

# Title

**Abstract:**

**Keywords:**

## 1 Formulation

Consider a random variable  $\mathbf{X} \in \mathbb{R}^p$  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:

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

with  $L$  hidden layers, and  $\varphi(\cdot)$  being a pointwise known transformation (e.g. ReLU, sigmoid, tanh). When  $\Sigma_x$  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  $\Sigma_x$  non-diagonal (but sparse), while others diagonal.

The negative log-likelihood function is

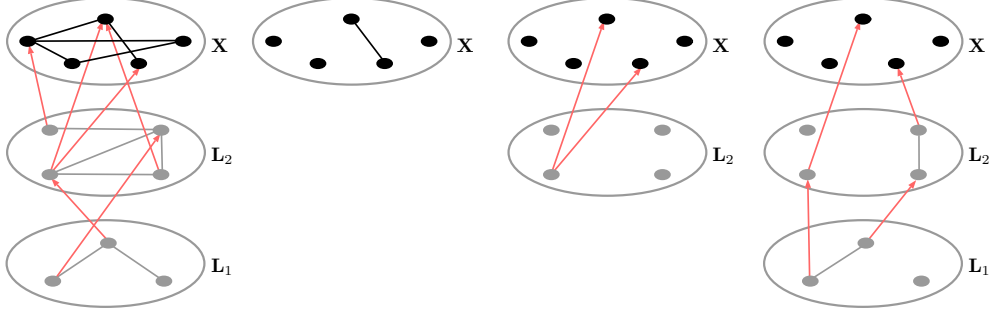
$$-\ell(\mathbf{X}|\mathcal{H}, \mathcal{B}, \Omega) = \frac{n}{2} \left[ \text{Tr}(\mathbf{S}_x \Omega_x) - \log \det \Omega_x + \sum_{l=1}^L \{ \text{Tr}(\mathbf{S}_l \Omega_l) - \log \det \Omega_l \} \right]$$

where  $\mathbf{S}_x = \mathbf{E}_x^T \mathbf{E}_x / n$ ,  $\mathbf{S}_l = \mathbf{F}_l^T \mathbf{F}_l / n$  for  $l = 1, \dots, L-1$  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  $\mathcal{M} := \{\mathbf{M}_1, \dots, \mathbf{M}_L\}$ ,  $\mathcal{S} := \{\mathbf{S}_1, \dots, \mathbf{S}_L\}$ . Now we have the variational lower-bound

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



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

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

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)

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

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

## 1.1 Objective function

Our goal is to solve a penalized version of the variational lower bound in (1.1):

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

We take the greedy strategy of solving two-layer problems successively. This means monte-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 convergence is reached, go to the next layer. Similar to Bengio et al. (2007); Hinton and Salakhutdinov (2006).

The objective function for a two-layer model, with the top layer being the observed data  $\mathcal{X}$ , is:

$$\text{Tr} \left[ \frac{1}{n} (\mathbf{X} - \varphi(\mathbf{H})\mathbf{B})^T (\mathbf{X} - \varphi(\mathbf{H})\mathbf{B}) \Omega_x \right] + \log \det \Omega_x + \lambda_n \|\mathbf{B}\|_1 + \gamma_n \|\Omega_x\|_{1,\text{off}}.$$

When  $\varphi$  is identity, this becomes a sparse factor model (refs). However for non-linear  $\varphi$  optimizing this becomes difficult. To tackle this, we use a variational approximation of the objective function. We assume the following hierarchical generative model for the hidden variables and associated variational parameters:

$$\begin{aligned} h_{ij} &\sim N(\mu_{ij}, \sigma_{ij}^2), \\ \mu_{ij} &\sim \sum_{k=1}^K I_{mk} N(\mu_{mk}, \sigma_{mk}^2); \quad I_{mk} = \text{Ber}(\pi_{mk}) \\ \log \sigma_{ij} &\sim \sum_{k=1}^K I_{sk} N(\mu_{sk}, \sigma_{sk}^2); \quad I_{sk} = \text{Ber}(\pi_{sk}). \end{aligned} \tag{1.4}$$

Thus in total there are  $6K$  generative parameters. Following the framework of hierarchical variational models (refs), we replace the negative loglikelihood  $-l(\mathbf{x}; \theta)$  by a hierarchical evidence lower bound (HELBO)  $-\bar{l}(\mathbf{x}; \boldsymbol{\vartheta})$ , defined as:

$$\bar{l}(\mathbf{x}; \boldsymbol{\eta}) = \mathbb{E}_2 l(\mathbf{x}, \mathbf{h}_\varphi; \boldsymbol{\eta}) + \mathbb{E}_1 \text{KL}(q(\mathbf{m}, \mathbf{s} | \mathbf{x}) \| p(\mathbf{m}, \mathbf{s})) + \text{KL}(q(\boldsymbol{\vartheta} | \mathbf{z}) \| p(\boldsymbol{\vartheta})),$$

with  $\mathbb{E}_1$  and  $\mathbb{E}_2$  denoting expectations over the distributions of  $(\mathbf{m}, \mathbf{s})$  only and both the latent data and variational parameter distributions, respectively. Our desired set of parameter estimates (variational as well as generative) are defined as the minimizer of a penalized version of this HELBO:

$$\hat{\boldsymbol{\eta}} = \arg \min_{\boldsymbol{\eta}} \left[ \frac{1}{n} \sum_{i=1}^n \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) + \lambda_n \|\mathbf{B}\|_1 + \gamma_n \|\Omega_x\|_{1,\text{off}} \right]. \tag{1.5}$$

## 1.2 Computational algorithm

Our objective function in (1.5) is not convex in general. However, with the variational specifications in (1.4) it is bi-convex with respect to the variational and generative components of our parameter vector.

**Proposition 1.1.** *The objective function in (1.5) is bi-convex with respect to the variational parameters  $\boldsymbol{\vartheta}$  and generative parameters  $\boldsymbol{\theta}$ .*

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

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

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

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

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

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

**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:

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

## 2 Theoretical properties

### 2.1 Concentration bounds for a two-layer model

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

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

define the negative hierarchical ELBO by  $\bar{l}(\cdot)$ . We consider a  $\ell_1$ -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

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

for  $T \geq 1, \lambda_0 > 0$ . Then for the solution  $\hat{\boldsymbol{\eta}}$  defined in **tbd** that is calculated using a fixed sample  $\mathcal{X} \in \mathcal{T}$ , we have

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \frac{\lambda - 2T\lambda_0}{2} \|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq [(\lambda + 2T\lambda_0)(\alpha\sqrt{s_{\beta}} + (1 - \alpha)\sqrt{s_{\omega}})c_0]^2$$

for some constant  $c_0 > 0$ .

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

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

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  $\lambda_0$ :

$$\lambda_0 = (20 + c_1)^{1/2} \frac{\log n}{\sqrt{n}} \left[ \frac{\sqrt{2c_2K}}{m_{\alpha}} + \log \left( \frac{n\sqrt{c_2K} \log n}{m_{\alpha}} \right) \sqrt{\log(2d)} \right], \quad (2.1)$$

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

$$1 - c_3 \log n \exp \left[ -\frac{nT^2\lambda_0^2}{c_4^2 K \log n} \right] - \frac{c_5 \sqrt{K}}{n^{3/2} \log n}, \quad (2.2)$$

for some constants  $c_1, c_2, c_3, c_4, c_5 > 0$ , and sample size condition  $n \log n \geq m_{\alpha}(c_2K)^{-1/2}$ .

Using theorems **2.1** and **2.2** a concentration bound on both the excess risk and weights of irrelevant nodes is now immediate.

**Corollary 2.3.** Suppose conditions 1-4 are satisfied, and we have sample size  $n$  so that  $n \log n \geq m_{\alpha}(c_2K)^{-1/2}$ . Then for random  $\mathcal{X}$  and  $\lambda_n \geq 2\lambda_0$ , with  $\lambda_0$  defined as (2.1), we have

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \frac{\lambda_n - 2\lambda_0}{2} \|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq c_0^2 m_{\alpha}^2 (\lambda_n + 2\lambda_0)^2 s$$

with probability larger than or equal to

$$1 - c_3 \log n \exp \left[ -\frac{n\lambda_0^2}{c_4^2 K \log n} \right] - \frac{c_5 \sqrt{K}}{n^{3/2} \log n}.$$

## 2.2 Multi-layer sparse latent model

### 3 Proofs of main results

*Proof of Proposition 1.1.* Since our penalties are convex, we only need to prove the bi-convexity of the HELBO. For this we first simplify its squared error part. For  $j, j' \in \mathcal{I}_p$ , consider  $\mathbf{X}_j$  to be the  $j^{\text{th}}$  column of  $\mathbf{X}$ ,  $\mathbf{b}_j$  to be the  $j^{\text{th}}$  row of  $\mathbf{B}$ , and  $\omega_{jj'}$  to be the  $(j, j')^{\text{th}}$  element of  $\Omega_x$ .

$$\frac{1}{n} \text{Tr}[(\mathbf{X} - \mathbf{H}_\varphi \mathbf{B})(\mathbf{X} - \mathbf{H}_\varphi \mathbf{B})^T \Omega_x] = \frac{1}{n} \sum_{j, j'=1}^p \omega_{jj'} (\mathbf{X}_j - \mathbf{H}_\varphi \mathbf{b}_j)^T (\mathbf{X}_{j'} - \mathbf{H}_\varphi \mathbf{b}_{j'}) \quad (3.1)$$

Now substitute  $\mathbf{H}_\varphi = \mathbf{M} + \mathbf{S} \odot \mathbf{E}_h$ , where the elements of  $\mathbf{E}_h$  are independently drawn from  $N(0, 1)$ , and  $\odot$  denotes element-wise multiplication. Simplifying, we have

$$\begin{aligned} (\mathbf{X}_j - \mathbf{H}_\varphi \mathbf{b}_j)^T (\mathbf{X}_{j'} - \mathbf{H}_\varphi \mathbf{b}_{j'}) &= \mathbf{X}_j^T \mathbf{X}_{j'} - 2\mathbf{X}_j^T (\mathbf{M} + \mathbf{S} \odot \mathbf{E}_h) \mathbf{b}_{j'} + \\ &\quad \mathbf{b}_j^T \mathbf{M}^T \mathbf{M} \mathbf{b}_{j'} + 2\mathbf{b}_j^T \mathbf{M}^T (\mathbf{S} \odot \mathbf{E}_h) \mathbf{b}_{j'} + \mathbf{b}_j^T (\mathbf{S} \odot \mathbf{E}_h)^T (\mathbf{S} \odot \mathbf{E}_h) \mathbf{b}_{j'} \end{aligned}$$

Taking expectation over the first-level variational distribution,

$$\mathbb{E}_1[(\mathbf{X}_j - \mathbf{H}_\varphi \mathbf{b}_j)^T (\mathbf{X}_{j'} - \mathbf{H}_\varphi \mathbf{b}_{j'})] = \mathbf{X}_j^T \mathbf{X}_{j'} - 2\mathbf{X}_j^T \mathbf{M} \mathbf{b}_{j'} + \mathbf{b}_j^T \mathbf{M}^T \mathbf{M} \mathbf{b}_{j'} + \sum_{i=1}^n \mathbf{b}_j^T \mathbf{S}_i^T \mathbf{S}_i \mathbf{b}_{j'}$$

Now substitute

$$\begin{aligned} \mathbf{M} &= \sum_{k=1}^K I_{mk} (\mu_{mk} \mathbf{J}_{n \times q} + \sigma_{mk} \mathbf{J}_{n \times q} \odot \mathbf{E}_{mk}), \\ \mathbf{S} &= \exp \left[ \sum_{k=1}^K I_{sk} (\mu_{sk} \mathbf{J}_{n \times q} + \sigma_{sk} \mathbf{J}_{n \times q} \odot \mathbf{E}_{sk}) \right], \end{aligned}$$

and take expectation over the second-level variational distributions.

$$\begin{aligned} \mathbb{E}_2[(\mathbf{X}_j - \mathbf{H}_\varphi \mathbf{b}_j)^T (\mathbf{X}_{j'} - \mathbf{H}_\varphi \mathbf{b}_{j'})] &= \mathbf{X}_j^T \mathbf{X}_{j'} - 2 \sum_k \pi_{mk} \mu_{mk} \mathbf{X}_j^T \mathbf{J}_{n \times q} \mathbf{b}_{j'} \\ &\quad + n \mathbf{b}_j^T \mathbf{b}_{j'} \sum_k \pi_{mk} (\mu_{mk}^2 + \sigma_{mk}^2) \\ &\quad + n \mathbf{b}_j^T \mathbf{b}_{j'} \sum_k \pi_{sk} \exp[2(\mu_{sk} + \sigma_{sk}^2)]. \end{aligned} \quad (3.2)$$

For the last term we use the fact that if  $x \sim N(\mu, \sigma^2)$  then  $\mathbb{E}e^{tx} = e^{t\mu + t^2\sigma^2/2}$ .

The right hand side of (3.2) is quadratic in  $\mathbf{B}$ , and the quadratic terms all have positive coefficients. Thus, in conjunction with (3.1) we can easily conclude that the negative log-likelihood is convex in  $(\mathbf{B}, \Omega_x)$  for fixed  $\boldsymbol{\vartheta}$ . Other terms in the HELBO do not depend either  $\mathbf{B}$  or  $\Omega_x$ , the same holds for it as well.  $\square$

*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

$$\begin{aligned} \mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda P_\alpha(\hat{\boldsymbol{\theta}}) &\leq |V_n(\boldsymbol{\eta}_0) - V_n(\hat{\boldsymbol{\eta}})| + \lambda P_\alpha(\boldsymbol{\theta}_0) \\ &\leq T\lambda_0 \left( \lambda_0 \vee (P_\alpha(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2) \right) + \lambda P_\alpha(\boldsymbol{\theta}_0) \end{aligned} \quad (3.3)$$

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 (3.3) 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

$$\begin{aligned} \mathcal{E}(\hat{\boldsymbol{\eta}}) + (\lambda - T\lambda_0)P_\alpha(\hat{\boldsymbol{\theta}}_{S^c}) &\leq T\lambda_0\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2 + T\lambda_0P_\alpha(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}) + \lambda(P_\alpha(\boldsymbol{\theta}_{0,S}) - P_\alpha(\hat{\boldsymbol{\theta}}_S)) \\ &\leq T\lambda_0\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2 + (\lambda + T\lambda_0)P_\alpha(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}) \end{aligned}$$

□

*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  $|V_n(\boldsymbol{\eta}_0^\eta) - V_n(\boldsymbol{\eta})|$  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

$$\begin{aligned} |\bar{V}_n(\boldsymbol{\eta}) - \bar{V}_n(\boldsymbol{\eta}_0)| &\leq \mathbb{E}[|\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}) \leq M_n)] + \\ &\quad \frac{1}{n} \sum_{i=1}^n |\bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}_0)|\mathbb{I}(G(\mathbf{x}_i) \leq M_n) \end{aligned} \quad (3.4)$$

To get an upper bound on the right hand side of (3.4), we start by bounding the entropy of the functional class  $\mathcal{E}_r, r > 0$ :

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

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

**Lemma 3.1.** For a collection of functions  $\mathcal{H}$  taking values in  $\mathcal{X}$ , denote its metric entropy by  $H(\cdot, \mathcal{H}, \|\cdot\|_{P_n})$ . Then for any  $u, r, M_n > 0$  and some  $c_0 > 0$  the following holds:

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

Leveraging the bound in Lemma 3.1, we now prove that a symmetrized version of the truncated empirical process is small with high probability.

**Lemma 3.2.** 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

$$\delta = (5 + c_0)^{1/2} \frac{M_n}{\sqrt{n}} \left[ \frac{\sqrt{2c_1 K}}{m_\alpha} + \log \left( \frac{n\sqrt{c_1 K} M_n}{m_\alpha} \right) \sqrt{\log(2d)} \right] \quad (3.5)$$

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

$$\begin{aligned} P \left( \sup_{\boldsymbol{\eta} \in \Theta_r} \left| \frac{1}{n} \sum_{i=1}^n W_i (\bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}_0)) \mathbb{I}(G(\mathbf{x}) \leq M_n) \right| \geq Tr\delta \right) \\ \leq C \exp \left[ - \frac{nT^2 \delta^2 m_\alpha^2 (r^2 \vee 1)}{C_1^2 K M_n^2} \right] \end{aligned}$$

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

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

$$P \left( \sup_{\boldsymbol{\eta} \in \Theta_r} |\bar{V}_n(\boldsymbol{\eta}) - \bar{V}_n(\boldsymbol{\eta}_0)| \geq Tr\delta \right) \leq 5C \exp \left[ - \frac{nT^2 \delta^2 m_\alpha^2 (r^2 \vee 1)}{16C_1^2 K M_n^2} \right]. \quad (3.6)$$

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  $\Theta$ .

**Lemma 3.3.** Let  $\lambda_0 = 2\delta$ . Then for any  $T \geq 1$  we have

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

**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

$$\begin{aligned}
& P \left( \sup_{\boldsymbol{\eta}} \frac{|V_n(\boldsymbol{\eta}_0) - V_n(\boldsymbol{\eta})| \mathbb{I}(G(\mathbf{x}) > M_n)}{\lambda_0 \vee (P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2)} \geq T\lambda_0 \right) \\
& \leq P \left( \frac{\sqrt{K}}{n} \sum_{i=1}^n k_3(\|\boldsymbol{\epsilon}_i\| + k_4) \mathbb{I}(\|\boldsymbol{\epsilon}_i\| > M_n - k_5) + \mathbb{E}[k_3(\|\boldsymbol{\epsilon}\| + k_4) \mathbb{I}(\|\boldsymbol{\epsilon}\| > M_n - k_5)] \geq T\lambda_0 \right) \\
& \leq \frac{(\sqrt{K} + 1) \mathbb{E}[k_3(\|\boldsymbol{\epsilon}\| + k_4) \mathbb{I}(\|\boldsymbol{\epsilon}\| > M_n - k_5)]}{T\lambda_0} \tag{3.7}
\end{aligned}$$

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  $\Omega_x$  only. Now take  $M_n = \log n$ . Using (3.5) and  $\lambda_0 = 2\delta$  it is easy to see that  $\lambda_0 \geq \log n / \sqrt{n}$ . Putting everything back in (3.7) we thus have a probability bound on the tail of the empirical process:

$$P \left( \sup_{\boldsymbol{\eta}} \frac{|V_n(\boldsymbol{\eta}_0) - V_n(\boldsymbol{\eta})| \mathbb{I}(G(\mathbf{x}) > M_n)}{\lambda_0 \vee (P_\alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2)} \geq T\lambda_0 \right) \leq \frac{k_8(\sqrt{K} + 1)\sqrt{n}}{n^2 \log n} \leq \frac{k_9 \sqrt{K}}{n^{3/2} \log n}. \tag{3.8}$$

Now, we combine the conclusions of parts I and II, taking  $M_n = \log n$  to conclude the proof.  $\square$

## 4 Proofs of auxiliary lemmas

*Proof of Lemma 3.1.* For any  $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \Theta$ , due to the mean value theorem there exists  $\boldsymbol{\eta}''$  so that

$$\|\nabla_{\boldsymbol{\eta}} \bar{l}(\mathbf{x}; \boldsymbol{\eta})|_{\boldsymbol{\eta}=\boldsymbol{\eta}''}\|_{\infty} = \frac{|\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}')|}{\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_1}. \tag{4.1}$$

Define  $e_{\boldsymbol{\eta}}(\mathbf{x}) = |\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}_0)| \mathbb{I}(G(\mathbf{x}) \leq M_n)$ . Then, combining (4.1) with Condition (4) we get

$$\begin{aligned}
|e_{\boldsymbol{\eta}}(\mathbf{x}) - e_{\boldsymbol{\eta}'}(\mathbf{x})| & \leq |\bar{l}(\mathbf{x}; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}; \boldsymbol{\eta}')| \mathbb{I}(G(\mathbf{x}) \leq M_n) \\
& \leq M_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_1) \\
& \leq M_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 + \sqrt{6K}\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2) \tag{4.2}
\end{aligned}$$

so that for  $u > 0$ ,

$$\begin{aligned}
H(u, \mathcal{E}_r, \|\cdot\|_{P_n}) & \leq H \left( \frac{u}{M_n}, \left\{ \boldsymbol{\vartheta} : \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \leq \frac{r}{\sqrt{6K}} \right\}, \|\cdot\|_2 \right) + \\
& H \left( \frac{u}{M_n}, \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \leq r\}, \|\cdot\|_2 \right) \tag{4.3}
\end{aligned}$$

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

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

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

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

for some  $c_0 > 0$ . Now putting everything back in (4.3) we have the needed.  $\square$

*Proof of Lemma 3.2.* Continuing from the right hand side of (4.2), we have for any  $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \Theta_r$ ,

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

Now applying second part of Condition 4 we get

$$\frac{1}{n} \sum_{i=1}^n |\bar{l}(\mathbf{x}_i; \boldsymbol{\eta}) - \bar{l}(\mathbf{x}_i; \boldsymbol{\eta}')|^2 \mathbb{I}(G(\mathbf{x}) \leq M_n) \leq \frac{c_1 KM_n^2 (r^2 \wedge 1)}{m_\alpha^2} := R_n^2. \quad (4.4)$$

Also, say  $\tilde{R}_n^2 = c_1 KM_n^2 r^2 / m_\alpha^2$ . Now using Lemma 3.1 we have

$$\begin{aligned} \int_{r/n}^{R_n} H^{1/2}(u, \mathcal{E}_r, \|\cdot\|_{P_n}) du &\leq \int_{r/n}^{\tilde{R}_n} H^{1/2}(u, \mathcal{E}_r, \|\cdot\|_{P_n}) du \\ &\leq (5 + c_0)^{1/2} \int_{r/n}^{\tilde{R}_n} \frac{rM_n}{u} \log^{1/2}\left(1 + \frac{du^2}{r^2 M_n^2}\right) du \end{aligned} \quad (4.5)$$

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

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

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

$$\begin{aligned}
\int_{r/n}^{\tilde{R}_n} \frac{1}{u} \log^{1/2} \left( 1 + \frac{du^2}{r^2 M_n^2} \right) du &= \sqrt{\log d} \int_{r/n}^{\tilde{R}_n} \frac{du}{u} + \int_{r/n}^1 \frac{1}{u} \log^{1/2} \left( 1 + \frac{u^2}{r^2 M_n^2} \right) du + \\
&\quad \int_1^{\tilde{R}_n} \frac{1}{u} \log^{1/2} \left( 1 + \frac{u^2}{r^2 M_n^2} \right) du \\
&\leq \sqrt{\log(2d)} \log \left( \frac{n\tilde{R}_n}{r} \right) + \frac{\sqrt{2}\tilde{R}_n}{r M_n}
\end{aligned}$$

Now  $n > m_\alpha(M_n\sqrt{c_1\bar{K}})^{-1}$  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:

$$\begin{aligned}
\int_{r/n}^{R_n} H^{1/2}(u, \mathcal{E}_r, \|\cdot\|_{P_n}) du &\leq (5 + c_0)^{1/2} \left[ \sqrt{2}\tilde{R}_n + rM_n \log \left( \frac{n\sqrt{c_1\bar{K}}M_n}{m_\alpha} \right) \sqrt{\log(2d)} \right] \\
&= (5 + c_0)^{1/2} rM_n \left[ \frac{\sqrt{2c_1\bar{K}}}{m_\alpha} + \log \left( \frac{n\sqrt{c_1\bar{K}}M_n}{m_\alpha} \right) \sqrt{\log(2d)} \right]
\end{aligned} \tag{4.6}$$

Take  $\delta$  as in (3.5). We now apply Lemma 3.2 in [van de Geer \(2000\)](#) on the functional class  $\mathcal{E}_r$ . Combining (4.4) and (4.6), 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  $\delta$  by  $Tr\delta$  and substitute for  $R_n^2$  from right hand side of (4.4) to get the needed.  $\square$

*Proof of Lemma 3.3.* 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.  $\square$

## 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 Non-parametric 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).  $\ell_1$ -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.