ICML 2026 Poster

Model Monotonicity in Autobidding Auctions

When Do Better Predictions Lead to Better Outcomes?

Ashwinkumar Badanidiyuru

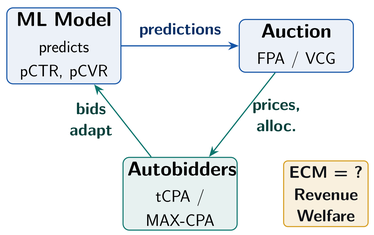
Uber Technologies, Inc. · San Francisco, CA, USA · ashwinkumarbv@uber.com

Read the paper

arXiv:2605.31036

1. The Problem

Modern ad platforms couple **three adaptive components**. Improving one (the model) need not improve the system.



Platforms constantly ship better **pCTR/pCVR** models, *assuming* better predictions ⇒ better business metrics. **This intuition can fail.**

Central Question

When does a **better prediction model** guarantee a better platform **Evaluation Criteria Metric** (ECM) — revenue, welfare, or liquid welfare?

ECM monotonicity:

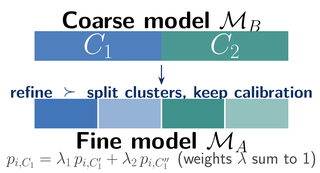
$$\mathcal{M}_A \succeq \mathcal{M}_B \;\Rightarrow\; \mathrm{ECM}(\mathcal{M}_A) \geq \mathrm{ECM}(\mathcal{M}_B)$$

2. Framework: Models as Partitions

A model $\mathcal{M}$ partitions users $\mathcal{U}$ into clusters and predicts the **calibrated** cluster average

$$p_{i,C} = \tfrac{1}{|C|} \textstyle\sum_{u \in C} p_i(u).$$

Refinement (inspired by filtrations): $\mathcal{M}_A \succeq \mathcal{M}_B$ if every cluster of $\mathcal{M}_A$ sits inside a cluster of $\mathcal{M}_B$ — *finer = more information.*



Calibration is preserved: each coarse prediction is the weighted average of its refined predictions (a **mean-preserving split**).

3. Two Autobidder Types

- **tCPA:** maximize conversions s.t. average cost/conversion $\leq t_i$ (we set value $v_i = t_i$).
- **MAX-CPA:** maximize value − payment, value v_i per conversion.

Uniform bidding: $b_i(C) = \mu_i t_i p_{i,C}$ (tCPA) or $\mu_i v_i p_{i,C}$ (MAX-CPA).

Auctions studied: first-price (FPA), VCG (= second-price for a single item), and a centralized LP benchmark.

Main Result — Monotonicity is *Rare*

Across every setting studied, only **three mechanisms** guarantee that a better model never hurts the metric (green): FPA (tCPA), VCG (MAX-CPA welfare), and the LP benchmark (covers both bidder types under budgets).

Bidder Type	Auction	Revenue	Welfare
tCPA	FPA	✔	✔
tCPA	VCG (=SPA)	✘	✘
tCPA + Budget	FPA	✘	✘
tCPA + Budget	LP*	✔	✔
MAX-CPA	FPA†	✘	✘
MAX-CPA	VCG	✘	✔
MAX-CPA + Budget	FPA†	✘	✘
MAX-CPA + Budget	LP*	–	✔

For budgeted rows, *welfare* means **liquid welfare**. *LP = centralized (not incentive-compatible) benchmark; tCPA "revenue" = surrogate LP payment. †MAX-CPA FPA compares designated feasible profiles, not a full equilibrium.

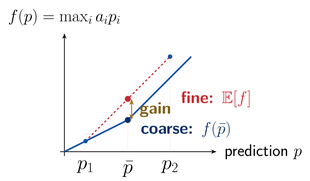
Why It Works: Convexity + Calibration ⇒ Jensen

Both *auction* guarantees share **one argument**. Per-cluster value

$$f(p) = \max_i a_i p_i \qquad (a_i = t_i \text{ or } v_i)$$

is **convex** (a max of linear functions). Refinement is a mean-preserving spread of predictions, so **Jensen's inequality** gives

$$\underbrace{\mathbb{E}[f(p_{\text{fine}})]}_{\text{finer model}} \;\geq\; \underbrace{f(\mathbb{E}[p])}_{\text{coarser model}} .$$



⇒ **FPA revenue** (tCPA) and **VCG welfare** (MAX-CPA). The **LP** benchmark is monotone by the *same* calibration identity, via a lifting argument (copy each coarse allocation to its sub-clusters).

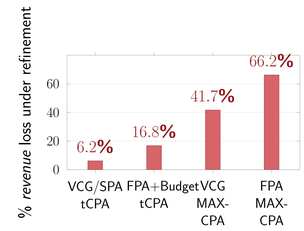
Contributions

- A **filtration-based** formalization of "model improvement" for calibrated prediction systems.
- A **systematic characterization** of ECM monotonicity over bidder types × auctions × budgets — with proofs for the positive cases and explicit numerical counterexamples for the rest.

Why It Breaks: Counterexamples

Once payments stop being a convex function of predictions, Jensen no longer protects the metric:

- **VCG/SPA** pays the *second* bid (externality), not convex ⇒ tCPA revenue & welfare drop.
- **Budgets** force pacing ($\mu \neq 1$) and bid shading ⇒ FPA revenue drops.
- **MAX-CPA** shading creates *market segmentation* that removes competition ⇒ revenue collapses.



Model refinement can *cut revenue* by up to **66%**.

A Concrete Failure (VCG/SPA, tCPA)

Two tCPA bidders (t_A=10, t_B=1), four single-impression auctions. True conversion probabilities:

	Auc 0	Auc 1	Auc 2	Auc 3
A	0.20	0.05	0.01	0.01
B	0.08	0.30	0.03	0.70

- **Coarse** $\{0,1\},\{2,3\}$: A wins $\{0,1\}$, B wins $\{2,3\}$ ⇒ Rev = Welfare = **3.23**.
- **Fine** (singletons): A wins $\{0\}$, B wins $\{1,2,3\}$ ⇒ Rev = Welfare = **3.03**.

The finer model hands high-value *Auction 1* from high-t bidder A to low-t bidder B — **both metrics fall 6.2%.**

Key Takeaways

- **Don't assume** better models give better metrics — monotonicity is the exception.
- Among *strategic* auctions, only **FPA** (tCPA, no budget) gives *both* revenue & welfare; **VCG** (MAX-CPA, no budget) gives *welfare only*.
- **Budgets** broadly break monotonicity.
- A fundamental **tension**: incentive compatibility *vs.* monotonicity under budgets.

Open Questions

- Any mechanism that is *both* incentive-compatible *and* monotone under budgets?
- **Approximate** monotonicity — can we bound the size of violations?
- How often do these failures arise in **deployed** systems?