Algorithmic Aspects of Machine Learning: Problem Set # 2

Instructor: Ankur Moitra

Due: April 6th

You can work with other students, but you must write-up your solutions by yourself and indicate at the top who you worked with!

Problem 1

In the *multi-reference alignment* problem we observe many noisy copies of the same unknown signal $x \in \mathbb{R}^d$, but each copy has been circularly shifted by a random offset (as pictured).

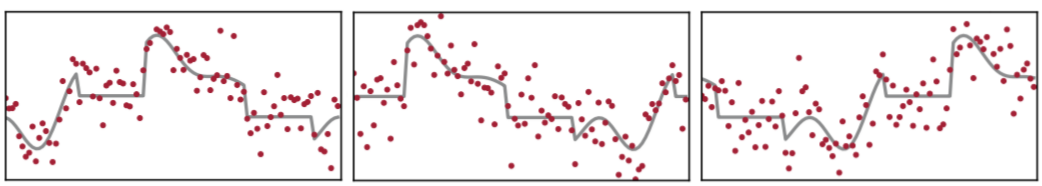

Figure: Three shifted copies of the true signal x are shown in gray. Noisy samples y_i are shown in red.

Formally, for $i = 1, 2, \ldots, n$ we observe

$$y_i = R_{\ell_i} x + \xi_i$$

where: the ℓ_i are drawn uniformly and independently from $\{0, 1, \ldots, d-1\}$; R_ℓ is the operator that circularly shifts a vector by ℓ indices; $\xi_i \sim \mathcal{N}(0, \sigma^2 I_{d \times d})$ with $\{\xi_i\}_i$ independent; and $\sigma > 0$ is a known constant. Think of d, x and σ as fixed while $n \to \infty$. The goal is to recover x (or a circular shift of x).

(a) Consider the tensor $T(x) = \frac{1}{d} \sum_{\ell=0}^{d-1} (R_\ell x) \otimes (R_\ell x) \otimes (R_\ell x)$. Show how to use the samples y_i to estimate T (with error tending to zero as $n \to \infty$). Take extra care with the entries that have repeated indices (e.g. T_{aab}, T_{aaa}).

(b) Given $T(x)$, prove that Jennrich's algorithm can be used to recover x (up to circular shift). Assume that x is *generic* in the following sense: let $x' \in \mathbb{R}^d$ be arbitrary and let x be obtained from x' by adding a small perturbation $\delta \sim \mathcal{N}(0, \epsilon)$ to the first entry. *Hint:* form a matrix with rows $\{R_\ell x\}_{0 \le \ell < d}$, arranged so that the diagonal entries are all x_1.

Problem 2

In this problem we will study the task of learning the high-dimensional parameters of a mixture model from its one-dimensional projections. Suppose we have a d-dimensional mixture model

$$\mathcal{F} = \frac{1}{2}F_1 + \frac{1}{2}F_2$$

where F_1 and F_2 have means μ_1 and μ_2 respectively. Suppose are given access to an oracle, that when we query a direction r, outputs the set

$$\{r^T\mu_1 \pm \gamma, r^T\mu_2 \pm \gamma\}$$

That is, we learn the means of the projected distributions up to accuracy γ but we do not know which estimates correspond to which components.

(a) Explain how to interpret the oracle as giving noisy estimates of quadratic forms on the matrix

$$M = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$$

Why doesn't it matter that we don't know which estimates corresponds to which components?

(b) Give an algorithm that makes $O(d^2)$ queries and recovers the entries of M up to accuracy $O(\gamma)$.

(c) Explain how to use your estimate of M to estimate μ_1 and μ_2 up to permutation. A sketch of the algorithm and how you would go about analyzing it is enough, since the details can get messy.

Problem 3

In this problem we will show that a monotone adversary can break even very powerful approaches for community detection. Consider the planted bisection model on n nodes with inter- and intra-connection probabilities p and q respectively. Suppose we look for a bisection $(A, V \setminus A)$ with the property that every node u has more neighbors on its side of the bisection than on the other side. We will call such a bisection *admissible*. Show how a monotone adversary can construct an admissable bisection that is far from the planted one.