Monte Carlo with control variate for the Sliced-Wasserstein distance

Anatole Gallouët

Ongoing work with J. Delon, J. Digne and N. Bonneel

March 2025

Introduction

Optimal transport and Wasserstein distance

- A distance between densities (or point clouds)
- Many applications: Computer graphics, Data science, Physics, Geometry...

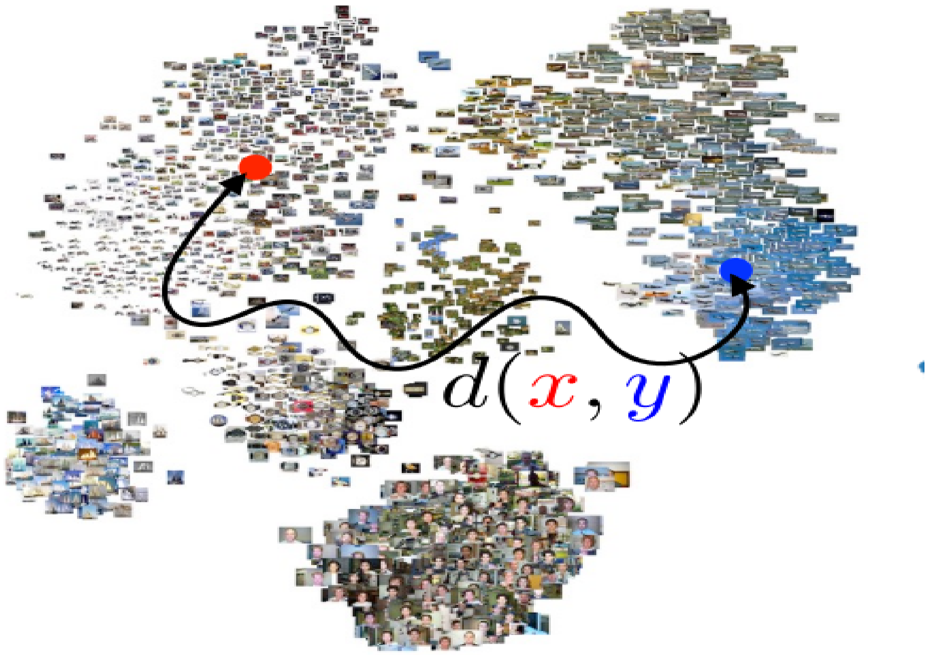
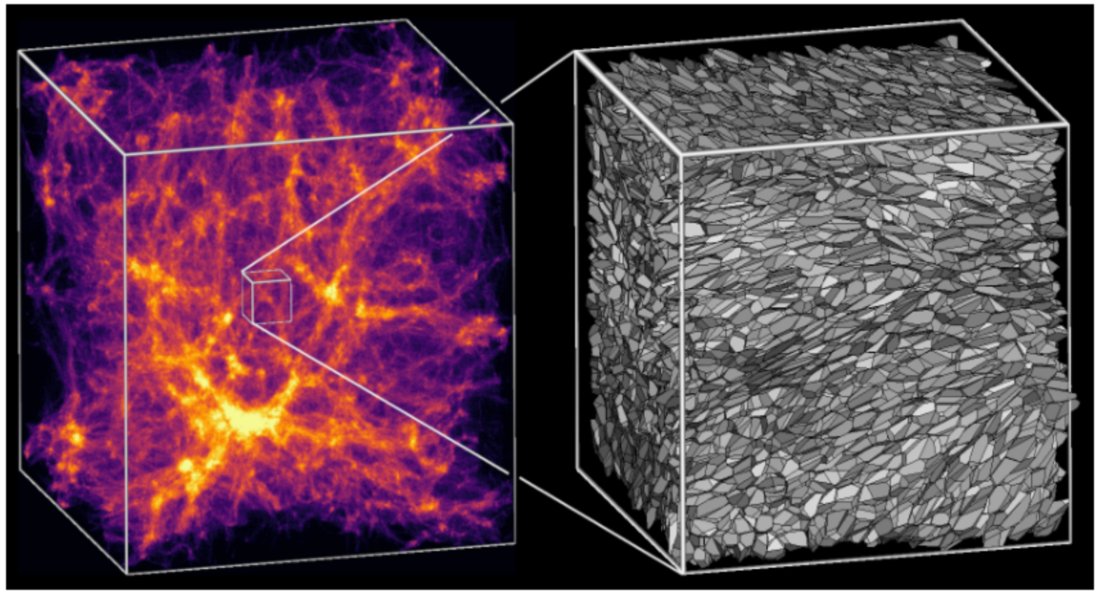
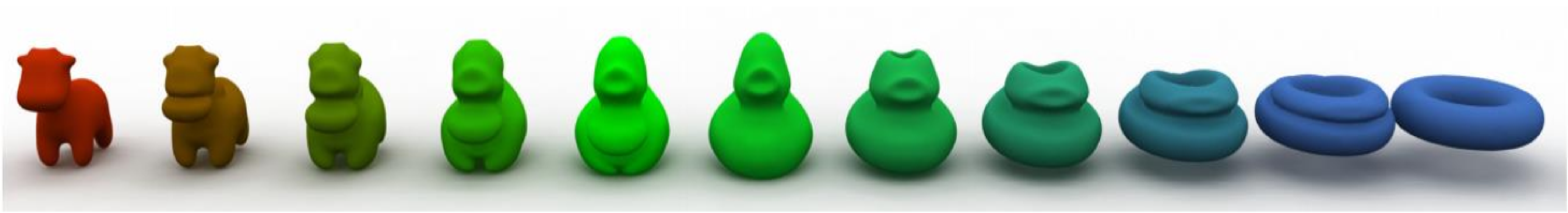


Image from *G. Peyré*.



Monge-Ampère Gravity *[Levy et al. '24]*



Wasserstein barycenters *[Solomon et al. '15]*

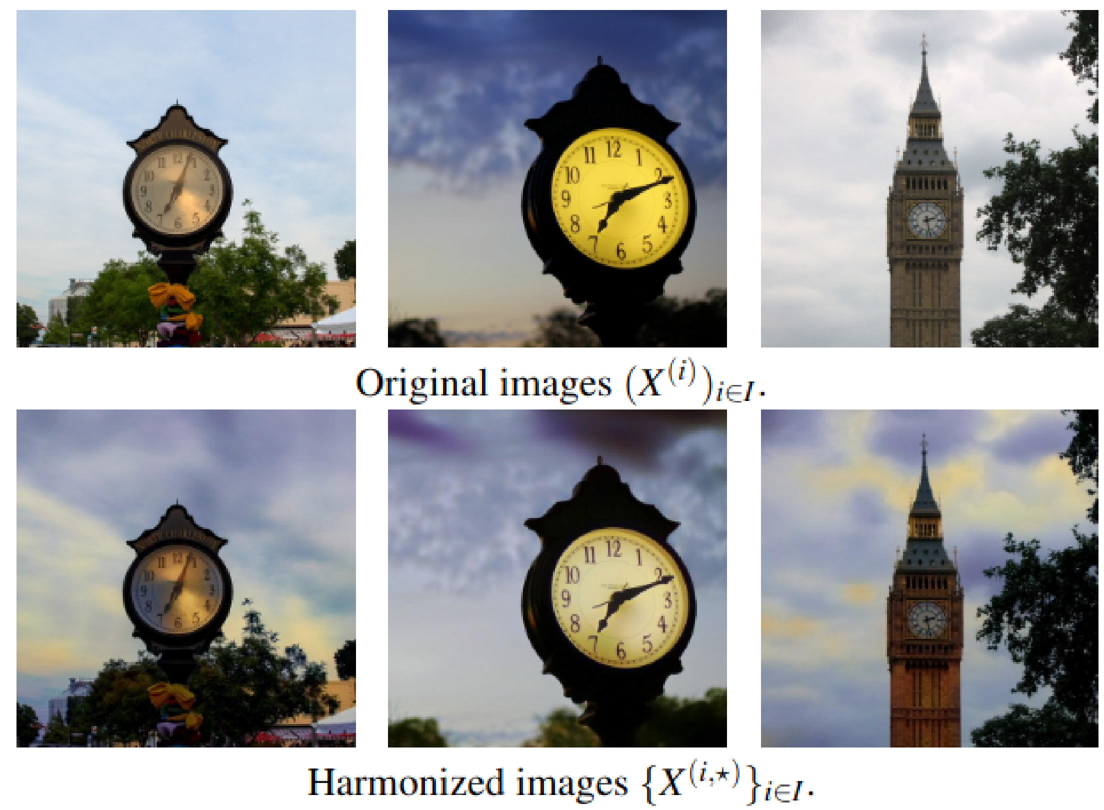
Introduction

Sliced optimal transport (Introduced by *[Rabin et al. '12]*)

- Projected 1-D optimal transport, better computational complexity
- Useful for large scale problems or high dimension (Machine Learning, Image...)

Generated samples from the
LSUN bedrooms dataset

SW GANs *[Desphande et al. 18]*



SW barycenters *[Bonneel et al. '15]*

The Wasserstein distance

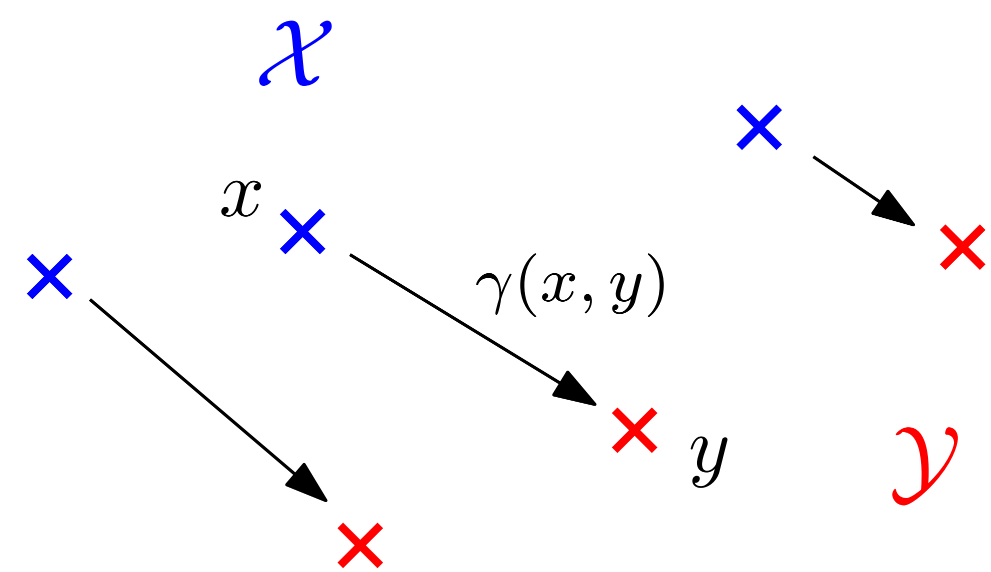
Definition: (Wasserstein distance)
The Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{W}_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|_p^p \, \mathrm{d}\,\gamma(x, y)$$

where $\Pi(\mu, \nu)$ is the set of transport plans (or couplings) between μ and ν.

Discrete example:

$$\mu = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}$$



$$\nu = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i}$$

- Optimal transport problem between μ and ν *[Monge 1781]*.
- Linear problem on couplings γ *[Kantorovich '42]*.
  - $\mathrm{O}(m^3)$ complexity on disrete measures.
  - $m \sim \frac{1}{\varepsilon^d}$ for ε like error when sampling densities *[Dudley '68]*

The Wasserstein distance

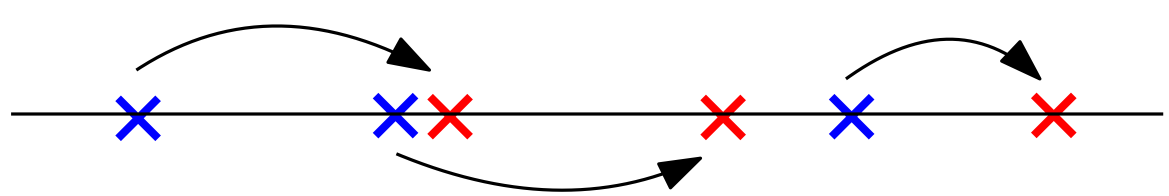
Definition: (Wasserstein distance)
The Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{W}_p^p(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|_p^p \,\mathrm{d}\,\gamma(x,y)$$

where $\Pi(\mu,\nu)$ is the set of transport plans (or couplings) between μ and ν.

Discrete example:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}$$



$$\nu = \frac{1}{m}\sum_{i=1}^{m} \delta_{y_i}$$

sorted points $(x_{\sigma(i)})$ and $(y_{\kappa(i)})$ $\longrightarrow$ $\mathrm{O}(m\log(m))$

The 1-D case:

when $\mu, \nu \in \mathcal{P}(\mathbb{R})$, we have

$$\mathrm{W}_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p \,\mathrm{d}\,t = \sum_{i=1}^{m} \|x_{\sigma(i)} - y_{\kappa(i)}\|^p$$

where F_μ (resp. F_ν) is the c.d.f of μ (resp ν).

The Sliced-Wasserstein distance

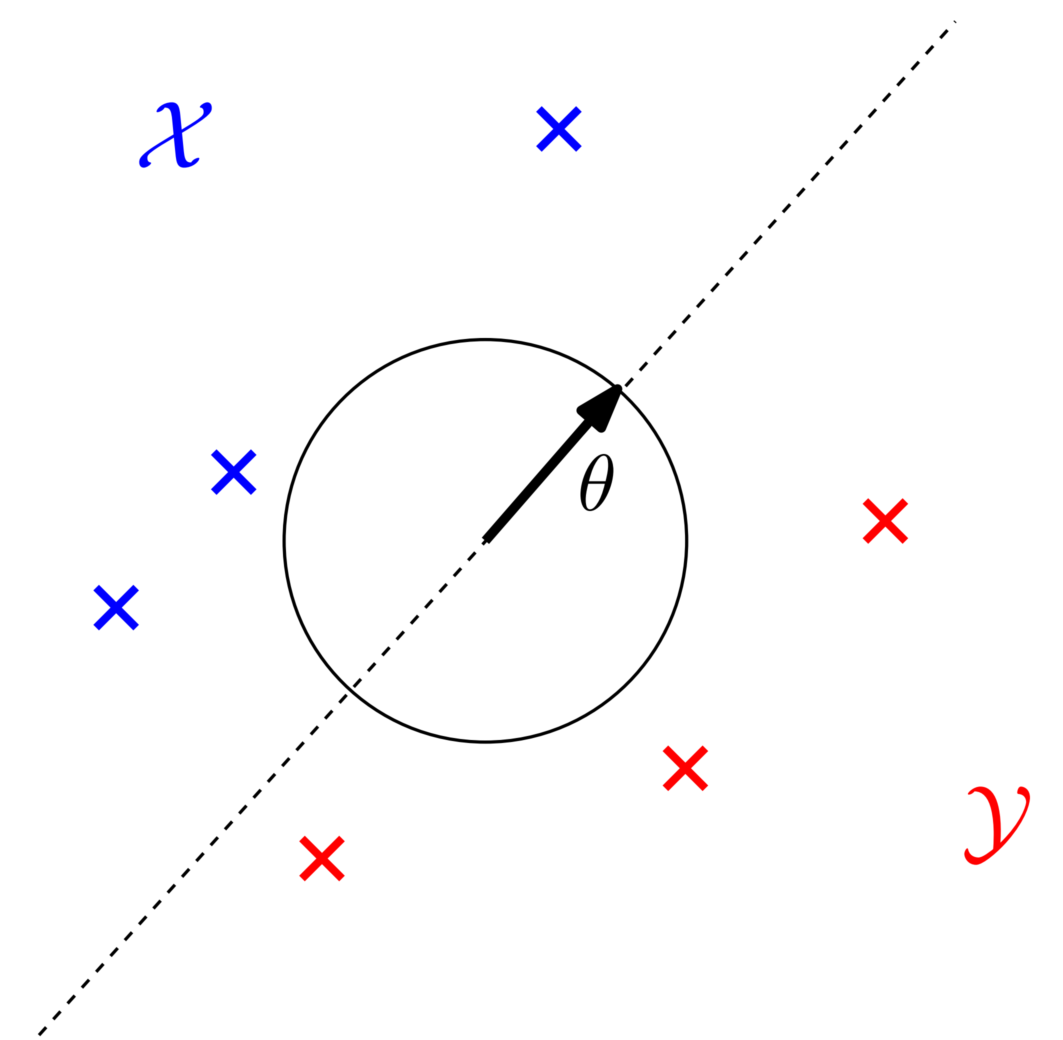
Definition: (Sliced-Wasserstein distance)
The Sliced- Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is

$$\mathrm{SW}_p^p(\mu, \nu) = \int_{\mathcal{S}^{d-1}} \mathrm{W}_p^p(\theta^*_\# \mu, \theta^*_\# \nu) \, \mathrm{d}\, \theta$$

where $\theta^*_\# \mu$ is the image measure (or pushforward) of μ by $\theta^* = \langle \cdot | \theta \rangle$ and the image measure is defined for $B \subset \mathbb{R}$ by $\theta^*_\# \mu(B) = \mu(\theta^{*-1}(B))$

Discrete example:

- $\mu = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i} \quad \nu = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i}$



The Sliced-Wasserstein distance

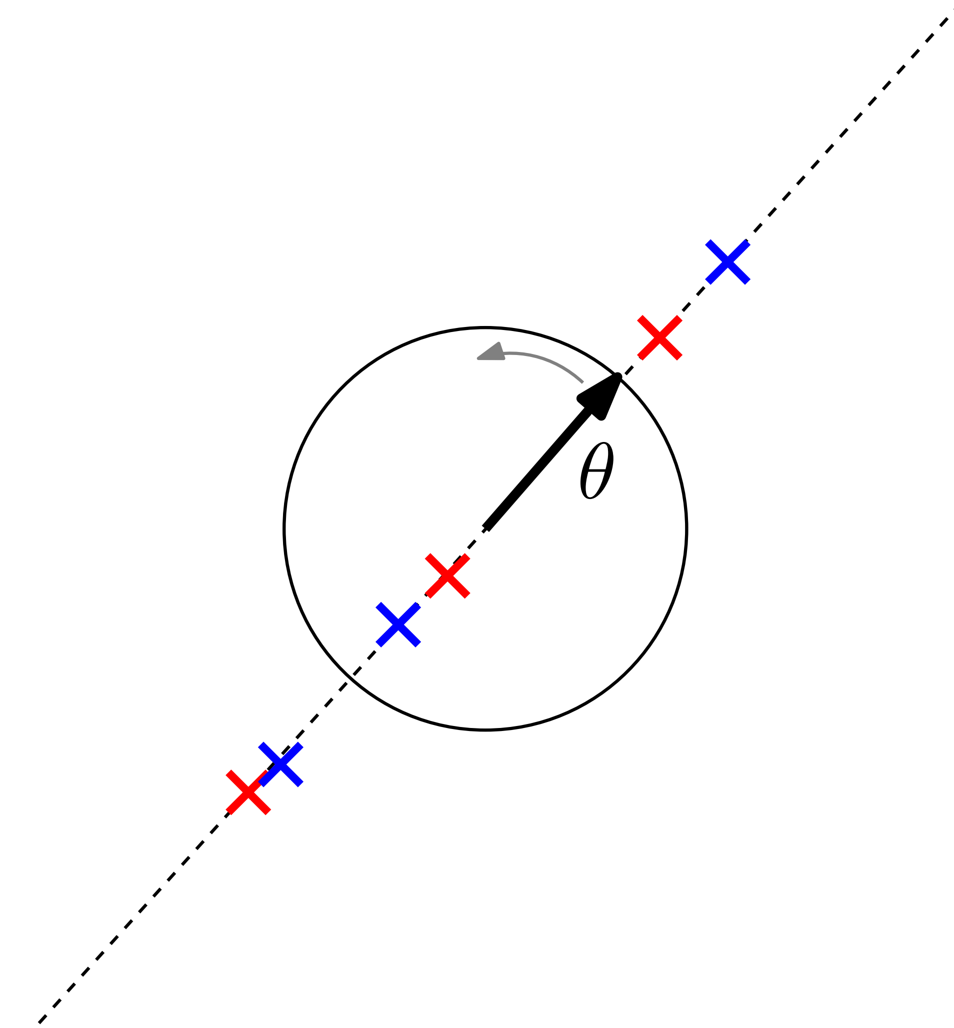
Definition: (Sliced-Wasserstein distance)
The Sliced- Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is

$$\mathrm{SW}_p^p(\mu, \nu) = \int_{\mathcal{S}^{d-1}} \mathrm{W}_p^p(\theta^*_\# \mu, \theta^*_\# \nu) \,\mathrm{d}\,\theta$$

where $\theta^*_\# \mu$ is the image measure (or pushforward) of μ by $\theta^* = \langle \cdot | \theta \rangle$ and the image measure is defined for $B \subset \mathbb{R}$ by $\theta^*_\# \mu(B) = \mu(\theta^{*-1}(B))$

Discrete example:

- $\mu = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i} \quad \nu = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i}$
- Project μ and ν along θ
- Compute $f_{\mu,\nu}(\theta) = \mathrm{W}_p^p(\theta^*_\# \mu, \theta^*_\# \nu)$
- Integrate on all directions $\theta \in \mathcal{S}^{d-1}$



The Sliced-Wasserstein distance

Definition: (Sliced-Wasserstein distance)
The Sliced- Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is

$$\mathrm{SW}_p^p(\mu, \nu) = \int_{\mathcal{S}^{d-1}} \mathrm{W}_p^p(\theta^*_\# \mu, \theta^*_\# \nu) \, \mathrm{d}\, \theta$$

where $\theta^*_\# \mu$ is the image measure (or pushforward) of μ by $\theta^* = \langle \cdot | \theta \rangle$ and the image measure is defined for $B \subset \mathbb{R}$ by $\theta^*_\# \mu(B) = \mu(\theta^{*-1}(B))$

Theoretical results:

- SW is indeed a distance.
- $\mathrm{SW}_p^p(\mu, \nu) \leq \mathrm{W}_p^p(\mu, \nu)$ for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.
- $\mathrm{W}_p^p(\mu, \nu) \leq C_{d,p,r} \, \mathrm{SW}_p^{\frac{1}{d+1}}(\mu, \nu)$ for $\mu, \nu \in \mathcal{P}(B(0, r))$. *[Bonnotte '13]*
- No curse of dimensionality: $\mathrm{O}\left(n \, m \log(m)\right)$

For ε error:
$n \sim \frac{1}{\varepsilon^2}$ Monte Carlo error
$m \sim \frac{1}{\varepsilon^{2p}}$ for sampling *[Nadjahi et al. '20]*

Monte Carlo and control variates

Objective: Integral $I(f) = \int_\Omega f(x)\,\mathrm{d}\,\mathrm{P}(x)$

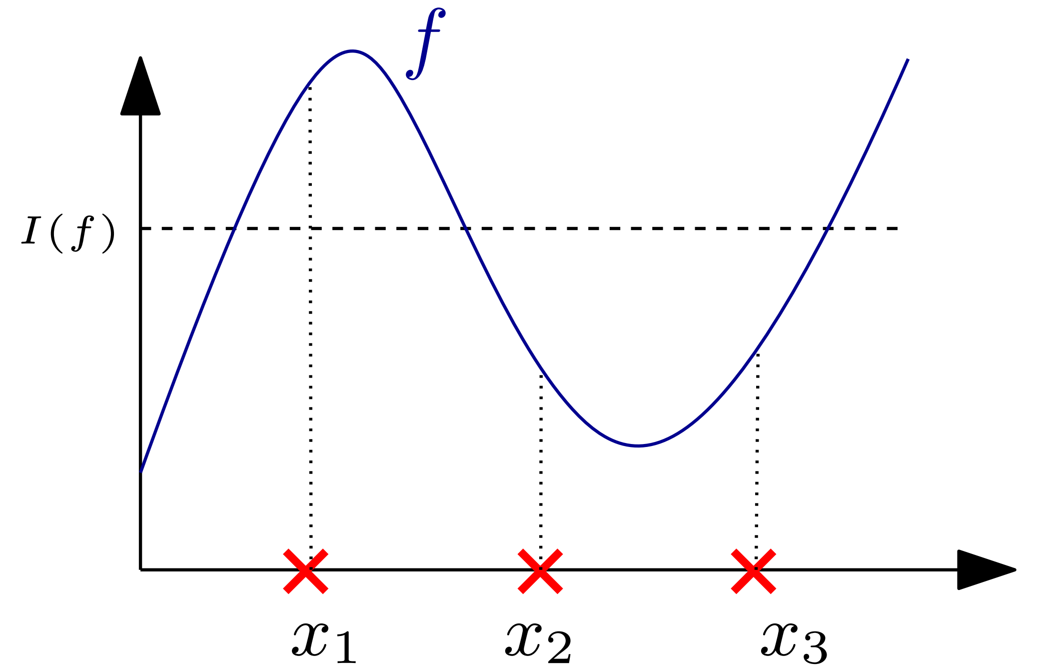
Monte Carlo:

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \text{ with } (x_i) \sim \mathrm{P}$$

Convergence rate:

$$\mathrm{Var}(\widehat{I}_n) = \mathbb{E}((\widehat{I}_n - I)^2) = \frac{\mathrm{Var}(f)}{n}$$

$\mathrm{O}(n^{-1/2})$ rate



Monte Carlo and control variates

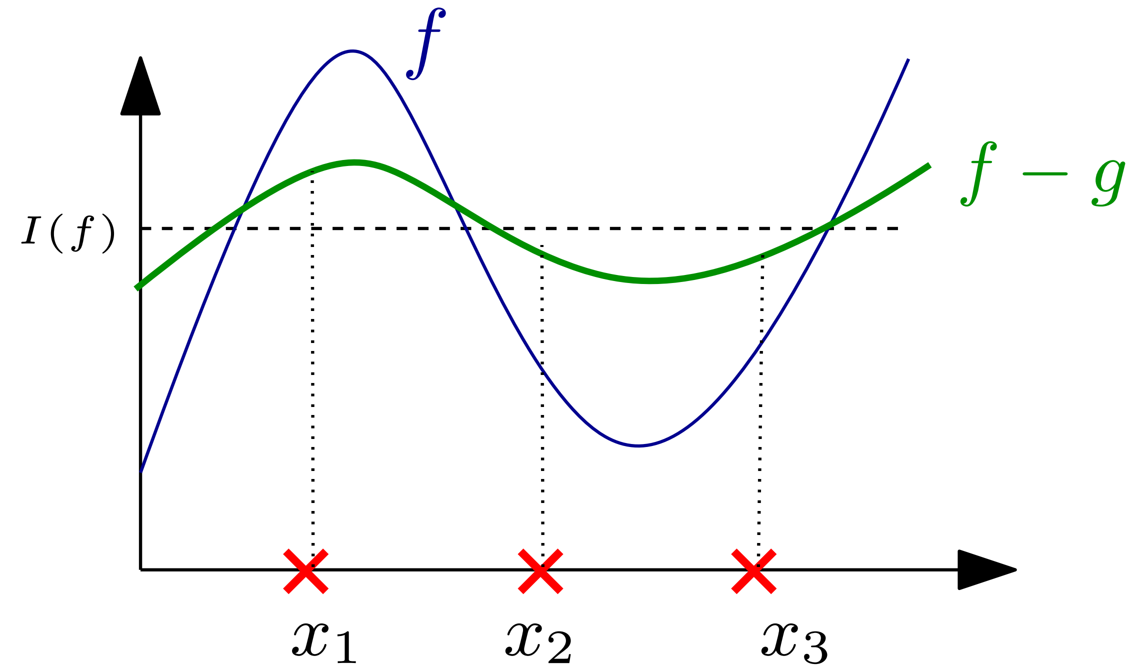
Objective: Integral $I(f) = \int_\Omega f(x)\,\mathrm{d}\,\mathrm{P}(x)$

Monte Carlo:

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \text{ with } (x_i) \sim \mathrm{P}$$

Convergence rate:

$$\mathrm{Var}(\widehat{I}_n) = \mathbb{E}((\widehat{I}_n - I)^2) = \frac{\mathrm{Var}(f)}{n}$$



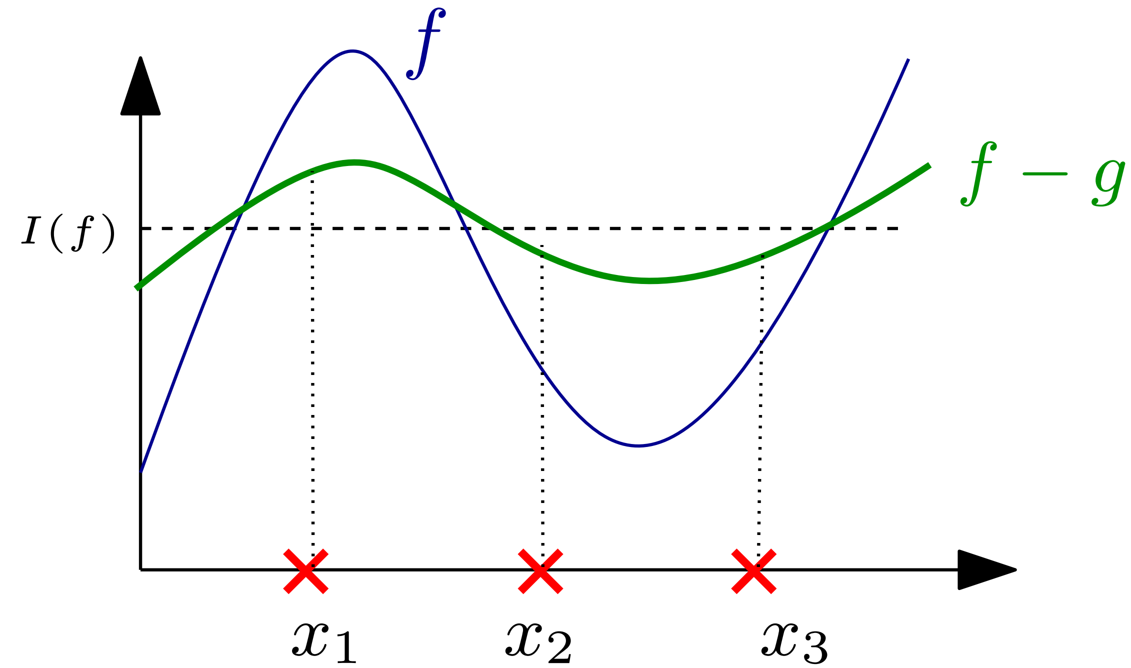
Definition: (Control variate) A control variate is a function $g : \Omega \to \mathbb{R}$ such that $\mathrm{Var}(f-g) \leq \mathrm{Var}(f)$ and $\int_\Omega g$ is known.
The control variate estimator is:

$$\widehat{ICV_n} = \frac{1}{n}\sum_{i=1}^{n}(f-g)(x_i) + \int_\Omega g$$

- Unbiased: $\mathbb{E}(\widehat{ICV_n}) = \mathbb{E}(\widehat{I}_n) = I(f)$
- Variance reduction: $\mathrm{Var}(\widehat{ICV_n}) = \dfrac{\mathrm{Var}(f-g)}{n} \leq \mathrm{Var}(\widehat{I}_n)$

Monte Carlo and control variates

Objective: Integral $I(f) = \int_\Omega f(x)\,\mathrm{d}\,\mathrm{P}(x)$



Monte Carlo:

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \text{ with } (x_i) \sim \mathrm{P}$$

Convergence rate:

$$\mathrm{Var}(\widehat{I}_n) = \mathbb{E}((\widehat{I}_n - I)^2) = \frac{\mathrm{Var}(f)}{n}$$

Definition: (Control variate) A control variate is a function $g : \Omega \to \mathbb{R}$ such that $\mathrm{Var}(f-g) \leq \mathrm{Var}(f)$ and $\int_\Omega g$ is known.
The control variate estimator is:

$$\widehat{ICV_n} = \frac{1}{n}\sum_{i=1}^{n}(f-g)(x_i) + \int_\Omega g$$

- Unbiased: $\mathbb{E}(\widehat{ICV_n}) = \mathbb{E}(\widehat{I}_n) = I(f)$
- Variance reduction: $\mathrm{Var}(\widehat{ICV_n}) = \frac{\mathrm{Var}(f-g)}{n} \leq \mathrm{Var}(\widehat{I}_n)$

Cv rate doesn't change

Goal: Find control variates for SW i.e. $f_{\mu,\nu}(\theta) = \mathrm{W}_p^p(\theta^*_\# \mu, \theta^*_\# \nu)$

A naïve control variate for SW_2.

- Explicit formula for W_2 with centered measures:

$$\mathrm{W}_2^2(\alpha,\beta) = \mathrm{W}_2^2(\bar\alpha,\bar\beta) + \|m_\alpha - m_\beta\|^2 \qquad \text{with} \qquad \begin{array}{l} m_\alpha = \int x \,\mathrm{d}\,\alpha(x) \\ \bar\alpha = T_{m_\alpha \#}\alpha. \end{array}$$

- Then $\mathrm{SW}_2^2(\mu,\nu) = \displaystyle\int_{\mathcal{S}^{d-1}} \underbrace{\mathrm{W}_2^2(\theta^*_\#\bar\mu, \theta^*_\#\bar\nu)}_{= f_{\bar\mu,\bar\nu}(\theta)} + \underbrace{\|m_{\theta^*_\#\mu} - m_{\theta^*_\#\nu}\|^2}_{= \langle\theta|m_\mu - m_\nu\rangle^2} \mathrm{d}\,\theta$

$$= \mathrm{SW}_2^2(\bar\mu,\bar\nu) + \frac{1}{d}\|m_\mu - m_\nu\|^2$$

[Nadjahi et al. '22]

Lemma: Using the projected means as control variate amounts to compute SW_2^2 on centered measures:

$$\widehat{I}_n(f_{\bar\mu,\bar\nu}) + \tfrac{1}{d}\|m_\mu - m_\nu\|_2^2 = \widehat{ICV}_n(f_{\mu,\nu}, g)$$

with control variate $g(\theta) = \langle\theta|m_\mu - m_\nu\rangle^2$ and $\int g = \frac{1}{d}\|m_\mu - m_\nu\|_2^2$.

This control variate was introduced as LCV by *[Nguyen, Ho '24]* using a Gaussian approximation. This lemma shows that it is not necessary to compute $\langle\theta_i|m_\mu - m_\nu\rangle^2$ for each sample θ_i.

QNET: A neural network for integrals

Main idea: Train a network g_w with known integral to approximate $f_{\mu,\nu}$.

QNET: A neural network for integrals

Main idea: Train a network g_w with known integral to approximate $f_{\mu,\nu}$.

Q-NETS *[Subr '21]* are shallow neural networks with explicit integral

- $$g_w : \mathbb{R}^d \to \mathbb{R}$$
  $$x \mapsto w_2^T \sigma(w_1 x + b_1) + b_2$$

  sigmoid activation:
  $\sigma(x) = \frac{1}{1+e^{-x}}$

  weights:
  $w_1 \in \mathbb{R}^{k \times d}$
  $b_1 \in \mathbb{R}^k$
  $w_2 \in \mathbb{R}^k$
  $b_2 \in \mathbb{R}$

- $$\int_{[0,1]^d} g_w = w_2^T v + 2^d b_2$$
  where v can be computed using k evaluations of the polylogarithm function of order d.

- Main observation from *[Subr '21]*:

  $$\int g_w$$
  can be evaluated on any interval using a neural network with fixed weights

- Useful when same integrand over several domains.

- Shallow architecture gives limited approximation precision.

Auto-integrable neural network

Neural control variate with automatic integration *[Li et al. '24]*

Train on derivative :

- Neural network G_w with any architecture
- Compute by autodifferentation $g_w = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} G_w$
- Train g_w to match $f_{\mu,\nu}$
- Integrate using $\int_{[-1,1]^d} g_w = \sum_{x_i \in \{-1,1\}^d} \pm G_w(x_i)$

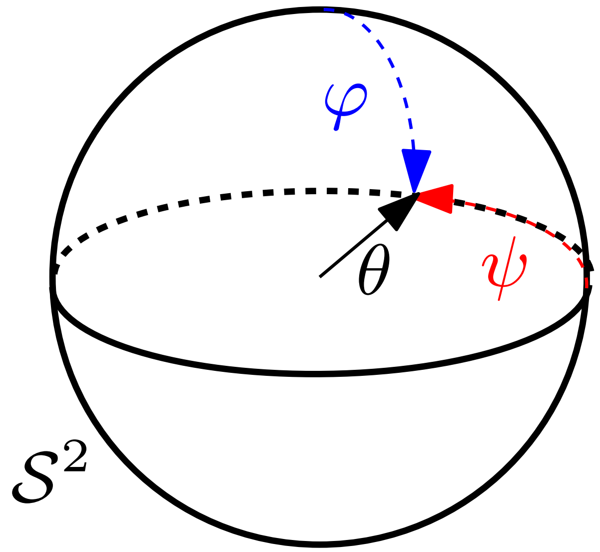
- Architecture choice gives better approximation properties.
- In practice, the architecture chosen is SIREN which uses periodic activation functions. *[Sitzmann et al . '20, Li et al. '24]*

Integration on the sphere

Problem : We want to integrate $f_{\mu,\nu}$ on the sphere $\mathcal{S}^{d-1}$.

Spherical coordinates:

When $d = 3$ (for simplicity), $\theta \in \mathcal{S}^2$ is parametrized by angles $\varphi \in [0, \pi]$, $\psi \in [0, 2\pi[$



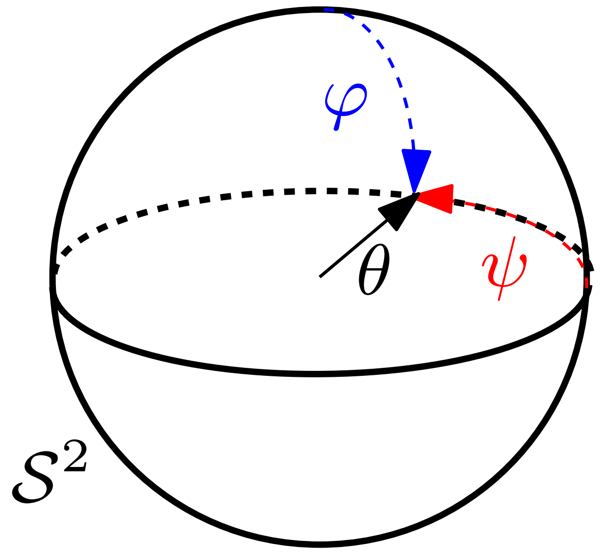
Change of variable:

$$\int_{\mathcal{S}^2} f(\theta)\,\mathrm{d}\,\theta = \int_0^{\pi}\int_0^{2\pi} f(\varphi,\psi)\sin(\varphi)\,\mathrm{d}\,\psi\,\mathrm{d}\,\varphi$$

Integration on the sphere

Problem : We want to integrate $f_{\mu,\nu}$ on the sphere $\mathcal{S}^{d-1}$.

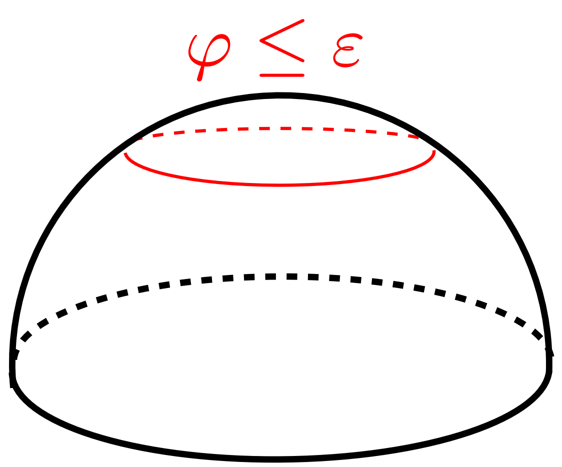
Spherical coordinates: When $d = 3$ (for simplicity), $\theta \in \mathcal{S}^2$ is parametrized by angles $\varphi \in [0, \pi]$, $\psi \in [0, 2\pi[$



Change of variable:

$$\int_{\mathcal{S}^2} f(\theta)\,\mathrm{d}\,\theta = \int_0^{\pi}\int_0^{2\pi} \underbrace{f(\varphi,\psi)\sin(\varphi)}_{\approx g_w(\varphi,\psi)}\,\mathrm{d}\,\psi\,\mathrm{d}\,\varphi$$

- Neural networks integrate on (hyper)-rectangles so we train g_w to match $f_{\mu,\nu}(\varphi,\psi)\sin(\varphi)$
- Control variate is $\dfrac{g_w(\varphi,\psi)}{\sin(\varphi)}$ Numerically unstable
- $f_{\mu,\nu}$ is even so we can integrate on $\underbrace{\text{one hemisphere}}_{\varphi \le \frac{\pi}{2}}$
- Set $g_w(\varphi,\psi) = 0$ for $\varphi \le \varepsilon$.



Integrate on D_ε

Neural network control variate (NNCV)

- Draw k sample $(\tilde{\theta}_j) \in \mathcal{S}^{d-1}$ for training
- Draw L sample $(\theta_i) \in \mathcal{S}^{d-1}$ for Monte Carlo

$\Big\}$ $n = k + L$ total samples.

- Train network on $g_w(\tilde{\theta}_j)$ with objective $f_{\mu,\nu}(\tilde{\varphi}_j, \tilde{\psi}_j)\sin(\tilde{\varphi}_j)$.

  → Gradient descent w/r to parameters w.

$\widehat{\mathrm{NNCV}}$ estimator:

$$\widehat{\mathrm{NNCV}} = \sum_{i=1}^{L}\left(f_{\mu,\nu}(\theta_i) - \underbrace{1_{D_\varepsilon}(\theta_i)}\frac{g_w(\theta_i)}{\sin(\varphi_i)}\right) + \int_{\underbrace{D_\varepsilon}} g_w(\varphi,\psi)\,\mathrm{d}(\varphi,\psi)$$

Restriction to D_ε for numerical stability

- Two estimators $\widehat{\mathrm{NNCV}}_{\mathrm{AI}}$ and $\widehat{\mathrm{NNCV}}_{\mathrm{QN}}$ for Auto integrable and Qnet.

[Subr '21, Li et al. '24]

Spherical Harmonics control variates

Definition: (Spherical Harmonics)
Spherical harmonics $(\phi_{\ell,k})_{\ell \geq 0}^{k \leq N_\ell^d}$ are harmonic homogenous polynomials of degree ℓ restricted to $\mathcal{S}^{d-1}$. They form an orthonormal Hilbert basis of $L^2(\mathcal{S}^{d-1})$.

Properties: For $i \neq 1$ $\int_{\mathcal{S}^{d-1}} \phi_i = 0$ (zero mean) $\quad \int_{\mathcal{S}^{d-1}} \phi_i \phi_j = 0$ (orthogonal)

Ordinary least squares Monte Carlo: (for s harmonics)

$(f_{\mu,\nu}(\theta_i))_{1 \leq i \leq n} \in \mathbb{R}^n$ $\qquad$ $\Phi_{i,j} = \phi_j(\theta_i) \in \mathbb{R}^{n \times s}$.

$$\widehat{SHCV} \in \underset{\alpha \in \mathbb{R},\, \beta \in \mathbb{R}^s}{\arg\min} \; \|f_{\mu,\nu}^n - \alpha \mathbf{1}_n - \Phi \beta\|_2^2$$

[Leluc et al. '24]

Dist. estim. $\qquad$ coeff. of $f_{\mu,\nu}$ on harmonics.

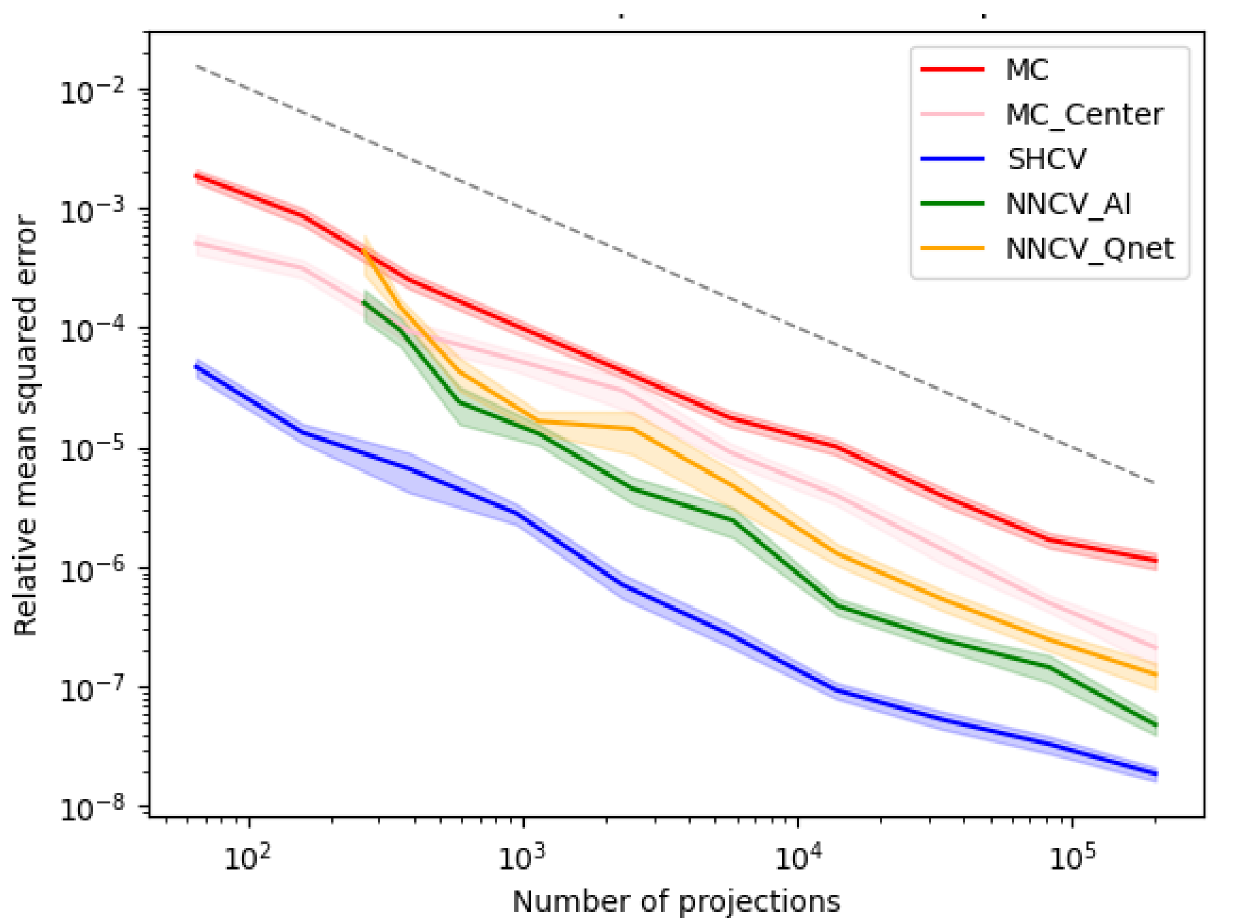
- $\widehat{SHCV} = \langle v | f_{\mu,\nu}^n \rangle$ for $\underbrace{v \text{ indep. of } f_{\mu,\nu}}$ (involving $(\Phi^T \Phi)^{-1}$)

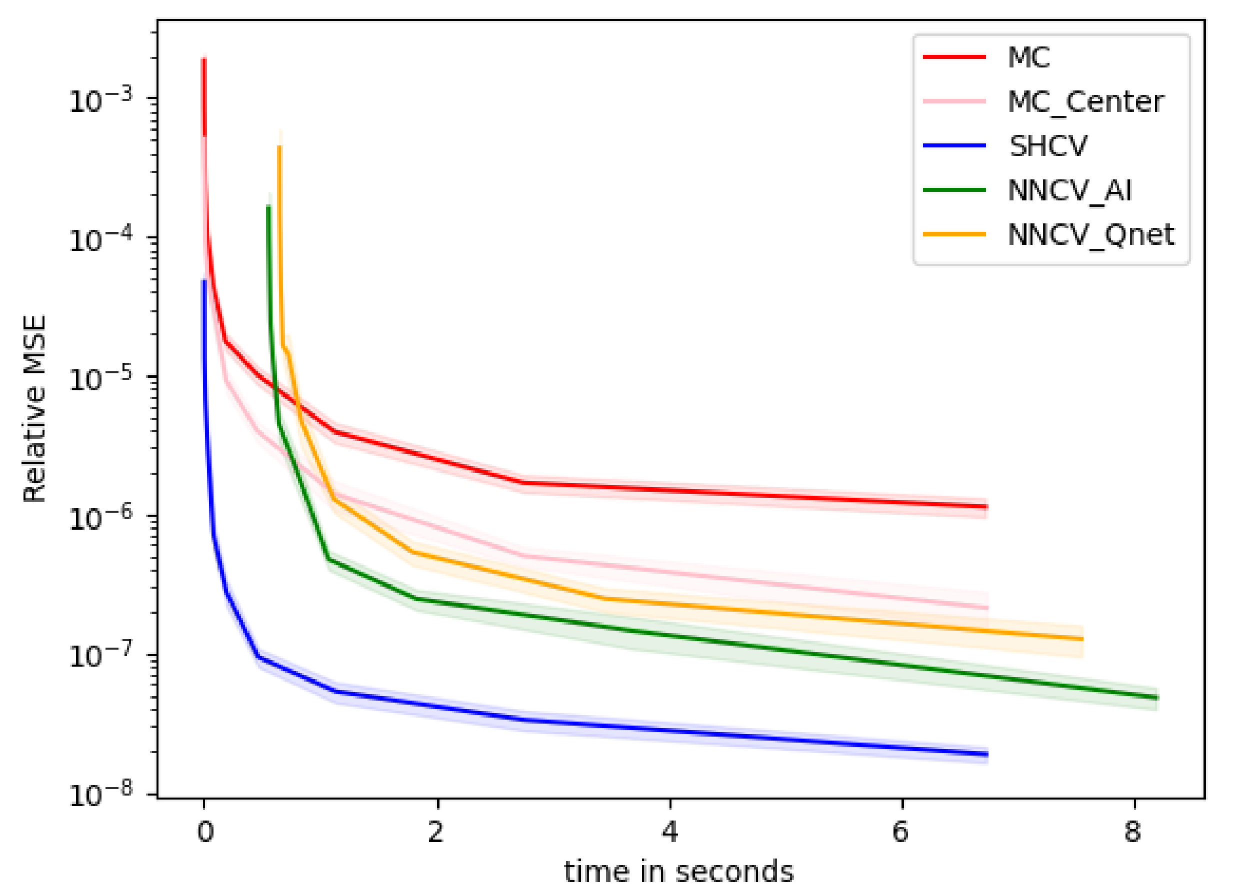
  Efficient for multiple integrals over the same directions (θ_i)
- $\mathbb{E}(|\widehat{SHCV}_{n,L} - \mathrm{SW}(\mu,\nu)|) = \mathrm{O}(L^{-1} n^{-1/2})$ for max degree $L = o(n^{1/2(d-1)})$

Numerical experiments

Dimension $d = 3$

GPU implementation (pytorch) of different Monte Carlo estimator for SW_2^2



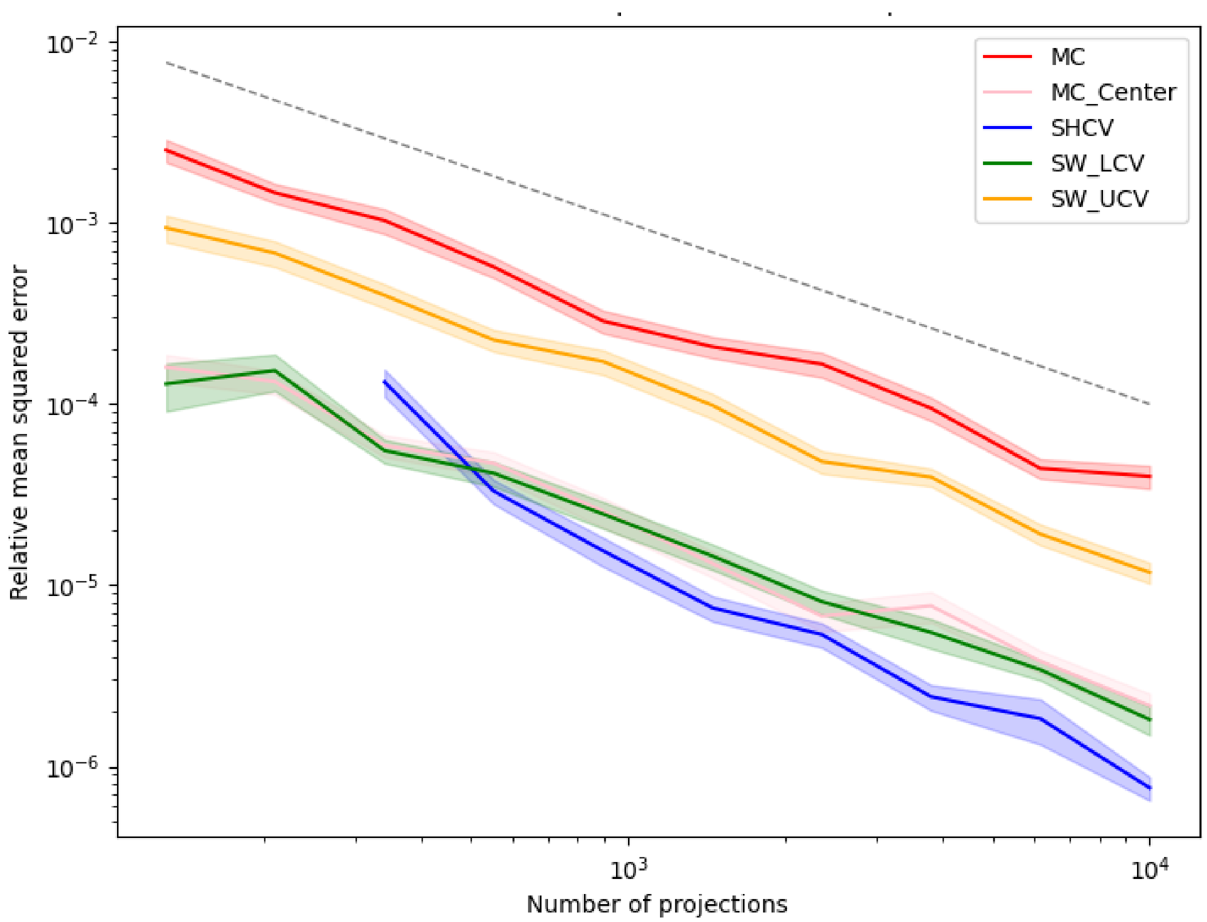


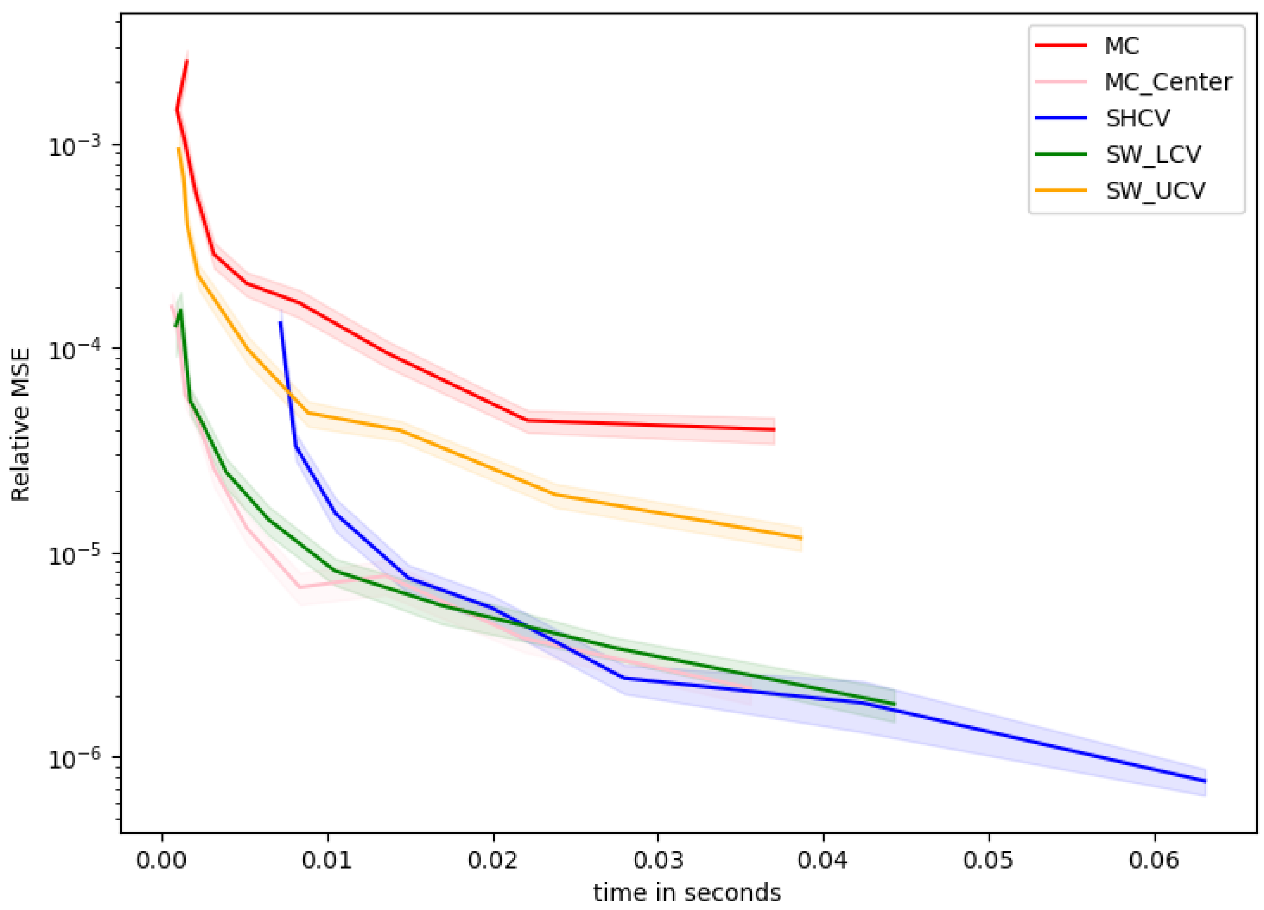
Relative MSE w/r to time and number of projection.

Measures are sampled from mixture of 5 multivariate Gaussians supported on $m = 10000$ diracs in dimension $d = 3$.

Numerical experiments

Dimension $d = 20$





Relative MSE w/r to time and number of projection.

Measures are sampled from mixture of 5 multivariate Gaussians supported on $m = 1000$ diracs in dimension $d = 20$.

Conclusion

Contribution

- GPU implementation state of the art control variates for SW_2
- Test of Neural control variates
- Centering measures gives a simple control variate.

Best control variate to compute SW:

- In low dimension ($d \leq 20$): Spherical Harmonics *[Leluc et al. '24]*
- In high dimension ($d \geq 20$): Centered measures