Lecture 12: Robust statistics via Sum-of-Squares

November 5, 2025

Recall from last lecture that the SoS hierarchy is an automatic, algorithmic way to find solutions to systems of polynomial constraints. Of course, this comes with caveats, since this problem is hard in the worst case. But loosely speaking, SoS allows us to find "pseudo-distributions" over solutions to such systems of polynomial constraints, in the sense that it can find these pseudo-expectation operators that satisfy the systems of polynomial constraints.

How exactly does this relate to robust statistics? Recall that in the bounded second moments setting, the key idea was to find the top eigenvector of the empirical covariance matrix, that is, given a corrupted dataset $X_1, \ldots, X_n$ with empirical mean $\widehat{\mu}$, we needed to certify a bound on $\|\widehat{\Sigma}\|_\infty$, where $\widehat{\Sigma}$ is the empirical covariance, which is equivalent to certifying an upper bound on

$$\max_{\|v\|_2=1} \frac{1}{n} \sum_{i=1}^{n} \langle X_i - \widehat{\mu}, v \rangle^2 \ . \tag{1}$$

Moreover in the case where this quantity was large, we also needed to find the vector v which maximized its value. In this setting, this is easy, because we know that the optimizer is given by the top eigenvector of $\widehat{\Sigma}$ and the optimal value is given by its top eigenvalue.

But what if we have a distribution with bounded k-th moments, for some even $k > 2$, i.e., what if our distribution D satisfies

$$\sup_{\|v\|_2=1} \mathop{\mathbb{E}}_{X \sim D} [\langle X - \mu, v \rangle^k] \leq 1 \ ? \tag{2}$$

In this case, we can still use the bounded second moment techniques (since such a distribution will also have bounded second moments), but the error we will be able to achieve is $O(\sqrt{\varepsilon})$. On the other hand, by information-theoretic arguments, we know that one should be able to achieve error $O(\varepsilon^{1-1/k})$ for this setting. The natural analog of what we would want to do is to take the solution to the natural degree-k generalization of (1):

$$\max_{\|v\|_2=1} \frac{1}{n} \sum_{i=1}^{n} \langle X_i - \widehat{\mu}, v \rangle^k \ . \tag{3}$$

One can show that if we could certify that the quantity in (3) was appropriately bounded, then the empirical mean of the dataset would have error at most $O(\varepsilon^{1-1/k})$, as desired. Simultaneously, if (3) was too large, if we filtered based on the scores given by

$$\tau_i = \langle X_i - \widehat{\mu}, v \rangle^k \ , \tag{4}$$

we would have the usual safety guarantee that we remove more bad points than good points. In other words, solving (3) would allow us to solve the robust mean estimation problem to the correct error guarantee in this setting.

Unfortunately, this is where we hit a snag, as (3) is computationally hard to solve in the worst case, for any $k > 2$. This is because this is more or less equivalent to computing the injective norm of a k-tensor:

Definition 0.1 (Injective norms of tensors)**.** Let T be an order-k tensor. The *injective* norm of T is defined to be

$$\|T\|_{\text{inj}} = \sup_{\|v\|=1} |T(v, \ldots, v)| \ .$$

To see the connection, notice that if we let

$$\hat{T} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^{\otimes k} \ ,$$

we have that $\hat{T}(v, \ldots, v)$ is exactly the quantity in (3). Thus, finding the optimizer of (3) is equivalent to finding the maximizer of the injective norm of $\hat{T}$. Unfortunately, computing injective norms, and indeed, most tensor optimization problems, are hard in the worst-case for any $k > 2$, see e.g. [1].

However, not all hope is lost, as the empirical moment tensors $\hat{T}$ we encounter have some structure: namely, most of the points in the dataset are drawn i.i.d. from some nice distribution. So maybe we won't encounter the worst-case hard instances?

1 Injective norm certification via SoS

Indeed, this is what we will show, at least under some additional assumptions on our distribution D. The key point is that, for many natural distributions, there will be an SoS proof of the fact that $\|T\|_{\mathrm{inj}}$ is bounded. Formally:

Definition 1.1. Let k be even. We say a distribution D has *certifiably bounded k-th* moments with bound C_k if

$$\vdash_{O(k)} C_k \|v\|_2^k - \mathop{\mathbb{E}}_{X\sim D}[\langle X - \mu, v\rangle^k] \geq 0 \ ,$$

where we treat the RHS as a polynomial in the formal variable v. That is, there is an SoS proof of (2).

There is a nice way of succinctly describing these sorts of quantities. Namely, one can define the *sos norm* of the tensor T, denoted $\|T\|_{\mathrm{sos}}$, to be the infimum over C so that

$$\vdash_k C\|v\|_2^k - T(v, \ldots, v) \geq 0 \ . \tag{5}$$

One can verify that this is indeed a norm. Then for instance, the first lemma is equivalent to saying that $\|\mathbb{E}_{X\sim N(0,I)}[X^{\otimes k}]\|_{\mathrm{sos}} \leq k^{k/2}$.

Equivalently by convex duality, (under some mild regularity conditions), note that this norm can also be defined to be

$$\sup\left\{\widetilde{\mathbb{E}}[T(v, \ldots, v)] : \widetilde{\mathbb{E}} \models_k \|v\|_2^2 = 1\right\} \ ,$$

where the supremum is taken over degree k pseudo-expectations.

How strict is certifiable moments? As a brief tangent, it is a natural question to ask how restrictive an assumption this is. Under plausible complexity theoretic assumptions, it is likely not true that all distributions with bounded higher moments (in the sense of (2)) will have certifiably bounded moments [2]. However, many natural classes of distributions do. For instance:

Lemma 1.1. *A standard Gaussian has certifiably bounded k-th moments with bound $C_k \leq k^{k/2}$.*

Proof. Let us expand out the definition: for some vector of indeterminants u, we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,I)}\left[\langle X, v\rangle^k\right] &= \sum_{\alpha} v^{\alpha} \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,I)}[X^{\alpha}] \\ &= \sum_{\alpha} v^{\alpha} \prod_{i=1}^{d} \mathop{\mathbb{E}}_{X\sim N(0,1)}[X^{\alpha_i}] \\ &= \sum_{\alpha \text{ even}} v^{\alpha} \prod_{i=1}^{d} \mathop{\mathbb{E}}_{X\sim N(0,1)}[X^{\alpha_i}] \ ,\end{aligned}$$

where the second line follows from the independence of the coordinates of the Gaussian, and in the third line, we say that α is even if α_i is even for all $i \in [d]$, and this follows since all odd terms are 0. Note that the condition that α is even is equivalent to the condition that v^α is a square of a polynomial! We next apply the following univariate estimate, which is more or less tight:

$$\mathop{\mathbb{E}}_{X \sim N(0,1)} [X^t] \leq t^{t/2} \; .$$

From this, we obtain that

$$\begin{aligned} \vdash_k \mathop{\mathbb{E}}_{X \sim \mathcal{N}(0,I)} \left[\langle X, v \rangle^k \right] &\leq t^{t/2} \sum_{\alpha \text{ even}} u^\alpha \\ &= t^{t/2} \|u\|_2^k \; , \end{aligned}$$

which completes the proof. □

So this implies that Gaussians are certifiably bounded, and this bound $C_k \leq k^{k/2}$ is pretty much the best bound one can hope for, since even in one dimension, this is the best bound you can achieve. More generally, we say that a univariate distribution D with mean μ is *sub-gaussian* (with variance proxy 1) if

$$\mathop{\mathbb{E}}_{X \sim D} [(X - \mu)^k] \leq k^{k/2} \; ,$$

and by following this proof almost exactly, we can also demonstrate the following:

Lemma 1.2. *Let D be any distribution over $\mathbb{R}^d$ whose coordinates are independent sub-gaussian random variables. Then D is certifiably bounded.*

One can also show that rotation does not affect certifiable boundedness, so this also applies to rotations of such random variables. In fact, it was recently shown that any sub-gaussian distribution must be certifiably so [3]! However, this phenomena seems very specific to the sub-gaussian setting in particular.

One can even go further: in fact, one can show that any distribution satisfying the so-called Poincaré condition is certifiably sub-gaussian [4]. In particular, by Chen's proof of the KLS conjecture in convex geometry, this also implies that any log-concave distribution is certifiably sub-gaussian with variance proxy which is logarithmic in the dimension [5].

2 Filtering via SoS

With this sort of tool, it is not hard to see that one can easily adapt the machinery of filtering to work with SoS-based generalizations of spectral tests. Suppose we have $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ as usual, and let μ be the uncorrupted mean. When you have bounded k-th moments, you can show a generalization of the spectral signatures lemma of the following form: let $\widehat{\mu}$ be the empirical mean of S, and let $\widehat{T}$ denote the empirical k-th moment tensor of S. If the distribution of the good points has k-th moment tensor with injective norm at most C_k, then (up to issues of concentration which we will ignore), we have that

$$\|\widehat{\mu} - \mu\|_2 \leq \varepsilon^{1-1/k} \cdot C_k^{1/k} + \varepsilon^{1-1/k} \|\widehat{T}\|_{\text{inj}}^{1/k} \; .$$

We'll do one term in the analysis to demonstrate how such an expression arises. Recall that if we let $\Delta = \widehat{\mu} - \mu$, then a key term we have to bound in order to upper bound $\|\widehat{\mu} - \mu\|_2^2$ is

$$\frac{1}{|S|} \sum_{i \in S_{\text{bad}}} \langle \Delta, X_i - \mu \rangle \; .$$

By applying Jensen's inequality, we can write:

$$\begin{aligned}\left(\frac{1}{|S_{\text{bad}}|}\sum_{i\in S_{\text{bad}}}\langle \Delta, X_i-\mu\rangle\right)^k &\leq \frac{1}{|S_{\text{bad}}|}\sum_{i\in S_{\text{bad}}}\langle \Delta, X_i-\mu\rangle^k\\ &\leq \varepsilon^{-1}\sum_{i\in S}\langle \Delta, X_i-\mu\rangle^k\\ &\leq \varepsilon^{-1}\|\widehat{T}\|_{\text{inj}}\cdot\|\Delta\|_2^k\ ,\end{aligned}$$

and so this term's contribution to $\|\Delta\|_2^2$ is at most

$$\frac{1}{|S|}\sum_{i\in S_{\text{bad}}}\langle \Delta, X_i-\mu\rangle \leq \varepsilon^{1-1/k}\|\widehat{T}\|_{\text{inj}}^{1/k}\cdot\|\Delta\|_2\ ,$$

which is of the order that we need.

By following the logic of before, but once again using these higher-order Jensen's inequality style proofs (or equivalently, using Hölder's inequalities), one can also show that if you filter with the scores given by (4), then we maintain the safety conditions we need for the filter to succeed.

But again, we cannot perform such filtering procedures efficiently, because we cannot optimize $\|\cdot\|_{\text{inj}}$! But the key point is that in fact, if the distribution has certifiably bounded k-th moments with bound C_k, then (once again ignoring issues of concentration), one can show an analog of this result. Namely, one can more or less show that if S_{good} is a set of samples from such a distribution, then:

$$\|\widehat{\mu}-\mu\|_2 \leq \varepsilon^{1-1/k}\cdot C_k^{1/k}+\varepsilon^{1-1/k}\|\widehat{T}\|_{\text{sos}}^{1/k}\ .$$

But now it remains: how do we filter, using this lemma? We need to do something when $\|\widehat{T}\|$ is much larger than C_k to decrease its value without removing too many inliers. Recall the key property of the score in (4) is that we need non-negative scores τ_i so that $\sum_{i\in S_{\text{good}}}\tau_i \ll \sum_{i\in S}\tau_i$. The key insight will be to use the pseudo-expectation interpretation of the sos norm. Namely, let $\widetilde{\mathbb{E}}$ be a degree-k pseudo-expectation that satisfies $\|v\|_2^2=1$ which maximizes

$$\widetilde{\mathbb{E}}\left[\frac{1}{|S|}\sum_{i\in S}\langle X_i-\widehat{\mu}, v\rangle^k\right]\ .$$

We know that the value of this expression is $\|\hat{T}\|_{\text{sos}}$, which is supposed to be large. So, similar to before, we can just define

$$\tau_i=\widetilde{\mathbb{E}}[\langle X_i-\widehat{\mu}, v\rangle^k]\ .$$

The key point is the following: by linearity, we know that the sum of the τ_i over S is $\|\hat{T}\|_{\text{sos}}$, but if we only take the sum over the good points, their sum should be bounded since the good distribution is certifiably bounded. We leave working out all of the details of this as a good exercise for the reader.

3 An end-to-end SoS algorithm

There is an alternative formalism for using SoS which sounds more complicated to state in this context, but which turns out to be useful in other settings. Recall that one of the implications of the spectral-signatures style results is that if we could find a subset of points of size $(1-\varepsilon)n$ so that all k-th moments were suitably bounded, then the empirical mean of this set of points must be close to the true mean. In fact, we can encode the problem of finding such a "bounded core" completely in SoS! So we don't even need to plug SoS into some filtering algorithm: we can just read off the solution from the solution of the SoS program.

Here is the idea: we encode the subset selection problem by adding in auxiliary program variables w_i, $i = 1, \ldots, n$, which we enforce to have the constraint $w_i^2 = w_i$, so that the only solutions to these constraints are $w_i \in \{0, 1\}$. Then the constraint that we should select a subset of size $(1-\varepsilon)n$ can be encoded as the polynomial constraint $\sum_{i=1}^n w_i = (1-\varepsilon)n$. We can also then easily encode the empirical mean of the subset we chose as a linear constraint on the w's, or alternatively just define a new program variable μ so that

$$\widehat{\mu} = \frac{1}{(1-\varepsilon)|S|} \sum_{i \in S} w_i X_i \; .$$

The only annoying thing is that the constraint we would naturally want to enforce about this set of points is that its injective norm is bounded, which we can't easily encode in SoS directly. This is because it is a "for-all v" constraint, which SoS doesn't do well with natively. However, what SoS can fairly easily enforce is the constraint that the sos-norm of the set of points is bounded. This is because the set of SoS proofs (by last lecture) is a nice convex set: in fact, it is captured by the set of polynomials of the form $(x^{\otimes k/2})^\top M (x^{\otimes k/2})$, for M being PSD. So the idea is to add M itself as a program variable, so that the SoS program searches for an SoS proof of the boundedness of the subset of points chosen. Putting this all together, we arrive at the following polynomial program $\mathcal{A}$, which is a program in program variables $w_1, \ldots, w_n$, μ, and $M \in \mathbb{R}^{d^{k/2}} \times \mathbb{R}^{d^{k/2}}$, and it is specified by:

$$\mathcal{A} = \begin{cases} w_i^2 = w_i \text{for all } i \in [n] \\ \widehat{\mu} = \frac{1}{(1-\varepsilon)|S|} \sum_{i \in S} w_i X_i \\ M \succeq 0 \\ (1-\varepsilon)n \cdot C_k \|v\|_2^k - \sum_{i \in S} w_i \langle X_i - \widehat{\mu}, v \rangle^k \geq (1-\varepsilon)n \cdot (v^{\otimes k/2})^\top M (v^{\otimes k/2}) \text{ for all } v \in \mathbb{R}^d \end{cases} \tag{6}$$

We emphasize that while this program seems to have program variables v, again the program constraint is just really the constraint that there is an SoS proof of this fact, which is encoded *solely by M*.

The punchline is that once you do so, one can observe that the entire proof of the spectral signatures lemma can be lifted almost verbatim to SoS, because all we're using is simple facts like Hölder's inequality, which can be proven in SoS. In particular, what you can show is that

$$\mathcal{A} \vdash_{O(k)} \|\widehat{\mu} - \mu\|_2^2 \leq C_k^{2/k} \varepsilon^{2-2/k} \; .$$

So now, let $\widetilde{\mathbb{E}}$ be any pseudoexpectation that satisfies $\mathcal{A}$. But it still remains to somehow actually round this pseudoexpectation to an actual solution. By the above, we know that $\widetilde{\mathbb{E}}\left[\|\widehat{\mu} - \mu\|_2^2\right] \leq \varepsilon^{2-2/k}$. We claim that this implies that $\|\widetilde{\mathbb{E}}[\widehat{\mu}] - \mu\|_2^2 \varepsilon^{2-2/k}$, so that our solution can just be $\widetilde{\mathbb{E}}[\widehat{\mu}]$. Indeed, we note that

$$\begin{aligned} \|\widetilde{\mathbb{E}}[\widehat{\mu}] - \mu\|_2^2 &= \sup_{v : \|v\|_2 = 1} \langle v, \widetilde{\mathbb{E}}[\widehat{\mu}] - \mu \rangle^2 & (7) \\ &= \sup_{v : \|v\|_2 = 1} \widetilde{\mathbb{E}}[\langle v, \widehat{\mu} - \mu \rangle]^2 \; . & (8) \end{aligned}$$

But observe that for any polynomial f, we have that

$$0 \leq \widetilde{\mathbb{E}}[(f - \widetilde{\mathbb{E}}[f])^2] = \widetilde{\mathbb{E}}[f^2] - \widetilde{\mathbb{E}}[f]^2 \; ,$$

and so this implies that $\widetilde{\mathbb{E}}[\langle v, \widehat{\mu} - \mu \rangle]^2 \leq \widetilde{\mathbb{E}}[\langle v, \widehat{\mu} - \mu \rangle^2] \leq \widetilde{\mathbb{E}}[\|\widehat{\mu} - \mu\|_2^2] \leq C_k^{2/k} \varepsilon^{2-2/k}$, where the last line follows since we can prove Cauchy-Schwarz in SoS. Hence, we obtain that

$$\|\widetilde{\mathbb{E}}[\widehat{\mu}] - \mu\|_2^2 \leq C_k^{2/k} \varepsilon^{2-2/k} \; ,$$

which implies that we can just take our solution to be $\widetilde{\mathbb{E}}[\widehat{\mu}]$,as claimed.

References

[1] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

[2] Samuel B Hopkins and Jerry Li. How hard is robust mean estimation? *arXiv preprint arXiv:1903.07870*, 2019.

[3] Ilias Diakonikolas, Samuel B Hopkins, Ankit Pensia, and Stefan Tiegel. Sos certifiability of subgaussian distributions and its algorithmic applications. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1689–1700, 2025.

[4] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

[5] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.