4.7.1 Oblivious Sparse Norm-Approximating Projections

There is an interesting random matrix problem related to Oblivious Sparse Norm-Approximating Projections [NN], a form of dimension reduction useful for fast linear algebra. In a nutshell, The idea is to try to find random matrices Π that achieve dimension reduction, meaning $\Pi \in \mathbb{R}^{m\times n}$ with $m \ll n$, and that preserve the norm of every point in a certain subspace [NN], moreover, for the sake of computational efficiency, these matrices should be sparse (to allow for faster matrix-vector multiplication). In some sense, this is a generalization of the ideas of the Johnson-Lindenstrauss Lemma and Gordon's Escape through the Mesh Theorem that we will discuss next Section.

Open Problem 4.4 (OSNAP [NN]) *Let $s \leq d \leq m \leq n$.*

1. *Let $\Pi \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. entries*

$$\Pi_{ri} = \frac{\delta_{ri}\sigma_{ri}}{\sqrt{s}},$$

where σ_{ri} is a Rademacher random variable and

$$\delta_{ri} = \begin{cases} \frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{m} \\ 0 & \text{with probability} \quad 1 - \frac{s}{m} \end{cases}$$

Prove or disprove: there exist positive universal constants c_1 and c_2 such that

For any $U \in \mathbb{R}^{n\times d}$ for which $U^T U = I_{d\times d}$

$$\text{Prob}\left\{ \left\| (\Pi U)^T (\Pi U) - I \right\| \geq \varepsilon \right\} < \delta,$$

for $m \geq c_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon^2}$.

2. *Same setting as in (1) but conditioning on*

$$\sum_{r=1}^{m} \delta_{ri} = s, \quad \text{for all } i,$$

meaning that each column of Π has exactly s non-zero elements, rather than on average. The conjecture is then slightly different:

Prove or disprove: there exist positive universal constants c_1 and c_2 such that

For any $U \in \mathbb{R}^{n\times d}$ for which $U^T U = I_{d\times d}$

$$\text{Prob}\left\{ \left\| (\Pi U)^T (\Pi U) - I \right\| \geq \varepsilon \right\} < \delta,$$

for $m \geq c_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon}$.

3. *The conjecture in (1) but for the specific choice of* U*:*

$$U = \begin{bmatrix} I_{d\times d} \\ 0_{(n-d)\times d} \end{bmatrix}.$$

In this case, the object in question is a sum of rank 1 *independent matrices. More precisely,* $z_1, \ldots, z_m \in \mathbb{R}^d$ *(corresponding to the first* d *coordinates of each of the* m *rows of* Π*) are i.i.d. random vectors with i.i.d. entries*

$$(z_k)_j = \begin{cases} -\frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{2m} \\ 0 & \text{with probability} \quad 1-\frac{s}{m} \\ \frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{2m} \end{cases}$$

Note that $\mathbb{E} z_k z_k^T = \frac{1}{m} I_{d\times d}$*. The conjecture is then that, there exists* c_1 *and* c_2 *positive universal constants such that*

$$\text{Prob}\left\{\left\|\sum_{k=1}^{m}\left[z_k z_k^T - \mathbb{E} z_k z_k^T\right]\right\| \geq \varepsilon\right\} < \delta,$$

for $m \geq c_1 \frac{d+\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ *and* $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon^2}$.

I think this would is an interesting question even for fixed δ*, for say* $\delta = 0.1$*, or even simply understand the value of*

$$\mathbb{E}\left\|\sum_{k=1}^{m}\left[z_k z_k^T - \mathbb{E} z_k z_k^T\right]\right\|.$$

Reference

[NN] J. Nelson and L. Nguyen. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. *Available at arXiv:1211.1002 [cs.DS].*