SOS and the unit sphere: Sparse vectors, tensor decomposition, dictionary learning, and quantum separability

So far our main focus has been on optimizing polynomials over the Boolean cube, but as we've seen, the sos algorithm can be applied in more general settings. In particular several very interesting problems can be phrased as relate to the task of maximizing a polynomial over the *unit sphere*. That is, given some polynomial $p: \mathbb{R}^n \to \mathbb{R}$, compute or approximate $\max_{\|x\|=1} p(x)$ and/or find an input x that (approximately) achieves this maximum.[1]

[1] This problem is also sometimes known as the problem of computing the *injective tensor norm*, see (Harrow and Montanaro [2010]).

Some examples that we will discuss include the following:

Tensor PCA and tensor decomposition

In **principal component analysis (PCA)** we are given samples $x^1, \ldots, x^m$ of some distribution X over $\mathbb{R}^n$, and want to find the direction $v \in \mathbb{R}^n$ that maximizes $\sum_{i,j} v_i v_j M_{i,j}$ where $M_{i,j} = \frac{1}{m}\sum_{k=1}^m x_i^k x_j^k$. The idea is that if X has the form $v_0 + Y$ with $v_0 \in \mathbb{R}^n$ and Y a mean zero "noise variable" then $\mathbb{E}\, XX^\top = v_0 v_0^\top$ and hence we expect the direction v that maximizes this correlation to be proportional to v_0.

PCA only involves maximizing a *quadratic* polynomial which amounts to an efficiently solvable maximum eigenvalue computation. However, if for example X is a *mixture* with equal weights of two distributions of the form $v_0 + Y$ and $v_1 + Y'$ (where v_0, v_1 are unit vectors, and assume they are orthogonal for simplicity) then one can see that $\mathbb{E}\, XX^\top$ would be proportional to the identity linear operator on the subspace $Span\{v_0, v_1\}$. Hence, even if we had an infinite number of samples and could get the expectation of the second moment matrix $\mathbb{E}\, XX^\top$ precisely, we would still not be able to recover v_0 and v_1 but rather only the subspace that they span. However, it can be shown that in this case v_0 and v_1 will be the only two global maximizers of $\sum_{T_{i,j,k}} x_i x_j x_k$ where $T_{i,j,k} = \mathbb{E}\, X_i X_j X_k$. So, recovering these centers reduces to maximizing this polynomial.

More generally the **tensor PCA** problem (see (Richard and Montanari [2014])) is defined as follows: we are given the d level moments $\mathbb{E}\, X^{\otimes d}$ of some random variable X over $\mathbb{R}^n$, and our goal is to find a vector x maximizing the value of $p(x) = \mathbb{E}\langle X, x\rangle^d$. It is an extremely useful generalization of PCA though unfortunately it is NP hard on average.

If tensor PCA is the higher degree analog of computing the max-

imum eigenvalue, the **tensor decomposition problem** (see (Kolda and Bader [2009])) is the higher degree analog of *singular value decomposition*. Namely, given some tensor $T \in \mathbb{R}^{n^d}$ we want to find the smallest r and dr n-dimensional vectors $\{v_{i,j}\}_{i\in[r],j\in[d]}$ so we can write T as $\sum_{i=1}^{r} v_{i,1} \otimes \cdots \otimes v_{i,d}$. This is an incredibly useful primitive and many of the known algorithms and heuristics for it are obtained by iteratively running subroutines for tensor PCA and then "peeling off" the resulting vectors.[2]

[2] This viewpoint of tensor decomposition being a generalization of tensor PCA to multiple vectors is somewhat inaccurate. The two problems are somewhat orthogonal. Tensor decomposition generalizes the problem of finding the smallest matrices B and C such that $A = BC$ for a given matrix A.

Finding sparse vectors in subspaces

In the **sparse vector problem** we are given an n dimensional subspace $V \subseteq \mathbb{R}^m$ and want to find the nonzero vector $v \in V$ that is the *sparsest* possible, in the sense of having as few nonzero entries as possible (or in the sense that a few of the entries have large magnitude and the rest very small one). This is a natural problem, that can be thought of as a continuous variant of finding the *shortest codeword* in a linear code. It also arises in a variety of applications (including the sparse coding and small set expansion applications mentioned below), see Demanet and Hand [2014] . In particular in the context of *compressed sensing* certifying that a subspace does not contain a sparse vector is closely related to certifying the *restricted isometry property* (Candès and Tao [2005]) of subspaces.

The sparse vector problem does not immediately translate into maximizing polynomial over the sphere, but it turns out that sparseness of a vector v can be approximated by the relation of $\|v\|_q / \|v\|_p$ for $q > p$. Alas, there appears to be a subtle tradeoff between the quality of this approximation and the tractability of computing this ratio. For example, for $p = 1$ and $q = \infty$ this can be efficiently computed (**exercise**) but only yields an $\tilde{O}(\sqrt{n})$ approximation. One can get very good approximation guarantees by considering $p = 1$ and $q = 2$ but no efficient algorithm is known for this case. For $p = 2$ and $q = 4$ the problem becomes one of computing $\max_{\|x\|_2=1} p(x)$ where $p(x) = \|Ax\|_4^4$ and $A\colon \mathbb{R}^n \to \mathbb{R}^m$ is a generating matrix for V. It turns out that this formulation yields non-trivial guarantees, and while it is NP hard in general, it can be efficiently computed via SOS in several important cases.

Sparse coding

In the **sparse coding** problem (also known as *dictionary learning*) we are again given samples $x^1, \ldots, x^m$ of some distribution over $\mathbb{R}^n$ and our goal is to find a basis $A = \{a_1, \ldots, a_n\}$ for $\mathbb{R}^n$ which maximizes the average *sparsity* of the vectors $\{(\langle a_1, x^i \rangle, \ldots, \langle a_n, x^i \rangle)\}_{i=1,\ldots,m}$.[3] The intuition behind this is that the sparse representation is often the "right" one, just as sounds tend to be sparser when represented in the Fourier base, images in the Wavelet base, etc... . Indeed, while for a generic basis $A = \{a_1, \ldots, a_n\}$, a sample x would have most of the values $\langle a_i, x \rangle$ be of similar magnitude, in a *sparse* representations we can interpret the a_i's as meaningful *features* that are turned on or off depending on this magnitude.

[3] Often we consider also A's that are not bases of $\mathbb{R}^n$ but merely span it, and hence also consider the case $|A| > n$ which is known as the *overcomplete* case.

Indeed, Olshausen and Field [1997] suggested that sparse coding may be used as a strategy for the visual cortex and many deep neural networks use sparse coding as a way to generate their bottom-most layer. One way to solve the sparse coding problem is to try to recover the elements of A one vector at a time by trying to find sparse vectors in the subspace $V = Image(X)$ of $\mathbb{R}^m$ where X is the $m \times n$ matrix whose columns are $x^1, \ldots, x^m$.

Quantum information theory, quantum entanglement and QMA(2)

In quantum information theory, a system with N classical states (such as a system of $\log n$ bits) is modeled as an $N \times N$ positive semidefinite matrix ρ of trace 1 known as the *density matrix* of the state. A *classical* state, which can be thought of as a distribution $(p_1, \ldots, p_N)$ over the N different outcomes corresponds to the special case when ρ is diagonal. A quantum *pure* state corresponds to the case where ρ is rank one- $\rho = vv^\top$ for some unit vector $v \in \mathbb{R}^N$. A *mixed* state corresponds to the more general case where the matrix is not rank one. Note that by singular value decomposition, any mixed state ρ can be written as $\rho = \sum_{i=1}^N p_i v_i v_i^\top$ where the v_i's are unit and the p_i's are non-negative and sum up to one.[4]

[4] Quantum states are generally *complex* vectors but essentially all of their interesting phenomena already arise in the real case in which they can be thought of as "probabilities with negative numbers". In this lecture we will restrict attention to the real case but everything we say can be generalized to the complex case by interpreting $v^\top$ as a the complex adjoint of v and interpreting quantities such as x^2 as $xx^\top = |x|^2$.

If we have two systems each of M states, then we can think of the combined system as a system of $N = M^2$ states (e.g., two one-bit systems are the same as a single two-bit system). One of the most mysterious phenomena of quantum states is that they can create *entanglement* between the different subsystems. A pure state $v \in \mathbb{R}^N$ is *separable* (i.e., non entangled) if $v = ww^\top$ for some $w \in \mathbb{R}^M$.[5] In

[5] Actually this is the definition for a state that is both *symmetric* and *separable*: a separable state would have the form $ww'^\top$ for some w, w'. In other words, a separable state on an M^2 state system can be interpreted as a rank one $M \times M$ matrix while we make the additional requirement that this matrix is symmetric (or Hermitian in the complex case). We restrict to the symmetric case for notational convenience and easier relation to other problems though it does not make much of a difference.

other words the density matrix has the form $w^{\otimes 4}$. A mixed state ρ is separable if it is a convex combination of pure separable states. That is, $\rho = \sum_i p_i w_i^{\otimes 4}$ for non-negative p_i's that sum of up to one.

One of the ways that the complexity of entanglement is manifested is that it is NP hard to determine if a state is separable or not. But one can ask how hard is it to *approximate*. A *quantum measurement* on an N-state system is a p.s.d. $N \times N$ matrix $\mathcal{M}$ such that $\mathcal{M} \preceq I$. The probability that $\mathcal{M}$ accepts a state ρ is simpy $\langle \mathcal{M}, \rho \rangle = \mathrm{Tr}(\mathcal{M}\rho)$. (Can you see why this number is always between zero and one?.) Two states ρ, ρ' are identical if and only if they are accepted with the same probability for all possible measurements (**exercise**). We say that a state ρ is ϵ-separable if there is some separable ρ' such that $|\mathrm{Tr}(\mathcal{M}\rho) - \mathrm{Tr}(\mathcal{M}\rho')| \leq \epsilon$ for all measurements $\mathcal{M}$. The *quantum separability problem* with parameter ϵ is to distinguish, given a density matrix ρ, between the YES case that ρ is separable and the NO case that ρ is not ϵ separable. Assuming the exponential time hypothesis, the quantum separability problem requires at least $N^{\Omega(\log N)}$ time for every constant $\epsilon > 0$ (Harrow and Montanaro [2010]). Doherty et al. [2004] proposed using sos for this problem, but we still do not know the degree required for this problem in the worst-case. Brandão et al. [2011] showed that it does work in $O(\log N)$ degree if we consider a relaxed notion of ϵ-separability that only restricts attention to particular types of measurements (known as local operations and one-way communication or one-way LOCC).

A related problem is the *Best Separable State (BSS)* problem where one is given a measurement $\mathcal{M}$ and paramters $1 > c > s > 0$ as input and needs to distinguish between the case that there is a separable state ρ with $\mathrm{Tr}(\mathcal{M}\rho) \geq c$ and the case that every separable state ρ satisfies $\mathrm{Tr}(\mathcal{M}\rho) \leq s$. (A natural setting of parameters is $c = 1$ and $s = 1 - \epsilon$.) This problem is also known as finding the acceptance probability of a *quantum arthur merlin verifier with two provers*, since we can think of $\mathcal{M}$ as a verfiying algorithm that receives two quantum states from two provers that are guaranteed to be non-entangled. A quasipolynomial time algorithm for this problem would correspond to placing $QMA(2)$ in EXP while the best known upper bounds is $QMA(2) \in EE = Dtime(2^{O(N)})$ (in this problem N corresponds to 2^n where n is the number of qubits in this proofs). The problem of determining the complexity of $QMA(2)$ is of intense interest to quantum information theorists and recently University of Maryland's QuiCS center held a weeklong workshop dedicated solely to this problem.

It is not hard to prove that there is always a *pure* state maximizing the acceptance probability of any measurement $\mathcal{M}$ and

hence the BSS problem corresponds to finding the maximum of $p(x) = \mathrm{Tr}(\mathcal{M}(xx^\intercal)^{\otimes 2})$ over all unit vectors x.[6]

[6] Once again, recall that we restrict attention to the real symmetric case. In the literature this problem is typically written as maximizing $\mathrm{Tr}(\mathcal{M}(x \otimes y)(x \otimes y)^*)$ over all unit complex x, y or in quantum notation maximizing $\langle x, y|\mathcal{M}|x, y\rangle$.

What does this have to do with sos?

It turns out that all these problems share similar characteristics:

- They are NP hard to solve exactly in the worst-case. Indeed, that's true for almost all problems involving tensors (Hillar and Lim [2013]).
- The algorithm with the best known rigorous guarantees for these problems is the sos hierarchy.
- Often we can *prove* that the sos algorithm provides non-trivial guarantees in the worst-case and/or average-case setting that go beyond what other algorithms can achieve.
- We typically do not have tight bounds on the performance of the sos algorithm for these problems in neither the worst-case nor the average-case setting.
- Some of these problems have natural heuristics that people apply in practice. More often than not we do not know how to analyze the performance of thee heuristics.

Beyond these superficial similarities, it turns out that there are some deep and fascinating technical connections between these problems as well as other important questions in theoretical CS such as the *Unique Games Conjecture* Khot [2002] and the *Log Rank Conjecture* Lovász and Saks [1988] . We will see some of these results and connections in the next lectures.

References

Fernando G. S. L. Brandão, Matthias Christandl, and Jon Yard. A quasipolynomial-time algorithm for the quantum separability problem. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 343–352, 2011. doi: 10.1145/1993636.1993683. URL http://doi.acm.org/10.1145/1993636.1993683.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Information Theory*, 51(12):4203–4215, 2005.

Laurent Demanet and Paul Hand. Scaling law for recovering the sparsest element in a subspace. *Information and Inference*, page iau007, 2014.

Andrew C Doherty, Pablo A Parrilo, and Federico M Spedalieri. Complete family of separability criteria. *Physical Review A*, 69(2): 022308, 2004.

Aram Wettroth Harrow and Ashley Montanaro. An efficient test for product states with applications to quantum merlin-arthur games. In *FOCS*, pages 633–642. IEEE Computer Society, 2010.

Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *J. ACM*, 60(6):Art. 45, 39, 2013. ISSN 0004-5411. doi: 10.1145/2512329. URL http://dx.doi.org/10.1145/2512329.

Subhash Khot. On the power of unique 2-prover 1-round games. In *IEEE Conference on Computational Complexity*, page 25. IEEE Computer Society, 2002.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

László Lovász and Michael E. Saks. Lattices, möbius functions and communication complexity. In *FOCS*, pages 81–90. IEEE Computer Society, 1988.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23): 3311–3325, 1997.

Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *NIPS*, pages 2897–2905, 2014.