Nested Expectations with Kernel Quadrature

Zonghao (Hudson) Chen

Department of Computer Science
University College London

March 14, 2025

Background: Quadrature

- In mathematics, statistics, machine learning, etc., people run into intractable expectations / integrals.

$$I = \mathbb{E}_{X\sim\pi}[h(X)] = \int_{\mathcal{X}} h(x)\pi(x)dx, \quad h : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}.$$

- π is some probability measure.
- How to approximate I given samples $x_1, \ldots, x_N$ and function evaluations $h(x_1), \ldots, h(x_N)$?
- Quadrature:

$$\hat{I} = \sum_{n=1}^{N} w_i h(x_i), \quad \hat{I} \approx I$$

- Monte Carlo!
  - $w_1 = \ldots = w_N = \frac{1}{N}$ are uniform weights.
  - Suppose $h \in L_2(\pi)$. When $x_1, \ldots, x_N$ are iid samples from π, we have

$$|\hat{I}_{MC} - I| = \mathcal{O}_P(N^{-1/2}).$$

Background: Quadrature

- Monte Carlo is relatively slow.
- In some applications, getting one sample / function evaluation would take hours.
  - Weather forecasts.

Does there exist a faster way of quadrature than Monte Carlo? Yes!

- By choosing smart weights w_i, $|\hat{I} - I| \to 0$ faster than Monte Carlo as $N \to \infty$.

$$\hat{I} = \sum_{n=1}^{N} w_i h(x_i).$$

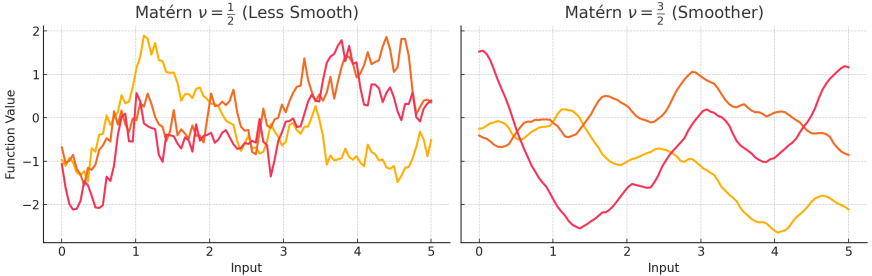
- Provided that h has some nice properties - smoothness!
- How to describe smoothness? RKHS!

Background: reproducing kernel Hilbert space

- Suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric positive definite function.
- There exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with k such that 1) $k(x, \cdot) \in \mathcal{H}$. 2) the reproducing property $\langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$.
- 'Intuitively', RKHS is a space of continuous functions of certain smoothness.
- Matern-ν kernels ($\nu = 1/2$, $\nu = 3/2$)

$$\exp(-\ell^{-1}\|x - x'\|), \quad (1 + \sqrt{3}\ell^{-1}\left\|x - x'\right\|)\exp(-\sqrt{3}\ell^{-1}\left\|x - x'\right\|).$$

Background: reproducing kernel Hilbert space



- Matern-ν RKHS is 'norm equivalent' to the Sobolev space $W_2^{\nu+d/2}(\mathcal{X})$.

$$W_2^s(\mathcal{X}) := \left\{ h \in L_2(\mathcal{X}) : D^\beta h \in L_2(\mathcal{X}) \text{ for all } \beta \in \mathbb{N}^d \text{ with } |\beta| \leq s \right\}, \quad s \in \mathbb{N}^+$$

- $f \in W_2^s(\mathcal{X})$ indicates that it has derivatives up to order s.
- We call them Sobolev reproducing kernels of order s if their RKHS 'is' $W_2^s(\mathcal{X})$.

Background: Quadrature

- The task is to approximate an intractable integral $I = \mathbb{E}_{X\sim\pi}[h(X)] = \int_{\mathcal{X}} h(x)\pi(x)dx$.
- Suppose $h \in W_2^s(\mathcal{X})$, a Sobolev space and also a RKHS $\mathcal{H}$ with reproducing kernel k.
- Denote $\hat{I} = \sum_{n=1}^{N} w_i h(x_i)$.

$$\begin{aligned}\left|I - \hat{I}\right| &= \left|\mathbb{E}_{X\sim\pi}[h(X)] - \sum_{n=1}^{N} w_n h(x_n)\right| \\ &= \left\langle h, \mathbb{E}_{X\sim\pi}[k(X,\cdot)] - \sum_{n=1}^{N} w_n k(x_n,\cdot)\right\rangle_{\mathcal{H}} \\ &\leq \|h\|_{\mathcal{H}} \left\|\mathbb{E}_{X\sim\pi}[k(X,\cdot)] - \sum_{n=1}^{N} w_n k(x_n,\cdot)\right\|_{\mathcal{H}}.\end{aligned}$$

- Optimal weights minimize $R(w_{1:N}) = \|\mathbb{E}_{X\sim\pi}[k(X,\cdot)] - \sum_{n=1}^{N} w_n k(x_n,\cdot)\|_{\mathcal{H}}$.

$$w_{1:N} = \mathbb{E}_{X\sim\pi}[k(X, x_{1:N})]\left(K(x_{1:N}, x_{1:N}) + N\lambda \boldsymbol{I}_N\right)^{-1}.$$

- Here $w_{1:N} = [w_1, \ldots, w_N]^\top \in \mathbb{R}^N$. $\lambda \geq 0$ helps improve numerical stability.

Background: Quadrature

- Kernel quadrature estimator $\hat{I}_{KQ}$ takes weighted average with optimal weights.

$$\hat{I}_{KQ} = \sum_{i=1}^{N} w_i h(x_i) = \mathbb{E}_{X\sim\pi}[k(X, x_{1:N})] \left(K(x_{1:N}, x_{1:N}) + N\lambda \boldsymbol{I}_N\right)^{-1} h(x_{1:N}).$$

- $|\hat{I}_{KQ} - I| = \mathcal{O}_P(N^{-\frac{s}{d}})$ provided that $h \in W_2^s(\mathcal{X})$ and $\mathcal{X} \subset \mathbb{R}^d$.
  - In contrast, standard Monte Carlo $|\hat{I}_{MC} - I| = \mathcal{O}_P(N^{-1/2})$.
- KQ is faster than MC when $s > \frac{d}{2}$.
  - KQ rate is minimax optimal: $|\hat{I}_{KQ} - I| = \Theta_P(N^{-s/d})$.
  - KQ suffers from curse of dimensionality.
  - KQ requires the tractable form of $\mathbb{E}_{X\sim\pi}[k(X, x_{1:N})]$. 'change of variable'
- Intuition: KQ achieves faster rate because 1) It utilizes smoothness of h. 2) There is no observation noise in quadrature settings.
- KQ can be extended to Bayesian quadrature.

Nested Expectation

- Let $\mathcal{X} \subseteq \mathbb{R}^{d_{\mathcal{X}}}$ and $\Theta \subseteq \mathbb{R}^{d_\Theta}$.

$$I = \mathbb{E}_{\theta \sim \mathbb{Q}} \left[f \left(\mathbb{E}_{X \sim \mathbb{P}_\theta} \left[g(X, \theta) \right] \right) \right], \quad g : \mathcal{X} \times \Theta \to \mathbb{R}, \quad f : \mathbb{R} \to \mathbb{R}$$

- $\mathbb{P}_\theta$ is any probability measure on $\mathcal{X}$ parameterized by θ. A simple case is the conditional distribution $\mathbb{P}(\cdot \mid \theta)$.
- I consists of two expectations. The inner expectation $J(\theta) = \mathbb{E}_{X \sim \mathbb{P}_\theta} [g(X, \theta)]$, and the outer expectation $I = \mathbb{E}_{\theta \sim \mathbb{Q}}[f(J(\theta))]$.
- When f is linear $f(t) = ct$ for $c \in \mathbb{R}$, I reduces to $c\mathbb{E}_{X,\theta} [g(X, \theta)]$.

Nested Expectation: Why do we care?

- Bayesian experimental design.

$$\int_{\mathcal{Y}} p(y \mid d) \log \left(\int_{\Theta} p(y \mid \theta, d) d\theta \right) dy.$$

- Acquisition function in Bayesian optimization.

$$\alpha(z; \mathcal{D}) := \mathbb{E}_{f_{|\mathcal{D}}} \left[g(f_{|\mathcal{D}}, z) + \max_{z'} \mathbb{E}_{f_{|\mathcal{D}'}} \left[g\left(f_{|\mathcal{D}'}, z'\right) \right] \right].$$

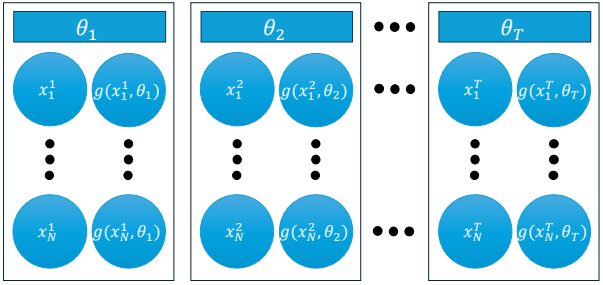
- Statistical divergences.
- Financial risk management .
- ...

Nested expectation

- Samples and function evaluations to estimate $I = \mathbb{E}_{\theta\sim\mathbb{Q}}\left[f\left(\mathbb{E}_{X\sim\mathbb{P}_\theta}\left[g(X,\theta)\right]\right)\right]$.

$$\theta_{1:T} := [\theta_1, \ldots, \theta_T]^\top \in \Theta^T,$$
$$x_{1:N}^{(t)} := \left[x_1^{(t)}, \ldots, x_N^{(t)}\right] \in \mathcal{X}^N,$$
$$g\left(x_{1:N}^{(t)}, \theta_t\right) := \left[g\left(x_1^{(t)}, \theta_t\right), \ldots, g\left(x_N^{(t)}, \theta_t\right)\right] \in \mathbb{R}^N,$$

- Total $\approx N \times T$ number of samples / function evaluations.



Nested Expectation

- The cost of getting samples / function evaluations is the dominating cost.
  - The computational complexity of a method is small by comparison.
- Efficiency: To reach Δ error, $|\hat{I} - I| \leq \Delta$, how many samples / function evaluations we need?

NMC	$\mathcal{O}(\Delta^{-4})$	NQMC	$\mathcal{O}(\Delta^{-2.5})$
MLMC	$\mathcal{O}(\Delta^{-2})$	**NKQ**	$\mathcal{O}\left(\Delta^{-\frac{d_{\mathcal{X}}}{s_{\mathcal{X}}}-\frac{d_{\Theta}}{s_{\Theta}}}\right)$

- Smaller exponents r in Δ^{-r} indicate a cheaper method.
- NKQ is the most efficient when the smoothness parameters $s_{\mathcal{X}}, s_{\Theta}$ are large.

Nested Monte Carlo

- $I = \mathbb{E}_{\theta \sim \mathbb{Q}} \left[f \left(\mathbb{E}_{X \sim \mathbb{P}_\theta} \left[g(X, \theta) \right] \right) \right]$.
- Nested Monte Carlo

$$\hat{I}_{\mathrm{NMC}} := \frac{1}{T} \sum_{t=1}^{T} f \left(\frac{1}{N} \sum_{n=1}^{N} g(x_n^{(t)}, \theta_t) \right)$$

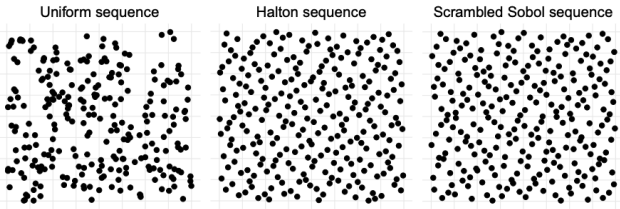
- It is proved that

$$\mathbb{E}[|I - \hat{I}_{\mathrm{NMC}}|] = \mathcal{O}\left(N^{-\frac{1}{2}} + T^{-\frac{1}{2}} \right)$$

- Expectation is taken over the randomness of samples.
- To reach error smaller than Δ, we need $N = \mathcal{O}(\Delta^{-2})$ and $T = \mathcal{O}(\Delta^{-2})$ so the total cost is $N \times T = \mathcal{O}(\Delta^{-4})$.
- For example, to reach error smaller than 0.01, we need 10^8 number of samples / function evaluations.

Nested Quasi Monte Carlo

- Given a uniform distribution $\pi = \mathbb{U}[0,1]^d$, Quasi Monte Carlo samples can cover the domain more uniformly than i.i.d samples.



- It is proved that the efficiency of *NQMC* is $\Delta^{-2.5}$, i.e to reach error smaller than Δ, NQMC needs $\mathcal{O}(\Delta^{-2.5})$ number of samples / function evaluations.
- QMC samples only work for domains like $[0,1]^d$ or simple transformations thereof.

Nested Kernel Quadrature

Does there exist a faster way than Nested (Quasi) Monte Carlo? Yes!

$$\hat{I}_{\mathrm{NMC}} := \frac{1}{T}\sum_{t=1}^{T} f\left(\frac{1}{N}\sum_{n=1}^{N} g(x_n^{(t)}, \theta_t)\right), \quad \hat{I}_{\mathrm{NKQ}} := \sum_{t=1}^{T} w_t^{\Theta} f\left(\sum_{n=1}^{N} w_{n,t}^{\mathcal{X}} g(x_n^{(t)}, \theta_t)\right)$$

- NMC used uniform weights $w_{n,t}^{\mathcal{X}} = \frac{1}{N}$ and $w_t^{\Theta} = \frac{1}{T}$.
- Kernel quadrature weights.

Nested Kernel Quadrature

- **Stage I**: For each $t \in \{1, \ldots, T\}$, we estimate the inner conditional expectation J evaluated at θ_t with N observations $x_{1:N}^{(t)}$ and $g(x_{1:N}^{(t)}, \theta_t)$ using a KQ estimator:

$$\hat{J}_{KQ}(\theta_t) := \mathbb{E}_{X \sim \mathbb{P}_{\theta_t}}[k_{\mathcal{X}}(X, x_{1:N}^{(t)})](\boldsymbol{K}_{\mathcal{X}}^{(t)} + N\lambda_{\mathcal{X}} \boldsymbol{I}_N)^{-1} g(x_{1:N}^{(t)}, \theta_t).$$

  - The inner expectation $J(\theta) = \mathbb{E}_{X \sim \mathbb{P}_\theta}[g(X, \theta)]$ is a standard expectation.
  - $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$. $\boldsymbol{K}_{\mathcal{X}}^{(t)}$ is the $N \times N$ Gram matrix.
  - $\mathcal{O}(N^{-\frac{s_{\mathcal{X}}}{d_{\mathcal{X}}}})$ rate of convergence for each $J(\theta_t)$ with $t = 1, \ldots, T$.

- **Stage II**: We use KQ again by treating $\hat{F}_{KQ}(\theta_t) = f(\hat{J}_{KQ}(\theta_t))$ as the observation for the outer expectation $I = \mathbb{E}_{\theta \sim \mathbb{Q}}[f(J(\theta))]$.

$$\hat{I}_{NKQ} := \mathbb{E}_{\theta \sim \mathbb{Q}}[k_\Theta(\theta, \theta_{1:T})](\boldsymbol{K}_\Theta + T\lambda_\Theta \boldsymbol{I}_T)^{-1} \hat{F}_{KQ}(\theta_{1:T}).$$

  - $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ. $\boldsymbol{K}_\Theta$ is the $T \times T$ Gram matrix.
  - The same rate of convergence $\mathcal{O}(T^{-\frac{s_\Theta}{d_\Theta}})$ holds even with $\hat{F}_{KQ}(\theta_t)$ rather than $F(\theta_t)$.

Nested Kernel Quadrature

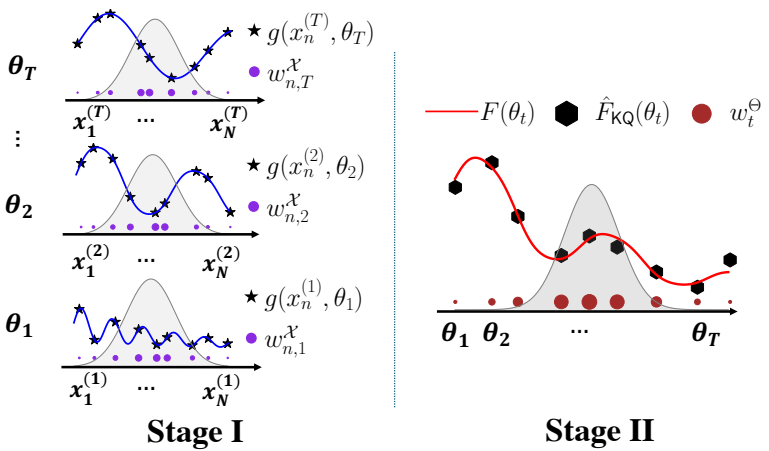


Figure: *Illustration of NKQ.* In stage I, we estimate $J(\theta_t)$ using $\hat{J}_{KQ}(\theta_t) = \sum_{n=1}^{N} w_{n,t}^{\mathcal{X}} g(x_n^{(t)}, \theta_t)$ for all $t \in \{1, \ldots, T\}$. In stage II, we estimate I with $\hat{I}_{NKQ} = \sum_{t=1}^{T} w_t^{\Theta} \hat{F}_{KQ}(\theta_t)$ where $\hat{F}_{KQ}(\theta_t) = f(\hat{J}_{KQ}(\theta_t))$. The shaded areas depict $\mathbb{P}_\theta$ (for stage I) and $\mathbb{Q}$ (for stage II).

Nested Kernel Quadrature: Theory

Theorem

Let $\mathcal{X} = [0,1]^{d_\mathcal{X}}$ and $\Theta = [0,1]^{d_\Theta}$. Suppose that $k_\mathcal{X}$ and k_Θ are Sobolev kernels of smoothness $s_\mathcal{X} > d_\mathcal{X}/2$ and $s_\Theta > d_\Theta/2$, and that the following conditions hold,

(1) *For any $\theta \in \Theta$ and any $\beta \in \mathbb{N}^{d_\Theta}$ with $|\beta| \leq s_\Theta$, $D_\theta^\beta g(\cdot, \theta) \in W_2^{s_\mathcal{X}}(\mathcal{X})$.*

(2) *For any $x \in \mathcal{X}$, $g(x, \cdot) \in W_2^{s_\Theta}(\Theta)$ and $\theta \mapsto p_\theta(x) \in W_2^{s_\Theta}(\Theta)$.*

(3) *$f \in C^{s_\Theta+1}(\mathbb{R})$.*

$$\left| I - \hat{I}_{NKQ} \right| \leq \tau \left(C_1 N^{-\frac{s_\mathcal{X}}{d_\mathcal{X}}} (\log N)^{\frac{s_\mathcal{X}+1}{d_\mathcal{X}}} + C_2 T^{-\frac{s_\Theta}{d_\Theta}} (\log T)^{\frac{s_\Theta+1}{d_\Theta}} \right),$$

holds with probability at least $1 - 4e^{-\tau}$.

- Convergence in high probability!
- The stage II observations $\{\hat{F}_{KQ}(\theta_t)\}_{t=1}^T$ can be treated as noiseless. The additional error it introduces is the same order as the stage I error $\tilde{\mathcal{O}}(N^{-\frac{s_\mathcal{X}}{d_\mathcal{X}}})$ and is therefore subsumed by it.

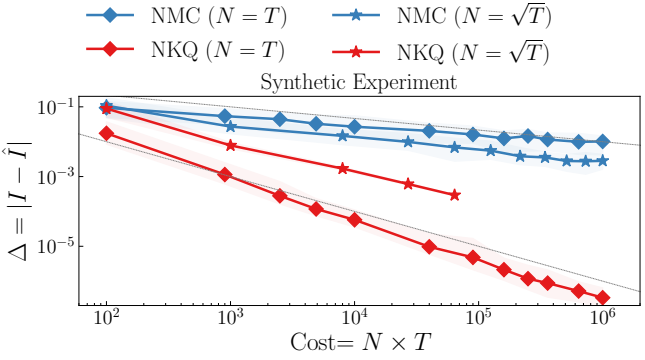
Nested Kernel Quadrature

$$\left| I - \hat{I}_{NKQ} \right| = \tilde{\mathcal{O}}_P \left(N^{-\frac{s_{\mathcal{X}}}{d_{\mathcal{X}}}} + T^{-\frac{s_\Theta}{d_\Theta}} \right), \quad \text{Efficiency} = \tilde{\mathcal{O}}(\Delta^{-\frac{d_{\mathcal{X}}}{s_{\mathcal{X}}} - \frac{d_\Theta}{s_\Theta}})$$

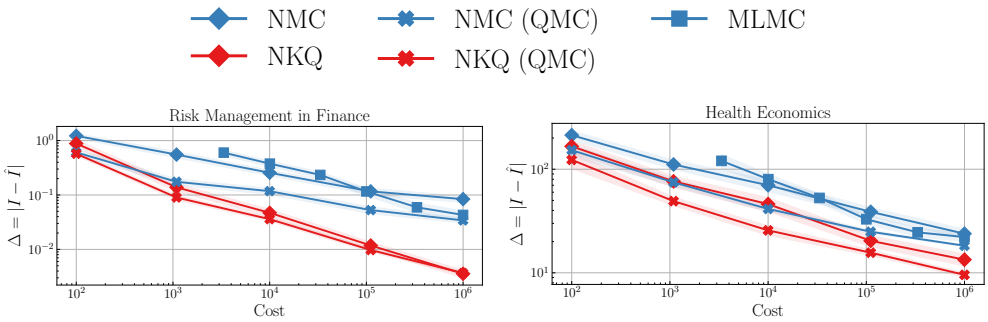
- To reach error smaller than Δ, we need $N = \tilde{\mathcal{O}}(\Delta^{-\frac{d_{\mathcal{X}}}{s_{\mathcal{X}}}})$ and $T = \tilde{\mathcal{O}}(\Delta^{-\frac{d_\Theta}{s_\Theta}})$ so the total cost is $N \times T = \tilde{\mathcal{O}}(\Delta^{-\frac{d_{\mathcal{X}}}{s_{\mathcal{X}}} - \frac{d_\Theta}{s_\Theta}})$.
- NKQ can be extended to its Bayesian counterpart to obtain finite-sample uncertainty.
  - Question: How to propagate Stage I uncertainty to Stage II?
  - Answer: Linearization of f.

Synthetic Experiment

- $\mathbb{Q} = \mathrm{U}[0,1]$, $\mathbb{P}_\theta = \mathrm{U}[0,1]$, $g(x,\theta) = x^{\frac{5}{2}} + \theta^{\frac{5}{2}}$, and $f(z) = z^2$
- In this case $I = 0.4115$ can be computed analytically.
- $s_{\mathcal{X}} = s_\Theta = 2$, $d_{\mathcal{X}} = d_\Theta = 1$.
- The grey line represents the theory.

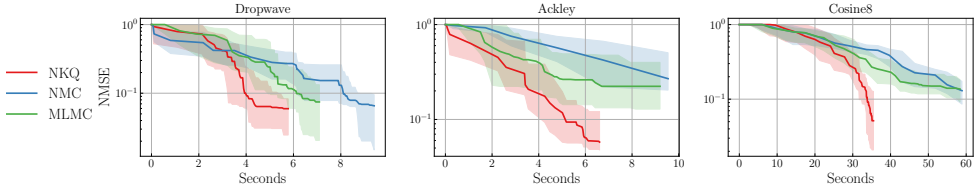


Other Experiments



- Detailed settings can be found in the paper.
- NKQ (QMC) represents nested kernel quadrature with Quasi-Monte Carlo samples.

Bayesian Optimization



- Bayesian Optimization with look-ahead acquisition functions.

$$\alpha(z; \mathcal{D}) := \mathbb{E}_{f_{|\mathcal{D}}}\left[g(f_{|\mathcal{D}}, z) + \max_{z'} \mathbb{E}_{f_{|\mathcal{D}'}}\left[g\left(f_{|\mathcal{D}'}, z'\right)\right]\right],$$

- $f_{|\mathcal{D}}, f_{|\mathcal{D}'}$ are the GP posteriors.
- The reward g is the q-expected improvement with $q = 2$.

$$g(f_{|\mathcal{D}}, z) = \max_{j=1,2}(f_{|\mathcal{D}}(z_j) - r_{\max}), 0), \quad z = (z_1, z_2).$$

- The integral is computed with respect to a 2 dimensional Gaussian distribution, hence NKQ works pretty well!

Conclusions

- To approximate nested expectation

$$I = \mathbb{E}_{\theta \sim \mathbb{Q}} \left[f \left(\mathbb{E}_{X \sim \mathbb{P}_\theta} \left[g(X, \theta) \right] \right) \right]$$

- We propose a new method nested kernel quadrature (NKQ).
- We prove a faster rate of convergence than baselines when the problem has sufficient amount of smoothness.
- The theory is confirmed in several experiments.

Multi-level Monte Carlo

- Decompose the challenging problem of estimating I into the sum of easier problems of multiple fidelity level $\ell \in \{0, \ldots, L\}$.
- At each level ℓ, we are given T_ℓ samples $\theta_{1:T_\ell}$ sampled i.i.d from $\mathbb{Q}$ and we have N_ℓ samples $x_{1:N_\ell}^{(\theta_t)}$ sampled i.i.d from $\mathbb{P}_{\theta_t}$ for each $t = 1, \ldots, T_\ell$.

$$\hat{I}_{MLMC} := \sum_{l=1}^{L} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} (f(J_{\ell,t}) - f(J_{\ell-1,t})) + \frac{1}{T_0} \sum_{t=1}^{T_0} f(J_{0,t})$$

$$\text{where } J_{\ell,t} := \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} g(x_n^{(t)}, \theta_t) \text{ for } \ell \in \{0, \ldots, L\},$$

Multi-level Monte Carlo

- The total cost and expected absolute error

$$\text{Cost} = \mathcal{O}\left(\sum_{\ell=0}^{L} N_\ell \times T_\ell\right), \qquad \mathbb{E}|I - \hat{I}_{MLMC}| = \mathcal{O}\left(\sum_{\ell=0}^{L} N_\ell^{-1} \times T_\ell^{-\frac{1}{2}}\right).$$

- In order to reach error threshold Δ, one can take $N_\ell \propto 2^\ell$ and $T_\ell \propto 2^{-2\ell}\Delta^{-2}$. Therefore, one has

$$\mathbb{E}|I - \hat{I}_{MLMC}| = \mathcal{O}(\Delta), \quad \text{Efficiency} = \mathcal{O}(\Delta^{-2}).$$

- NKQ can be combined with multi-level construction as well (MLKQ).

$$\mathbb{E}|I - \hat{I}_{MLKQ}| = \mathcal{O}(\Delta), \quad \text{Efficiency} = \mathcal{O}(\Delta^{-1-\frac{d_\mathcal{X}}{2s_\mathcal{X}}-\frac{d_\Theta}{2s_\Theta}}).$$

- The proved theoretical fast rate is not verfied in experiments.

Future Work / Directions

- Extend NKQ to its Bayesian counterpart and use active learning to further improve efficiency.
- More practical exploration of Multi-level kernel quadrature (MLKQ).