Ultra Fast Medoid Identification via Correlated Sequential Halving

Tavor Z. Baharav, David N. Tse
Stanford University

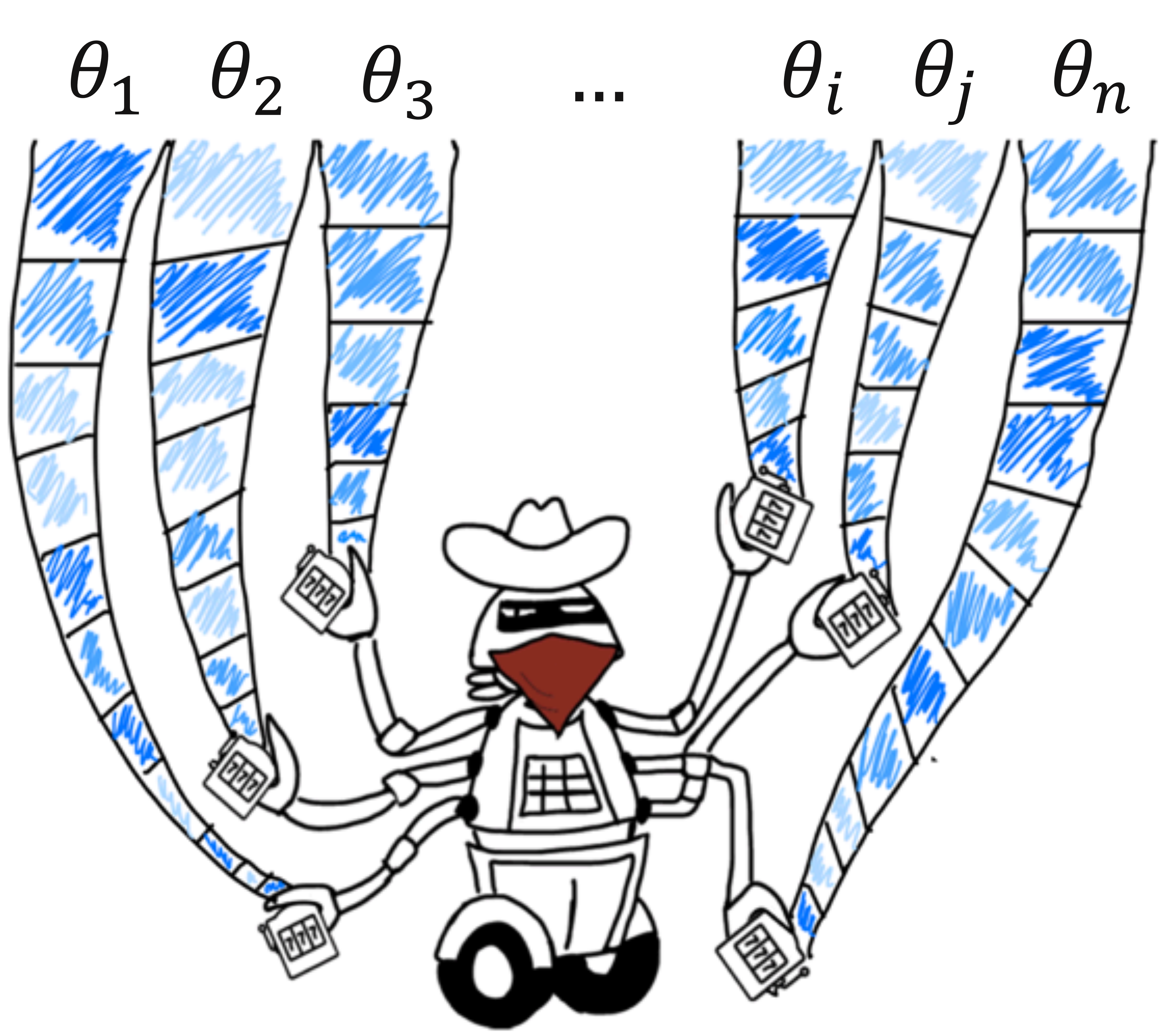
Problem Formulation

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

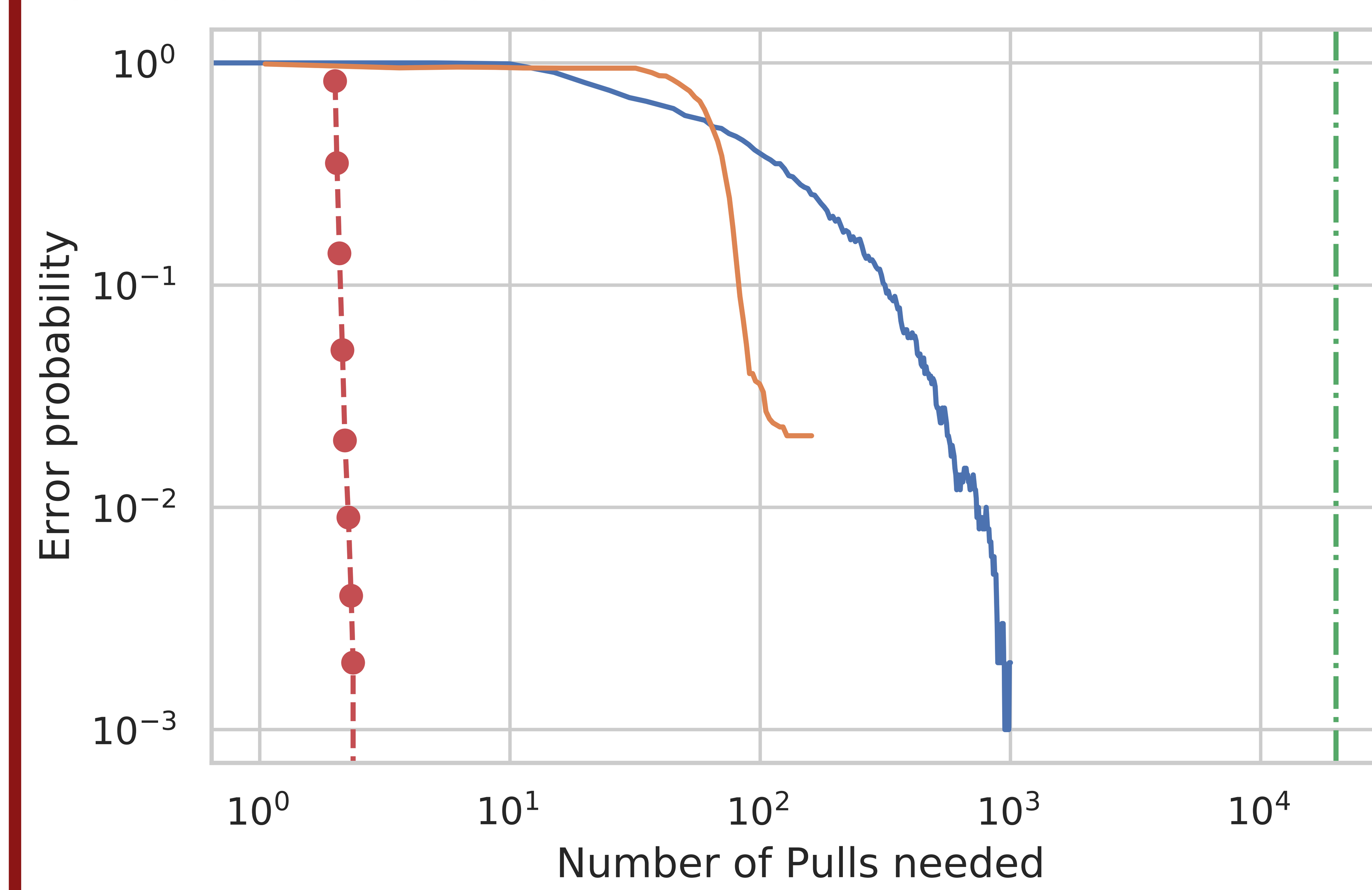
- Generalization of median
- Unlike mean, medoid is inside dataset
- Interpretable, works for general distances
- Robust against outliers

Computation → Estimation

- Naïve: compute each θ_i exactly by averaging $d(x_i, x_j)$
- Idea: estimate θ_i via random sampling
  $$\mathbb{E}\{d(x_i, x_J)\} = \theta_i \qquad J \sim \text{Unif}([n])$$
- Game, get to choose which θ_i to measure at each step
- 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 (Meddit): sample adaptively [1]
  - UCB based sampling strategy



Simulation Results



Dataset, Metric	n, d
RNA-Seq 20k, ℓ_1	20k, 28k
RNA-Seq 100k, ℓ_1	109k, 28k
Netflix 20k, cos	20k, 18k
Netflix 100k, cos	100k, 18k
MNIST Zeros, ℓ_2	6424, 784

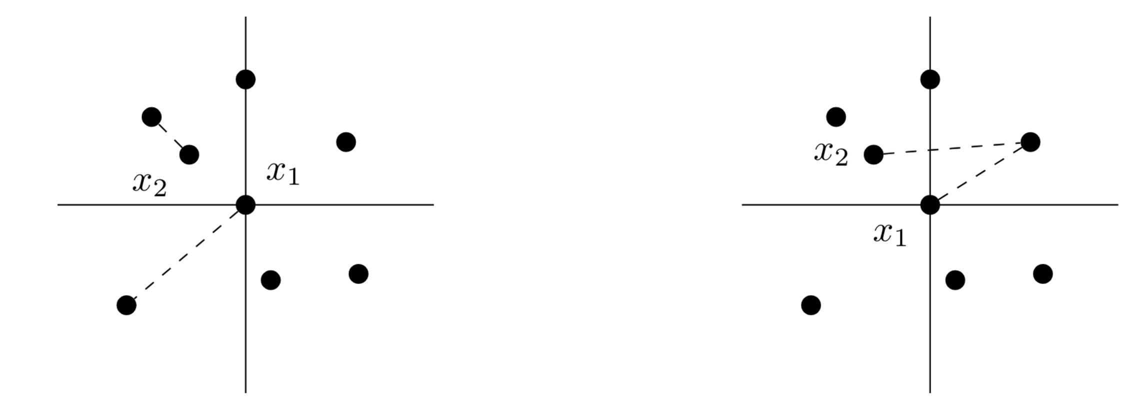
- Similar performance across many datasets and metrics
- Drastic improvements in wall clock time

Theory

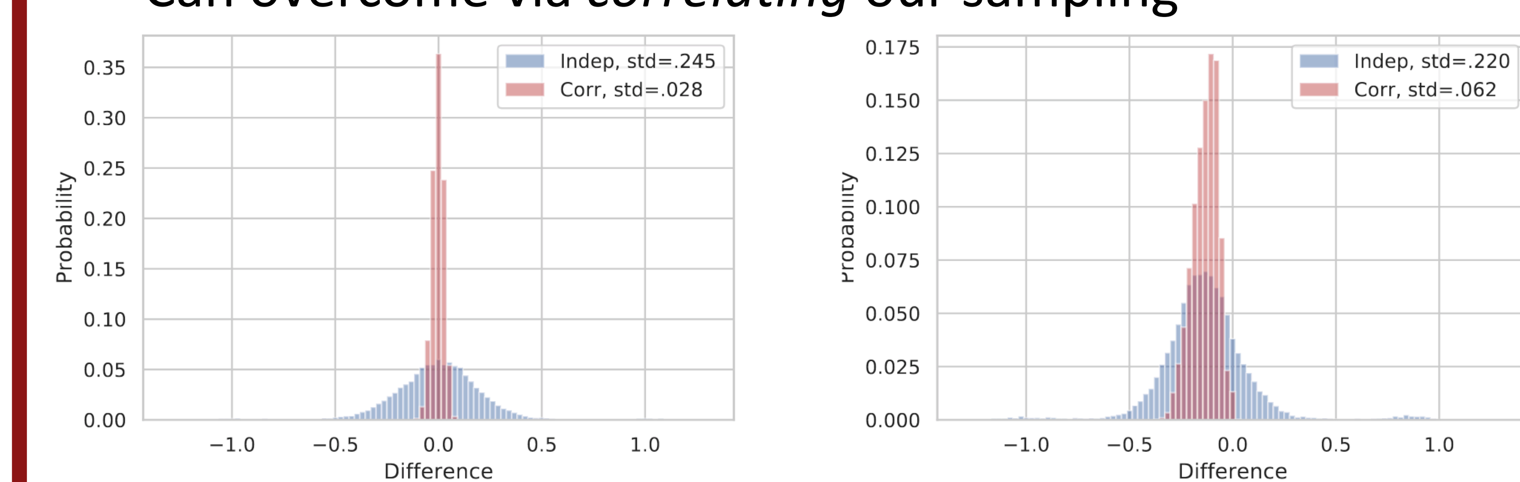
- **Assumption:** $d(x_1, x_J) - d(x_i, x_J)$ is $\sigma\rho_i$–subgaussian
- **Theorem:** corrSH identifies the medoid with probability at least 1-δ after computing
  $$T = O\left(\sigma^2 \log n \log\left(\frac{\log n}{\delta}\right) \max_{i \geq \frac{T}{n \log_2 n}} \left[\frac{i\rho_{(i)}^2}{\Delta_{(i)}^2}\right]\right)$$
  distance evaluations
- Proof follows similarly to [3]
- Traditional lower bounds do not work when access to θ_i is not independent [4]

Our contributions

- Naïve bandit reduction to statistical estimation ignores structure of problem

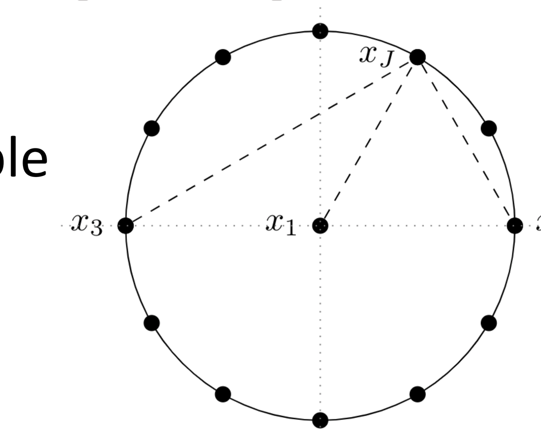


- Can overcome via *correlating* our sampling



(a) Comparison of top 2 points (i=2) (b) Comparison of top and mid (i=10000)

- Additional improvements realizable beyond pairwise



Summary

- Can convert computational problems to statistical estimation ones

References
[1] V. Bagaria, G. Kamath, V. Ntranos, M. Zhang, and D. Tse, "Medoids in 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] D. LeJeune, R. Heckel, and R. Baraniuk, "Adaptive estimation for approximate k-nearest-neighbor computations," in Proceedings of Machine Learning Research, pp. 3099–3107, 2019.