Linear-cost unbiased posterior estimates for crossed effects and matrix factorization models

via couplings

Paolo Ceriani & Giacomo Zanella

Bocconi University

May 31, 2024

Bocconi

Main Contributions

Context: Bayesian estimation of posterior quantities via MCMC

Main Contributions

Context: Bayesian estimation of posterior quantities via MCMC

- Provide a methodology for unbiased estimation of posterior quantities with linear computational cost in many models of interest (namely crossed random effect and matrix factorization models), leveraging couplings.

Main Contributions

Context: Bayesian estimation of posterior quantities via MCMC

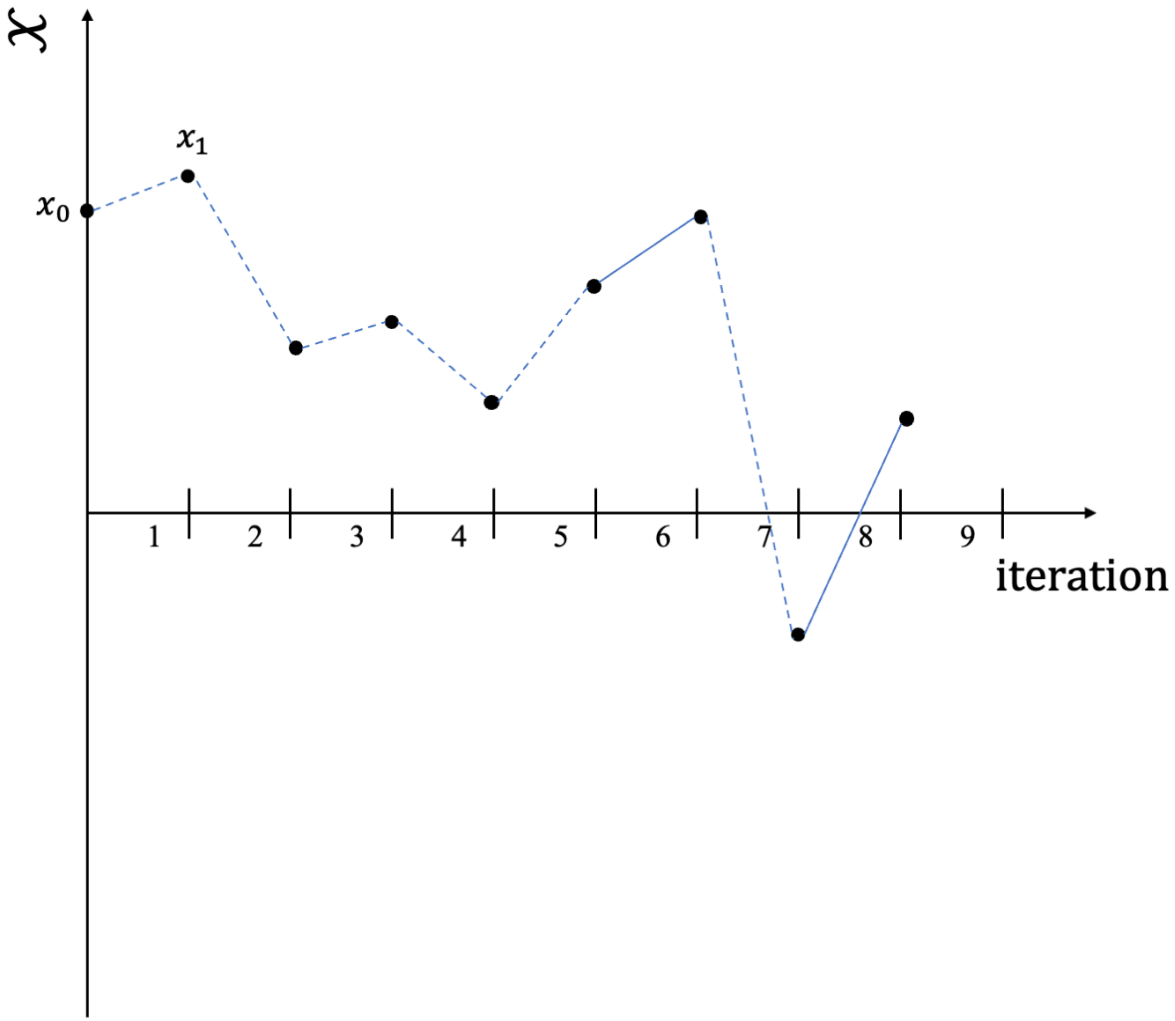
- Provide a methodology for unbiased estimation of posterior quantities with linear computational cost in many models of interest (namely crossed random effect and matrix factorization models), leveraging couplings.
- Find a **bound** on the expected number of iterations needed for the chains to meet, when coupled under previous strategy, and hence, on their computational cost.

Main Contributions

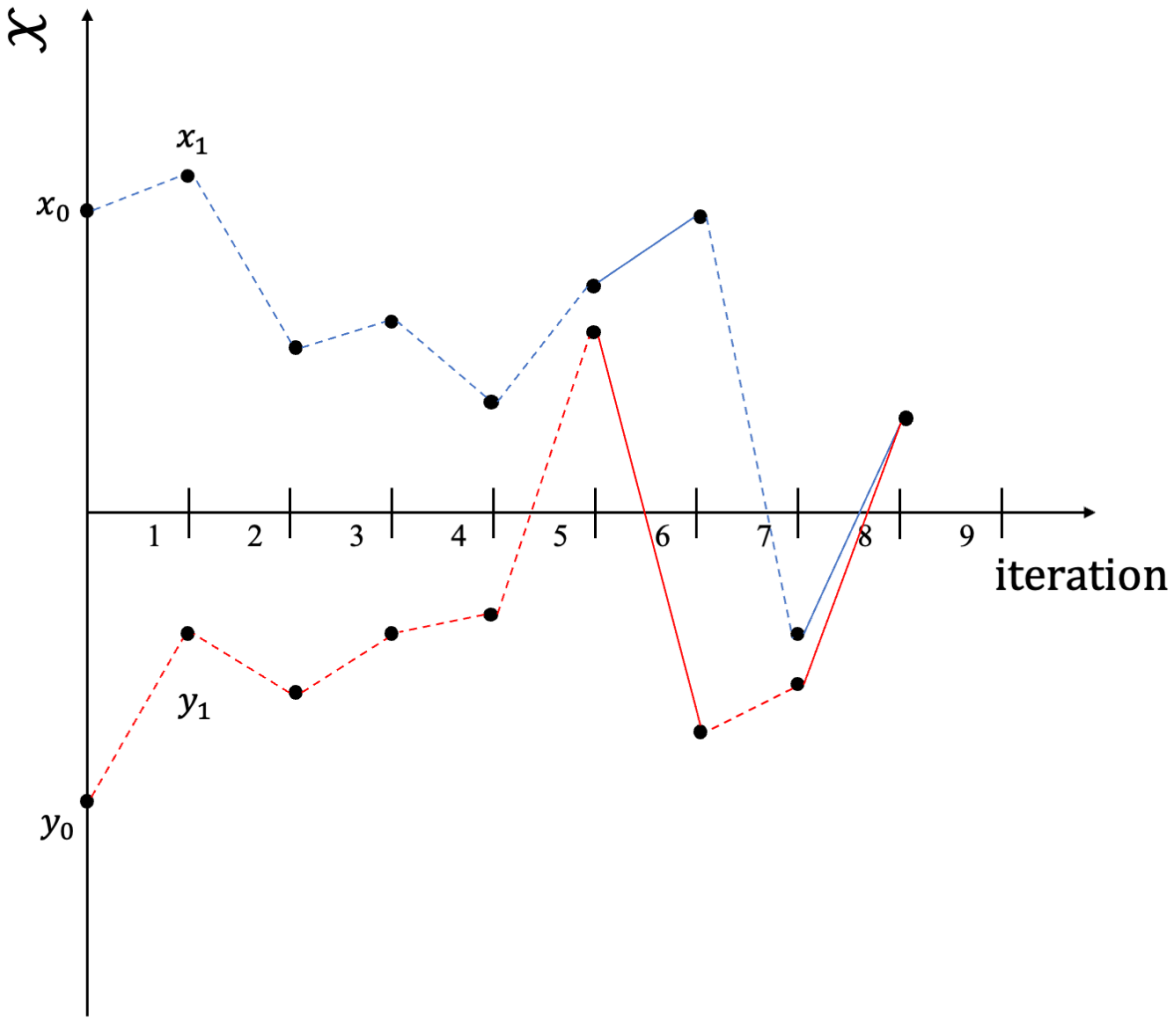
Context: Bayesian estimation of posterior quantities via MCMC

- Provide a methodology for unbiased estimation of posterior quantities with linear computational cost in many models of interest (namely crossed random effect and matrix factorization models), leveraging couplings.
- Find a **bound** on the expected number of iterations needed for the chains to meet, when coupled under previous strategy, and hence, on their computational cost.
- Extensive simulations on crossed random effects and probabilistic matrix factorization models, proving goodness of theory and methodology.

Unbiased estimates via couplings - Idea



Unbiased estimates via couplings - Idea



Couplings for estimation

We are interested in approximating expectations of the form

$$\mathbb{E}_\pi[h] = \int_{\mathcal{X}} h(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}),$$

where $\pi \in \mathcal{P}(\mathcal{X})$ is the target probability distribution and $h : \mathcal{X} \to \mathbb{R}^d$.

Couplings for estimation

We are interested in approximating expectations of the form

$$\mathbb{E}_\pi[h] = \int_{\mathcal{X}} h(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}),$$

where $\pi \in \mathcal{P}(\mathcal{X})$ is the target probability distribution and $h : \mathcal{X} \to \mathbb{R}^d$.

Jacob, O'Leary, and Atchadé 2020 proposed to use couplings to obtain **unbiased** estimates from (biased) MCMCs.
Let $\{\mathbf{X}^t\}_{t=1}^T, \{\mathbf{Y}^t\}_{t=1}^T$ be coupled (i.e. correlated) chains evolving with π-invariant kernel P. Initialize $(\boldsymbol{X}^0, \boldsymbol{Y}^0) \sim (\pi_0 P) \otimes \pi_0$ for some π_0. It follows $\mathbf{X}^{t-1} =^d \mathbf{Y}^t$.

Couplings for estimation

We are interested in approximating expectations of the form

$$\mathbb{E}_\pi[h] = \int_{\mathcal{X}} h(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}),$$

where $\pi \in \mathcal{P}(\mathcal{X})$ is the target probability distribution and $h : \mathcal{X} \to \mathbb{R}^d$.

Jacob, O'Leary, and Atchadé 2020 proposed to use couplings to obtain **unbiased** estimates from (biased) MCMCs.
Let $\{\mathbf{X}^t\}_{t=1}^T, \{\mathbf{Y}^t\}_{t=1}^T$ be coupled (i.e. correlated) chains evolving with π-invariant kernel P. Initialize $(\boldsymbol{X}^0, \boldsymbol{Y}^0) \sim (\pi_0 P) \otimes \pi_0$ for some π_0. It follows $\mathbf{X}^{t-1} =^d \mathbf{Y}^t$.

Under some regularity assumptions, if $T = \inf_t\{\mathbf{X}^t = \mathbf{Y}^t\}$, then an unbiased estimate of $\mathbb{E}_\pi[h(\mathbf{X})]$ is

$$H_k(\boldsymbol{X}, \boldsymbol{Y}) = h\left(\boldsymbol{X}^k\right) + \sum_{t=k+1}^{T-1} \left(h(\boldsymbol{X}^t) - h(\boldsymbol{Y}^t)\right).$$

Heuristic

$$
\begin{aligned}
\mathbb{E}_\pi[h(\mathbf{X})] &= \mathbb{E}\left[\lim_{t\to+\infty} h(\mathbf{X}^t)\right] = \mathbb{E}\left[h(\mathbf{X}^k) + \sum_{t=k+1}^{\infty} h(\mathbf{X}^t) - h(\mathbf{X}^{t-1})\right] \\
&= \mathbb{E}[h(\mathbf{X}^k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(\mathbf{X}^t) - h(\mathbf{X}^{t-1})] = \mathbb{E}[h(\mathbf{X}^k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(\mathbf{X}^t)] - \mathbb{E}[h(\mathbf{Y}^t)] \\
&= \mathbb{E}\left[h(\mathbf{X}^k) + \sum_{t=k+1}^{\infty} (h(\mathbf{X}^t) - h(\mathbf{Y}^t))\right] \\
&= \mathbb{E}\left[h(\mathbf{X}^k) + \sum_{t=k+1}^{T-1} (h(\mathbf{X}^t) - h(\mathbf{Y}^t))\right].
\end{aligned}
$$

Where we used $\mathbf{Y}^t =^d \mathbf{X}^{t-1}$, and that $T = \inf_t\{t \geq 0 : \mathbf{Y}^t = \mathbf{X}^t\} < +\infty$.
Unbiased $\forall k \geq 0$, but the cost and variance depends on it.

Actually

It is possible to improve the above estimator computing $H_k(\mathbf{X}, \mathbf{Y})$ for several values of k from the same realization and take the average. For $k \geq m$ consider:

$$
\begin{aligned}
H_{k:m}(\mathbf{X}, \mathbf{Y}) &= \frac{1}{m-k+1} \sum_{l=k}^{m} H_l(\mathbf{X}, \mathbf{Y}) \\
&= \frac{1}{m-k+1} \sum_{l=k}^{m} h(\mathbf{X}^l) + \sum_{l=k+1}^{T-1} \min\left(1, \frac{l-k}{m-k+1}\right) \left(h(\boldsymbol{X}^t) - h(\boldsymbol{Y}^t)\right)
\end{aligned}
$$

Couplings and Notation

Definition 1

Given $p, q \in \mathcal{P}(\mathcal{X})$, a coupling of p, q is a joint distributions on $\mathcal{X} \times \mathcal{X}$ whose first and second marginals are, respectively, p and q. We denote the space of such couplings as $\Gamma(p, q)$. We also write $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p, q)$ for random vectors $(\boldsymbol{X}, \boldsymbol{Y})$ s.t. $\mathbf{X} \sim p, \mathbf{Y} \sim q$.

Couplings and Notation

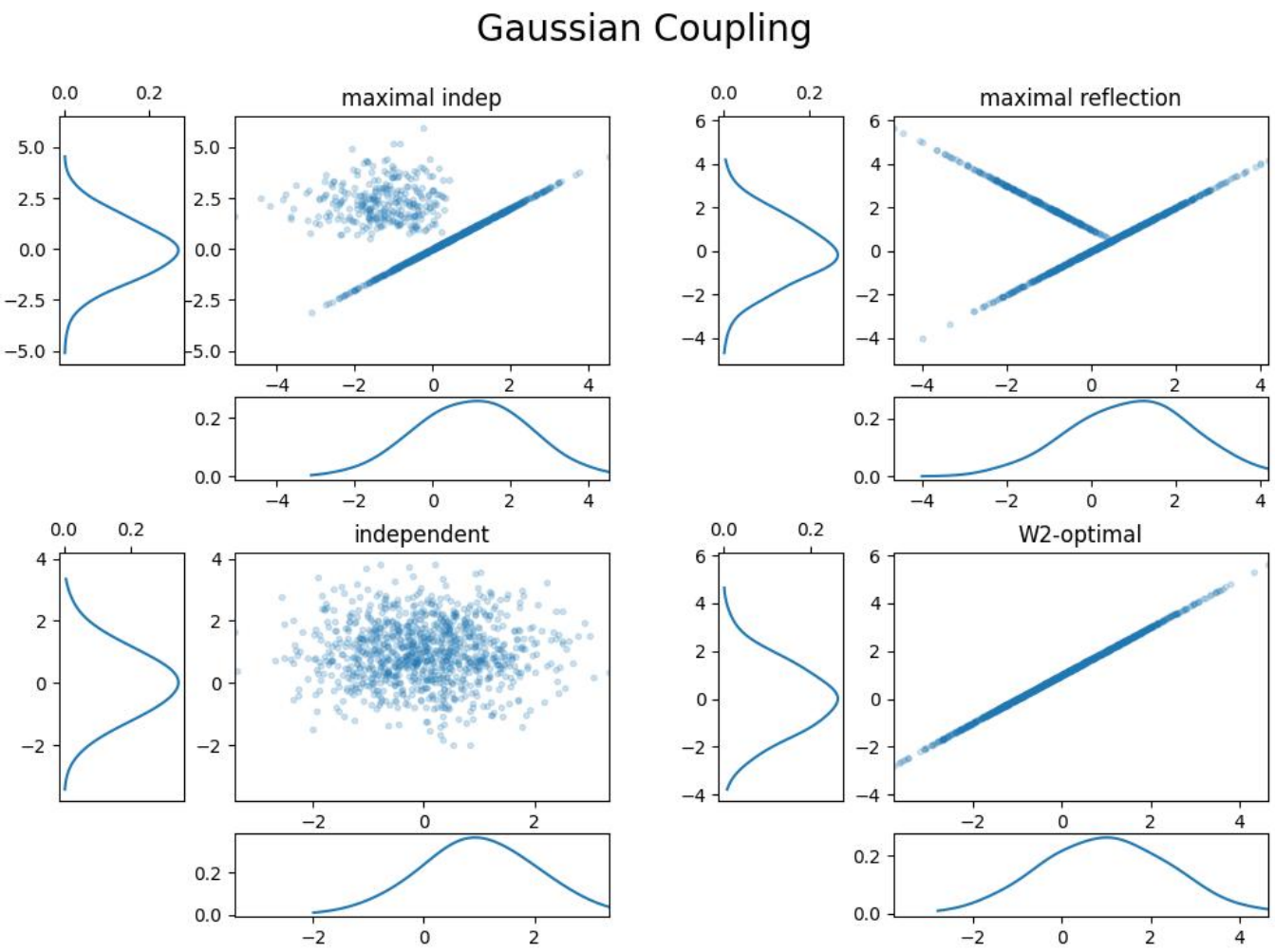
Definition 1

Given $p, q \in \mathcal{P}(\mathcal{X})$, a coupling of p, q is a joint distributions on $\mathcal{X} \times \mathcal{X}$ whose first and second marginals are, respectively, p and q. We denote the space of such couplings as $\Gamma(p, q)$. We also write $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p, q)$ for random vectors $(\boldsymbol{X}, \boldsymbol{Y})$ s.t. $\mathbf{X} \sim p, \mathbf{Y} \sim q$.

Definition 2

Consider a transition kernel $P : \mathcal{X} \times \mathcal{F} \to [0, 1]$, we denote $\bar{P}[P]$ a distribution on $\mathcal{X} \times \mathcal{X}$ such that $\bar{P}[P]((\boldsymbol{x}, \boldsymbol{y}), \cdot) \in \Gamma\left(P(\boldsymbol{x}, \cdot), P(\boldsymbol{y}, \cdot)\right)$ for every $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{X}$.

Example


Gaussian Coupling
maximal indep
maximal reflection
independent
W2-optimal

Optimal strategies for coupling chains

Coupling of Gibbs Chains

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, ..., \boldsymbol{\theta}_{(K)})$. In a Gibbs sampler we iteratively sample from $\pi\left(\boldsymbol{\theta}_{(k)} | \boldsymbol{\theta}_{(-k)}\right)$ for $k = 1, ..., K$ up to convergence. The resulting Gibbs Sampler kernel P can be written as the following composition of K kernels

$$P = P_K \cdots P_1, \tag{1}$$

$$P_k(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \pi(d\boldsymbol{\theta}'_{(k)} | \boldsymbol{\theta}_{(-k)}) \delta_{\boldsymbol{\theta}_{(-k)}}(d\boldsymbol{\theta}'_{(-k)}) \qquad k = 1, \ldots, K, \;\; \boldsymbol{\theta} \in \mathcal{X}\,. \tag{2}$$

Coupling of Gibbs Chains

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, ..., \boldsymbol{\theta}_{(K)})$. In a Gibbs sampler we iteratively sample from $\pi\left(\boldsymbol{\theta}_{(k)} | \boldsymbol{\theta}_{(-k)}\right)$ for $k = 1, ..., K$ up to convergence. The resulting Gibbs Sampler kernel P can be written as the following composition of K kernels

$$P = P_K \cdots P_1, \tag{1}$$

$$P_k(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \pi(d\boldsymbol{\theta}'_{(k)} | \boldsymbol{\theta}_{(-k)}) \delta_{\boldsymbol{\theta}_{(-k)}}(d\boldsymbol{\theta}'_{(-k)}) \qquad k = 1, \ldots, K, \;\; \boldsymbol{\theta} \in \mathcal{X}\,. \tag{2}$$

A strategy is to sequentially compose a coupling of each full conditional, i.e.

$$\bar{P}((\boldsymbol{x}, \boldsymbol{y}), \cdot) := \bar{P}[P_K] \cdots \bar{P}[P_1]\left((\boldsymbol{x}, \boldsymbol{y}), \cdot\right) \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}, \tag{3}$$

Coupling of Gibbs Chains

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, ..., \boldsymbol{\theta}_{(K)})$. In a Gibbs sampler we iteratively sample from $\pi\left(\boldsymbol{\theta}_{(k)} | \boldsymbol{\theta}_{(-k)}\right)$ for $k = 1, ..., K$ up to convergence. The resulting Gibbs Sampler kernel P can be written as the following composition of K kernels

$$P = P_K \cdots P_1, \tag{1}$$

$$P_k(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \pi(d\boldsymbol{\theta}'_{(k)} | \boldsymbol{\theta}_{(-k)}) \delta_{\boldsymbol{\theta}_{(-k)}}(d\boldsymbol{\theta}'_{(-k)}) \qquad k = 1, \ldots, K, \;\; \boldsymbol{\theta} \in \mathcal{X}\,. \tag{2}$$

A strategy is to sequentially compose a coupling of each full conditional, i.e.

$$\bar{P}((\boldsymbol{x}, \boldsymbol{y}), \cdot) := \bar{P}[P_K] \cdots \bar{P}[P_1]\left((\boldsymbol{x}, \boldsymbol{y}), \cdot\right) \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}, \tag{3}$$

$$\neq \bar{P}[P_K \cdots P_1]((\mathbf{x}, \mathbf{y}), \cdot)$$

Coupling of Gibbs Chains

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, ..., \boldsymbol{\theta}_{(K)})$. In a Gibbs sampler we iteratively sample from $\pi\left(\boldsymbol{\theta}_{(k)} | \boldsymbol{\theta}_{(-k)}\right)$ for $k = 1, ..., K$ up to convergence. The resulting Gibbs Sampler kernel P can be written as the following composition of K kernels

$$P = P_K \cdots P_1, \tag{1}$$

$$P_k(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \pi(d\boldsymbol{\theta}'_{(k)} | \boldsymbol{\theta}_{(-k)}) \delta_{\boldsymbol{\theta}_{(-k)}}(d\boldsymbol{\theta}'_{(-k)}) \qquad k = 1, \ldots, K, \;\; \boldsymbol{\theta} \in \mathcal{X}\,. \tag{2}$$

A strategy is to sequentially compose a coupling of each full conditional, i.e.

$$\bar{P}((\boldsymbol{x}, \boldsymbol{y}), \cdot) := \bar{P}[P_K] \cdots \bar{P}[P_1]\left((\boldsymbol{x}, \boldsymbol{y}), \cdot\right) \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}, \tag{3}$$

$$\neq \bar{P}[P_K \cdots P_1]((\mathbf{x}, \mathbf{y}), \cdot)$$

Remark

For BGS with c.i. blocks, univariate updates are equivalent to block updates. In general it is not the same for couplings.

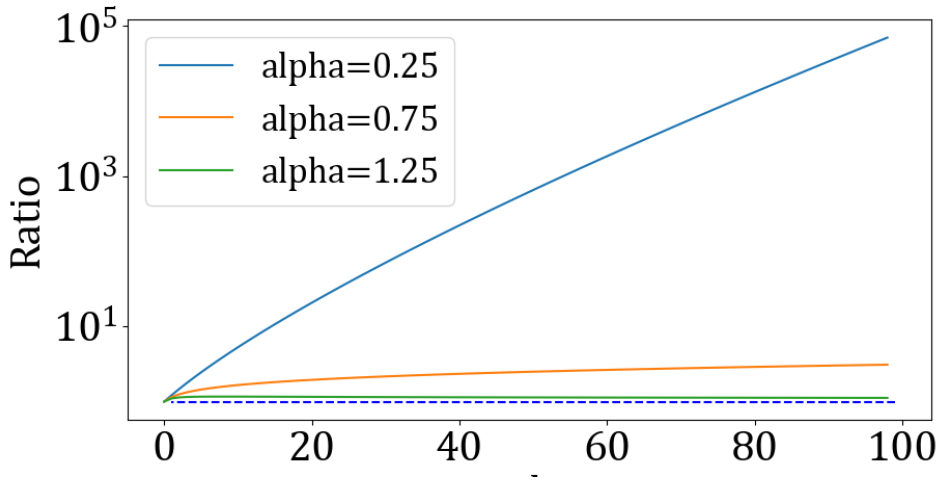
Gibbs couplings

Suppose that

$$P_k(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \pi(d\boldsymbol{\theta}'_{(k)}|\boldsymbol{\theta}_{(-k)})\delta_{\boldsymbol{\theta}_{(-k)}}(d\boldsymbol{\theta}'_{(-k)}) = \prod_i \pi(d\theta'_{(k),i}|\boldsymbol{\theta}_{(-k)})\delta_{\boldsymbol{\theta}_{(-k)}}(d\boldsymbol{\theta}'_{(-k)})$$

Let $p_i = \pi(dX_{(k),i}|\boldsymbol{X}_{(-k)})$, $q_i = \pi(dY_{(k),i}|\boldsymbol{Y}_{(-k)})$, then it holds:

$$\min_i Pr_{max}(p_i, q_i)) \geq Pr_{max}(p, q) \geq \prod_i Pr_{max}(p_i, q_i). \tag{4}$$



Two step Coupling of Markov Chains

We propose a two step technique as in Biswas et al. 2022: if the chains are "far away" in the space implement a contractive coupling, if "close enough", implement a maximal coupling.

$$\bar{P}[P]((\boldsymbol{x},\boldsymbol{y}),\cdot) = \begin{cases} \bar{P}^c[P]((\boldsymbol{x},\boldsymbol{y}),\cdot) & \text{if } d(\boldsymbol{x},\boldsymbol{y}) > \varepsilon \\ \bar{P}^m[P]((\boldsymbol{x},\boldsymbol{y}),\cdot) & \text{if } d(\boldsymbol{x},\boldsymbol{y}) \leq \varepsilon, \end{cases} \tag{5}$$

where $\bar{P}^m$ is a maximal coupling of the kernels within brackets, and $\bar{P}^c$ is a (hopefully optimal) contracting one.

Two step couplings

Algorithm 1: Two-step coupling algorithm

Input: initial distribution π_0, kernels P, $\bar{P}^c$, $\bar{P}^m$

Sample $\boldsymbol{X}^{-1} \sim \pi_0$, $\boldsymbol{Y}^0 \sim \pi_0$ and $\boldsymbol{X}^0 \sim P(\boldsymbol{X}^{-1}, \cdot)$;

while $\boldsymbol{X}^t \neq \boldsymbol{Y}^t$ **do**

  if $d(\boldsymbol{X}^t, \boldsymbol{Y}^t) > \varepsilon$ **then**

    $(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}) \sim \bar{P}^c[P]((\boldsymbol{X}^t, \boldsymbol{Y}^t), \cdot)$

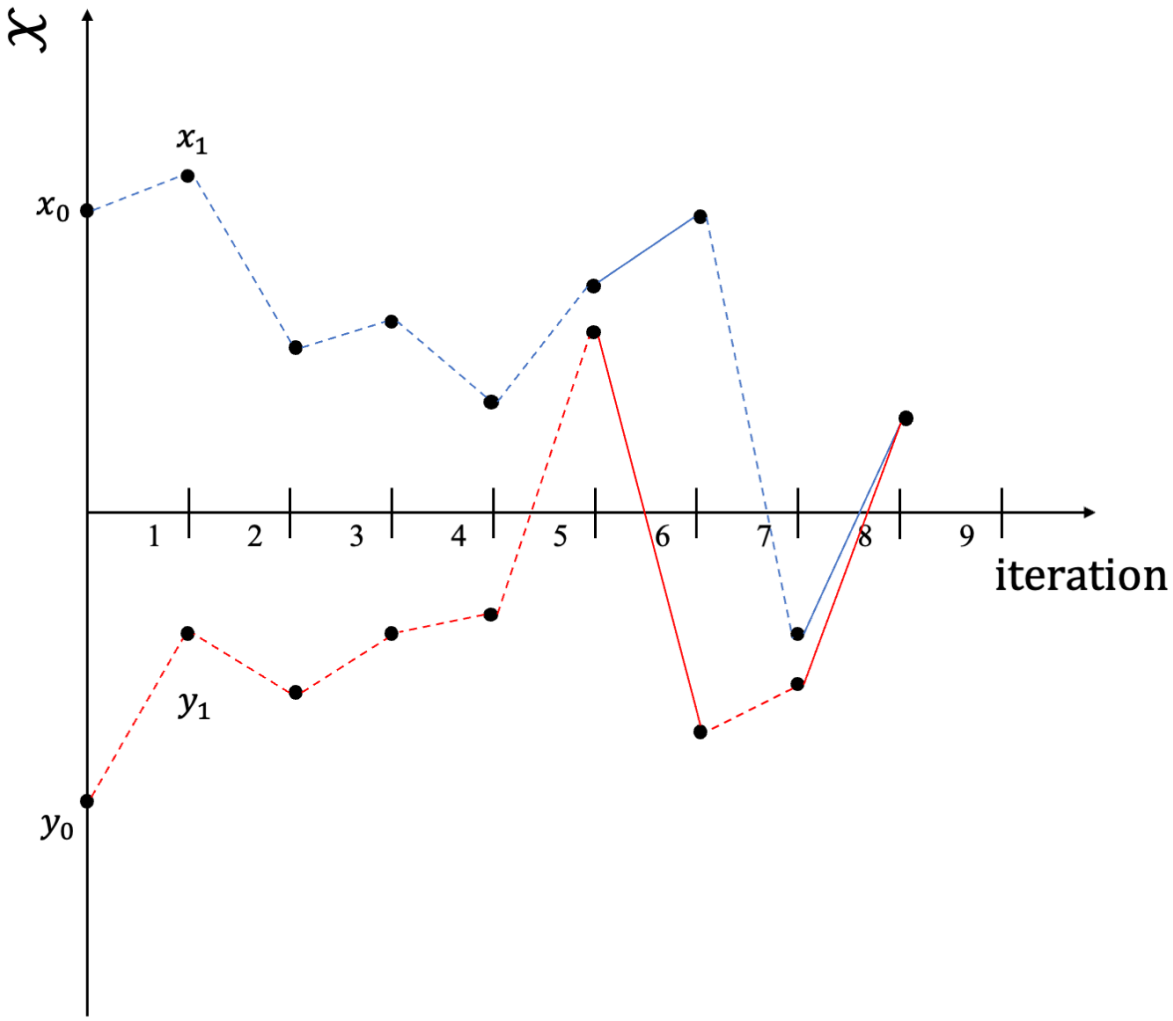
  else

    $(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}) \sim \bar{P}^m[P]((\boldsymbol{X}^t, \boldsymbol{Y}^t), \cdot)$

  $t \leftarrow t + 1$

Output: trajectory $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t \in \{0, \ldots, T\}}$

- again -



Bound on meeting time, π-reversible

Consider the forward-backward-scan kernel $P^{(FB)}$ defined as

$$P^{(FB)} = P_1 \cdots P_{K-1} P_K P_{K-1} \cdots P_1 \, .$$

Bound on meeting time, π-reversible

Consider the forward-backward-scan kernel $P^{(FB)}$ defined as

$$P^{(FB)} = P_1 \cdots P_{K-1} P_K P_{K-1} \cdots P_1 \,.$$

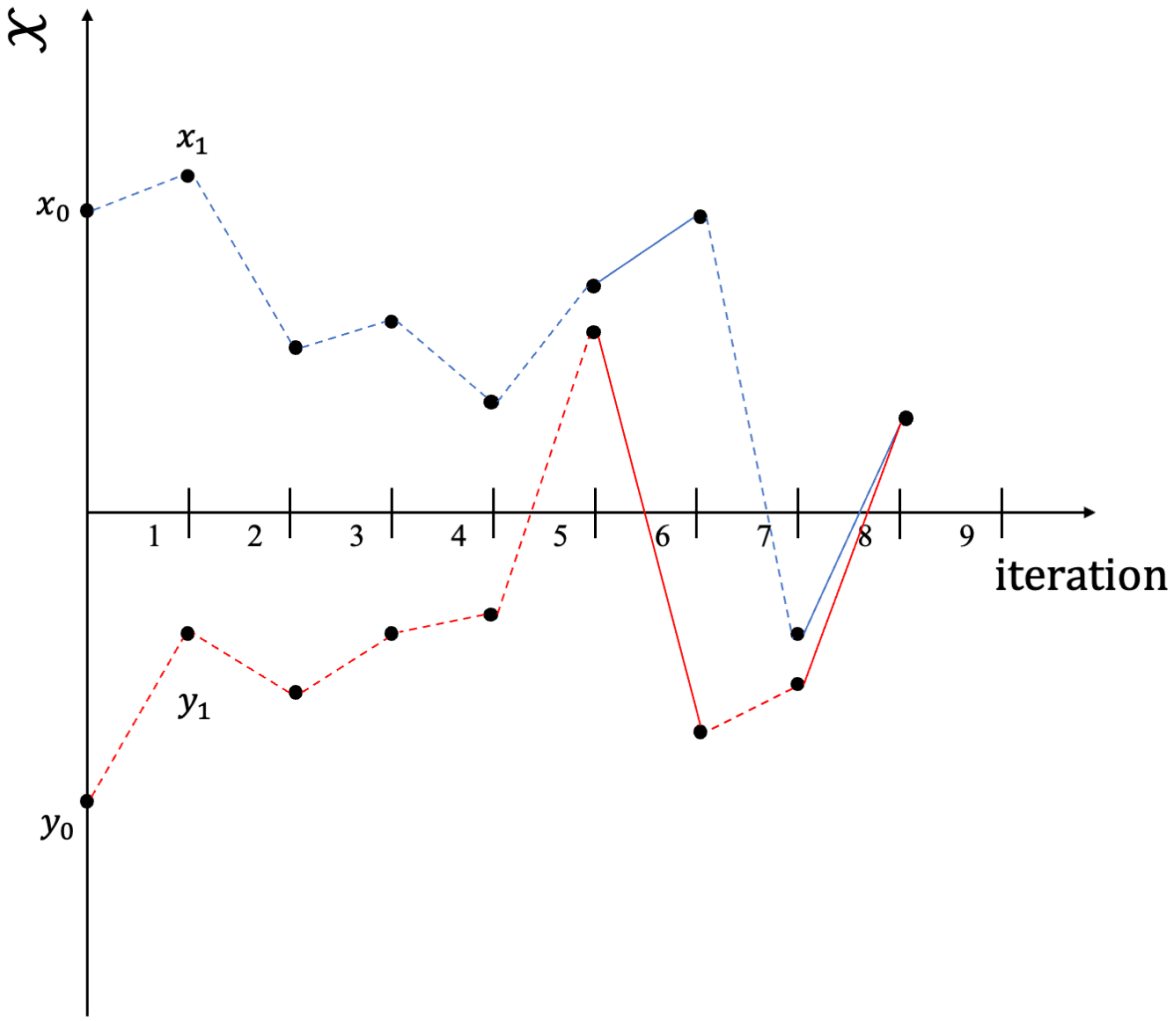
Theorem 3 (Bound for reversible chains)

Let $\pi = N(\boldsymbol{\mu}, \Sigma)$ and $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t \geq 0}$ be Markov chain marginally evolving with $P^{(FB)}$ and coupled via Algorithm 1. Let $T := \min\{t \geq 0 \mid \boldsymbol{X}^t = \boldsymbol{Y}^t\}$. Then

$$\mathbb{E}[T | \boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 4 + T_{rel} \left[\frac{1}{2} \ln(T_{rel}) + C_0 + C_\varepsilon \right], \tag{6}$$

where, $C_0 := \ln\left(\|L^{-1}(\boldsymbol{X}^0 - \boldsymbol{Y}^0)\|\right)$ with L s.t. $LL^\top = \Sigma$, and C_ε a constant depending on the fixed parameter ε of Algorithm 1, $T_{rel} := \frac{1}{1-\rho(B)}$ with B autoregressive matrix of Lemma 1 G. O. Roberts and Sahu 1997.

Sketch of the proof



Sketch of the proof

Let t_k be the k-th time at which $d_{tv}(\mathcal{L}(\boldsymbol{X}^{t+1}|\boldsymbol{X}^t), \mathcal{L}(\boldsymbol{Y}^{t+1}|\boldsymbol{Y}^t)) < \varepsilon$, i.e.

$$t_k := \min\{t > t_{k-1} : d_{tv}(\mathcal{L}(\boldsymbol{X}^{t+1}|\boldsymbol{X}^t), \mathcal{L}(\boldsymbol{Y}^{t+1}|\boldsymbol{Y}^t)) < \varepsilon\} \qquad k \geq 1, \tag{7}$$

with $t_0 := -1$ by convention. By the form of Algorithm 1, it follows we try maximal couplings only at iterations t_k. Also, let A_k be a binary variable indicating whether the maximal coupling attempt at t_k is successful, i.e.

$$A_k := \begin{cases} 1 & \text{if } \boldsymbol{X}^{t_k+1} = \boldsymbol{Y}^{t_k+1} \\ 0 & \text{otherwise} \end{cases}, \qquad k \geq 1. \tag{8}$$

By faithfulness, $A_k = 1$ implies that $\boldsymbol{X}^t = \boldsymbol{Y}^t$, $\forall t \geq t_k + 1$ and by convention $A_{k'} = 1$ for all $k' > k$. Thus, T can be written as

$$T = t_1 + 1 + \sum_{k=1}^{+\infty} (1 - A_k)(t_{k+1} - t_k). \tag{9}$$

MCMC convergence properties

Lemma 4

A Markov chain targeting a K-blocks $N(\boldsymbol{\mu}, \Sigma)$, can be written as (G. O. Roberts and Sahu 1997):

$$\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t \sim N\left(B\boldsymbol{\theta}_t + \boldsymbol{b}, \Sigma - B\Sigma B'\right), \tag{10}$$

where $Q = \Sigma^{-1} b = (I - B)\boldsymbol{\mu}$.

It follows:

$$\boxed{T_{rel} \approx \frac{1}{1-\rho(B)}.}$$

Bound for two blocks Gibbs

Consider now a two blocks Gibbs kernel, i.e.

$$P^{(2b)} = P_2 P_1 \qquad P_i(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \pi\left(d\boldsymbol{\theta}'_{(j)} | \boldsymbol{\theta}_{(i)}\right) \delta_{\boldsymbol{\theta}_{(i)}}(d\boldsymbol{\theta}'_{(i)}) \text{ for } i, j = 1, 2$$

Theorem 5

Let $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t \geq 0}$ be Markov chain marginally evolving with P with $K = 2$ blocks, coupled via Algorithm 1, let $T := \min\{t \geq 0 \mid \boldsymbol{X}^t = \boldsymbol{Y}^t\}$ as before. It holds that

$$\mathbb{E}[T | \boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 5 + T_{rel}\left[C_0 + C_\varepsilon\right], \tag{11}$$

where C_0, C_ε as in Theorem 3 and $T_{rel} = \frac{1}{1-\rho(B)}$ for B the autoregressive matrix as in Lemma 1 G. O. Roberts and Sahu 1997.

Unbiased estimates of crossed random effect models

Crossed Random Effect models

Models for recommending systems: y_n is a rating given by costumer $i[n]$ to film $j[n]$, and

$$\mathcal{L}(y_n|\mu, \mathbf{a}, \tau) = N\left(\mu + a_{i[n]} + a_{j[n]}, \tau^{-1}\right) \quad i = 1, ..., I_1 \; j = 1, ..., I_2.$$

Crossed Random Effect models

Models for recommending systems: y_n is a rating given by costumer $i[n]$ to film $j[n]$, and

$$\mathcal{L}(y_n|\mu, \mathbf{a}, \tau) = N\left(\mu + a_{i[n]} + a_{j[n]}, \tau^{-1}\right) \quad i = 1, ..., I_1 \; j = 1, ..., I_2.$$

Generally, **additive models** that relates a **response** variable to K **categorical** ones, whose effects are unknown and need to be estimated.

- K categorical variables, each with I_k different levels for $k = 1...K$,
- The effect of the j-th level of the k-th factor is described by an unknown random variable $a_j^{(k)}$.

$$y_n|\mu, \mathbf{a}, \boldsymbol{\tau} \sim N\left(\mu + \sum_{k=1}^{K} a_{i_k[n]}^{(k)}, \frac{1}{\tau_0}\right) \text{ for } j = 1, ..., N$$

Vanilla algorithms

Simple models whose computational cost can be overwhelmingly high:

Frequentist estimation : either via OLS (inefficient) or GLS

COMPUTATIONAL COST

$O(N^{\frac{3}{2}})$ [Ghosh, Hastie, and A. B. Owen 2022]

Vanilla Gibbs sampler: exploit block updates

for $t=1,\dots,T$ **do**

$\mu \sim \mathcal{L}(\mu|y, \mathbf{a}, \boldsymbol{\tau})$

for $k = 1, \dots, K$ **do**

$$\mathbf{a}^{(k)} \sim \mathcal{L}(\mathbf{a}^{(k)}|y, \mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}) = \otimes\mathcal{L}(a_i^{(k)}|y, \mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau})$$

COMPUTATIONAL COST

$O(N) \cdot O(\sqrt{N})$ [$l_1 = l_2$,Gao and A. Owen 2016]

Frequentist & Bayesian estimation

State of the art algorithms:

Backfitting (GLS): iterative algorithm maximizing $p(\mu, \mathbf{a}|y)$ via coordinate wise ascent. On Gaussians:

$$p(x) \propto \exp\{-x^T Q x/2 + x^T b\}$$

$$m^{(k)} \leftarrow -(Q^{(k,k)})^{-1} \sum_{l \neq k} Q^{(k,l)} m^{(l)} + (Q^{(k,k)})^{-1} b^{(k)}$$

COMPUTATIONAL COST

$O(1) \cdot O(N)$ [Ghosh, Hastie, and A. B. Owen 2022]

Collapsed Gibbs sampler: integrate μ out;

for *t=1,...,T* **do**
  for *k=1,...,K* **do**
    $\mu \sim \mathcal{L}(\mu|y, \mathbf{a}^{(-k)}, \boldsymbol{\tau})$
    $\mathbf{a}^{(k)} \sim \mathcal{L}(\mathbf{a}^{(k)}|y, \mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau})$

COMPUTATIONAL COST

$O(1) \cdot O(N)$ [Papaspiliopoulos, G O Roberts, and Zanella 2019][1]

[2] *for balanced cells design or balanced levels and K=2*

Theoretical results

Combining Theorem 5 with the results in Omiros Papaspiliopoulos, Stumpf-Fétizon, and Giacomo Zanella 2021 , we obtain the following bound for the expected meeting times.

Theorem 6

Let $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t\geq 0}$ be as in Theorem 5 and let $\pi = N(\boldsymbol{\mu}, \Sigma)$ be the posterior distribution of CREM with $K = 2$ factors, fixed $\boldsymbol{\tau}$ and design $(n_{ij})_{i,j}$ picked uniformly at random from $\mathcal{D}(n, d_1, d_2)$. Then

$$\Pr\left(\mathbb{E}[T|\boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 5 + C\left(1 + \frac{2}{\sqrt{\min\{d_1, d_2\} - 2}} + \epsilon\right)[C_0 + C_\varepsilon]\right) \to 1,$$

as $N \to +\infty$, where C_ε, C_0 as in Theorem 5, where the probability is with respect to the randomness of the design.

Asymptotic regimes

We study the behaviour of coalescence time and the previous bounds of Theorem 1 in two different asymptotic regimes: both with K=2, but different missingness patterns.

1. an observation of given combination of two factor levels i, j is seen with probability $p = 0.1$, and we let the level number grows to infinity:

$$Z_{ij} \sim Bern(p)$$

$$I = O(\sqrt{N})$$

2. the probability of observing an observation decreases as I increases:

$$Z_{ij} \sim Bern(10/I)$$

$$I = O(N)$$

Outfill regime 1

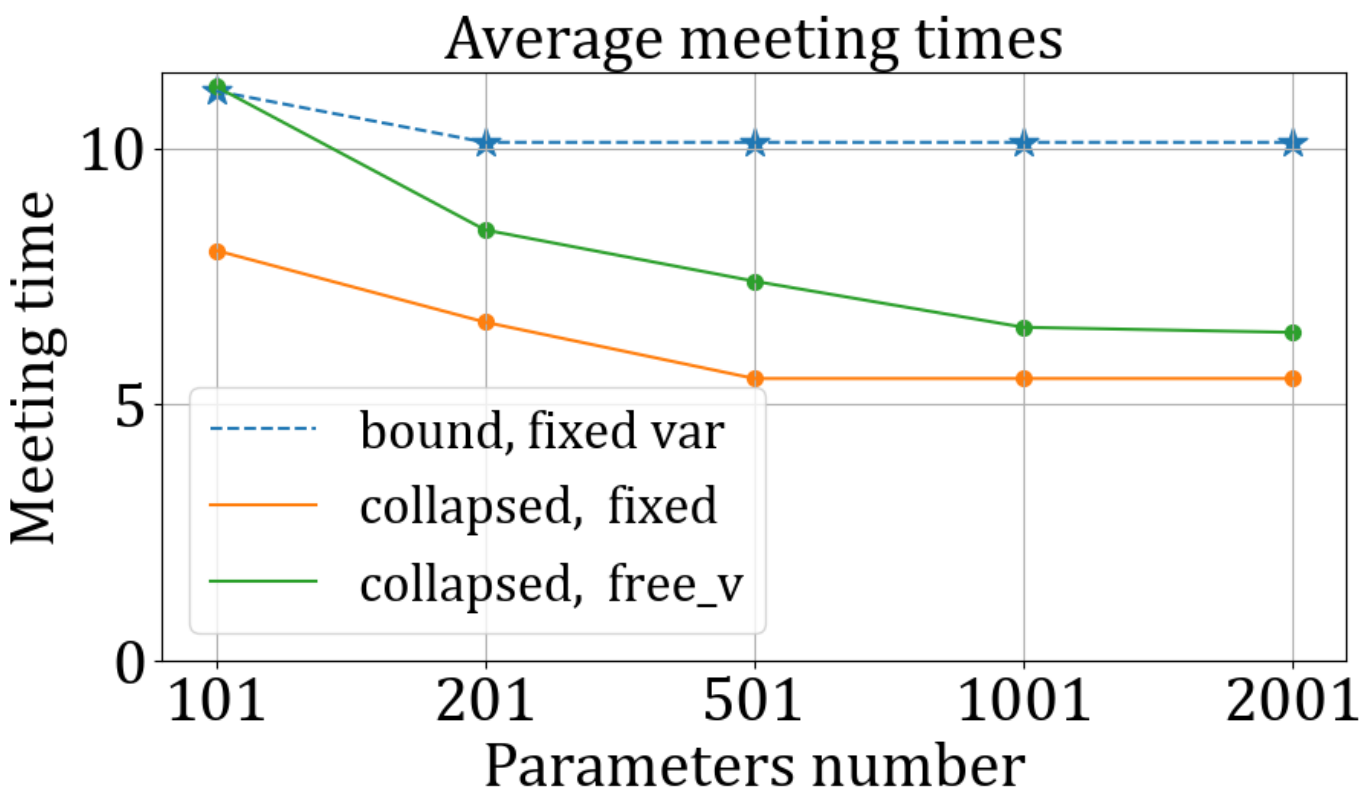


Figure 1: estimated mean number of iterations and bounds for $K = 2, I = \{50, 100, 250, 500, 750, 1000\}, \tau_1 = \tau_2 = 1$. Observing probability $p = 0.1$, log scale.

Outfil regime 2

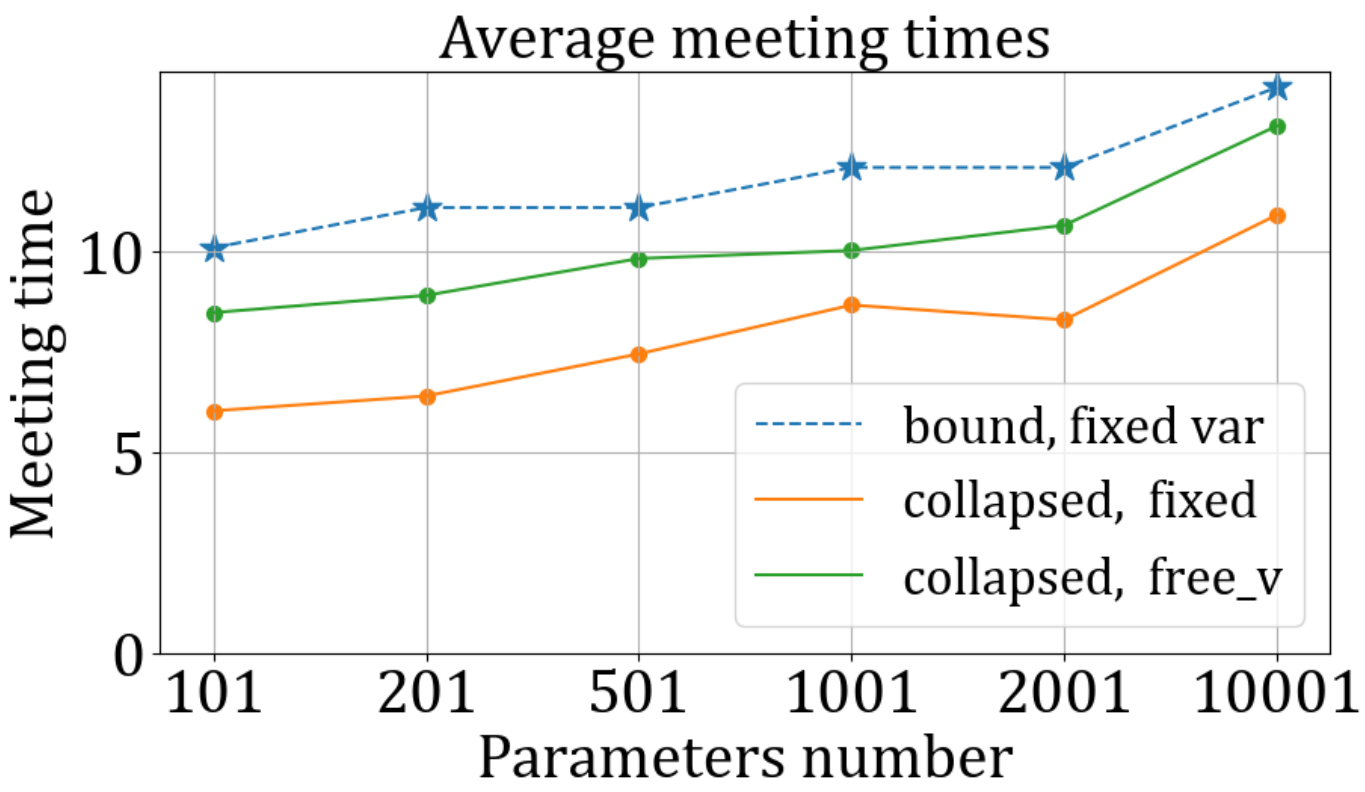


Figure 2: estimated mean number of iterations and bounds for $K = 2, I = \{50, 100, 250, 500, 750, 1000\}, \tau_1 = \tau_2 = 1$. Observing probability $p = 10/I$, log scale.

InstEval Dataset

Dataset containing university lecture evaluations by students at ETH Zurich. It contains 73421 observations, each corresponding to a score ranging from 1 to 5, assigned to a lecture together with 6 factors potentially impacting such score, such as identity of the student giving the rating or department that offers the course. We have $N = 73421, K = 6$ and $(I_1, ..., I_K) = (2972, 1128, 4, 6, 2, 14)$. The results are shown in the table below:

	Factor number	mean #iter
col-	[1,2]	8.1
	[1,6]	7.53
vanilla	[1,2]	39.3
	[1,6]	127.6

Non-Gaussian case

If non-gaussian response, then no collapsed is possible, and local centering within each block as in Omiros Papaspiliopoulos, Gareth O. Roberts, and Sköld 2007:

$$(\mu, \mathbf{a}^{(k)}) \to (\mu, \boldsymbol{\xi}^{(k)}),\ \boldsymbol{\xi}^{(k)} = \mu + \mathbf{a}^{(k)}.$$

We exploit algorithm in Omiros Papaspiliopoulos, Stumpf-Fétizon, and Giacomo Zanella 2021:

Algorithm 2: Gibbs sampler with local centering for non Gaussian likelihoods

for *k= 1,...,K* **do**

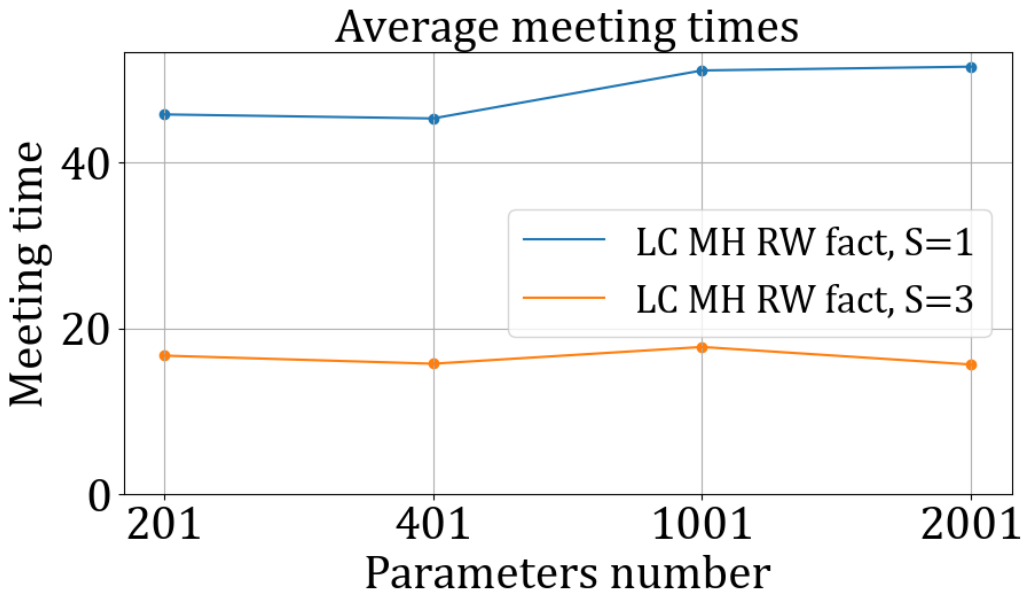
- Reparametrize $(\mu, \mathbf{a}^{(k)}) \to (\mu, \boldsymbol{\xi}^{(k)})$
- Draw μ from $\mathcal{L}(\mu|\boldsymbol{\xi}^{(k)}) = N\left(\frac{\tau_0\mu_0+\tau_k\sum_{i=1}^{l_k}\xi_i^{(k)}}{l_k\tau_k}, \frac{1}{\tau_0+l_k\tau_k}\right)$
- **for** i=$1, \ldots, l_k$ **do**
  - draw $\xi_i^{(k)}$ from $\mathcal{L}(\xi_i^{(k)}|\mathbf{y}, \tau_1, \ldots, \tau_k, \mu, a^{-(k)})$
- Reparametrize $(\mu, \boldsymbol{\xi}^{(k)}) \to (\mu, \mathbf{a}^{(k)})$

Sampling from $\mathcal{L}(\xi_i^{(k)}|\mathbf{y}, \tau_1, \ldots, \tau_k, \mu, a^{-(k)})$ requires MwG.

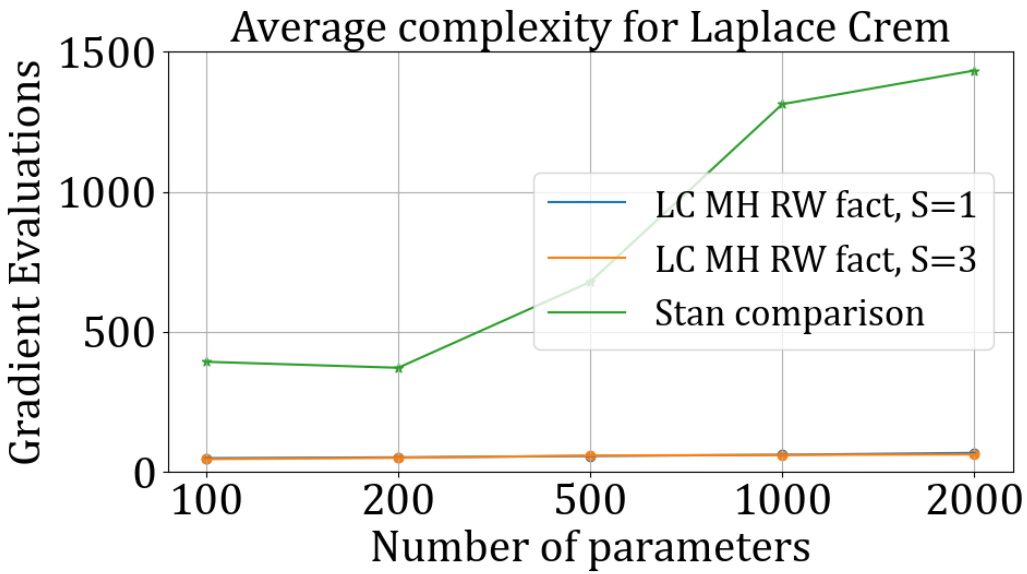
Laplace response

$$y_n|\mu, \mathbf{a} \sim \textit{Laplace}\left(\mu + \sum_{k=1}^{K} a_{i_{k[n]}}^{(k)}, b\right) \; n = 1, \ldots, N$$

Below the estimated mean number of iterations for
$K = 2, I = \{50, 100, 250, 500\}$, $\tau_1 = \tau_2 = 1, b = 1$ with Laplace response.

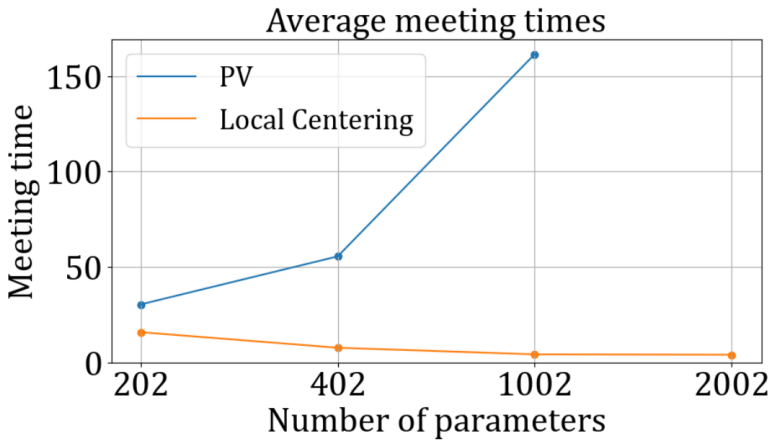


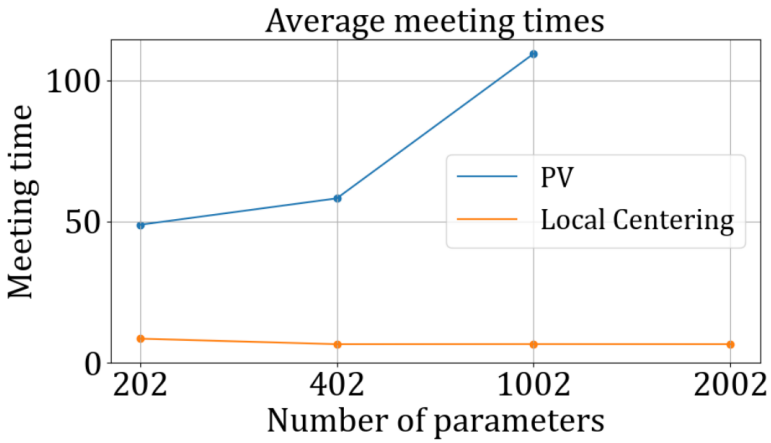
Stan Comparison



Probabilistic Matrix factorization

$$
\begin{aligned}
y_n &\sim N(\rho \mathbf{u}_{i_1[n]} \mathbf{v}_{i_2[n]}, \tau_0^{-1}), & n = 1, \ldots, N \qquad (12)\\
\mathbf{u}_i, \mathbf{v}_j &\sim N(\mathbf{0}, 1_d), & i = 1, \ldots, I_1;\ j = 1, \ldots, I_2\\
\tau_0 \sim \Gamma(c, d), \rho^{-\frac{1}{2}} &\sim \Gamma(a, b),
\end{aligned}
$$





Two Step in detail

W2 optimal coupling

They minimize the expected square distance between draws from p and q. i.e.:

$$\Gamma_{W_2} = argmin_{\Gamma \in \Gamma(p,q)} \ \mathbb{E}_{(X,Y)\sim\Gamma} \left[\|X - Y\|^2 \right]$$

W2 optimal coupling

They minimize the expected square distance between draws from p and q. i.e.:

$$\Gamma_{W_2} = argmin_{\Gamma \in \Gamma(p,q)} \ \mathbb{E}_{(X,Y)\sim\Gamma} \left[\|X - Y\|^2 \right]$$

For every univariate distributions we have:

Lemma 1

Common random number coupling is optimal for any cost $c(x, y)$ of the form $c(x, y) = h(x - y)$ for $h(\cdot)$ convex.

Sampling from a W_2 optimal coupling

Lemma 7

Let $p = N(\boldsymbol{\xi}, \Sigma_1)$ and $q = N(\boldsymbol{\nu}, \Sigma_2)$ be d-dimensional Gaussian, with $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$. Define

$$\Gamma_{W_2}(p,q) := N\left(\begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \Sigma_1 & FG^\top \\ GF^\top & \Sigma_2 \end{pmatrix}\right), \tag{13}$$

where $FF^\top = \Sigma_1, GG^\top = \Sigma_2$. Then $\Gamma_{W_2}(p,q)$ is a W_2-optimal coupling of p and q.

Note that the variance covariance matrix of Γ_{W_2} above is singular. In order to sample:

$$\begin{aligned} &\mathbf{Z} \sim N(\mathbf{0}_d, 1_d) \\ &\begin{cases} \mathbf{X} = \boldsymbol{\mu} + F\mathbf{Z} \\ \mathbf{Y} = \boldsymbol{\nu} + G\mathbf{Z}. \end{cases} \end{aligned} \tag{14}$$

W_2 optimality

It is possible to show that the previous optimal coupling still remains optimal if iterated for n steps in a Markov chain:

Lemma 2

Consider $(\boldsymbol{X}_t)_{t\geq 1}, (\boldsymbol{Y}_t)_{t\geq 1}$, chains arising from Gibbs sampler targeting Gaussian distribution. Iterating n steps of Γ_{W_2} on $\mathcal{L}(\boldsymbol{X}_{(k)}|\boldsymbol{X}_{(-k)}), \mathcal{L}(\boldsymbol{Y}_{(k)}|\boldsymbol{Y}_{(-k)}), k = 1, .., K$ is W_2 optimal, i.e.

$$\mathbb{E}[\|\boldsymbol{X}_{t+n} - \boldsymbol{Y}_{t+n}\|^2 | \boldsymbol{X}_t, \boldsymbol{Y}_t] = W_2^2 \left(\mathcal{L}(\boldsymbol{X}_{t+n}|\boldsymbol{X}_t), \mathcal{L}(\boldsymbol{Y}_{t+n}|\boldsymbol{Y}_t)\right),$$

where $W_2(\cdot,\cdot)$ indicates the Wasserstein 2 distance between distributions.

Conclusions

- Explicit bound on number of iteration (hence on computational cost) for Gaussian Gibbs sampler, of the order of $T_{rel} \log T_{rel}$.
- Methodology matching state of the art techniques, providing unbiased estimates.
- Insights on designing scalable strategies for general coupling algorithms.

Thank you

Bibliography I

Biswas, Niloy et al. (Mar. 2022). “Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84. DOI: 10.1111/rssb.12495.

Gao, Katelyn and Art Owen (Jan. 2016). “Efficient moment calculations for variance components in large unbalanced crossed random effects models”. In: *Electronic Journal of Statistics* 11. DOI: 10.1214/17-EJS1236.

Ghosh, Swarnadip, Trevor J. Hastie, and Art B. Owen (2022). “Backfitting for large scale crossed random effects regressions”. In: *The Annals of Statistics*.

Jacob, Pierre E., John O’Leary, and Yves F. Atchadé (2020). “Unbiased Markov chain Monte Carlo methods with couplings”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.3, pp. 543–600. DOI: https://doi.org/10.1111/rssb.12336. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12336. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12336.

Bibliography II

Papaspiliopoulos, O, G O Roberts, and G Zanella (Nov. 2019). "Scalable inference for crossed random effects models". In: *Biometrika* 107.1, pp. 25–40. ISSN: 0006-3444. DOI: 10.1093/biomet/asz058. eprint: https://academic.oup.com/biomet/article-pdf/107/1/25/32450844/asz058.pdf. URL: https://doi.org/10.1093/biomet/asz058.

Papaspiliopoulos, Omiros, Gareth O. Roberts, and Martin Sköld (Feb. 2007). "A General Framework for the Parametrization of Hierarchical Models". In: *Statistical Science* 22.1. DOI: 10.1214/088342307000000014. URL: https://doi.org/10.1214%2F088342307000000014.

Papaspiliopoulos, Omiros, Timothée Stumpf-Fétizon, and Giacomo Zanella (2021). *Scalable computation for Bayesian hierarchical models*. DOI: 10.48550/ARXIV.2103.10875. URL: https://arxiv.org/abs/2103.10875.

Bibliography III

Roberts, G. O. and S. K. Sahu (1997). "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 59.2, pp. 291–317. ISSN: 00359246. URL: http://www.jstor.org/stable/2346048.

Bound for invariant chains

Theorem 8

Let $\pi = N(\boldsymbol{\mu}, \Sigma)$ and $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t\geq 0}$ be Markov chain evolving with kernel P, coupled via Algorithm 1. For any $\delta > 0$, it holds that

$$\mathbb{E}[T|\boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 4+3\max\left(n_\delta^*, (1+\delta)T_{rel}\left[-\frac{1}{2}\ln\big(1-\lambda_{\min}(NN^\top)\big) + C_0 + C_\varepsilon\right]\right), \tag{15}$$

with $N = L^{-1}BL$, $LL^\top = \Sigma$, $C_0, C_\varepsilon, L, \lambda_{min}$ as in Theorem 3 and

$$n_\delta^* := \inf_{n_0}\left\{n_0 \geq 1 : \forall n \geq n_0 \;\; 1 - \|N^n\|_2^{\frac{1}{n}} \geq \frac{1-\rho(N)}{1+\delta}\right\}.$$

Coupling of distributions

Definition 9

Given $p, q \in \mathcal{P}(\mathcal{X})$, a coupling of p, q is a joint distributions on $\mathcal{X} \times \mathcal{X}$ whose first and second marginals are, respectively, p and q. We denote the space of such couplings as $\Gamma(p, q)$. We also write $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p, q)$ for random vectors $(\boldsymbol{X}, \boldsymbol{Y})$ s.t. $\mathbf{X} \sim p, \mathbf{Y} \sim q$.

Coupling of distributions

Definition 9

Given $p, q \in \mathcal{P}(\mathcal{X})$, a coupling of p, q is a joint distributions on $\mathcal{X} \times \mathcal{X}$ whose first and second marginals are, respectively, p and q. We denote the space of such couplings as $\Gamma(p, q)$. We also write $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p, q)$ for random vectors $(\boldsymbol{X}, \boldsymbol{Y})$ s.t. $\mathbf{X} \sim p, \mathbf{Y} \sim q$.

Consider $X \sim Bern(p)$, $Y \sim Bern(q)$, then infinitely many couplings are possible. If the table below shows the joint frequencies, then:

X \ Y	0	1	
0	a	b	1-p
1	c	d	p
	1-q	q	

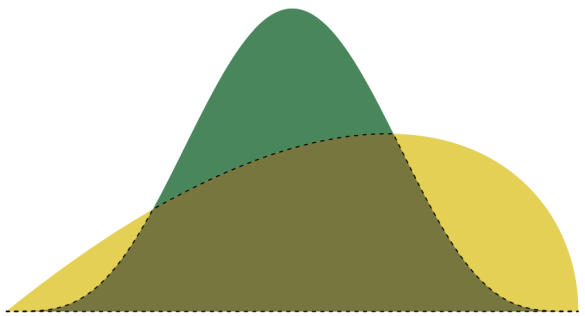
$$\begin{cases} a + b = 1 - p \\ a + c = 1 - q \\ b + d = q \end{cases}$$

system of 3 equations with 4 unknowns (the fourth equation comes from the others)

Maximal independent coupling

Algorithm 4: Rejection Maximal Coupling

```
Sample X ~ p;
Sample W ~ U(0,1);
if W p(X) ≤ q(X) then
    set Y = X
else
    Sample Y* ~ q;
    W* ~ U(0,1);
    while W* q(Y*) > p(Y*) do
        Sample Y* ~ q;
        W* ~ U(0,1);
    set Y = Y*
```



Computational cost:

$$\mathbb{E}[cost] = (1 - d_{tv}(p,q)) \times 1 + d_{tv}(p,q)(1 + 1/d_{tv}(p,q)) = 2$$

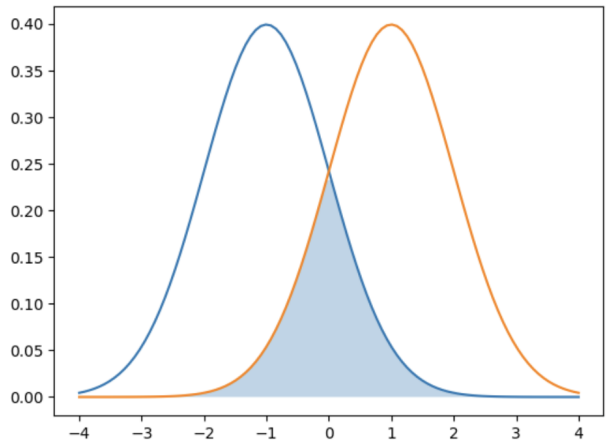
but $var \to +\infty$ as $d_{tv}(p,q) \to 0$ (since variance of $Geom(p) = (1-p)/p^2$).

Sampling from reflection coupling

Available only for Gaussian rvs with same variance covariance matrices:

Algorithm 5: Reflection Maximal Coupling

set $\mathbf{z} := \Sigma^{-1/2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$, $\mathbf{e} = \mathbf{z}/||\mathbf{z}||$;
sample $\dot{\mathbf{X}} \sim N_d(\mathbf{0}, \mathbf{1}_d)$, $W \sim U(0,1)$;
if $s(\dot{\mathbf{X}})W \leq s(\dot{\mathbf{X}} + \mathbf{z})$ **then**
  set $\dot{\mathbf{Y}} := \dot{\mathbf{X}} + \mathbf{z}$
else
  $\dot{\mathbf{Y}} := \dot{\mathbf{X}} - \mathbf{2}(\mathbf{e}'\dot{\mathbf{X}})\mathbf{e}$
$\mathbf{X} = \Sigma^{1/2}\dot{\mathbf{X}} + \mu_{\mathbf{1}}$;
$\mathbf{Y} = \Sigma^{1/2}\dot{\mathbf{Y}} + \mu_{\mathbf{2}}$;



In the univariate case it can be written as: sample $X \sim N(\mu_1, \sigma^2)$, sample W, if accept, set $Y = X$, else set $Y = \mu_2 - (X - \mu_1)$.
Computational cost: deterministically 2.