Adaptive Wage Setting

A Prior-Free Theory of Adverse Selection and Monopsony Markets

Carlos Gonzalez Perez*

May 2023

Abstract

We consider the problem of wage setting in a monopsony market with unknown productivity and reservation wage of workers. We develop a simple algorithm which ensures convergence to equilibrium strategies under realistic limited feedback, without any prior knowledge on the joint distribution of target variables and for any arbitrary distribution of outcomes. We show an upper bound on the regret of the algorithm of $\tilde{\mathcal{O}}(K^{\frac{2}{3}})$, and matching lower bounds on the problem which ensure near-optimality of the algorithm under stochastic and adversarial considerations. This new approach to the problem of the firm unveils learning dynamics in the presence of adverse selection and imperfect competition under minimal information requirements. In addition, it facilitates the examination of counterfactual policy analysis like minimum wage and limited information processing capacity. Simulations verify the theoretical predictions of our wage-setting strategy. Finally, this study introduces relevant concepts to the adaptive policy design literature, including *asymmetric feedback* structures and the welfare implications of greedy parameter selection.

Key Words: Adaptive Policy Design, Hannan Consistency, Online Learning, Monopsony, Wage-Setting

*University of Oxford, Department of Economics. Manor Road Building, Manor Rd, Oxford OX1 3UQ. Email: carlos.gonzalezperez@economics.ox.ac.uk. I would like to thank my supervisor, Professor Max Kasy, for his invaluable feedback and support that led to this thesis.

1 Introduction

Characterisation of equilibrium dynamics under imperfect information and adverse selection mechanisms is of utmost interest in the fields of mechanism design and public policy evaluation. Intuitively, equilibrium existence is futile in realistic environments if agents cannot implement simple strategies which gets them close to those equilibria [Hart and Mas-Colell, 2013]. Using a novel online learning technology, this paper develops a simple strategy which ensures convergence to some notion of equilibrium (best action) in the generalised monopolist problem (GMP)[1]. More specifically, we show constructively the existence of Hannan Consistent (HC) policies in the GMP with limited feedback, without any information requirements (priors) on the learner, and for any arbitrary distribution of outcomes. HC means that, asymptotically, the average reward of the learner is as large as the maximum reward she could have obtained by playing the best fixed action.

Our algorithm can be relevant in many public policy frameworks which suffer from adverse selection, including wage setting with unknown productivity and reservation wage distribution functions. Although this study focuses on wage setting, the theory developed in this paper is generally applicable and can be extended to any market with quality differentiation and asymmetric information, such as insurance markets, investment markets, or discrimination in labour markets.

Contribution to the Literature. This study imports technologies from the machine learning and statistics literature into the fields of microeconomics and policy design. From a positive perspective, we first establish a connection between some notion of equilibrium (best response) and Hannan Consistency. In particular, we show that in one-player games with finite number of actions and bounded rewards, the limiting distribution of empirical probabilities over actions induced by Hannan Consistent policies converge to the best response equilibrium. This result is simply an instance of well-established theory in adaptive policy design [Hart and Mas-Colell, 2001a].

The main contribution of this paper is showing the existence of HC policies in the GMP with limited feedback, without demanding any priors on the learner, and for any arbitrary distribution of rewards. We do so constructively, by creating an

[1]See Section 2 for a description of the problem.

algorithm which induces a Hannan Consistent policy. The GMP is a multi-variable canonical problem in adverse selection and imperfect information environments, hence the development of simple strategies which converge to optimal actions is very relevant in applied and theory settings.

From an online learning standpoint, we demonstrate that our algorithm is near-optimal under stochastic and adversarial specifications with regret bounds of $\tilde{\mathcal{O}}(K^{\frac{2}{3}})$. The GMP is weakly harder than similar problems in the literature like the adaptive welfare maximisation problem outlined in [Cesa-Bianchi et al., 2022], where the authors report bounds of the same order up to constant factors. As a result, our evidence extends their results. Moreover, our problem introduces a new class of feedback, *asymmetric feedback*, which we compare in terms of welfare to two prevalent feedback structures in the literature, *full feedback* and *realistic feedback* [Cesa-Bianchi et al., 2021]. A wide array of simulations validate our theoretical findings.

Finally, from a public policy standpoint we show the potential risk of greedy parameter selection and its connection to the variance of the data generating process (DGP). To the best of our knowledge, this is the first empirical evidence of this phenomenon in the adaptive policy design literature. In addition, this paper identifies a relevant economic application of bandit learning: wage setting with unknown productivity and reservation wage distributions of workers. As such, the practitioner may use this paper as a tool for structural economic analysis, including the evaluation of counterfactual policies like minimum wage or limited information processing capacity[2]. We show that when relying on our algorithm, the behaviour of the firm is compatible with classic theories of imperfect competition, and empirical evidence. In addition, we discover new results like the notion of *information gains* associated to Minimum Wage policies. Interestingly, all these prescriptions can be established under much weaker information requirements than mainstream theory results in Economics.

Paper Structure. The remaining of this paper is organised as follows. Section 2 introduces the notion of regret, relevant notation, and the probabilistic framework for the problem of interest. Section 3 summarises previous results in adaptive policy design, convergence to equilibrium, and adverse selection. Section 4 presents classic theory in

[2]In the extended version of this paper, we also explore the impact of productivity shocks. For space considerations, we have not included such results in this manuscript.

offline monopoly adverse selection problems and introduces its online analogues (GMP). In addition, it discusses the concept of Hannan Consistency and its connection to equilibrium convergence. This Section also establishes difficulty relations across feedback structures and information requirements. Section 5 presents Algorithm 1 which is Hannan Consistent at a near-optimal rate in the GMP. Section 6 offers a wide array of simulations which validate our theoretical prescriptions. It further discusses the hazards of sub-optimal (greedy) parameter selection. Section 7 derives suited theory for structural policy analysis, and compares these results to mainstream economic theory and empirical evidence. Section 8 concludes.

2 Set-Up, Notation and Regret

Set-Up. Consider an economy with K workers who are characterised by pairs (u_i, v_i), where u_i and v_i are the ith realisations of the random variables $U \in \Omega_U$ (productivity) and $V \in \Omega_V$ (reservation wage). Of particular interest is the joint distribution $F_{U,V}$. There is a policymaker with the mandate of selecting K wages $x_i \in \Omega_X$. Without much loss of generality assume that $\Omega_U = \Omega_V = \Omega_X = [0, 1]$. The policymaker owns the firm, so the words firm and policymaker are used interchangeably. The policymaker selects a policy π to maximise total welfare $\sum_i^K S_i(x_i)$, where $S_i(\cdot)$ is a (possibly random) function of wages. To do so, she may rely on a sequence of information sets H_i, referred as the *history*, which contain all the information available to the firm after observing workers $\{1, \dots, i-1\}$. In our setting, workers are perfectly rational individuals who select action $J_i = \mathbb{1}\{x_i \geq v_i\}$. For convenience, define $J^w = \{i : J_i = w\}$.

Literature typically characterises $v = r(u)$, where $r(\cdot)$ is a measurable map $\Omega_U \mapsto \Omega_V$, such that the joint distribution $F_{U,V}$ is degenerate on V [Mas-Colell et al., 1995]. However, we do not restrict $F_{U,V}$ in this way, nor in any other meaningful way. Following the online learning literature, this study considers two limiting cases: The *stochastic* framework, where the sequence $\{(u_i, v_i)\}_{i=1}^K$ is conformed by K *iid* samples from a fixed distribution $F_{U,V}$, and the *adversarial* framework, where the sequence $\{(u_i, v_i)\}_{i=1}^K$ can be selected arbitrarily. We formalise this argument below.

Regret and Policy Analysis. The (expected) regret $\mathcal{R}$, which is the main object of interest in the field of bandit optimisation, can be defined as the difference in cumulative welfare between a sequence $\{x_i\}_{i=1}^K$ and the best fixed policy x^* in hindsight. Formally,

$$\mathcal{R}(\pi, \nu)_K = \mathbb{E}[\sup_{x \in \Omega_X} \sum_i^K S_i(x) - \sum_i^K S_i(x_i)] \tag{1}$$

where ν, also referred as the environment, is the class of DGP available to the adversary. As mentioned earlier, the stochastic environment is defined by the class of DGP such that each pair (u_i, v_i) is an *iid* realisation of $F_{U,V}$, while the (oblivious) adversarial environment is characterised by the set of arbitrary DGP which may even depend on the algorithm (policy) of the learner. It follows that any non-convex deterministic algorithm necessarily incurs in linear regret under adversarial specifications. This class is only limited by the fact that outcome realisations cannot depend on the history of actions and outcomes. In the stochastic environment, expectations are taken with respect to the randomness in the DGP process (and possibly, any randomness in the algorithm of the learner). In the (oblivious) adversarial framework, the sequence of outcomes is considered to be fixed, so expectations are taken only with respect to the randomness in the algorithm of the learner. We say that a policy is sub-linear of $\mathcal{O}(K^r)$ iff $\limsup_{K \to \infty} \frac{\mathcal{R}}{.K^r} \leq C$ for some $C < \infty$. Certainly, under bounded rewards, any policy achieves $\mathcal{O}(K)$, thus non-trivial bounds require $0 < r < 1$.

Online Convex Optimisation. There exists an interesting connection between (adversarial) bandit problems and Online Convex Optimisation (OCO) [Hazan et al., 2016]. In fact, Bandit Optimisation can be reduced to general OCO frameworks, and, consequently, OCO solving techniques like Stochastic Gradient Descent (SGD) become immediately available to the bandit learner. Even when dealing with a finite number of arms in the policy space, convexity of the decision set can be established by relying on the convexity of the probability space of selecting a particular action (arm). Hence, to ensure convexity in finite arms settings, the policy of the learner must rely once again on some degree of randomisation. In Section 4, we explore this equivalence in further detail and demonstrate that Algorithm 1 can be interpreted as a penalised SGD.

3 Literature Review

In his seminal work, [Akerlof, 1978] characterised the notion of competitive equilibrium in markets with quality differentiation and incomplete information. His findings highlighted how information asymmetry could lead to market unravelling through adverse selection mechanisms. Similar insights can be applied to monopoly markets as in [Mas-Colell et al., 1995]. Benchmark results consider that the monopolist does not observe the variable realisations (like productivity and reservation wage of workers) but holds some prior over the join distribution of outcomes $F_{U,V}$ and best responds to it. Static (offline) equilibrium results can be established under wrong beliefs, but in repeated games we might demand posteriors to be consistent with the empirical distribution of outcomes. Under reasonable feedback assumptions, asymptotically, firm's best response should be connected to the convergence of posterior to the true joint distribution of outcomes $F_{U,V}$.

Convergence to equilibrium in repeated normal games is a well studied problem in Economics [Hart and Mas-Colell, 2001a] [Hart and Mas-Colell, 2013], and Online Learning [Cesa-Bianchi and Lugosi, 2006]. Hart and Mas-Collel established convergence to correlated equilibrium of Hannan Consistent policies [Hart and Mas-Colell, 2001a], using Blackwell's approachability theorem [Blackwell, 1956], and they showed that their *regret matching* strategy was indeed Hannan Consistent [Hart and Mas-Colell, 2000]. Later on, they proved that a small variation of *regret matching* was in fact HC in the bandit framework, where the learner only has access to the outcome of the action she played, but not to conterfactual outcomes [Hart and Mas-Colell, 2001b].

Their results have been extended and improved by the Online Learning literature as in [Auer et al., 2002] [Cesa-Bianchi and Lugosi, 2006] [Bubeck et al., 2012]. However, convergence to equilibrium under realistic notions of feedback is still an open question in the adaptive policy literature. For instance, [Kleinberg and Leighton, 2003] proposed an optimal algorithm with regret of $\mathcal{O}(K^{\frac{1}{2}})$ for the monopolist pricing problem where a monopolist had to maximise profits of the form $x \cdot \mathbb{1}(x \geq v)$, with v equal the reservation price. At the end of the period, only $\mathbb{1}(x \geq v)$, and not v, is revealed to the learner. Similar (more complicated) problems have gained attention in the last few years. For instance, [Cesa-Bianchi et al., 2021] studied the bilateral trade problem $\mathbb{1}(x \leq v^b) \cdot \max(x - v^s, 0) + \mathbb{1}(x \geq v^s) \cdot \max(v^b - x, 0)$ where v^b and v^s are the reservation prices

of the buyer and the seller, respectively. They showed that under *realistic feedback*[3] no policy could achieve sub-linear regret without strong assumptions on the DGP. Finally, [Cesa-Bianchi et al., 2022] presented a near-optimal algorithm of $\mathcal{O}(K^{\frac{2}{3}})$ for the adaptive welfare maximisation problem, with applications to optimal tax design. Their problem $x \cdot \mathbb{1}(x \leq v) + \lambda \cdot \max(v - x, 0)$ is a weakly easier version of the GMP, thus, by showing regret bounds of the same order (up to constant factors), our results extend their work. We come back to these problems later in the text and explain their different feedback structures and regret bounds.

Furthermore, our algorithm gives a better understanding of adverse selection effects on dynamic games. The behaviour of economic agents exposed to adverse selection has only been studied in stylised settings with little attention to evolution dynamics, like the one described by Esponda [Esponda, 2008] or Bjorkegren et al. [Björkegren et al., 2020]. Our work allows for a deeper investigation of transition patterns in a broader class of games. Finally, imperfect competition in dynamic games has also been studied in the theoretical Industrial Organisation literature (like the famous [Ericson and Pakes, 1995] framework) under a Markov Process approach. We depart from these models in several ways. In particular we assume no effect of present actions on future states, we impose minimal primitives on the model, and provide simple strategies to best response (unlike EP complicated numerical optimisation solutions). I refer to [Doraszelski and Pakes, 2007] for a review of the topic.

4 Offline and Online Adverse Selection

Building on the set-up introduced in Section 2, this Section describes standard offline monopoly equilibrium under partial information and introduces an online learning analogue of the problem. For simplicity, we focus on games where the firm holds correct beliefs on the distribution of target variables.

[3]See Section 4.2 for a detailed description of feedback structures.

4.1 Offline Monopolist Equilibria

Consider a generalised version of the offline monopolist problem[4], where a single wage-setting firm maximises welfare $\mathbb{S}$ given a continuum of workers,

$$\mathbb{S} = \int_{\Omega_U \times \Omega_V} \mathbb{1}(x \geq v) \cdot (u - x + \lambda(x - v)) \; dF_{U,V} \tag{2}$$

where $(x - v)$ is the workers' surplus, $\lambda < 1$ captures the preferences of the monopolist towards workers' welfare, and $F_{U,V}$ is the true distribution of target variables. We refer to the $\lambda > 0$ case as the generalised monopoly problem (GMP), and to the $\lambda = 0$ especial case as the simplified monopolist problem (SMP), where welfare is equivalent to firm profits. We write $x(u, v) = x$ for notation simplicity. We further assume that in case of indifference, both the worker and the firm prefer working to not working.

Full Information vs Partial Information. The offline framework is characterised by the absence of learning, such that no information revealed in period $i \geq 1$ is carried over to the following periods. In this context, the notion of equilibrium is always defined as the belief-conditional profit maximising action. For concreteness, the reader may characterise such equilibrium as a one-player Bayesian Nash Equilibrium (BNE) where beliefs are given by the sequence of histories H_i, and the action space equals $[0, 1]^K$. The Full Information (F) case is defined by the inclusion of $(u_i, v_i) \in H_i$ for all i. Equilibrium is simply given by the set of strategies $\boldsymbol{x}^{\mathrm{F}} : x^{\mathrm{F}}_{\mathrm{GMP},i} = x^{\mathrm{F}}_{\mathrm{SMP},i} = v_i$ for $\{i : u_i \geq v_i\}$, $x^{\mathrm{F}}_{\mathrm{GMP},i} < v_i, x^{\mathrm{F}}_{\mathrm{SMP},i} < v_i$ otherwise. Another object of interest is the employment equilibrium, which we define as $J^{1,\mathrm{F}}_{\mathrm{GMP}} = J^{1,\mathrm{F}}_{\mathrm{SMP}} = \{i : u_i \geq v_i\}$. Unemployment equilibrium is defined analogously. The Full Information case is a textbook example of first-degree price discrimination where the monopolist is able to capture all workers' surplus (WS).

$$\int_{\Omega_U \times \Omega_V} \mathbb{1}(x \geq v) \cdot (x - v) \, dF_{U,V} = \int_{\mathbb{1}(u >= v)} (v - v) \, dF_V + \int_{\mathbb{1}(u < v)} 0 \cdot (x - v) \, dF_V = 0 \tag{3}$$

The Partial Information case (P), where only $F_{U,V} \in H_i$ for all i, is of particular

[4]Technically, this is a monopsony and not a monopoly problem. We use the word monopoly for convenience and make reference to monopsony theory when relevant.

interest. In this context, equilibrium is given by $x^{\mathrm{P}}_{\mathrm{SMP}} = \arg\max_x \mathbb{E}_V\big[\mathbb{1}(x \geq v)\cdot\big(\mathbb{E}_U[u \mid v \leq x] - x\big)\big]$ and $x^{\mathrm{P}}_{\mathrm{GMP}} = \arg\max_x \mathbb{E}_{V,U}\big[\mathbb{1}(x \geq v)\cdot\big(u - x + \lambda\cdot(x-v)\big)\big]$. $J^{1,\mathrm{P}}_{\mathrm{GMP}}$ and $J^{1,\mathrm{P}}_{\mathrm{SMP}}$ are defined as expected. Despite the ambiguity in the effect on employment, the WS is weakly larger in the Partial Information setting, and the overall welfare, weakly smaller. This decay in welfare (profits) is driven by adverse selection considerations and market unravelling dynamics. For instance, in a scenario with positive correlation between U and V, to attract highly productive individuals, the firm must increase the wage x for all employees, including low-productivity ones, causing a decline in profitability (and possibly employment). From the firm's perspective, F is preferred over P in the Blackwell-informative sense.

There are two final points which are worth mentioning. Firstly, in the Partial Information setting, the firm is not constrained to set equal wages. Nonetheless, since workers' heterogeneity remains unobserved, the firm can only establish i-dependent wages through randomised strategies. It can be formally proven that no randomised strategy can generate strictly higher profits than the optimal deterministic strategy. Secondly, even if equilibrium is well characterised as the solution to the conditional expectation above, the implementation of such policy, i.e. the derivation of closed form solutions, remains a challenging task. This is one of the reasons why the literature typically relies on simplifying assumptions of the kind $v = r(u)$. Observe that this problem is generally a well-behaved convex optimisation problem, so optimisation techniques like SGD become available. In Section 5 we demonstrate that Algorithm 1 is in fact a variation of penalised SGD. From a deployment perspective, convex optimisation methods stand as a transparent simple alternative to analytically limited theory.

Lower Information Requirements and Learning. Although this framework elegantly captures complex market dynamics under adverse selection, it feels incomplete. It is unreasonable to assume perfect knowledge of the joint distribution of target variables $F_{U,V}$, thus, under incorrect priors, the solution concept to the monopoly problem above is not optimal. One may impose some degree of learning on the firm's strategy, for instance, consistency between the posterior and the empirical distribution of outcomes. However, under limited feedback, it is unclear that the strategy of the firm can be as good as the optimal strategy under partial information considerations, where $F_{U,V}$ is the true distribution of target variables. By limited feedback, we mean that the firm only recovers $\mathbb{1}(x \geq v_i)$ for the action x it selects in period i, and conditional on

$\mathbb{1}(x \geq v_i)$ it recovers the productivity u_i of the worker. Otherwise u_i remains unknown.

We show that policies converging to correct-beliefs Partial Information equilibrium are indeed available. We do so constructively, by presenting an algorithm which converges to equilibrium outcomes at an optimal rate under the feedback considerations above, without any prior knowledge and for any arbitrary distribution of outcomes. Because the equilibrium strategy in the Partial Information setting is given by the best fixed strategy in hindsight, convergence to equilibrium actions is equivalent to Hannan Consistency of the policy. We establish this equivalence formally below.

Definition 1 *Hannan Consistency.* A policy is said to be Hannan consistent if the sequence of actions $\{x_i\}_i^K$ induced by the policy

$$\limsup_{K\to\infty} \frac{1}{K}\Big(\sup_{x\in\Omega_X} \sum_i^K S_i(x) - \sum_i^K S_i(x_i)\Big) = 0 \text{ with probability 1} \tag{4}$$

where the probability is taken with respect to the randomness in the policy. To show convergence of Hannan Consistent policies to equilibrium strategies in the GMP, we first need to introduce the online analogue of the monopolist wage setting problem.

4.2 Adverse Selection in Adaptive Settings

Timeline. At the beginning of period i, the firm offers wage x_i to worker i, possibly relying on the history of previous actions and rewards H_i. Crucially, $F_{U,V} \notin H_i$ for any period i. Worker i evaluates the job offer based on their private information v_i, and accepts if $x_i \geq v_i$. Shall $x_i \geq v_i$, the worker works, the firm observes his productivity u_i, and the worker obtains surplus $x_i - v_i$. In the contrary event, u_i remains unknown and welfare is set equal to zero. In both cases, the policymaker does not observe v_i, but a binary statistic $\mathbb{1}(x_i \geq v_i)$, some times referred as the *demand function*.

Single Period Online Monopolist Welfare. Consider,

$$S_i^{\text{GMP}}(x_i) = \mathbb{1}(x_i \geq v_i)\big((u_i - x_i) + \lambda(x_i - v_i)\big) \tag{5}$$

where $S_i(x)$ is the welfare recovered in round i, given wage x. Define $x^{\pi,K}$ as the K-sequence of actions $\{x_i\}_{i=1}^K$ induced by policy $\pi(H)$. Similarly, define x_i^π as the action x induced by policy π in period i. Finally, let $p_x^{\pi,K} = \frac{1}{K}\sum_i^K \mathbb{1}(x_i = x)$ be the vector of empirical probabilities of selecting actions $\boldsymbol{x}_b$ with $b \in \{1, \ldots, B+1\}$ induced by policy π over K periods.

Proposition 2. Fix an arbitrary sequence $(u_i, v_i)_{i=1}^K$ and a policy space Ω_X with a discrete number of arms $B+1$. Let policy π be Hannan Consistent for a one player game with bounded rewards $S_i(x)$, then $p_{x^*}^{\pi,K} = \frac{1}{K}\sum_i^K \mathbb{1}(x_i = x^*) \xrightarrow{p} 1$ as K goes to ∞ where $x^* \in \arg\max_x \limsup_{K\to\infty} \sum_i^K S_i(x)$.

Proof: Assume $p_{x^*}^{\pi,K} \overset{p}{\not\to} 1$, then $\lim_{K\to\infty} \mathbb{P}\big(1 - \frac{1}{K}\sum_i^K \mathbb{1}(x_i^\pi = x^*) > \epsilon\big) > 0$ for some $\epsilon > 0$. Refer to this event with positive probability as E. Define the set $K \supseteq I^* = \{i : x_i^{\pi,K} \neq x^*\}$. Rewrite $\mathbb{P}(E)$ as $\mathbb{P}\big(\frac{|I^*|}{K} > \epsilon\big) > 0$. It follows that $|I^*| > K \cdot \epsilon$ with positive probability. From now on, let event E hold. Define $x' = \arg\max_{x\neq x^*} \limsup_{K\to\infty} \sum_{i\in I^*} S_i(x)$. By construction of x^*, $\limsup_{K\to\infty} \frac{1}{|I^*|}\sum_{i\in I^*} S_i(x^*) - \frac{1}{|I^*|}\sum_{i\in I^*} S_i(x') = S_{\min} > 0$. Finally, apply the definition of Hannan consistency to see that

$$\begin{aligned} &\limsup_{K\to\infty} \frac{1}{K}\Big(\sup_{x\in\Omega_X} \sum_i^K S_i(x) - \sum_i^K S_i(x_i)\Big) \geq \limsup_{K\to\infty} \frac{1}{K}\Big(\sum_{i\in I^*} S_i(x^*) - \sum_{i\in I^*} S_i(x')\Big) = \\ &\limsup_{K\to\infty} \frac{1}{K}\cdot|I^*|\cdot S_{\min} \geq \limsup_{K\to\infty} \frac{1}{K}\cdot K\cdot\epsilon\cdot S_{\min} = \epsilon\cdot S_{\min} > 0 \end{aligned} \tag{6}$$

Where all inequalities hold with positive probability. It now follows that $\mathbb{P}\Big(\limsup_{K\to\infty} \frac{1}{K}\big(\sum_i^K S_i(x^*) - \sum_i^K S_i(x_i)\big) > \delta = \epsilon\cdot S_{\min}/2\Big) > 0$. We have just shown that NOT A $\implies$ NOT B, it follows that B $\implies$ A. $\square$

Interpretation. In an online game which satisfies the conditions in Proposition 2, the implementation of a Hannan Consistent strategy ensures that the actions of the agent converge in probability to the optimal set of actions x^* in the Partial Information context with correct beliefs. In fact, this result is a particular instance of more general results in the adaptive learning literature [Hart and Mas-Colell, 2001a] [Cesa-Bianchi and Lugosi, 2006]. To move now into the normative realm and establish the main result of this paper, it remains to prove that Hannan Consistent policies are in

fact available under limited feedback, for any arbitrary sequence $\{u_i, v_i\}$ of outcomes, and without prior knowledge of $F_{U,V}$. We refer to these policies as HC* policies. To do so, we make the following observation

Observation 3. Under equation (1), it follows that adversarial sub-linear regret of a prior-free policy under the feedback structure in the GMP is sufficient for HC*. Consequently, let policy π embedded in Algorithm A be sub-linear under adversarial specifications, then the π-induced strategy converges in probability to the offline equilibrium strategy with $F_{U,V} \in H_i$ for all i. We present a near-optimal algorithm for the GMP with sub-linear regret in Section 5.

Before introducing this algorithm, we explore two dimensions of the online monopolist wage setting problem which are very related to each other and which are of interest in the field of bandit optimisation, namely the feedback structure and the information requirements.

Feedback Asymmetry. The structure of feedback ψ is of ultimate interest in online learning problems because $H_i = H_{i-1} \cup \psi_i$. So far, the literature in adaptive policy design has focused on two limiting cases, the *full* feedback case and the *realistic* feedback case [Cesa-Bianchi et al., 2021]. Feedback is said to be realistic for variable Z if $\mathbb{1}(x \geq z)$ is recovered at the end of every period, and it is said to be full if $Z = z$ is recovered at the end of every period. Feedback in the monopoly pricing problem [Kleinberg and Leighton, 2003] and the adaptive welfare maximisation problem [Cesa-Bianchi et al., 2022] is realistic, while authors in the bilateral trade problem [Cesa-Bianchi et al., 2021] discuss their results under both specifications.

The feedback structure in the GMP is arguably different from both of these structures and very relevant in many economic settings. Consider the set function $\psi^{\emptyset}(A, u)$ which returns u if $\mathbb{1}_A = 1$, $\emptyset$ otherwise. Characterise feedback ψ as a function which returns a tuple ψ_i at the end of period i, specifically, $\psi : (x, u, v) \mapsto (x, \mathbb{1}(x \geq v), \psi^{\emptyset}((x \geq v), u))$. This feedback structure imposes some asymmetry with respect to variable U. While feedback on variable V is always realistic (i.e. $\mathbb{1}(x \geq v)$), feedback on variable U is full whenever $\mathbb{1}(x \geq v)$ and empty otherwise. We say that U feedback is $\mathbb{1}(x \geq v)$-asymmetric. In Section 5, we show that Algorithm 1 circumvents this issue by relying only on the construction of unbiased estimates of $y_i = \mathbb{1}(x_i \geq v_i) \cdot u_i$,

which is effectively observed in every period under the following feedback structure $\phi_y : (x, u, v) \mapsto (x, \mathbb{1}(x \geq v), y)$.

There exists some interesting connections between feedback classes. First, observe that, setting approximation errors aside, under full feedback considerations (i.e. $u_{i-1}, v_{i-1} \in H_i$ for all i), the GMP becomes an instance of adversarial bandits, for which Tempered Exp3 Algorithm is known to be near-optimal with regret $\mathcal{O}(\sqrt{KB \ln B}) << \mathcal{O}(K^{\frac{2}{3}})$ [Cesa-Bianchi and Lugosi, 2006]. Similarly, we conjecture that under realistic feedback on both variables U and V in all periods our problem incurs in linear regret $\mathcal{O}(K)$. To justify our intuition, observe that the positive part of $u_i - x_i$ must be re-written as $\int_0^x \mathbb{1}(u \geq x)dx$. It now follows that the gradient of the welfare function depends globally on x and, as shown in [Cesa-Bianchi et al., 2021] and [Cesa-Bianchi et al., 2022], no algorithm can obtain sub-linear regret without strong assumptions on the DGP. Although realistic feedback yields higher regret in the GMP, more generally, there is no clear difficulty relationship between realistic feedback and some classes of asymmetric feedback. To see this, consider the talent-hunter monopoly problem (TMP)

$$S_i^{\text{TMP}}(x_i) = \mathbb{1}(x_i \geq v_i)(u_i - x_i) + \lambda_2 \cdot u_i \cdot x \tag{7}$$

In this context, the firm receives extra utility from making high salary offers to high productivity individuals regardless their employment decision. Standard proof devices[5] rely on the derivation of unbiased estimates of rewards at the end of every period. In this sense, it is not clear which feedback structure yields more efficient estimates: Observing $\mathbb{1}(u_i \geq x_i)$ every period or observing u_i fully conditional on $\mathbb{1}(x_i \geq v_i)$. We conjecture that previous and future research in adaptive policy design can benefit from asymmetric feedback considerations, especially considering its ubiquity in economics. For instance, [Cesa-Bianchi et al., 2021] characterises upper bounds for the bilateral trade problem under full and realistic feedback regimes. While the order of regret is very low $\mathcal{O}(\sqrt{K})$ in the full-feedback case, it is only linear $\mathcal{O}(K)$ in the realistic regime (in the absence of strong assumptions on the DGP). This gap across information structures hints that some asymmetric feedback classes may yield interesting regret bounds in canonical problems.

[5] See Appendix 9.1.

Integral Form and Information Requirements. Setting feedback considerations aside, the difficulty of the problem can also be analysed in terms of the information requirements on the learner[6]. In particular, the difficulty of the problem is connected to the dependence of the objective and its gradient to the policy x. To conduct this analysis, we adopt the notation in [Cesa-Bianchi et al., 2022]. Consider re-writing the GMP and SMP as

$$S_i^{\text{GMP}} = G_i^v(x_i) \cdot (u_i - x_i) + \lambda \int_0^x G_i^v(x') \; dx' \qquad S_i^{\text{SMP}} = G_i^v(x_i) \cdot (u_i - x_i) \tag{8}$$

where $G_i^v(x_i) = \mathbb{1}(x_i \geq v_i)$ is the *demand function*. We note that $\mathbb{1}(x_i \geq v_i)(x_i - v_i) = \max(x_i - v_i, 0) = \int_0^x G_i^v(x') \; dx'$. While the objective in the SMP depends pointwise in x, the objective in the GMP depends globally on x because the integral component relies on values x' away from x. The gradients for the expressions in equation (8) are given by

$$\nabla S_i^{\text{GMP}} = G^{v'}(x) \cdot (u - x) - (1 - \lambda) \cdot G^v(x) \qquad \nabla S_i^{\text{SMP}} = G^{v'}(x) \cdot (u - x) - G^v(x) \tag{9}$$

Both gradients depend locally on x. As a result, the SMP shares information requirements with the monopolist pricing problem in [Kleinberg and Leighton, 2003], while the GMP is most similar to the optimal taxation problem in [Cesa-Bianchi et al., 2022]. However, GMP and SMP remain weakly harder due to their multi-variable nature. We describe this difficulty relation precisely in Appendix 9.2, but intuitively the environment class of DGP in GMP/SMP is larger than the benchmark analogues, hence our problems are more difficult. In Section 5 we show that Algorithm 1 incurs in regret of order $\tilde{\mathcal{O}}(K^{\frac{2}{3}})$, the same bounds up to constant factors than the ones presented in [Cesa-Bianchi et al., 2022], thus our results extend the literature.

An Online Convex Optimisation Interpretation of the GMP. To conclude this section, we show briefly that GMP and SMP can be re-interpreted as OCO problems [Hazan et al., 2016]. In OCO, the learner is interested in minimising the regret loss,

$$\mathcal{R}_K^{\text{OCO}} = \sum_i^K f_i(x_i) - \min_{x \in \mathcal{K}} \sum_i^K f_i(x) \tag{10}$$

[6]Appendix 9.2 introduces a rigorous definition of difficulty in the Embedding Lemma sense.

where $\mathcal{K}$ is a convex decision set and f_i is a sequence of (possibly adversarial) convex loss-function realisations. Rewards and losses can be interchanged without loss of generality, hence the only concern comes from the convexity of the decision set provided some discretisation of the policy space Ω_X. This convexity can be established by redefining the policy space as the B arm simplex Δ_B. Under this transformation, expected losses can be defined as $\mathbb{E}[f_i(x)] = \sum_b p_{ib} \cdot f_{ib}(x)$, where p_{ib} is the probability of selecting arm b and f_{ib} is its associated one-period loss. This re-interpretation of GMP highlights again the necessity of introducing some degree of randomisation in the learner's policy. Moreover, it enables the importation of OCO results, especially the powerful heuristics of SGD-based algorithms. Unfortunately, OCO techniques (like SGD) typically rely on the gradient of the loss functions sequence $\nabla f_i(x_i)$, which may not be accessible in bandit problems where only $f_i(x_i)$ is recoverable[7].

To derive unbiased estimates of the gradient, the key insight is, once again, that the decision set is given by the simplex of the policy space. So,

$$\tilde{\nabla}_{ib} = \frac{1}{p_{ib}} \cdot \nabla_p f_i(x_i) = \frac{1}{p_{ib}} \cdot \nabla_p (f_{ib} \cdot p_{ib}) = f_{ib}(x_i) \frac{\mathbb{1}(x_i = x_b)}{p_{ib}} \tag{11}$$

It can be shown that $\tilde{\nabla}_i$ is an unbiased estimate of the true gradient of $f_i(x_i)$. We develop further intuitions in Section 5.2 and Appendix 9.1.

5 Bounds on Adaptive Adverse Selection

This Section introduces a near-optimal Hannan consistent algorithm for the GMP without previous knowledge of the DGP. As outlined in Section 4.2, this problem is most similar to the adaptive welfare maximisation problem [Cesa-Bianchi et al., 2022]. Hence, both the algorithm and proof devices rely heavily on their findings. Nevertheless, there are some technical intricacies which are discussed in Appendix 9.1.

Algorithm 1 Tempered Exp3 for the GMP

Input B, λ, η, γ
Set $x_b = (b-1)/B$ for $b \in \{1, 2, \ldots, B+1\}$, $\widehat{\mathbb{G}}_{1b} = 0$, $\widehat{\mathbb{U}}_{1b} = 0$ for all b
for $i = 1, 2, \ldots, K$
  for $b = 1, \ldots, B+1$
    Set $\widehat{\mathbb{S}}_{ib} = \widehat{\mathbb{U}}_{ib} - x_b \widehat{\mathbb{G}}_{ib} + \frac{\lambda}{B} \sum_{b'<b} \widehat{\mathbb{G}}_{ib'}$, $p_{ib} = (1-\gamma) \frac{\exp(\eta\, \widehat{\mathbb{S}}_{ib})}{\sum_{b'} \exp(\eta\, \widehat{\mathbb{S}}_{ib'})} + \frac{\gamma}{B+1}$
  end for
  Sample $b_i \sim p_{ib}$ and observe $\mathbb{1}(x_{b_i} \geq v_i)$
  If $\mathbb{1}(x_{b_i} \geq v_i) = 1$ observe u_i
  for $b = 1, \ldots, B+1$
    Update $\widehat{\mathbb{G}}_{i+1,b} = \widehat{\mathbb{G}}_{ib} + \mathbb{1}(x_{b_i} \geq v_i) \frac{\mathbb{1}(b_i = b)}{p_{ib}}$, $\widehat{\mathbb{U}}_{i+1,b} = \widehat{\mathbb{U}}_{ib} + u_i \mathbb{1}(x_{b_i} \geq v_i) \frac{\mathbb{1}(b_i = b)}{p_{ib}}$
  end for
end for

5.1 An Algorithm for the GMP

Algorithm 1 is a modification of standard Tempered Exp-3 algorithms. To give some heuristics, in each iteration Algorithm 1 calculates an unbiased estimate of the expected cumulative welfare $\mathbb{S}_{ib} = \sum_{j \leq i} S_{ib}$ for each arm x_b, using as inputs the sequence of $\mathbb{1}(x_i \geq v_i)$, $y_i = \mathbb{1}(x_i \geq v_i) \cdot u_i$. Then, Algorithm 1 selects an action x_b, based on the probability distribution p_{ib} that is induced by the welfare estimates. Note that this distribution is tempered, ensuring that every arm b is explored at least $\frac{\gamma}{B+1} \cdot K$ times in expectation. This exploration boost is necessary due to the integral element in the GMP, which is discussed in detail in Section 5.3 and [Cesa-Bianchi et al., 2022].

Algorithm 1 as Penalised Stochastic Gradient Descent. Standard (constrained) SGD algorithms perform the update step by setting $z_{i+1} = x_i + \eta_i \cdot \tilde{\nabla}_i$ where η_i is the learning rate in period i and $\tilde{\nabla}_i$ is the ith realisation of the random variable $\tilde{\nabla}$ such that $\mathbb{E}[\tilde{\nabla}_i] = \nabla_i$. Finally, it projects back into the convex decision set by setting $x_{i+1} = \Pi_{\mathcal{K}}(z_{i+1})$ where $\Pi_{\mathcal{K}}$ stands for the projection into the set $\mathcal{K}$. In GMP context, the decision set is just the probability simplex, thus, a sensible projection is the *soft-max* function such that $x_{i+1,b} = \frac{x_{ib} \cdot \exp(-\eta_i \tilde{\nabla}_{ib})}{\sum_b \exp(-\eta_i \tilde{\nabla}_{ib})}$. The update step in Algorithm 1 follows these intuitions with $-\tilde{\nabla}_{ib} = \widehat{\mathbb{S}}_{ib} \frac{\mathbb{1}(b_i = b)}{p_{ib}}$ and $\eta_i = \eta$ (up to a regularisation term $\frac{\gamma}{B+1}$). Appendix 9.1 shows that $\widehat{\mathbb{S}}_i$ is in fact an unbiased estimate of the true gradient.

[7] In fact, in GMP the learner does not even recover $f_i(x_i)$, but feedback ψ_i. In Section 5.2 we show that feedback ψ_i can be used to recover unbiased estimates of $\nabla f_i(x_i)$.

5.2 Upper Bounds on the Algorithm

Theorem 3. *Adversarial Upper Bound on Regret of Algorithm 1 for the GMP.* Consider a sequence $\{x_i\}_{i=1}^K$ as given by Algorithm 1 with parameters $\gamma = c_1 \cdot \left(\frac{log(K)}{K}\right)^{\frac{1}{3}}$, $\eta = c_2 \cdot \gamma^2$ and $B = \frac{c_3}{\gamma}$ for some $c_1, c_2, c_3 \in \mathbb{R}$. It now follows that for any arbitrary sequence $\{(u_i, v_i)\}_{i=1}^K$, there exist a constant $c_4 < \infty$ such that

$$\mathbb{E}[\sup_x \sum_i^K S_i(x) - \sum_i^K S_i(x_i)] \leq c_4 \cdot \log(K)^{\frac{1}{3}} \cdot K^{\frac{2}{3}} \tag{12}$$

Intuition for the proof builds closely to the arguments in [Cesa-Bianchi and Lugosi, 2006], [Lattimore and Szepesvári, 2020] and [Cesa-Bianchi et al., 2022]. I refer the reader to Appendix 9.1 for a full version of the proof. In a nutshell, setting approximation errors aside, the derivation is based on bounding welfare S_i in terms of welfare estimates $\hat{S}_i$. We then establish bounds on the first and second order moments of welfare estimates $\hat{S}_{ib}$. Once all bounds are set, inequalities are all put together, and parameters are tuned accordingly to achieve the desired result.

Theorem 4. *Stochastic Upper Bound on Regret of Algorithm 1 for the GMP.* Consider a sequence $\{x_i\}_{i=1}^K$ as given by Algorithm 1 with parameters $\gamma = c_1 \cdot \left(\frac{log(K)}{K}\right)^{\frac{1}{3}}$, $\eta = c_2 \cdot \gamma^2$ and $B = \frac{c_3}{\gamma}$ for some $c_1, c_2, c_3 \in \mathbb{R}$. It now follows that for any sequence $\{(u_i, v_i)\}_{i=1}^K$ sampled *iid* from $F_{U,V}$, there exist a constant $c_4 < \infty$ such that

$$\mathbb{E}[\sup_x \sum_i^K S_i(x) - \sum_i^K S_i(x_i)] \leq c_4 \cdot \log(K)^{\frac{1}{3}} \cdot K^{\frac{2}{3}} \tag{13}$$

Proof: Intuitively, *iid* sampling is a (very) special case of arbitrary sequence, hence any upper bound on adversarial regret immediately returns an upper bound on the stochastic problem. Equivalently, the maximum is larger than the average. Formally, define the stochastic regret in Theorem 4 as $\mathcal{R}(\pi, \nu)$, and the adversarial regret as $\mathcal{R}(\pi, \nu')$, with $\nu \subset \nu'$

$$\mathcal{R}(\pi, \nu) = \sup_x \mathbb{E}\Big[\sum_i^K S_i(x) - S_i(x_i)\Big] \leq \mathbb{E}\Big[\sup_x \sum_i^K S_i(x) - S_i(x_i)\Big] = \mathcal{R}(\pi, \nu') \tag{14}$$

where the first equality serves as definition of the stochastic regret and the inequality is given by Jensen's inequality. Theorem 4 follows as a corollary of Theorem 3, provided equation (14). $\square$

5.3 Lower Bound on the GMP

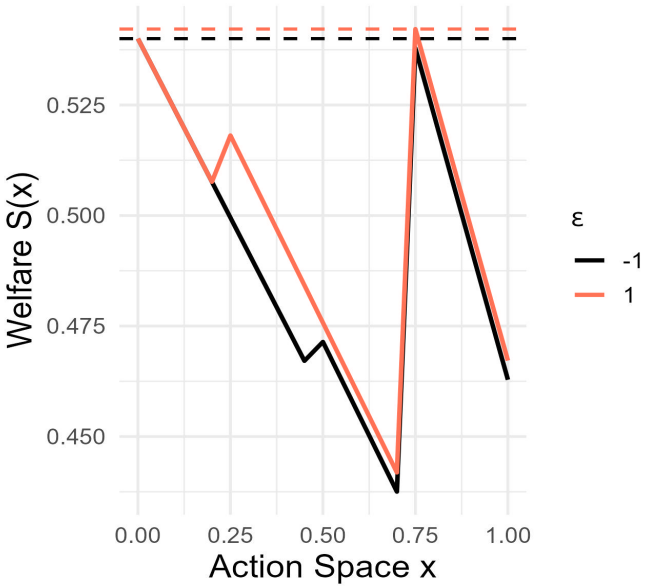
Theorem 5. *Stochastic Lower Bound on Regret of the GMP.* Consider the problem of sequentially choosing $\{x_i\}_{i=1}^K$ where rewards are given by $S_i(x) = \mathbb{1}(x_i \geq v_i)(u_i - x_i + \lambda(x_i - v_i))$ and the set-up described in Section 2. There exists a constant $C > 0$ such that for any policy π, and horizon $K \in \mathbb{N}$ there exists a distribution $F_{U,V}$ over outcomes (u_i, v_i) such that $\mathbb{E}_{F_{U,V}}[\mathcal{R}(\pi)] \geq C \cdot K^{\frac{2}{3}}$.

Excess of regret of the GMP in comparison to canonical adversarial bandit frameworks $\mathcal{O}(\sqrt{K})$ is attributed to the global information requirement of the objective function (i.e. the integral component). Intuitively, the adversary could create a distribution $F_{U,V}^{\epsilon}$ that necessitates the policymaker to infer the sign of ϵ to determine the optimality between policies x' and x''. In particular, $F_{U,V}^{\epsilon}$ can be defined in a way such that the sign of ϵ can only be inferred by sampling from a clearly sub-optimal region. Shall the policymaker not investigate this policy space, she will incur in linear regret. Figure 1 illustrates this phenomenon, where the only two candidates to optimal policy are $x = 0$ and $x = 3/4$, but the policymaker needs to sample from the sub-optimal region $[1/4, 1/2]$ to infer the sign of ϵ. I refer the reader to [Cesa-Bianchi et al., 2022] for an extensive discussion on the topic.

For borrowing some of the results in the literature, we first need to transform the two variable framework into a canonical one variable setting. This can be achieved by a simple application of the Embedding Lemma. The Embedding Lemma claims that, provided a difficulty relationship between two games, the easier game has uniformly no higher regret than the difficult game. In this case, we show that the GMP is more difficult in the Embedding Lemma sense than a comparable one variable game. I refer the reader to Appendix 9.2 for a full description of the proof.

Corollary 6. *Adversarial Lower Bound on Regret of the GMP.* Consider the problem of sequentially choosing $\{x_i\}_{i=1}^K$ where rewards are given by $S_i(x) = \mathbb{1}(x_i \geq v_i)(u_i - x_i +$

Figure 1: ϵ-dependent Lower Bound for the GMP with $U = 1, V \sim F_v^\epsilon$



F_v^ϵ is parameterised with $b = (1-\lambda)/(48-34\lambda), a = (3b\lambda+1)/(4-3\lambda)$, $\lambda = 0.7$.

$\lambda(x_i - v_i)$) and the set-up described in Section 2. There exists a constant $C > 0$ such that for any policy π, and horizon $K \in \mathbb{N}$ there exists an arbitrary sequence $(u_i, v_i)_{i=1}^K$ such that $\mathcal{R}(\pi) \geq C \cdot K^{\frac{2}{3}}$. *Proof:* This result follows from Theorem 5 and equation (14). Alternatively, this result can be derived as a trivial application of the Embedding Lemma in Appendix 9.2. □

Corollary 7. *Hannan Consistency of Algorithm 1.* Policy π embedded in Algorithm 1 is Hannan Consistent for any arbitrary distribution $\{u_i, v_i\}_{i=1}^K$. *Proof:*

$$\limsup_{K\to\infty} \frac{1}{K}\Big(\sup_{x\in\Omega_X} \sum_i^K S_i(x) - \sum_i^K S_i(x_i)\Big) \leq \limsup_{K\to\infty} \frac{1}{K} \cdot c_4 \cdot \log(K) \cdot K^{\frac{2}{3}} = 0 \qquad (15)$$

where the inequality is given by equation (12) and the equality holds with probability 1 as $K \to \infty$. □

Corollary 8. *Convergence in Probability to the Monopolist Equilibrium with Partial Information and Correct Beliefs.* The empirical probability $p_{x^*}^\pi$ of playing the BNE x^*

by an agent who implements policy π embedded in Algorithm 1 $\xrightarrow{p} 1$ as $K \to \infty$. *Proof:* This Corollary follows from Corollary 7 and Proposition 2. $\square$

Observe that Corollary 8 is an asymptotic result and, consequently, it does necessarily hold for fixed K and fixed parameter values. In fact, the empirical distribution of probabilities $p_x^{\pi,K} \to \propto \exp(\mathbb{S}_{ib})$ with $p_{x^*}^{\pi,K} \to \propto \exp(\mathbb{S}_{ib^*}) < 1$ for any $K < \infty$.

Corollary 9. *Optimality of Algorithm 1.* We have presented matching upper and lower bounds, hence Algorithm 1 is essentially unimprovable (up to logarithmic factors) under adversarial and stochastic specifications in the GMP.

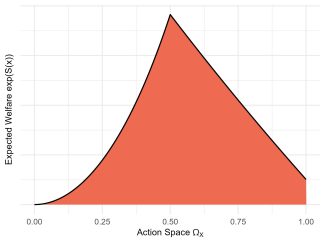
6 Simulations

This section contains an extensive simulation testing of Algorithm 1 with a focus on the welfare implications of sub-optimal parameter selection. To examine its empirical performance, we first need to identify the best fixed policy in hindsight analytically. As discussed earlier, identification remains obscure even in some stylised stochastic contexts. This fact is not surprising, as one of the motivations for the adoption of SGD based algorithms to best response identification was the analytical intractability of conventional theory. We attempt to convince the reader of the robustness of Algorithm 1, and to extend that confidence beyond analytically tractable DGP.

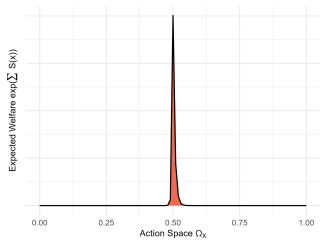
Uniform Linear Degenerate Case (ULDG). In this section, we explore the stochastic ULDG case with $U \sim \mathcal{U}[0,1]$ and $v = 0.5 \cdot U$. I refer the reader to Appendix 9.4 for further analyses under a broader range of DGP. The reader may notice that the theoretical upper bounds derived in Section 5.2 are not tight to the stochastic ULDG context, thus, in principle, there may exist convergence rates faster than $K^{-\frac{1}{3}}$. The best fixed policy in hindsight is given by

$$\arg\max_x \mathbb{E}_{U,V}\big[\mathbb{1}(x \geq v) \cdot \big((u-x) + \lambda(x-v)\big)\big] = \int_0^{2x} u - x + \lambda\Big(x - \frac{1}{2}\Big) du \qquad (16)$$

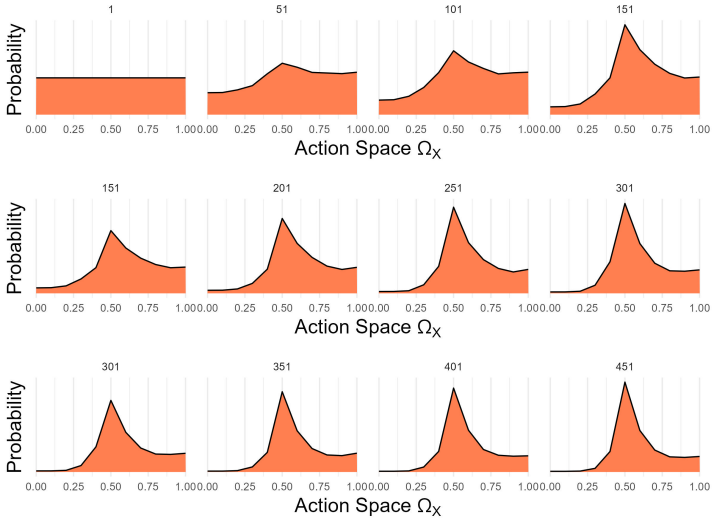
Figure 2: Algorithm 1 with $U \sim \mathcal{U}[0,1]$, $v = 0.5u$



(a) $\mathbb{E}_{U,V}[\exp(S_i(x))]$



(b) $\mathbb{E}_{U,V}[\exp(\mathbb{S}_i(x))]$



(c) Empirical Distribution of Probabilities across K

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$.

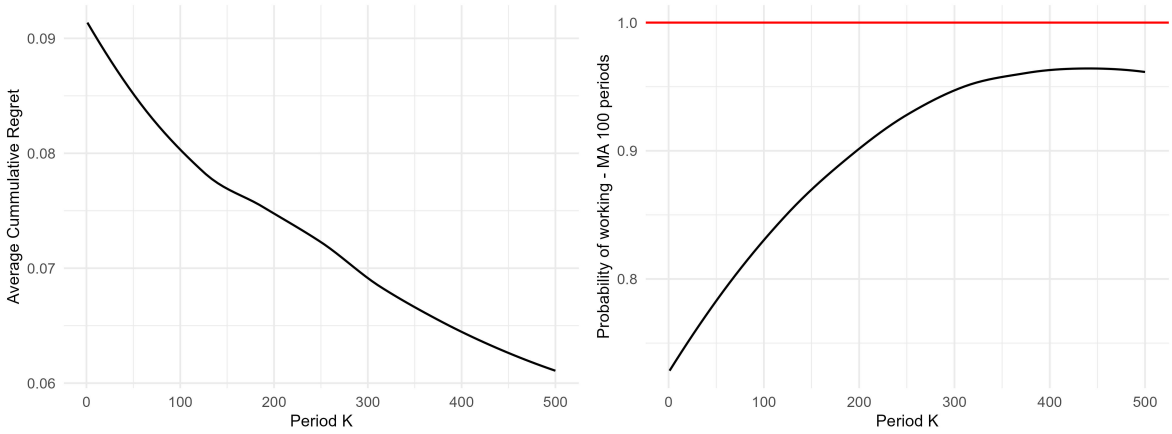
Consider separately the cases where $x \in (1/2, 1]$ and $x \in [0, 1/2]$,

$$\mathbb{E}[S_i(x)] = \begin{cases} \lambda x^2 & x \in [0, 1/2] \\ 1/2 - x + \lambda \cdot x - \lambda/4 & x \in (1/2, 1] \end{cases} \tag{17}$$

It follows that $x^* = 1/2$ (see Figure 2a). As discussed in the observations to Corollary 9, for fixed K and fixed parameters we expect some degree of convergence (up to the discretisation error) between the empirical distribution of probabilities and $\exp(\mathbb{S}_{ib})$ (see Figures 2b and 2c).

Empirical Performance Analysis of Algorithm 1. Figure 3a shows a decrease in average regret in line with the theoretical prescriptions of Section 5.2. Nevertheless, empirical rates could differ from theory bounds for two reasons. The upper bound was established for an adversarial sequence of outcomes, hence it might not be tight to the ULDC. Moreover, Algorithm 1 was tuned with sub-optimal parameters. Overall, simulations could reflect faster convergence rates due to the upper bounds being overly generous or slower convergence rates due to sub-optimal parameter selection.

Figure 3: Algorithm 1 given $U \sim \mathcal{U}[0,1]$, $v = 1/2u$



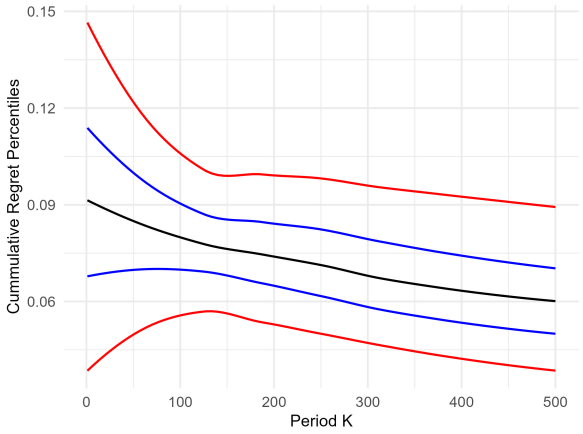
(a) Average Cumulative Regret (b) Probability of Work $\mathbb{P}(v \leq x)$

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$. Moving Average = 100 periods.

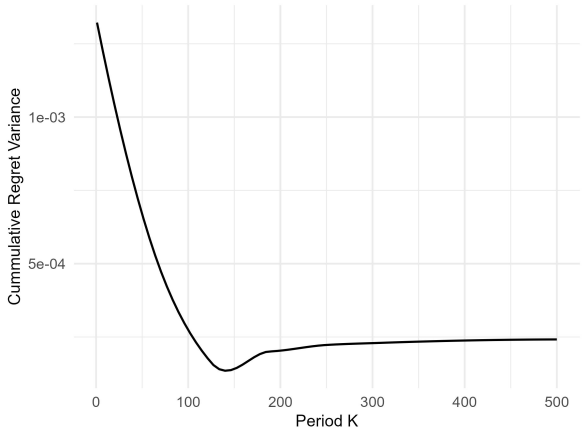
We expect similar convergence dynamics between the induced empirical probability of working and the expected probability of working at the optimal wage $\mathbb{P}(v \leq x^*)$. This is because the decision of working J_i is a bounded degenerate function of x conditional on target variables, thus the same regret order bounds hold for $1/K \sum_i^K J_i(x^*) - 1/K \sum_i^K J_i(x_i)$. Findings in Figure 3b are consistent with these prescriptions. The simulations confirm that the distribution over actions induced by Algorithm 1 converge to the equilibrium strategy in the GMP with Partial Information and correct beliefs without prior knowledge of the joint distribution of target variables.

Further Analysis. One may analyse alternative moments of the empirical cumulative regret beyond the expectation, like the q-percentile regret (Figure 4a). Our findings suggest that the cumulative regret is not only higher during the initial periods, but it also exhibits higher variance. This is due to the gradual updating process of the sampling probability distribution, which is regulated by parameter η and the softmax function. As a result, Algorithm 1 effectively samples at random during the first periods, leaving the performance of the algorithm to chance. We confirm these intuitions by plotting the cross-simulation variance in Figure 4b.

Figure 4: Further Analyses of Algorithm 1 given $U \sim \mathcal{U}[0,1]$, $v = 0.5u$



(a) Average Regret, p5-p95 (red), p25-p75 (blue), p50 (black)



(b) Regret Variance across Simulations

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$. Moving Average = 100 periods.

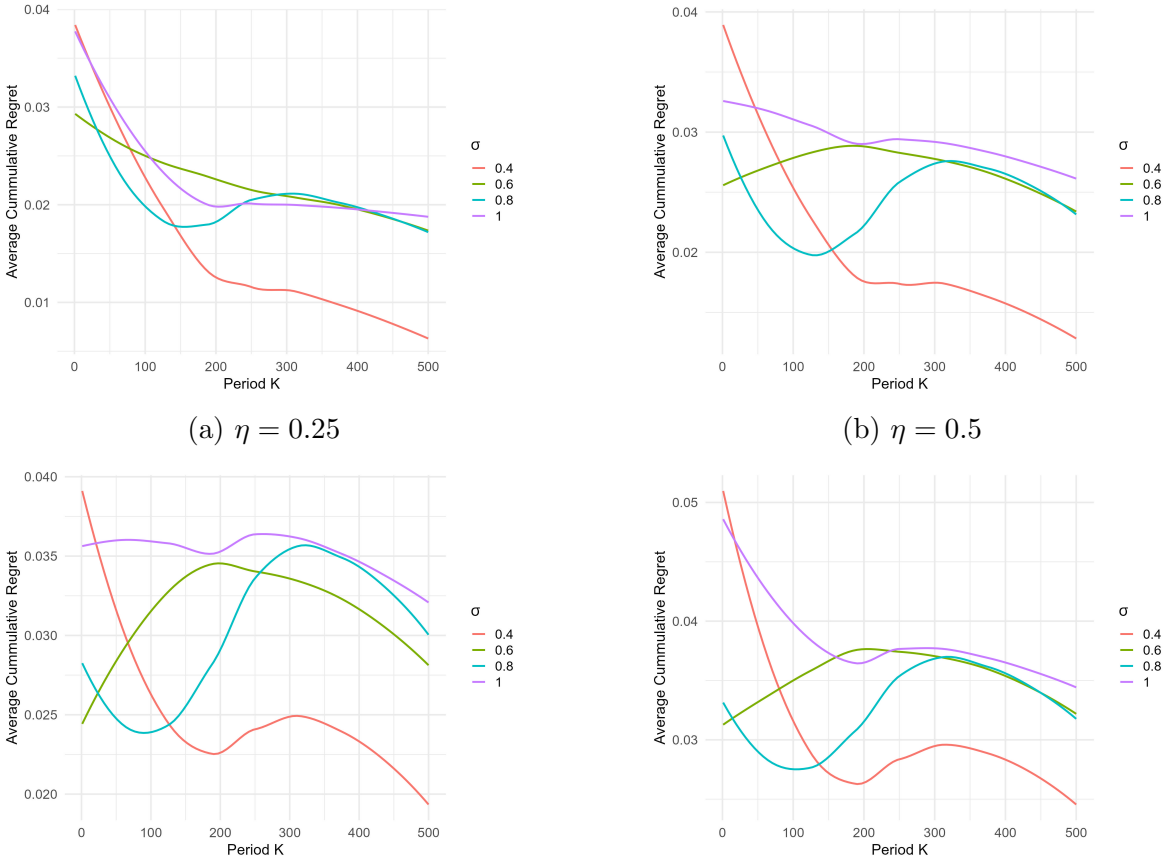
Welfare implications of sub-optimal parameter selection. To conclude this section, it is worth mentioning that the policymaker should be careful with parameter selection. The tuned parameters outlined in Appendix 9.1 ensure asymptotic optimality, but in practice, they are overly conservative. As a result the policymaker might be tempted to "speed-up" the convergence process in some especial contexts. For instance, empirical performance of the algorithm in the ULDC with $K = 500$ benefited from a parameter $\eta = 0.132$ much bigger than the asymptotic optimal $\eta^* \leq 0.0027$. However, this greedy acceleration can go badly shall the policymaker not be aware of the variance of the DGP. From a theory standpoint, this intuition turns out to be correct provided that upper-bound derivation, mediated by the learning rate η, relies on bounding the second order moment of the gradient. In words, in very noisy DGP, it is optimal to restrict the effect of a single realisation on the sampling probabilities. To illustrate this phenomenon, consider the case where $U \sim \mathcal{U}[0,1]$ and $V = U + \phi_\sigma$ where $\phi_\sigma \sim N(0, \sigma^2)$. It follows that

$$\mathbb{E}[S_i(x)] = -\frac{1}{20}(1-\lambda)(2x-1)\texttt{erf}\Big(\frac{x}{\sqrt{2}\cdot\sigma}\Big) + \frac{\lambda\cdot\sigma\cdot(e^{\frac{-x^2}{2\sigma^2}}-1)}{5\sqrt{2\pi}} \tag{18}$$

where `erf` is the error function. Figure 5 shows average cumulative regret across

four different ϕ_σ distributions and four different η values. We observe how an increase in η leads to a uniform decrease in performance of the algorithm. In addition, more volatile regimes do not only perform worse with high η values but their performance seems disproportionately affected by an increase in η. For instance we do not observe a decrease in average cumulative regret for $\phi_\sigma = 0.6$ and $\phi_\sigma = 0.8$ in the $\eta = 1$ and $\eta = 2$ cases. We conclude that sub-optimal greedy parameter selection can lead to important reductions in welfare.

Figure 5: Algorithm 1 given $U \sim \mathcal{U}[0,1]$, $V = u + \phi_\sigma$



(a) $\eta = 0.25$

(b) $\eta = 0.5$

(c) $\eta = 1$

(d) $\eta = 2$

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $B = 10$, $\gamma = 0.029$

We refer the reader to Appendix 9.4 for additional empirical testing of Algorithm 1 under alternative DGP. More specifically, we show empirical convergence in generalised $Beta[\alpha, \beta]$ distributions (of which $\mathcal{U}[0,1]$ is a special case with $\alpha = 1, \beta = 1$), discrete distributions, independent distributions, and non-linearly dependent distribu-

tions. Theoretical prescriptions are matched under sub-optimal parameter selection for all DGP considered in Appendix 9.4.

7 Economic Modelling and Policy Implications

Our model is subject to significant limitations when it comes to studying the problem of the firm. This is due to a number of implicit assumptions, such as the linearity of the production function with respect to the sum of workers' productivity, the existence of an infinite pool of workers, no meaningful constraints to firm's budget, and the strict monopsony nature of the firm. Nevertheless, even in this highly stylised setting, we can analyse some theoretical prescriptions of the model and compare them with mainstream offline theory and empirical evidence in Economics. In particular, we examine the welfare implications of two shocks which are very relevant within the Labour Economics and Applied Macroeconomics literature: Minimum Wage policies and Limited Information Processing Capacity (LIPC) [Dube, 2019b] [Maćkowiak et al., 2023][8]. We acknowledge that these two events can be modelled in different ways, hence, our theory prescriptions are not final nor robust. Rather, the reader should approach this Section as a demonstration of the economic modelling potential of online learning tools within economically relevant environments in the presence of uncertainty.

7.1 Minimum Wage Analysis

Minimum Wage in Monopsony Labour Markets. There is a large body of evidence pointing at very small (if any) employment effects following an increase in minimum wage [Cengiz et al., 2019] [Dube, 2019a] [Wolfson and Belman, 2019]. This literature challenges the prescriptions of perfect competition labour markets and hints that firms may possess monopsony wage setting power. Indeed, the Applied Macroeconomics and Industrial Organisation literature have extensively documented the existence of monopsony power in labour markets [Furman and Orszag, 2018] [Manning, 2003] [Manning, 2021]. In addition, there has been a growing interest in the impact of uncer-

[8]In the extended version of this paper, we also explore the impact of productivity shocks, and analyse the recovery time of firms when the shock hits at different periods in time. For space considerations, we have not included such results in this manuscript.

tainty on poverty and inequality in non-competitive labour markets [Dube et al., 2016] [Card et al., 2012]. Despite its simplicity, our model stands as an obvious choice for analysing the impact of minimum wage policies in the presence of uncertainty and monopsonic behaviour. To the best of our knowledge, it is the first model that leverages online learning technologies to provide theoretical prescriptions without imposing any information requirements on the firm about the labour supply curve of workers. We examine the theoretical prescriptions of our model and compare them to standard offline theory results and empirical evidence.

Problem Characterisation. In this context, we define a minimum wage shock as the reduction of the policy space from Ω_X to $\Omega_{X,\text{MW}}$, where $\Omega_{X,\text{MW}} = [\min \Omega_{X,\text{MW}}, 1]$ with $\min \Omega_{X,\text{MW}} > 0$. Because the modified policy space satisfies the conditions described in Section 2, Algorithm 1 remains near optimal for the MW case, under unknown regret bounds. Welfare implications of minimum wage policies are discussed in light of the variation in adversarial regret bounds compared to the benchmark case. Nonetheless, the analysis distinguishes between different environments where the policy shock is expected to have varying levels of impact, either more or less damaging.

As shown in Appendix 9.1, setting approximation error aside, upper bound derivation for Algorithm 1 relies on bounding first and second order moments for the welfare estimates $\widehat{\mathbb{S}}_i$. Bounds and estimates remain unchanged, thus, the only modification comes from bounding

$$\begin{aligned} |\sup_x \tilde{S}_i(x) - \max_{x_b \in \Omega_{X,\text{MW}}} \tilde{S}_i(x_b)| &\leq |\sup_x \tilde{S}_i(x) - \max_{x_b \in \Omega_X} \tilde{S}_i(x_b)| + |\max_{x_b \in \Omega_X} \tilde{S}_i(x_b) - \max_{x_b \in \Omega_{X,\text{MW}}} \tilde{S}_i(x_b)| \\ &\leq 1/B + |\max_{x_b \in \Omega_X} \tilde{S}_i(x_b) - \max_{x_b \in \Omega_{X,\text{MW}}} \tilde{S}_i(x_b)| = 1/B + D_i \end{aligned} \tag{19}$$

where the last equality serves as the definition of D_i.

Context Specific Bounds. To derive tight bounds for D_i one can distinguish across four period types. For simplicity we assume that both regimes Ω_X and $\Omega_{X,\text{MW}}$ are discretised under the same B (i.e. assume that the MW restriction of the policy space is produced **after** the discretisation of the policy space). For notation simplicity, define $x_{ib^*} = \arg\max_{x_{ib} \in \Omega_X} \tilde{S}_i(x_{ib})$, and $x_{ib^*,\text{MW}} = \arg\max_{x_{ib} \in \Omega_{X,\text{MW}}} \tilde{S}_i(x_{ib})$. Consider now the

following cases,

- $x_{ib^*} \in \Omega_{X,\text{MW}}$. It follows that $S_i(x_{ib^*}) = S_i(x_{ib^*,\text{MW}})$. $D_1 = 0$.
- $x_{ib^*} \notin \Omega_{X,\text{MW}}$. In this case, we distinguish between $x_{ib^*} < v_i$ and $x_{ib^*} \geq v_i$. In the first case, we distinguish between
  - $x_{ib^*,\text{MW}} < v_i$. It now follows that $S_i(x_{ib^*,\text{MW}}) = S_i(x_{ib^*}) = 0$ and $D_2 = 0$
  - $x_{ib^*,\text{MW}} \geq v_i$ It follows that $S_i(x_{ib^*,\text{MW}}) = u_i - x_{ib^*,\text{MW}} + \lambda(x_{ib^*,\text{MW}} - v_i) \leq 0$, otherwise $S_i(x_{ib^*,\text{MW}}) > S_i(x_{ib^*}) = 0$. $D_i = S_i(x_{ib^*,\text{MW}})$ can be bounded by $D_3 = x_{ib^*,\text{MW}}$, with $u_i = 0$, $v_i = x_{ib^*,\text{MW}}$.

  In the $x_{ib^*} \geq v_i$ case, observe that $x_{ib^*,\text{MW}} \geq x_{ib^*}$, hence $\mathbb{1}(x_{ib^*,\text{MW}} \geq v_i) = \mathbb{1}(x_{ib^*} \geq v_i) = 1$. D_i can be bounded by $D_4 = (1-\lambda)(x_{b^*,\text{MW}} - x_{b^*})$. Define $D_{\max} = \max\{D_1, D_2, D_3, D_4\}$. Under adversarial specifications, we can rewrite equation (32) in Appendix 9.1 as

$$\sup_x \mathbb{S}(x) - \mathbb{E}[\mathbb{S}(x_i)] \leq$$
$$K\left(\gamma + D_{\max} + \frac{(1+\lambda)}{B} + \eta(e-2)\frac{B}{B+1}\left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma}\right)\right) + \frac{\log(B+1)}{\eta} + 1 \tag{20}$$

Analysis of Theory Prescriptions. The expected difference in cumulative welfare between Ω_X and $\Omega_{X,\text{MW}}$ scenarios is of order $K \cdot D_{\max}$. The analysis distinguishes among four different possibilities. In two of them (D_1 and D_2), the minimum wage policy does not bite, so they yield zero loss, and, in the other two (D_3 and D_4) loss is strictly positive. In D_3 welfare loss is driven by failing to exclude non-profitable workers from the labour market (those with $u_i < v_i$), and in D_4 by not squeezing workers surplus as much as desired from the firm's perspective. Loss of profits associated to D_3 could be interpreted as undesirable, while loss of profits associated to D_4 as desirable. However, we leave that discussion for future research and analyse both sources of welfare loss as equally concerning. If $D_{\max} > 0$ under adversarial specifications, no parameter tuning can obtain sub-linear regret, therefore cumulative welfare loss is unbounded as $K \to \infty$. This is to be expected. Proposition 2 demonstrates that the sequence of actions induced by a policy needs to converge in probability to the equilibrium strategy, so if this strategy is not within the policy space, then convergence cannot occur.

Our analysis reveals that MW increases workers' surplus monotonically (through employment and salary mechanisms). This prescription differs from the standard MW monopsony theory, where the increase in workers' welfare is non-monotonic in MW. Nonetheless, this monotonicity is simply an artificial byproduct of our model, as the firm is always compelled to make a wage offer $x \geq \min \Omega_{X,\text{MW}}$ without any shut-down conditions. Notably, empirical evidence shows very small employment effects following large increase in MW (above 66% of the median wage), hence our model is compatible with these findings [Dube, 2019b].

On the other hand, because our definition of welfare $\mathbb{S}$ is firm-oriented ($\lambda < 1$), MW decreases overall welfare also in a weakly monotonic fashion. Shall alternative definitions of welfare be considered, this might not necessarily be the case. Regardless, we can analyse from a stochastic perspective the scenarios which minimise profit loss. In other words, we can identify under which populations $F_{U,V}$, the firm would suffer less following an increase in MW. To do so, we start by noting that setting discretisation errors aside, and assuming that $x_{ib^*} = v_i - \epsilon$ for $\epsilon > 0$, whenever $v_i > u_i$, the $\mathbb{P}(x_{ib^*} \in \Omega_{X,\text{MW}}) \approx \mathbb{P}(v_i \in \Omega_{X,\text{MW}})$[9].

Favourable cases (D_1 and D_2) are given by events $x_{ib^*} \in \Omega_{X,\text{MW}}$ and $(x_{ib^*} \notin \Omega_{X,\text{MW}} \cap x_{ib^*} < v_i \cap x_{ib^*,\text{MW}} < v_i)$. So, $\mathbb{P}(D_i = D_1) = \mathbb{P}(v_i \notin \Omega_{X,\text{MW}})$. For the second case, $\mathbb{P}(x_{ib^*} \notin \Omega_{X,\text{MW}} \cap x_{ib^*} < v_i \cap x_{ib^*,\text{MW}} < v_i) \approx \mathbb{P}(v_i \notin \Omega_{X,\text{MW}}) \cdot \mathbb{P}(x_{ib^*} < v_i \mid v_i \notin \Omega_{X,\text{MW}}) \cdot \mathbb{P}(x_{ib^*,\text{MW}} < v_i \mid v_i \notin \Omega_{X,\text{MW}}, x_{ib^*} < v_i) = 0$. Where the last equality follows from $\mathbb{P}(x_{b^*,\text{MW}} < v_i \mid v_i \notin \Omega_{X,\text{MW}}) = 0$. In summary, the event $D_i = D_2$ has probability zero, and the probability of the event $D_i = D_1$ is increasing in the density of high reservation wage values V.

We follow the same strategy to examine the comparative statics of D_3 and D_4. $\mathbb{P}(D_i = D_3) = \mathbb{P}(x_{ib^*} \notin \Omega_{X,\text{MW}} \cap x_{ib^*} < v_i \cap x_{ib^*,\text{MW}} \geq v_i) \approx \mathbb{P}(v_i \notin \Omega_{X,\text{MW}}) \cdot \mathbb{P}(x_{ib^*} < v_i \mid v_i \notin \Omega_{X,\text{MW}}) = \mathbb{P}(v_i \notin \Omega_{X,\text{MW}}) \cdot \mathbb{P}(u_i < v_i \mid v_i \notin \Omega_{X,\text{MW}})$. The probability of this event decreases in U and in the worker's productivity to reservation wage gap (PRWG) $(U - V)$. Nonetheless, V enters in a non-linear way. While high V reduces the probability of the first term, it increases the probability of the second one. In addition, remember that $D_3 = (1 - \lambda) \cdot x_{b^*,MW} - u_i + \lambda v_i$. So, while an increase in U leads to a decrease in the probability of the event, there is an ambiguous effect on

[9] If $v_i \in \Omega_{X,\text{MW}}$ and $v_i \geq u_i$, then $x_{ib^*} = v_i$ so $x_{ib^*} \in \Omega_{X,\text{MW}}$. Shall $v_i \in \Omega_{X,\text{MW}}$ and $v_i < u_i$, then $x_{ib^*} = v_i - \epsilon$ by assumption, so $x_{ib^*} \in [\min \Omega_{X,\text{MW}} - \epsilon, 1]$

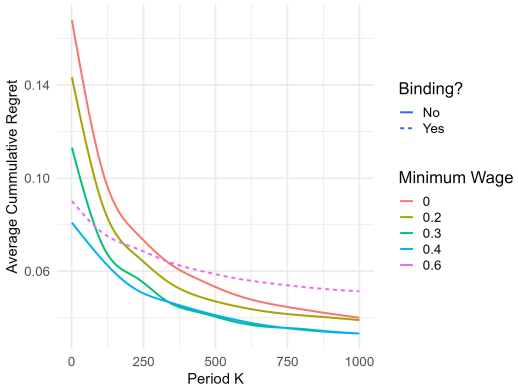
$\mathbb{P}(D_i = D_3) \cdot D_3$. Finally, $\mathbb{P}(D_i = D_4) = \mathbb{P}(x_{ib^*} \notin \Omega_{X,\text{MW}} \cap x_{ib^*} \geq v_i) = \mathbb{P}(v_i \notin \Omega_{X,\text{MW}}) \cdot \mathbb{P}(u_i \geq v_i \mid v_i \notin \Omega_{X,\text{MW}})$. In this case, the value of D_4 does not depend on target variables. $\mathbb{P}(D_i = D_4)$ decreases in V, and increases in U and PRWG. Overall, the size of the expected welfare loss in the two damaging cases relies heavily on the context specific $F_{U,V}$.

Comparative statics of $F_{U,V}$ on MW damages remain obscure, but, in general we might expect a negative correlation between reservation wages and MW-related welfare loss. Mathematically, an increase in population reservation wages V augments the $\mathbb{P}(D_i = D_1)$, lowers $\mathbb{P}(D_i = D_4)$, and has an ambiguous effect on the welfare loss associated to D_3. To give some intuition on why this might be the case, remember that in our model all workers are offered a wage, so low wage offers operate as entry barriers for low PRWG workers, preventing them from entering the labour force. MW policies remove these barriers, leading to an increase in employment and workers' surplus at the expense of monopolist profits. In this context, population's high reservation value serve as natural barriers which keep low productivity workers away from employment. This prescription aligns with standard theory in monopsony labour markets. To see this, consider an upward shift of the labour supply curve driven by a general increase in population reservation wages, which raises monopsony equilibrium wage from x_0^* to x_1^*. As a result, the loss in monopsonist welfare following a further increase in equilibrium wage driven by a MW policy will affect the monopolist less severely in the high V case compared to the low V one. In terms of workers' productivity (conditional on reservation wage), the model remains ambiguous. Empirically, we do not observe variation in employment stocks nor labour force composition following an increase in MW, most likely due to small extensive elasticities among potential MW earners [Cengiz et al., 2022].

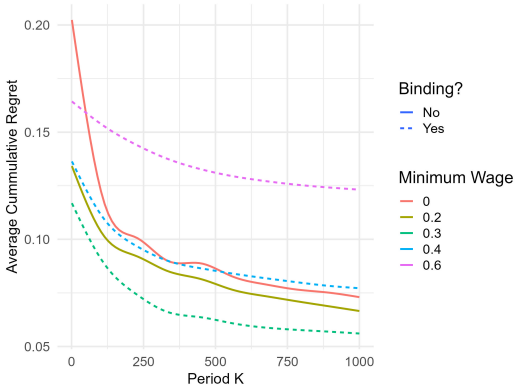
Simulation Analysis. Figure 6 summarises average cumulative regret across five different Minimum Wage policies ($0, 0.2, 0.3, 0.4$ and 0.6) and four different joint distributions $F_{U,V}$. Minimum wage policies are binding ($x_{ib^*} \in \Omega_{X,MW}$) or not depending on the distribution $F_{U,V}$. Joint distributions have been selected to differentiate across high/low U and high/low V scenarios.

Simulations follow theoretical prescriptions in at least three dimensions. (i) When MW is not binding ($D_i = 0$ cases) there is asymptotically no welfare loss associated

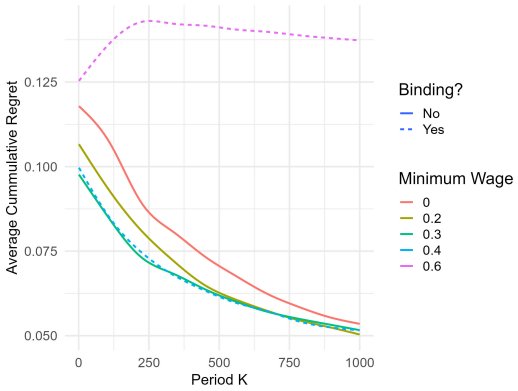
Figure 6: Algorithm 1 under MW restrictions



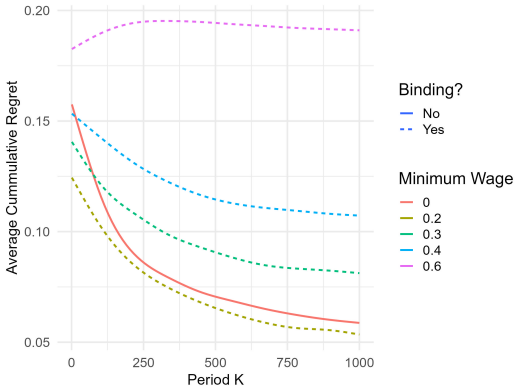
(a) $U = \mathcal{U}[1/2, 1], V = 0.5u, x_{b^*} = 1/2$



(b) $U = \mathcal{U}[1/2, 1], V = 0.25u, x_{b^*} = 1/4$



(c) $U = \mathcal{U}[1/4, 3/4], V = 0.5u, x_{b^*} = 3/8$



(d) $U = \mathcal{U}[1/4, 3/4], V = 0.25u, x_{b^*} = 3/16$

Average across 1,000 simulations. $\lambda = 0.7$. $K = 1000$. $B = 10$, $\eta = 0.132$ $\gamma = 0.029$

to the implementation of this policy. (ii) The implementation of large MW policies is associated to significant decrease in profits. This finding is aligned with standard theory and empirical evidence [Draca et al., 2011]. (iii) Firms exposed to populations with higher reservation wages V suffer less following an implementation of MW (Figure 6a vs. Figure 6b, and Figure 6c vs. Figure 6d). In addition, simulations show that productivity U plays an ambiguous role in welfare loss. The most interesting result comes from the implementation of MW with small "policy bite" (i.e. $x_{b^*} \notin \Omega_{X,\text{MW}}$ but $x_{b^*} > \min \Omega_{X,\text{MW}} - \epsilon$ for some small ϵ). Asymptotically we expect not-binding policies to outperform binding ones. However, in finite samples welfare loss associated to sub-optimal action selection is compensated by the fact that the policymaker cannot explore arms below x_{b^*}, which also turn out to be sub-optimal. This result could be

interpreted as "information gains" of MW implementation, and suggests that in some contexts it might be beneficial not to allow the firm to "explore" excessively low wages.

Overall, the theoretical prescriptions of our model are confirmed via simulations. In addition, most of its implications match classic monopsony theory and current empirical findings. An interesting takeaway is the increase in welfare associated to MW information gains. Note that these results have been established without any priors on the labour supply curve of workers, what constitutes an extension to labour economic theory. A finer characterisation of the problem of the firm is likely to provide more accurate prescriptions.

7.2 Limited Information Processing Capacity Analysis

A growing interest has emerged around the role of imperfect information and costly information acquisition in many areas of Economics including macroeconomics [Sims, 2003], labour economics [Acharya and Wee, 2020], finance [Van Nieuwerburgh and Veldkamp, 2010] and others. Notably, models which shape information acquisition as a strategic costly decision, like rational inattention, have gained traction as promient tools within the field [Maćkowiak et al., 2023]. These models assume that despite information's availability, its incorporation within decision processes can be costly or restricted.

Modelling Information Constraints. There are many ways limited information can be incorporated into the GMP. A non exhaustive list includes: Blackwell less-informative feedback, imperfect recall (predictions $\widehat{\mathbb{S}}_{ib}$ are built using a number of periods $i' < i$), and managerial rigidity (the firm can only update $\widehat{\mathbb{S}}_{ib}$ in some periods). All these examples are connected by the notion of firm's limited information processing capacity (LIPC), which can be justified from behavioural and rational inattention perspectives. For instance, the firm might find costly querying the productivity of every worker, or keep book records of all its salary offers. In this case, we model LIPC as an instance of the "label efficient" problem [Cesa-Bianchi and Lugosi, 2006].

Consider a binary variable $\Theta \sim \mathcal{B}ern(p_\theta)$, which is not controlled by the adversary. Assume that, at the end of every period, the learner observes $\Theta_i = \{0, 1\}$. Shall $\Theta_i = 1$, $\psi_i = \psi_i$, otherwise $\psi_i = \{x_i\}$. This means that the policymaker recovers the

asymmetric feedback with probability p_θ and recovers her action but no information on U, V with probability $1 - p_\theta$. Implicitly, the higher p_θ, the lower the limitation in processing capacity. Our modelling assumption differs from standard label efficient procedures in that information querying is not a strategic decision, but the result of an exogenous process. There is, however, small loss in this assumption, as optimal strategic interactions in adversarial specifications rely on similar randomisation devices [Cesa-Bianchi and Lugosi, 2006]. Unlike the MW case, Algorithm 1 does not remain optimal for the LIPC case. In fact, $\hat{U}_i$, $\hat{G}_i$ are not even defined whenever $\Theta_i = 0$. We start this section by showing that a small variation of Algorithm 1 is in fact near optimal with matching bounds of $\mathcal{O}(K^{\frac{2}{3}}/p_\theta^2)$.

Proposition 10. *Optimality of Algorithm 2.* Algorithm 2 is optimal up to logarithmic factors in the adversarial and stochastic environments of the label efficient monopolist wage setting problem (LE-GMP) with regret of $\mathcal{O}(K^{\frac{2}{3}}/p_\theta^2)$. In addition, the cumulative average regret of Algorithm 2 is sub-linear whenever $p_\theta \geq \mathcal{O}\left(\left(\frac{\log(K)^{\frac{1}{3}}}{K^{\frac{1}{3}}}\right)^{1/2}\right)$.

Proof: To verify the second statement in Proposition 10 simply note that,

$$c_5 \cdot \log(K)^{\frac{1}{3}} \cdot K^{\frac{2}{3}} \cdot \frac{1}{p_\theta^2} \leq c_6 \cdot K \implies p_\theta \geq \left(\frac{c_5}{c_6}\frac{\log(K)^{\frac{1}{3}}}{K^{\frac{1}{3}}}\right)^{1/2} \tag{21}$$

Hence, the amount of information needed to achieve sub-linear regret decreases with K. Empirically, we show that very low values of p_theta still guarantee decreasing average regret. It now follows that if the regret of Algorithm 2 is $\leq \mathcal{O}(K^{\frac{2}{3}}/p_\theta^2)$, the second statement holds. To prove the first statement in Proposition 10, we start by presenting Algorithm 2. Then we suitably modify the arguments in Appendix 9.1 and Appendix 9.2 to incorporate p_θ within the analysis and show matching upper and lower bounds.

Intuitively, Algorithm 2 does not update any arm whenever $\Theta_i = 0$, while it creates unbiased estimates of $\tilde{\mathbb{G}}_{ib}$, $\tilde{\mathbb{U}}_{ib}$ by factoring in the probability of recovering feedback. Unbiasedness can be shown by

Algorithm 2 Tempered Exp3 for the LE-GMP

Input B, λ, η, γ
Set $x_b = (b-1)/B$ for $b \in \{1, 2, \ldots, B+1\}$, $\widehat{\mathbb{G}}_{1b} = 0$, $\widehat{\mathbb{U}}_{1b} = 0$ for all b
for $i = 1, 2, \ldots, K$
  for $b = 1, \ldots, B+1$
    Set $\widehat{\mathbb{S}}_{ib} = \widehat{\mathbb{U}}_{ib} - x_b \widehat{\mathbb{G}}_{ib} + \frac{\lambda}{B} \sum_{b'<b} \widehat{\mathbb{G}}_{ib'}$, $p_{ib} = (1-\gamma) \frac{\exp(\eta\, \widehat{\mathbb{S}}_{ib})}{\sum_{b'} \exp(\eta\, \widehat{\mathbb{S}}_{ib'})} + \frac{\gamma}{B+1}$
  end for
  Sample $b_i \sim p_{ib}$
  Observe $\Theta_i = \theta$. **If** $\Theta_i = 1$
    Observe $\mathbb{1}(x_{b_i} \geq v_i)$, **If** $\mathbb{1}(x_{b_i} \geq v_i) = 1$ **Observe** u_i
    for $b = 1, \ldots, B+1$
      Update $\widehat{\mathbb{G}}_{i+1,b} = \widehat{\mathbb{G}}_{ib} + \mathbb{1}(x_{b_i} \geq v_i) \frac{\mathbb{1}(b_i = b)}{p_{ib} \cdot p_\theta}$, $\widehat{\mathbb{U}}_{i+1,b} = \widehat{\mathbb{U}}_{ib} + u_i \cdot \mathbb{1}(x_{b_i} \geq v_i) \frac{\mathbb{1}(b_i = b)}{p_{ib} \cdot p_\theta}$
    end for
  else $\widehat{\mathbb{G}}_{i+1,b} = \widehat{\mathbb{G}}_{ib}$, $\widehat{\mathbb{U}}_{i+1,b} = \widehat{\mathbb{U}}_{ib}$
end for

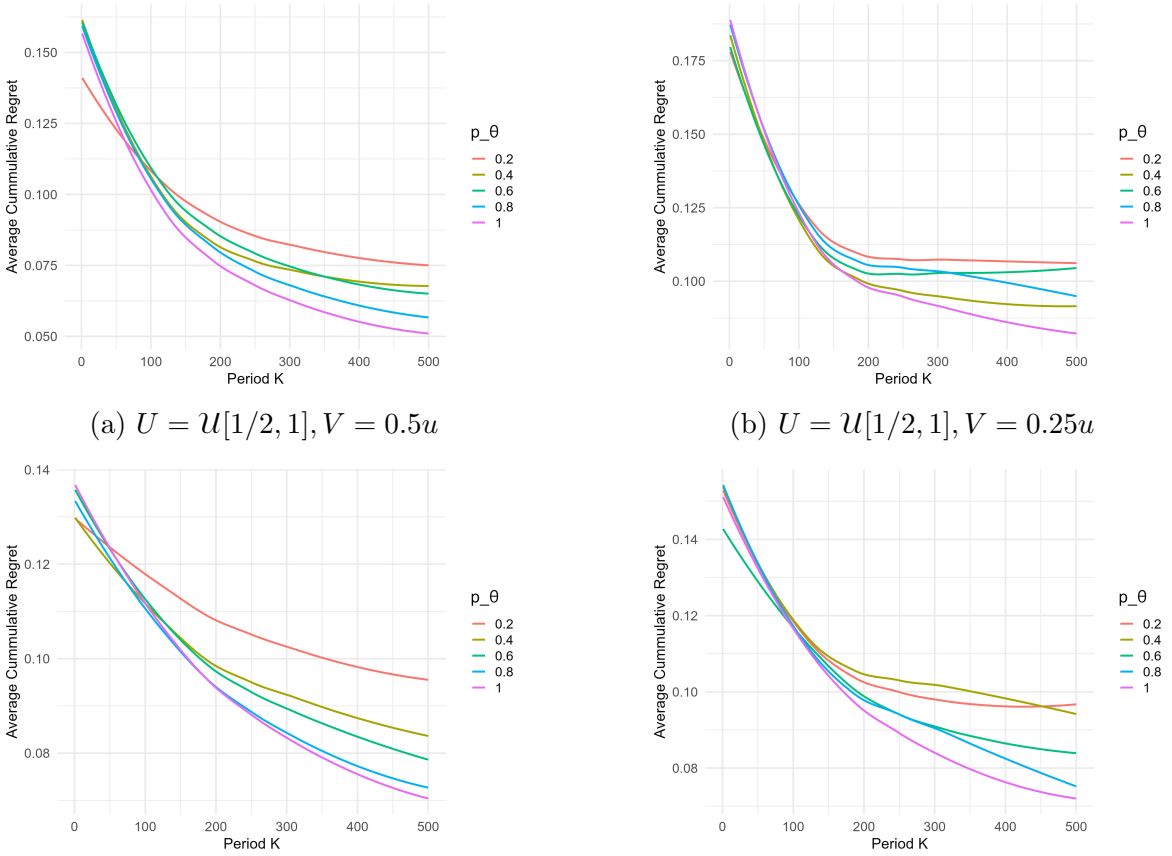
$$\mathbb{E}[\hat{G}_{ib}] = \sum_{\Theta \in \{0,1\}} \sum_b p_\Theta \cdot p_{ib} \hat{G}_{ib} = \sum_b p_\theta \cdot p_{ib} \frac{\mathbb{1}(x_{b_i} \geq v_i)\mathbb{1}(b_i = b)}{p_{ib} \cdot p_\theta} = \sum_b p_\theta \cdot p_{ib} \frac{\mathbb{1}(x_{b_i} \geq \tilde{v}_i)\mathbb{1}(b_i = b)}{p_{ib} \cdot p_\theta} = \mathbb{1}(x_{b_i} \geq \tilde{v}_i) \tag{22}$$

It now follows that $\mathbb{E}[\widehat{\mathbb{G}}_{ib}] = \mathbb{E}[\sum_{j \leq i} \hat{G}_{jb}] = \tilde{\mathbb{G}}_{ib}$. Similarly, $\mathbb{E}[\widehat{\mathbb{U}}_{ib}] = \tilde{\mathbb{U}}_{ib}$. Finally, $\mathbb{E}[\widehat{\mathbb{S}}_{ib}] = \tilde{\mathbb{S}}_{ib}$. We delay the rest of the proof for the upper bound of Algorithm 2 to Appendix 9.3. Lower bounds on the LE-GMP are established in a similar way than benchmark results in Appendix 9.2. In a nutshell, the proof relied on building policy relations between $\mathbb{E}_\epsilon[S_i(x)]$ for a very particular DGP, $F^\epsilon_{U,V}$. This DGP remains unchanged in the LE-GMP, so updates concentrate in the application of the Embedding Lemma, the adaptation of Claim A.2 in [Cesa-Bianchi et al., 2022] (Lemma 11 in Appendix 9.2), and the inequalities which follow thereafter. The full-proof is deferred to Appendix 9.3. □

Simulation Analysis and Welfare Implications. Figure 7 shows the evolution of average cumulative regret of Algorithm 2 for different DGP and different values of p_θ. According to the theory, we expect the asymptotic convergence rate of average cumulative regret to decrease quadratically on the inverse of p_θ. In particular, we expect welfare losses to be unbounded as $p_\theta \to 0$. These prescriptions are matched

empirically with an asymptotic regret ordering based on p_θ. Surprisingly, we observer decreasing average regret with as little feedback as $p_\theta = 0.2$. In addition, we do not see any significant difference in terms of regret associated to the interaction between the DGP and the LIPC.

Figure 7: Algorithm 2 under LIPC



(a) $U = \mathcal{U}[1/2, 1], V = 0.5u$

(b) $U = \mathcal{U}[1/2, 1], V = 0.25u$

(c) $U = \mathcal{U}[1/4, 3/4], V = 0.5u$

(d) $U = \mathcal{U}[1/4, 3/4], V = 0.25u$

Figure 8: Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $B = 10$, $\eta = 0.132$ $\gamma = 0.029$

Overall, this section elucidates the welfare implications of limited information processing capacity. These results are particularly relevant in an era where information is generally available, but costly to incorporate and to adapt business strategies accordingly [Maćkowiak et al., 2023]. In addition, LIPC can also be interpreted as a shock. For instance, [Coen, 2021] studies the information loss associated to firms' exit or firms' bankruptcy. A general macroeconomic shock is likely to generate exogenous firms' ex-

its, and consequently a reduction in the information processing capacity of the affected industries. Welfare reductions associated to information loss may call for public bailout of failing companies with high information positive externalities.

8 Conclusion

This paper has discussed the equilibrium convergence of monopsony strategies to optimal wage, under reasonable, economically relevant feedback structures. We have shown that such convergence is possible using a simple strategy without any prior on the learner and for any arbitrary distribution of outcomes. In fact, we have demonstrated that our strategy, guided by Algorithm 1, is near-optimal under stochastic and adversarial specifications. In addition, we have presented complementary side topics which expand our understanding of adaptive policy design like the notion of asymmetric feedback and the welfare implications of sub-optimal parameter selection.

Despite these noteworthy technical contributions, this paper highlights the economic modelling potential of bandit learning. In particular, this paper explored the role of adverse selection and imperfect competition within the learning process. In addition, the new tools developed in this paper allow for counterfactual policy analysis like Minimum Wage policies and LIPC. In fact, an extended version of this paper, also includes some analyses on productivity shocks and firms' recovery time. The prescriptions of our extremely stylised model are compatible with sophisticated theory, and it does so under no information requirements at all. Interestingly, our model also unveils new theory like information gains associated to Minimum Wage and the (limited) welfare loss associated to reductions in the information processing capacity.

We identify three future lines of research following this manuscript. (i) Increase in the sophistication of the problem of the firm. For instance, recent work by [Gonzalez, 2023] discusses a similar problem under Concave Bandits/Convex Knapsacks specifications which allow for budget constraints in the GMP. (ii) Consider alternative workers-oriented definitions of welfare ($\lambda > 1$). In this case, we had to restrict $\lambda < 1$, for encoding meaningful wage setting dynamics (otherwise $x^* = \infty$). However, under budget constraints, workers-oriented objective functions can be analysed. Finally, (iii) our tools can be used to study alternative economically relevant problems or alternative counterfactual public policies.

References

[Acharya and Wee, 2020] Acharya, S. and Wee, S. L. (2020). Rational inattention in hiring decisions. *American Economic Journal: Macroeconomics*, 12(1):1–40.

[Akerlof, 1978] Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

[Björkegren et al., 2020] Björkegren, D., Blumenstock, J. E., and Knight, S. (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

[Blackwell, 1956] Blackwell, D. (1956). An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–18.

[Bubeck et al., 2012] Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

[Card et al., 2012] Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review*, 102(6):2981–3003.

[Cengiz et al., 2022] Cengiz, D., Dube, A., Lindner, A., and Zentler-Munro, D. (2022). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. *Journal of Labor Economics*, 40(S1):S203–S247.

[Cengiz et al., 2019] Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

[Cesa-Bianchi et al., 2021] Cesa-Bianchi, N., Cesari, T. R., Colomboni, R., Fusco, F., and Leonardi, S. (2021). A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 289–309.

[Cesa-Bianchi et al., 2022] Cesa-Bianchi, N., Colomboni, R., and Kasy, M. (2022). Adaptive maximization of social welfare.

[Cesa-Bianchi and Lugosi, 2006] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[Coen, 2021] Coen, P. (2021). Information loss over the business cycle.

[Doraszelski and Pakes, 2007] Doraszelski, U. and Pakes, A. (2007). A framework for applied dynamic analysis in io. *Handbook of industrial organization*, 3:1887–1966.

[Draca et al., 2011] Draca, M., Machin, S., and Van Reenen, J. (2011). Minimum wages and firm profitability. *American economic journal: applied economics*, 3(1):129–151.

[Dube, 2019a] Dube, A. (2019a). Impacts of minimum wages: review of the international evidence. *Independent Report. UK Government Publication*, pages 268–304.

[Dube, 2019b] Dube, A. (2019b). Making a case for a higher minimum wage. *Essay in The Milken Review, 2019*.

[Dube et al., 2016] Dube, A., Lester, T. W., and Reich, M. (2016). Minimum wage shocks, employment flows, and labor market frictions. *Journal of Labor Economics*, 34(3):663–704.

[Ericson and Pakes, 1995] Ericson, R. and Pakes, A. (1995). Markov-perfect industry dynamics: A framework for empirical work. *The Review of economic studies*, 62(1):53–82.

[Esponda, 2008] Esponda, I. (2008). Behavioral equilibrium in economies with adverse selection. *American Economic Review*, 98(4):1269–1291.

[Furman and Orszag, 2018] Furman, J. and Orszag, P. (2018). 1. a firm-level perspective on the role of rents in the rise in inequality. In *Toward a Just Society*, pages 19–47. Columbia University Press.

[Gonzalez, 2023] Gonzalez, C. (2023). Hiring decisions with knapsack bandits.

[Hart and Mas-Colell, 2000] Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150.

[Hart and Mas-Colell, 2001a] Hart, S. and Mas-Colell, A. (2001a). A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54.

[Hart and Mas-Colell, 2001b] Hart, S. and Mas-Colell, A. (2001b). *A reinforcement procedure leading to correlated equilibrium*. Springer.

[Hart and Mas-Colell, 2013] Hart, S. and Mas-Colell, A. (2013). *Simple adaptive strategies: from regret-matching to uncoupled dynamics*, volume 4. World Scientific.

[Hazan et al., 2016] Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

[Kleinberg and Leighton, 2003] Kleinberg, R. and Leighton, T. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE.

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[Maćkowiak et al., 2023] Maćkowiak, B., Matějka, F., and Wiederholt, M. (2023). Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273.

[Manning, 2003] Manning, A. (2003). The real thin theory: monopsony in modern labour markets. *Labour economics*, 10(2):105–131.

[Manning, 2021] Manning, A. (2021). Monopsony in labor markets: A review. *ILR Review*, 74(1):3–26.

[Mas-Colell et al., 1995] Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory*, volume 1. Oxford university press New York.

[Sims, 2003] Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.

[Van Nieuwerburgh and Veldkamp, 2010] Van Nieuwerburgh, S. and Veldkamp, L. (2010). Information acquisition and under-diversification. *The Review of Economic Studies*, 77(2):779–805.

[Wolfson and Belman, 2019] Wolfson, P. and Belman, D. (2019). 15 years of research on us employment and the minimum wage. *Labour*, 33(4):488–506.

9 Appendix

9.1 Proof of an Upper Bound on Regret for Algorithm 1

The following proof proceeds in a sequential manner. Setting approximation errors aside, we first bound welfare S_i in terms of welfare estimates $\hat{S}_i$ under the premise that the maximum is smaller than the sum. We make a second order approximation to welfare estimates and bound independently their first and second order moments. Then we put all inequalities together, and tune parameters accordingly to achieve the desired result.

Approximation error. Define $\tilde{v}_i = \min_b\{x_b : x_b \geq v_i\}$ and consider $\tilde{S}_i(x) = \mathbb{1}(x \geq v_i)(u_i - x) + \lambda \max(x - \tilde{v}_i, 0)$. Similarly, define $\tilde{\mathbb{S}}_{ib} = \tilde{\mathbb{S}}_i(x_b) = \sum_{j \leq i} \tilde{S}_i(x_b)$. Clearly, $\tilde{S}_i(x) \leq S_i(x)$ for all x and all i. Moreover,

$$\sup_x \mid S_i(x) - \tilde{S}_i(x) \mid \leq \frac{\lambda}{B} \tag{23}$$

given our initialisation of the grid. We see that $\mid \sup_x \tilde{S}_i(x) - \max_b \tilde{S}_i(x_b) \mid \leq 1/B$. This step is central in the MW analysis of Section 7.1. It now follows

$$\max_b \tilde{\mathbb{S}}_{ib} \geq \sup_x \tilde{\mathbb{S}}_i(x) - \frac{i}{B} \geq \sup_x \mathbb{S}_i(x) - \frac{i(1+\lambda)}{B} \tag{24}$$

Unbiasedness. We now show that $\mathbb{E}[\widehat{\mathbb{S}}_{ib}] = \tilde{\mathbb{S}}_{ib}$, where $\widehat{\mathbb{S}}_{ib} = \sum_{j<i} \hat{S}_{ib}$. Define its one period counterpart as $\hat{S}_{ib} = \hat{U}_{ib} - x_b \cdot \hat{G}_{ib} + \frac{\lambda}{B} \sum_{b'<b} \hat{G}_{ib'}$, where $\hat{U}_{ib} = u_i \cdot \mathbb{1}(x_b \geq v_i) \frac{\mathbb{1}(b_i = b)}{p_{ib}}$ and $\hat{G}_{ib} = \mathbb{1}(x_b \geq v_i) \frac{\mathbb{1}(b_i = b)}{p_{ib}}$. It then suffices to show that at the **end** of every period i, $\mathbb{E}[\hat{G}_{ib}] = \tilde{G}_{ib}$ for all b.

$$\mathbb{E}[\hat{G}_{ib}] = \sum_b p_{ib} \hat{G}_{ib} = \sum_b p_{ib} \frac{\mathbb{1}(x_{b_i} \geq v_i)\mathbb{1}(b_i = b)}{p_{ib}} = \sum_b p_{ib} \frac{\mathbb{1}(x_{b_i} \geq \tilde{v}_i)\mathbb{1}(b_i = b)}{p_{ib}} = \mathbb{1}(x_{b_i} \geq \tilde{v}_i) \tag{25}$$

But then $\mathbb{E}[\mathbb{G}_{ib}] = \mathbb{E}[\sum_{j \leq i} \hat{G}_{jb}] = \tilde{\mathbb{G}}_{ib}$. Similarly, $\mathbb{E}[\mathbb{U}_{ib}] = \tilde{\mathbb{U}}_{ib}$. Finally, $\mathbb{E}[\widehat{\mathbb{S}}_{ib}] = \tilde{\mathbb{S}}_{ib}$. We have just shown that our welfare estimate is unbiased for $\tilde{\mathbb{S}}_{ib}$ after the update

in period i.

Upper bound on maximum welfare. Now we are in position to set bounds for the welfare estimates. This derivation is quite standard in Exp-3-like upper bounds derivation, and it is based on the principle that the sum is larger than the maximum. Define $W_i = \sum_b \exp(\eta \cdot \widehat{\mathbb{S}}_{ib})$, it follows that

$$\begin{aligned}
&\mathbb{E}[\log W_K] = \mathbb{E}[\log \sum_b \exp(\eta \cdot \widehat{\mathbb{S}}_{Kb})] = \eta \cdot \mathbb{E}[\log \sum_b \exp(\widehat{\mathbb{S}}_{Kb})] \geq \\
&\eta \cdot \mathbb{E}[\max_b \log \exp(\widehat{\mathbb{S}}_{Kb})] = \eta \cdot \mathbb{E}[\max_b \widehat{\mathbb{S}}_{Kb}] \geq \eta \cdot \max_b \mathbb{E}[\widehat{\mathbb{S}}_{Kb}] = \eta \cdot \max_b \tilde{\mathbb{S}}_{Kb} \\
&\mathbb{E}[\log W_K] = \sum_i \mathbb{E}\left[\log\left(\frac{W_{i+1}}{W_i}\right)\right] + \log(W_0) = \sum_i \mathbb{E}\left[\log\left(\frac{W_{i+1}}{W_i}\right)\right] + \log(B+1)
\end{aligned} \tag{26}$$

Lower bound on estimated analogues. We now rewrite the expression above in terms of the estimated analogues, and make a second order approximation. To do so, we define $q_{ib} = \exp(\eta \cdot \widehat{\mathbb{S}}_{ib})/W_i$ as opposed to p_{ib}. In particular note that $p_{ib} = (1-\gamma)q_{ib} + \gamma/(B+1)$ or equivalently $q_{ib} = \left(p_{ib} - \frac{\gamma}{B+1}\right)\frac{1}{1-\gamma}$. We may now rewrite

$$\log\left(\frac{W_{i+1}}{W_i}\right) = \log\left(\frac{\sum_b \exp(\eta \cdot \sum_{j \leq i} \hat{S}_{jb})}{\sum_b \exp(\eta \cdot \sum_{j<i} \hat{S}_{jb})}\right) = \log\Big(\sum_b q_{ib} \cdot \exp(\eta \cdot \hat{S}_{ib})\Big) \tag{27}$$

From Algorithm 1, $\gamma/(B+1) \leq p_{ib} \leq 1$ for all b and all $i \implies \hat{S}_{ib} \in [-(B+1)/\gamma, (B+1)/\gamma] \implies \eta \cdot \hat{S}_{ib} \leq \eta \cdot (B+1)/\gamma \leq 1$ by construction of η (see the *Verifying Parameters* step for details). We can now use the following inequalities $\exp(a) \leq 1 + a + (e-2) \cdot a^2$ for $a \leq 1$ and $\log(1+a) \leq a$ to establish that

$$\begin{aligned}
\exp(\eta \cdot \hat{S}_{ib}) &\leq 1 + \eta \cdot \hat{S}_{ib} + (e-2) \cdot \eta^2 \cdot \hat{S}_{ib}^2 \log\left(\frac{W_{i+1}}{W_i}\right) \\
&\leq \log\Big(\sum_b q_{ib}\big(1 + \eta \cdot \hat{S}_{ib} + (e-2) \cdot \eta^2 \cdot \hat{S}_{ib}^2\big)\Big) \\
&\leq \eta \sum_b q_{ib} \hat{S}_{ib} + (e-2) \cdot \eta^2 \cdot \sum_b q_{ib} \hat{S}_{ib}^2
\end{aligned} \tag{28}$$

Bounding first and second order terms. Once that we have established an upper bound on welfare in terms of first and second order moments of the welfare estimates, we proceed to bound those terms. Using our definition of q_{ib} in terms of p_{ib}, it follows that

$$\mathbb{E}[\sum_b q_{ib} \cdot \hat{S}_{ib}] = \frac{1}{1-\gamma}\mathbb{E}[\sum_b p_{ib} \cdot \hat{S}_{ib}] - \frac{\gamma}{(1-\gamma)(B+1)}\mathbb{E}[\sum_b \hat{S}_{ib}] \leq \frac{1}{1-\gamma}(\mathbb{E}[\tilde{S}_i(x_i)] + 1) \tag{29}$$

Where we used that $\hat{S}_{ib} \in [-(B+1)/\gamma, (B+1)/\gamma]$ for every i and b, and so does $\sum_b \hat{S}_{ib}$. Moreover, we used equation (25) for setting $\mathbb{E}[\hat{S}_{ib}] = \sum_b p_{ib} \cdot \hat{S}_{ib} = \tilde{S}_{ib}$. To bound the second order term $\mathbb{E}[\sum_b q_{ib} \cdot \hat{S}_{ib}^2]$, we make the following observations

$$\sum_b q_{ib} \cdot \hat{S}_{ib}^2 \leq \frac{1}{1-\gamma} \sum_b p_{ib} \cdot \hat{S}_{ib}^2 \tag{30}$$

and $\hat{S}_{ib} = \frac{\mathbb{1}(x_b \geq v_i)}{p_{ib_i}}(\mathbb{1}(b_i = b)(u_i - x_b) + \frac{\lambda}{B}\sum_{b'<b}\mathbb{1}(b_i < b'))$. At most one of the two elements $\mathbb{1}(b_i = b)(u_i - x_i)$ OR $\sum_{b'<b}\mathbb{1}(b_i < b)$ is different from zero, so $\mathbb{E}[\hat{S}_{ib}^2] \leq u_i^2/p_{ib} + x_b^2/p_{ib} + (\frac{\lambda}{B})^2 \sum_{b'<b}\frac{1}{p_{ib'}}$. Using $u_i^2 \leq 1$,

$$\begin{aligned}
\mathbb{E}[\sum_b p_{ib} \cdot \hat{S}_{ib}^2] &\leq \sum_b u_i^2 + \sum_b x_b^2 + \Big(\frac{\lambda}{B}\Big)^2 \sum_b \sum_{b'<b} \frac{p_{ib}}{p_{ib'}} \\
&\leq B u_i^2 + \sum_b \Big(\frac{b}{B}\Big)^2 + \Big(\frac{\lambda}{B}\Big)^2 \sum_b p_{ib} \sum_{b' \neq b} \frac{B+1}{\gamma} \\
&\leq B + \frac{B(B+1)(2B+1)}{6B^2} + \frac{\lambda^2}{\gamma}\frac{B+1}{B} \\
&\leq B + \frac{B}{B+1}\Big(\frac{2B+1}{6} + \frac{\lambda^2}{\gamma}\Big) = \frac{B}{B+1}\Big(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma}\Big)
\end{aligned} \tag{31}$$

Wrapping-up. We can simply put all inequalities together such that

$$
\begin{aligned}
&\eta \cdot \Big(\sup_x \mathbb{S}(x) - \frac{K(1+\lambda)}{B} \Big) \leq \eta \cdot \big(\max_b \tilde{\mathbb{S}}(x_b) \big) = \max_b \eta \cdot \big(\mathbb{E}[\widehat{\mathbb{S}}(x_b)] \big) \leq \\
&\mathbb{E}[\log W_K] = \log(B+1) + \sum_i^K \mathbb{E}\left[\log\Big(\frac{W_{i+1}}{W_i} \Big) \right] \leq \\
&\log(B+1) + \sum_i^K \mathbb{E}\left[\eta \cdot \sum_b q_{ib} \cdot \hat{S}_{ib} + (e-2) \cdot \eta^2 \cdot \sum_b q_{ib} \hat{S}_{ib}^2 \right] \leq \\
&\log(B+1) + \frac{\eta}{1-\gamma} (\mathbb{E}[\tilde{\mathbb{S}}(x_i)] + 1) + (e-2) \cdot \frac{\eta^2}{1-\gamma} \cdot K \cdot \frac{B}{B+1} \left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \right)
\end{aligned}
\tag{32}
$$

Finally, we solve for $\sup_x \mathbb{S}(x) - \mathbb{E}[\tilde{\mathbb{S}}(x_i)]$. To do so, we first solve for $\mathbb{E}[\tilde{\mathbb{S}}_i]$ and then we add $\gamma \cdot \sup_x \mathbb{S}(x)$ at both sides. Here, we do not need to worry about the negative sign because $S(0) = 0$ for any arbitrary sequence of (u_i, v_i), so $\gamma \cdot \sup_x \mathbb{S}(x) \geq 0$.

$$
\begin{aligned}
&\sup_x \mathbb{S}(x) - \mathbb{E}[\mathbb{S}(x_i)] \leq \sup_x \mathbb{S}(x) - \mathbb{E}[\tilde{\mathbb{S}}(x_i)] \\
&\leq 1 + (1-\gamma) \frac{K(1+\lambda)}{B} + \gamma \cdot \sup_x \mathbb{S}(x) + \frac{1-\gamma}{\eta} \log(B+1) + (e-2) \cdot \eta \cdot K \cdot \frac{B}{B+1} \left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \right) \\
&\leq 1 + \frac{K(1+\lambda)}{B} + \gamma \cdot K + \frac{\log(B+1)}{\eta} + (e-2) \cdot \eta \cdot K \frac{B}{B+1} \left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \right) \\
&= K \left(\gamma + \frac{(1+\lambda)}{B} + \eta(e-2) \frac{B}{B+1} \left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \right) \right) + \frac{\log(B+1)}{\eta} + 1
\end{aligned}
\tag{33}
$$

Tuning parameters. We can now infer optimal parameter values in terms of K by solving the FOCs in the expression above. Consider setting $\gamma = c_1 \Big(\frac{log(K)}{K} \Big)^{\frac{1}{3}}$, $\eta = c_2 \cdot \gamma^2$ and $B = \frac{c_3}{\gamma}$. It follows that for some $c_4 < \infty$

$$
\begin{aligned}
&\sup_x \mathbb{S}(x) - \mathbb{E}[\mathbb{S}(x_i)] \leq \\
&K \left(\gamma + \frac{(1+\lambda) \cdot \gamma}{c_3} + c_2 \cdot \gamma^2 \cdot (e-2) \cdot \frac{c_3}{c_3 + \gamma} \left(\frac{8 \cdot \frac{c_3}{\gamma} + 7}{6} + \frac{\lambda^2}{\gamma} \right) \right) + \frac{\log(\frac{c_3}{\gamma} + 1)}{c_2 \cdot \gamma^2} + 1 =
\end{aligned}
$$

$$\log(K)^{\frac{1}{3}}K^{\frac{2}{3}}\left(c_1+(1+\lambda)\frac{c_1}{c_3}+c_2c_1^2(e-2)\frac{c_3}{c_3+o(1)}\Big(\frac{4}{3}\frac{c_3}{c_1}+o(1)+\frac{\lambda^2}{c_1}\Big)+\frac{\log\left(\frac{c_3}{c_1}\big(\frac{K}{\log(K)}\big)^{\frac{1}{3}}+1\right)}{c_2\log(K)}+o(1)\right)=$$

$$\log(K)^{\frac{1}{3}}\cdot K^{\frac{2}{3}}\left(c_1+(1+\lambda)\frac{c_1}{c_3}+c_2\cdot c_1^2\cdot(e-2)\cdot\frac{c_3}{c_3+o(1)}\left(\frac{4}{3}\frac{c_3}{c_1}+o(1)+\frac{\lambda^2}{c_1}\right)+o(1)\right)\leq c_4\cdot\log(K)^{\frac{1}{3}}\cdot K^{\frac{2}{3}}$$

Verify parameters. It remains to prove that tuned parameter η is compatible with our claim in equation (28). This claim requires $\eta=\{\eta:\eta\cdot(B+1)/\gamma\leq 1\}$. Considering the parameters above, we can verify that

$$\eta\cdot(B+1)/\gamma=c_2\cdot\gamma\cdot\Big(\frac{c_3}{\gamma}+1\Big)=c_2\cdot c_3+c_2/\gamma=c_2\cdot c_3+\frac{c_2}{c_1}\Big(\frac{K}{\log(K)}\Big)^{\frac{1}{3}}\tag{34}$$

Consider setting $c_2\leq\frac{c_1}{c_1\cdot c_3+(K/\log(K))^{1/3}}$ such that the inequality above holds. In addition, c_1 and c_3 must be chosen in such a way that parameters γ and B remain plausible. In this context $\gamma\in(0,1)$ and $B>1$. The following conditions ensure such plausibility,

$$0<c_1\leq\left(\frac{K}{\log(K)}\right)^{\frac{1}{3}}\qquad\qquad c_3\geq\left(\frac{K}{\log(K)}\right)^{\frac{1}{3}}>0\tag{35}$$

This completes the proof of Theorem 3. □

9.2 Proof of a Lower Bound for the GMP

A two step procedure. As introduced in Section 5, this proof relies on a two-step procedure. First, we use the Embedding Lemma introduced in [Cesa-Bianchi et al., 2021] to establish a difficulty relation between the GMP and an easier problem. We then lower bound this "easy" problem following closely the intuitions in [Cesa-Bianchi et al., 2022] for the adaptive welfare maximisation problem.

The aim of the Embedding Lemma is to establish difficulty relations across problems. Provided a game is "easier" it will have uniformly no higher regret (almost surely) than the initial game. Consequently, establishing a lower bound on the easy problem

immediately guarantees a lower bound on the original problem. In particular, the Embedding Lemma states that a game $\mathcal{G}_1$ is easier than a game $\mathcal{G}_2$ if (i) the optimal strategy in $\mathcal{G}_2$ obtains weakly higher returns (lower regret) in $\mathcal{G}_1$ than the optimal policy in $\mathcal{G}_1$, (ii) sub-optimal policies in $\mathcal{G}_1$ obtain a reward at least as high as sub-optimal policies in $\mathcal{G}_2$, and (iii) feedback is not smaller in the the easy game $\mathcal{G}_1$ compared to the difficult game $\mathcal{G}_2$. I refer the reader to Appendix B in [Cesa-Bianchi et al., 2021] for a formal characterisation (and proof) of the Embedding Lemma.

An application of the Embedding Lemma. I introduce now some useful notation. Define a game $\mathcal{G}$ as a tuple $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \psi, \mathcal{P}\}$ where $\mathcal{X}$ is the action space, $\mathcal{Y}$ is the adversary space, $\mathcal{Z}$ is the feedback space, $\rho : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the reward function, $\psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is the feedback function and $\mathcal{P}$ is the probability space of the adversary's behaviour ($\mathcal{Y}^K$). Consider the GMP $\mathcal{G} = \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \psi, \mathcal{P}\}$ where $\mathcal{X} = [0,1]$, $\mathcal{Y} = [0,1]^2$, $\mathcal{Z} = [0,1] \times \{0,1\} \times [0,1] \cup \emptyset$, $\rho : (x,u,v) \mapsto S(x,u,v)$, $\psi : (x,u,v) \mapsto (x, \mathbb{1}(x \geq v), \psi^{\emptyset}((x \geq v), u))$ and $\mathcal{P} = F_{U,V}$.

Consider now some $\epsilon \in [-1,1]$ and define $\mathcal{P}_\epsilon^1 = F_{U,V}^\epsilon = F_V^\epsilon \otimes F_U^1$ where F_V^ϵ is characterised by some $f_v^\epsilon \coloneqq (a, b\cdot(1+\epsilon), b\cdot(1-\epsilon), 1-a-2b)$ with support $\text{Supp}_1(V) = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$. Observe $\text{Supp}_1(V) \subset \Omega_V$. Moreover, consider $f_U^1 : \mathbb{1}(u = 1)$. Based on this characterisation of $\mathcal{P}_\epsilon^1$, define $\mathcal{G}_1 = \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \psi, \mathcal{P}_\epsilon^1\}$. By the Embedding Lemma with f and g as identities and $\hbar$ as the inclusion, $\mathcal{R}(\mathcal{G}_1) \leq \mathcal{R}(\mathcal{G})$.

An ϵ dependent distribution. Notice that for $\mathcal{G}_1$, $\mathbb{E}_\epsilon[S_i(X)] = \sum_{v \in \text{Supp}_1(V)} \mathbb{P}(V = v)\mathbb{1}(x \geq v)[1 - x + \lambda(x - v)]$. Consider now the following parameters and definitions

$$a \coloneqq \frac{3 \cdot b \cdot \lambda + 1}{4 - 3 \cdot \lambda}, \quad b \coloneqq \frac{1 - \lambda}{2 \cdot (24 - 17\lambda)} \tag{36}$$

$$\begin{gathered} c_1 \coloneqq -\frac{\lambda \cdot b}{4}; \quad c_2 \coloneqq \frac{1-\lambda}{32 - 24\lambda} - \frac{3 - \lambda \cdot b}{2}; \quad c_3 \coloneqq b \cdot \sqrt{\frac{2}{a \cdot (1 - a - 2 \cdot b)}} \\ C \coloneqq \min\left(\frac{c_1^2 \cdot c_3^2}{c_2}, \frac{c_2}{2}, \frac{1}{16}\sqrt{\frac{c_1^2 \cdot c_2}{c_3^2}}\right) \end{gathered} \tag{37}$$

$$n_1 = \sum_{i \in K} \mathbb{1}(x_i \in [1/4, 1/2)); \quad n_2 = \sum_{i \in K} \mathbb{1}(x_i \in [1/2, 1]); \quad n_3 = \sum_{i \in K} \mathbb{1}(x_i \in [0, 1/4)) \tag{38}$$

Under $\mathscr{P}_\epsilon^1$,

$$\begin{aligned} \mathbb{E}_\epsilon[S_i(0)] &= a & \mathbb{E}_\epsilon[S_i(1/4)] &= a \cdot \frac{3+\lambda}{4} + 3 \cdot b \cdot \frac{(1+\epsilon)}{4} \\ \mathbb{E}_\epsilon[S_i(1/2)] &= a \cdot \frac{1+\lambda}{2} + b \cdot \frac{4+\lambda+\lambda\cdot\epsilon}{2} & \mathbb{E}_\epsilon[S_i(3/4)] &= a \cdot \frac{3\cdot\lambda}{4} + b \cdot \frac{\lambda\cdot(3+\epsilon)}{4} + \frac{1}{4} \end{aligned} \tag{39}$$

Similarly,

$$\begin{aligned} &\max_{x\in[1/2,1]} \mathbb{E}_\epsilon[S_i(x)] = \mathbb{E}_\epsilon[S_i(3/4)] && \max_{x\in[1/4,1/2)} \mathbb{E}_\epsilon[S_i(x)] = \mathbb{E}_\epsilon[S_i(1/4)] \\ &\max_{x\in[0,1/4)} \mathbb{E}_\epsilon[S_i(x)] = \mathbb{E}_\epsilon[S_i(0)] \end{aligned} \tag{40}$$

Note that the results above hold uniformly for all $\epsilon \in [-1, 1]$. Second and third expressions in expression (40) are trivial. Intuitively, conditional on being on the same distribution support bracket, $S_i(x)$ is decreasing in x uniformly for all ϵ. Global maxima are then attained at $x = \min\{x : x \in \text{Bracket}_V\}$ for all $a, b : \{a, b > 0 : a+b < 1\}$. First result however, imposes some pressure on the selected parameters a, b. It is immediately true that $\max_{x\in[3/4,1]} \mathbb{E}_\epsilon[S_i(x)] = \mathbb{E}_\epsilon[S_i(3/4)]$. Similarly $\max_{x\in[1/2,3/4)} \mathbb{E}_\epsilon[S_i(x)] = \mathbb{E}_\epsilon[S_i(1/2)]$. Remains to prove that $\mathbb{E}_\epsilon[S_i(3/4)] \geq \mathbb{E}_\epsilon[S_i(1/2)]$ for all ϵ.

Proof of Equation (40.1). The proposition above holds iff

$$a \cdot \frac{1+\lambda}{2} + b \cdot \frac{4+\lambda+\lambda\cdot\epsilon}{2} \leq a \cdot \frac{3\lambda}{4} + b \cdot \frac{\lambda\cdot(3+\epsilon)}{4} + \frac{1}{4} \ \forall\epsilon \implies a \leq \frac{1-8b}{2-\lambda} \tag{41}$$

Set a and b as defined in equation (36). Observe that proposition holds iff $\frac{3\cdot b\cdot\lambda+1}{4-3\cdot\lambda} \leq \frac{1-8\cdot b}{2-\lambda}$. Equivalently, $b \leq \frac{2-2\cdot\lambda}{32-18\cdot\lambda-3\cdot\lambda^2}$ which holds uniformly across λ given the parameterisation of b. Result follows. □ Further algebra shows that,

$$\mathbb{E}_\epsilon[S_i(0)] - \mathbb{E}_\epsilon[S_i(3/4)] = c_1 \cdot \epsilon \tag{42}$$

Given $c_1 (< 0)$, a sufficient condition for this result to hold is $a = \frac{3\cdot b\lambda+1}{4-3\cdot\lambda}$, which holds by construction of a. Intuitively, $x^* = 0$ for $\epsilon \in [-1, 0]$ and $x^* = 3/4$ for $\epsilon \in (0, 1]$. As a corollary of the results above, one may show that

$$
\begin{aligned}
&\min_{\epsilon\in[-1,1]} \min\Big(\mathbb{E}_\epsilon[S_i(3/4), \mathbb{E}_\epsilon[S_i(0)]\Big) = \mathbb{E}_{\epsilon=-1}[S_i(3/4)] \\
&\max_{\epsilon\in[-1,1]} \max_{x\in[1/4,1/2)} \mathbb{E}_\epsilon[S_i(x)] = \mathbb{E}_{\epsilon=1}[S_i(1/4)] \\
&\mathbb{E}_{\epsilon=-1}[S_i(3/4)] - \mathbb{E}_{\epsilon=1}[S_i(1/4)] = c_2 > 0
\end{aligned} \tag{43}
$$

Proof of Equation (43).

- $\mathbb{E}_\epsilon[S_i(0)] = a \perp\!\!\!\perp \epsilon$. Moreover, $\min_\epsilon \mathbb{E}_\epsilon[S_i(3/4)] = \mathbb{E}_{\epsilon=-1}[S_i(3/4)] = a \cdot \frac{3\cdot\lambda}{4} + \frac{b\cdot\lambda}{2} + \frac{1}{4}$. For the selected a, b it follows that $\mathbb{E}_{\epsilon=-1}[S_i(3/4)] < \mathbb{E}_\epsilon[S_i(0)]$ for all λ.
- Second equality: Given monotonicity of $\mathbb{E}_\epsilon[S_i]$ on x for all ϵ within the support bracket, it follows that $\epsilon = \arg\max_{\epsilon'} \mathbb{E}_{\epsilon'}[S_i(1/4)] = 1$.
- Third equality: This result follows by construction of b and c_2. $\square$

Intuitively, our characterisation of a and b allowed us to distinguish across an "optimal region" given by policies $x \in \{0, 3/4\}$ and a "sub-optimal region" given by policies $x \in [1/4, 1/2)$. Even under the worst ϵ scenario, any sampling of the optimal region outperforms uniformly those in the sub-optimal region. Nonetheless, to avoid linear regret (i.e. to learn the optimal policy), the player must infer the sign of ϵ. ϵ can only be learned in the sub-optimal region what forces the learner to sample in this policy space leading to a regret $\mathcal{O}(K^{\frac{2}{3}})$.

From now on, let $\epsilon \in [-1, 0]$. Positive values of ϵ will be referred to either explicitly (i.e. $\epsilon = 1/2$) or more generally as $-\epsilon$. Using our results in equations (39), (40), (42) and (43) one may bound the ϵ-dependent regret of any policy π such that

$$
\begin{aligned}
\mathcal{R}(\pi \mid \epsilon) &= \mathbb{E}_\epsilon[K \cdot \sup_{x\in[0,1]} S_i(x) - \sum_i^K S_i(x_i)] = K \cdot S_i(0) - \sum_i^K \mathbb{E}_\epsilon[S_i(x_i)] \\
&\geq K S_i(0) - \sum_i^K \mathbb{E}_\epsilon\Big[S_i(0)\mathbb{1}(x_i \in [0, 1/4)) + S_i(1/4)\mathbb{1}(x_i \in [1/4, 1/2)) + S_i(3/4)\mathbb{1}(x_i \in [1/2, 1])\Big] \\
&= \big(S_i(0) - \mathbb{E}_\epsilon[S_i(1/4)]\big) \cdot \mathbb{E}_\epsilon[n_1] + \big(S_i(0) - \mathbb{E}_\epsilon[S_i(3/4)]\big) \cdot \mathbb{E}_\epsilon[n_2] \\
&\geq \mathbb{E}_{\epsilon=-1}[S_i(3/4)] - \mathbb{E}_{\epsilon=1} S_i(1/4)]\mathbb{E}[n_1] + \mathbb{E}_\epsilon[S_i(0)] - \mathbb{E}_\epsilon S_i(3/4)]\mathbb{E}[n_2] \\
&= c_2 \cdot \mathbb{E}_\epsilon[n_1] + c_1 \cdot \epsilon \cdot E[n_2]
\end{aligned} \tag{44}
$$

A similar result holds for $-\epsilon$. In particular,

$$\mathcal{R}(\pi \mid \epsilon) \geq -c_1 \cdot (-\epsilon) \cdot \mathbb{E}_\epsilon[n_3] + c_2 \cdot \mathbb{E}[n_1] \geq -c_1 \cdot \epsilon \cdot \mathbb{E}_\epsilon[n_3] \tag{45}$$

To proceed, we now restore on a Lemma proven in Annex A.2. in [Cesa-Bianchi et al., 2022]. A GMP adapted version of such Lemma can be summarised as follows.

Lemma 11. For every $\epsilon \in [-1, 0]$, $\mathbb{E}_{-\epsilon}[n_3] \geq \mathbb{E}_\epsilon[n_3] - c_3 \cdot \epsilon \cdot K\sqrt{\mathbb{E}_\epsilon[n_1]}$

Proof of Lemma 11 is delayed to the end of this Section. We can now use Lemma 11 to show that for $C < \infty$ uniformly for all $\epsilon \in [-1, 1]$, $\mathcal{R}(\pi \mid \epsilon) \geq C \cdot K^{\frac{2}{3}}$. Suppose by contradiction that $\mathcal{R}(\pi \mid \epsilon) < C \cdot K^{\frac{2}{3}}$. Start by presenting the claim below, which follows directly from equations (44) and (45)

$$\begin{aligned} \mathbb{E}_\epsilon(n_1) &\leq \frac{\mathcal{R}(\pi \mid \epsilon) - c_1 \cdot \epsilon \cdot \mathbb{E}_\epsilon[n_2]}{c_2} \leq \frac{\mathcal{R}(\pi \mid \epsilon)}{c_2} \leq \frac{C}{c_2} \cdot K^{\frac{2}{3}} \\ \mathbb{E}_\epsilon(n_2) &\leq \frac{\mathcal{R}(\pi \mid \epsilon) - c_2 \cdot \mathbb{E}_\epsilon[n_1]}{c_1 \cdot \epsilon} \leq \frac{C}{c_1 \cdot \epsilon} \cdot K^{\frac{2}{3}} \end{aligned} \tag{46}$$

We can now rewrite equation (45) using "$-\epsilon$ notation" and Lemma 11 such that

$$\begin{aligned} \mathcal{R}(\pi \mid \epsilon) &\geq c_1 \cdot \epsilon \cdot \mathbb{E}_{-\epsilon}[n_3] \geq c_1 \cdot \epsilon \cdot \left(\mathbb{E}_\epsilon[n_3] - c_3 \cdot \epsilon \cdot K\sqrt{\mathbb{E}_\epsilon[n_1]}\right) \\ &\geq c_1 \cdot \epsilon \cdot \left(K - \mathbb{E}_\epsilon[n_1] - \mathbb{E}_\epsilon[n_2] - c_3 \cdot \epsilon \cdot K \cdot \sqrt{\mathbb{E}_\epsilon[n_1]}\right) \\ &\geq c_1 \cdot \epsilon \cdot \left(K - \frac{C}{c_2}K^{\frac{2}{3}} - \frac{C}{c_1 \cdot \epsilon}K^{\frac{2}{3}} - c_3 \cdot \epsilon \cdot K \cdot \sqrt{\frac{C}{c_2}K^{\frac{2}{3}}}\right) \\ &= K \cdot c_1 \cdot \epsilon\left(1 - \frac{C}{c_2}K^{-\frac{1}{3}} - \frac{C}{c_1\epsilon}K^{-\frac{1}{3}} - c_3 \cdot \epsilon \cdot K^{\frac{1}{3}}\sqrt{\frac{C}{c_2}}\right) \end{aligned} \tag{47}$$

Now set $\epsilon = -K^{\frac{1}{3}}\sqrt{\frac{\sqrt{C \cdot c_2}}{-c_1 \cdot c_3}}$. By (37) $\epsilon \in [-1, 0)$. Plug in the value of ϵ in the expression above, and we immediately obtain

$$\begin{aligned} \mathcal{R}(\pi \mid \epsilon) &\geq \sqrt{\frac{\sqrt{C \cdot c_2 \cdot (-c_1)}}{c_3}} \cdot \left(1 - \frac{C}{c_2}K^{\frac{-1}{3}} - 2 \cdot \sqrt{\frac{c_3}{-c_1\sqrt{c_2}}} \cdot C^{\frac{3}{4}}\right)K^{\frac{2}{3}} \\ &\geq \frac{1}{4} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot (-c_1)}{c_3}} \cdot K^{\frac{2}{3}} \end{aligned} \tag{48}$$

Simplifying in the equation above leads to a contradiction of the kind $C > C$. We have proved that there exists a finite constant C, given by expression (37), such that for some ϵ, $\mathcal{R}(\pi \mid \epsilon) \geq C \cdot K^{\frac{2}{3}}$ for any policy π, hence $\mathcal{R}^{\text{GMP}}(\cdot) \geq C \cdot K^{\frac{2}{3}}$ $\square$

Sketch of a Proof of Lemma 11. We just include this proof for completeness. In fact, some steps of this proof have been omitted whenever results presented by [Cesa-Bianchi et al., 2022] are problem independent. The reader is referred to the relevant Annex for a complete characterisation of the proof, as well as for a detailed explanation of the notation below. To see why this result holds in our context recall that $\mathscr{P}_\epsilon^1$ is degenerate in V. Consequently, there is no regret loss associated to the restriction of the feedback function $\psi : (x, u, v) \mapsto (x, \mathbb{1}(x \geq v), \psi^{\emptyset}((x \geq v), u))$ to $\psi_2 : (x, v) \mapsto (x, \mathbb{1}(x \geq v))$. In fact, this result is a trivial implementation of the Simulation Lemma discussed in [Cesa-Bianchi et al., 2021]. A game $\mathcal{G}_2$ can be defined accordingly. Once this notion is clear, we may simply reproduce their intuitions with $\psi_x : [0,1] \to \{0,1\}, d \mapsto \mathbb{1}_{[3/4,1]}(x) + \mathbb{1}_{[1/2,3/4)}(x) \cdot \mathbb{1}_{[0,a+2b]}(d) + \mathbb{1}_{[0,1/4)}(x) \cdot \mathbb{1}_{[0,a]}(d)$. It now follows that

$$\mathbb{P}_1^\epsilon(z_i = 1|x_i) = \mathbb{P}_1^\epsilon(\tilde{z}_i = 1|\tilde{x}_i) = \begin{cases} 1, & x_i \in [3/4, 1] \\ a + 2b, & x_i \in [1/2, 3/4) \\ a + (1+\epsilon)b, & x_i \in [1/4, 1/2) \\ a, & x_i \in [0, 1/4) \end{cases} \tag{49}$$

Define $\tilde{n}_j$ for $j \in \{1, 2, 3\}$ analogously to equation (38). Replace the sub-optimal region of interest $(1/2, 3/4]$ in [Cesa-Bianchi et al., 2022] by $[1/4, 1/2)$. In particular define $I_i \coloneqq \{i \in \{1, \ldots, i\} \mid \tilde{x}_i \in [1/4, 1/2)\}$. Moreover, consider

$$Z_{i,s} \coloneqq \begin{cases} \emptyset & \text{if } s \notin I_i \\ \mathbb{1}(v_s < 1/2) & \text{if } s \in I_i \end{cases} \tag{50}$$

Observe that given the support of $v_s = \text{Supp}(V)$ it remains true that $\mathbb{1}(v_s < 1/2) = \mathbb{1}(x = v_s)$. Finally, we may simply write Pinker's Inequality for $Q^\epsilon(\tilde{x}_i \in [0, 1/4))$. Although $Z_{i,s}$ has been defined using the set $(1/2 > v_s)$ rather than $(1/2 < v_s)$ (as in [Cesa-Bianchi et al., 2022]), our characterisation of the inequality below for the measure

Q^ϵ over our set of interest remains valid

$$\sum_{z\in\{0,1\}} \log\Big(\frac{Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z)}{Q^{-\epsilon}(\mathbb{1}(1/2 > v_{i+1}) = z)}\Big) \cdot Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z) \leq 2 \cdot c_3^2 \cdot \epsilon^2 \tag{51}$$

This is because

$$\begin{aligned} &\sum_{z\in\{0,1\}} \log\Big(\frac{Q^\epsilon(\mathbb{1}(1/2 < v_{i+1}) = z)}{Q^{-\epsilon}(\mathbb{1}(1/2 < v_{i+1}) = z)}\Big) \cdot Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z) \\ = &\sum_{z\in\{0,1\}} \log\Big(\frac{Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z)}{Q^{-\epsilon}(\mathbb{1}(1/2 > v_{i+1}) = z)}\Big) \cdot Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z) \end{aligned} \tag{52}$$

Equations (45), (46) and (47) in [Cesa-Bianchi et al., 2022] now hold over our sub-optimal region of interest $[1/4, 1/2)$ and so does equation (48) and (49) in $[0, 1/4)$. The rest of the proof is not context specific and, hence, remains true for $\mathcal{G}_2$. □

9.3 Proof of Proposition 10: Optimality of Algorithm 2

This section shows the optimality of Algorithm 2 by presenting an upper bound on the algorithm and a matching lower bound on the LE-GMP. We start with the characterisation of an upper bound on the algorithm. For space considerations, we only include the steps which differ substantially from those of Algorithm 1 in Appendix 9.1.

Lower bound on estimated analogues. Assume that $\eta \leq \frac{p_\theta \cdot \gamma}{B+1}$. It follows that $\eta \cdot \hat{S}_{ib} \leq \eta \cdot (B+1)/(\gamma \cdot p_\theta) \leq 1$, where we used that $\hat{S}_{ib} \in [-(B+1)/(\gamma \cdot p_\theta), (B+1)/(\gamma \cdot p_\theta)]$ for all i and b. The inequalities in equation (28) remain correct.

Bounding first and second order terms. We can bound the first order term by

$$\mathbb{E}\Big[\sum_b q_{ib} \cdot \hat{S}_{ib}\Big] \leq \frac{1}{1-\gamma}(\mathbb{E}[\tilde{S}_i(x_i)] + 1/p_\theta) \tag{53}$$

Where we used the unbiasedness of $\hat{S}_{ib}$ and, again, that $\hat{S}_{ib} \leq (B+1)/(\gamma \cdot p_\theta)$,

and so does $\sum_b \hat{S}_{ib}$. For the second order term, we start by noting that $\mathbb{E}[\hat{S}_{ib}^2] \leq u_i^2/(p_{ib} \cdot p_\theta^2) + x_b^2/p_{ib} + (\frac{\lambda}{B})^2 \sum_{b'<b} \frac{1}{p_{ib'}}$. As a result,

$$\mathbb{E}[\sum_b p_{ib} \cdot \hat{S}_{ib}^2] \leq \frac{B}{p_\theta^2 \cdot (B+1)} \Big(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \Big) \tag{54}$$

Wrapping-up. Collect inequalities and replace the updated bounds for first and second order terms wherever is necessary,

$$\begin{aligned} &\eta \cdot \Big(\sup_x \mathbb{S}(x) - \frac{K(1+\lambda)}{B} \Big) \leq \\ &\log(B+1) + \frac{\eta}{1-\gamma} (\mathbb{E}[\tilde{\mathbb{S}}(x_i)] + \frac{1}{p_\theta}) + (e-2) \cdot \frac{\eta^2}{1-\gamma} \cdot \frac{K}{p_\theta^2} \cdot \frac{B}{B+1} \left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \right) \end{aligned} \tag{55}$$

Finally, we can solve for $\sup_x \mathbb{S}(x) - \mathbb{E}[\tilde{\mathbb{S}}(x_i)]$, such that

$$\begin{aligned} &\sup_x \mathbb{S}(x) - \mathbb{E}[\mathbb{S}(x_i)] \leq \\ &= K \left(\gamma + \frac{(1+\lambda)}{B} + \eta(e-2) \frac{1}{p_\theta^2} \frac{B}{B+1} \left(\frac{8B+7}{6} + \frac{\lambda^2}{\gamma} \right) \right) + \frac{\log(B+1)}{\eta} + \frac{1}{p_\theta} \end{aligned} \tag{56}$$

Tuning Parameters. Label efficient considerations do not affect FOCs for optimal parameter selection, so B, η, γ remain unchanged compared to GMP. It then follows that

$$\sup_x \mathbb{S}(x) - \mathbb{E}[\mathbb{S}(x_i)] \leq c_5 \cdot \log(K)^{\frac{1}{3}} \cdot K^{\frac{2}{3}} \cdot \frac{1}{p_\theta^2} \tag{57}$$

This result concludes the proof on an upper bound for Algorithm 2. □

Proof for the lower bound follows a similar approach as we only present the steps which differ substantially from benchmark results in Appendix 9.2. In addition, we make numerous references to the work of [Cesa-Bianchi et al., 2022], hence, we recommend the reader to go through this section in parallel to Proof of Claim A.2. in [Cesa-Bianchi et al., 2022].

Game Characterisation and Embedding Lemma. Simply consider the new game $\mathcal{G}' = \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}', \rho, \psi', \mathcal{P}, \mathcal{Q}\}$ where $\mathcal{Z}' = [0,1] \times \{0,1\} \cup \emptyset \times [0,1] \cup \emptyset \times \{0,1\}$, $\psi' : (x, u, v, \theta) \mapsto (x, \psi^{\emptyset}(\theta = 1, (\mathbb{1}(x \geq v), \psi^{\emptyset}((x \geq v), u))))$, and $\mathcal{Q} = \mathcal{B}e\imath n(\Theta)$. Define now $\mathcal{G}'_1 = \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}', \rho, \psi', \mathcal{P}^1_\epsilon, \mathcal{Q}\}$ where $\mathcal{P}^1_\epsilon$ is as defined in Appendix (9.2). Applying the Embedding Lemma as in Appendix 9.2, it now follows that $\mathcal{R}(\mathcal{G}'_1) \leq \mathcal{R}(\mathcal{G}')$.

Adapting Lemma 11. Lemma 11, which is a restatement of Claim A.2. in [Cesa-Bianchi et al., 2022], is the only piece within the proof of the lower bound that establishes a connection between the DGP $\mathcal{P}^1_\epsilon$ and the information structure in the game. Thus, it is the only item which needs to be modified for accommodating the new feedback structure. We do so in Lemma 12.

Lemma 12. Consider the setting, variables and parameters described in Appendix 9.2 up to equation (45), it follows that for every $\epsilon \in [-1, 1]$

$$\mathbb{E}_{-\epsilon}[n_3] \geq \mathbb{E}_{\epsilon}[n_3] - c_3 \cdot \epsilon \cdot K\sqrt{\mathbb{E}_{\epsilon}[n_1] \cdot p_\theta} \tag{58}$$

Proof: Consider the new simplified feedback structure $\psi : [0,1] \times \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\} \cup \emptyset \times \{0,1\} \to \{0, 1, \emptyset\}$, defined via $(x, v, \theta) \mapsto \psi^{\emptyset}(\theta = 1, \mathbb{1}(x \geq v))$. Characterisation of x_i remains as in [Cesa-Bianchi et al., 2022], but we define $z_1 := \psi(x_1, v_1, \theta_1)$ and $z_{i+1} := \psi(x_{i+1}, v_{i+1}, \theta_{i+1})$. Characterise ψ_x as in the benchmark case, namely: $\psi_x : [0,1] \to \{0,1\}$ defined via $d \mapsto \mathbb{1}_{[3/4,1]}(x) + \mathbb{1}_{[1/2,3/4)}(x) \cdot \mathbb{1}_{[0,a+2b]}(d) + \mathbb{1}_{[0,1/4)}(x) \cdot \mathbb{1}_{[0,a]}(d)$. $\tilde{x}_i$ remains unchanged compared to [Cesa-Bianchi et al., 2022], but we need to re-define $\tilde{z}_1 := \psi(\tilde{x}_1, \psi_{\tilde{x}_1}(\xi(\omega_1)), \theta_1)$

$$\tilde{z}_{i+1} := \begin{cases} \psi(\tilde{x}_{i+1}, v_{i+1}, \theta_{i+1}) & \tilde{x}_{i+1} \in [1/4, 1/2) \\ \psi(\tilde{x}_{i+1}, \psi_{\tilde{x}_{i+1}}(\xi(\omega_{i+1})), \theta_{i+1}) & \text{otherwise} \end{cases} \tag{59}$$

We can now see that

$$\mathbb{P}_1^\epsilon(z_i = 1|x_i) = \mathbb{P}_1^\epsilon(\tilde{z}_i = 1|\tilde{x}_i) = \begin{cases} p_\theta, & x_i \in [3/4, 1] \\ p_\theta \cdot (a + 2b), & x_i \in [1/2, 3/4) \\ p_\theta \cdot (a + (1+\epsilon)b), & x_i \in [1/4, 1/2) \\ p_\theta \cdot a, & x_i \in [0, 1/4) \end{cases} \tag{60}$$

Proceed as expected with the characterisation of $\tilde{n}_j$. Redefine set $I_i := \{i \in \{1, \dots, i\} \mid x_i \in [1/4, 1/2), \theta_i = 1\}$ to account for LIPC, and define $Z_{i,s}$ as in the benchmark case, under the updated "informative region" I_i

$$Z_{i,s} := \begin{cases} \emptyset & \text{if } s \notin I_i \\ \mathbb{1}(v_s < 1/2) & \text{if } s \in I_i \end{cases} \tag{61}$$

Measurability considerations and decomposition of conditional probabilities through Pinker's inequality remain true under the σ-algebra generated by the index updated $\overline{Z}_{K-1}$. So we only need to update

$$\sum_{z \in \{0,1\}} \log\left(\frac{Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z)}{Q^{-\epsilon}(\mathbb{1}(1/2 > v_{i+1}) = z)}\right) \cdot Q^\epsilon(\mathbb{1}(1/2 > v_{i+1}) = z) \leq 2 \cdot p_\theta \cdot c_3^2 \cdot \epsilon^2 \tag{62}$$

using the transformed probabilities defined in equation (60). Because Θ is independent from present and previous outcomes and actions, the history-conditional probability of selecting actions $\tilde{x}$ remains unchanged such that

$$\mathcal{D}_{\mathrm{KL}}(Q^\epsilon_{\overline{Z}_i} || Q^{-\epsilon}_{\overline{Z}_i}) \leq 2 \cdot p_\theta \cdot c_3^2 \cdot \epsilon^2 \cdot \mathbb{E}_\epsilon[\tilde{n}_1 \mid \cdot] \tag{63}$$

Further algebra, as in [Cesa-Bianchi et al., 2022], shows that $\mathbb{E}^{-\epsilon}(n_3) \geq \mathbb{E}_\epsilon[n_3] - c_3 \cdot \epsilon \cdot K \cdot \sqrt{\mathbb{E}_\epsilon[n_1] \cdot p_\theta}$. $\square$

Final Inequalities. We can finally update the inequalities in the benchmark case and use Lemma 12 to obtain the desired result. Commence by assuming for contradiction that $\mathcal{R}_K(\pi, \epsilon) < C \cdot K^{\frac{2}{3}}/p_\theta^2$, it follows that

$$\mathbb{E}_\epsilon(n_1) \leq \frac{C}{c_2 \cdot p_\theta^2} \cdot K^{\frac{2}{3}} \qquad \mathbb{E}_\epsilon(n_2) \leq \frac{C}{c_1 \cdot \epsilon \cdot p_\theta^2} \cdot K^{\frac{2}{3}} \tag{64}$$

Now use equation (64) and Lemma 12,

$$\begin{aligned}\mathcal{R}_K(\pi,\epsilon) &\geq c_1\epsilon \cdot (\mathbb{E}_\epsilon[n_3] - c_3\epsilon \cdot K\sqrt{\mathbb{E}_\epsilon[n_1] \cdot p_\theta} \\ &K \cdot c_1 \cdot \epsilon\left(1 - \frac{C}{c_2 \cdot p_\theta^2} \cdot K^{\frac{-1}{3}} - \frac{C}{c_1 \cdot \epsilon \cdot p_\theta^2} \cdot K^{\frac{-1}{3}} - c_3 \cdot \epsilon \cdot K^{\frac{1}{3}}\sqrt{\frac{C}{c_2} \cdot p_\theta}\right)\end{aligned} \tag{65}$$

Set ϵ as in the benchmark case, and assume again that $\mathcal{R}_K(\pi,\epsilon) < \frac{C \cdot K^{\frac{2}{3}}}{p_\theta^2}$,

$$\begin{aligned}\frac{C \cdot K^{\frac{2}{3}}}{p_\theta^2} > \mathcal{R}_K(\pi,\epsilon) &\geq \frac{1}{p_{\theta^2}} K^{\frac{2}{3}}\sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - \frac{C}{c_2} \cdot K^{\frac{-1}{3}} - \frac{C}{c_1 \cdot \epsilon} \cdot K^{\frac{-1}{3}} - c_3 \cdot \epsilon \cdot K^{\frac{1}{3}} \cdot p_\theta^{\frac{3}{2}}\sqrt{\frac{C}{c_2}}\right) \\ &\geq \frac{1}{p_{\theta^2}} K^{\frac{2}{3}}\sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - \frac{C}{c_2} \cdot K^{\frac{-1}{3}} - (1 + p_\theta^{\frac{3}{2}}) \cdot C^{\frac{3}{4}}\sqrt{\frac{c_3}{\sqrt{c_2} \cdot c_1}}\right)\end{aligned} \tag{66}$$

By bounding $p_\theta^{\frac{3}{2}} \leq 1$ (provided $p_\theta \leq 1$), we can recover the same expression as in equation (38) of [Cesa-Bianchi et al., 2022]. $1/p_\theta^2$ terms cancel out and a contradiction of the type $C > C$ follows. It must be the case then that $\mathcal{R}_K(\pi,\epsilon) \geq \frac{C \cdot K^{\frac{2}{3}}}{p_\theta^2}$ for any $p_\theta^2 \in (0,1]$. $\square$

9.4 Further Simulation Analysis

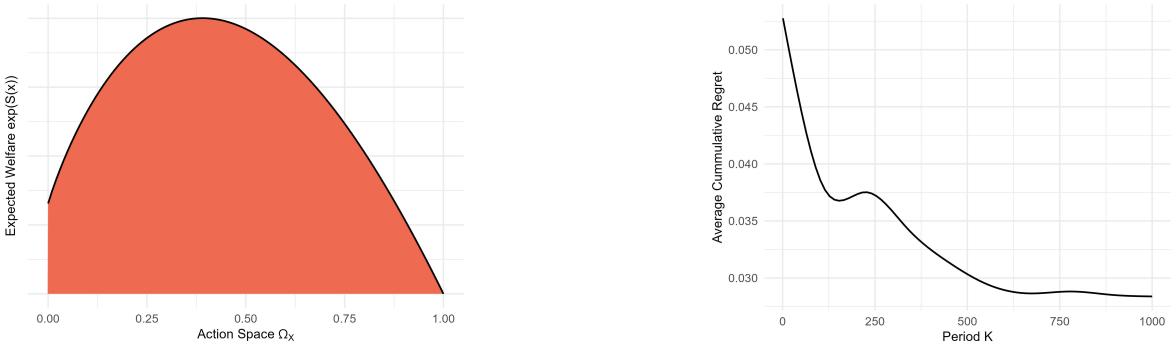
The following section provides additional simulation evidence on the theoretical guarantees presented in Section 6. Observe that we make specific emphasis on non-independent distributions, as these are the ones which are prompt to adverse selection dynamics.

Uniform non-linear degenerate case $U \sim \mathcal{U}[0,1], V = u^2$. This DGP introduces non-linear dependence in $F_{U,V}$, which yields a not-centred optimal policy x^*.

$$\mathbb{E}_U[S_i(x)] = \int_0^{\sqrt{x}} u - x + \lambda(x - u^2)du = \frac{x}{2} - x^{\frac{3}{2}} + \lambda x^{\frac{3}{2}} - \lambda\frac{x^{\frac{3}{2}}}{3} = \frac{x}{2} - x^{\frac{3}{2}}(1 - \frac{2\lambda}{3}) \tag{67}$$

For $\lambda = 0.7$, the expression above is maximised at ≈ 0.391.

Figure 9: Algorithm 1 given $U \sim \mathcal{U}[0,1]$, $V = u^2$



(a) $\mathbb{E}_{U,V}[S_i(x)]$ (b) Average Cumulative Regret

Average across 1,000 simulations. $\lambda = 0.7$. $K = 1000$. $\eta = 0.025$, $B = 10$, $\gamma = 0.029$.

Uniform-Uniform independent case $U \sim \mathcal{U}[0,1], V \sim \mathcal{U}[0,1]$. This DGP imposes independence across target variables, what should remove adverse selection and market unravelling dynamics.

$$\begin{aligned}\mathbb{E}_U[S_i(x)] &= \int_0^1 \int_0^1 \mathbb{1}(x \geq v)(u - x + \lambda(x - v))du\, dv = \int_0^1 \mathbb{1}(x \geq v)\big(\frac{1}{2} - x + \lambda x - \lambda v\big)dv = \\ &= \int_0^x \big(\frac{1}{2} - x + \lambda x - \lambda v\big)dv = \frac{1}{2}x + x^2\big(-1 + \frac{\lambda}{2}\big)\end{aligned} \tag{68}$$

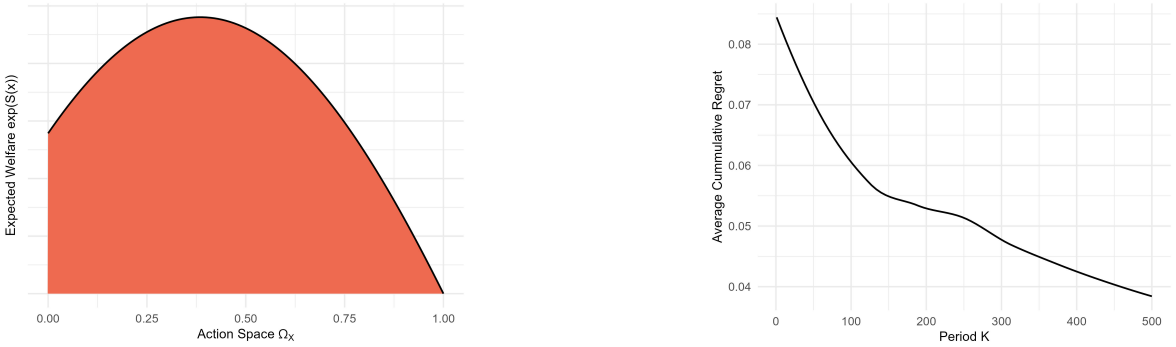
It follows that $x^* = 1/(4 - 2\lambda)$.

Bernouilli linear degenerate case $U \sim \mathcal{B}(p)$, $V = \frac{1}{2}u$. This DGP shows the consistency of Algorithm 1 with discrete distributions and unbounded densities.

$$\mathbb{E}_U[S_i(x)] = \mathbb{P}_U(u \leq 2x)\big(\mathbb{E}_U[u \mid u \leq 2x] - x\big) + \lambda \mathbb{P}_U(u \leq 2x) \cdot \big(x - \frac{1}{2}\mathbb{E}_U[u \mid u \leq 2x]\big) \tag{69}$$

$$\mathbb{E}_U[S_i(x)] = \begin{cases} (p - x) + \lambda(x - \frac{1}{2}p) = p\big(1 - \frac{\lambda}{2}\big) - (1 - \lambda)x & \text{if } x \in [1/2, 1] \\ (1 - p)(0 - x) + \lambda(1 - p)(x - 0) = -(1 - p)(1 - \lambda)x & \text{if } x \in [0, 1/2) \end{cases} \tag{70}$$

Figure 10: Algorithm 1 given $U \sim \mathcal{U}[0,1]$, $V =\sim \mathcal{U}[0,1]$

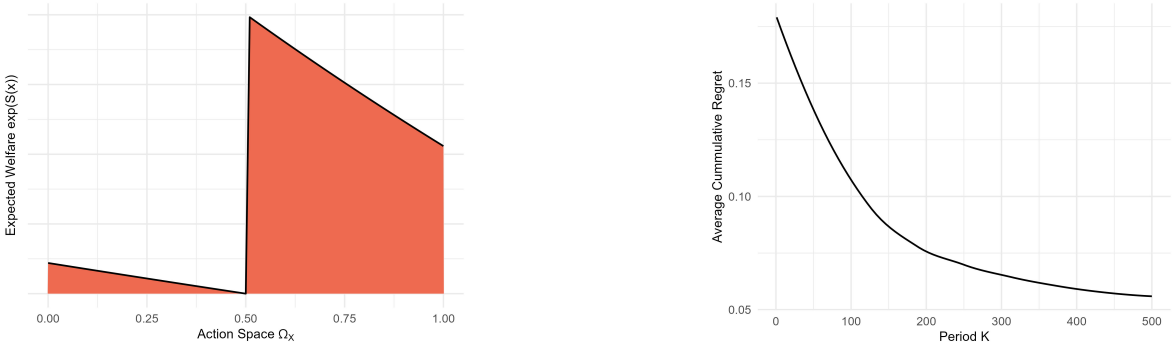


(a) $\mathbb{E}_{U,V}[S_i(x)]$ (b) Average Cumulative Regret

Average across 1,000 simulations. $\lambda = 0.7$. $K = 1000$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$.

It follows that $x^* = 1/2$ if $p \geq (1 - 2\lambda)$, $x^* = 0$ otherwise.

Figure 11: Algorithm 1 given $U \sim \mathcal{B}(p)$, $V = \frac{1}{2}u$



(a) $\mathbb{E}_{U,V}[S_i(x)]$ (b) Average Cumulative Regret

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$.

Beta linear degenerate case, $U \sim \mathcal{B}eta[\alpha, \beta], V = \frac{1}{2}u$. This DGP extends the ULDG to the Beta family of distributions, of which ULDC is a special case. Define $I_{f(x)}(\alpha, \beta) = \frac{1}{B(\alpha,\beta)} \int_0^{f(x)} t^{\alpha-1}(1-t)^{\beta-1} dt$.

$$\begin{aligned}\mathbb{E}_U[S_i(x)] &= \frac{1}{B(\alpha,\beta)}\int_0^{2x}(u-x+\lambda(x-1/2u))u^{\alpha-1}(1-u)^{\beta-1}du \\ &= \begin{cases} I_{2x}(\alpha+1,\beta)\big(1-\frac{\lambda}{2}\big)-(1-\lambda)xI_{2x}(\alpha,\beta) & \text{if } x\leq 1/2 \\ \frac{\alpha}{\alpha+\beta}\big(1-\frac{\lambda}{2}\big)-(1-\lambda)x & \text{if } x>1/2 \end{cases}\end{aligned} \tag{71}$$

Consider the special cases of $\mathcal{B}eta[2,1]$ and $\mathcal{B}eta[1/2,1/2]$. For $x<1/2$

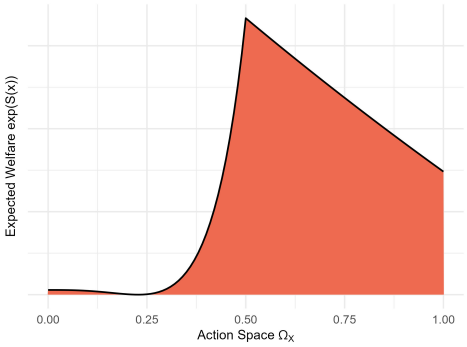
$$I_{2x}(\alpha,\beta) = \begin{cases} 4x^2 & \text{if } \alpha=2,\beta=1 \\ 1-\frac{2}{\pi}\arccos\sqrt{2}\sqrt{x} & \text{if } \alpha=\beta=1/2 \end{cases} \tag{72}$$

$$I_{2x}(\alpha+1,\beta) = \begin{cases} \frac{16x^3}{3} & \text{if } \alpha=2,\beta=1 \\ -\frac{1}{\pi}\sqrt{2-4x}\sqrt{x}+\frac{1}{\pi}\arcsin\sqrt{2}\sqrt{x} & \text{if } \alpha=\beta=1/2 \end{cases} \tag{73}$$

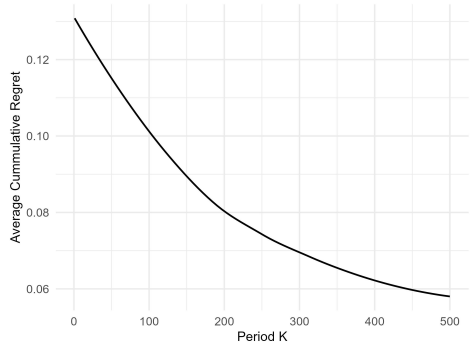
FOC reveals that $x^*=1/2$ for both cases.

Beta non-linear degenerate case $U\sim\mathcal{B}eta[2,1]$, $V=u^2$. This DGP extends the uniform non-linear case to the Beta family of distributions. Algebra in this example is surprisingly simple because we can rewrite $\mathbb{E}_{U,V}[S_i(x)]=\frac{2}{3}x^{\frac{3}{2}}+x^2(-1+\lambda/2)$, which is maximised at approximately $x^*=0.592$ for $\lambda=0.7$.

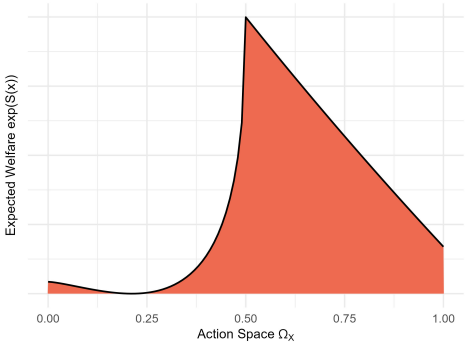
Figure 12: Algorithm 1 given $U \sim \mathcal{B}eta[\alpha, \beta]$, $V = \frac{1}{2}u$



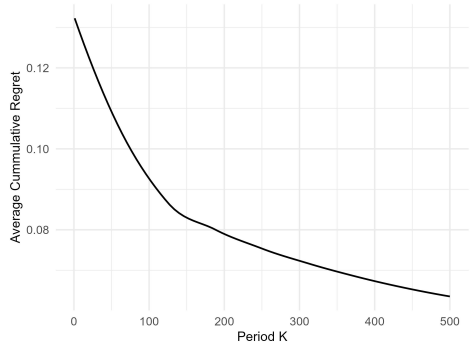
(a) $\mathbb{E}_{U,V}[S_i(x)], \alpha = 2, \beta = 1$



(b) Average Cumulative Regret, $\alpha = 2, \beta = 1$



(c) $\mathbb{E}_{U,V}[S_i(x)], \alpha = 0.5, \beta = 0.5$

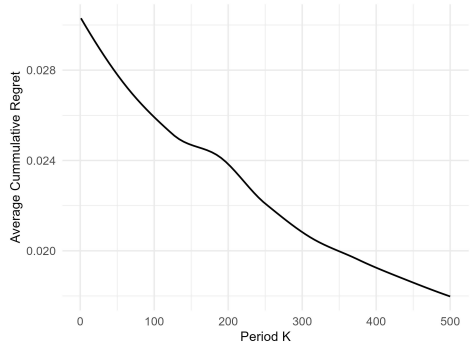


(d) Average Cumulative Regret, $\alpha = 0.5, \beta = 0.5$

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$.

Figure 13: Algorithm 1 given $U \sim \mathcal{B}eta[2, 1]$, $V = u^2$

(a) $\mathbb{E}_{U,V}[S_i(x)]$



(b) Average Cumulative Regret

Average across 1,000 simulations. $\lambda = 0.7$. $K = 500$. $\eta = 0.132$, $B = 10$, $\gamma = 0.029$.