PhD in Statistics, Department of Statistics

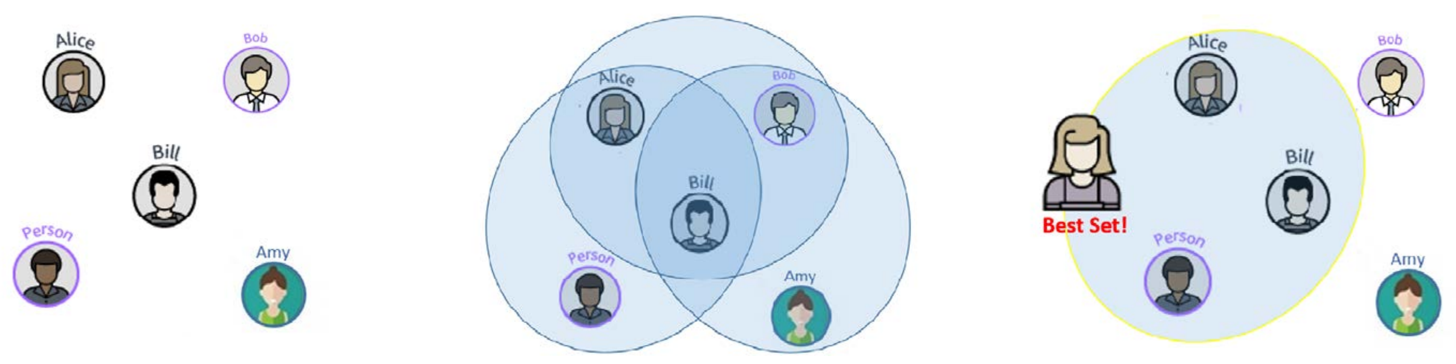
Sketching stochastic valuation functions

Student: Yiliu Wang

Supervisor: Milan Vojnovic

Motivations

A typical example is the crowding-sourcing platform.



Workers are assigned to teams based on scores.

- **Items of interest** are workers.
- The agent **approximate** possible team performances
- Evaluations can be used for comparing different sets or selecting the best team.

Other applications follow the same logic:
i) Gaming ii) Digital advertising iii) Online searching

Problem setup

Assume we have

- **Indep random variables** $X_1,\ldots X_n$ with distributions $p_1,\ldots p_n$
- **Set utility function** $u(S) = E[f(X_i, i \in S)]$
  where $|S| = k$ and f is a submodular value function

Given a ground set items Ω, a set function $u: 2^{\Omega} \to R^+$ is **submodular** if u satisfies the diminishing returns property:

$$u(T \cup i) - u(T) \le u(S \cup i) - u(S) \text{ s.t. } S \subseteq T$$

We are interested in finding compact representation $q_1, \ldots q_n$ of item distributions for approximation of set utility functions.

We have two specific goals:

- **Approximate set function everywhere**: Find a sketch set function v s.t.
  $$av(S) \le u(S) \le v(S) \ \forall S \subseteq \Omega$$
- **Best set selection:** Find set A of size k s.t. for some const c,
  $$u(A) \ge c \max\{u(S): |S| = k\}$$

Algorithms and main results

Discretization algorithm

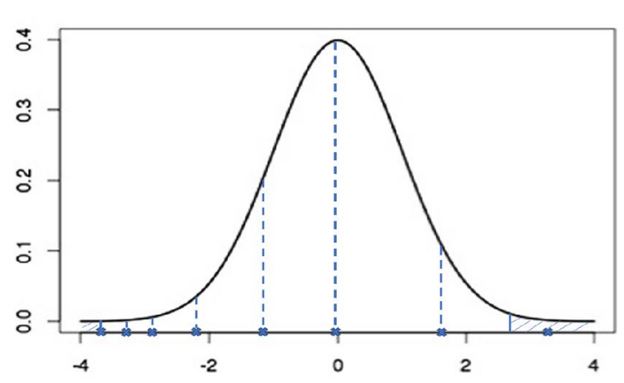
For each $i \in \Omega$:
Let τ_i be the top $1 - \varepsilon$ quantile X_i and

$$H_i = E[f(X_i \mid X_i > \tau_i)]$$

- Let $\widehat{X}_i$ be a new random variable s.t. $\widehat{X}_i = X_i$ if $X_i \le \tau_i$ and $\widehat{X}_i = f^{-1}(H_i)$ o.w.
- Assign values of $\widehat{X}_i \le a\tau_i$ to 0
- Transform $\widehat{X}_i$ using an exponential binning of the interval $[a\tau_i, \tau_i]$.

Exponential Binning: Partition the range into l intervals $I_1, \ldots I_l$ where $l = log_{1-\epsilon}(a)$ and map each value in a bin to the lower boundary of the bin.

$$I_j = \left[\frac{a\tau}{(1-\varepsilon)^{j-1}}, \frac{a\tau}{(1-\varepsilon)^{j}}\right]$$



Main results

Our work is the first step towards understanding approximation of stochastic valuation functions everywhere. Existing related work focused instead on optimization problems only, or approximation schemes using one-dimensional item value distribution representations (test scores).

A function f is **weakly homogeneous** with degree d and tolerance η over a set $\Theta \subseteq R$ if $(1/\eta)\theta f(x) \le f(\theta x) \le \theta^d f(x)$ for all x in the domain of f and $\theta \in \Theta$. Several commonly used valuation functions are weakly homogenous with degree $d = 1$ and tolerance $\eta = 1$, e.g. maximum value function.

Theorem 1 (approximate set function everywhere) Assume that f is a monotone subadditive or submodular function and is weakly homogeneuous with degree d and tolerance η over $[0,1]$. Then the discretization algorithm guarantees that for every set $S \subseteq \Omega$ such that $|S| \le k$,

$$\frac{1}{2}(1-\varepsilon)^{k-1} v(S) \le u(S) \le 2\eta \frac{1+a^d k/\varepsilon}{(1-\varepsilon)^k} v(S)$$

- This theorem implies a **constant-factor** approximation guarantee.
- The **support size** of each discretized distribution is $O((1/d)k \log k)$.

A **greedy algorithm** start from an empty set and sequentially selects items that yields the largest marginal value $u(S \cup i) - u(S)$.

Theorem 2 (best set selection) For the class of functions satisfying conditions above, and by taking $\epsilon = c/k$, the greedy algorithm has the approximation ratio

$$\frac{1}{4\eta}\left(1-\frac{1}{e}\right)\frac{e^{-4c/(1-c)}}{1+c} \quad \text{arbitrarily close to } \frac{1}{4\eta}\left(1-\frac{1}{e}\right) \text{ by taking } c \text{ small enough}$$

Extensions

We further show that similar approximation guarantees hold under other conditions so the results extend to a wide range of functions. We have the following two corollaries of Theorem 1.

A monotone subadditive and concave function f on R_+^n is said to have an extension on R^n if there exists a function f^* on R^n s.t. $f^*(x) = f(x)$ for all $x \in R_+^n$ and monotone subadditive and concave.

Corollary 1 Assume that f is **extendable concave**, then the algorithm guarantees that for every set S such that $|S| \le k$ we have

$$\frac{1}{2}(1-\varepsilon)^{k-1} v(S) \le u(S) \le 2\frac{1+ak/\varepsilon}{(1-\varepsilon)^k} v(S)$$

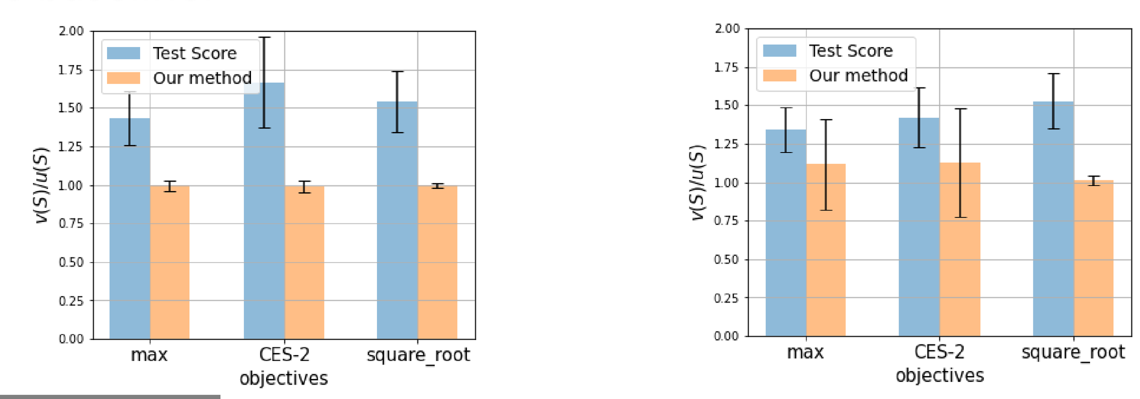
Corollary 2 Similar approximation guarantees can be established for:

- **Coordinate-wise** weakly homogeneous functions
- Same set of functions with random variables **transformed** by a continuous and strictly increasing function.

Numerical results

Synthetic data

Three types of set utility functions and two parametric families of item value distributions (left: exponential, right: Pareto). Our sketch outperforms the test score baseline.



Real-world data

Three real-world datasets: YouTube, Stack Exchange and New York Times. Our sketch provides good approximation in most cases.