Geometrically Coupled Monte Carlo Sampling

Mark Rowland*, Krzysztof Choromanski†, François Chalus*, Aldo Pacchiano‡, Tamás Sarlós♯, Richard E. Turner*, Adrian Weller*¶

*University of Cambridge, †Google Brain Robotics, ‡University of California Berkeley, ♯Google Research, ¶Alan Turing Institute

Random sampling in ML

- Approximate Bayesian inference: use Markov chain Monte Carlo in order to sample from complex untractable posterior.
- Reinforcement learning: Monte Carlo gradient estimation.
- Variational autoencoders: outputs generated from random samples.

Main ideas

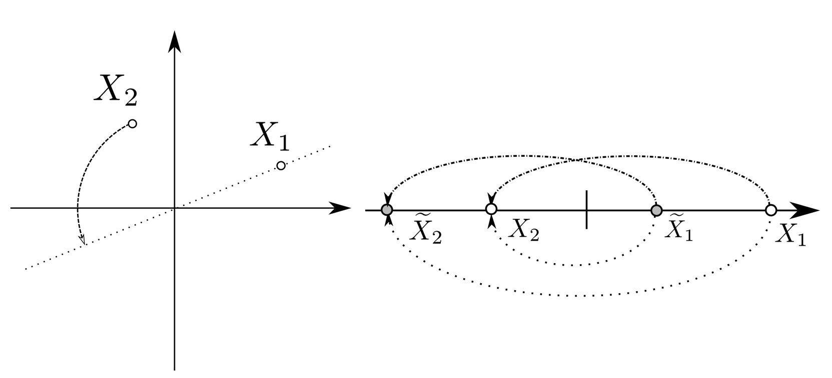
When faced with expensive downstream applications, Monte Carlo samples need to be of high quality and diversity.

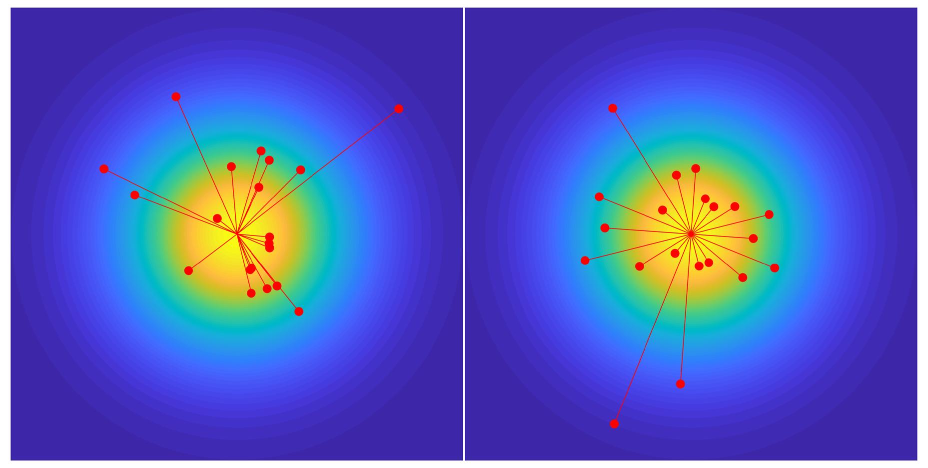
Improve sample diversity by enforcing geometric conditions on random samples:

- Orthogonal samples,
- Antithetic samples,
- Samples with coupled norms.

Contributions

- Formulation of optimal coupling problem as a multi-marginal transport problem.
- Example solutions derived from those problems.
- Comparison with classical QMC low-discrepancy methods.
- Theoretical bounds on estimating gradients of function smoothings when coupling samples.
- Experimental results on learning navigation policies with evolution strategies and ELBO estimation.





Key results

Optimal coupling problem

Aim: Estimate $I_f = \mathbb{E}_{X\sim\eta}[f(X)]$ with Monte Carlo estimator $\frac{1}{m}\sum_{i=1}^m f(X_i)$.

Problem: Optimal coupling when f is unknown?

Solution: Model $f \sim \mathrm{GP}(0, K)$, find distribution μ solving

$$\min_{\mu} \mathbb{E}_{f\sim\mathrm{GP}(0,K)}\left[\mathbb{E}_{X_{1:m}\sim\mu}\left[\left(\frac{\sum_{i=1}^m f(X_i)}{m} - I_f\right)^2\right]\right],$$

where marginals of μ are all equal to η. Solutions to this problem are called a *K-optimal couplings*.

Link to optimal transport problem

Multi-marginal transport formulation: A joint distribution μ is a K-optimal coupling if and only if it minimizes

$$\mathbb{E}_{X_{1:m}\sim\mu}\left[\sum_{i\neq j} K(X_i, X_j)\right].$$

Repulsive costs: Unlike many optimal transport problems in machine learning, here the cost is *repulsive*, encouraging diversity of samples.

Example solutions

- **Antithetic norm coupling.** When the marginal η is radially symmetric and K is a RBF kernel $K(x,y) = \Phi(||x-y||)$ with Φ decreasing and convex, then the optimal transport problem with $m = 2$ is solved when

$$X_2 = -F_\eta^{-1}(1 - F_\eta(||X_1||))\frac{X_1}{||X_1||},$$

or in other words

$$F_\eta(||X_1||) + F_\eta(||X_2||) = 1.$$

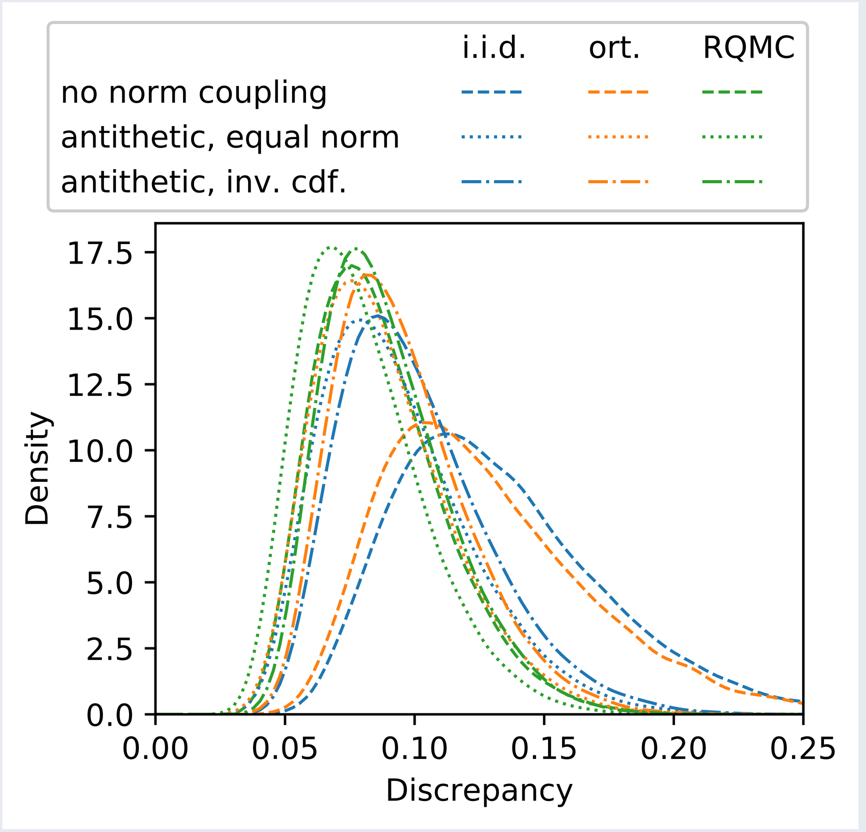
- **Orthogonal directions coupling.** When the marginal η is radially symmetric and K is a RBF kernel

$$K(x,y) = \Phi(||x-y||^2)$$

with Φ decreasing and convex, then the optimal transport problem is solved when $\langle X_i, X_j\rangle = 0$. This supports the experimental results shown in [1].

Discrepancy of GCMC samples

Discrepancy: $D_\eta^*(S) = \sup_{u\in[0,1]}\left|u - \frac{|\{i : X_i \leq u\}|}{|S|}\right|$ where $S = \{X_1, \ldots, X_{|S|}\}$.



Experimentally, the discrepancy is lower for antithetically coupled samples. Naturally, randomized quasi-Monte Carlo sampling also achieves low discrepancy.

This leads to a concentration of the error towards 0 thanks to the Koksma-Hlawka inequality:

$$\left|\frac{1}{m}\sum_{i=1}^m f(X_i) - I_f\right| \leq V_{\mathrm{HK}}(f) D_\eta^*(S).$$

Application to ELBO estimation

In training VAEs, one estimates the ELBO and its gradient by passing random samples through the network. We use coupled samples and observe:

- improved training speed,
- best performance when combining antithetic and orthogonal samples.

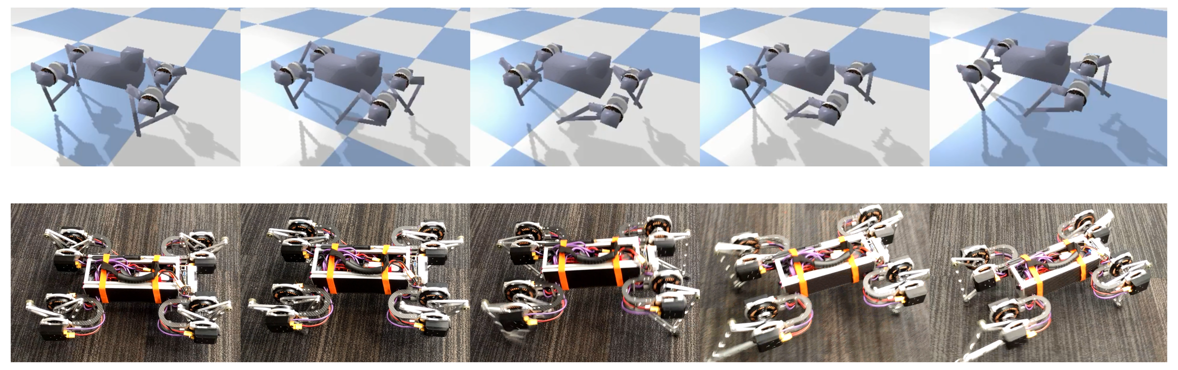
Application to gradient estimation of function smoothings

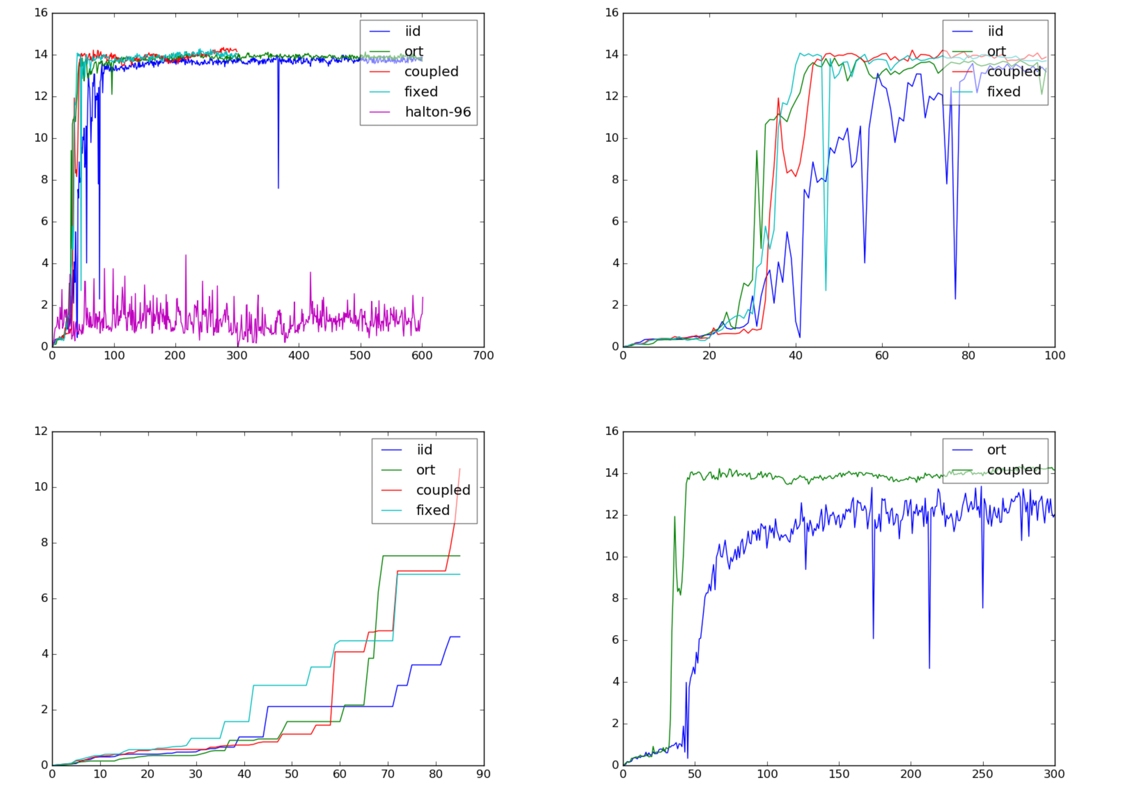
Aim: estimate the gradient of a function smoothing through sampling.

Observation: using orthogonal samples leads to a sub-Gaussian estimator which is more concentrated than the one using i.i.d. samples.

Application to policy learning

- **Aim:** maximize $J(\theta) = \mathbb{E}_{X\sim N(\theta,\sigma I)}[F(X)]$ with respect to θ where F is only available through function evalutions.
- **Strategy:** use coupled samples to estimate $\nabla J(\theta)$ by $\frac{1}{m\sigma}\sum_{i=1}^m F(\theta + \sigma\varepsilon_i)\varepsilon_i$, e.g. antithetic and orthogonal samples or samples of fixed lengths.
- Allows to learn walkable policies for simulated and real robots.





Conclusion

- Orthogonal and antithetic sampling can be motivated by a multi-marginal transport problem.
- The observed increase in performance can be explained by a lower discrepancy of the samples.
- Orthogonal samples can be applied in a wide range of domains where diversity of samples matters.

References

[1] F. Yu, A. Suresh, K. Choromanski, D. Holtmann-Rice, and S. Kumar, "Orthogonal random features," in *Neural Information Processing Systems (NIPS)*, 2016.

[2] K. Choromanski, M. Rowland, T. Sarlos, V. Sindhwani, R. Turner, and A. Weller, "The geometry of random features," in *Artificial Intelligence and Statistics (AISTATS)*, 2018.

[3] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," in *arXiv*, 2017.