Online k-means Clustering

Vincent Cohen-Addad Benjamin Guedj

Varun Kanade Guy Rom

The Online k-means clustering problem

- A stream of point in a d-dimensional bounded box is revealed incrementally,
  - one per step, for T steps. $X_T = x_1, x_2, \ldots, x_T$
- An algorithm must commit to k cluster centers at the beginning of each step
  - possibly different sets per step. $C_T = c_1^{(k)}, c_2^{(k)}, \ldots, c_T^{(k)}$
- The loss is the squared distance from the new data point to the closest available cluster center at t

$$\sum_t \min_{p \in c_t^{(k)}} ||x_t - p||^2$$

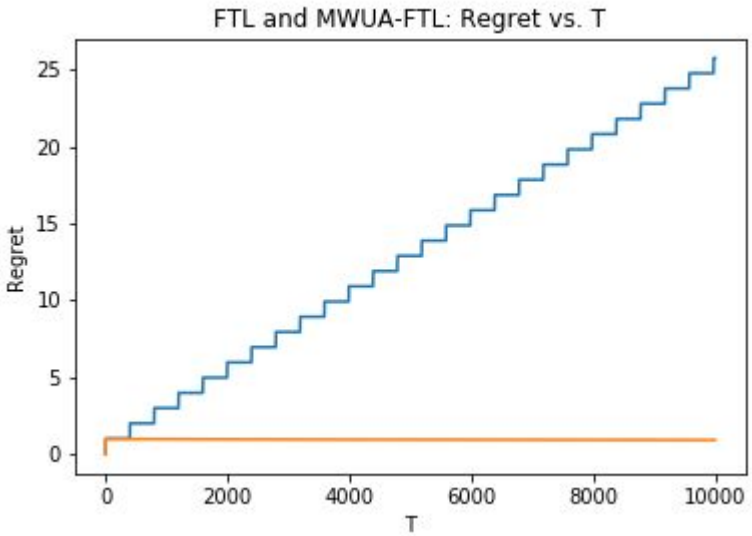
- Algorithms are compared to the best fixed set of k centers, i.e. additive regret

APX-Hardness and Intractable no-regret algorithm

- We present an online-to-offline reduction by uniformly sampling any offline kmeans instance
  - The online algorithm yields a small set that must contain an approx of the optimal offline solution
- An exponential grid discretization with multiplicative weight update algorithm (MWUA) yield an intractable no-regret algorithm

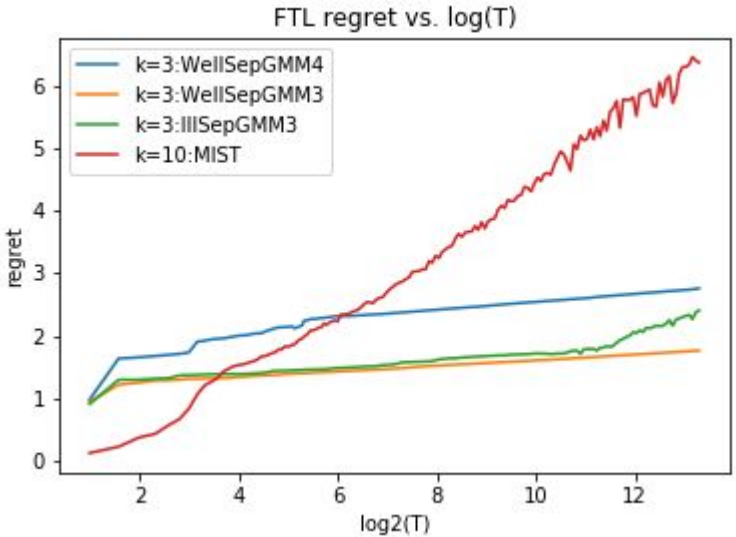
Follow-The-Leader (FTL) with Oracle has Linear Regret

We present a family of counter examples for any ε that makes FTL incur a fixed loss every f(ε) steps - linear regret



FTL Simulation on GMMs performs well

Running FTL with an oracle proxy (lloyd algorithm) on Gaussian Mixture Models yields logarithmic regret, which suggests FTL performs well under some data constraints



FPT Algorithm for Approximate Regret Minimization

- Incremental ε-coreset construction of size poly-log(T) $X_{1:t} \rightarrow \text{Coreset}_t$
- Data Adaptive Hierarchical Region Decomposition $\text{Coreset}_t \rightarrow \text{RegionDecomposition}_t$
- Adaptive Multiplicative Weight Update Algorithm (MWUA) with non-fixed expert set

$$\text{RegionDecomposition}_{1:t} \rightarrow \text{ExpertSet}_t$$