Bounding Wasserstein distance with couplings

Niloy Biswas[1] and Lester Mackey[2]

1: Harvard University (niloy_biswas@g.harvard.edu)

2: Microsoft Research New England (lmackey@microsoft.com)

Motivation: Assess quality of asymptotically biased Monte Carlo methods

- Large data applications have catalyzed interest in sampling methods such as **approximate MCMC** and **variational inference**.

- Such methods are **asymptotically biased**: they converge to a distribution Q that is different to the original distribution of interest P.

- We introduce estimators based on **couplings** of Markov chains to compute upper bounds for the **Wasserstein distance**,

$$\mathcal{W}_\mathbf{p}(\mathbf{P},\mathbf{Q}) = \inf_{\mathbf{X}\sim\mathbf{P},\mathbf{Y}\sim\mathbf{Q}} \mathbb{E}[\mathbf{c}(\mathbf{X},\mathbf{Y})^\mathbf{p}]^{\mathbf{1/p}}.$$

  - $\mathcal{W}_p(P,Q)$ can control difference between p^{th}-moment of P and Q.

Bounding Wasserstein distance with couplings

- Consider Markov chains $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ with marginal transition kernels K_1 and K_2 and invariant distributions P and Q.

- Construct kernel $\bar{\mathbf{K}}$ on the joint space such that for all $x,y\in\mathcal{X}$,

$$\bar{\mathbf{K}}\big((\mathbf{x},\mathbf{y}),(\cdot,\mathcal{X})\big) = \mathbf{K_1}(\mathbf{x},\cdot) \text{ and } \bar{\mathbf{K}}\big((\mathbf{x},\mathbf{y}),(\mathcal{X},\cdot)\big) = \mathbf{K_2}(\mathbf{y},\cdot).$$

- We propose the coupling upper bound (CUB) estimate

$$\mathbf{CUB_p} \triangleq \Big(\frac{\mathbf{1}}{\mathbf{I(T-S)}}\sum_{\mathbf{i=1}}^{\mathbf{I}}\sum_{\mathbf{t=S+1}}^{\mathbf{T}} \mathbf{c}(\mathbf{X_t^{(i)}},\mathbf{Y_t^{(i)}})^\mathbf{p}\Big)^{\mathbf{1/p}}.$$

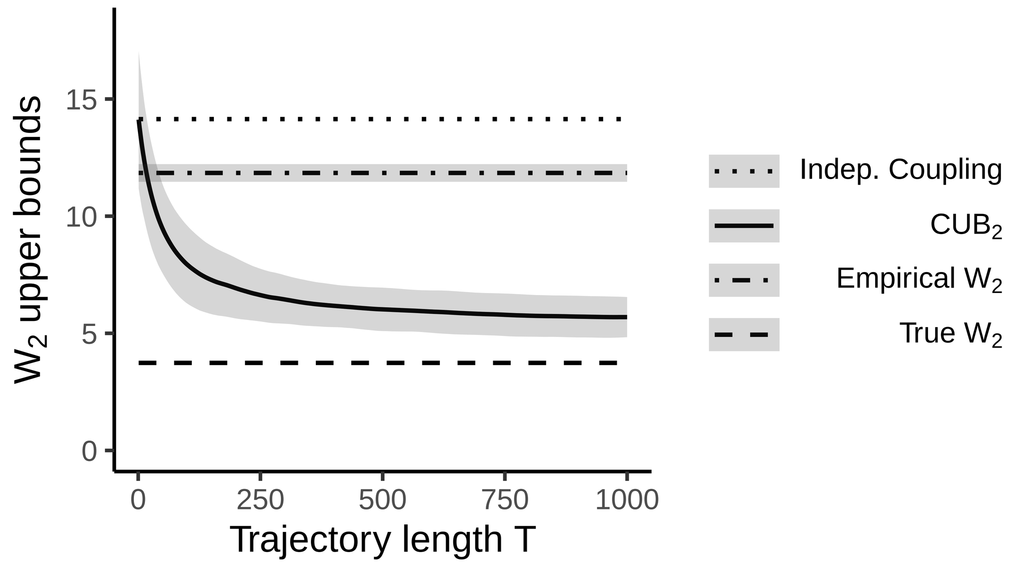
where $(X_t,^{(i)} Y_t^{(i)})_{t\geq 0}$ are independent chains sampled using $\bar{K}$.

A Stylized Example

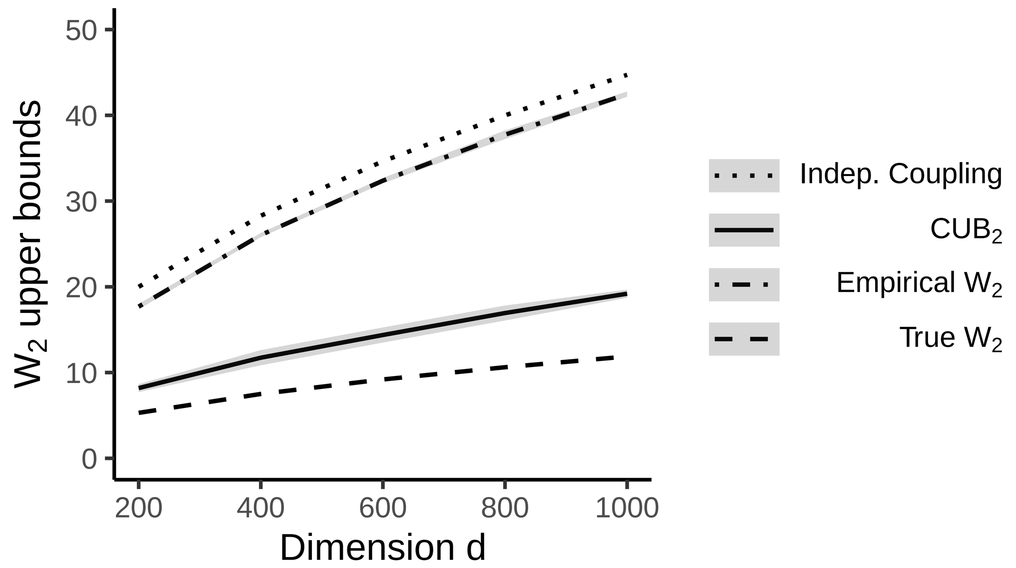
- $\mathcal{W}_2(P,Q)$ on $\mathbb{R}^d$ for

$$P=\mathcal{N}(0,\Sigma) \text{ where } \Sigma_{i,j}=0.5^{|i-j|}, Q=\mathcal{N}(0,I_d).$$

- We calculate CUB_2 using **common random numbers** coupling of marginal MALA kernels targeting P and Q.
  - Dimension $d=100$: tighter bounds for larger trajectory length T.



  - Higher dimensions d: favorable performance.



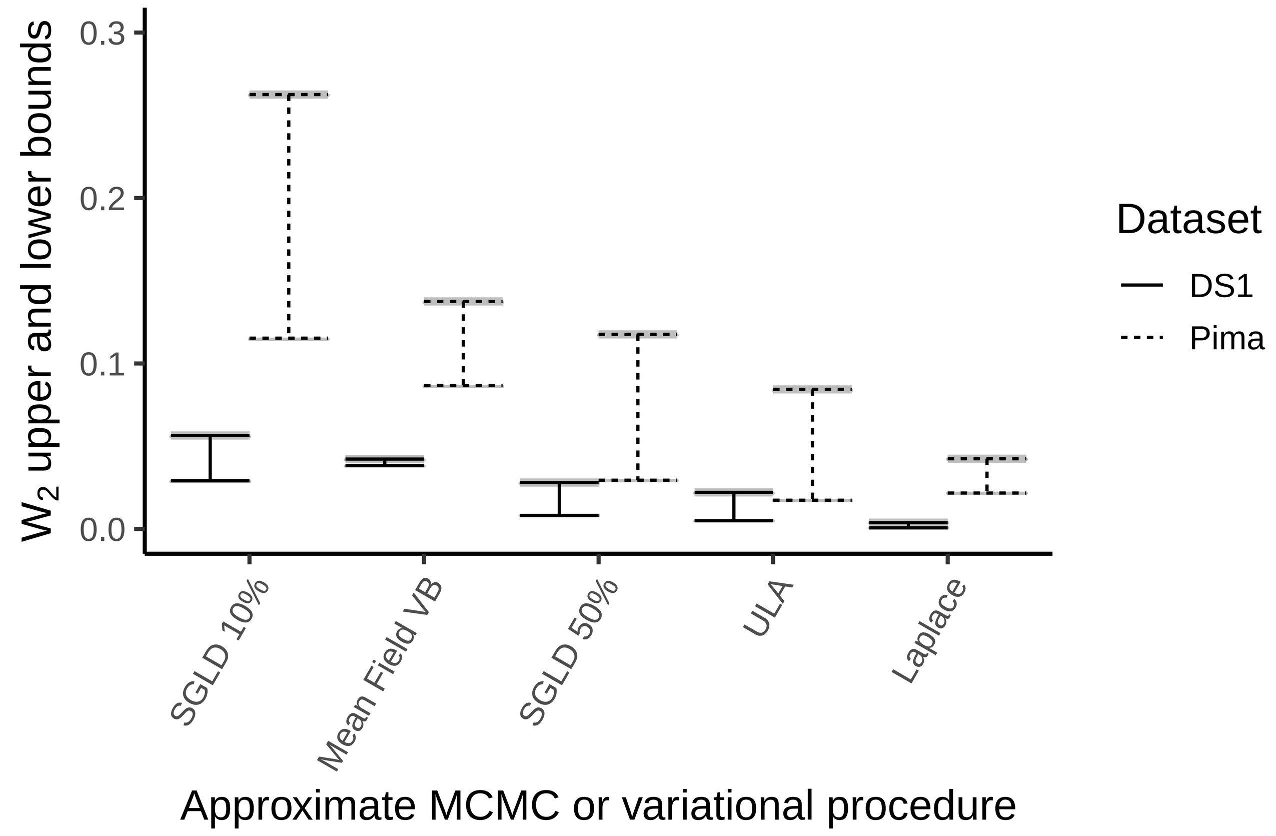
Consistency

Proposition 1. *Let $(X_t^{(i)},Y_t^{(i)})_{t\geq 0}$ for $i=1,\dots,I$ denote independent coupled chains generated using $\bar{\mathbf{K}}$. Suppose the marginal distributions $(P_t)_{t\geq 0}$ and $(Q_t)_{t\geq 0}$ converge in p-Wasserstein distance to distributions P and Q which have finite moments of order p. Then, for all $\epsilon>0$ there exists some $S\geq 1$ such that for all $T\geq S$, the estimator CUB$_p$ has finite moments of order p, and as $I\to\infty$,*

$$\text{CUB}_\text{p}^\text{p} \overset{\text{a.s.},\mathbf{L}^1}{\to} \mathbb{E}\big[\text{CUB}_\text{p}^\text{p}\big] \geq \mathcal{W}_\mathbf{p}(\mathbf{P},\mathbf{Q})^\mathbf{p}-\epsilon.$$

Stochastic Gradient MCMC and variational inference for tall data

- Baysian logistic regression with Gaussian priors:
  - DSI dataset: $n=26732$ observations and $d=10$ covariates
  - Pima Indians dataset: $n=768$ observations and $d=8$ covariates



Approximate MCMC for high-dimensional linear regression

- Baysian linear regression with Half-t(ν) priors:
  - Exact MCMC kernel: $\mathcal{O}(n^2 d)$ computation cost
  - ϵ-approximate MCMC based on matrix approximations

$$X\,\text{Diag}(\xi\eta_t)^{-1}X^\top \approx X\,\text{Diag}\big((\xi^{-1}\eta_j^{-1}\mathrm{I}_{\{\xi^{-1}\eta_j^{-1}>\epsilon\}})_{j=1}^p\big)\,X^\top$$

  - Synthetic dataset: $n=500$ observations and $d=50000$ covariates.