High-dimensional dynamical linear models: information-theoretic and algorithmic limits

D. Tieplova[1], **S. Lahiry**[2], **P. Sur**[3], **and J. Barbier**[4]

[1,4] *The Abdus Salam International Centre for Theoretical Physics, Italy*
[2,3] *Department of Statistics, Harvard University, USA*

The analysis of high-dimensional time series datasets is becoming increasingly crucial, driven by recent technological advancements. These datasets find applications in diverse fields such as finance, neuroscience, and environmental science, showcasing both cross-sectional and temporal dependencies, which often present analytical challenges.

While many scenarios can be framed as linear regression problems for signal estimation, traditional high-dimensional statistical approaches typically assume sparsity in the signals. However, over the last decade, a new paradigm influenced by statistical physics and information theory has emerged, relaxing these stringent sparsity assumptions. This paradigm examines scenarios where both the number of features and samples grow proportionally.

Within this framework, a significant focus lies in computing Minimum Mean Square Estimation (MMSE) in a Bayesian setting, akin to determining the normalized mutual information between signal and measurement. This involves establishing a single-letter formula for normalized mutual information in the large sample limit, deriving MMSE from it, and constructing an asymptotically optimal estimator.

While such formulas have been rigorously established for i.i.d. Gaussian setups and some correlated random ensembles, they do not encompass dependence structures inherent in certain time series models. In this talk, we aim to bridge this gap by establishing the single-letter formula for dynamic linear models, where the measurement matrix rows follow an AR(1) process.

More precisely, we define a p –dimensional, centered, stationary Gaussian process x_t as

$$x_{t+1} = Ax_t + \xi_t, \quad t \in \mathbb{Z}, \tag{1}$$

where A is a $p \times p$ deterministic matrix and ξ_t – i.i.d. p dimensional vectors of $\mathcal{N}(0, \sigma_1^2 I_p)$. We have N observations (x_t, y_t) where y_t is given by the following linear model:

$$y_t = \frac{1}{\sqrt{p}} x_t^T \beta + \omega_t, \quad t = 1, \dots, N \tag{2}$$

where β is a signal vector drawn from prior distribution P_β and ω_t is i.i.d. noise $\mathcal{N}(0, \sigma_2^2)$. We consider the regime where N and p both tend to infinity at the same rate, i.e. $N/p = c_N \to c > 0$. For this model, we derive a single-letter formula for the normalized mutual information between the measurements and the signal using replica method and later prove it using adaptive interpolation. We also discuss perspectives of utilizing the Vector Approximate Message Passing algorithm to estimate signal β in our setting.