Moment-based Uniform Deviation Bounds for k-means and Friends

Matus Telgarsky and Sanjoy Dasgupta

Goal

Given the output of some clustering heuristic on a finite sample, what performance is expected over the source distribution? Specifically:

k-means. Given a family of sets of centers $\mathcal{P}$ (e.g., $\mathcal{P} \ni P = \{p_i\}_{i=1}^k$), what can be said about the k-means cost deviations

$$\mathcal{D}_{\text{km}} := \sup_{P \in \mathcal{P}} \left| \int \min_{p \in P} \|p - x\|^2 d\mu(x) - \frac{1}{n} \sum_{j=1}^n \min_{p \in P} \|p - x_j\|^2 \right| \quad ?$$

Gaussian Mixture. Given a family of Gaussian mixture parameters $\mathcal{G}$ (e.g., $\mathcal{G} \ni (\alpha, \Theta) = \{(\alpha_i, \mu_i, \Sigma_i)\}_{i=1}^k$), what can be said about the log-likelihood deviations

$$\mathcal{D}_{\text{gm}} := \sup_{(\alpha,\Theta) \in \mathcal{G}} \left| \int \ln \left(\sum_{i=1}^k \alpha_i p_{\mu_i,\Sigma_i}(x) \right) d\mu(x) - \frac{1}{n} \sum_{j=1}^n \ln \left(\sum_{i=1}^k \alpha_i p_{\mu_i,\Sigma_i}(x_j) \right) \right| \quad ?$$

The setting for these questions, as well as its rationale, is as follows.

1. These costs are well-defined only if the source distribution μ has ≥ 2 moments. The results here require ≥ 4 and ≥ 8 moments.
2. The parameter families $\mathcal{P}$ and $\mathcal{G}$ should model heuristics, and in particular be unbounded. The results here consider parameters beating some fixed cost c (e.g., the variance).
3. No identifiability assumptions. The results here control cost, and additional assumptions grant parameter recovery.

Results

k-means. Given a fixed cost c and source distribution with $p \geq 4$ moments, (w.h.p.) deviations over $P \in \mathcal{P}$ with cost at most c are bounded as

$$\mathcal{D}_{\text{km}} \leq n^{-1/2+\min\{1/4,2/p\}} \mathcal{O}\left(c\sqrt{dk \ln(n/\delta)} + (1/\delta)^{4/p} \right).$$

Gaussian Mixtures. Given a fixed cost c and source distribution with $p \geq 8$ moments, deviations over $(\alpha, \Theta) \in \mathcal{G}$ with cost at most c and fixed bounded spectrum $\sigma_1 I \preceq \Sigma_i \preceq \sigma_2 I$ are bounded as

$$\mathcal{D}_{\text{gm}} \leq n^{-1/2+3/p} \text{poly}(c, d, k) \mathcal{O}\left(\sqrt{\ln(n/\delta)} + (1/\delta)^{4/p} \right).$$

Remarks

- $(1/\delta)^{4/p}$ is an artifact of Rosenthal-type moment inequalities.
- As $p \to \infty$, the dependence on n approaches $\widetilde{\mathcal{O}}(n^{-1/2})$.
- Heuristics can be forced to meet the conditions on $\mathcal{P}$ and $\mathcal{G}$ (e.g., if they fail to beat the variance score, output the mean).
- The writeup also presents bounds adapted to cluster structure.
- The Gaussian spectrum conditions are perhaps unnecessary.
- The core proof idea is due to Pollard (1981, k-means consistency). There, identifiability assumptions are made, and exact centers are recovered; the proofs diverge somewhat after controlling one center.

Other Approaches

- If the distribution is bounded, standard covering techniques suffice.
- There are a few results derived from Vapnik's work on unbounded losses...

Proof – Key Lemma

Far away centers and probability mass are irrelevant: given any $\epsilon > 0$, there exist balls B_3 and B_4 such that (w.h.p.)

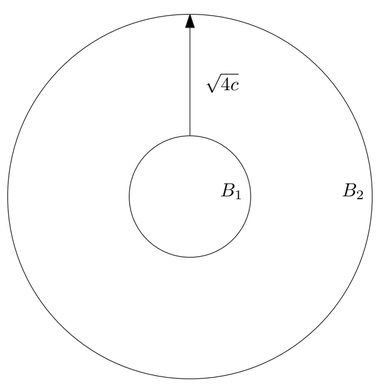
$$\left| \int \min_{p \in P} \|p - x\|^2 d\nu(x) - \int_{B_3} \min_{p \in P \cap B_4} \|p - x\|^2 d\nu(x) \right| \leq \epsilon,$$

where $\nu \in \{\mu, \hat{\mu}\}$. Remarks:

- Remainder trivial: kill outer reaches then cover inner regions.
- Various "there exist" statements quantifiable via moments.
- k-means sketched below, but Gaussian mixtures similar.
- This killing off is motivated by the following: given p, p' with $\|p - p'\|$ small, unclear how to reason about $\|p - x\|$ vs. $\|p' - x\|$ for distant x!

Lemma Proof – Step 1 – Controlling One Center

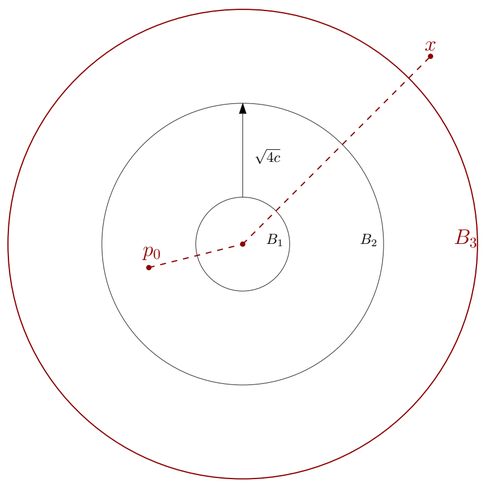
Choose B_1 with $\nu(B_1) > 1/4$; if $P \cap B_2 = \emptyset$, then



$$\begin{aligned} &\int \min_{p \in P} \|x - p\|^2 d\nu(x) \\ &\geq \int_{B_1} \min_{p \in P} \|x - p\|^2 d\nu(x) \\ &\geq \int_{B_1} \left(\sqrt{4c}\right)^2 d\nu(x) \\ &> c. \end{aligned}$$

Lemma Proof – Step 2 – Killing Remote Mass

Choose $B_3 \supseteq B_2$ with $\int_{B_3^c} 4\|x - \mathbb{E}(X)\|^2 d\nu(X) \leq \epsilon$; keeping in mind any $p_0 \in P \cap B_2$, then for any $x \in B_3^c$,



$$\begin{aligned} &\min_{p \in P} \|x - p\|^2 \\ &\leq \min_{p \in P} 2\|x - \mathbb{E}(X)\|^2 + 2\|\mathbb{E}(X) - p\|^2 \\ &\leq 4\|x - \mathbb{E}(X)\|^2, \end{aligned}$$

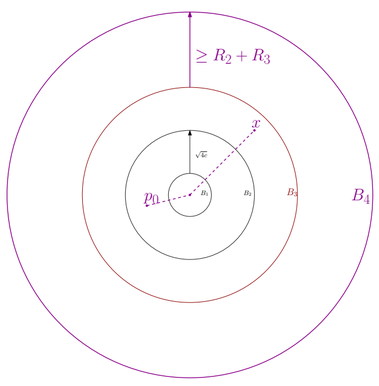
thus

$$0 \leq \int_{B_3^c} \min_{p \in P} \|x - p\|^2 d\nu(x) \leq \epsilon.$$

(The writeup calls this step *outer bracketing with* $x \mapsto 4\|x - \mathbb{E}(X)\|^2$.)

Lemma Proof – Step 3 – Killing Remote Centers

Choose B_4 with $R_4 := \text{radius}(B_4) \geq R_2 + 2R_3$. Then for any $x \in B_3$,



$$\begin{aligned} &\inf_{p \in B_4^c} \|x - p\|^2 \\ &\geq (R_2 + R_3)^2 \\ &\geq (\|x - \mathbb{E}(X)\| + \|\mathbb{E}(X) - p_0\|)^2 \\ &\geq \|x - p_0\|^2 \\ &\geq \min_{p \in P \cap B_4} \|x - p\|^2. \end{aligned}$$

Together,

$$\int \min_{p \in P} \|x - p\|^2 d\nu(x) \geq \int_{B_3} \min_{p \in P} \|x - p\|^2 d\nu(x) = \int_{B_3} \min_{p \in P \cap B_4} \|x - p\|^2 d\nu(x),$$

and

$$\int_{B_3} \min_{p \in P \cap B_4} \|x-p\|^2 d\nu(x) = \int_{B_3} \min_{p \in P} \|x-p\|^2 d\nu(x) \geq \int \min_{p \in P} \|x-p\|^2 d\nu(x) - \epsilon.$$