Lecture 6: Global correlation rounding

So far, in most of the examples we have seen, the "rounding" step, in which we take the pseudoexpectation operator and use it to produce a solution, was trivial. However, this is not always the case, especially in non-parametric settings where there is no parameter or "planted" solution that we are looking for. In this lecture we'll introduce a powerful and general technique called "global correlation rounding." Some bibliographic remarks will be deferred to the end.

These notes have not been reviewed with the same scrutiny applied to formal publications. There may be errors.

1 Rounding pseudoexpectations

In the applications we have seen so far (robust mean estimation, recovering communities in stochastic block models, clustering mixtures of Gaussians), our goal was to recover some specific parameter or planted structure; call this parameter θ^*. We gave an SoS proof of strong identifiability for these parameters under conditions which we encoded as polynomial constraints on variables θ: the proof argues that the only choice of θ which satisfies these constraints must be very close to θ^*. From this we were able to argue that $\theta^* \approx \widetilde{\mathbf{E}}[\theta]$, or sometimes that θ^* was an obvious function of $\widetilde{\mathbf{E}}[\theta\theta^\top]$.

In another type of application, we use SoS to solve a *variational* or *optimization* problem. In such applications, we have some variables X, and we are interested in finding the choice of X which maximizes or minimizes some function, subject to constraints. We have already seen one example in which solving a variational problem is helpful for inference: in stochastic block models with average degree $\Omega(\log n)$, we reduced the clustering/community recovery problem to finding a partition which minimized (or maximized) the number of edges cut. As another example, we could try to find only one block model community at a time, by considering the following variational problem:

Problem 1.1 (Densest-m-subgraph)**.** We are given a social network graph $G = (V, E)$ with $|V| = n$. We'd like to find a subset of m of the nodes/people in the network which are most well-connected. We can encode this problem with the following polynomial system:

$$\max_x \ X^\top A_G X, \qquad \text{subject to} \quad X_i^2 = X_i \ \forall i \in [n], \quad \sum_i X_i = m.$$

In this problem, there may be several dense communities, and the maximizing X may not be unique. Even if we had access to an *actual* distribution D over solutions to the program, via low-degree moments $\mathbf{E}_X : \mathbb{R}[X] \to \mathbb{R}$, it is not clear how we might use it to find a maximizing X. If we don't know all of the moments of D, how can we sample $X \sim D$, or even find one X in the support of D?

Global correlation rounding is a technique which, under favorable conditions, can use a bounded-degree moment operator to find some x such that $c(x) \approx \mathbf{E}_X[c(X)]$, when D is a distribution over X taking values in a finite alphabet. In this lecture, we'll focus exclusively on the case when $X \in \{0,1\}^n$. Furthermore, the proof of global correlation rounding applies equally well to pseudodistributions, so we can apply it to round SoS semidefinite relaxations to polynomial optimization problems.

In this lecture we will apply global correlation rounding to the densest subgraph problem; in a later lecture we may see it applied to mean-field approximations in Ising models [JKR19].

2 Independent rounding and local correlation

One way to produce a solution from (or "round") a degree-$\geqslant 1$ moment operator $\mathbf{E}_X$ satisfying the axioms $\{X_i^2 = X_i\}_{i\in[n]}$ is using *independent rounding*:

Definition 2.1 (Independent rounding)**.** Given access to a (pseudo)moment oracle $\mathbf{E}$ of degree $\geqslant 1$ for $X_1, \dots, X_n$ satisfying the axioms $\{X_i^2 = X_i\}_{i\in[n]}$, the *independent rounding* algorithm produces a solution $Y_1, \dots, Y_n$ by sampling $Y_i \sim \mathsf{Ber}(\mathbf{E}[X_i])$.

How good is the solution produced by independent rounding? If we care about the value of a linear function of X, then by linearity of expectation, $\mathbf{E}_Y[\langle c, Y\rangle] = \mathbf{E}_X[\langle c, X\rangle]$, so at least for linear functions the expected value is good. But in the densest-m-subgraph problem, our objective function is quadratic. In this case, our error is measured by the *"local correlation:"*

Definition 2.2 (Local correlation)**.** For a graph $G = (V, E)$ with adjacency matrix A_G and a (pseudo)distribution with (pseudo)moments $\mathbf{E}_X$, the local correlation is the quantity

$$\mathrm{loc}(\mathbf{E}_X, G) = \mathop{\mathbf{E}}_{(i,j)\sim\mathsf{Unif}(E)} \left(\mathbf{E}_X[X_iX_j] - \mathbf{E}_X[X_i]\,\mathbf{E}_X[X_j]\right).$$

In words, it is the average (pseudo)covariance across edges of G.

The local correlation precisely quantifies the error in the objective function $X^\top A_G X$ under independent rounding:

Claim 2.3. If $Y \in \{0,1\}^n$ is produced from $\mathbf{E}_X$ via independent rounding, then

$$\frac{1}{2|E(G)|}\left(\mathbf{E}_X[X^\top A_G X] - \mathbf{E}_Y[Y^\top A_G Y]\right) = \mathrm{loc}(G, \mathbf{E}_X).$$

Proof. This follows from linearity, using that when $i \neq j$, $\mathbf{E}_Y[Y_iY_j] = \mathbf{E}_Y[Y_i]\,\mathbf{E}_Y[Y_j] = \mathbf{E}_X[X_i]\,\mathbf{E}_X[X_j]$, and for the diagonal $i = j$ terms, $\mathbf{E}_Y[Y_i^2] = \mathbf{E}_Y[Y_i] = \mathbf{E}_X[X_i] = \mathbf{E}_X[X_i^2]$. □

Hence if we can guarantee that the local correlation is small, then we have a bound on the error of independent rounding. Unfortunately, the local correlation of $\mathbf{E}_X$ may be quite high. In what follows, we'll describe a general procedure for decreasing what is called the *global correlation*, and how in some cases global and local correlation can be related.

3 Global correlation

A priori we have no control over the local correlation of $\mathbf{E}_X$. However, we can control a different quantity, which we call the *global correlation*:

Definition 3.1 (Global correlation)**.** For a (pseudo)distribution over $X \in \mathbb{R}^n$ with (pseudo)moments $\mathbf{E}_X$, the *global correlation* is the quantity

$$\mathrm{glob}(\mathbf{E}_X) = \mathop{\mathbf{E}}_{i,j\sim\mathsf{Unif}([n])} \left[(\mathbf{E}_X[X_iX_j] - \mathbf{E}_X[X_i]\,\mathbf{E}_X[X_j])^2\right].$$

The global correlation measures the average covariance under D_X among pairs of variables chosen uniformly from $[n]$. If $\mathrm{loc}(G, \mathbf{E}_X)$ is close to zero, then we conclude that a randomly chosen pair of variables is not too correlated on average. Notice that this is the square of the average covariance, whereas in $\mathrm{loc}(G, \mathbf{E})$ we do not have the square. This discrepancy in the definitions is a matter of technical convenience.

To gain some intuition for why the global covariance could be large, consider the following situation: think of the densest-(n/k)-subgraph problem, and suppose that G comes from a k-community block model, as defined in Lecture 3, in the case when the inside-community probability is larger than the outside-community probability. If we take D_X to be uniformly distributed over the communities, then

$$\mathrm{glob}(\mathbf{E}_X) = (1 - \tfrac{1}{k}) \cdot \left(0 - \frac{1}{k^2}\right)^2 + \tfrac{1}{k} \cdot \left(\frac{1}{k} - \frac{1}{k^2}\right)^2 \sim \frac{1}{k^3},$$

which is non-negligible. The reason that these global correlations are occuring is that the distribution is actually a mixture over simpler distributions (one for each community). However if we can re-weight the distribution so that X is supported only on one of the communities, then the average correlation between variables drops (if X is fixed on one community, then X_i and X_j are always independent).

The block model example above hints at the following fact: one can always reduce the global correlation by conditioning. Before we see this, let's see why local and global correlation may be related.

3.1 Relating local and global correlation

A priori, there is no reason for us to suspect that global and local correlation are related; neighboring variables in G may be much more correlated than uniformly chosen pairs of variables. However, we'll show that under certain structural assumptions, local and global correlation can be related.

Lemma 3.2. *Suppose that G is a d-regular graph,[1] and suppose as well that its normalized adjacency matrix $\frac{1}{d}A_G$ has at most k eigenvalues exceeding some threshold $\tau \in [1, 0]$. Define $\mathbf{Var}(\mathbf{E}_X) = \mathbf{E}_{i\sim[n]} \mathbf{Var}(X_i)$, where the variance is taken with respect to D_X. Then*

$$\mathrm{loc}(\mathbf{E}_X, G) \leqslant (1 - \tau) \cdot \sqrt{k \cdot \mathrm{glob}(\mathbf{E}_X)} + \tau \cdot \mathbf{Var}(\mathbf{E}_X)$$

A graph with at most 1 eigenvalue close to 1 is an *expander*, which roughly means that there are no too-sparse cuts. Within an expander, it makes sense that large local correlation (and bounded average variance) might imply some nontrivial global correlation, since the expansion means that all variables are well-connected; this might offer some intuition for the case when $k = 1$. For larger k, intuitively, a graph with at most k eigenvalues exceeding τ can be partitioned into $f(k)$ subgraphs for some function f, where the partition does not cut too many edges, and where each piece is an expander inside; the lemma above shows that the local-to-global phenomenon exists in such graphs as well, albeit the effect is weaker due to the possible presence of sparse cuts.

Proof of Lemma 3.2. Let $\Sigma_X = \mathbf{E}_X[XX^\top] - \mathbf{E}_X[X]\,\mathbf{E}_X[X^\top]$ be the covariance matrix of D_X. By definition,

$$\mathrm{loc}(\mathbf{E}_X, G) = \frac{1}{dn} \langle \Sigma_X, A_G \rangle .$$

By virtue of being a covariance matrix, $\Sigma_X \succeq 0$. Since $\frac{1}{d}A_G$ is a normalized adjacency matrix, all of its eigenvalues lie in $[1, -1]$. Writing Π as the projector to the span of the top k eigenvalues of $\frac{1}{d}A_G$, we have

[1] This requirement can be relaxed, but we make it here for simplicity.

by the positive-semidefiniteness of Σ_X,

$$\begin{aligned}\mathrm{loc}(\mathbf{E}_X, G) = \frac{1}{n}\left\langle \Sigma_X, \tfrac{1}{d}A_G \right\rangle &\leqslant \frac{1}{n}\left(\langle \Sigma_X, \Pi\rangle + \tau \cdot \langle \Sigma_X, I - \Pi\rangle\right) \\ &= \frac{1}{n}(1-\tau)\cdot\langle \Sigma_X, \Pi\rangle + \frac{\tau}{n}\cdot \mathsf{Tr}(\Sigma_X)\end{aligned}$$

And now applying Cauchy-Schwarz to the first term and noting that the second term is a scaling of $\mathbf{Var}(\mathbf{E}_X)$,

$$\leqslant \frac{1}{n}(1-\tau)\cdot\|\Sigma_X\|_F \cdot \|\Pi\|_F + \tau\cdot\mathbf{Var}(\mathbf{E}_X).$$

But now by inspection, $\mathrm{glob}(\mathbf{E}_X) = \left(\frac{1}{n}\|\Sigma_X\|_F\right)^2$. Further, since Π is a projector to a dimension-k subspace, $\|\Pi\|_F = \sqrt{k}$. Thus,

$$= (1-\tau)\sqrt{k\cdot\mathrm{glob}(\mathbf{E}_X)} + \tau\mathbf{Var}(\mathbf{E}_X),$$

which gives the result □

3.2 Conditioning reduces global correlation

The reason that Lemma 3.2 is useful algorithmically is that, as we will see below, we can lower the global correlation by *conditioning*.

Lemma 3.3. *Let $X_1, \ldots, X_n$ be random variables taking values in $\{0,1\}$. Suppose we sample $i_1, \ldots, i_\ell \sim \mathsf{Unif}([n])$, and then sequentially set*

$$x_{i_t} \sim \mathsf{Ber}\left(\mathbf{E}[X_{i_t} \mid X_{i_1} = x_{i_1}, \ldots, X_{i_{t-1}} = x_{i_{t-1}}]\right).$$

Then there exists some $t \leqslant \ell$ such that

$$\mathop{\mathbf{E}}_{x_{i_1},\ldots,x_{i_t}} \mathrm{glob}(\mathbf{E}[\cdot \mid X_{i_1} = x_{i_1}, \ldots, X_{i_t} = x_{i_t}]) \leqslant \frac{2\log 2}{\ell}.$$

What this lemma says is that conditioning lets us decompose our measure into a mixture of measures that are "simpler," in that they are closer to product measures. Of course, one can write any discrete measure as a mixture over point masses (that is, by conditioning on the values of all of the variables), and point masses are trivially product measures. The interesting thing here is that we can get close to a product measure much before conditioning on everything, where we quantify the distance to a product measure in terms of the global correlation. See also [Eld20] for a different take on decomposing measures.

Proof of Lemma 3.3. For a joint distribution μ on variables Z_1, Z_2, let μ_i denote the marginal on Z_i, and let $\mu_1 \otimes \mu_2$ denote the product of μ_1 and μ_2. The proof makes elegant use of some information-theoretic inequalities. In particular, we'll exploit the relationship between covariance and *mutual information*:

Definition 3.4. If Z_1, Z_2 are jointly distributed random variables with joint distribution μ, the *mutual information of Z_1, Z_2* is the quantity

$$I(Z_1; Z_2) = \mathrm{D}(\mu \| \mu_1 \otimes \mu_2) = \mathop{\mathbf{E}}_{Z_1, Z_2 \sim \mu} \log \frac{\mathbf{Pr}_\mu[Z_1, Z_2]]}{\mathbf{Pr}_{\mu_1}[Z_1]\mathbf{Pr}_{\mu_2}[Z_2]}.$$

Claim 3.5. If Z_1, Z_2 are random variables taking values in a set $\Sigma \subset \mathbb{R}$ with $\max_{s\in\Sigma} |s| \leqslant M$,

$$\mathbf{Cov}(Z_1, Z_2)^2 \leqslant 2M^4 \cdot I(Z_1; Z_2).$$

Proof. The proof uses Pinsker's inequality, which states that for measures π, ν, $\mathrm{d_{TV}}(\nu, \pi) \leqslant \sqrt{\frac{1}{2}\mathrm{D}(\nu\|\pi)}$. We simply related the total variation distance and the covariance,

$$\begin{aligned}
\mathrm{d_{TV}}(\mu, \mu_1 \otimes \mu_2) &= \frac{1}{2}\sum_{z_1,z_2} |\mathbf{Pr}[Z_1 = z_1, Z_2 = z_2] - \mathbf{Pr}[Z_1 = z_1]\mathbf{Pr}[Z_2 = z_2]| \\
&\geqslant \frac{1}{2}\sum_{z_1,z_2} \left|\frac{z_1}{M}\frac{z_2}{M}\right| |\mathbf{Pr}[Z_1 = z_1, Z_2 = z_2] - \mathbf{Pr}[Z_1 = z_1]\mathbf{Pr}[Z_2 = z_2]| \\
&\geqslant \frac{1}{2M^2}\left|\sum_{z_1,z_2} z_1 z_2 (\mathbf{Pr}[Z_1 = z_1, Z_2 = z_2] - \mathbf{Pr}[Z_1 = z_1]\mathbf{Pr}[Z_2 = z_2])\right| \\
&= \frac{1}{2M^2}|\mathbf{Cov}(Z_1, Z_2)|,
\end{aligned}$$

where the first inequality uses that $z_i \leqslant M$ and the second inequality is the triangle inequality. The conclusion follows by applying Pinsker's inequality to μ vs. $\mu_1 \otimes \mu_2$. □

We'll use the entropy to bound the mutual information.

Definition 3.6. If Z_1, Z_2 are discrete random variables, the *entropy* of Z_1 is defined as the quantity

$$H(Z_1) = \mathop{\mathbf{E}}_{z_1\sim\mu} \log \frac{1}{\mathbf{Pr}[Z_1 = z_1]},$$

and the *entropy of Z_1 conditioned on Z_2* is the quantity:

$$H(Z_1 \mid Z_2) = \mathop{\mathbf{E}}_{z_2\sim\mu_2} H(Z_1 \mid Z_2 = z_2)$$

Claim 3.7. The mutual information may be related to the entropy as follows:

$$I(Z_1; Z_2) = H(Z_1) - H(Z_1 \mid Z_2).$$

Proof. Using properties of the logarithm and our definitions, we have the chain of equalities,

$$\begin{aligned}
I(Z_1; Z_2) &= \mathop{\mathbf{E}}_{z_1,z_2\sim\mu} \log \frac{\mathbf{Pr}[Z_1 = z_1, Z_2 = z_2]}{\mathbf{Pr}[Z_1 = z_1]\mathbf{Pr}[Z_2 = z_2]} \\
&= \mathop{\mathbf{E}}_{z_1,z_2\sim\mu} \left(\log \frac{\mathbf{Pr}[Z_1 = z_1, Z_2 = z_2]}{\mathbf{Pr}[Z_2 = z_2]} + \log \frac{1}{\mathbf{Pr}[Z_1 = z_1]}\right) \\
&= \mathop{\mathbf{E}}_{z_1,z_2\sim\mu} \left(-\log \frac{1}{\mathbf{Pr}[Z_1 = z_1 \mid Z_2 = z_2]} + \log \frac{1}{\mathbf{Pr}[Z_1 = z_1]}\right) \\
&= H(Z_1) - H(Z_1 \mid Z_2).
\end{aligned}$$

□

Finally, define the conditional mutual information

$$I(X_{i_1}; X_{i_t} \mid X_{i_2}, \ldots, X_{i_{t-1}}) = \mathop{\mathbf{E}}_{x_{i_2},\ldots,x_{i_{t-1}}\sim D_X} I(X_{i_1}; X_{i_t} \mid X_{i_2} = x_{i_2}, \ldots, X_{i_{t-1}} = x_{i_{t-1}}).$$

Our proof now proceeds to bound the conditional mutual information as follows. By Claim 3.7, for any $t \geqslant 2 \in \mathbb{Z}$ and any $i_1, \dots, i_t \in [n]$,

$$I(X_{i_1}; X_{i_t} \mid X_{i_2}, \dots, X_{i_{t-1}}) = H(X_{i_1} \mid X_{i_2}, X_{i_2}, \dots, X_{i_{t-1}}) - H(X_{i_1} \mid X_{i_1}, \dots, X_{i_{t-1}}, X_{i_t}).$$

In particular, we can average this equality over $t \in \{2, \dots, \ell + 2\}$,

$$\begin{aligned}\frac{1}{\ell}\sum_{t=2}^{\ell+2} I(X_{i_1}; X_{i_t} \mid X_{i_2}, \dots, X_{i_{t-1}}) &= \frac{1}{\ell}\sum_{t=2}^{\ell+2} \left(H(X_{i_1} \mid X_{i_2}, X_{i_2}, \dots, X_{i_{t-1}}) - H(X_{i_1} \mid X_{i_1}, \dots, X_{i_{t-1}}, X_{i_t})\right) \\ &= \frac{1}{\ell}(H(X_{i_1}) - H(X_{i_1} \mid X_{i_1}, \dots, X_{i_{\ell+2}})),\end{aligned}$$

where we have used that the sum telescopes. But now, since the entropy is non-negative always, and since $H(X_{i_1}) \leqslant \ln 2$ for X_{i_1} taking values in $\{0, 1\}$,

$$\leqslant \frac{\ln 2}{\ell}.$$

Taking the expectation over $i_1, \dots, i_t \sim [n]$,

$$\mathop{\mathbf{E}}_{i_1,\dots,i_{\ell+2}\sim[n]} \frac{1}{\ell}\sum_{t=2}^{\ell+2} I(X_{i_1}; X_{i_t} \mid X_{i_2}, \dots, X_{i_{t-1}}) \leqslant \frac{\ln 2}{\ell}.$$

Applying the fact that the indices $i_1, \dots, i_t$ are exchangeable, we also have

$$\mathop{\mathbf{E}}_{i_1,\dots,i_{\ell+2}\sim[n]} \frac{1}{\ell}\sum_{t=2}^{\ell+2} I(X_{i_1}; X_{i_2} \mid X_{i_3}, \dots, X_{i_t}) \leqslant \frac{\ln 2}{\ell}.$$

Hence there must be some $t \in \{2, \dots, \ell + 2\}$ for which

$$\mathop{\mathbf{E}}_{i_3,\dots,i_t\sim[n]} \left[\mathop{\mathbf{E}}_{i_1,i_2\sim[n]} I(X_{i_1}; X_{i_t} \mid X_{i_2}, \dots, X_{i_{t-1}})\right] \leqslant \frac{\ln 2}{\ell}.$$

Finally, applying Claim 3.5 pointwise with the definition of global correlation gives the result. □

4 SoS algorithms via global correlation rounding

Global correlation rounding, the algorithm. If we had access to the degree-$\ell + 2$ moments of a distribution over solutions to a variational problem, such as densest-m-subgraph, we now have an algorithm which finds a solution. Then we run the following algorithm:

1. **Condition.** Sample $i_1, \dots, i_\ell \sim [n]$, then for each $t \leqslant \ell$, sample a value
$$x_{i_{t+1}} \sim \mathsf{Ber}(\mathbf{E}[X_{i_{t+1}} \mid X_{i_1} = x_{i_1}, \dots, X_{i_t} = x_{i_t}]).$$
2. **Choose a value of t.** For each $t \in \{0, \dots, \ell\}$, check if $\mathrm{glob}(\mathbf{E}[\cdot \mid X_{i_1} = x_{i_1}, \dots, X_{i_t} = x_{i_t}])$ has global correlation at most $4\frac{\ln 2}{\ell}$. If yes, fix this value of t.
3. **Independent rounding.** Sample $Y_i \sim \mathsf{Ber}(\mathbf{E}[X_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_t} = x_{i_t}])$ independently for each $i \in [n]$.

If step 2 fails, and the conditioning did not drop the global correlation, then we can try the whole procedure again—Markov's inequality guarantees that the global correlation for the minimizing t will exceed $4 \ln 2/\ell$ with probability at most $\frac{1}{2}$. Also, in expectation, our objective value can be bounded from below (using Claim 2.3):

$$\frac{1}{|E(G)|} \mathop{\mathbf{E}}_{Y, x_{i_1}, \ldots, x_{i_\ell}} [Y^\top A Y] = \frac{1}{|E(G)|} \mathop{\mathbf{E}}_{X} [X]^\top A \mathop{\mathbf{E}}_{X} [X] = \frac{1}{|E(G)|} \mathop{\mathbf{E}}_{X} [X^\top A X] - \mathrm{loc}(\mathop{\mathbf{E}}_{X}, G),$$

and we can show that this quantity concentrates (it is a degree-2 polynomial in independent random variables).

Suppose now that $\frac{1}{d} A_G$ has at most k eigenvalues of eigenvalue larger than τ, as in the setting of Lemma 3.2. In fact, the k-community stochastic block model is one example of such a graph: as we showed in a prior lecture, when the degree is sufficiently large, the normalized adjacency matrix has k eigenvalues close to $\frac{1}{k}$, and all other eigenvalues $o(1)$. Applying the dense-(n/k)-subgraph optimization in such a setting, we would have that our error is bounded by

$$\left| \frac{1}{|E(G)|} \mathop{\mathbf{E}}_{Y, x_{i_1}, \ldots, x_{i_\ell}} [Y^\top A Y] - \frac{1}{|E(G)|} \mathop{\mathbf{E}}_{X} [X^\top A X] \right| \leqslant \mathrm{loc}(\mathbf{E}_X, G) \leqslant (1 - o(1)) \sqrt{k \cdot \frac{4 \ln 2}{\ell}} + o(\tfrac{1}{k}),$$

where we have used Lemma 3.2 to relate the local and global correlation, and then Lemma 3.3 that the average variance $\mathbf{Var}(\mathbf{E}_X)$ is at most $\frac{1}{k}$ from our program constraints. Here you can see that **we can choose ℓ as large as we want**, at the expense of more computation, **to get as small an error bound as we please**.

Using pseudodistributions. Of course, we do not necessarily have access to actual moments of a distribution over solutions to our optimization problem.

Note that the proof of Lemma 3.2 only used that the covariance matrix, Σ_X, is a positive semidefinite matrix; this is equally true of pseudocovariance matrices when the pseudodistribution has degree at least 2. So at least Lemma 3.2 works for pseudodistributions as well.

Similarly, the proof of Lemma 3.3 can be modified to work for pseudodistributions of degree at least $\ell + 2$. In order for this to make sense, we need a notion of a *conditional pseudodistribution.*

Definition 4.1. Let $\widetilde{\mathbf{E}}$ be a degree-d pseudodistribution over variables $X_1, \ldots, X_n, Z_1, \ldots, Z_m$[2] satisfying the axioms $X_i^2 = X_i$ for all $i \in [n]$. Then for any $i \in [n]$ where $\widetilde{\mathbf{E}}[X_i] > 0$, we can define the conditional pseudodistribution $\widetilde{\mathbf{E}}[\cdot \mid X_i = 1]$ of degree-$\ell - 1$ by setting:

$$\widetilde{\mathbf{E}}[f(X, Z) \mid X_i = 1] = \frac{\widetilde{\mathbf{E}}[f(X, Z) X_i]}{\widetilde{\mathbf{E}}[X_i]},$$

for any polynomial $f(X, Z)$ of degree at most $d - 1$. Similarly, if $\widetilde{\mathbf{E}}[X_i] < 1$ we can define the conditional pseudodistribution $\widetilde{\mathbf{E}}[\cdot \mid X_i = 0]$ of degree-$d - 1$ by setting

$$\widetilde{\mathbf{E}}[f(X, Z) \mid X_i = 0] = \frac{\widetilde{\mathbf{E}}[f(X, Z)(1 - X_i)]}{\widetilde{\mathbf{E}}[1 - X_i]}.$$

[2]The Z_j are meant to emphasize that the notion of conditional distributions does not require Boolean constraints for all of the variables.

One may check that if $\widetilde{\mathbf{E}}$ is a valid pseudoexpectation, then so is the conditional pseudoexpectation (albeit of one less degree).

Since in the proof of Lemma 3.3 we only used facts about subsets of $\ell + 2$ variables at a time, the following fact ensures that so long as we work with a pseudoexpectation $\widetilde{\mathbf{E}}$ of degree at least $\ell + 2$, the conclusion of Lemma 3.3 is valid for $\widetilde{\mathbf{E}}$ (and conditionings thereof).

Fact 4.2. *If $\widetilde{\mathbf{E}}$ is a degree-d pseudoexpectation over $X_1, \dots, X_n$ satisfying the axioms $X_i^2 = X_i$ for all $i \in [n]$, then for any $S \subset [n]$ with $|S| \leqslant d$, the pseudomoments $\{\widetilde{\mathbf{E}}[x^\alpha]\}_{\alpha \subset S}$ are consistent with some actual distribution over $\{0,1\}^S$.*

Proof. Any event on $\{0,1\}^S$ is expressible as a sum of indicators that the variables in X take some value $z \in \{0,1\}^S$. These indicators can be written as degree $|S|$ polynomials in X:

$$\mathbf{1}[X_S = z] = \mathbf{1}[X_i = z_i \ \forall i \in S] = \prod_{i \in S} (X_i z_i + (1 - X_i)(1 - z_i)).$$

The booleanity axioms ensure that $\widetilde{\mathbf{E}}[\mathbf{1}[X_S = z]] \geqslant 0$, and also that $\sum_{z \in \{0,1\}^S} \widetilde{\mathbf{E}}[\mathbf{1}[X_s] = z]] = 1$. Hence, the moments $\widetilde{\mathbf{E}}$ are consistent with the distribution on $\{0,1\}^S$ that sets $\mathbf{Pr}[X_S = z] = \widetilde{\mathbf{E}}[\mathbf{1}[X_S = z]]$. □

Since the inequalities used in the proof of Lemma 3.3 only use facts about marginal distributions on $\ell + 2$ variables, and since a degree-$\ell + 2$ pseudodistribution must have valid marginal distributions on subsets of $\leqslant \ell + 2$ variables by Fact 4.2, Lemma 3.3 applies to pseudoexpectations as well. Thus, we can use global correlation rounding as a technique for rounding SoS relaxations of variational problems, trading off computation time for higher degree $\ell + 2$ and thus better rounding error for quadratic objectives (scaling like $\ell^{-1/2}$ for, say, an optimization problem where the objective is defined over edges of an expanding graph).

5 Conclusion

Bibliographic remarks. Lemma 3.3 was developed independently in several parallel works. On the SoS side, Barak, Raghavendra and Steurer [BRS11] first suggested the technique of global correlation rounding as an approach to the Unique Games problem and other constraint satisfaction problems. The technique was later developed to deal with situations where global constraints are present [RT12], or higher-degree objective functions are considered [MR17], and some extensions exist even for non-boolean variables [BKS17]. Another interesting application in the context of statistics is mean-field approximations to Ising models [JKR19].

Independently, the same lemma was discovered by Montanari [Mon08] as a decomposition result for probability measures into simpler measures, in the context of the analysis of Belief Propagation. In fact, the idea of decreasing correlation by conditioning or "pinning" appeared much earlier in the literature of statistical physics [**?**]. The idea of decomposing measures into (interesting) mixtures of simpler measures is a powerful one, and conditioning on discrete-valued random variables is not the only way to do this. See, for example, [Eld20] if you are interested.

Contact. Comments are welcome at tselil@stanford.edu.

References

[BKS17] Boaz Barak, Pravesh K Kothari, and David Steurer. Quantum entanglement, sum of squares, and the log rank conjecture. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 975–988, 2017. 8

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *2011 IEEE 52nd annual symposium on foundations of computer science*, pages 472–481. IEEE, 2011. 8

[Eld20] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020. 4, 8

[JKR19] Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236, 2019. 1, 8

[Mon08] Andrea Montanari. Estimating random variables from random sparse observations. *European Transactions on Telecommunications*, 19(4):385–403, 2008. 8

[MR17] Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense csps. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. 8

[RT12] Prasad Raghavendra and Ning Tan. Approximating CSPs with global cardinality constraints using sdp hierarchies. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 373–387. SIAM, 2012. 8