Adaptive maximization of social welfare

(Preliminary draft)

Nicolò Cesa-Bianchi[*] Roberto Colomboni[†] Maximilian Kasy[‡]

September 14, 2023

Abstract

We consider the problem of repeatedly choosing policy parameters, such as tax or transfer rates, in order to maximize social welfare. Social welfare is a weighted sum of private utility and public revenue. The outcomes of earlier policy choices inform later choices. In contrast to multi-armed bandit models, utility is not observed, but needs to be indirectly inferred from the integral of the response function. In contrast to standard optimal tax theory, response functions need to be learned through policy choices.

We derive a lower bound on regret for this problem, and a matching adversarial upper bound on regret, for a variant of the Exp3 algorithm. In both cases, cumulative regret grows at a rate of $T^{2/3}$, up to a logarithmic term. This implies that (i) the social welfare maximization problem is harder than the multi-armed bandit problem (with a rate of $T^{1/2}$ for finite policy sets), and (ii) our proposed algorithm achieves the optimal rate. If we restrict attention to the stochastic setting and assume that social welfare is concave, however, we can achieve a rate of $T^{1/2}$ (for continuous policy sets), using a dyadic search algorithm.

While our discussion is initially restricted to a minimal, stylized optimal tax problem, we conclude the paper with an extension to nonlinear income taxation, and sketch an extension commodity taxation. We also compare the social welfare maximization problem to two related learning problems, monopoly pricing (which is easier), and price setting for bilateral trade (which is harder).

[*] Dipartimento di Informatica, Università degli Studi di Milano. nicolo.cesa-bianchi@unimi.it.

[†] Dipartimento di Informatica, Università degli Studi di Milano. roberto.colomboni@unimi.it.

[‡] Department of Economics, University of Oxford. maximilian.kasy@economics.ox.ac.uk. Maximilian Kasy was supported by the Alfred P. Sloan Foundation, under the grant "Social foundations for statistics and machine learning."

1 Introduction

Consider a policymaker who aims to maximize social welfare, defined as a weighted sum of utility across individuals. The policymaker can choose a policy parameter such as a sales tax rate, an unemployment benefit level, a health-insurance copay rate, etc. The policymaker does *not* directly observe the welfare resulting from their policy choices. They do, however, observe behavioral outcomes such as consumption of the taxed good, labor market participation, or health care expenditures. They can revise their policy choices over time in light of observed outcomes. How should such a policymaker act? This is the question that we study. To address this question, we bring together insights from welfare economics (in particular optimal taxation, Ramsey 1927; Mirrlees 1971; Baily 1978; Saez 2001) with insights from machine learning (in particular online learning and multi-armed bandits, Bubeck and Cesa-Bianchi 2012; Slivkins 2019; Lattimore and Szepesvári 2020).

In our baseline model, individuals arrive sequentially and make a single binary decision. In each period the policymaker chooses a tax rate that applies to this binary decision, and then observes the individual's response. They do not observe the individual's private utility. Social welfare is given by a weighted sum of private utility and public revenue. Later, we extend our model to nonlinear income taxation, where welfare weights vary as a function of individual earnings capacity, and sketch an extension to commodity taxation, where individual decisions involve a continuous consumption vector.

Our goal is to give guidance to the policymaker. We propose algorithms to maximize cumulative social welfare, and we provide guarantees for the performance of these algorithms. In doing so, we also show that welfare maximization is a harder learning problem than reward maximization in the multi-armed bandit setting. Private utility in our baseline model is equal to consumer surplus, which is given by the integral of demand. In order to learn this integral, we need to learn demand for counterfactual, suboptimal tax rates. This drives the difficulty of the learning problem.

A lower bound on regret Our main theorems provide lower and upper bounds on cumulative regret. Cumulative regret is defined as the difference in welfare between the *chosen* sequence of policies and the *best* possible constant policy. We consider both stochastic and adversarial regret. The former assumes that preference parameters are drawn i.i.d. from some distribution, whereas the latter allows for arbitrary sequences of preference parameters.

We first prove a stochastic (and thus also adversarial) lower bound on regret, for any possible algorithm. Our proof of this bound constructs a family of possible distributions for preferences. This family is such that there are two candidate policies which are potentially optimal. The difference in welfare between these two policies depends on the integral of demand over intermediate policy values. In order to learn which of the two candidate policies is optimal, we need to learn behavioral responses for intermediate policies, which are strictly suboptimal. Because of the need to probe these suboptimal policies sufficiently often, we obtain a lower bound on regret which grows at a rate of $T^{2/3}$, even if we restrict our attention to settings with finite, known support for preference parameters and policies. This rate is worse than the worst-case rate for bandits of $T^{1/2}$.

A matching upper bound on adversarial regret for modified Exp3 We next propose an algorithm for the adaptive maximization of social welfare. Our algorithm is a modification of the well-known Exp3 algorithm (Auer et al., 2002). Exp3 is based on an unbiased estimate of cumulative welfare for each policy. The probability of choosing a given policy is proportional to the exponential of this estimate of cumulative welfare, times some rate parameter. Relative to Exp3, we require two modifications for our setting. First, we need to discretize the continuous policy space. Second, and more interestingly, we need additional exploration of counterfactual policies, including some policies that are clearly sub-optimal, in order to learn welfare for the policies which are contenders for the optimum. This need for additional exploration again arises because of the dependence of welfare on the integral of demand over counterfactual policy choices. For our modified Exp3 algorithm, we prove an adversarial (and thus also stochastic) upper bound on regret. We show that, for an appropriate choice of tuning parameters, worst case cumulative regret over all possible sequences of preference parameters grows at a rate of $T^{2/3}$, up to a logarithmic term. The algorithm thus achieves the best possible rate.

Since stochastic regret (averaged over sequences of willingness to pay) is always less or equal than adversarial regret (for the worst-case sequence), the stochastic lower bound immediately implies a corresponding adversarial lower bound, and the adversarial upper bound implies a corresponding stochastic upper bound. Since the rates for our stochastic lower and adversarial upper bound coincide, up to a logarithmic term, we have a complete characterization of learning rates for the welfare maximization problem.

Improved stochastic bounds for concave social welfare The proof of our lower bound on regret is based on the construction of a distribution of preferences which delivers a non-concave social welfare function. If we restrict attention to the stochastic setting, where preferences are i.i.d. over time, and if we assume that social welfare is concave, then we can improve upon this bound on regret. We prove a lower bound on stochastic regret, under the assumption of concavity, which grows at the rate of $T^{1/2}$. We then propose a dyadic search algorithm which achieves this rate, up to logarithmic terms. This dyadic search algorithm maintains an "active interval," containing the optimal policy with high probability, which is narrowed down over time. Only policies within the active interval are sampled.

Extensions to non-linear income taxation and to commodity taxation Our discussion up to this point focuses on a minimal, stylized case of an optimal tax problem, where individual actions are binary, and the policy imposes a tax on this binary action. Our arguments generalize, however, to more complicated and practically relevant settings. This includes optimal nonlinear income taxation, as in Mirrlees (1971); Saez (2001), and commodity taxation for a bundle of goods, as in Ramsey (1927). For nonlinear income taxation, different tax rates apply at different income levels, and welfare weights depend on individual earnings capacity. In Section 5, we discuss an extension of our tempered Exp3 algorithm to nonlinear income taxation, and characterize its regret. For commodity taxation, different tax rates apply to different goods, and consumption decisions are continuous vectors. In Appendix A we sketch an extension of our algorithm to commodity taxation, but leave its characterization for future research.

Roadmap The rest of this paper proceeds as follows. We conclude this introduction with a discussion of some related work and relevant references. Section 2 introduces our setup, formally defines the adversarial and stochastic settings, and compares our setup to related learning problems. Section 3 provides lower and upper bounds on regret in the adversarial and stochastic settings. Section 4 restricts attention to the stochastic setting with concave social welfare, and provides improved regret bounds for this setting. Section 5 discusses an extension of our baseline model to non-linear income taxation. Appendix A sketches another extension of our baseline model to commodity taxation. All proofs can be found in Appendix B.

1.1 Background and literature

To put our work in context, it is useful to contrast our framework with the standard approach in public finance and optimal tax theory, and with the frameworks considered in machine learning and the multi-armed bandit literature.

Optimal taxation Optimal tax theory, and optimal policy theory more generally, is concerned with the maximization of social welfare, where social welfare is understood as a (weighted) sum of subjective utility across individuals (Ramsey, 1927; Mirrlees, 1971; Baily, 1978; Saez, 2001; Chetty, 2009). A key tradeoff in such models is between, first, *redistribution* to those with higher welfare weights, and second, the efficiency cost of behavioral responses to tax increases. Such *behavioral responses* might reduce the tax base.

Optimal tax problems are defined by normative parameters (such as welfare weights for different individuals), as well as empirical parameters (such as the elasticity of the tax base with respect to tax rates). The typical approach in public finance uses historical or experimental variation to estimate the relevant empirical parameters (causal effects, elasticities). These estimated parameters are then plugged into formulas for optimal policy choice, which are derived from theoretical models. The implied optimal policies are finally implemented, without further experimental variation. Kasy (2018) made a case for directly choosing the optimal policy based on the data, rather than plugging elasticity estimates into optimal policy formulas. The proposed approach uses Gaussian process priors.

Multi-armed bandits The standard approach of public finance, which separates elasticity estimation from policy choice, contrasts with the adaptive approach that characterizes decision-making in many branches of AI, including online learning, multi-armed bandits, and reinforcement learning. Multi-armed bandit algorithms, in particular, trade off *exploration* and *exploitation* over time (Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020). Exploration here refers to the acquisition of information for better future policy decisions, while exploitation refers to the use of currently available information for optimal policy decisions at the present moment. The goal of bandit algorithms is to maximize a stream of rewards, which requires an optimal balance between exploration and exploitation. Bandit algorithms for the stochastic setting are characterized by optimism in the face of uncertainty: Policies with uncertain payoff should be tried until their expected payoff is clearly suboptimal.

Bandit algorithms (and similarly, adaptive experimental designs for informing policy choice, as in Russo 2020; Kasy and Sautmann 2021) are not directly applicable to social welfare maximization problems, such as those of optimal tax theory. The reason is that bandit algorithms maximize a stream of *observed* rewards. By contrast, social welfare as conceived in welfare economics is based on *unobserved* subjective utility.

Bandit-type approaches have been applied to a number of other economic and financial scenarios in the literature where rewards *are* observable. These include monopoly pricing (Kleinberg and Leighton, 2003) (see also the survey den Boer 2015), second-price auctions (Cesa-Bianchi et al., 2015; Weed et al., 2016; Cesa-Bianchi et al., 2017), first-price auctions (Han et al., 2020b,a)—see also (Kolumbus and Nisan, 2022; Feng et al., 2021)—combinatorial auctions (Daskalakis and Syrgkanis, 2022), bilateral trading Cesa-Bianchi et al. (2021), and the newsvendor problem (Lugosi et al., 2022).

2 Setup

At each time $i = 1, 2, \dots, T$, one individual arrives who is characterized by an unknown willingness to pay $v_i \in [0, 1]$. This individual is exposed to a tax rate x_i, and makes a binary decision $y_i = \mathbf{1}(x_i \leq v_i)$. The implied public revenue is $x_i \cdot y_i$. The implied private welfare is $\max(v_i - x_i, 0)$. We define social welfare as a weighted sum of public revenue and private welfare, with a weight λ for the latter. Social welfare for time period i is therefore given by

$$U_i(x_i) = \underbrace{x_i \cdot \mathbf{1}(x_i \leq v_i)}_{\text{Public revenue}} \quad + \quad \lambda \cdot \underbrace{\max(v_i - x_i, 0)}_{\text{Private welfare}}. \tag{1}$$

After period i, we observe y_i and the tax rate x_i, but nothing else. In particular, we do *not* observe welfare $U_i(x_i)$.

We can rewrite social welfare $U_i(x)$ as follows. Denote $G_i(x) = \mathbf{1}(v_i \geq x)$, so that $y_i =$

$G_i(x_i)$. This is the individual demand function. Then private welfare can be written as $\max(v_i - x, 0) = \int_x^1 G_i(x')dx'$. That is, due to the absence of income effects, private utility, compensating variation, and equivalent variation coincide with consumer surplus, given by integrated demand. This implies

$$U_i(x) = \underbrace{x \cdot G_i(x)}_{\text{Public revenue}} + \underbrace{\lambda \cdot \int_x^1 G_i(x')dx'}_{\text{Private welfare}}. \tag{2}$$

We consider algorithms for the choice of x_i which might depend on the observable history $(x_j, y_j)_{j=1}^{i-1}$, as well as possibly a randomization device.

Notation For the *adversarial* setting, we will consider cumulative demand and welfare, denoted by blackboard bold letters, summing across $j = 1, \ldots, i$. In particular,

$$\mathbb{G}_i(x) = \sum_{j \leq i} G_i(x), \qquad \mathbb{U}_i(x) = \sum_{j \leq i} U_i(x), \qquad \mathbb{U}_i = \sum_{j \leq i} U_j(x_j).$$

$\mathbb{G}_i(x)$ and $\mathbb{U}_i(x)$ are cumulative demand and welfare for a counterfactual, fixed policy x. $\mathbb{U}_i$, without an argument, is the cumulative welfare for the policies x_j actually chosen.

For the *stochastic* setting, we will analogously consider expected demand and expected welfare, denoted by boldface letters. The expectation is taken across some stationary distribution μ of v_i, where v_i is statistically independent of x_i, and of v_j for $j \neq i$. In particular,

$$\boldsymbol{G}(x) = E[G_i(x)], \qquad \boldsymbol{U}(x) = E[U_i(x)].$$

2.1 Regret

The adversarial case Following the literature, we consider regret for both the adversarial and the stochastic setting. In the adversarial setting, we allow for arbitrary sequences of willingness to pay, $\{v_i\}_{i=1}^T$. We compare the expected performance of any given algorithm for choosing $\{x_i\}_{i=1}^T$ to the performance of the best possible constant policy x. This comparison yields cumulative expected regret, which is given by

$$\mathcal{R}_T(\{v_i\}_{i=1}^T) = \sup_x E\left[\mathbb{U}_T(x) - \mathbb{U}_T \middle| \{v_i\}_{i=1}^T\right]. \tag{3}$$

The expectation in this expression is taken over any possible randomness in the tax rates x_i chosen by the algorithm; there is no other source of randomness.

The stochastic case We also consider the stochastic setting. In this setting, we add structure by assuming that the v_i are i.i.d. draws from some distribution μ on $[0, 1]$, with implied demand function $\boldsymbol{G}(x) = P(v_i \geq x)$. This demand function is identified by the regression

$$\boldsymbol{G}(x) = E[y_i | x_i = x].$$

The expectation in this expression is taken over the distribution of v_i, which is presumed to be independent of the tax rate x. Expected welfare for this distribution of v_i is given by

$$\boldsymbol{U}(x) = x \cdot \boldsymbol{G}(x) + \lambda \int_x^1 \boldsymbol{G}(x')dx'.$$

Cumulative expected regret in the stochastic case equals

$$\begin{aligned} \mathcal{R}_T(\boldsymbol{G}) &= \sup_x E\left[\mathbb{U}_T(x) - \mathbb{U}_T\right] \\ &= T \cdot \sup_x \boldsymbol{U}(x) - E\left[\sum_{i \leq T} \boldsymbol{U}(x_i)\right]. \end{aligned} \tag{4}$$

The expectation in this expression is taken over both any possible randomness in the tax rates x_i, and the i.i.d. draws v_i.

Lower and upper bounds Below, we will derive lower and upper bounds for adversarial and stochastic regret. A lower bound on adversarial regret has to hold for any algorithm and any sequence $\{v_i\}_{i=1}^T$. A lower bound on stochastic regret has to hold for any algorithm and any stationary distribution μ of v_i. A lower bound on stochastic regret immediately implies a lower bound on adversarial regret, since the supremum over sequences $\{v_i\}_{i=1}^T$ exceeds the expectation over such sequences, generated from any distribution μ.

An adversarial upper bound on regret has to hold for a given algorithm and any sequence $\{v_i\}_{i=1}^T$. Such an adversarial upper bound again immediately implies a stochastic upper bound on regret, by the same argument as above. When an adversarial upper bound coincides with a stochastic lower bound, in terms of rates of regret, it follows that the proposed algorithm is rate efficient, for both stochastic and adversarial regret.

2.2 Comparison to related learning problems

Before proceeding with our analysis of regret, we take a step back, and compare our learning problem to two related problems that have received some attention in the literature. The first of these is the adaptive **monopoly pricing** problem; see for instance Kleinberg and Leighton (2003). This problem is equivalent to our setting when we set $\lambda = 0$, interpret x as a price, and U_i^{MP} as monopolist profits:

$$U_i^{\mathrm{MP}}(x) = x_i \cdot \mathbf{1}(x_i \leq v_i) = \underbrace{x \cdot G_i(x).}_{\text{Monopolist revenue}} \tag{5}$$

Table 1: Regret rates for different learning problems

Model	Policy space		Objective function	
	Discrete	Continuous	Pointwise	One-sided Lipschitz
Monopoly price setting	$T^{1/2}$	$T^{2/3}$	Yes	Yes
Optimal taxation	$T^{2/3}$	$T^{2/3}$	No	Yes
Bilateral trade	$T^{2/3}$	T	No	No

Notes: This table shows the efficient rates of regret for different learning problems. Rates are up to logarithmic terms, and apply to both the stochastic and the adversarial setting. Regret rates are shown for the discrete case, where the space of policies x is restricted to a finite set, and the continuous case, where x can take any value in $[0,1]$. The columns on the right describe the properties of the objective function in each problem, which drive the differences in regret rates.
Rates for the continuous monopoly price setting case are from Kleinberg and Leighton (2003); the discrete case reduces to a standard bandit problem. Rates for the continuous bilateral trade case are from Cesa-Bianchi et al. (2021); the discrete case is discussed in forthcoming work by some of the authors. Rates for the optimal taxation case are proven in this paper.

As in our adaptive taxation setting, the feedback received at the end of period i is

$$y_i = G_i(x_i) = \mathbf{1}(x_i \leq v_i).$$

Another related problem is price setting for **bilateral trade**, see for instance Cesa-Bianchi et al. (2021). In this problem, welfare $U_i^{\mathrm{BT}}(x)$ is given by the sum of seller and buyer welfare. Trade happens if and only if both sides agree to transact at the proposed price. Buyer willingness to pay is given by v_i^b, while the seller is willing to trade at prices above v_i^s.

$$\begin{aligned} U_i^{\mathrm{BT}}(x) &= \mathbf{1}(v_i^b \geq x) \cdot \max(x - v_i^s, 0) &+& \quad \mathbf{1}(v_i^s \leq x) \cdot \max(v_i^b - x, 0) \\ &= G_i^b(x) \cdot \underbrace{\int_0^x G_i^s(x')dx'}_{\text{Seller welfare}} &+& \quad G_i^s(x) \cdot \underbrace{\int_x^1 G_i^b(x')dx'}_{\text{Buyer welfare}}. \end{aligned} \tag{6}$$

Feedback in this case is a little richer: We observe both whether the buyer b would have accepted the posted price, and whether the seller would have accepted this price,

$$y_i^b = G_i^b(x_i) = \mathbf{1}(x_i \leq v_i^b) \qquad \text{and} \qquad y_i^s = G_i^s(x_i) = \mathbf{1}(x_i \geq v_i^s).$$

Lipschitzness and information requirements The difficulty of the learning problem in each of these models critically depends on (i) the Lipschitz properties of the welfare function, and (ii) the information required to evaluate welfare at a point. We say that a generic welfare function $W : [0,1] \to \mathbb{R}$ is one-sided Lipschitz if $W(x+\varepsilon) \leq W(x) + \varepsilon$ for all $0 \leq x \leq 1$ and all $0 \leq \varepsilon \leq 1 - x$. We say that learning $W(\cdot)$ requires only pointwise information if $W(x)$ is a function of $G(x)$, and does not depend on $G(\cdot)$ otherwise. One-sided Lipschitzness allows us to bound the approximation error of a learning algorithm operating on a finite subset of the set of policies. Pointwise information allows us to avoid exploring policies that are clearly suboptimal, when we aim to learn the optimal policy.

Table 1 summarizes the Lipschitz properties and information requirements in each of the three models; the following justifies the claims made in Table 1:

1. For **monopoly pricing**, welfare $U_i^{\mathrm{MP}}(x)$ is one-sided Lipschitz and only depends on $G_i(x)$ pointwise.

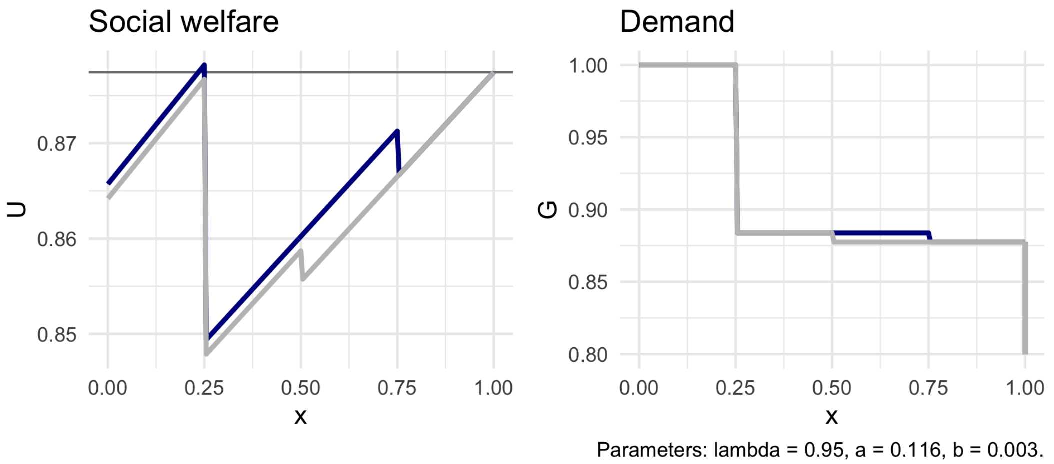
2. For **optimal taxation**, welfare $U_i(x)$ is one-sided Lipschitz and depends on both $G_i(x)$ at the given x (pointwise), and on an integral of $G_i(x')$ for a range of values of x' (non-pointwise).

3. For **bilateral trade**, welfare $U_i^{\mathrm{BT}}(x)$ is not one-sided Lipschitz and depends on both $G_i^b(x)$ and $G_i^s(x)$ (pointwise), as well as the integrals of $G_i^b(x')$ and $G_i^s(x')$ (non-pointwise).

These properties suggest a ranking in terms of the difficulty of the corresponding learning problems, and in particular in terms of the rates of divergence of cumulative regret: The information requirements of optimal taxation are stronger than those of monopoly pricing, but its continuity properties are more favorable than those of bilateral trade. This intuition is correct, as shown by Table 1. The rates for monopoly pricing and for bilateral trade are known from the literature. In this paper we prove corresponding rates for optimal taxation.

In comparing optimal taxation and monopoly pricing to conventional multi-armed bandits, it is worth emphasizing that there are two distinct reasons for the slower rate of convergence. First, the continuous support of x, as opposed to a finite number of arms, which is shared by optimal taxation and monopoly pricing. Second, the requirement of additional exploration of sub-optimal policies for the optimal tax problem. As shown in Table 1, the continuous support alone is enough to slow down convergence, with no extra penalty for the additional exploration requirement, in terms of rates. If, however, we restrict our attention to a discrete set of feasible policies x, then monopoly pricing reduces to a multi-armed bandit problem, with a minimax regret rate of $T^{1/2}$. The optimal tax problem, by contrast, still has a rate of $T^{2/3}$, even if we restrict our attention to the case of finite known support for v and x, as shown by the proof of Theorem 1 below.

Figure 1: Construction for proving the lower bound on regret



Notes: This figure illustrates our construction for proving the lower bound on regret. The relative social welfare of policies 1 and .25 depends on the sign of ϵ. The dark line corresponds to $\epsilon = -1$, the bright line to $\epsilon = 1$. In order to distinguish between these two, we must learn demand in the intermediate interval $[.5, .75]$.

3 Stochastic and adversarial regret bounds

We now turn to our main theoretical results, lower and upper bounds on stochastic and adversarial regret for the problem of social welfare maximization. We first prove a lower bound on stochastic regret, which applies to any algorithm, and which immediately implies a lower bound on adversarial regret. We then introduce the algorithm Tempered Exp3 for Social Welfare. We show that, for an appropriate choice of tuning parameters, this algorithm achieves the rates of the lower bound on regret, up to a logarithmic term. Formal proofs of these bounds can be found in Appendix B.

3.1 Lower bound

Theorem 1 (Lower bound on regret)**.** *Consider the setup of Section 2. There exists a constant $C > 0$ such that, for any randomized algorithm for the choice of $x_1, x_2, \ldots$ and any time horizon $T \in \mathbb{N}$, the following holds.*

1. *There exists a distribution μ on $[0,1]$ with associated demand function $\boldsymbol{G}$ for which the stochastic cumulative expected regret $\mathcal{R}_T(\boldsymbol{G})$ is at least $C \cdot T^{2/3}$.*

2. *There exists a sequence $(v_1, \ldots, v_T)$ for which the adversarial cumulative expected regret $\mathcal{R}_T(\{v_i\}_{i=1}^T)$ is at least $C \cdot T^{2/3}$.*

The proof of Theorem 1 can be found in Appendix B. The adversarial lower bound follows immediately from the stochastic lower bound, since worst case regret (over possible sequences of v_i) is bounded below by average regret (over i.i.d. draws of v_i), for any distribution of v_i.

Sketch of proof To prove the stochastic lower bound we construct a family of distributions for v_i that is indexed by a parameter $\epsilon \in [-1, 1]$. The distributions in this family have four points of support, $(1/4, 1/2, 3/4, 1)$. The probability of these points is given by

$$(a, (1+\epsilon)b, (1-\epsilon)b, 1-a-2b)\,.$$

The values of a and b are chosen such that (i) the two middle points $1/2, 3/4$ are far from optimal, for any value of ϵ, and (ii) learning which of the two end points $(1/4, 1)$ is optimal requires sampling from the middle.[1] For each $\epsilon \in [-1, 1]$, denote the demand function associated to μ^ϵ by $\boldsymbol{G}^\epsilon$, and the expected social welfare associated to $\boldsymbol{G}^\epsilon$ by $\boldsymbol{U}^\epsilon$. Property (ii) holds because of the integral term $\int_{\frac{1}{4}}^{1} \boldsymbol{G}^\epsilon(x')dx'$, which shows up in $\boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(1/4)$. This construction is illustrated in Figure 1. This figure shows plots of $\boldsymbol{G}^\epsilon$ and of $\boldsymbol{U}^\epsilon$ for $\lambda = .95$ and $\epsilon \in \{\pm 1\}$.

The difference in welfare $\boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(1/4)$ of the two candidates optimal policies $1/4$ and 1 depends on the sign of ϵ. In order not to suffer linear expected regret, any learning algorithm needs to sample policies from points that are informative about this sign. The only points that are informative are those in the region $(1/2, 3/4]$, where welfare is bounded away from optimal welfare. Exploring in this sub-optimal region forces us to accumulate regret along the way which grows at a rate of at least $T^{2/3}$.

3.2 An algorithm that achieves the lower bound

We next introduce an algorithm that allows us to essentially achieve the lower bound on regret, in terms of rates. Algorithm 1 is a modification of the well-known Exp3 algorithm. Conventional Exp3, for the multi-armed bandit setting, uses inverse probability weighting to construct an unbiased estimator $\widehat{\mathbb{U}}_k$ of the cumulative payoff of each arm k. A given arm is then chosen with probability proportional to $\exp(\eta \cdot \widehat{\mathbb{U}}_{ik})$, where η is a tuning parameter.

Modifications relative to standard Exp3 Relative to this standard algorithm, we require three modifications. First, we discretize the continuous support $[0, 1]$ of x, restricting attention to the grid of policy values $\tilde{x}_k = (k-1)/K$. Second, since welfare $U_i(x)$ is not directly observed for the chosen policy x, we need to estimate it indirectly. In particular, we first form an estimate $\widehat{\mathbb{G}}_{ik}$ of cumulative demand for each of the policy values $\tilde{x}_k$, using inverse probability weighting. We then use this estimated demand, interpolated using a step-function, to form estimates of cumulative social welfare, $\widehat{\mathbb{U}}_{ik} = \tilde{x}_k \cdot \widehat{\mathbb{G}}_{ik} + \frac{\lambda}{K} \cdot \sum_{k'>k} \widehat{\mathbb{G}}_{ik'}$. Third, we introduce some additional exploration, relative to Exp3. Since social welfare depends on counterfactual

[1] Specifically, $a := \frac{(1-\lambda)\cdot(136-99\cdot\lambda)}{2\cdot(4-3\cdot\lambda)\cdot(24-17\cdot\lambda)}$, and $b := \frac{1-\lambda}{2\cdot(24-17\cdot\lambda)}$. These two constants are strictly greater than zero, and satisfy $1 - a - 2 \cdot b > 0$.

Algorithm 1 Tempered Exp3 for Social Welfare

```
Require: Tuning parameters $K$, $\gamma$ and $\eta$.
Calculate evenly spaced grid-points $\tilde{x}_k = (k-1)/K$,
and initialize $\widehat{\mathbb{G}}_{1k} = 0$ and $\widehat{\mathbb{U}}_{1k} = 0$ for $k = 1, \ldots, K+1$.
for individual $i = 1, 2, \ldots, T$ do
    For all $k = 1, 2, \ldots, K+1$, set                                  {Assignment probabilities}
        $p_{ik} = (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_{ik})}{\sum_{k'} \exp(\eta \cdot \widehat{\mathbb{U}}_{ik'})} + \frac{\gamma}{K+1}.$   (7)
    Choose $k_i$ at random according to the probability distribution $(p_{i,1}, \ldots, p_{i,K+1})$.
    Set $x_i = \tilde{x}_{k_i}$, and query $y_i$ accordingly.
    For all $k = 1, 2, \ldots, K+1$, set                                  {Estimated demand}
        $\widehat{\mathbb{G}}_{i+1,k} = \widehat{\mathbb{G}}_{i,k} + y_i \cdot \frac{\mathbf{1}(k_i = k)}{p_{ik}}.$   (8)
    For all $k = 1, 2, \ldots, K+1$, set                                  {Estimated welfare}
        $\widehat{\mathbb{U}}_{i+1,k} = \tilde{x}_k \cdot \widehat{\mathbb{G}}_{i+1,k} + \frac{\lambda}{K} \cdot \sum_{k' > k} \widehat{\mathbb{G}}_{i+1,k'}.$   (9)
end for
```

policy choices, we need to explore policies that are away from the optimum, in order to learn the relative welfare of approximately optimal policy choices. This is achieved in our algorithm by mixing the Exp3 assignment distribution with a uniform distribution, with a mixing weight γ that is another tuning parameter.

Theorem 2 (Adversarial upper bound on regret of Tempered Exp3 for Social Welfare)**.** *Consider the setup of Section 2, and Algorithm 1. Assume that $(K+1)\eta < \gamma$.*
Then for any sequence $(v_1, \ldots, v_T)$ expected regret $\mathcal{R}_T(\{v_i\}_{i=1}^T)$ is bounded above by

$$\left(\gamma + \eta \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2K+1}{6} + \tfrac{\lambda^2}{\gamma}\right) + \tfrac{\lambda}{K}\right) \cdot T + \tfrac{\log(K+1)}{\eta}. \tag{10}$$

Suppose additionally that $\gamma = c_1 \cdot \left(\frac{\log(T)}{T}\right)^{1/3}$, $\eta = c_2 \cdot \gamma^2$, and $K = c_3/\gamma$, for some constants c_1, c_2, c_3. Then expected regret $\mathcal{R}_T(\{v_i\}_{i=1}^T)$ is bounded above by

$$c_4 \cdot \log(T)^{1/3} T^{2/3}, \tag{11}$$

for some constant c_4.

Corollary 1 (Stochastic upper bound on regret of Tempered Exp3 for Social Welfare)**.** *Under the assumptions of Theorem 2, suppose additionally that v_i is drawn i.i.d. from some distribution with associated demand function $\boldsymbol{G}$. Then expected regret $\mathcal{R}_T(\boldsymbol{G})$ is bounded above by the* *same expressions as in Theorem 2.*

The proof of Theorem 2 can again be found in Appendix B.

Tuning The statement of the theorem leaves the constants c_1, c_2, c_3 in the definition of the tuning parameters unspecified. Suppose we wish to choose the tuning parameters so as to optimize the upper bound obtained in Theorem 2. An approximate solution to this problem is given by

$$\begin{aligned}\eta &= 1/a \cdot (\log(T)/T)^{2/3} \\ \gamma &= \lambda\sqrt{(e-2)/a} \cdot (\log(T)/T)^{1/3} \\ K &= \sqrt{3\lambda a/(e-2)} \cdot (T/\log(T))^{1/3}\end{aligned}$$

where

$$a = (9(e-2))^{1/3}\,(\sqrt{\lambda/3} + \lambda)^{2/3}.$$

This solution is obtained by taking the upper bound in Equation (23), approximating $(K+1)/K \approx 1$ and $(2K+1)/6 \approx K/3$, and solving the first order conditions with respect to the three tuning parameters. This approximation, and the tuning parameters specified above, then yield an approximate upper bound on regret of $6 \cdot \log(T)^{1/3} T^{2/3}$.

Unknown time horizon Note that the proposed tuning depends crucially on knowledge of the time horizon T at which regret is to be evaluated. In order to extend our rate results to the case of unknown time horizons, we can use the so-called doubling trick; cf. Section 2.3 of Cesa-Bianchi and Lugosi (2006): Consider a sequence of epochs (intervals of time-periods) of exponentially increasing length, and re-run Algorithm 1 for each time-period separately, tuning the parameters over the current epoch length. This construction converts Algorithm 1 into an "anytime algorithm" which enjoys the same regret guarantees of Theorem 2, up to a multiplicative constant factor. Another more efficient strategy to achieve the same goal is to modify Algorithm 1, allowing the parameters η and γ to change at each iteration, and splitting each bin associated with the discretization parameter K whenever more precision is required.

4 Stochastic regret bounds for concave social welfare

Theorem 1 in Section 3 provides a lower bound proportional to $T^{2/3}$, for adversarial and stochastic regret for social welfare maximization. The proof of this lower bound constructs a distribution for the v_i. This distribution is such that expected social welfare $\boldsymbol{U}(x)$ is non-concave, as a function of x; two global optima are separated by a region of lower welfare. In order to learn which of two candidates for the globally optimal policy is actually optimal, it is necessary to sample policies in between. These intermediate policies yield lower welfare, and sampling them contributes to cumulative regret. This construction is illustrated in Figure 1.

Given that the construction relies on non-concavity of expected social welfare, could we achieve lower regret if we knew that social welfare is actually concave? The answer turns out to be yes, for the stochastic setting. In the adversarial setting, cumulative welfare is necessarily non-concave.

For the stochastic setting with concave social welfare, we present an algorithm that achieves a bound on regret of order $T^{1/2}$, up to logarithmic terms. Before describing our proposed algorithm, Dyadic Search for Social Welfare, let us formally state the improved regret bounds. The proofs of these lower and upper bounds can again be found in Appendix B.

Theorem 3 (Lower bound on regret for the concave case)**.** *Consider the setup of Section 2. There exists a constant $C > 0$ such that, for any randomized algorithm for the choice of $x_1, x_2, \ldots$ and any time horizon $T \in \mathbb{N}$, the following holds:*

There exists a distribution μ on $[0,1]$ with associated demand function $\boldsymbol{G}$, where $\boldsymbol{G}$ is concave, for which the stochastic cumulative expected regret $\mathcal{R}_T(\boldsymbol{G})$ is at least $C \cdot T^{1/2}$.

Theorem 4 (Stochastic upper bound on regret of Dyadic Search for Social Welfare)**.** *Consider the stochastic setup of Section 2, and Algorithm 2, with confidence parameter $\delta = \frac{1}{T^{5/2}}$. Suppose that μ is such that $\boldsymbol{G}$ is concave. Then, for any time horizon $T \in \mathbb{N}$, expected regret $\mathcal{R}_T(\boldsymbol{G})$ is of order at most $T^{1/2}$, up to logarithmic terms.*

Dyadic search Our algorithm is based on a modification of dyadic search, as discussed in (Bachoc et al., 2022a,b). At any point in time, this algorithm maintains an active interval I_τ, which contains the optimal policy with high probability. Only policies within this interval are sampled going forward. As evidence accumulates, this interval is trimmed down, by excluding policies that are sub-optimal with high probability.

The algorithm proceeds in epochs τ. At the start of each epoch, a sub-interval $[l, r] \subset I_\tau$ is formed, with mid-point $c = (l + r)/2$. The points l, c, r are in a dyadic grid, that is, they are of the form $k/2^m$. After sampling from $[l, r]$, we calculate confidence intervals $J_t(l, c)$, $J_t(c, r)$, and $J_t(l, r)$ for the welfare differences $\Delta(l, c)$, $\Delta(c, r)$, and $\Delta(l, r)$, where $\Delta(x, x') = \boldsymbol{U}(x') - \boldsymbol{U}(x)$.

If the confidence interval $J_t(l, c)$ or $J_t(l, r)$ lies above 0, concavity implies that the optimal policy cannot lie to the left of l; we can thus trim the active interval I_τ by dropping all points

Algorithm 2 Dyadic Search for Social Welfare

```
Require: A confidence parameter $\delta \in (0,1)$.
$I_1 = [0,1]$, $t_0 = 0$, $k = 0$
for epochs $\tau = 1, 2, \ldots$ do
    Let $c = (\sup I_\tau + \inf I_\tau)/2$, and $d = \sup I_\tau - \inf I_\tau$.        {Subinterval for sampling}
    if $\tau$ is odd then
        Let $l = c - \frac{1}{4}d$, $r = c + \frac{1}{4}d$.
    else
        Let $l = c - \frac{1}{6}d$, $r = c + \frac{1}{6}d$.
    end if
    for $t = t_{\tau-1} + 1, t_{\tau-1} + 2, \ldots$ do
        Select $w \in \operatorname{argmax}_{w' \in \{l,c,r,(l,c),(c,r)\}} \Gamma_{t-1}(w')$,        {Sampling}
            breaking ties following the order $l, c, r, (l,c), (c,r)$
        if $w \in \{l, c, r\}$ then
            Set $x_t = w$.
        else
            Set $x_t = w_1 + (w_2 - w_1) \cdot \frac{k+1/2}{n_{t-1}(w_1,w_2)+1}$, and $k = (k+1) \mod n_{t-1}(w_1,w_2) + 1$.
        end if
        Calculate $J_t(l,c)$, $J_t(c,r)$, and $J_t(l,r)$, as in Equations (15) and (16).        {Inference}
        if $\inf\big(J_t(l,c)\big) \geq 0$ or $\inf\big(J_t(l,r)\big) \geq 0$ then
            let $I_{\tau+1} = I_\tau \cap [l,1]$ and $t_\tau = t$ and break        {Shrinking the active interval}
        else if $\sup\big(J_t(c,r)\big) \leq 0$ or $\sup\big(J_t(l,r)\big) \leq 0$ then
            let $I_{\tau+1} = I_\tau \cap [0,r]$ and $t_\tau = t$ and break
        end if
    end for
end for
```

to the left of l. Symmetrically, if the confidence interval $J_t(c,r)$ or $J_t(l,r)$ lies below 0, we can trim I_τ by dropping all points to the right of r.

Confidence intervals for welfare differences This procedure requires the construction of confidence intervals for welfare differences of the form

$$\Delta(x,x') = \boldsymbol{U}(x') - \boldsymbol{U}(x) = x' \cdot \boldsymbol{G}(x') - x \cdot \boldsymbol{G}(x) - \lambda \int_x^{x'} \boldsymbol{G}(x'')\mathrm{d}x''. \tag{12}$$

At time t, we estimate demand $\boldsymbol{G}(x)$, for policies x chosen in previous periods, as[2]

$$\widehat{\boldsymbol{G}}_t(x) = \frac{1}{n_t(x)} \sum_{i \leq t} y_i \cdot \mathbf{1}(x_i = x), \qquad\qquad n_t(x) = \sum_{i \leq t} \mathbf{1}(x_i = x).$$

We similarly estimate integrated demand $\int_x^{x'} \boldsymbol{G}(x'')\mathrm{d}x''$ by $(x' - x)$ times the average of realized demand y_i for observations x_i in the open interval (x,x'). We have to be careful, however, to

[2] We use the convention $0/0 = 0$ and $a/0 = +\infty$ whenever $a > 0$. Furthermore, every summation over an empty set of indices is understood to have value 0.

use a sample of x_i that is (approximately) uniformly distributed over this interval. This can be achieved for our dyadic search procedure, as specified in Algorithm 2, by truncating the time index used to estimate this average.[3] Let

$$s(x, x', t) = \max \left\{ s \le t : \ \log_2 \left(1 + \sum_{i \le s} \mathbf{1}(x_i \in (x, x')) \right) \in \mathbb{N} \right\}.$$

We define

$$\widehat{\boldsymbol{G}}_t(x, x') = \frac{1}{n_t(x, x') + 1} \sum_{i \le s(x,x',t)} y_i \cdot \mathbf{1}(x_i \in (x, x')), \quad n_t(x, x') = \sum_{i \le s(x,x',t)} \mathbf{1}(x_i \in (x, x')).$$

At each round, Algorithm 2 maintains estimates for welfare differences among three points l, c, r (for left, center and right, respectively). The estimate of the welfare difference between $x' = c$ and $x = l$ (or between $x' = r$ and $x = c$) is given by

$$\widehat{\Delta}_t(x, x') = x' \cdot \widehat{\boldsymbol{G}}_t(x') - x \cdot \widehat{\boldsymbol{G}}_t(x) - \lambda \cdot (x' - x) \cdot \widehat{\boldsymbol{G}}_t(x, x'). \tag{13}$$

while the estimate of the welfare difference between r and l is given by

$$\widehat{\Delta}_t(l, r) = \widehat{\Delta}_t(r, c) + \widehat{\Delta}_t(c, l). \tag{14}$$

To construct confidence intervals for $\Delta(x, x')$, we also need to quantify the uncertainty of our demand estimates. We use the following interval half-lengths for confidence intervals for tax revenue at x, and for the private welfare difference between x' and x:

$$\Gamma_t(x) = x \cdot \sqrt{\tfrac{1}{2n_t(x)} \log\left(\tfrac{2}{\delta}\right)}, \quad \Gamma_t(x, x') = \lambda \cdot (x' - x) \cdot \left(\sqrt{\tfrac{1}{2\left(n_t(x,x')+1\right)} \log\left(\tfrac{2}{\delta}\right)} + \tfrac{2}{n_t(x,x')+1} \right).$$

Using the shorthand $a \pm b = [a - b, a + b]$, our confidence interval for $\Delta(x, x')$, where $x' = c$ and $x = l$ (or $x' = r$ and $x = c$) is given by

$$J_t(x, x') = \widehat{\Delta}_t(x, x') \pm \left(\Gamma_t(x') + \Gamma_t(x) + \Gamma_t(x, x')\right), \tag{15}$$

while our confidence interval for $\Delta(l, r)$ is given by

$$J_t(r, l) = \widehat{\Delta}_t(r, l) \pm \left(\Gamma_t(r) + \Gamma_t(l) + \Gamma_t(l, c) + \Gamma_t(c, r)\right), \tag{16}$$

With these preliminaries, we are now ready to state our algorithm, Dyadic Search for Social Welfare, in Algorithm 2.

[3] The sampling procedure in Algorithm 2 samples sequentially from the dyadic grid in the active interval, refining the grid in subsequent iterations. $s(x, x', t)$ provides a truncation of the time index such that one round of such dyadic sampling has been completed.

5 Income taxation

We discuss two extensions of the baseline model of optimal taxation that we introduced in Section 2. These extensions incorporate features that are important in more realistic models of optimal taxation. For both of these extensions, we propose a properly modified version of Algorithm 1. The first extension, discussed in this section, is a variant of the Mirrlees model of optimal income taxation (Mirrlees, 1971; Saez, 2001). This extension allows for a taste for redistribution between different taxpayers, based on their earnings capacity. The second extension, discussed in Appendix A is a variant of the Ramsey model of commodity taxation (Ramsey, 1927). This extension allows for multidimensional actions (consumption choices), with separate taxes for different commodities.

Our model of income taxation generalizes our baseline model by allowing for heterogeneous wages w_i, welfare weights $\omega(w_i)$, extensive-margin labor supply responses determined by the cost of participation v_i, and non-linear income taxes $x_i = x(w_i)$. Two simplifications are maintained in this model, relative to a more general model of income taxation. First, only extensive margin responses (participation decisions) by individuals are allowed; there are no intensive margin responses (hours adjustments). Second, as in the baseline model of Section 2, there are no income effects. In imposing these assumptions, our model mirrors the model of optimal income taxation discussed in Section II.2 of Saez (2002).

Setup At each time $i = 1, 2, \ldots, T$, one individual arrives who is characterized by (i) a potential wage $w_i \in [0, 1]$, and (ii) an unknown cost of participation $v_i \in [0, 1]$. This individual makes a binary labor supply decision y_i. If they participate in the labor market ($y_i = 1$), they earn w_i, but pay a tax according to the tax rate $x_i = x(w_i)$ on their earnings w_i. They furthermore incur a non-monetary cost of participation v_i.

Their optimal labor supply decision is therefore given by $y_i = \mathbf{1}(v_i \leq w_i \cdot (1 - x_i))$, and private welfare equals $\max(w_i \cdot (1 - x_i) - v_i, 0)$. The implied public revenue is equal to the tax on earnings $x_i \cdot w_i$ if $y_i = 1$, and 0 otherwise.

We define social welfare as a weighted sum of public revenue and private welfare, with a weight $\omega(w_i)$ for the latter. Typically, ω is a decreasing function of w, reflecting a preference for redistribution towards those with lower earnings potential, cf. Saez and Stantcheva (2016). Social welfare for time period i, as a function of the tax schedule $x(\cdot)$, is therefore given by

$$U_i(x(\cdot)) = \underbrace{x(w_i) \cdot w_i \cdot \mathbf{1}(v_i \leq w_i \cdot (1 - x(w_i)))}_{\text{Public revenue}} \quad + \quad \omega(w_i) \cdot \underbrace{\max(w_i \cdot (1 - x(w_i)) - v_i, 0)}_{\text{Private welfare}}. \tag{17}$$

After period i, we observe y_i and the tax schedule $x_i(\cdot)$. If $y_i = 1$, we also observe w_i. Nothing else is observed.[4]

[4] It should be noted that in this model we take the transfer x_0 for individuals without other income as given. The effective tax bill of an employed individual equals $x(w_i) \cdot w_i - x_0$. The "unconditional basic income" x_0 does not affect labor supply, given our assumption that there are no income effects, and it enters social welfare

Piecewise constant tax schedules We next construct a generalization of Algorithm 1 based on piecewise constant tax schedules, with tax rates changing at the grid-points $\mathcal{W} \subset [0,1]$. Formally, define $\tilde{w}(w) = \max\{w' \in \mathcal{W} : \; w' \leq w\}$, rounding the wage w down to the nearest grid-point in $\mathcal{W}$, and $\tilde{w}_i = \tilde{w}(w_i)$. Denote $H = |\mathcal{W}|$, and let

$$\mathcal{X}_{\mathcal{W}} = \{x(\cdot) : \; x(w) = x(\tilde{w}(w)) \; \forall w\}.$$

For $w \in \mathcal{W}$ and any $x \in [0,1]$, denote

$$G_i(w,x) = \mathbf{1}(v_i \leq w_i \cdot (1-x)) \cdot \mathbf{1}(\tilde{w}_i = w),$$

so that $y_i = G_i(w_i, x_i(w_i))$. $G_i(w,x)$ is the individual labor supply function, interacted with an indicator for whether their wage w_i falls into the tax bracket starting at w. With this notation, and still assuming piecewise constant tax rates $x(\cdot)$, we can rewrite

$$\max(w_i \cdot (1-x) - v_i, 0) = w_i \cdot \int_x^1 G_i(\tilde{w}_i, x')dx',$$

and

$$U_i(x(\cdot)) = \sum_{w \in \mathcal{W}} \left[x(w) \cdot G_i(w, x(w)) + \omega(w_i) \cdot w_i \cdot \int_{x(w)}^1 G_i(w,x')dx' \right]. \tag{18}$$

Cumulative social welfare is then given by $\mathbb{U}_i = \sum_{j \leq i} U_i(x_i(\cdot))$, and we correspondingly define cumulative expected regret, in the adversarial setting, as

$$\mathcal{R}_T(\{v_i\}_{i=1}^T) = \sup_{x(\cdot) \in \mathcal{X}_{\mathcal{W}}} E\left[\mathbb{U}_T(x(\cdot)) - \mathbb{U}_T \middle| \{v_i\}_{i=1}^T \right].$$

The supremum here is taken over all tax schedules $x(\cdot)$ that are piecewise constant between the gridpoints $w \in \mathcal{W}$.

Algorithm Algorithm 3 generalizes Algorithm 1 to this setting. As before, we form an unbiased estimate $\widehat{G}_i$ of G_i using inverse probability weighting, map this estimate into a corresponding estimate $\widehat{U}_i$ of U_i, based on Equation (18), and cumulate across time periods to obtain $\widehat{\mathbb{U}}_i$. Note that w_i is observed whenever $y_i = 1$. This implies that the estimate $\widehat{G}_i$ is in fact a function of observables, and the same holds for $\widehat{U}_i$.

Algorithm 3 keeps track of estimated demand and social welfare for each bin ("tax bracket"), as defined by the gridpoints $w \in \mathcal{W}$. The algorithm then constructs a distribution $p_i(x|w)$ over tax rates $x \in \mathcal{X}$ given w, using the tempered Exp3 distribution. The tax schedule $x(\cdot)$ is sampled according to these (marginal) distributions of tax rates for each bracket. Though

additively. It is therefore without loss of generality to omit x_0 from our model.

Algorithm 3 Tempered Exp3 for optimal income taxation

```
Require: Tuning parameters $K$, $\gamma$ and $\eta$, and set of gridpoints $\mathcal{W} \subset [0,1]$.
Calculate evenly spaced grid-points $\mathcal{X} = \{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\}$.
All sums in the following are taken over elements $x'$ of $\mathcal{X}$.
Initialize $\widehat{\mathbb{G}}_1(w,x) = 0$ and $\widehat{\mathbb{U}}_1(w,x) = 0$ for all $w \in \mathcal{X}$ and all $x \in \mathcal{X}$.
for individual $i = 1, 2, \dots, T$ do
    For all $x, w \in \mathcal{X}$ with $x \leq w$, set $\tilde{w} = \max\{w' \in \mathcal{W} : \; w' \leq w\}$, and
                                                        {Assignment probabilities}
        $$p_i(x|w) = (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_i(x,\tilde{w}))}{\sum_{x'=0}^{1} \exp(\eta \cdot \widehat{\mathbb{U}}_i(x',\tilde{w}))} + \frac{\gamma}{K+1}. \tag{19}$$
    Draw $A_i \sim U[0,1]$. For all $w \in [0,1]$, set
        $$x_i(w) = \max\left\{ x \in \mathcal{X} : \sum_{x'=0}^{x-1/k} p_i(x'|w) \leq A_i \right\}, \tag{20}$$
    and query $y_i$ accordingly.
    For all $w \in \mathcal{W}$ and $x \in \mathcal{X}$, set               {Estimated labor supply}
        $$\widehat{G}_i(x,w) = y_i \cdot \frac{\mathbf{1}(\tilde{w}_i = w, x_i(w_i) = x)}{p_i(x|w)}. \tag{21}$$
    For all $w \in \mathcal{W}$ and $x \in \mathcal{X}$, set               {Estimated welfare}
        $$\widehat{\mathbb{U}}_{i+1}(x,w) = \widehat{\mathbb{U}}_i(x,w) + x \cdot w_i \cdot \widehat{G}_i(x,w) + \frac{\omega(w_i) \cdot w_i}{K} \cdot \sum_{x'=x}^{1} \widehat{G}_i(x',w). \tag{22}$$
end for
```

immaterial for the following theorem, we choose the perfectly correlated coupling of these marginal distributions, which is implemented using the random variable A_i in Algorithm 3.

Theorem 5 (Adversarial upper bound on regret of Tempered Exp3 for optimal income taxation)**.** *Consider the setup of Section 5, and Algorithm 3. Assume that* $(K+1)\eta < \gamma$. *Then for any sequence* $(v_1, \dots, v_T)$ *expected regret* $\mathcal{R}_T(\{v_i\}_{i=1}^T)$ *is bounded above by*

$$\left(\gamma + \eta \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2K+1}{6} + \tfrac{1}{\gamma}\right) + \tfrac{1}{K}\right) \cdot T + \tfrac{H \log(K+1)}{\eta}. \tag{23}$$

Suppose additionally that $K = c_1 \cdot (T/H)^{1/3}$, $\gamma = c_2/(K+1)$, *and* $\eta = c_3/(K+1)^2$, *for some constants* c_1, c_2, c_3. *Then expected regret* $\mathcal{R}_T(\{v_i\}_{i=1}^T)$ *is bounded above by*

$$c_4 \cdot H^{1/3} \cdot \log(T)^{1/3} T^{2/3}, \tag{24}$$

for some constant c_4.

References

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

Bachoc, F., Cesari, T., Colomboni, R., and Paudice, A. (2022a). A near-optimal algorithm for univariate zeroth-order budget convex optimization.

Bachoc, F., Cesari, T., Colomboni, R., and Paudice, A. (2022b). Regret analysis of dyadic search. *arXiv preprint arXiv:2209.00885*.

Baily, M. (1978). Some aspects of optimal unemployment insurance. *Journal of Public Economics*, 10(3):379–402.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

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, N., Gaillard, P., Gentile, C., and Gerchinovitz, S. (2017). Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, pages 465–481. PMLR.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2015). Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 1(61):549–564.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Chetty, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics*, 1(1):451–488.

Daskalakis, C. and Syrgkanis, V. (2022). Learning in auctions: Regret is hard, envy is easy. *Games and Economic Behavior*.

den Boer, A. V. (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*.

Feng, Z., Guruganesh, G., Liaw, C., Mehta, A., and Sethi, A. (2021). Convergence analysis of no-regret bidding algorithms in repeated auctions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5399–5406.

Han, Y., Zhou, Z., Flores, A., Ordentlich, E., and Weissman, T. (2020a). Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568*.

Han, Y., Zhou, Z., and Weissman, T. (2020b). Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795*.

Kasy, M. (2018). Optimal taxation and insurance using machine learning – sufficient statistics and beyond. *Journal of Public Economics*, 167.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132.

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

Kolumbus, Y. and Nisan, N. (2022). Auctions between regret-minimizing agents. In *Proceedings of the ACM Web Conference 2022*, pages 100–111.

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

Lugosi, G., Markakis, M., and Neu, G. (2022). On the hardness of learning from censored demand. *Available at SSRN 3509255*.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601.

Mirrlees, J. (1971). An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, pages 175–208.

Ramsey, F. P. (1927). A contribution to the theory of taxation. *The economic journal*, 37(145):47–61.

Russo, D. (2020). Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647.

Saez, E. (2001). Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229.

Saez, E. (2002). Optimal income transfer programs: intensive versus extensive labor supply responses. *The Quarterly Journal of Economics*, 117(3):1039–1073.

Saez, E. and Stantcheva, S. (2016). Generalized social welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45.

Slivkins, A. (2019). Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*.

Thomas M. Cover, J. A. T. (2006). *Elements of Information Theory*. Wiley-Interscience.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated.

Weed, J., Perchet, V., and Rigollet, P. (2016). Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR.

Williams, D. (1991). *Probability with martingales.* Cambridge University Press.

A Commodity taxation

In this appendix, we generalize our baseline model of optimal taxation to a model of commodity taxation with multiple goods $j \in \{1, \ldots, k\}$ and continuous demand functions $y_i(x) \in [0,1]^k$, where $x \in [0,1]^k$ is a vector of tax rates. We again assume that there are no income effects. Our setup is a version of the classic Ramsey model (Ramsey, 1927). We propose a generalization of Tempered Exp3 for Social Welfare to this setting. We leave a regret analysis of this generalization for future research. In the following, we use $\langle x, y \rangle$ to denote the inner product between x and y.

Setup At each time $i = 1, 2, \ldots, T$, one individual arrives who is characterized by a utility function $u_i : [0,1]^k \to \mathcal{R}$. This individual is exposed to a tax vector $x_i \in [0,1]^k$, and makes a continuous consumption decision y_i. Public revenue is given by $\langle x_i, y \rangle$. Agent utility is given by $u_i(y_i)$ plus their consumption of a numeraire good, which has price normalized to 1 and enters utility additively. The individual consumption choice y_i costs $\langle x_i + p, y \rangle$, where p is the (exogenously given) vector of pre-tax prices. This cost reduces consumption of the numeraire good. The optimal individual decision is therefore given by

$$y_i = G_i(x_i) = \underset{y \in [0,1]^k}{\operatorname{argmax}} \left[u_i(y) - \langle x_i + p, y \rangle \right]. \tag{25}$$

Defining v_0 as individual utility when $y = 0$, the implied private welfare is

$$v_i(x) = v_0 + \max_{y \in [0,1]^k} \left[u_i(y) - \langle x_i + p, y \rangle \right],$$

We choose the constant v_0 such that $v_i(0) = 0$; this is just a normalization to simplify notation.

We define social welfare as a weighted sum of public revenue and private welfare, with a weight λ for the latter. Social welfare for time period i, as a function of the tax vector x, is therefore given by

$$U_i(x_i) = \underbrace{\langle x_i, y_i \rangle}_{\text{Public revenue}} + \lambda \cdot \underbrace{v_i(x_i)}_{\text{Private welfare}}. \tag{26}$$

After period i, we observe y_i and the tax vector x_i. Nothing else is observed.

Mapping demand to welfare By the envelope theorem (Milgrom and Segal, 2002),

$$\nabla_x v_i(x) = y_i = G_i(x).$$

Let $\mathcal{V}$ be the set of differentiable functions v on $[0,1]^k$ such that $\nabla_x v \in L^2$, and such that $v(0) = 0$. Consider the following operator, mapping the demand function G into the corresponding

Algorithm 4 Tempered Exp3 for commodity taxation

```
Require: Tuning parameters $K$, $\gamma$ and $\eta$.
Calculate the set of evenly spaced grid-points $\mathcal{X} = [0, \frac{1}{K}, \ldots, 1]^k$
  and initialize $\widehat{\mathbb{G}}_1(x) = 0$ for all grid points.
for individual $i = 1, 2, \ldots, T$ do
    For all $x \in \mathcal{X}$, set                                {Estimated welfare}
        $\widehat{\mathbb{U}}_i(x) = \langle x_i, \widehat{\mathbb{G}}_i \rangle + \lambda \cdot \widehat{v}_i(x_i).$      (28)
    For all $x \in \mathcal{X}$, set                                {Assignment probabilities}
        $p_i = (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_i(x))}{\sum_{x'} \exp(\eta \cdot \widehat{\mathbb{U}}_i(x'))} + \frac{\gamma}{(K+1)^k}.$      (29)
    Choose $x_i$ at random according to the probability distribution $p_i$, and query $y_i$ accordingly.
    For all $x \in \mathcal{X}$, set                                {Estimated demand}
        $\widetilde{\mathbb{G}}_{i+1}(x) = \widehat{\mathbb{G}}_i(x) + \frac{y_i}{p_i}$      (30)
        $\widehat{v}_{i+1}(x) = \Pi(\widetilde{\mathbb{G}}_{i+1})$      (31)
        $\widehat{\mathbb{G}}_{i+1} = \nabla_x \widehat{v}_{i+1}.$      (32)
end for
```

indirect utility function v.

$$\Pi(G(\cdot)) = \underset{v(\cdot)\in\mathcal{V}}{\operatorname{argmin}} \int_{[0,1]^k} \|\nabla_x v(x) - G(x)\|^2 \, dx \tag{27}$$

We can think of the operator Π as combining two operators. First, the function G is projected on the subspace of functions on $[0,1]^k$ which can be written as the gradient of some function v. Second, the projected G is then integrated to get $v(x)$ for any x. Integration here is along some curve in $[0,1]^k$ from 0 to x. Given the first projection, the choice of curve does not matter for the resulting function v.

B Proofs

B.1 Theorem 1 (Lower bound on regret)

Proof of Theorem 1.

Defining a family of distributions for v Recall that, for each $\epsilon \in [-1,1]$, the probability distribution μ^ϵ is defined as the probability measure supported on $(1/4, 1/2, 3/4, 1)$ with masses $\big(a, (1+\epsilon)\cdot b, (1-\epsilon)\cdot b, 1-a-2\cdot b\big)$, where

$$a := \frac{(1-\lambda)\cdot(136-99\cdot\lambda)}{2\cdot(4-3\cdot\lambda)\cdot(24-17\cdot\lambda)} \,, \qquad b := \frac{1-\lambda}{2\cdot(24-17\cdot\lambda)} \,.$$

Furthermore, for each $\epsilon \in [-1,1]$, recall that $\boldsymbol{G}^\epsilon$ and $\boldsymbol{U}^\epsilon$ are respectively the demand function and the expected social welfare associated to μ^ϵ. Figure 1 illustrates. Let $v_1, v_2, \dots \in [0,1]$ be the sequence of individual valuations. For each $\epsilon \in [-1,1]$, consider a distribution P^ϵ such that the individual valuations $v_1, v_2, \dots$ form a P^ϵ-i.i.d. sequence (independent of the randomization used by the algorithm) with common distribution μ^ϵ.

Explicit lower bound on regret that will be proven Define

$$c_1 := \frac{\lambda}{4}\cdot b \,, \quad c_2 := \frac{1}{8}\cdot\frac{1-\lambda}{4-3\cdot\lambda} \,, \quad c_3 := b\cdot\sqrt{\frac{2}{a\cdot(1-a-2\cdot b)}} \,.$$

We will prove that, for any randomized algorithm and any time horizon $T \in \mathbb{N}$, there exists $\epsilon \in [-1,1]$ such that

$$\mathcal{R}_T(\boldsymbol{G}^\epsilon) \ge C \cdot T^{2/3} \,,$$

where

$$C := \min\left(\frac{c_1^2\cdot c_3^2}{c_2}, \frac{c_2}{2}, \frac{1}{16}\cdot\sqrt[3]{\frac{c_1^2\cdot c_2}{c_3^2}}\right) = \min\Bigg(\frac{\lambda^2\cdot(4-3\cdot\lambda)^3}{8\cdot(136-99\cdot\lambda)\cdot(26-19\cdot\lambda)}, \\ \frac{\lambda^{2/3}\cdot(1-\lambda)^{4/3}\cdot(136-99\cdot\lambda)^{1/3}\cdot(26-19\cdot\lambda)^{1/3}}{128\cdot(4-3\cdot\lambda)\cdot(24-17\cdot\lambda)^{4/3}}\Bigg) > 0 \tag{33}$$

Fix a randomized algorithm to choose the policies $x_1, x_2, \dots$, and fix a time horizon $T \in \mathbb{N}$.

Number of mistakes and lower bound on regret We need to count the random number of times the algorithm has played in the regions $(1/2, 3/4]$, $[0, 1/2]$ and $(3/4, 1]$ up to time T. This can be done relying on the following random variables:

$$n_1 := \sum_{i=1}^T \mathbf{1}_{(1/2,3/4]}(x_i) \,, \qquad n_2 := \sum_{i=1}^T \mathbf{1}_{[0,1/2]}(x_i) \,, \qquad n_3 := \sum_{i=1}^T \mathbf{1}_{(3/4,1]}(x_i) \,.$$

Notice that since the intervals $(1/2, 3/4], [0, 1/2]$ and $(3/4, 1]$ form a partition of $[0,1]$, we have that

$$n_1 + n_2 + n_3 = T \tag{34}$$

For each $\epsilon \in [-1,1]$, denote by E^ϵ the expectation taken with respect to the distribution P^ϵ. Notice that, for each $\epsilon \in [-1,1]$, the expected regret when the underlying distribution is P^ϵ equals

$$\mathcal{R}_T(\boldsymbol{G}^\epsilon) = T \cdot \sup_{x\in[0,1]} \boldsymbol{U}^\epsilon(x) - \sum_{i=1}^T E^\epsilon\big(\boldsymbol{U}^\epsilon(x_i)\big) \,. \tag{35}$$

Algebraic calculations show that, for each $\epsilon \in [-1,1]$

$$\max_{x \in (\nicefrac{1}{2},\nicefrac{3}{4}]} \boldsymbol{U}^{\epsilon}(x) = \boldsymbol{U}^{\epsilon}(\nicefrac{3}{4}) \ , \quad \max_{x \in [0,\nicefrac{1}{2}]} \boldsymbol{U}^{\epsilon}(x) = \boldsymbol{U}^{\epsilon}(\nicefrac{1}{4}) \ , \quad \max_{x \in (\nicefrac{3}{4},1]} \boldsymbol{U}^{\epsilon}(x) = \boldsymbol{U}^{\epsilon}(1) \ , \tag{36}$$

$$\text{and} \qquad \boldsymbol{U}^{\epsilon}(1) - \boldsymbol{U}^{\epsilon}(\nicefrac{1}{4}) = c_1 \cdot \epsilon \ . \tag{37}$$

Further calculations show also that

$$\min_{\epsilon \in [-1,1]} \min\big(\boldsymbol{U}^{\epsilon}(\nicefrac{1}{4}), \boldsymbol{U}^{\epsilon}(1)\big) = \boldsymbol{U}^{1}(\nicefrac{1}{4}) \ , \quad \max_{\epsilon \in [-1,1]} \max_{x \in (\nicefrac{1}{2},\nicefrac{3}{4}]} \boldsymbol{U}^{\epsilon}(x) = \boldsymbol{U}^{-1}(\nicefrac{3}{4}) \ , \tag{38}$$

$$\text{and} \qquad \boldsymbol{U}^{1}(\nicefrac{1}{4}) - \boldsymbol{U}^{-1}(3/4) = c_2 \ . \tag{39}$$

Equations (36), (37), (38), and (39) imply that

$$\sup_{x \in [0,1]} \boldsymbol{U}^{\epsilon}(x) = \boldsymbol{U}^{\epsilon}(1) \ , \qquad \text{if } \epsilon \in [0,1] \ . \tag{40}$$

It follows that, if $\epsilon \in [0,1]$,

$$\begin{aligned}
\mathcal{R}_T(\boldsymbol{G}^{\epsilon}) &\overset{(35)}{=} T \cdot \sup_{x \in [0,1]} \boldsymbol{U}^{\epsilon}(x) - \sum_{i=1}^{T} E^{\epsilon}\big(\boldsymbol{U}^{\epsilon}(x_i)\big) \overset{(40)}{=} T \cdot \boldsymbol{U}^{\epsilon}(1) \\
&\quad - \sum_{i=1}^{T} E^{\epsilon}\Big(\boldsymbol{U}^{\epsilon}(x_i) \cdot \big(\mathbf{1}_{(\nicefrac{1}{2},\nicefrac{3}{4}]}(x_i) + \mathbf{1}_{[0,\nicefrac{1}{2}]}(x_i) + \mathbf{1}_{(\nicefrac{3}{4},1]}(x_i)\big)\Big) \\
&\overset{(36)}{\geq} T \cdot \boldsymbol{U}^{\epsilon}(1) - \sum_{i=1}^{T} E^{\epsilon}\Big(\boldsymbol{U}^{\epsilon}(\nicefrac{3}{4}) \cdot \mathbf{1}_{(\nicefrac{1}{2},\nicefrac{3}{4}]}(x_i) \\
&\quad + \boldsymbol{U}^{\epsilon}(\nicefrac{1}{2}) \cdot \mathbf{1}_{[0,\nicefrac{1}{2}]}(x_i) + \boldsymbol{U}^{\epsilon}(1) \cdot \mathbf{1}_{(\nicefrac{3}{4},1]}(x_i)\Big) \\
&\overset{(34)}{=} \big(\boldsymbol{U}^{\epsilon}(1) - \boldsymbol{U}^{\epsilon}(\nicefrac{3}{4})\big) \cdot E^{\epsilon}(n_1) + \big(\boldsymbol{U}^{\epsilon}(1) - \boldsymbol{U}^{\epsilon}(\nicefrac{1}{4})\big) \cdot E^{\epsilon}(n_2) \\
&\overset{(38)}{\geq} \big(\boldsymbol{U}^{1}(\nicefrac{1}{4}) - \boldsymbol{U}^{-1}(\nicefrac{3}{4})\big) \cdot E^{\epsilon}(n_1) + \big(\boldsymbol{U}^{\epsilon}(1) - \boldsymbol{U}^{\epsilon}(\nicefrac{1}{4})\big) \cdot E^{\epsilon}(n_2) \\
&\overset{(39)}{=} c_2 \cdot E^{\epsilon}(n_1) + \big(\boldsymbol{U}^{\epsilon}(1) - \boldsymbol{U}^{\epsilon}(\nicefrac{1}{4})\big) \cdot E^{\epsilon}(n_2) \\
&\overset{(37)}{=} c_2 \cdot E^{\epsilon}(n_1) + c_1 \cdot \epsilon \cdot E^{\epsilon}(n_2)
\end{aligned} \tag{41}$$

Notice that inequality (41) quantifies how much regret the algorithm is going to suffer in terms of the expected number of times it plays in the wrong regions, when the demand function is $\boldsymbol{G}^{\epsilon}$ and $\epsilon > 0$.

In the same way inequality (41) was proven, we can prove that, if $\epsilon \in [0,1]$,

$$\mathcal{R}_T(\boldsymbol{G}^{-\epsilon}) \geq c_2 \cdot E^{-\epsilon}(n_1) + c_1 \cdot \epsilon \cdot E^{-\epsilon}(n_3) \geq c_1 \cdot \epsilon \cdot E^{-\epsilon}(n_3) \ , \tag{42}$$

which again quantifies how much regret the algorithm is going to suffer in terms of the expected number of times it plays in the wrong regions, when the demand function is $\boldsymbol{G}^{-\epsilon}$ and $\epsilon > 0$.

Intuition for the remainder of the proof At high level, inequalities (41) and (42) tell us that, if $|\epsilon|$ is not negligible, the algorithm has to play a substantially different number of times in the region $(\nicefrac{3}{4}, 1]$ depending on the sign of ϵ not to suffer significant regret when the demand function is $\boldsymbol{G}^{\epsilon}$. The crucial idea is that the only way for the algorithm to present this different behavior is by playing in the only informative region about the sign of ϵ, i.e., the region $(\nicefrac{1}{2}, \nicefrac{3}{4}]$. However, as shown in (41), selecting policies in this region comes at a cost in terms of regret. To relate quantitatively the number of times the algorithm has to play in this costly region with the difference in the expected number of times the algorithm selects policies in the region $(3/4, 1]$ is the last missing ingredient that we can obtain relying on information theoretic techniques: It can be proved (and a formal proof is provided at the end of the current proof) that, for each $\epsilon \in [0, 1]$,

$$E^{-\epsilon}(n_3) \geq E^{\epsilon}(n_3) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^{\epsilon}(n_1)}\ . \tag{43}$$

Now, if the algorithm is going to suffer low regret when $\epsilon > 0$, then by (41) we have an upper bound on the number of times the algorithm plays in the region $(1/2, 3/4]$ and a lower bound on the number of times it plays in the region $(3/4, 1]$, whenever $\epsilon > 0$. In turn, by (43), this gives a lower bound on the number of times the algorithm plays in the sub-optimal region $(3/4, 1]$ when $\epsilon < 0$. Then, relying on (42), we have an explicit lower bound on how much regret the algorithm is going to suffer when $\epsilon < 0$. We will now carry out this plan —and prove the theorem— as follows.

Low regret cannot be achieved for both positive and negative ϵ To get a contradiction, suppose that

$$\forall \epsilon \in [-1, 1] \qquad \mathcal{R}_T(\boldsymbol{G}^{\epsilon}) < C \cdot T^{2/3}\ . \tag{44}$$

It follows from (41) that, for each $\epsilon \in [0, 1]$,

$$E^{\epsilon}(n_1) \overset{(41)}{\leq} \frac{\mathcal{R}_T(\boldsymbol{G}^{\epsilon})}{c_2} \overset{(44)}{\leq} \frac{C}{c_2} \cdot T^{2/3}\ , \qquad E^{\epsilon}(n_2) \overset{(41)}{\leq} \frac{\mathcal{R}_T(\boldsymbol{G}^{\epsilon})}{c_1 \cdot \epsilon} \overset{(44)}{\leq} \frac{C}{c_1 \cdot \epsilon} \cdot T^{2/3}\ . \tag{45}$$

This implies, relying also on (42) and (43), that for each $\epsilon \in [0, 1]$ we have

$$\begin{aligned}
\mathcal{R}_T(\boldsymbol{G}^{-\epsilon}) &\overset{(42)}{\geq} c_1 \cdot \epsilon \cdot E^{-\epsilon}(n_3) \overset{(43)}{\geq} c_1 \cdot \epsilon \cdot \left(E^{\epsilon}(n_3) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^{\epsilon}(n_1)}\right) \\
&\overset{(34)}{=} c_1 \cdot \epsilon \cdot \left(T - E^{\epsilon}(n_1) - E^{\epsilon}(n_2) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^{\epsilon}(n_1)}\right) \\
&\overset{(45)}{\geq} c_1 \cdot \epsilon \cdot \left(T - \frac{C}{c_2} \cdot T^{2/3} - \frac{C}{c_1 \cdot \epsilon} \cdot T^{2/3} - c_3 \cdot \epsilon \cdot T \cdot \sqrt{\frac{C}{c_2} \cdot T^{2/3}}\right) \\
&= c_1 \cdot \epsilon \cdot \left(1 - \frac{C}{c_2} \cdot T^{-1/3} - \frac{C}{c_1 \cdot \epsilon} \cdot T^{-1/3} - c_3 \cdot \epsilon \cdot T^{1/3} \cdot \sqrt{\frac{C}{c_2}}\right) \cdot T\ .
\end{aligned} \tag{46}$$

Pick $\epsilon := T^{-1/3} \cdot \sqrt{\frac{\sqrt{C \cdot c_2}}{c_1 \cdot c_3}}$. First, note that since $0 < C \overset{(33)}{\leq} \frac{c_1^2 \cdot c_3^2}{c_2}$ we have that $\epsilon \in (0, 1]$. Plugging this value of ϵ in (46) leads to

$$\begin{aligned}
C \cdot T^{2/3} &\overset{(44)}{>} \mathcal{R}_T(\boldsymbol{G}^{-\epsilon}) \\
&\overset{(46)}{\geq} \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - \frac{C}{c_2} \cdot T^{-1/3} - 2 \cdot \sqrt{\frac{c_3}{c_1 \cdot \sqrt{c_2}}} \cdot C^{3/4}\right) \cdot T^{2/3} \\
&\overset{(33)}{\geq} \frac{1}{2} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - 4 \cdot \sqrt{\frac{c_3}{c_1 \cdot \sqrt{c_2}}} \cdot C^{3/4}\right) \cdot T^{2/3} \\
&\overset{(33)}{\geq} \frac{1}{4} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot T^{2/3}\ ,
\end{aligned} \tag{47}$$

where the second to last inequality follows from $C \leq \frac{c_2}{2}$, while the last inequality follows from $C \leq \frac{1}{16} \sqrt[3]{\frac{c_1^2 \cdot c_2}{c_3^2}}$. Rearranging inequality (47) leads to the contradiction

$$C \overset{(47)}{>} \left(\frac{1}{4} \cdot \sqrt{\frac{c_1 \cdot \sqrt{c_2}}{c_3}}\right)^{4/3} = \frac{1}{8} \cdot \sqrt[3]{\frac{2 \cdot c_1^2 \cdot c_2}{c_3^2}} > \frac{1}{16} \cdot \sqrt[3]{\frac{c_1^2 \cdot c_2}{c_3^2}} \overset{(33)}{\geq} C\ .$$

Since (44) leads to a contradiction, it follows that there exists $\epsilon \in [-1,1]$ such that $\mathcal{R}_T(\boldsymbol{G}^\epsilon) \geq C \cdot T^{2/3}$. Given that the time horizon T and the randomized algorithm were arbitrarily fixed, the theorem is proved. $\square$

B.2 Claim (43) (Relating choice probabilities for positive and negative ϵ)

Proof of the claim (43).
Let $w_1, w_2, \dots \in [0,1]$ be the randomization seeds to be used by the algorithm. In the light of the Skorokhod representation theorem (Williams, 1991, Section 17.3), we may assume without (much) loss of generality that, for each $\epsilon \in [-1,1]$, these seeds form a sequence of P^ϵ-i.i.d. $[0,1]$-valued uniform random variables. In particular, this implies,

$$P^\epsilon_{(w_i)_{i\in\mathbb{N}}} = P^{-\epsilon}_{(w_i)_{i\in\mathbb{N}}} \;, \qquad \forall \epsilon \in [0,1] \;. \tag{48}$$

Recall that a sequence of functions $\alpha \coloneqq (\alpha_i)_{i\in\mathbb{N}}$ is called a randomized algorithm if

$$\alpha_1 \colon [0,1] \to [0,1] \;, \qquad \forall i \in \mathbb{N}, \quad \alpha_{i+1} \colon [0,1]^{i+1} \times \{0,1\}^i \to [0,1] \;.$$

The feedback function associated to our problem is

$$\varphi \colon [0,1] \times \{1/4, 1/2, 3/4, 1\} \to \{0,1\} \;, \qquad (x,v) \mapsto \mathbf{1}(x \leq v) \;.$$

Now, a randomized algorithm α generates a sequence of choices $x_1, x_2, \dots$ using the randomization seeds $w_1, w_2, \dots$ and the received feedback $z_1, z_2, \dots \in \{0,1\}$ in the following inductive way on $i \in \mathbb{N}$

$$\begin{aligned} x_1 &\coloneqq \alpha_1(w_1) \;, & z_1 &\coloneqq \varphi(x_1, v_1) \;, \\ x_{i+1} &\coloneqq \alpha_{i+1}(w_1, \dots, w_{i+1}, z_1, \dots, z_i) \;, & z_{i+1} &\coloneqq \varphi(x_{i+1}, v_{i+1}) \;. \end{aligned}$$

For each $a \in [0,1]$, fix a binary representation $0.a_1a_2a_3\dots$ and define $\xi(a) \coloneqq 0.a_1a_3a_5\dots$ and $\zeta(a) \coloneqq 0.a_2a_4a_6\dots$. Notice that $\xi, \zeta \colon [0,1] \to [0,1]$ are independent with respect to the Lebesgue measure on $[0,1]$ and that their (common) distribution is a uniform on $[0,1]$. For each $x \in [0,1]$, define $\psi_x \colon [0,1] \to \{0,1\}, u \mapsto \mathbf{1}_{[0,1/4]}(x) + \mathbf{1}_{(1/4,1/2]}(x) \cdot \mathbf{1}_{[0,1-a]}(u) + \mathbf{1}_{(3/4,1]}(x) \cdot \mathbf{1}_{[0,1-a-2\cdot b]}(u)$. Define by induction on $i \in \mathbb{N}$ the following process

$$\begin{aligned} \tilde{x}_1 &\coloneqq \alpha_1\big(\zeta(w_1)\big) \;, \\ \tilde{z}_1 &\coloneqq \varphi\Big(\tilde{x}_1, \psi_{\tilde{x}_1}\big(\xi(w_1)\big)\Big) \;, \\ \tilde{x}_{i+1} &\coloneqq \alpha_{i+1}\big(\zeta(w_1), \dots, \zeta(w_{i+1}), \tilde{z}_1, \dots, \tilde{z}_i\big), \\ \tilde{z}_{i+1} &\coloneqq \begin{cases} \varphi(\tilde{x}_{i+1}, v_{i+1}), & \tilde{x}_{i+1} \in (1/2, 3/4] \\ \varphi\Big(\tilde{x}_{i+1}, \psi_{\tilde{x}_{i+1}}\big(\xi(w_{i+1})\big)\Big), & \text{otherwise.} \end{cases} \end{aligned}$$

Since, for each $\epsilon \in [-1,1]$ and each $i \in \mathbb{N}$,

$$P^\epsilon(z_i = 1 \mid x_i) = \begin{cases} 1 & x_i \in [0, \frac{1}{4}] \\ 1-a & x_i \in (\frac{1}{4}, \frac{1}{2}] \\ 1-a-(1+\epsilon)\cdot b & x_i \in (\frac{1}{2}, \frac{3}{4}] \\ 1-a-2\cdot b & x_i \in (\frac{3}{4}, 1] \end{cases},$$

$$P^\epsilon(\tilde{z}_i = 1 \mid \tilde{x}_i) = \begin{cases} 1 & \tilde{x}_i \in [0, \frac{1}{4}] \\ 1-a & \tilde{x}_i \in (\frac{1}{4}, \frac{1}{2}] \\ 1-a-(1+\epsilon)\cdot b & \tilde{x}_i \in (\frac{1}{2}, \frac{3}{4}] \\ 1-a-2\cdot b & \tilde{x}_i \in (\frac{3}{4}, 1] \end{cases}$$

it follows that, for each $\epsilon \in [-1,1]$ and each $i \in \mathbb{N}$, the random variable $\tilde{x}_i$ has the same distribution as the random choice x_i made by the randomized algorithm α at time i when the underlying distribution is P^ϵ, i.e.,

$$P^\epsilon_{\tilde{x}_i} = P^\epsilon_{x_i} \ . \tag{49}$$

As with the process $x_1, x_2, \dots$, we have to count the number of times the process $\tilde{x}_1, \tilde{x}_2, \dots$ lands in the regions $(1/2, 3/4]$, $[0, 1/2]$ and $(3/4, 1]$ up to the time T. This can be done relying on the following random variables

$$\tilde{n}_1 \coloneqq \sum_{i=1}^{T} \mathbf{1}_{(1/2,3/4]}(\tilde{x}_i) \ , \qquad \tilde{n}_2 \coloneqq \sum_{i=1}^{T} \mathbf{1}_{[0,1/2]}(\tilde{x}_i) \ , \qquad \tilde{n}_3 \coloneqq \sum_{i=1}^{T} \mathbf{1}_{(3/4,1]}(\tilde{x}_i) \ .$$

Since, for each $\epsilon \in [-1,1]$ and each $j \in \{1,2,3\}$,

$$E^\epsilon(\tilde{n}_j) = \sum_{i=1}^{T} P^\epsilon_{x_i}\big((1/2, 3/4]\big) \overset{(49)}{=} \sum_{i=1}^{T} P^\epsilon_{\tilde{x}_i}\big((1/2, 3/4]\big) = E^\epsilon(n_j) \ ,$$

to prove the claim (43), it is enough to prove that, for each $\epsilon \in [-1,1]$,

$$E^{-\epsilon}(\tilde{n}_3) \geq E^\epsilon(\tilde{n}_3) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^\epsilon(\tilde{n}_1)} \ .$$

We first prove the result when the sequence of randomization seeds is fixed, i.e., we suppose first that $\bar{w}_1, \bar{w}_2, \dots$ are such that $w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots$. For each $\epsilon \in [-1,1]$, we consider the associated probability distribution Q^ϵ, defined as the conditional probability distribution $P^\epsilon(\cdot \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots)$. For each $t \in \mathbb{N}$, let $I_t \coloneqq \big\{ i \in \{1, \dots, t\} \mid \tilde{x}_i \in (1/2, 3/4] \big\}$, and for each $s \in \{1, \dots, t\}$, let

$$Z_{t,s} \coloneqq \begin{cases} \emptyset & \text{if } s \notin I_t \ , \\ \mathbf{1}(1/2 < v_s) & \text{if } s \in I_t \ . \end{cases}$$

Notice that for each $t_1, t_2 \in \mathbb{N}$ and each $s \in \{1, \dots, \min(t_1, t_2)\}$, we have that $Z_{t_1,s} = Z_{t_2,s}$. Then, for each $s \in \mathbb{N}$, it is well defined the random variable $Z_s \coloneqq Z_{t,s}$, where $t \in \mathbb{N}$ is any number $t \geq s$. Define, for each $t \in \mathbb{N}$, the random vector $\bar{Z}_t \coloneqq (Z_1, \dots, Z_t)$. Notice that, given that the sequence of randomization seeds is fixed and that, for each $s \in \mathbb{N}$, we have that $v_s \in \{1/4, 1/2, 3/4, 1\}$ (hence, for each $x \in (1/2, 3/4]$, it holds that $\mathbf{1}(1/2 < v_s) = \mathbf{1}(x = v_s)$), the random vector $(\tilde{x}_1, \dots, \tilde{x}_T)$ is measurable with respect to the σ-algebra generated by $\bar{Z}_{T-1}$. Hence, for each $\epsilon \in [0,1]$ and each $i \in \{1, \dots, T\}$, we can deduce from Pinsker's inequality (see, e.g., (Tsybakov, 2008, Lemma 2.5)) that

$$Q^{\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) \leq Q^{-\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) + \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_{T-1}} \mid\mid Q^{-\epsilon}_{\bar{Z}_{T-1}}\big)}\ , \tag{50}$$

where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence. Now, for each $t \in \mathbb{N}$ and each $\epsilon \in [0, 1]$, by the chain rule for Kullback-Leibler divergence (see, e.g., (Thomas M. Cover, 2006, Theorem 2.5.3)), we have

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_{t+1}} \mid\mid Q^{-\epsilon}_{\bar{Z}_{t+1}}\big) = \mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{(\bar{Z}_t, Z_{t+1})} \mid\mid Q^{-\epsilon}_{(\bar{Z}_t, Z_{t+1})}\big) &= \mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_t} \mid\mid Q^{-\epsilon}_{\bar{Z}_t}\big) \\ &\quad + \sum_{(\bar{z}, z) \in \{\emptyset, 0, 1\}^t \times \{\emptyset, 0, 1\}} \log\Bigg(\bigg(\frac{Q^{\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}{Q^{-\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}\bigg) \\ &\quad \cdot Q^{\epsilon}\big(\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\big)\Bigg)\ .\end{aligned} \tag{51}$$

Notice that, for each $t \in \mathbb{N}$ and each $\epsilon \in [0, 1]$ we have

$$\begin{aligned}&\sum_{(\bar{z}, z) \in \{\emptyset, 0, 1\}^t \times \{\emptyset, 0, 1\}} \log\left(\frac{Q^{\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}{Q^{-\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}\right) \cdot Q^{\epsilon}\big(\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\big) \\ &\qquad = \sum_{\substack{(\bar{z}, z) \in \{\emptyset, 0, 1\}^t \times \{\emptyset, 0, 1\} \\ t+1 \in I_{t+1}}} \log\left(\frac{Q^{\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}{Q^{-\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}\right) \cdot Q^{\epsilon}\big(\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\big) \\ &\qquad = \left(\sum_{\substack{\bar{z} \in \{\emptyset, 0, 1\}^t \\ t+1 \in I_{t+1}}} Q^{\epsilon}(\bar{Z}_t = \bar{z})\right) \\ &\qquad\qquad \cdot \sum_{z \in \{0, 1\}} \log\left(\frac{Q^{\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}{Q^{-\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}\right) \cdot Q^{\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big) \\ &\qquad = Q^{\epsilon}\big(\tilde{x}_{t+1} \in (1/2, 3/4]\big) \\ &\qquad\qquad \cdot \sum_{z \in \{0, 1\}} \log\left(\frac{Q^{\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}{Q^{-\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}\right) \cdot Q^{\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)\ .\end{aligned} \tag{52}$$

Algebraic calculations show that, for each $t \in \mathbb{N}$ and each $\epsilon \in [0,1]$,

$$
\begin{aligned}
\sum_{z\in\{0,1\}} \log\left(\frac{Q^{\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}{Q^{-\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}\right) \cdot Q^{\epsilon}(\mathbf{1}(1/2 < v_{t+1}) = z) \\
&= \log\left(\frac{Q^{\epsilon}\big(\frac{1}{2} < v_{t+1}\big)}{Q^{-\epsilon}\big(\frac{1}{2} < v_{t+1}\big)}\right) \cdot Q^{\epsilon}\left(\frac{1}{2} < v_{t+1}\right) \\
&\quad + \log\left(\frac{Q^{\epsilon}\big(\frac{1}{2} \geq v_{t+1}\big)}{Q^{-\epsilon}\big(\frac{1}{2} \geq v_{t+1}\big)}\right) \cdot Q^{\epsilon}\left(\frac{1}{2} \geq v_{t+1}\right) \\
&= \log\left(\frac{1-a-(1+\epsilon)\cdot b}{1-a-(1-\epsilon)\cdot b}\right) \cdot \big(1-a-(1+\epsilon)\cdot b\big) \\
&\quad \log\left(\frac{a+(1+\epsilon)\cdot b}{a+(1-\epsilon)\cdot b}\right) \cdot \big(a+(1+\epsilon)\cdot b\big) \\
&\leq \frac{4\cdot b^2\cdot\epsilon^2}{\big(1-a-(1-\epsilon)\cdot b\big)\cdot\big(a+(1-\epsilon)\cdot b\big)} \leq \frac{4\cdot b^2\cdot\epsilon^2}{a\cdot(1-a-2b)} = 2\cdot c_3^2\cdot\epsilon^2\ . \qquad (53)
\end{aligned}
$$

Putting (51), (52) and (53) together, we obtain that, for each $t \in \mathbb{N}$ and each $\epsilon \in [0,1]$,

$$
\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_{t+1}} \,||\, Q^{-\epsilon}_{\bar{Z}_{t+1}}\big) \leq \mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{Z_1} \,||\, Q^{-\epsilon}_{Z_1}\big) + 2\cdot c_3^2\cdot\epsilon^2\cdot\sum_{s=1}^{t} Q^{\epsilon}\big(\tilde{x}_{s+1} \in (1/2, 3/4]\big)\ . \tag{54}
$$

With the same technique used above, for each $\epsilon \in [0,1]$, we can prove that

$$
\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{Z_1} \,||\, Q^{-\epsilon}_{Z_1}\big) \leq 2\cdot c_3^2\cdot\epsilon^2\cdot Q^{\epsilon}\big(\tilde{x}_1 \in (1/2, 3/4]\big)\ . \tag{55}
$$

For each $t \in \{1,\dots,T\}$, putting (54) and (55) together, we obtain

$$
\begin{aligned}
\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_t} \,||\, Q^{-\epsilon}_{\bar{Z}_t}\big) &\overset{(54)+(55)}{\leq} 2\cdot c_3^2\cdot\epsilon^2\cdot\sum_{s=1}^{t} Q^{\epsilon}\big(\tilde{x}_s \in (1/2, 3/4]\big) \\
&\leq \quad 2\cdot c_3^2\cdot\epsilon^2\cdot E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)\ . \qquad (56)
\end{aligned}
$$

Now, (50) and (56) imply that, for each $\epsilon \in [0,1]$ and each $i \in \{1,\dots,T\}$,

$$
Q^{\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) \leq Q^{-\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) + c_3\cdot\epsilon\cdot\sqrt{E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)}\ . \tag{57}
$$

Taking the sum of (57) over $i \in \{1,\dots,T\}$, we obtain that for each $\epsilon \in [0,1]$,

$$
\begin{aligned}
&E^{-\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \\
&\qquad \geq E^{\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) - c_3\cdot\epsilon\cdot T\cdot\sqrt{E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)}\ . \qquad (58)
\end{aligned}
$$

Now, since the sequence $\bar{w}_1, \bar{w}_2, \dots$ of randomization seeds has been arbitrarily chosen, for each $\epsilon \in [0,1]$, using the fact that $P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}} = P^{-\epsilon}_{(w_t)_{t\in\mathbb{N}}}$ and Jensen's inequality, we have that

$$\begin{aligned}
E^{-\epsilon}(\tilde{n}_3) &= \int_{[0,1]^{\mathbb{N}}} E^{-\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{-\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\overset{(48)}{=} \int_{[0,1]^{\mathbb{N}}} E^{-\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\overset{(58)}{\geq} \int_{[0,1]^{\mathbb{N}}} E^{\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\quad - c_3 \cdot \epsilon \cdot T \cdot \int_{[0,1]^{\mathbb{N}}} \sqrt{E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)} \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
\text{(by Jensen)} \qquad &\geq \int_{[0,1]^{\mathbb{N}}} E^{\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\quad - c_3 \cdot \epsilon \cdot T \cdot \sqrt{\int_{[0,1]^{\mathbb{N}}} E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots)} \\
&= E^{\epsilon}(\tilde{n}_3) - c_3 \cdot \epsilon \cdot \sqrt{E^{\epsilon}(\tilde{n}_1)}\,. \qquad \square
\end{aligned}$$

B.3 Theorem 2 (Adversarial upper bound on regret)

The proof of this theorem builds upon the proof of Theorem 6.5 in Cesa-Bianchi and Lugosi (2006). Relative to this theorem, we need to additionally consider the discretization error introduced by Algorithm 1, and explicitly control the variance of estimated welfare.

Proof of Theorem 2.
Recall our notation $\mathbb{U}$ and $\mathbb{U}(x)$ for realized cumulative welfare, and for cumulative welfare for the counterfactual, fixed policy x. We further abbreviate $\mathbb{U}_{Tk} = \mathbb{U}(\tilde{x}_k)$. Throughout this proof, the sequence $\{v_i\}_{i=1}^{T}$ is given and conditioned on in any expectations.

1. **Discretization**
Recall that $U_i(x) = x \cdot \mathbf{1}(x \leq v_i) + \lambda \cdot \max(v_i - x, 0)$. Let

$$\tilde{v}_i = \max_k \{\tilde{x}_k : \ \tilde{x}_k \leq v_i\}$$

(this is v_i rounded down to the next gridpoint $\tilde{x}_k$), and denote

$$\begin{aligned}
\tilde{U}_i(x) &= x \cdot \mathbf{1}(x \leq v_i) + \lambda \cdot \max(\tilde{v}_i - x, 0), \\
\tilde{\mathbb{U}}_i(x) &= \sum_{j \leq i} \tilde{U}_j(x),
\end{aligned}$$

as well as $\tilde{\mathbb{U}}_{ik} = \tilde{\mathbb{U}}_i(\tilde{x}_k)$. Then it is immediate that $\tilde{U}_i(x) \leq U_i(x)$,

$$\sup_x |\tilde{U}_i(x) - U_i(x)| \leq \frac{\lambda}{K},$$

and argmax $_x\, \tilde{\mathbb{U}}_i(x) \in \{\tilde{x}_k\}$, and therefore

$$\max_k \tilde{\mathbb{U}}_{ik} \geq \sup_x \mathbb{U}_i(x) - i \cdot \frac{\lambda}{K}$$

2. **Unbiasedness**

At the end of period i, $\widehat{G}_k$ is an unbiased estimator of $\sum_{j \leq i} \mathbf{1}(\tilde{x}_k \leq v_j)$ for all k. Therefore, $E\left[\widehat{\mathbb{U}}_{ik}\right] = \tilde{\mathbb{U}}_{ik}$ for all i and k.

3. **Upper bound on optimal welfare**
Define $W_i = \sum_k \exp(\eta \cdot \widehat{\mathbb{U}}_{ik})$, and $q_{ik} = \exp(\eta \cdot \widehat{\mathbb{U}}_{ik})/W_i$.
It is immediate that,

$$E[\log W_T] \geq \eta \cdot E[\max_k \widehat{\mathbb{U}}_{Tk}] \geq \eta \cdot \max_k E[\widehat{\mathbb{U}}_{Tk}] = \eta \cdot \max_k \tilde{\mathbb{U}}_{Tk}.$$

Furthermore

$$E[\log W_T] = \sum_{0 \leq i < T} E\left[\log\left(\frac{W_{i+1}}{W_i}\right)\right] + \log(W_0).$$

Given our initialization of the algorithm, $\log(W_0) = \log(K+1)$.

4. **Lower bound on estimated welfare**
Denote $\widehat{U}_{ik} = \tilde{x}_k \cdot \widehat{H}_k + \frac{\lambda}{K} \cdot \sum_{k' > k} \widehat{H}_{k'}$, where $\widehat{H}_k = \frac{y_i}{p_{ik}} \cdot \mathbf{1}(k_i = k)$,
so that $\widehat{\mathbb{U}}_{ik} = \sum_{j < i} \widehat{U}_{jk}$, and $E[\widehat{U}_{jk}] = U_i(\tilde{x}_k)$.
By definition of W_i,

$$\log\left(\frac{W_{i+1}}{W_i}\right) = \log\left(\sum_k q_{ik} \cdot \exp(\eta \cdot \widehat{U}_{ik})\right).$$

Since $p_k \geq \gamma/(K+1)$ for all k, $\widehat{U}_{ik} \in [0, 1/\gamma]$ for all i and k, and therefore $\eta \cdot \widehat{U}_{ik} \leq (K+1) \cdot \eta/\gamma \leq 1$ (where the last inequality holds by assumption). Using $\exp(a) \leq 1 + a + (e-2)a^2$ for any $a \leq 1$ yields

$$\exp\left(\eta \widehat{U}_{ik}\right) \leq 1 + \eta \cdot \widehat{U}_{ik} + (e-2) \cdot \left(\eta \cdot \widehat{U}_{ik}\right)^2.$$

Therefore,

$$\begin{aligned}\log\left(\frac{W_{i+1}}{W_i}\right) \leq & \log\left(\sum_k q_{ik} \cdot \left(1 + \eta \cdot \widehat{U}_{ik} + (e-2) \cdot \left(\eta \cdot \widehat{U}_{ik}\right)^2\right)\right) \\ \leq & \eta \cdot \sum_k q_{ik} \cdot \widehat{U}_{ik} + (e-2) \cdot \eta^2 \cdot \sum_k q_{ik} \cdot \widehat{U}_{ik}^2\end{aligned}$$

The second inequality follows from $\log(1+x) \leq x$.

5. **Connecting the first order term to welfare**
Note that, by definition, $q_{ik} = \left(p_{ik} - \frac{\gamma}{K+1}\right)/(1-\gamma)$. Therefore

$$\sum_k q_{ik} \cdot \widehat{U}_{ik} = \frac{1}{1-\gamma} \sum_k p_{ik} \cdot \widehat{U}_{ik} - \tfrac{\gamma}{(1-\gamma)(K+1)} \cdot \sum_k \widehat{U}_{ik},$$

and thus

$$E\left[\sum_k q_{ik} \cdot \widehat{U}_{ik}\right] \leq \frac{1}{1-\gamma} E\left[\tilde{U}_i(x_i)\right],$$

where we have used the fact that $0 \leq \tilde{U}_k \leq 1$ for all k, given our definition of $\tilde{U}$, and the fact that k_i is distributed according to p_{ik}, by construction.

6. **Bounding the second moment of estimated welfare**
It remains to bound the term $E\left[\sum_k q_{ik}\cdot\widehat{U}_{ik}^2\right]$. As in the preceding item, we have

$$\sum_k q_{ik}\cdot\widehat{U}_{ik}^2 \leq \frac{1}{1-\gamma}\sum_k p_{ik}\cdot\widehat{U}_{ik}^2.$$

We can rewrite

$$\widehat{U}_{ik} = \left(\tilde{x}_k\cdot\mathbf{1}(k_i=k) + \tfrac{\lambda}{K}\cdot\mathbf{1}(k_i>k)\right)\cdot\frac{y_i}{p_{ik_i}}.$$

Bounding $y_i \leq 1$ immediately gives

$$E_i\left[\widehat{U}_{ik}^2\right] \leq \frac{\tilde{x}_k^2}{p_{ik}} + \left(\tfrac{\lambda}{K}\right)^2\cdot\sum_{k'>k}\frac{1}{p_{ik'}},$$

and therefore

$$\begin{aligned}
E_i\left[\sum_k p_{ik}\cdot\widehat{U}_{ik}^2\right] &\leq \sum_k \tilde{x}_k^2 + \left(\tfrac{\lambda}{K}\right)^2\cdot\sum_k\sum_{k'>k}\frac{p_{ik}}{p_{ik'}}\\
&\leq \sum_k\left(\tfrac{k}{K}\right)^2 + \left(\tfrac{\lambda}{K}\right)^2\cdot\sum_k p_{ik}\sum_{k'\neq k}\tfrac{K+1}{\gamma}\\
&= \tfrac{K(K+1)(2K+1)}{6K^2} + \tfrac{\lambda^2}{\gamma}\tfrac{K+1}{K}\\
&= \tfrac{K+1}{K}\cdot\left(\tfrac{2K+1}{6}+\tfrac{\lambda^2}{\gamma}\right).
\end{aligned}$$

7. **Collecting inequalities**
Combining the preceding items, we get

$$\begin{aligned}
&\eta\cdot\left(\sup_x\mathbb{U}(x) - T\cdot\frac{\lambda}{K}\right)\\
\leq&\eta\cdot\max_k\tilde{\mathbb{U}}_{Tk} \leq E[\log W_T] && \text{(Item 1)}\\
=&\sum_{0\leq i<T}E\left[\log\left(\frac{W_{i+1}}{W_i}\right)\right] + \log(K+1) && \text{(Item 3)}\\
\leq&\frac{\eta}{1-\gamma}\cdot E\left[\tilde{\mathbb{U}}\right] + (e-2)\cdot\frac{\eta^2}{1-\gamma}\sum_{1\leq i\leq T}\sum_k E\left[p_{ik}\cdot\widehat{U}_{ik}^2\right] + \log(K+1) && \text{(Item 4 and 5)}\\
\leq&\frac{\eta}{1-\gamma}\cdot E\left[\tilde{\mathbb{U}}\right] + (e-2)\cdot\frac{\eta^2}{1-\gamma}T\cdot\tfrac{K+1}{K}\cdot\left(\tfrac{2K+1}{6}+\tfrac{\lambda^2}{\gamma}\right) + \log(K+1). && \text{(Item 6)}
\end{aligned}$$

Multiplying by $(1-\gamma)$ and dividing by η, adding $\gamma\sup_x\mathbb{U}(x)+T\frac{\lambda}{K}$ to both sides and subtracting $E\left[\tilde{\mathbb{U}}\right]$, bounding $\sup_x\mathbb{U}(x)\leq T$, and $E\left[\tilde{\mathbb{U}}\right]\leq E\left[\mathbb{U}\right]$ (from Item 1), yields

$$\begin{aligned}
&\sup_x\mathbb{U}(x) - E\left[\mathbb{U}\right]\\
&\leq\left(\gamma + \eta\cdot(e-2)\tfrac{K+1}{K}\cdot\left(\tfrac{2K+1}{6}+\tfrac{\lambda^2}{\gamma}\right)+\tfrac{\lambda}{K}\right)\cdot T + \tfrac{\log(K+1)}{\eta}.
\end{aligned}\tag{59}$$

This proves the first claim of the theorem.

8. **Optimizing tuning parameters**
   Suppose now that we choose the tuning parameters as follows:

$$\gamma = c_1 \cdot \left(\tfrac{\log(T)}{T}\right)^{1/3}, \qquad \eta = c_2 \cdot \gamma^2, \qquad K = c_3/\gamma.$$

   Plugging in we get

$$\begin{aligned}
&\sup_x \mathbb{U}(x) - E\left[\mathbb{U}\right] \\
&\leq \left(\gamma + c_2 \cdot \gamma^2 \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2c_3/\gamma+1}{6} + \tfrac{\lambda^2}{\gamma}\right) + \lambda \cdot \gamma/c_3\right) \cdot T + \tfrac{\log(K+1)}{c_2 \cdot \gamma^2} \\
&\quad = \log(T)^{1/3} T^{2/3} \cdot \left(c_1 + (e-2)\tfrac{K+1}{K} \cdot c_1 c_2 \left(\tfrac{c_3}{3} + \lambda^2 + \tfrac{\gamma}{6}\right) + \lambda \tfrac{c_1}{c_3} + \frac{\log(T^{1/3}\log(T)^{-1/3} c_3/c_1 + 1)}{c_1^2 \log(T)}\right) \\
&\quad = \log(T)^{1/3} T^{2/3} \cdot \left(c_1 + (e-2) \cdot c_1 c_2 \left(\tfrac{c_3}{3} + \lambda^2\right) + \lambda \tfrac{c_1}{c_3} + \frac{1}{3c_1^2} + o(1)\right).
\end{aligned}$$

   The second claim of the theorem follows.

□

B.4 Theorem 3 (Lower bound on regret for the concave case)

Proof of Theorem 3.

Defining a family of distributions for v Define $\bar{h} := \frac{1-\sqrt{1-\lambda}}{2}$ and notice that $0 < \bar{h} < \frac{1}{2}$. Define $\bar{\eta} := \left(\bar{h} \cdot (1-\bar{h})^{1-\lambda} \cdot (1-\lambda)\right)^{-1}$ and $\bar{\epsilon} := \frac{1}{2} \cdot \min(\bar{\eta}, \frac{2}{3} \cdot 2^{-\lambda})$. For each $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$ and each $x \in [0,1]$, define

$$f^\epsilon(x) := \bar{c} \cdot \left(\left(2^{2-\lambda} - 8 \cdot \bar{h} \cdot \epsilon\right)\right) \cdot x \cdot \mathbf{1}_{[0,\frac{1}{2})}(x) + \frac{1}{x^{2-\lambda}} \cdot \mathbf{1}_{[\frac{1}{2},1-\bar{h}]}(x) + (\bar{\eta} + \epsilon) \cdot \mathbf{1}_{(1-\bar{h},1]}(x)\Bigg) ,$$

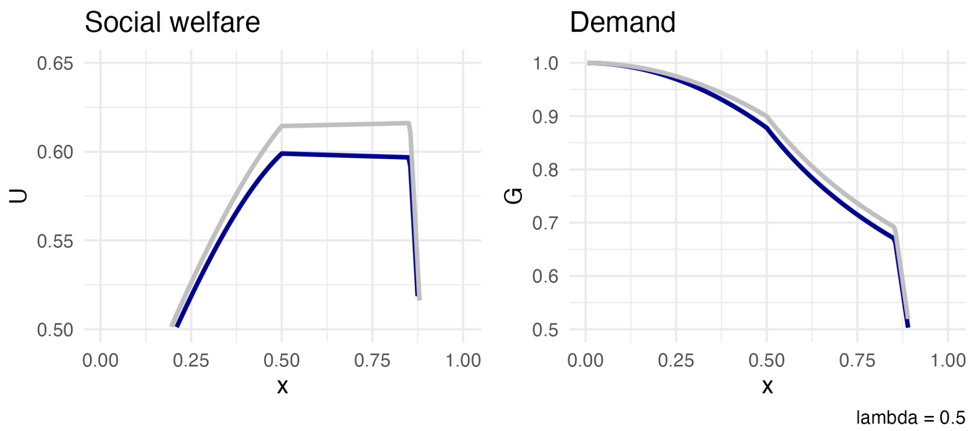
where $\bar{c}$ is such that $\int_0^1 f^0(x)\mathrm{d}x = 1$. For each $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$, note that f^ϵ is a density function on $[0,1]$, i.e., a non-negative function whose integral is 1. For each $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$, let μ^ϵ be the probability measure whose density is f^ϵ, and define $\boldsymbol{G}^\epsilon$ and $\boldsymbol{U}^\epsilon$ as the demand function and the expected social welfare associated to μ^ϵ, respectively. Figure 2 illustrates.

Properties of $\boldsymbol{U}$ Define also $\bar{x} := \frac{1}{2} \cdot \left(\frac{1}{2} + (1-\bar{h})\right) = \frac{3}{4} - \frac{\bar{h}}{2}$ and $\bar{m} := \frac{1-\sqrt{1-\lambda}}{8} \cdot (1-\lambda)^{3/2}$. Notice that, for all $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$, we have:

- $\boldsymbol{U}^\epsilon$ is continuous and concave.
- $\boldsymbol{U}^\epsilon$ is strictly increasing in $[0, \frac{1}{2}]$, linear in $[\frac{1}{2}, 1-\bar{h}]$ with slope $(1-\lambda) \cdot \bar{h} \cdot \epsilon$, and strictly decreasing on $[1-\bar{h}, 1]$, which in particular implies that the maximum of $\boldsymbol{U}^\epsilon$ is at $1-\bar{h}$ if $\epsilon > 0$, and at $\frac{1}{2}$ if $\epsilon < 0$.
- If $\epsilon > 0$, then $\boldsymbol{U}^\epsilon(1-\bar{h}) - \max_{x \in [0,\bar{x}]} \boldsymbol{U}^\epsilon(x) = \bar{m} \cdot |\epsilon| = \boldsymbol{U}^{-\epsilon}(\frac{1}{2}) - \max_{x \in [\bar{x},1]} \boldsymbol{U}^{-\epsilon}(x)$.

Now, consider the sequence of individual valuations $v_1, v_2, \dots \in [0,1]$, and assume that, for each $\epsilon \in (-\bar{\epsilon}, \bar{\epsilon})$, when the underlying distribution is P^ϵ, this sequence is i.i.d. (independent of the randomization used by the algorithm) with common distribution μ^ϵ. The previous list of properties implies that, for each $\epsilon \in (0, \bar{\epsilon})$ (resp., $\epsilon \in (-\bar{\epsilon}, 0)$), when the underlying distribution is P^ϵ, the expected instantaneous regret at time t is at least $\bar{m} \cdot |\epsilon|$ if the learner plays in the region $\bar{I} := [0, \bar{x}]$ (resp., in the region $\bar{J} := (\bar{x}, 1]$). It follows that, in order not to suffer linear regret, the learner has to discriminate the sign of ϵ.

Figure 2: Construction for proving the lower bound on regret for the concave case



Intuition for the proof Now, the high-level idea is that in order to discriminate the sign of ϵ, the learner needs on the order of $\frac{1}{\epsilon^2}$ observations. Therefore, for a number of periods on the order of $\frac{1}{\epsilon^2}$, the algorithm is playing "in the dark," and thus suffers an expected regret on the order of $\epsilon \cdot T$, or, equivalently, on the order of $\sqrt{T}$, when the underlying distribution is between P^{ϵ} or $P^{-\epsilon}$.

Defining constants We now formalize this idea. Let

$$\gamma \coloneqq \left(\int_0^{1/2} \left(\frac{16\bar{h}}{2^{2-\lambda} - 8\bar{h}\bar{\epsilon}} \right)^2 f^{-\bar{\epsilon}}(x)\mathrm{d}x + \int_{1-\bar{h}}^{1} \left(\frac{2}{\bar{\eta} - \bar{\epsilon}} \right)^2 f^{\bar{\epsilon}}(x)\mathrm{d}x \right)^{1/2} > 0$$

Let $\bar{M} > 0$ such that $2 \cdot \sqrt{\frac{\sqrt{2}}{3} \cdot \frac{\gamma \cdot \bar{M}}{\bar{m}}} = 1$. Let $M \in (0, \bar{M})$ such that

$$k \coloneqq \sqrt{\frac{\frac{M}{\bar{m}}}{\frac{\sqrt{2}}{3} \cdot \gamma}} < \bar{\epsilon} \, .$$

From now on, fix a time horizon $T \in \mathbb{N}$ and let $\epsilon \coloneqq \frac{k}{\sqrt{T}}$. In the following we use the notation E^{ϵ} (resp., $E^{-\epsilon}$) to denote the expectation with respect to the probability measure P^{ϵ} (resp., $P^{-\epsilon}$). Let $x_1, x_2, \dots$ be the policies chosen by the algorithm. Note that, since the algorithm bases its decision at time t only on the (partial) knowledge of $v_1, \dots, v_{t-1}$ and some independent randomization, there exists a (measurable) function $\varphi_t \colon [0,1]^{t-1} \to [0,1]$ such that

$$E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I}) \mid v_1, \dots, v_{t-1}\big) = \varphi_t(v_1, \dots, v_{t-1}) = E^{-\epsilon}\big(\mathbf{1}(x_t \in \bar{I}) \mid v_1, \dots, v_{t-1}\big) \, .$$

Then, for each time t, it holds

$$\begin{aligned} \big| E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) - E^{-\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) \big| &= \big| E^{\epsilon}\big(\varphi_t(v_1, \dots, v_{t-1})\big) - E^{-\epsilon}\big(\varphi_t(v_1, \dots, v_{t-1})\big) \big| \\ &\le \left\| \bigotimes_{s=1}^{t-1} \mu^{\epsilon} - \bigotimes_{s=1}^{t-1} \mu^{-\epsilon} \right\|_{\mathrm{TV}} = (\star) \end{aligned}$$

Relating choice probabilities for positive and negative ϵ By Pinsker's inequality and the fact that the Kullback-Leibler divergence is upper bounded by the χ^2-divergence, it follows that

$$(\star) \leq \sqrt{\frac{\mathcal{D}_{\mathrm{KL}}\left(\bigotimes_{s=1}^{t-1} \mu^{-\epsilon}, \bigotimes_{s=1}^{t-1} \mu^{\epsilon}\right)}{2}} = \sqrt{\frac{(t-1) \cdot \mathcal{D}_{\mathrm{KL}}\left(\mu^{-\epsilon}, \mu^{\epsilon}\right)}{2}} \leq \sqrt{\frac{(t-1) \cdot \mathcal{D}_{\chi^2}\left(\mu^{-\epsilon}, \mu^{\epsilon}\right)}{2}}$$
$$= \sqrt{\frac{t-1}{2} \int_0^1 \left| \frac{f^{\epsilon}(x)}{f^{-\epsilon}(x)} - 1 \right|^2 f^{-\epsilon}(x) \mathrm{d}x} = (\star\star)$$

Now, noticing that

$$\int_0^1 \left| \frac{f^{\epsilon}(x)}{f^{-\epsilon}(x)} - 1 \right|^2 f^{-\epsilon}(x) \mathrm{d}x = \int_0^{1/2} \left(\frac{16\bar{h}\epsilon}{2^{2-\lambda} + 8\bar{h}\epsilon} \right)^2 f^{-\epsilon}(x) \mathrm{d}x + \int_{1-\bar{h}}^1 \left(\frac{2\epsilon}{\bar{\eta} - \epsilon} \right)^2 f^{-\epsilon}(x) \mathrm{d}x$$
$$\leq \left(\int_0^{1/2} \left(\frac{16\bar{h}}{2^{2-\lambda} - 8\bar{h}\bar{\epsilon}} \right)^2 f^{-\bar{\epsilon}}(x) \mathrm{d}x + \int_{1-\bar{h}}^1 \left(\frac{2}{\bar{\eta} - \bar{\epsilon}} \right)^2 f^{\bar{\epsilon}}(x) \mathrm{d}x \right) \cdot \epsilon^2 = \gamma^2 \cdot \epsilon^2 \ ,$$

it follows that

$$(\star\star) \leq \gamma \cdot \epsilon \cdot \sqrt{\frac{t-1}{2}} \ .$$

Summing over $t = 1, 2, \ldots, T$, we obtain

$$\left| E^{\epsilon}\left(\sum_{t=1}^T \mathbf{1}(x_t \in \bar{I}) \right) - E^{-\epsilon}\left(\sum_{t=1}^T \mathbf{1}(x_t \in \bar{I}) \right) \right| \leq \frac{\sqrt{2}}{3} \cdot \gamma \cdot \epsilon \cdot T^{3/2} = \frac{\sqrt{2}}{3} \cdot \gamma \cdot k \cdot T \ .$$

Upper bound on regret for $\epsilon > 0$ implies lower bound on regret for $-\epsilon$. Now, suppose that in the scenario determined by P^{ϵ} the algorithm suffer a regret $R_T^{\epsilon} \leq M \cdot \sqrt{T}$. Then

$$M \cdot \sqrt{T} \geq R_T^{\epsilon} \geq \bar{m} \cdot \epsilon \cdot \sum_{t=1}^T E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) = \bar{m} \cdot \frac{k}{\sqrt{T}} \cdot \sum_{t=1}^T E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) \ .$$

and rearranging

$$\sum_{t=1}^T E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) \leq \frac{M \cdot T}{\bar{m} \cdot k}$$

It follows that the expected number of times the algorithm plays in the (correct) region I when the underlying scenario is determined by $P^{-\epsilon}$ is

$$\sum_{t=1}^T E^{-\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) = \left(\sum_{t=1}^T E^{-\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) - \sum_{t=1}^T E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big) \right) + \sum_{t=1}^T E^{\epsilon}\big(\mathbf{1}(x_t \in \bar{I})\big)$$
$$\leq \left(\frac{\sqrt{2}}{3} \cdot \gamma \cdot k + \frac{M}{\bar{m} \cdot k} \right) \cdot T$$

The last inequality implies that the expected number of times that the algorithm plays in the (wrong) region $J = I^c$ when the underlying scenario is determined by $P^{-\epsilon}$ is lower bounded by

$$\sum_{t=1}^T E^{-\epsilon}\big(\mathbf{1}(x_t \in \bar{J})\big) = \sum_{t=1}^T E^{-\epsilon}\big(\mathbf{1}(x_t \notin \bar{I})\big) \geq \left(1 - \left(\frac{\sqrt{2}}{3} \cdot \gamma \cdot k + \frac{M}{\bar{m} \cdot k} \right) \right) \cdot T \ ,$$

which implies that the regret the algorithm suffers in the scenario determined by $P^{-\epsilon}$ is lower bounded by

$$
\begin{aligned}
R_T^{-\epsilon} &\geq \bar{m} \cdot \epsilon \cdot \sum_{t=1}^{T} E^{-\epsilon}\left(\mathbf{1}(x_t \in J)\right) = \bar{m} \cdot \frac{k}{\sqrt{T}} \cdot \sum_{t=1}^{T} E^{-\epsilon}\left(\mathbf{1}(x_t \in J)\right) \\
&\geq \bar{m} \cdot \frac{k}{\sqrt{T}} \cdot \left(1 - \left(\frac{\sqrt{2}}{3} \cdot \gamma \cdot k + \frac{M}{\bar{m} \cdot k}\right)\right) \cdot T = \bar{m} \cdot k \cdot \left(1 - 2\sqrt{\frac{\sqrt{2} \cdot \gamma \cdot M}{3 \cdot \bar{m}}}\right) \cdot \sqrt{T} \,.
\end{aligned}
$$

Putting everything together, any algorithm has to suffer regret of at least $\min\left(M, \bar{m} \cdot k \cdot \left(1 - 2 \cdot \sqrt{\frac{\sqrt{2} \cdot \gamma \cdot M}{3 \cdot \bar{m}}}\right)\right) \cdot \sqrt{T}$ in at least one scenario between the ones determined by P^{ϵ} and $P^{-\epsilon}$. Recalling that our choice of M implies $1 - 2\sqrt{\frac{\sqrt{2} \cdot \gamma \cdot M}{3 \cdot \bar{m}}} > 0$, the conclusion follows. □

B.5 Theorem 4

For the sake of simplicity, we assume that $\boldsymbol{U}$ admits a unique minimizer $x^\star \in [0,1]$ (the other cases can be treated similarly, and actually imply better constants).

For each epoch $\tau = 1, 2, \dots,$ we refer to the three current l (left), c (center) and r (right) points of the corresponding epoch τ using l_τ, c_τ and r_τ, respectively. For any time t, the epoch to which the time t belongs is denoted τ_t. The length of an interval J is denoted $|J|$, while the number of elements in a finite set A is denoted $\# A$.

Consider a family $(v_{x,i})_{x \in [0,1], i \in \mathbb{N}}$ of random variables such that, for each $x \in [0,1]$, the sequence $(v_{x,i})_{i \in \mathbb{N}}$ is i.i.d. with the same distribution as $(v_i)_{i \in \mathbb{N}}$. With these random variables, we can define the auxiliary family $(y_{x,i})_{x \in [0,1], i \in \mathbb{N}} \coloneqq \big(\mathbf{1}(x \leq v_{x,i})\big)_{x \in [0,1], i \in \mathbb{N}}$. We assume that, whenever we select a policy $x \in [0,1]$ at time t, we observe $\mathbf{1}(x \leq v_{x, n_t(x)})$ (recall that $n_t(x) = \sum_{s=1}^{t} \mathbf{1}(x_s = x)$) instead of $\mathbf{1}(x \leq v_t)$. This does not change anything in expectation, but will be useful in what follows.

The next lemma states that Algorithm 2 maintains confidence intervals containing the differences of the welfare function (among left, center and right points) with high probability.

Lemma 1 (Confidence intervals contain true welfare differences with high probability)**.** *There exists a constant $\tilde{C} \in (0, 20]$ such that, for every time horizon T and any $\delta \in (0,1)$, if the learner runs Algorithm 2 with confidence parameter δ, then the probability of the event*

$$
\mathcal{E} \coloneqq \bigcap_{t=1}^{T} \Big(\big\{ \boldsymbol{U}(c_{\tau_t}) - \boldsymbol{U}(l_{\tau_t}) \in J_t(l_{\tau_t}, c_{\tau_t}) \big\} \cap \big\{ \boldsymbol{U}(r_{\tau_t}) - \boldsymbol{U}(c_{\tau_t}) \in J_t(c_{\tau_t}, l_{\tau_t}) \big\} \cap \big\{ \boldsymbol{U}(r_{\tau_t}) - \boldsymbol{U}(l_{\tau_t}) \in J_t(r_{\tau_t}, l_{\tau_t}) \big\} \Big)
$$

is lower bounded by $1 - \tilde{C} \cdot T^2 \cdot \delta$.

Proof. For each $n \in \mathbb{N}$, let $\mathcal{D}_n \coloneqq \{ k \cdot 2^{-n} \mid k \in \mathbb{Z} \}$, let $\mathcal{D}_n^\star \coloneqq \{ x_{n,1}, \dots, x_{n,10} \} \subset \mathcal{D}_n$ such that

$$
x_{n,1} < \dots < x_{n,5} \leq x^\star \leq x_{n,6} < \dots < x_{n,10}
$$

and $x_{n,j+1} - x_{n,j} \leq 2^{-n}$, for all $j \in \{1, \dots, 9\}$. Define $\mathcal{D} \coloneqq \bigcup_{n=1}^{T} \mathcal{D}_n^\star \cap (0,1)$. Consider the following events

$$
\mathcal{E}' \coloneqq \bigcap_{\substack{n,t \in \{1,\dots,T\} \\ j \in \{1,\dots,10\}}} \left\{ \left| \frac{1}{t} \sum_{s=1}^{t} y_{x_{n,j}, s} - \boldsymbol{G}(x_{n,j}) \right| \leq \sqrt{\frac{1}{2t} \log\left(\frac{2}{\delta}\right)} \right\}
$$

$$
\mathcal{E}'' \coloneqq \bigcap_{\substack{n \in \{1,\dots,T\} \\ m \in \{1,\dots,\lfloor \log_2(T) \rfloor\} \\ j \in \{1,\dots,9\}}} \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m - 1} y_{x_{n,j} + \frac{i}{2^{n+m}}, 1} - \frac{1}{x_{n,j+1} - x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \boldsymbol{G}(x) \mathrm{d}x \right| \leq \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} + \frac{2}{2^m} \right\}
$$

and note that $\mathcal{E} \subset \mathcal{E}' \cup \mathcal{E}''$, since, in the event $\mathcal{E}' \cup \mathcal{E}''$, Algorithm 2 will query only points in $\mathcal{D}^\star$, using a subset of those estimates to build its own estimates (in particular, due to the ties breaking rules, to estimate the integral terms it will only use the first query of the relevant dyadic points). Now, notice that for each $n \in \{1, \dots, n\}$, each $m \in \{1, \dots, \lfloor \log_2(T) \rfloor\}$ and each $j \in \{1, \dots, 9\}$ we have

$$
\begin{aligned}
&\left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j}+\frac{i}{2^{n+m}},1} - \frac{1}{x_{n,j+1}-x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \boldsymbol{G}(x) \mathrm{d}x \right| > \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} + \frac{2}{2^m} \right\} \\
&\subset \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j}+\frac{i}{2^{n+m}},1} - \frac{1}{2^m} \sum_{i=1}^{2^m-1} \boldsymbol{G}\left(x_{n,j} + \frac{i}{2^{n+m}}\right) \right| > \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} \right\} \\
&\quad \cup \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} \boldsymbol{G}\left(x_{n,j} + \frac{i}{2^{n+m}}\right) - \frac{1}{x_{n,j+1}-x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \boldsymbol{G}(x) \mathrm{d}x \right| > \frac{2}{2^m} \right\} \\
&= \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j}+\frac{i}{2^{n+m}},1} - \frac{1}{2^m} \sum_{i=1}^{2^m-1} \boldsymbol{G}\left(x_{n,j} + \frac{i}{2^{n+m}}\right) \right| > \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} \right\}
\end{aligned}
$$

where the last equality follows from

$$
\begin{aligned}
&\left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} \boldsymbol{G}\left(x_{n,j} + \frac{i}{2^{n+m}}\right) - \frac{1}{x_{n,j+1}-x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \boldsymbol{G}(x) \mathrm{d}x \right| \\
&\le \sum_{i=1}^{2^m-1} \int_{x_{n,j}+\frac{i-1}{2^{n+m}}}^{x_{n,j}+\frac{i}{2^{n+m}}} \left(\boldsymbol{G}(x) - \boldsymbol{G}\left(x_{n,j} + \frac{i}{2^{n+m}}\right) \right) \mathrm{d}x + \frac{1}{2^m} \\
&\le \sum_{i=1}^{2^m-1} \int_{x_{n,j}+\frac{i-1}{2^{n+m}}}^{x_{n,j}+\frac{i}{2^{n+m}}} \left(\boldsymbol{G}\left(x_{n,j} + \frac{i-1}{2^{n+m}}\right) - \boldsymbol{G}\left(x_{n,j} + \frac{i}{2^{n+m}}\right) \right) \mathrm{d}x + \frac{1}{2^m} \\
&\le \frac{1}{2^m} \cdot \left(\boldsymbol{G}\left(x_{n,j}\right) - \boldsymbol{G}\left(x_{n,j+1}\right) \right) + \frac{1}{2^m} \le \frac{2}{2^m}
\end{aligned}
$$

By De Morgan's laws, a union bound and Hoeffding's inequality, we have $P(\mathcal{E}^c) \le P\big((\mathcal{E}')^c\big) + P\big((\mathcal{E}'')^c\big) \le 20 \cdot T^2 \cdot \delta$. □

The following lemma establishes the rate of shrinking of the length of the confidence intervals as the length of an epoch increases.

Lemma 2 (Confidence intervals shrink with epoch length)**.** *For any $\delta \in (0,1)$, if the learner runs Algorithm 2 with confidence parameter δ then, for any time t,*

$$
\max\left(\left| J_t(l_{\tau_t}, c_{\tau_t}) \right|, \left| J_t(c_{\tau_t}, r_{\tau_t}) \right|, \left| J_t(l_{\tau_t}, r_{\tau_t}) \right| \right) \le \frac{\tilde{c}_\delta}{\sqrt{t - t_{\tau_t - 1}}} \ , \tag{60}
$$

whenever $t - t_{\tau_t-1} \ge \tilde{n}$, where $\tilde{n} = 10$ and $\tilde{c}_\delta = 72 \cdot \sqrt{10} \cdot \left(\sqrt{2 \log(2/\delta)} + 4 \right)$.

We break the proof of Lemma 2 in several steps. Let $d_1, d_2, d_3, d_4, d_5 > 0$ be constants. For each $k \in \{1, 2, 3\}$, define

$$
f_k : \{0, 1, 2, \dots\} \to [0, +\infty], \qquad n \mapsto \frac{d_k}{\sqrt{n}}
$$

and for each $k \in \{4, 5\}$ define

$$
f_k : \{0, 1, 2, \dots\} \to [0, +\infty], \qquad n \mapsto \frac{d_4}{\sqrt{2^{\lfloor \log_2(n+1) \rfloor} - 1}} + \frac{d_5}{2^{\lfloor \log_2(n+1) \rfloor}} \ ,
$$

Algorithm 5 Index selection

```
for $s = 1, 2, \dots$ do
    Let $k_s = \min\Big(\operatorname{argmax}_{k\in[5]} f_k\big(m_k(s-1)\big)\Big)$
    $m_{k_s}(s) = m_{k_s}(s-1) + 1$
    for $i \in [5] \setminus \{k_s\}$ do
        $m_i(s) = m_i(s-1)$
    end for
end for
```

with the usual convention that $a/0 = +\infty$, for any $a > 0$. Suppose that $m_1(0), m_2(0), m_3(0), m_4(0), m_5(0) \in \{0, 1, 2, \dots\}$ and consider the following algorithm.

The following lemma holds.

Lemma 3. *Consider Algorithm 5 and the notation defined therein. For each $s \in \mathbb{N}$ there exists an index $i \in [5]$ for which $m_i(s) \geq \lceil s/5 \rceil$*

Proof. Let $s \in \mathbb{N}$ and suppose by contradiction that for each $k \in [5]$ it holds that $m_k(s) < s/5$. Then

$$s \leq \sum_{k=1}^{5} m_k(s) \leq 5 \cdot \max_{k\in[5]} m_k(s) < 5 \cdot \frac{s}{5} = s \ ,$$

which is a contradiction. It follows that there exists $k \in [5]$ for which $m_k(s) \geq s/5$, which also implies $m_k(s) \geq \lceil s/5 \rceil$. Given that s was arbitrarily chosen, the conclusion follows. □

Notice that, for each $n \in \{0, 1, 2, \dots\}$, we have

$$\frac{d_4}{\sqrt{n}} \leq \frac{d_4}{\sqrt{2^{\lfloor \log_2(n+1) \rfloor} - 1}} \leq \frac{2d_4}{\sqrt{n}}$$

and

$$0 \leq \frac{d_5}{\sqrt{2^{\lfloor \log_2(n+1) \rfloor}}} \leq \frac{2d_5}{n} \ ,$$

which implies that, for each $k \in [5]$ and each $n \in \{0, 1, 2, \dots\}$

$$\frac{d_k}{\sqrt{n}} \leq f_k(n) \leq \frac{D_k}{\sqrt{n}}$$

where $D_1 = d_1, D_2 = d_2, D_3 = d_3, D_4 = D_5 = 2(d_4 + d_5)$.

The following lemma holds.

Lemma 4. *Consider Algorithm 5 and the notation defined therein. For any $i, j \in [5]$ and any $s \in \mathbb{N}$ it holds*

$$m_i(s) \geq \left(\frac{d_i}{D_j}\right)^2 (m_j(s) - 1) \ .$$

Proof. Let $i, j \in [5]$. Suppose by contradiction that the conclusion does not hold. Then there exists a smallest $s \in \{0, 1, 2, \dots\}$ for which

$$m_i(s) < \left(\frac{d_i}{D_j}\right)^2 (m_j(s) - 1) \ ,$$

which we call s_0. Notice that $s_0 \neq 0$. Then, the fact that

$$m_i(s_0 - 1) \geq \left(\frac{d_i}{D_j}\right)^2 (m_j(s_0 - 1) - 1) \ ,$$

implies that at time s_0 the algorithm selected $k_{s_0} = j$, which in turn implies that $m_i(s_0 - 1) = m_i(s_0)$ and $m_j(s_0 - 1) = m_j(s_0) - 1$. It follows that

$$\left(\frac{d_i}{D_j}\right)^2 m_j(s_0 - 1) = \left(\frac{d_i}{D_j}\right)^2 \big(m_j(s_0) - 1\big) > m_i(s_0) = m_i(s_0 - 1) \ ,$$

Rearranging, we get

$$m_j(s_0 - 1) > \left(\frac{D_j}{d_i}\right)^2 m_i(s_0 - 1) \ .$$

from which it follows that

$$f_j\big(m_j(s_0 - 1)\big) \le \frac{D_j}{\sqrt{m_j(s_0 - 1)}} < \frac{d_i}{\sqrt{m_i(s_0 - 1)}} \le f_i\big(m_i(t_0 - 1)\big) \ .$$

This last inequality implies that at time s_0 the algorithm should have chosen the index i and not the index j, which is a contradiction. □

Combining the last two lemmas we can prove the following result.

Lemma 5. *Consider Algorithm 5 and the notation defined therein. Then, for any $s \ge 5$ it holds that*

$$\max_{k \in [5]} f_k\big(m_k(s)\big) \le \frac{D}{\sqrt{s - 5}}$$

where $D = \sqrt{5} \cdot \big(\max_{j \in [5]} D_j\big) \cdot \big(\max_{k \in [5]} \frac{D_k}{d_k}\big)$.

Proof. Let $s \ge 5$. Pick $j \in [5]$ such that $m_j(s) \ge \lceil s/5 \rceil$ (which does exist by Lemma 3). Then, by Lemma 4

$$\begin{aligned}\max_{k \in [5]} f_k\big(m_k(s)\big) &\le \max_{k \in [5]} \frac{D_k}{\sqrt{m_k(s)}} \le \max_{k \in [5]} \frac{D_k}{\sqrt{\left(\frac{d_k}{D_j}\right)^2 \big(m_j(s) - 1\big)}} \\ &= D_j \cdot \max_{k \in [5]} \left(\frac{D_k}{d_k}\right) \frac{1}{\sqrt{m_j(s) - 1}} \le D_j \cdot \max_{k \in [5]} \left(\frac{D_k}{d_k}\right) \frac{1}{\sqrt{\lceil s/5 \rceil - 1}} \le \frac{D}{\sqrt{s - 5}} \ .\end{aligned}$$ □

We are now ready for the proof of Lemma 2.

Proof of Lemma 2. It is enough to notice that Algorithm 2 with confidence parameter $\delta \in (0,1)$ relies, inside each epoch, on the same routine given by Algorithm 5 with $d_1 = l \cdot \sqrt{\frac{\log(2/\delta)}{2}}, d_2 = c \cdot \sqrt{\frac{\log(2/\delta)}{2}}, d_3 = r \cdot \sqrt{\frac{\log(2/\delta)}{2}}, d_4 = \lambda \cdot (c - l) \cdot \sqrt{\frac{\log(2/\delta)}{2}}, d_5 = 2 \cdot \lambda \cdot (c - l)$, with the convention that l correspond to 1, c corresponds to 2, r corresponds to 3, (l, c) corresponds to 4 and (c, r) corresponds to 5, the correspondence between times is given by $s = t - t_{\tau_t - 1}$, and, for each $s \in \{0, 1, 2, \dots\}$, $m_1(s) = n_{s + t_{\tau_t - 1}}(l)$, $m_2(s) = n_{s + t_{\tau_t - 1}}(c)$, $m_3(s) = n_{s + t_{\tau_t - 1}}(r)$, $m_4(s) = \sum_{i \le s + t_{\tau_t - 1}} \mathbf{1}\big(x_i \in (l, c)\big)$, $m_5(s) = \sum_{i \le s + t_{\tau_t - 1}} \mathbf{1}\big(x_i \in (c, r)\big)$. With these conventions, in Lemma 5 we have that $D \le 9 \cdot \sqrt{5} \cdot \left(\sqrt{2 \log(2/\delta)} + 4\right)$ and, for example (the other cases can be proved analogously)

$$\begin{aligned}\big|J_t(l_{\tau_t}, r_{\tau_t})\big| &\le 2 \cdot (\Gamma_t(r) + \Gamma_t(l) + \Gamma_t(l, c) + \Gamma_t(c, r)) \le 2 \cdot 4 \cdot \max_{k \in [5]} f_k\big(m_k(s)\big) \\ &\le 8 \cdot \frac{D}{\sqrt{t - t_{\tau_t - 1} - 5}} \le \frac{\tilde{c}_\delta}{\sqrt{2}} \cdot \frac{1}{\sqrt{t - t_{\tau_t - 1} - 5}} \le \frac{\tilde{c}_\delta}{\sqrt{t - t_{\tau_t - 1}}}\end{aligned}$$

where in the last inequality we used the fact that $t - t_{\tau_t - 1} \ge 10$. □

Lemma 1 and Lemma 2 allow us to prove Theorem 4, which closely follows the proof given in (Bachoc et al., 2022b).

Proof of Theorem 4. Define τ_T as the last epoch, $t_0 = 0$ and (if not already defined) $t_{\tau_T} = T$.

Due to Lemma 1, we may (and do!) assume that for each $t \in \{1, \dots, T\}$ it holds

$$\big(\boldsymbol{U}(c_{\tau_t}) - \boldsymbol{U}(l_{\tau_t}) \in J_t(l_{\tau_t}, c_{\tau_t})\big) \wedge \big(\boldsymbol{U}(r_{\tau_t}) - \boldsymbol{U}(c_{\tau_t}) \in J_t(c_{\tau_t}, l_{\tau_t})\big) \wedge \big(\boldsymbol{U}(r_{\tau_t}) - \boldsymbol{U}(l_{\tau_t}) \in J_t(r_{\tau_t}, l_{\tau_t})\big) \ .$$

This is because, given our choice $\delta = \frac{1}{T^{5/2}}$, assuming these conditions costs us in the expected regret a further additive term which is no greater than $T \cdot \tilde{C} \cdot T^2 \cdot \delta = \tilde{C} \cdot \sqrt{T}$.

Under these assumptions, notice that for each $\tau \in [\tau_T]$ we have that $x^\star \in I_\tau$. In fact, if the confidence intervals are guaranteed to contain the corresponding differences in the expected welfare, every time Algorithm 2 shrinks the active interval is because all the discarded points are guaranteed to be suboptimal.

For each epoch $\tau \in \{1, \dots, \tau_T\}$, define

$$B_\tau \coloneqq (t_\tau - 1) - t_{\tau-1} \ .$$

Now, for each epoch $\tau \in \{1, \dots, \tau_T\}$ if $B_\tau \geq \tilde{n}$, then

$$\max_{x \in [l_\tau, r_\tau]} \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\big) \leq 2 \cdot \tilde{c}_\delta \cdot \sqrt{\frac{1}{B_\tau}} \ .$$

In fact, assume that $x^\star > r_\tau$ (the other cases have similar proofs). Then, leveraging concavity, and recalling that $\inf\big(J_{t_\tau - 1}(l_\tau, r_\tau)\big) < 0$ and that $x^\star \in I_\tau$ (which implies $\frac{x^\star - l_\tau}{r_\tau - l_\tau} \leq 2$), we have

$$\begin{aligned}
\max_{x \in [l_\tau, r_\tau]} \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\big) &= \boldsymbol{U}(x^\star) - \boldsymbol{U}(l_\tau) = \frac{\boldsymbol{U}(x^\star) - \boldsymbol{U}(r_\tau)}{x^\star - r_\tau}(x^\star - r_\tau) + \boldsymbol{U}(r_\tau) - \boldsymbol{U}(l_\tau) \\
&\leq \frac{\boldsymbol{U}(r_\tau) - \boldsymbol{U}(l_\tau)}{r_\tau - l_\tau}(x^\star - r_\tau) + \boldsymbol{U}(r_\tau) - \boldsymbol{U}(l_\tau) = \frac{x^\star - l_\tau}{r_\tau - l_\tau} \cdot \big(\boldsymbol{U}(r_\tau) - \boldsymbol{U}(l_\tau)\big) \\
&\leq 2 \cdot \big(\boldsymbol{U}(r_\tau) - \boldsymbol{U}(l_\tau)\big) \leq 2 \cdot \sup(J_{t_\tau - 1}(l_\tau, r_\tau)) \leq 2 \cdot |J_{t_\tau - 1}(l_\tau, r_\tau)| \\
&\leq 2 \cdot \tilde{c}_\delta \cdot \sqrt{\frac{1}{B_\tau}} \ ,
\end{aligned}$$

where the final inequality follows by Lemma 2.

Let $\tau^\star$ be the first epoch from which it holds $x^\star \in [l_\tau, r_\tau]$. If $\tau^\star \geq 2$, then for each $\tau \in \{2, \dots, \tau^\star - 1\}$ it holds that

$$\max_{x \in [l_\tau, r_\tau]} \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\big) \leq \frac{3}{4} \cdot \max_{x \in [l_{\tau-1}, r_{\tau-1}]} \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\big) \ .$$

In fact, either for all $\tau \in \{1, \dots, \tau^\star - 1\}$ it holds that $r_\tau < x^\star$, or for all $\tau \in \{1, \dots, \tau^\star - 1\}$ it holds that $l_\tau > x^\star$. In the first case, for all $\tau \in \{1, \dots, \tau^\star - 1\}$, leveraging concavity and recalling that $x^\star \in I_\tau$ (which implies $\frac{x^\star - l_\tau}{x^\star - l_{\tau-1}} \leq \frac{3}{4}$), we have

$$\begin{aligned}
\max_{x \in [l_\tau, r_\tau]} \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\big) &= \boldsymbol{U}(x^\star) - \boldsymbol{U}(l_\tau) = \frac{\boldsymbol{U}(x^\star) - \boldsymbol{U}(l_\tau)}{x^\star - l_\tau} \cdot (x^\star - l_\tau) \leq \frac{\boldsymbol{U}(x^\star) - \boldsymbol{U}(l_{\tau-1})}{x^\star - l_{\tau-1}} \cdot (x^\star - l_\tau) \\
&\leq \frac{3}{4} \cdot \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(l_{\tau-1})\big) = \frac{3}{4} \cdot \max_{x \in [l_{\tau-1}, r_{\tau-1}]} \big(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\big) \ ,
\end{aligned}$$

while the second case can be deduced analogously.

For each $m \in \mathbb{N}$, let $A_m \coloneqq \big\{x \in (0,1) : \exists k \in \{1, \dots, 2^m - 1\}, x = k/2^m\big\}$ be the dyadic mesh in $(0,1)$ of index m. For any epoch $\tau \in \mathbb{N}$, let $m_\tau \coloneqq -\log_2(c_\tau - l_\tau)$ be the index of the dyadic mesh in $(0,1)$ at epoch τ of Algorithm 2 (note that $m_\tau \geq 2$ for all $\tau \in \mathbb{N}$ because Algorithm 2 begins with a step-size of $1/4$).

Let $m^\star := \min\bigl\{m \in \mathbb{N} : \#\bigl(A_m \cap (0, x^\star]\bigr) \geq 4 \text{ and } \#\bigl(A_m \cap [x^\star, 1)\bigr) \geq 4\bigr\}$ be the smallest index of the dyadic mesh in $(0,1)$ such that there are at least 4 points of the dyadic mesh in $(0,1)$ to the right and to the left of $x^\star$. For each $m \geq m^\star$ let $x_1^m < x_2^m < x_3^m < x_4^m \leq x^\star$ be the four points of $A_m \cap (0, x^\star]$ closest to $x^\star$ and $x^\star \leq x_5^m < x_6^m < x_7^m < x_8^m$ be the four points of $A_m \cap [x^\star, 1)$ closest to $x^\star$. Observe that, for all epochs $\tau \geq \tau^\star + 3$, Algorithm 2 selects policies only in the closed interval $[x_1^{m_\tau}, x_8^{m_\tau}]$. Observe further that, for each $m \geq m^\star + 1$, it holds

$$\max_{x \in [x_1^m, x_8^m]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) \leq \frac{4}{7} \cdot \max_{x \in [x_1^{m-1}, x_8^{m-1}]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) .$$

In fact, either $\max_{x \in [x_1^m, x_8^m]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) = \boldsymbol{U}(x^\star) - \boldsymbol{U}(x_1^m)$ or $\max_{x \in [x_1^m, x_8^m]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) = \boldsymbol{U}(x^\star) - \boldsymbol{U}(x_8^m)$. In the first case, leveraging concavity and observing that $\frac{x^\star - x_1^m}{x^\star - x_1^{m-1}} \leq \frac{4}{7}$, we have

$$\begin{aligned}\max_{x \in [x_1^m, x_8^m]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) &= \boldsymbol{U}(x^\star) - \boldsymbol{U}(x_1^m) = \frac{\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_1^m)}{x^\star - x_1^m} \cdot (x^\star - x_1^m) \leq \frac{\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_1^{m-1})}{x^\star - x_1^{m-1}} \cdot (x^\star - x_1^m) \\ &\leq \frac{4}{7} \cdot \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_1^{m-1})\bigr) \leq \frac{4}{7} \cdot \max_{x \in [x_1^{m-1}, x_8^{m-1}]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) .\end{aligned}$$

The second case can be worked out similarly.

Define $\tau^\# := \bigl\lfloor 4 + 2\log_{4/3}(\sqrt{T})\bigr\rfloor$ so that

$$\left(\frac{3}{4}\right)^{\left\lfloor \frac{\tau^\# - 1}{2} \right\rfloor} = \left(\frac{3}{4}\right)^{\left\lfloor \frac{\left\lfloor 4 + 2\log_{4/3}(\sqrt{T})\right\rfloor - 1}{2} \right\rfloor} \leq \left(\frac{3}{4}\right)^{\log_{4/3}(\sqrt{T})} = \frac{1}{\sqrt{T}} .$$

Assume that $\tau^\# < \tau^\star$ and $\tau^\star + 2 + \tau^\# < \tau_T$ (the other cases can be treated analogously, omitting terms which are not there anymore). Then, the expected regret can be decomposed as follows:

$$\begin{aligned}\sum_{t=1}^{T} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr) &= \sum_{\tau=1}^{\tau^\#} \sum_{t=t_{\tau-1}+1}^{t_\tau} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr) + \sum_{\tau=\tau^\#+1}^{\tau^\star - 1} \sum_{t=t_{\tau-1}+1}^{t_\tau} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr) \\ &+ \sum_{\tau=\tau^\star}^{\tau^\star+2} \sum_{t=t_{\tau-1}+1}^{t_\tau} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr) + \sum_{\tau=\tau^\star+3}^{\tau^\star+2+\tau^\#} \sum_{t=t_{\tau-1}+1}^{t_\tau} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr) + \sum_{\tau=\tau^\star+3+\tau^\#}^{\tau_T} \sum_{t=t_{\tau-1}+1}^{t_\tau} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr).\end{aligned}$$

We analyze these five terms individually.

For the first one, we further split the sum into two terms, depending on whether or not $B_\tau := t_\tau - 1 - t_{\tau-1} \geq \tilde{n}$. Recalling that for each $\tau \in \{1, \dots, \tau_T\}$ and for each $t \in \{t_{\tau-1}+1, \dots, t_\tau\}$ Algorithm 2 selects the policy x_t in the closed interval $[l_\tau, r_\tau]$, we have that

$$\begin{aligned}\sum_{\substack{\tau=1 \\ B_\tau \geq \tilde{n}}}^{\tau^\#} \sum_{t=t_{\tau-1}+1}^{t_\tau} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x_t)\bigr) &\leq \sum_{\substack{\tau=1 \\ B_\tau \geq \tilde{n}}}^{\tau^\#} (B_\tau + 1) \cdot \max_{x \in [l_\tau, r_\tau]} \bigl(\boldsymbol{U}(x^\star) - \boldsymbol{U}(x)\bigr) \\ &\leq \sum_{\substack{\tau=1 \\ B_\tau \geq \tilde{n}}}^{\tau^\#} (B_\tau + 1) \cdot 2 \cdot \tilde{c}_\delta \cdot \sqrt{\frac{\log(2/\delta)}{B_\tau}} \\ &\leq 4 \cdot \tilde{c}_\delta \cdot \sum_{\substack{\tau=1 \\ B_\tau \geq \tilde{n}}}^{\tau^\#} \sqrt{B_\tau} \leq 4 \cdot \tilde{c}_\delta \cdot \tau^\# \cdot \sqrt{T}.\end{aligned}$$

On the other hand, we also have that

$$\sum_{\substack{\tau=1\\B_\tau\le(\tilde{n}-1)}}^{\tau^{\#}}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x_t)\big)\le(\tilde{n}-1)\sum_{\tau=0}^{\infty}\big({}^3\!/\!{}_4\big)^\tau=4\cdot(\tilde{n}-1).$$

Thus, the first term is upper bounded by $4\cdot\tilde{c}_\delta\cdot\tau^{\#}\cdot\sqrt{T}+4\cdot(\tilde{n}-1)$.

For the second term, leveraging the definition of $\tau^{\#}$, we obtain

$$\begin{aligned}\sum_{\tau=\tau^{\#}+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x_t)\big)&\le\sum_{\tau=\tau^{\#}+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big({}^3\!/\!{}_4\big)^{\tau-1}\le\big({}^3\!/\!{}_4\big)^{\tau^{\#}-1}\cdot\sum_{\tau=\tau^{\#}+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}1\\&\le\big({}^3\!/\!{}_4\big)^{\left\lfloor\frac{\tau^{\#}-1}{2}\right\rfloor}\cdot\sum_{\tau=\tau^{\#}+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}1\le\sqrt{T}.\end{aligned}$$

For the third term, we further split the sum into two terms, depending on whether or not $B_\tau\ge\tilde{n}$. Proceeding exactly as for the first term, we obtain

$$\sum_{\tau=\tau^\star}^{\tau^\star+2}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x_t)\big)\le3\cdot4\cdot\tilde{c}_\delta\cdot\sqrt{T}+3\cdot(\tilde{n}-1).$$

For the fourth term, we split again the sum into two terms, depending on whether or not $B_\tau\ge\tilde{n}$. If $B_\tau\ge\tilde{n}$, proceeding exactly as for the corresponding part of the first term, we obtain

$$\sum_{\substack{\tau=\tau^\star+3\\B_\tau\ge\tilde{n}}}^{\tau^\star+2+\tau^{\#}}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x_t)\big)\le4\cdot\tilde{c}_\delta\cdot\tau^{\#}\cdot\sqrt{T}.$$

Instead, if $B_\tau\le(\tilde{n}-1)$, we get

$$\begin{aligned}\sum_{\substack{\tau=\tau^\star+3\\B_\tau\le(\tilde{n}-1)}}^{\tau^\star+2+\tau^{\#}}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x_t)\big)&\le(\tilde{n}-1)\cdot\sum_{\substack{\tau=\tau^\star+3\\B_\tau\le(\tilde{n}-1)}}^{\tau^\star+2+\tau^{\#}}\max_{x\in[l_\tau,r_\tau]}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x)\big)\\&\le(\tilde{n}-1)\cdot\sum_{\substack{\tau=\tau^\star+3\\B_\tau\le(\tilde{n}-1)}}^{\tau^\star+2+\tau^{\#}}\max_{x\in\left[x_1^{m_\tau},x_8^{m_\tau}\right]}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x)\big)\le2\cdot(\tilde{n}-1)\cdot\sum_{\tau=0}^{\infty}\big({}^4\!/\!{}_7\big)^\tau\\&\le\frac{14}{3}\cdot(\tilde{n}-1)\,.\end{aligned}$$

For the last term, we have

$$\begin{aligned}\sum_{\tau=\tau^\star+3+\tau^{\#}}^{\tau_T}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x_t)\big)&\le\sum_{\tau=\tau^\star+3+\tau^{\#}}^{\tau_T}\sum_{t=t_{\tau-1}+1}^{t_\tau}\max_{x\in\left[x_1^{m_\tau},x_8^{m_\tau}\right]}\big(\boldsymbol{U}(x^\star)-\boldsymbol{U}(x)\big)\\&\le\sum_{\tau=\tau^\star+3+\tau^{\#}}^{\tau_T}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big({}^4\!/\!{}_7\big)^{\left\lfloor\frac{\tau-(\tau^\star+3)-1}{2}\right\rfloor}\\&\le\big({}^3\!/\!{}_4\big)^{\left\lfloor\frac{\tau^{\#}-1}{2}\right\rfloor}\sum_{\tau=\tau^\star+3+\tau^{\#}}^{\tau_T}\sum_{t=t_{\tau-1}+1}^{t_\tau}1\le\sqrt{T}\,.\end{aligned}$$

Putting everything together, and recalling the definition of $\tau^{\#}$, the conclusion follows. □