Quasi-Monte Carlo Features for Kernel Approximation

Zhen Huang

Department of Statistics, Columbia University

ICML 2024

Joint work with Jiajin Sun and Yian Huang

Introduction: kernel method

- Kernel method: mathematically well-founded, practically powerful modeling framework
- Remarkably effective in small and medium size problems with certain optimal statistical results (Kimeldorf & Wahba, 1970; Scholkopf et al., 2001; Caponnetto & De Vito, 2007)
- Infeasible for large scale problems due to its time and memory requirements

Introduction

- Example: Kernel ridge regression (KRR)
  - space complexity $O(n^2)$; time complexity $O(n^3)$
- Various approximation techniques: Nyström (Williams & Seeger, 2000); Smola (2000); incomplete Cholesky decomposition (Bach & Jordan, 2003); random features (Rahimi & Recht, 2007) ...
- Focus on: random features (Rahimi & Recht, 2007)
  - based on Monte Carlo method
  - KRR: space complexity $O(nM)$; time complexity $O(nM^2 + M^3)$ with small $M \ll n$
  - well-understood theoretically (Sutherland & Schneider, 2015; Sriperumbudur & Szabo, 2015; Choromanski et al., 2018; Jacot et al., 2020; Lanthaler & Nelsen, 2023)

Goal: Further improve random features with Quasi-Monte Carlo method in place of Monte Carlo method

Random features: Preliminary

Many kernels on $\mathcal{X} \subset \mathbb{R}^d$ have an integral representation:

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega),$$

π: probability measure over some space Ω
$\psi(\cdot, \cdot)$: a function on $\mathcal{X} \times \Omega$.

Bochner's theorem: For any shift-invariant kernel $K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$, $\exists$ finite non-negative symmetric Borel measure μ s.t.

$$h(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} e^{-i(\mathbf{x}-\mathbf{x}')^\top \omega} \mathrm{d}\mu(\omega)$$
$$= \int_{\mathbb{R}^d} \int_0^{2\pi} \frac{1}{\pi} \cos\left(\mathbf{x}^\top \omega + b\right) \cos\left((\mathbf{x}')^\top \omega + b\right) \mathrm{d}b \, \mathrm{d}\mu(\omega).$$

Some popular shift-invariant kernels

$$h(\mathbf{x}-\mathbf{x}') = \int_{\mathbb{R}^d} e^{-i(\mathbf{x}-\mathbf{x}')^\top \omega} \mathrm{d}\mu(\omega)$$

1. Gaussian kernel $e^{-\|\sigma(\mathbf{x}-\mathbf{x}')\|_2^2/2}$: $\mu \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_d)$.
2. Laplacian kernel $e^{-\|\gamma(\mathbf{x}-\mathbf{x}')\|_1}$: μ has Lebesgue density $\prod_{i=1}^d \frac{1}{\pi\gamma(1+(\omega_i/\gamma)^2)}$ (Cauchy distribution).
3. Cauchy kernel $\prod_{i=1}^d \frac{1}{1+(x_i-x_i')^2/\lambda^2}$: μ has Lebesgue density $\frac{\lambda}{2}e^{-\lambda\|\omega\|_1}$ (Laplace distribution).

Random features

Given the kernel function has integral representation

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega),$$

$K(\mathbf{x}, \mathbf{x}')$ can be approximated by

$$K_M(\mathbf{x}, \mathbf{x}') = \frac{1}{M}\sum_{i=1}^{M} \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i),$$

with $\omega_1, \ldots, \omega_M$ i.i.d. from π (Monte Carlo method)

Computation: Reduce KRR complexity to that of usual ridge regression (as K_M is an inner product on $\mathbb{R}^M$)

Approximation error: $|K(x, x') - K_M(x, x')| = O_P(1/\sqrt{M})$

RF approximation error: $|K(x,x') - K_M(x,x')| = O_P(1/\sqrt{M})$

Limitation:

- non-deterministic error bound
- error rate $\frac{1}{\sqrt{M}}$ decays slowly

Goal: Replace **MC** sequence $\omega_1, \omega_2, \ldots$ with **QMC** sequence to yield

- deterministic error bound
- error rate $\frac{1}{M}$ (up to log factors)

Quasi-Monte Carlo (QMC) method

- QMC: Powerful tool in numerical integration
- Focus: Approximate $\int_{[0,1]^d} f(\mathbf{x})\mathrm{d}\mathbf{x}$ with $\frac{1}{M}\sum_{i=1}^{M} f(\mathbf{x}_i)$ for some well-chosen deterministic sequence $\{\mathbf{x}_i\}_{i=1}^{M}$ that are spread out more 'uniformly' in some sense.

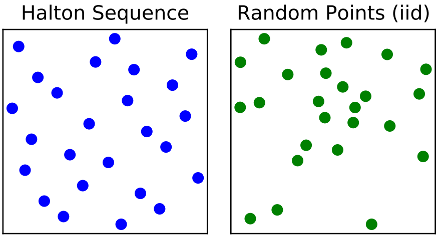

Figure: Left: the first 25 points of the two-dimensional Halton sequence. Right: 25 i.i.d. random points from $\mathrm{Unif}[0,1]^2$.

QMC targets functions with finite variation:

Koksma-Hlawka inequality (Hlawka, 1961)

Suppose $f : [0,1]^d \to \mathbb{R}$ has finite variation in the sense of Hardy and Krause $V_{\mathrm{HK}}(f)$. Then for any $\mathbf{x}_1, \ldots, \mathbf{x}_M \in [0,1]^d$, we have

$$\left| \int_{[0,1]^d} f(\mathbf{x}) \mathrm{d}\mathbf{x} - \frac{1}{M} \sum_{i=1}^{M} f(\mathbf{x}_i) \right| \leq V_{\mathrm{HK}}(f) \mathcal{D}^*(\{\mathbf{x}_i\}_{i=1}^M),$$

where $\mathcal{D}^*(\{\mathbf{x}_i\}_{i=1}^M)$ is the *star discrepancy*[a] of the point set $\{\mathbf{x}_i\}_{i=1}^M$.

[a] $\mathcal{D}^*(\{\mathbf{x}_i\}_{i=1}^M) := \sup_{\mathbf{t} \in [0,1]^d} \left| \mathrm{Vol}(J_{\mathbf{t}}) - \frac{|\{i \in \{1,\ldots,M\} : \mathbf{x}_i \in J_{\mathbf{t}}\}|}{M} \right|$, where $J_{\mathbf{t}} := [0, t_1) \times [0, t_2) \times \cdots \times [0, t_d)$ and $\mathrm{Vol}(J_{\mathbf{t}}) := \prod_{i=1}^{d} t_i$ is the volume.

Halton sequence (a QMC sequence): $\mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^M) \leq C_H(d) (\log M)^d / M$

Question: Can we directly apply QMC inequality when approximating

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega)$$

with

$$K_M(\mathbf{x}, \mathbf{x}') = \frac{1}{M}\sum_{i=1}^{M} \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i) \quad ?$$

Question: Can we directly apply QMC inequality when approximating

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega)$$

with

$$K_M(\mathbf{x}, \mathbf{x}') = \frac{1}{M}\sum_{i=1}^{M} \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i) \quad ?$$

Negative result (Avron et al., 2016): For all shift-invariant kernels, the integral representation from Bochner's theorem has infinite variation (when written as the integral over the unit cube)

Question: Can we directly apply QMC inequality when approximating

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega)$$

with

$$K_M(\mathbf{x}, \mathbf{x}') = \frac{1}{M}\sum_{i=1}^{M} \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i) \quad ?$$

Negative result (Avron et al., 2016): For all shift-invariant kernels, the integral representation from Bochner's theorem has infinite variation (when written as the integral over the unit cube)

Our contribution: For a class of shift-invariant kernels (including Gaussian kernel), even though the integrand has infinite variation, the singularity is mild, so the approximation error can still be well controlled:

$$|K_M(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')| \lesssim \frac{1}{M} \qquad \text{(up to log factors)}$$

Methodology for shift-invariant kernel

Assume μ from Bochner's theorem is a probability measure with independent components, with the *i*-th component having c.d.f. $\Phi_i(t)$

$\mathbf{\Phi}(\mathbf{t}) := (\Phi_1(\mathbf{t}), \ldots, \Phi_d(\mathbf{t}))^\top$; $\mathbf{\Phi}^{-1}(\mathbf{t}) := (\Phi_1^{-1}(\mathbf{t}), \ldots, \Phi_d^{-1}(\mathbf{t}))^\top$

By a change of variable,

$$K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}') = \int_{[0,1]^{d+1}} 2\cos\left(\mathbf{x}^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 2\pi b\right) \cos\left((\mathbf{x}')^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 2\pi b\right) \mathrm{d}b\mathrm{d}\mathbf{t}.$$

$\omega := (\mathbf{t}, b) \sim \mathrm{Unif}[0,1]^{d+1}$; $\psi(\mathbf{x}, \omega) := \sqrt{2}\cos\left(\mathbf{x}^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 2\pi b\right)$.

Our QMC features: Set $\omega_1, \ldots, \omega_M$ as the first M points in the Halton sequence (instead of M i.i.d. points), and define the approximate kernel $K_M(\cdot, \cdot) := \frac{1}{M}\sum_{i=1}^M \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i)$ as in classical random features.

Mild singularity condition for $1/M$ error bound

QMC Condition 1

$K(\cdot,\cdot)$ is shift invariant with marginal c.d.f. Φ_i $(i = 1, \ldots, d)$ satisfying $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_i^{-1}(t) \leq \frac{C_i}{\min(t, 1-t)}$ for some constant $C_i > 0$ and all $t \in (0, 1)$. $\mathcal{X}$ is compact.

- Gaussian kernel and Cauchy kernel over a compact domain satisfy QMC Condition 1.

Mild singularity condition for $1/M$ error bound

QMC Condition 1

$K(\cdot,\cdot)$ is shift invariant with marginal c.d.f. Φ_i $(i = 1, \ldots, d)$ satisfying $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_i^{-1}(t) \le \frac{C_i}{\min(t,1-t)}$ for some constant $C_i > 0$ and all $t \in (0,1)$. $\mathcal{X}$ is compact.

- Gaussian kernel and Cauchy kernel over a compact domain satisfy QMC Condition 1.
- They are examples of *universal kernels* (Micchelli et al., 2006): the associated function class (RKHS) can approximate any continuous function arbitrarily well
- Particularly useful in ML applications such as kernel ridge regression

QMC: Improved approximation error

Theorem (Approximation error of QMC features)

Suppose $K(\cdot, \cdot)$ satisfies QMC Condition 1. Then there exists a constant $C > 0$ (depending on $\mathcal{X} \subset \mathbb{R}^d$ and K) such that for any $x, x' \in \mathcal{X}$ and $M \geq 2$,

$$|K_M(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')| \leq \frac{C(\log M)^{2d+1}}{M}.$$

QMC: Improved approximation error

Theorem (Approximation error of QMC features)

Suppose $K(\cdot,\cdot)$ satisfies QMC Condition 1. Then there exists a constant $C > 0$ (depending on $\mathcal{X} \subset \mathbb{R}^d$ and K) such that for any $x, x' \in \mathcal{X}$ and $M \geq 2$,

$$|K_M(\mathbf{x},\mathbf{x}') - K(\mathbf{x},\mathbf{x}')| \leq \frac{C(\log M)^{2d+1}}{M}.$$

Proof idea:

1. Singularity near the boundary is mild when QMC Condition 1 holds
2. Halton sequence avoids the boundary of the unit cube (Owen, 2006)

Non-shift invariant kernel

Bochner's theorem no longer applicable.

Whether $K(\cdot, \cdot)$ has an integral representation

$$K(\mathbf{x}, \mathbf{x}') = \int_{[0,1]^p} \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega), \tag{1}$$

needs to be considered on a case-by-case basis.

Non-shift invariant kernel

Bochner's theorem no longer applicable.

Whether $K(\cdot,\cdot)$ has an integral representation

$$K(\mathbf{x},\mathbf{x}') = \int_{[0,1]^p} \psi(\mathbf{x},\omega)\psi(\mathbf{x}',\omega)\mathrm{d}\pi(\omega), \tag{1}$$

needs to be considered on a case-by-case basis.

QMC Condition 2: If (1) exists, **and** $\forall\ \mathbf{x},\mathbf{x}' \in \mathcal{X}$, $g(\omega) = \psi(\mathbf{x},\omega)\psi(\mathbf{x}',\omega)$ is of bounded variation $V_{\mathrm{HK}}(g) \leq C_0$, then QMC features yields

$$|K_M(\mathbf{x},\mathbf{x}') - K(\mathbf{x},\mathbf{x}')| \leq C_0 C_H(p) \cdot \frac{(\log M)^p}{M}.$$

Examples

Non-shift invariant kernels to which QMC applies:

1. **Min kernel**: $K(u, v) = \min\{u, v\} = \int_0^1 1_{t<u} 1_{t<v} \mathrm{d}t$
2. **Brownian bridge**:
   $K(u, v) = \min\{u, v\} - uv = \int_0^1 (1_{t<u} - u)(1_{t<v} - v)\mathrm{d}t$
3. **Iterative kernel** (Courant & Hilbert, 1953): $K_1(\cdot, \cdot)$: a 'smooth' kernel; μ: positive integrable function. Iterative kernel:

$$K_2(\mathbf{x}, \mathbf{z}) := \int_{[0,1]^d} K_1(\mathbf{x}, \mathbf{t}) K_1(\mathbf{z}, \mathbf{t}) \mu(\mathbf{t}) \mathrm{d}\mathbf{t}.$$

4. **Natural cubic spline**: $K(u, v) = \int_0^1 (u \wedge t - ut)(v \wedge t - vt)\mathrm{d}t$
5. **Product kernels**

- Exact kernel ridge regression (KRR)
  - space complexity $O(n^2)$; time complexity $O(n^3)$
- RF-KRR & QMCF-KRR
  - space complexity $O(nM)$; time complexity $O(nM^2 + M^3)$

Question: How large should M be?

- Exact kernel ridge regression (KRR)
  - space complexity $O(n^2)$; time complexity $O(n^3)$
- RF-KRR & QMCF-KRR
  - space complexity $O(nM)$; time complexity $O(nM^2 + M^3)$

Question: How large should M be?

Short answer: Our QMC features require a smaller M.

To achieve the **same** error rate as the exact KRR:

1. RF-KRR: $M \asymp n^{\frac{2r}{2r+1}}$ (up to log factors)
2. QMCF-KRR: $M \asymp n^{\frac{1}{2r+1}}$ (up to log factors)

($r \in [1/2, 1]$: smoothness parameter of regression function)

Substantial improvement in smoother cases!

Notations

$\mathcal{H}$: Reproducing kernel Hilbert space (space of function consisting of $\mathrm{span}\{K(\mathbf{x},\cdot) : \mathbf{x} \in \mathcal{X}\}$ and their limits)

Integral operator $L : L^2(P_{\mathbf{X}}) \to L^2(P_{\mathbf{X}})$:

$$Lf(\mathbf{x}) := \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[K(\mathbf{X},\mathbf{x})f(\mathbf{X})\right].$$

Fact: $\operatorname{ran} L^{1/2} = \mathcal{H}$

Notations

$\mathcal{H}$: Reproducing kernel Hilbert space (space of function consisting of $\mathrm{span}\{K(\mathbf{x},\cdot) : \mathbf{x} \in \mathcal{X}\}$ and their limits)

Integral operator $L : L^2(P_{\mathbf{X}}) \to L^2(P_{\mathbf{X}})$:

$$Lf(\mathbf{x}) := \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[K(\mathbf{X},\mathbf{x})f(\mathbf{X})\right].$$

Fact: $\mathrm{ran}\, L^{1/2} = \mathcal{H}$

Assume: The true regression function is in $\mathrm{ran}\, L^r$ for some $r \in [1/2, 1]$.

(r: smoothness parameter)

Theorem: QMCF-KRR error rate

Assume

1. QMC condition holds: $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} |K(\mathbf{x},\mathbf{x}') - K_M(\mathbf{x},\mathbf{x}')| \leq C \cdot \frac{\log^a M}{M}$
2. Continuity conditions on the kernel
3. Standard Bernstein condition on the response Y
4. True regression $f_{\mathcal{H}} \in \arg\min_{f\in\mathcal{H}} \mathcal{E}(f)$ is in ran L^r, $r \in [1/2, 1]$

Let $\lambda = \tilde{C} n^{-\frac{1}{2r+1}} \in (0, e^{-1}]$. Then $M = \frac{\log^a(1/\lambda)}{\lambda} = n^{\frac{1}{2r+1}} \log^a(n^{\frac{1}{2r+1}}/\tilde{C})/\tilde{C}$ is enough to guarantee that, for any $\delta \in (0, 1]$, there exists n_0, such that when $n \geq n_0$, with probability at least $1 - \delta$, the QMCF-KRR excess risk

$$\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq C_1 n^{-\frac{2r}{2r+1}} \log^2 \frac{6}{\delta}.$$

$n^{-\frac{2r}{2r+1}}$: same error rate as in exact KRR (Caponnetto & De Vito, 2007) and RF-KRR (Rudi & Rosasco, 2017)

Summary of the paper

- **Goal:** Faster approximate computation of kernel methods using quasi-Monte Carlo methods.
- **Main Methodology:** Replace the Monte Carlo sequence in the random features approach (Rahimi & Recht, 2007) by quasi-Monte Carlo sequence.
- **Theoretical Guarantee:** With M features, the approximation error can be improved from $O_P(1/\sqrt{M})$ to $O(1/M)$ (up to logarithmic factors), for a class of kernels including Gaussian kernels.