High-dimensional statistics 2: assignment 4

Due Weds Mar 8, 2017

1. **Matrix sampling and concentration**

Theorem 1 (Tropp's matrix Bernstein inequality)**.** *Suppose that $X \in \mathbb{R}^{d_1 \times d_2}$ is a random matrix with $\|X\| \leq R$ always (here $\|X\|$ denotes the matrix operator norm) and with*

$$\nu = \max\left\{\left\|\mathbb{E}\left[X^\top X\right]\right\|, \left\|\mathbb{E}\left[XX^\top\right]\right\|\right\} .$$

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$. Then

$$\mathbb{E}\left[\left\|\sum_{t=1}^{n}(X_t - \mathbb{E}[X])\right\|\right] \leq \sqrt{n} \cdot \sqrt{2\nu \log(d_1 + d_2)} + \frac{2}{3} R \log(d_1 + d_2) .$$

We will use this theorem to study the behavior of the random sampling operator for a matrix completion problem. Recall that for uniform sampling, in class we discussed the matrix "mask" $\Omega \in \mathbb{R}^{d_1 \times d_2}$ which has a 1 in every observed location and a 0 elsewhere. This is equivalent to drawing a sample of size n, without replacement, from the set of all possible locations in the matrix. We can approximate Ω with another matrix where we instead sample with replacement (to get things to be iid):

$$M = \sum_{t=1}^{n} X_t$$

where X_t is a matrix with a 1 in a single location drawn uniformly at random, and a 0 elsewhere.

(a) Calculate $\mathbb{E}[M]$, and use the theorem to prove a bound on $\mathbb{E}[\|M - \mathbb{E}[M]\|]$.

(b) Suppose that the sampling is no longer uniformly at random, but instead each X_t has a single 1 in location (i, j) with probability p_{ij} (where $\sum_i \sum_j p_{ij} = 1$ gives a probability distribution over the set of locations). Suppose that the row and column probabilities are bounded as $\max_i \sum_j p_{ij} \leq \frac{L}{d_1}$, $\max_j \sum_i p_{ij} \leq \frac{L}{d_2}$. Calculate $\mathbb{E}[M]$ again. Calculate $\mathbb{E}[\|M - \mathbb{E}[M]\|]$ again now in terms of L. (Note: when L is very large (e.g. scaling with dimension rather than with a constant), this will be extremely large, and so instead of using nuclear norm for matrix completion we would want to use a reweighted penalty; see `http://arxiv.org/abs/1106.4251` and `http://arxiv.org/abs/1009.2118` if you're interested in this case.)

2. **Gaussian width**

Recall that the Gaussian width for a cone $S \subset \mathbb{R}^d$ is given by

$$\omega(S) = \mathbb{E}\left[\sup_{x \in S \cap \mathbb{S}^{d-1}} \langle x, g \rangle\right]$$

where $g \sim N(0, \mathbf{I}_d)$.

Let S = {all k-sparse vectors}. Prove that $\omega(S) \lesssim \sqrt{k \log(d/k)}$. You can use the χ^2 tail bound

$$\mathbb{P}\left\{\chi^2_m \geq m + 2\sqrt{mt} + 2t\right\} \leq e^{-t}.$$

Roughly speaking this bound on $\omega(S)$ means that the sample complexity for learning a k-sparse vector is $n \sim k \log(d/k)$.

3. **Rademacher complexity & Gaussian complexity**

Recall that the *Rademacher complexity* of a class of functions $\mathcal{F}$ is defined by

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(Z_i) \right| \right)$$

where $\sigma_1, \ldots, \sigma_n$ are Rademacher random variables. The expectation is taken over both the σ_i's and the sample (the Z_i's).

Define also the Gaussian complexity

$$\mathfrak{G}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n g_i f(Z_i) \right| \right)$$

where $g_i \overset{\text{iid}}{\sim} N(0,1)$.

(a) Prove that $\mathfrak{R}_n(\mathcal{F}) \leq C \cdot \mathfrak{G}_n(\mathcal{F})$ where C is an explicit constant you should calculate. (Hint: let $\sigma_i = \operatorname{sign}(g_i)$.)

(b) It is known that Gaussian complexity is bounded by Rademacher complexity, up to a log factor. We will not prove this but will prove something intuitively similar. First, we will show that for any fixed $w_1, \ldots, w_n \in [-1, 1]$, the "weighted Rademacher complexity" is no larger than the Rademacher complexity:

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n w_i \cdot \sigma_i f(Z_i) \right| \right) \leq \mathfrak{R}_n(\mathcal{F}).$$

To see why, let $\tau_i \in \{\pm 1\}$ be independent such that $\mathbb{E}[\tau_i] = w_i$, and consider instead bounding the complexity measure

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \cdot \sigma_i f(Z_i) \right| \right) \leq \mathfrak{R}_n(\mathcal{F}),$$

then use this to prove the "weighted Rademacher" bound above.

(c) As the next step, take W_i to be *random* independent and symmetric variables with $W_i \in [-B, B]$. Prove that

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n W_i f(Z_i) \right| \right) \leq B\mathfrak{R}_n(\mathcal{F})$$

using your work in the part above.

4. **Simulation: approximate isometries via non-Gaussian matrices**

It is known that if the matrix $A \in \mathbb{R}^{n \times p}$ has i.i.d. $N(0,1)$ entries, for $p \gg n$, then with high probability it acts as an approximate isometry on any sufficiently restricted class of vectors, e.g. sparse vectors, meaning that $\frac{1}{\sqrt{n}} \|Ax\|_2 \approx \|x\|_2$ is true simultaneously for all restricted vectors x; this allows for much of the theory for high dimensional sparse regression and other statistical problems.

In this problem, we'll see whether the same phenomenon holds when the entries of A are i.i.d. but non-Gaussian. In particular, it may be more efficient to store the matrix A if it is highly sparse, so we will consider the case that

$$A_{ij} = \begin{cases} \frac{1}{\sqrt{\rho}}, & \text{with probability } \rho/2, \\ -\frac{1}{\sqrt{\rho}}, & \text{with probability } \rho/2, \\ 0, & \text{with probability } 1-\rho. \end{cases}$$

where $\rho \in [0,1]$ controls the amount of sparsity. (The reason for signs is that $\mathbb{E}[A_{ij}] = 0$; the reason for the scaling on the nonzero entries is so that $\mathbb{E}\left[A_{ij}^2\right] = 1$.)

(a) Write a function that approximates the "maximum sparse eigenvalue" of A: given a sparsity level k, our goal is to calculate

$$S_k(A) \coloneqq \max\left\{\|Au\|_2 : u \in \mathbb{R}^p, \|u\|_2 \le 1, \|u\|_0 \le k\right\} .$$

We can rewrite this as:

$$S_k(A) \coloneqq \max\left\{v^\top A u : u \in \mathbb{R}^p, \|u\|_2 \le 1, \|u\|_0 \le k; v \in \mathbb{R}^n, \|v\|_2 \le 1\right\} .$$

To approximate this, we can use an alternating maximization:

- Initialize with a random unit vector $u \in \mathbb{R}^p$
- Update v: $v = \frac{Au}{\|Au\|_2}$
- Update u: $u = \arg\max\left\{u^\top (A^\top v) : \|u\|_2 \le 1, \|u\|_0 \le k\right\}$. (Hint: sort the entries of $A^\top v$.)
- Iterate the two above steps.
- Then repeat $\sim$100 times with different initial random choices for u, and take the largest output (since this is an alternating maximization algorithm, it's nonconvex, and is actually extremely sensitive to the initialization.)

(b) Fix $p \gg n$ large. Plot the average "maximum sparse eigenvalue" of A against k, for each of your distributions on A—that is, for Gaussian A, and for signed Bernoulli A as described above with various values of ρ (we are particularly interested in how the behavior degrades as ρ gets extremely small). Recall from class that the scaling for a Gaussian matrix A should be something like $\sqrt{n} + \sqrt{k \log(p)}$ (where the second term comes from Gaussian width) — so if you fix n and p, try plotting $S_k(A)$ against $\sqrt{k}$ to look for linearity. You can also investigate how the empirical outcome varies with n, p, k jointly.