Optimally-weighted Estimators of the Maximum Mean Discrepancy for Likelihood-free Inference

Ayush Bharti[1], Masha Naslidnyk[2], Oscar Key[2], Samuel Kaski[1,3], François-Xavier Briol[2,4]

Department of Computer Science, Aalto University, Finland[1]
Department of Statistical Science, University College London, UK[2]
Department of Computer Science,University of Manchester, UK[3]
The Alan Turing Institute, UK[4]

June 4, 2023

Outline

Setting

Inference for simulator-based models with intractable likelihoods.

- Data $\{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Simulator $\mathcal{P}_\theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$, characterised through generative process $(G_\theta, \mathbb{U})$, where $G_\theta : \mathcal{U} \to \mathcal{X}$ and $\mathbb{U}$ is a distribution on $\mathcal{U} \subset \mathbb{R}^s$
- The likelihood associated to $\mathbb{P}_\theta$ is unknown
- We can sample $y \sim \mathbb{P}_\theta$ by
    1. Sampling $u \sim \mathbb{U}$, $\mathbb{U}$ being uniform or Gaussian distribution
    2. Applying the generator $y = G_\theta(u)$

Setting

Inference for simulator-based models with intractable likelihoods.

- Data $\{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Simulator $\mathcal{P}_\theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$, characterised through generative process $(G_\theta, \mathbb{U})$, where $G_\theta : \mathcal{U} \to \mathcal{X}$ and $\mathbb{U}$ is a distribution on $\mathcal{U} \subset \mathbb{R}^s$
- The likelihood associated to $\mathbb{P}_\theta$ is unknown
- We can sample $y \sim \mathbb{P}_\theta$ by
    1. Sampling $u \sim \mathbb{U}$, $\mathbb{U}$ being uniform or Gaussian distribution
    2. Applying the generator $y = G_\theta(u)$

Discrepancy-based Likelihood-free Inference

Suppose $\mathcal{D}$ is a discrepancy measure between probability distributions.

Approximate Bayesian computation
Allows sampling from the approximate posterior $p(\theta|\mathcal{D}(\mathbb{P}_\theta, \mathbb{Q}^n) < \epsilon)$.

Minimum distance estimation:
Solve the optimisation problem

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

However, we can't compute $\mathcal{D}(\mathbb{P}_\theta, \mathbb{Q}^n)$ but can only estimate it given samples $\{y_i\}_{i=1}^m$ from $\mathbb{P}_\theta$
$\Rightarrow$ estimating $\mathcal{D}$ accurately in few samples is key!

Discrepancy-based Likelihood-free Inference

Suppose $\mathcal{D}$ is a discrepancy measure between probability distributions.

Approximate Bayesian computation
Allows sampling from the approximate posterior $p(\theta|\mathcal{D}(\mathbb{P}_\theta, \mathbb{Q}^n) < \epsilon)$.

Minimum distance estimation:
Solve the optimisation problem

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

However, we can't compute $\mathcal{D}(\mathbb{P}_\theta, \mathbb{Q}^n)$ but can only estimate it given samples $\{y_i\}_{i=1}^m$ from $\mathbb{P}_\theta$
$\Rightarrow$ estimating $\mathcal{D}$ accurately in few samples is key!

Choice of discrepancy $\mathcal{D}$

We want discrepancy measure that can be estimated efficiently from samples

- $\mathcal{D}(\mathbb{P}_\theta, \mathbb{Q}^n) \approx \mathcal{D}(\mathbb{P}_\theta^m, \mathbb{Q}^n)$
- Efficient in terms of *sample complexity*

Popular discrepancies for likelihood-free inference:

- KL divergence [Jiang, 2018]
- Wasserstein distance [Bernton et al., 2019]
- Sinkhorn divergence [Genevay et al., 2019]
- Classification accuracy [Gutmann et al., 2017]
- **Maximum mean discrepancy** [Gretton et al., 2012]

Choice of discrepancy $\mathcal{D}$

We want discrepancy measure that can be estimated efficiently from samples

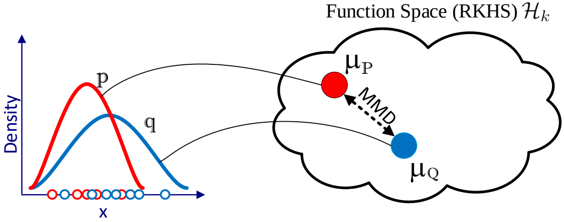
- $\mathcal{D}(\mathbb{P}_\theta, \mathbb{Q}^n) \approx \mathcal{D}(\mathbb{P}_\theta^m, \mathbb{Q}^n)$
- Efficient in terms of *sample complexity*

Popular discrepancies for likelihood-free inference:

- KL divergence [Jiang, 2018]
- Wasserstein distance [Bernton et al., 2019]
- Sinkhorn divergence [Genevay et al., 2019]
- Classification accuracy [Gutmann et al., 2017]
- **Maximum mean discrepancy** [Gretton et al., 2012]

Maximum mean discrepancy (MMD)

Maximum mean discrepancy (MMD) is a notion of distance between probability distributions.

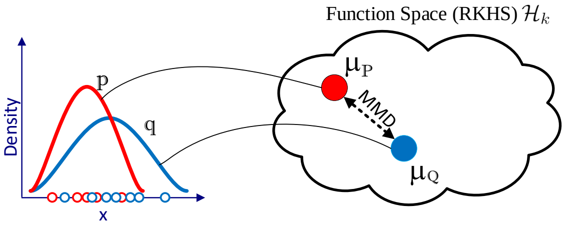


Advantages of MMD:

- Sample complexity of $\mathcal{O}(m^{-1/2})$, better than its alternatives
- Desirable properties — leads to consistent and robust estimators
- Applicable on any data type for which a kernel can be defined
- Hence, it is used in many likelihood-free inference frameworks, e.g. [Park et al., 2015, Briol et al., 2019, Dellaporta et al., 2022]

Maximum mean discrepancy (MMD)

Maximum mean discrepancy (MMD) is a notion of distance between probability distributions.



Advantages of MMD:

- Sample complexity of $\mathcal{O}(m^{-1/2})$, better than its alternatives
- Desirable properties — leads to consistent and robust estimators
- Applicable on any data type for which a kernel can be defined
- Hence, it is used in many likelihood-free inference frameworks, e.g. [Park et al., 2015, Briol et al., 2019, Dellaporta et al., 2022]

Estimating Maximum Mean Discrepancy

For a reproducing kernel k, the MMD between distributions $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathrm{MMD}_k^2(\mathbb{P}^m, \mathbb{Q}^n) = \mathbb{E}_{y,y' \sim \mathbb{P}}[k(y, y')] - 2\mathbb{E}_{y \sim \mathbb{P}, x \sim \mathbb{Q}}[k(x, y)] + \mathbb{E}_{x,x' \sim \mathbb{Q}}[k(x, x')]$$

V-statistic estimator for MMD computed using samples $\{y_i\}_{i=1}^m \sim \mathbb{P}$ and $\{x_i\}_{i=1}^n \sim \mathbb{Q}$:

$$\mathrm{MMD}_k^2(\mathbb{P}^m, \mathbb{Q}^n) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(y_i, y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j)$$

Sample complexity: $\mathcal{O}(m^{-1/2} + n^{-1/2})$

- n: no. of observed data samples (fixed)
- To reduce error by half, need four times the no. of samples from $\mathbb{P}_\theta$

Estimating Maximum Mean Discrepancy

For a reproducing kernel k, the MMD between distributions $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathrm{MMD}_k^2(\mathbb{P}^m, \mathbb{Q}^n) \quad = \quad \mathbb{E}_{y,y' \sim \mathbb{P}}[k(y, y')] \quad - \quad 2\mathbb{E}_{y \sim \mathbb{P}, x \sim \mathbb{Q}}[k(x, y)] \quad + \quad \mathbb{E}_{x,x' \sim \mathbb{Q}}[k(x, x')]$$

V-statistic estimator for MMD computed using samples $\{y_i\}_{i=1}^m \sim \mathbb{P}$ and $\{x_i\}_{i=1}^n \sim \mathbb{Q}$:

$$\mathrm{MMD}_k^2(\mathbb{P}^m, \mathbb{Q}^n) \quad = \quad \frac{1}{m^2} \sum_{i,j=1}^{m} k(y_i, y_j) \quad - \quad \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j) \quad + \quad \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j)$$

Sample complexity: $\mathcal{O}(m^{-1/2} + n^{-1/2})$

- n: no. of observed data samples (fixed)
- To reduce error by half, need four times the no. of samples from $\mathbb{P}_\theta$

Estimating Maximum Mean Discrepancy

For a reproducing kernel k, the MMD between distributions $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathrm{MMD}_k^2(\mathbb{P}^m, \mathbb{Q}^n) = \mathbb{E}_{y,y' \sim \mathbb{P}}[k(y, y')] - 2\mathbb{E}_{y \sim \mathbb{P}, x \sim \mathbb{Q}}[k(x, y)] + \mathbb{E}_{x,x' \sim \mathbb{Q}}[k(x, x')]$$

V-statistic estimator for MMD computed using samples $\{y_i\}_{i=1}^m \sim \mathbb{P}$ and $\{x_i\}_{i=1}^n \sim \mathbb{Q}$:

$$\mathrm{MMD}_k^2(\mathbb{P}^m, \mathbb{Q}^n) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j)$$

Sample complexity: $\mathcal{O}(m^{-1/2} + n^{-1/2})$

- n: no. of observed data samples (fixed)
- To reduce error by half, need four times the no. of samples from $\mathbb{P}_\theta$

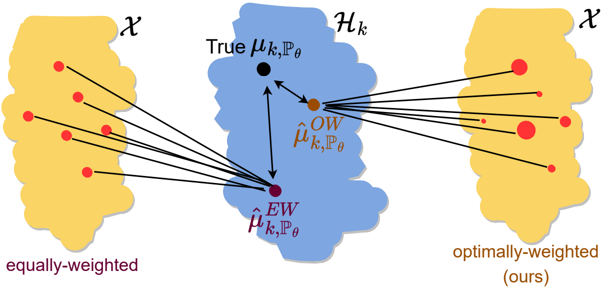
Outline

Proposed Optimally-weighted Estimator of MMD

We estimate the MMD as

$$\mathrm{MMD}_k^2(\mathbb{P}_\theta^{m,w}, \mathbb{Q}^n) = \sum_{i,j=1}^{m} w_i w_j k(y_i, y_j) - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_j k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j),$$

where the weights w_i chosen *optimally*.



Deriving the optimal weights

Let $\mathbb{P}_\theta^{m,w} = \sum_{i=1}^m w_i \delta_{y_i} = \sum_{i=1}^m w_i \delta_{G_\theta(u_i)}$.

Using the reverse triangle inequality, we get

$$\left|\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})\right| \leq \mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}).$$

Let $c : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be a reproducing kernel s.t. $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$.

Then, the error upper bounded can be written as (see paper for proof):

$$\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}) = K \times \mathrm{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right).$$

The weights minimising this upper bound are:

$$w^* = \operatorname*{arg\,min}_{w \in \mathbb{R}^m} \mathrm{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right)$$

Weights can be obtained in closed-form as $\mathbb{U}$ is a simple distribution.

Deriving the optimal weights

Let $\mathbb{P}_\theta^{m,w} = \sum_{i=1}^m w_i \delta_{y_i} = \sum_{i=1}^m w_i \delta_{G_\theta(u_i)}$.

Using the reverse triangle inequality, we get

$$\left|\text{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \text{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})\right| \leq \text{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}).$$

Let $c : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be a reproducing kernel s.t. $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$.

Then, the error upper bounded can be written as (see paper for proof):

$$\text{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}) = K \times \text{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right).$$

The weights minimising this upper bound are:

$$w^* = \arg\min_{w \in \mathbb{R}^m} \text{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right)$$

Weights can be obtained in closed-form as $\mathbb{U}$ is a simple distribution.

Deriving the optimal weights

Let $\mathbb{P}_\theta^{m,w} = \sum_{i=1}^m w_i \delta_{y_i} = \sum_{i=1}^m w_i \delta_{G_\theta(u_i)}$.

Using the reverse triangle inequality, we get

$$\left| \mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q}) \right| \leq \mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}).$$

Let $c : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be a reproducing kernel s.t. $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$.

Then, the error upper bounded can be written as (see paper for proof):

$$\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}) = K \times \mathrm{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right).$$

The weights minimising this upper bound are:

$$w^* = \underset{w \in \mathbb{R}^m}{\arg\min}\ \mathrm{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right)$$

Weights can be obtained in closed-form as $\mathbb{U}$ is a simple distribution.

Theoretical guarantees

Assumptions:

- c is a Matérn kernel of order ν_c on $\mathcal{U} \subset \mathbb{R}^s$
- k is Matérn or squared-exponential kernel of order ν_k
- $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ holds

Sample complexity result for our estimator:

$$\left|\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})\right| = \mathcal{O}(m^{-\frac{\nu_c}{s} - \frac{1}{2}}).$$

- Our method has improved sample complexity over V-statistic for any ν_c and s

Choice of kernel c:

- depends on smoothness of kernel k and generator G_θ
- as smooth as possible, but not smoother than G_θ or k

Theoretical guarantees

Assumptions:

- c is a Matérn kernel of order ν_c on $\mathcal{U} \subset \mathbb{R}^s$
- k is Matérn or squared-exponential kernel of order ν_k
- $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ holds

Sample complexity result for our estimator:

$$\left|\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})\right| = \mathcal{O}(m^{-\frac{\nu_c}{s} - \frac{1}{2}}).$$

- Our method has improved sample complexity over V-statistic for any ν_c and s

Choice of kernel c:

- depends on smoothness of kernel k and generator G_θ
- as smooth as possible, but not smoother than G_θ or k

Theoretical guarantees

Assumptions:

- c is a Matérn kernel of order ν_c on $\mathcal{U} \subset \mathbb{R}^s$
- k is Matérn or squared-exponential kernel of order ν_k
- $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ holds

Sample complexity result for our estimator:

$$\left|\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})\right| = \mathcal{O}(m^{-\frac{\nu_c}{s}-\frac{1}{2}}).$$

- Our method has improved sample complexity over V-statistic for any ν_c and s

Choice of kernel c:

- depends on smoothness of kernel k and generator G_θ
- as smooth as possible, but not smoother than G_θ or k

Computational cost

Total cost of the method is

1. cost of simulating from the model $\mathcal{O}(mC_{\text{gen}})$
   - often the bottleneck
2. the cost of estimating MMD
   - V-statistic: $\mathcal{O}(m^2 + mn + n^2)$
   - Optimally-weighted: $\mathcal{O}(m^3 + mn + n^2)$

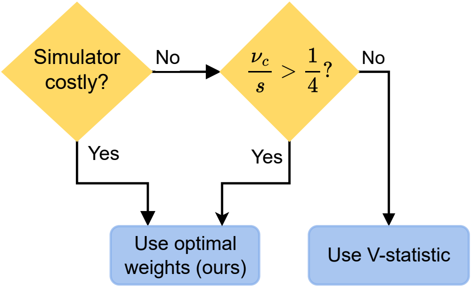


Figure: When to use our optimally-weighted estimator over the V-statistic.

Outline

Benchmarking on popular simulators

We fix θ for each model and estimate the MMD^2 between $\mathbb{P}_\theta^m$ and $\mathbb{P}_\theta^n$ with $n = 10,000$ and $m = 256$.

Model	s	d	IID V-stat	IID OW (ours)
g-and-k	1	1	2.25 (1.52)	**0.086** (0.049)
Two moons	2	2	2.36 (1.94)	**0.057** (0.054)
Bivariate Beta	5	2	2.13 (1.17)	**0.555** (0.227)
MA(2)	12	10	2.42 (0.796)	**0.705** (0.107)
M/G/1 queue	10	5	2.52 (1.19)	**1.71** (0.568)
Lotka-Volterra	600	2	2.13 (1.10)	**2.04** (0.956)

- Our estimator achieves the lowest error for all the models when $\{u_i\}_{i=1}^m$ are taken to be iid uniforms.
- Magnitude of this improvement reduces as s (the dimension of $\mathcal{U}$) increases.

Varying dimensions s and d

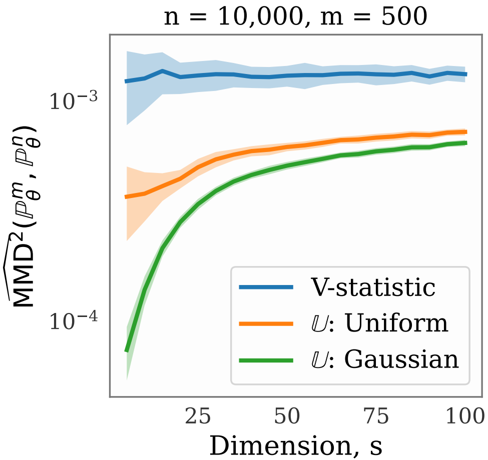
Multivariate g-and-k distribution

Two formulation of the model:

1. $(G_\theta, \mathbb{U}_\theta)$ where $\mathbb{U} = \mathcal{N}(0, I_s)$
2. $(\tilde{\mathbb{U}}, \tilde{G}_\theta)$ where $\tilde{\mathbb{U}} = \text{Unif}(0,1)^s$

Observations:

- Our estimator performs better than the V-statistic even in dimensions as high as 100.
- Gaussian embedding is better than uniform for this model.



Varying dimensions s and d

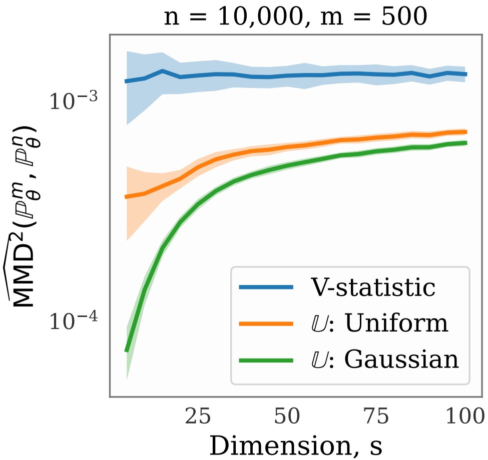
Multivariate g-and-k distribution

Two formulation of the model:

1. $(G_\theta, \mathbb{U}_\theta)$ where $\mathbb{U} = \mathcal{N}(0, I_s)$
2. $(\tilde{\mathbb{U}}, \tilde{G}_\theta)$ where $\tilde{\mathbb{U}} = \text{Unif}(0,1)^s$

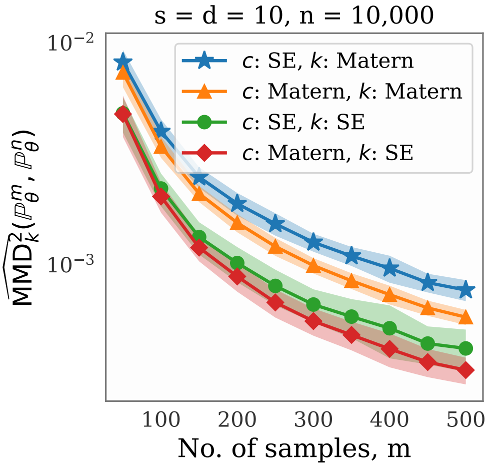
Observations:

- Our estimator performs better than the V-statistic even in dimensions as high as 100.
- Gaussian embedding is better than uniform for this model.



Varying choice of kernels k and c

Multivariate g-and-k distribution



Observations:

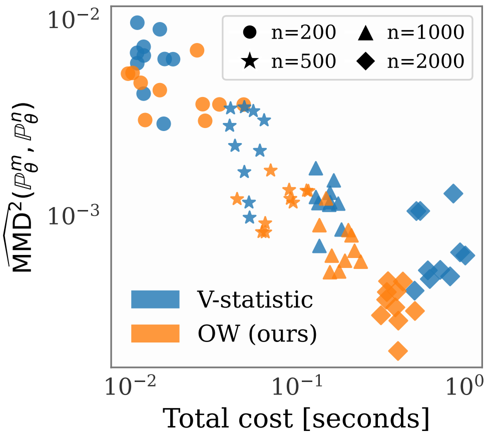
- Our method performs best when k is the squared-exponential (SE) kernel, i.e., when it is infinitely smooth.
- Combination of c as SE and k as the Matérn kernel is the worst.
- From a computational viewpoint, it is always beneficial to take k to be very smooth.

Performance vs. computational cost

Multivariate g-and-k distribution

We compare estimators for a fixed computational budget.

- We vary n and take $m = n$ for the V-statistic and $m = 2n^{2/3}$ for the OW estimator.
- Our estimator achieves lower error on average than the V-statistic.
- It is preferable to use the OW estimator even for a computationally cheap simulator like the multivariate g-and-k.

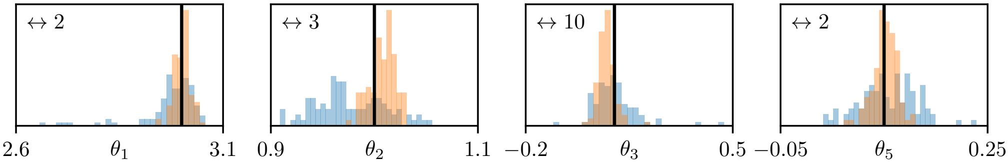


Composite goodness-of-fit test based on MMD^2

Multivariate g-and-k distribution

Suppose we have iid draws from distribution $\mathbb{Q}$.

- Null hypothesis: $\mathbb{Q}$ **is** an element of model $\{\mathbb{P}_\theta : \theta \in \Theta\}$
- Alternate hypothesis: $\mathbb{Q}$ **is not** an element of $\{\mathbb{P}_\theta : \theta \in \Theta\}$
- $\mathbb{Q}$ is multivariate g-and-k with $\theta_4 = 0.1$ (null) or $\theta_4 = 0.5$ (alternative)
- Test requires performing two steps repeatedly:
  1. estimating parameters $\hat{\theta}$ using MMD
  2. estimating MMD^2 between $\mathbb{Q}$ and $\mathbb{P}\hat{\theta}$



Composite goodness-of-fit test based on MMD^2

Multivariate g-and-k distribution

Suppose we have iid draws from distribution $\mathbb{Q}$.

- Null hypothesis: $\mathbb{Q}$ **is** an element of model $\{\mathbb{P}_\theta : \theta \in \Theta\}$
- Alternate hypothesis: $\mathbb{Q}$ **is not** an element of $\{\mathbb{P}_\theta : \theta \in \Theta\}$
- $\mathbb{Q}$ is multivariate g-and-k with $\theta_4 = 0.1$ (null) or $\theta_4 = 0.5$ (alternative)
- Test requires performing two steps repeatedly:
  1. estimating parameters $\hat{\theta}$ using MMD
  2. estimating MMD^2 between $\mathbb{Q}$ and $\mathbb{P}\hat{\theta}$

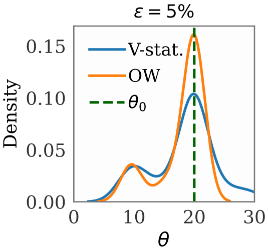
Table: Fraction of repeats for which the null was rejected. An ideal test would have 0.05 when the null holds, and 1 otherwise.

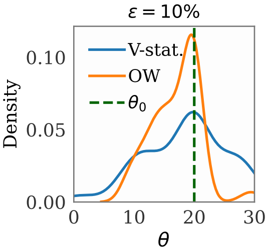
Cases	IID V-stat	IID OW (ours)
$\theta_4 = 0.1$ (null holds)	0.040	0.047
$\theta_4 = 0.5$ (alternative holds)	0.040	0.413

Large scale offshore wind farm model

We apply ABC to a low-order wake model [Kirby et al., 2023]

- Simulates estimate of the farm-averaged local turbine thrust coefficient
- Parameter θ is the angle (in degrees) at which the wind is blowing
- Challenge: simulating one data point takes ≈ 2 mins
- Generating 1000 datasets with $m = 10$ took ≈ 245 hours
- Our method can achieve similar performance as the V-statistic with much smaller m, saving hours of computation time.





Conclusion

- We proposed an optimally-weighted MMD estimator which has improved sample complexity than the V-statistic.
- Our estimator requires fewer data points than alternatives in this setting, making it especially advantageous for computationally expensive simulators.
- Limitations and open questions:
  - Parameterisation of a simulator through generator G_θ and measure $\mathbb{U}$ is usually not unique.
  - We focus on the MMD and not its gradient.
  - Our ideas can potentially translate to other distances, such as the Wasserstein distance and Sinkhorn divergence.

Conclusion

- We proposed an optimally-weighted MMD estimator which has improved sample complexity than the V-statistic.
- Our estimator requires fewer data points than alternatives in this setting, making it especially advantageous for computationally expensive simulators.
- Limitations and open questions:
  - Parameterisation of a simulator through generator G_θ and measure $\mathbb{U}$ is usually not unique.
  - We focus on the MMD and not its gradient.
  - Our ideas can potentially translate to other distances, such as the Wasserstein distance and Sinkhorn divergence.

References

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019).
Approximate Bayesian computation with the Wasserstein distance.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(2):235–269.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019).
Statistical inference for generative models with maximum mean discrepancy.
arXiv:1906.05944.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022).
Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap.
In Proceedings of the International Conference in Artificial Intelligence and Statistics, pages 943–970.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019).
Sample complexity of Sinkhorn divergences.
In International Conference on Artificial Intelligence and Statistics.

Gretton, A., Borgwardt, K., Rasch, M. J., and Scholkopf, B. (2012).