Supplementary material - Sample complexity bounds for localized sketching

1 Proof of Theorem 1

The fundamental property of a distribution of matrices $\mathcal{D}$ that enables any $\mathbf{S} \sim \mathcal{D}$ to satisfy (8, main paper) is the subspace embedding moment property, defined in [Avron et al., 2016]:

$$\mathop{\mathrm{E}}_{\mathbf{S}\sim\mathcal{D}} \left\|(\mathbf{SU})^T(\mathbf{SU}) - \mathbf{I}\right\|^l \leq \epsilon^l \delta, \tag{1}$$

for some $l \geq 2$, where ϵ and δ are tolerance parameters that determine the sample complexity and $\mathbf{U}$ is any orthobasis for the span of the columns of $\mathbf{W}$ and $\mathbf{Y}$. Thus, our main goal is to prove the subspace embedding moment property holds for block diagonal sketching matrices.

Our methods differ from the common ϵ-net argument, since using union bound for block diagonal matrices results in a suboptimal sample complexity. The main tools we use are the estimates for the suprema of chaos processes found in [Krahmer et al., 2014] and an entropy estimate from the study of restricted isometry properties of block diagonal matrices computed in [Eftekhari et al., 2015]. We first establish tail bounds on the spectral norm of the matrix

$$\mathbf{\Delta} = (\mathbf{S}_D\mathbf{U})^T(\mathbf{S}_D\mathbf{U}) - \mathbf{I}, \tag{2}$$

where $\mathbf{U}$ is an orthobasis for a subspace of dimension d and then bound its moments to establish the subspace embedding moment property.

1.1 Suprema of chaos processes

We briefly state here the main result from [Krahmer et al., 2014] that provides a uniform bound on the deviation of a Gaussian quadratic form from its expectation. Obtaining a tail bound on the spectral norm of $\mathbf{\Delta}$ is just a particular application of this general framework.

For a given set of matrices $\mathcal{P}$, we define the spectral radius $d_2(\mathcal{P})$, the Frobenius norm radius $d_F(\mathcal{P})$, and the Talagrand functional $\gamma_2(\mathcal{P}, \|\cdot\|_2)$ as

$$\begin{aligned} d_2(\mathcal{P}) &= \sup_{\mathbf{P}\in\mathcal{P}} \|\mathbf{P}\|, \\ d_F(\mathcal{P}) &= \sup_{\mathbf{P}\in\mathcal{P}} \|\mathbf{P}\|_F, \\ \gamma_2(\mathcal{P}, \|\cdot\|_2) &= \int_0^{d_2(\mathcal{P})} \sqrt{\log N(\mathcal{P}, \|\|_2, u)}du, \end{aligned}$$

where $N(\mathcal{P}, \|\|_2, u)$ denotes the covering number of the set $\mathcal{P}$ with respect to balls of radius u in the spectral norm. The main result of [Krahmer et al., 2014] then is the following theorem.

Theorem 1 *[Theorem 3.1, [Krahmer et al., 2014]] Let $\mathcal{P}$ be a set of matrices and let ϕ be a vector of i.i.d. standard normal entries. Then for $t \geq 0$,*

$$\mathbb{P}\left(\sup_{\mathbf{P}\in\mathcal{P}} |\|\mathbf{P}\phi\|^2 - \mathrm{E}\,\|\mathbf{P}\phi\|^2| > c_1 E + t\right) \leq 2\mathrm{e}^{-c_2 \min\{\frac{t^2}{V^2}, \frac{t}{U}\}} \tag{3}$$

where

$$\begin{aligned} E &= \gamma_2(\mathcal{P})[\gamma_2(\mathcal{P}) + d_F(\mathcal{P})] + d_2(\mathcal{P})d_F(\mathcal{P}), \\ V &= d_2(\mathcal{P})[\gamma_2(\mathcal{P}) + d_F(\mathcal{P})], \\ U &= d_2^2(\mathcal{P}). \end{aligned}$$

A similar approach of using the results from [Krahmer et al., 2014] to analyze block diagonal random matrices was first used in [Eftekhari et al., 2015] in the context of compressed sensing. However, we target a different set of problems that result in different theoretical considerations and proof techniques.

1.2 Tail bound on the spectral norm of the matrix $\mathbf{\Delta}$

We first express $\|\mathbf{\Delta}\|$ as

$$\|\mathbf{\Delta}\| = \sup_{\substack{\mathbf{z}\in\mathbb{R}^d \\ \|\mathbf{z}\|=1}} \left|\mathbf{z}^T(\mathbf{S}_D\mathbf{U})^T(\mathbf{S}_D\mathbf{U})\mathbf{z} - 1\right| \tag{4}$$

$$= \sup_{\substack{\mathbf{z}\in\mathbb{R}^d \\ \|\mathbf{z}\|=1}} \left|\|\mathbf{S}_D\mathbf{U}\mathbf{z}\|^2 - \mathrm{E}\,\|\mathbf{S}_D\mathbf{U}\mathbf{z}\|^2\right|. \tag{5}$$

For the matrices $\mathbf{S}_j$, let $\mathrm{vec}(\mathbf{S}_j)$ denote their vectorized versions, obtained by stacking the columns one below the other. Let $\mathbf{S}_v$ =

$[\mathrm{vec}(\mathbf{S}_1)^T\ \mathrm{vec}(\mathbf{S}_2)^T\ \cdots \mathrm{vec}(\mathbf{S}_J)^T]^T$ be the vector containing all of the $\mathrm{vec}(\mathbf{S}_j)$'s. Note that $\mathbf{S}_v$ is a vector with entries drawn from $\mathcal{N}(0,1)$. We can then express (4) as

$$\|\mathbf{\Delta}\| = \sup_{\mathbf{P}_z \in \mathcal{P}} \left| \|\mathbf{P}_z \mathbf{S}_v\|^2 - \mathrm{E}\,\|\mathbf{P}_z \mathbf{S}_v\|^2 \right|$$

where $\mathcal{P}$ is defined

$$\mathcal{P} = \left\{ \mathbf{P}_z = \begin{bmatrix} \mathbf{P}_1(z) & 0 & \cdots & 0 \\ 0 & \mathbf{P}_2(z) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{P}_J(z) \end{bmatrix} \right\},$$

$$\mathbf{P}_j(z) = \frac{1}{\sqrt{M_j}} \begin{bmatrix} (U_1 z)^T & 0 & \cdots & 0 \\ 0 & (U_1 z)^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (U_1 z)^T \end{bmatrix}$$

where $z \in \mathbb{R}^d$ and $\|z\| = 1$. Observe that $\|\mathbf{\Delta}\|$ is then the supremum of the deviation of a Gaussian quadratic form from its expectation, taken over the set $\mathcal{P}$.

We can then compute the corresponding quantities $d_2(\mathcal{P})$, $d_F(\mathcal{P})$ and $\gamma_2(\mathcal{P}, \|\cdot\|_2)$ as follows.

The spectral radius $d_2(\mathcal{P})$ is defined as

$$\begin{aligned} \sup_{\mathbf{P}_z \in \mathcal{P}} \|\mathbf{P}_z\| &= \max_{j, \|z\|_2 = 1} \frac{\|U_j z\|}{\sqrt{M_j}} \\ &\le \min\left(\frac{\sqrt{N}\,\|U_j\|_\infty \|z\|_1}{\sqrt{M_j}}, \frac{\|U_j\|\,\|z\|_2}{\sqrt{M_j}} \right) \\ &\le \min\left(\frac{\sqrt{N}\,\|U_j\|_\infty \|z\|_1}{\sqrt{M_j}}, \frac{\|U_j\|\,\|z\|_1}{\sqrt{M_j}} \right) \\ &\le \|z\|_1 / \sqrt{M_0} \le \frac{d_1}{\sqrt{M_0}} \end{aligned}$$

where the fourth line follows from the definition of M_j.

The radius in the Frobenius norm $d_F(\mathcal{P})$ is defined as

$$\sup_{\mathbf{P}_z \in \mathcal{P}} \|\mathbf{P}_z\|_F = \sum_j \|U_j z\|^2 = 1.$$

The upper bound for $\gamma_2(\mathcal{P}, \|\cdot\|)$ can be obtained from the Equation (34) in Eftekhari et al., 2015).. In their derivation, they consider a full orthobasis and the set of d-sparse vectors. This bound also holds for a fixed d-dimensional subspace. Hence,

$$\gamma_2(\mathcal{P}, \|\cdot\|) \lesssim \sqrt{\frac{d}{M_0}} \log d \log \widetilde{M} \tag{6}$$

Plugging these quantities into Theorem 1, we can obtain Lemma 1.

Lemma 1 *For any orthonormal matrix $\mathbf{U} \in R^{\widetilde{N} \times d}$ and a block diagonal matrix $\mathbf{S}_D$ as in Theorem 1, there exists a constant c such that*

$$\mathbb{P}\left(\|\mathbf{\Delta}\| \le c \sqrt{\frac{d \log(2/\delta)}{M_0}} \right) \ge 1 - \delta. \tag{7}$$

For a desired tolerance ϵ, if $M_0 = \Omega\left(\frac{d \log(2/\delta)}{\epsilon^2}\right)$, $\mathbb{P}(\|\mathbf{\Delta}\| \le \epsilon) \ge 1 - \delta$. This is similar to a subspace embedding guarantee. We now show that this tail bound naturally induces a bound on the moments of $\|\mathbf{\Delta}\|$, from which the main theorems in Section 2 can be proved.

1.3 Moment bound on $\|\mathbf{\Delta}\|$

Tail bounds for certain random variables can be translated into bounds on their moments using the following result:

Lemma 2 (Proposition 7.13, [Foucart and Rauhut, 2013]) *Suppose that a random variable q satisfies, for some $\gamma > 0$,*

$$\mathbb{P}\left(|q| \ge \mathrm{e}^{1/\gamma} \alpha u \right) \le \beta \mathrm{e}^{-u^\gamma/\gamma}$$

for all $u > 0$. Then, for $p > 0$,

$$\mathrm{E}\,|q|^p \le \beta \alpha^p (\mathrm{e}\gamma)^{p/\gamma} \Gamma\left(\frac{p}{\gamma} + 1 \right)$$

where $\Gamma(\cdot)$ is the Gamma function.

To adapt this result to bound the moments of the spectral norm of the random matrix $\mathbf{\Delta}$, we can choose $q = \|\mathbf{\Delta}\|$, $\gamma = 2$, $\beta = 1$ and $\mathrm{e}^{-u^2/2} = \delta$. We can then obtain the following result.

Lemma 3 *For any orthonormal matrix $\mathbf{U} \in R^{\widetilde{N} \times d}$ and a block diagonal matrix $\mathbf{S}_D$ as in Theorem 1 and $M_0 = \Omega\left(\frac{d \log(2/\delta)}{\epsilon^2}\right)$, then*

$$\mathrm{E}\,\|\mathbf{\Delta}\|^p \le \epsilon^p \delta \tag{8}$$

for $p = (\frac{\log(1/\delta)}{\epsilon^2})$.

1.4 Approximate matrix product guarantee

With the moment bound established above, we can now use the framework given by [Cohen et al., 2015] to establish (8, main paper). However, we cannot use their proof directly, since the sample complexity $\widetilde{M}$ in the moment bound in (8) is not oblivious to the matrix $\mathbf{U}$. However, once we fix the data matrix, we can adapt the argument used in [Cohen et al., 2015] to show that (8, main paper) holds.

Let $\mathbf{W}$ and $\mathbf{Y}$ be as in (8, main paper). As explained in [Cohen et al., 2015], we can assume that they have orthogonal columns. For a given k as in (8, main paper), let $\mathbf{W}$ and $\mathbf{Y}$ be partitioned into groups of k columns, with $\mathbf{W}_l$ and $\mathbf{Y}_{l'}$ denoting the l^{th} groups. [Cohen et al., 2015] then use the following result in their argument, which follows from (8):

$$\mathrm{E}\left\|(\mathbf{S}\mathbf{W}_l)^T(\mathbf{S}\mathbf{Y}_{l'}) - \mathbf{W}_l^T\mathbf{Y}_{l'}\right\|^p \leq \epsilon^p \left\|\mathbf{W}_l\right\|^p \left\|\mathbf{Y}_{l'}\right\|^p \delta \tag{9}$$

for all pairs (l, l'). This holds since in their setting, the sketching matrices are oblivious to the data matrices.

Although block diagonal matrices are not oblivious, this result holds with for $M_0 = \Omega\left(\frac{2k\log(2/\delta)}{\epsilon^2}\right)$. This is because of the observation that if $\mathbf{U}$ is an orthobasis for the span of $\mathbf{W}$ and $\mathbf{Y}$ and $\mathbf{U}^{l,l'}$ is an orthobasis for the span of $\mathbf{W}_l$ and $\mathbf{Y}_{l'}$, then

$$\Gamma(\mathbf{U}_j^{l,l'}) \leq \Gamma(\mathbf{U}_j) \tag{10}$$

for all pairs (l, l'). Hence, a given block diagonal sketching matrix $\mathbf{S}_D$ can satisfy (9) as well. The rest of the proof remains the same as [Cohen et al., 2015]. This concludes the proof for Theorem 1. Extending this to prove Theorem 2 is straightforward, with $\mathbf{S}_D$ being a particular case of their framework.

2 Algorithm for estimation of the incoherence parameters $\Gamma(\mathbf{U}_j)$

Our algorithm for estimating the block incoherence parameters is inspired by the algorithms for leverage score estimation in the row sampling literature [Drineas et al., 2012, Woodruff, 2014] and from randomized SVD algorithms [Halko et al., 2011].

The main idea is the following: suppose we had access to the QR factorization of the data matrix $\mathbf{A} \in \widetilde{N} \times d$:

$$\mathbf{A} = \mathbf{Q}\mathbf{R}. \tag{11}$$

Then, an orthobasis can be obtained by computing $\mathbf{Q} = \mathbf{A}\mathbf{R}^{-1}$. However, computing the QR-factorization is as expensive as the matrix multiplication or ridge regression problems. We use a similar approach, but we only aim to capture the row space of $\mathbf{A}$ in a distributed fashion. However, we take random projections in an iterative fashion, until the row space of the sketch "converges". we estimate the QR factorization from this resulting sketch. Our algorithm is described in Algorithm 1. Note that we only aim to compute a constant factor approximation of the QR factors. Hence, computing the $\mathbf{R}$ takes, in the worst case, $O(JdN\log N) = O(\widetilde{N}d\log N)$ time. The QR factorization in each iteration can be updated from its previous estimates efficiently. Computing the final estimate takes about $O(Jd^3)$ time. Finally computing $\hat{\Gamma}(\mathbf{U}_j)$'s takes $O(\widetilde{N}d)$ time, resulting in a total worst case time complexity of $O(\widetilde{N}d\log N)$.

Algorithm 1 Estimation of incoherence parameters up to constant factor error

Input: Blocks $\mathbf{A}_j$.
Initialize $\Omega \in \mathbb{R}^{O(1)\times N}, \mathbf{Q} = 0, \mathbf{R} = 0$, $\hat{\mathbf{A}} = 0$ where Ω is drawn from any subsampled randomized FJLT.
rank(R) not converged Compute $\hat{\mathbf{A}}_j = \Omega\mathbf{A}$.
Aggregate $\hat{\mathbf{A}} = [\hat{\mathbf{A}}_1^\top\ \hat{\mathbf{A}}_2^\top \cdots \hat{\mathbf{A}}_J^\top]^\top$ at the central processing unit with previous estimate
Update $\mathbf{Q}\mathbf{R} = \mathrm{qr}(\hat{\mathbf{A}})$
Draw a new independent realization of Ω
Compute $\hat{\Gamma}(\mathbf{U}_j) = \left\|\mathbf{A}\mathbf{R}^{-1}\right\|_F^2$

Output: Normalized estimates $\hat{\Gamma}(\mathbf{U}_j)/\sum_j \hat{\Gamma}(\mathbf{U}_j)$

References

[Avron et al., 2016] Avron, H., Clarkson, K. L., and Woodruff, D. P. (2016). Sharper bounds for regularized data fitting. *arXiv preprint arXiv:1611.03225*.

[Cohen et al., 2015] Cohen, M. B., Nelson, J., and Woodruff, D. P. (2015). Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*.

[Drineas et al., 2012] Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506.

[Eftekhari et al., 2015] Eftekhari, A., Yap, H. L., Rozell, C. J., and Wakin, M. B. (2015). The restricted isometry property for random block diagonal matrices. *Applied and Computational Harmonic Analysis*, 38(1):1–31.

[Foucart and Rauhut, 2013] Foucart, S. and Rauhut, H. (2013). *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel.

[Halko et al., 2011] Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

[Krahmer et al., 2014] Krahmer, F., Mendelson, S., and Rauhut, H. (2014). Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904.

[Woodruff, 2014] Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157.