Title

Abstract:

Keywords:

1 Formulation

Consider a random variable $\mathbb{X} \in \mathbb{R}^p$ that has a sparse dependency structure among its features. This graph structure is potentially non-linear, and we want to infer the structure from a data matrix $\mathbf{X} \in \mathbb{M}(n, p)$.

We assume a multi-layer generative model for the structure:

$$\begin{aligned}
\mathbf{X} &= \varphi(\mathbf{H}_1)\mathbf{B}_1 + \mathbf{E}_x; \quad \mathbb{E} \sim \mathcal{N}_p(\mathbf{0}, \Sigma_x), \\
\mathbf{H}_1 &= \varphi(\mathbf{H}_2)\mathbf{B}_2 + \mathbf{F}_1; \quad \mathbb{F}_1 \sim \mathcal{N}_{p_1}(\mathbf{0}, \Sigma_1), \\
&\cdots \\
\mathbf{H}_{L-1} &= \varphi(\mathbf{H}_L)\mathbf{B}_L + \mathbf{F}_{L-1}; \quad \mathbb{F}_{L-1} \sim \mathcal{N}_{p_{L-1}}(\mathbf{0}, \Sigma_{L-1}), \\
\mathbb{H}_L &\sim \mathcal{N}_{p_L}(\mathbf{0}, \Sigma_L).
\end{aligned}$$

with L hidden layers, and $\varphi(\cdot)$ being a pointwise known transformation (e.g. ReLU, sigmoid, tanh). When Σ_x and $\Sigma_l, l \in \mathcal{I}_L$ are diagonal, it is the Non-linear Gaussian Belief Network of Frey and Hinton (1999). In our case, we keep Σ_x non-diagonal (but sparse), while others diagonal.

The negative log-likelihood function is

$$-\ell(\mathbf{X}|\mathcal{H}, \mathcal{B}, \Omega) = \frac{n}{2}\left[\mathrm{Tr}\,(\mathbf{S}_x\Omega_x) - \log\det\Omega_x + \sum_{l=1}^{L}\left\{\mathrm{Tr}\,(\mathbf{S}_l\Omega_l) - \log\det\Omega_l\right\}\right]$$

where $\mathbf{S}_x = \mathbf{E}_x^T\mathbf{E}_x/n, \mathbf{S}_l = \mathbf{F}_l^T\mathbf{F}_l/n$ for $l = 1, \ldots, L-1$ and $\mathbf{S}_L = \mathbf{H}_L^T\mathbf{H}_L/n$. Inferring the distribution of the hidden variables is difficult so we assume pointwise variational approximations:

$$h_{ij,l} \sim N(\mu_{ijl}, s_{ijl}); \quad i \in \mathcal{I}_n, j \in \mathcal{I}_{p_l}, l \in \mathcal{I}_L.$$

Collect the variational parameters in $\mathcal{M} := \{\mathbf{M}_1, \ldots, \mathbf{M}_L\}, \mathcal{S} := \{\mathbf{S}_1, \ldots, \mathbf{S}_L\}$. Now we have the variational lower-bound

$$\ell(\mathbf{X}|\mathcal{H}, \mathcal{B}, \Omega) \geq \mathbb{E}_q\ell(\mathbf{X}, \mathcal{H}|\mathcal{B}, \Omega, \mathcal{M}, \mathcal{S}) - \mathbb{E}_q \log q(\mathcal{H}|\mathbf{X}, \mathcal{B}, \Omega, \mathcal{M}, \mathcal{S}) \tag{1.1}$$

Denote this lower bound by $\ell_q(\mathbf{X}|\mathcal{B}, \Omega, \mathcal{M}, \mathcal{S})$. Under the simplified model $\Sigma_l = \boldsymbol{\sigma}_l \mathbf{I}$ for $l \in \mathcal{I}_L$, the second term becomes (Frey and Hinton, 1999)

$$\mathbb{E}_q \log q(\mathcal{H}|\mathbf{X}, \mathcal{B}, \Omega, \mathcal{M}, \mathcal{S}) = \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{j=1}^{p_l}\sum_{l=1}^{L} \log \frac{s_{ijl}}{\sigma_{jl}} - \frac{s_{ijl}}{\sigma_{jl}} + n \log \det \Omega_x + \text{constant}\right]. \tag{1.2}$$

For the first term we have

$$\begin{aligned}\mathbb{E}_q \ell(\mathbf{X}, \mathcal{H}|\mathcal{B}, \Omega, \mathcal{M}, \mathcal{S}) &= \frac{n}{2}\mathbb{E}_q\left[\text{Tr}(\mathbf{S}_x \Omega_x) + \sum_{l=1}^{L} \text{Tr}(\mathbf{S}_l \Omega_l)\right] \\ &= \end{aligned}$$

which simplifies to (Frey and Hinton, 1999)

$$-\left[\mathbb{E}_q \text{Tr}(\mathbf{E}_x^T \mathbf{E}_x \Omega_x) + \sum_{i=1}^{n}\sum_{j=1}^{p_l}\sum_{l=1}^{L-1} \frac{1}{\sigma_{jl}}\left\{(\mu_{ijl} - b_{ij,l+1} m_{ij,l+1})^2 + b_{ij,l+1}^2 v_{ij,l+1}\right\} + \text{const}\right] \tag{1.3}$$

where $m_{ijl} = \mathbb{E}_q \varphi(h_{ijl}), v_{ijl} = \mathbb{E}_q(\varphi(h_{ijl}) - m_{ijl})^2$.

1.1 Objective function

We shall solve a penalized version of the variational lower bound in (1.1):

$$-\frac{2}{n}\ell_q(\mathbf{X}|\mathcal{B}, \Omega, \mathcal{M}, \mathcal{S}) + \sum_{l=1}^{L} \|\mathbf{B}_l\|_1 + \|\Omega_x\|_{1,\text{off}} + P(\mathcal{M}) + Q(\mathcal{S})$$

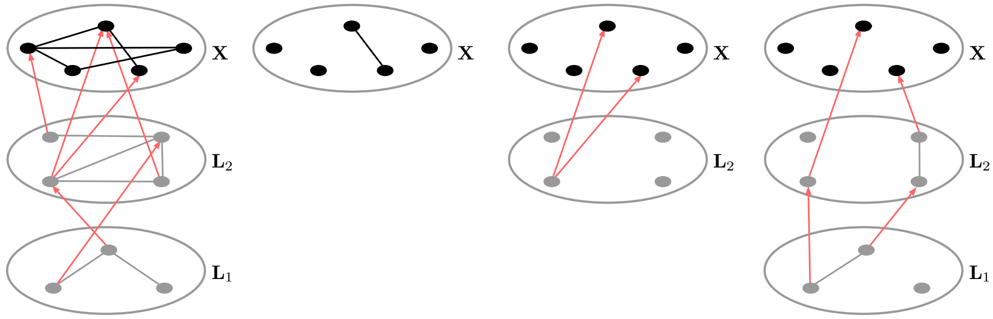
with P, Q being penalties over the variational parameters. We solve this using a variational (monte-carlo?) EM algorithm-
E -step: Given values of $\mathcal{B}, \Omega_x, \boldsymbol{\sigma}_l$, solve for the variational parameters by solving

$$-\frac{2}{n}\ell_q(\mathbf{X}|\mathcal{B}, \Omega, \mathcal{M}, \mathcal{S}) + P(\mathcal{M}) + Q(\mathcal{S})$$

M -step: Given the variational parameters, solve for the model parameters by solving an ℓ_1-penalized version of (1.3).

We take the greedy strategy of solving two-layer problems successively. This means molte-carlo *sequential* EM: first solve for the variational parameters $(\mathbf{M}_1, \mathbf{S}_1) = ((\mu_{ij,1}, s_{ij,1}))$, in the E step, then solve for $(\mathbf{B}_1, \Omega_x)$ in the M step, and continue until convergence. After that only go to the next layer. Similar to Bengio et al. (2007); Hinton and Salakhutdinov (2006). We assume a rank-1 representation for $\mathbf{M} \equiv \mathbf{M}_1$ and $\mathbf{S} \equiv \mathbf{S}_1$:

$$\begin{aligned}\mathbf{M} &= \mathbf{ab}^T, \mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^q, q \equiv p_1, \\ \mathbf{S} &= \mathbf{cd}^T, \mathbf{c} \in \mathbb{R}^n, \mathbf{d} \in \mathbb{R}^q\end{aligned}$$



We further assume generative parameters for $\mathbf{a}$: $a_i \sim N(\mu, \sigma^2)$ for $i \in \mathcal{I}_n$.

Can calculate gradients of E-step using chain rule and Appendix A of Frey and Hinton (1999).

Now the objective function for a two-layer model becomes:

$$\mathrm{Tr}\left[\frac{1}{n}(\mathbf{X} - \varphi(\mathbf{H})\mathbf{B})^T(\mathbf{X} - \varphi(\mathbf{H})\mathbf{B})\Omega_x\right] + \log\det\Omega_x + \|\mathbf{B}\|_1$$

We assume the following hierarchical structures for the hidden variables and associated variational parameters:

$$\begin{aligned} h_{ij} &\sim N(\mu_{ij}, \sigma_{ij}^2), \\ \mu_{ij} &\sim \sum_{k=1}^{K} I_{mk} N(\mu_{mk}, \sigma_{mk}^2); \quad I_{mk} = \mathrm{Ber}(\pi_{mk}), \\ \sigma_{ij} &\sim \sum_{k=1}^{K} I_{sk} N(\mu_{sk}, \sigma_{sk}^2); \quad I_{sk} = \mathrm{Ber}(\pi_{sk}), \end{aligned}$$

Thus in total there are $6K$ variational parameters.

E-step: we solve for the variational parameters by minimizing the following (take $\mathbf{H}_\varphi \equiv \varphi(\mathbf{H})$)

$$\begin{aligned} \mathcal{F}(\mathbf{M}, \mathbf{S}) &= \mathbb{E}_q \,\mathrm{Tr}\left[\frac{1}{n}(\mathbf{X} - \mathbf{H}_\varphi\mathbf{B})^T(\mathbf{X} - \mathbf{H}_\varphi\mathbf{B})\Omega_x\right] \\ &= \mathrm{Tr}\left[\left\{\frac{1}{n}(\mathbf{X} - \mathbf{M}_\varphi\mathbf{B})^T(\mathbf{X} - \mathbf{M}_\varphi\mathbf{B}) + \mathbf{B}^T\mathbf{V}_\varphi\mathbf{B}\right\}\Omega_x\right] \\ &= \left[\sum_{j=1}^{p}\sum_{j'=1}^{p}\omega_{jj'}\left\{\frac{1}{n}(\mathbf{X}_j - \mathbf{M}_\varphi\mathbf{B}_j)^T(\mathbf{X}_{j'} - \mathbf{M}_\varphi\mathbf{B}_{j'}) + \mathbf{B}_j^T\mathbf{V}_\varphi\mathbf{B}_{j'}\right\}\right] \\ &= \sum_{j,j'=1}^{p}\omega_{jj'}\left\{-\frac{2}{n}\mathbf{X}_j^T\mathbf{M}_\varphi\mathbf{B}_{j'} + \mathbf{B}_j^T\left(\frac{1}{n}\mathbf{M}_\varphi^T\mathbf{M}_\varphi + \mathbf{V}_\varphi\right)\mathbf{B}_{j'}\right\} + c \end{aligned}$$

where $(\mathbf{M}_\varphi)_{ik} = \mathbb{E}_q \varphi(h_{ik})$ for $i \in \mathcal{I}_n, k \in \mathcal{I}_q$, and $\mathbf{V}_\varphi = \mathbb{E}_q[(\mathbf{H}_\varphi - \mathbf{M}_\varphi)^T (\mathbf{H}_\varphi - \mathbf{M}_\varphi)/n]$. Differentiating with respect to entries of $\mathbf{M}$ we now have

$$\frac{\partial \mathcal{F}}{\partial \mu_{ik}} = \sum_{j,j'=1}^{p} \omega_{jj'} \left[-\frac{2}{n} x_{ij} \frac{dm_{ik}}{d\mu_{ik}} b_{j'k} + \frac{1}{n} \left\{ 2b_{jk} \left(2m_{ik} + \sum_{i' \neq i} m_{i'k} \right) \frac{dm_{ik}}{d\mu_{ik}} b_{j'k} \right\} + \textbf{tbd} \right]$$

Using chain rule, we get the derivatives with respect to the component vectors:

$$\frac{\partial \mathcal{F}}{\partial \mu} = \mathbf{b}^T \frac{\partial \mathcal{F}}{\partial (\mu \mathbf{b})}; \quad \frac{\partial \mathcal{F}}{\partial \sigma} = \textbf{tbd}; \quad \frac{\partial \mathcal{F}}{\partial b_k} = \mathbf{a}^T \frac{\partial \mathcal{F}}{\partial (b_k \mathbf{a})}.$$

M-step: First generate data $\mathbf{H}_\varphi$ using the variational parameters $(\mathbf{M}, \mathbf{S})$. Then obtain $\mathbf{B}, \Omega_x$ by solving a penalized LS problem:

$$\{\hat{\mathbf{B}}, \hat{\Omega}_x\} = \underset{\mathbf{B}, \Omega_x}{\arg\min}\, \text{Tr}(\mathbf{S}_x^\varphi \Omega_x) + \log\det \Omega_x + \|\mathbf{B}\|_1 + \|\Omega_x\|_{\text{off},1}.$$

2 Theoretical properties

Define equivalence classes, $\boldsymbol{\theta} = \text{vec}(\mathbf{B}, \Omega_{x,off})$, $\boldsymbol{\vartheta}$ denoting the variational parameters, $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\vartheta})$. Then we are minimizing

$$\mathbb{E}_q \left[l(\mathbf{x}; \mathbf{z}, \boldsymbol{\eta}) + \text{KL}(q(\mathbf{z}|\boldsymbol{\vartheta}_1) \| p(\mathbf{z})) + \text{KL}(r(\boldsymbol{\vartheta}_1 | \mathbf{z}; \boldsymbol{\vartheta}) \| q(\boldsymbol{\vartheta}_1; \boldsymbol{\vartheta})) \right] + P(\boldsymbol{\theta}).$$

define the negative hierarchical ELBO by $\bar{l}(\cdot)$. We consider a ℓ_1-penalty

$$P(\boldsymbol{\theta}) = \rho_1 \|\boldsymbol{\beta}\|_1 + \rho_2 \|\boldsymbol{\omega}\|_1 = \lambda P_\alpha(\boldsymbol{\theta})$$

by reparameterizing the penalties: $\lambda = \rho_1 + \rho_2, \alpha = \rho_1/\lambda$.

Conditions 1, 2, 3 same as those in SPINN paper.

Define $V_n(\boldsymbol{\eta}) = \mathbb{E}\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{X}; \boldsymbol{\eta})$, $\mathcal{E}(\boldsymbol{\eta}|\boldsymbol{\eta}_0), \bar{\mathcal{E}}(\boldsymbol{\eta}|\boldsymbol{\eta}_0)$ as in Städler et al. (2010).

Theorem 2.1. *Define the event*

$$\mathcal{T} = \left\{ \sup_{\boldsymbol{\eta}} \frac{|V_n(\boldsymbol{\eta}_0^{\boldsymbol{\eta}}) - V_n(\boldsymbol{\eta})|}{\lambda_0 \vee (P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0^{\boldsymbol{\eta}}) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0^{\boldsymbol{\eta}}\|_2)} \leq T\lambda_0 \right\}$$

for $T \geq 1, \lambda_0 > 0$. Then for the solution $\hat{\boldsymbol{\eta}}$ defined in **tbd**, *we have*

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \frac{\lambda - 2T\lambda_0}{2} \|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq \left[(\lambda + 2T\lambda_0)(\alpha\sqrt{s_\beta} + (1-\alpha)\sqrt{s_\omega}) C_0 \right]^2$$

Proof of Theorem 2.1. Just prove an equivalent lemma of Städler et al. (2010). Details **tbd**.

Other details similar to Thm 1 of Städler et al. (2010).

By definition we now have that

$$\bar{l}(\mathbf{X}; \hat{\boldsymbol{\eta}}) + \lambda P_\alpha(\hat{\boldsymbol{\theta}}) \leq \bar{l}(\mathbf{X}; \boldsymbol{\eta}_0) + \lambda P_\alpha(\boldsymbol{\theta}_0)$$

for any $\boldsymbol{\eta}_0 \in \mathcal{Q}_0$. Adding $\mathcal{E}(\hat{\boldsymbol{\eta}}) = \mathbb{E}\bar{l}(\mathbf{x}; \hat{\boldsymbol{\eta}}) - \mathbb{E}\bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)$ on both sides, we get

$$\begin{aligned} \mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda P_\alpha(\hat{\boldsymbol{\theta}}) &\leq |V_n(\boldsymbol{\eta}_0) - V_n(\hat{\boldsymbol{\eta}})| + \lambda P_\alpha(\boldsymbol{\theta}_0) \\ &\leq T\lambda_0 \left(\lambda_0 \vee (P_\alpha(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2) \right) + \lambda P_\alpha(\boldsymbol{\theta}_0) \qquad (2.1) \end{aligned}$$

on the set $\mathcal{T}$. There are three cases now.

Case I. Suppose $\lambda_0 \geq P_\alpha(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2$. Then rearranging the terms in (2.1) we have

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda P_\alpha(\hat{\boldsymbol{\theta}}_{S^c}) \leq T\lambda_0^2 + \lambda P_\alpha(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}) \leq T\lambda_0^2 + \lambda\lambda_0$$

since $\lambda_0 \geq P_\alpha(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S})$. □

Case II. Suppose $\lambda_0 < P_\alpha(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2$. Then after some rearrangement we get

$$\begin{aligned}\mathcal{E}(\hat{\boldsymbol{\eta}}) + (\lambda - T\lambda_0)P_\alpha(\hat{\boldsymbol{\theta}}_{S^c}) &\leq T\lambda_0\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2 + T\lambda_0 P_\alpha(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}) + \lambda(P_\alpha(\boldsymbol{\theta}_{0,S}) - P_\alpha(\hat{\boldsymbol{\theta}}_S)) \\ &\leq T\lambda_0\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2 + (\lambda + T\lambda_0)P_\alpha(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S})\end{aligned}$$

Condition 4. The gradient of $\bar{l}(\cdot)$ with respect to the model parameters is bounded above:

$$\left\|\nabla_{\boldsymbol{\eta}}\bar{l}(\mathbf{x};\boldsymbol{\eta})\right\|_\infty \leq G(\mathbf{x})$$

for some function $G : \mathbb{R}^p \mapsto \mathbb{R}^+$. Further, there exists $c' > 0$ such that

$$|\bar{l}(\mathbf{x};\boldsymbol{\eta}) - \bar{l}(\mathbf{x},\boldsymbol{\eta}')|\mathbb{I}(G(\mathbf{x}) \leq M)) \leq c'$$

for any $M \geq 0$ and $\boldsymbol{\eta}, \boldsymbol{\eta}'$.

Theorem 2.2. *For the choice of λ_0:*

$$\lambda_0 = 2(5 + c_0)^{1/2}\frac{\log n}{\sqrt{n}}\left[\frac{\sqrt{2c_1K}}{m_\alpha} + \log\left(\frac{n\sqrt{c_1K}\log n}{m_\alpha}\right)\sqrt{\log(2d)}\right],$$

and any $T \geq 1$, the event $\mathcal{T}$ happens with probability $\geq$

$$1 - C_2\log n\exp\left[-\frac{nT^2\delta^2}{C_3^2K\log n}\right] - \frac{k_9\sqrt{K}}{n^{3/2}\log n}$$

Proof of Theorem 2.2. We follow an approach similar to Städler et al. (2010) and Feng and Simon (2017) to obtain probability bounds for truncated versions and tails of the quantity $|V_n(\boldsymbol{\eta}_0^{\boldsymbol{\eta}}) - V_n(\boldsymbol{\eta})|$ after proper scaling.

Part I: Bounding truncated parts. Define the following:

$$\bar{V}_n(\boldsymbol{\eta}) := \mathbb{E}[\bar{l}(\mathbf{x};\boldsymbol{\eta})\mathbb{I}(G(\mathbf{x}) \leq M_n)] - \frac{1}{n}\sum_{i=1}^n \bar{l}(\mathbf{x}_i;\boldsymbol{\eta})\mathbb{I}(G(\mathbf{x}_i) \leq M_n)$$

so that

$$\begin{aligned}|\bar{V}_n(\boldsymbol{\eta}) - \bar{V}_n(\boldsymbol{\eta}_0)| \leq \mathbb{E}[|\bar{l}(\mathbf{x};\boldsymbol{\eta}) - \bar{l}(\mathbf{x};\boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}) \leq M_n)] + \\ \frac{1}{n}\sum_{i=1}^n |\bar{l}(\mathbf{x}_i;\boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i;\boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}_i) \leq M_n)\end{aligned} \tag{2.2}$$

To get an upper bound on the right hand side of (2.2), we start by bounding the entropy of the functional class $\mathcal{E}_r, r > 0$:

$$\Theta_r := \{\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\vartheta}) : P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \le r\}$$
$$\mathcal{E}_r := \left\{\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)\mathbb{I}(G(\mathbf{x}) \le M_n) : \boldsymbol{\eta} \in \Theta_r\right\}$$

with respect to the empirical norm $\|h\|_{P_n} = \sqrt{\sum_{i=1}^n h^2(\mathbf{x}_i)/n}$.

Lemma 2.3. *For a collection of functions $\mathcal{H}$ taking values in $\mathcal{X}$, denote its metric entropy by $H(\cdot, \mathcal{H}, \|.\|_{P_n})$. Then for any $u, r, M_n > 0$ and some $c_0 > 0$ the following holds:*

$$H(u, \mathcal{E}_r, \|.\|_{P_n}) \le \frac{(5 + c_0) r^2 M_n^2}{u^2} \log\left(1 + \frac{du^2}{r^2 M_n^2}\right)$$

Proof of Lemma 2.3. For any $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \Theta$, due to the mean value theorem there exists $\boldsymbol{\eta}''$ so that

$$\left\|\nabla_{\boldsymbol{\eta}} \bar{l}(\mathbf{x}; \boldsymbol{\eta})|_{\boldsymbol{\eta} = \boldsymbol{\eta}''}\right\|_\infty = \frac{|\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}')|}{\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_1}. \tag{2.3}$$

Define $e_{\boldsymbol{\eta}}(\mathbf{x}) = |\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}) \le M_n)$. Then, combining (2.3) with Condition (4) we get

$$\begin{aligned} |e_{\boldsymbol{\eta}}(\mathbf{x}) - e_{\boldsymbol{\eta}'}(\mathbf{x})| &\le |\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}')|\mathbb{I}(G(\mathbf{x}) \le M_n) \\ &\le M_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_1) \\ &\le M_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 + \sqrt{6K}\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2) \end{aligned} \tag{2.4}$$

so that for $u > 0$,

$$\begin{aligned} H(u, \mathcal{E}_r, \|.\|_{P_n}) \le H &\left(\frac{u}{M_n}, \left\{\boldsymbol{\vartheta} : \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \le \frac{r}{\sqrt{6K}}\right\}, \|.\|_2\right) + \\ &H\left(\frac{u}{M_n}, \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \le r\}, \|.\|_2\right) \end{aligned} \tag{2.5}$$

The first term is bounded above by $6K \log(5rM_n/(\sqrt{6K}u))$ (Städler et al., 2010). Also $\log x/x \le 1/e$ for any $x > 0$ implies that

$$6K \log\left(\frac{5rM_n}{\sqrt{6K}u}\right) = 3K \log\left(\frac{25r^2M_n^2}{6Ku^2}\right) \le \frac{25r^2M_n^2}{2eu^2} \le \frac{5r^2M_n^2}{u^2}$$

For the second term, we use Lemma 2.6.11 in van der Vaart and Wellner (1996) to obtain

$$H\left(\frac{u}{M_n}, \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \le r\}, \|.\|_2\right) \le \frac{c_0 r^2 M_n^2}{u^2} \log\left(1 + \frac{du^2}{r^2 M_n^2}\right)$$

for some $c_0 > 0$. Now putting everything back in (2.5) we have the needed. □

Leveraging the bound in Lemma 2.3, we now prove that a symmetrized version of the truncated empirical process is small with high probability.

Lemma 2.4. *Assume fixed* $\mathbf{X}$*, and Rademacher random variables* $W_i, i \in \mathcal{I}_n$ *(defined as* $P(W_i = 1) = P(W_i = -1) = 1/2$*). Also define*

$$\delta = (5 + c_0)^{1/2} \frac{M_n}{\sqrt{n}} \left[\frac{\sqrt{2c_1 K}}{m_\alpha} + \log\left(\frac{n\sqrt{c_1 K} M_n}{m_\alpha} \right) \sqrt{\log(2d)} \right] \tag{2.6}$$

for constants $c_0, c_1 > 0$*, and* $m_\alpha := \alpha \vee (1 - \alpha)$*. Then for any* $r > 0, T \geq 1$ *we have*

$$P\left(\sup_{\boldsymbol{\eta} \in \Theta_r} \left| \frac{1}{n} \sum_{i=1}^n W_i (\bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}_0)) \mathbb{I}(G(\mathbf{x}) \leq M_n) \right| \geq Tr\delta \right)$$

$$\leq C \exp\left[- \frac{nT^2 \delta^2 m_\alpha^2 (r^2 \vee 1)}{C_1^2 K M_n^2} \right]$$

for constants $C, C_1 > 0$ *and sample size* $n > m_\alpha (M_n \sqrt{c_1 K})^{-1}$*.*

Proof of Lemma 2.4. Continuing from the right hand side of (2.4), we have for any $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \Theta_r$,

$$\begin{aligned} |\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}')|^2 \mathbb{I}(G(\mathbf{x}) \leq M_n) &\leq 6KM_n^2 \left(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2 \right)^2 \\ &\leq \frac{6KM_n^2}{m_\alpha^2} \left(P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|_2 + \right. \\ &\qquad \left. P_\alpha(\boldsymbol{\theta}' - \boldsymbol{\theta}_0) + \|\boldsymbol{\eta}' - \boldsymbol{\eta}_0\|_2 \right)^2 \\ &\leq \frac{24KM_n^2 r^2}{m_\alpha^2} \end{aligned}$$

Now applying second part of Condition 4 we get

$$\frac{1}{n} \sum_{i=1}^n |\bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}')|^2 \mathbb{I}(G(\mathbf{x}) \leq M_n) \leq \frac{c_1 K M_n^2 (r^2 \wedge 1)}{m_\alpha^2} := R_n^2. \tag{2.7}$$

Also, say $\tilde{R}_n^2 = c_1 K M_n^2 r^2 / m_\alpha^2$. Now using Lemma 2.3 we have

$$\begin{aligned} \int_{r/n}^{R_n} H^{1/2}(u, \mathcal{E}_r, \|.\|_{P_n}) du &\leq \int_{r/n}^{\tilde{R}_n} H^{1/2}(u, \mathcal{E}_r, \|.\|_{P_n}) du \\ &\leq (5 + c_0)^{1/2} \int_{r/n}^{\tilde{R}_n} \frac{rM_n}{u} \log^{1/2}\left(1 + \frac{du^2}{r^2 M_n^2} \right) du \end{aligned} \tag{2.8}$$

There are three cases now.

Case I: If $\tilde{R}_n / rM_n < 1$, we have

$$\int_{r/n}^{\tilde{R}_n} \frac{1}{u} \log^{1/2}\left(1 + \frac{du^2}{r^2 M_n^2} \right) du \leq \sqrt{\log(2d)} \log\left(\frac{n\tilde{R}_n}{r} \right)$$

Case II: If $1 < (nM_n)^{-1}$, we have

$$\begin{aligned}\int_{r/n}^{\tilde{R}_n} \frac{1}{u} \log^{1/2}\left(1 + \frac{du^2}{r^2 M_n^2}\right) du &\leq \sqrt{\log d} \int_{r/n}^{\tilde{R}_n} \frac{du}{u} + \int_{r/n}^{\tilde{R}_n} \frac{1}{u} \log^{1/2}\left(\frac{2u^2}{r^2 M_n^2}\right) du \\ &\leq \sqrt{\log(d)} \log\left(\frac{n\tilde{R}_n}{r}\right) + \int_{r/n}^{\tilde{R}_n} \frac{1}{u} \frac{\sqrt{2}u}{rM_n} du \\ &\leq \sqrt{\log(d)} \log\left(\frac{n\tilde{R}_n}{r}\right) + \frac{\sqrt{2}\tilde{R}_n}{rM_n}\end{aligned}$$

using the fact that $\log x/x < 1/e < 1$ for any $x > 0$.

Case II: If $(nM_n)^{-1} < 1 < \tilde{R}_n/rM_n$, we have

$$\begin{aligned}\int_{r/n}^{\tilde{R}_n} \frac{1}{u} \log^{1/2}\left(1 + \frac{du^2}{r^2 M_n^2}\right) du &= \sqrt{\log d} \int_{r/n}^{\tilde{R}_n} \frac{du}{u} + \int_{r/n}^{1} \frac{1}{u} \log^{1/2}\left(1 + \frac{u^2}{r^2 M_n^2}\right) du + \\ &\int_{1}^{\tilde{R}_n} \frac{1}{u} \log^{1/2}\left(1 + \frac{u^2}{r^2 M_n^2}\right) du \\ &\leq \sqrt{\log(2d)} \log\left(\frac{n\tilde{R}_n}{r}\right) + \frac{\sqrt{2}\tilde{R}_n}{rM_n}\end{aligned}$$

Now $n > m_\alpha (M_n \sqrt{c_1 K})^{-1}$ implies $n\tilde{R}_n/r > 1$ so $\log(n\tilde{R}_n/r) > 0$. Thus we can combine all three cases to get the common upper bound:

$$\begin{aligned}\int_{r/n}^{R_n} H^{1/2}(u, \mathcal{E}_r, \|.\|_{P_n}) du &\leq (5 + c_0)^{1/2} \left[\sqrt{2}\tilde{R}_n + rM_n \log\left(\frac{n\sqrt{c_1 K} M_n}{m_\alpha}\right) \sqrt{\log(2d)}\right] \\ &= (5 + c_0)^{1/2} r M_n \left[\frac{\sqrt{2c_1 K}}{m_\alpha} + \log\left(\frac{n\sqrt{c_1 K} M_n}{m_\alpha}\right) \sqrt{\log(2d)}\right]\end{aligned} \tag{2.9}$$

Take δ as in (2.6). We now apply Lemma 3.2 in van de Geer (2000) on the functional class $\mathcal{E}_r$. Combining (2.7) and (2.9), all conditions of the lemma are satisfied. Thus we obtain

$$P\left(\sup_{\boldsymbol{\eta} \in \Theta_r} \left|\frac{1}{n} \sum_{i=1}^{n} W_i(\bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}_0)) \mathbb{I}(G(\mathbf{x}) \leq M_n)\right| \geq \delta\right) \leq C \exp\left[-\frac{n\delta^2}{C^2 R_n^2}\right]$$

by applying the lemma. Now replace δ by $Tr\delta$ and substitute for R_n^2 from right hand side of (2.7) to get the needed. □

Using (2.2) and Corollary 3.4 in van de Geer (2000), we now obtain the bound:

$$P\left(\sup_{\boldsymbol{\eta} \in \Theta_r} |\bar{V}_n(\boldsymbol{\eta}) - \bar{V}_n(\boldsymbol{\eta}_0)| \geq Tr\delta\right) \leq 5C \exp\left[-\frac{nT^2\delta^2 m_\alpha^2 (r^2 \vee 1)}{16 C_1^2 K M_n^2}\right]. \tag{2.10}$$

Finally, the following lemma bounds a scaled version of $|\bar{V}_n(\boldsymbol{\eta}) - \bar{V}_n(\boldsymbol{\eta}_0)|$ over the full parameter space Θ.

Lemma 2.5. *Let $\lambda_0 = 2\delta$. Then for any $T \geq 1$ we have*

$$P\left(\sup_{\boldsymbol{\eta}} \frac{|\bar{V}_n(\boldsymbol{\eta}_0) - \bar{V}_n(\boldsymbol{\eta})|}{\lambda_0 \vee (P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2)} \geq T\lambda_0\right) \leq C_2 \log n \exp\left[-\frac{nT^2\delta^2}{C_3^2 K M_n^2}\right].$$

Proof of Lemma 2.5. The proof follows using a peeling argument similar to the proof of Lemma 4 in Feng and Simon (2017) and Lemma 2 in Städler et al. (2010). We leave the details to the reader. □

Pat II: Bounding the tails. Using taylor expansion, we have

$$\begin{aligned}
|\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}) > M_n) &\leq G(\mathbf{x})\mathbb{I}(G(\mathbf{x}) > M_n)(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_1 + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_1) \\
&\leq \frac{\sqrt{6K}}{m_\alpha} G(\mathbf{x})\mathbb{I}(G(\mathbf{x}) > M_n)(P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2) \\
\Rightarrow \frac{|\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}) > M_n)}{P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2} &\leq \frac{\sqrt{6K}}{m_\alpha} G(\mathbf{x})\mathbb{I}(G(\mathbf{x}) > M_n)
\end{aligned}$$

Now since $\mathbf{x} = \varphi(\mathbf{z})^T\mathbf{B} + \boldsymbol{\epsilon}$ and $\varphi(\cdot)$ is bounded, we have the bound

$$G(\mathbf{x}) \leq k_1(\|\boldsymbol{\epsilon}\| + k_2); \quad k_1, k_2 > 0,$$

so that

$$\begin{aligned}
&P\left(\sup_{\boldsymbol{\eta}} \frac{|V_n(\boldsymbol{\eta}_0) - V_n(\boldsymbol{\eta})|\mathbb{I}(G(\mathbf{x}) > M_n)}{\lambda_0 \vee (P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2)} \geq T\lambda_0\right) \\
&\leq P\left(\frac{\sqrt{K}}{n}\sum_{i=1}^n k_3(\|\boldsymbol{\epsilon}_i\| + k_4)\mathbb{I}(\|\boldsymbol{\epsilon}_i\| > M_n - k_5) + \mathbb{E}[k_3(\|\boldsymbol{\epsilon}\| + k_4)\mathbb{I}(\|\boldsymbol{\epsilon}\| > M_n - k_5)] \geq T\lambda_0\right) \\
&\leq \frac{(\sqrt{K}+1)\mathbb{E}[k_3(\|\boldsymbol{\epsilon}\| + k_4)\mathbb{I}(\|\boldsymbol{\epsilon}\| > M_n - k_5)]}{T\lambda_0} \qquad (2.11)
\end{aligned}$$

using Markov inequality.

Since $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Omega_x^{-1})$, i.e. sub-gaussian, we have the following tail bound:

$$\mathbb{E}[k_3(\|\boldsymbol{\epsilon}\| + k_4)\mathbb{I}(\|\boldsymbol{\epsilon}\| > M_n - k_5)] \leq k_6 \exp(-k_7 M_n^2),$$

using constants $k_6, k_7 > 0$ depending on Ω_x only. Now take $M_n = \log n$. Using (2.6) and $\lambda_0 = 2\delta$ it is easy to see that $\lambda_0 \succeq \log n/\sqrt{n}$. Putting everything back in (2.11) we thus have a probability bound on the tail of the empirical process:

$$P\left(\sup_{\boldsymbol{\eta}} \frac{|V_n(\boldsymbol{\eta}_0) - V_n(\boldsymbol{\eta})|\mathbb{I}(G(\mathbf{x}) > M_n)}{\lambda_0 \vee (P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2)} \geq T\lambda_0\right) \leq \frac{k_8(\sqrt{K}+1)\sqrt{n}}{n^2 \log n} \leq \frac{k_9\sqrt{K}}{n^{3/2}\log n}.$$

Combining the conclusions of parts I and II and taking $M_n = \log n$, theorem 2.2 is now immediate. □

References

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy Layer-Wise Training of Deep Networks. In *Advances in Neural Information Processing Systems 19 (NIPS06)*, pages 153–160. MIT Press.

Csiszár, I. and Shields, P. (2004). *Information Theory and Statistics: A Tutorial*. Now Publishers Inc.

Feng, J. and Simon, N. (2017). Sparse-Input Neural Networks for High-dimensional Nonparametric Regression and Classification. https://arxiv.org/abs/1711.07592.

Frey, B. J. and Hinton, G. E. (1999). Variational learning in nonlinear Gaussian belief networks. *Neural Comput.*, 11(1):193–213.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *Test*, 19:209–256.

van de Geer, S. (2000). *Applications of Empirical Process Theory*. Cambridge University Press.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer, Berlin.