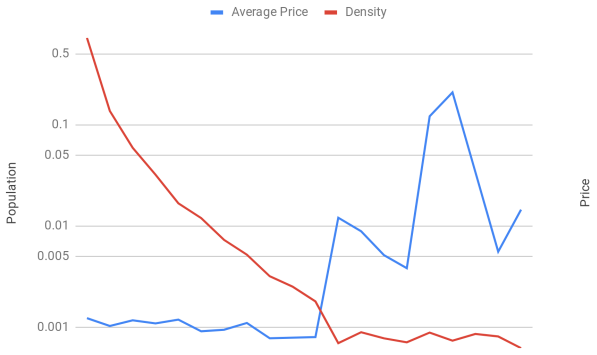

Figure 2: Due to the inverse relationship between population density and average prices, optimizing toward greater equality will squeeze out higher priced sellers

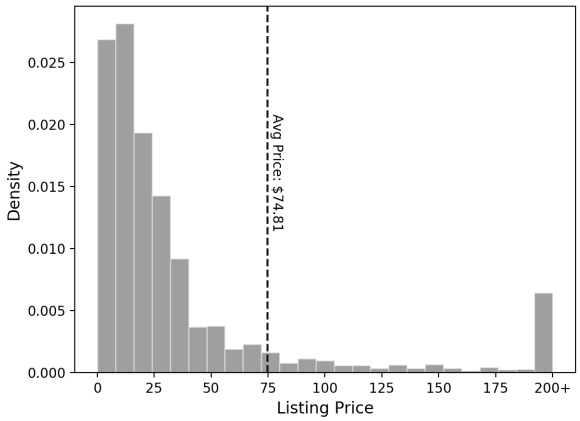

Figure 3: Prices in the market place skew heavily left. Incentivizing higher priced items benefits shops selling more labor intensive, higher quality items as well as less price-sensitive buyers.

5 EXPERIMENTS

In this section, we first discuss the data used in experiments in §5.1 and then we discuss our baselines and chosen implementation details in §5.2 and finally we analyze results in §5.3.

5.1 Data

Our training set is extracted from production search logs, containing the top 5,000 queries observed over an eight day period, with documents human annotated with $R \in \{1, 5\}$. While generally head and torso queries, they exhibit the properties we wish to test while ensuring reasonable seller competition: large shifts in query importance (due to pronounced power law distribution), wealth inequality, and varied topicality (significant volume is dedicated to occasion gifting, such as wedding gifts). We use over 200 features from our production ranking systems which are a combination of query relevancy, historical performance, and taxonomic information. We further augment the dataset with a variety of metadata:

Population, Wealth, Observation: Seller population scores are based on GMV vigintiles: Etsy, like most E-commerce sites, exhibits a heavy power law distribution with respect to dollar shares. We compute the observation model by summing the number of purchases over a week at each rank position and dividing by the total number of purchases observed. Wealth is calculated by summing the total number of purchases per query over the same week as the observation model.

Diversity: Listings taxonomy is used as the source of diversity in queries. Overall, it has 175 different categories with a large class imbalance.

Incentives: We binarize our product prices by the mean listing value in our sampled search results, factoring in the importance of high priced outliers such as Rolexes. Price has a sharp left skew (3) toward lower cost items, often times washing out higher quality, more labor intensive products. We add an incentive toward premium prices to boost high quality listings closer to the top of the rankings.

Gini vs Price: As can be seen in figure (2), listing price and population counts are inversely correlated, illustrating a common problem faced by E-commerce: improving the Gini Index naturally results in a reduction in listing price.

5.2 Algorithms

Variant	Relevance	Group Diversity	Gini Index	Incentive
-{0,0}	1.00	0.00	0.00	0.00
-{0.05,0.05}	0.85	0.05	0.05	0.05
-{0.1,0.1}	0.70	0.10	0.10	0.10
-{0.17,0.17}	0.49	0.17	0.17	0.17
-{0.25,0.25}	0.25	0.25	0.25	0.25
-{0.3,0.3}	0.10	0.30	0.30	0.30
-{0.05,0}	0.90	0.00	0.05	0.05
-{0.1,0}	0.80	0.00	0.10	0.10
-{0.25,0}	0.50	0.00	0.25	0.25
-{0.33,0}	0.33	0.00	0.33	0.33
-{0.4,0}	0.20	0.00	0.40	0.40

Table 2: Weighted fitness function weights to understand how tradeoffs in importance lead to different outcomes

Baselines: We compare two different baselines to affirm efficacy of our approach. We first look at the venerable Maximal Marginal Relevance [5] where the relevance model is learned using LambdaMART [4] optimized for `NDCG@10`. Document similarity is based on the Jaccard of each document's taxonomy. We tune the blend parameter λ by maximizing the weighted sum of scores between all indicators. The second model is optimized using the same generalized ES algorithm above, however we replace the greedy policy with a standard pointwise inference policy, sorting documents based on their learned scores.

Evolutionary Strategies: In our experiments we compare ES policies trained against a variety of different metrics:

- Relevancy scores measured as `NDCG@10`

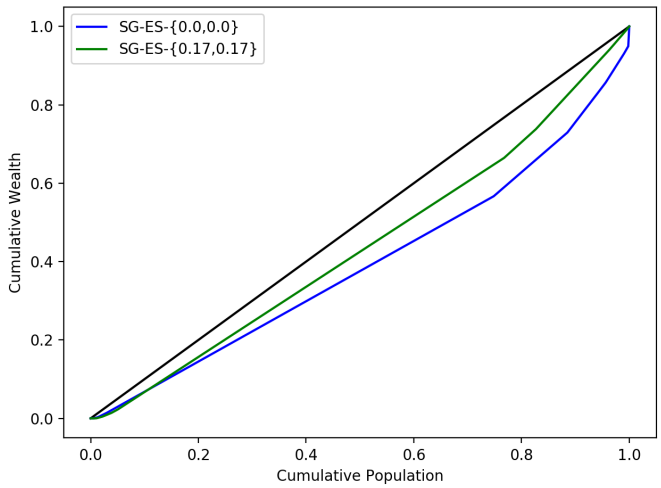

Figure 4: Stochastic greedy policies are able to shift the Lorenz Curve towards more equality when incentivized.

- Groupwise diversity using `ERR-IA@10` (1) across different taxonomic groups
- Market indicators: Gini_1 and Incentives_1 (4)

We combine these into our fitness function, F, via a weighted linear combination in EQ 7 with weights described in Table 2. We explore two different policies: a standard point-wise baseline and a greedy policy (1). We use the generalized form of Canonical Evolutionary Strategies, fixing $\lambda = 768$ and $\mu = 50$ for all tests. We set *update* to True and fix (p) at 0.05. We compare three different variations of our proposed models: Greedy with a static value function, Greedy with a stochastic value function, and Pointwise with a stochastic value function. For Φ, we use a small fully connected neural net $(20 \rightarrow 20 \rightarrow 1)$ utilizing the ReLU[17] non-linearity. When utilizing stochastic value functions, we evaluate the test set 5 times with different random seeds to determine expected performance.

5.3 Experimental Analysis

To understand how different weights impact the overall model, we first examine how market indicator constraints impact relevancy across the different policies. We follow it up by examining to what degree relevancy, group-level, and indicators can jointly optimize all metrics simultaneously. We finally compare stochastic vs static policies and to what degree it improves the model. All results are verified offline against a held-out test set.

Comparison to Baselines: Unsurprisingly, we see in Table 3 that while MMR does best with respect to ERR-IA and NDCG, it is at the expense of both Incentives and Gini Index which it is unable to optimize. Similarly, while the Pointwise-ES baseline does well on Gini and Incentives, it performs worse of all the policies on group level diversity, which is not surprising due to document independence assumptions. Importantly, forall ES policies were able to optimize for all metrics compared to the baseline, providing evidence of its efficacy.

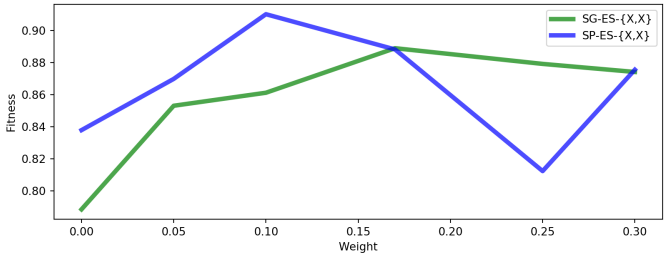

Figure 5: Gini - SG-ES vs. SP-ES

Table 3: Policy Comparison of Baselines Across All Metrics

Variant (2)	Metric	Validation	Test Mean	Test Std
MMR-LambdaMART	ERR-IA	0.487	0.480	-
P-ES-{0.17,0.17}	ERR-IA	0.481	0.467	-
SP-ES-{0.17,0.17}	ERR-IA	0.484	0.468	0.001
G-ES-{0.17,0.17}	ERR-IA	0.489	0.475	-
SG-ES-{0.17,0.17}	ERR-IA	0.478	0.471	0.002
MMR-LambdaMART	Gini	0.795	0.800	-
P-ES-{0.17,0.17}*	Gini	0.925	0.891	-
SP-ES-{0.17,0.17}*	Gini	0.894	0.888	0.029
G-ES-{0.17,0.17}*	Gini	0.883	0.881	-
SG-ES-{0.17,0.17}*	Gini	0.911	0.889	0.011
MMR-LambdaMART	Incentive	0.396	0.403	-
P-ES-{0.17,0.17}*	Incentive	0.466	0.525	-
SP-ES-{0.17,0.17}*	Incentive	0.466	0.518	0.002
G-ES-{0.17,0.17}*	Incentive	0.470	0.543	-
SG-ES-{0.17,0.17}*	Incentive	0.466	0.543	0.009
MMR-LambdaMART	NDCG	0.692	0.679	-
P-ES-{0.17,0.17}	NDCG	0.655	0.637	-
SP-ES-{0.17,0.17}	NDCG	0.652	0.634	0.001
G-ES-{0.17,0.17}	NDCG	0.662	0.651	-
SG-ES-{0.17,0.17}	NDCG	0.652	0.642	0.001

* indicates stat. sig. compared to MMR (P < 0.005).

Influence on Market Indicators: We evaluated variants by adjusting the importance weight for the Gini Index and Incentives in 5. Compared to the baselines, we were able to progressively and smoothly improve both the Gini Index and Incentive indicators with both stochastic and static Greedy variants. We find that weight does indeed improve the overall equality of the system compared to its unconstrained form 4.

Table 4: Gini Index: Stochastic vs. Static (Variant ES-{0.4,0})

Variant	Validation	Test Mean	Test Std
G-ES-{0.4,0}	0.906	0.798	0.000
SG-ES-{0.4,0}	0.922	0.903	0.008

Joint Optimization of all Metrics: We found we were able to optimize multiple metrics simultaneously as seen in Figure 6. Despite the conflicting nature of the metrics, ES was able to find policies that improved the underlying metrics compared to the baselines.

Stochastic Features: We find that the additional noise had a few benefits: first, it helped regularize the networks; we find the difference between train and test scores were narrower than the static variants. Furthermore, the stochastic variants were smoother: they had lower fitness variance as importance weighting increased. Table 4 exemplifies the differences when considering only market indicators and relevance. Table 5 shows how the stochastic policy is more

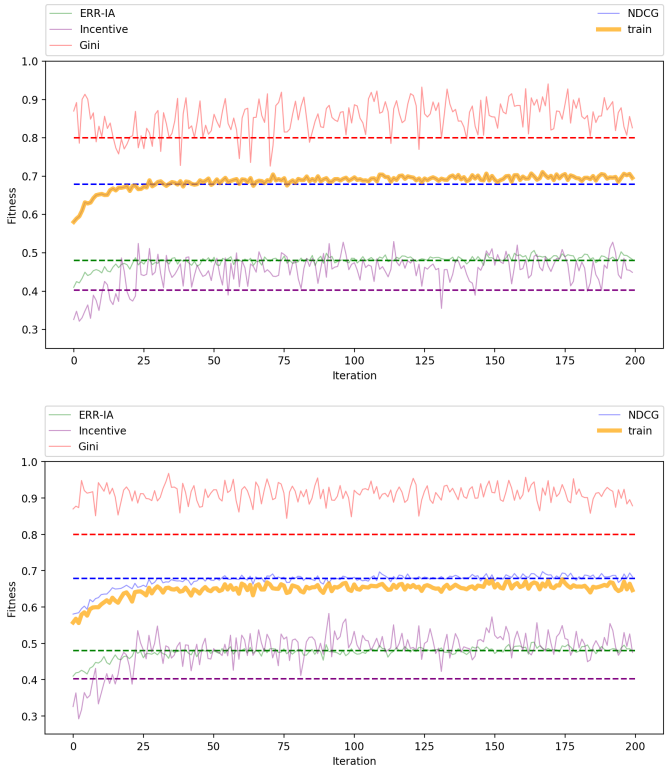

Figure 6: Fitness scores from each iteration - SG-ES-{0.0,0.0} (top) and SG-ES-{0.17,0.17} (bottom). The horizontal dashed lines display the baseline MMR-LambdaMART scores for each metric matched by color.

reliable with generalization from validation to the test set. Unlike the greedy variant, SVFs applied to pointwise models were unable to stably improve market indicators. Figure 5 compares stochastic pointwise and greedy algorithms on different importance weightings - while greedy is fairly smooth, the pointwise model displays high variance both within policy and across different weightings.

Table 5: Gini Index: Stochastic vs. Static

Variant (2)	Validation	Test Mean	Difference
G-ES-{0,0}	0.740	0.849	-0.109
SG-ES-{0,0}	0.734	0.789	-0.065
G-ES-{0,0.05}	0.815	0.883	-0.068
SG-ES-{0,0.05}	0.854	0.853	0.001
G-ES-{0,0.1}	0.825	0.899	-0.074
SG-ES-{0,0.1}	0.881	0.861	0.020
G-ES-{0,0.17}	0.883	0.881	0.002
SG-ES-{0,0.17}	0.911	0.889	0.012
G-ES-{0,0.25}	0.924	0.861	0.063
SG-ES-{0,0.25}	0.901	0.879	0.022
G-ES-{0,0.3}	0.945	0.876	0.069
SG-ES-{0,0.3}	0.925	0.874	0.051

6 CONCLUSION

In this paper we defined types of market indicators critical for creating healthy, two-sided marketplaces and proposed strategies for learning policies to jointly maximize those desired market characteristics. We showed that we can influence these market-level metrics via trained policies, evaluated on offline data, resulting in a method for imposing business needs and eliminating many of the common forms of interventions that lead to sub-par search experiences.

We further release our production code, capable of scaling to tens of millions of examples, on github[1] to ensure both reproducible results and provide a flexible framework for solving real-world production ranking problems. We endeavor to also publish our training datasets to establish more modern baselines in the difficult E-commerce space.

While the results hold for popular head queries in an offline environment, follow-up work is needed show the effects translate into online systems. Future work also intends to focus on alternative models beyond our greedy, neural network approach: as our optimizer is shown capable of learning non-differentiable models, there is opportunity to explore more efficient greedy architectures. As it relates to the types of objectives, additional work incorporating supply and demand signals in international markets is of significant interest. Similarly, the findings from Moshary[21] and Blake[3] are likely to be fertile grounds for future direction. Finally, while the intent of this work is to understand the impact of a standalone ranking policy, there is likely significant opportunities in ensembling expert models within the framework to boost underlying metric performance.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*. ACM, 5–14.
[2] Franco Berbeglia and Pascal Van Hentenryck. 2017. Taming the Matthew Effect in Online Markets with Social Influence.. In *AAAI*. 10–16.
[3] Thomas Blake, Chris Nosko, and Steven Tadelis. 2016. Returns to consumer search: Evidence from ebay. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. 531–545.
[4] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.
[5] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.
[6] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-Objective Ranking Optimization for Product Search Using Stochastic Label Aggregation. In *Proceedings of The Web Conference 2020*. 373–383.
[7] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval* 14, 6 (2011), 572–592.
[8] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 621–630.
[9] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. 2018. Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari. *arXiv preprint arXiv:1802.08842* (2018).
[10] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 659–666.
[11] W.S. Cooper. 1971. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. (1971).

[1] https://github.com/etsy/Evokit

[12] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.

[13] Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 65–74.

[14] Corrado Gini. 1912. Italian: Variabilità e Mutabilità (Variability and Mutability). *Cuppini, Bologna* (1912).

[15] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application. *arXiv preprint arXiv:1803.00710* (2018).

[16] Osman Ali Sadek Ibrahim and Dario Landa-Silva. 2018. An evolutionary strategy with machine learning for learning to rank in information retrieval. *Soft Computing* 22, 10 (2018), 3171–3185.

[17] Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, et al. 2009. What is the best multi-stage architecture for object recognition?. In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2146–2153.

[18] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 41–48.

[19] Joel Lehman, Jay Chen, Jeff Clune, and Kenneth O Stanley. 2018. ES is more than just a traditional finite-difference approximator. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 450–457.

[20] Michinari Momma, Alireza Bagheri Garakani, and Yi Sun. 2019. Multi-objective Relevance Ranking. In *Proceedings of the SIGIR 2019 Workshop on eCommerce (ECOM 19)*.

[21] Sarah Moshary, T Blake, K Sweeney, and S Tadelis. 2017. Price salience and product choice. (2017).

[22] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*. ACM, 784–791.

[23] Karthik Raman, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Learning to diversify from implicit feedback. In *WSDM Workshop on Diversity in Document Retrieval*.

[24] Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* 33, 4 (1977), 294–304.

[25] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. 2017. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864* (2017).

[26] Rodrygo LT Santos, Craig Macdonald, Iadh Ounis, et al. 2015. Search result diversification. *Foundations and Trends® in Information Retrieval* 9, 1 (2015), 1–90.

[27] Rodrygo LT Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit search result diversification through sub-queries. In *European Conference on Information Retrieval*. Springer, 87–99.

[28] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*. 5427–5437.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[30] Krysta M Svore, Maksims N Volkovs, and Christopher JC Burges. 2011. Learning to rank with multiple objective functions. In *Proceedings of the 20th international conference on World wide web*. ACM, 367–376.

[31] Pascal Van Hentenryck, Andrés Abeliuk, Franco Berbeglia, Felipe Maldonado, and Gerardo Berbeglia. 2016. Aligning Popularity and Quality in Online Cultural Markets.. In *ICWSM*. 398–407.

[32] Daan Wierstra, Tom Schaul, Jan Peters, and Juergen Schmidhuber. 2008. Natural evolution strategies. In *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*. IEEE, 3381–3387.

[33] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 113–122.

[34] ChengXiang Zhai, William W Cohen, and John Lafferty. 2015. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR Forum*, Vol. 49. ACM, 2–9.