Discrepancy-based Inference for Intractable Generative Models using QMC

Ziang Niu, Johanna Meier, François-Xavier Briol

contact: ziangniu6@gmail.com **code:** `https://github.com/johannnamr/Discrepancy-based-inference-using-QMC`

Statistical Inference for Intractable Generative Models

__Intractable Generative Models:__ Intractable generative models are models for which the **likelihood is unavailable** but **sampling is possible**. One is required to compute some discrepancy between the data and the generative model when doing inference.

__Minimum Distance Estimators:__ Once a discrepancy is defined, one can easily obtain the **Minimum Distance Estimators (MDE)**. Given the dataset $\{y_j\}_{j=1}^m \overset{IID}{\sim} \mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and generator G_θ such that $x = G_\theta \sim \mathbb{P}_\theta \in \mathcal{P}(\mathcal{X})$, one can construct an estimator through the framework of MDE:

$$\hat{\theta}_m^D = \arg\min_{\theta\in\Theta} D(\mathbb{P}_\theta, \mathbb{Q}^m)$$

where $\mathbb{Q}^m = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}(x)$. A common approach is to solve the optimisation problem through evaluations of $\widehat{D}(\mathbb{P}_\theta^n, \mathbb{Q}^m)$ instead of the unknown optimisation problem. A closely related discrepancy family is **Integral Probability Metrics (IPMs)**. Given a set of functions $\mathcal{F}$, an IPM is a probability metric which takes the form:

$$D_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \int_{\mathcal{X}} f(x)\mathbb{Q}(dx) \right|$$

Popular metrics include Maximum Mean Discrepancy (MMD) and Wasserstein Distance:

1. **MMD:** Let $\mathcal{F} = f : \mathcal{X} \to \mathbb{R} : \|f\|_{\mathcal{H}_k} \le 1$, the unit-ball of a RKHS $\mathcal{H}_k$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
2. **$p-$Wasserstein Distance:** When $p = 1$, Wasserstein distance is an IPM with $\mathcal{F} = \{f : X \to \mathbb{R} \text{ s.t.} \forall x, y \in \mathcal{X}, |f(x) - f(y)| \le c(x, y)\}$.

Other popular divergences include **Sinkhorn divergence**($S_{c,p,\lambda}$), a regularized version of Wasserstein distance and **Sliced Wasserstein distance**($SW_{c,p}$), which works better for high-dimensional setting.

__Sample Complexity:__ Consider D is a metric

$$\begin{aligned}|D(\mathbb{P}_\theta^n, \mathbb{Q}^m) - D(\mathbb{P}_\theta, \mathbb{Q})| &\le |D(\mathbb{P}_\theta^n, \mathbb{Q}^m) - D(\mathbb{P}_\theta, \mathbb{Q}^m)| \\ &\quad + |D(\mathbb{P}_\theta, \mathbb{Q}^m) - D(\mathbb{P}_\theta, \mathbb{Q})| \\ &\le D(\mathbb{P}_\theta^n, \mathbb{P}_\theta) + D(\mathbb{Q}, \mathbb{Q}^m)\end{aligned}$$

Sample complexity $D(\mathbb{P}_\theta^n, \mathbb{P}_\theta)$ plays a key role here!

__Issue with Previous Method:__ $D(\mathbb{P}_\theta^n, \mathbb{Q}^m)$ invovles choosing $\mathbb{P}_\theta^n$ and a usual choice is **Monte Carlo** estimand, i.e. sampling IID data $\{x_i\}_{i=1}^n$ from $\mathbb{P}_\theta$. The sample complexity for MC is $D(\mathbb{P}_\theta^n, \mathbb{P}_\theta) = \boldsymbol{O_p(n^{-1/2})}$, which can be expensive when requiring high accuracy.

Enhancing Sample Complexity via Quasi-Monte Carlo

__(Randomized) Quasi-Monte Carlo:__ The essence of **(R)QMC** sampling is to generate a more "diverse" set of samples from the model (see right figures).

__Faster Convergence Rate:__ A nice theoretical result can be obtained if the integrand f is smooth enough and that domain $\mathcal{U}$ is regular: for any $\epsilon > 0$

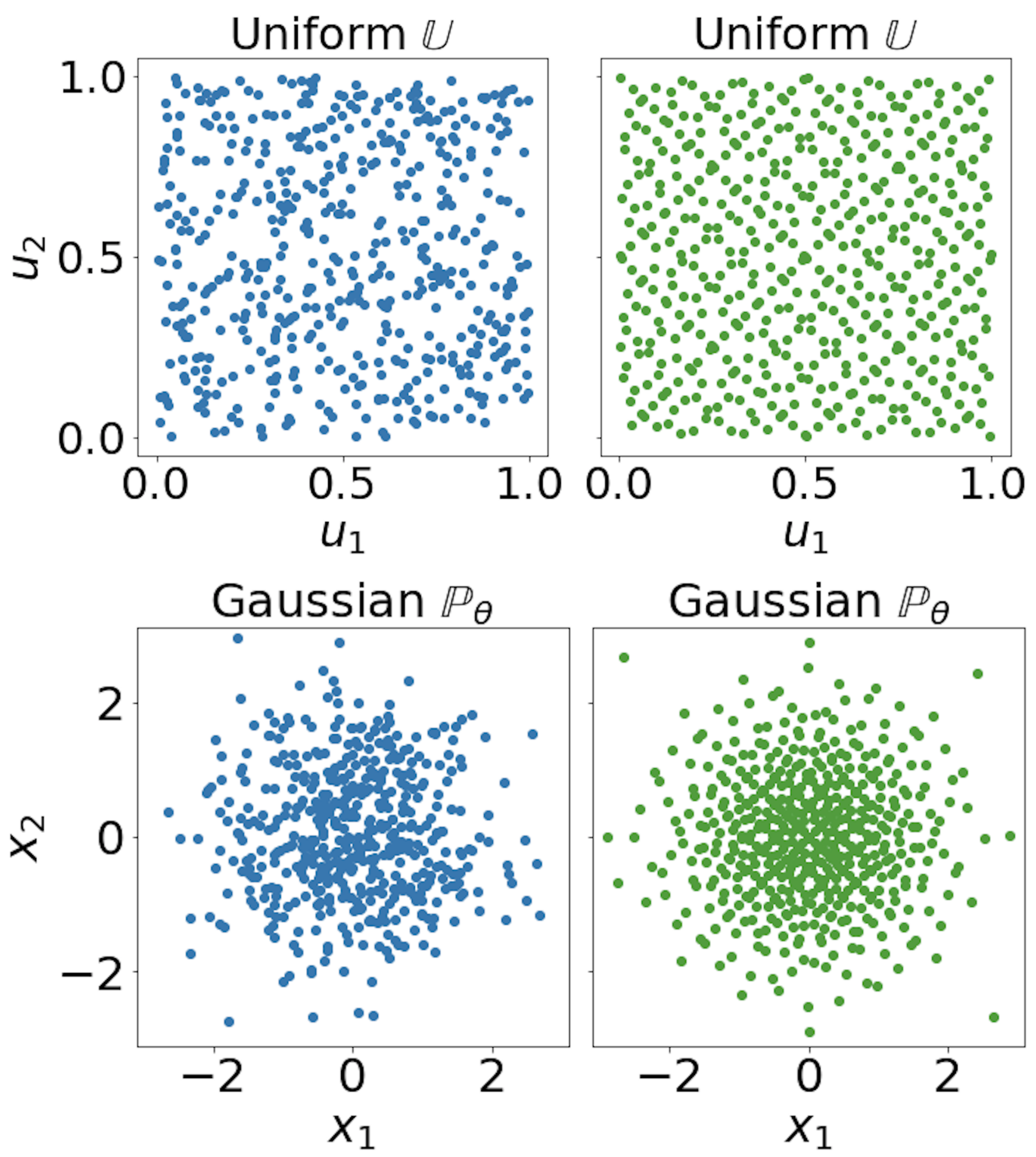
$$\left| \int_{\mathcal{U}} f(u)du - \frac{1}{n}\sum_{i=1}^n f(u_i) \right| = \boldsymbol{O(n^{-1+\epsilon})}$$

where $\{u_i\}_{i=1}^n$ is a **low discrepancy point set**.

__Sample Complexity Improvement:__

IDEA: *Replace MC points to estimate discrepancies with QMC/RQMC points.*

Consider the generator G_θ is smooth enough and $\mathcal{X}$ is regular enough, we could expect $D(\mathbb{P}_\theta, \mathbb{P}_\theta^n) = O(n^{-1+\epsilon})$, which is a great improvement compared with MC.



Numerical Results

__Bivariate Beta Distributions:__ Let $\lfloor x \rfloor$ as the integer part of some $x \in \mathbb{R}$ and consider

$$G_\theta^1 := \frac{\tilde{u}_1 + \tilde{u}_3}{\tilde{u}_1 + \tilde{u}_3 + \tilde{u}_4 + \tilde{u}_5}, G_\theta^2 := \frac{\tilde{u}_2 + \tilde{u}_4}{\tilde{u}_2 + \tilde{u}_3 + \tilde{u}_4 + \tilde{u}_5}$$

where

$$\tilde{u}_i = -\sum_{k=1}^{\lfloor\theta_i\rfloor} \ln(u_{ik}) + u_{i0}, u_{i0} \sim \text{Gamma}(\theta_i - \lfloor\theta_i\rfloor, 1)$$

and

$$u = (u_{11}, \ldots, u_{1\lfloor\theta_1\rfloor}, u_{21}, \ldots, u_{5,\lfloor\theta_5\rfloor}) \sim \text{Unif}([0,1]^s)$$

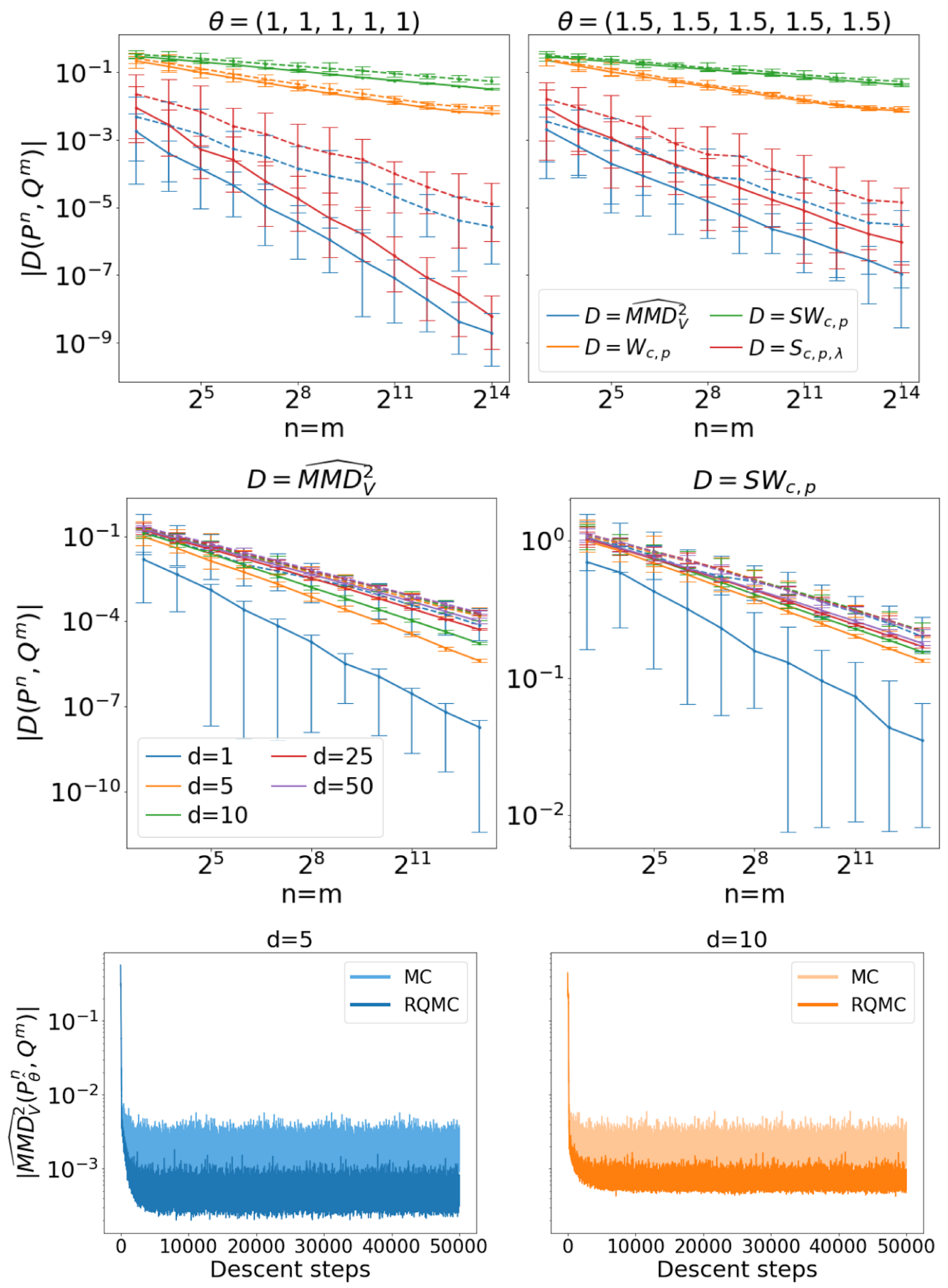
where $s = \sum_{i=1}^5 \lfloor\theta_i\rfloor$.

__Inference for Multivariate g-and-k Models:__ The generator for g-and-k model is

$$G_\theta(u) := \theta_1 + \theta_2\left(1 + 0.8\frac{(1 - \exp(-\theta_3 z))}{(1 + \exp(-\theta_3 z))}\right)(1 + z^2)^{\theta_4} z$$

where $z = \Sigma^{\frac{1}{2}}\Phi^{-1}(u)^\top, u \sim \text{Unif}([0,1]^d)$. Σ is a symmetric Toepliz matrix with diagonal entries equal to 1 and subdiagonals equal to θ_5 and Φ^{-1} is the inverse CDF of Gaussian.

WARNING: The performance gets worse as dimension grows due to the convergence $(\log(n)^s)n^{-1}$



Theoretical Results

Assumption 1. *Given a model $\mathbb{P}_\theta$ with $(G_\theta, [0,1]^s)$, we have access to $x_i = G_\theta(u_i), i = 1, \ldots, n$ where $\{u_i\}_{i=1}^n \subset [0,1]^s$ form a QMC or RQMC point set.*

Assumption 2. *Suppose that $\mathcal{X} \subset \mathbb{R}^d$ is a compact domain and that $G_\theta : [0,1]^s \to \mathcal{X}$ satisfies:*

- *$\partial^{(1,\ldots,1)}(G_\theta)_j \in \mathcal{C}([0,1]^s)$ for all $j = 1, \ldots, d$.*
- *$\partial^v(G_\theta)_j(\cdot, : 1_{-v}) \in L^{p_j}([0,1]^{|v|})$ for all $j = 1, \ldots, d$ and $v \in \{0,1\}^s \backslash (0, \ldots, 0)$, where $p_j \in [1, \infty]$ and $\sum_{j=1}^d p_j^{-1} \le 1$.*

Theorem 1 (MMD). *Let $k \in \mathcal{C}^{s\times s}(\mathcal{X}), \mathbb{P}_\theta \in \mathcal{P}_k(\mathcal{X})$ and suppose Assumption 1-2 hold. Then,*

$$\text{MMD}(\mathbb{P}_\theta, \mathbb{P}_\theta^n) = O(n^{-1+\epsilon}) \quad \forall\epsilon > 0$$

Theorem 2 (Wasserstein). *Let $\mathbb{P}_\theta \in \mathcal{P}_{c,1}(\mathcal{X})$ where c is a metric on $\mathcal{X}$ and suppose Assumption 1 holds with $s = d = 1$. Further, assume $V_{HK}(G_\theta) < \infty$. Then,*

$$W_{c,1}(\mathbb{P}_\theta, \mathbb{P}_\theta^n) = O(n^{-1+\epsilon}) \quad \forall\epsilon > 0$$

Theorem 3 (Sinkhorn). *Let $c \in \mathcal{C}^{\infty,\infty}(\mathcal{X} \times \mathcal{X})$ and suppose $\mathbb{P}_\theta, \mathbb{Q} \in \mathcal{P}_{c,p}(\mathcal{X})$. Further, suppose Assumptions 1-2 hold. Then*

$$|S_{c,p,\lambda}(\mathbb{P}_\theta, \mathbb{Q}) - S_{c,p,\lambda}(\mathbb{P}_\theta^n, \mathbb{Q})| = O(n^{-1+\epsilon}) \quad \forall\epsilon > 0$$

More technical details and experiments can be found in the paper: https://arxiv.org/abs/2106.11561