Data Generating Process

Objective for DGPs:

- Create multiple clusters, where distributions of covariates are different across clusters
- Each cluster has a different arm that produces highest reward
- Generate "lumpy" reward functions that cannot be straightforwardly recovered by a linear model
- Allow levers to move:
  - number of covariates used to define clusters
  - relative size of clusters
  - *Heterogeneity ratio* (value of best contextual/best fixed policy)

Data Generating Process

Requirements for DGPs:

- The difference between the best contextual policy & the control is fixed across DGPs
  - → Differences in power curves between DGPs are based on ability of agent to learn the DGP, not differences in effect sizes
- The best fixed arm is always the same arm across DGPs

Data Generating Process

Generate Baseline Dataset ($N = 10000, p = 15$)

- Parameter
  - Number of useful covariates $p' \in \{3, 5, 10\}$
  - Largest cluster size ratio $c \in \{0.4, 0.6, 0.8\}$
  - Heterogeneity ratio $h \in \{1.05, 1.5, 1.95\}$
- Generate large covariate matrix X using correlated multivariate normal distribution, covariance matrix generated from $\frac{Beta(2,2)-0.5}{2}$
- Use iterative KNN to group covariates observed into $k = 3$ clusters with cluster size $[N * c, \frac{N*(1-c)}{2}, \frac{N*(1-c)}{2}]$, where $c =$ largest cluster size ratio

Data Generating Process

Reward Generation

- Generate reward for best arm for each cluster $c_i, i = 1, .., 3$,

$$R_{w_{best,i}, c_i} = 0.6 - X_{1,c_i}$$

- Reward for 2nd best arm for each cluster $c_i, i = 1, .., 3$,

$$R_{w_{best_2,i}, c_i} = R_{w_{best,i}, c_i} + \epsilon$$

  where $\epsilon \sim \mathcal{N}(\mu, 0.01), \mu <= 0$.

- Reward for the rest of the arm

$$R_{w,c_i} = -X_{1,c_i} \text{ for } w \notin \{w_{best,i}, w_{best_2,i}\}, i = 1, .., 3,$$

- We vary μ to search for the level of desired heterogeneity ratio h.

Data Generating Process

Note

- 0.6 was chosen to simulate an average treatment effect between best-fixed policy and control policy; assume control policy is not the best arm for any cluster. (See PAP for justification of magnitude.)
- In the simulations, arm 0 is chosen as the best arm for the largest cluster. It is also chosen as the 2nd best arm for the other two clusters to ensure it is the best-fixed policy.