The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings

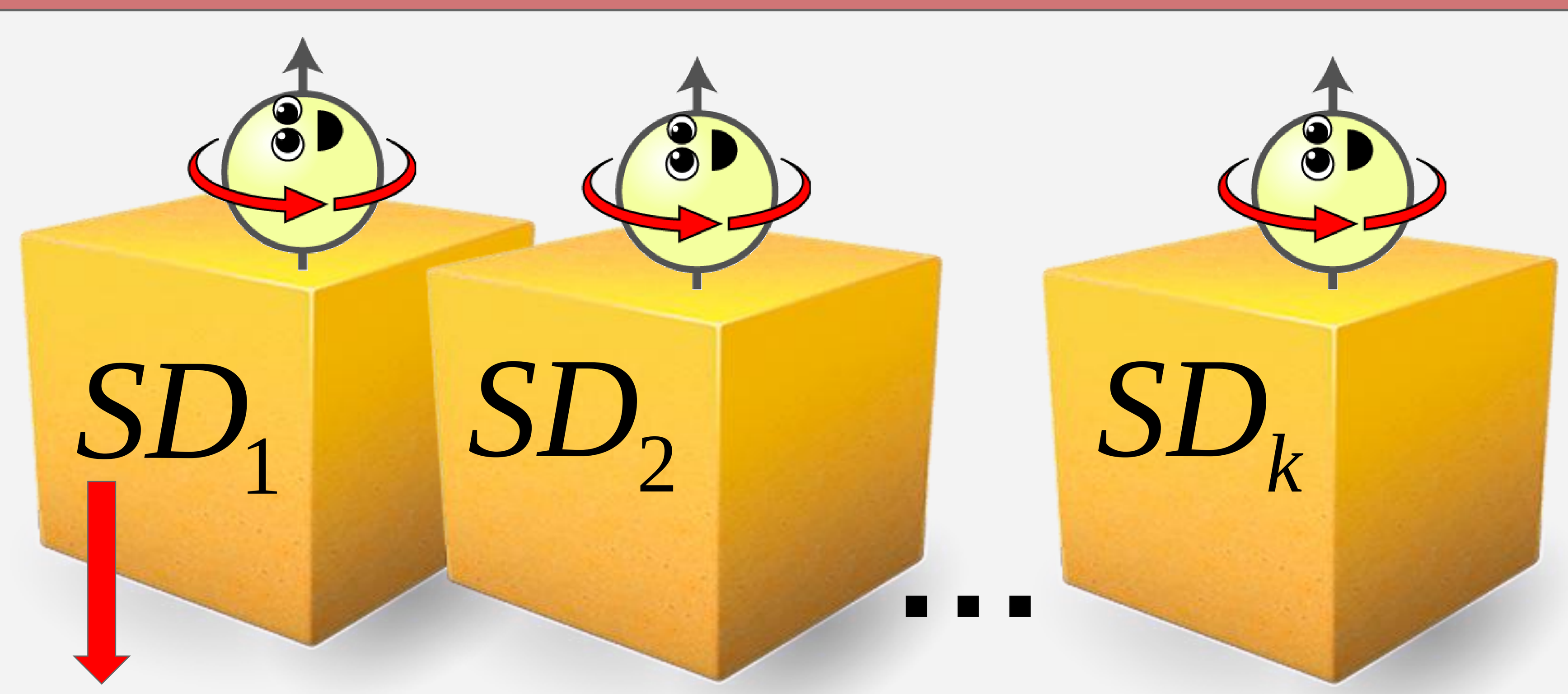
Krzysztof Choromanski, Mark Rowland, Adrian Weller

Why orthogonal random features ?

- ➢ **In practice they are very effective**
  - applied successfully in many ML settings: LSH, kernel ridge regression, RNNs and more…
- ➢ **Better time and space complexity**
  - certain discrete variants provide computational speedups and lower space complexity even though it was **not known** whether they offer also accuracy improvements
- ➢ **Improving accuracy over state-of-the-art**
  - improved accuracy proved previously **only** for the Gaussian kernel and **only** asymptotically for large n

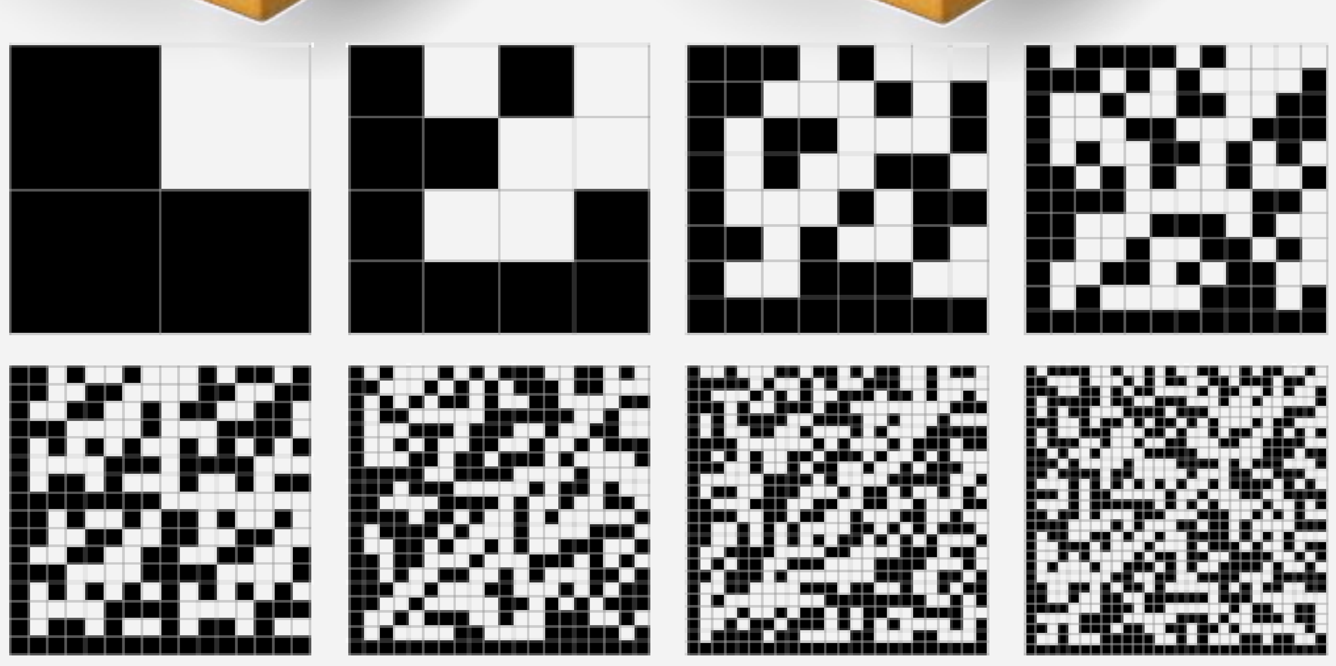
ROM Matrices

Discrete constructions



$s_{i,j} \in \{-\frac{1}{\sqrt{n}}, +\frac{1}{\sqrt{n}}\}$

- orthogonal rows
- many examples: Kronecker-product matrices, Walsh-Hadamard matrices, quadratic residue constructions



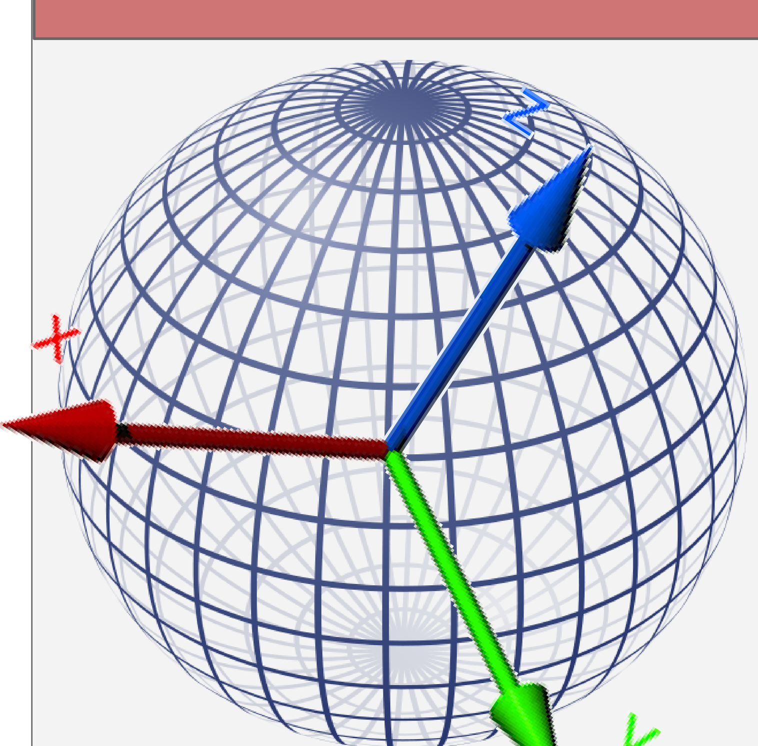
$$\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$\lambda_i \sim Unif\{-1,+1\}$

$\lambda_i \sim Unif\{-1,+1,-i,+i\}$

$\lambda_i \sim Unif(S^1) \subset \mathbb{C}$ ➡ random complex matrices

$\mathbf{G}_{\mathrm{ort}}$ Continuous constructions



- ➢ **Gaussian orthogonal matrices**
  - correspond to truly random rotations in n-dimensional spaces
  - can be easily generated from unstructured Gaussian matrices (with one-time extra cost) via the Gram-Schmidt orthogonalization

Our theoretical results

Free-lunch JLTs

Estimators:

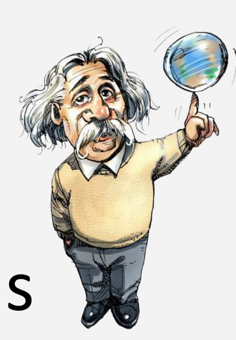
$$\widehat{K}^{\mathrm{base}}_m(\mathbf{x},\mathbf{y}) = \frac{1}{m}(\mathbf{G}\mathbf{x})^\top(\mathbf{G}\mathbf{y})$$ ➡ previous state-of-the-art

$$\widehat{K}^{\mathrm{ort}}_m(\mathbf{x},\mathbf{y}) = \frac{1}{m}(\mathbf{G}_{\mathrm{ort}}\mathbf{x})^\top(\mathbf{G}_{\mathrm{ort}}\mathbf{y})$$ our estimators

$$\widehat{K}^{(k)}_m(\mathbf{x},\mathbf{y}) = \frac{1}{m}\left(\mathbf{M}^{(k),\mathrm{sub}}_{\mathbf{S}\mathcal{R}}\mathbf{x}\right)^\top\left(\mathbf{M}^{(k),\mathrm{sub}}_{\mathbf{S}\mathcal{R}}\mathbf{y}\right)$$

$$\mathbf{M}^{(k)}_{\mathbf{S}\mathcal{R}} = \prod_{i=1}^{k}\mathbf{S}\mathbf{D}^{(\mathcal{R})}_i \rightarrow |\lambda_i| = 1$$

- ➢ **sub: different subsampling strategies for row selection to reduce dimensionality**
  - first/last m rows
  - sample uniformly at random with repetitions
  - sample uniformly at random without repetitions



THEOREM

$$\mathrm{MSE}(\widehat{K}^{(k)}_m(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left(\frac{n-m}{n-1}\right)\Big(((\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{n^r}(2(\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{n^{k-1}}\sum_{i=1}^{n}x_i^2 y_i^2\Big).$$

- ➢ **Implies that discrete ROMs provide more accurate JLT mechanisms than state-of-the-art, with better time and space complexity**
- ➢ **x2 smaller MSEs for complex discrete ROMs**
- ➢ **strictly better accuracy also for the angular kernel approximation with Gaussian orthogonal matrices**

$$K^{\mathrm{ang}}(\mathbf{x},\mathbf{y}) = 1 - \frac{2\theta_{\mathbf{x},\mathbf{y}}}{\pi}$$ ➡ kernel definition

$$\widehat{K}_m(\mathbf{x},\mathbf{y}) = \frac{1}{m}\mathrm{sgn}(\mathbf{M}\mathbf{x})^\top\mathrm{sgn}(\mathbf{M}\mathbf{y})$$ ➡ general estimator

THEOREM

$$\mathrm{MSE}(\widehat{K}^{\mathrm{ang,ort}}_m(\mathbf{x},\mathbf{y})) < \mathrm{MSE}(\widehat{K}^{\mathrm{ang,base}}_m(\mathbf{x},\mathbf{y}))$$

Experiments