Theorem 1. *For the canonical* MambaZero *model with dimensions $d = N = 2^{k+1} = 4$, $e = 1$, and convolution window $w = 2$, there is a choice of parameters such that the model prediction is exactly equal to the Laplacian estimator for random first-order Markov chains. More formally, for any $\beta > 0$, there exists a set of parameters $\boldsymbol{\theta}$ such that, for all sequences $(x_t)_{t\geq 1}$ and all $t \geq 1$,*

$$D_{\mathrm{KL}}\left(\mathbb{P}_\beta^{(1)}(\cdot \mid x_1^t) \| \mathbb{P}_{\boldsymbol{\theta}}\left(\cdot \mid x_1^t\right)\right) = 0.$$

Proof. Let $\beta > 0$ be the constant of the considered add-constant estimator. Let us fix $a = 0$ and $\Delta_t = 1$, so that $a_t = 1$, for all $t \geq 1$. This can be done by picking, e.g., $\boldsymbol{w}_\Delta = \mathbf{0}$ and δ such that $\mathrm{softplus}(\delta) = 1$. Note that one can

$$\mathrm{conv}_X = \begin{pmatrix} \alpha_{00} & \alpha_{01} \\ \alpha_{10} & \alpha_{11} \end{pmatrix}, \qquad \mathrm{conv}_B = \begin{pmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{pmatrix} \tag{1}$$

where each row corresponds to the kernel weights applied time-wise to each coordinate of the input sequence $(\boldsymbol{x}_t)_{t\geq 1}$. Since the window for conv_C is $w_C = 1$, we can simply assume w.l.o.g. that $C_t = W_C \boldsymbol{x}_t$.

Let us take the embedding vectors to be $\boldsymbol{e}_0 = (1, 0, 0, 0)^\top$ and $\boldsymbol{e}_1 = (0, 1, 0, 0)^\top$, and take

$$W_X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{2}$$

Take also

$$\mathrm{conv}_X = \begin{pmatrix} 1 & 1 \\ 3 & -1 \\ 1 & 1 \\ 3 & -1 \end{pmatrix}, \tag{3}$$

so that one has, after the application of W_X and conv_X,

$$X^{(00)} = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \qquad X^{(01)} = \begin{pmatrix} 1 \\ 3 \\ 1 \\ -1 \end{pmatrix}, \qquad X^{(10)} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 3 \end{pmatrix}, \qquad X^{(11)} = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 2 \end{pmatrix}. \tag{4}$$

Take also $\mathrm{conv}_B = \mathrm{conv}_X$ and $W_B = W_X$, so that

$$B_0 = W_B \boldsymbol{e}_0 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad B_1 = W_B \boldsymbol{e}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \tag{5}$$

and take $W_C = \frac{1}{4} W_B$, so that $C^{(0)} = \frac{1}{4} B_0$ and $C^{(1)} = \frac{1}{4} B_1$.

The final logit vector is in general equal to

$$\mathrm{logit}_t = W_\ell \boldsymbol{x}_t + W_\ell W_o X^{(x_1)} B^{(x_1)^\top} C^{(x_t)} + \sum_{ij} n_{ij} W_\ell W_o X^{(ij)} B^{(ij)\top} C^{(x_t)}. \tag{6}$$

Note that the application of convolution to a given sequence of vectors $\boldsymbol{z}_1^t$ can be rewritten as a linear matrix-form operation. For example, for conv_X, one has that

$$\mathrm{conv}_X(\boldsymbol{z}_t) = D_X^{(0)}\boldsymbol{z}_{t-1} + D_X^{(1)}\boldsymbol{z}_t \tag{7}$$

where $D_X^{(0)} = \mathrm{diag}(\alpha_{00}, \alpha_{10})$ and $D_X^{(1)} = \mathrm{diag}(\alpha_{01}, \alpha_{11})$ are diagonal matrices. The same holds for conv_B, with corresponding diagonal matrices $D_B^{(0)}$ and $D_B^{(1)}$. Using this fact, we can rewrite the logit formula as

$$\begin{aligned}\mathrm{logit}_t &= W_\ell \boldsymbol{x}_t + W_\ell W_o D_X^{(1)} X_{x_1} B_{x_1}^\top D_B^{(1)} C^{(x_t)} \\ &\quad + \sum_{ij} n_{ij} W_\ell W_o (D_X^{(0)} X_i + D_X^{(1)} X_j)(B_i^\top D_B^{(0)} C^{(x_t)} + B_j^\top D_B^{(1)} C^{(x_t)}).\end{aligned} \tag{8}$$

Note that, with the choice of parameters above, one has

$$D_X^{(1)} X_{x_1} B_{x_1}^\top D_B^{(1)} C^{(i)} = \mathbf{0} \tag{9}$$

and also,

$$B_0^\top D_B^{(0)} C^{(1)} + B_0^\top D_B^{(1)} C^{(1)} = 0 \tag{10}$$
$$B_0^\top D_B^{(0)} C^{(1)} + B_1^\top D_B^{(1)} C^{(1)} = 0 \tag{11}$$
$$B_1^\top D_B^{(0)} C^{(0)} + B_0^\top D_B^{(1)} C^{(0)} = 0 \tag{12}$$
$$B_1^\top D_B^{(0)} C^{(0)} + B_1^\top D_B^{(1)} C^{(0)} = 0, \tag{13}$$

that is, only the desired counts are kept in the final logit, depending on the current symbol. Furthermore, for the relevant counts, one has

$$B_0^\top D_B^{(0)} C^{(0)} + B_0^\top D_B^{(1)} C^{(0)} = 1 \tag{14}$$
$$B_0^\top D_B^{(0)} C^{(0)} + B_1^\top D_B^{(1)} C^{(0)} = 1 \tag{15}$$
$$B_1^\top D_B^{(0)} C^{(1)} + B_0^\top D_B^{(1)} C^{(1)} = 1 \tag{16}$$
$$B_1^\top D_B^{(0)} C^{(1)} + B_1^\top D_B^{(1)} C^{(1)} = 1. \tag{17}$$

Hence, the final logit becomes

$$\mathrm{logit}_t = W_\ell \boldsymbol{e}_0 + \sum_j n_{0j} W_\ell W_o (D_X^{(0)} X_0 + D_X^{(1)} X_j) \tag{18}$$

for $x_t = 0$, and

$$\mathrm{logit}_t = W_\ell \boldsymbol{e}_1 + \sum_j n_{1j} W_\ell W_o (D_X^{(0)} X_1 + D_X^{(1)} X_j) \tag{19}$$

for $x_t = 1$. Finally, take

$$W_\ell = \begin{pmatrix} \beta & \beta & 1 & 0 \\ \beta & \beta & 0 & 1 \end{pmatrix} \tag{20}$$

and

$$W_o = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & 1 & -\frac{1}{2} \end{pmatrix}. \tag{21}$$

With this choice, we get, for all $t \geq 1$,

$$\mathrm{logit}_t = \begin{pmatrix} \beta \\ \beta \end{pmatrix} + n_{x_t,0} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + n_{x_t,1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{22}$$

After the normalization, we finally get

$$f_{\boldsymbol{\theta}}(x_1^t) = \left(\frac{n_{00} + \beta}{n_{00} + n_{01} + 2\beta} \, , \, \frac{n_{01} + \beta}{n_{00} + n_{01} + 2\beta} \right)^\top \tag{23}$$

if $x_t = 0$, and

$$f_{\boldsymbol{\theta}}(x_1^t) = \left(\frac{n_{10} + \beta}{n_{10} + n_{11} + 2\beta} \, , \, \frac{n_{11} + \beta}{n_{10} + n_{11} + 2\beta} \right)^\top \tag{24}$$

if $x_t = 1$. This is precisely the required add-β Laplacian estimator. □