

1 Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be independent Rademacher random variables (that is, $\mathbf{x}_i = \pm 1$ with equal probability) and $\mathbf{a} = (a_1, a_2, \dots, a_n)$ be a sequence of real numbers.

- (a) Show that $\mathbb{E}e^{s\mathbf{x}_i} = \cosh(s)$.
- (b) Prove that $\cosh(s) \leq e^{s^2/2}$.
- (c) Use (a), (b), and Markov's inequality to prove that

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i \mathbf{x}_i\right| \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\|\mathbf{a}\|_2^2}}$$

■

2^a In the k -means clustering problem we are given some input vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathbb{R}^n$ and a positive integer k , and we'd like to output a partition P of $\{1, 2, \dots, m\}$ into k disjoint subsets P_1, P_2, \dots, P_k and cluster centers $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k \in \mathbb{R}^n$. which minimize the function

$$f(\{P_i\}, \{\mathbf{y}_i\}; \{\mathbf{x}_i\}) = \sum_{j=1}^k \sum_{i \in P_j} \|\mathbf{x}_i - \mathbf{y}_j\|_2^2.$$

That is, the x_i are clustered into k clusters according to P . This problem is NP-hard, but good approximation algorithms exist which can return almost-optimal clusterings.

- (a) For a fixed partition P , show that the optimal $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k$ is where for every nonempty $P_i \in P$,

$$\mathbf{y}_i = \frac{1}{|P_i|} \sum_{j \in P_i} \mathbf{x}_j$$

is just the average of the points in P_i . Thus we can restrict ourselves to optimizing over partitions P .

- (b) Prove that for a given cluster P_i , the optimal cost is just the average pairwise distance between points in that cluster.
- (c) Using the Johnson-Lindenstrauss lemma, show that for any $0 < \epsilon < \frac{1}{2}$ there is a linear map $\mathbf{S} \in \mathbb{R}^{m \times n}$, $m = \mathcal{O}(\epsilon^{-2} \log m)$ such that for all partitions P simultaneously

$$(1 - \epsilon)f(\{P_i\}; \{\mathbf{x}_i\}) \leq f(\{P_i\}; \{\mathbf{S}\mathbf{x}_i\}) \leq (1 + \epsilon)f(\{P_i\}; \{\mathbf{x}_i\})$$

and where \mathbf{S} can be found efficiently with a randomized algorithm with small failure probability. Thus if one does not mind worsening the quality of our clusters by a factor $1 + \epsilon$, without loss of generality one can assume that the input vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathbb{R}^n$ are in dimension $n = \mathcal{O}(\epsilon^{-2} \log m)$, which can be *much* smaller than the original dimension n . *Hint:* Use the Johnson-Lindenstrauss lemma.

^aJelani Nelson, 2013

■