Proof that the ADDA and DA Markov chains have the same stationary distribution

Consider a Data Augmentation setting where $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2)$ denotes the collection of latent variables partitioned into two blocks ($\mathcal{D}_1$ and $\mathcal{D}_2$), and θ denotes the parameter of interest. Consider the setting where $r = 0.5$, so the asynchronous version ADDA updates exactly one of the two latent variable blocks at any given iteration. *At any given iteration, we allow the probability that the first latent variable block will be updated to depend on the current parameter value.* Let $\pi_{\mathcal{D},\theta}$ denote the posterior density of interest (this is the stationary density of the corresponding DA algorithm). The dependence of this posterior density on the observed data has been suppressed for notational convenience. We use $\pi_{\mathcal{D}|\theta}, \pi_{\theta|\mathcal{D}}, \pi_\theta, \pi_\mathcal{D}$ to denote the associated conditional and marginal densities. For ease of exposition, we will assume that $\mathcal{D}$ and θ are supported on a discrete space (hence the densities above are with respect to an appropriate discrete measure). The arguments below can be extended in a straightforward way to a general setting with more than two latent variable blocks, and to non-discrete settings.

Let $(\mathcal{D}^{(0)}, \theta^{(0)})$ denote the starting value for the ADDA chain, and assume that it is drawn from the desired posterior $\pi_{\mathcal{D},\theta}$. Let $(\mathcal{D}^{(1)}, \theta^{(1)})$ denote the next iterate generated by the ADDA chain. Our goal is to show that $(\mathcal{D}^{(1)}, \theta^{(1)}) \sim \pi_{\mathcal{D},\theta}$. As we will see below, *a key assumption which enables this is the conditional independence assumption which implies* $\pi_{\mathcal{D}|\theta}(\tilde{d} \mid \tilde{\theta}) = \pi_{\mathcal{D}_1|\theta}(\tilde{d}_1 \mid \tilde{\theta})\pi_{\mathcal{D}_2|\theta}(\tilde{d}_2 \mid \tilde{\theta})$ (recall that $\tilde{d}_1$ and $\tilde{d}_2$ denote the two blocks of $\tilde{d}$).

Since $\theta^{(1)}$ given $\mathcal{D}^{(1)}$ is a draw from $\pi_{\theta|\mathcal{D}}$, it follows that

$$\begin{aligned} P((\mathcal{D}^{(1)}, \theta^{(1)}) = (\tilde{d}, \tilde{\theta})) &= P(\theta^{(1)} = \tilde{\theta} \mid \mathcal{D}^{(1)} = \tilde{d}) P(\mathcal{D}^{(1)} = \tilde{d}) \\ &= \pi_{\theta|\mathcal{D}}(\tilde{\theta} \mid \tilde{d}) P(\mathcal{D}^{(1)} = \tilde{d}). \end{aligned}$$

Hence, to prove the desired result, it is enough to show that $P(\mathcal{D}^{(1)} = \tilde{d}) = \pi_\mathcal{D}(\tilde{d})$. Note that

$$P(\mathcal{D}^{(1)} = \tilde{d}) = \sum_{d', \theta'} P(\mathcal{D}^{(1)} = \tilde{d} \mid (\mathcal{D}^{(0)}, \theta^{(0)}) = (d', \theta')) \pi_{\mathcal{D},\theta}(d', \theta').$$

Let us recall how $\mathcal{D}^{(1)}$ is sampled given $(\mathcal{D}^{(0)}, \theta^{(0)}) = (d', \theta')$. With probability say $c_1(\theta')$, only the first latent variable block $\mathcal{D}_1^{(1)}$ is obtained using a draw from $\pi_{\mathcal{D}_1|\theta}(\cdot \mid \theta')$ and the second block is left unchanged at d'_2, and with probability $c_2(\theta') = 1 - c_1(\theta')$, only the second latent variable block $\mathcal{D}_2^{(1)}$ is obtained using $\pi_{\mathcal{D}_2|\theta}(\cdot \mid \theta')$ and the first block is left unchanged

at d_1'. It follows that

$$\begin{aligned}
& P(\mathcal{D}^{(1)} = \tilde{d}) \\
= & \sum_{d',\theta'} c_1(\theta')\pi_{\mathcal{D}_1|\theta}(\tilde{d}_1 \mid \theta')1_{\{d_2'=\tilde{d}_2\}}\pi_{\mathcal{D},\theta}(d',\theta') + \sum_{d',\theta'} c_2(\theta')\pi_{\mathcal{D}_2|\theta}(\tilde{d}_2 \mid \theta')1_{\{d_1'=\tilde{d}_1\}}\pi_{\mathcal{D},\theta}(d',\theta')
\end{aligned}$$

Using $\pi_{\mathcal{D},\theta}(d',\theta') = \pi_{\mathcal{D}_1|\theta}(d_1' \mid \theta')\pi_{\mathcal{D}_2|\theta}(d_2' \mid \theta')\pi_\theta(\theta')$ (by conditional independence of $\mathcal{D}_1$ and $\mathcal{D}_2$ given θ), we get

$$\begin{aligned}
& P(\mathcal{D}^{(1)} = \tilde{d}) \\
= & \sum_{\theta'}\sum_{d_1'}\sum_{d_2'} c_1(\theta')\pi_{\mathcal{D}_1|\theta}(\tilde{d}_1 \mid \theta')1_{\{d_2'=\tilde{d}_2\}}\pi_{\mathcal{D}_1|\theta}(d_1' \mid \theta')\pi_{\mathcal{D}_2|\theta}(d_2' \mid \theta')\pi_\theta(\theta') + \\
& \sum_{\theta'}\sum_{d_1'}\sum_{d_2'} c_2(\theta')\pi_{\mathcal{D}_2|\theta}(\tilde{d}_2 \mid \theta')1_{\{d_1'=\tilde{d}_1\}}\pi_{\mathcal{D}_1|\theta}(d_1' \mid \theta')\pi_{\mathcal{D}_2|\theta}(d_2' \mid \theta')\pi_\theta(\theta') \\
= & \sum_{\theta'} c_1(\theta')\pi_{\mathcal{D}_1|\theta}(\tilde{d}_1 \mid \theta')\pi_\theta(\theta') \left(\sum_{d_1'} \pi_{\mathcal{D}_1|\theta}(d_1' \mid \theta')\right)\left(\sum_{d_2'} \pi_{\mathcal{D}_2|\theta}(d_2' \mid \theta')1_{\{d_2'=\tilde{d}_2\}}\right) + \\
& \sum_{\theta'} c_2(\theta')\pi_{\mathcal{D}_2|\theta}(\tilde{d}_2 \mid \theta')\pi_\theta(\theta') \left(\sum_{d_2'} \pi_{\mathcal{D}_2|\theta}(d_2' \mid \theta')\right)\left(\sum_{d_1'} \pi_{\mathcal{D}_1|\theta}(d_1' \mid \theta')1_{\{d_1'=\tilde{d}_1\}}\right) \\
= & \sum_{\theta'} c_1(\theta')\pi_{\mathcal{D}_1|\theta}(\tilde{d}_1 \mid \theta')\pi_\theta(\theta')\pi_{\mathcal{D}_2|\theta}(\tilde{d}_2 \mid \theta') + \sum_{\theta'} c_2(\theta')\pi_{\mathcal{D}_2|\theta}(\tilde{d}_2 \mid \theta')\pi_\theta(\theta')\pi_{\mathcal{D}_1|\theta}(\tilde{d}_1 \mid \theta') \\
= & \sum_{\theta'} (c_1(\theta') + c_2(\theta'))\pi_{\mathcal{D},\theta}(\tilde{d},\theta').
\end{aligned}$$

The last step again uses conditional independence of $\mathcal{D}_1$ and $\mathcal{D}_2$ given θ. Since $c_1(\theta')+c_2(\theta') = 1$, it follows that

$$P(\mathcal{D}^{(1)} = \tilde{d}) = \sum_{\theta'} \pi_{\mathcal{D},\theta}(\tilde{d},\theta') = \pi_{\mathcal{D}}(\tilde{d}).$$

Note that the above argument considers the pure asynchronous ADDA ($\epsilon = 0$). But the ADDA kernel with positive ϵ is a mixture of the DA and pure asynchronous ADDA kernels, so the result immediately follows for such settings as well. For a setting with more than two blocks and a general value of $r \in (0,1)$, we will have $J = \binom{K}{\lceil Kr \rceil}$ terms in the derivation above instead of two terms. The j^{th} term, with essentially the same manipulations as above, will simplify to $c_j(\theta')\pi_{\mathcal{D},\theta}(\tilde{d},\theta')$, where $c_j(\theta')$ denotes the probability of choosing the relevant subset of latent variable blocks. Since $\sum_{j=1}^{J} c_j(\theta') = 1$, the result will follow. $\square$