Plan:

1. Univer. quantization via Random rotation.

2. MatMul & IPRD problem
   - → worst case $R<\infty$ quant perf $\gtrsim n^2$
   - → sketching $\gtrsim n^2$
   - → our thm 2 $\lesssim n$!

3. Gaussian. IPRD.

4. Open problems.

(1) Recall $D_{\text{vec}}(R; P_X) \leq 2^{-2R}$

Random rotation: Any $X \to$ gsn + ε)

$$X_G^n \to \hat{X}_G^n$$

$$Y^n$$

$$\mathbb{E}\|Y^n - X_G^n\|_2^2 \leq n\varepsilon \Rightarrow \mathbb{E}\|\hat{X}_{G}^n - Y^n\|_2^2 \leq n(D+\varepsilon+2\sqrt{D\varepsilon})$$

$$\mathbb{E}\|a+b\|^2 = \mathbb{E}\|a\|^2 + \mathbb{E}\|b\|^2 + 2\mathbb{E}\langle a,b\rangle.$$

$$\leq nD + n\varepsilon + 2n\sqrt{D\varepsilon}$$

take: $Y^N = UX^N$

Given x^N : $W_2(Y^n, X_G^n) \leq f(N,n)$

$$(Y_1^n, Y_{n+1}^{2n}, \ldots, Y_{N-n+1}^N)$$

$$(\hat{Y}_1^n, \ldots, \hat{Y}_{N-n+1}^N)$$

we know $\mathbb{E}_U \|\hat{Y}_1^n - Y_1^n\|_2^2 \leq n(D+\delta)$

$$\Downarrow$$

$$\mathbb{E}_U \|Y^N - \hat{Y}^N\|_2^2 \leq N(D+\delta)$$

$$\Updownarrow$$

$$\mathbb{E}_U \|U^{-1}Y^N - U^{-1}\hat{Y}^N\|_2^2 \leq N(D+\delta)$$

$$\Updownarrow$$

$$\mathbb{E}_U \|X^N - U^{-1}\hat{Y}^N\|_2^2 \leq N(D+\delta)$$

$$\Downarrow$$

$$\exists U \text{ s.t. } \mathbb{E}_{X^N} \|X^N - U^{-1}\hat{Y}^N\|_2^2 \leq N(D+\delta)$$

Final Quantizer:

- Send $\|X^N\|$
- UX^N
- Compress n-blocks via Gsn Quant.

Note: this requires sending $\|x^N\|$ separately and then redefining $x^N \leftarrow \frac{x^N}{\|x^N\|}$

+ explain how outliers are removed:

$X \sim (1-\delta)\, U_{[-1,1]} + \delta/2\, \delta_{1/\sqrt{\delta}} + \delta/2\, \delta_{-1/\sqrt{\delta}} \Rightarrow \mathcal{N}(0, 1/3)$

(2) Mat Mul:

(a) **Setup:**

$$\mathbb{R}^{n\times a} \ni A \xrightarrow{f_1} w_A \in \{0,1\}^{naR}$$

$$\mathbb{R}^{n\times b} \ni B \xrightarrow{f_2} w_B \in \{0,1\}^{nbR}.$$

$$(w_A, w_B) \xrightarrow{g} \widehat{A^T B} \in \mathbb{R}^{a\times b}.$$

WANT: R - small

$\forall A, B$ $\quad \|A^T B - g(w_A, w_B)\|^2$ — small

(B) What if we try to directly quantize A & B?

$$(A^T B - \hat{A}^T \hat{B})^2 \underset{\hat{B}=B \text{ best case}}{=} \left((A-\hat{A})^T B\right)^2 \leq \|A-\hat{A}\|^2 \|B\|^2 \simeq n^2.$$

Cannot do anything better than C.-S.!

$\boxed{\text{Prop}}$ Suppose $A, B \in \sqrt{n} S^{n-1}$. Then $\forall$ finite rate deterministic quantizers f_1, f_2, g we must have $|A^T B - \widehat{A^T B}|^2 \gtrsim n^2$ for some A, B.

TODO: need to show that also $|A^T B|^2 \leq n$

Later: this is only to compare vs Thm 1a, but Thm 1b is ok

__Pf:__

Take $\sqrt{n}\delta$-net on S^{n-1} with 2^{nR+1} pts.

$\exists \; A, \tilde{A}$ s.t. $q(A) = q(\tilde{A})$

$\forall B$: $\widehat{A^T B} = \widehat{\tilde{A}^T B}$ $\Rightarrow$ taking $B = \frac{(A-\tilde{A})\sqrt{n}}{\|A-\tilde{A}\|}$ we get

$$\sqrt{n\delta} \lesssim \|A-\tilde{A}\| = |A^T B - \tilde{A}^T B| \frac{1}{\sqrt{n}}$$
$$\leq \left(|A^T B - \widehat{A^T B}| + |\widehat{\tilde{A}^T B} - \tilde{A}^T B|\right)\frac{1}{\sqrt{n}}$$

$\Rightarrow$ At least one of the terms is $\gtrsim \frac{n\sqrt{\delta}}{2}$.

$\Rightarrow$ QED.

(Skip for lecture)

$\boxed{\text{Prop}}$ Dithered $\mathbb{Z}^n$ ach's $\Theta(n)$ error

Consider dithered $\mathbb{Z}^n$ quantizer. $\hat{A}_1 = [A + Z_A]_\varepsilon - Z_A \overset{(d)}{=} A + Z_A$ $\quad$ Unif$[-\varepsilon/2, \varepsilon/2]$

$$\mathbb{E}(\hat{A}^T\hat{B} - A^TB)^2 = \mathbb{E}\left((A+Z_A)^T(B+Z_B) - A^TB\right)^2 = \mathbb{E}\left(Z_A^T Z_B + Z_A^T B + Z_B^T A\right)^2 = \operatorname{tr}\Sigma_A\Sigma_B + \operatorname{tr}\Sigma_A BB^T + \operatorname{tr}\Sigma_B AA^T = \left(\frac{\varepsilon^2}{12}\right)^2 \operatorname{tr} I_n + \frac{\varepsilon^2}{12}\|B\|^2 + \frac{\varepsilon^2}{12}\|A\|^2 \asymp n\varepsilon^2$$

But notice that we need $\sqrt{n}/\varepsilon$ pts per coordinate

$\Rightarrow$ Rate is $\Omega(\log n)$ ☹

However, if $A_i \overset{iid}{\sim} P_A$, then

If we truncate $\hat{A} = \operatorname{clip}(A, M, -M)$ then we introduce error $\sqrt{n}$: $\quad \|A - \hat{A}\| \le \varepsilon\sqrt{n}$ w.h.p.

$$\Rightarrow |\hat{A}^T\hat{B} - A^TB| \le (\|A\| + \|B\|)\varepsilon\sqrt{n} \lesssim \varepsilon n.$$

$\Rightarrow$ quantizing to $R = \log\frac{2M}{\varepsilon}$ uniformly with dither achieves

$$\mathbb{E}\|A^TB - \widehat{A^TB}\|^2 \lesssim \varepsilon n$$

but not $\forall A, B \in \sqrt{n}\,S^{n-1}$!

(c) What about sketching? (randomization)

Let $u \sim \mathcal{N}(0, I_n)$, shared between A, B

$$\hat{u}_A = q_\varepsilon(u^T A) \quad = \overline{u^T A} + Z_A$$

dithering

$$\hat{u}_B = q_\varepsilon(u^T B). \quad = u^T B + Z_B$$

<u>Note</u>: $\mathbb{E}(u^T A)(u^T B) = \mathbb{E} A^T u u^T B = A^T B$!

So approximate:

$$\widehat{A^T B} := \hat{u}_A \hat{u}_B \quad \text{- unbiased}$$

$$\mathbb{E}\left(\hat{u}_A \hat{u}_B - A^T B\right)^2 = \mathbb{E}\Big((u,A)(u,B) + Z_A (u,B) + Z_B (u,A) + Z_A Z_B - A^T B\Big)^2$$

$$= \left(\mathbb{E} u_A^2 + \mathbb{E} u_B^2\right)\frac{\varepsilon^2}{12} + \left(\frac{\varepsilon^2}{12}\right)^2 + \underbrace{\mathbb{E}\left(u_A u_B - A^T B\right)^2}_{\mathbb{E}(u_A u_B)^2}$$

$$\mathbb{E}\, u_A^2 = \mathbb{E}\, A^T u u^T A = \|A\|^2.$$

where $A = \|A\| e_1$, $B = \|B\|(\rho e_1 + \sqrt{1-\rho^2} e_2)$

$$\mathbb{E}\, u_A^2 u_B^2 = \|A\|^2 \|B\|^2 \; (*)$$

$$(*)\; \mathbb{E}\left(u_1^2 \left(\rho u_1 + \sqrt{1-\rho^2}\, u_2\right)^2\right)$$

Actually, let's do $u \sim \mathcal{N}(0, I_n)$.

$$= \|A\|^2 \|B\|^2 \mathbb{E}\left[\rho^2 u_1^4 + u_1^2 (1-\rho^2) u_2^2\right]$$

$$= \|A\|^2 \|B\|^2 \left(3\rho^2 + (1-\rho^2)\right)$$

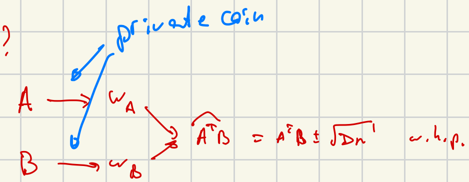
$$= \|A\|^2 \|B\|^2 \left(1 + 2\rho^2\right)$$

$$\rho \triangleq \cos\angle(A, B)$$

$\Rightarrow$ Again $\mathbb{E}\left(A^T B - \widehat{A^T B}\right)^2 \asymp n^2$ if $\|A\| \asymp \|B\| \asymp \sqrt{n}$

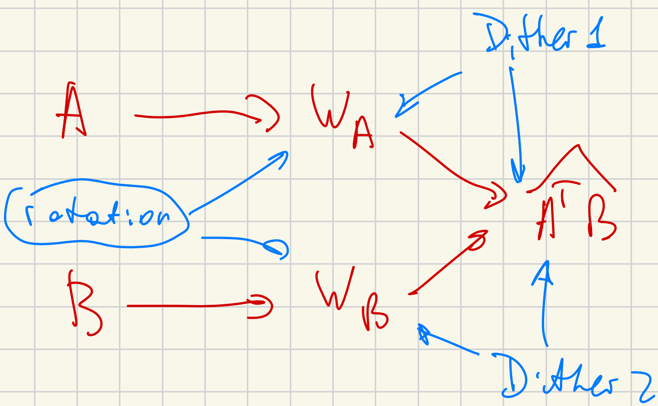
$\boxed{\text{Open Q:}}$ Does non-shared randomness help?

$w \sim \text{Unif}$ $\quad q(u, w)$



$\boxed{\text{Open Q:}}$ How about C.R. between A and dec, B and dec, but not between A, B?

Our solution uses 3 types of randomness



(a) Nevertheless, we proved.

(b) $\boxed{\text{Thm 1}}$ $\forall\, \varepsilon > 0$, $R > 0$ $\forall\, n \geq n_0(\varepsilon)$ $\exists\, f_1, f_2, g$ (sharing common randomness) s.t.

$\forall$ reasonable A, B

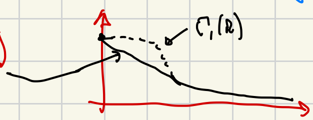
a) $$\mathbb{E}\, \|A^T B - g(f_1(A), f_2(B))\|_F^2 \leq \|A^T B\|_F^2 \left(\Gamma^2(R) + \varepsilon\right) + \frac{\|A\|_F^2\, \|B\|_F^2}{n} \left(\Gamma - \Gamma^2 + \varepsilon\right)$$

Only show this $\rightarrow$

b) $$\leq \left(2D - D^2 + \varepsilon\right) \frac{\|A\|_F^2\, \|B\|_F^2}{n} \qquad D = \frac{1}{1 + 2^R}$$

Notes:

- $\mathbb{E}$ over shared randomness
- In practice, just use seeds and store them with the matrix
- Guarantee is for every entry.
- $\Gamma(R)$



$\Gamma_1(R) = 2^{-2R+1} - 2^{-4R}$

- b) is slightly worse for Gsn but gives $\sqrt{n}$ rate for all pairs A, B even if $A^T B \approx n$!

③ Gaussian case is key to understand.

Setting: • $A \perp\!\!\!\perp B \sim \mathcal{N}(0, I_n)$.

• compress each to nR bits

• **Goal:** $\min \mathbb{E}\left(A^T B - g(f(A), f(B))\right)^2$

$\boxed{\text{Th}}$ 1) $\forall f_1, f_2, g$:

$$\mathbb{E}\left(A^T B - \widehat{A^T B}\right)^2 \geq n\,\Gamma(R)$$

2) $\forall \varepsilon$ $\exists$ compressors $f_1 = f_2, g$:

$$\mathbb{E}\left(A^T B - \widehat{A^T B}\right)^2 \leq n\left(\Gamma(R) + \varepsilon\right)$$

⑥ Preparatory

Consider some quantizer $f: \mathbb{R}^n \to \{0,1\}^{nR}$

(It can be shown WLOG $f_1 = f_2$)

Given $f(A), f(B)$ $\quad \widehat{A^T B} = \mathbb{E}[A^T B \mid f(A), f(B)]$

$$= \left(\mathbb{E}[A \mid f(A)]\right)^T \left(\mathbb{E}[B \mid f(B)]\right)$$

Define $\hat{A} = \mathbb{E}[A|f(A)]$

$\Sigma_A = \mathbb{E}(A-\hat{A})(A-\hat{A})^T \Rightarrow \mathbb{E}\hat{A}\hat{A}^T = I - \Sigma_A.$

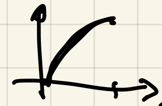
Note: $u \perp\!\!\!\perp v$
$\mathbb{E}(u^T v)^2 = \mathrm{tr}\, \Sigma_u \Sigma_v.$
$\Sigma_u = \mathbb{E} u u^T$

$\boxed{\text{Prop}}$ $\mathbb{E}\left(A^T B - \widehat{A^T B}\right)^2 = \mathbb{E}(A^T B)^2 - \mathbb{E}(\hat{A}^T \hat{B})^2 = \mathrm{tr}\, I - \mathrm{tr}\,(I - \Sigma_A)(I - \Sigma_B)$

$= \mathrm{tr}\,\Sigma_A + \mathrm{tr}\,\Sigma_B - \mathrm{tr}\,\Sigma_A \Sigma_B$

$= 2\,\mathrm{tr}\,\Sigma_A - \mathrm{tr}\,\Sigma_A^2.$

$= \sum_{i=1}^{n} \varphi(d_i)$ $\quad d_i = \mathrm{eig}(\Sigma_A)$

$\varphi(x) = 2x - x^2 =$ 

Thus, designing good IPRD quantizer $\Leftrightarrow \min \sum_i \varphi(d_i)$

E.g.: scalar quantizer has $d_i = D$

$\Rightarrow$ best scalar quantizer for IPRD is same as best S.Q. for RD!

(c) Lower Bound:

$\Gamma_1(R) \triangleq \varphi\left(2^{-2R}\right).$ $\quad R_i \triangleq \frac{1}{2}\ln\frac{1}{d_i}$

$nR \geq I(A; f(A)) = h(A) - h(A|f(A)).$

$\geq \frac{n}{2}\log(2\pi e) - \frac{1}{2}\log\left[(2\pi e)^n \det \Sigma_A\right]$

Fact: $\mathrm{Cov}(X|Y) = \Sigma$
$\Rightarrow h(X|Y) \leq h(\mathcal{N}(0, \Sigma))$

$= \sum_{i=1}^{n} \frac{1}{2}\log\frac{1}{d_i}$

So: $R \geq \frac{1}{n}\sum_{i=1}^{n} R_i$

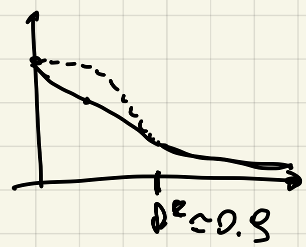
$D = \frac{1}{n}\sum_{i=1}^{n} \Gamma_1(R_i) \geq \frac{1}{n}\sum_{i=1}^{n} \Gamma(R_i) \geq \Gamma\left(\frac{1}{n}\sum_i R_i\right) \geq \Gamma(R).$

QED.

(d) Upper Bound:

So suppose we have some Gaussian quantizer
$A \to \{0,1\}^{nR}$ and $\mathrm{tr}\, \Sigma_A \approx n 2^{-2R}$.

$$\mathbb{E}(A^T B - \hat{A}^T \hat{B})^2 = 2\,\mathrm{tr}\,\Sigma_A - \mathrm{tr}\,\Sigma_A^2 \qquad \frac{1}{n}\mathrm{tr}\,\Sigma_A^2 \geq \left(\frac{1}{n}\mathrm{tr}\,\Sigma_A\right)^2$$
$$\leq 2\,\mathrm{tr}\,\Sigma_A - \frac{1}{n}(\mathrm{tr}\,\Sigma_A)^2$$
$$= n\,\varphi(2^{-2R}) = n\Gamma_1(R)$$



Q: How to Improve to $\Gamma(R)$?

A: Compress fraction of coord at rate 0 (i.e. repl. with 0)
and other fraction at rate $R^* \approx 0.9$

then $\Sigma_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \\ 0 & & \boxed{\Sigma_A'} \end{bmatrix}$ $\qquad \mathrm{tr}\,\Sigma_A^{n'} = n' 2^{-2R^*}$.

Overall distortion $n'\varphi(2^{-2R^*}) + (n-n')\,\varphi(1)$ $\approx 1.$

Overall rate: $\dfrac{n' \cdot R^* + (n-n')\cdot 0}{n}$.

(f) For moving these results to $\forall A, B$
need to tackle case where $A_i, B_i \sim \mathcal{N}(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$ - correl.!

Solution: New result about $\exists$ of lattices

[specifically, random rot. creates $(A_i, B_i) \overset{iid}{\approx} \mathcal{N}(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$ and we cannot apply Gsn IPRD thm since it is for $\rho = 0$!]

(4) Conclusions & open problems.

- Two regimes : tokenization ($R \ll 1$)
  quantization ($R \gtrsim 1$)
- Vector Q. yields advantages
- History:
  - For audio / image / video
    critique: loss $\neq$ MSE
    input $\neq$ iid
  - turned out input is approx $\perp\!\!\!\perp$ $\mathcal{N}$ in wavelet / DCT domain $\Rightarrow$ waterfilling suggests what to keep / kill

<u>Open:</u> Same for LLM matrices? what is the right str-re?

- One interesting conclusion from VQVAE story is that (despite decades of prevailing opinion) good compression <u>can</u> be found by info bottleneck

$$\boxed{x} \longrightarrow [K]^{\otimes m} \longrightarrow \boxed{\hat{x}}$$

and vanilla MSE $\|x-\hat{x}\|^2$

Somehow training this on 1B images gives good lossy compression

Note: VQGAN also adds a loss term for local patch statistics

$$KL(P_{\hat{x}_{patch}} \,\|\, P_{x_{patch}}) \to \min.$$

Open problems: **in weight-only quant.**

- Sparsification: $Y = WX$ is solved faster if there are 0's in W => planting many 0's is a form of quantization (w/o any rigorous analysis)

„theory of LDLQ"

- New form of ε-net.

metric $= (W - \hat{W})^T H (W - \hat{W})$

Problem: metric is only known at encoder:

C is ε-net for S^{n-1} if:

$\forall H > 0 \ (\mathrm{tr}\, H \leq n) \ \forall x \in S^{n-1} \ \exists\, c \in C$ s.t.

$$\| x - c \|_H \leq \varepsilon \cdot \sqrt{n}$$

Q: $\min |C| = ?$

Prelim: $R(D; M)$ function is worst for $M = I_n$ => suspect that regular ε-net is ok for this?

Second tier problem is:

Let $\xi^*(M, R)$ – best ξ for M-metric and $|C| = 2^{nR}$. Want universal quant C s.t.

$$\mathrm{red}_n = \sup_x \sup_M \min_{c \in C} \|x - c\|_M - \xi^*(M, R)$$

„price of universality"

„adapting to W"

- Adapting to correlations in rows.

Suppose $w^{(1)}, \dots, w^{(m)} \in \mathbb{R}^n$ are rows of wt matrix.

Suppose $w^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \Sigma_i)$, but Σ_i unknown.

If Σ_i were known we'd communicate at distortion $D = \frac{1}{n}\sum_i \sigma_i^2 \wedge T$

$$R = \frac{1}{n}\sum_i \log_+ \frac{T}{\sigma_i}$$

But we only have empirical statistics of weights.

What is the price of not knowing Σ_1?

$$\left.\begin{matrix} w^{(1)} \\ \vdots \\ w^{(m)} \end{matrix}\right\} \rightarrow \lceil 2^{nmR} \rceil \text{ bits} \rightarrow \begin{matrix} \hat{w}^{(1)} \\ \vdots \\ \hat{w}^{(m)} \end{matrix}$$

$$\text{redundancy}_m \triangleq \sup_{\Sigma_1} \frac{1}{nm} \sum_i \mathbb{E}\|w^{(i)} - \hat{w}^{(i)}\|^2 - D(R;\Sigma_1).$$

optimal distortion when Σ_1 is known ahead of time (water filling)

Refs: classical work and more recent (Wagner...) shows red_m for the case of discrete source w and $n=1$. Here we ask abt Gsn and $n>1$.

Note: Main goal of these open problems is to understand how to allocate bits between explaining top PCA directions of Σ_1, M vs compressing them

Open Questions in weight + act quantizations (mat mul).

1) $A \perp B \longrightarrow UA,\ UB \xrightarrow{\{\mathbb{Z}\}^n} [UA]_\varepsilon, [UB]_\varepsilon$

(random rotation) (dithered unif. quant.)

$\longrightarrow$ Clip $\longrightarrow \hat{A}, \hat{B}$

(truncation to $[-100;100]$)

This yields finite rate quant.

Q1: Show $\mathbb{E}(\hat{A}^T\hat{B} - A^TB)^2 \lesssim n$.

(here idea is that $UA \approx \mathcal{N}(0, I_n \frac{\|A\|^2}{n})$)

Q2: Show the same w/o dithering.

2) Optimal IPRD for $A_i, B_i \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$

3) [Hard] Quantizing $\prod_{i=1}^{k} A^{(i)}$ for $k \geq 2$? $A^{(i)}_{\ell,m} \overset{iid}{\sim} \mathcal{N}(0,1)$

Draft notes

Sketching: $A, B \in \mathbb{R}^n$, $u \sim \mathrm{Unif}\ S^{n-1}\sqrt{n}$, $\|A\|, \|B\| \lesssim \sqrt{n}$.

$$|(u,A)| \le \sqrt{n}\cdot \|A\| \lesssim n \Rightarrow \quad \hat{u}_A = (u,A) + Z$$

$Z \sim \mathrm{Unif}(-\delta/2, \delta/2)$.

$\#\text{bits} = \log_2 \frac{2n}{\delta}$.

$$\mathbb{E}\,(u,A)(u,B) = \mathbb{E}\,A^T u u^T B = A^T B \ !$$

$$\mathbb{E}\left(\hat{u}_A \hat{u}_B - A^T B\right)^2 = \mathbb{E}\big((u,A)(u,B) + Z_A(u,B) + Z_B(u,A) + Z_A Z_B - A^T B\big)^2$$

$$= \left(\mathbb{E}\,u_A^2 + \mathbb{E}\,u_B^2\right)\frac{\delta^2}{12} + \left(\frac{\delta^2}{12}\right)^2 + \underbrace{\mathbb{E}\left(u_A u_B - A^T B\right)^2}_{\mathbb{E}(u_A u_B)^2}$$

$$\mathbb{E}\,u_A^2 = \mathbb{E}\,A^T u u^T A = \|A\|^2.$$

wlog $A = \|A\| e_1$, $B = \|B\|(\rho e_1 + \sqrt{1-\rho^2}\, e_2)$

$$\mathbb{E}\,u_A^2 u_B^2 = \|A\|^2 \|B\|^2 \cdot \mathbb{E}\left(u_1^2 \left(\rho u_1 + \sqrt{1-\rho^2}\, u_2\right)^2\right)$$

$$= \|A\|^2 \|B\|^2\, \mathbb{E}\left[\rho^2 u_1^4 + u_1^2 (1-\rho^2) u_2^2\right]$$

$$= \|A\|^2 \|B\|^2 \left(3\rho^2 + (1-\rho^2)\right)$$

$$= \|A\|^2 \|B\|^2 \left(1 + 2\rho^2\right)$$

$\boxed{\text{Actually, let's do } u \sim \mathcal{N}(0, I_n)!}$

$$\rho \triangleq \cos\angle(A,B)$$

Truncating Gsn for $\mathbb{Z}^d$ quant. method

$$A \to \tilde{A} = \{A_i \vee -M \wedge M.\}$$

$$\mathbb{E} \; \|A - \tilde{A}\|_2^2 = \mathbb{E} \sum_i \mathbb{1}\{|A_i| > M\}(|A_i| - M)^2 \leq \delta n.$$

$$|\tilde{A}^T \tilde{B} - A^T B + \tilde{A}^T B - \tilde{A}^T B| \leq$$

$$\leq \|\tilde{A}\| \cdot \|B - \tilde{B}\| + \|B\| \, \|A - \tilde{A}\|$$

$$\leq 2 \tau n \quad \text{w.h.p.}$$

$$\Rightarrow \tilde{A}$$