A SYNCHRONOUS GIBBS SAMPLER DIVERGENCE

We illustrate how the Synchronous sampler can converge to the wrong distribution using two simple cases. The first case uses two Gaussian variables while the second case uses two discrete binary variables. In both cases we observe that the resulting distribution preserves the marginals.

A.1 Synchronous Gaussian Sampling

Suppose we draw $\left(X_1^{(t-1)}, X_2^{(t-1)}\right)$ from the multivariate normal distribution:

$$\begin{pmatrix} X_1^{(t-1)} \\ X_2^{(t-1)} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \right]$$

and then use the Synchronous Gibbs sampler to generate $\left(X_1^{(t)}, X_2^{(t)}\right)$ given $\left(X_1^{(t-1)}, X_2^{(t-1)}\right)$. One can easily show that:

$$\begin{pmatrix} X_1^{(t)} \\ X_2^{(t)} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \frac{\rho^3}{\sigma_1^2 \sigma_2^2} \\ \frac{\rho^3}{\sigma_1^2 \sigma_2^2} & \sigma_2^2 \end{pmatrix} \right]$$

where the covariance term is not preserved. Note, however that the marginals remain consistent.

A.2 Discrete Case

In the discrete case, consider the two binary variable discrete model $P(X_1, X_2)$ where

$$P(X_1 = x_1, X_2 = x_2) \propto \begin{cases} \epsilon & x_1 \neq x_2 \\ 1 - \epsilon & x_1 = x_2 \end{cases}$$

for some $0 < \epsilon < 1$.

The transition matrix of the Synchronous sampler which samples from both variables simultaneously is therefore:

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	$(1-\epsilon)^2$	$(1-\epsilon)\epsilon$	$(1-\epsilon)\epsilon$	ϵ^2
(0,1)	$(1-\epsilon)\epsilon$	ϵ^2	$(1-\epsilon)^2$	$(1-\epsilon)\epsilon$
(1,0)	$(1-\epsilon)\epsilon$	$(1-\epsilon)^2$	ϵ^2	$(1-\epsilon)\epsilon$
(1,1)	ϵ^2	$(1-\epsilon)\epsilon$	$(1-\epsilon)\epsilon$	$(1-\epsilon)^2$

which has the incorrect stationary distribution:

$$P(X_1 = x_1, X_2 = x_2) \propto 1$$

Observe however, that the marginals $P(X_1)$ and $P(X_2)$ remain consistent.

B Proof of Theorem 4.1

The proof of Theorem 4.1 follows closely the original proof of the Gibbs sampler provided by Geman and Geman [1984] but with the modification of the Splash blocking. We prove Theorem 4.1 in three parts. First we show that π is the invariant distribution:

$$\pi\left(X^{(t+1)}\right) = \sum_x K\left(X^{(t+1)} \,|\, X^{(t)} = x\right) \pi\left(X^{(t)} = x\right) \tag{B.1}$$

of the Splash sampler. Second, we show that the Splash sampler forgets its starting state:

$$\sup_{x,y,z} \left|\left| \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1 \leq \gamma^{t/n} \tag{B.2}$$

Finally, we show that the sampler draws from π in the limit:

$$\lim_{t\to\infty} \sup_{y,x} \left|\left| \mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = x\right) - \pi(y) \right|\right|_1 = 0 \tag{B.3}$$

B.1 π Invariance

Because vanishing adaptation is used we show π invariance for the Splash sampler without adaptation: where the choice of Splash $\mathcal{S}^{(t)}$ does not depend on the state of the chain. However, we do allow the choice of the Splash to depend on all previous Splash choices $\left\{\mathcal{S}^{(0)}, \ldots, \mathcal{S}^{(t)}\right\}$.

The transition kernel for the Splash sampler given the Splash $\mathcal{S}$ is defined as:

$$\mathbf{P}\left(X^{(t+1)} \,|\, x^{(t)}\right) = \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, \mathcal{S}^{(0)} \ldots \mathcal{S}^{(t-1)}\right) \mathbf{1}\left[X^{(t+1)}_{-\mathcal{S}} = X^{(t)}_{-\mathcal{S}}\right] \pi\left(X^{(t+1)}_{\mathcal{S}} \,|\, X^{(t+1)}_{-\mathcal{S}}\right) \tag{B.4}$$

Substituting Eq. (B.4) into Eq. (B.1) we obtain:

$$\begin{aligned}
\mathbf{P}\left(X^{(t+1)}\right) &= \sum_{X^{(t)}} \mathbf{P}\left(X^{(t+1)} \,|\, X^{(t)}\right) \pi\left(X^{(t)}\right) && \text{(B.5)}\\
\mathbf{P}\left(X^{(t+1)}\right) &= \sum_{X^{(t)}} \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, \mathcal{S}^{(0)} \ldots \mathcal{S}^{(t)}\right) \mathbf{1}\left[x^{(t+1)}_{-\mathcal{S}} = x^{(t)}_{-\mathcal{S}}\right] \pi\left(X^{(t+1)}_{\mathcal{S}} \,|\, X^{(t+1)}_{-\mathcal{S}}\right) \pi\left(X^{(t)}\right) && \text{(B.6)}\\
&= \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, \mathcal{S}^{(0)} \ldots \mathcal{S}^{(t)}\right) \pi\left(X^{(t+1)}_{\mathcal{S}} \,|\, X^{(t+1)}_{-\mathcal{S}}\right) \sum_{X^{(t)}} \mathbf{1}\left[x^{(t+1)}_{-\mathcal{S}} = x^{(t)}_{-\mathcal{S}}\right] \pi\left(X^{(t)}\right) && \text{(B.7)}\\
&= \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, \mathcal{S}^{(0)} \ldots \mathcal{S}^{(t)}\right) \pi\left(X^{(t+1)}_{\mathcal{S}} \,|\, X^{(t+1)}_{-\mathcal{S}}\right) \pi\left(X^{(t+1)}_{-\mathcal{S}}\right) && \text{(B.8)}\\
&= \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, \mathcal{S}^{(0)} \ldots \mathcal{S}^{(t)}\right) \pi\left(X^{(t+1)}\right) && \text{(B.9)}\\
&= \pi\left(X^{(t+1)}\right) && \text{(B.10)}
\end{aligned}$$

B.2 Dependence on Starting State

We now show that the Splash sampler forgets its initial starting state. This is done by first showing that there is a positive probability of reaching any state after a bounded number of Splash operations. Then we use this strong notion of irreducibility to setup a recurrence which bounds the dependence on the starting state.

The Splash algorithm sweeps across the roots ensuring that after $\Delta t \leq n$ tree updates all variables are sampled at least once. Define the time t_i as the last time variable X_i was updated. Without loss of generality let's assume that $t_1 < t_2 < \ldots < t_n$ (we can rearrange the variable ordering to achieve this). We then can bound the probability that after n Splashes we reach state x given we were initially at state y,

$$\mathbf{P}\left(X^{(n)} = x \,|\, X^{(0)} = y\right) \geq \prod_{i=1}^{n} \inf_{x_{-i}} \pi\left(X_i = x_i \,|\, X_{-i} = x_{-i}\right). \tag{B.11}$$

Define the smallest conditional probability,

$$\delta = \inf_{i,x} \pi\left(X_i = x_i \,|\, X_{-i} = x_{-i}\right). \tag{B.12}$$

Then

$$\mathbf{P}\left(X^{(n)} = x \,|\, X^{(0)} = y\right) \geq \delta^n. \tag{B.13}$$

Effectively, we are stating that π is irreducible since all the conditionals are positive. Using this we will now show that there exists an γ such that $0 \leq \gamma < 1$ and

$$\sup_{x,y,z} \left|\left| \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1 \leq \gamma^{t/n}. \tag{B.14}$$

This is trivially true for $t = 0$. We can rewrite the left side of the above equation as

$$\sup_{x,y,z} \left|\left| \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1 = \tag{B.15}$$

$$\sup_{x} \left|\left| \sup_{y} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \inf_{z} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1 . \tag{B.16}$$

Now we will bound the inner sup and inf terms. For $t > n$ we can introduce a probability measure μ such that

$$\sup_{y} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) = \sup_{y} \sum_{w} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right) \mathbf{P}\left(X^{(n)} = w \,|\, X^{(0)} = y\right) \tag{B.17}$$

$$\leq \sup_{\mu \geq \delta^n} \sum_{w} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right) \mu(w) \tag{B.18}$$

Then by defining the w^* as

$$w^* = \arg\sup_{w} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right), \tag{B.19}$$

we can easily construct the maximizing μ which places minimal mass δ^n mass on w except w^* and places the remaining mass on $1 - (|\Omega| - 1)\delta^n$ on w^*. This leads to

$$\sup_{y} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) \leq (1 - (|\Omega| - 1)\delta^n) \, \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w^*\right) + \tag{B.20}$$

$$\delta^n \sum_{w \neq w^*} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right). \tag{B.21}$$

Similarly we define w_* as the minimizing element in Ω

$$w_* = \arg\inf_{w} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right). \tag{B.22}$$

We can then construct the lower bound,

$$\inf_{z} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \geq (1 - (|\Omega| - 1)\delta^n) \, \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w_*\right) + \tag{B.23}$$

$$\delta^n \sum_{w \neq w^*} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right). \tag{B.24}$$

Taking the difference, we get

$$\sup_{x,y,z} \left|\left| \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1$$

$$= \sup_{x} \left|\left| \sup_{y} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \inf_{z} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1 \tag{B.25}$$

$$\leq \sup_{x} || \, (1 - (|\Omega| - 1)\delta^n) \, \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w^*\right) + \delta^n \sum_{w \neq w^*} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right) -$$

$$(1 - (|\Omega| - 1)\delta^n) \, \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w_*\right) - \delta^n \sum_{w \neq w_*} \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = w\right) ||_1. \tag{B.26}$$

We can bound the sum terms

$$\delta^n \sum_{w \neq w^*} \mathbf{P}\left(X(t) = x \,|\, X(n) = w\right) - \delta^n \sum_{w \neq w_*} \mathbf{P}\left(X(t) = x \,|\, X(n) = w\right) \leq$$

$$\delta^n \sum_{w} \mathbf{P}\left(X(t) = x \,|\, X(n) = w\right) - \delta^n \sum_{w} \mathbf{P}\left(X(t) = x \,|\, X(n) = w\right) = 0,$$

and then simplify Eq. (B.26) to obtain

$$\sup_{x,y,z} \left|\left| \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right) \right|\right|_1 \leq$$

$$(1 - (|\Omega| - 1)\delta^n) \sup_{x,y,z} \left|\left| \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = z\right) \right|\right|_1 . \tag{B.27}$$

We can repeat this procedure t/n times to the term $\left|\left|\mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(n)} = z\right)\right|\right|_1$ on the right side of the above equation to obtain the desired result,

$$\sup_{x,y,z} \left|\left|\mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = y\right) - \mathbf{P}\left(X^{(t)} = x \,|\, X^{(0)} = z\right)\right|\right|_1 \leq (1 - (|\Omega| - 1)\delta^n)^{t/n}. \tag{B.28}$$

B.3 Convergence in distribution

We can use Eq. (B.28) along with π invariance to finish the proof,

$$\begin{aligned}
&\lim_{t\to\infty} \sup_{y,x} \left|\left|\mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = x\right) - \pi(y)\right|\right|_1 \\
&\quad = \lim_{t\to\infty} \sup_{y,x} \left|\left|\mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = x\right) - \sum_z \pi(y \,|\, z)\,\pi(z)\right|\right|_1 && \text{(B.29)} \\
&\quad = \lim_{t\to\infty} \sup_{y,x} \left|\left|\mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = x\right) - \sum_z \mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = z\right)\pi(z)\right|\right|_1 && \text{(B.30)} \\
&\quad = \lim_{t\to\infty} \sup_{y,x} \left|\left|\sum_z \pi(z)\left(\mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = x\right) - \mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = z\right)\right)\right|\right|_1 && \text{(B.31)} \\
&\quad \leq \lim_{t\to\infty} \sup_{y,x,z} \left|\left|\mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = x\right) - \mathbf{P}\left(X^{(t)} = y \,|\, X^{(0)} = z\right)\right|\right|_1 && \text{(B.32)} \\
&\quad \leq \lim_{t\to\infty} (1 - (|\Omega| - 1)\delta^n)^{t/n} && \text{(B.33)} \\
&\quad = 0. && \text{(B.34)}
\end{aligned}$$

C CONTINUOUS ADAPTATION IS NOT π-INVARIANT

Given the flexibility in scheduling Geman and Geman [1984] ascribed to the Gibbs sampler, one may consider the possibility that the Gibbs sampler could still be π invariant if we allowed the choice (shape) of Splashes to be adapted according to the state of the sampler. Such a conjecture would seem reasonable since any individual Splash is invariant. Indeed, Levine and Casella [2006] make a similar claim about Algorithm 7.1. Unfortunately, we can demonstrate, by means of a simple counter example, that tuning the Splash (blocking) in general breaks π invariance even when we include all variables with positive probability.

Consider the case of two uniform independent binary variables X_1 and X_2 with constant probability mass function

$$\pi(x_1, x_2) \propto 1$$

We define an adaptive sampling procedure which picks a variable to sample uniformly at random if both variables have the same assignment ($x_1 = x_2$). However, if the variables have different assignments ($x_1 \neq x_2$), we pick the variable with assignment 1 with 90% probability. Let the choice of variable to sample (Splash) be represented by $\mathbf{P}(\mathcal{S} \,|\, X_1, X_2)$ where $\mathcal{S} \subseteq \{1, 2\}$.

Let (X_1, X_2) be the current state. Let (X_1', X_2') be the new state after one step of the procedure. Then we can define the kernel transition $\mathbf{P}(X_1', X_2' | X_1, X_2, \mathcal{S})$ as

$$\begin{aligned}
\mathbf{P}(X_1', X_2' \,|\, \mathcal{S} = \{1\}, X_1, X_2) &\propto \mathbf{1}[x_2' = x_2] \\
\mathbf{P}(X_1', X_2' \,|\, \mathcal{S} = \{2\}, X_1, X_2) &\propto \mathbf{1}[x_1' = x_1]
\end{aligned}$$

We can derive the distribution of (X_1', X_2') after taking one adaptive Splash step,

$$\begin{aligned}
\mathbf{P}\left(X_1' = x_1', X_2' = x_2'\right) &= \sum_{x_1,x_2} \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, X_1, X_2\right) \mathbf{P}\left(X_1', X_2' \,|\, \mathcal{S}, X_1, X_2\right) \mathbf{P}\left(X_1, X_2\right) \\
&\propto \sum_{x_1,x_2} \sum_{\mathcal{S}} \mathbf{P}\left(\mathcal{S} \,|\, X_1, X_2\right) \mathbf{P}\left(X_1', X_2' \,|\, \mathcal{S}, X_1, X_2\right) \\
&\propto \sum_{x_1,x_2} \mathbf{P}\left(\mathcal{S} = \{1\} \,|\, X_1, X_2\right) \mathbf{1}\left[x_2' = x_2\right] + \mathbf{P}\left(\mathcal{S} = \{2\} \,|\, X_1, X_2\right) \mathbf{1}\left[x_1' = x_1\right] \\
&\propto \left(\sum_{x_1,x_2} \mathbf{P}\left(\mathcal{S} = \{1\} \,|\, X_1, X_2\right) \mathbf{1}\left[x_2' = x_2\right]\right) + \left(\sum_{x_1,x_2} \mathbf{P}\left(\mathcal{S} = \{2\} \,|\, X_1, X_2\right) \mathbf{1}\left[x_1' = x_1\right]\right) \\
&\propto \left(\sum_{x_1} \mathbf{P}\left(\mathcal{S} = \{1\} \,|\, X_1, X_2 = x_2'\right)\right) + \left(\sum_{x_2} \mathbf{P}\left(\mathcal{S} = \{2\} \,|\, X_1 = x_1', X_2\right)\right) \\
&= \mathbf{1}\left[x_2' = 1\right](0.5 + 0.1) + \mathbf{1}\left[x_2' = 0\right](0.9 + 0.1) \\
&\quad + \mathbf{1}\left[x_1' = 1\right](0.5 + 0.1) + \mathbf{1}\left[x_1' = 0\right](0.9 + 0.1),
\end{aligned}$$

where the last step of the process is computing by substituting in the values of $\mathbf{P}\left(\mathcal{S} \,|\, X_1, X_2\right)$ as defined by the adaptation process.

The resulting distribution of $\mathbf{P}\left(X_1', X_2'\right)$ is

x_1'	x_2'	$\mathbf{P}\left(X_1' = x_1', X_2' = x_2'\right)$
0	0	0.35
0	1	0.25
1	0	0.25
1	1	0.15

which clearly is not π. The procedure is therefore not π ergodic.

This counter example refutes the proof of Algorithm 7.1 in Levine and Casella [2006], and places conditions upon π-invariance of the Gibbs sampler described in [Geman and Geman, 1984]. Specifically, simply sampling from every variable infinitely frequently is insufficient for π-invariance. It is also necessary that the choice of variable to sample from not depend on the state of the sampler.