APPENDIX – SUPPLEMENTARY MATERIAL

A Proofs

Here we will show that the joint pdf of the entire field conditioned on its margin (Figure 9, gray area) equals the product of the predictive conditional distributions.

Proof of Proposition 2.3. To simplify notation, we assign index numbers $i \in 1, \dots N$ to the space-time grid $(\mathbf{s}, t)$ to ensure that the PLC of i_1 cannot contain X_{i_2} if $i_2 > i_1$. We can do this by iterating through space and time in increasing order over time (and, for fixed t, any order over space):

$$\begin{aligned}(\mathbf{s}, t), \mathbf{s} \in \mathbf{S}, t \in \mathbb{T} \rightarrow \left(i_{(t-1)\cdot|\mathbf{S}|+1}, \dots, i_{(t-1)\cdot|\mathbf{S}|+|\mathbf{S}|}\right) \\ = (t-1)\cdot|S| + (1, \dots, |S|).\end{aligned} \tag{31}$$

We assume that the process we observed is part of a larger field on an extended grid $\tilde{\mathbf{S}} \times \tilde{\mathbb{T}}$, with $\tilde{\mathbf{S}} \supset \mathbf{S}$ and $\tilde{\mathbb{T}} = \{-(h_p - 1), \dots, 0, 1, \dots T$, for a total of $\tilde{N} > N$ space-time points, $X_1, \dots, X_{\tilde{N}}$. The margin $\mathbf{M}$ are all points $X(\mathbf{s}, u), (\mathbf{s}, u) \in \tilde{\mathbf{S}} \times \tilde{\mathbb{T}}$ that do not have a fully observed past light cone. Formally,

$$\mathbf{M} = \{X(\mathbf{s}, u) \mid \ell^{-}(\mathbf{s}, u) \notin \{X(\mathbf{r}, t), (\mathbf{r}, t) \in \mathbf{S} \times \mathbb{T}\}\}, \tag{32}$$

The size of $\mathbf{M}$ depends on the past horizon h_p as well as the speed of propagation c, $\mathbf{M} = \mathbf{M}(h_p, c)$.

In Figure 9, the part of the field with fully observed PLCs are marked in red. Points in the margin, in gray, have PLCs extending into the fully unobserved region (blue). Points in the red area have a PLC that lies fully in the red or partially in the gray, but never in the blue area. As can be see in Fig. 9 the margin at each t is a constant fraction of space, thus overall $\mathbf{M}$ grows linearly with T; it does not grow with an increasing $\mathbf{S}$, but stays constant.

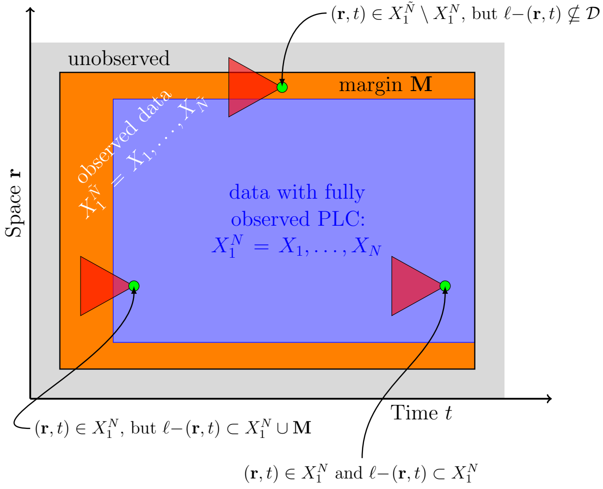
$(\mathbf{r}, t) \in X_1^{\tilde{N}} \setminus X_1^N$, but $\ell{-}(\mathbf{r}, t) \nsubseteq \mathcal{D}$
unobserved
margin $\mathbf{M}$
observed data $X_1^{\tilde{N}} = X_1, \dots, X_{\tilde{N}}$
data with fully observed PLC: $X_1^N = X_1, \dots, X_N$
Space $\mathbf{r}$
Time t
$(\mathbf{r}, t) \in X_1^N$, but $\ell{-}(\mathbf{r}, t) \subset X_1^N \cup \mathbf{M}$
$(\mathbf{r}, t) \in X_1^N$ and $\ell{-}(\mathbf{r}, t) \subset X_1^N$

Figure 9: Margin of spatio-temporal field in $(1+1)D$.

For simplicity, assume that $X_1, \dots, X_N$ are from the truncated (red) field, such that all their PLCs are observed (they may lie in $\mathbf{M}$), and the remaining $\tilde{N} - N$ X_js lie in $\mathbf{M}$ (with a PLC that is only partially unobserved). Furthermore let $X_1^k := \{X_1, \dots, X_k\}$. Thus

$$\begin{aligned}\mathbb{P}\left(\{X(\mathbf{s}, t) \mid (\mathbf{s}, t) \in \tilde{\mathbf{S}} \times \tilde{\mathbb{T}}\}\right) &= \mathbb{P}\left(X_1^{\tilde{N}}\right) \\ &= \mathbb{P}\left(X_1^N, \mathbf{M}\right) \\ &= \mathbb{P}\left(X_1^N \mid \mathbf{M}\right) \mathbb{P}(\mathbf{M})\end{aligned}$$

The first term factorizes as

$$\begin{aligned}&\mathbb{P}\left(X_1^N \mid \mathbf{M}\right) \\ &= \mathbb{P}\left(X_N \mid X_1^{N-1}, \mathbf{M}\right) \mathbb{P}\left(X_1^{N-1} \mid \mathbf{M}\right) \\ &= \mathbb{P}\left(X_N \mid \ell_N^{-} \cup \{X_1^{N-1}, \mathbf{M}\} \setminus \{\ell_N^{-}\}\right) \mathbb{P}\left(X_1^{N-1} \mid \mathbf{M}\right) \\ &= \mathbb{P}\left(X_N \mid \ell_N^{-}\right) \mathbb{P}\left(X_1^{N-1} \mid \mathbf{M}\right)\end{aligned}$$

where the second-to-last equality follows since by (31), $\ell_N^{-} \subset \{X_k \mid 1 \leq k < N\} \cup \mathbf{M}$, and the last equality holds since X_i is conditional independent of the rest given its own PLC (due to limits in information propagation over space-time).

By induction,

$$\begin{aligned}\mathbb{P}(X_1, \dots, X_N \mid \mathbf{M}) &= \prod_{j=0}^{N-1} \mathbb{P}\left(X_{N-j} \mid \ell_{N-j}^{-}\right) &(33) \\ &= \prod_{i=1}^{N} \mathbb{P}\left(X_i \mid \ell_i^{-}\right) &(34)\end{aligned}$$

This shows that the conditional log-likelihood maximization we use for our estimators is equivalent (up to a constant) to full joint maximum likelihood estimation. □

B Predictive States and Mixture Models

Another way to understand predictive states is as the extremal distributions of an optimal mixture model (Lauritzen, 1974, 1984).

To predict any variable L^{+}, we have to know its distribution $\mathbb{P}(L^{+})$. If, as often happens, that distribution is very complicated, we may try to decompose it into a mixture of simpler "base" or "extremal" distributions, $\mathbb{P}(L^{+} \mid \theta)$, with mixing weights $\pi(\theta)$,

$$\mathbb{P}\left(L^{+}\right) = \int \pi(\theta) \mathbb{P}\left(L^{+} \mid \theta\right) d\theta\ . \tag{35}$$

The familiar Gaussian mixture model, for instance, makes the extremal distributions to be Gaussians (with θ indexing both expectations and variances), and makes the mixing weights $\pi(\theta)$ a combination of delta functions, so that $\mathbb{P}(L^+)$ becomes a weighted sum of finitely-many Gaussians.

The conditional predictive distribution of $L^+ \mid \ell^-$ in (35) is a weighted average over the extremal conditional distributions $\mathbb{P}(L^+ \mid \theta, \ell^-)$,

$$\mathbb{P}\left(L^+ \mid \ell^-\right) = \int \pi(\theta|\ell^-)\mathbb{P}\left(L^+ \mid \theta, \ell^-\right) d\theta \qquad (36)$$

This only makes the forecasting problem harder, unless $\mathbb{P}(L^+ \mid \theta, \ell^-)\, \pi(\theta \mid \ell^-) = \mathbb{P}\left(L^+ \mid \hat{\theta}(\ell^-)\right) \delta(\theta - \hat{\theta}(\ell^-))$, that is, unless $\hat{\theta}(\ell^-)$ is a predictively sufficient statistic for L^+. The most parsimonious mixture model is the one with the minimal sufficient statistic, $\theta = \epsilon(\ell^-)$. This shows that predictive states are the best "parameters" in (35) for optimal forecasting. Using them,

$$\mathbb{P}\left(L^+\right) = \sum_{j=1}^{K} \mathbb{P}\left(\epsilon(\ell^-) = s_j\right) \mathbb{P}\left(L^+ \mid \epsilon(\ell^-) = s_j\right) \qquad (37)$$

$$= \sum_{j=1}^{K} \pi_j(\ell^-) \cdot \mathbb{P}_j\left(L^+\right) , \qquad (38)$$

where $\pi_j(\ell^-)$ is the probability that the predictive state of ℓ^- is s_j, and $\mathbb{P}_j(L^+) = \mathbb{P}(L^+ \mid S = s_j)$. Since each light cone has a unique predictive state,

$$\pi_j(\ell^-) = \begin{cases} 1, & \text{if } \epsilon(\ell^-) = s_j, \\ 0 & \text{otherwise.} \end{cases} \qquad (39)$$

Thus the predictive distribution given ℓ_i^- is just

$$\mathbb{P}\left(L^+ \mid \ell_i^-\right) = \sum_{j=1}^{K} \pi_j(\ell_i^-) \cdot \mathbb{P}_j\left(L^+\right) = \mathbb{P}_{\epsilon(\ell_i^-)}\left(L^+\right). \qquad (40)$$

Now the forecasting problem simplifies to mapping ℓ_i^- to its predictive state, $\epsilon(\ell_i^-) = s_j$; the appropriate distribution-valued forecast is $p_j(L^+)$, and point forecasts are derived from it as needed.

This mixture-model point of view highlights how prediction benefits from grouping points by their predictive consequences, rather than by spatial proximity (as a Gaussian mixture would do). For us, this means clustering past light-cone configurations according to the similarity of their predictive distributions, not according to (say) the Euclidean geometry. Mixed LICORS thus learns a new geometry for the system, which is optimized for forecasting.