Unifying Orthogonal Monte Carlo Methods

Krzysztof Choromanski[1*], Mark Rowland[2*], Wenyu Chen[3], Adrian Weller[2,4]

[1]Google Brain, [2]University of Cambridge, [3]Massachusetts Institute of Technology, [4]The Alan Turing Institute, *equal contribution

Monte Carlo (MC) Methods

Consider an expectation of the form:

$$\mathbb{E}_{X\sim\mu}\left[f(X)\right],$$

where: $\mu \in \mathscr{P}(\mathbb{R}^d)$ is an isotropic probability distribution, $f : \mathbb{R}^d \to \mathbb{R}$ is measurable, and μ-integrable. A **standard MC estimator** is given by:

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

Applications in Machine Learning:

1. dimensionality reduction techniques (JLTs)
2. scaling kernel methods via random features
3. locality sensitive hashing algorithms (LSH)
4. ES methods for Reinforcement Learning
5. and many more...

How to improve **computational complexity** and **accuracy** of the above baseline ?

Approximate MC Methods

To improve computational complexity of orthogonal MC estimators, several constructions related to sampling from distributions "close" to Haar distribution on $\mathscr{O}(d)$ were proposed:

Hadamard-Rademacher Chains:

$$\mathbf{X}_T = \prod_{i=1}^{T} \mathbf{HD}_i,$$

with i.i.d. random diagonal matrices $(\mathbf{D}_i)_{i=1}^{T}$ with i.i.d. diagonal entries Unif($\{\pm 1\}$), and where $\mathbf{H} \in \mathbb{R}^{2^L \times 2^L}$ stands for the normalized *Kronecker-product* Hadamard matrix defined as:

$$\mathbf{H} = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}}_{L \text{ times}},$$

for $\otimes$ denoting the Kronecker-product operator.

Computational Complexity: Due to Fast Walsh-Hadamard Transform, $\mathbf{X}_T\mathbf{v}$ for any vector $\mathbf{v} \in \mathbb{R}^{2^L}$ can be computed in time: $\mathcal{O}(T \times L \times 2^L)$.

Many related constructions: e.g. Butterfly matrices generalizing Hadamard-Rademacher chains.

Orthogonal MC Estimators

Consider the following orthogonal MC estimator:

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

Note that for $N \leq d$:

- if μ is isotropic, it can be constructed via **Gram-Schmidt orthogonalization**,
- **the constructions are expensive: $\mathcal{O}(d^3)$ time**
- condition: $X_i^{\mathrm{ort}} \sim \mu$ implies **unbiasedness**.
- $(X_i^{\mathrm{ort}})_{i=1}^{N}$ are renormalized rows of a matrix sampled from Haar measure on $\mathscr{O}(d)$.

Statistical improvements: It often holds

$$\mathrm{MSE}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i^{\mathrm{ort}})\right) < \mathrm{MSE}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i^{\mathrm{iid}})\right).$$

Givens Rotations Matrices

A d-dimensional **Givens rotation** is an orthogonal matrix specified by two distinct indices $i, j \in [d]$, and an angle $\theta \in [0, 2\pi)$. The Givens rotation is then given by the matrix $\mathbf{G}[i, j, \theta]$ satisfying

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

A Givens rotation $\mathbf{G}[i, j, \theta]$ composed on the right with a reflection in the j coordinate will be termed a **Givens reflection** and written $\widetilde{\mathbf{G}}[i, j, \theta]$. Givens rotations and reflections will be generically referred to as **Givens transformations**.

Lemma 1 (Pillai, Smith 2016) *There exists $C > 0$ such that the total variation distance* TV *between distribution $\mathcal{D}^d_{\mathrm{Giv}}$ on the d-sphere induced by the product of $Cd\log(d)$ independent Givens random rotations acting on the L_2-normalized input vector $\mathbf{x}$ and the distribution $\mathcal{D}^d_{\mathrm{Haar}}$ related to Haar measure on that sphere satisfies:*

$$\lim_{d\to\infty} \mathrm{TV}(\mathcal{D}^d_{\mathrm{Giv}}, \mathcal{D}^d_{\mathrm{Haar}}) = 0.$$

On the Hunt for Unifying Approximate Orthogonal MCs

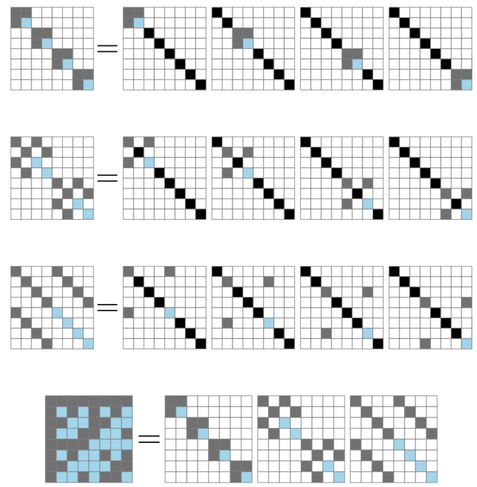

Figure 1: Row 1: the matrix $\widetilde{\mathbf{F}}^{1,3}$ expressed as a commuting product of Givens reflections. Row 2: As above, but for matrix $\widetilde{\mathbf{F}}^{2,3}$. Row 3: the matrix $\widetilde{\mathbf{F}}^{3,3}$ expressed as a product of commuting Givens rotations. Row 4: the normalised Hadamard matrix $\mathbf{H}_3$ written as a product of $\widetilde{\mathbf{F}}^{1,3}$, $\widetilde{\mathbf{F}}^{2,3}$ and $\widetilde{\mathbf{F}}^{3,3}$. White/black represent $0/1$ elements and grey/blue - elements in $(0, 1)$ and $(-1, 0)$.

Kac's random walk matrices (random I_t,J_t, θ_t):

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

The Hadamard-Rademacher Chains: Each block $\mathbf{HD}_i$ of the chain can be rewritten as:

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

where:

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

for the canonical basis $\mathbf{e}_1, \ldots, \mathbf{e}_L$ of $\mathbb{F}_2^L$.
In this expression, we may interpret $\widetilde{\mathbf{F}}^{L,L}\mathbf{D}_t$ as a product of random Givens transformations with a deterministic, structured choice of rotation axes.

Hadamard-MultiRademacher matrices:

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

Denote the kernel estimator using Kac's random walk matrices with k Givens rotations as: $\widehat{K}^k_{\mathrm{kac}}$ and the unstructured baseline as $\widehat{K}_{\mathrm{base}}$. We have:

Theorem 1 (Kac's random walk for RBF kernels) *Let $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel. Then there exists a constant $C > 0$ such that for $\mathbf{x}, \mathbf{y}$, $k = C \cdot d\log d$ and d large enough we have:*

$$\mathrm{MSE}(\widehat{K}^k_{\mathrm{kac}}(\mathbf{x}, \mathbf{y})) < \mathrm{MSE}(\widehat{K}_{\mathrm{base}}(\mathbf{x}, \mathbf{y})).$$

Experiments: From Kernel Approximation to RL

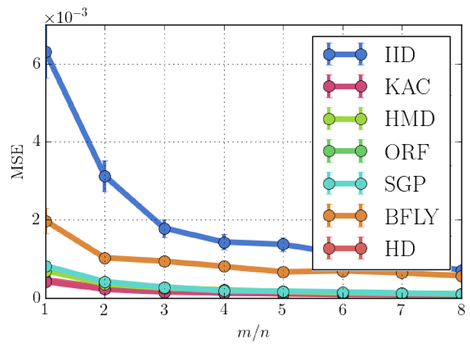

Figure 2: Comparison of different Monte Carlo methods on the task of Gaussian kernel approximation.

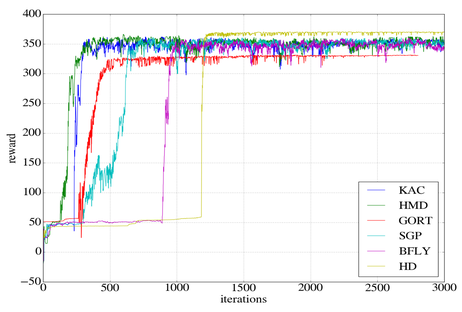

Figure 3: Comparison of different Monte Carlo methods for gradient estimation in ES algorithms for RL.