Community detection in networks and Grothendieck's inequality*

11.4., 12.4., 18.4., 19.4.2024 (lectures)

David Steurer (lecture)

Contents

1 Introduction

We observe a graph G on n nodes with a latent community structure. Suppose that pairs of nodes are more likely to be connected by an edge in G if both are

*students – staff – index — versions: web / pdf / gitlab

members of the same community, and less likely to be connected by an edge if they are not in the same community. The goal is to approximately recover the latent community structure underlying the observed graph G.

2 Stochastic block model

We consider the following widely-studied statistical model for graphs with latent community structure, known as the *stochastic block model*. For simplicity, we restrict ourselves to the most basic case of two disjoint communities that affect the model symmetrically.

This model has two parameters: The *bias parameter* $\varepsilon \in [0, 1]$ determines the amount of information observed edges carry about community memberships. The *degree parameter* $d > 1$ determines the average number of edges observed per node.

We represent the latent community structure of the n nodes as a vector of labels[1] $x^\star \in \{\pm 1\}^n$.

We observe a graph $\boldsymbol{G}$ on n nodes where for every pair i, j of distinct nodes we decide independently whether to create an edge or not according to the following probabilities,

$$\mathbb{P}\{ij \in E(\boldsymbol{G})\} = \begin{cases} (1+\varepsilon) \cdot \frac{d}{n} & \text{if } x_i^\star = x_j^\star \\ (1-\varepsilon) \cdot \frac{d}{n} & \text{if } x_i^\star \neq x_j^\star \,. \end{cases}$$

Since the two vectors of labels $x^\star$ and $-x^\star$ give rise to the same distribution over graphs, we can hope to recover the vector of labels only up to sign. It turns out that the following goal is meaningful and allows us to recover the correct communities up to at most a 0.1 fraction of nodes:[2]

Goal. Given $\boldsymbol{G}$, compute an n-by-n matrix $\hat{\boldsymbol{X}}$ in polynomial time such that with high probability,[3]

$$\|\hat{\boldsymbol{X}} - X^\star\|_F^2 \leq 0.001 \cdot n^2 \,, \tag{1}$$

where $X^\star = x^\star (x^\star)^\mathsf{T}$.

Overview In sec. 4, we discuss and analyze an algorithm based on singular-value decompositions that achieves our goal (1) whenever $d \gg \varepsilon^{-2} \cdot \log n$. This algorithm follows the same strategy that we have used for matrix completion in previous lectures. Its analysis boils down to a spectral norm bound for certain

[1] Here, the set of nodes i with $x_i^\star = 1$ form one community and the set of nodes i with $x^\star = -1$ form another community.

[2] See this exercise

[3] The off-diagonal entries of the matrix $X^\star = x^\star x^{\star\mathsf{T}}$ indicate if two nodes are in the same community or not.

sparse random matrices, which we prove by (a variant) of the Matrix Bernstein inequality (see sec. 3).

In sec. 6 and sec. 7, we discuss and analyze a more sophisticated algorithm that achieves our goal (1) whenever $d \gg \varepsilon^{-2}$. This bound improves over the previous one by a logarithmic factor. Remarkably, this bound is tight up to a constant factor in the sense that for $d \ll \varepsilon^{-2}$, no algorithm can possibly achieve the goal (1). This algorithm is based on semidefinite programming and its analysis makes use of Grothendieck's inequality, which we prove in sec. 8.

3 Interlude: Matrix Bernstein inequality (variant)

The following version of the Matrix Bernstein inequality is an important ingredient of our analysis of the algorithm in sec. 4. The crucial feature of this version of the inequality is that we need to bound the variance of the sum of matrices as opposed to the variances of the individual summands. For non-identically distributed summands, this difference can be significant.

Theorem. Let $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_m$ be independent n-by-n random matrices. Suppose these random matrices are centered $\mathbb{E}\,\boldsymbol{Z}_i = 0$ and symmetric $\boldsymbol{Z}_i = \boldsymbol{Z}_i^{\mathsf{T}}$. Choose $\sigma, b > 0$ such that $\|\boldsymbol{Z}_i\| \le b$ for all $i \in [m]$ and

$$\left\| \sum_{i=1}^{m} \mathbb{E}\, \boldsymbol{Z}_i \boldsymbol{Z}_i^{\mathsf{T}} \right\| \le \sigma^2 .$$

Then, with probability at least $1 - n^{-100}$, it holds

$$\left\| \sum_{i=1}^{m} \boldsymbol{Z}_i \right\| \lesssim \sigma \cdot \sqrt{\log n} + b \cdot \log n . \tag{2}$$

4 Recovery guarantees via spectral norm bounds

In this section, we consider an algorithm that follows the strategy we have used for matrix completion:

- directly compute an n-by-n random matrx $\boldsymbol{Y}$ with the property $\mathbb{E}\,\boldsymbol{Y} = X^\star$,
- output a best rank-1 approximation $\hat{\boldsymbol{X}}$ of $\boldsymbol{Y}$.

Since the diagonal entries of $X^\star$ are all ones, we can choose the diagonal entries of $\boldsymbol{Y}$ to be 1.

We choose the off-diagonal entries $\boldsymbol{Y}_{ij}$ of $\boldsymbol{Y}$ based on whether the nodes $i \neq j$ are adjacent in $\boldsymbol{G}$ or not:

$$\boldsymbol{Y}_{ij} = \begin{cases} \alpha_1 := \frac{1}{\varepsilon} \cdot (\frac{n}{d} - 1) & \text{if } ij \in E(\boldsymbol{G})\,, \\ \alpha_0 := -\frac{1}{\varepsilon} & \text{if } ij \notin E(\boldsymbol{G})\,. \end{cases} \tag{3}$$

Since $\mathbb{E}\,\boldsymbol{Y}_{ij} = \mathbb{P}\,\{ij \in E(\boldsymbol{G})\} \cdot (\alpha_1 - \alpha_0) + \alpha_0$ and $\mathbb{P}\,\{ij \in E(\boldsymbol{G})\} = (1 + \varepsilon \cdot x_i^\star \cdot x_j^\star) \cdot \frac{d}{n}$, it follows that for our choice of α_0, α_1,

$$\begin{aligned} \mathbb{E}\,\boldsymbol{Y}_{ij} &= \tfrac{d}{n} \cdot (\alpha_1 - \alpha_0) + \alpha_0 + \varepsilon \cdot x_i^\star \cdot x_j^\star \cdot (\alpha_1 - \alpha_0) \cdot \tfrac{d}{n} \\ &= x_i^\star \cdot x_j^\star\,. \end{aligned}$$

Theorem. Suppose $d \gg \varepsilon^{-2} \log n$. Then, with high probability, the matrix $\hat{\boldsymbol{X}} = \arg\min\{\|X - \boldsymbol{Y}\|_F^2 \mid \operatorname{rank}(X) = 1\}$ satisfies the recovery guarantee (1).

Proof.

Suppose $n \geq 10$, $0 < \varepsilon < 1$ and $d \geq 10 \cdot \varepsilon^{-2} \cdot \log n$. As in the analyses of the estimators for the single-spike model and for matrix completion, the error for the above estimator satisfies

$$\|\hat{\boldsymbol{X}} - X^\star\|_F^2 \leq 8 \cdot \|\boldsymbol{Y} - X^\star\|^2\,.$$

Hence, in order to prove the above theorem, it is enough to show that with high probability, $\|\boldsymbol{Y} - X^\star\| \leq 0.001 \cdot n$. Let $\boldsymbol{Z}_{ij} := (\boldsymbol{Y}_{ij} - X_{ij}^\star) \cdot (e_i \cdot e_j^{\mathsf{T}} + e_j \cdot e_i^{\mathsf{T}})$. By construction,

$$\boldsymbol{Y} - X^\star = \sum_{i<j} \boldsymbol{Z}_{ij}\,.$$

Furthermore, the random matrices $\{\boldsymbol{Z}_{ij}\}$ are independently distributed. The spectral norms of these matrices are determinisitcally upper bounded by $b := \frac{n}{\varepsilon d}$,

$$\begin{aligned} \|\boldsymbol{Z}_{ij}\| &= |\boldsymbol{Y}_{ij} - X_{ij}^\star| \\ &\leq 1 + \tfrac{1}{\varepsilon} \cdot (\tfrac{n}{d} - 1) \leq b\,. \end{aligned}$$

The first step follows from the fact that $\boldsymbol{Z}_{ij}$ has two singular values equal to $|\boldsymbol{Y}_{ij} - X_{ij}^\star|$ and all other singular values equal to 0. Moreover, we can choose $\sigma^2 := \frac{2n^2}{\varepsilon^2 d}$ as a variance proxy for $\{\boldsymbol{Z}_{ij}\}$,

$$\begin{aligned} \mathbb{E}\,\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^{\mathsf{T}} &= \mathbb{E}(\boldsymbol{Y}_{ij} - X_{ij}^\star)^2 \cdot (e_i e_i^{\mathsf{T}} + e_j e_j^{\mathsf{T}}) \\ &\leq \tfrac{2n}{\varepsilon^2 d} \cdot (e_i e_i^{\mathsf{T}} + e_j e_j^{\mathsf{T}})\,. \end{aligned}$$

Here, we use

$$\begin{aligned} \mathbb{E}(\boldsymbol{Y}_{ij} - X_{ij}^\star)^2 &\leq \mathbb{E}\,\boldsymbol{Y}_{ij}^2 \\ &\leq (1 + \varepsilon)\tfrac{d}{n} \cdot \alpha_1^2 + \alpha_0^2 \\ &\leq (1 + \varepsilon)\tfrac{n}{\varepsilon^2 d} \leq \tfrac{2n}{\varepsilon^2 d}\,. \end{aligned}$$

It follows that

$$\sum_{i<j} \mathbb{E}\, \boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\mathsf{T} \preceq \tfrac{2n}{\varepsilon^2 d} \cdot \sum_{i<j} (e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T}) \\ \preceq \tfrac{2n^2}{\varepsilon^2 d} \cdot I_n = \sigma^2 \cdot I_n \,.$$

Thus, by the Matrix Bernstein inequality (2), it holds with probability at least $1 - n^{-100}$,

$$\|\boldsymbol{Y} - X^\star\| \lesssim \sigma \cdot \sqrt{\log n} + b \cdot \log n \\ = \sqrt{\tfrac{2n^2 \log n}{\varepsilon^2 d}} + \tfrac{n \log n}{\varepsilon d} \lesssim \sqrt{\frac{n^2 \log n}{\varepsilon^2 d}} \,.$$

Here, we use that the first term dominates up to a constant factor (as $d \geq \log n$).

We conclude that with probability at least $1 - n^{-100}$, the estimation error satisfies the upper bound

$$\|\hat{\boldsymbol{X}} - X^\star\|_F^2 \lesssim \frac{n^2 \log n}{\varepsilon^2 d} \,.$$

In particular, this error bound is smaller than the desired bound $0.001 \cdot n^2$ as long as $d \geq O(\varepsilon^{-2} \log n)$.

5 Interlude: Bernstein inequality (for scalars)

The analysis of the next algorithm we discuss relies on the following version of the Bernstein inequalilty for scalar valued random variables.

Theorem. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ be independent real-valued random variables. Suppose $\mathbb{E}\, \boldsymbol{Z}_1 = \cdots = \mathbb{E}\, \boldsymbol{Z}_N = 0$. Let $\sigma, b > 0$ satisfy $\mathbb{E}\, \boldsymbol{Z}_1^2 + \cdots + \mathbb{E}\, \boldsymbol{Z}_N^2 \leq \sigma^2$ and $|\boldsymbol{Z}_1|, \ldots, |\boldsymbol{Z}_N| \leq b$. Then, for all $\delta > 0$, the following event has probability at least $1 - \delta$,

$$|\boldsymbol{Z}_1 + \cdots + \boldsymbol{Z}_N| \lesssim \sigma \cdot \sqrt{\log(1/\delta)} + b \cdot \log(1/\delta) \,.$$

6 Recovery guarantees via cut norm bounds

In this section, we improve the guarantees of the algorithm in sec. 4 by a logarithmic factor and develop an estimator that approximately recovers the underlying communities with high probability (in the sense of (1)) whenever $d \gg \varepsilon^{-2}$. (As discussed before, this bound on d is best possible up to constant factors.)

Recall that the key ingredient for our previous analysis is a spectral norm bound on the "noise matrix" $\boldsymbol{Y} - X^*$. Indeed, this logarithmic factor is necessary for this bound (and not merely an artifact of our particular analysis). Moreover, the algorithm in sec. 4 actually fails to achieve the goal (1) when $d \ll \varepsilon^{-2} \cdot \log n$.

In this light, a natural idea for improving our previous estimator is to choose a norm that is better suited for the problem at hand than the spectral norm. To this end, we consider the following family of estimators parameterized by a set of matrices $C \subseteq \mathbb{R}^{n \times n}$,

$$\hat{\boldsymbol{X}} := \arg\min \left\{ \|X - \boldsymbol{Y}\|_F^2 \mid X \in C \right\} . \tag{4}$$

We recover the estimator used in sec. 4 by choosing C to be the set of all rank-1 matrices. (We choose $\boldsymbol{Y}$ as before (3).)

In order to obtain lower error bounds, we should choose C to be as small as possible while still including $X^\star$. In this sense, a natural choice for C would be the set of all possible choices for $X^\star$ (symmetric rank-1 matrices with all entries 1 or -1). For important technical reasons that we discuss later, we choose C to also include non-symmetric matrices,

$$C := \left\{ uv^\mathsf{T} \mid u, v \in \{\pm 1\}^n \right\} . \tag{5}$$

This set C is a strict subset of the set of all rank-1 matrices. We can view the matrices in C as (combinatorial) 2-by-2 block matrices with all ones in the diagonal blocks and all minus ones in the off-diagonal blocks.

Theorem. As before, let $\boldsymbol{G}$ be an n-vertex graph drawn from the stochastic block model distribution with degree parameter d and bias parameter ε. Then, the estimator in (4) for the choice of C in (5) satisfies the following error bound with high probability,

$$\|\hat{\boldsymbol{X}} - X^\star\|_F^2 \lesssim \frac{n^2}{\varepsilon\sqrt{d}} .$$

We emphasize that the estimator presented in (the proof of) the above theorem is a-priori not computationally efficient. We discuss in the next section sec. 7 how to make it computationally efficient (polynomial running time).

Proof. Since $X^\star$ is a feasible solution for the optimization problem that $\hat{\boldsymbol{X}}$ optimizes, we have the following (familiar) inequality,

$$\|\hat{\boldsymbol{X}} - X^\star\|_F^2 \le 2\langle \hat{\boldsymbol{X}} - X^\star, \boldsymbol{Y} - X^\star \rangle .$$

To upper bound this inner product, we consider the following matrix norm, called *cut norm*,

$$\|M\|_{cut} := \max_{W \in C} \langle W, M \rangle = \max_{u,v \in \{\pm 1\}^n} \langle u, Mv \rangle .$$

Since $\hat{\boldsymbol{X}}, X^\star \in C$, we have $\langle \hat{\boldsymbol{X}} - X^\star, \boldsymbol{Y} - X^\star \rangle \le 2\|\boldsymbol{Y} - X^\star\|_{cut}$. Plugging this inner-product bound into the above inequality, we get

$$\|\hat{\boldsymbol{X}} - X^\star\|_F^2 \le 4\|\boldsymbol{Y} - X^\star\|_{cut} .$$

It remains to show that with high probability, $\|\boldsymbol{Y} - X^\star\|_{cut} \lesssim \frac{n^2}{\varepsilon\sqrt{d}}$. To this end, consider an arbitrary matrix $W \in \mathcal{C}$. Let $\boldsymbol{Z}_{ij} := (\boldsymbol{Y}_{ij} - X^\star_{ij}) \cdot W_{ij}$. Then,

$$\langle W, \boldsymbol{Y} - X^\star \rangle = 2 \sum_{i<j} \boldsymbol{Z}_{ij} \,.$$

Since the random variables $\{\boldsymbol{Z}_{ij} \mid i < j\}$ are independent and have expectation $\mathbb{E}\,\boldsymbol{Z}_{ij} = 0$, we can apply Bernstein inequality to bound the tail probability for their sum. Each variable $\boldsymbol{Z}_{ij}$ satisfies the following uniform upper bound,

$$|\boldsymbol{Z}_{ij}| \le \tfrac{1}{\varepsilon} \cdot (\tfrac{n}{d} - 1) + 1 \le \tfrac{n}{\varepsilon d} \,.$$

At the same time, we can bound the variance of each variable $\boldsymbol{Z}_{ij}$,

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{Z}_{ij}^2 &\le \mathbb{E}\,\boldsymbol{Y}_{ij}^2 \\
&= (1 + \varepsilon X^\star_{ij}) \tfrac{d}{n} \cdot \left(\tfrac{1}{\varepsilon^2} \cdot (\tfrac{n}{d} - 1)^2 - \tfrac{1}{\varepsilon^2} \right) + \tfrac{1}{\varepsilon^2} \\
&\le (1 + \varepsilon) \tfrac{n}{\varepsilon^2 d} + \tfrac{1}{\varepsilon^2} \le \tfrac{3n}{\varepsilon^2 d} \,.
\end{aligned}$$

Choose $\delta := 2^{-3n}$, $b := \frac{n}{\varepsilon d}$ and $\sigma > 0$ such that $\sigma^2 = n^2 \cdot \frac{3n}{\varepsilon^2 d}$. Then, by the version of Bernstein inequality in sec. 5, with probability at least $1 - \delta$, it holds

$$\begin{aligned}
\langle W, \boldsymbol{Y} - X^\star \rangle &\lesssim \sigma \cdot \sqrt{\log_2(1/\delta)} + b \cdot \log_2(1/\delta) \\
&= \tfrac{\sqrt{3} \cdot n^{3/2}}{\varepsilon\sqrt{d}} \cdot \sqrt{3n} + \tfrac{n}{\varepsilon d} \cdot 3n \lesssim \tfrac{n^2}{\varepsilon\sqrt{d}} \,.
\end{aligned}$$

Hence, we can choose $t \lesssim \frac{n^2}{\varepsilon^2 d}$ such that $\mathbb{P}\{\langle W, M \rangle > t\} \le \delta$ for all $W \in \mathcal{C}$. By the union bound,

$$\begin{aligned}
\mathbb{P}\left\{\|\boldsymbol{Y} - X^\star\|_{cut} > t\right\} &\le \sum_{W \in \mathcal{C}} \mathbb{P}\left\{\langle W, \boldsymbol{Y} - X^\star \rangle > t\right\} \\
&\le 2^{2n} \cdot 2^{-3n} = 2^{-n} \,.
\end{aligned}$$

7 Polynomial-time algorithm via Grothendieck's inequality

Recall the optimization problem underlying the estimator in the previous section sec. 6:

> Given a matrix Y, the goal is to find the closest rank-1 matrix X with all entries 1 or -1.

This optimization turns out to be NP-hard (for worst-case Y). In this section, we show how to obtain a polynomial-time estimator with the same error bound (up to a constant factor) based on approximation algorithm for the cut norm.[4]

Concretely, we consider the following set of symmetric matrices,

$$\mathcal{E} := \left\{ Z \;\middle|\; Z \succeq 0, \operatorname{diag}(Z) = I_{2n} \right\} .$$

This set of matrices is convex and has a polynomial-time separation oracle.[5] We consider the following matrix norm, which we call *Grothendieck norm*, defined in terms of the set $\mathcal{E}$,

$$\|M\|_G := \max \left\{ \left\langle Z, \begin{pmatrix} 0 & M \\ 0 & 0 \end{pmatrix} \right\rangle \;\middle|\; Z \in \mathcal{E} \right\} .$$

The above inner product is equal to $\langle Z_{1,2}, M \rangle$, where $Z_{1,2}$ is the upper right n-by-n block of the matrix Z. Since the set $\mathcal{E}$ has a polynomial-time separation oracle, there is a polynomial-time algorithm for computing the Grothendieck norm using standard convex optimization techniques (ellipsoid method).

The following theorem shows that the Grothendieck norm is within a constant factor of the cut norm.

Theorem. (Grothendieck's inequality) For all matrices M,

$$\|M\|_{cut} \le \|M\|_G \le k_G \cdot \|M\|_{cut} ,$$

where $k_G \in [1, 2c]$ is an absolute constant.

For the following choice of C, we obtain a polynomial-time computable estimator in the family (4),

$$C := \{ Z_{1,2} \mid Z \in \mathcal{E} \} . \tag{6}$$

This choice of C turns out to be a super set of the choice in sec. 6.[6] Note that this directly implies that for any matrix M it holds that $\|M\|_{cut} \le \|M\|_G$. Nevertheless, we will be able to show the same error bound as in sec. 6 (up to constant factors). The reason is that we will be able to bound the error in terms of Grothendieck norm of the "noise matrix". By Grothendieck's inequality, this norm is within a constant factor of the cut norm, which we bounded already in sec. 6.

Theorem. As before, let G be an n-vertex graph drawn from the stochastic block model distribution with degree parameter d and bias parameter ε. Then

[4] It is interesting that we need here an approximation guarantee for the norm that appeared in our previous analysis as opposed to an approximation guarantee for the optimization problem underlying the estimator.

[5] Given a symmetric matrix $Z \notin \mathcal{E}$, we can find a hyperplane separating it from $\mathcal{E}$ as follows: If $\operatorname{diag} Z \neq I_{2n}$, then one of the matrices $e_i e_i^T$ provides a separating hyperplane. Otherwise, we have $Z \not\succeq 0$. In this case, Z has an eigenvector v with negative eigenvalue and the matrix vv^T provides a separating hyperplane.

[6] This is left as an exercise.

for the choice of C in (6), the estimator in (4) satisfies the following error bound with high probability,

$$\|\hat{X} - X^\star\|_F^2 \lesssim \frac{n^2}{\varepsilon\sqrt{d}}\,.$$

Proof. As usual, we bound the norm of the error in terms of the inner product of the error matrix and the noise matrix,

$$\begin{aligned}\|\hat{X} - X^\star\|_F^2 &\le 2|\langle \hat{X} - X^\star, Y - X^\star\rangle| \\ &\le 4\|Y - X^\star\|_G \\ &\le 4k_G \cdot \|Y - X^\star\|_{cut}\,.\end{aligned}$$

The second step uses that $\hat{X}, X^\star \in \mathcal{E}$. The third step uses Grothendieck's inequality.

Using the bound on the cut norm in sec. 6, we conclude that our estimator satisfies with high probability,

$$\|\hat{X} - X^\star\|_F^2 \lesssim \tfrac{n^2}{\varepsilon\sqrt{d}}\,.$$

8 Proof of Grothendieck's inequality

We will now prove Grothendieck's inequality via a technique that is know as randomized rounding. Let M be any matrix. Recall that $\|M\|_{cut} \le \|M\|_G$ because for the Grothendieck norm we are optimizing the same objective over a larger set. It remains to prove that

$$\|M\|_G \le k_G \cdot \|M\|_{cut}$$

We first consider the following alternate forms for the cut norm and the Grothendieck norm. Recall that we defined the cut norm as $\|M\|_{cut} = \max_{x,y\in\{\pm1\}^n}\langle x, My\rangle$. We claim that this is equivalent to

$$\max_{\|x\|_\infty\le 1,\|y\|_\infty\le 1}\langle x, My\rangle = \max_{x,y\in\mathbb{R}^n\setminus\{0\}}\frac{\langle x, My\rangle}{\|x\|_\infty\|y\|_\infty}$$

This follows since $\langle x, My\rangle$ is linear in every coordinate of x and y (individually), and linear functions attain their maximum at the extreme points. The second equality follows by considering $\tilde{x} = \frac{x}{\|x\|_\infty}$ and $\tilde{y} = \frac{y}{\|y\|_\infty}$. We also claim that the Grothendieck norm has the following equivalent formulation:

$$\|M\|_G = \max_{u_1,\dots,u_n,v_1,\dots,v_n\in B^{2n}} M_{i,j}\langle u_i, v_j\rangle\,,$$

where $B^{2n} := \{w \in \mathbb{R}^{2n} \mid \|w\|_2 \le 1\}$. See this exercise for a proof. In the same exercise you will also show that we can replace the unit ball by the unit sphere, i.e., with the set $\{w \in \mathbb{R}^{2n} \mid \|w\|_2 = 1\}$

8.1 Randomized Rounding

For a matrix M, let $u_1, \dots, u_n, v_1, \dots v_n \in \mathbb{R}^{2n}$ be unit vectors such that

$$\|M\|_G = \sum_{i,j} M_{i,j} \langle u_i, v_j \rangle .$$

We use a technique refered to randomized rounding: Draw a random vector $\boldsymbol{g} \sim N(0, I_{2n})$ and define

$$\boldsymbol{x}_i := \langle \boldsymbol{g}, u_i \rangle , \boldsymbol{y}_j := \langle \boldsymbol{g}, v_j \rangle .$$

Note that $\boldsymbol{g}$ is the same in both definitions above. Also notice that $\boldsymbol{x}_i \sim N(0,1)$ and $\boldsymbol{y}_j \sim N(0,1)$ for all $i, j \in [n]$. Furthermore, observe that

$$\begin{aligned} \mathbb{E}\langle \boldsymbol{x}, M\boldsymbol{y} \rangle &= \mathbb{E}\left[\sum_{i,j} M_{i,j} \boldsymbol{x}_i \boldsymbol{y}_j \right] = \sum_{i,j} M_{i,j} \langle u_i, \mathbb{E}\left[\boldsymbol{g}\boldsymbol{g}^T \right] v_j \rangle \\ &= \sum_{i,j} M_{i,j} \langle u_i, v_j \rangle = \|M\|_G . \end{aligned}$$

That is, we expressed $\|M\|_G$ as the expectation of $\langle \boldsymbol{x}, M\boldsymbol{y} \rangle$. This enables us to analyze the former quantity via properties of the standard Gaussian distribution. In particular, it is now enough to show that $\mathbb{E}[\langle \boldsymbol{x}, M\boldsymbol{y} \rangle] \leq k_G \|M\|_{cut}$. Indeed, we will achieve this goal by showing that

$$\|M\|_G = \mathbb{E}\langle \boldsymbol{x}, M\boldsymbol{y} \rangle \leq a\|M\|_G + c\|M\|_{cut}$$

for absolute constants $a < 1$ and $c > 0$. We can then deduce Grothendieck's inequality by rearranging. In this proof, we will not try to obtain the best-possible value for k_G.

The key idea behind the proof will be to truncate the large entries of $\boldsymbol{x}, \boldsymbol{y}$ and bound the resulting error. To this end, we define for $i, j \in [n]$,

$$\begin{aligned} \boldsymbol{x}_i^{\leq \tau} &:= \mathrm{clamp}_\tau(\boldsymbol{x}_i) , \boldsymbol{x}_i^{> \tau} := \boldsymbol{x}_i - \boldsymbol{x}_i^{\leq \tau} , \\ \boldsymbol{y}_j^{\leq \tau} &:= \mathrm{clamp}_\tau(\boldsymbol{y}_j) , \boldsymbol{y}_i^{> \tau} := \boldsymbol{y}_i - \boldsymbol{y}_i^{\leq \tau} , \end{aligned}$$

where

$$\mathrm{clamp}_\tau(z) := \begin{cases} z & \text{if } |z| \leq \tau , \\ \tau \cdot \mathrm{sign}(z) & \text{otherwise.} \end{cases}$$

Therefore we can express

$$\begin{aligned} \|M\|_G &= \mathbb{E}\langle \boldsymbol{x}, M\boldsymbol{y} \rangle \\ &= \mathbb{E}\left[\langle \boldsymbol{x}^{\leq \tau}, M\boldsymbol{y}^{\leq \tau} \rangle + \langle \boldsymbol{x}^{> \tau}, M\boldsymbol{y}^{\leq \tau} \rangle + \langle \boldsymbol{x}^{\leq \tau}, M\boldsymbol{y}^{> \tau} \rangle + \langle \boldsymbol{x}^{> \tau}, M\boldsymbol{y}^{> \tau} \rangle \right] . \end{aligned}$$

Now observe that $\langle \boldsymbol{x}^{\leq\tau}, M\boldsymbol{y}^{\leq\tau}\rangle \leq \tau^2\|M\|_{cut}$ since the entries of $\boldsymbol{x}^{\leq\tau}$ and $\boldsymbol{y}^{\leq\tau}$ are at most τ. Note that we are already making progress, since we have bounded $\|M\|_G$ by some constant times $\|M\|_{cut}$ plus some additional terms. Indeed as stated above, we will now argue that these additional terms together contribute at most a small fraction of $\|M\|_G$ which will prove the theorem. We will now focus on the remaining terms.

Define $\boldsymbol{R} := \langle \boldsymbol{x}^{>\tau}, M\boldsymbol{y}^{\leq\tau}\rangle + \langle \boldsymbol{x}^{\leq\tau}, M\boldsymbol{y}^{>\tau}\rangle + \langle \boldsymbol{x}^{>\tau}, M\boldsymbol{y}^{>\tau}\rangle$

Claim: There exists $W \in \mathbb{R}^{2n\times 2n}$ such that $W \succeq 0$, all diagonal entries of W are strictly positive, and

$$\mathbb{E}[\boldsymbol{R}] = \left\langle W, \begin{pmatrix} 0 & M \\ 0 & 0 \end{pmatrix} \right\rangle \leq W_{11}\|M\|_G$$

where W_{11} refers to the first diagonal element of W. Notice once again that we are making progress. As long as we can show that $W_{11} < 1$, we are done.

Let $\lambda \geq 1$ to be decided later and define:

$$\boldsymbol{w_1} := \begin{pmatrix} \lambda \boldsymbol{x}^{>\tau} \\ \lambda^{-1}\boldsymbol{y}^{\leq\tau} \end{pmatrix} \quad \boldsymbol{w_2} := \begin{pmatrix} \lambda^{-1}\boldsymbol{x}^{\leq\tau} \\ \lambda \boldsymbol{y}^{>\tau} \end{pmatrix} \quad \boldsymbol{w_3} := \begin{pmatrix} \boldsymbol{x}^{>\tau} \\ \boldsymbol{y}^{>\tau} \end{pmatrix}$$

In this exercise you will verify that the following matrix

$$W := \mathbb{E}\left[\boldsymbol{w_1}\boldsymbol{w_1}^T + \boldsymbol{w_2}\boldsymbol{w_2}^T + \boldsymbol{w_3}\boldsymbol{w_3}^T\right]$$

satisfies the claim (for every choice of λ). It remains to show that $W_{11} < 1$. For all $i \in [n]$ it holds that

$$W_{ii} = \lambda^{-2}\mathbb{E}\left[\mathrm{clamp}_\tau(\boldsymbol{x}_i^2)\right] + (\lambda^2+1)\mathbb{E}\left[(|x_i|-\tau)_+^2\right]$$

where $(x)_+ := \max\{x, 0\}$. Note that we can write the above without loss of generality for just $\boldsymbol{x}_i$ since $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ have the same distributions. We will show that $\mathbb{E}\left[(|x_i|-\tau)_+^2\right] \leq e^{-\tau^2/2}$. In the exercises we will show that there exists a choice of λ such that the remaining terms are upper bounded by $e^{-\tau^2/4}$. Together this implies that $W_{11} \leq 5e^{-\tau^2/4}$ When choosing τ to be a large enough constant, this is smaller than 1. We write

$$\mathbb{E}\left[(|x_i|-\tau)_+^2\right] = 2\int_\tau^\infty (x-\tau)^2\gamma(x)dx\,,$$

where $\gamma(x)$ is the density function of the one-dimensional gaussian distribution.

Now we have,

$$\begin{aligned}\int_\tau^\infty (x-\tau)^2\gamma(x)\,dx &= \int_\tau^\infty (x-\tau)^2\gamma(x-\tau)\frac{\gamma(x)}{\gamma(x-\tau)}\,dx \\ &\leq \max_{y\geq\tau}\frac{\gamma(y)}{\gamma(y-\tau)}\int_\tau^\infty (x-\tau)^2\gamma(x-\tau)\,dx\,.\end{aligned}$$

Since $\int_\tau^\infty (x-\tau)^2 \gamma(x-\tau)\, dx = \frac{1}{2}$, it follows that

$$2\int_\tau^\infty (x-\tau)^2 \gamma(x)\, dx \le \max_{x\ge\tau} \frac{\gamma(x)}{\gamma(x-\tau)} = \max_{x\ge\tau} \frac{e^{-x^2/2}}{e^{-(x-\tau)^2/2}} \le e^{-\tau^2/2}\,.$$

9 Comments

links to relevant handwritten notes

- first lecture
- second lecture
- third lecture
- fourth lecture