Partial Disentanglement via Mechanism Sparsity

Sébastien Lachapelle[1], Simon Lacoste-Julien[1,2]

[1]Mila, Université de Montréal [2]Canada CIFAR AI Chair

Contributions

- [7] showed how **mechanism sparsity regularization** can identify causal latent factors from high-dimensional observations based on the assumption that the **ground-truth causal graph connecting the latent factors is typically sparse** [1, 3].
  - Objects usually interact sparsely with each others.
  - Actions typically affect only a few factors of variations.
- [7] introduced a **graphical criterion** guaranteeing complete disentanglement.
- This work generalizes [7] by dropping the graphical criterion and instead **characterizes qualitatively how disentangled the learned representation is expected to be** given the specific form of the ground-truth causal graph.
- To do so, we introduce a novel equivalence relation over models we call **consistency**.
- This equivalence relation captures which variables are expected to remain entangled and which are not, hence the term **partial disentanglement**.
- The graphical criterion of [7] can be derived from our more general theory.
- We follow [7] by leveraging VAE and gumbel-sigmoid masks, but replace sparsity regularization by a **sparsity constraint**, as argued for in [2].
- Illustrations of our theory with synthetic data.

1- Background

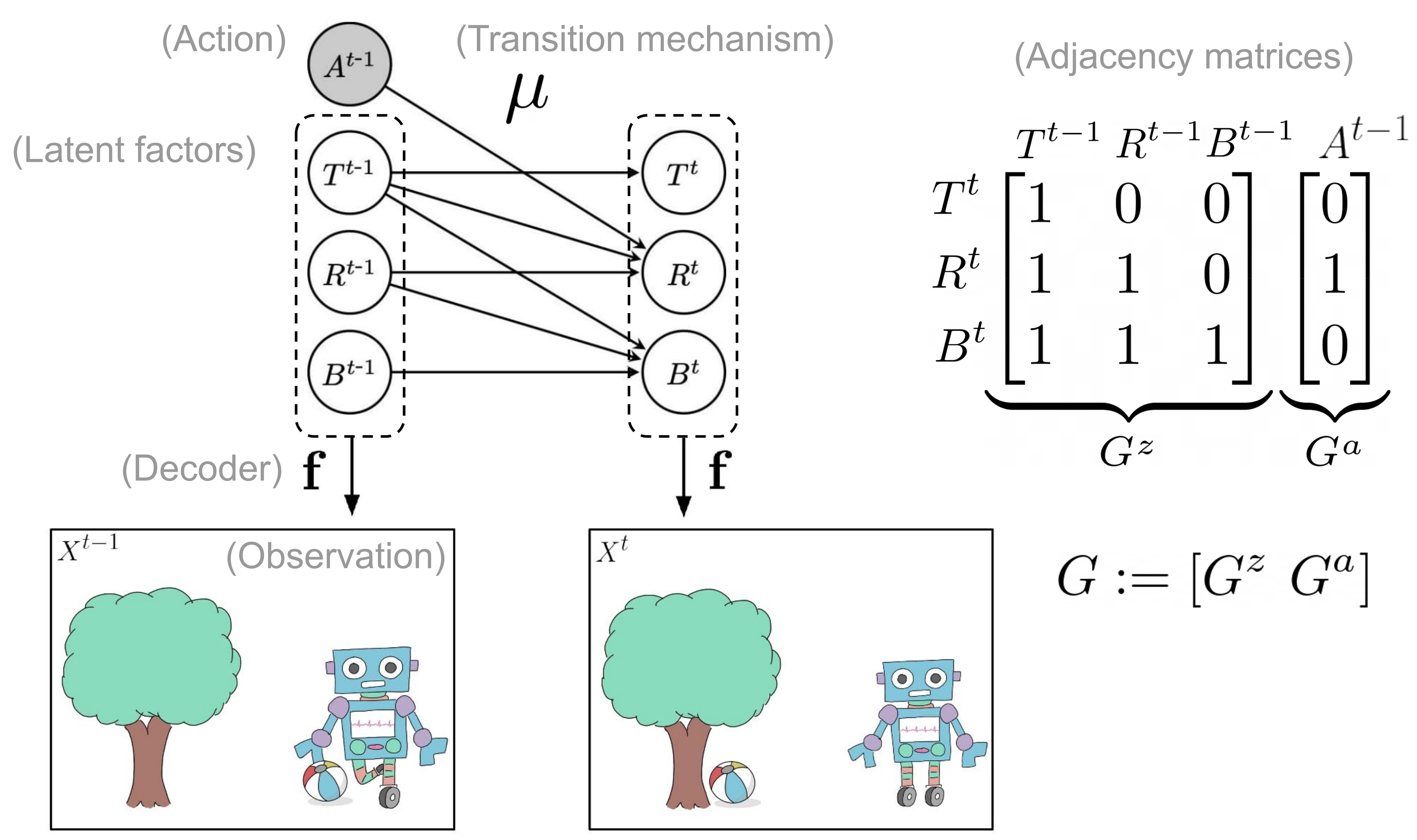


Figure 1: Model in the context of the motivating example of [7]

1.1 Model (following [7])

- We observe sequences $\{X^t\}_{t=1}^T$ and $\{A^t\}_{t=1}^T$ and have

$$X^t = \mathbf{f}(Z^t) + \texttt{noise}^t \text{ with } \mathbf{f}: \mathbb{R}^{d_z} \to \mathcal{X} \subset \mathbb{R}^{d_x} \text{ (diffeomorphism and } d_z \leq d_x)$$

- We follow [5] and assume the Z_i^t **are independent given** $Z^{<t}$ **and** $A^{<t}$

$$p(z^t \mid z^{<t}, a^{<t}) = \prod_{i=1}^{d_z} p(z_i^t \mid z^{<t}, a^{<t}),$$

and, for simplicity, assume each factor is Gaussian with a fixed variance i.e.

$$p(z_i^t \mid z^{<t}, a^{<t}) = \mathcal{N}(z_i^t; \mu_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t}), 1).$$

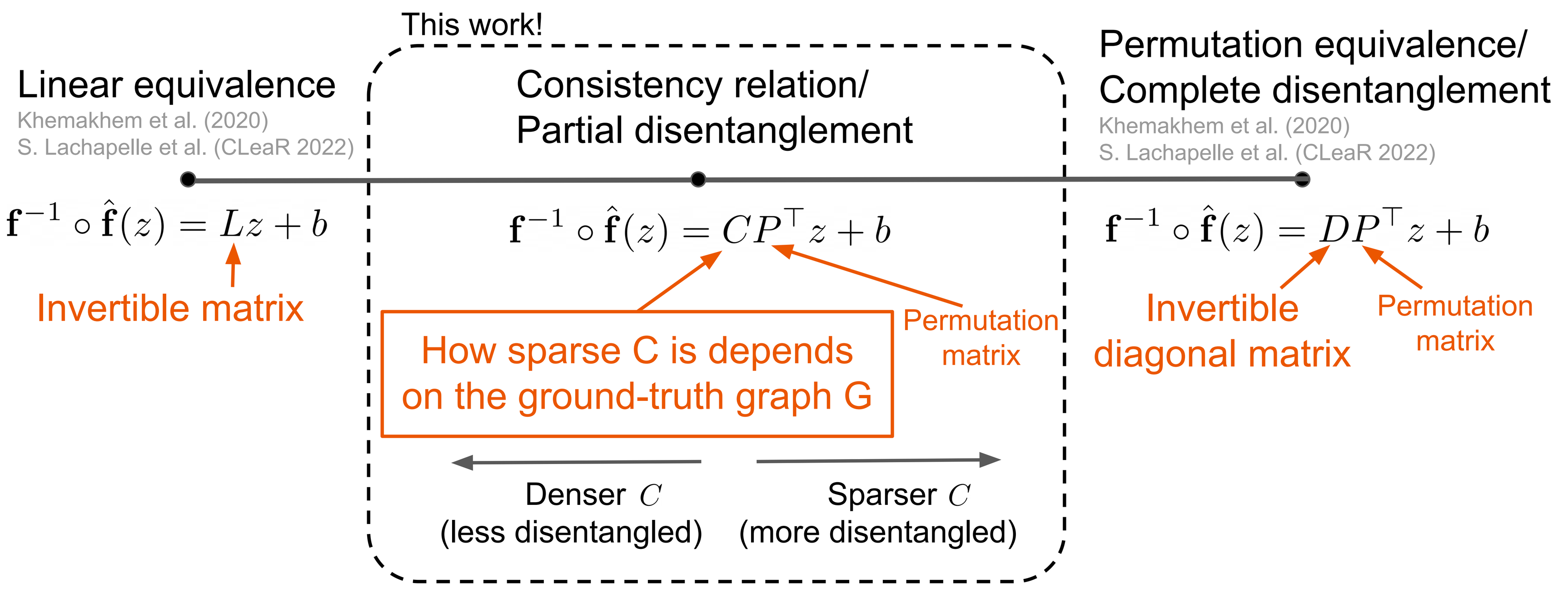
Paper is more general and applies to the **exponential family** with 1d sufficient statistics.

- μ_i's are the *transition functions*/*mechanisms* (e.g. parameterized by a NN).
- $G = [G^z G^a]$ is the adjacency matrix of the causal graph.
- **Learnable parameters** are $\theta := (\mathbf{f}, \mu, G)$

1.2 Anatomy of identifiability results

- Postulate a family of distributions over observations $\mathbb{P}_\theta$ parameterized by θ
- Make assumptions about the ground-truth model θ
- Prove guarantee of the form: $\mathbb{P}_\theta = \mathbb{P}_{\hat{\theta}} \implies \theta \sim \hat{\theta}$, where $\sim$ is some equivalence relation that is more or less strong, depending on the assumptions.

1.2 Continuum of equivalence relations



2- Identifiability result

Theorem (Partial disentanglement via mechanism sparsity) Suppose we have two models with parameters $\theta = (\mathbf{f}, \mu, G)$ and $\hat{\theta} = (\hat{\mathbf{f}}, \hat{\mu}, \hat{G})$ representing the same distribution, i.e. $\mathbb{P}_{X^{\leq T} | a^{<T}; \theta} = \mathbb{P}_{X^{\leq T} | a^{<T}; \hat{\theta}}$ for all $a^{<T}$. Assume

1. **[Variability]** The mechanisms $\boldsymbol{\lambda}_i$ are "sufficiently complex" (see paper)
2. **[Sparsity]** $||\hat{G}||_0 \leq ||G||_0$

Then, $\hat{\theta}$ is consistent with θ, i.e. $\theta \sim_{\text{con}} \hat{\theta}$ (see next definition).

Definition ($\sim_{\text{con}}$-equivalence) Two models $\theta := (\mathbf{f}, \mu, G)$ and $\tilde{\theta} := (\tilde{\mathbf{f}}, \tilde{\mu}, \tilde{G})$ are **consistent**, denoted $\theta \sim_{\text{con}} \tilde{\theta}$, if and only if there exists a permutation matrix P such that

1. $G^z = P^\top \tilde{G}^z P$ and $G^a = P^\top \tilde{G}^a$, and
2. $\mathbf{f}^{-1} \circ \hat{\mathbf{f}}(z) = CP^\top$, where the matrix C is invertible, G^z-consistent, $(G^z)^\top$-consistent and G^a-consistent (see next definition).

Definition (S-consistency) Given a binary matrix $S \in \{0,1\}^{m \times n}$, a matrix $C \in \mathbb{R}^{m \times m}$ is **S-consistent** when

$$\forall i, j, \; [\mathbb{1} - S(\mathbb{1} - S)^\top]_{i,j}^+ = 0 \implies C_{i,j} = 0\,, \qquad (1)$$

where $[\cdot]^+ := \max\{0, \cdot\}$ and $\mathbb{1}$ is a matrix filled with ones.

- We showed that, **for any binary matrix S, the set of invertible S-consistent matrices forms a group under matrix multiplication**. This allowed us to show that the relation $\sim_{\text{con}}$ is indeed an equivalence relation.
- Recall the graphical criterion of [7]:
$\forall 1 \leq i \leq d_z, \left(\bigcap_{j \in \mathbf{Ch}_i^z} \mathbf{Pa}_j^z\right) \cap \left(\bigcap_{j \in \mathbf{Pa}_i^z} \mathbf{Ch}_j^z\right) \cap \left(\bigcap_{\ell \in \mathbf{Pa}_i^a} \mathbf{Ch}_\ell^a\right) = \{i\}$,
where $\mathbf{Pa}_i^z$ and $\mathbf{Ch}_i^z$ are the sets of parents and children of node z_i in G^z, respectively, while $\mathbf{Ch}_\ell^a$ is the set of children of a_ℓ in G^a.
- We proved that if the ground-truth graph happens to satisfy the criterion of [7] (above), the matrix C must be diagonal i.e. we have complete disentanglement.

4- Experiments

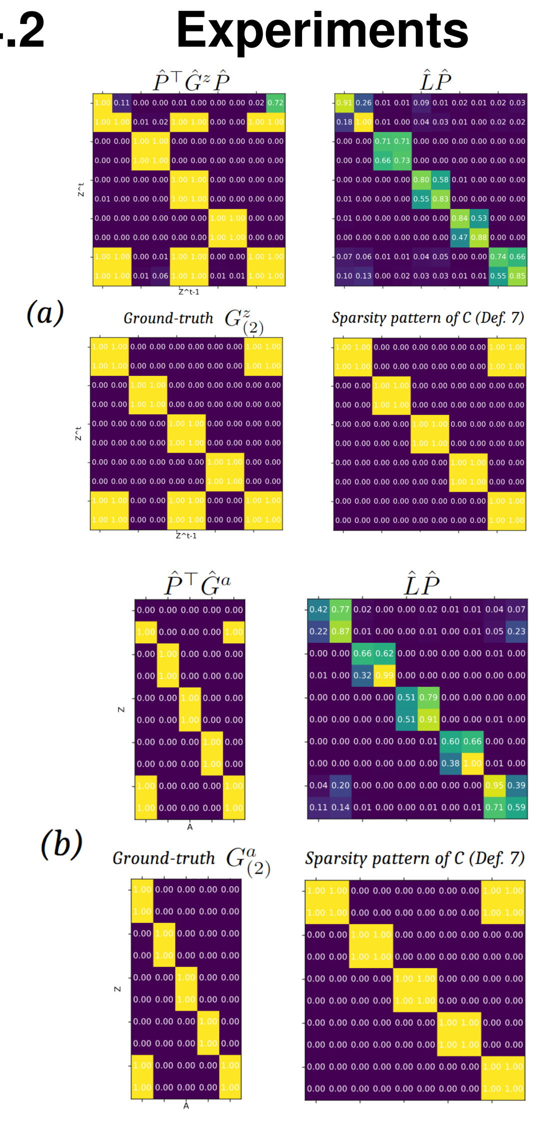
4.1 Learning with VAE + sparsity constraint

- We use the **VAE framework** [6] with a variational approximate posterior given by $q(z^{\leq T} \mid x^{\leq T}, a^{<T}) := \prod_{t=1}^T q(z^t \mid x^t)$.
- This yields the following **evidence lower bound (ELBO)**: $\log p(x^{\leq T} | a^{<T}) \geq$

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{Z^t \sim q(\cdot | x^t)} [\log p(x^t \mid Z^t)] + \mathop{\mathbb{E}}_{Z^{t-1} \sim q(\cdot | x^{t-1})} KL(q(Z^t \mid x^t) || p(Z^t \mid Z^{t-1}, a^{t-1}))\,.$$

- We model the mechanisms $\boldsymbol{\lambda}_i$ with MLPs
- Instead of adding a sparsity penalty as in [7], we add sparsity constraint, following [2].
- The binary masks G^z and G^a are treated as random to allow for optimization via SGD using the **Gumbel-Softmax trick** [8, 4].

4.2 Experiments on synthetic data $(d_x = 20, \; d_z = 10)$



Sparsity	SHD (# edge errors)	MCC (permutation eq.)	R_{con} (consistency)	R (linear eq.)
No	—	.68±.03	.78±.02	.98±.00
Yes	**5.6**±5.0	**.86**±.02	**.99**±.01	**1.0**±.00

Table 1: Performance with and without the sparsity constraint on **synthetic dataset with temporal dependencies**. **Left:** Ground-truth graph and a learned graph of a typical run.

Sparsity	SHD (# edge errors)	MCC (permutation eq.)	R_{con} (consistency)	R (linear eq.)
No	—	.69±.05	.83±.02	.95±.00
Yes	**1.6**±1.7	**.81**±.06	**.98**±.03	**.99**±.01

Table 2: Performance with and without the sparsity constraint on **synthetic dataset with actions**. **Left:** Ground-truth graph and a learned graph of a typical run.

References

[1] Y. Bengio. "The Consciousness Prior". In: *arXiv preprint arXiv:1709.08568* (2019).

[2] J. Gallego-Posada, J. Ramirez De Los Rios, and A. Erraqabi. "Flexible Learning of Sparse Neural Networks via Constrained L0 Regularization". In: *NeurIPS 2021 Workshop LatinX in AI*. 2021.

[3] A. Goyal et al. "Recurrent Independent Mechanisms". In: *ICLR*. 2021.

[4] E. Jang, S. Gu, and B. Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *ICML* (2017).

[5] I. Khemakhem et al. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". In: *AISTATS*. 2020.

[6] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *ICLR*. 2014.

[7] S. Lachapelle et al. "Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA". In: *CLeaR*. 2022.

[8] C. J. Maddison, A. Mnih, and Y. W. Teh. "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables". In: *ICML* (2017).