A APPENDIX

A.1 DERIVATION OF COROLLARY 1 & 2

Corollary 1. *Consider a set of independent q-dimensional Gaussian random vectors which are pairwise ϵ-orthogonal with probability $1-\nu$, then the number of such Gaussian random vectors is bounded by*

$$N \leq \sqrt[4]{\frac{\pi}{2q}}\, \mathrm{e}^{\frac{\epsilon^2 q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}. \tag{A.1}$$

Proof. Recall that, in the case of Gaussian distributed random vectors, the pdf of ρ is

$$g(\rho) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} (1-\rho^2)^{\frac{q-3}{2}}.$$

This directly yields that $\omega := \sqrt{q}\rho$ has the density function

$$f(\omega) = \frac{1}{\sqrt{q}} \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} \left(1 - \frac{\omega^2}{q}\right)^{\frac{q-3}{2}} \to \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\omega^2}{2}} \tag{A.2}$$

as $q \to \infty$, using the fact that $\frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} \sim \sqrt{\frac{q}{2}}$. Therefore the probability that two random Gaussian vectors are not ϵ-orthogonal is upper bounded by

$$\begin{aligned}\Pr(|\rho| \geq \epsilon) = \Pr(|\omega| \geq \sqrt{q}\epsilon) &= 2\int_{\sqrt{q}\epsilon}^{\sqrt{q}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\omega^2}{2}}\, \mathrm{d}\omega \\ &< \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{q\epsilon^2}{2}} (\sqrt{q} - \sqrt{q}\epsilon) < \sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{q\epsilon^2}{2}}.\end{aligned} \tag{A.3}$$

To estimate the probability that ϵ-orthogonality is satisfied for a set of N independent Gaussian random vectors, let us consider the following quantity

$$\mathcal{P}(\epsilon, N) := \prod_{k=1}^{N-1} \left[1 - k\Pr(|\rho| \geq \epsilon)\right]. \tag{A.4}$$

The above estimation has clear meaning. Given one Gaussian random vector $\mathbf{X}_1$, the probability that an independently sampled random vector $\mathbf{X}_2$ which is not ϵ-orthogonal to $\mathbf{X}_1$ is $\Pr(|\rho| > \epsilon)$. Similarly, given k i.i.d. Gaussian random vectors $\mathbf{X}_1, \cdots, \mathbf{X}_k$, the probability that an independently drawn Gaussian random vector $\mathbf{X}_{k+1}$ which is not ϵ-orthogonal to $\mathbf{X}_1, \cdots, \mathbf{X}_k$ is upper bounded by $k\Pr(|\rho| > \epsilon)$. Therefore, we have the estimate in Eq. A.4 for N independent random vectors. Using Eq. A.3, $\mathcal{P}(\epsilon, N)$ can be computed as follows

$$\begin{aligned}\mathcal{P}(\epsilon, N) &> \prod_{k=1}^{N-1} (1 - k\sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}}) \\ &> (1 - N\sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}})^N \sim \mathrm{e}^{-N^2\sqrt{\frac{2q}{\pi}}\mathrm{e}^{-\frac{\epsilon^2 q}{2}}},\end{aligned}$$

for sufficiently large N and q satisfying $N\sqrt{\frac{2q}{\pi}}\mathrm{e}^{-\frac{\epsilon^2 q}{2}} < 1$. If we require $\mathcal{P}(\epsilon, N) \geq 1-\nu$, then the number of pairwise ϵ-orthogonal i.i.d. Gaussian random vectors is bounded from above by

$$\begin{aligned}&\mathrm{e}^{-N^2\sqrt{\frac{2q}{\pi}}\mathrm{e}^{-\frac{\epsilon^2 q}{2}}} \geq 1 - \nu \quad \Rightarrow \\ &N \leq \sqrt[4]{\frac{\pi}{2q}}\, \mathrm{e}^{\frac{\epsilon^2 q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}\end{aligned}$$

■

Corollary 2. *Consider a set of n q-dimensional random Gaussian vectors, we have*

$$\lambda_{\mathrm{G}} := \mathbb{E}[|\rho_{\mathrm{G}}|] = \sqrt{\frac{2}{\pi q}}. \tag{A.5}$$

Proof. Given the $g(\rho_{\mathrm{G}})$ in Theorem 1, we have

$$\begin{aligned}\mathbb{E}[|\rho_{\mathrm{G}}|] &= \int_{-1}^{1} |\rho| g(\rho)\, \mathrm{d}\rho = \sqrt{\frac{2q}{\pi}} \int_0^1 \rho(1-\rho^2)^{\frac{q-3}{2}}\, \mathrm{d}\rho \\ &= -\sqrt{\frac{2q}{\pi}} \frac{(1-\rho^2)^{\frac{q-1}{2}}}{q-1}\Bigg|_0^1 = \sqrt{\frac{2}{\pi q}},\end{aligned}$$

for large q. ■

A.2 DISCUSSION ON CONJECTURE 1

In this section, we derive the approximations stated in Conjecture 1 and verify them with empirical simulations.

According to the central limit theorem, the sum of independently and identically distributed random variables with finite variance converges weakly to a normal distribution as the number of random variables approaches infinity. Our derivation relies on the generalized central limit theorem proven by Gnedenko and Kolmogorov in 1954 [Gnedenko et al. 1954].

Theorem A 1. *(**Generalized Central Limit Theorem** [Gnedenko et al. 1954]) Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d random variables drawn from the distribution with probability density function $f(x)$ with the following asymptotic behaviour*

$$f(x) \simeq \begin{cases} c_+ x^{-(\alpha+1)} & \textit{for} \quad x \to \infty \\ c_- |x|^{-(\alpha+1)} & \textit{for} \quad x \to -\infty, \end{cases} \tag{A.6}$$

where $0 < \alpha < 2$, and c_+, c_- are real positive numbers. Define random variable S_n as a superposition of $X_1, \cdots, X_n$

$$S_n = \frac{\sum\limits_{i=1}^{n} X_i - C_n}{n^{\frac{1}{\alpha}}}, \quad \textit{with}$$

$$C_n = \begin{cases} 0 & \textit{if} \quad 0 < \alpha < 1 \\ n^2 \Im \ln(\phi_X(1/n)) & \textit{if} \quad \alpha = 1 \\ n\mathbb{E}[X] & \textit{if} \quad 1 < \alpha < 2, \end{cases}$$

where ϕ_X is the characteristic function of a random variable X with probability density function $f(x)$, $\mathbb{E}[X]$ is the expectation value of X, $\Im$ denotes the imaginary part of a variable. Then as the number of summands n approaches infinity, the random variables S_n converge in distribution to a unique stable distribution $S(x; \alpha, \beta, \gamma, 0)$, that is

$$S_n \xrightarrow{d} S(\alpha, \beta, \gamma, 0), \quad \textit{for} \quad n \to \infty,$$

where, α characterizes the power-law tail of $f(x)$ as defined above, and parameters β and γ are given as:

$$\beta = \frac{c_+ - c_-}{c_+ + c_-},$$
$$\gamma = \left[\frac{\pi(c_+ + c_-)}{2\alpha \sin(\frac{\pi\alpha}{2})\Gamma(\alpha)}\right]^{\frac{1}{\alpha}}. \qquad \text{(A.7)}$$

To be self-contained, we give the definition of stable distributions after [Nolan 2003; Mandelbrot 1960].

Definition A 1. *A random variable X follows a stable distribution if its characteristic function can be expressed as*

$$\phi(t; \alpha, \beta, \gamma, \mu) = \mathrm{e}^{i\mu t - |\gamma t|^{\alpha}(1 - i\beta \ \mathrm{sgn}(t)\Phi(\alpha, t))}, \qquad \text{(A.8)}$$

with $\Phi(\alpha, t)$ defined as

$$\Phi(\alpha, t) = \begin{cases} \tan(\frac{\pi\alpha}{2}) & \textit{if} \quad \alpha \neq 1 \\ -\frac{2}{\pi}\log|t| & \textit{if} \quad \alpha = 1. \end{cases}$$

Then the probability density function $S(x; \alpha, \beta, \gamma, \mu)$ of the random variable X is given by the Fourier transform of its characteristic function

$$S(x; \alpha, \beta, \gamma, \mu) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \phi(t; \alpha, \beta, \gamma, \mu) \, \mathrm{e}^{-ixt} \, \mathrm{d}x.$$

The parameter α satisfying $0 < \alpha \leq 2$ characterizes the power-law asymptotic limit of the stable distribution, $\beta \in [-1, 1]$ measures the skewness, $\gamma > 0$ is the scale parameter, and $\mu \in \mathbb{R}$ is the shift parameter. Note that the normal distribution is a typical stable distribution. Other examples with analytical expression include the Cauchy distribution and the Lévy distribution. For the later use, we give the analytical form of the Lévy distribution.

Remark A 1. *The probability density function of the Lévy distribution is given by*

$$f(x; \gamma, \mu) = \sqrt{\frac{\gamma}{2\pi}} \frac{\mathrm{e}^{-\frac{\gamma}{2(x-\mu)}}}{(x-\mu)^{\frac{3}{2}}}, \quad x \geq \mu, \qquad \text{(A.9)}$$

where μ is the shift parameter and γ is the scale parameter. The Lévy distribution is a special case of the stable distribution $S(x; \alpha, \beta, \gamma, \mu)$ with $\alpha = \frac{1}{2}$ and $\beta = 1$. This can be seen from its characteristic function, which can be written as

$$\phi(t; \gamma, \mu) = \mathrm{e}^{i\mu t - |\gamma t|^{1/2}(1 - i \ \mathrm{sgn}(t))}$$

To derive $g(\rho_{\mathrm{C}})$ for Cauchy random vectors, we first need the distribution function of X^2 given that the random variable X has a Cauchy distribution.

Lemma A 1. *Let X be a Cauchy random variable having the probability density function $f_X(x) = \frac{1}{\pi}\frac{\zeta}{x^2+\zeta^2}$, where $\zeta > 0$ is the scale parameter. Then the squared variable $Y := X^2$ has the pdf:*

$$f_Y(y) = \begin{cases} \frac{1}{\pi}\frac{\zeta}{\sqrt{y}(\zeta^2+y)} & \textit{for} \quad y \geq 0, \\ 0 & \textit{otherwise}. \end{cases} \qquad \text{(A.10)}$$

Proof. $f_Y(y)$ can be derived from $f_X(x)$ by a simple variable transformation $y = g(x) = x^2$. In particular, utilizing the symmetry of $f_X(x)$, we have

$$\begin{aligned} f_Y(y) &= 2\left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right| f_X(g^{-1}(y)) \\ &= \frac{1}{\pi}\frac{\zeta}{\sqrt{y}(\zeta^2 + y)}. \end{aligned}$$

■

In the following Lemma we derive the probability density function for $z_{\mathbf{X},\mathbf{Y}}$, which is defined as $z_{\mathbf{X},\mathbf{Y}} := \frac{1}{q^2}\frac{X_2^2+\cdots X_q^2}{X_1^2}$.

Lemma A 2. *Let $X_1, \cdots, X_q$ be a sequence of i.i.d. random variables drawn from $\mathcal{C}(0, 1)$. Then the random variable $Z_q := \frac{1}{q^2}\frac{X_2^2+\cdots+X_q^2}{X_1^2}$ converges in distribution to*

$$f(z) = -\frac{1}{\pi^2}\frac{1}{z^{\frac{3}{2}}}\left[\mathrm{e}^{\frac{1}{\pi z}} \ \mathrm{Ei}\left(-\frac{1}{\pi z}\right)\right], \qquad \text{(A.11)}$$

as $q \to \infty$, where $\mathrm{Ei}(x)$ denotes the exponential integral.

Proof. The numerator in Z_q can be regarded as a sum of independent random variables with density function $f_{Y:=X^2}(y) = \frac{1}{\pi}\frac{1}{\sqrt{y}(1+y)}$, see Eq. A.10 with $\zeta = 1$. Thus, we can use the generalized central limit theorem to obtain the density function $g(\frac{1}{q^2}\sum_{i=2}^{q} X_i^2)$ for the numerator, as $q \to \infty$.

Note that $f_Y(y) \sim \frac{1}{\pi}y^{-\frac{3}{2}}$ as $y \to +\infty$. From this asymptotic behaviour we can extract that $c_+ = \frac{1}{\pi}$, $c_- = 0$, and $\alpha = \frac{1}{2}$. Moreover, Eq. A.7 with $\beta = 1$ yields $\gamma = \left[\frac{1}{\sin(\frac{\pi}{4})\,\Gamma(\frac{1}{2})}\right]^2 = \frac{2}{\pi}$. In summary, $g(\frac{1}{q^2}\sum_{i=2}^{q} X_i^2)$ converges to a unique stable distribution $S(\alpha = \frac{1}{2}, \beta = 1, \gamma = \frac{2}{\pi}, \mu = 0)$, which is exactly the Lévy distribution shown in Remark A 1. Hence, we have

$$g(\frac{1}{q^2}\sum_{i=2}^{q} X_i^2) \xrightarrow{d} S(x; \frac{1}{2}, 1, \frac{2}{\pi}, 0) = \frac{1}{\pi}\frac{\mathrm{e}^{-\frac{1}{\pi x}}}{x^{\frac{3}{2}}},$$
$$\text{as} \quad q \to \infty. \tag{A.12}$$

Next, we consider the quotient distribution of two random variables in order to derive the pdf of Z_q. To be more specific, let X and Y be independent non-negative random variables with corresponding probability density function $f_X(x)$ and $f_Y(y)$ over the domains $x \geq 0$ and $y \geq 0$, respectively. Then the cumulative distribution function $F_Z(z)$ of $Z := \frac{Y}{X}$ can be computed by

$$\begin{aligned} F_Z(z) &= \Pr(\frac{Y}{X} \leq z) = \Pr(Y \leq zX) \\ &= \int_0^\infty \left[\int_0^{y=zx} f_Y(y)\mathrm{d}y\right] f_X(x)\mathrm{d}x. \end{aligned}$$

Differentiating the cumulative distribution function yields

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z}F_Z(z) = \int_0^\infty x\, f_Y(zx)\, f_X(x)\, \mathrm{d}x.$$

Following the above procedure, we can obtain the pdf for Z_q as $q \to \infty$ in case the density functions of the numerator and the denominator are given by Eq. A.12 and Eq. A.10, respectively. That yields

$$\begin{aligned} f(z) &= \frac{1}{\pi^2}\int_0^\infty x\,\frac{\mathrm{e}^{-\frac{1}{\pi z x}}}{(zx)^{\frac{3}{2}}}\,\frac{1}{\sqrt{x}(1+x)}\,\mathrm{d}x \\ &= \frac{1}{\pi^2}\frac{1}{z^{\frac{3}{2}}}\left[-\mathrm{e}^{\frac{1}{\pi z}}\,\mathrm{Ei}\left(-\frac{x+1}{\pi z x}\right)\right]\Bigg|_{x=0}^{\infty} \\ &= -\frac{1}{\pi^2}\frac{1}{z^{\frac{3}{2}}}\left[\mathrm{e}^{\frac{1}{\pi z}}\,\mathrm{Ei}\left(-\frac{1}{\pi z}\right)\right]. \end{aligned}$$

■

In the following we discuss why the density function $g(\rho_{\mathrm{C}})$ can only be approximated by taking the limit as $q \to \infty$.

Suppose $\mathbf{X} = (X_1, \cdots, X_q)$ and $\mathbf{Y} = (Y_1, \cdots, Y_q)$ are Gaussian random variables. To derive $g(\rho_{\mathbf{X},\mathbf{Y}})$ in Lemma 1, [Cai et al. 2012; Muirhead 2009] compute the density function of $\frac{\boldsymbol{\alpha}^\intercal \cdot \mathbf{X}}{||\mathbf{X}||}$ instead, where $\boldsymbol{\alpha}^\intercal \cdot \boldsymbol{\alpha} = 1$, and $\boldsymbol{\alpha} := \frac{\mathbf{Y}}{||\mathbf{Y}||}$. In particular, without loss of generality, they assume $\boldsymbol{\alpha} = (1, 0, \cdots, 0)$. The justification for this assumption is that the random variable $\mathbf{X}^{'} := \frac{\mathbf{X}}{||\mathbf{X}||}$ is uniformly distributed on the $(q-1)$-dimensional sphere (see Theorem 1.5.6 in [Muirhead 2009]).

In our case, the distributional uniformity of $\frac{\mathbf{X}}{||\mathbf{X}||}$ is not superficial, since the density function of $\mathbf{X}^{'}$ doesn't depend on $\mathbf{X}^{'}$ only through the value of $\mathbf{X}^{'\intercal}\mathbf{X}^{'}$. To see this, in the following Lemma, we discuss the distribution function of the normalization $\frac{\mathbf{X}}{||\mathbf{X}||}$.

Lemma A 3. *Consider a q-dimensional random vector $\mathbf{X} = (X_1, \cdots, X_q)$, where $X_1, \cdots, X_q$ are independently and identically drawn from a Cauchy distribution $\mathcal{C}(0, 1)$. Then, as $q \to \infty$, the normalized random vector $\frac{\mathbf{X}}{||\mathbf{X}||} = (X_1^{'}, \cdots, X_q^{'})$ has a joint density function, in which the random variables $X_1^{'}, \cdots, X_q^{'}$ are all independent from each other.*

Proof. Without loss of generality, we study the pdf of $X_1^{'} = \frac{X_1}{\sqrt{X_1^2 + \cdots + X_q^2}}$. Similar to the proof of Lemma A 2, the random variable $Z_q := \frac{1}{q^2}\frac{X_2^2 + \cdots + X_q^2}{X_1^2}$ converges weakly to the distribution with pdf given by Eq. A.11 as $q \to \infty$, which is independent of the other random variables due to the generalized central limit theorem. Hence, $X_1^{'}$ can be treated as an independent random variable as $q \to \infty$. In addition, we obtain the pdf of $X_1^{'}$ given by

$$f_{X_1^{\prime}}(x_1^{\prime}) = -\frac{2}{\pi^2 q^2 x_1^{\prime 3}}\frac{1}{z_1^{\frac{3}{2}}}\left[\mathrm{e}^{\frac{1}{\pi z_1}}\,\mathrm{Ei}\left(-\frac{1}{\pi z_1}\right)\right], \tag{A.13}$$

where z_1 is defined as $z_1 := \frac{1}{q^2}\left(\frac{1}{x_1^{\prime 2}} - 1\right)$. The arguments can be easily generalized to $X_2^{'}, \cdots, X_q^{'}$. ■

The pdf of the joint distribution $f_{\mathbf{X}^{\prime}}(x_1^{\prime}, \cdots, x_q^{\prime})$ can be written as a product of marginals, that is

$$f_{\mathbf{X}^{\prime}}(x_1^{\prime}, \cdots, x_q^{\prime}) = \prod_{i=1}^{q} f_{X_i^{\prime}}(x_i^{\prime}),$$

as $q \to \infty$. The density function of $\mathbf{X}^{'}$ is not invariant under an arbitrary rotation. Thus, it is not uniformly distributed on S^{q-1}.

The above density function of normalized Cauchy random vectors leads to the following Remark.

Remark A 2. *The normalized Cauchy random vector $\mathbf{X}' = \frac{\mathbf{X}}{||\mathbf{X}||}$ is sparse in the sense that the density function of its elements can be approximated by a δ-function.*

Fig. 1 shows the empirical elements distribution of 1000 normalized Cauchy random vectors. This indicates that in sufficiently high-dimensional spaces the density function of the normalized entries converges to a δ-function. To explain this, recall the Laurent expansion of the density function given in Eq. A.13,

$$f_{X_1'}(x_1') = \frac{2}{\pi q x_1'^2} - \frac{2}{q^3 x_1'^4} + \frac{4\pi}{q^5 x_1'^6} + \mathcal{O}\left(\frac{1}{q^7 x_1'^8}\right). \tag{A.14}$$

This expansion converges to zero almost everywhere expect for $x_1' = 0$ as $q \to \infty$.

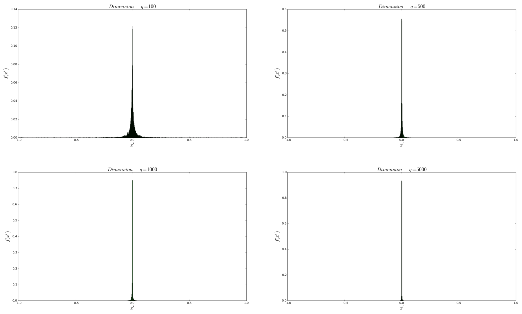


Figure 1: Empirical distributions of 10000 normalized Cauchy random vectors with dimensions $q = 100$, 500, 1000, 5000.

In the following, we provide a full derivation of $g(\rho_{\mathrm{C}})$ proposed in the Conjecture 1.

Conjecture 1. *Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent q-dimensional random vectors whose elements are independently and identically drawn from a Cauchy distribution $\mathcal{C}(0,1)$. Let Θ_{ij} be the angle between $\mathbf{X}_i$ and $\mathbf{X}_j$. Then, as $q \to \infty$, $\rho_{ij} := \cos\Theta_{ij} \in [-1,1]$, $1 \le i < j \le n$ are pairwise i.i.d. with density function approximated by*

$$g(\rho_{\mathrm{C}}) = -\frac{2}{\pi^2 q^2 \rho_{\mathrm{C}}^3} \cdot \frac{1}{z^{\frac{3}{2}}} \left[\mathrm{e}^{\frac{1}{\pi z}}\, \mathrm{Ei}\left(-\frac{1}{\pi z}\right)\right], \tag{A.15}$$

where $z := \frac{1}{q^2}\left(\frac{1}{\rho_{\mathrm{C}}^2} - 1\right)$.

Given two Cauchy random vectors $\mathbf{X} = (X_1, \cdots, X_q)$ and $\mathbf{Y} = (Y_1, \cdots, Y_q)$, $\rho_{\mathbf{X},\mathbf{Y}}$ is approximated by $\rho_{\mathbf{X},\mathbf{Y}} \approx \frac{X_1}{\sqrt{X_1^2 \cdots + X_q^2}}$.

Furthermore, we introduce the new variable $z_{\mathbf{X},\mathbf{Y}} := \frac{1}{q^2}(\frac{1}{\rho_{\mathbf{X},\mathbf{Y}}} - 1)$. From Lemma A 2 we have the density function $\hat{g}(z_{\mathbf{X},\mathbf{Y}})$. Then, $g(\rho_{\mathbf{X},\mathbf{Y}})$ can be directly obtained from $\hat{g}(z_{\mathbf{X},\mathbf{Y}})$ by a variable transform, that is $g(\rho_{\mathbf{X},\mathbf{Y}}) = \left|\frac{\mathrm{d}z}{\mathrm{d}\rho}\right| \hat{g}(z_{\mathbf{X},\mathbf{Y}})$. With $\left|\frac{\mathrm{d}z}{\mathrm{d}\rho}\right| = \frac{2}{q^2\rho^3}$ we immediately get Eq. A.15 as the density function for $\rho_{\mathbf{X},\mathbf{Y}}$.

Assume that Eq. A.15 is valid as $q \to \infty$. In the following we show that $\{\rho_{ij} | 1 \le i < j \le n\}$ are i.i.d random variables. First, notice that ρ_{ij} and ρ_{kl} are independent if $\{i,j\} \cap \{k,l\} = \emptyset$. It is left to prove that $\rho_{\mathbf{X},\mathbf{Y}}$ and $\rho_{\mathbf{X},\mathbf{Z}}$ are independent, given that $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are independent random variables.

To prove the independence, consider $\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})h_2(\rho_{\mathbf{X},\mathbf{Z}})]$, where h_1 and h_2 are arbitrary bounded functions. Since $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are independent,

$$\begin{aligned}\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) \cdot & h_2(\rho_{\mathbf{X},\mathbf{Z}})] \\ &= \mathbb{E}\left[\, \mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) \cdot h_2(\rho_{\mathbf{X},\mathbf{Z}}) | \mathbf{X}]\,\right] \\ &= \mathbb{E}\left[\, \mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) | \mathbf{X}] \cdot \mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}}) | \mathbf{X}]\,\right].\end{aligned}$$

Given $\mathbf{X}$, the probability density function of $\rho_{\mathbf{X},\mathbf{Y}}$ is independent of $\mathbf{X}$. Thus, $\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})|\mathbf{X}] = \int_{-1}^{1} h_1(\rho_{\mathbf{X},\mathbf{Y}}) g(\rho_{\mathbf{X},\mathbf{Y}})\ \mathrm{d}\rho = \mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})]$, and similarly $\mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}})|\mathbf{X}] = \mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}})]$. It gives,

$$\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) \cdot h_2(\rho_{\mathbf{X},\mathbf{Z}})] = \mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})] \cdot \mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}})],$$

This concludes that $\rho_{\mathbf{X},\mathbf{Y}}$ and $\rho_{\mathbf{X},\mathbf{Z}}$ are also independent. ■

Recall that the derivation of Eq. A.15 uses the generalized central limit theorem which requires the limiting condition $q \to \infty$. Therefore it is important to check how the dimensionality q effects the quality of the prediction.

Fig. 2 displays the empirical distribution of ρ, that is $g(\rho) = \sum\limits_{1 \le i < j \le n} \delta_{\rho_{ij}}$, and the theoretical prediction in Eq. A.15 for various dimensions q. For the simulation, $n = 10000$ random vectors are drawn independently from $\mathcal{C}(0,1)$. We use the leading orders of the Laurent series of Eq. A.15 to represent the theoretical predictions.

It can be seen that for a sufficiently high-dimensional space, say $q = 2000$, the theoretical prediction fits the simulation very well. Moreover, the pairwise angles among Cauchy random vectors converge to $\frac{\pi}{2}$ as the dimensionality increases.

It implies that in high-dimensional spaces the distributional uniformity of normalized Cauchy random vectors could be tenable. We explain this in an intuitive way. According to Remark A 2, each element in the normalized variable converges independently in distribution to a Dirac δ-function, which can be constructed as the limit of a sequence of zero-centered normal distribution

$$f_{X_i'}(x_i') = \frac{1}{a\sqrt{\pi}} \mathrm{e}^{-\frac{x_i'^2}{a^2}} \quad \text{for} \quad a \to 0^+.$$

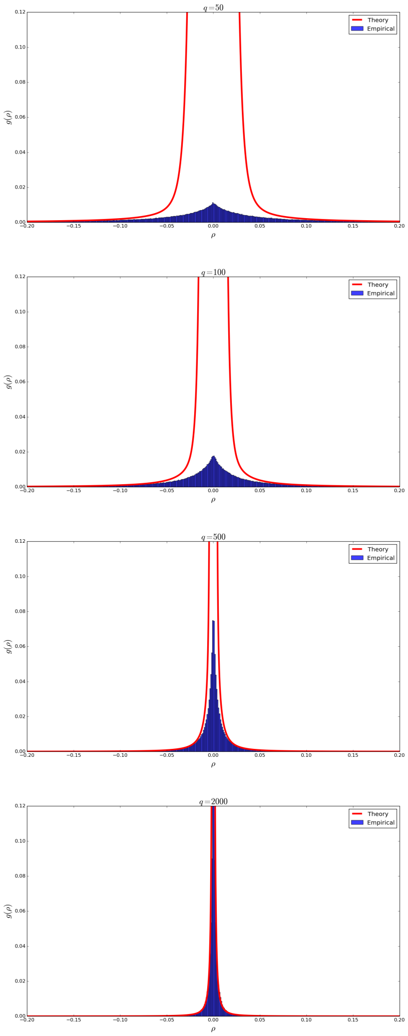


Figure 2: Comparisons between empirical distributions and theoretical predictions of ρ_{C} for various dimensions, $q = 50, 100, 500, 2000$.

Thus, following Lemma A 3, the density function of $f_{\mathbf{X}'}(x'_1, \cdots, x'_q)$ can be approximated by

$$f_{\mathbf{X}'}(x'_1, \cdots, x'_q) = \left(\frac{1}{a\sqrt{\pi}}\right)^q \mathrm{e}^{-\frac{\mathbf{x}'^{\intercal}\mathbf{x}'}{a^2}} \quad \text{for} \quad a \to 0^+.$$

This joint distribution is invariant under an arbitrary orthogonal rotation. Thus, it is a spherical distribution, as well as a uniform distribution on S^{q-1}. A rigorous proof of this result is still necessary. However, it is beyond the scope of this work.

A.3 DERIVATION OF COROLLARY 3

Corollary 3. *Consider a set of independent q-dimensional Cauchy random vectors which are pairwise ϵ-orthogonal with probability $1 - \nu$. Then the number of such Cauchy random vectors is bounded by*

$$N \leq \sqrt{\frac{\pi\epsilon q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}. \qquad \text{(A.16)}$$

Proof. The derivation of this bound is similar to that of Corollary 2. The probability, that two random vectors whose elements are independently and identically Cauchy distributed are not ϵ-orthogonal, is bounded from above by

$$\Pr(|\rho| \geq \epsilon) = 2\int_{\epsilon}^{1} \frac{2}{\pi q \rho^2}\,\mathrm{d}\rho < \frac{4}{\pi q}\frac{1}{\epsilon},$$

where only the leading order Laurent expansion of Eq. A.15 is considered. Then the quantity $\mathcal{P}(\epsilon, N)$ can be estimated as follows,

$$\mathcal{P}(\epsilon, N) := \prod_{k=1}^{N-1}\left[1 - k\Pr(|\rho| \geq \epsilon)\right] > \prod_{k=1}^{N-1}(1 - k\frac{4}{\pi\epsilon q})$$
$$> (1 - N\frac{4}{\pi\epsilon q})^N \sim \mathrm{e}^{-N^2\frac{4}{\pi\epsilon q}},$$

for sufficiently large N, and $q \to \infty$, with $N\frac{4}{\pi\epsilon q} < 1$. If we require $\mathcal{P}(\epsilon, N) \geq 1 - \nu$, then the number of pairwise ϵ-orthogonal i.i.d. Cauchy random vectors is upper bounded by

$$\mathrm{e}^{-N^2\frac{4}{\pi\epsilon q}} \geq 1 - \nu \;\Rightarrow\; N \leq \sqrt{\frac{\pi\epsilon q}{4}}\left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}$$

■

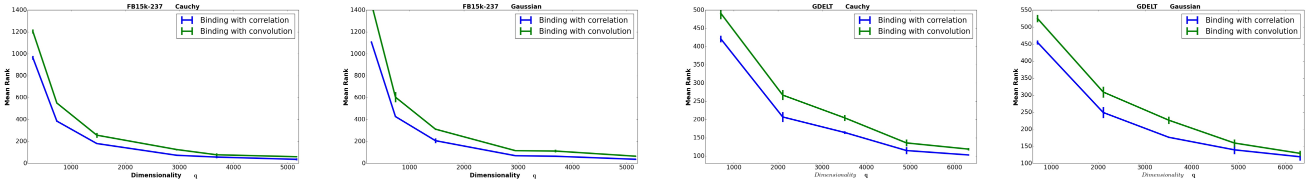
A.4 BINDING WITH CORRELATION OR CONVOLUTION

The filtered mean rank scores with different binding operations are compared in Fig. 3.

Now we give a heuristic explanation. For the sake of simplicity, consider only one semantic triple (s, p, o). For the binding with circular correlation the holistic representations are given by $\mathbf{h}_s^{\mathrm{corr}} = \mathbf{r}_p \star \mathbf{r}_o + \xi\mathbf{r}_s$, $\mathbf{h}_p^{\mathrm{corr}} = \mathbf{r}_s \star \mathbf{r}_o + \xi\mathbf{r}_p$, and $\mathbf{h}_o^{\mathrm{corr}} = \mathbf{r}_p \star \mathbf{r}_s + \xi\mathbf{r}_o$.

On the other hand, for the binding with convolution, the holistic representations given by: $\mathbf{h}_s^{\mathrm{conv}} = \mathbf{r}_p * \mathbf{r}_o + \xi\mathbf{r}_s$, $\mathbf{h}_p^{\mathrm{conv}} = \mathbf{r}_s * \mathbf{r}_o + \xi\mathbf{r}_p$, and $\mathbf{h}_o^{\mathrm{conv}} = \mathbf{r}_p * \mathbf{r}_s + \xi\mathbf{r}_o$.

Suppose that the subject needs to be retrieved and recalled using holistic representations only. To quantify the retrieval quality, a similarity $s^{\mathrm{corr/conv}}$ is introduced for different binding operators. In particular, for binding with circular correlation $s^{\mathrm{corr}} := \mathbf{h}_s^{\mathrm{corr}\,\intercal}(\mathbf{h}_p^{\mathrm{corr}} * \mathbf{h}_o^{\mathrm{corr}})$,



(a) (b) (c) (d)

Figure 3: Comparison of the filtered MR scores for binding with convolution and binding with correlation (a) for FB15k-237 with Cauchy initialization, (b) for FB15k-237 with Gaussian initialization, (c) for GDELT dataset with Cauchy initialization, (d) for GDELT with Gaussian initialization

while for binding with circular convolution $s^{\text{conv}} := \mathbf{h}_s^{\text{conv}\,\intercal}(\mathbf{h}_p^{\text{conv}} \star \mathbf{h}_o^{\text{conv}})$.

Before any further derivations, recall that circular correlation can be computed in log-linear complexity via

$$\mathbf{a} \star \mathbf{b} = \mathcal{F}^{-1}\left(\overline{\mathcal{F}(\mathbf{a})} \odot \mathcal{F}(\mathbf{b})\right),$$

where $\mathcal{F}(\cdot)$ denotes the *fast Fourier transform* and $\mathcal{F}^{-1}(\cdot)$ its inverse, and the bar denotes the complex conjugate of a complex-valued vector. Moreover, circular convolution can also be computed via *fast Fourier transforms*

$$\mathbf{a} * \mathbf{b} = \mathcal{F}^{-1}\left(\mathcal{F}(\mathbf{a}) \odot \mathcal{F}(\mathbf{b})\right).$$

First we compute the similarity s^{corr}

$$\begin{aligned} s^{\text{corr}} &= \mathbf{h}_s^{\text{corr}\,\intercal}(\mathbf{h}_p^{\text{corr}} * \mathbf{h}_o^{\text{corr}}) \\ &= (\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal[(\mathbf{r}_s \star \mathbf{r}_o + \xi \mathbf{r}_p) * (\mathbf{r}_p \star \mathbf{r}_s + \xi \mathbf{r}_o)] \\ &= (\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal[\underbrace{(\mathbf{r}_s \star \mathbf{r}_o) * (\mathbf{r}_p \star \mathbf{r}_s)}_{\textcircled{1}} + \\ &\quad \xi \underbrace{(\mathbf{r}_s \star \mathbf{r}_o) * \mathbf{r}_o}_{\textcircled{2}} + \xi \underbrace{\mathbf{r}_p * (\mathbf{r}_p \star \mathbf{r}_s)}_{\textcircled{3}} + \xi^2 \mathbf{r}_p * \mathbf{r}_o]. \end{aligned}$$

Using that

$$\begin{aligned} &\textcircled{1} = \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \mathcal{F}(\mathbf{r}_o) \odot \overline{\mathcal{F}(\mathbf{r}_p)} \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_p \star \mathbf{r}_o, \\ &\textcircled{2} = \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \mathcal{F}(\mathbf{r}_o) \odot \mathcal{F}(\mathbf{r}_o)\right] = \text{Noise}, \\ &\textcircled{3} = \mathcal{F}^{-1}\left[\mathcal{F}(\mathbf{r}_p) \odot \overline{\mathcal{F}(\mathbf{r}_p)} \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_s, \end{aligned}$$

yields

$$\begin{aligned} s^{\text{corr}} &\approx (\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal[\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s + \text{Noise}] \\ &\approx (1 + \xi^2) + \text{Noise}. \end{aligned}$$

The similarity s^{conv} can be computed in a similar way,

$$\begin{aligned} s^{\text{conv}} &= \mathbf{h}_s^{\text{conv}\,\intercal}(\mathbf{h}_p^{\text{conv}} \star \mathbf{h}_o^{\text{conv}}) \\ &= (\mathbf{r}_p * \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal[(\mathbf{r}_s * \mathbf{r}_o + \xi \mathbf{r}_p) \star (\mathbf{r}_p * \mathbf{r}_s + \xi \mathbf{r}_o)] \\ &= (\mathbf{r}_p * \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal[\underbrace{(\mathbf{r}_s * \mathbf{r}_o) \star (\mathbf{r}_p * \mathbf{r}_s)}_{\textcircled{1}} + \\ &\quad \xi \underbrace{(\mathbf{r}_s * \mathbf{r}_o) \star \mathbf{r}_o}_{\textcircled{2}} + \xi \underbrace{\mathbf{r}_p \star (\mathbf{r}_p * \mathbf{r}_s)}_{\textcircled{3}} + \xi^2 \mathbf{r}_p \star \mathbf{r}_o]. \end{aligned}$$

Moreover, using that

$$\begin{aligned} &\textcircled{1} = \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \overline{\mathcal{F}(\mathbf{r}_o)} \odot \mathcal{F}(\mathbf{r}_p) \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_o \star \mathbf{r}_p, \\ &\textcircled{2} = \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \overline{\mathcal{F}(\mathbf{r}_o)} \odot \mathcal{F}(\mathbf{r}_o)\right] \approx \mathbf{r}_s, \\ &\textcircled{3} = \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_p)} \odot \mathcal{F}(\mathbf{r}_p) \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_s, \end{aligned}$$

leads to

$$\begin{aligned} s^{\text{conv}} &\approx (\mathbf{r}_p * \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal[\mathbf{r}_o \star \mathbf{r}_p + 2\xi \mathbf{r}_s + \text{Noise}] \\ &\approx 2\xi^2 + \text{Noise}. \end{aligned}$$

The optimal hyper-parameter requires $\xi < 1$ which in turn yields $s^{\text{corr}} > s^{\text{conv}}$. From the derivation of s^{corr}, we have that the subject-object association pair stored in $\mathbf{h}_p^{\text{corr}}$ contributes the most in $s^{\text{corr}} \approx 1 + \xi^2$ via the term $\textcircled{1}$.

A.5 APPROXIMATION OF $\rho_{\mathbf{r}'_{\mathrm{o}},\mathbf{h}_{\mathrm{o}}}$

Here we provide a heuristic study on the relations between hyper-parameter ξ, $\lambda_{\mathrm{G/C}}$, and the average number of association pairs N_{a}. Recall that ξ was introduced for holistic representations, and $\lambda_{\mathrm{G/C}}$ is defined as $\lambda_{\mathrm{G/C}} := \mathbb{E}[|\rho_{\mathrm{G/C}}|]$.

Consider a subject s. The predicate-object pair (p, o) is stored in the holistic representation $\mathbf{h}_{\mathrm{s}}$ along with the other $N_{\mathrm{a}} - 1$ pairs. This means

$$\mathbf{h}_{\mathrm{s}} = \xi N_{\mathrm{a}} \mathbf{r}_{\mathrm{s}} + \mathbf{r}_{\mathrm{p}} \star \mathbf{r}_{\mathrm{o}} + \sum_{i=2}^{N_{\mathrm{a}}} \mathbf{r}_{p_i} \star \mathbf{r}_{o_i}.$$

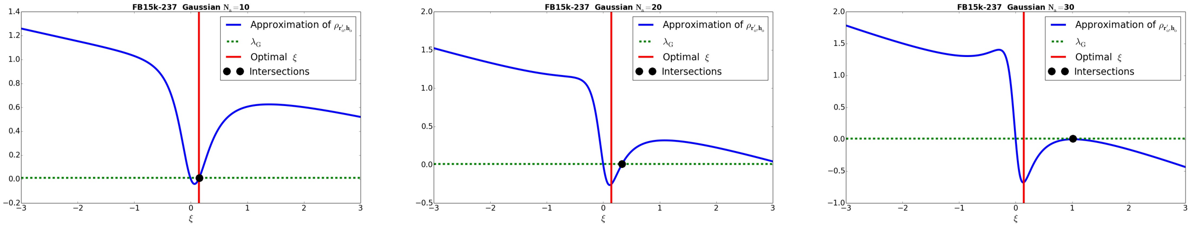


Figure 4: Approximations of $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}}(\xi, N_\mathrm{a})$ in the case of Gaussian holistic representations with (a): $N_\mathrm{a} = 10$ (b): $N_\mathrm{a} = 20$ (c): $N_\mathrm{a} = 30$. We use the experiment setting with dimnsionality $q = 5200$, $\lambda_\mathrm{G} = 0.0111$, and optimal $\xi = 0.14$.

Suppose that we aim to identify the object in the triple $(\mathrm{s}, \mathrm{p}, \cdot)$ via $\mathbf{h}_\mathrm{s}$ and $\mathbf{h}_\mathrm{p}$, where $\mathbf{h}_\mathrm{p}$ is the holistic representation for the predicate p. We further assume that up to N_a subject-object pairs can be stored in $\mathbf{h}_\mathrm{p}$ having high enough fidelity, then

$$\mathbf{h}_\mathrm{p} = \xi N_\mathrm{a}\mathbf{r}_\mathrm{p} + \sum_{k=1}^{N_\mathrm{a}} \mathbf{r}_{s_k} \star \mathbf{r}_{o_k}.$$

To retrieve the object o, the decoding via circular convolution is obtained as follows

$$\begin{aligned}
&\mathbf{r}'_\mathrm{o} = \mathbf{h}_\mathrm{p} * \mathbf{h}_\mathrm{s} \\
&\approx \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \xi^2 N_\mathrm{a}^2(\mathbf{r}_\mathrm{p} * \mathbf{r}_\mathrm{s}) + \xi N_\mathrm{a} \sum_{i=2}^{N_\mathrm{a}} [\mathbf{r}_\mathrm{p} * (\mathbf{r}_{p_i} \star \mathbf{r}_{o_i})] \\
&+ \xi N_\mathrm{a} \sum_{k=1}^{N_\mathrm{a}} [(\mathbf{r}_{s_k} \star \mathbf{r}_{o_k}) * \mathbf{r}_\mathrm{s}] + \sum_{k=1}^{N_\mathrm{a}} [(\mathbf{r}_{s_k} \star \mathbf{r}_{o_k}) * (\mathbf{r}_\mathrm{p} \star \mathbf{r}_\mathrm{o})] \\
&+ \sum_{k=1, i=2}^{N_\mathrm{a}, N_\mathrm{a}} [(\mathbf{r}_{s_k} \star \mathbf{r}_{o_k}) * (\mathbf{r}_{p_i} \star \mathbf{r}_{o_i})] \\
&= \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \xi^2 N_\mathrm{a}^2 \mathbf{b}_1 + \xi N_\mathrm{a} \sum_{i=2}^{N_\mathrm{a}} \mathbf{b}_i + \xi N_\mathrm{a} \sum_{k=1}^{N_\mathrm{a}} \mathbf{c}_k \\
&+ \sum_{k=1}^{N_\mathrm{a}} \mathbf{d}_k + \sum_{k=1, i=2}^{N_\mathrm{a}, N_\mathrm{a}} \mathbf{e}_{ki},
\end{aligned}$$

where $\mathbf{b}_i$, $\mathbf{c}_k$, $\mathbf{d}_k$, and $\mathbf{e}_{ki}$ with $i, k = 1, \cdots, N_\mathrm{a}$ are approximately normalized Gaussian/Cauchy random vectors. This is due to the fact that in high-dimensional spaces both circular correlation and circular convolution of two normalized Gaussian/Cauchy random vectors is approximately a normalized Gaussian/Cauchy random vectors.

After decoding with circular convolutions, the decoded noisy version of the object needs to be recalled with $\mathbf{h}_\mathrm{o}$ which is the holistic representation of o. As before, N_a predicate-subject association pairs are assumed to be stored in the holistic representation of o, with

$$\mathbf{h}_\mathrm{o} = \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \sum_{j=1}^{N_\mathrm{a}} \mathbf{r}_{p_j} \star \mathbf{r}_{s_j} = \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \sum_{j=1}^{N_\mathrm{a}} \mathbf{f}_j,$$

where $\mathbf{f}_j$, $j = 1, \cdots, N_\mathrm{a}$ are approximately normalized Gaussian/Cauchy random vectors.

In order to recall the object successfully, the angle between $\mathbf{r}'_\mathrm{o}$ and $\mathbf{h_o}$ should be smaller than the expected absolute angle between two arbitrary vectors, namely $\theta_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}} < \mathbb{E}[|\theta_{\mathrm{G/C}}|]$. Given the definition of λ, equivalently, it requires $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}} > \lambda_{\mathrm{G/C}}$.

Now we turn to approximate the numerator of $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}}$, that is $\mathbf{r}'^{\intercal}_\mathrm{o}\mathbf{h}_\mathrm{o}$. Recall that, in general, the expectation of the dot product of two normalized, independent random vectors equals 0 due to the symmetry of the density function $g(\rho_{\mathrm{G/C}})$. Therefore, in the following approximation we only consider noisy terms which are directly related to $\mathbf{r}_\mathrm{o}$ as adverse effects to a successful retrieval and treat other terms as white noisy with zero expectation. This yields,

$$\begin{aligned}
&\mathbf{r}'^{\intercal}_\mathrm{o}\mathbf{h}_\mathrm{o} \\
&\approx \xi^2 N_\mathrm{a}^2 + \xi N_\mathrm{a} \sum_{j=1}^{N_\mathrm{a}} (\mathbf{r}_\mathrm{o}^\intercal \mathbf{f}_j) + \xi^3 N_\mathrm{a}^3 (\mathbf{r}_\mathrm{o}^\intercal \mathbf{b}_1) + \xi^2 N_\mathrm{a}^2 \sum_{i=2}^{N_\mathrm{a}} (\mathbf{r}_\mathrm{o}^\intercal \mathbf{b}_i) \\
&\quad + \xi^2 N_\mathrm{a}^2 \sum_{k=1}^{N_\mathrm{a}} (\mathbf{r}_\mathrm{o}^\intercal \mathbf{c}_k) + \xi N_\mathrm{a} \sum_{k=1}^{N_\mathrm{a}} (\mathbf{r}_\mathrm{o}^\intercal \mathbf{d}_k) + \xi N_\mathrm{a} \sum_{k=1, i=2}^{N_\mathrm{a}, N_\mathrm{a}} (\mathbf{r}_\mathrm{o}^\intercal \mathbf{e}_{ki}) \\
&> \xi^2 N_\mathrm{a}^2 - (\xi N_\mathrm{a}^2 + \xi^3 N_\mathrm{a}^3 + \xi^2 N_\mathrm{a}^2 (N_\mathrm{a} - 1) + \xi^2 N_\mathrm{a}^3 \\
&\quad + \xi N_\mathrm{a}^2 + \xi N_\mathrm{a}^2 (N_\mathrm{a} - 1))\lambda_{\mathrm{G/C}} \\
&= \xi^2 N_\mathrm{a}^2 - (\xi^3 N_\mathrm{a}^3 + 2\xi^2 N_\mathrm{a}^3 - \xi^2 N_\mathrm{a}^2 + \xi N_\mathrm{a}^2 + \xi N_\mathrm{a}^3)\lambda_{\mathrm{G/C}}.
\end{aligned}$$

Furthermore, the denominator of $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}}$ can be approximated in the same way. More concretely, we have

$$\begin{aligned}
||\mathbf{r}'_\mathrm{o}|| \cdot ||\mathbf{h}_\mathrm{o}|| < \xi^2 N_\mathrm{a}^2 + N_\mathrm{a} + 2\xi N_\mathrm{a}^2 \lambda_{\mathrm{G/C}} \\
+ N_\mathrm{a}(N_\mathrm{a} - 1)\lambda_{\mathrm{G/C}}.
\end{aligned}$$

Combining these results, a sufficient condition to retrieve the object correctly is given by

$$\rho_{\mathbf{r}'_{\mathrm{o}},\mathbf{h}_{\mathrm{o}}} > \\ \frac{\xi^2 N_{\mathrm{a}}^2 - (\xi^3 N_{\mathrm{a}}^3 + 2\xi^2 N_{\mathrm{a}}^3 - \xi^2 N_{\mathrm{a}}^2 + \xi N_{\mathrm{a}}^2 + \xi N_{\mathrm{a}}^3)\lambda_{\mathrm{G/C}}}{\xi^2 N_{\mathrm{a}}^2 + N_{\mathrm{a}} + 2\xi N_{\mathrm{a}}^2 \lambda_{\mathrm{G/C}} + N_{\mathrm{a}}(N_{\mathrm{a}}-1)\lambda_{\mathrm{G/C}}} \\ > \lambda_{\mathrm{G/C}}. \tag{A.17}$$

Consider the experimental setting for the memorization task on the FB15k-237 dataset: The dimensionality of the holistic representations is $q = 5200$, $\lambda_{\mathrm{G}}(q = 5200) = 0.0111$, and $\lambda_{\mathrm{C}}(q = 5200) = 0.00204$. Fig. 4 displays the above approximation of $\rho_{\mathbf{r}'_{\mathrm{o}},\mathbf{h}_{\mathrm{o}}}(\xi, N_{\mathrm{a}})$ for Gaussian initializations.

After performing grid search, the optimal ξ is found to be close to the intersection of the curve $\rho_{\mathbf{r}'_{\mathrm{o}},\mathbf{h}_{\mathrm{o}}}(\xi, N_{\mathrm{a}} = 10)$ and the threshold λ_{G}. However, for $N_{\mathrm{a}} > 30$, no intersection points on $\xi > 0$ exists. This explains why Gaussian holistic representations have lower memory capacity compared to Cauchy holistic representations.

More comparisons between Gaussian and Cauchy initializations can be found in Fig. 5.

A.6 HOLISTIC ENCODING ALGORITHM

Algorithm 1 Holistic Encoding

```
Require: hyper-parameter $\xi$
for $i = 1, \cdots, N_e$ do
    Draw $\tilde{\mathbf{r}}_{e_i}^{\mathrm{G/C}}$ from Gaussian or Cauchy
    $\mathbf{r}_{e_i}^{\mathrm{G/C}} \leftarrow \mathbf{Norm}(\tilde{\mathbf{r}}_{e_i}^{\mathrm{G/C}})$
for $i = 1, \cdots, N_p$ do
    Draw $\tilde{\mathbf{r}}_{p_i}^{\mathrm{G/C}}$ from Gaussian or Cauchy
    $\mathbf{r}_{p_i}^{\mathrm{G/C}} \leftarrow \mathbf{Norm}(\tilde{\mathbf{r}}_{p_i}^{\mathrm{G/C}})$
for $i = 1, \cdots, N_e$ do
    Extract $\in \mathcal{S}^s(e_i), \mathcal{S}^o(e_i)$ from Database
    $\mathbf{h}_{e_i}^s \leftarrow \sum_{(p,o)\in\mathcal{S}^s(e_i)} [\mathbf{Norm}(\mathbf{r}_p \star \mathbf{r}_o) + \xi \mathbf{r}_{e_i}]$
    $\mathbf{h}_{e_i}^o \leftarrow \sum_{(s,p)\in\mathcal{S}^o(e_i)} [\mathbf{Norm}(\mathbf{r}_p \star \mathbf{r}_s) + \xi \mathbf{r}_{e_i}]$
    $\mathbf{h}_{e_i} \leftarrow \mathbf{h}_{e_i}^s + \mathbf{h}_{e_i}^o$
for $i = 1, \cdots, N_p$ do
    Extract $\mathcal{S}(p_i)$ from Database
    $\mathbf{h}_{p_i} \leftarrow \sum_{(s,o)\in\mathcal{S}(p_i)} [\mathbf{Norm}(\mathbf{r}_s \star \mathbf{r}_o) + \xi \mathbf{r}_{p_i}]$
```

Remark:

Normalizing initial random vectors can assist the analysis of memory capacities via different sampling schemes. For example, for the derivation of retrieval condition Eq. A.17 we heavily relay on the fact that the dot product of two random vectors - say $\mathbf{r}_i \cdot \mathbf{r}_j$, where $\mathbf{r}_i$ and $\mathbf{r}_j$ are randomly sampled and normalized - is just ρ_{ij}. In the memorization task, since triples are recalled by comparing the angles (a.k.a cosine similarity) between decoded noisy vector and all other holistic vectors, normalization does not effect the recall scores.

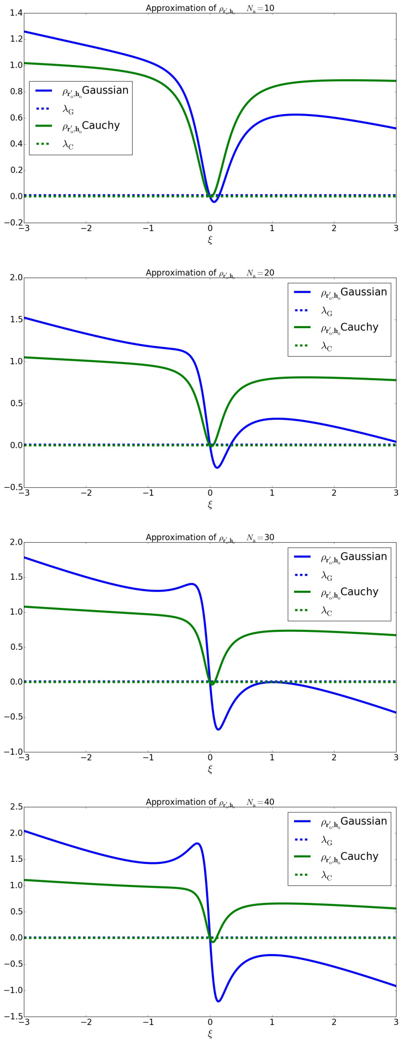


Figure 5: Comparison of $\rho_{\mathbf{r}'_{\mathrm{o}},\mathbf{h}_{\mathrm{o}}}(\xi, N_{\mathrm{a}})$ for Gaussian (blue) and Cauchy (green) holistic representations with (a): $N_{\mathrm{a}} = 10$ (b): $N_{\mathrm{a}} = 20$ (c): $N_{\mathrm{a}} = 30$ (d): $N_{\mathrm{a}} = 40$.

A.7 NOTATIONS

In Table 1 and Table 2, we summary important notations introduced in Section 3 and 4, respectively.

A.8 FURTHER EXPERIMENTAL DETAILS

After searching for the optimal hyper-parameter ξ for holistic encoding, holistic representations with superior

Table 1: Notations for ϵ-orthogonality

Symbol	Meaning
$\mathbf{X}$	q-dimensional random variable with elements drawn from Gaussian or Cauchy distribution
Θ_{ij}	Angle between two random variables $\mathbf{X}_i$ and $\mathbf{X}_j$
ρ_{ij}	Cosine of the angle between random variables $\mathbf{X}_i$ and $\mathbf{X}_j$
$g(\rho_{\mathrm{G}})$	Asymptotic density function of ρ_{ij} given an ensemble of Gaussian random variables $\mathbf{X}_i$, $i = 1, \cdots, n$, with $n \to \infty$
$g(\rho_{\mathrm{C}})$	Asymptotic density function of ρ_{ij} given an ensemble of Cauchy random variables $\mathbf{X}_i$, $i = 1, \cdots, n$, with $n \to \infty$
λ_{G}	Expectation value of $\lvert\rho_{\mathrm{G}}\rvert$
λ_{C}	Expectation value of $\lvert\rho_{\mathrm{C}}\rvert$

memory capacity will be fixed and applied to the next inference tasks.

The architecture is a simple 2-layered fully-connected neural network, which map high-dimensional holistic representations ($q = 3600$) of subjects, predicates, and objects to low-dimensional ($h_2 = 256$) representations, separately. We choose ReLU as the activation function for faster training, and batch normalization after the hidden-layer for regularization. In order to reduce the number of trainable parameters, the network has a bottleneck structure with the dimensionality of the hidden-layer $h_1 = 64$. The extracted low-dimensional features are then combined via tri-linear dot-product, similar to DistMult.

In summary, given a triple $(\mathrm{s}, \mathrm{p}, \mathrm{o})$ the scoring function η_{spo} takes the following form:

$$\begin{aligned}\eta_{\mathrm{spo}} =& \langle \mathrm{BN}(\mathrm{ReLU}(\mathbf{h}_{\mathrm{s}}\mathbf{W}_1^e))\mathbf{W}_2^e,\\ & \mathrm{BN}(\mathrm{ReLU}(\mathbf{h}_{\mathrm{p}}\mathbf{W}_1^p))\mathbf{W}_2^p,\\ & \mathrm{BN}(\mathrm{ReLU}(\mathbf{h}_{\mathrm{o}}\mathbf{W}_1^e))\mathbf{W}_2^e\rangle,\end{aligned}$$

where $\mathbf{h}_{\mathrm{s}}$, $\mathbf{h}_{\mathrm{s}}$ are the holistic representations for the subject s and object o; $\mathbf{h}_{\mathrm{p}}$ is the holistic representation for the predicate p. Note that there are two separate networks for extracting low-dimensional features of entities and predicates, respectively. In particular, $\mathbf{W}_1^e \in \mathbb{R}^{q \times h_1}$ and $\mathbf{W}_2^e \in \mathbb{R}^{h_1 \times h_2}$ are shared weights for entities, including subjects and objects; $\mathbf{W}_1^p \in \mathbb{R}^{q \times h_1}$ and $\mathbf{W}_2^p \in \mathbb{R}^{h_1 \times h_2}$ are shared weights for predicates.

For training the model, we minimize the following binary cross-entropy loss with l_2 regularization:

Table 2: Notations for holistic representations

Symbol	Meaning
$*$	Circular convolution
$\star$	Circular correlation
Norm	Normalization operator, $\mathrm{Norm}(\mathbf{r}) := \frac{\mathbf{r}}{\lvert\lvert\mathbf{r}\rvert\rvert}$
N_e	Number of entities in the KG
N_p	Number of predicates in the KG
N_a	Average number of association pairs encoded in holistic representations of entities
$\mathbf{r}_{e_i}^{\mathrm{G/C}}$	Random initialization of entity e_i with elements drawn from Gaussian or Cauchy distribution
$\mathbf{r}_{p_i}^{\mathrm{G/C}}$	Random initialization of predicate p_i with elements drawn from Gaussian or Cauchy distribution
$\mathbf{h}_{e_i}^{s}$	Holistic representation of entity e_i as subject
$\mathbf{h}_{e_i}^{o}$	Holistic representation of entity e_i as object
$\mathbf{h}_{e_i}$	Overall holistic representation of entity e_i
$\mathbf{h}_{p_i}$	Holistic representation of predicate p_i
ξ	Hyper-parameter for holistic encoding

$$\begin{aligned}\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}(y_i \cdot \log(\sigma(\eta_{x_i})) +\\ (1 - y_i) \cdot \log(1 - \sigma(\eta_{x_i}))) + \lambda||\mathcal{A}||_2^2,\end{aligned}$$

where the label vector y_i has dimension $\{0, 1\}^{1 \times N}$ for 1-N scoring to accelerate the link prediction tasks. To be more specific, during the training given a triple $(\mathrm{s}, \mathrm{p}, \mathrm{o})$, we take the subject-predicate pair (s, p) and and rank it against all object entities $o \in \mathcal{E}$; take the predicate-object pair (p, o) and rank it against all subject entities $s \in \mathcal{E}$ simultaneously as well.

Hyper-parameters in the HolNN$_{\mathrm{G}}$ and HolNN$_{\mathrm{C}}$ are optimized via grid search with respect to the mean reciprocal rank (MRR). The ranges for grid search are as follows - learning rate $\{0.001, 0.003, 0.005\}$, $l2$ regularization parameter $\{0., 0.01, 0.05\}$, decay parameter in the batch normalization $\{0.99, 0.9, 0.8, 0.7\}$, and batch size $\{1000, 3000, 5000\}$.

References

Cai, T Tony and Tiefeng Jiang (2012). "Phase transition in limiting distributions of coherence of high-dimensional random matrices". In: *Journal of Multivariate Analysis* 107, pp. 24–39.

Gnedenko, B.V. and A.N. Kolmogorov (1954). *Limit distributions for sums of independent random variables*. Addison-Wesley. URL: `https://books.google.de/books?id=7qVyAQAACAAJ`.

Mandelbrot, Benoit (1960). "The Pareto-Levy law and the distribution of income". In: *International Economic Review* 1.2, pp. 79–106.

Muirhead, Robb J (2009). *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons.

Nolan, John (2003). *Stable distributions: models for heavy-tailed data*. Birkhauser New York.