Problem Formulation

- Generalization of median
- Unlike mean, medoid is inside dataset
- Special instance of AMO framework, also works for k-NN

Computation → Estimation

- Estimate $\hat{\theta}_i$ via random sampling
- RAND: estimate each $\hat{\theta}_i$ to same degree of accuracy
  $\hat{\theta}_i = \frac{1}{|J|} \sum_{j \in J} d(x_i, x_j)$
- Medoid Bandit (Med-dit): sample adaptively (UCB) [1]

Our contributions [4]

- Naïve bandit reduction to statistical estimation ignores structure of problem
- Can overcome via correlating our sampling
  Sample rows of $D, D_{i,j} = d(x_i, x_j)$
- Need to prove $\hat{\theta}_1 < \hat{\theta}_i$
  Control $\hat{\theta}_1, \hat{\theta}_i$ instead of $\hat{\theta}_1, \hat{\theta}_i$

Simulation Results

<table>
<thead>
<tr>
<th>Dataset, Metric</th>
<th>n, d</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq 20k, $\ell_1$</td>
<td>20k, 28k</td>
</tr>
<tr>
<td>RNA-Seq 100k, $\ell_1$</td>
<td>109k, 28k</td>
</tr>
<tr>
<td>Netflix 20k, cos</td>
<td>20k, 18k</td>
</tr>
<tr>
<td>Netflix 100k, cos</td>
<td>100k, 18k</td>
</tr>
<tr>
<td>MNIST Zeros, $\ell_2$</td>
<td>6424, 784</td>
</tr>
</tbody>
</table>

Theorem Statement

- Assumption: $d(x_i, x_j) - d(x_i, x_j)$ is $\sigma_1$-subgaussian
- Theorem: $\text{corrSH} [4]$ identifies the medoid with probability at least $1 - \delta$ after computing
  $$T = O\left(\sigma^2 \log n \log \left(\frac{\log n}{\delta}\right) \max_{i \neq j} \left| \frac{\sigma_1}{\sigma_2} \right|^2 \right)$$
  distance evaluations

Another Application: k-NN

- Works for any $\ell_p$ (separable) distance
- Randomly rotate data for better subgaussian constant
- $O(n^2 \log n)$ time under distributional assumption
- 100x gain in theory on ImageNet, 25x wall clock speedup

Summary

- Convert computational problem to statistical estimation
- Fast randomized algorithms
- Incorporating structure of the computational problem in this reduction can yield massive gains

References