Algorithmic Risk Aversion and Recommendation-Mediated Demand*

Andreas Haupt† Aroon Narayanan‡

March 21, 2025

Abstract

Recommendation algorithms shape market outcomes by selecting which offers to make salient to consumers. Such algorithms learn to recommend from repeatedly interacting with consumers, using *online learning algorithms*. This paper presents an algorithmic property with important impacts on market outcomes. We show that algorithms are sensitive not only to the quality of observable measures of match quality, but also noise. Concretely, we show that a popular sequential learning mechanism sharing features with current recommendation systems, ε-Greedy, exhibits *emergent risk aversion*. The algorithm prefers actions with less volatile feedback. When presented with actions of the same expectation, under a wide range of conditions, ε-Greedy chooses the lower-variance action with probability approaching one. The risk aversion is a result of asymmetric exploration in the algorithm's learning. We show that optimistic algorithms and a particular reweighting of past data are risk-neutral algorithms.

1 Introduction

In many settings in the digital economy, a recommendation algorithm shapes which product is highlighted to a user. This paper shows one channel of how *feedback*, that is, measurement of user-product match, affects which product is highlighted to a user under realistic recommendation algorithms.

Many algorithms for solving problems in recommendation keep an estimate of how well a recommendation has performed in the past. While most deployed algorithms are guaranteed to recommend optimally in expectation in the limit (the *no-regret* criterion, see Lattimore and Szepesvári (2020)), as we show, they are sensitive, and *averse* to noise, without being specifically defined to have this property. Risk aversion for two products to highlight can be of the following form: For a sequence of one-off interactions with consumers, either highlight product A and get a match estimate normalized to zero, or highlight product B and get a stochastic measurement of 1 or -1, with equal probability. At any point in time, contingent on past observations, the algorithm chooses to highlight a product. What is the probability that the learning algorithm recommends each product after t rounds of interaction? A risk-averse recommendation algorithm chooses the product with perfectly measured match value more often than the other; a risk-neutral learning algorithm chooses them with equal probability. We prove that a classic algorithm, ε-Greedy, is risk-averse. This property holds without ε-Greedy being explicitly designed to be risk-averse. It is an emergent property of the algorithm.

Risk aversion has important impacts in the digital economy beyond. In particular, if some features of feedback are easier to measure, e.g., because data is more accurate, for an offer of a platform, yet has the same *expected* value, risk averse algorithms will favour the platform's own brand, and hence leading to another channel of self-preferencing, see, *e.g.*, Musolff 2022; Hartzell and Haupt 2025. They may also systematically favour incumbents, if incumbent match value is more accurately measured.

Algorithmic risk aversion is important beyond the digital economy. For example, consider a firm making credit decisions using a risk-averse algorithm. Underrepresented minorities often have wide variances in their credit scores, shaped partly by historic inequities in access to good credit opportunities:

*We thank seminar audiences at Harvard and MIT and the mathoverflow.com user fedja for helpful conversations.

†Stanford University

‡Massachusetts Institute of Technology

> *As the white suburbs and black inner cities diverged in their mortgage access, two different credit markets emerged in both zones. Lower-risk mortgages led to higher wealth and stability in the white suburbs. These conditions also led to a healthy consumer credit market. In the redlined black ghettos, the economic climate was radically different.* (Baradaran 2018, p. 893)

When faced with minority applicants with higher variability in credit history, a risk-averse algorithm may decide to systematically deny them *even if* it would have approved privileged applicants with similar expected repayment probability but features that are correlated with less variability in credit repayment. They hence perpetuate centuries of iniquity. Yet another setting in which risk aversion can play a significant role is recommendation systems. The choice of recommended products and search results are determined by how valuable the recommendation is deemed to be. Here, too, risk aversion can lead to the recommendation system suppressing "noisier" content—which, in most cases, will be the less mainstream, more marginalized content—even when its deployers do not find such bias desirable. In the long run, this bias can also lead to the homogenization of the content on these platforms, as more divergent content is not recommended as frequently by the algorithm.

Our first formal result is that the ε-Greedy algorithm exhibits risk aversion, preferring deterministic over non-deterministic actions of the same variance. The intuition for the emergence of risk aversion lies in the way ε-Greedy estimates the payoff of each action. If an algorithm's estimate of each action is the simple average of the observed payoffs in the past from this action, which is what ε-Greedy does, its estimates will be biased because the algorithm undersamples actions after low payoff realizations. We then discuss two corrections to the algorithm that enable it to be risk-neutral. Our first correction, which we call the *Reweighted ε-Greedy*, counters the undersampling propensity by adjusting its estimate to account for the probability with which an action was chosen. We show that Reweighted ε-Greedy is risk neutral. We also propose another correction for a broader class of settings: the *Optimistic ε-Greedy*. It adds an optimism term to the estimate that corrects bias asymptotically, see (Auer 2002; Auer, Cesa-Bianchi, and Fischer 2002; Lattimore and Szepesvári 2020). Our third formal result shows that this correction also makes ε-Greedy risk neutral. We use simulations to explore the necessity of conditions we make in our theoretical analysis and the transient persistence of risk behavior even with unequal expected values for the actions.

1.1 Related literature

There is a large literature on learning from feedback in economics. We highlight the papers Bolton and Harris (1999), Keller, Rady, and Cripps (2005), Klein and Rady (2011), Baek and Farias (2021) as theoretical contributions, Bergemann and Välimäki (2017) is a survey. The paper Bardhi, Guo, and Strulovici (2020) demonstrates that even arbitrarily small differences in early-career discrimination can be highly consequential later in life. Our results complement this literature by showing that algorithmic learning can exhibit unintended discrimination with strong consequences in the long run. A second branch of literature studies learning by economic agents in empirical settings. Farber and Gibbons (1996), Altonji and Pierret (2001), and Baek and Makhdoumi (2023) study learning by employers, providing testable predictions for wage dynamics. Crawford and Shum (2005) applies learning to demand for pharmaceutical drugs. Recently several papers have considered the behavior of learning algorithms in simulations, particularly in relation to collusion. Calvano et al. (2020), Musolff (2022), Brown and MacKay (2023), and Banchio and Mantegazza (2022) find in different game-theoretic settings that pricing algorithms learn to play collusive equilibria, raising antitrust concerns about the use of such algorithms in pricing. Banchio and Mantegazza (2022) shows that in games, spontaneous collusion can arise because of correlation of play and asymmetric sampling. The intuition presented here relies on a similar causal channel—asymmetric sampling—but considers different algorithm classes and stresses synchronization as opposed to distributional questions. It also considers a setting with a single algorithm making decisions, as opposed to multiple algorithms interacting in a game.

Our paper is connected to the computer science literature on the effect of biased payoff estimates in recommendation systems. Marlin and Zemel (2009) observes that online learning in recommendation systems leads to confounding of average user scores in recommendation systems and proposes algorithmic interventions to correct this bias. Chaney, Stewart, and Engelhardt (2018) proposes a model of recommendation and shows that recommendation systems' biased estimates of user preferences can increase homogeneity and decrease user utility. Our study focuses on the effect of noise and the propensity of taking particular

actions and does not directly consider the bias in estimates.

We also relate to the study of fairness in bandit problems. While Joseph et al. (2016) considers fairness (which is a finite-time variant of our notion of risk neutrality) as a constraint for algorithm design and constructs algorithms that approximately satisfy it, this paper provides evidence on the risk preferences of an existing algorithm, ε-Greedy, and proposes two ways to mitigate risk preferences; see also Patil et al. (2021) and Liu et al. (2017) for treatments of fairness in bandit problems. Dai et al. (2024) considers the relationship of sampling rates to the regret of algorithms and provides improved regret bounds for algorithms that are sampling actions in a more balanced fashion.

Our work also relates to the study of notions of "rationality" for algorithms (Raman et al. 2024; Rahwan et al. 2019). This literature aims to understand in which environments algorithms behave according to behavioral axioms that were developed for humans.

Finally, we relate to the regret analysis of bandit algorithms under diffusion scaling. Kalvit and Zeevi (2021a) studies this for the Upper Confidence Bound algorithm (compare Theorem 3). Fan and Glynn (2021) derives the limit action distribution of Thompson sampling as a solution to a random ordinary differential equation.

1.2 Outline

The structure of the rest of this paper is as follows. In Section 2, we introduce our online learning setup and our definition of risk aversion, along with formal definitions of our algorithms. The result on ε-Greedy's risk aversion is presented in Section 3. We discuss two corrections of risk aversion in Section 4. We complement our theoretical analysis with simulations in Section 5. We conclude in Section 6. An appendix contains additional simulations.

2 Model

In a bandit problem, a decision maker repeatedly takes an action from a finite set A, $|A| < \infty$. Each action a is associated to a sub-Gaussian distribution $F_a \in \Delta(\mathbb{R})$, $a \in A$ with expectation μ_a and variance proxy σ_a^2.[1] An algorithm π generates (potentially random) sequences of *actions* $(a_t)_{t\in\mathbb{N}}$ and *rewards* or *payoffs* $(r_t)_{t\in\mathbb{N}}$. For each $t \in \mathbb{N}$, repeatedly, the algorithm chooses an action $a_t \sim \pi_t$ and receives a reward $r_t \sim F_{a_t}$. That is, an *algorithm* is a function $\pi\colon \bigcup_{t=1}^{\infty}(A\times\mathbb{R})^t \to \Delta(A)$. We denote action-reward histories by $(a_{1:t}, r_{1:t})$ and the probability that action $a \in A$ is chosen in round t by $\pi_{at} = \pi(a_{1:t}, r_{1:t})_a$. Denote $N_a(t) \coloneqq |\{1 \le t' \le t : a_t = a\}|$ the number of times action a has been chosen up to time t.

The main concept in this paper is a notion of *risk aversion* of algorithms. An algorithm is risk-neutral if it chooses (asymptotically) uniformly from amongst actions of equal expectation. In the long run, risk-averse algorithms prefer less risky actions than others of the same expectation. In the extreme case where the algorithm exclusively chooses (asymptotically) the least risky action among those of the same expectation, we call them *perfectly* risk averse.

Definition 1. We call π *risk-neutral* if for any actions $a, a' \in A$ such that $\mu_a = \mu_{a'}$,

$$\lim_{t\to\infty} \mathbb{P}[a_t = a] = \lim_{t\to\infty} \mathbb{P}[a_t = a'].$$

We call an algorithm *risk-averse* if for all $a, a' \in A$ such that $\mu_a = \mu_{a'}$ and $F_{a'} \prec_{\text{SOSD}} F_a$, it holds that[2]

$$\lim_{t\to\infty} \mathbb{P}[a_t = a] > \lim_{t\to\infty} \mathbb{P}[a_t = a'].$$

An algorithm is *perfectly risk averse* if for any instance for which there is $a \in A$ such that either $\mu_{a'} < \mu_a$ or $F_{a'} \prec_{\text{SOSD}} F_a$ for all $a' \in A \setminus \{a\}$,

$$\lim_{t\to\infty} \mathbb{P}[a_t = a] = 1.$$

[1] A distribution is sub-Gaussian if $\int e^{\lambda X}\,\mathrm{d}F_a(x) \le \exp(\frac{\lambda^2\sigma_a^2}{2})$. In this case, σ_a^2 is called the *variance proxy*.

[2] Distribution F dominates F' in second-order stochastic dominance, $F \succeq_{\text{SOSD}} F'$ if $\int u\,\mathrm{d}F \ge \int u\,\mathrm{d}F'$ for all concave, non-decreasing functions u, with a strict inequality for some such function u.

This paper considers the ε-Greedy algorithm and two variants of ε-Greedy with different statistics.

Definition 2 (ε-Greedy)**.** Let $(\varepsilon_t)_{t\in\mathbb{N}}$ be a $[0,1]$-valued sequence. ε-Greedy chooses the empirically best action with probability $1-\varepsilon_t$, and randomizes between all the actions with probability ε_t, *i.e.*

$$\pi_{at} = \begin{cases} \text{Unif}(\arg\max_{a\in A} \mu_a(t-1)) & \text{w.p. } 1-\varepsilon_t \\ \text{Unif}(A) & \text{w.p. } \varepsilon_t, \end{cases}$$

where $\mu_a(t-1)$ is historical average payoff

$$\mu_a(t) \coloneqq \frac{1}{N_a(t)} \sum_{\substack{1\le t'\le t \\ a_{t'}=a}} r_{t'}.$$

When ε-Greedy takes an action to maximize $\mu_a(t-1)$, we say it *exploits* or *takes an exploitation step*. Otherwise, it *explores* or *takes an exploration step*. We also call $\mu_a(t-1)$ ε-Greedy's *statistic*.

The first variant reweighs data points to change their importance.

Definition 3 (Reweighted ε-Greedy)**.** Reweighted ε-Greedy uses a reweighted payoff estimate as a statistic:

$$\mu_{a,r}(t) = \frac{1}{N_a(t)} \sum_{\substack{1\le t'\le t \\ a_t=a}} \frac{r_{t'}}{\sqrt{\pi_{at'}}}.$$

A second intervention adds an optimism term to the statistic of ε-Greedy.

Definition 4 (Optimistic ε-Greedy)**.** Optimistic ε-Greedy adds an optimism term to its statistic:

$$\mu_{a,o}(t) = \mu_a(t) + \rho\sqrt{\frac{\log(t)}{N_a(t)}}.$$

If action a has not been chosen until time t, $\mu_{a,o}(t) = \infty$.

3 Risk aversion of ε-Greedy

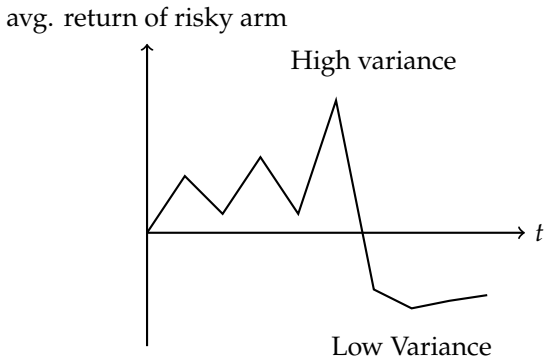
We first show that ε-Greedy is perfectly risk-averse.

Theorem 1. *Let $(\varepsilon_t)_{t\in\mathbb{N}}$ such that $\varepsilon_t \to 0$ and $\sum_{t=1}^{\infty} \varepsilon_t = \infty$. If there is a deterministic, centered dominant action, and all other actions have symmetric continuous distributions, ε-Greedy is perfectly risk-averse.*

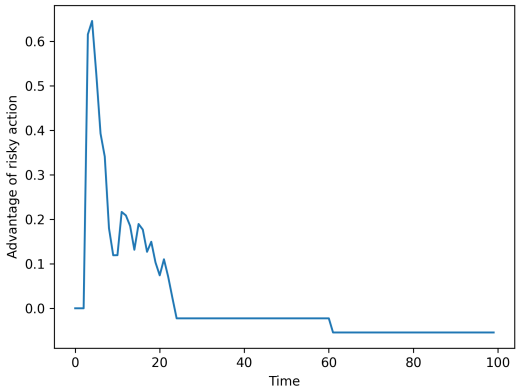
A discussion on the conditions in this result is in order before we move on to the proof. Both hypotheses on the exploration rates are necessary to yield a *no-regret algorithm*, compare Lattimore and Szepesvári (2020), and are standard in the literature. We show in our simulations in Section 5 that relaxing the requirement of determinism of a dominant action leads to risk aversion, but not perfect risk aversion, as does relaxing the symmetry requirement.

The intuition behind this result lies in the sampling bias of ε-Greedy. Upon receiving a low payoff realization for an action, it becomes less likely to choose that action and, hence, less likely to receive data to correct its estimate. This means that it keeps a pessimistic estimate of reward. In contrast, for a high payoff realization, the algorithm frequently samples this action, leading the algorithm to correct its estimate.

Proof of Theorem 1. We first observe that we can restrict to bandit problems of two actions with reward distributions of the same expectation, one deterministic dominant action a and another action a' with a continuous, symmetric distribution. Under the assumptions on the exploration rate made in the theorem, ε-Greedy chooses dominated actions with vanishing probability. We prove this in Lemma 1 in the appendix. As this probability is low, we may consider instances of actions of equal expectation. In addition, a union bound shows that if the probability that the algorithm chooses action a over any single action a' converges to



(a) Illustration of the advantage walk



(b) One realization of the advantage walk for ε-Greedy where the safe action has distribution $\mathbb{1}_{\{0\}}$ while the risky action has distribution $U[-1,1]$

Figure 1: The advantage walk for ε-Greedy. The main intuition for risk aversion in online algorithms is that a random walk with non-uniform variance spends more time in places with lower variance.

1, this implies that this action will be chosen with probability one among all actions of the same probability. Hence, it is without loss to restrict to two-action bandit problems.

We also observe that the result is trivial for two deterministic actions with the same expectation. Hence, we may assume that $A = \{a, a'\}$, $\mathrm{Var}(F_a)$ has a positive variance, and $F_{a'} = \mathbb{1}_{\{0\}}$. Furthermore, it is without loss to assume that the deterministic action is centered: ε-Greedy is invariant to the addition of a constant to all reward distributions As a final reduction step, as $\varepsilon_t \to 0$, it is sufficient to show that it becomes unlikely that ε-Greedy chooses a' in exploitation steps, or

$$\mathbb{P}[\mu_a(t) < \mu_{a'}(t)] \to 0.$$

We express this event as a property of a stochastic process. The sum of the payoffs of the non-deterministic action, denoted by $(X_t)_{t\in\mathbb{N}}$, is a sufficient statistic for the dynamics of the algorithm. $(X_t)_{t\in\mathbb{N}}$ is a lazy random walk starting at the origin, $X_0 = 0$, with transition kernel

$$X_{t+1} = \begin{cases} X_t + r & \text{with probability } (1-\frac{\varepsilon_t}{2})\mathbb{1}_{X_t>0} + \frac{1}{2}\mathbb{1}_{X_t=0} + \frac{\varepsilon_t}{2}\mathbb{1}_{X_t<0}, \\ X_t & \text{else,} \end{cases}$$

where $r \sim F_a$. We call this the *advantage walk* and depict it in Figure 1. Consider the time since the last time that the random walk crossed zero, $\tau_t := t - \max\{1 \le t' \le t | X_{t'-1} \le 0 \le X_{t'}\}$. We claim that $\tau_t \to \infty$ as $t \to \infty$, in probability.

To prove this claim, it suffices to show that for any $c \ge 0$ and $\varepsilon > 0$, there is t' such that $\mathbb{P}[\tau_{t'} \le c] \le \varepsilon$ for all $t' \ge t$. Observe that for any $c \ge 0$, there is $C > 0$ and $t \in \mathbb{N}$ such that for all $t' \ge t$

$$\mathbb{P}[\tau_{t'} \le c] \le \mathbb{P}[|X_{t'-c}| < C] + \frac{\varepsilon}{2} \le \varepsilon.$$

The first inequality is a consequence of the sub-Gaussianity of F_a. The second inequality is a result of $\mathrm{Var}(F_a) > 0$, the conditional independence of increments, and $\sum_{t'=1}^{t-c} \varepsilon_{t'} \to \infty$ as $t \to \infty$.

We also define the distribution of the number of steps taken since τ_t on the positive side. These are distributed as

$$P_t \sim \sum_{t'=t-\tau_t}^{t} Z_{t'}, \quad Z_t \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(1-\varepsilon_t/2).$$

As P_t is a sum of i.i.d. random variables, by Hoeffding's inequality, P_t is close to its expectation $\mathbb{E}[P_t|\tau_t]$ for large enough τ_t. Conditioning on τ_t,

$$\mathbb{E}[P_{\tau_t}|\tau_t] = \sum_{t'=t-\tau_t}^{t} 1 - \frac{\varepsilon_{t'}}{2}.$$

We have

$$\mathbb{P}[X_t > 0] \leq c\mathbb{P}[Y_1, Y_2, \ldots, Y_{P_t} > 0],$$

where $Y_0 = 0$ and $(Y_t)_{t\in\mathbb{N}}$ is a standard random walk with increment distribution F_a. For this inequality, we can choose $c \geq 1/\mathbb{P}[r > X_{t-\tau_t}]$, where $r \sim F_a$ is independent of $(X_t)_{t\in\mathbb{N}}$. This follows as a single step from zero could lead from 0 to $X_{t-\tau_t}$, or a higher value. Because $X_{t'-\tau_{t'}}$ is reached from $X_{t'-\tau_{t'}-1} < 0$, $\mathbb{P}[r > X_{t'-\tau_{t'}}] > 0$ must be positive, and hence c is well-defined.

Hence, for any $\delta > 0$, there is $t' \in \mathbb{N}$ such that for all $t \geq t'$,

$$\mathbb{P}[X_t > 0|\tau_t] \leq \frac{c}{\sqrt{\pi(\sum_{t=t-\tau_t}^{t} 1 - \frac{\varepsilon_{t'}}{2})}}(1+\delta). \tag{1}$$

This inequality uses the well-known property that the probability of a random walk stays positive until time t with probability approximately $\frac{1}{\sqrt{\pi t}}$ (U. Frisch and H. Frisch 1995, Eqn. 35). (This property also holds for Rademacher-distributed increments and is the only place where we use the assumption that our distribution is continuous and symmetric.)

Observe that (1) approaches 0 as $\tau_t \to \infty$ and recall that $\tau_t \to \infty$ as $t \to \infty$, in probability. Given these two facts,

$$\mathbb{P}[X_t > 0] \xrightarrow[t\to\infty]{\mathbb{P}} 0,$$

which concludes the proof. □

We highlight that a similar argument applies to actions that are dominated by others in expectation. For two actions a, a' of the same expectation, such that a is deterministic but a' is not, and a third action a'' such that $\mu_{a''} > \mu_a, \mu_{a'}$, the probability that a is taken conditional a or a' are taken converges to one.

4 Achieving risk neutrality

Next, we propose variants to the ε-Greedy statistic to achieve risk neutrality. These corrections are motivated by the technical analysis of the previous theorem as well as known algorithmic ideas (*e.g.*, Auer (2002)). Our corrections highlight the mechanisms leading to bias, and their redressal.

4.1 A reweighting approach to risk neutrality

The first approach stems from our analysis of variance: we should reweight data to achieve risk neutrality. A reweighted ε-Greedy provably is risk-neutral if exploration is sufficiently high.

Theorem 2 (Reweighted ε-Greedy)**.** *Let $\varepsilon_t = t^{\frac{1}{2}+\kappa}, \kappa \in (0, \frac{1}{2})$. Reweighted ε-Greedy is risk-neutral for two centered actions, one of which is deterministic.*

The reweighting proposed here means that payoffs resulting from currently unfavored actions are weighted more highly in the statistic of the algorithm. In the credit scoring setting, if the option of rejecting the loan is currently favored, the outcome of any credit that is given (due to exploration) is weighted more

highly in the statistic. Thus, this correction tells us that we should assign more importance to outcomes that resulted from choices that seemed apriori less attractive.

It is worth noting that this reweighting does *not* lead to an unbiased estimator of action rewards. We show in simulations in an appendix that an unbiased estimator leads to a risk-loving algorithm, Section 8. This is a result of what the rescaling does to the algorithm's statistic: The goal of the algorithm proposed here is to equalize variance, which is at odds with producing an unbiased estimator.

Several comments on the theorem are in order. The first assumption on exploration means that a sufficient amount of exploration is needed for this algorithm to be risk-neutral. We provide a simulation in Section 5 showing that this condition is important for risk neutrality. The assumption of centeredness and determinism is needed for our proof, but risk aversion seems to hold beyond them, as we show in Section 5. On the other hand, the requirement that there are only two actions with the same expectation is crucial, as we show in another simulation in Section 5. In an appendix, we discuss the regret properties of this algorithm.

Proof. Note that if both actions are deterministic, the conclusion of the theorem holds trivially. Otherwise, denote the non-deterministic action by $a \in A$. The evolution of choices can be expressed only based on the stochastic process

$$Y_t = \sum_{\substack{1 \le t' \le t \\ a_t = a}} \frac{r_t}{\sqrt{\pi_t}}.$$

If $Y_t > 0$, action a is chosen with probability $1 - \varepsilon_t/2$, if $Y_t = 0$, it is chosen with probability $1/2$, and if $Y_t < 0$, then it is chosen with probability $\varepsilon/2$. We define the random array

$$X_{t't} = \frac{1}{\sqrt{t}} Y_{t'}$$

This random array has the following properties:

Martingale It is an L^2-martingale array, *i.e.* $(X_{t't})_{1 \le t' \le t}$ is a square-integrable martingale with respect to its natural filtration.

Asymptotic Variance The conditional variances of martingale increments are constant (and deterministic).

$$\begin{aligned}
\sum_{t'=1}^{t} \mathbb{E}[(X_{t't} - X_{(t'-1)t})^2 | X_{(t'-1)t}] &= \sum_{t'=1}^{t} \sum_{a \in A} \pi_{a(t')} \mathbb{E}\left[\frac{r_a^2}{\sqrt{t \pi_{a(t')}}^2} \middle| X_{(t'-1)t} \right] \\
&= \sum_{t'=1}^{t} \frac{1}{t} \sum_{a \in A} \sigma_a^2 \\
&= \sigma_a^2.
\end{aligned}$$

In particular, as $n \to \infty$, $\mathbb{E}[(X_{t't} - X_{(t'-1)t})^2 | X_{(t'-1)t}] \to \sigma_a^2$ in probability.

Lindeberg Condition For any $\varepsilon > 0$, we have that

$$\begin{aligned}\sum_{t'=1}^{t}\mathbb{E}[(X_{t't}-X_{(t'-1)t})^2\mathbb{1}_{|X_{t't}-X_{(t'-1)t}|\geq\varepsilon}|X_{(t'-1)t}] &\leq 2\sum_{t'=1}^{t}\frac{(t')^{1-2\kappa}}{t}\mathbb{E}[r^2\mathbb{1}_{|r|\geq t^{\frac{1}{2}}(t')^{\frac{1}{2}-\kappa}\varepsilon}]\\ &\leq \frac{2}{t}\sum_{t'=1}^{t}(t')^{1-2\kappa}\sigma_a^2 e^{\frac{\lambda^2\sigma_a^2}{2}-\lambda t^{\frac{1}{2}}(t')^{\frac{1}{2}-\kappa}\varepsilon}\\ &\leq \frac{2}{t}\sum_{t'=1}^{t}t^{1-2\kappa}\sigma_a^2 e^{\frac{\lambda^2\sigma_a^2}{2}-\lambda t^{\frac{1}{2}}\varepsilon}\\ &\to 0.\end{aligned}$$

The first inequality plugs in definitions. The second uses a Chernoff bound. The last uses $1 \leq t \leq t'$. The convergence follows as exponential decay dominates polynomial growth and as convergence of a sequence implies convergence of the Cesàro mean.

Given these conditions, we can apply a Martingale Central Limit Theorem (Hall and Heyde 1980, Corollary 3.1) and conclude that the distribution of X_{tt} converges to $N(0, \sigma_a^2)$. This means that

$$\mathbb{P}[X_{tt} > 0], \mathbb{P}[X_{tt} < 0] \xrightarrow{t\to\infty} \frac{1}{2},$$

and hence

$$\mathbb{P}[a_t = a] = \frac{1}{2}.$$ □

While this algorithm works for two actions, another approach to risk neutrality, *optimism*, allows us to guarantee risk neutrality for an arbitrary number of actions of equal expectation. We discuss in the next subsection.

4.2 An optimism approach to risk neutrality

Another way to modify the statistic is not to reweight but to explicitly favor alternatives that have not previously been chosen as frequently in the past. Conventionally, this is referred to as *optimism* in the multi-armed bandit literature, compare Slivkins et al. (2019, Section 1.3.3). We show that it ensures risk neutrality.

Theorem 3 (Optimistic ε-Greedy)**.** *There exists $\rho_0 > 1$ such that for any $\rho \geq \rho_0$ and any $(\varepsilon_t)_{t\in\mathbb{N}}$ with $\varepsilon_t \to 0$, Optimistic ε-Greedy is risk-neutral.*

Proof Sketch. Note first that for exploitation steps of Optimistic ε-Greedy, the policy is the same as Upper Confidence Bound with exploration coefficient ρ, compare Auer (2002). We adapt the proof of Kalvit and Zeevi (2021b, Theorem 2) for our variant of ε-Greedy. Theorem 2 in Kalvit and Zeevi (2021b) shows that an optimistic policy without exploration has the property that

$$\lim_{t\to\infty}\mathbb{P}[a_t = a] \to \frac{1}{|\arg\max_{a\in A}\mu_a|}$$

for all a such that $a \in \arg\max_{a\in A}\mu_a$ and $\rho > \rho_0$. As Optimistic ε-Greedy does not incur regret as we show in Section 7, this property implies that the algorithm does not incur regret.

The full proof can be found in Kalvit and Zeevi (2021a, Appendix D). The proof goes as follows. First, show that $N_i(t)/t > 1/2|\arg\max_{a\in A}\mu_a|$ with probability approaching 1 in the limit, *i.e.* the fraction of times any action is chosen can be bounded below in probability. This is proved by means of a union bound and a Hoeffding bound. The operative equation that gets to this lower bound is Kalvit and Zeevi (2021a, Eqn. 40), which depends on Kalvit and Zeevi (2021a, Eqns. 35 and 39). Using this lower bound, we show that $|N_i(t)-N_j(t)|$ is small, *i.e.* the difference in the number of times any two actions are chosen is small as the time

goes to infinity. This again uses Markov's inequality, along with the Law of the Iterated Logarithm. Now, consider optimistic ε-Greedy. Since it implements the Upper Confidence Bound policy with probability $1 - \varepsilon_t$ and randomizes otherwise, in either case, it must be eventually choosing uniformly from amongst the highest mean actions, except for the vanishing probability with which it chooses dominated actions. □

It is worth noting why the mathematical intuition from the earlier result on ε-Greedy breaks. In this proof, the main object was a random walk with different variances for positive and negative values of the statistic. Optimism may be seen as introducing a drift towards the origin. The proof shows that this drift is strong enough to correct risk aversion.

This result demonstrates that one way to build a fairer world is with a particular type of optimism. Linking back to the credit decisions example we referenced in the introduction, credit scoring algorithms should evaluate applicants in the best possible light, adjusting for the risk profile of minority applicants by accounting for the fact that there might be less information on them.

5 Experiments

For the final section of this paper, we use experiments to confirm our theoretical results and investigate how far our results extend beyond their conditions. Unless noted otherwise, our experiments consider $\varepsilon_t = t^{-1}$, and report confidence bands that are Gaussian 90 % confidence intervals from 100 independent runs.

5.1 On risk aversion of ε-Greedy

Our initial set of experiments relate to the conditions in Theorem 1. The first experiment considers a setting that satisfies the conditions of Theorem 1. We simulate an ε-Greedy algorithm for an instance with three actions—a safe action that has distribution $\mathbb{1}_{\{0\}}$, a riskier action that has payoff distribution $U[-0.5, 0.5]$ while the riskiest action has payoff distribution $U[-1, 1]$. The results of this experiment can be found in Figure 2a ε-Greedy converges very quickly to selecting only the safe action.

The next experiment investigates a setting where arm rewards are not symmetric. ε-Greedy chooses between a safe action that has distribution $\mathbb{1}_{\{10\}}$ and a risky action that has distribution an exponential distribution with rate 10 *i.e.* Exp(10). The results are in Figure 2b. Even for the asymmetric exponential distribution, we observe perfectly risk-averse behavior.

Next, we consider settings where reward distributions are not centered around 0. We do this by setting the safe action to have distribution $\mathbb{1}_{\{0.5\}}$ while the risky action has distribution $U[2, -1]$. The results from this experiment are in Figure 2c. Again, we find that the safe arm is chosen with high probability.

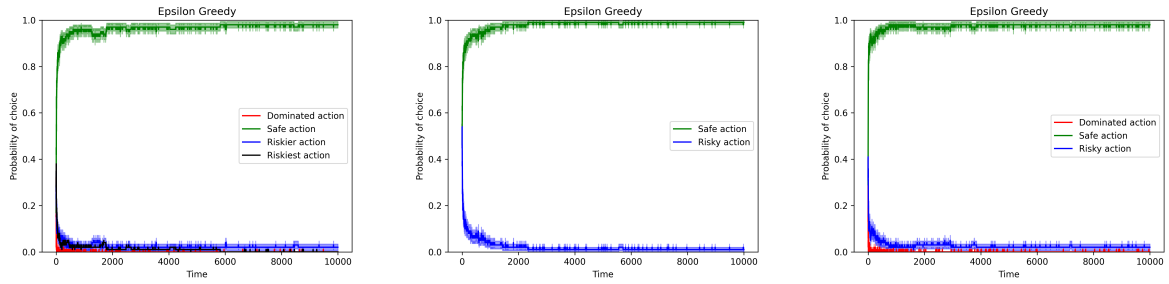
In all of the experiments so far, we had a perfectly safe action with constant reward. Our next experiment explores what happens when there is no optimal deterministic action. We consider an instance with a dominated action with reward distribution $\mathbb{1}_{\{-1\}}$, a safer action with distribution $U[-0.25, 0.25]$, a riskier action to have payoff distribution $U[-0.5, 0.5]$, and the riskiest action to have payoff distribution $U[-1, 1]$. The results are in Figure 3a. We find that ε-Greedy still chooses the safer action with significantly higher probability than the riskier one, which in turn is chosen with significantly higher probability than the riskiest one.

An important natural question that arises here is whether emergent risk aversion has implication beyond actions with equal mean payoffs. We show in Figure 3b that ε-Greedy's risk aversion has large transient effects before asymptotic guarantees kick in. This clarifies that the bias we identify can persist for a long time—for example, credit decisions can continue to be significantly discriminatory with such an algorithm *even if* the minority candidate has a strictly higher likelihood of repaying the loan.

5.2 On risk neutrality of Reweighted ε-Greedy

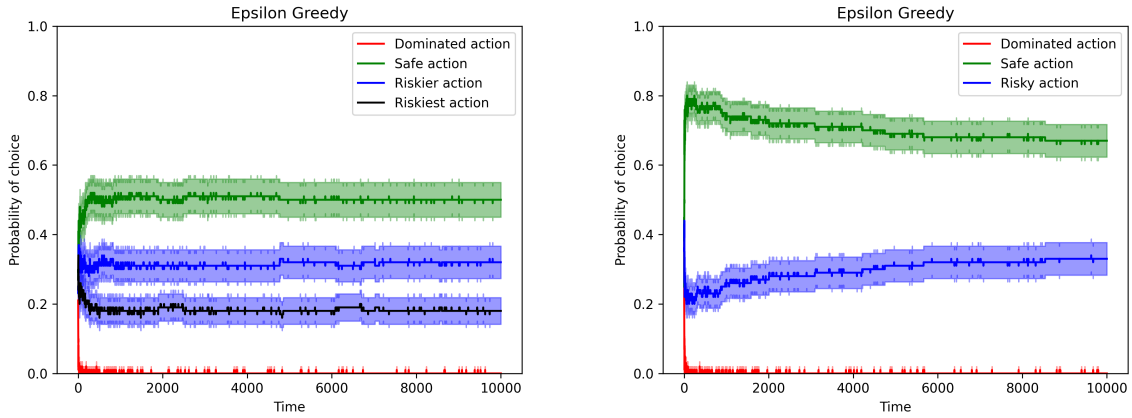
Our second set of experiments consider Reweighted ε-Greedy. To explore the limits of Theorem 2, we run simulations that vary exploration rate and the number of actions. We find that Reweighted ε-Greedy is risk-neutral for two actions across different reward distributions as long as the exploration rate is sufficiently high.

Our first experiment chooses a higher exploration rate than before ($\varepsilon_t = t^{-0.49}$) as required by Theorem 2. The instance has a safe action with payoff distribution $\mathbb{1}_{\{0\}}$ and a risky action with payoff distribution $U[-1, 1]$.



(a) Perfect risk aversion. (b) Perfect risk aversion. (c) Perfect risk aversion.

Figure 2: Plots of the behaviour of ε-Greedy, under and outside the conditions of Theorem 1. (a) Three actions with the same mean, ordered in second-order stochastic dominance, and including a deterministic arm. The deterministic arm is chosen most of the time. (b) Two arms with the same expectation, one determinstic, and one having an asymmetric distribution (exponential). While outside of the conditions of Theorem 1, the safe action is chosen most of the time. (c) Non-centered arms, ordered in second-order stochastic dominance. While outside of the conditions of Theorem 1, the safe action is chosen most of the time.



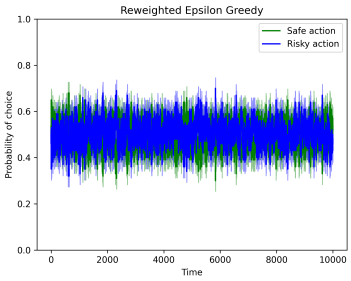
(a) Risk aversion (b) Transient risk aversion

Figure 3: Results of additional experiments investigating the generality of ε-Greedy's risk aversion. (a) For three centered arms that are ordered in second-order stochastic dominance, the safe arm is chosen most of the time. (b) For a dominated safe action, after 1,000 steps, the safe action is still chosen with probability more than a half.

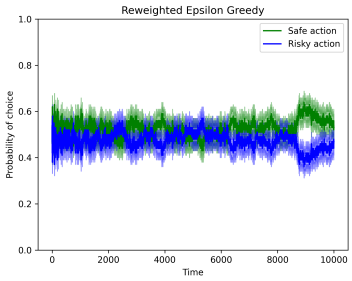
The second experiment considers again more exploration $\varepsilon_t = t^{-0.49}$. We consider runs of the algorithm on an instance with a safer action with reward distribution $U[0.25, 0.75]$ and a risker action with payoff distribution $U[0, 1]$.

The third instance considers the same instance as the first, but with an exploration rate of $\varepsilon_t = t^{-1}$. Figure 4a, Figure 4b, and Figure 4c respectively show the results. They show that the conclusions of Theorem 2 extend beyond the conditions of the Theorem as long as exploration is sufficiently high.

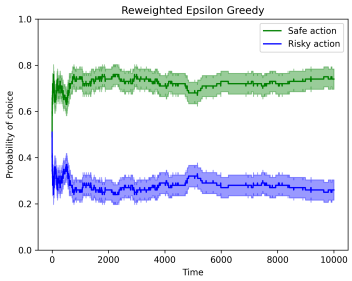
However, the restriction to two actions is crucial. To show this, we run a simulation of Reweighted ε-Greedy where we add a third action with distribution $\mathbb{1}_{\{-1\}}$ to the setup of where the safe action has



(a) Risk neutrality



(b) Risk neutrality



(c) Risk aversion

Figure 4: Plots showing that the Reweighted ε-Greedy is risk-neutral for two actions as long as exploration is sufficiently high. (a) For the case of two arms that are ordered in second-order stochastic dominance, one of which is deterministic, both arms are chosen equally often, as predicted by Theorem 2. (b) While outside of the conditions of Theorem 2, this seems to continue to hold for distributions that are ordered in second-order stochastic dominance but no action that is deterministic. (c) for less exploration, Reweighted ε-Greedy fails to choose the arms with the same probability, favouring the safer action.

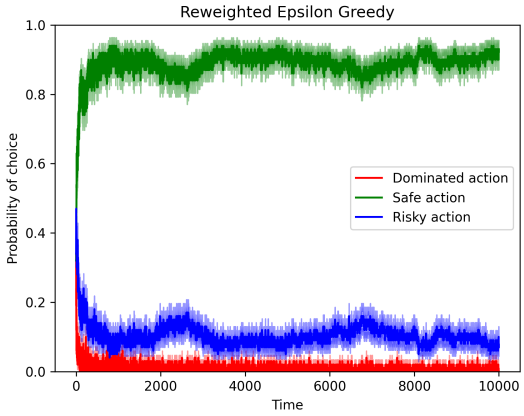


Figure 5: Reweighted ε-Greedy with a third dominated arm. The addition of a third arm leads the safe arm to be chosen more frequently.

payoff distribution $\mathbb{1}_{\{0\}}$ while the risky action has payoff distribution $U[-1,1]$ and set $\varepsilon_t = t^{-0.49}$. Figure 5 shows that risk neutrality need not hold in such a setting.

5.3 On risk neutrality of Optimistic ε-Greedy

Finally, we consider the assumptions of Theorem 3. We do this by running three simulations. In the first one, we consider a high exploration coefficient $\rho = 2$. We consider the behavior of the algorithm on an instance with payoff distributions $U[-0.25, 0.25]$ and $U[-1,1]$, and a dominated action with reward $\mathbb{1}_{\{-1\}}$.

The second experiment considers a high exploration coefficient $\rho = 2$ once more, but an instance consisting of a non-centered reward distributions $U[0.25, 0.75]$, $U[0,1]$, and a dominated action with reward $\mathbb{1}_{\{-1\}}$.

The final experiments consider a lower exploration coefficient $\rho = 0.02$ and an instance with reward distributions $\mathbb{1}_{\{0\}}$ and $U[-1,1]$, and a dominated action $\mathbb{1}_{\{-1\}}$.

The results are in Figure 6a, Figure 6b, and Figure 6c, respectively. While the first two experiments show risk neutrality, the last experiment shows the necessity of a sufficient optimism coefficient for the conclusions in Theorem 3 to hold.

(a) Risk neutrality (b) Risk neutrality (c) Risk aversion

Figure 6: Plots showing that the Optimistic ε-Greedy is quite generally risk-neutral, as long as exploration coefficient ρ is sufficiently high. (a) For two centered arms and a dominated arm, one of which is deterministic, we find that the safer action is chosen most of the time. (b) Also, for two non-centered arms that are ordered in second-order stochastic dominance and a dominated arm, the safe arm is chosen most of the time. (c) For a lower exploration rate and two centered arms ordered in second-order stochastic dominance, one of which is deterministic, risk neutrality fails to hold.

6 Conclusion

Learning algorithms can have unintended emergent risk behavior, leading to outcomes that may be at odds with the objectives of those who deploy them. The basic intuition behind this is that exploration policies often use simple statistics such as the mean to keep track of the estimated value of each option while not using the fact that other properties of their data, such as noise, can affect their learning process. As a consequence, higher variance actions can end up being shunned purely because they yield a bad outcome early on. Corrections to the algorithm's statistic can reinstate risk neutrality.

7 Regret properties of algorithms

This section provides evidence that the algorithms we consider do not incur regret.

Definition 5. An algorithm π *is no-regret* or *does not incur regret* if for all instances such that F_a is sub-Gaussian for all $a \in A$ and for all $a, a' \in A$, $\mu_a < \mu_{a'}$, we have

$$\mathbb{P}[a_t = a] \xrightarrow{t\to\infty} 0.$$

We provide a proof sketch for the following statement, which implies that ε-Greedy does not incur regret.

Lemma 1. *Let $\varepsilon_t \to 0$, $\sum_{t=1}^{\infty} \varepsilon_t = \infty$. Also let $a \in A$, $\mu_a < \max_{a\in A} \mu_a$. Then, for any $\delta > 0$, there is $t \in \mathbb{N}$ such that for all $t' \geq t$,*

$$\pi_{at'} \leq \delta.$$

Proof Sketch. Choose $t' \in \mathbb{N}$ such that for all $\tilde{t} \geq t'$, $\varepsilon_t \leq \delta/2$. That is, the probability of exploration steps is small. It is sufficient to show that for some t'' and $\tilde{t} \geq t''$, in exploitation steps,

$$\mathbb{P}[\mu_a(\tilde{t}) - \mu_{a'}(\tilde{t}) \geq 0] \leq \frac{\delta}{2}.$$

By Hoeffding's inequality, with high probability in $\tilde{t}$, both actions have been chosen at least $\frac{1}{3}\sum_{t=1}^{t''} \varepsilon_t$ times. Conditional on this event,

$$\mathbb{P}[\mu_a(\tilde{t}) - \mu_{a'}(\tilde{t}) \geq 0] \xrightarrow[t\to\infty]{} 0.$$

In particular,

$$\mathbb{P}[\mu_a(\tilde{t}) - \mu_{a'}(\tilde{t}) \geq 0] \leq \frac{\delta}{2},$$

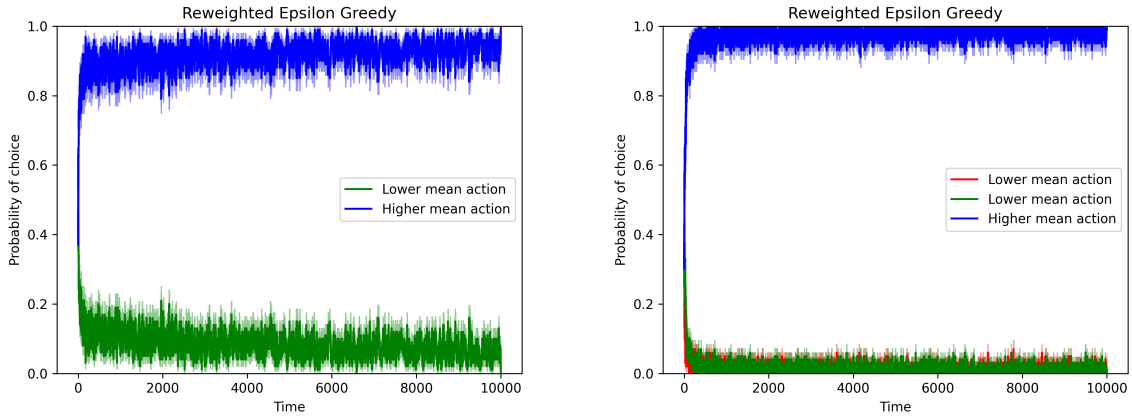
for some $t'' \in \mathbb{N}$ and $\tilde{t} \geq t''$. Choosing $t \geq \max\{t', t''\}$ yields the claim. □

Corollary 1. *ε-Greedy does not incur regret.*

Next, we provide a proof sketch that Optimistic ε-Greedy does not incur regret.

Proposition 1. *Optimistic ε-Greedy does not incur regret.*

Proof Sketch. This proof is similar to the proof that Upper Confidence Bound does not incur regret (see, *e.g.*, Auer (2002)). The only difference between the Upper Confidence Bound algorithm and the Optimistic ε-Greedy is exploration, which vanishes in the limit. It remains the case that the confidence bands are valid, and hence, there is a logarithmic upper bound for the probability that the bandit algorithm chooses a sub-optimal action in an exploitation step. As in the original proof of the Upper Confidence Band, this amounts to a sub-linearly growing probability of choosing a sub-optimal action and hence no regret. □



(a) No regret

(b) No regret

Figure 7: Plots illustrating the regret of Reweighted ε-Greedy. (a) For two arms, one of which with higher variance and one of which with higher expectation, the higher-expectation action is chosen most of the time. (b) For three actions with a deterministic action, a non-deterministic action of the same expectation, and a yet-higher-expectation dominant arm, we find that the dominant arm is chosen most of the time.

Finally, we test whether Reweighted ε-Greedy incurs regret. We present two simulations. One with an instance with payoff distribution Exp(1) and Exp(2) + $N(0,1)$, and another where two lower mean actions have payoff distributions $\mathbb{1}_{\{0.35\}}$ and $U[0.25, 0.75]$ while a higher mean action has payoff distribution $U[-1,3]$. Figure 7 shows the results. We find that for a non-deterministic, non-centered dominant action, Reweighted ε-Greedy appears not to incur regret.

8 Additional simulations

Both ε-Greedy's and Reweighted ε-Greedy's payoff estimates are biased estimators, compare Lattimore and Szepesvári (2020, Section 11.2). It is natural to ask whether debiasing the payoff estimates can address emergent risk preferences. To answer this, we run a simulation of a ε-Greedy with the statistic

$$\mu_{a,d}(t) = \frac{1}{N_a(t)} \sum_{\substack{1 \leq t' \leq t \\ a_t = a}} \frac{r_t}{\pi_{at}}.$$

This statistic leads to an unbiased estimate of the reward of an arm (Lattimore and Szepesvári 2020). The safe action in our simulation has reward distribution $\mathbb{1}_{\{0\}}$ while the risky action has reward distribution $U[-1,1]$. Figure 8 reports the results. The debiasing leads to risk-*loving* behavior as opposed to risk-neutral behavior. One intuition for this uses monotonicity of risk attitude in probability normalization: The division by the square root of the choice probability in Reweighted ε-Greedy led to a correction that makes the algorithm exactly risk-neutral. Debiasing, which divides the statistic by the choice probability, a smaller number, leads to a stronger correction in the direction of risk-loving behavior.

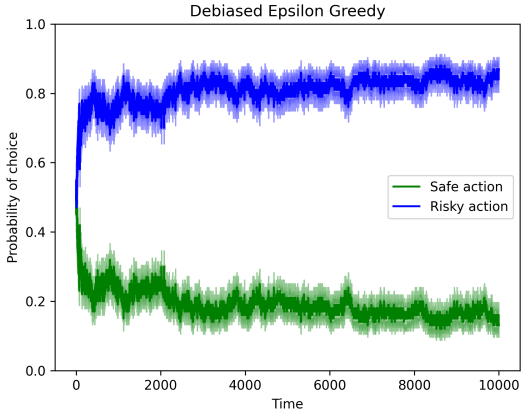


Figure 8: ε-Greedy with a debiased reward estimate. We find that for two centered actions ordered in second-order stochastic dominance, one of which is deterministic, the non-deterministic action is chosen most of the time.