Adaptive maximization of social welfare

(Preliminary draft)

Nicolò Cesa-Bianchi* Roberto Colomboni† Maximilian Kasy‡

September 30, 2022

Abstract

We consider the problem of repeatedly choosing policy parameters, such as tax or transfer rates, in order to maximize social welfare, defined as the 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 $T^{2/3}$. This implies that (i) the social welfare maximization problem is harder than the multi-armed bandit problem (with a rate of $T^{1/2}$), and (ii) that our proposed algorithm achieves the optimal rate.

Simulations confirm these results, as well as the viability of the proposed algorithm. 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). We lastly discuss extensions to nonlinear income taxation, and to commodity taxation.

1 Introduction

Consider a policymaker who aims to maximize social welfare, defined as a weighted sum of utility. They 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) with insights from machine learning (in particular online learning and multi-armed bandits).

Optimal taxes and multi-armed bandits 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 (Chetty, 2009). Optimal tax problems are defined by normative parameters (such as welfare weights for different

*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."

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 estimates 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.

Such an approach contrasts with the adaptive approach characterizing 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 the currently available information for optimal policy decisions for the present moment. The goal of bandit algorithms is to maximize a stream of rewards, which requires an optimal balance between exploration and exploitation. Most bandit algorithms are characterized by optimism in the face of uncertainty: policies with uncertain payoff should be tried, even when their estimated payoff is not optimal.

Bandit algorithms (and similarly, adaptive experimental designs for informing policy choice, as in 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. In this paper, we consider the problem of adaptive policy choice for the objective of maximizing the stream of social welfare based on unobserved utility. We focus on the problem of choosing a tax rate or insurance rate, in a canonical and minimal model in the tradition of Ramsey (1927); Mirrlees (1971); Baily (1978); Saez (2001). The tradeoff in our model, interpreted as a model of taxation, is between, first, raising public revenue (for 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. The policymaker needs to learn the magnitude of the response of the tax base to the tax rate. In our setting, the policymaker can repeatedly choose the tax rate, observe responses, update their beliefs, and adjust the tax rate again.

While utility is unobserved, it can be indirectly inferred by integrating response functions over counterfactual policy choices. Absent income effects, in particular, utility is equivalent to consumer surplus, which is given by an integral of demand (or supply) over different prices or tax rates. A key feature of our model, and of social welfare maximization problems more generally, is therefore that welfare at a particular tax rate depends on behavioral responses at *other* tax rates. This dependence on counterfactuals contrasts with multi-armed bandits, where the reward for a particular arm (policy choice) depends only on the observed outcomes for that arm. It is this dependence on outcomes for counterfactual policies which makes social welfare maximization harder than bandit problems.

Regret bounds Our main results 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 proceeds by constructing a family of possible distributions for preferences. This family is such that, in order to learn which of two possible policies is optimal in terms of welfare, we need to learn behavioral responses for intermediate policies, which are strictly suboptimal. This is because the difference in welfare between the two policies depends on

the integral of demand over intermediate policy values. Because of the need to probe these suboptimal arms 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 is worse than the worst-case rate for bandits of $T^{1/2}$.

We next propose an algorithm for the adaptive maximization of social welfare. Our algorithm is a modification of the well-known Exp3 algorithm. Exp3 is based on an unbiased estimate of cumulative welfare for each arm. The probability of choosing a given arm is then 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 that are clearly sub-optimal, in order to learn welfare for the policies which are contenders for the optimum. This need for additional exploration 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 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), again.

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.

Further literature Regret minimization approaches have been applied to a number of economic and financial scenarios in the literature. This includes 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). Our minimal model of optimal taxation has been discussed in a static setting in Kasy (2018).

Roadmap The rest of this paper proceeds as follows. 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 provides simulation evidence. Section 5 discusses extensions of our baseline model, including non-linear income taxation (cf. Mirrlees, 1971) and commodity taxation (cf. Ramsey, 1927). All proofs can be found in Appendix A. Additional simulation results can be found in the supplementary appendix.

2 Setup

At each time $i = 1, 2, \ldots, T$, one agent arrives who is characterized by an unknown willingness to pay $v_i \in [0, 1]$. This agent 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

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

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 the values of v_i for $j = 1, \ldots, i$. In particular,

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

The last expression defines $\mathbb{U}_i$ 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, taking an expectation 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

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

The expectation in this expression is taken over any possible randomness in the tax rates x_i. We will consider worst case cumulative expected regret, which is attained by the sequence $\{v_i\}_{i=1}^T$ which maximizes regret for any given algorithm. We will derive a lower bound for worst-case cumulative expected regret, which holds for any algorithm, and a corresponding upper bound for a given algorithm (to be defined below) which achieves the lower bound, up to a logarithmic term.

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

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

Expected welfare for this distribution of v_i is given by

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

The expectation in this expression is taken over the distribution of v_i, which is presumed to be independent of the tax rate x. 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. We will again consider worst case cumulative expected regret, which is now attained by the demand function $\boldsymbol{G}(\cdot)$ which maximizes regret for any given algorithm and time horizon T.

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:

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

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, the seller is willing to trade at prices above v_i^s.

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

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,

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

Table 1: Properties of different learning problems

Model	Pointwise	One-sided Lipschitz
Monopoly price setting	Yes	Yes
Optimal tax	No	Yes
Bilateral trade	No	No

Notes: See Subsection 2.2 for a formal justification of this table.

Table 2: Comparison of minimax regret rates

Model	Discrete	Continuous
Monopoly price setting	$T^{1/2}$	$T^{2/3}$
Optimal tax	$T^{2/3}$	$T^{2/3}$
Bilateral trade	$T^{2/3}$	T

Notes: Rates are up to logarithmic terms, and apply to both the stochastic and the adversarial setting. 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 tax case are proven in this paper.

Lipschitzness and information requirements The difficulty of the learning problem in each of these models critically depends on the Lipschitz properties of the welfare function, and on the information requirements needed to evaluate it. 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 also 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 cover of the set of feasible policies. Pointwise information allows us to avoid exploring policies that are clearly suboptimal, when we aim to learn the optimal policy.

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

1. 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).

2. For **monopoly pricing**, welfare $U_i^{\mathrm{MP}}(x)$ is one-sided Lipschitz and only depends on $G_i(x)$ 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 2. 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 2, 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.

3 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 Tempered Exp3 Algorithm 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 are relegated to Appendix A.

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 A. 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.

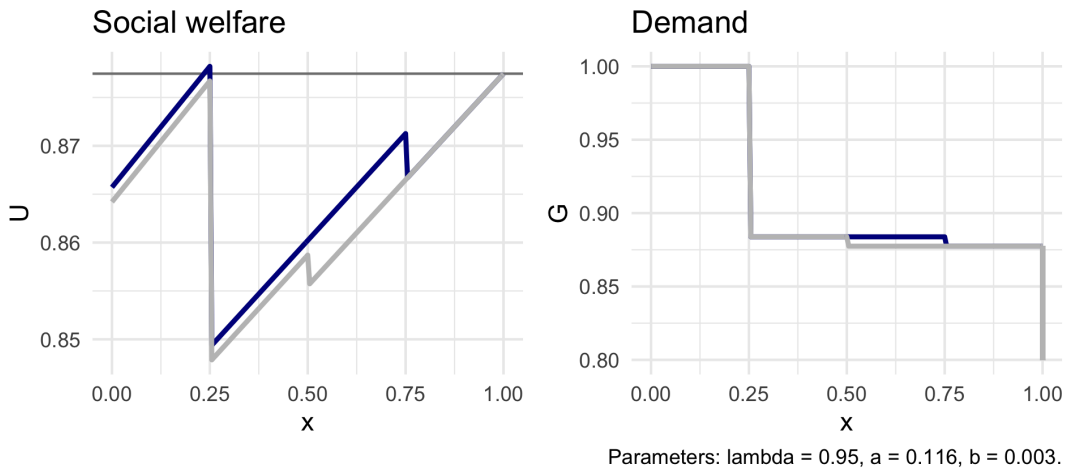
To prove the stochastic lower bound we construct a family of distributions for v_i that is indexed by a parameter $\epsilon \in [-1, 1]$. The distributions in this family have four points of support, $(1/4, 1/2, 3/4, 1)$. The probability of these points is given by

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

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

[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 $1 - a - 2 \cdot b > 0$.

Figure 1: Construction for proving the lower bound on regret



Notes: This figure illustrates our construction for proving the lower bound on cumulative 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 $[.5, .75]$.

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 at least $\Omega(T^{2/3})$ regret along the way.

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.

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 policy choices, we potentially need to explore policies

Algorithm 1 Tempered Exp3 for social welfare

Require: Tuning parameters K, γ and η.

Calculate evenly spaced grid-points $\tilde{x}_k = (k-1)/K$, and initialize $\widehat{\mathbb{G}}_{1k} = 0$ for $k = 1, \ldots, K+1$.

for individual $i = 1, 2, \ldots, T$ **do**

For all $k = 1, 2, \ldots, K+1$, set

$$\widehat{\mathbb{U}}_{ik} = \tilde{x}_k \cdot \widehat{\mathbb{G}}_{ik} + \tfrac{\lambda}{K} \cdot \sum_{k'>k} \widehat{\mathbb{G}}_{ik'}. \tag{7}$$

For all $k = 1, 2, \ldots, K+1$, set

$$p_{ik} = (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_{ik})}{\sum_{k'} \exp(\eta \cdot \widehat{\mathbb{U}}_{ik'})} + \frac{\gamma}{K+1}. \tag{8}$$

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

$$\widehat{\mathbb{G}}_{i+1,k} = \widehat{\mathbb{G}}_{i,k} + y_i \cdot \frac{\mathbf{1}(k_i = k)}{p_{ik}}. \tag{9}$$

end for

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)**.** *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

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

for some constants c_1, c_2, c_3*. Then expected regret* $\mathcal{R}_T(\{v_i\}_{i=1}^T)$ *is bounded above by*

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

for some constant c_4*.*

Corollary 1 (Stochastic upper bound on regret of tempered Exp3)**.** *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 A.

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 (10), 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 $6 \cdot \log(T)^{1/3} T^{2/3}$.

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. The analysis of this anytime version of Algorithm 1 requires one to deal with technical nuances which we omit for clarity of presentation. We however discuss some simulations with time-varying tuning parameters η and γ in Algorithm 1 in Section 4 below.

4 Simulations

We next consider a series of simulations, to check and verify our theoretical predictions for Algorithm 1, the Tempered Exp3 Algorithm for Social Welfare. Throughout these simulations we consider the stochastic case, with $v_i \sim^{iid} \text{Beta}(\alpha, \beta)$. As a first example, we consider $\alpha = \beta = 1$, which implies $v_i \sim U[0,1]$. For this case, we can derive demand and social welfare in closed form:

$$\begin{aligned}\boldsymbol{G}(x) &= P(v \geq x) = 1 - x \\ \boldsymbol{U}(x) &= x \cdot \boldsymbol{G}(x) + \lambda \cdot \int_x^1 \boldsymbol{G}(x')dx' = x - x^2 + \lambda \cdot \left[\tfrac{1}{2} - x + \tfrac{1}{2}x^2\right] \\ &= \lambda/2 + x \cdot (1-\lambda) + x^2 \cdot (\lambda/2 - 1)\,.\end{aligned}$$

Figure 2 shows the simulation results for this case, with $\lambda = .7$. The figure on the bottom right shows true expected social welfare, $\boldsymbol{U}(x)$, as a function of the policy choice x.

The figure on top shows average cumulative regret over time. Our theoretical characterizations in Theorem 2 imply that a worst-case bound (over possible distributions for v_i), with tuning parameters chosen optimally, converges at a rate of $T^{-1/3}$, up to logarithmic terms.

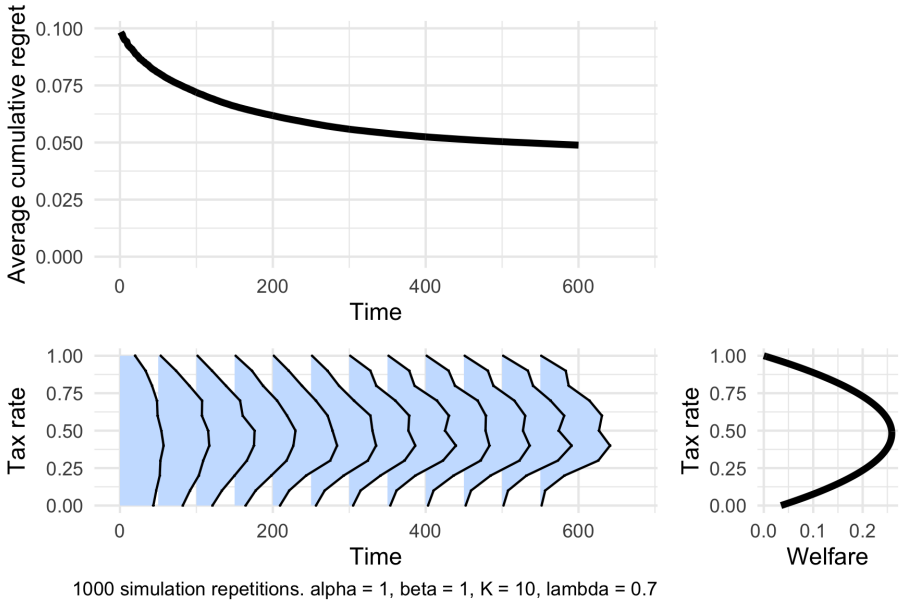
Note that this is different from the scenario here, where we plot cumulative regret for a fixed distribution of v_i, and fixed tuning parameters. A fixed distribution of v_i suggests that we might converge faster than $T^{-1/3}$, since we are not in the worst-case scenario for each T. Fixed tuning parameters, on the other hand, imply a lower bound for cumulative average regret over time, because of (i) fixed discretization error (finite tuning parameter K), and (ii) a fixed expected regret due to the exploration term corresponding to $\gamma > 0$.

The figure on the bottom left, lastly, provides a more fine-grained representation of our simulation results. It shows, for each batch of 50 successive time-periods, the distribution across policy choices x made by the algorithm. This figure shows how the distribution across policy choices starts to mimick the shape of expected social welfare $\boldsymbol{U}(x)$ over time. This is as expected, given that our algorithm chooses these policy options with probability $p_{ik} = (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_{ik})}{\sum_{k'} \exp(\eta \cdot \widehat{\mathbb{U}}_{ik'})} + \frac{\gamma}{K+1}$.

Figure 3 presents similar plots, but for the case of time-dependent tuning parameters of the form $\eta_t = t^{-2/3} \cdot 10$ and $\gamma_t = t^{-1/3}/\sqrt{10}$. While our theoretical results do not explicitly cover such varying tuning parameters, they do improve performance in the stochastic case. Effectively, they help by "front-loading" exploration, thereby leading to more rapid improvements in performance in initial periods.

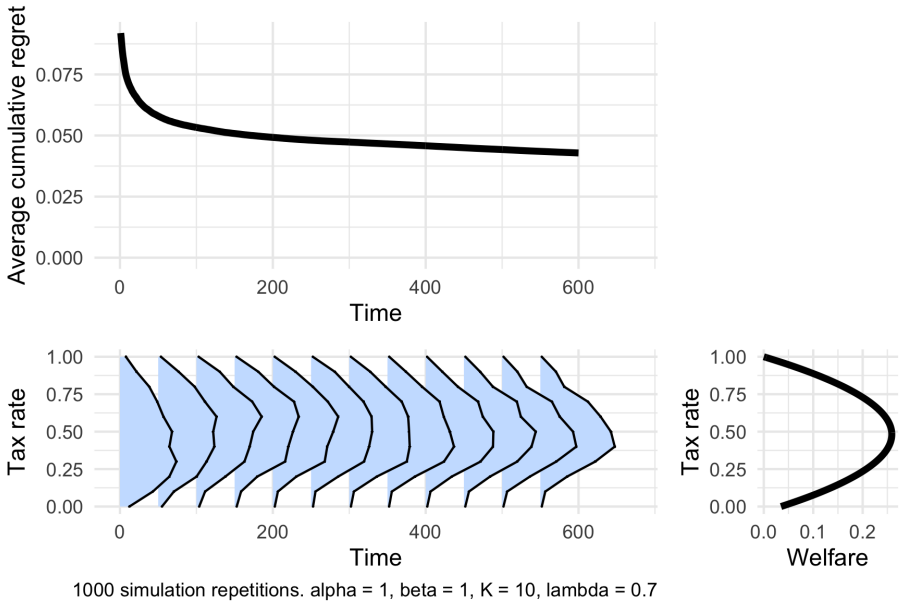
The online supplement presents a series of additional simulation results, for a wide range of parameters α, β, and different values for K. The behavior of our algorithm stays largely the same across all these different scenarios.

Figure 2: Average regret and distribution of policy choices



Notes: The top figure shows the average regret across 1,000 simulations of the tempered Exp3 algorithm for social welfare. Willingness to pay v_i is drawn from the uniform (i.e., Beta$(1, 1)$) distribution. Simulations for alternative distributions are shown in the supplementary appendix.

Figure 3: Time-dependent tuning parameters



Notes: This figure replicates Figure 2, but with time-dependent tuning parameters of the form $\eta_t = t^{-2/3} \cdot 10$ and $\gamma_t = t^{-1/3}/\sqrt{10}$.

5 Extensions and model variations

We next discuss two extensions of the baseline model of optimal taxation that we introduced in Section 2. These extensions incorporate some 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 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 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.

5.1 Income taxation

In this subsection, we generalize our baseline model of optimal taxation to a model of income taxation with heterogeneous wages w_i, welfare weights $\omega = \omega(w_i)$, extensive-margin labor supply responses determined by the cost of participation v_i, and (potentially progressive) income taxation $x_i = x(w_i)$.

Two simplifications are maintained in our model, relative to a fully 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. Both of these assumptions are empirically realistic, but not without loss of generality.

Setup At each time $i = 1, 2, \ldots, T$, one agent arrives who is characterized by (i) a potential wage $w_i \in \{1, \ldots, \bar{w}\}$, and (ii) an unknown cost of participation $v_i \in \{1, \ldots, \bar{w} + 1\}$. Discrete support is maintained here for simplicity of exposition. This agent 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 $x_i = x(w_i)$ on their earnings. They furthermore incur a non-monetary cost of participation v_i. Their optimal labor supply decision is therefore given by $y_i = \mathbf{1}(x_i \leq w_i - v_i)$, and agent private welfare equals $\max(w_i - v_i - x_i, 0)$. The implied public revenue is equal to the tax on earnings $x(w_i)$ if $y_i = 1$, and 0 otherwise, that is, $x_i \cdot y_i$.

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

$$U_i(x(\cdot)) = \underbrace{x(w_i) \cdot \mathbf{1}(x(w_i) \leq w_i - v_i)}_{\text{Public revenue}} \quad + \quad \omega(w_i) \cdot \underbrace{\max(w_i - v_i - x(w_i), 0)}_{\text{Private welfare}}. \tag{11}$$

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.[2] Denote now

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

[2]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 $x(w_i) - x_0$. 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 additively. It is therefore without loss of generality to omit x_0 from the model.

Algorithm 2 Tempered Exp3 for optimal income taxation

Require: Tuning parameters γ and η, and a set of policies $\mathscr{X} = \{x(\cdot)\}$.

1: Initialize $\widehat{\mathbb{G}}_1(w, x) = 0$ for all w, x.

2: **for** individual $i = 1, 2, \ldots, T$ **do**

3: For all $x(\cdot) \in \mathscr{X}$, set

$$\widehat{\mathbb{U}}_i(x(\cdot)) = \sum_{w=1}^{\bar{w}} \left[x(w) \cdot \widehat{\mathbb{G}}_i(w, x(w)) + \omega(w) \cdot \sum_{x'=x(w)+1}^{\bar{w}} \widehat{\mathbb{G}}_i(w, x') \right]. \tag{13}$$

4: For all $x(\cdot) \in \mathscr{X}$, set

$$p_i(x(\cdot)) = (1 - \gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_i(x(\cdot)))}{\sum_{x'(\cdot) \in \mathscr{X}} \exp(\eta \cdot \widehat{\mathbb{U}}_i(x'(\cdot)))} + \frac{\gamma}{|\mathscr{X}|}. \tag{14}$$

5: Choose $x_i(\cdot) \in \mathscr{X}$ at random according to the probability distribution p_i.

6: For all w, x, set

$$\widehat{\mathbb{G}}_{i+1}(w, x) = \widehat{\mathbb{G}}_i(w, x) + y_i \cdot \frac{\mathbf{1}(w_i = w, x_i(w_i) = x)}{p_i(x_i(w) = x)}, \tag{15}$$

where $p_i(x_i(w) = x)$ is the marginal probability implied by the distribution $p_i(x_i(\cdot))$.

7: **end for**

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 their wage w_i. With this notation, we can rewrite

$$\max(w - v_i - x, 0) = \int_x^{\bar{w}} G_i(w, x')dx' = \sum_{x'=x+1}^{\bar{w}} G_i(w, x').$$

The last equality holds because of the assumed discrete support of (w, v). It follows that

$$U_i(x(\cdot)) = \sum_{w=1}^{\bar{w}} \left[x(w) \cdot G_i(w, x(w)) + \omega(w) \cdot \sum_{x'=x(w)+1}^{\bar{w}} G_i(w, x') \right]. \tag{12}$$

Algorithm Algorithm 2 generalizes Algorithm 1 to this setting. As before, we form an unbiased estimate $\widehat{G}_i$ of G_i using inverse probability weighting, and cumulate across time periods to obtain $\widehat{\mathbb{G}}_i$. The inverse probability weighting estimator is based on the *marginal* distribution $p_i(x_i(w) = x)$ of $x_i(w)$ at $w = w_i$, rather than the distribution of the function $x_i(\cdot)$. This yields a more efficient unbiased estimate than would be obtained when weighting by $p_i(x(\cdot))$. Note also that w_i is observed whenever $y_i = 1$, so that the estimate $\widehat{\mathbb{G}}_i$ is in fact a function of observables.

Plugging this estimate into the cumulated version of Equation (12), defining $\mathbb{U}_i$, we obtain an unbiased estimate $\widehat{\mathbb{U}}_i$ of cumulative social welfare. Algorithm 2 chooses the policy x_i from a set of policies $\mathscr{X} = \{x(\cdot)\}$ that might be restricted. The distribution p_i over this set of policies is again given by the tempered Exp3 distribution, as in our baseline model.

5.2 Commodity taxation

In this subsection, 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). 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 agent arrives who is characterized by a utility function $u_i : [0,1]^k \to \mathbb{R}$. This agent 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 agent 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 agent decision is therefore given by

$$y_i = G_i(x_i) = \operatorname*{argmax}_{y \in [0,1]^k} \left[u_i(y) - \langle x_i + p, y \rangle \right]. \tag{16}$$

Defining v_0 as agent 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

$$U_i(x_i) = \underbrace{\langle x_i, y_i \rangle}_{\text{Public revenue}} + \lambda \cdot \underbrace{v_i(x_i)}_{\text{Private welfare}}. \tag{17}$$

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

By the envelope theorem (Milgrom and Segal, 2002),

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

Let $\mathscr{V}$ 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 indirect utility function v.

$$\Pi(G(\cdot)) = \operatorname*{argmin}_{v(\cdot) \in \mathscr{V}} \int_{[0,1]^k} \left\| \nabla_x v(x) - G(x) \right\|^2 dx \tag{18}$$

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.

Algorithm 3 Tempered Exp3 for commodity taxation

Require: Tuning parameters K, γ and η.

Calculate the set of evenly spaced grid-points $\mathbb{X} = [0, \frac{1}{K}, \ldots, 1]^k$ and initialize $\widehat{\mathbb{G}}_1(x) = 0$ for all grid points.

for individual $i = 1, 2, \ldots, T$ **do**

For all $x \in \mathbb{X}$, set

$$\widehat{\mathbb{U}}_i(x) = \langle x_i, \widehat{\mathbb{G}}_i \rangle + \lambda \cdot \widehat{v}_i(x_i). \tag{19}$$

For all $x \in \mathbb{X}$, set

$$p_i = (1 - \gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_i(x))}{\sum_{x'} \exp(\eta \cdot \widehat{\mathbb{U}}_i(x'))} + \frac{\gamma}{(K+1)^k}. \tag{20}$$

Choose x_i at random according to the probability distribution p_i, and query y_i accordingly.

For all $x \in \mathbb{X}$, set

$$\widetilde{\mathbb{G}}_{i+1}(x) = \widehat{\mathbb{G}}_i(x) + \frac{y_i}{p_i} \tag{21}$$

$$\widehat{v}_{i+1}(x) = \Pi(\widetilde{\mathbb{G}}_{i+1}) \tag{22}$$

$$\widehat{\mathbb{G}}_{i+1} = \nabla_x \widehat{v}_{i+1}. \tag{23}$$

end for

References

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.

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

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 Proofs

A.1 Theorem 1

Proof of Theorem 1.
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

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

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. Let $v_1, v_2, \cdots \in [0,1]$ be the sequence of agent valuations. For each $\epsilon \in [-1,1]$, consider a distribution P^ϵ such that the agent valuations $v_1, v_2, \ldots$ form a P^ϵ-i.i.d. sequence (independent of the randomization used by the algorithm) with common distribution μ^ϵ. Define

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

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

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

where

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

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

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:

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

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

$$n_1 + n_2 + n_3 = T \tag{25}$$

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

$$\mathcal{R}_T(\boldsymbol{G}^\epsilon) = T\cdot\sup_{x\in[0,1]} \boldsymbol{U}^\epsilon(x) - \sum_{i=1}^{T} E^\epsilon\big(\boldsymbol{U}^\epsilon(x_i)\big)\,. \tag{26}$$

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

$$\max_{x\in(1/2,3/4]} \boldsymbol{U}^\epsilon(x) = \boldsymbol{U}^\epsilon(3/4)\,,\ \max_{x\in[0,1/2]} \boldsymbol{U}^\epsilon(x) = \boldsymbol{U}^\epsilon(1/4)\,,\ \max_{x\in(3/4,1]} \boldsymbol{U}^\epsilon(x) = \boldsymbol{U}^\epsilon(1)\,, \tag{27}$$

$$\text{and} \qquad \boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(1/4) = c_1\cdot\epsilon\,. \tag{28}$$

Further calculations show also that

$$\min_{\epsilon\in[-1,1]} \min\big(\boldsymbol{U}^\epsilon(1/4), \boldsymbol{U}^\epsilon(1)\big) = \boldsymbol{U}^1(1/4)\,,\ \max_{\epsilon\in[-1,1]} \max_{x\in(1/2,3/4]} \boldsymbol{U}^\epsilon(x) = \boldsymbol{U}^{-1}(3/4)\,, \tag{29}$$

$$\text{and} \qquad \boldsymbol{U}^1(1/4) - \boldsymbol{U}^{-1}(3/4) = c_2 \ . \tag{30}$$

Equations (27), (28), (29), and (30) imply that

$$\sup_{x\in[0,1]} \boldsymbol{U}^\epsilon(x) = \boldsymbol{U}^\epsilon(1) \ , \qquad \text{if } \epsilon \in [0,1] \ . \tag{31}$$

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

$$\begin{aligned}
\mathcal{R}_T(\boldsymbol{G}^\epsilon) &\overset{(26)}{=} T \cdot \sup_{x\in[0,1]} \boldsymbol{U}^\epsilon(x) - \sum_{i=1}^{T} E^\epsilon\big(\boldsymbol{U}^\epsilon(x_i)\big) \overset{(31)}{=} T \cdot \boldsymbol{U}^\epsilon(1) \\
&\quad - \sum_{i=1}^{T} E^\epsilon\Big(\boldsymbol{U}^\epsilon(x_i) \cdot \big(\mathbf{1}_{(1/2,3/4]}(x_i) + \mathbf{1}_{[0,1/2]}(x_i) + \mathbf{1}_{(3/4,1]}(x_i)\big)\Big) \\
&\overset{(27)}{\geq} T \cdot \boldsymbol{U}^\epsilon(1) - \sum_{i=1}^{T} E^\epsilon\Big(\boldsymbol{U}^\epsilon(3/4) \cdot \mathbf{1}_{(1/2,3/4]}(x_i) \\
&\quad + \boldsymbol{U}^\epsilon(1/2) \cdot \mathbf{1}_{[0,1/2]}(x_i) + \boldsymbol{U}^\epsilon(1) \cdot \mathbf{1}_{(3/4,1]}(x_i)\Big) \\
&\overset{(25)}{=} \big(\boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(3/4)\big) \cdot E^\epsilon(n_1) + \big(\boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(1/4)\big) \cdot E^\epsilon(n_2) \\
&\overset{(29)}{\geq} \big(\boldsymbol{U}^1(1/4) - \boldsymbol{U}^{-1}(3/4)\big) \cdot E^\epsilon(n_1) + \big(\boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(1/4)\big) \cdot E^\epsilon(n_2) \\
&\overset{(30)}{=} c_2 \cdot E^\epsilon(n_1) + \big(\boldsymbol{U}^\epsilon(1) - \boldsymbol{U}^\epsilon(1/4)\big) \cdot E^\epsilon(n_2) \\
&\overset{(28)}{=} c_2 \cdot E^\epsilon(n_1) + c_1 \cdot \epsilon \cdot E^\epsilon(n_2)
\end{aligned} \tag{32}$$

Notice that inequality (32) 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 (32) was proven, we can prove that, if $\epsilon \in [0,1]$,

$$\mathcal{R}_T(\boldsymbol{G}^{-\epsilon}) \geq c_2 \cdot E^{-\epsilon}(n_1) + c_1 \cdot \epsilon \cdot E^{-\epsilon}(n_3) \geq c_1 \cdot \epsilon \cdot E^{-\epsilon}(n_3) \ , \tag{33}$$

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 $\epsilon > 0$.

At high level, inequalities (32) and (33) tell us that, if $|\epsilon|$ is not negligible, the algorithm has to play a substantially different number of times in the region $(3/4, 1]$ 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 $(1/2, 3/4]$. However, as shown in (32), 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 $\epsilon \in [0,1]$,

$$E^{-\epsilon}(n_3) \geq E^\epsilon(n_3) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^\epsilon(n_1)} \ . \tag{34}$$

Now, if the algorithm is going to suffer low regret when $\epsilon > 0$, then by (32) 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 (34), 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 (33), 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.

To get a contradiction, suppose that

$$\forall \epsilon \in [-1,1] \qquad \mathcal{R}_T(\boldsymbol{G}^\epsilon) < C \cdot T^{2/3} \ . \tag{35}$$

It follows from (32) that, for each $\epsilon \in [0,1]$,

$$E^{\epsilon}(n_1) \overset{(32)}{\leq} \frac{\mathcal{R}_T(\boldsymbol{G}^{\epsilon})}{c_2} \overset{(35)}{\leq} \frac{C}{c_2} \cdot T^{2/3} \,, \qquad E^{\epsilon}(n_2) \overset{(32)}{\leq} \frac{\mathcal{R}_T(\boldsymbol{G}^{\epsilon})}{c_1 \cdot \epsilon} \overset{(35)}{\leq} \frac{C}{c_1 \cdot \epsilon} \cdot T^{2/3} \,. \tag{36}$$

This implies, relying also on (33) and (34), that for each $\epsilon \in [0,1]$ we have

$$\begin{aligned}
\mathcal{R}_T(\boldsymbol{G}^{-\epsilon}) &\overset{(33)}{\geq} c_1 \cdot \epsilon \cdot E^{-\epsilon}(n_3) \overset{(34)}{\geq} c_1 \cdot \epsilon \cdot \left(E^{\epsilon}(n_3) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^{\epsilon}(n_1)}\right) \\
&\overset{(25)}{=} c_1 \cdot \epsilon \cdot \left(T - E^{\epsilon}(n_1) - E^{\epsilon}(n_2) - c_3 \cdot \epsilon \cdot T \cdot \sqrt{E^{\epsilon}(n_1)}\right) \\
&\overset{(36)}{\geq} c_1 \cdot \epsilon \cdot \left(T - \frac{C}{c_2} \cdot T^{2/3} - \frac{C}{c_1 \cdot \epsilon} \cdot T^{2/3} - c_3 \cdot \epsilon \cdot T \cdot \sqrt{\frac{C}{c_2} \cdot T^{2/3}}\right) \\
&= c_1 \cdot \epsilon \cdot \left(1 - \frac{C}{c_2} \cdot T^{-1/3} - \frac{C}{c_1 \cdot \epsilon} \cdot T^{-1/3} - c_3 \cdot \epsilon \cdot T^{1/3} \cdot \sqrt{\frac{C}{c_2}}\right) \cdot T \,.
\end{aligned} \tag{37}$$

Pick $\epsilon \coloneqq T^{-1/3} \cdot \sqrt{\frac{\sqrt{C \cdot c_2}}{c_1 \cdot c_3}}$. First, note that since $0 < C \overset{(24)}{\leq} \frac{c_1^2 \cdot c_3^2}{c_2}$ we have that $\epsilon \in (0,1]$. Plugging this value of ϵ in (37) leads to

$$\begin{aligned}
C \cdot T^{2/3} &\overset{(35)}{>} \mathcal{R}_T(\boldsymbol{G}^{-\epsilon}) \\
&\overset{(37)}{\geq} \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - \frac{C}{c_2} \cdot T^{-1/3} - 2 \cdot \sqrt{\frac{c_3}{c_1 \cdot \sqrt{c_2}}} \cdot C^{3/4}\right) \cdot T^{2/3} \\
&\overset{(24)}{\geq} \frac{1}{2} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - 4 \cdot \sqrt{\frac{c_3}{c_1 \cdot \sqrt{c_2}}} \cdot C^{3/4}\right) \cdot T^{2/3} \\
&\overset{(24)}{\geq} \frac{1}{4} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot T^{2/3} \,,
\end{aligned} \tag{38}$$

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 (38) leads to the contradiction

$$C \overset{(38)}{>} \left(\frac{1}{4} \cdot \sqrt{\frac{c_1 \cdot \sqrt{c_2}}{c_3}}\right)^{4/3} = \frac{1}{8} \cdot \sqrt[3]{\frac{2 \cdot c_1^2 \cdot c_2}{c_3^2}} > \frac{1}{16} \cdot \sqrt[3]{\frac{c_1^2 \cdot c_2}{c_3^2}} \overset{(24)}{\geq} C \,.$$

Since (35) 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. □

A.2 Claim (34)

Proof of the claim (34).
Let $w_1, w_2, \cdots \in [0,1]$ 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,

$$P^{\epsilon}_{(w_i)_{i \in \mathbb{N}}} = P^{-\epsilon}_{(w_i)_{i \in \mathbb{N}}} \,, \qquad \forall \epsilon \in [0,1] \,. \tag{39}$$

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.,

$$P^\epsilon_{\tilde{x}_i} = P^\epsilon_{x_i}\,. \tag{40}$$

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 $j \in \{1,2,3\}$,

$$E^\epsilon(\tilde{n}_j) = \sum_{i=1}^{T} P^\epsilon_{x_i}\big((1/2, 3/4]\big) \overset{(40)}{=} \sum_{i=1}^{T} P^\epsilon_{\tilde{x}_i}\big((1/2, 3/4]\big) = E^\epsilon(n_j)\,,$$

to prove the claim (34), 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

$$Q^\epsilon\big(\tilde{x}_i \in (3/4, 1]\big) \leq Q^{-\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) + \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\big(Q^\epsilon_{\bar{Z}_{T-1}} \,||\, Q^{-\epsilon}_{\bar{Z}_{T-1}}\big)} \ , \tag{41}$$

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

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}\big(Q^\epsilon_{\bar{Z}_{t+1}} \,||\, Q^{-\epsilon}_{\bar{Z}_{t+1}}\big) = \mathcal{D}_{\mathrm{KL}}\big(Q^\epsilon_{(\bar{Z}_t, Z_{t+1})} \,||\, Q^{-\epsilon}_{(\bar{Z}_t, Z_{t+1})}\big) &= \mathcal{D}_{\mathrm{KL}}\big(Q^\epsilon_{\bar{Z}_t} \,||\, Q^{-\epsilon}_{\bar{Z}_t}\big) \\ &\quad + \sum_{(\bar{z}, z) \in \{\emptyset, 0, 1\}^t \times \{\emptyset, 0, 1\}} \log\Bigg(\bigg(\frac{Q^\epsilon(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}{Q^{-\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}\bigg) \\ &\quad \cdot Q^\epsilon\big(\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\big)\Bigg) \ . \end{aligned} \tag{42}$$

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

$$\begin{aligned} &\sum_{(\bar{z}, z) \in \{\emptyset, 0, 1\}^t \times \{\emptyset, 0, 1\}} \log\left(\frac{Q^\epsilon(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}{Q^{-\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}\right) \cdot Q^\epsilon\big(\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\big) \\ &\quad = \sum_{\substack{(\bar{z}, z) \in \{\emptyset, 0, 1\}^t \times \{\emptyset, 0, 1\} \\ t+1 \in I_{t+1}}} \log\left(\frac{Q^\epsilon(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}{Q^{-\epsilon}(Z_{t+1} = z \mid \bar{Z}_t = \bar{z})}\right) \cdot Q^\epsilon\big(\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\big) \\ &\quad = \left(\sum_{\substack{\bar{z} \in \{\emptyset, 0, 1\}^t \\ t+1 \in I_{t+1}}} Q^\epsilon(\bar{Z}_t = \bar{z})\right) \\ &\qquad \cdot \sum_{z \in \{0, 1\}} \log\left(\frac{Q^\epsilon\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}{Q^{-\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}\right) \cdot Q^\epsilon(\mathbf{1}(1/2 < v_{t+1}) = z) \\ &\quad = Q^\epsilon\big(\tilde{x}_{t+1} \in (1/2, 3/4]\big) \\ &\qquad \cdot \sum_{z \in \{0, 1\}} \log\left(\frac{Q^\epsilon\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}{Q^{-\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}\right) \cdot Q^\epsilon(\mathbf{1}(1/2 < v_{t+1}) = z) \ . \end{aligned} \tag{43}$$

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

$$\begin{aligned}
&\sum_{z\in\{0,1\}} \log\left(\frac{Q^{\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}{Q^{-\epsilon}\big(\mathbf{1}(1/2 < v_{t+1}) = z\big)}\right) \cdot Q^{\epsilon}(\mathbf{1}(1/2 < v_{t+1}) = z) \\
&\qquad = \log\left(\frac{Q^{\epsilon}\big(\frac{1}{2} < v_{t+1}\big)}{Q^{-\epsilon}\big(\frac{1}{2} < v_{t+1}\big)}\right) \cdot Q^{\epsilon}\left(\frac{1}{2} < v_{t+1}\right) \\
&\qquad\quad + \log\left(\frac{Q^{\epsilon}\big(\frac{1}{2} \geq v_{t+1}\big)}{Q^{-\epsilon}\big(\frac{1}{2} \geq v_{t+1}\big)}\right) \cdot Q^{\epsilon}\left(\frac{1}{2} \geq v_{t+1}\right) \\
&\qquad = \log\left(\frac{1-a-(1+\epsilon)\cdot b}{1-a-(1-\epsilon)\cdot b}\right) \cdot \big(1-a-(1+\epsilon)\cdot b\big) \\
&\qquad\quad \log\left(\frac{a+(1+\epsilon)\cdot b}{a+(1-\epsilon)\cdot b}\right) \cdot \big(a+(1+\epsilon)\cdot b\big) \\
&\qquad \leq \frac{4\cdot b^2\cdot\epsilon^2}{\big(1-a-(1-\epsilon)\cdot b\big)\cdot\big(a+(1-\epsilon)\cdot b\big)} \leq \frac{4\cdot b^2\cdot\epsilon^2}{a\cdot(1-a-2b)} = 2\cdot c_3^2\cdot\epsilon^2\,. \qquad (44)
\end{aligned}$$

Putting (42), (43) and (44) together, we obtain that, for each $t \in \mathbb{N}$ and each $\epsilon \in [0,1]$,

$$\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_{t+1}} \,||\, Q^{-\epsilon}_{\bar{Z}_{t+1}}\big) \leq \mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{Z_1} \,||\, Q^{-\epsilon}_{Z_1}\big) + 2\cdot c_3^2\cdot\epsilon^2\cdot\sum_{s=1}^{t} Q^{\epsilon}\big(\tilde{x}_{s+1} \in (1/2, 3/4]\big)\,. \qquad (45)$$

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

$$\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{Z_1} \,||\, Q^{-\epsilon}_{Z_1}\big) \leq 2\cdot c_3^2\cdot\epsilon^2\cdot Q^{\epsilon}\big(\tilde{x}_1 \in (1/2, 3/4]\big)\,. \qquad (46)$$

For each $t \in \{1,\dots,T\}$, putting (45) and (46) together, we obtain

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}\big(Q^{\epsilon}_{\bar{Z}_t} \,||\, Q^{-\epsilon}_{\bar{Z}_t}\big) &\overset{(45)+(46)}{\leq} 2\cdot c_3^2\cdot\epsilon^2\cdot\sum_{s=1}^{t} Q^{\epsilon}\big(\tilde{x}_s \in (1/2, 3/4]\big) \\
&\leq 2\cdot c_3^2\cdot\epsilon^2\cdot E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)\,. \qquad (47)
\end{aligned}$$

Now, (41) and (47) imply that, for each $\epsilon \in [0,1]$ and each $i \in \{1,\dots,T\}$,

$$Q^{\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) \leq Q^{-\epsilon}\big(\tilde{x}_i \in (3/4, 1]\big) + c_3\cdot\epsilon\cdot\sqrt{E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)}\,. \qquad (48)$$

Taking the sum of (48) over $i \in \{1,\dots,T\}$, we obtain that for each $\epsilon \in [0,1]$,

$$\begin{aligned}
&E^{-\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \\
&\qquad \geq E^{\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) - c_3\cdot\epsilon\cdot T\cdot\sqrt{E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)}\,. \qquad (49)
\end{aligned}$$

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

$$\begin{aligned}
E^{-\epsilon}(\tilde{n}_3) &= \int_{[0,1]^{\mathbb{N}}} E^{-\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{-\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\overset{(39)}{=} \int_{[0,1]^{\mathbb{N}}} E^{-\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\overset{(49)}{\geq} \int_{[0,1]^{\mathbb{N}}} E^{\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\quad - c_3 \cdot \epsilon \cdot T \cdot \int_{[0,1]^{\mathbb{N}}} \sqrt{E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big)} \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
\text{(by Jensen)} \qquad &\geq \int_{[0,1]^{\mathbb{N}}} E^{\epsilon}\big(\tilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\quad - c_3 \cdot \epsilon \cdot T \cdot \sqrt{\int_{[0,1]^{\mathbb{N}}} E^{\epsilon}\big(\tilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big) \mathrm{d}P^{\epsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots)} \\
&= E^{\epsilon}(\tilde{n}_3) - c_3 \cdot \epsilon \cdot \sqrt{E^{\epsilon}(\tilde{n}_1)}\,. \qquad \square
\end{aligned}$$

A.3 Theorem 2

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

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

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

$$\begin{aligned} &\sup_x \mathbb{U}(x) - E\left[\mathbb{U}\right] \\ \leq &\left(\gamma + \eta \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2K+1}{6} + \tfrac{\lambda^2}{\gamma}\right) + \tfrac{\lambda}{K}\right) \cdot T + \tfrac{\log(K+1)}{\eta}. \end{aligned} \tag{50}$$

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

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

The second claim of the theorem follows.

□