Ultra Fast Medoid Identification via Correlated Sequential Halving

Tavor Z. Baharav, David N. Tse
{tavorb,dntse}@stanford.edu

Problem Formulation

$x_1, \dots, x_n \in \mathbb{R}^d \qquad i^* = \arg\min_{i \in [n]} \theta_i \qquad \theta_i \triangleq \frac{1}{n} \sum_{j=1}^{n} d(x_i, x_j)$

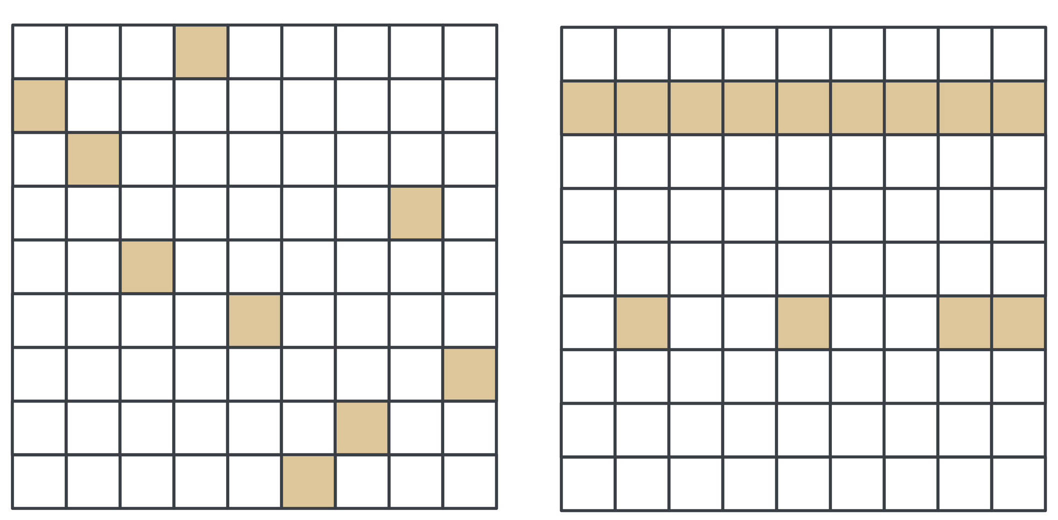
- High dimensional generalization of median
- Unlike mean, medoid is inside dataset

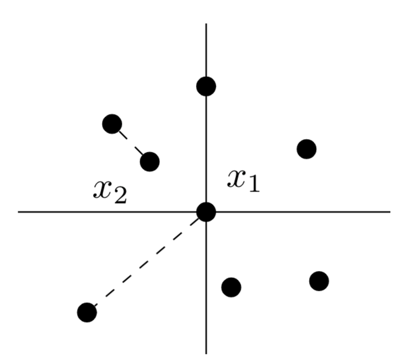
Computation → Estimation: Bandits!

- Estimate θ_i via random sampling
  $\mathbb{E}\{d(x_i, x_J)\} = \theta_i \qquad J \sim \text{Unif}([n])$
- RAND: estimate each θ_i to same degree of accuracy
  $\hat{\theta}_i = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} d(x_i, x_j)$
- Medoid Bandit (Med-dit): sample adaptively (UCB) [1]
  - Can we do better than UCB?

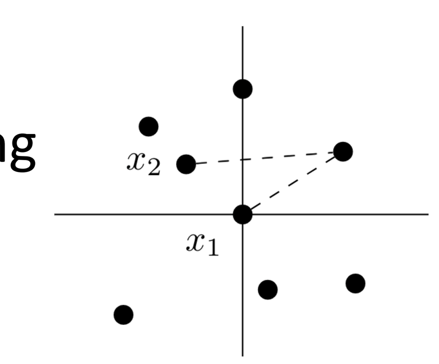
Intuition

- UCB ignores structure of problem: consider dist matrix D

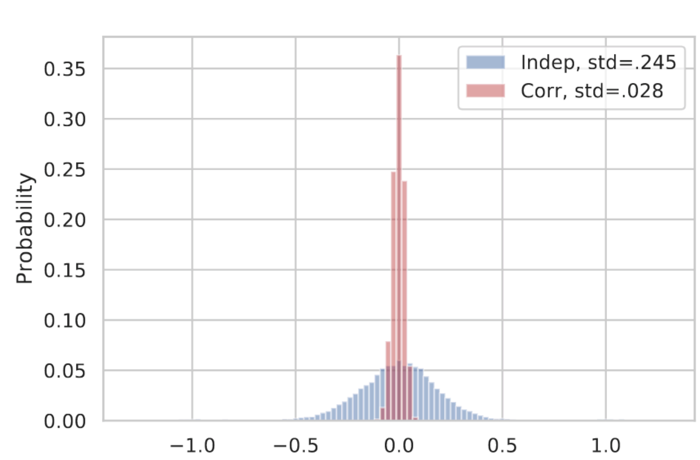




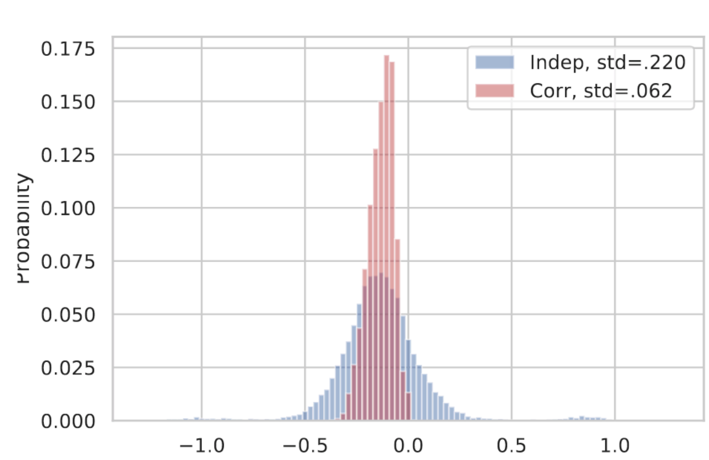
- Can overcome via *correlating* our sampling
  - Sample rows of D, $D_{i,j} = d(x_j, x_i)$



- Need to prove $\widehat{\theta_i} < \widehat{\theta_1}$
  - Control $\widehat{\theta_i - \theta_1}$ instead of $\widehat{\theta_i}, \widehat{\theta_1}$



(a) Comparison of top 2 points (i=2)



(b) Comparison of top and mid (i=10000)

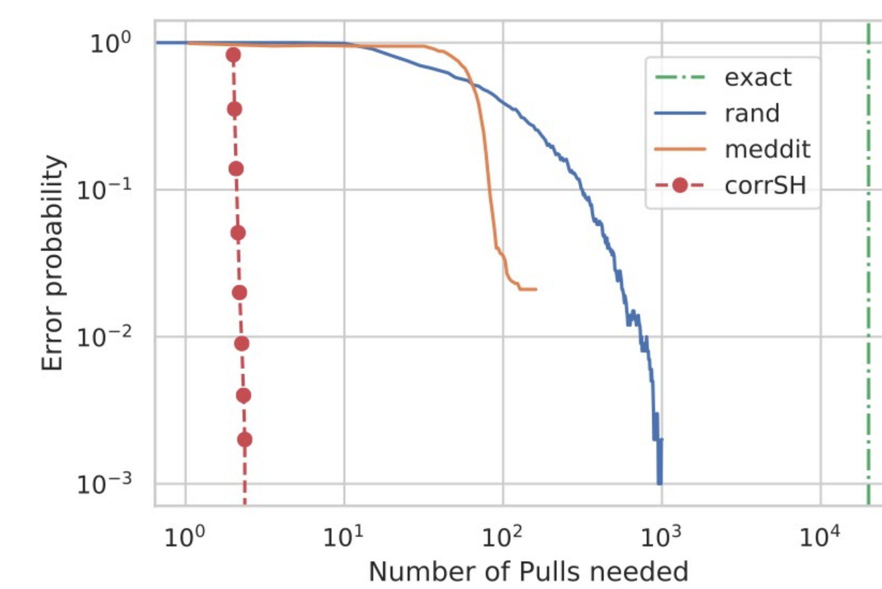
Input: Budget T
initialize $S_0 \leftarrow [n]$
for r=0 **to** $\lceil \log_2 n \rceil - 1$ **do**
  select a set $\mathcal{J}_r$ of t_r reference points uniformly at random without replacement from $[n]$ where

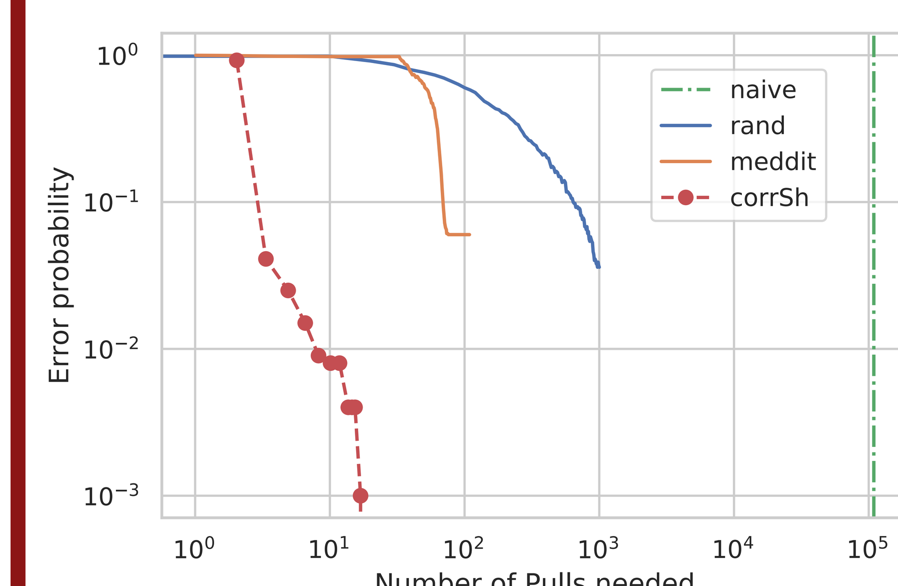
$$t_r = \left\{ 1 \vee \left\lfloor \frac{T}{|S_r| \lceil \log_2 n \rceil} \right\rfloor \right\} \wedge n$$

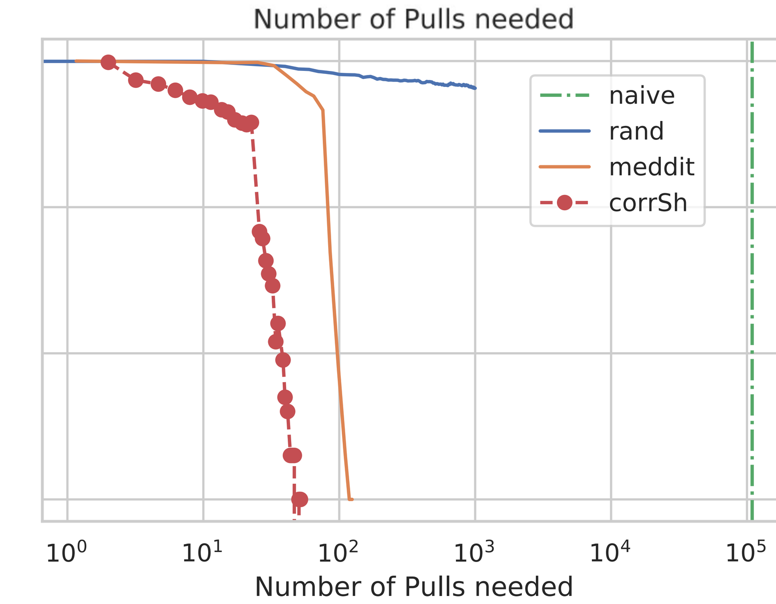
  For each $i \in S_r$ set $\hat{\theta}_i^{(r)} = \frac{1}{t_r} \sum_{j \in \mathcal{J}_r} d(x_i, x_j)$
  if $t_r = n$ **then**
    Output arm in S_r with the smallest $\hat{\theta}_i^{(r)}$
  else
    Let S_{r+1} be the set of $\lceil |S_r|/2 \rceil$ arms in S_r with the smallest $\hat{\theta}_i^{(r)}$
  end if
end for
return arm in $S_{\lceil \log_2 n \rceil}$

Simulation Results

Dataset, Metric	n, d
RNA-Seq 20k, ℓ_1	20k, 28k
Netflix 100k, cos	100k, 18k
MNIST Zeros, ℓ_2	6424, 784







- Figures arranged top to bottom, left to right, following the table

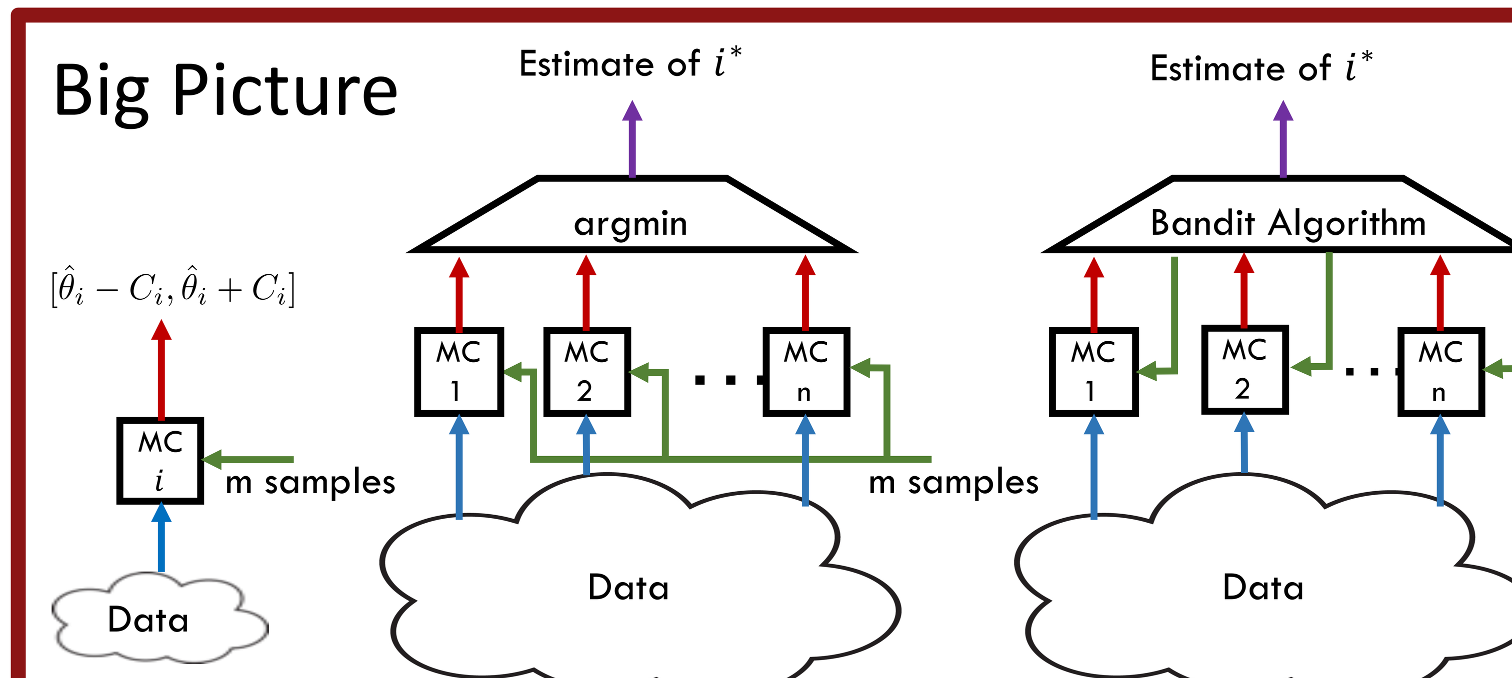
dataset, metric	$n,\ d$		**corrSH**	Med-dit	Rand	Exact Comp.
RNA-Seq 20k, ℓ_1	20k, 28k	time # pulls	**10.9** **2.43**	246 121 (2.1%)	2131 1000 (.1%)	40574 20000
RNA-Seq 100k, ℓ_1	109k, 28k	time # pulls	**64.2** **2.10**	5819 420	10462 1000 (.5%)	- 100000
Netflix 20k, cosine dist	20k, 18k	time # pulls	**6.82** **15.0**	593 85.8	70.2 1000 (.6%)	139 20000
Netflix 100k, cosine dist	100k, 18k	time # pulls	**53.4** **18.5**	6495 90.5 (6%)	959 1000 (3.6%)	- 100000
MNIST Zeros, ℓ_2	6424, 784	time # pulls	**1.46** **47.9**	151 91.2 (.1%)	65.7 1000 (65.2%)	22.8 6424

Theorem Statement

- **Notation:** $d(x_1, x_J) - d(x_i, x_J)$ is $\sigma\rho_i$–subgaussian
- **Theorem:** corrSH identifies the medoid within T distance computations with probability at least

$$1 - 3\log n \exp\left(-\frac{T}{16\sigma^2 \log n} \cdot \min_{i \geq \frac{T}{n \log n}} \left[\frac{\Delta_{(i)}^2}{i\rho_{(i)}^2} \right] \right)$$

Big Picture



Summary

- Convert computational problem to statistical estimation
- Fast randomized algorithm for data science primitive
- Incorporating structure of the computational problem in this reduction can yield massive gains
- Similar approach can work for k-NN

Paper

Code

References

[1] V. Bagaria, G. Kamath, V. Ntranos, M. Zhang, and D. Tse, "Almost-linear time via multi-armed bandits," in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, pp. 500–509, 2018.
[2] V. Bagaria, G. M. Kamath, and D. N. Tse, "Adaptive monte-carlo optimization," arXiv preprint arXiv:1805.08321, 2018.
[3] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in International Conference on Machine Learning, pp. 1238–1246, 2013.
[4] Baharav, Tavor Z., and David N. Tse. "Ultra Fast Medoid Identification via Correlated Sequential Halving." NeurIPS 2019.