Dimension-free mixing times of coordinate-wise MCMC schemes for Bayesian hierarchical models

Filippo Ascolani and Giacomo Zanella

Bocconi University

November 24, 2023

Bayesian computation

- **Computational scalability** is crucial for Bayesian statistics applications.
- **Exact** posterior sampling is usually **not available**.
- Approximations needed.

Bayesian computation

- **Computational scalability** is crucial for Bayesian statistics applications.
- **Exact** posterior sampling is usually **not available**.
- Approximations needed.
- Task: approximate complex and high-dimensional posterior distributions

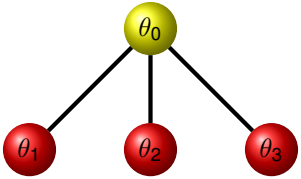
$$p(\theta \mid Y)$$

for **observations** Y and **parameters**/latent variables θ

High dimensional hierarchical models

General formulation:

$$\boxed{\begin{aligned} Y_j \mid \theta_j &\overset{\text{ind.}}{\sim} f(Y_j \mid \theta_j) \\ \theta_j \mid \psi &\sim p(\theta_j \mid \psi), \\ \psi &\sim p_0(\psi), \end{aligned}}$$

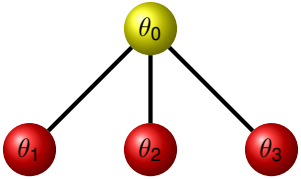


- Example: prediction of **real estate prices** at high spatial resolution.
- Problems with high number of datapoints n and parameters p.

High dimensional hierarchical models

General formulation:

$$\boxed{\begin{aligned} Y_j \mid \theta_j &\overset{\text{ind.}}{\sim} f(Y_j \mid \theta_j) \\ \theta_j \mid \psi &\sim p(\theta_j \mid \psi), \\ \psi &\sim p_0(\psi), \end{aligned}}$$



- Example: prediction of **real estate prices** at high spatial resolution.
- Problems with high number of datapoints n and parameters p.

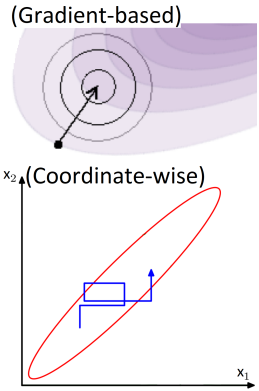
1. Which computational schemes are **scalable** for both n and p?
2. What about **non Gaussian** likelihood?
3. What about a **non Gaussian** prior?

Complexity of Bayesian computation

Main Markov chain Monte Carlo (MCMC) approaches:

1. **Gradient-based**: MALA, HMC,. . .
2. **Coordinate-wise**: Gibbs (single-site, blocked, collapsed,. . .), Metropolis-within-Gibbs, . . .

- How do these **scale** with n and p for important classes of statistical models?
- Which algorithm is good for which class of models?



Motivation

Hierarchical models:

- Data are divided in **groups**.
- Each group has its own "local" parameters.

High dimensional regime: large n and large p.

Motivation

Hierarchical models:

- Data are divided in **groups**.
- Each group has its own "local" parameters.

High dimensional regime: large n and large p.

Coordinate-wise schemes:

- Can naturally **exploit model structure** (conditional conjugacy, conditional independence, . . .) resulting in appealing features (cheap block-updating, no tuning, . . .)
- **Default choice** in many applications; empirical evidences of competitive/state-of-the-art performances

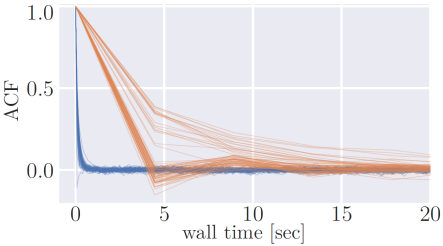


Figure: HMC (orange) vs collapsed-Gibbs (blue) for a high-d hierarchical model (Papaspiliopoulos et al., 2021).

Goal

A lot of interest in **scalability** of MCMC methods:

- Many recent advances in gradient-based techniques[1]. Increasingly sharper understanding. State-of-the-art Black-Box schemes.
- Relatively less results for Gibbs-type schemes. Arguments and proofs are quite **case–specific**.[2]

[1] Dalalyan (2017), Durmus and Moulines (2017), Chen et al. (2020), . . .
[2] see e.g. Diaconis, Khare, Saloff-Coste (2008) or Qin and Hobert (2021)
[3] Belloni and Chernozhukov (2009)

Goal

A lot of interest in **scalability** of MCMC methods:

- Many recent advances in gradient-based techniques[1]. Increasingly sharper understanding. State-of-the-art Black-Box schemes.
- Relatively less results for Gibbs-type schemes. Arguments and proofs are quite **case–specific**.[2]

Goals: contribute to complexity theory by

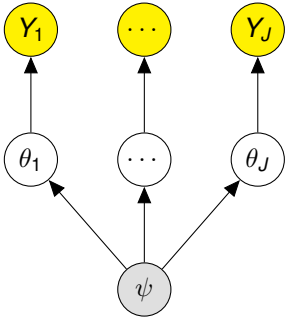
- Use the tools of **Bayesian asymptotics** to provide average-case complexity results[3] for coordinate-wise MCMC on hierarchical/structured models.
- Seek general proof technique (e.g. w.r.t. likelihood)

[1] Dalalyan (2017), Durmus and Moulines (2017), Chen et al. (2020), . . .
[2] see e.g. Diaconis, Khare, Saloff-Coste (2008) or Qin and Hobert (2021)
[3] Belloni and Chernozhukov (2009)

Motivating example: Hierarchical models

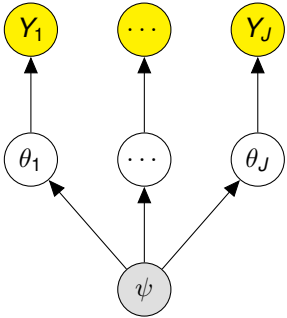
$$
\begin{aligned}
Y_j \mid \theta_j &\sim f(Y_j \mid \theta_j) \quad && j = 1, \dots, J \\
\theta_j \mid \psi &\overset{\text{iid}}{\sim} p(\theta_j \mid \psi) \quad && j = 1, \dots, J \\
\psi &\sim p_0(\psi).
\end{aligned}
$$

$Y_j = (Y_{j1}, \dots, Y_{jm}) \in \mathbb{R}^m$
$\theta_j \in \mathbb{R}^\ell$
$\psi \in \mathbb{R}^d$



Motivating example: Hierarchical models

$$
\begin{aligned}
Y_j \mid \theta_j &\sim f(Y_j \mid \theta_j) \quad && j = 1, \dots, J \\
\theta_j \mid \psi &\overset{\text{iid}}{\sim} p(\theta_j \mid \psi) \quad && j = 1, \dots, J \\
\psi &\sim p_0(\psi).
\end{aligned}
$$



$Y_j = (Y_{j1}, \dots, Y_{jm}) \in \mathbb{R}^m$
$\theta_j \in \mathbb{R}^\ell$
$\psi \in \mathbb{R}^d$

Gibbs sampler (GS):

$$
\begin{cases}
\psi \sim p\left(\psi \mid Y_{1:J}, \theta\right) \\
\theta_j \sim p\left(\theta_j \mid Y_j, \psi\right) \text{ for } j = 1, \dots, J
\end{cases}
$$

$n = mJ$ **datapoints**
$p = d + \ell J$ **parameters**

How does GS perform as J grows?

Numerical illustration. Toy logistic hierarchical model:

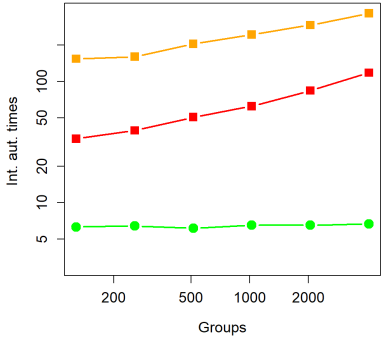
$$Y_{ji} \mid \theta_j \sim \text{Ber}\left(\text{logit}^{-1}(\theta_j)\right), \quad \theta_j \mid \mu, \tau \sim N\left(\mu, \tau^{-1}\right), \quad (\mu, \tau) \sim p(\mu, \tau)$$

Simulation with $\mu^* = \tau^* = 1$ and 10 observations per group.

Numerical illustration. Toy logistic hierarchical model:

$$Y_{ji} \mid \theta_j \sim \text{Ber}\left(\text{logit}^{-1}(\theta_j)\right), \quad \theta_j \mid \mu, \tau \sim N\left(\mu, \tau^{-1}\right), \quad (\mu, \tau) \sim p(\mu, \tau)$$

Simulation with $\mu^* = \tau^* = 1$ and 10 observations per group.



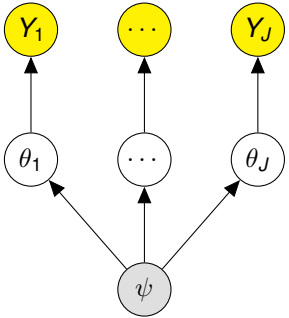
iterations for each posterior sample for:

- Gibbs sampler (GS) alternating updates of θ and (μ, τ)
- Langevin Monte Carlo (MALA) with optimal preconditioning.
- Hamiltonian Monte Carlo (HMC) with default Stan implementation.

$\rightsquigarrow$ numerics suggest GS enjoys **dimension-free convergence**

Motivating example: Hierarchical models

$$\boxed{\begin{aligned} Y_j \mid \theta_j &\sim f(Y_j \mid \theta_j) \quad j = 1, \dots, J \\ \theta_j \mid \psi &\overset{\text{iid}}{\sim} p(\theta_j \mid \psi) \quad j = 1, \dots, J \\ \psi &\sim p_0(\psi). \end{aligned}}$$



$Y_j = (Y_{j1}, \dots, Y_{jm}) \in \mathbb{R}^m$
$\theta_j \in \mathbb{R}^\ell$, $\psi \in \mathbb{R}^d$

Metropolis-within-Gibbs (MwG):

$$\begin{cases} \psi \sim p\left(\psi \mid Y_{1:J}, \theta\right) \\ \theta_j^{(t)} \sim P\left(\theta_j \mid Y_j, \psi, \theta_j^{(t-1)}\right) \end{cases}$$

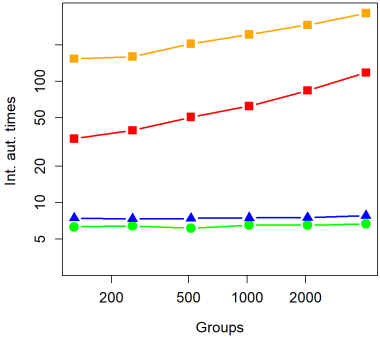
$n = mJ$ **datapoints**
$p = d + \ell J$ **parameters**

How does MwG perform as J grows?

Numerical illustration. Toy logistic hierarchical model:

$$Y_{ji} \mid \theta_j \sim \text{Ber}\left(\text{logit}^{-1}(\theta_j)\right), \quad \theta_j \mid \mu, \tau \sim N\left(\mu, \tau^{-1}\right), \quad (\mu, \tau) \sim p(\mu, \tau)$$

Simulation with $\mu^* = \tau^* = 1$ and 10 observations per group.



iterations for each posterior sample for:

- Metropolis-within-Gibbs sampler (MwG)
- Gibbs sampler (GS) alternating updates of θ and (μ, τ)
- Langevin Monte Carlo (MALA) with optimal preconditioning.
- Hamiltonian Monte Carlo (HMC) with default Stan implementation.

$\rightsquigarrow$ numerics suggest GS and MwG enjoy **dimension-free convergence**

Gibbs Sampler and asymptotics

Sequence of **posterior distributions**

$$\boxed{\pi_n(x) := p(x \mid Y_{1:n})}, \qquad n \geq 1$$

Gibbs kernels

$$x = (x_1, \ldots, x_K)$$

parameters partitioned in K blocks

$$\boxed{G_n = \frac{1}{K} \sum_{i=1}^{K} G_{n,i}},$$

$G_{n,i} :=$ "sample $x_i \sim \pi_n(x_i | x_{-i})$"

Metropolis-within-Gibbs kernels

$$\boxed{P_n = \frac{1}{K} \sum_{i=1}^{K} P_{n,i}},$$

$P_{n,i}$ invariant with respect to $\pi_n(x_i | x_{-i})$

Mixing times from warm starts

Worst case **mixing time** from M**-warm starts**:

$$t_{mix}^{(n)}(P_n, \varepsilon, M) = \sup_{\mu_n \in \mathcal{M}(\pi_n, M)} \min \left\{ t \geq 1 \; : \; \left|\left| \mu_n P_n^t - \pi_n \right|\right|_{TV} < \varepsilon \right\}$$

Interpretation: number of iterations to sample from π_n up to error ε.

Mixing times from warm starts

Worst case **mixing time** from M**-warm starts**:

$$t_{mix}^{(n)}(P_n, \varepsilon, M) = \sup_{\mu_n \in \mathcal{M}(\pi_n, M)} \min \left\{ t \geq 1 \ : \ \left|\left| \mu_n P_n^t - \pi_n \right|\right|_{TV} < \varepsilon \right\}$$

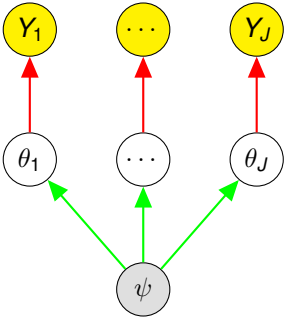
Interpretation: number of iterations to sample from π_n up to error ε.

Warm starts:

$$\mathcal{M}(\pi, M) := \left\{ \mu \ : \ \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) \leq M \text{ for all } x \right\} \qquad M \in [1, \infty)$$

$\Rightarrow$ starting point.

Exponential-family priors, generic likelihood



Assumptions:
Likelihood $f(Y_j|\theta_j)$ **generic**

Prior in the **exponential family**:

$$\boxed{p(\theta_j \mid \psi) = h(\theta_j)\exp\left\{\eta^T(\psi)T(\theta_j) - A(\psi)\right\}}$$

Implications:

$$\mathcal{L}\left(\psi \mid \theta, Y_{1:J}\right) = \mathcal{L}\left(\psi \mid T(\theta), Y_{1:J}\right)$$

but

$$\mathcal{L}(Y_{1:J} \mid \theta, \psi) \neq \mathcal{L}(Y_{1:J} \mid T(\theta), \psi)\,.$$

Dimensionality reduction

Lemma

Let

$$(\psi^{(t)}, \theta^{(t)})_{t\geq 1} \sim GS(\mathcal{L}(\psi, \theta | Y_{1:J})).$$

Then $(\psi^{(t)}, T(\theta^{(t)}))_{t\geq 1}$ *is also a Markov chain*[4]*, it has transition kernel* $GS(\mathcal{L}(\psi, T | Y_{1:J}))$ *and its mixing times* $\hat{t}_{mix}^{(J)}$ *satisfy*

$$\sup_{\hat{\mu}_J \in \mathcal{M}(\mathcal{L}(\psi, T | Y_{1:J}), M)} \hat{t}_{mix}^{(J)}(\varepsilon, \hat{\mu}_J) = \sup_{\mu_J \in \mathcal{M}(\mathcal{L}(\psi, \theta | Y_{1:J}), M)} t_{mix}^{(J)}(\varepsilon, \mu_J)\,.$$

- This is a Gibbs sampler of **fixed dimensionality**.
- The law $\mathcal{L}(\psi, T | Y_{1:J})$ and corresponding *GS* **not available** in closed form.

[4] a de-initializing one, see e.g. Roberts & Rosenthal (2001) Markov chains and de-initializing processes

Asymptotic convergence for reduced model

Assume data generated from the **exact likelihood** with fixed ψ^*

$$Y_j \mid \theta_j \sim f(Y_j \mid \theta_j), \quad \theta_j \sim p(\theta_j \mid \psi^*) \qquad\qquad j = 1, 2, \ldots .$$

Consider the Gibbs sampler on (ψ, T).

Asymptotic convergence for reduced model

Assume data generated from the **exact likelihood** with fixed ψ^*

$$Y_j \mid \theta_j \sim f(Y_j \mid \theta_j), \quad \theta_j \sim p(\theta_j \mid \psi^*) \qquad\qquad j = 1, 2, \dots .$$

Consider the Gibbs sampler on (ψ, T).

Approach:

1. A rescaling of (ψ, T) converges to a Gaussian distribution!
   - **Bernstein - von Mises** Theorem for ψ.
   - **Conditional CLT** in Total Variation for T.
2. Gibbs samplers on Gaussian distributions behave well.

Asymptotic convergence for reduced model

Assume data generated from the **exact likelihood** with fixed ψ^*

$$Y_j \mid \theta_j \sim f(Y_j \mid \theta_j), \quad \theta_j \sim p(\theta_j \mid \psi^*) \qquad\qquad j = 1, 2, \dots .$$

Consider the Gibbs sampler on (ψ, T).

Approach:

1. A rescaling of (ψ, T) converges to a Gaussian distribution!
   - **Bernstein - von Mises** Theorem for ψ.
   - **Conditional CLT** in Total Variation for T.
2. Gibbs samplers on Gaussian distributions behave well.

General idea: asymptotic behaviour of π_n can be translated in the asymptotic behaviour of $t_{mix}^{(n)}(G_n, \varepsilon, M)$.

Dimension-free convergence of GS for hierarchical models

Theorem

Assume:

1) $Y_j \overset{i.i.d.}{\sim} g_{\psi^*}$

2) Regularity assumptions for BvM (testability, non-singular Fisher information, . . .)

Then for every $M \geq 1$ *and* $\varepsilon > 0$ *the GS mixing time satisfies*

$$t_{mix}^{(J)}(G_J, \varepsilon, M) = \mathcal{O}_P(1) \qquad \text{as } J \to \infty.$$

Dimension-free convergence of GS for hierarchical models

Theorem

Assume:

1) $Y_j \overset{i.i.d.}{\sim} g_{\psi^*}$

2) Regularity assumptions for BvM (testability, non-singular Fisher information, . . .)

Then for every $M \geq 1$ *and* $\varepsilon > 0$ *the GS mixing time satisfies*

$$t_{mix}^{(J)}(G_J, \varepsilon, M) = \mathcal{O}_P(1) \qquad \text{as } J \to \infty.$$

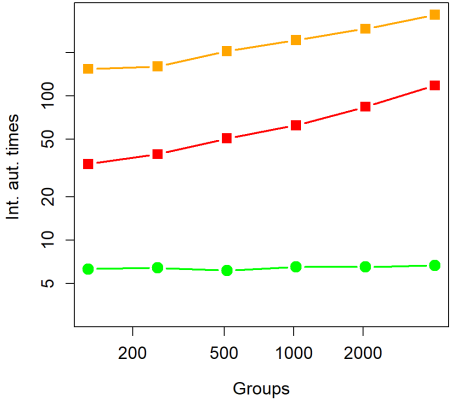
Mixing times are **bounded** with respect to the number of groups.

Ascolani, F. and *Zanella, G.* (2023+) Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models. *Submitted.*

Numerical illustration. Binomial hierarchical model:

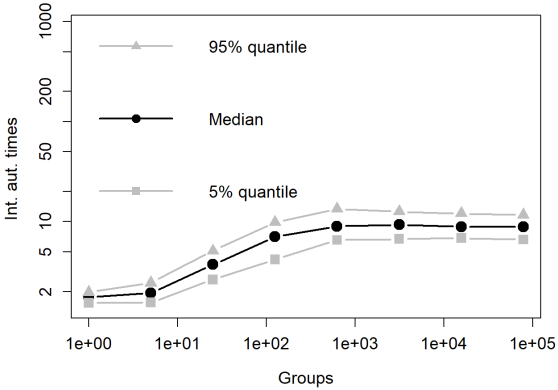
$$Y_{ji} \mid \theta_j \sim \text{Ber}\left(\text{logit}^{-1}(\theta_j)\right), \quad \theta_j \mid \mu, \tau \sim N\left(\mu, \tau^{-1}\right), \quad (\mu, \tau) \sim p(\mu, \tau)$$

Simulation with $\mu^* = \tau^* = 1$ and 10 observations per group.



- Gibbs sampler.
- MALA with optimal tuning and preconditioning.
- HMC with default Stan implementation.

Numerical illustration (using *integrated autocorrelation times* as proxy of mixing times)
Variability refers to randomness over different datasets.



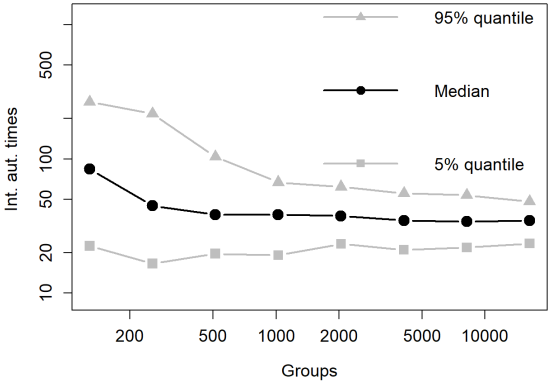


Figure: Left: hierarchical linear model; Right: hierarchical binomial model

What about MwG? Issues

Various tools to study **Gibbs samplers**:

- Drift minorization (Rosenthal, 1995), orthogonal decomposition (Diaconis et al., 2008), de-initializing sequences (Roberts and Rosenthal, 2001), . . .
- Sharp results for specific distributions, e.g. Gaussian (Amit, 1996 and Roberts and Sahu, 1997).
- Asymptotic analysis: Yang and Rosenthal (2021), Jin and Hobert (2022), Ascolani and Zanella (2023+), . . .

What about MwG? Issues

Various tools to study **Gibbs samplers**:

- Drift minorization (Rosenthal, 1995), orthogonal decomposition (Diaconis et al., 2008), de-initializing sequences (Roberts and Rosenthal, 2001), . . .
- Sharp results for specific distributions, e.g. Gaussian (Amit, 1996 and Roberts and Sahu, 1997).
- Asymptotic analysis: Yang and Rosenthal (2021), Jin and Hobert (2022), Ascolani and Zanella (2023+), . . .

Difficult to translate them for generic **coordinate-wise** schemes! Challenges:

1. Passing from Gibbs to generic coordinate-wise $\Leftrightarrow$ from **exact** to **invariant** updates.
2. Exploit the **asymptotic characterization** of the posterior.

From MwG to Gibbs

Common tool to study MCMC convergence: **conductance**

$$\Phi_0(P) = \inf \left\{ \frac{\int_A P(x, A^c)\, \pi(\mathrm{d}x)}{\pi(A)\pi(A^c)} \; : \; \pi(A) > 0 \right\}.$$

From MwG to Gibbs

Common tool to study MCMC convergence: **conductance**

$$\Phi_0(P) = \inf \left\{ \frac{\int_A P(x, A^c)\, \pi(\mathrm{d}x)}{\pi(A)\pi(A^c)} \; : \; \pi(A) > 0 \right\}.$$

$\Phi_0(P) > 0 \Rightarrow$ **exponentially fast convergence**:

$$\boxed{\left\| \mu P^t - \pi \right\|_{TV} \leq M \left(1 - \frac{1}{2} \Phi_0^2(P) \right)^t,}$$

with M-warm start μ.

From MwG to Gibbs

How to measure the **goodness** of conditional updates?

$$\boxed{\kappa\left(P_i^{x_{-i}}\right) = \Phi_0\left(P_i(\cdot \mid x_{-i})\right)} \qquad \textbf{Conditional conductance}$$

It is the **conductance of the i-th update**, conditional on $\mathbf{x}_{-i}$.

From MwG to Gibbs

How to measure the **goodness** of conditional updates?

$$\boxed{\kappa\left(P_i^{x_{-i}}\right) = \Phi_0\left(P_i(\cdot \mid x_{-i})\right)} \qquad \textbf{Conditional conductance}$$

It is the **conductance of the *i*-th update**, conditional on $\mathbf{x}_{-i}$.

(A - Conditional updates): assume $\boxed{\kappa := \min_i \inf_x \kappa\left(P_i^{x_{-i}}\right) > 0}$ ($\Longleftarrow$ *to be relaxed*)

From MwG to Gibbs

How to measure the **goodness** of conditional updates?

$$\boxed{\kappa\left(P_i^{x_{-i}}\right) = \Phi_0\left(P_i(\cdot \mid x_{-i})\right)} \qquad \textbf{Conditional conductance}$$

It is the **conductance of the *i*-th update**, conditional on $\mathbf{x}_{-i}$.

(A - Conditional updates): assume $\boxed{\kappa := \min_i \inf_x \kappa\left(P_i^{x_{-i}}\right) > 0}$ ($\Longleftarrow$ *to be relaxed*)

Theorem

Under (A) we have $\boxed{\Phi_0(P) \geq \kappa\, \Phi_0(G)}$ $\rightarrow$ *connection between GS and MwG!*

Sketch of the proof

Let $\boxed{P(\partial A) = \int_A P(x, A^c)\pi(\mathrm{d}x)}$.

Sketch of the proof

Let $\boxed{P(\partial A) = \int_A P(x, A^c)\pi(\mathrm{d}x)}$.Then:

$$P_i(\partial A) \geq \int_A \kappa\left(P_i^{x_{-i}}\right) G_i(x, A^c)\pi(\mathrm{d}x)$$

Sketch of the proof

Let $\boxed{P(\partial A) = \int_A P(x, A^c)\pi(\mathrm{d}x)}$.Then:

$$P_i(\partial A) \geq \int_A \kappa\left(P_i^{x_{-i}}\right) G_i(x, A^c)\pi(\mathrm{d}x) \geq \kappa \int_A G_i(x, A^c)\pi(\mathrm{d}x)$$

Sketch of the proof

Let $\boxed{P(\partial A) = \int_A P(x, A^c)\pi(\mathrm{d}x)}$.Then:

$$P_i(\partial A) \geq \int_A \kappa \left(P_i^{x_{-i}}\right) G_i(x, A^c)\pi(\mathrm{d}x) \geq \kappa \int_A G_i(x, A^c)\pi(\mathrm{d}x) \geq \kappa G_i(\partial A)$$

Sketch of the proof

Let $\boxed{P(\partial A) = \int_A P(x, A^c)\pi(\mathrm{d}x)}$.Then:

$$P_i(\partial A) \geq \int_A \kappa\left(P_i^{x_{-i}}\right) G_i(x, A^c)\pi(\mathrm{d}x) \geq \kappa \int_A G_i(x, A^c)\pi(\mathrm{d}x) \geq \kappa G_i(\partial A)$$

Remark: If $\boxed{\kappa(C) := \min_i \inf_{x \in C} \kappa\left(P_i^{x_{-i}}\right)}$, then

$$P_i(\partial A) \geq \kappa(C) G_i(\partial A) - \kappa(C)\pi(A \cup C^c)$$

It suffices to choose C large enough!

Asymptotic characterization

(B - Posterior convergence): assume $\tilde{\pi}$ such that $\tilde{\pi}_n := \mathcal{L}(\varphi_n(x)|Y_{1:n})$ satisfies

$$\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0 \qquad \text{as } n \to \infty\,,$$

in probability under $Y_{1:n} \sim Q_n$, with $(\varphi_n)_{n\geq 1}$ *coordinate-wise* and *injective* transformations.

We require **concentration** of the posterior distribution, e.g. **Bernstein - von Mises Theorem**.

Asymptotic characterization

(B - Posterior convergence): assume $\tilde{\pi}$ such that $\tilde{\pi}_n := \mathcal{L}(\varphi_n(x)|Y_{1:n})$ satisfies

$$\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0 \qquad \text{as } n \to \infty\,,$$

in probability under $Y_{1:n} \sim Q_n$, with $(\varphi_n)_{n\geq 1}$ *coordinate-wise* and *injective* transformations.

We require **concentration** of the posterior distribution, e.g. **Bernstein - von Mises Theorem**.

Problem: asymptotic characterization of the posterior does **not** control the conductance.

Let $\tilde{G}$ be the Gibbs sampler on $\tilde{\pi}$. Under (B), we have

$$\boxed{\Phi_0(\tilde{G}) > 0} \quad \not\Longrightarrow \quad \boxed{\liminf_n \Phi_0(G_n) > 0}$$

Nice **conductance of the limit** $\not\Longrightarrow$ nice **limit of conductances.**

Asymptotic characterization

Approximate conductance

$$\Phi_s(P) = \inf \left\{ \frac{\int_A P(x, A^c)\, \pi(\mathrm{d}x)}{\pi(A) - s},\ s < \pi(A) \leq \frac{1}{2} \right\}, \quad s > 0.$$

$\Phi_s(P) > 0$ for every $s > 0 \Rightarrow$ still **control of the TV distance**.

Asymptotic characterization

Approximate conductance

$$\Phi_s(P) = \inf \left\{ \frac{\int_A P(x, A^c)\, \pi(\mathrm{d}x)}{\pi(A) - s},\ s < \pi(A) \leq \frac{1}{2} \right\}, \quad s > 0.$$

$\Phi_s(P) > 0$ for every $s > 0 \Rightarrow$ still **control of the TV distance**.

(B): $\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0$ in probability under $Y_{1:n} \sim Q_n$

Theorem

Assume (B) with GS $\tilde{G}$ on $\tilde{\pi}$ such that $\Phi_0(\tilde{G}) > 0$. Then for every $s > 0$ we have

$$\boxed{\liminf_n \Phi_s(G_n) > 0}$$

Nice **conductance of the limit** $\implies$ nice **limit of approximate conductances.**

Convergence of Metropolis-within-Gibbs

(A): $\liminf_n \inf_{x \in C_n} \kappa\left(P_{n,i}^{x_{-i}}\right) > 0$, for $i = 1, \ldots, K$ and large enough C_n.
(B): $\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0$ in probability under $Y_{1:n} \sim Q_n$

[5] $\lim_{c \to \infty} \sup_{n \geq 1} Pr(t_{mix}^{(n)}(\varepsilon, M) > c) = 0$

Convergence of Metropolis-within-Gibbs

(A): $\liminf_n \inf_{x \in C_n} \kappa\left(P_{n,i}^{x_{-i}}\right) > 0$, for $i = 1, \ldots, K$ and large enough C_n.
(B): $\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0$ in probability under $Y_{1:n} \sim Q_n$

Theorem

Assume (A)-(B) with GS $\tilde{G}$ on $\tilde{\pi}$ such that $\Phi_0(\tilde{G}) > 0$. Then for every $\varepsilon \in (0,1)$ and $M \geq 1$

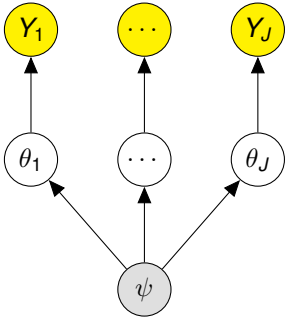
$$t_{mix}^{(n)}(P_n, \varepsilon, M) = \mathcal{O}_p(1) \qquad \text{as } n \to \infty\,,$$

i.e. the induced sequence of mixing times is bounded in probability (or uniformly tight[5]).

[5] $\lim_{c \to \infty} \sup_{n \geq 1} Pr(t_{mix}^{(n)}(\varepsilon, M) > c) = 0$

Motivating example: Hierarchical models

$$\boxed{\begin{aligned} Y_j \mid \theta_j &\sim f(Y_j \mid \theta_j) \quad j = 1, \dots, J \\ \theta_j \mid \psi &\overset{\text{iid}}{\sim} p(\theta_j \mid \psi) \quad j = 1, \dots, J \\ \psi &\sim p_0(\psi). \end{aligned}}$$



$Y_j = (Y_{j1}, \dots, Y_{jm}) \in \mathbb{R}^m$
$\theta_j \in \mathbb{R}^\ell$, $\psi \in \mathbb{R}^d$

Metropolis-within-Gibbs (MwG):

$$\begin{cases} \psi \sim p\left(\psi \mid Y_{1:J}, \theta\right) \\ \theta_j^{(t)} \sim P\left(\theta_j \mid Y_j, \psi, \theta_j^{(t-1)}\right) \end{cases}$$

$n = mJ$ **datapoints**
$p = d + \ell J$ **parameters**

How does MwG perform as J grows?

Recap: general strategy

Two steps:

1. Show that the **Gibbs sampler** is asymptotically well-behaved $\Longleftarrow$ first part of the talk!

Recap: general strategy

Two steps:

1. Show that the **Gibbs sampler** is asymptotically well-behaved $\Longleftarrow$ first part of the talk!
2. Check that **conditional updates** are "good enough" $\Longleftarrow$ conditional conductance.

Recap: general strategy

Two steps:

1. Show that the **Gibbs sampler** is asymptotically well-behaved $\Longleftarrow$ first part of the talk!
2. Check that **conditional updates** are "good enough" $\Longleftarrow$ conditional conductance.

We ask that there exists Ψ neighborhood of ψ^* such that

$$\boxed{\inf_{Y_j} \inf_{\psi \in \Psi} \Phi_0 \left(P(\theta_j^{(t)} \mid Y_j, \psi, \theta_j^{(t-1)}) \right) > 0.} \quad (\star)$$

Conditional updates need to be good "around" ψ^*.

Dimension-free convergence of MwG for hierarchical models

Theorem

Assume:

1) $Y_j \overset{i.i.d.}{\sim} g_{\psi^*}$

2) Regularity assumptions for BvM (testability, non-singular Fisher information, . . .)

3) Control of the conditional updates as in $(\star)$.

Then for every $M \geq 1$ *and* $\varepsilon > 0$ *the MwG mixing time satisfies*

$$t_{mix}^{(J)}(P_J, \varepsilon, M) = \mathcal{O}_P(1) \qquad \text{as } J \to \infty.$$

Dimension-free convergence of MwG for hierarchical models

Theorem

Assume:

1) $Y_j \overset{i.i.d.}{\sim} g_{\psi^*}$

2) Regularity assumptions for BvM (testability, non-singular Fisher information, . . .)

3) Control of the conditional updates as in $(\star)$.

Then for every $M \geq 1$ *and* $\varepsilon > 0$ *the MwG mixing time satisfies*

$$t_{mix}^{(J)}(P_J, \varepsilon, M) = \mathcal{O}_P(1) \qquad \text{as } J \to \infty.$$

Mixing times are **bounded** with respect to the number of groups.

Ascolani, F., *Roberts, G. O.* and *Zanella, G.* Asymptotic behaviour of Metropolis-within-Gibbs schemes through conditional conductance. *In preparation.*

Starting distribution

Feasible start. It can be obtained in a natural way:

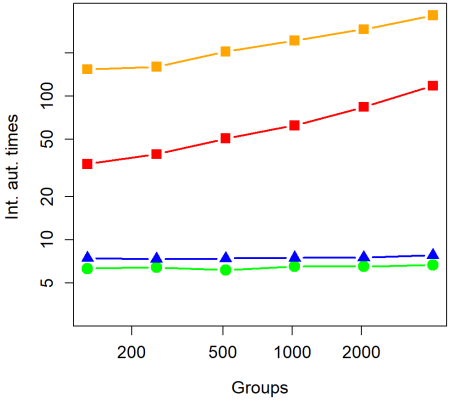
1. sample ψ from
$$Unif(\hat{\psi}_{MML} - cJ^{-1/2}, \hat{\psi}_{MML} + cJ^{-1/2})$$
with $c > 0$ where $\hat{\psi}_{MML}$ is the maximum marginal likelihood estimator.
2. run the conditional updates of $\{\theta_j\}_j$ for $\approx \log(J)$ iterations.

Overall computational cost: $\mathcal{O}_P(J \log(J))$.

Non-conjugate models

The technique allows to study various **non-conjugate** models.

1. Hierarchical models with Y_j supported on **finite** space (e.g. binary or categorical data) and *arbitrary* likelihood.



- Metropolis-within-Gibbs
- Gibbs sampler.
- MALA with optimal tuning and preconditioning.
- HMC with default Stan implementation.

Non-conjugate models

The technique allows to study various **non-conjugate** models.

1. Hierarchical models with Y_j supported on **finite** space (e.g. binary or categorical data) and *arbitrary* likelihood.
2. More general hierarchical models with **multiple blocks**
   $\implies$ multilevel, crossed, ...
3. **Gaussian processes** with non-gaussian likelihood, e.g. binary data.

Non-conjugate models

The technique allows to study various **non-conjugate** models.

1. Hierarchical models with Y_j supported on **finite** space (e.g. binary or categorical data) and *arbitrary* likelihood.
2. More general hierarchical models with **multiple blocks**
   $\implies$ multilevel, crossed, ...
3. **Gaussian processes** with non-gaussian likelihood, e.g. binary data.
4. **Beyond hierarchical models**: regression problems, conditionally log-concave distributions...

Conclusions

Summary:

- Develop general theory to derive **asymptotic complexity** of generic coordinate-wise MCMC schemes.
- Connection between **Bayesian computation** and **Bayesian asymptotics**.
- Application to **hierarchical models** with conjugate global-local priors and general likelihood.

Conclusions

Summary:

- Develop general theory to derive **asymptotic complexity** of generic coordinate-wise MCMC schemes.
- Connection between **Bayesian computation** and **Bayesian asymptotics**.
- Application to **hierarchical models** with conjugate global-local priors and general likelihood.

What's next?

- Connection between GS and MwG can exploited much beyond hierarchical models.
- What happens under **misspecification**?

Conclusions

Summary:

- Develop general theory to derive **asymptotic complexity** of generic coordinate-wise MCMC schemes.
- Connection between **Bayesian computation** and **Bayesian asymptotics**.
- Application to **hierarchical models** with conjugate global-local priors and general likelihood.

What's next?

- Connection between GS and MwG can exploited much beyond hierarchical models.
- What happens under **misspecification**?

Thanks for listening!