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: \( E_X \sim \mu [f(X)] \)

isotropic distribution (e.g. Gaussian)

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...

Standard MC approach:
\[
\frac{1}{N} \sum_{i=1}^{N} f(X_i), \text{ where } (X_i)_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mu.
\]
The Phenomenon of Orthogonal Monte Carlo Estimators

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

\[
G_{\text{ort}} = \begin{pmatrix}
g_{1,1} & g_{1,2} & \cdots & g_{1,n} \\
g_{2,1} & g_{2,2} & \cdots & g_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
g_{m,1} & g_{m,2} & \cdots & g_{m,n}
\end{pmatrix}
\]

isotropic distribution (e.g. Gaussian)

The Orthogonal Trick: guarantees unbiasedness

\[
\frac{1}{N} \sum_{i=1}^{N} f(X_{i}^{\text{ort}}), \text{ where } (X_{i}^{\text{ort}}) \sim \mu \text{ and } X_{i}^{\text{ort}} \perp X_{j}^{\text{ort}}.
\]

# of samples of the MC estimator $\leq \dim$

Sampling from the Haar measure on the O(d) group

Expensive: $O(n^3 \text{ time})$
Towards Computational Efficiency:
The Zoo of Approximate MCs
Towards Computational Efficiency: The Zoo of Approximate MCs

\[ M^{(k)}_{SR} = \prod_{i=1}^{k} SD^{(R)}_i \rightarrow |\lambda_i| = 1 \]

\[ \lambda_i \sim Unif\{-1, +1\} \]

\[ S_4 D = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \]
Towards Computational Efficiency:
The Zoo of Approximate MCs

\[
\prod_{i=1}^{k} S^R D_i \rightarrow |\lambda_i| = 1
\]

\[
\lambda_i \in \text{Unif}\{-1, +1\}
\]

\[
S_0, S_1, S_2, S_3, S_4
\]

\[
S_4^D = \begin{pmatrix}
\lambda_1 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & \cdots \\
0 & 0 & \lambda_3 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\]
Towards Computational Efficiency: The Zoo of Approximate MCs

\[ \prod_{i=1}^{k} SD_i^{(R)} \Rightarrow |\lambda_i| = 1 \]

\[ Unif \{-1, +1\} \]

\[ D = \begin{pmatrix}
(\lambda_1 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & \cdots \\
0 & 0 & \lambda_3 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix} \]

\[ R_{Cn} \log(n) \]

\[ R_{1} \]

\[ R_{i,j} = \begin{cases}
1, & \text{if } i = j \text{ and } i \notin \{I, J\} \\
0, & \text{if } i \neq j \text{ and } \{i, j\} \neq \{I, J\} \\
\cos \Theta, & \text{if } i = j \text{ and } i \in \{I, J\} \\
\sin \Theta, & \text{if } i = J, j = I \\
-\sin \Theta, & \text{if } i = I, j = J
\end{cases} \]
Towards Computational Efficiency: The Zoo of Approximate MCs

\[ \prod_{i=1}^{k} \mathcal{SD}_{i}^{(R)} \Rightarrow |\lambda_i| = 1 \]

\[ \text{Unif}\{-1, +1\} \]

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

\[ R_{\text{core}} \log(n) \]

\[ j = \begin{cases}
1, & \text{if } i = j \text{ and } i \notin \{I, J\} \\
0, & \text{if } i \neq j \text{ and } \{i, j\} \neq \{I, J\} \\
\cos \Theta, & \text{if } i = j \text{ and } i \in \{I, J\} \\
\sin \Theta, & \text{if } i = J, j = I \\
-\sin \Theta, & \text{if } i = I, j = J
\end{cases} \]
Towards Computational Efficiency:
The Zoo of Approximate MCs

\[ \prod_{i=1}^{k} \mathbb{SD}^{(R)}_i \rightarrow |\lambda_i| = 1 \]

\[ \text{Unif}\{-1, +1\} \]

\[
\begin{pmatrix}
\lambda_1 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & \cdots \\
0 & 0 & \lambda_3 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\]
Towards Computational Efficiency: The Zoo of Approximate MCs

Constraints:

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

Size $N \times N$

Size $N/2 \times N/2$
Towards Computational Efficiency: The Zoo of Approximate MCs

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

Constraints:
- \( C_{n-1}^2 + S_{n-1}^2 = I \)
- \( C_{n-1}S_{n-1} = S_{n-1}C_{n-1} \)

\( N = 2^n \)
On the Hunt for the Unifying Theory: The World of Givens Reflections and Rotations

**Givens rotations**

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

**Givens reflections**

\[ \tilde{G}[i,j,\theta] \]

**Reflection in the jth coordinate**

**Kac's random walk matrices**

\[ K_T = \prod_{t=1}^{T} G[I_t, J_t, \theta_t] \]

**Hadamard-Rademacher Chains**

\[ X_T = \prod_{t=1}^{T} HD_t \]
On the Hunt for the Unifying Theory:
The World of Givens Reflections and Rotations

Kac’s random walk matrices

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

Hadamard-Rademacher Chains

$$X_T = \prod_{t=1}^{T} HD_t$$

$$\tilde{F}_{j,L} = \prod_{\lambda \in \mathbb{F}_2^L, \lambda_j = 0} \tilde{G}[\lambda, \lambda + e_j, \pi/4] \in \mathcal{O}(2^L)$$

$$HD_t = \left( \prod_{i=1}^{L-1} \tilde{F}_{i,L} \right) (\tilde{F}_{L,L} D_t)$$
On the Hunt for the Unifying Theory:
The World of Givens Reflections and Rotations

Hadamard-MultiRademachers

\[ M_L = \prod_{i=1}^{L} \left( \tilde{F}^i, L D_i \right) \]

Butterfly Matrices

\[ \tilde{F}^{j, L} = \prod_{\lambda \in \mathbb{F}_2^L, \lambda_j = 0} \tilde{G}[\lambda, \lambda + e_j, \pi/4] \in \mathcal{O}(2^L) \]

\[ B_L = \prod_{i=1}^{L} F^{i, L} \left[ (\theta_i, \mu) \mu \in \mathbb{F}_2^{L-i} \right] \]
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 $\text{diam}(\mathcal{B}) \leq B$ 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 $x, y \in \mathcal{B} \setminus \mathcal{S}(\epsilon)$ and $d$ large enough we have:

$$\text{MSE}(\hat{K}_{\text{kac}}^{\phi,m,k}(x, y)) < \text{MSE}(\hat{K}_{\text{base}}^{\phi,m}(x, y)),$$

where $k = C \cdot d \log d$ and $m = \ell d$ for some $\ell \in \mathbb{N}$. 
First Theoretical Results for Free-Lunch Phenomenon in the Nonlinear Regime

Theorem (Kac’s kernels). Let $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a universal constant $B$ for some $l$ (a constant $S'(\epsilon)$ and $d$

\[ \text{MSE}(\hat{K}_{kac}^{\phi,m,k}(x, y)) < \text{MSE}(\hat{K}_{base}^{\phi,m}(x, y)), \]

Still more accurate estimator than unstructured MC baseline. 

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). Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed samples from the Gaussian kernel and let $\varepsilon > 0$ be an arbitrary parameter. Assume that $\text{diam}(B) \leq B$ for some universal $B > 0$. Then, for $n \to \infty$, the following holds:

\[ C = C(\varepsilon) \exists k \left( \forall x, y \in B \setminus S(\varepsilon) \text{ and } d \right), \]

where $k = C \cdot d \log d$ 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

$$\text{MSE}(\hat{Y}) = \mathbb{E}[(Y - \hat{\mu})^2] = \int_0^\infty \mathbb{P}[|Y - \mu| > \sqrt{t}]dt$$

_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

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

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

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

$$\text{MSE}(\bar{Y}) = \mathbb{E}[(Y - \bar{\mu})^2] = \int_0^\infty \mathbb{P}[|Y - \mu| > \sqrt{t}]dt$$

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

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

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

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

More careful analysis of the LHS
How Does It Work In Practice?

Kernel Approximation via Random Features

Maximum Mean Discrepancy Experiment

- Butterfly
- Hadamard-MultiRademacher
- Hadamard-Rademacher(1)
- Hadamard-Rademacher(2)
- Hadamard-Rademacher(3)
- Kac-d
- Kac-dlogd
- Structured Givens Product
- Uniform

Reinforcement Learning via ES-methods

Accuracy

Computational Efficiency
How Does It Work In Practice?

Kernel Approximation via Random Features

Maximum Mean Discrepancy Experiment

Reinforcement Learning via ES-methods

Accuracy

Computational Efficiency
Thank you for your attention!