# A non-parametric approach for estimating valuation distributions in Second Price Auctions

Sourav Mukherjee

University of Florida

November 17, 2021

# Plan for the talk

1. Introduction
2. Details of Non-parametric approach
3. Numerical illustrations
4. Analysis of Xbox data
5. Summary

Joint work with Drs. Kshitij Khare and Rohit Patra

# Introduction: What is a Second Price Auction?

- There is a reserve price set by the seller at the beginning of the auction.
- Bids are made sequentially.
- The person with the highest bid (often called, winning bidder) wins the item if the corresponding bid is higher than the reserve price.
- But he/she pays the price of the maximum of the second highest bid and the reserve price, plus a small increment (e.g., $0.01).

# Bids based on True consumer valuation distribution $F$

- Assume, Bidders can only bid once.
- Unlike Final price auctions, bidders are more comfortable in bidding their true private value in a second price auction.
- Widely used on E-commerce platforms e.g. eBay, Rokt, and online ad exchanges such as Xandr.

# Example of a second price auction

- Consider A, B, C, bidding on a certain item with reserve/starting price \$1.
- A, B, C are willing to pay \$3, \$5, and \$10 respectively ($A \to B \to C$).
- A bids \$3 $\longrightarrow$ Current selling price = \$1.01, Current leader is A.
  B bids \$5 $\longrightarrow$ Current selling price = \$3.01, Current leader is B.
  C bids \$10 $\longrightarrow$ Current selling price = \$5.01, Current leader is C.
- Assuming no more bids $\implies$ C is the Winner, and pays final selling price of \$5.01.
- Note that, the highest bid of \$10 (C's bid) is always unobserved.
- Though all bids except highest bid are observed above (\$3 and \$5), this is not true in general.

# Another Example

- Some of the bids apart from the highest bid may also be unobserved.
- Take the same set of bidders (A: \$3; B: \$5; C: \$10).
- However, assume they place their bids in the order $B \to C \to A$ (instead of $A \to B \to C$).
- B bids \$5 $\longrightarrow$ Current selling price = \$1.01, Current leader is B.
  C bids \$10 $\longrightarrow$ Current selling price = \$5.01, Current leader is C.
  A enters $\longrightarrow$ Unable to place a bid, as Current selling price (\$5.01) $>$ \$3.
- A does not bid, and \$3 is unobserved. Also, the highest bid of \$10 (C's bid) remains unobserved. Here, only the bid \$5 is observed.

# Why is it important to learn the valuation distribution of the population of buyers/consumers?

- Helps to determine the optimal reserve price which maximizes the revenue.

  Too low $\implies$ adversely affect the profit
  Too high $\implies$ item may remain unsold.

  Hence, choice of right reserve price is crucial for the seller.

- Companies can implement other marketing strategies.

# Difference with the existing approaches in the literature

- Previous non-parametric approaches [George & Hui, AOAS, 2011] only use the final selling price and assume knowledge of the total number of bidders.

- For many online second price auctions, the seller does not see all the bids or the actual number of bidders.

- Has access only to the current selling price values (also defined as Observed bids) throughout the course of the auction.

- Our approach uses the entire sequence of current selling prices and not just the final selling price to get a better estimate.

# Notations

$r \rightarrow$ reserve price / initial asking price.

$N \rightarrow$ (random) number of potential bidders.

$M \rightarrow$ number of observed bids.

$\tau \rightarrow$ length of total bid window (also called, Auction window).

$\lambda \rightarrow$ arrival rate of the bidders.

(In our case, we assumed constant rate of arrival throughout the course of the auction).

# Notations (Cont'd.)

$\{X_i\}_{i=1}^{M} \;\rightarrow\;$ observed bids, $\qquad X_0 \;\rightarrow\;$ reserve price $r$.

$X_M \;\rightarrow\;$ Highest observed bid $=$ Second-highest bid.

# Notations (Cont'd.)

$\{X_i\}_{i=1}^{M} \rightarrow$ observed bids, $\quad X_0 \rightarrow$ reserve price $r$.

$X_M \rightarrow$ Highest observed bid = Second-highest bid.

$T_i \rightarrow$ time past between $i^{th}$ and $(i+1)^{th}$ observed bid, $1 \leq i \leq M-1$

$T_0 \rightarrow$ time we had to wait for the first observed bid.

$T_M \rightarrow$ time remaining after the last observed bid until bid window $(\tau)$ closes.

i.e. $T_M = \tau - \sum_{i=0}^{M-1} T_i$.

# Non-parametric Methodologies

- **Target:**

  Estimate $F$, the true valuation distribution of the population of buyers.

- **Assumption:**

$$\text{Number of potential bidders } (N) \text{ in any auction } \sim \text{ Poisson}\,(\lambda\tau)$$

# Joint density of $\{X_i\}_{i=1}^{M}$, $\{T_i\}_{i=0}^{M-1}$, and $M$

For any auction, there can be only two possibilities:

- **When the item has been sold**.

$$\exp\left(-\lambda\tau\right)\lambda^{M+1}2^M T_0\Big(1-F(X_M)\Big)\exp\left(\lambda\sum_{i=0}^{M}F(X_i)\,T_i\right)\left(\prod_{i=1}^{M}f(X_i)\right)^{1_{\{M>0\}}}$$

- **When the item has not been sold**.

$$M=0,\quad X_0=r=X_M,\quad T_0=\tau=T_M.$$

$$\exp\left(-\lambda\tau\right)\exp\left(\lambda\sum_{i=0}^{M}F(X_i)\,T_i\right)\left(\prod_{i=1}^{M}f(X_i)\right)^{1_{\{M>0\}}}$$

# Likelihood Function for Second Price Auction

- $K$ many independent second price auctions of identical copies of an item (with possibly random reserve prices)

- Observed data: $\left\{ \left( r^k, T_0^k \right) \ , \ \left\{ \left( X_m^k, T_m^k \right) \right\}_{m=1}^{M^k} \right\}_{k=1}^{K}$.

  $r^k \ \rightarrow$ reserve price the $k^{th}$ auction, $1 \leq k \leq K$.
  $M^k \ \rightarrow$ number of observed bids for the $k^{th}$ auction, $1 \leq k \leq K$.

- Above data set has $(L + K)$ many pairs, where, $L = \sum_{k=1}^{K} M^k =$ total number of observed bids from all the auctions.

- $(z_1, \tilde{t}_1), (z_2, \tilde{t}_2), \ldots\ldots, (z_{L+K}, \tilde{t}_{L+K})$ is a rearrangement of the Observed data such that $z_1 < z_2 < \ldots < z_{L+K}$.

# Likelihood Function for Second Price Auction

Define,

- $\bar{S}$ = Set of Ranks of the highest observed bids $\left(X^k_{M^k}\right)$ in $\underset{\sim}{Z} = (z_1, z_2, \ldots, z_{L+K})^T$, for those auctions where the item is sold.

- $\underset{\sim}{U} := \ (u_1, u_2, \ldots, u_L)^T$, be the vector of ranks of all the observed bids from all auctions in $\underset{\sim}{Z}$, such that $u_1 < u_2 < \ldots < u_L$.

- Notice that, $u_L$ is the rank of the highest observed bid (from all the auctions) in $\underset{\sim}{Z}$.

- However, $u_L$ may not be equal to $(L + K)$, since there could be some auctions where the item is unsold and the corresponding reserve prices are higher than the highest observed bid.

# Reparameterization of F

- $G(z_i) = 1 - F(z_i) \; \forall \; 1 \leq i \leq (L+K), \qquad G(z_0) = 1, \text{ with } z_0 = 0.$

  [Note, $G(z_i) \leq G(z_{i-1})$]

- $\underset{\sim}{\theta} := (\theta_1, \theta_2, \ldots, \theta_{L+K})^T$, defined as

  $$\theta_i = \frac{G(z_i)}{G(z_{i-1})}, \quad \forall \; 1 \leq i \leq (L+K).$$

  [Note, $\theta_i \leq 1$]

- Can retrieve $F(z_i)$ at any time using the equality:

  $$F(z_i) = 1 - G(z_i) = 1 - \prod_{j=1}^{i} \theta_j \;, \qquad \forall \; 1 \leq i \leq (L+K).$$

# Estimation of $\lambda$ - Generalized method of moments

- Computed a consistent estimator of $\lambda$, and used it in the likelihood expression.

- Can be proved: $E\left[\bar{M}\right] = 2E\left[\sum_{k=1}^{N} \frac{1}{k}\right] - 2,$

  where, $\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M^k =$ Mean number of observed bids.
  and, $N =$ Number of bidders in an auction.

- $\gamma := \lim_{n\to\infty} \left(-\ln n + \sum_{k=1}^{n} \frac{1}{k}\right) = 0.5772156649 =$ Euler's constant.

- For large values of $n$, $\sum_{k=1}^{n} \frac{1}{k}$ can be well approximated by $(\ln n + \gamma)$.

# Consistent estimator of $\lambda$

- $E\left[\sum_{k=1}^{N}\frac{1}{k}\right] \approx E\left[ln\ \ N+\gamma\right] = E\left[ln\ \ N\right]+\gamma.$

  which gives,

$$
\begin{aligned}
E\left[\bar{M}\right] &= 2E\left[\ln N\right] \ + \ 2\gamma \ - \ 2 \\
&= 2\sum_{n=2}^{\infty}\frac{\ln n\ \exp(-\lambda\tau)\ \frac{(\lambda\tau)^n}{n!}}{1-\exp(-\lambda\tau)-\lambda\tau\exp(-\lambda\tau)} \ + \ 2\gamma \ - \ 2 \\
&:= g\left(\lambda\right)
\end{aligned}
$$

- $\hat{\lambda} = g^{-1}\left(\bar{M}\right)$

- $\hat{\lambda}$ is consistent.

$$
\begin{aligned}
\bar{M} \to g(\lambda) \text{ a.s., } g^{-1} \text{ continuous } &\implies g^{-1}\left(\bar{M}\right) \ \to \ \lambda \text{ a.s.} \\
&\implies g^{-1}\left(\bar{M}\right) \ \underset{\to}{P} \ \lambda
\end{aligned}
$$

# Likelihood Function for Second Price Auction

Then, the likelihood can be written as:

$$L\left(\underset{\sim}{\theta} \mid \underset{\sim}{Z}\right)$$

$$= \hat{\lambda}^{L+|K_s|} \, 2^L \left(\prod_{i \in K_s} T_0^i\right) \left(\prod_{i \in \bar{S}} \prod_{j=1}^{i} \theta_j\right) \exp\left(-\hat{\lambda} \sum_{i=1}^{L+K} \tilde{t}_i \left(\prod_{j=1}^{i} \theta_j\right)\right)$$

$$\times \left[\left(1 - \prod_{j=1}^{u_1} \theta_j\right) \times \prod_{l=2}^{L} \left\{\left(1 - \prod_{j=u_{l-1}+1}^{u_l} \theta_j\right) \prod_{j=1}^{u_{l-1}} \theta_j\right\}\right]^{1_{\{L>0\}}}$$

where, $K_s =$ Set of auction indexes for which the item has been sold.

# Coordinate-wise Maximization algorithm

- Used Coordinate ascent on log-likelihood function to successively maximize along coordinate directions of $\underset{\sim}{\theta}$ to find the maximizer of the function.

- Recall, $\bar{S} =$ Set of Ranks of the highest observed bids $\left(X^k_{M^k}\right)$ in $\underset{\sim}{Z} = (z_1, z_2, \ldots, z_{L+K})^T$, for those auctions where the item is sold.

- Define,

$$\begin{aligned} Q_k &:= \left\{ j \in \bar{S} : j \geq k \right\} \quad , \quad 1 \leq k \leq L+K \\ &= \text{ Set of } j \in \bar{S} \text{ which are greater than or equal to } k. \\ &\left(e.g., Q_1 = \bar{S}\right) \end{aligned}$$

Recall,

- $K =$ Number of independent second price auctions of identical copies of an item.
- $L = \sum_{k=1}^{K} M^k =$ total number of observed bids from all the auctions.
- $u_L =$ rank of the highest observed bid (from all the auctions) in $\underset{\sim}{Z}$.
- $u_L$ may not be equal to $(L + K)$. i.e., in general, $u_L < L + K$.

  since, there could be some auctions where the item is unsold and the corresponding reserve prices are higher than the highest observed bid.

# Coordinate-wise Maximization Steps for $\underset{\sim}{\theta}$

Recall, $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_{L+K})^T$.

- When $1 \leq k \leq u_L$,

  The derivative of the log-likelihood w.r.t. $\theta_k$ is a quadratic polynomial in $\theta_k$.

$$A_k C_k \theta_k^2 \; - \; \left(A_k + B_k C_k + 1_{\{L>0\}} C_k\right) \theta_k \; + \; B_k \; = \; 0$$

which gives,

$$\theta_k = \frac{\left(A_k + B_k C_k + 1_{\{L>0\}} C_k\right) \pm \sqrt{\left(A_k + B_k C_k + 1_{\{L>0\}} C_k\right)^2 - 4 A_k B_k C_k}}{2 A_k C_k}$$

- However, the root with "$+$" sign is $> 1$.

# Coordinate-wise Maximization Steps for $\underset{\sim}{\theta}$

Hence,

$$\hat{\theta}_k = \min\Bigg\{1,$$

$$\frac{\left(A_k + B_k C_k + 1_{\{L>0\}} C_k\right) - \sqrt{\left(A_k + B_k C_k + 1_{\{L>0\}} C_k\right)^2 - 4 A_k B_k C_k}}{2 A_k C_k}\Bigg\}$$

$$\text{where,} \qquad A_k = \hat{\lambda} \sum_{i=k}^{L+K} \tilde{t}_i \left( \prod_{\substack{j=1 \\ j \neq k}}^{i} \theta_j \right) \; > \; 0$$

$$B_k = |Q_k| + 1_{\{L>0\}} \left( L - l_k \right) \; > \; 0$$

$$C_k = \prod_{\substack{j=u_{l-1}+1 \\ j \neq k}}^{u_l} \theta_j \; > \; 0$$

with, $l_k = l \quad \text{if} \quad u_{l-1} + 1 \leq k \leq u_l$ , for $k \in \{1, 2, \ldots, u_L\}, l \in \{1, 2, \ldots, L\}$.
and, $l_k = L \quad \text{if} \quad u_L + 1 \leq k \leq L + K$.

# Coordinate-wise Maximization Step for $\underset{\sim}{\theta}$

- When $u_L + 1 \leq k \leq L + K$,

The derivative of the log-likelihood w.r.t. $\theta_k$ is a linear expression in $\theta_k$.

$$- A_k \; + \; \frac{|Q_k|}{\theta_k} = 0$$

which gives,

$$\hat{\theta}_k = \min\left\{1 \; , \; \frac{|Q_k|}{A_k}\right\}$$

# Initial Estimate of $\theta$ and $F$ for the Coordinate-wise Maximization Algorithm

- Construct the initial estimate of F ($F_{init}$) based on the empirical distribution functions of the final selling prices and the first observed bids, respectively, of those auctions whose reserve prices are considerably lower. ($<$ an user specified cutoff)

- The initial value of $\underset{\sim}{\theta}$:

$$\theta_i^{(0)} = \frac{1 - F_{init}(z_i)}{1 - F_{init}(z_{i-1})}, \quad 1 \leq i \leq L + K$$

- Get the MLE of $\underset{\sim}{\theta}$ as $\hat{\underset{\sim}{\theta}}_{MLE}$

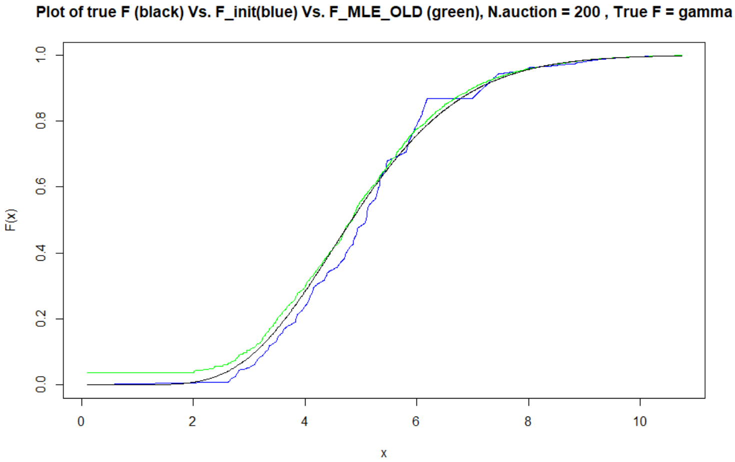
- Retrieve $F_{MLE}(z_i) = 1 - \prod_{j=1}^{i} \hat{\theta}_{j,\ MLE}$ , $\forall\ 1 \leq i \leq (L + K)$. Do linear interpolation for all other values in the x - coordinate.

# Instability of $F_{MLE}$ near 0

- Near 0 (in the region below min(set of all observed bids)), $F_{MLE}$ seems to overestimate the truth (observed through simulations).
- Instability of non-parametric MLE near the boundaries is common.

# Instability of $F_{MLE}$ near 0

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Gamma (shape = 10, rate = 2) distribution.
- min(set of all observed bids) = 2.015451.



# Constrained approach

- Initial estimate is more stable in this region [below min(set of all observed bids)].

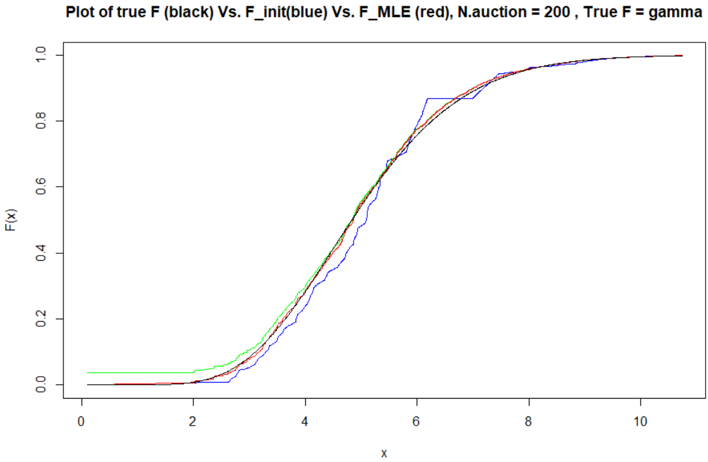
- Let, $i_0$ be the index such that $z_{i_0} = \min(\text{set of all observed bids})$.

- For any $1 \leq k \leq i_0$, take $\hat{\theta}_{k,\ MLE} = \theta_k^{(0)}$, where $\underset{\sim}{\theta}^{(0)}$ is the initial value of $\theta$.

- For any $(i_0 + 1) \leq k \leq (L + K)$, take $\hat{\theta}_{k,\ MLE}$ as the usual MLE got from the coordinate ascent algorithm.

# Constrained approach

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Gamma (shape = 10, rate = 2) distribution.
- min(set of all observed bids) = 2.015451.



# Numerical illustrations

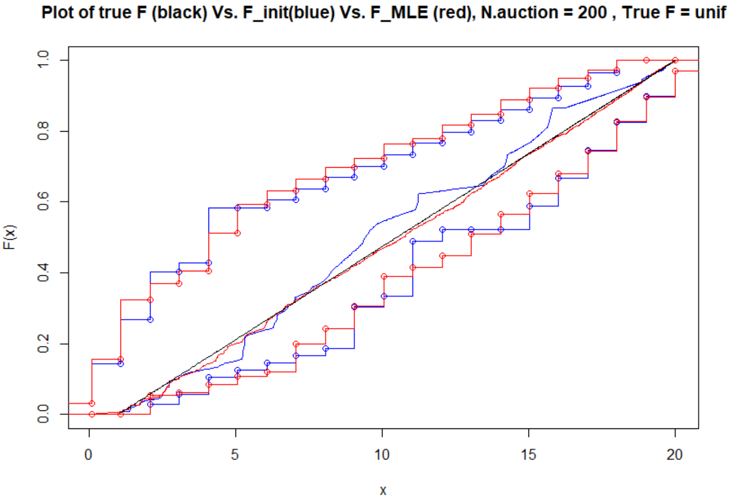
- Simulation studies with various choices of true underlying valuation distribution.
- e.g., Uniform, Pareto, Gamma, Beta, and Piecewise uniform.
- Generate $K$ (e.g., $K = 200,\ 300$) many independent auctions.
- Use the proposed methodology to get $F_{init}$ and $F_{MLE}$.

# Confidence regions

- Construct confidence regions, called HulC (Kuchibhotla et al., 2021), using the convex hull of estimates constructed from subsets of the data.
- Unlike classical methods which are based on estimating the (limiting) distribution of an estimator, the HulC effectively bypasses this step.
- The validity of the HulC requires knowledge of the (asymptotic) median-bias of the estimators (computed from each subset).
- Construct only the 90% HulC confidence regions.

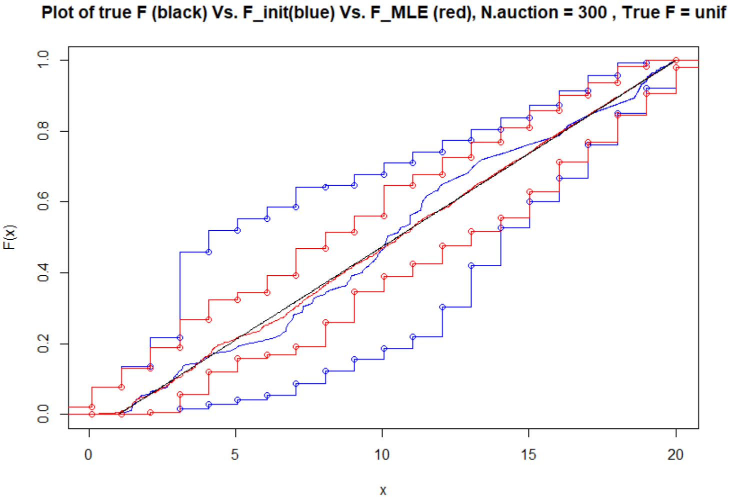
# Uniform distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True valuation distribution is Uniform$(1, 20)$.



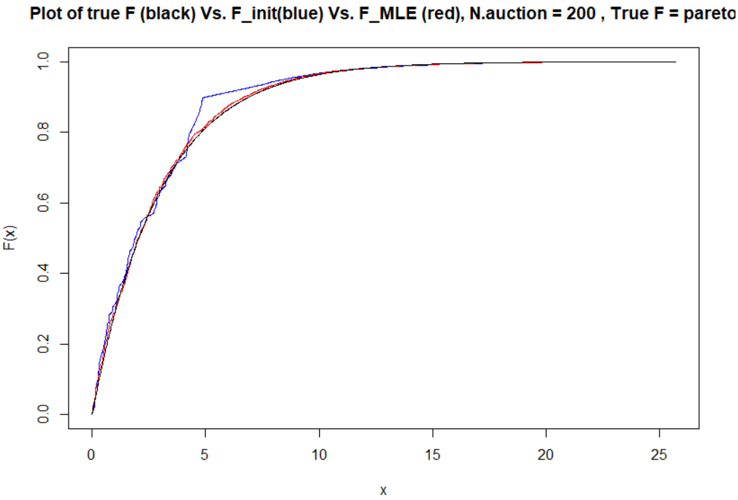
# Uniform distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True valuation distribution is Uniform$(1, 20)$.



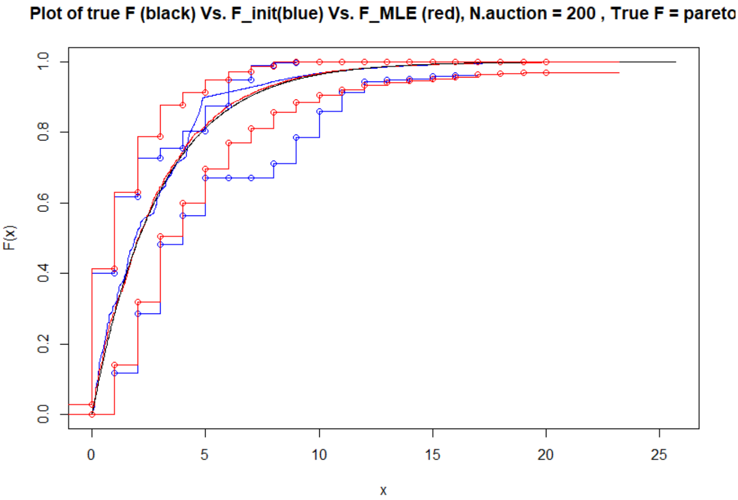
# Pareto distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Pareto with location parameter = 3, dispersion = 100.



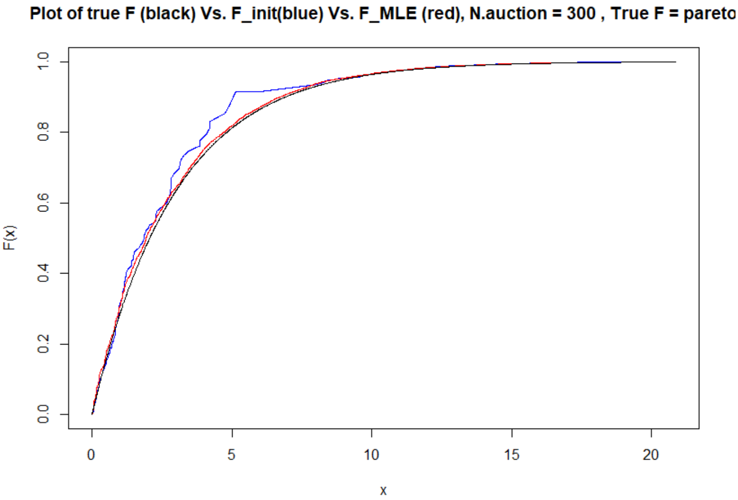
# Pareto distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Pareto with location parameter = 3, dispersion = 100.



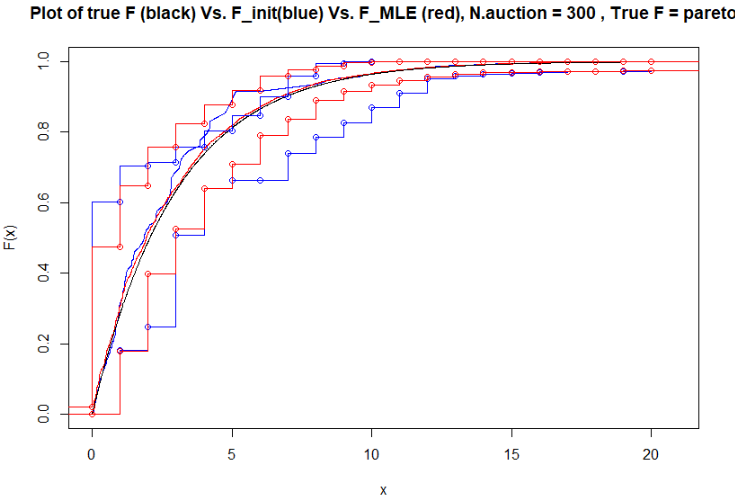
# Pareto distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is Pareto with location parameter = 3, dispersion = 100.



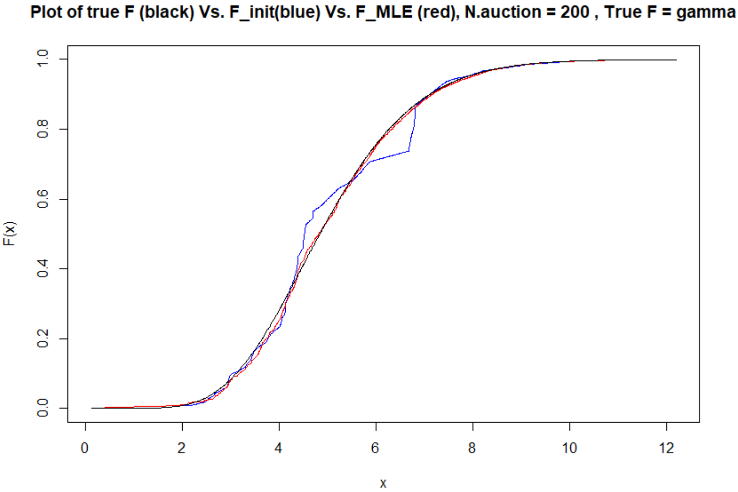
# Pareto distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is Pareto with location parameter = 3, dispersion = 100.



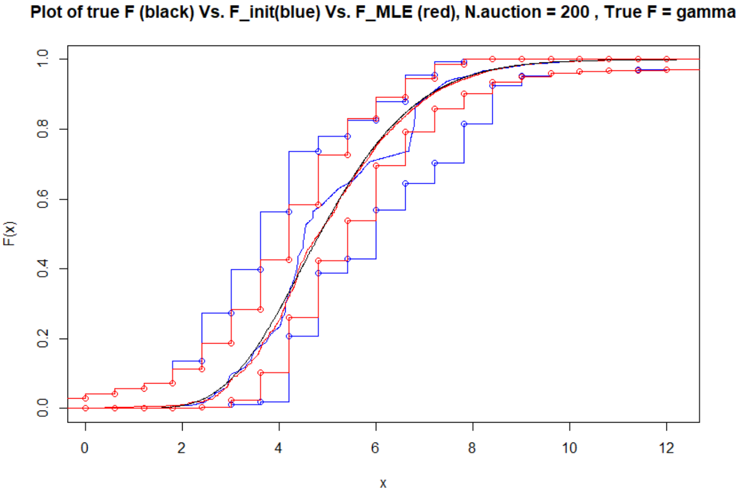
# Gamma distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Gamma (shape = 10, rate = 2) distribution.



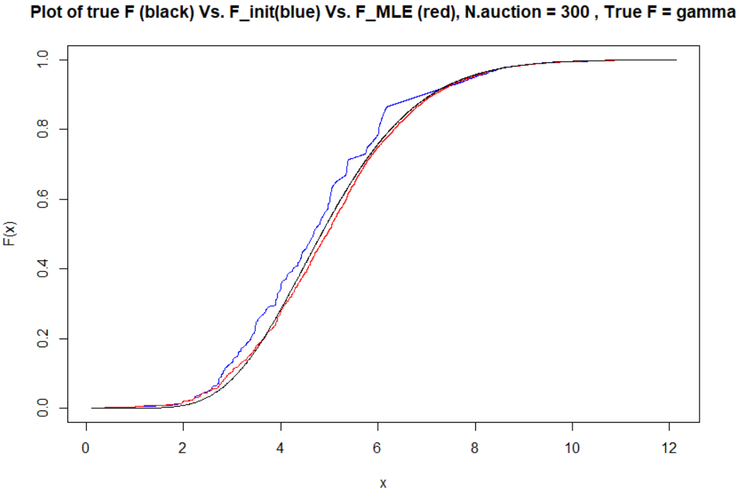
# Gamma distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Gamma (shape = 10, rate = 2) distribution.



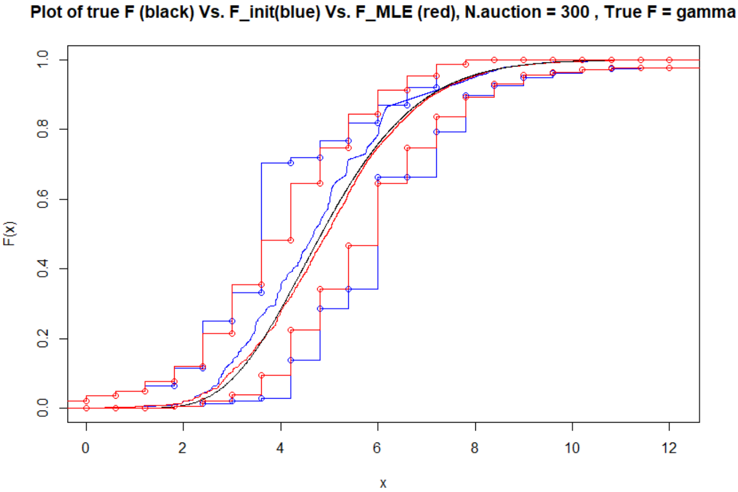
# Gamma distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is Gamma (shape = 10, rate = 2) distribution.



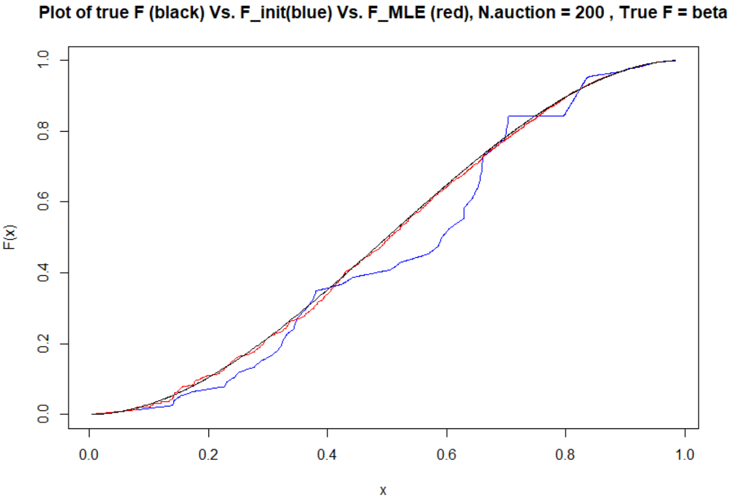
# Gamma distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is Gamma (shape = 10, rate = 2) distribution.



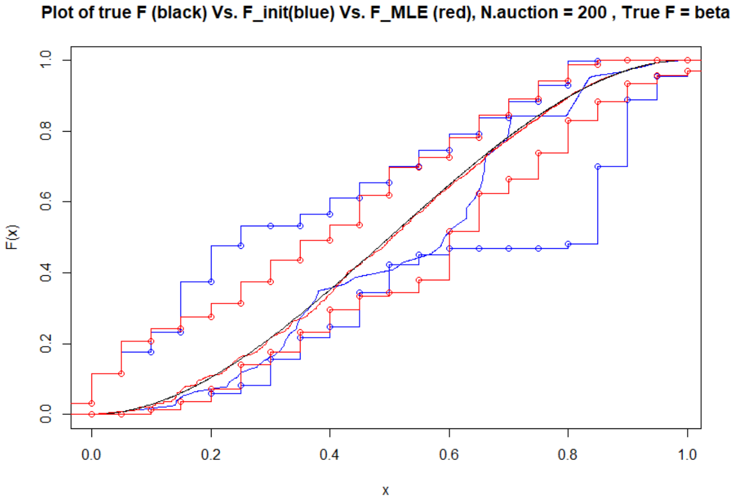
# Beta distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Beta(2, 2) distribution.



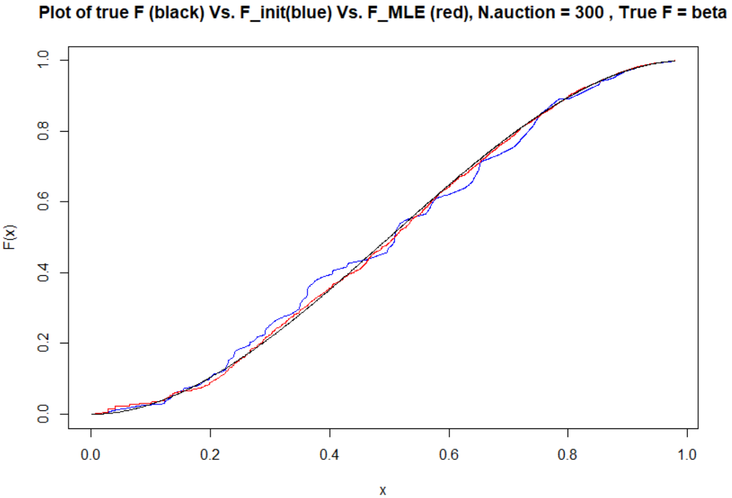
# Beta distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is Beta(2, 2) distribution.



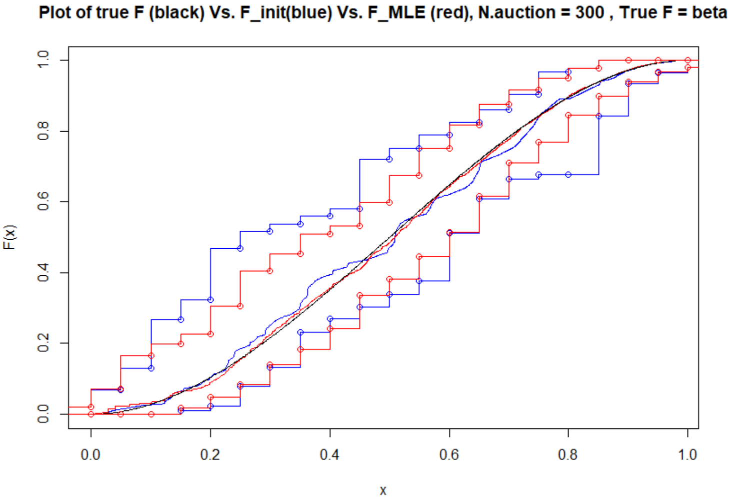
# Beta distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is Beta(2, 2) distribution.



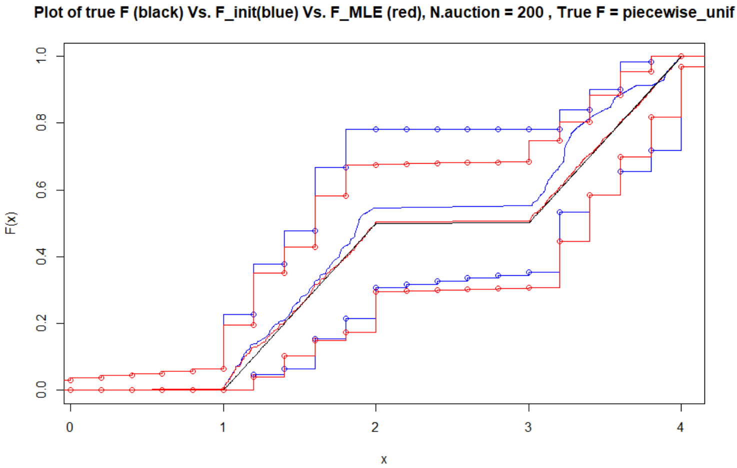
# Beta distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is Beta(2, 2) distribution.



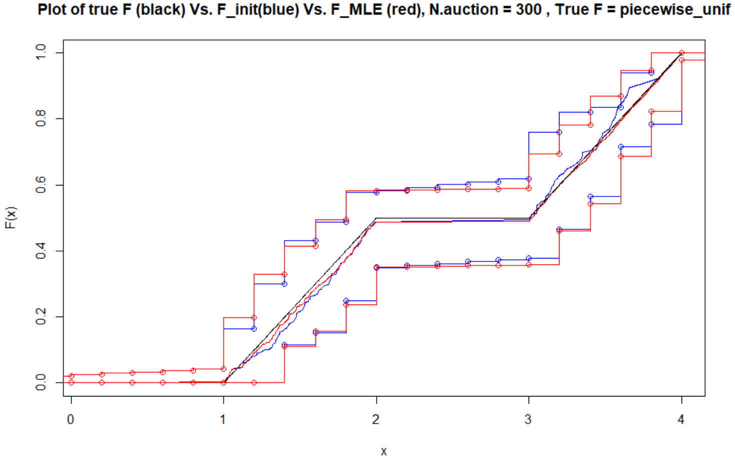
# Piecewise uniform distribution - illustration

- Number of independent auctions = 200; Auction window = 100 unit.
- True F is a piecewise uniform distribution, where two parts are Uniform$(1, 2)$ and Uniform$(3, 4)$.



# Piecewise uniform distribution - illustration

- Number of independent auctions = 300; Auction window = 100 unit.
- True F is a piecewise uniform distribution, where two parts are Uniform$(1, 2)$ and Uniform$(3, 4)$.



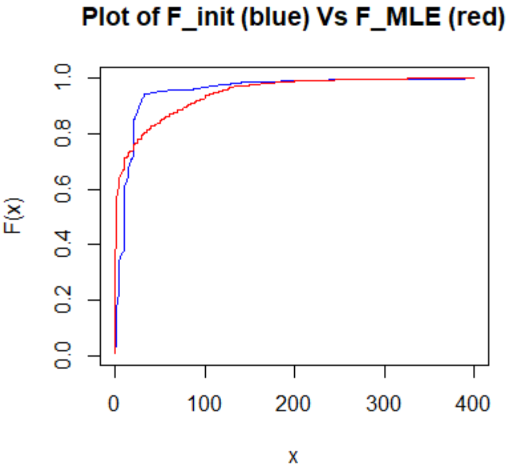
# Analysis of Xbox data - 7 day auctions

- 93 auctions in the online auction data set of Xbox game consoles on eBay [Modeling Online Auctions, Jank & Shmueli, Wiley & sons., 2010].

- Each auction lasts for 7 days.

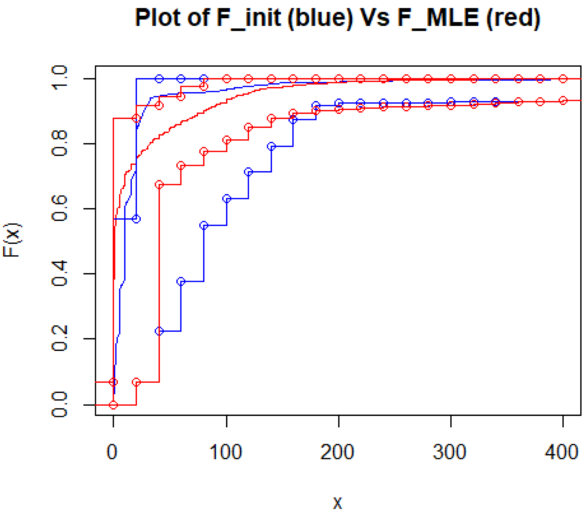
- Added a small random noise to all the bids in the data set to avoid equality of bid values from different bidders.

# Analysis of Xbox data - 7 day auctions




- Wasserstein distance between $F_{init}$ and $F_{mle} = 12.6797$.

# Analysis of Xbox data - 7 day auctions

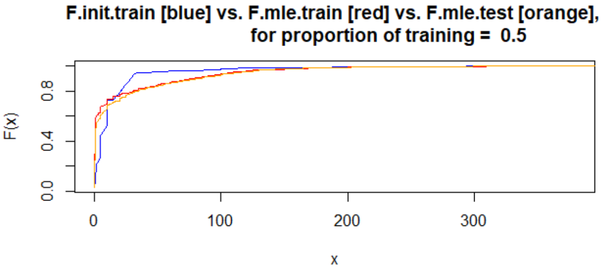


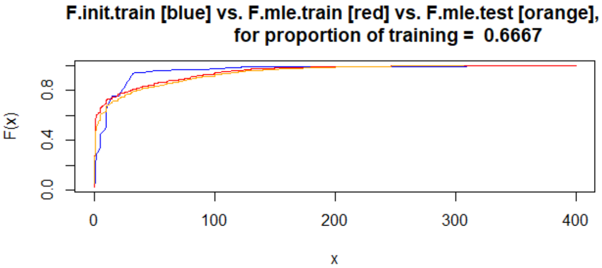
# Performance evaluation

- To evaluate the performance of both $F_{init}$ and $F_{MLE}$, we split the dataset into training and test sets (e.g., $1:1$ and $2:1$ split).

- Evaluate both $F_{init,\ train}$ and $F_{MLE,\ train}$ based on the training data, and compare each of them with $F_{MLE,\ test}$ based on the test data.

# Performance evaluation



# Wasserstein distance

| **Train:Test** | **W(F.init.train, F.mle.test)** | **W(F.mle.train, F.mle.test)** |
|---|---|---|
| 1:1 | 8.828918 | 3.532927 |
| 2:1 | 8.883618 | 3.866656 |

Table: Wasserstein distance between $F.mle.test$ and both of $F.init.train$, $F.mle.test$, averaged over 1000 replications of the random splits with same proportion.

# Summary

- Proposed non-parametric approach uses the current selling price values throughout the course of the auctions and not just the final selling prices.

- Significantly better performance than estimators based on just final selling prices or first observed bids.

- It does not require knowledge of all the bids or the total number of bidders.

- It is free of tuning parameters.

- The resulting estimate $F_{MLE}$ can be used by the seller to compute the optimal reserve price which maximizes revenue, and also for determining other marketing strategies.

# Thank You!