Africa Summer School in Econometrics, Abidjan:
Adaptive field experiments

Adaptive maximization of social welfare

Maximilian Kasy

Department of Economics, University of Oxford

June 2024

Introduction

How should a policymaker act,

- who aims to maximize social welfare,

  Weighted sum of utility.

  $\Rightarrow$ Tradeoff redistribution vs. cost of behavioral responses.

- and needs to learn agent responses to policy choices?

  Adaptively updated policy choices.

  $\Rightarrow$ Tradeoff exploration vs. exploitation.

Introduction

How should a policymaker act,

- who aims to maximize social welfare,

  Weighted sum of utility.

  $\Rightarrow$ Tradeoff redistribution vs. cost of behavioral responses.

- and needs to learn agent responses to policy choices?

  Adaptively updated policy choices.

  $\Rightarrow$ Tradeoff exploration vs. exploitation.

Introduction

How should a policymaker act,

- who aims to maximize social welfare,

  Weighted sum of utility.

  $\Rightarrow$ Tradeoff redistribution vs. cost of behavioral responses.

- and needs to learn agent responses to policy choices?

  Adaptively updated policy choices.

  $\Rightarrow$ Tradeoff exploration vs. exploitation.

Taxes and bandits

- **Optimal tax theory**
  - Mirrlees (1971); Saez (2001); Chetty (2009)
- **Multi-armed bandits**
  - Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2020)
- This talk: **Merging bandits and welfare economics**.
  - Unobserved welfare, as in optimal taxation.
  - Unknown response functions (treatment effects), as in multi-armed bandits.
  - Coauthors: Nicolò Cesa-Bianchi and Roberto Colomboni.

Review: Optimal taxation

- Social welfare = weighted sum of individual utilities.
- Welfare weights:

  Relative value of a marginal lump-sum $ across individuals.

  $\approx$ Distributional preferences (rich vs. poor, healthy vs. sick,...)
- Envelope theorem:
  - Behavioral responses to marginal tax changes don't affect individual utilities.
  - They only impact public revenue (absent externalities).

  $\Rightarrow$ Impact on revenue is a sufficient statistic.
- Absent income effects:

  Consumer surplus

  $=$ Equivalent variation

  $=$ integrated response function.

Review: Adversarial bandits

- Canonical bandit problems:
  - Assign treatment sequentially.
  - Observe previous outcomes before the next assignment.
- Regret:

  How much worse is an algorithm

  than the best alternative in a given comparison set (e.g., fixed treatments).
- Two approaches for analyzing bandits:
  1. Stochastic: Potential outcomes are i.i.d. draws from some distribution.
  2. Adversarial: Potential outcomes are an arbitrary sequence.
- Adversarial regret guarantees:
  - Bound regret for arbitrary sequences.
  - We can do that because the stable comparison set substitutes for the stable data generating process.

Setup: Tax on a binary choice

Each time period $i = 1, 2, \ldots, T$:

- Policymaker (algorithm):
  - Chooses tax rate $x_i \in [0, 1]$.
- Agent i:
  - Willingness to pay: $v_i \in [0, 1]$.
  - Response function: $G_i(x) = \mathbf{1}(x \leq v_i)$
  - Binary agent decision: $y_i = G_i(x_i)$.
- Observability:
  - After period i, we observe y_i.
  - We do *not* observe welfare $U_i(x_i)$.

Social welfare

Weighted sum of public revenue and private welfare:

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

We can rewrite private welfare as an integral (consumer surplus):

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

Cumulative demand, welfare and regret

- Cumulative demand:
$$\mathbb{G}_T(x) = \sum_{i \leq T} G_i(x).$$
- Cumulative welfare for a constant policy x:
$$\mathbb{U}_T(x) = \sum_{i \leq T} \mathbb{U}_i(x) = x \cdot \mathbb{G}_T(x) + \lambda \int_x^1 \mathbb{G}_T(x')dx'.$$
- Cumulative welfare for the policies x_i actually chosen:
$$\mathbb{U}_T = \sum_{i \leq T} \mathbb{U}_i(x_i).$$
- Adversarial regret:
$$\mathcal{R}_T(\{v_i\}_{i=1}^T) = \sup_x E\left[\mathbb{U}_T(x) - \mathbb{U}_T \middle| \{v_i\}_{i=1}^T\right].$$

The structure of observability

Choice x_i reveals $G_i(x_i)$. But

$$U_i(x) - U_i(x') = \left[x \cdot G_i(x) - x' \cdot G_i(x')\right] + \lambda \int_x^{x'} G_i(x'')dx''$$

depends on values of $G_i(x'')$ for $x'' \in [x, x']$!

Different from standard adaptive decision-making problems:

- Multi-armed bandits:
  Observe welfare for the choice made.
- Online learning:
  Observe welfare for all possible choices.
- Online convex optimization:
  Observe gradient of welfare for the choice made.

Introduction

Setup

Lower and upper bounds on regret

Related learning problems and extensions

Lower bound on regret

Theorem

There exists a constant $C > 0$ such that,
for any algorithm for the choice of $x_1, x_2, \ldots$
and any time horizon $T \in \mathbb{N}$:

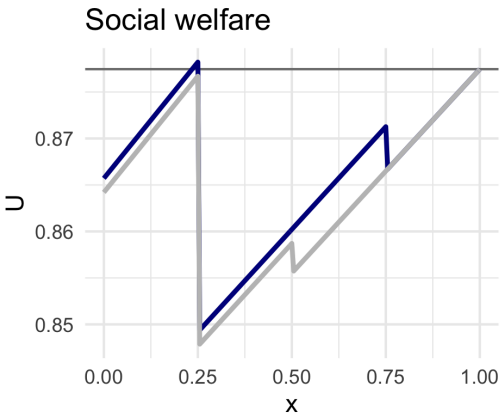
There exists a sequence $(v_1, \ldots, v_T)$ for which

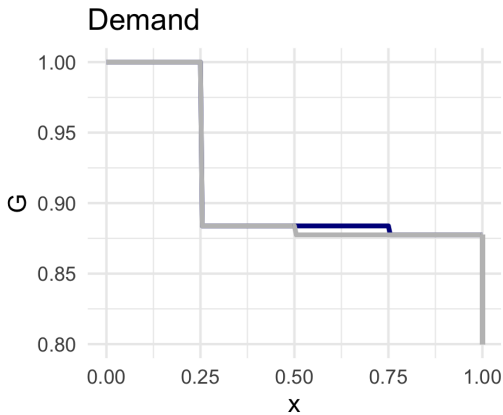
$$\mathcal{R}_T(\{v_i\}_{i=1}^T) \geq C \cdot T^{2/3}.$$

Sketch of proof: Lower bound on regret

- Stochastic regret $\leq$ adversarial regret.
  (Since average $\leq$ maximum.)
- Construct a distribution for $\boldsymbol{v}$ with 4 points of support, e.g. $(\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1)$.
- Choose the probability of each of these points such that
  1. The two middle points are far from optimal.
  2. Learning which of the two end points is optimal
     requires sampling from the middle.
     (Because of the integral term.)

Construction for the proof of the lower bound





Parameters: lambda = 0.95, a = 0.116, b = 0.003.

Tempered Exp3 for social welfare

Require: Tuning parameters K, γ and η.

1: Set $\tilde{x}_k = (k-1)/K$, initialize $\widehat{\mathbb{G}}_{1k} = 0$ for $k = 1, \ldots, K+1$.

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

3: $\forall\, k$, set

$$\widehat{\mathbb{U}}_{ik} = \tilde{x}_k \cdot \widehat{\mathbb{G}}_{ik} + \frac{\lambda}{K} \cdot \sum_{k' > k} \widehat{\mathbb{G}}_{ik'}. \tag{1}$$

4: $\forall\, k$, set

$$p_{ik} = (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\mathbb{U}}_{ik})}{\sum_{k'} \exp(\eta \cdot \widehat{\mathbb{U}}_{ik'})} + \frac{\gamma}{K+1}. \tag{2}$$

5: Sample $k_i \sim (p_{i,1}, \ldots, p_{i,K+1})$. Set $x_i = \tilde{x}_{k_i}$.

6: $\forall\, k$, set

$$\widehat{\mathbb{G}}_{i+1k} = \widehat{\mathbb{G}}_{i,k_i} + y_i \cdot \frac{\mathbf{1}(k_i = k)}{p_{ik}}. \tag{3}$$

7: **end for**

Upper bound on regret

Theorem

Consider the algorithm "Tempered Exp3 for social welfare."
There exists a constant C' and choices for K, γ, η such that,
for any sequence $(v_1, \ldots, v_T)$,

$$\mathcal{R}_T(\{v_i\}_{i=1}^T) \leq C' \cdot \log(T)^{1/3} \cdot T^{2/3}.$$

Note:

- Same rate as the lower bound, up to the logarithmic term.
- Upper bounds on adversarial regret
  are closely related to "Blackwell approachability."

Sketch of proof: upper bound on regret

- Discretize to balance the approximation error
  against the cost of having to learn $\mathbb{G}_i$ on more points.
- $\widehat{\mathbb{G}}$ is an unbiased estimator for cumulative demand $\mathbb{G}_i$.
  $\widehat{\mathbb{U}}$ is an unbiased estimator for cumulative discretized welfare.
- Consider $W_i = \sum_k \exp(\eta \cdot \widehat{\mathbb{U}}_{ik})$.
  - $E[\log W_T]$ is bounded below by η times optimal constant policy welfare.
  - $E\left[\log\left(\frac{W_i}{W_{i-1}}\right)\right]$ is bounded above by a combination of expected $\mathbb{U}_i$,
    and a term based on the second moment of $\widehat{\mathbb{U}}_i$.
- Bounding this second moment, and optimizing tuning parameters,
  yields the bound on adversarial regret.

Introduction

Setup

Lower and upper bounds on regret

Related learning problems and extensions

Related learning problems and extensions

- **Monopoly pricing**:
  - Monopolist profits:

$$U_i^{MP}(x) = \underbrace{x \cdot G_i(x).}_{\text{Monopolist revenue}}$$

  - Easier – like a continuous multi-armed bandit.

- **Bilateral trade**:
  - Buyer plus seller welfare:

$$U_i^{BT}(x) = G_i^b(x) \cdot \underbrace{\int_0^x G_i^s(x')dx'}_{\text{Seller welfare}} + G_i^s(x) \cdot \underbrace{\int_x^1 G_i^b(x')dx'}_{\text{Buyer welfare}}.$$

  - Harder – even gradients depend on global information.

Comparison of regret rates

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

- Rates are up to logarithmic terms.
- They reflect:
  1. Information structures:
     Pointwise (like bandit) vs. global (require exploration away from optimum).
  2. Smoothness properties:
     One-sided Lipschitzness allows us to bound the discretization error.

Extensions

1. **Concave welfare** functions:
   - Dyadic search algorithm.
   - Improved rate: $T^{1/2}$ (up to logarithmic terms).
2. **Non-linear income taxation**
   - Tax rate and welfare weights vary by income level.
   - Tempered Exp3 for welfare separately by tax brackets.
3. **Commodity taxation**:
   - Consumer choice in $\mathbb{R}^k$.
   - Regret rates: Future work.

Conclusion

- A canonical economic problem:

  Choosing policies to maximize social welfare,

  while needing to learn behavioral responses.

- More difficult than canonical bandits, monopoly pricing:

  Learning the optimal policy

  requires exploration of sub-optimal policies.

- Broader agenda:

  1. Adapt tools from machine learning for the purpose of public good.
     (Vs. profit maximization – monopoly pricing, ad click maximization...)
  2. Unify insights from (welfare) economics and computer science.
  3. Span the range from theoretical performance guarantees
     to practical implementation.

Thank you!