Optimal and Robust Price Experimentation: Learning by Lottery

Christopher R. Dance **Onno Zoeter**
Xerox Research Centre Europe

Abstract

This paper studies optimal price learning for one or more items. We introduce the Schrödinger price experiment (SPE) which superimposes classical price experiments using lotteries, and thereby extracts more information from each customer interaction. If buyers are perfectly rational we show that there exist SPEs that in the limit of infinite superposition learn optimally *and* exploit optimally. We refer to the new resulting mechanism as the hopeful mechanism (HM) since although it is incentive compatible, buyers can deviate with extreme consequences for the seller at very little cost to themselves. For real-world settings we propose a robust version of the approach which takes the form of a Markov decision process where the actions are functions. We provide approximate policies motivated by the best of sampled set (BOSS) algorithm coupled with approximate Bayesian inference. Numerical studies show that the proposed method significantly increases seller revenue compared to classical price experimentation, even for the single-item case.

1 INTRODUCTION

We consider the problem of learning while selling. A seller does not know the distribution of valuations of his customer base, and strategically experiments with his prices to learn this distribution and maximize his profit.

This problem is conventionally addressed as a Markov decision problem (MDP) where the actions are real numbers, called prices. We argue that it is interesting to consider it as a *functional MDP* where the actions are *functions* corresponding to offering different probabilities of obtaining an item for different prices. Such functional MDPs allow us to adopt a new approach to exploration-exploitation tradeoffs which conducts several explorations *in parallel*.

The conventional approach to learning while selling a single item is a *censored price experiment (CPE)*. This is an instance of Bayesian reinforcement learning [1] where the seller obtains censored observations that a buyer's valuation is larger or smaller than a price. There is much previous work on learning from a CPE, notably [2, 3].

We consider exploration-exploitation tradeoffs for selling multiple items. In a two-item setting, it is necessary to learn from *region-censored observations* that a buyer's valuation falls in some two-dimensional set. Inference from such observations has been studied rarely [4]. In the absence of learning, it was recently discovered that optimal methods for selling multiple items involve *lotteries* where a buyer is allocated items with probability other than 0 or 1 [5, 6]. Lotteries form an important part of these methods: for two items, price-only methods can lose up to a factor of three in revenue relative to optimal lottery methods; for more than three items, price-only methods can lose an arbitrary factor [7].

Contributions

In Section 2 we introduce the Schrödinger price experiment (SPE) which exploits lotteries to provably learn faster than traditional censored price experiments. Our main contribution, in Section 3 is to demonstrate that a suitable SPE with infinite options can learn optimally *and* exploit optimally at the same time. In other words the censored observations in traditional price learning can be replaced by a mechanism that makes crisp observations by making use of lotteries. This extends the growing body of literature on incentive compatible learning [8, 9] and MDPs in dy-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

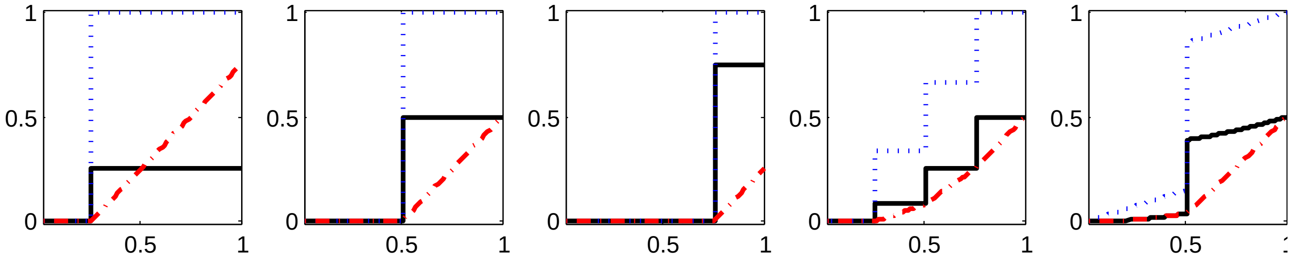

Figure 1: Example Mechanisms. Each plot has buyer valuation v on the x-axis and shows the price (black, solid), probability (blue, dotted), surplus (red, dot-dashed). From left to right, CPE for first n_1 days, CPE for next n_2 days, CPE for next n_3 days, SPE, variable lottery in Lemma 3.

namic mechanism design [10, 11, 12].

We demonstrate in Section 3.1 that a clever use of insurance can make the method applicable to settings where buyers are risk averse.

We say a mechanism is *incentive compatible (IC)* when it is in a participant's best interest to act in the way prescribed by the mechanism in all rounds (see [13] p. 218).

We refer to our optimal learning method as the *hopeful mechanism (HM)* since it is not robust: while it is IC, it is also possible for buyers to deviate far from truthful behaviour at little or no cost to themselves. Such deviations (errors, lies, or other unmodelled friction) can greatly harm the learning process.

The second part of the paper is therefore dedicated to formulating a robust mechanism learning problem that could be used in a practical setting where it is not safe to rely on perfect rationality. The general idea is to exploit lotteries to learn faster, but not take this to the theoretical limit to avoid the risk implied by possible imperfections in the agent choices. Only in the hopeful mechanism are exploration and exploitation maximal at the same time. A robust mechanism with a finite number of options requires a balance between exploration and exploitation. In Section 4 we introduce a robust learning algorithm that finds optimal lottery menus that bracket customer groups to balance exploration and exploitation while being sufficiently robust against buyers that pick suboptimal options.

We use exploration methods that sample from posterior distributions [14] to approximately solve the (PO)MDP and use assumed density filtering [15] to approximate the censored updates.

Finally we evaluate the optimal learning method and our algorithm on one- and two-item settings with a simple buyer deviation model.

2 SCHRÖDINGER PRICE EXPERIMENT (SPE)

In a classical censored price experiment (CPE), the seller sets prices $(p_k)_{k=1}^N$ for $(n_k)_{k=1}^N$ days respectively, with $p_k, n_k \in \mathbb{R}^+$ for $k = 1, 2, \ldots, N-1$. Without loss of generality for the current argument we assume that $p_k < p_{k+1}$. An alternative approach, that we call a *SPE*, is illustrated in Figure 1. A SPE sets a single menu of N lotteries for the whole period of $T := \sum_{k=1}^N n_k$ days. Each lottery $(z_k, \bar{p}_k)$ is given by the probability z_k that the buyer obtains the item given that they pay price $\bar{p}_k$. We set

$$z_k := T^{-1} \sum_{t=1}^{k} n_t, \ \ \bar{p}_k := T^{-1} \sum_{t=1}^{k} p_t n_t.$$

One could view the set of lotteries as a supply curve: that is, a mapping from quantity z_k to price $\bar{p}_k$. Buyers who value the item highly, will purchase in higher quantities and for higher prices.

Lemma 1 *For iid valuations $v \in \mathbb{R}$, expected buyer surplus and revenue from CPE and SPE are the same.*

Proof. For the SPE a buyer selects the lottery k maximizing their surplus $S_k(v) := z_k v - \bar{p}_k = T^{-1} \sum_{t=1}^{k} n_t (v - p_t)$ and if $S_k(v) < 0, \forall k$ they do not buy. It is readily verified that for $v \in [p_k, p_{k+1})$ we have $S_k(v) \geq S_j(v)$ for all j. Indeed the definition of lottery $(z_k, \bar{p}_k)$ is exactly the probability that a randomly-arriving buyer confronted by CPE with $v \in [p_k, p_{k+1})$ would buy, and the average amount that such a buyer would pay. Therefore surplus and revenue for CPE and SPE are the same. □

Interestingly SPE learns faster than CPE. That is, if a new price must be selected on the basis of observations made by SPE, or observations made by CPE, then the expected revenue from a new buyer with the same valuation distribution as previous buyers is larger for the price based on SPE than for the price based on

CPE. This is because observations by SPE carve up the space of valuations more finely than observations by CPE. To formalize this intuition, we say a partition $\mathcal{A} := \{A_1, A_2, \dots\}$ is a *refinement* of a partition $\mathcal{B} := \{B_1, B_2, \dots\}$ if they are partitions of the same set and for all $A \in \mathcal{A}$ there is a $B \in \mathcal{B}$ such that $A \subseteq B$. Let $\mathcal{M}$ be a mechanism selected from some set $\Omega_{\mathcal{M}}$. Let $W(\mathcal{M}, v)$ be a real function (a reward to the seller) of $\mathcal{M}$ and a valuation $v \in \mathbb{R}^n$ for some $n \in \mathbb{Z}^+$. We denote an optimal mechanism for valuation v_2 given an observation that v_1 is in some set C of a partition $\mathcal{C}$ by

$$\mathcal{M}_{\mathcal{C}|v_1} \in \mathrm{argmax}_{\mathcal{M}\in\Omega_{\mathcal{M}}} \mathbb{E}_{v_2}[W(\mathcal{M}, v_2)|v_1 \in C].$$

Lemma **2** *Let v_1, v_2 be valuation vectors with joint density $\mathbb{P}(v_1, v_2)$. If $\mathcal{A}, \mathcal{B}$ are partitions of the space of valuations and $\mathcal{A}$ is a refinement of $\mathcal{B}$, then optimal mechanisms for observations from these partitions satisfy*

$$\mathbb{E}_{v_1,v_2}[W(\mathcal{M}_{\mathcal{A}|v_1}, v_2)] \geq \mathbb{E}_{v_1,v_2}[W(\mathcal{M}_{\mathcal{B}|v_1}, v_2)].$$

Proof. The definition of $\mathcal{M}_{\mathcal{A}|v_1}$ gives $\mathbb{E}_{v_2}[W(\mathcal{M}_{\mathcal{A}|v_1}, v_2)|v_1 \in A] \geq \mathbb{E}_{v_2}[W(\mathcal{M}_{\mathcal{B}|v_1}, v_2)|v_1 \in A]$ for any $A \in \mathcal{A}$. Note that $\mathcal{M}_{\mathcal{A}|v_1}$ is independent of v_1 given $v_1 \in A$ and expand the joint density:

$$\mathbb{P}(v_1, v_2) = \sum_{A\in\mathcal{A}} \mathbb{P}(v_1 \in A)\mathbb{P}(v_1|v_2, v_1 \in A)\mathbb{P}(v_2|v_1 \in A)$$

to get

$$\begin{aligned} &\mathbb{E}_{v_1,v_2}[W(\mathcal{M}_{\mathcal{A}|v_1}, v_2)] \\ &= \sum_{A\in\mathcal{A}} \mathbb{P}(v_1 \in A)\mathbb{E}_{v_2}[W(\mathcal{M}_{\mathcal{A}|v_1}, v_2)|v_1 \in A] \\ &\geq \sum_{A\in\mathcal{A}} \mathbb{P}(v_1 \in A)\mathbb{E}_{v_2}[W(\mathcal{M}_{\mathcal{B}|v_1}, v_2)|v_1 \in A] \\ &= \mathbb{E}_{v_1,v_2}[W(\mathcal{M}_{\mathcal{B}|v_1}, v_2)]. \ \Box \end{aligned}$$

It is possible to refine the partition employed by a SPE until it becomes a continuum. The following Lemma shows that this can have interesting consequences (see Figure 1, Plot 5).

Lemma **3** *There exist mechanisms that are arbitrarily close to price-only mechanisms yet for which it is strictly IC for a buyer to identify their valuation.*

Proof. Consider a mechanism for a valuation $v \in [0, 1]$ that gives probability $z(v) := \bar{\kappa}\{v \geq q\} + \kappa v$ of obtaining an item for price $p(v) := \bar{\kappa}q\{v \geq q\} + \kappa v^2/2$, where κ and q are in the range $(0, 1]$, $\bar{\kappa} := 1 - \kappa$ and $\{\cdot\}$ is the indicator function. If a buyer with valuation v acts as if their valuation were $v + x$, for $x^2 > 0$, then their surplus is

$$\begin{aligned} w(v, v + x) &:= z(v + x)v - p(v + x) \\ &= \bar{\kappa}(v - q)(\{v + x \geq q\} - \{v \geq q\}) \\ &\quad - \kappa x^2/2 + w(v, v) \\ &< w(v, v) . \end{aligned}$$

Therefore the mechanism is strictly IC. Furthermore, the difference from the price-only mechanism for price q is $|z(v) - \{v \geq q\}| \leq \kappa$, $|p(v) - q\{v \geq q\}| \leq \kappa/2$, $\forall v, q \in (0, 1]$. $\Box$

Intuitively, Lemma 3 converts a price-only mechanism into a strictly IC mechanism by mixing the price-only mechanism with a strictly IC mechanism. The buyer's surplus under the price-only mechanism is a piecewise constant function of the amount by which they deviate. In contrast, the surplus under proposed mixture mechanism has a unique maximum. While the proof of Lemma 3 depends on the valuation distribution having bounded support, Lemma 5 (in the supplementary material) shows how to produce strictly IC mixtures even for distributions with unbounded support.

3 OPTIMAL IC LEARNING

Model Definition. At each time $t \in \{1, 2, \dots, T\}$, one buyer from a set of buyers B interacts with a seller. Each buyer can interact multiple times. We refer to the buyer at time t as b_t. Buyer $b \in B$ arrives at the *set* of times τ_b. The buyer at time t has a valuation $v_t \in \mathbb{R}^n$ with one component for each of $n \in \mathbb{Z}^+$ items. All valuations v_t are iid with distribution $f(v|\theta)$ where θ are parameters over which the seller has prior belief $g(\theta)$. After the statement of Theorem 1, we discuss how this model may be generalized to situations where valuations are not iid.

Generally a *mechanism* $\mathcal{M} := \langle\mathcal{A}, \mathcal{P}\rangle$ is defined by an allocation rule $\mathcal{A}$ and a payment rule $\mathcal{P}$. The allocation and payment are each functions of the valuation v'_t that a buyer b_t with true valuation v_t reports to the mechanism. At time t, the seller offers a mechanism $\mathcal{M}_t = (z_t, p_t)$ given by an n-vector of multinomial probabilities $z_t(v_t)$ that the buyer gets each item (the allocation rule), and a purchase price $p_t(v_t)$. The seller observes the chosen lottery and identity b_t of each buyer. We wish to choose a *policy* π which is a mapping from the observation history $\mathcal{F}_t$ up to and not including time t, and the current buyer's identity b_t to a mechanism $\mathcal{M}_t = \pi(\mathcal{F}_t, b_t)$. Full notation would therefore have all offered probabilities $z_t(\mathcal{F}_t, b_t, v_t)$ and prices $p_t(\mathcal{F}_t, b_t, v_t)$ explicitly dependent on history $\mathcal{F}_t$ and identity b_t. For brevity we write these as $z_t(v_t)$, and $p_t(v_t)$ and denote a full history of observations by $\mathcal{F} := \mathcal{F}_{T+1}$.

The space of feasible policies Ω_π is given by the following constraints

For all buyers $b \in B$, for all times $t \in \tau_b$,
for all possible true valuations v_t and deviations v'_t:

- IR Buyers must find it individually rational to participate: $\sum_{t\in\tau_b}(z_t(v_t)\cdot v_t - p_t(v_t)) \geq 0$.
- IC Incentive compatibility: $\sum_{t\in\tau_b}(z_t(v_t)\cdot v_t - p_t(v_t)) \geq \sum_{t\in\tau_b}(z_t(v'_t)\cdot v_t - p_t(v'_t))$.
- B1 The buyer only wishes to buy at most one item at a time (unit buyer): $1_n \cdot z_t(v_t) \leq 1, z_t(v_t) \geq 0$ where 1_n is the n-vector with components all one.

Note that via the dependence of z_t and p_t on histories $\mathcal{F}_t$ and buyer identities b_t the forall quantifiers imply that the above three constraints need to hold across multiple time steps. For example the IC constraint requires that manipulating in earlier rounds cannot be beneficial for agents in later rounds. Lemma 4 characterizes the implications for mechanisms for this IC through time notion.

We consider a general notion of welfare parameterized by weights α, β with $\alpha \geq \beta \geq 0$. If c_t is the vector of seller costs, then the *welfare* at time t is

$$\begin{aligned} &W_t(\mathcal{M}_t, v_t) := \\ &\qquad \alpha(p_t(v_t) - c_t \cdot z_t(v_t)) + \beta(v_t \cdot z_t(v_t) - p_t(v_t)). \end{aligned}$$

Setting $\alpha = 1, \beta = 0$ corresponds to seller profit, and $\alpha = 1, \beta = 1$ corresponds to social welfare. The *IC learning problem* is then to maximize the expected total welfare

$$\max_{\pi\in\Omega_\pi} \mathbb{E}_\theta \mathbb{E}_{v_1,v_2,\ldots,v_T|\theta}\left[\sum_{t=1}^{T} W_t(\pi(\mathcal{F}_t, b_t), v_t)\right]. \qquad (1)$$

Hopeful Mechanism (HM). The *hopeful mechanism* is a specific policy for the IC learning problem. Given a bound on the range of valuations $v_i \leq V$ for each component $i = 1, 2, \ldots, n$, the seller selects a small parameter $\kappa \geq 0$. The seller then repeats the following steps:

1. The seller computes the posterior belief about parameters θ on the basis of $\mathcal{F}_t^{-b_t}$, which is defined as the observed history $\mathcal{F}_t$ *excluding* information from buyer b_t (see Lemma 4).

2. The seller computes the IR, IC mechanism $\mathcal{M}_t$ that maximizes expected welfare in the current step given this posterior belief. This is the *myopic mechanism.* It is the solution to

$$\max_{\mathcal{M}} \mathbb{E}_{\theta, v_t}\left[W_t(\mathcal{M}, v_t)|\mathcal{F}_t^{-b_t}\right]$$

   where $\mathcal{M} = (z, p)$, subject to

$$\begin{aligned} 1_n \cdot z(v_t) &\leq 1\,, \\ z(v_t) &\geq 0\,, \\ z(v_t)\cdot v_t - p(v_t) &\geq 0\,, \\ z(v_t)\cdot v_t - p(v_t) &\geq z(v'_t)\cdot v - p(v'_t),\ \forall\ v_t, v'_t\,. \end{aligned}$$

3. The seller posts the mechanism $\mathcal{M}_t$ with $z_t = \bar{\kappa} z(v) + \kappa v/V, p_t(v) = \bar{\kappa} p(v) + \kappa |v|^2/(2V)$.

4. The buyer makes a purchase decision for this mechanism.

Lemma 4 *Policies for the IC learning problem are IC and IR over time (in the sense of Ω_π) if and only if they are IC and IR at each time (in the sense of the constraints of Step 2 of the definition of HM), and observations from any buyer b at one time do not affect mechanisms offered to b at other times, up to a redistribution of the payments of buyer b over time.*

In this Lemma we define two policies $(z_t, p_t)_{t=1}^T$ and $(z'_t, p'_t)_{t=1}^T$ as *equivalent up to a redistribution of payments over time* if for all histories $\mathcal{F}$, $z_t(\mathcal{F}) = z'_t(\mathcal{F})$ and for all $b \in B$ the total revenue satisfies $\sum_{t\in\tau_b} p_t(\mathcal{F}) = \sum_{t\in\tau_b} p'_t(\mathcal{F})$. A detailed proof of this Lemma is given in the Appendix (supplementary material).

Theorem 1 *As $\kappa \to 0$, HM becomes an optimal IC learning policy.*

Optimal here means in the sense of Eq. 1 where the restriction to the set Ω_π implies IR, IC and B1.

Proof. HM is clearly IC at any one time, because the myopic mechanism is IC by definition and the argument of Lemma 3 shows that adding the terms in κ to the price and probability make do not affect IC. By Lemma 4, HM is IC across times and loses no welfare by offering mechanisms based only on $\mathcal{F}_t^{-b_t}$. To demonstrate that HM is IR at any one time, one uses the same approach. This shows that $z_t(v_t)\cdot v_t - p_t(v_t) \geq 0$, hence $\sum_{t\in\tau_b} z_t(v_t)\cdot v_t - p_t(v_t) \geq 0$ showing that HM is IR across time.

If the seller's observations were some functions of valuations, such as a censored observation that the buyers' valuation lies in some region, then by Lemma 2, the seller does not gain any welfare relative what they would have obtained by HM in the limit $\kappa \to 0$.

Given truthful observations of valuations, the seller moves through a sequence of belief states that is independent of which mechanisms are offered. The optimal solution to this belief MDP is therefore myopic. This myopic behaviour is exactly what HM does in the limit $\kappa \to 0$. □

Until now we assumed that valuations were iid given a parameter θ. An examination of the proof of Theorem 1 shows that HM remains optimal under more general assumptions on the valuation distribution. All that is required is that the seller moves through a sequence of belief states that is independent of which mechanisms are offered, given truthful observations. This requirement is true if for all t

$$\begin{aligned}\mathbb{P}((v_s)_{s=t}^T, (b_s)_{s=t}^T | (v_s)_{s=1}^{t-1}, (b_s)_{s=1}^{t-1}, (z_s)_{s=1}^{t-1}, (p_s)_{s=1}^{t-1}) \\ = \mathbb{P}((v_s)_{s=t}^T, (b_s)_{s=t}^T | (v_s)_{s=1}^{t-1}, (b_s)_{s=1}^{t-1}).\end{aligned}$$

This alternative assumption includes settings where valuations are not identically distributed, but may instead depend on the buyers identity or follow some trend over time.

Repeat Buyers. HM requires the maintainance of a separate *leave-one-out* belief for all buyers. This could require computation of belief updates that is quadratic in the number of buyers and the solution of a number of mechanism design problems that is linear in the number of buyers. If this is a problem, a practical solution is to draw an analogy with n-fold cross-validation [16]. The set of buyers is partitioned into n_f sets called *folds.* A separate belief is then maintained for each fold on the basis of reports from buyers in other folds. Buyers who are suspected to collude may beneficially be put in the same fold.

Maintaining a separate belief for each buyer requires reliable *buyer identification.* While this may be difficult to implement, as the proof of Lemma 4 shows, this is a necessary condition for IC for any method where prices are learned.

In the *buy-many* model of Briest *et al* [7], buyers optimize their surplus by purchasing several lotteries. This makes designing optimal lotteries computationally challenging and reduces profits. Our model does not preclude time steps being so close together that a buyer is essentially purchasing several lotteries at the same time. So why is HM still optimal, yet simple? The root of this apparent paradox is that our model assumes a buyer's valuations at different times are independent. More sophisticated models would have to account for the time-dependence of valuations.

3.1 Risk Aversion

Risk-averse buyers would typically value non-deterministic lotteries less than deterministic lotteries. This can be handled by offering *insurance*, as follows. If a buyer has valuation v_i for item $i = 1, \ldots, n$, and selects lottery probabilities z_i and price p, then they are charged a price p_i when the lottery outcome is i:

$$p_i \quad := \quad p + v_i - \sum_{j=1}^{n} z_j v_j, \text{ for } i = 1, \ldots, n \, .$$

If risk aversion is described by a monotonically-increasing strictly concave function $C(u)$ of a risk-neutral utility u, then it is straightforward to show that: (i) prices p_i make buyers risk neutral; (ii) the average price paid is p.

4 ROBUST LEARNING MECHANISMS

The hopeful mechanism (HM) for $\kappa = 0$ is IC but not strictly IC. Consider for instance the following simple deviation scheme: "when buying, claim the highest possible valuation; when not buying, claim a valuation of zero." Given a set of buyers deviating in this fashion and a sufficiently flexible family of beliefs, HM would construct a posterior valuation distribution with a peak at the highest possible valuation and a peak at zero. If the objective were revenue maximization then HM with $\kappa = 0$ would then only ever set a price equal to the highest possible valuation, thereby making very few sales and very little revenue.

Thus it is possible for buyers to deviate to extreme valuations at no cost to themselves. This would corrupt the learning process, rendering HM impractical. Thus, price-learning problems are not problems of IC learning, but of *robust* IC learning. We discuss how such problems might be formulated, describe a heuristic algorithm and a choice of belief update that will be used in experiments.

Formulation. How might we formulate this problem when the manner in which buyers lie is unknown? We identify the following three stationary models in which a buyer with valuation v deviates to some valuation $\mathfrak{D}$, given the mechanism $\mathcal{M}$:

1. **Essentially-True.** Given the true valuations, the deviation is iid and independent of the mechanism: $\mathbb{P}(\mathfrak{D}|v, \mathcal{M}) = \mathbb{P}(\mathfrak{D}|v)$. The learning problem is then equivalent to one in which the true distribution of buyer valuations v is replaced by a distribution over deviations $\mathfrak{D}$. HM remains optimal.
2. **Surplus-Dependent.** Given the true valuation and the loss of surplus resulting from the deviation for the given mechanism, the deviation is iid and otherwise independent of the mechanism. If the buyer's surplus when deviating from v to $\mathfrak{D}$ is $w(v, \mathfrak{D}|\mathcal{M})$, then such models satisfy $\mathbb{P}(\mathfrak{D}|v, \mathcal{M}) = \mathbb{P}(\mathfrak{D}|v, w(v, v|\mathcal{M}) - w(v, \cdot|\mathcal{M}))$.

This includes the case of quantal response equilibrium [17].

3. **General Stationary.** The deviation is iid given the valuation and mechanism. This family of deviations is rather large and includes the continuum-armed bandit problem with one arm per possible mechanism [18, 19].

Buyers would have to have a rather loose bound on their bounded rationality to act according to most general stationary models. Therefore we restrict our attention to the following surplus-dependent model. A buyer draws a *loss-of-surplus parameter* ϵ from distribution $\mathbb{P}(\epsilon)$ and deviates to some valuation

$$\mathfrak{D} \in A^{\epsilon} := \{v' : w_{\mathcal{M}}(v, v') \geq w_{\mathcal{M}}(v, v) - \epsilon\}.$$

One can envisage a corresponding robust optimal IC learning problem where the seller observes the sequence of deviations $\mathfrak{D}_t$ for $t = 1, 2, \ldots, T$ and attempts to learn about the distributions of v, ϵ and $\mathfrak{D}|\mathfrak{D} \in A^{\epsilon}$. Solutions to this problem might involve complex mechanisms that would be difficult to communicate and for buyers to decide between. It is not even obvious how buyers without detailed knowledge of other buyers should deviate within A^{ϵ}, as illustrated by the following result.

Lemma 5 *Deviating up and deviating down can both be optimal strategies for a surplus-dependent deviator having the intent of lowering price.*

Proof. Consider a sufficiently-smooth one-dimensional valuation distribution. Taylor expand the seller's revenue in the price p around the optimum price p_0: $p\mathbb{P}(v \geq p) =: c_0 - c_1(p-p_0)^2/2 + O((p-p_0)^3)$. If buyers deviate by a small amount ϵ, the revenue becomes $J(p,\epsilon) := p\mathbb{P}(v \geq p+\epsilon)$. The maximum of $J(p,\epsilon)$ for small ϵ is at

$$p = p_0 + \epsilon\left(\frac{c_0}{c_1 p_0^2} - 1\right) + O(\epsilon^2).$$

Thus the price decreases with ϵ if $c_0 < p_0^2 c_1$ and increases if $c_0 > p_0^2 c_1$. □

An alternative is to explore simpler settings based on the SPE that involve a limited number m of lotteries for buyers to select between. The seller might believe that the loss-of-surplus parameter is at most some small constant ϵ for most of the population. The given form for the deviation then makes the observation history $\mathcal{F}_t$ the sequence of censoring sets A_s^{ϵ} and buyers b_s for $s = 1, 2, \ldots, t-1$. This leads to the following robust learning problem:

$$\max_{\pi \in \Omega_\pi} \mathbb{E}_\theta \mathbb{E}_{v_1, v_2, \ldots, v_T|\theta}\left[\sum_{t=1}^{T} W_t(\pi(\mathcal{F}_t, b_t), v_t)\right]$$

over the space of policies where each mechanism involves at most m distinct choices of lottery.

Algorithm. The above problem is a special case of Bayesian reinforcement learning. In particular, it could be viewed as an extension of the linearly parameterized bandit model of [20] to a setting where the arms are functions. We explore a heuristic for the exploration-exploitation trade-off that is related to the best of sampled sets (BOSS) method [14]. BOSS samples multiple models from the posterior belief and selects actions optimistically. While it is optimistic, like the upper confidence bound approaches to learning MDPs [21], BOSS is clearer about the role of posterior beliefs.

Our heuristic samples parameters of the valuation distribution $\theta_1, \ldots, \theta_K$ from a posterior with hyperparameters α. It then identifies an optimal mechanism $\mathcal{M}_k$ for each sample θ_k. Some such mechanisms have a high expected welfare $\mathbb{E}_{v|\theta_k} W(v, \mathcal{M}_k)$ on their sampled valuation distribution relative to the expected welfare $\mathbb{E}_{v|\alpha} W(v, \mathcal{M}^*)$ of the myopic mechanism $\mathcal{M}^*$ on the current posterior. This can happen in two ways. Either sample θ_k is "lucky" and both mechanisms $\mathcal{M}_k, \mathcal{M}^*$ perform well on it, or mechanism $\mathcal{M}_k$ is substantially different from the myopic mechanism $\mathcal{M}^*$. In the second case, the myopic mechanism is a risky choice and it is imperative to explore alternatives such as $\mathcal{M}_k$.

Thus we identify a specific exploratory mechanism $\mathcal{M}^X$ which achieves the maximum over k of

$$\mathbb{E}_{v|\theta_k}[W(v, \mathcal{M}_k) - W(v, \mathcal{M}^*)].$$

The two mechanisms $\mathcal{M}^X = (z^X, p^X)$ and $\mathcal{M}^* = (z^*, p^*)$ are then superposed to create the mechanism $\mathcal{M}_t = (z, p)$ offered to the buyer. For simplicity, we linearly combine the probabilities and prices via a parameter λ (similar to the previous parameter κ), via

$$z := \lambda z^X + \bar{\lambda} z^*, \; p := \lambda p^X + \bar{\lambda} p^*.$$

Figure 2 shows an example of a myopic and exploratory mechanism for a two-dimensional valuation distribution, as well as the result of superposing them.

There are two steps needed to select λ. The first step is to choose an appropriate form for the sellers prior beliefs about the nature of deviations. For instance, experimental economics widely uses the notion of Quantal Response Equilibrium (QRE) [17] to model deviations from rational behavior. The second step involves numerically searching for a value of λ that maximizes the average welfare or revenue using a simulation of buyers responses given the prior beliefs about valuations and deviations.

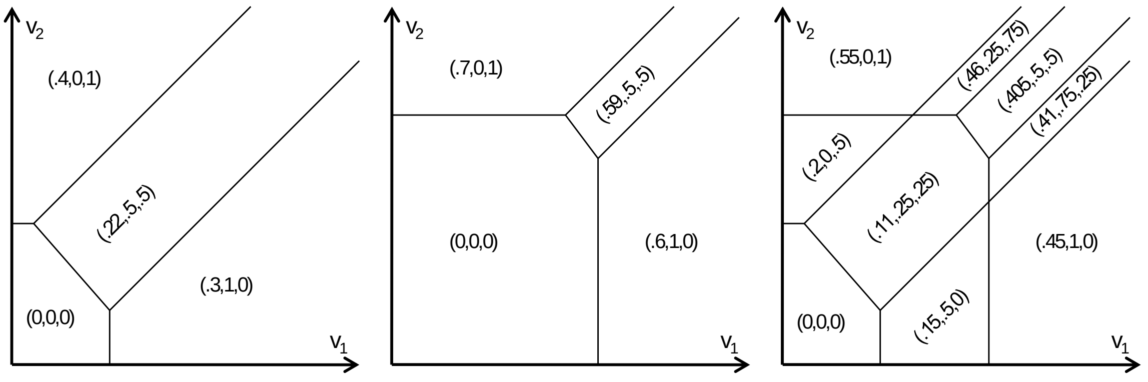

Figure 2: Each plot shows the partition of a two-dimensional valuation space induced by a mechanism. The quantities in parentheses are (p, z_1, z_2): the price, the probability of obtaining item 1 and the probability of obtaining item 2. From left to right: a myopic mechanism $\mathcal{M}^*$; an exploratory mechanism $\mathcal{M}^X$; the result of superposing them with parameter $\lambda = 1/2$.

Belief and Update. We assume that valuations are drawn from a multinomial distribution with parameters θ_k satisfying $\sum_{i=1}^N \theta_k = 1$ over a set $V := \{v_1, v_2, \ldots, v_N\}$. This is a flexible choice and the assumption of a discrete set could be motivated by a discretization of the set of monetary units. The seller does not know θ but instead has a Dirichlet prior with parameters $\alpha_1, \ldots, \alpha_N$ over it, given by $\mathbb{P}(\theta|\alpha) := \Gamma(\sum_{i=1}^N \alpha_i) \prod_{i=1}^N \left(\theta_i^{\alpha_i - 1}/\Gamma(\alpha_i)\right)$. The Dirichlet is conjugate to the multinomial. It is also general enough to model *any* specific valuation probabilities, and makes sense if valuation distributions are driven by discontinuous effects of budget constraints and competing outside options. However we should expect that this flexibility would come at the price of slow learning, relative to a setting where a seller is confident that the distribution of valuations is from a more restricted set. Studying refinements seems worthwhile.

We lose conjugacy to the Dirichlet density if a valuation v is not directly observed but rather comes from a set $S \subseteq V$. Bayes's rule gives the posterior as a *mixture* of Dirichlets $\mathbb{P}(\theta|v \in S, \alpha) = \sum_{i \in S} \mathbb{P}(\theta|v_i, \alpha)\mathbb{P}(v_i|\alpha)/\sum_{j \in S} \mathbb{P}(v_j|\alpha)$. If we want the belief to remain in a simple family, we could apply assumed density filtering (ADF), expectation propagation (EP), or improvements thereof [15, 22, 23]. Here, we apply ADF which computes the full posterior after each observation and then approximates it by projecting it in the sense of minimum Kullback-Leibler (KL) divergence, into the simple family. For exponential probability density families, the minimum KL approximation is one that matches the *natural moments.* The natural moments for the Dirichlet are $\mathbb{E}_{\theta|\alpha} \log \theta_k = \psi(\alpha_k) - \psi(\alpha_0)$, where $\psi(z) := \frac{d}{dz} \log \Gamma(z)$. Using the expression for the exact posterior and matching moments, the corresponding ADF update reduces to solving

$$\psi(\alpha'_i) - \psi(\sum_{j=1}^N \alpha'_j) = \psi(\alpha_i) - \psi(1 + \sum_{j=1}^N \alpha_j) + \{i \in S\} \frac{\sum_{j=1}^N \alpha_j}{\sum_{j \in S} \alpha_j}$$

for new hyperparameters α'_j. Newton's method exploiting sparsity of the Jacobian is effective here.

Solving for Mechanisms. In the case of discrete one-dimensional valuations, the myopic and exploratory mechanisms are both price-only mechanisms which are easily enumerated. For multi-dimensional valuations, the problem can be cast either as a linear program (LP) or a semi-definite program [24]. The LP can be difficult to solve directly as there are a large number of IC constraints: one for each valuation v and possible deviation v'. We worked with valuations from a two-dimensional evenly-spaced $n \times n$ grid. We constructed a relaxed LP by only including IC constraints for the 5×5 square of deviations around a given valuation v. For such valuations we found that solutions to the relaxed LP repeatably satisfied the full IC constraint. A constraint sampling approach, as developed for approximate linear programming [25] could also be adopted for such problems. The proposed approach to selecting exploratory mechanisms $\mathcal{M}^X$ requires a large number of LP solutions. We therefore cached a large number of LP solutions prior to learning, corresponding to 1000 random samples from the initial belief distribution.

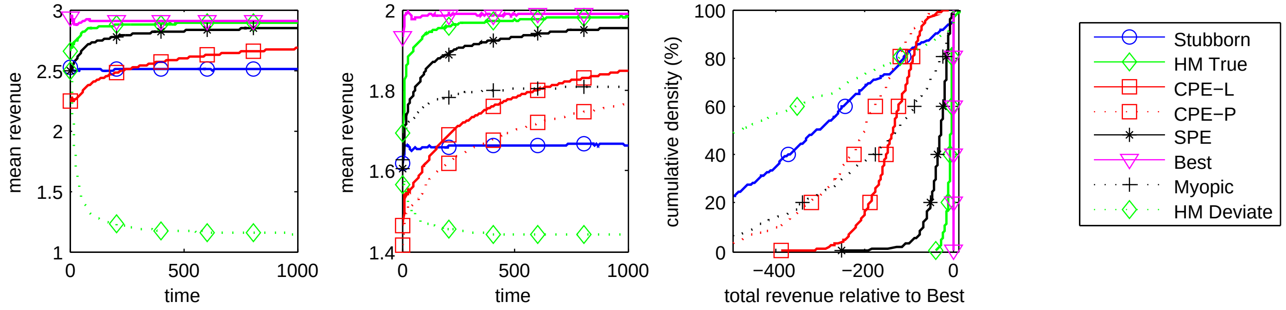

Figure 3: Left to Right: Mean revenue per unit time for 1D valuations; Mean revenue for 2D valuations; Cumulative density of total revenue after 1000 steps minus total revenue of Best for 2D valuations. Note that results for Best fluctuate slightly due to the fact it is computed using (still many) samples.

5 RESULTS

We explore the performance of six policies based on the method of the previous Section, with $K = 4$ posterior samples, and revenue as welfare. The first three methods, *SPE*, *Myopic* and *CPE*, are the method described in the previous section, with different parameters. Note that for the purposes of these results, SPE and CPE only refer to one specific implementation of the general concepts of SPE and CPE as introduced in Section 2. These methods are parameterized as follows: (i) SPE has $\lambda = 1/6$; (ii) Myopic has $\lambda = 0$ so that only the myopic mechanism $\mathcal{M}^*$ is used; (iii) CPE has $\lambda = 1$ so that only the exploratory mechanism $\mathcal{M}^X$ is used. For 2D valuations, *CPE-L* uses all possible mechanisms and *CPE-P* only uses price-only mechanisms. The remaining three methods are as follows: (iv) *HM* is the optimal method with no belief robustness (*i.e.* in the limit $\kappa \to 0$ of Theorem 1); (v) *Stubborn* is the optimal mechanism for the seller's initial belief; (vi) *Best* is the optimal mechanism if the valuation distribution is known (which is inachievable in practice).

Figure 3 shows the performance of these methods against a simple deviation scheme: "when buying claim the highest valuation ($1_n.\mathfrak{D}$) consistent with the purchase; when not buying, claim the lowest possible valuation." Buyers deviating this way lose no surplus in any individual buying decision. Aside from HM (curve HM Deviate) other methods are insensitive to this deviation as they make censored updates, if any. Thus the results for other methods are the same in the presence or absence of deviation. Each result is an average of 500 runs of length 1000, consistent with other literature on Bayesian reinforcement learning. At each step, all policies were confronted with an identical true valuation.

For 1D valuations, all methods were initialized with a Dirichlet belief over valuations in the set $\{1, 2, \ldots, 8\}$ with $\alpha_k = 1$. For 2D valuations, the belief was over the grid $\{0, 1, 2, \ldots, 5\}^2$ with $\alpha_k = 1/6$. We use a smaller α_k in 2D since if we were to set $\alpha_k = 1$, then the average gap between revenues for Best and Stubborn would be at most 7.3% (results not shown for space reasons), so there would be little for any policy to learn.

As expected, on average: HM outperforms others given the truth, but degrades rapidly given liars; given enough samples both CPE and SPE outperform Stubborn, but SPE learns faster; in 2D, CPE-L outperforms CPE-P. The cumulative density shows that Myopic performs variably: it is often lucky and beats CPE-L; but it can also "get stuck" due to lack of exploration. Other experiments with larger (e.g. 8 × 8) grids (not shown) indicate that CPE-L with $K = 4$ often only shows a benefit over Stubborn or Myopic after 2000 or more time steps, given the highly-flexible Dirichlet prior coupled with censored observations.

6 CONCLUSIONS

We introduced the hopeful mechanism (HM) and showed that it is an optimal method for incentive compatible learning of multi-dimensional economic mechanisms. HM is not robust to irrational buyer behaviour. Therefore we formulated a robust version of the problem. Finally we compared heuristic mechanism learning methods with optimal methods in the presence of potentially irrational buyers. This comparison demonstrated that the proposed method learns faster than alternatives and is robust. Future work on mechanism learning with alternative priors on valuations, models of deviations and of the temporal evolution of valuations would be of substantial interest, but of less interest than a general study of exploration with functional MDPs.

References

[1] C. Dimitrakakis. Complexity of stochastic branch and bound methods for belief tree search in Bayesian reinforcement learning. In *2nd ICAART*, pages 259–264, 2010.

[2] M. Rothschild. A two-armed bandit theory of market pricing. *J. Economic Theory*, 9(2):185–202, 1974.

[3] V. Farias and B. Van Roy. Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29, 2010.

[4] P. Bulla, P. Muliere, and S. Walker. Bayesian nonparametric estimation of a bivariate survival function. *Statist. Sinica*, 17:427–444, 2007.

[5] J. Thanassoulis. Haggling over substitutes. *J. Economic Theory*, 117(2):217–245, August 2004.

[6] A. Manelli and D. Vincent. Multidimensional mechanism design: Revenue maximization and the multiple-good monopoly. *J. Economic Theory*, 137(1):153–185, 2007.

[7] P. Briest, S. Chawla, R. Kleinberg, and S. Weinberg. Pricing randomized allocations. In *21st SODA*, 2010.

[8] A. Procaccia O. Dekel, F. Fischer. Incentive compatible regression learning. In *19th SODA*, pages 277–286, 2008.

[9] R. Meir, A. Procaccia, and J. Rosenschein. Strategyproof classification with shared inputs. In *21st IJCAI*, pages 220–225, 2009.

[10] D. Parkes and S. Singh. An MDP-based approach for online mechanism design. In *17th NIPS*, 2003.

[11] A. Pavan, I. Segal, and J. Toikka. Dynamic mechanism design: Incentive compatibility, profit maximization and information disclosure. Technical report, Stanford University, 2008.

[12] S. Kakade, I. Lobel, and H. Nazerzadeh. An optimal dynamic mechanism for multi-armed bandit processes. Technical report, arXiv:1001.4598, 2010.

[13] N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[14] J. Asmuth, L. Li, M. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *25th UAI*, pages 19–26, 2009.

[15] T. Minka. Expectation propagation for approximate Bayesian inference. In *17th UAI*, pages 362–369, 2001.

[16] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[17] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

[18] R. Kleinberg and F. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th FOCS*, pages 594–605, 2003.

[19] Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In *21st NIPS*, pages 1729–1736, 2008.

[20] P. Rusmevichientong and J. Tsitsiklis. Linearly parameterized bandits. *CoRR*, abs/0812.3465, 2008.

[21] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *21st NIPS*, pages 89–96, 2008.

[22] T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *18th UAI*, pages 216–223, 2002.

[23] M. Opper, U. Paquet, and O. Winther. Improving on expectation propagation. In *21st NIPS*, pages 1241–1248, 2008.

[24] N. Aguilera and P. Morin. On convex functions and the finite element method. *SIAM J. Numerical Analysis*, 47(4):3139–3157, 2009.

[25] V. Desai, V. Farias, and C. Moallemi. A smoothed approximate linear program. In *22nd NIPS*, pages 459–467. 2009.