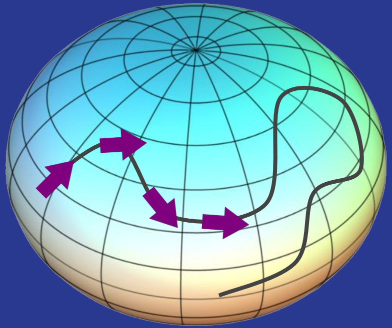
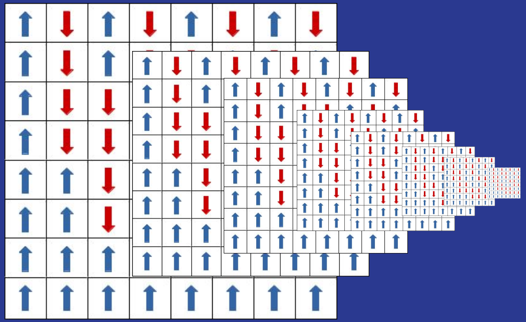
Unifying Orthogonal Monte Carlo Methods

From Kac's Random Walks To Hadamard Multi Rademachers



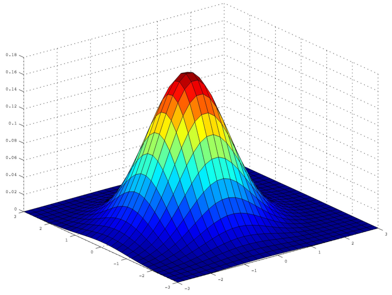


Krzysztof Choromanski, Mark Rowland
Wenyu Chen, Adrian Weller

The Phenomenon of Orthogonal Monte Carlo Estimators

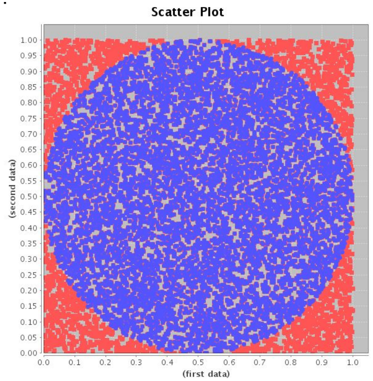
Estimation task: $\mathbb{E}_{X \sim \boxed{\mu}}[f(X)]$

isotropic distribution
(e.g. Gaussian)



Standard MC approach:

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i)\,, \text{ where } (X_i)_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mu\,.$$



Applications:

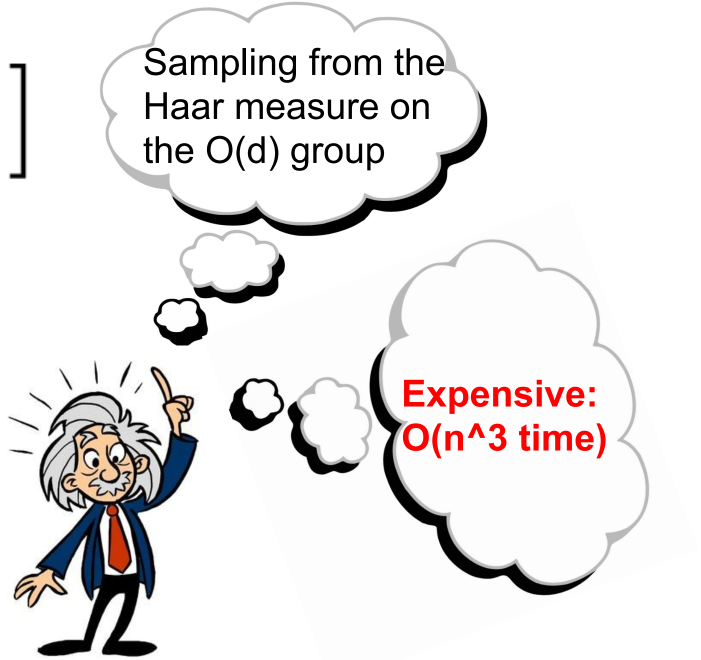
- dimensionality reduction (JLT-mechanisms)
- scaling kernel methods (random feature maps)
- hashing algorithms (e.g. LSH)
- (sliced) Wasserstein distances (WGANs, autoencoders...)
- reinforcement learning (ES algorithms)
- and many, many more...

The Phenomenon of Orthogonal Monte Carlo Estimators

Estimation task: $\mathbb{E}_{X \sim \boxed{\mu}}[f(X)]$

$$\mathbf{G}_{\mathrm{ort}} = \begin{pmatrix} g_{1,1}^{\mathrm{ort}} & g_{1,2}^{\mathrm{ort}} & \cdots & g_{1,n}^{\mathrm{ort}} \\ g_{2,1}^{\mathrm{ort}} & g_{2,2}^{\mathrm{ort}} & \cdots & g_{2,n}^{\mathrm{ort}} \\ \cdots & \cdots & \cdots & \cdots \\ g_{m,1}^{\mathrm{ort}} & g_{m,n}^{\mathrm{ort}} & \cdots & g_{m,n}^{\mathrm{ort}} \end{pmatrix}$$

isotropic distribution (e.g. Gaussian)



The Orthogonal Trick: **guarantees unbiasedness**

$$\frac{1}{\boxed{N}} \sum_{i=1}^{N} f(X_i^{\mathrm{ort}}), \quad \text{where } \boxed{(X_i^{\mathrm{ort}}) \sim \mu} \text{ and } \boxed{X_i^{\mathrm{ort}} \perp X_j^{\mathrm{ort}}}.$$

# of samples of the MC estimator <= dim

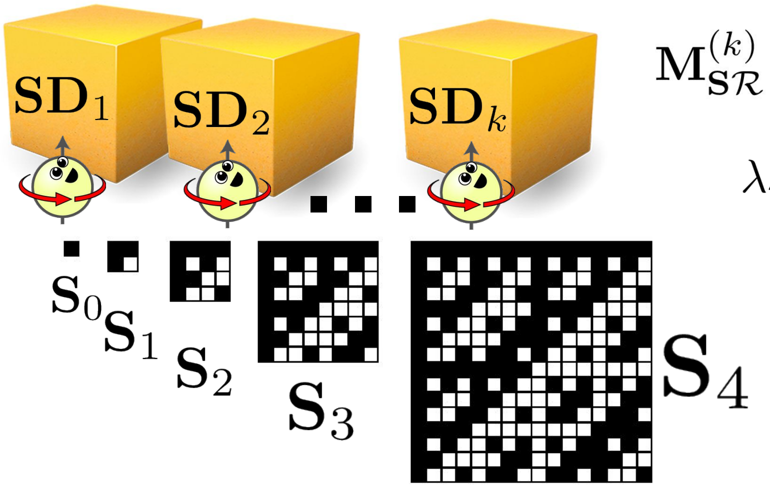
often implies better accuracy

Towards Computational Efficiency:

The Zoo of Approximate MCs

Towards Computational Efficiency:

The Zoo of Approximate MCs



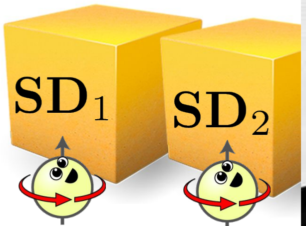
$$\mathbf{M}_{\mathbf{S}\mathcal{R}}^{(k)} = \prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

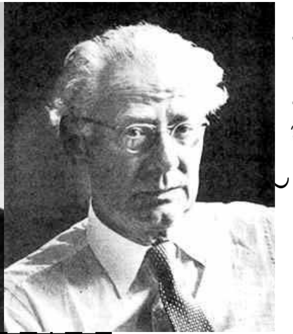
$$\lambda_i \sim Unif\{-1, +1\}$$

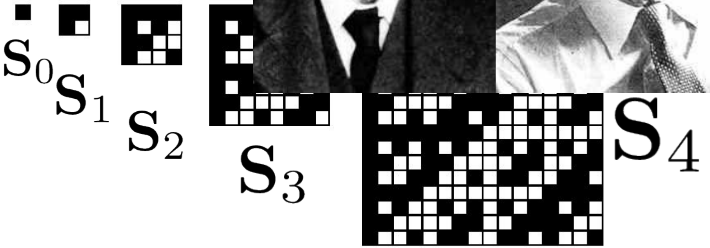
$$\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Towards Computational Efficiency:

The Zoo of Approximate MCs







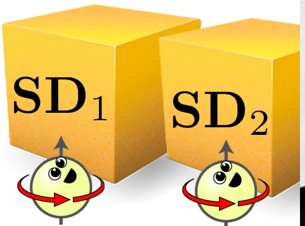
$$\prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

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

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

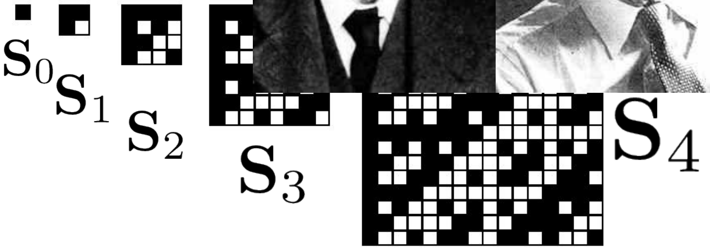
Towards Computational Efficiency:

The Zoo of Approximate MCs

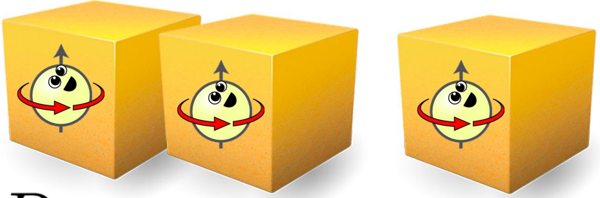


$$\prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

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



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

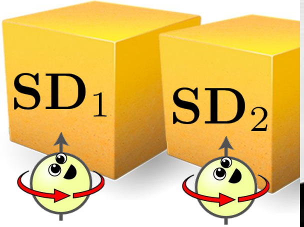


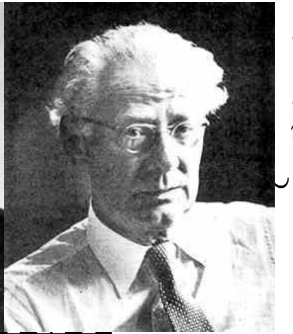
$R_{Cn\log(n)}$ - - - R_1

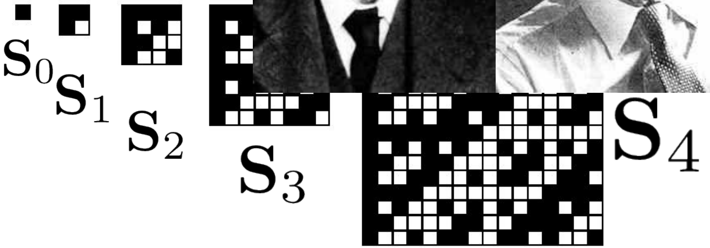
$$\mathbf{R}[i,j] = \begin{cases} 1, & \text{if } i = j \text{ and } i \notin \{I,J\} \\ 0, & \text{if } i \neq j \text{ and } \{i,j\} \neq \{I,J\} \\ \cos\Theta, & \text{if } i = j \text{ and } i \in \{I,J\} \\ \sin\Theta, & \text{if } i = J, j = I \\ -\sin\Theta, & \text{if } i = I, j = J \end{cases}$$

Towards Computational Efficiency:

The Zoo of Approximate MCs



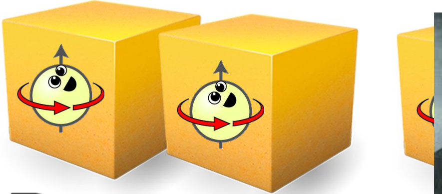




$$\prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

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

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

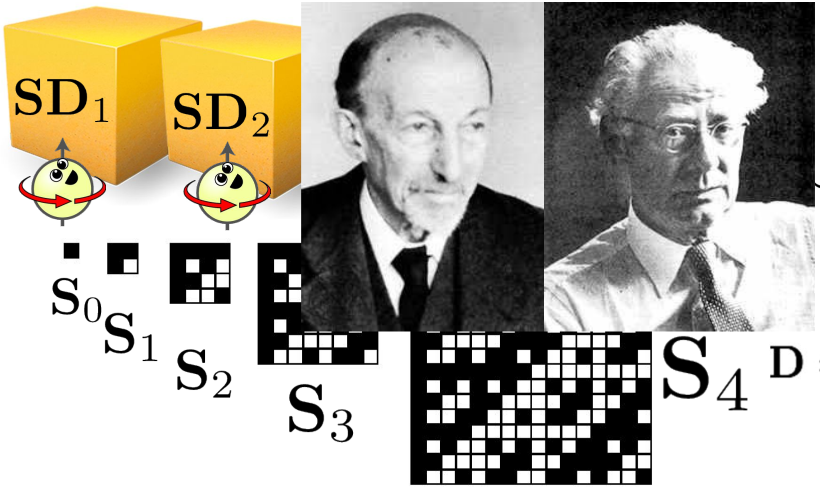


$R_{Cn\log(n)}$

$$, j] = \begin{cases} 1, & \text{if } i = j \text{ and } i \notin \{I, J\} \\ 0, & \text{if } i \neq j \text{ and } \{i, j\} \neq \{I, J\} \\ \cos\Theta, & \text{if } i = j \text{ and } i \in \{I, J\} \\ \sin\Theta, & \text{if } i = J, j = I \\ -\sin\Theta, & \text{if } i = I, j = J \end{cases}$$

Towards Computational Efficiency:

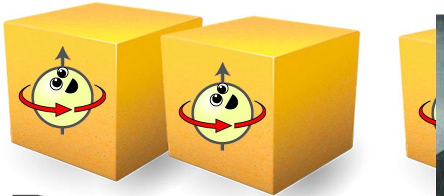
The Zoo of Approximate MCs



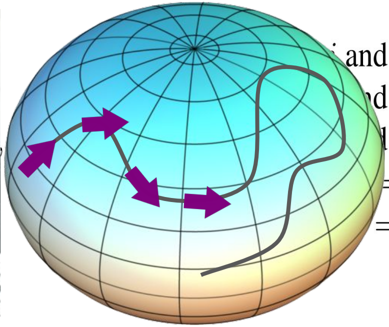
$$\prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

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

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



$R_{Cn\log(n)}$



and $i \notin \{I, J\}$

d $\{i, j\} \neq \{I, J\}$

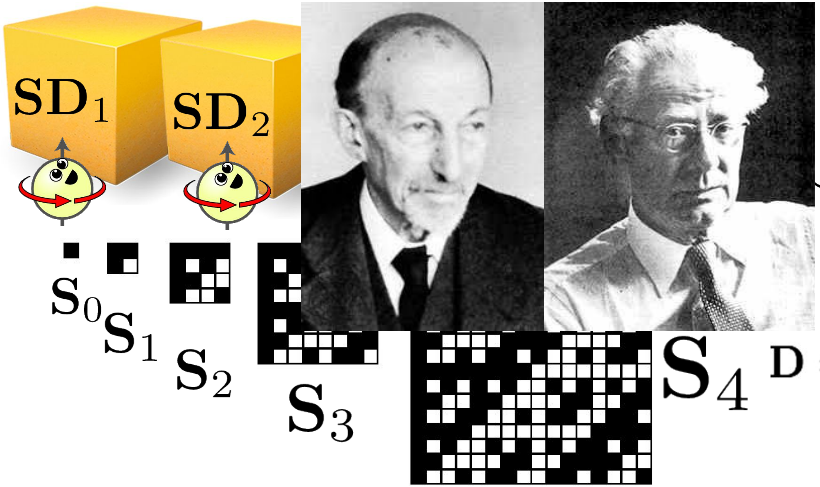
$i \in \{I, J\}$

$= I$

$= J$

Towards Computational Efficiency:

The Zoo of Approximate MCs



$$\prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

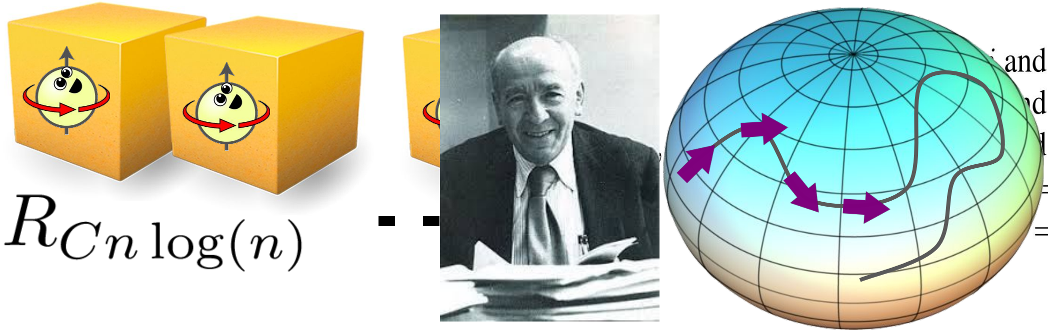
$$\sim Unif\{-1,+1\}$$

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

$$B^{(N)} = \begin{bmatrix} \boxed{A_1^{(N/2)}} C_{n-1} & A_2^{(N/2)} S_{n-1} \\ -A_1^{(N/2)} S_{n-1} & \boxed{A_2^{(N/2)}} C_{n-1} \end{bmatrix}$$

size N x N

size N/2 x N/2



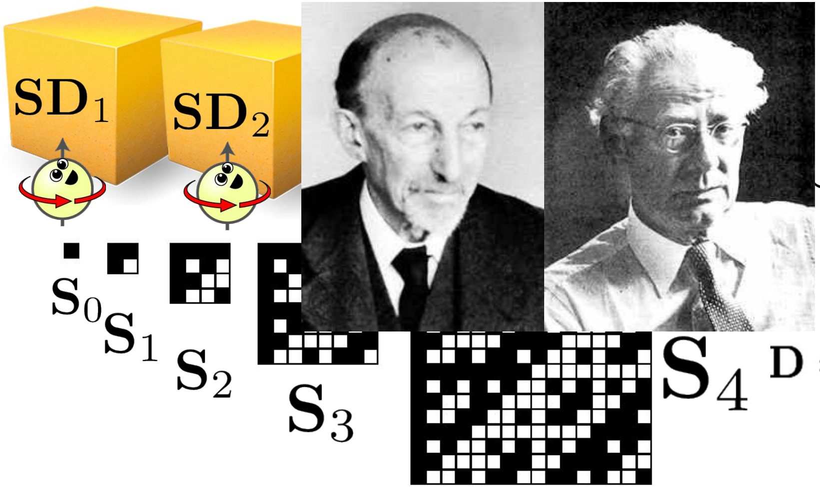
$$N = 2^n$$

Constraints:

- $C_{n-1}^2 + S_{n-1}^2 = I$
- $C_{n-1} S_{n-1} = S_{n-1} C_{n-1}$

Towards Computational Efficiency:

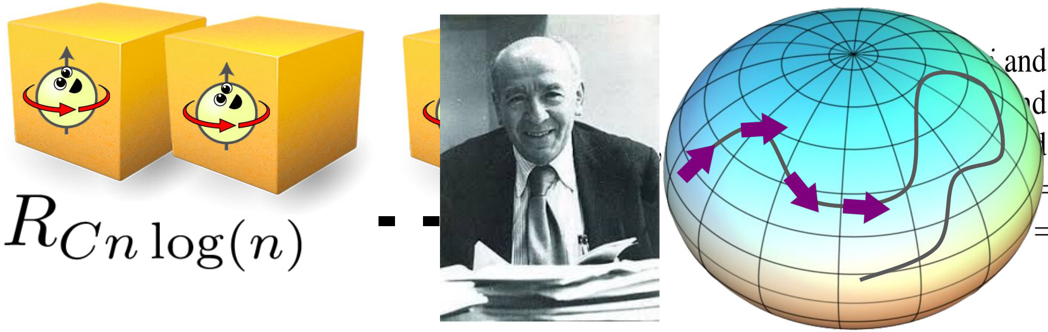
The Zoo of Approximate MCs



$$\prod_{i=1}^{k} \mathbf{SD}_i^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$$

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

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



$$B^{(N)} = \begin{bmatrix} \boxed{A_1^{(N/2)}} C_{n-1} & A_2^{(N/2)} S_{n-1} \\ -A_1^{(N/2)} S_{n-1} & \boxed{A_2^{(N/2)}} C_{n-1} \end{bmatrix}$$

size N x N

size N/2 x N/2

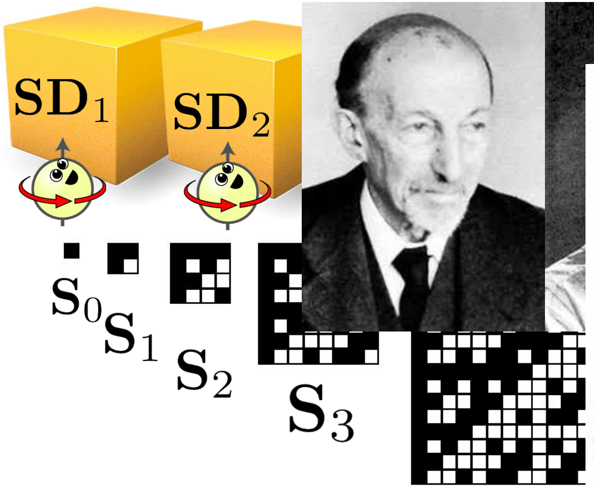
$$N = 2^n$$

Constraints:

- $C_{n-1}^2 + S_{n-1}^2 = I$
- $C_{n-1}S_{n-1} = S_{n-1}C_{n-1}$

Towards Computational Efficiency:

The Zoo of Approximate MCs

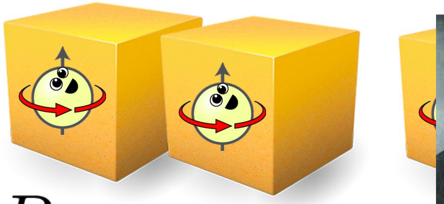


$\prod^{k} SD^{(\mathcal{R})} \rightarrow |\lambda_i| = 1$

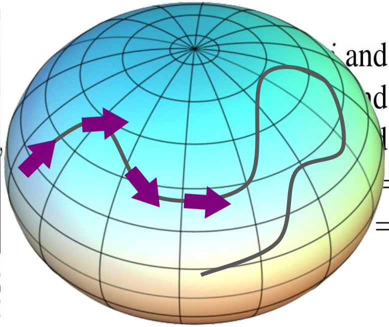
$$) = \begin{bmatrix} \boxed{A_1^{(N/2)}} C_{n-1} & A_2^{(N/2)} S_{n-1} \\ -A_1^{(N/2)} S_{n-1} & \boxed{A_2^{(N/2)}} C_{n-1} \end{bmatrix}$$

size
N/2 x N/2

$N = 2^n$



$R_{Cn\log(n)}$



and $i \notin \{I, J\}$
d $\{i, j\} \neq \{I, J\}$
$i \in \{I, J\}$
$= I$
$= J$

Constraints:

- $C_{n-1}^2 + S_{n-1}^2 = I$
- $C_{n-1} S_{n-1} = S_{n-1} C_{n-1}$

On the Hunt for the Unifying Theory:

The World of Givens Reflections and Rotations

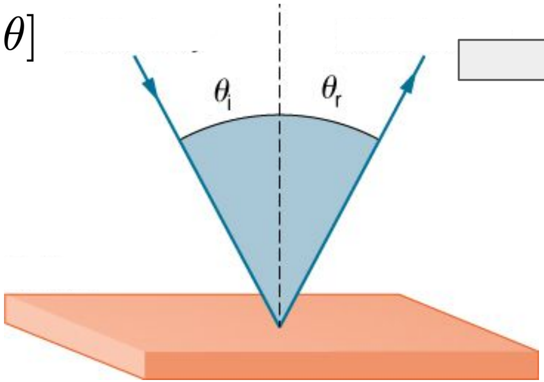
Givens rotations

$$\mathbf{G}[i,j,\theta]_{k,l} = \begin{cases} \cos(\theta) & \textit{if } k = l \in \{i,j\} \\ -\sin(\theta) & \textit{if } k = i, l = j \\ \sin(\theta) & \textit{if } k = j, l = i \\ 1 & \textit{if } k = l \notin \{i,j\} \\ 0 & \textit{otherwise}\,. \end{cases}$$

+

Givens reflections

$$\widetilde{\mathbf{G}}[i,j,\theta]$$



reflection in the jth coordinate

Kac's random walk matrices

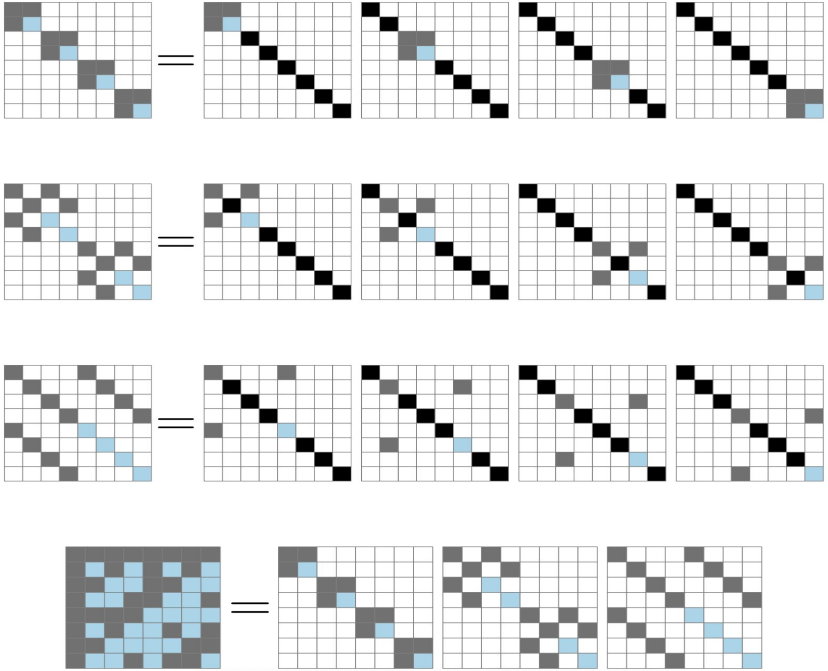
$$\mathbf{K}_T = \prod_{t=1}^{T} \mathbf{G}[I_t, J_t, \theta_t]$$

Hadamard-Rademacher Chains

$$\mathbf{X}_T = \prod_{t=1}^{T} \mathbf{HD}_t$$

On the Hunt for the Unifying Theory:

The World of Givens Reflections and Rotations



$$\widetilde{\mathbf{F}}^{j,L} = \prod_{\lambda \in \mathbb{F}_2^L, \lambda_j = 0} \widetilde{\mathbf{G}}[\boldsymbol{\lambda}, \boldsymbol{\lambda} + \mathbf{e}_j, \pi/4] \in \mathcal{O}(2^L)$$

Kac's random walk matrices

$$\mathbf{K}_T = \prod_{t=1}^{T} \mathbf{G}[I_t, J_t, \theta_t]$$

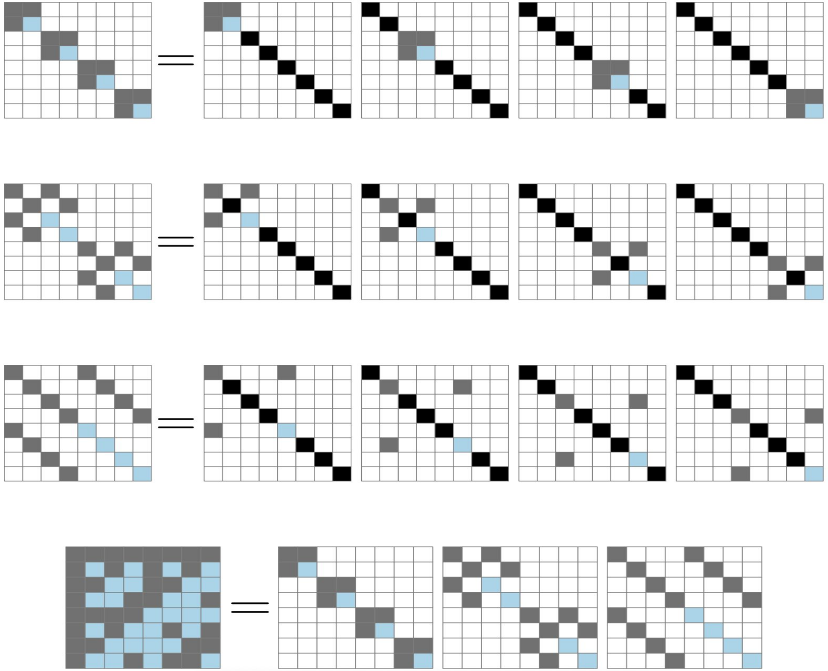
Hadamard-Rademacher Chains

$$\mathbf{X}_T = \prod_{t=1}^{T} \mathbf{HD}_t$$

$$\boxed{\mathbf{HD}_t = \left(\prod_{i=1}^{L-1} \widetilde{\mathbf{F}}^{i,L} \right) (\widetilde{\mathbf{F}}^{L,L} \mathbf{D}_t)}$$

On the Hunt for the Unifying Theory:

The World of Givens Reflections and Rotations



$$\widetilde{\mathbf{F}}^{j,L} = \prod_{\lambda\in\mathbb{F}_2^L,\lambda_j=0} \widetilde{\mathbf{G}}[\boldsymbol{\lambda}, \boldsymbol{\lambda}+\mathbf{e}_j, \pi/4] \in \mathcal{O}(2^L)$$

Hadamard-MultiRademachers

$$\mathbf{M}_L = \prod_{i=1}^{L} \left(\widetilde{\mathbf{F}}^{i,L} \mathbf{D}_i \right)$$

Butterfly Matrices

$$\mathbf{F}^{j,L}[(\theta_{j,\boldsymbol{\mu}})_{\boldsymbol{\mu}\in\mathbb{F}_2^{L-j}}] = \prod_{\substack{\boldsymbol{\lambda}\in\mathbb{F}_2^L \\ \boldsymbol{\lambda}_j=0}} \mathbf{G}[\boldsymbol{\lambda}, \boldsymbol{\lambda}+\mathbf{e}_j, \theta_{j,\boldsymbol{\lambda}_{j+1:L}}] \in \mathcal{O}(2^L)$$

$$((\theta_{i,\boldsymbol{\mu}})_{\boldsymbol{\mu}\in\mathbb{F}_2^{L-i}})_{i=1}^{L} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,2\pi))$$

$$\mathbf{B}_L = \prod_{i=1}^{L} \mathbf{F}^{i,L}[(\theta_{i,\boldsymbol{\mu}})_{\boldsymbol{\mu}\in\mathbb{F}_2^{L-i}}]$$

First Theoretical Results for Free-Lunch Phenomenon in the Nonlinear Regime

Theorem (Kac's random walk estimators of RBF kernels). *Let* $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be the Gaussian kernel and let* $\epsilon > 0$. *Let* $\mathcal{B}$ *be a set satisfying* $\mathrm{diam}(\mathcal{B}) \leq B$ *for some universal constant* B *that does not depend on d (*$\mathcal{B}$ *might be for instance a unit sphere). Then there exists a constant* $C = C(B, \epsilon) > 0$ *such that for every* $\mathbf{x}, \mathbf{y} \in \mathcal{B} \backslash S(\epsilon)$ *and d large enough we have:*

$$\mathrm{MSE}(\widehat{K}^{\phi,m,k}_{\mathrm{kac}}(\mathbf{x}, \mathbf{y})) < \mathrm{MSE}(\widehat{K}^{\phi,m}_{\mathrm{base}}(\mathbf{x}, \mathbf{y})),$$

where $k = C \cdot d \log d$ *and* $m = ld$ *for some* $l \in \mathbb{N}$.

First Theoretical Results for Free-Lunch Phenomenon in the Nonlinear Regime

Theorem (Kac's ... f RBF ker-
nels). *Let* $K_d : \mathbb{R}^d \times$ ... *kernel and*
let $\epsilon > 0$. *Let* $\mathcal{B}$ *be a* ... $\mathcal{B}$ *for some*
universal constant B ... *d* ($\mathcal{B}$ *might*
be for instance a unit ... *a constant*
$C = C(B, \epsilon) > 0$ *suc* ... $S(\epsilon)$ *and d*
large enough we have

Still more accurate estimator than unstructured MC baseline

$$\boxed{\mathrm{MSE}(\widehat{K}^{\phi,m,k}_{\mathrm{kac}}(\mathbf{x},\mathbf{y})) < \mathrm{MSE}(\widehat{K}^{\phi,m}_{\mathrm{base}}(\mathbf{x},\mathbf{y}))},$$

where $k = C \cdot d \log d$ *and* $m = ld$ *for some* $l \in \mathbb{N}$.

First Theoretical Results for Free-Lunch Phenomenon in the Nonlinear Regime

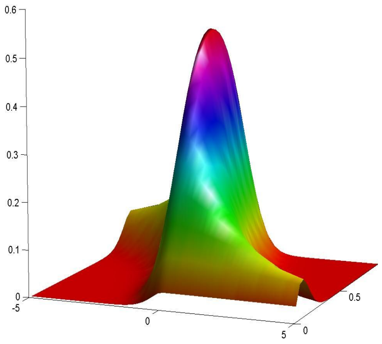
Theorem (Kac's random walk estimators of RBF kernels). *L* ... *he Gaussian kernel and*
let $\epsilon > ($ ... $\mathrm{diam}(\mathcal{B}) \leq B$ *for some*
univers ... *t depend on d* ($\mathcal{B}$ *might*
be for i ... *n there exists a constant*
$C = C($ ... *ry* $\mathbf{x}, \mathbf{y} \in \mathcal{B} \backslash S(\epsilon)$ *and d*
large en ...

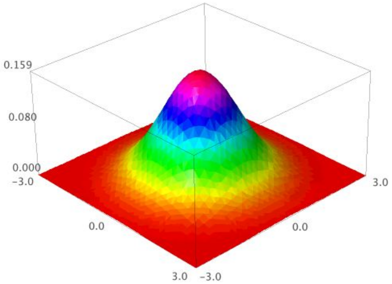
$$\mathrm{M} \ldots \mathrm{E}(\widehat{K}^{\phi,m}_{\mathrm{base}}(\mathbf{x},\mathbf{y})),$$

where $\boxed{k = C \cdot d \log d}$ *and* $m = ld$ *for some* $l \in \mathbb{N}$.

Log-Linear Time Complexity (unstructured MC baseline has quadratic)

First Theoretical Results for Free-Lunch Phenomenon in the Nonlinear Regime





Analysis of the Total Variation Distance between Haar measure on d-sphere and measure induced by standard Kac's random walk on d-sphere

$$\mathrm{MSE}(\underbrace{\boxed{Y}}_{\text{estimator}}) = \mathbb{E}[(Y - \underbrace{\boxed{\mu}}_{\text{estimated value}})^2] = \int_0^\infty \mathbb{P}[|Y-\mu| > \sqrt{t}]dt$$

Pillai, Smith 2016

Kac's random walk on d-sphere mixes in O(d log d) steps

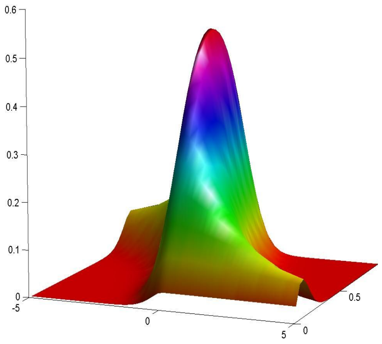
Theorem *Fix* $C_1 < \frac{1}{2}$ *and* $C_2 > 200$. *If the sequence of times* $\{T_1(n)\}_{n\in\mathbb{N}}$ *satisfies* $T_1(n) < C_1 n\log(n)$ *for all* n, *then*

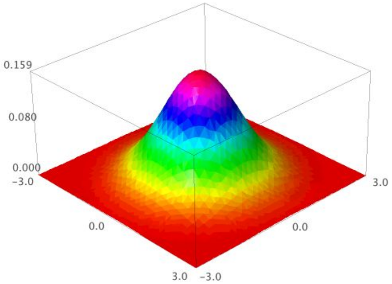
$$\lim_{n\to\infty} \inf_{X_0\in\mathbb{S}^{n-1}} \|\mathcal{L}(X_{T_1(n)}) - \mu\|_{\mathrm{TV}} = 1.$$

If the sequence of times $\{T_2(n)\}_{n\in\mathbb{N}}$ *satisfies* $T_2(n) > C_2 n\log(n)$ *for all* n, *then*

$$\lim_{n\to\infty} \sup_{X_0\in\mathbb{S}^{n-1}} \|\mathcal{L}(X_{T_2(n)}) - \mu\|_{\mathrm{TV}} = 0.$$

First Theoretical Results for Free-Lunch Phenomenon in the Nonlinear Regime





Analysis of the Total Variation Distance between Haar measure on d-sphere and measure induced by standard Kac's random walk on d-sphere

$$\mathrm{MSE}(\underbrace{\boxed{Y}}_{\text{estimator}}) = \mathbb{E}[(Y - \underbrace{\boxed{\mu}}_{\text{estimated value}})^2] = \int_0^\infty \mathbb{P}[|Y-\mu| > \sqrt{t}]dt$$

Pillai, Smith 2016

Kac's random walk on d-sphere mixes in O(d log d) steps

Theorem *Fix $C_1 < \frac{1}{2}$ and $C_2 > 200$. If the sequence of times $\{T_1(n)\}_{n\in\mathbb{N}}$ satisfies $T_1(n) < C_1 n \log(n)$ for all n, then*

$$\lim_{n\to\infty} \inf_{X_0 \in \mathbb{S}^{n-1}} \|\mathcal{L}(X_{T_1(n)}) - \mu\|_{\mathrm{TV}} = 1.$$

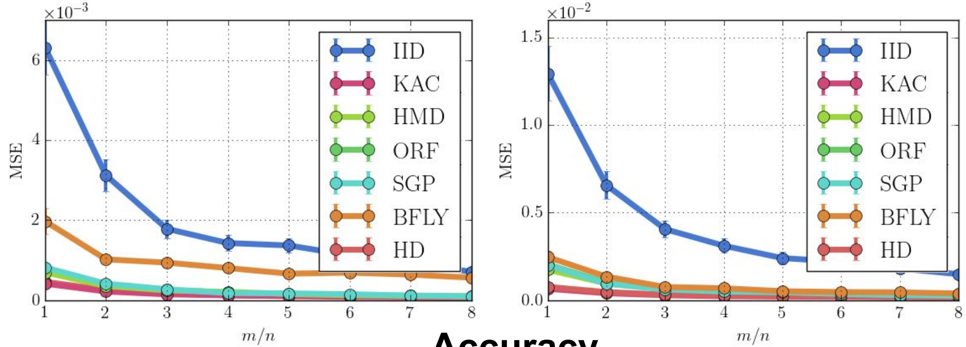
If the sequence of times $\{T_2(n)\}_{n\in\mathbb{N}}$ satisfies $T_2(n) > C_2 n \log(n)$ for all n, then

$$\boxed{\lim_{n\to\infty} \sup_{X_0 \in \mathbb{S}^{n-1}} \|\mathcal{L}(X_{T_2(n)}) - \mu\|_{\mathrm{TV}} = 0.}$$

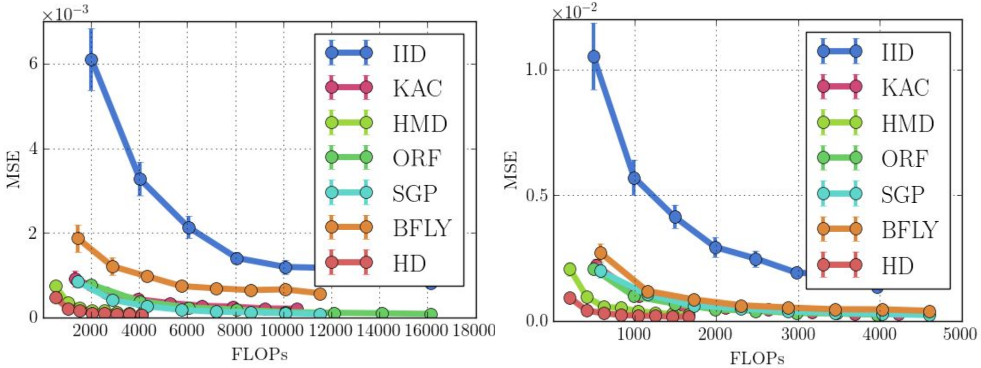
More careful analysis of the LHS

How Does It Work In Practice ?

Kernel Approximation via Random Features

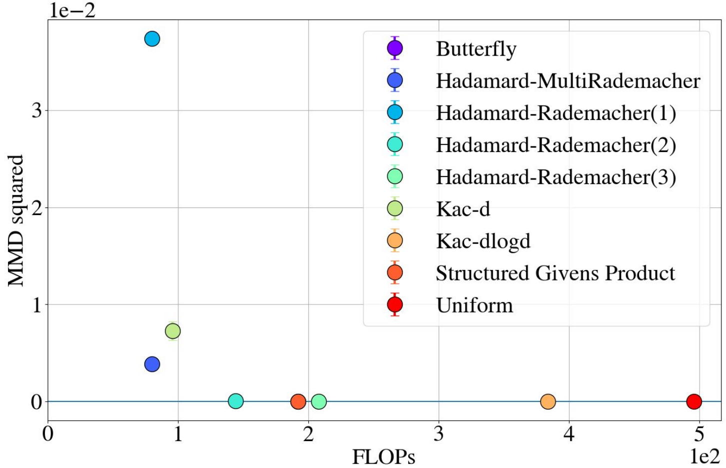


Accuracy

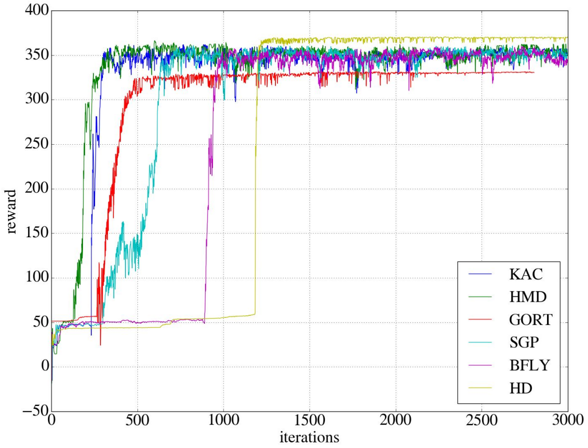


Computational Efficiency

Maximum Mean Discrepancy Experiment

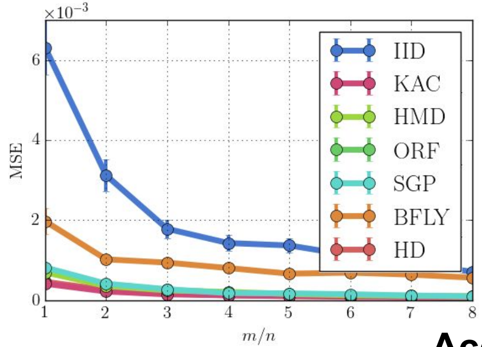


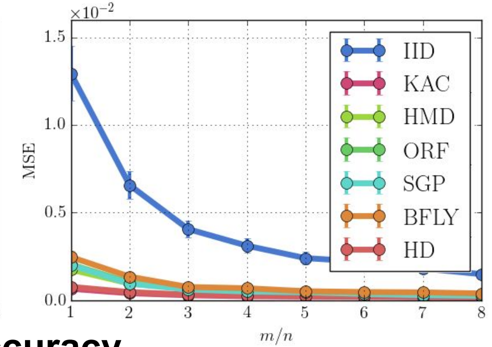
Reinforcement Learning via ES-methods



How Does It Work In Practice ?

Kernel Approximation via Random Features

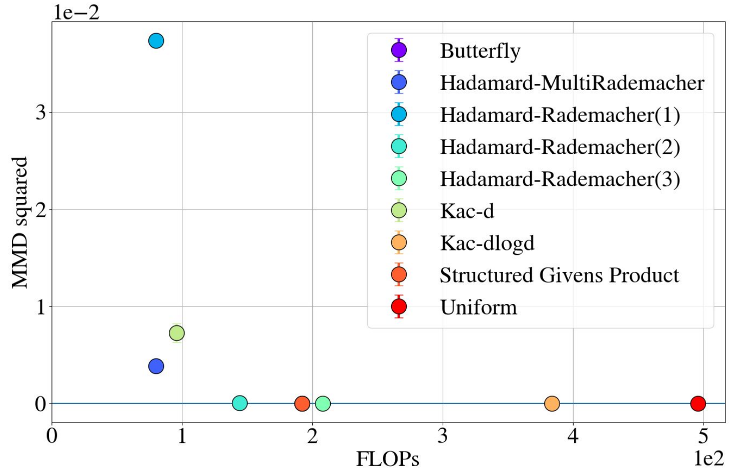




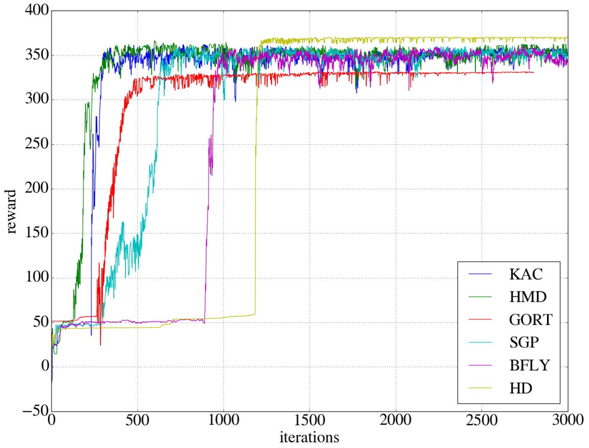
Accuracy

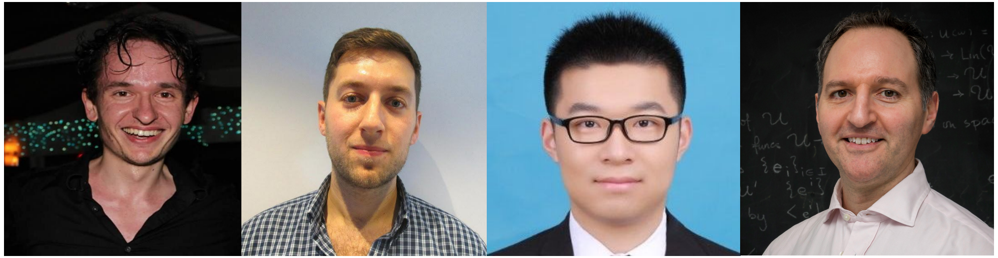
Computational Efficiency

Maximum Mean Discrepancy Experiment



Reinforcement Learning via ES-methods





The Alan Turing Institute

Thank you for your attention !