Structured Adaptive and Random Spinners for Fast Machine Learning Computations

Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Francois Fagan, Cedric Gouy-Pailler, Anne Morvan, Nourhan Sakr, Tamas Sarlos, Jamal Atif

Why structured projections ?

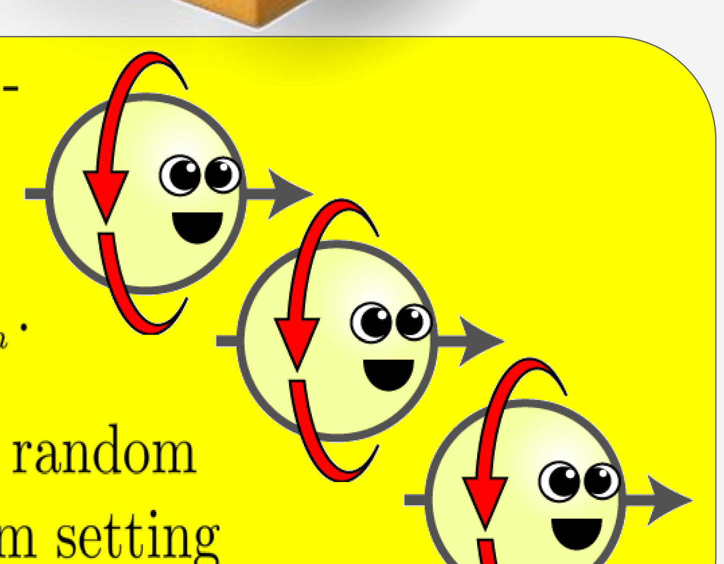
➢Replace unstructured matrices in various algorithms

- Kernel methods via random feature maps
- Neural networks and dimensionality reduction techniques
- Cross-polytope LSH methods and convex optimization
- Random projection trees
- Quasi-Monte Carlo techniques
- **Advantages:** Speed-ups. Storage compression. Almost no loss of accuracy.

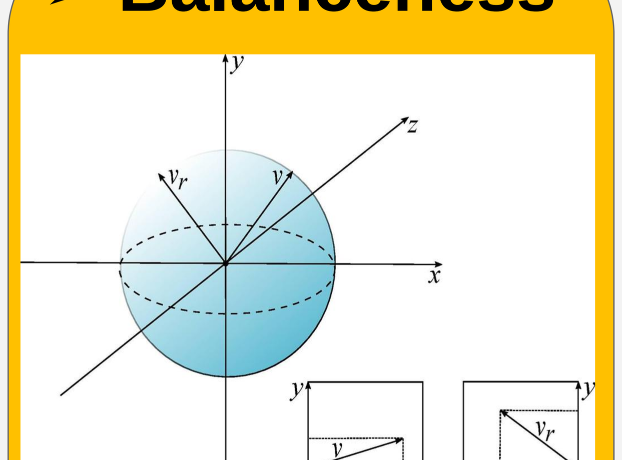
➢Provide tradeoff between required accuracy level and computational time/storage complexity

Structured Spinners

- Balanceness: $\mathbf{M}_1$ and $\mathbf{M}_2\mathbf{M}_1$ are $(\delta(n), p(n))$-balanced isometries.
- Decorrelation: $\mathbf{M}_2 = \mathbf{V}(\mathbf{W}^1, ..., \mathbf{W}^n)\mathbf{D}_{\rho_1,...,\rho_n}$.
- Budget: $\mathbf{M}_3 = \mathbf{C}(\mathbf{r}, n)$ for $\mathbf{r} \in \mathbb{R}^k$, where $\mathbf{r}$ is random Rademacher/Gaussian in the random setting and is learned in the adaptive setting.

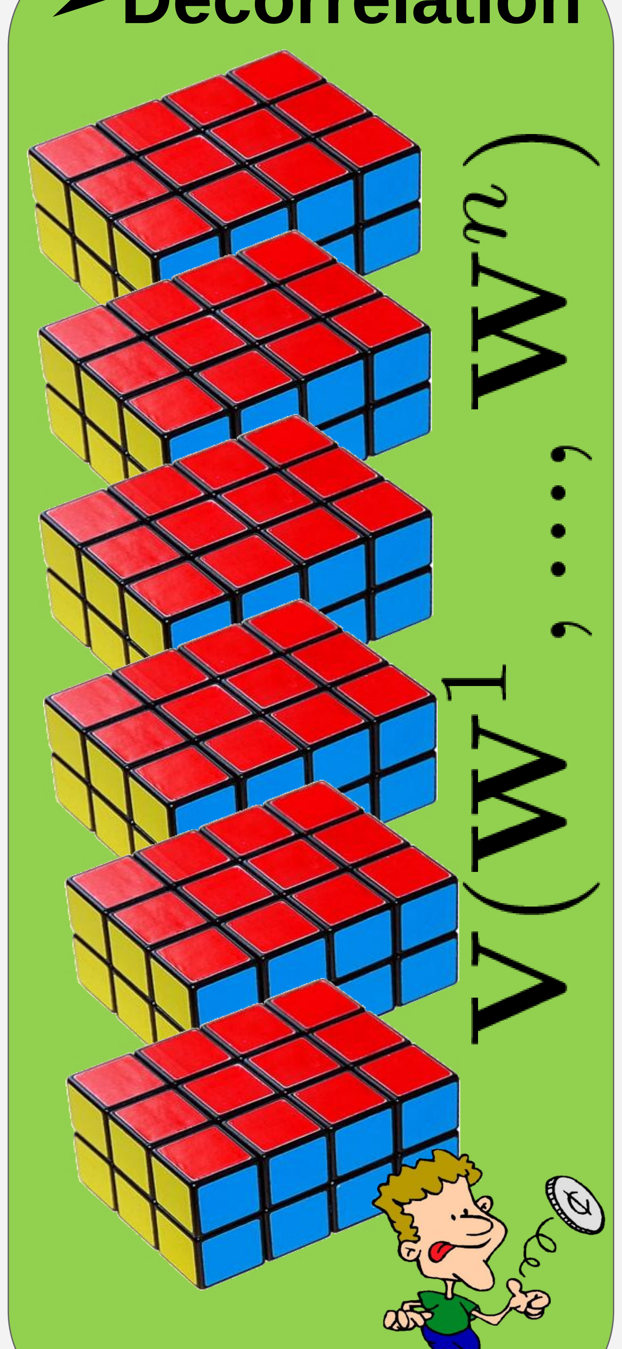


➢Balanceness

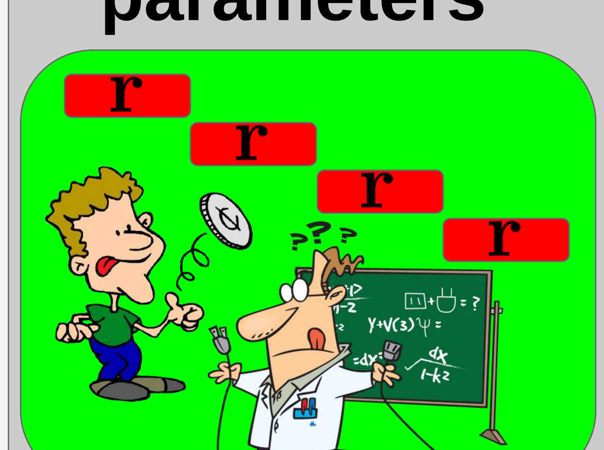


1	1	1	1	1	1	1	1
1	1	1	1	-1	-1	-1	-1
1	1	-1	-1	-1	-1	1	1
1	1	-1	-1	1	1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	-1	-1	1	-1	1	1	-1
1	-1	1	-1	-1	1	-1	1
1	-1	1	-1	1	-1	1	-1

➢Decorrelation



➢Budget or randomness/ learnable parameters



- Defines the capacity of the model
- From quadratic with no computational and storage gains to linear with **O(nlog(n))** time complexity

Smooth Sets of Matrices

Definition 1. *((Δ_F, Δ_2)-smooth sets):* *A deterministic set of matrices $\boldsymbol{W}^1, ..., \boldsymbol{W}^n \in \mathbb{R}^{k\times n}$ is (Δ_F, Δ_2)-smooth if:*

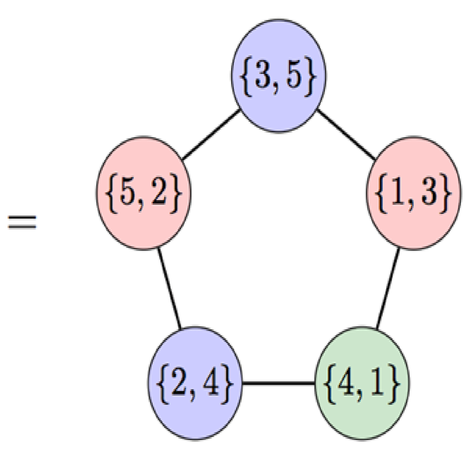
- $\|\boldsymbol{W}^i_1\|_2 = .. = \|\boldsymbol{W}^i_n\|_2$ *for* $i = 1, ..., n$, *where* $\boldsymbol{W}^i_j$ *is the* j^{th} *column of* $\boldsymbol{W}^i$,
- *for* $i \neq j$ *and* $l = 1, ..., n$ *we have:* $(\boldsymbol{W}^i_l)^T \cdot \boldsymbol{W}^j_l = 0$, → **Permutation matrices**
- $\max_{i,j} \|(\boldsymbol{W}^j)^T\boldsymbol{W}^i\|_F \leq \Delta_F$ *and* $\max_{i,j} \|(\boldsymbol{W}^j)^T\boldsymbol{W}^i\|_2 \leq \Delta_2$.

$$\mathbf{A}_{circ} = \begin{pmatrix} g_0 & g_1 & g_2 & g_3 & g_4 \\ g_4 & g_0 & g_1 & g_2 & g_3 \\ g_3 & g_4 & g_0 & g_1 & g_2 \\ g_2 & g_3 & g_4 & g_0 & g_1 \\ g_1 & g_2 & g_3 & g_4 & g_0 \end{pmatrix}$$

$$\mathbf{P}_1 = \begin{pmatrix} 1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1 \end{pmatrix} \quad \mathbf{P}_2 = \begin{pmatrix} 0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\\1&0&0&0&0 \end{pmatrix}$$

$$\mathbf{P}_3 = \begin{pmatrix} 0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\\1&0&0&0&0\\0&1&0&0&0 \end{pmatrix} \quad \mathbf{P}_4 = \begin{pmatrix} 0&0&0&1&0\\0&0&0&0&1\\1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0 \end{pmatrix}$$

$\mathcal{G}_{2,4} =$



$$\mathbf{P}_5 = \begin{pmatrix} 0&0&0&0&1\\1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0 \end{pmatrix}$$

Examples: **P-model**

- Circulant
- Toeplitz
- Hankel
- Low-displacement rank
- FastFood

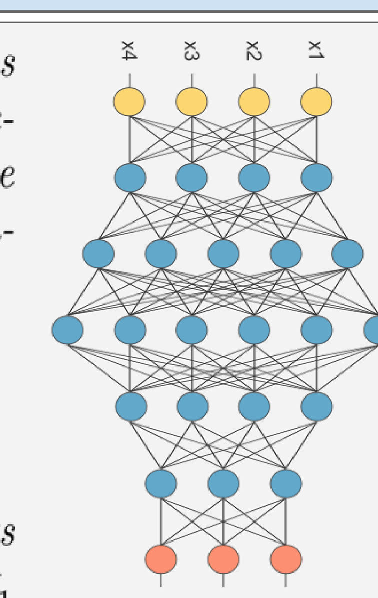
- **HDHDHD**
- **General Kroncecker products**

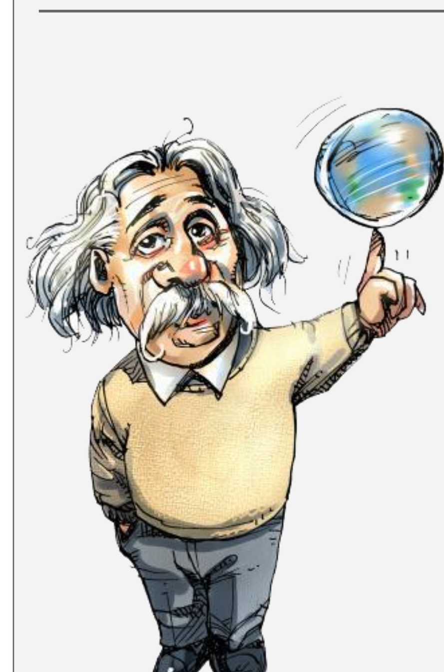
Theoretical results

Theorem 1 *Consider a matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ encoding the weights of connections between a layer l_0 of size n and a layer l_1 of size m in some learned unstructured neural network model. Assume that the input to layer l_0 is taken from the d-dimensional space $\mathcal{L}$ (although potentially embedded in a much higher dimensional space). Then with probability at least*

$$1 - 2p(n)d - 2\binom{md}{2}e^{-\Omega(\min(\frac{t^2n^2}{K^4\Lambda_F^2\delta^4(n)}, \frac{tn}{K^2\Lambda_2\delta^2(n)}))}$$

for $t = \frac{1}{md}$ and with respect to random choices of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, there exists a vector $\boldsymbol{r}$ defining $\boldsymbol{M}_3$ such that the structured spinner $\boldsymbol{M}^{struct} = \boldsymbol{M}_3\boldsymbol{M}_2\boldsymbol{M}_1$ equals to $\boldsymbol{M}$ on $\mathcal{L}$.





Lemma 1 (structured random setting theorem) *Let $\mathcal{A}$ be a randomized algorithm using unstructured Gaussian matrices $\boldsymbol{G}$ and let $\mathcal{A}^{TS}$ be its structured version obtained by replacing the unstructured matrix $\boldsymbol{G}$ by TripleSpinner with blocks of m rows each. Denote by d the dimensionality of the space on which $\mathcal{A}$ acts. Then for n large enough and $\epsilon = o_{md}(1)$ with probability p_{succ} at least:*

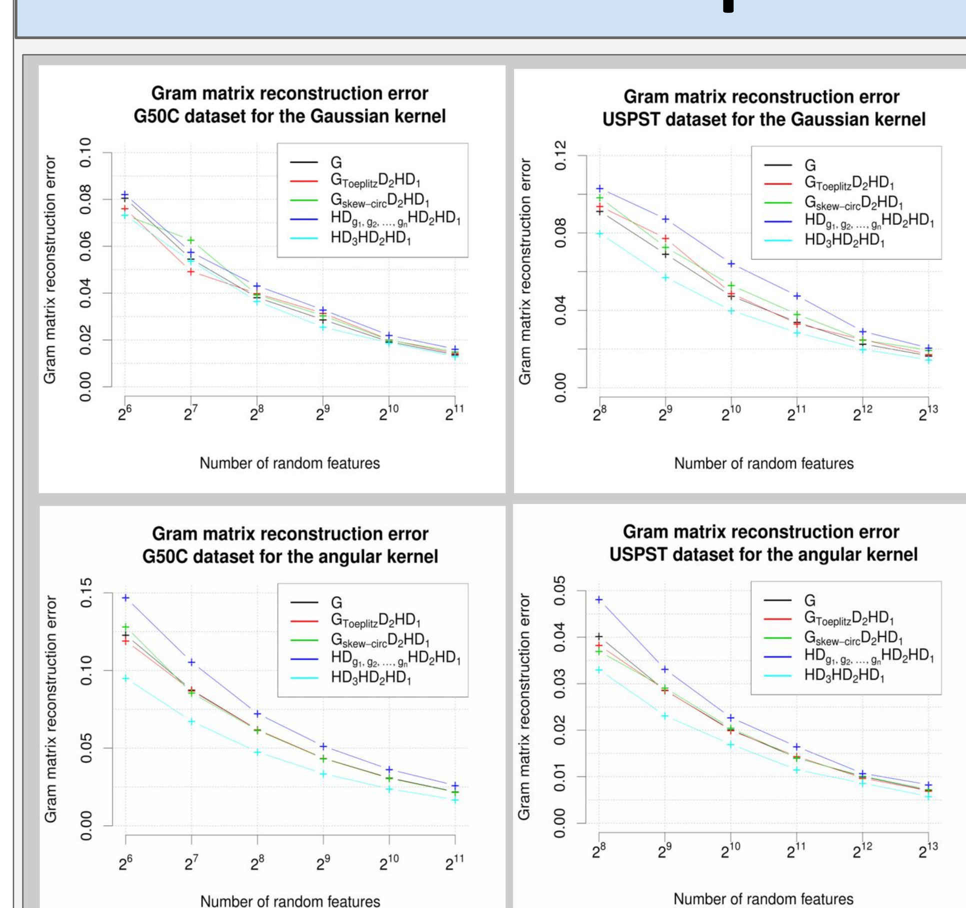
$$1 - 2p(n)d - 2\binom{md}{2}e^{-\Omega(\min(\frac{\epsilon^2n^2}{K^4\Lambda_F^2\delta^4(n)}, \frac{\epsilon n}{K^2\Lambda_2\delta^2(n)}))}$$

with respect to the random choices of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ the following holds for any $\mathcal{S}$ such that $\mathcal{A}^{-1}(\mathcal{S})$ is measurable and b-convex: Cross - polytope LSH

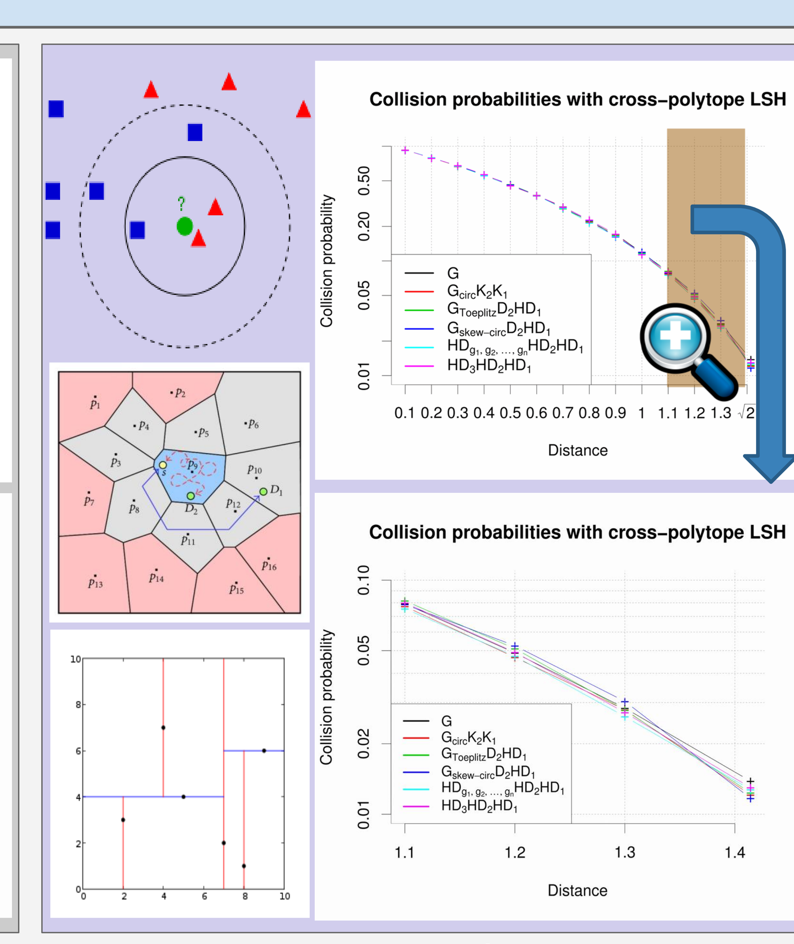
$$|\mathbb{P}[\mathcal{A}(\boldsymbol{q}) \in \mathcal{S}] - \mathbb{P}[\mathcal{A}^{TS}(\boldsymbol{q}) \in \mathcal{S}]| \leq b\eta,$$

where the the probabilities in the last formula are with respect to the random choice of $\boldsymbol{M}_3$ and $\eta = \frac{\delta^3(n)}{n^{\frac{2}{5}}}$.

Experiments



Kernel approximation with random feature maps



Cross-polytope LSH