# Generalization Analysis of Deep Non-linear Matrix Completion

**Antoine Ledent** [1]
**Rodrigo Alves** [2]

## Abstract

We provide generalization bounds for matrix completion with Schatten $p$ quasi-norm constraints, which is equivalent to deep matrix factorization with Frobenius constraints. In the uniform sampling regime, the sample complexity scales like $\widetilde{O}(rn)$ where $n$ is the size of the matrix and $r$ is a constraint of the same order as the ground truth rank in the isotropic case. In the distribution-free setting, the bounds scale as $\widetilde{O}\left(r^{1-\frac{p}{2}}n^{1+\frac{p}{2}}\right)$, which reduces to the familiar $\sqrt{r}n^{\frac{3}{2}}$ for $p=1$. Furthermore, we provide an analogue of the weighted trace norm for this setting which brings the sample complexity down to $\widetilde{O}(nr)$ in all cases. We then present a non-linear model, Functionally Rescaled Matrix Completion (FRMC) which applies a single trainable function from $\mathbb{R}\to\mathbb{R}$ to each entry of a latent matrix, and prove that this adds only negligible terms of the overall sample complexity, whilst experiments demonstrate that this simple model improvement already leads to significant gains on real data. We also provide extensions of our results to various neural architectures, thereby providing the first comprehensive uniform convergence PAC analysis of neural network matrix completion.

## 1. Introduction

Matrix Completion (MC), the problem which consists in estimating a ground truth matrix $\mathrm{G}\in\mathbb{R}^{m\times n}$ from a small number $N\ll mn$ of observations, is an important machine learning problem with applications in various fields such as recommender systems (Mazumder et al., 2010; Hastie et al., 2015; Zhang et al., 2018; Koren et al., 2009), community discovery (Qiaosheng et al., 2019) and drug interaction prediction (Li et al., 2015). To recover a ground truth matrix based on a small number of observations, it is necessary to assume that it has some structure. Accordingly, there are a wide set of constraints and regularizers which aim to indirectly induce *rank sparsity*. One of the most well-known examples is the nuclear norm $\|\cdot\|_*$, which is defined as the sum of the singular values (Mazumder et al., 2010).

The *Schatten $p$ quasi-norm* (for $p<1$) provides an alternative form of rank sparsity inducing constraint. The Schatten $p$ quasi-norm $\|Z\|_{\mathrm{sc},p}$ of a matrix $Z$ is defined as $\left[\sum_v \rho_v^p\right]^{\frac{1}{p}}$, where the $\rho_v$s are the singular values of $Z$. In particular, when $p$ approaches 0, $\|Z\|_{\mathrm{sc},p}^p=\sum_v\rho_v^p$ approaches the rank of $Z$. When $p=\frac{2}{d}$ for some integer $d$, this is known to be equivalent to the popular *deep matrix factorization* (DMF) framework (De Handschutter et al., 2021; Arora et al., 2019; Fan & Cheng, 2018), whose predictors take the form of a product of matrices $AD_1\dots D_{d-2}B^\top$, with a regularizer of the form $\mathcal{L}(A,D,B):=\sum\|D_v\|_{\mathrm{Fr}}^2+\|A\|_{\mathrm{Fr}}^2+\|B\|_{\mathrm{Fr}}^2$. Indeed, the minimum

$$\min\mathcal{L}(A,D,B)\qquad\text{s.t. } AD_1\dots D_{d-2}B^\top=Z \tag{1}$$

is $d\|Z\|_{\mathrm{sc},p}^p$ (see (Dai et al., 2021), or Theorem F.22). This equivalence with Schatten $p$ quasi-norm constrained MC is gathering substantial interest in recent years (Arora et al., 2019; Giampouras et al., 2020), and implications for sample complexity are not fully explored. Indeed, the early literature on deep matrix factorization is mostly concerned with algorithmic and optimization issues. It is also worth noting the equivalence has intriguing implications *beyond* matrix completion, to the study of the implicit regularization of depth in neural networks, which seen explosion of recent interest in the community (Jacot, 2022; Wang & Jacot, 2023).

The last few years have also witnessed a surge in the popularity of *non-linear* matrix completion models. For instance, a branch of the deep matrix factorization literature incorporates non-linear functions inside the product (Xue et al., 2017; Fan & Cheng, 2018; Fan, 2021), leading to predictors of the form $g_0(Ag_1(D_1\dots g_{d-2}(D_{d-2}g_{d-1}(B^\top))\dots)$, where the $g$s are activation functions. Moreover, many models are simply neural network architectures which take a (row, column) combination as input. Such models typically incorporate *learnable* row and column embeddings. This

[1]School of Computing and Information Sciences, Singapore Management University, Singapore [2]Department of Applied Mathematics, Czech Technical University in Prague, Prague, Czech Republic. Correspondence to: Antoine Ledent <aledent@smu.edu.sg>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

idea appears to date back to the *Neural Network Matrix Factorisation* model (NNMF) (Dziugaite & Roy, 2015). A relevant variant is (He et al., 2017), which involves a concatenation of Hadamard products of user and item embeddings and neural encodings followed by a linear layer.

Whilst existing research provides many new algorithms and insights into the optimization landscape of various DMF and NNMF methods, very few provide a *sample complexity analysis* of the associated function classes: to the best of our knowledge, most of the existing work in this direction is limited to MC without non-linearities and with $p = 1$ (Shamir & Shalev-Shwartz, 2011), with the exception of (Fan et al., 2020). In this paper, we study sample complexity of DMF with *and* without non-linear components. Our programme is to study a broad class of predictors: $g(i,j) = \phi(Z^1_{i,j}, \dots, Z^{\mathrm{m}}_{i,j}, \Psi(i,j))$, where $\phi$ and $\Psi$ are neural networks and the matrices $Z^1, \dots, Z^{\mathrm{m}}$ may be subject to various Schatten type constraints. We include a large variety of results for many such models in the supplementary material, but for the sake of simplicity, here, we focus our exposition on the following two much simpler cases: (1)**MC with Schatten $p$ Constraints:** $g_{i,j} = Z_{i,j}$, subject to $\|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M}$ for some constant $\mathcal{M}$; and (2) **Functionally Rescaled Matrix Completion (FRMC):** $g_{i,j} = f_\theta(Z_{i,j})$ subject to $\|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M}$, $\|Z\|_\infty \leqslant \mathcal{B}_0$ and $\|f_\theta\|_{\mathrm{lip}} \leqslant \mathrm{L_f}$ for some constants $\mathcal{M}$ and $\mathrm{L_f}$.

In addition, inspired by earlier work on the *weighted trace norm* (Foygel et al., 2011; Srebro & Salakhutdinov, 2010), we study alternative constraints based on the following weighted version of the Schatten quasi-norm: $\|\operatorname{diag}(\check{p})^{\frac{1}{2}} Z \operatorname{diag}(\check{q})^{\frac{1}{2}}\|_{\mathrm{sc},p}$, where the vectors $\check{p}, \check{q}$ are estimates of the marginal row and column probabilities. Throughout the paper, we use abbreviations such as FSd ("functionally rescaled Schatten-d") for the model with Schatten $\frac{2}{d}$ constraint and a rescaling function $f$, and other similar acronyms, which we summarize in the table of notation in Section A. Our contributions are as follows:

- For MC with a Schatten quasi-norm constraint in the uniform sampling regime, we show sample complexity bounds of $\widetilde{O}\left((m+n)r\right)$ where $r = \left[\frac{\mathcal{M}}{\sqrt{mn}}\right]^{\frac{2p}{2-p}}$ scales like the rank of the ground truth.
- In the distribution-free setting, we show a sample complexity bound of $\widetilde{O}\left(r^{1-\frac{1}{p}}(m+n)^{1+\frac{1}{p}}\right)$. This reduces to the classic rate of $\widetilde{O}\left(\sqrt{r}(m+n)^{\frac{3}{2}}\right)$ (cf. (Shamir & Shalev-Shwartz, 2011; 2014)) for $p = 1$.
- By considering the weighted version of the Schatten quasi-norm, we are able to bring the rate back to $\widetilde{O}\left(r(m+n)\right)$, analogously to the case $p = 1$ in (Foygel et al., 2011).
- As can be seen in Table 1 the Functionally Rescaled model in all of the cases above, we show that learning the function only brings a negligible cost to the sample complexity (it merely adds a constant which depends on the Lipschitz and boundedness parameters).
- We provide extensions of our results to the case of multiple latent matrices and neural encodings in the appendix. Cf. Subsection C.3 and Section G. In particular, some of our results apply to (Dziugaite & Roy, 2015; He et al., 2017), which we show (cf. Sec F.6) involve implicit Schatten $2/3$ regularization.
- Our proofs rely on low-level modifications of chaining arguments which may be of independent interest. In particular, we prove "multi-class chaining" Lemmas E.4 and E.3, which allow one to bound the Rademacher complexity of combinations of function classes without access to covering numbers for each individual class.
- In extensive synthetic and real life experiments, we evaluate the effects of the depth parameter $d$, the presence or absence of weights in the norm constraints, and the presence or absence of additional neural embeddings. We find that $p = \frac{2}{3}$ generally performs significantly better than $p = 1$, and our proposed weighted Schatten norm is slightly superior to their non-weighted counterparts.

Our results and the comparison to the related works can also be seen in Tables 1, 4 and 5. In all our results, we assume a bounded and Lipschitz loss function.

## 2. Related Works

**Approximate Recovery in Matrix Completion:** There is a substantial body of literature on the sample complexity of matrix completion with bounded Lipschitz losses and norm constraints. In particular, our work takes much inspiration from the pioneering works of (Foygel et al., 2011) and (Shamir & Shalev-Shwartz, 2011; 2014), which proved some particular cases of some of our results for MC, without a learnable function, in the case $p = 1$. The explicitly rank-restricted case was studied in classification settings in (Srebro et al., 2004; Srebro & Shraibman, 2005; Srebro & Jaakkola, 2005). Table 4 positions our work within the approximate recovery literature.

Beyond the examples above, we are not aware of any work on the approximate recovery for Schatten norm constrained matrix completion. However, similar problems have been studied with different losses or sampling regimes. In particular, (Fan et al., 2020; Fan, 2021) study approximate *tensor* recovery with **Schatten regularization** over the space $\mathcal{S}^{\perp}_{d,n}$ of order $d$ *tensors with orthogonal*

Table 1: Summary of our results for Functionally Rescaled Matrix completion (FRMC). The $\widetilde{O}$ notation hides logarithmic factors, including $N, m, n, r, \ell, \mathcal{B}$ the failure probability $\delta$ and the constraint on $\mathcal{B}_0$ on the maximum entry.

| $\mathbf{f(Z)}$ with | Sampling | Generalization bound | Result |
|---|---|---|---|
| $\left\|\frac{Z}{\sqrt{mn}}\right\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}$<br>$\|f\|_{\mathrm{Lip}} \leqslant \mathrm{L_f}, \|Z\|_\infty \leqslant \mathcal{B}_0$ | Uniform | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}+\sqrt{\frac{\mathcal{B}^2+\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}}{N}}\right)$ | Thm 3.4 |
| $\left\|\frac{Z}{\sqrt{mn}}\right\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}$<br>$\|f\|_{\mathrm{Lip}} \leqslant \mathrm{L_f}, \|Z\|_\infty \leqslant \mathcal{B}_0$ | Arbitrary | $\widetilde{O}\left(\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{N}}+\sqrt{\frac{\mathcal{B}^2+\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}}{N}}\right)$ | Thm 3.5 |
| $\left\|\widetilde{Z}\right\|_{\mathrm{sc},p} \leqslant r^{1-\frac{p}{2}}$<br>$\|f\|_{\mathrm{Lip}} \leqslant \mathrm{L_f}, \|Z\|_\infty \leqslant \mathcal{B}_0$ | Arbitrary | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}+\sqrt{\frac{\mathcal{B}^2+\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}}{N}}\right)$ | Thm 3.4 |

*CP factors*[1]. Since tensors are more general and generally more complex to study than matrices, the results go well beyond the more restricted setting of *matrix* completion which we study here. However, in the case of a 2-way tensor (i.e. a matrix), the results can be interpreted as a Lagrangian formulation of the empirical risk minimization problems we study. The loss function is the square loss and sampling is uniformly at random without replacement, which means the results are not quite directly comparable. Nonetheless, the achieved L2 excess risk bounds scale like $\sqrt[4]{(n^{\frac{2-2p}{2-p}}\,\mathcal{M}^{\frac{2p}{2-p}})/(Np)}$(cf. (Fan et al., 2020), Theorem 4), where $\mathcal{M}$ is an upper bound on the $\|\cdot\|_{\mathrm{sc},p}$ norm of the recovered matrix. Expressed in terms of our rank-like quantity $r$, this turns into $\sqrt[4]{(rn^{\frac{2}{2-p}})/(Np)}$. In contrast, our result is $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{(\mathcal{M}^{\frac{2p}{2-p}}\,n^{\frac{2-3p}{2-p}})/(Np)}\right)$, which translates to $\widetilde{O}\left(\sqrt{(rn)/(Np)}\right)$. Firstly, note both results scale like $\widetilde{O}(rn)$ when $p \to 0$ (though the constant blows up like $1/p$ in both cases). Secondly, our rate is uniformly tighter since $2/(2-p) > 1$. And lastly, the bound in (Fan et al., 2020) is vacuous for $p = 1$, scaling like $\widetilde{O}(rn^2)$ in that case, compared to $\widetilde{O}(rn)$ in our result.

Exact and perturbed recovery for matrix completion (and inductive matrix completion) is a very well-studied problem (Recht, 2011; Candes & Plan, 2010; Candès & Tao, 2010). In general, using nuclear norm constraints or regularization (which is equivalent to the case $p = 1$ from our study) results in a sample complexity of $\widetilde{O}(rn)$. We refer the reader to (Recht, 2011; Xu et al., 2013) for more details. There is also a substantial amount of work on other soft relaxations of the rank, such as the max norm. In particular, the early work of (Srebro & Shraibman, 2005) shows a sample complexity of $\widetilde{O}(nM^2)$, where $M$ is a constraint on the max norm. A low-noise recovery result was achieved for the max norm in the classic work of (Cai & Zhou, 2016), which was further extended in (Wang et al., 2021) to provide bounds on the *uniformly weighted* Frobenius error of the recovered matrix in the *non-uniform sampling* regime (under some approximate uniformity assumption on the sampling probabilities). For Schatten constraints with $p < 1$, there appears to be little to no existing work in the case of randomly sampled *entries*. However, there are several works on the sample complexity of *compressed sensing* for Schatten quasi-norm MC (Zhang et al., 2013; Arora et al., 2019; Liu et al., 2014; Recht et al., 2010). Nonetheless, compressed sensing is not directly comparable to matrix completion, especially in the arbitrary sampling regime we study. Cf Section I for more details.

Earlier works on **deep matrix factorization** often focus on the optimization and algorithmic aspects (Trigeorgis et al., 2016; Zhao et al., 2017) without providing sample complexity bounds, though some include non-linear components (Xue et al., 2017; Fan & Cheng, 2018; Wang et al., 2017; De Handschutter et al., 2021; Wei et al., 2020; Lara-Cabrera et al., 2020). Note that the non-linear components in those works are interspersed between each matrix in the product which implies the models are different from both our proposed FRMC and the analogous models we study.

The observation that deep matrix factorization is equivalent to Schatten norm regularization was made in other works, including (Arora et al., 2019), which studies the optimization landscape of the problem in a compressed sensing setting where the measurement matrices commute (which does not apply to indicator measurements). The implications this has on the implicit rank-restriction which occurs when training deep neural networks is currently the subject of a large amount of interest in the community (Dai et al., 2021; Jacot, 2022; Wang & Jacot, 2023). However, those works typically do not study sample complexity, perhaps it is only non trivial when the matrix is not flat, which implies a multi-*output* scenario in the neural network context. Nevertheless, the potential to generalize our results to that situation is a tantalizing direction for future work which may shed a different light on implicit rank-restriction in DNN training.

[1]This is a strict subset of the set of tensors of order $d$ when $d > 2$, but it coincides with the set of all matrices when $d = 2$.

## 3. Main Results

**Notation and Setting:** In line with much of the literature on approximate recovery in matrix completion (Shamir & Shalev-Shwartz, 2011; 2014; Foygel et al., 2011), we assume an i.i.d. sampling regime in a supervised learning setting where the input space is the set of entries $[m] \times [n]$: each sample/observation consists in a pair $(\xi, \widetilde{G})$ sampled i.i.d. from a joint distribution, where $\xi \in [m] \times [n]$, and $\widetilde{G}$ is a real number. We try to learn a model $g : [m] \times [n] \to \mathbb{R}$, whose performance is to be evaluated by a loss function $\mathrm{l}$, which can depend on $\xi$, $\widetilde{G}$ and the prediction of the model $g_\xi$. The loss function $\mathrm{l}$ is assumed to be $\ell$-Lipschitz w.r.t. the prediction $g(\xi)$ (for fixed $\xi, \widetilde{G}$), and uniformly bounded by a constant $\mathcal{B}$. For each fixed $\xi = (i, j)$, we choose $\mathrm{G}_\xi \in \arg\min(\mathrm{l}(g_\xi, \widetilde{G}_\xi, \xi))$. The resulting matrix $\mathrm{G} \in \mathbb{R}^{m \times n}$ is referred to as the *ground truth matrix*. For instance, if $\mathrm{l}(g, \widetilde{G}, \xi) = F(|g - \widetilde{G}|)$ where $F : \mathbb{R}^+ \to \mathbb{R}$ is a strictly increasing function and for each $(i, j) \in [m] \times [n]$ we have $\widetilde{G}_{(i,j)} = R_{i,j} + \zeta$ where $\zeta$ is generated via i.i.d. noise from a symmetric distribution with $\mathbb{E}(\zeta) = 0$, then $\mathrm{G} = R$. The marginal distribution over $\xi$ is a doubly stochastic matrix whose $(i, j)$th entry we denote by $p_{i,j}$. We also write $p_i$ and $q_j$ for the marginal probabilities of the $i$th row and $j$th column, respectively. Our training set $S$ consists of i.i.d. $N$ samples: $S = \left\{(\xi^1, \widetilde{G}_1), \ldots, (\xi^N, \widetilde{G}_N)\right\} \subset ([m] \times [n]) \times \mathbb{R}$. By abuse of notation, we sometimes omit the dependence of $\mathrm{l}$ on $\widetilde{G}$ and $\xi$ by writing the empirical expectation of a function $F$ as $\widehat{\mathbb{E}}(F(\xi)) = \frac{1}{N}\sum_{o=1}^N \mathrm{l}_o(F)$ instead of $\frac{1}{N}\sum_{o=1}^N \mathrm{l}(F_o, \widetilde{G}_o, \xi_o)$ and sometimes use notations such as $g_{i,j}$ and $g(i, j)$ interchangeably to denote the prediction made by predictor $g \in \mathbb{R}^{m \times n}$ for entry $(i, j)$. In addition, by further abuse of notation, we will often write $\mathbb{E}(\ell(Z))$ and $\widehat{\mathbb{E}}(\ell(Z))$ instead of the previous quantities. A table of notations 3 is available in Appendix A.

- For a predictor $g \in \mathbb{R}^{m \times n}$, the **empirical loss** is

$$\hat{\mathrm{l}}(g) := \widehat{\mathbb{E}}(\mathrm{l}(g(\xi), \widetilde{G}, \xi)) = \frac{1}{N}\sum_{o=1}^N \mathrm{l}(g_{\xi_0}, \widetilde{G}_o, \xi_o).$$

  In particular, if the ground truth matrix $\mathrm{G} \in \mathbb{R}^{m \times n}$ is observed without noise, the $N$ observations are distinct and the loss function $\mathrm{l}$ is the square loss, $\hat{\mathrm{l}}(g) =$

$$\frac{1}{N}\sum_{\substack{(i,j)\\ \in\Omega}} \mathrm{l}(g_{i,j}, \mathrm{G}_{i,j}, (i, j))) = \frac{1}{N}\sum_{\substack{(i,j)\\ \in\Omega}} |g_{i,j} - \mathrm{G}_{i,j}|^2$$

  where $\Omega \subset [m] \times [n]$ is the set of observed entries.

- The **population expected loss** is

$$\mathrm{l}(g) := \mathbb{E}(\mathrm{l}(g(\xi), \widetilde{G}, \xi)),$$

  where the expectation runs over a random joint draw of the entry $\xi \in [m] \times [n]$, and the observation $\widetilde{G}$. In particular, if the entries of the ground truth matrix $\mathrm{G}$ are observed without noise and the loss function $\mathrm{l}$ is the square loss, we have $\mathrm{l}(g) =$

$$\sum_{\substack{(i,j)\in\\ [m]\times[n]}} p_{i,j}\, \mathrm{l}(g_{i,j}, \mathrm{G}_{i,j}) = \sum_{(i,j)\in\Omega} p_{i,j}(g_{i,j} - \mathrm{G}_{i,j})^2,$$

  where $p_{i,j}$ denotes the marginal probability of sampling entry $(i, j)$. For uniform sampling, we further have $\mathrm{l}(g) = \frac{1}{mn}\|g - \mathrm{G}\|_{\mathrm{Fr}}^2$, where $\|\cdot\|_{\mathrm{Fr}}$ denotes the Frobenius norm.

- For a general predictor $g \in \mathbb{R}^{m \times n}$, the **generalization error** is $\mathrm{l}(g) - \hat{\mathrm{l}}(g)$. Given the class of matrices $\mathcal{F} \subset \mathbb{R}^{m \times n}$, the empirical risk minimizer $\hat{g} \in \mathcal{F}$ is defined by $\hat{g} \in \arg\min_{g \in \mathcal{F}}(\mathrm{l}(g))$. The **excess risk** is then

$$\mathrm{l}(\hat{g}) - \min_{g \in \mathcal{F}} \mathrm{l}(g). \tag{2}$$

**O and $\widetilde{O}$ Notations:** For simplicity, some of our results (e.g. Thms 3.4 and 3.5 and the results in the summary tables) are expressed in terms of a $\widetilde{O}$ notation which hides polylogarithmic factors in *all* variables, including $\mathcal{B}_0, \mathrm{L_f}$, the failure probability $\delta$, the constants $\ell, \mathcal{B}$ relative to the loss function, etc. Both the $O$ and $\widetilde{O}$ notations also assume that $\mathcal{B}_0, \mathcal{B} \geqslant 1$. The formal results including all polylogarithmic terms are in the appendix.

### 3.1. Excess Risk Bounds for Matrix Completion with the Schatten Quasi-norm

In this subsection, we present our results for matrix completion with Schatten quasi-norm constraints. A summary of our results is available in Appendix B.

**Notation for the Weighted Setting:** In the weighted setting, we require empirical estimates $\hat{p}_i = \frac{\sum_{o=1}^N 1_{(\xi^o)_1 = i}}{N}$ and $\hat{q}_j = \frac{\sum_{o=1}^N 1_{(\xi^o)_2 = j}}{N}$ of the quantities $p_i$ and $q_j$ respectively. Furthermore, similarly to the literature on the weighted trace norm (Foygel et al., 2011; Srebro & Salakhutdinov, 2010), we also work with the smoothed versions $\tilde{p}_i = \frac{1}{2}p_i + \frac{1}{2m}$ $\tilde{q}_j = \frac{1}{2}q_j + \frac{1}{2n}$ of the ground truth distribution, as well as the empirically evaluated analogues $\check{p}_i = \frac{1}{2}p_i + \frac{1}{2m}$ and $\check{q}_j = \frac{1}{2}q_j + \frac{1}{2n}$. By abuse of notation, we write $\mathrm{diag}(p)$ and $\mathrm{diag}(q)$ for the diagonal matrices with diagonal elements $p_1, \ldots, p_m$ and $q_1, \ldots, q_n$ respectively (and use similar notations for $\tilde{p}$ and $\tilde{q}$). For a matrix $Z$, we denote by $\widetilde{Z}$ the matrix $\mathrm{diag}(\tilde{p})^{\frac{1}{2}} Z\, \mathrm{diag}(\tilde{q})^{\frac{1}{2}}$, so that $\|\widetilde{Z}\|_*$ is the (smoothed) weighted trace norm (Foygel et al., 2011). Similarly, $\check{Z} = \mathrm{diag}(\check{p})^{\frac{1}{2}} Z\, \mathrm{diag}(\check{q})^{\frac{1}{2}}$.

**Remark on the Definition of the Rank-like Quantity $r$:** To better illustrate the implicit 'dimensional' dependence of

the bounds which arise from our norm-based constraints, we typically express our constraints on the Schatten norms of matrices in terms of the "rank-like" quantity $r = [\frac{\mathcal{M}}{\sqrt{mn}}]^{\frac{2p}{2-p}}$, where $\mathcal{M}$ is an upper bound constraint on $\|Z\|_{\mathrm{sc},p}$. In the case $p = 1$ this is a well-established convention in the literature (Foygel et al., 2011; Srebro & Salakhutdinov, 2010; Ledent et al., 2021b; Foygel et al., 2012).

Let us briefly explain the rationale behind this notation in the case of an arbitrary $p$. Suppose the entries of some matrix $Z$ are bounded above by some constant $C$: $|Z_{i,j}| \leqslant C$ (for all $i, j$). Then we have $\|Z\|_{\mathrm{Fr}}^2 = \sum |Z_{i,j}|^2 \leqslant C^2 mn$. Writing $\rho_1, \ldots, \rho_r$ for the singular values of $Z$ in decreasing order, we then have, by Holder's inequality:

$$\begin{aligned}\|Z\|_{SC,p}^p &= \sum_{o=1}^{r} \rho_o^p = \sum_{o=1}^{r} \rho_o^p .1 \\ &\leqslant \left[\sum_{o=1}^{r} [\rho_o^p]^{\frac{2}{p}}\right]^{\frac{p}{2}} \left[\sum_{o=1}^{r} 1^{\frac{1}{1-\frac{p}{2}}}\right]^{1-\frac{p}{2}} \\ &\leqslant [mnC^2]^{\frac{p}{2}} r^{1-\frac{p}{2}} = O(\sqrt{mn}^p r^{1-\frac{p}{2}}). \quad (3)\end{aligned}$$

Similarly, if we have $|Z_{i,j}| \geqslant C_0$ for all $i, j$, then $\|Z\|_{\mathrm{Fr}} \geqslant C_0\sqrt{mn}$. If the spectrum is homogeneous, i.e., $\rho_1/\rho_r := \kappa = O(1)$, then we also have $\|Z\|_{\mathrm{SC},p}^p =$

$$\begin{aligned}&\sum_{o=1}^{r} \rho_o^p \geqslant r\rho_r^p = [r\rho_r^2]^{\frac{p}{2}} r^{1-\frac{p}{2}} \\ &\geqslant \left[\sum_{o=1}^{r} \frac{\rho_o^2}{\kappa^2}\right]^{\frac{p}{2}} r^{1-\frac{p}{2}} \geqslant \kappa^{-p} \|Z\|_{\mathrm{Fr}}^{\frac{p}{2}} r^{1-\frac{p}{2}} \\ &\geqslant \kappa^{-p} r^{1-\frac{p}{2}} [C_0^2 mn]^{\frac{p}{2}} = \Omega(\sqrt{mn}^p r^{1-\frac{p}{2}}). \quad (4)\end{aligned}$$

Thus, if a matrix $Z$ has $\Omega(1)$ entries and an approximately uniform spectrum, then its Schatten quasi-norm is $\Omega(\sqrt{mn} r^{\frac{2-p}{2p}})$, which justifies that notation: enforcing the constraint $\left[\frac{\|Z\|_{\mathrm{sc},p}}{\sqrt{mn}}\right]^{\frac{2p}{2-p}} \leqslant r$ can be understood as a 'soft' analogue of restricting the rank to $r$ or less with a tolerance for additional singular values of very small magnitudes. The tolerance is greater for larger values of $p$. A similar argument can be easily derived for the **weighted case** by substituting the estimates of the Frobenius norms by the following: $C_0^2 \leqslant \|\tilde{Z}\|_{\mathrm{Fr}}^2 = \sum \tilde{p}_i \tilde{q}_j |Z_{i,j}|^2 \leqslant C^2$. This leads to the conclusion that $\|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant C^p r^{1-\frac{p}{2}}$ (and in the case of a uniform spectrum, $C_0^p r^{1-\frac{p}{2}} \leqslant \|\widetilde{Z}\|_{\mathrm{sc},p}^p$. This justifies the use of the notation $r$ for constraints imposed on $\|\widetilde{Z}\|_{\mathrm{sc},p}^{\frac{2p}{2-p}}$.

We provide the **following results** in this subsection:

- A sample complexity result of $\widetilde{O}((m+n)rp^{-1})$ for matrix completion with the Schatten norm $p \leqslant 1$ weighted with the smoothed ground truth marginals (Theorem 3.1). In particular, this result applies to the unweighted Schatten norm regularized matrix completion problem in the uniform sampling regime.
- A sample complexity result of $\widetilde{O}((m+n)^{1+\frac{p}{2}} r^{1-\frac{p}{2}} p^{-1})$ for the unweighted Schatten quasi-norm regularized matrix completion problem in the distribution-free setting.
- An excess risk bound corresponding to a sample complexity of $\widetilde{O}((m+n)r)$ for the empirically weighted Schatten quasi-norm regularized problem in the distribution-free setting under the assumption that $p = \frac{2}{d}$ for some integer $d$.
- Furthermore, the factors of $p$ can be removed at the cost of an additional factor of $\log(\mathcal{B}_0)$, where $\mathcal{B}_0$ is an upper bound imposed on the entries. See also Table 5.

**Theorem 3.1** (cf. Theorems C.1 and C.2)**.** *As in the rest of this paper, assume the loss function* $\mathrm{l}$ *is* $\ell$*-Lipschitz and bounded by* $\mathcal{B}$*. Let* $p > 0$ *be a fixed Schatten index and let* $r > 0$ *be a fixed real number. Consider the class* $\mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$ *of matrices with Schatten Quasi-norm bounded by* $\sqrt{mn} r^{\frac{2-p}{2p}}$*:*

$$\mathcal{F}_{\mathrm{t}}^{\mathrm{p}} = \left\{ Z \in \mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p}^p \leqslant \sqrt{mn}^p r^{1-\frac{p}{2}} \right\}. \quad (5)$$

*If the sampling distribution over entries is uniform, then for any* $\delta > 0$*, w.p.* $\geqslant 1-\delta$ *over the draw of the training set, every matrix* $Z \in \mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$ *satisfies the generalization error bound* $\mathbb{E}\,\mathrm{l}(Z_\xi, \widetilde{G}, \xi) - \widehat{\mathbb{E}}\,\mathrm{l}(Z_\xi, \widetilde{G}, \xi) \leqslant$

$$O\left[\mathcal{B}^{\frac{2-2p}{2-p}} \ell^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{Np}} \ln \frac{r_\star mnN\ell_\star}{\delta} + \mathcal{B}\sqrt{\frac{\ln\frac{1}{\delta}}{N}}\right],$$

*where* $\ell_\star = \ell + 1$ *and* $r_\star = r + 1$*. More generally, if the sampling distribution is arbitrary with smoothed marginals* $\tilde{p}$ *and* $\tilde{q}$*, the result holds for the class* $\widetilde{\mathcal{F}}_r^p = \left\{Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}\right\}$ *where* $\widetilde{Z} = \mathrm{diag}(\tilde{p})^{\frac{1}{2}} Z\, \mathrm{diag}(\tilde{q})^{\frac{1}{2}}$.

*Furthermore, if one incorporates an enforced upper bound on all the absolute values of the entries:*

$$\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p = \left\{Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}};\ \|Z\|_\infty \leqslant \mathcal{B}_0\right\},$$

*then we have instead (w.p.* $\geqslant 1-\delta$*),* $\mathbb{E}\,\mathrm{l}(Z) - \widehat{\mathbb{E}}\,\mathrm{l}(Z) \leqslant$

$$O\left[\mathcal{B}^{\frac{2-2p}{2-p}} \ell^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{N}} \ln \frac{mnNr_\star \ell_\star \mathcal{B}_0}{\delta} + \mathcal{B}\sqrt{\frac{\ln\frac{1}{\delta}}{N}}\right].$$

See Theorems C.1 and C.2 in the appendix for a full proof.

To illustrate the implications of Theorem 3.1, let us consider the idealized situation where the ground truth matrix G is

of low-rank $\tilde{\mathrm{r}}$, the loss function $\mathrm{l}$ is the truncated square loss $\mathrm{l}(a,b,(i,j)) = \min((a-b)^2,1)$, there is no noise in the observations and the sampling distribution is uniform. By Equation (3), we have $\|\mathrm{G}\|_{\mathrm{sc},p} \leqslant \sqrt{mn}\,\tilde{\mathrm{r}}^{\frac{2-p}{2p}}$. Thus, if we solve the following empirical risk minimization problem:

$$\begin{aligned}&\text{Minimize}_Z \frac{1}{N}\sum_{(i,j)\in\Omega}\min(|Z_{i,j}-\mathrm{G}_{i,j}|^2,1)\\ &\text{subject to}\quad \|Z\|_{\mathrm{sc},p}\leqslant r^{\frac{2-p}{p}}\sqrt{mn}\end{aligned}$$

for some $\tilde{\mathrm{r}} \leqslant r \leqslant O(\tilde{\mathrm{r}})$ where the set of observed entries $\Omega$ is counted with multiplicity, Theorem 3.1 implies a high probability error bound for the (global) minimizer $\widehat{Z}$:

$$\begin{aligned}&\frac{1}{mn}\sum_{(i,j)\in[m]\times[n]}\min\left[\left|\widehat{Z}_{i,j}-\mathrm{G}_{i,j}\right|^2,1\right]\\ &\leqslant O\left[\sqrt{\frac{\tilde{\mathrm{r}}(m+n)}{Np}}\log\left(\frac{r_\star mnN\ell_\star}{\delta}\right)+\mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\right].\end{aligned}$$

This matches the same sample complexity rate of $\widetilde{O}(n\,\tilde{\mathrm{r}})$ achieved by nuclear norm regularization. However, the result is more general since it is not necessary impose $\tilde{\mathrm{r}} \leqslant r$, only $\|\mathrm{G}\|_{\mathrm{sc},p} \leqslant \sqrt{mn}r^{\frac{2-p}{2p}}$: thus, the sample complexity can adapt to the approximate low-rank ness of the ground truth as expressed through its Schatten quasi-norm.

*Sketch of proof of Theorem 3.1.* The proof uses a novel technique we refer to as "parametric interpolation", which consists in interpolating between the regimes where $p \sim 0$ and $p \sim 1$. Since we need to use the boundedness of the loss function to get a tight bound on the parametric component, the combination also requires the refined "multi-class chaining" arguments from Lemma E.3, but we leave the details to the Appendix and focus on the intuition in this proof sketch. For simplicity, we treat $\mathcal{B}$ and $\ell$ as constants and absorb all logarithmic factors of $N, m, n, r$ into $\widetilde{O}$ notation. See Theorems C.1 and C.2 (and the results they rely on, such as Theorem D.2) for details.

At the left extreme ($p \to 0$), it is known that the class of matrices whose rank is explicitly restricted to some value $r_1$ exhibits a sample complexity of $\widetilde{O}(r_1(m+n))$ (see (Srebro & Shraibman, 2005; Srebro et al., 2004) for an early discussion of a nearly identical problem where the target matrix is assumed to be in $\{-1,1\}^{m\times n}$ and the distribution is uniform, see (Vandermeulen & Ledent, 2021) for a covering number of the class of low rank matrices, see Lemma D.1). This is in line with the fact that explicitly rank-restricted matrix completion is a parametric model, leading to a sample complexity of the same order as the number of parameters, omitting logarithmic factors of the magnitude of them.

More precisely, by Lemma D.1, the sample complexity of a bounded loss class associated to the set of matrices of rank $r_1$ is $\widetilde{O}(r_1(m+n)\log(\mathcal{B}_0))$, where the $\widetilde{O}$ notation hides logarithmic factors of $n, m, N$ and $r$. Similarly, the sample complexity of matrices $Z$ satisfying $\|\widetilde{Z}\|_* \leqslant \sqrt{r_2}$ is $\widetilde{O}((m+n)r_2)$ by the more recent results of (Foygel et al., 2011).

Next, for any matrix $Z$ with $\|\widetilde{Z}\|^p_{\mathrm{sc},p} \leqslant r^{1-\frac{p}{2}}$ (and $\|Z\|_\infty \leqslant \mathcal{B}_0$), we can write $\widetilde{Z} = \widetilde{Z}_1 + \widetilde{Z}_2$ where $\widetilde{Z}_1$ is the sum of the terms in the singular value decomposition of $\widetilde{Z}$ associated with a singular value greater than $\tau$ for some threshold $\tau$. Writing $\rho_1, \rho_2, \ldots$ for the singular values of $\widetilde{Z}$, since $\|\widetilde{Z}\|^p_{\mathrm{sc},p} = \sum \rho_v^p \leqslant r^{1-\frac{p}{2}}$, by Markov's inequality, we have

$$r_1 = \mathrm{rank}(Z_1) \leqslant \frac{r^{1-\frac{p}{2}}}{\tau^p}. \tag{6}$$

Furthermore, since all the singular values of $\widetilde{Z}_2$ are bounded above by $\tau$, the nuclear norm $\|\widetilde{Z}_2\|_* = \sum_{v=r_1+1}^n \rho_v$ can be controlled as $r_2 := \|\widetilde{Z}_2\|_* =$

$$\sum_{v=r_1+1}^{n}(\rho_v)^p(\rho_v)^{1-p} \leqslant \|\widetilde{Z}_2\|^p_{\mathrm{sc},p}\,\tau^{1-p} \leqslant r^{1-\frac{p}{2}}\,\tau^{1-p}.$$

Thus, the function class $\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$ is included in the function class $\mathcal{R}_\tau + \mathcal{T}_\tau$, where

$$\begin{aligned}\mathcal{R}_\tau &= \left\{Z_1 : \mathrm{rank}(Z_1) \leqslant \frac{r^{1-\frac{p}{2}}}{\tau^p}, \|Z_1\|_* \leqslant \sqrt{mn}\,\mathcal{B}_0\right\}\\ \mathcal{T}_\tau &= \left\{Z_2 : \|\widetilde{Z}_2\|_* \leqslant r^{1-\frac{p}{2}}\,\tau^{1-p}\right\}.\end{aligned}$$

Thus by Lemma D.1 (parameter counting bound of $\mathcal{R}_\tau$) and Proposition F.5 (norm-based bound on the set of low nuclear norm matrices), together with the sample complexity of $\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$ can be upper bounded by

$$\widetilde{O}\left((m+n)\left[\frac{r^{1-\frac{p}{2}}}{\tau^p}\log(\mathcal{B}_0) + r^{2-p}\,\tau^{2-2p}\right]\right).$$

Setting the threshold as $r^{-\frac{1}{2}}$ yields a sample complexity of $\widetilde{O}\left((m+n)r\log(\mathcal{B}_0)\right)$, as expected. When no separate upper bound is enforced on the entries, we can still upper bound $\|Z\|_\infty$ by $2\sqrt{mn}\|\widetilde{Z}\|_{\mathrm{sc},p} \leqslant 2\sqrt{mn}r^{\frac{2-p}{2p}}$, which implies the additional factor $\log(\mathcal{B}_0)$ becomes $\widetilde{O}(\log(r^{\frac{2-p}{2p}})) = \widetilde{O}(\frac{1}{p})$. □

Next, we also control the sample complexity of learning with the non-weighted trace norm under arbitrary sampling.

**Theorem 3.2** (Cf. Theorem C.3)**.** *Consider the following function class for* $0 < p \leqslant 1$*:*

$$\mathcal{F}^{\mathrm{p}}_{\mathrm{t}} = \left\{Z \in \mathbb{R}^{m\times n} : \|Z\|^p_{\mathrm{sc},p} \leqslant \mathcal{M}^p = \sqrt{mn}^p r^{1-\frac{p}{2}}\right\}.$$

*W.p.* $\geqslant 1-\delta$, *every* $Z$ *in* $\mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$ *satisfies* $\mathbb{E}\,\mathrm{l}(Z) - \widehat{\mathbb{E}}\,\mathrm{l}(Z) \leqslant$

$$O\left[\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}c_1}{Np}} + \mathcal{B}\sqrt{\frac{\ln(\frac{1}{\delta})}{N}}\right] \quad (7)$$

$$= O\left[\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{\mathcal{M}^p(m+n)^{1-\frac{p}{2}}c_2}{Np}} + \mathcal{B}\sqrt{\frac{\ln(\frac{1}{\delta})}{N}}\right],$$

*where* $c_1 = \ln(mnNr_\star\,\ell_\star)$ *and* $c_2 = \ln(mnN[\mathcal{M}^p+1]\,\ell_\star)$ *with* $\ell_\star = \ell+1$ *and* $r_\star = r+1$. *Thus (fixing* $\mathcal{B},\ell$*) the sample complexity is* $\widetilde{O}(r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}/p)$. *For the class* $\mathcal{F}^p_{r,\mathcal{B}_0} :=$

$$\left\{Z\in\mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p}\leqslant\mathcal{M}=\sqrt{mn}[r]^{\frac{2-p}{2p}},\ \|Z\|_\infty\leqslant\mathcal{B}_0\right\},$$

*then we have instead (w.h.p.)* $\mathbb{E}\,\mathrm{l}(Z) - \widehat{\mathbb{E}}\,\mathrm{l}(Z) \leqslant$

$$O\left[\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}c_3}{N}} + \mathcal{B}\sqrt{\frac{\ln(\frac{1}{\delta})}{N}}\right],$$

*where* $c_3 = \ln(mnN[\mathcal{B}_0+1][\ell+1])$.

**Remark:** The above result reverts to the classic $\widetilde{O}((m+n)^{\frac{3}{2}}r^{\frac{1}{2}})$ when $p=1$. As $p\to 0$, the first bound in (7) blows up due to the factor of $p^{-\frac{1}{2}}$, whilst the second yields a complexity of $\widetilde{O}\left((m+n)\,\mathcal{M}^p\right)$, in line with the parameter counting argument.

Finally, an excess risk bound can be shown in the more realistic case where the function class restriction relies on the empirical marginals instead of the true marginals.

**Theorem 3.3** (Cf. Theorem C.4)**.** *Assume* $p=\frac{2}{d}$ *for some integer* $d$ *and that the ground truth is realizable:* $\|\,\mathrm{G}\,\|_{\mathrm{sc},p}\leqslant r^{1-\frac{p}{2}}$. *Let* $\widehat{Z}\in\arg\min_Z\left(\widehat{E}(Z):\ \|\breve{Z}\|^p_{\mathrm{sc},p}\leqslant[2r]^{1-\frac{p}{2}}\right)$. *We have the following excess risk bound w.h.p. (where* $\ell_\star=\ell+1$ *and* $r_\star=r+1$*):* $\mathbb{E}(\mathrm{l}(\widehat{Z})) - \mathbb{E}(\mathrm{l}(\mathrm{G}))\leqslant$

$$O\left[\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\ln\frac{r_\star mnN\ell_\star}{\delta} + \mathcal{B}\sqrt{\frac{\ln\frac{1}{\delta}}{N}}\right].$$

The proof relies mostly on Lemma E.5, which is a generalization of Lemma 4 in (Foygel et al., 2011) to the case $p\neq 1$. This lemma shows that for large enough $N$, the Schatten quasi-norms of $\breve{Z}$ and $\widetilde{Z}$ (for any $Z$) are within a small ratio of each other (w.h.p.). This allows us to show that the class $\breve{\mathcal{F}}^p_{2r}$ contains the ground truth with high probability. Note the constraint in our result is $\|\breve{Z}\|^p_{\mathrm{sc},p}\leqslant[2r]^{1-\frac{p}{2}}$ rather than $\|\breve{Z}\|^p_{\mathrm{sc},p}\leqslant r^{1-\frac{p}{2}}$. This is in contrast to the case $p=1$ in (Foygel et al., 2011) with the more natural constraint $\|\breve{Z}\|_*\leqslant r$. However, for practical purposes, the presence of the factor of 2 in our result is not an issue, since it merely slightly increases the cross-validation cost of the constraint parameter. Also, the result only works for $p=\frac{2}{d}$, i.e., when the optimization problem can be reformulated as

$$\widehat{Z}\in\arg\min_Z\left(\mathbb{E}(\mathrm{l}(Z)) : \breve{Z} = A\prod_{v=1}^{d-2}D_vB^\top :\right.$$
$$\left.\|A\|^2_{\mathrm{Fr}} + \|B\|^2_{\mathrm{Fr}} + \sum_{v=1}^{d-2}\|D_v\|^2_{\mathrm{Fr}}\leqslant d[2r]^{1-\frac{p}{2}}\right). \quad (8)$$

**Remark:** Interestingly, reformulating the condition as above makes the factors of $p$ disappear from the bounds: if we reformulate that condition by writing $r' = d^{\frac{2}{2-p}}r$, so that $[2r']^{1-\frac{p}{2}}$ is an upper bound on $\|A\|^2_{\mathrm{Fr}} + \|B\|^2_{\mathrm{Fr}} + \sum_{v=1}^{d-2}\|D_v\|^2_{\mathrm{Fr}}$, the final sample complexity is $\widetilde{O}((m+n)rp^{-1}) = \widetilde{O}\left((m+n)r'd^{-\frac{2}{2-p}+1}\right)$, i.e. $\widetilde{O}\left((m+n)r'd^{-\frac{1}{d-1}}\right) = \widetilde{O}\left((m+n)r'\right)$. Of course, this applies to Theorems 3.1 and 3.2 as well.

### 3.2. Generalization Bounds for FRMC

We now move on to our results on a new class of models we refer to as "Functionally Rescaled Matrix Completion" (FRMC), where the predictors take the form $f_\theta(Z)$ where $f_\theta$ is a trainable function and $Z$ is a Schatten-constrained matrix. Thus, these models can be seen as an analogue of "generalized linear models" in Matrix Completion. In a nutshell, our results show that learning the rescaling function $f$ can be done at negligible cost to function class capacity and generalization performance. Indeed, our generalization error bounds take the form of a sum of two terms, one corresponding to learning the complexity of the matrix class, and another one corresponding to the function class $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ of bounded Lipschitz functions from $[-\mathcal{B}_0,\mathcal{B}_0]\to\mathbb{R}$, which has very small function class capacity thanks to the low dimensionality (see Proposition F.12 from (von Luxburg & Bousquet, 2004) and (Tikhomirov, 1993)).

**Theorem 3.4** (Cf. Theorem C.5)**.** *Let* $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} = \left\{f : [-\mathcal{B}_0,\mathcal{B}_0]\to\mathbb{R} : \|f\|_{lip}\leqslant\mathrm{L_f};\ \|f\|_\infty\leqslant\mathcal{B}_\mathrm{f}\right\}$. *Consider the following function class for our learning algorithm:*

$$\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} := \left\{g : [m]\times[n]\to\mathbb{R} :\right.$$
$$\left.\exists f\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f},\ Z\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} :\ g(i,j)=f(Z_{i,j})\right\}.$$

*With probability greater than* $1-\delta$ *over the draw of the training set, the following holds for all* $g\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$*:*

$$\mathbb{E}(\mathrm{l}(g)) - \widehat{E}(\mathrm{l}(g)) \leqslant \widetilde{O}\left[\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right. \quad (9)$$
$$\left.\mathcal{B}^{\frac{2-2p}{2-p}}[\ell\,\mathrm{L_f}]^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}} + \sqrt{\frac{\mathcal{B}^2+\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}}{N}}\right].$$

*Furthermore, similarly to Theorem 3.3 above, an excess risk result holds for the minimization problem over the empirically weighted class $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_{2r,\mathcal{B}_0}$ (cf. Thm C.7).*

In the distribution-free unweighted setting, we have:

**Theorem 3.5** (cf. Thm C.6)**.** *Consider the function class*

$$\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \mathcal{F}^p_{r,\mathcal{B}_0} := \Bigg\{ g : [m] \times [n] \to \mathbb{R} : \exists f \in \mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f},\ Z \in \mathcal{F}^p_{r,\mathcal{B}_0} :\ g(i,j) = f(Z_{i,j}) \Bigg\}.$$

*W.p. $\geqslant 1-\delta$, for each $g \in \mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \mathcal{F}^p_{r,\mathcal{B}_0}$, one has*

$$\mathbb{E}(\mathrm{l}(g)) - \widehat{E}(\mathrm{l}(g)) \leqslant \widetilde{O}\Bigg[ \mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B}^{1-\frac{p}{2}} \left[\ell\, \mathrm{L_f}\right]^{\frac{p}{2}} \sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{N}} + \sqrt{\frac{\mathcal{B}^2 + \mathcal{B}_0\, \mathrm{L_f}\, \ell\, \mathcal{B}}{N}} \Bigg].$$

Note that in all cases above, the incorporation of a learnable function $f \in \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ only contributes an additional term of $\sqrt{\frac{\mathcal{B}^2 + \mathcal{B}_0\, \mathrm{L_f}\, \ell\, \mathcal{B}}{N}}$ to the generalization error bound: there is no dependence on the architectural or norm-based parameters such as $m, n, r$ and the function $f$ only needs to be learned once for the whole dataset. The interaction between this learning task and the learning of the low rank latent matrix does not introduce any additional challenge: the complexities of both tasks are disentanglable.

It is worth noting that the proofs are far from being a trivial combination of the proofs from Subsection 3.1 above and Proposition F.12. Indeed, no sufficiently tight *covering number bound* is available for any of the function classes discussed in Section 3.1[2], not even for $p = 1$: of course, it is possible to obtain such a cover respect to the Frobenius or $L^\infty$ norms via covering numbers for linear function classes applied to the matrices $A, B, D_1, \ldots, D_{d-2}$, but this leads to loose covering number bounds that translate to vacuous results in terms of sample complexity. For instance, the error bound (applicable to a setting analogous to uniform sampling) for higher order tensors of (Fan et al., 2020; Fan, 2021) is based on a Frobenius covering number bound, but for the case of matrices and for $p = 1$, the log covering number scales as $\widetilde{O}\left(r(m+n)^2\right)$, which is vacuous. In fact, since the Rademacher complexities involved in the proofs of the results in Section 3.1 depend subtly on the sampling distribution, it is clear that the metric used in the cover must be carefully chosen. Moreover, the Frobenius norm doesn't seem to work well, not even in the uniform sampling case.

[2]except the parametric class $\mathcal{E}_{\mathrm{r,t}}$ of matrices with explicitly restricted rank

Our proof of the results of this section relies instead on multi-class generalization of classic "chaining" arguments. More specifically, in Section E, we establish two generalizations of Dudley's Entropy Theorem, Lemma E.4 and Lemma E.3, which allow one to bound the Rademacher complexity of the function class $F(\mathcal{F}_1, \mathcal{F}_2)$, where $F$ is a fixed function and the following two conditions are satisfied: (1) A covering number is available for $F(\mathcal{F}_1, f_2)$, uniformly over any choice of $f_2 \in \mathcal{F}_2$ and (2) A Rademacher complexity complexity bound is available for $F(f_1, \mathcal{F}_2)$, uniformly over any choice of $f_1 \in \mathcal{F}_1$. Results with some similarities can be traced back to (Golowich et al., 2018) (Thm. 4) and (Ledent et al., 2021b) (Prop. A.4.).

### 3.3. Generalization Bounds with Neural Encodings

In this section, we briefly describe some of our extended results for the Sd+NN setting, which includes an additional neural network encoding. Specifically, we consider neural encodings of the form $\Psi(i,j) = f(A^0(u_i, v_j)^\top)$ where $u_i$ is the embedding for row $i$, $v_j$ is the embedding for column $j$, $f$ is the neural network given by $f(x) = \sigma_L\left(A^L \mathrm{Relu}\left(\mathrm{Relu}\left(\ldots \mathrm{Relu}\left(A^1 x\right)\ldots\right)\right)\right)$, where the matrices $\mathbb{R}^{1\times w_{l-1}} \ni A^L, \ldots, A^1$ are the weight matrices. The predictors then take the form $g_{i,j} = Z_{i,j} + \Psi_{i,j}$ where $\Psi$ is the neural encoding and $Z$ is a matrix to which (potentially weighted) Schatten $p$ quasi-norm regularization is applied. In particular, for $p = \frac{2}{3}$, the model corresponds to the one presented in (He et al., 2017).

Our results (cf. Thm C.8, Thm C.9) show that the generalization error is bounded as a sum of terms corresponding to the matrix class and the neural encoding class.

**Extension to Multiple Latent Matrices:** In the Appendix, we extend our results to the case of models of the form $\phi \circ (Z^1, Z^2, \ldots, Z^{\underline{\mathrm{m}}}, \Psi)$ where $\phi$ and $\Psi$ are trainable networks ($\Psi$ a neural encoding) and the matrices $Z^1, \ldots, Z^{\underline{\mathrm{m}}}$ are constrained via various Schatten $p$ quasi-norms.

## 4. Experiments

**Synthetic Data Experiments:** We generated synthetic square data matrices in $\mathbb{R}^{n\times n}$ with a specified rank $r$. We varied the proportion of observed entries in the generated matrices ($\%\mathrm{obs} = \mathbb{E}[N/n^2]$), with a non-uniform sampling distribution. A summary of the results is provided in Figure 1. The results demonstrate, unsurprisingly, that FRMC achieves better performance than methods which do not incorporate a non-linearity. Going deeper, we observe that with sufficiently many observations, the model is able to recover the ground truth function nearly perfectly, together with the low rank latent matrix. We provide an example of the recovered functions in Figure 2 for $\%\mathrm{obs} \in \{0.14, 0.20\}$. Moreover, we observe that the weighted version of the

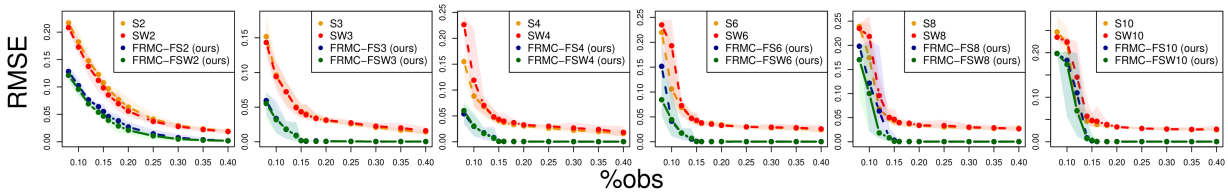


Figure 1: Summary of the results of the synthetic data experiments. Ground-truth generated by considering $f(x) = \sigma(x)$.

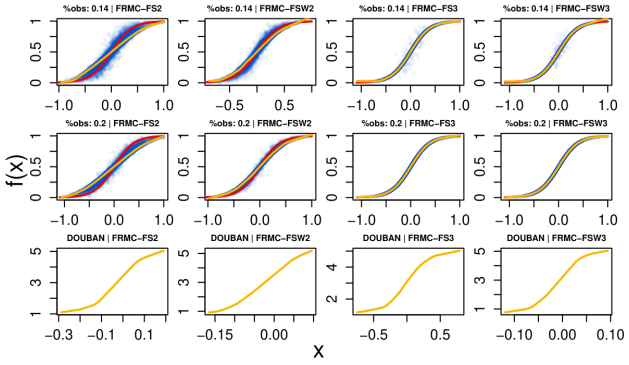


Figure 2: Learned function $f$ by our model (yellow curve). Red curves represent the ground truth. Blue dots are the predictions, which ideally should lie under the red curve.

regularizer works slightly better, especially with $d = 2$. Finally, an exponential performance improvement occurs when $d$ increases, especially from 2 to 3. This is in line with the expectation that lower values of $p$ induce more and more rank-sparsity, and with our theoretical results in the distribution-free case, which show much better sample complexities when $p$ is small (cf. our sample complexity of $\widetilde{O}\left(r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}\right)$). For additional results, such as a comparison with the identity function as ground truth, see Figure 3. For detailed information, including the generation procedure, parameter selection, and validation setup, refer to the appendix in Section H.1.

**Real Data Experiments:** MC can be applied to a wide range of domains, such as recommender systems, human event dynamics, and chemical engineering. We chose three standard datasets to evaluate our method in a real data scenario: DOUBAN and MovieLens 25M (MVL25) from the recommender systems domain, and LastFM, which stores listening habits of users in a music streaming platform. For descriptions of datasets and implementation details of the real-world strand, refer to Section H.2 in the Appendix. In Figure 2, we plot the functions learned by our model on real data. Interestingly, we see that the chosen functions look somewhat sigmoidal, probably to avoid out-of-range predictions and model the vanishing significance of increments between very high or very low ratings. Furthermore, we observe that in 2 out of 3 datasets, our mildly non-linear

Table 2: Test RMSE for the assessed methods. Notation: *Weighted* (W) models use weighted-norm regularization in the embeddings. Our methods learn a re-scaling *function* (FS). Thus, S2 refers to the traditional nuclear norm regularization, SW3 refers to weighted Schatten $2/3$ norm regularization and FRMC-FSW2 refers to the model $f(Z)$ with nuclear norm constraint on $Z$ and a trainable component-wise rescaling function $f$.

| Model | d | W | FS | Douban | LastFM | MVL25 |
|---|---|---|---|---|---|---|
| S2 | 2 | × | × | 0.8042 | 2.5885 | 0.8047 |
| SW2 | | ✓ | × | 0.7981 | 2.4980 | **0.7625** |
| FRMC-FS2 | | × | ✓ | 0.7627 | 1.0327 | 0.7795 |
| FRMC-FSW2 | | ✓ | ✓ | 0.7626 | 1.0091 | 0.7776 |
| S3 | 3 | × | × | 0.8050 | 2.0512 | 0.7786 |
| SW3 | | ✓ | × | 0.8030 | 2.0417 | 0.7876 |
| FRMC-FS3 | | × | ✓ | 0.7674 | 0.9952 | 0.7711 |
| FRMC-FSW3 | | ✓ | ✓ | **0.7616** | **0.9904** | 0.7799 |

model FRMC substantially outperforms traditional matrix completion. Furthermore, $p = 2/3$ outperforms $p = 1$ and the weighted version outperforms the unweighted version.

## 5. Conclusion

We studied matrix completion with Schatten $p$ quasi-norm constraints for $0 \leqslant p \leqslant 1$ in the approximate recovery setting. Ignoring the dependence on Lipschitz and boundedness constants, we provided sample complexity bounds of $\widetilde{O}(r(m+n))$ and $\widetilde{O}(r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}})$ in the uniform and arbitrary sampling regimes respectively. The results show the stronger rank-sparsity inducing properties of lower order Schatten $p$ quasi-norms, which we also observe in our experiments. Moreover, we showed that the use of the weighted Schatten $p$ quasi-norm can bring both rates back to $\widetilde{O}(r(m+n))$. We introduced a parsimonious non-linear model, Functionally Rescaled Matrix Completion (FRMC), which consists in applying a trainable function from $\mathbb{R} \to \mathbb{R}$ to the entries of a latent matrix. We show extensions of all of our results to the FRMC setting, which demonstrate that the addition of a learnable function from $\mathbb{R}$ to $\mathbb{R}$ negligibly increases function class capacity.

## Acknowledgements

Antoine Ledent's research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant. Rodrigo Alves thanks Recombee for supporting this research, specially while using the DGX Server.

## Impact Statement

Our work is mostly theoretical in nature, and is unlikely to have a negative societal impact.

## References

Alves, R., Ledent, A., Assunção, R., and Kloft, M. An empirical study of the discreteness prior in low-rank matrix completion. In *NeurIPS 2020 Workshop on Preregistration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*, pp. 111–125. PMLR, 11 Dec 2021.

Angluin, D. and Valiant, L. G. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and system Sciences*, 18(2):155–193, 1979.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. In Helmbold, D. and Williamson, B. (eds.), *Computational Learning Theory*, pp. 224–240, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6240–6249. Curran Associates, Inc., 2017.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. *Lecture Notes in Computer Science*, 3176: 208–240, 2004.

Cai, T. T. and Zhou, W.-X. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493 – 1525, 2016. doi: 10.1214/16-EJS1147.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

Candès, E. J. and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010.

Chen, Y. Incoherence-optimal matrix completion. *Information Theory, IEEE Transactions on*, 61, 10 2013. doi: 10.1109/TIT.2015.2415195.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Chiang, K.-Y., Hsieh, C.-J., and Dhillon, I. S. Matrix completion with noisy side information. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Chiang, K.-Y., Dhillon, I. S., and Hsieh, C.-J. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *Journal of Machine Learning Ressearch*, 2018.

Dai, Z., Karzand, M., and Srebro, N. Representation costs of linear neural networks: Analysis and design. *Advances in Neural Information Processing Systems*, 34:26884–26896, 2021.

De Handschutter, P., Gillis, N., and Siebert, X. A survey on deep matrix factorizations. *Computer Science Review*, 42:100423, 2021.

Dziugaite, G. K. and Roy, D. M. Neural network matrix factorization. *arXiv preprint arXiv:1511.06443*, 2015.

Fan, J. Multi-mode deep matrix and tensor factorization. In *international conference on learning representations*, 2021.

Fan, J. and Cheng, J. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34–41, 2018.

Fan, J., Ding, L., Yang, C., Zhang, Z., and Udell, M. Euclidean-norm-induced schatten-p quasi-norm regularization for low-rank tensor completion and tensor robust principal component analysis. *arXiv e-prints*, pp. arXiv–2012, 2020.

Foygel, R., Shamir, O., Srebro, N., and Salakhutdinov, R. R. Learning with the weighted trace-norm under arbitrary sampling distributions. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2133–2141. Curran Associates, Inc., 2011.

Foygel, R., Srebro, N., and Salakhutdinov, R. R. Matrix reconstruction with the local max norm. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Giampouras, P., Vidal, R., Rontogiannis, A., and Haeffele, B. A novel variational form of the schatten-$p$ quasi-norm. *Advances in Neural Information Processing Systems*, 33: 21453–21463, 2020.

Giné, E. and Guillou, A. On consistency of kernel density estimators for randomly censored data: Rates holding uniformly over adaptive intervals. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 37:503–522, 07 2001. doi: 10.1016/S0246-0203(01)01081-0.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Graf, F., Zeng, S., Rieck, B., Niethammer, M., and Kwitt, R. On measuring excess capacity in neural networks. *Advances in Neural Information Processing Systems*, 35: 10164–10178, 2022.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011. ISSN 1557-9654. doi: 10.1109/TIT.2011.2104999.

Guermeur, Y. Combinatorial and structural results for gamma-psi-dimensions. *arXiv preprint arXiv:1809.07310*, 2018.

Guermeur, Y. Rademacher complexity of margin multi-category classifiers. *Neural Computing and Applications*, 32(24):17995–18008, 2020.

Gui, Y., Barber, R., and Ma, C. Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36:4820–4844, 2023.

Hagerup, T. and Rüb, C. A guided tour of chernoff bounds. *Inf. Process. Lett.*, 33:305–308, 1990. URL https://api.semanticscholar.org/CorpusID:40984617.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16 (104):3367–3402, 2015. URL http://jmlr.org/papers/v16/hastie15a.html.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Jacot, A. Implicit bias of large depth networks: a notion of rank for nonlinear functions. *arXiv preprint arXiv:2209.15055*, 2022.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37, 2009.

Lara-Cabrera, R., González-Prieto, Á., and Ortega, F. Deep matrix factorization approach for collaborative filtering recommender systems. *Applied Sciences*, 10(14):4926, 2020.

Latała, R. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133 (5):1273–1282, 2005. ISSN 00029939, 10886826.

Ledent, A., Alves, R., and Kloft, M. Orthogonal inductive matrix completion. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2021a. doi: 10.1109/TNNLS.2021.3106155.

Ledent, A., Alves, R., Lei, Y., and Kloft, M. Fine-grained generalization analysis of inductive matrix completion. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25540–25552. Curran Associates, Inc., 2021b.

Ledent, A., Mustafa, W., Lei, Y., and Kloft, M. Norm-based generalisation bounds for deep multi-class convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8279–8287, May 2021c.

Ledent, A., Alves, R., Lei, Y., Guermeur, Y., and Kloft, M. Generalization bounds for inductive matrix completion in low-noise settings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8447–8455, Jun. 2023. doi: 10.1609/aaai.v37i7.26018.

Ledoux, M. and Talagrand, M. *Probability in Banach spaces : isoperimetry and processes*. Springer, Berlin [u.a.], 1991. ISBN 3540520139.

Li, R., Dong, Y., Kuang, Q., Wu, Y., Li, Y., Zhu, M., and Li, M. Inductive matrix completion for predicting adverse drug reactions (adrs) integrating drug–target interactions. *Chemometrics and Intelligent Laboratory Systems*, 144: 71 – 79, 2015. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2015.03.013.

Liu, L., Huang, W., and Chen, D.-R. Exact minimum rank approximation via schatten p-norm minimization. *Journal of Computational and Applied Mathematics*, 267: 218–227, 2014.

Long, P. M. and Sedghi, H. Size-free generalization bounds for convolutional neural networks. In *International Conference on Learning Representations*, 2020.

Ma, W. and Chen, G. H. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in neural information processing systems*, 32, 2019.

Mao, K., Zhu, J., Xiao, X., Lu, B., Wang, Z., and He, X. Ultragcn: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1253–1262, 2021.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11: 2287–2322, August 2010.

Meir, R. and Zhang, T. Generalization error bounds for bayesian mixture algorithms. *J. Mach. Learn. Res.*, 4 (null):839–860, dec 2003. ISSN 1532-4435.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 978-0-262-03940-6.

Mustafa, W., Lei, Y., Ledent, A., and Kloft, M. Fine-grained generalization analysis of structured output prediction. In Zhou, Z.-H. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2841–2847. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

Pal, S. and Jain, P. Online low rank matrix completion. In *The Eleventh International Conference on Learning Representations*, 2022.

Pal, S., Sai Suggala, A., Shanmugam, K., and Jain, P. Optimal algorithms for latent bandits with cluster structure. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7540–7577. PMLR, 25–27 Apr 2023.

Platen, E. Pollard, d.:convergence of stochastic processes. (springer series in statistics). springer-verlag, new york - berlin - heidelberg - tokyo 1984, 216 pp., 36 illustr., dm 82. *Biometrical Journal*, 28(5):644–644, 1986. doi: 10.1002/bimj.4710280516.

Qiaosheng, Zhang, Tan, V. Y. F., and Suh, C. Community Detection and Matrix Completion with Two-Sided Graph Side-Information. *arXiv e-prints*, art. arXiv:1912.04099, December 2019.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(null):3413–3430, December 2011. ISSN 1532-4435.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835.

Scott, C. Rademacher complexity. *Lecture Notes*, Statistical Learning Theory, 2014.

Shamir, O. and Shalev-Shwartz, S. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 661–678. PMLR, 2011.

Shamir, O. and Shalev-Shwartz, S. Matrix completion with the trace norm: Learning, bounding, and transducing. *Journal of Machine Learning Research*, 15:3401–3423, 2014.

Sinclair, A. *Lecture notes for the course "CS271 Randomness and computation"*. URL https://web.archive.org/web/20141031035717/http://www.cs.berkeley.edu/~sinclair/cs271/n13.pdf.

Sportisse, A., Boyer, C., and Josse, J. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020.

Srebro, N. Learning with matrix factorizations. *PhD Thesis*, 2004.

Srebro, N. and Jaakkola, T. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems 17*, pp. 5–27. MIT Press, 2005.

Srebro, N. and Salakhutdinov, R. R. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 2056–2064. Curran Associates, Inc., 2010.

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. In Auer, P. and Meir, R. (eds.), *Learning Theory*, pp. 545–560, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31892-7.

Srebro, N., Alon, N., and Jaakkola, T. Generalization error bounds for collaborative prediction with low-rank matrices. *Advances In Neural Information Processing Systems*, 17, 2004.

Talagrand, M. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994. ISSN 00911798.

Talagrand, M. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, Nov 1996. ISSN 1432-1297. doi: 10.1007/s002220050108.

Tikhomirov, V. M. *$\epsilon$-Entropy and $\epsilon$-Capacity of Sets In Functional Spaces*, pp. 86–170. Springer Netherlands, Dordrecht, 1993. ISBN 978-94-017-2973-4. doi: 10.1007/978-94-017-2973-4_7.

Trigeorgis, G., Bousmalis, K., Zafeiriou, S., and Schuller, B. W. A deep matrix factorization method for learning attribute representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):417–429, 2016.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Vandermeulen, R. A. and Ledent, A. Beyond smoothness: Incorporating low-rank analysis into nonparametric density estimation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12180–12193. Curran Associates, Inc., 2021.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

von Luxburg, U. and Bousquet, O. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5 (Jun):669–695, 2004.

Wang, J., Wong, R. K., Mao, X., and Chan, K. C. G. Matrix completion with model-free weighting. In *International Conference on Machine Learning*, pp. 10927–10936. PMLR, 2021.

Wang, Q., Sun, M., Zhan, L., Thompson, P., Ji, S., and Zhou, J. Multi-modality disease modeling via collective deep matrix factorization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1155–1164, 2017.

Wang, Z. and Jacot, A. Implicit bias of sgd in $l\_\{2\}$-regularized linear dnns: One-way jumps from high to low rank. *arXiv preprint arXiv:2305.16038*, 2023.

Wei, S., Wang, J., Yu, G., Domeniconi, C., and Zhang, X. Multi-view multiple clusterings using deep matrix factorization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6348–6355, 2020.

Wu, L., Ledent, A., Lei, Y., and Kloft, M. Fine-grained generalization analysis of vector-valued learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (12):10338–10346, May 2021. doi: 10.1609/aaai.v35i12.17238.

Xu, M., Jin, R., and Zhou, Z.-H. Speedup matrix completion with side information: Application to multi-label learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2301–2309, Red Hook, NY, USA, 2013. Curran Associates Inc.

Xue, H.-J., Dai, X., Zhang, J., Huang, S., and Chen, J. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pp. 3203–3209. Melbourne, Australia, 2017.

Zhang, M. and Chen, Y. Inductive matrix completion based on graph neural networks. In *International Conference on Learning Representations*, 2020.

Zhang, M., Huang, Z.-H., and Zhang, Y. Restricted $p$ -isometry properties of nonconvex matrix recovery. *IEEE Transactions on Information Theory*, 59(7):4316–4323, 2013. doi: 10.1109/TIT.2013.2250577.

Zhang, Q., Suh, G., Suh, C., and Tan, V. Y. F. Mc2g: An efficient algorithm for matrix completion with social and item similarity graphs. *IEEE Transactions on Signal Processing*, 70:2681–2697, 2022. doi: 10.1109/TSP.2022.3174423.

Zhang, X., Du, S., and Gu, Q. Fast and sample efficient inductive matrix completion via multi-phase procrustes flow. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5756–5765, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Zhao, H., Ding, Z., and Fu, Y. Multi-view clustering via deep matrix factorization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

# A. Table of Notations

Table 3: Table of notations for quick reference

| Notation | Meaning |
|---|---|
| **Sampling setting** | |
| $\mathrm{G} \in \mathbb{R}^{m\times n}$ | Ground truth matrix |
| $N$ | Number of samples |
| $S = \{(\xi^1, \widetilde{G}^1), \ldots, (\xi^N, \widetilde{G}^N)\}$ | Training set |
| $\zeta_o$ | Noise in $o$th observation |
| $\widetilde{G}_o = \mathrm{G}_{\xi^o} + \zeta_o$ | $o$th observation |
| $\xi \in [m] \times [n]$ (resp. $\xi^o [m] \times [n]$) | Observed entry ($o$th resp. observed entry) |
| $\mathrm{l}$ | Loss function |
| $\mathrm{l}(g_{\xi^0}, \widetilde{G}_o, \xi^o) = \mathrm{l}_o(g_{\xi^o})$ | Loss function at $o$th datapoint |
| $p_{i,j}$ | $\mathbb{P}(\xi = (i,j))$, (marginal) probability of observing entry $i,j$ |
| $p_i = \sum_j p_{i,j}$ | Row $i$ marginal probability |
| $q_j = \sum_i p_{i,j}$ | Column $j$ marginal probability |
| $\widehat{\mathbb{E}}(F(\xi, \widetilde{G})) = \frac{1}{N}\sum_{o=1}^N F(\xi^o, \widetilde{G}_o)$ | Empirical expectation of $F$ |
| $e_i \in \mathbb{R}^m$ | Indicator vector of $i$th row |
| $e_j \in \mathbb{R}^n$ | Indicator vector of $j$th column |
| **(Weighted) norms** | |
| $\|\cdot\|$ | Spectral norm |
| $\|\cdot\|_{\mathrm{Fr}}$ | Frobenius norm |
| $\|\cdot\|_*$ | Nuclear norm |
| $\|\cdot\|_{\mathrm{sc},p}$ | Schatten $p$ quasi-norm ($p \leqslant 1$) |
| $\|Z\|_{2,1}$ | $\sum_{j=1}^n \sqrt{\sum_{i=1}^m A_{i,j}^2}$ |
| $\hat{p}_i = \frac{\sum_{o=1}^N 1_{(\xi^o)_1 = i}}{N}$ | $i$th empirical row marginal |
| $\hat{q}_j = \frac{\sum_{o=1}^N 1_{(\xi^o)_2 = j}}{N}$ | $j$th empirical column marginal |
| $\tilde{p}_i = \frac{1}{2}p_i + \frac{1}{2m}$ | Smoothed row marginal |
| $\tilde{q}_j = \frac{1}{2}q_j + \frac{1}{2n}$ | Smoothed column marginal |
| $\check{p}_i = \frac{1}{2}p_i + \frac{1}{2m}$ | Smoothed empirical row marginal |
| $\check{q}_j = \frac{1}{2}q_j + \frac{1}{2n}$ | Smoothed empirical column marginal |
| $\widetilde{Z}$ | $\mathrm{diag}(\tilde{p})^{\frac{1}{2}} Z\, \mathrm{diag}(\tilde{q})^{\frac{1}{2}}$ |
| $\check{Z}$ | $\mathrm{diag}(\check{p})^{\frac{1}{2}} Z\, \mathrm{diag}(\check{q})^{\frac{1}{2}}$ |
| $\mathcal{M}$ (in Fn Class definitions) | Upper bound on $\|\cdot\|_{\mathrm{sc},p}$ |
| $r$ (in Fn Class definitions) | Upper bound on $\|\widetilde{Z}\|_{\mathrm{sc},p}^{\frac{2p}{2-p}}$, $\|\check{Z}\|_{\mathrm{sc},p}^{\frac{2p}{2-p}}$ or $\|Z\|_{\mathrm{sc},p}^{\frac{2p}{2-p}} \sqrt{mn}^{-\frac{2p}{2-p}}$ |
| $d$ | Depth of deep matrix factorization $A\prod D_i B^\top$ ($p = \frac{2}{d}$) |
| (also $d$) | width of 1st layer after embedding in $\Psi \in \mathcal{N}_{1,\mathrm{W}}$ |
| **Definitions of $r$** | |
| Unweighted Setting | $r = \left[\frac{\mathcal{M}}{\sqrt{mn}}\right]^{\frac{2p}{2-p}}$ |
| Weighted Setting | $\mathcal{M}^p = r^{1-\frac{p}{2}}\sqrt{mn}^p$<br>$\|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}$<br>$r \geqslant \|\widetilde{Z}\|_{\mathrm{sc},p}^{\frac{2p}{2-p}}$ |
| **Matrix Function Classes** | |
| $\mathcal{E}_{\mathrm{r,t}}$ | $\{R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant t,\ \mathrm{rank}(R) \leqslant r\}$ |
| $\widetilde{\mathcal{F}}_r^1$ | $\left\{\mathbb{R}^{m\times n} \ni Z : \|\widetilde{Z}\|_* \leqslant \sqrt{r}\right\}$ |
| $\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^1$ | $\left\{\mathbb{R}^{m\times n} \ni Z : \|\widetilde{Z}\|_* \leqslant \sqrt{r}, \|Z\|_\infty \leqslant \mathcal{B}_0\right\}$ |
| $\widetilde{\mathcal{F}}_r^p$ | $\left\{Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}\right\}$ |
| $\mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$ | $\left\{Z \in \mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M}\right\}$ |

| | |
|---|---|
| $\mathcal{F}^p_{r,\mathcal{B}_0}$ | $\left\{Z \in \mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M} = [r]^{\frac{2-p}{2p}}\sqrt{mn},\ \|Z\|_\infty \leqslant \mathcal{B}_0\right\}$ |
| $\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$ | $\left\{Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|^p_{\mathrm{sc},p} \leqslant r^{1-\frac{p}{2}};\ \|Z\|_\infty \leqslant \mathcal{B}_0\right\}$ |
| $\breve{\mathcal{F}}^p_r$ | $\left\{Z \in \mathbb{R}^{m\times n} : \|\breve{Z}\|^p_{\mathrm{sc},p} \leqslant [r]^{1-\frac{p}{2}}\right\}$ |
| $\breve{\mathcal{F}}^p_{r,\mathcal{B}_0}$ | $\left\{Z \in \mathbb{R}^{m\times n} : \|\breve{Z}\|^p_{\mathrm{sc},p} \leqslant [r]^{1-\frac{p}{2}}, \|Z\|_\infty \leqslant \mathcal{B}_0\right\}$ |
| **Other function classes** | |
| $\mathcal{L}_{\ell,\mathcal{B}}$ | Set of all loss functions bounded by $\mathcal{B}$ from $\mathbb{R}^2 \times ([m]\times[n])$ to $\mathbb{R}$, which are $\ell$-Lipschitz in the first argument |
| $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ | Set of all $\mathcal{B}_\mathrm{f}$-bounded, $\mathrm{L_f}$-Lipschitz functions |
| **DNN classes** | |
| $\phi : \mathbb{R}^{m+1} \to \mathbb{R}$ | Neural Network with final output |
| $\Psi : [m]\times[n] \to \mathbb{R}$ | Encoder Net taking users and items as input |
| $\mathcal{N}_{1,\mathrm{W}}$ | Networks satisfying conditions (261) |
| $\mathcal{N}_{2,W}$ | Networks satisfying conditions (265) |
| $\mathcal{N}_{0,\mathrm{W},\mathrm{c}}(a,s,c)$ | $\Big\{g : [m]\times[n] \to \mathbb{R}^1 \big\vert \exists f \in \mathcal{N}_{1,\mathrm{W}}(a,s),$ $U \in \mathbb{R}^{m\times\bar{\mathrm{m}}}, V \in \mathbb{R}^{n\times\bar{\mathrm{m}}}$ : $\|U\|^2_{\mathrm{Fr}} + \|V\|^2_{\mathrm{Fr}} \leqslant c^2\max(m,n),$ $\|A^0\| \leqslant s_0 :\ g(i,j) = f(A^0(u_i,v_j)^\top)\ \forall i,j\Big\}$ |
| $\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}(a,s,c)$ | $\Big\{g : [m]\times[n] \to \mathbb{R}^1 \big\vert \exists f \in \mathcal{N}_{1,\mathrm{W}}(a,s),$ $U \in \mathbb{R}^{m\times\bar{\mathrm{m}}}, V \in \mathbb{R}^{n\times\bar{\mathrm{m}}}$ : $\|\operatorname{diag}(\widetilde{p})^{\frac12}U\|^2_{\mathrm{Fr}} + \|\operatorname{diag}(\widetilde{q})^{\frac12}V\|^2_{\mathrm{Fr}} \leqslant c^2,$ $\|A^0\| \leqslant s_0 :\ g(i,j) = f(A^0(u_i,v_j)^\top)\ \forall i,j\Big\}$ |
| $\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}(a,s,c)$ | $\Big\{g : [m]\times[n] \to \mathbb{R}^1 \big\vert \exists f \in \mathcal{N}_{1,\mathrm{W}}(a,s),$ $U \in \mathbb{R}^{m\times\bar{\mathrm{m}}}, V \in \mathbb{R}^{n\times\bar{\mathrm{m}}}$ : $\|\operatorname{diag}(\widetilde{p})^{\frac12}U\|^2_{\mathrm{Fr}} + \|\operatorname{diag}(\widetilde{q})^{\frac12}V\|^2_{\mathrm{Fr}} \leqslant c, \|A^0\| \leqslant s_0 :$ $g(i,j) = f(A^0(u_i,v_j)^\top)\ \forall i,j\Big\}$ |
| $\mathcal{N}_{1,W,\mathrm{id}}(a,s)_{i,j}$ | $\tilde{\phi}(x_\xi)$ where $\tilde{\phi}$ is a network form (259) satisfying Cond. (265) and $x_{i,j} := \mathrm{concat}(e_i, e_j)$ |
| **Composite function classes** (illustrative examples) | |
| $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \mathcal{F}^p_{r,\mathcal{B}_0}$ | $\left\{g : [m]\times[n] \to \mathbb{R} : \exists f \in \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f},\ Z \in \mathcal{F}^p_{r,\mathcal{B}_0} :\ g(i,j) = f(Z_{i,j})\right\}$ |
| $\mathrm{Z}^0$ | Latent matrix in representation of ground truth G as $f \circ \mathrm{Z}^0$ |
| $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$ | $\left\{g : [m]\times[n] \to \mathbb{R} : \exists f \in \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f},\ Z \in \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} :\ g(i,j) = f(Z_{i,j})\right\}$ |
| $\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}$ | $\left\{g : [m]\times[n] \to \mathbb{R} : \exists Z \in \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} \wedge\ \Psi \in \mathcal{N}_{0,\mathrm{W},\mathrm{c}} :\ g(i,j) = Z_{i,j} + \Psi(i,j)\right\}$ |
| $\mathrm{l}\circ\mathcal{N}_{2,W}(a',s')(\widetilde{\mathcal{F}}^p_r, \mathcal{N}_{2,W}(a,s))$ | Set of functions $G$ written as $G(\xi,\widetilde{G}) = \mathrm{l}(\phi_1(Z+\phi_2), \widetilde{G}, \xi)$ for some $Z \in \widetilde{\mathcal{F}}^p_r,\ \mathcal{N}_{2,W}(a',s') \ni \phi_1 : \mathbb{R}^2 \to \mathbb{R},$ $\mathcal{N}_{2,W}(a,s) \ni \phi_2 : [m]\times[n] \to \mathbb{R}$ |
| **Constants** | |
| $\ell$ | Lipschitz constant of l |
| $\mathcal{B}$ | Bound on the loss function l |

| | |
|---|---|
| $C$ | Constant from (Latała, 2005) |
| $C_1$ | $\max(C, 1)$ |
| $\mathrm{L_f}$ | Bound on the Lipschitz constant of $f$ |
| $\mathcal{B}_\mathrm{f}$ | Bound on the values of $f$ |
| $\underline{\mathrm{m}}$ | Number of latent matrices |
| **Constants in Neural Networks** | |
| $W$ | Number of parameters (of DNN) |
| $L$ | Number of layers (of DNN) |
| $w_1, \ldots, w_L = 1$ | Layer widths |
| $s_1, \ldots, s_L$ | Constraints on $\|W_1\|, \ldots, \|W_L\|$ |
| $a_1, \ldots, a_L$ (in $\mathcal{N}_{1,\mathrm{W}}$) | Constraints on $\|(W^1 - M^1)^\top\|_{2,1}, \ldots, \|(W^1 - M^1)^\top\|_{2,1}$ |
| $a_1, \ldots, a_L$ (in $\mathcal{N}_{2,W}$) | Constraints on $\|W^1 - M^1\|, \ldots, \|W^1 - M^1\|$ |
| $W^1, \ldots, W^L$ | Weight matrices (of DNN) |
| $M^1, \ldots, M^L$ | Initialised weights (of DNN) |
| $s_0$ | Upper bound on $\|A^0\|$ |
| $c^2$ (with weights) | Upper bound on $\|\operatorname{diag}(\widetilde{p})^{\frac{1}{2}} U\|_{\mathrm{Fr}}^2 + \|\operatorname{diag}(\widetilde{q})^{\frac{1}{2}} V\|_{\mathrm{Fr}}^2$ |
| $c^2$ (without weights) | Upper bound on $\left[\|U\|_{\mathrm{Fr}}^2 + \|V\|_{\mathrm{Fr}}^2\right] \max(m, n)$ |
| $\mathrm{R_W}$ | $\sum_{\ell=1}^L 2^{3/2} \prod_{\ell=1}^L s_\ell \left[\sum_{\ell=1}^L \left[\frac{a_\ell}{s_\ell}\right]^{2/3}\right]^{3/2}$ |
| **Constants in log terms** | |
| $\Gamma_{\widetilde{\mathcal{F}}_r^p, \ell}$ | $\frac{6(m+n)N(\ell+1)(r+1)}{\delta}$ |
| $\Gamma_{\mathcal{F}_\mathrm{t}^\mathrm{p}, \ell}$ | $6N(m+n)(r+1)(\ell+1)$ |
| $\Gamma_{\mathcal{F}_{r,\mathcal{B}_0}^p, \ell}$ | $6N(m+n)(r+1)(\ell+1)(\mathcal{B}_0+1)$ |
| $\Gamma_{\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p, \ell}$ | $\frac{3Nmn^3[\mathcal{B}_0+1][\ell+1]+1}{\delta}$ |
| $\Gamma_{W,\epsilon}$ | $\frac{96 W s_0 (m+n)\sqrt{mn} \prod_{\ell=1}^L s_\ell}{\epsilon} + 1$ |
| **For multiple latent matrices** | |
| $\underline{\mathrm{m}}$ | Number of latent matrices |
| $p_v$ | $v$th latent matrix Schatten index |
| $r_v$ | Constraint on $\|Z\|$ |
| $\bar{r}$ | $\sum_{v=1}^{\underline{\mathrm{m}}} r_v$ |
| $\Gamma_{W,\underline{\mathrm{m}}}$ | $12N\left[\prod_{\ell=1}^L s_\ell + \underline{\mathrm{m}}\,\mathcal{B}_0\right]\left[\prod_{\ell=1}^L s'_\ell\right]\left[\prod_{\ell=1}^L s_\ell\right][\sum_\ell a'_\ell]\,[\sum_\ell a_\ell] + 1$ |
| $\underline{\Gamma}$ | $3Nmn^3[\mathcal{B}_0+1][\ell+1] + 1$ |
| $\mathcal{H}$ | $\mathcal{N}_{2,W}(a', s') \circ (\mathrm{concat}_{v=1}^{\underline{\mathrm{m}}}(\widetilde{\mathcal{F}}_{r_v,\mathcal{B}_0}^{p_v}), \mathcal{N}_{1,W,\mathrm{id}}(a, s))$ |
| **Model abbreviations** | Name and function class |
| **Sd** | Schatten matrix completion (MC)<br>$\mathcal{F}_\mathrm{t}^\mathrm{p}, \mathcal{F}_{r,\mathcal{B}_0}^p$ |
| **SWd** | Schatten weighted MC<br>$\widetilde{\mathcal{F}}_r^p, \widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p, \breve{\mathcal{F}}_r^p, \breve{\mathcal{F}}_{r,\mathcal{B}_0}^p$ |
| **FRMC-FSd** | Functionally rescaled Schatten MC<br>$\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \mathcal{F}_\mathrm{t}^\mathrm{p},\ \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \mathcal{F}_{r,\mathcal{B}_0}^p$ |
| **FRMC-FSWd** | Functionally rescaled Schatten weighted MC<br>$\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \breve{\mathcal{F}}_{r,\mathcal{B}_0}^p,\ \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \breve{\mathcal{F}}_{r,\mathcal{B}_0}^p$ |
| **FRMC-Sd+NN** | Sum of Sd and NN<br>$\mathcal{F}_\mathrm{t}^\mathrm{p} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}},\ \mathcal{F}_{r,\mathcal{B}_0}^p + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}$ |
| **FRMC-SWd+NN** | Sum of SWd and NN<br>$\widetilde{\mathcal{F}}_r^p + \mathcal{N}_{0,\mathrm{W},\mathrm{c}},\ \breve{\mathcal{F}}_r^p + \widetilde{\mathcal{N}_{0,W,2c}}, \widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}$ , etc. |

## B. Table summary of results

Table 4: Table of our results for the Schatten $p$ constrained matrix completion without linear components as compared to the previous works for $p = 1, p < 1$ and $p = 0$. For simplicity, we **omit polylogarithmic factors** in all relevant quantities such as $N, m, n, \mathcal{B}, \mathcal{B}_0, \mathcal{B}_{\mathrm{f}}, \ell$ and the failure probability $\delta$.). NC stands for "Not comparable", NDC stands for "Not directly comparable". The compressed sensing literature (Zhang et al., 2013; Recht et al., 2010; Liu et al., 2014) offers results which can loosely be compared to an exact recovery sample complexity of $\widetilde{O}(m+n)r$ where $r$ is the ground truth *rank* with uniform RIP measurements (e.g. Gaussian measurements, loosely analogous to uniform sampling). (Fan, 2021) includes somewhat different assumptions. The rank-restricted version ($p = 0$, c.f. Lemma D.1) is a simple consequence of parameter counting (see (Long & Sedghi, 2020; Graf et al., 2022; Mohri et al., 2018; Ledent et al., 2021b; Giné & Guillou, 2001; Platen, 1986; Talagrand, 1994; 1996)). There is also an analogous result for classification with uniform sampling (Srebro et al., 2004; Srebro & Shraibman, 2005; Srebro & Jaakkola, 2005).

| Constraint | Sampling | Our Bound | Previous Work | Comment |
|---|---|---|---|---|
| $\|Z\|_* \leqslant \sqrt{rmn} = \mathcal{M}$ | Uniform | $\mathcal{B}\sqrt{\frac{r(m+n)}{N}}$<br>$\mathcal{B}\mathcal{M}\sqrt{\frac{1}{N\min(m,n)}}$<br>(Thm 3.1) | $\mathcal{B}\mathcal{M}\sqrt{\frac{1}{N\min(m,n)}}$<br>$\mathcal{B}\sqrt{\frac{r(m+n)}{N}}$<br>(Foygel et al., 2011) | |
| $\|Z\|_* \leqslant \sqrt{rmn} = \mathcal{M}$ | Arbitrary | $\sqrt{\frac{\mathcal{B}\ell(m+n)^{\frac{3}{2}}\sqrt{r}}{N}}$<br>$\mathcal{M}\sqrt{\frac{\mathcal{B}\ell\mathcal{M}\sqrt{m+n}}{N}}$<br>(Thm 3.2) | $\sqrt{\frac{\mathcal{B}\ell(m+n)^{\frac{3}{2}}\sqrt{r}}{N}}$<br>$\mathcal{M}\sqrt{\frac{\mathcal{B}\ell\mathcal{M}\sqrt{m+n}}{N}}$<br>(Shamir & Shalev-Shwartz, 2011) | |
| $\|\widetilde{Z}\|_* \leqslant \sqrt{r}$ | Arbitrary | $\mathcal{B}\sqrt{\frac{r(m+n)}{N}}$<br>(Thm 3.1) | $\mathcal{B}\sqrt{\frac{r(m+n)}{N}}$<br>(Foygel et al., 2011) | |
| $\|Z\|_{\mathrm{sc},p}^p \leqslant \mathcal{M}^p = r^{1-\frac{p}{2}}\sqrt{mn}^p$ | Uniform | $\mathcal{B}^{\frac{2-2p}{2-p}}\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}$<br>$\mathcal{B}^{\frac{2-2p}{2-p}}\ell^{\frac{p}{2-p}}\sqrt{\frac{\mathcal{M}^{\frac{2p}{2-p}}(m+n)^{\frac{2-3p}{2-p}}}{Np}}$<br>(Thm 3.1) | $\sqrt[(4)]{\frac{rn^{\frac{2}{2-p}}}{Np}}$<br>$\sqrt[(4)]{\frac{n^{\frac{2-2p}{2-p}}\mathcal{M}^{\frac{2p}{2-p}}}{Np}}$<br>(Fan, 2021) | $\mathcal{B}, \ell$ constant;<br>$n = m$<br>no replacement |
| $\|Z\|_{\mathrm{sc},p}^p \leqslant \mathcal{M}^p = r^{1-\frac{p}{2}}\sqrt{mn}^p$ | Arbitrary | $\mathcal{B}^{1-\frac{p}{2}}\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{Np}}$<br>$\mathcal{B}^{1-\frac{p}{2}}\ell^{\frac{p}{2}}\sqrt{\frac{\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{Np}}$<br>(Thm 3.2) | N/A | NC to<br>Comp. sensing |
| $\|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}$ | Arbitrary | $\mathcal{B}^{\frac{2-2p}{2-p}}\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}$<br>(Thm 3.1) | N/A | |
| $\|Z\|_\infty \leqslant \mathcal{B}_0$<br>$\mathrm{rank}(Z) \leqslant r$ | Arbitrary | $\mathcal{B}\sqrt{\frac{r(m+n)}{N}\log(\ell\mathcal{B}_0)}$<br>Parameter counting<br>(Lemma D.1) | $\mathcal{B}\sqrt{\frac{r(m+n)}{N}\log(\ell\mathcal{B}_0)}$<br>Parameter counting<br>cf. also (Srebro, 2004)<br>(Mohri et al., 2018) | |

Table 5: Very short summary of our results for Schatten quasi-norm matrix completion including dependence on $p$. For simplicity, the $\widetilde{O}$ notation hides logarithmic factors of the relevant quantities, including of the failure probability $\delta$ and the constraint quantity $\mathcal{B}_0$.

| **Main constraint** | **Sampling** | $\|\mathbf{Z}\|_\infty$ **unconstrained** | $\|\mathbf{Z}\|_\infty \leqslant \mathcal{B}_0$ |
|---|---|---|---|
| $\|\frac{Z}{\sqrt{mn}}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}$ | Uniform | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\right)$ (Thm 3.1) | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\right)$ (Thm 3.1) |
| $\|\frac{Z}{\sqrt{mn}}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}}$ | Arbitrary | $\widetilde{O}\left(\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{Np}}\right)$ (Thm 3.2) | $\widetilde{O}\left(\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{N}}\right)$ (Thm 3.2) |
| $\|\widetilde{Z}\|_{\mathrm{sc},p} \leqslant r^{1-\frac{p}{2}}$ | Arbitrary | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\right)$ (Thm 3.1) | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\right)$ (Thm 3.1) |

Table 6: detailed tabular summary of our results in the supplementary, expressed in terms of generalization error bounds. The $\widetilde{O}$ notation hides polylogarithmic factors in all variables and constraints. Similar excess risk bounds hold for all function classes under the condition $p = \frac{2}{d}$ for some $d$. Similar excess risk bounds hold for the empirically weighted analogues (with a multiple of the constraint $r$) under both realisability assumptions and the assumption $p = \frac{2}{d}$. See Thms C.4, C.7 and C.10.

| Function class | Generalization Bound | Relevant Theorem |
|---|---|---|
| **SdMC** | | |
| $\mathrm{l} \circ \widetilde{\mathcal{F}}^{p}_{r}$ | $O\left(\mathcal{B}^{\frac{2-2p}{2-p}}\, \ell^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{Np}} \log(\frac{r_\star mnN\ell_\star}{\delta}) + \mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\right)$ | Thm C.1 |
| $\mathrm{l} \circ \widetilde{\mathcal{F}}^{p}_{r,\mathcal{B}_0}$ | $O\left(\mathcal{B}^{\frac{2-2p}{2-p}}\, \ell^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{N}} \log(\frac{mnNr_\star\, \ell_\star\, \mathcal{B}_0}{\delta}) + \mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\right)$ | Thm C.2 |
| $\mathrm{l} \circ \mathcal{F}^{\mathrm{p}}_{\mathrm{t}}$ | $O\left(\mathcal{B}^{1-\frac{p}{2}}\, \ell^{\frac{p}{2}} \sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}} \log(mnN\, \ell_\star\, r_\star)}{Np}} + \mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\right)$ | Thm C.3 |
| $\mathrm{l} \circ \mathcal{F}^{p}_{r,\mathcal{B}_0}$ | $O\left(\mathcal{B}^{1-\frac{p}{2}}\, \ell^{\frac{p}{2}} \sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}} \log(mnN\, \ell_\star\, r_\star\, \mathcal{B}_0)}{N}} + \mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\right)$ | Thm C.3 |
| **FRMC** | | |
| $\mathrm{l} \circ \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^{p}_{r,\mathcal{B}_0}$ | $O\Bigg(\mathcal{B}^{\frac{2-2p}{2-p}} [\mathrm{L_f}\, \ell]^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{N}} \log(N) \log(\frac{mn\, \mathcal{B}_0[\ell\, \mathrm{L_f}+1]}{\delta}) + \sqrt{\frac{\mathcal{B}_0\, \mathrm{L_f}\, \ell\, \mathcal{B} + \mathcal{B}^2}{N}} \log(N) + \mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\Bigg)$ | Thm C.5 |
| $\mathrm{l} \circ \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \mathcal{F}^{p}_{r,\mathcal{B}_0}$ | $O\Bigg(\mathcal{B}^{1-\frac{p}{2}} [\mathrm{L_f}\, \ell]^{\frac{p}{2}} \sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{N}} \log^{\frac{3}{2}}(Nmn\, \mathcal{B}_0[\ell\, \mathrm{L_f}+1]) + \sqrt{\frac{\mathcal{B}_0\, \mathrm{L_f}\, \ell\, \mathcal{B} + \mathcal{B}^2}{N}} \log(N) + \mathcal{B}\sqrt{\frac{\log(\frac{1}{\delta})}{N}}\Bigg)$ | Thm C.6 |
| **SdMC+NN** | | |
| $\mathrm{l} \circ (\widetilde{\mathcal{F}}^{p}_{r,\mathcal{B}_0} + \widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}})$ | $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\, \ell^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{N}} + \mathcal{B}\sqrt{\frac{d(m+n)}{N}} + \mathcal{B}\, \frac{s_0 c\, \mathrm{R_W}}{\sqrt{N}}\right)$ | Thm C.8 |
| $\mathrm{l} \circ (\mathcal{F}^{p}_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}})$ | $\widetilde{O}\Bigg(\mathcal{B}^{1-\frac{p}{2}}\, \ell^{\frac{p}{2}} \sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{N}} + \mathcal{B}\sqrt{\frac{d(m+n)}{N}} + \mathcal{B}\, \frac{s_0 c\sqrt{m+n}\, \mathrm{R_W}}{\sqrt{N}}\Bigg)$ | Thm C.9 |
| **Multi-latent extension** | | |
| $\mathcal{H} := \mathcal{N}_{2,W}(a', s') \circ (\mathrm{concat}^{\mathrm{m}}_{v=1}(\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}), \mathcal{N}_{1,W,\mathrm{id}}(a, s))$ | $\widetilde{O}\Bigg(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B}\sqrt{\frac{W+W'}{N}} + \mathcal{B}_0\, \mathcal{S}'\, \ell\sqrt{\frac{\mathrm{m}^2\, \bar{r}(m+n)}{N}} + \mathcal{B}_0\, \mathcal{S}'\, \ell\sqrt{\frac{\mathrm{m}^3}{N}}\Bigg)$, where $\mathcal{S}' = \left[\prod_{\ell=1}^{L} s'_\ell\right]$ | Thm G.5 |

## C. Generalization and Excess Risk Results

In this section, we prove our main results. Many of the key difficulties involved in the proofs have already been overcome in the proofs of the relevant partial results in Sections D, which itself relies on lower level tools from Section E.

### C.1. Generalization and Excess Risk Bounds Schatten Norm Matrix Completion (Sd and SWd)

In this subsection, we prove generalization and excess risk bounds for ordinary matrix completion (without a non-linear component) with Schatten norm regularization.

**Theorem C.1.** *Let $\mathrm{l} \in \mathcal{L}_{\ell,\mathcal{B}}$ be a loss function. We consider the function class $\widetilde{\mathcal{F}}_r^p := \left\{ Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}} \right\}$. Let $\hat{Z} := \min_{Z\in\widetilde{\mathcal{F}}_r^p} \widehat{E}(\mathrm{l}(Z_\xi, \widetilde{G}))$. With probability greater than $1-\delta$, we have the following excess risk bound:*

$$\mathbb{E}(\mathrm{l}(\hat{Z}_\xi, \widetilde{G}_\xi)) - \mathbb{E}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G}_\xi)) \leqslant 12\,\mathcal{B}\sqrt{\frac{\log(4/\delta)}{2N}} + 4\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 44\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\log(2\Gamma_{\widetilde{\mathcal{F}}_r^p,\ell})\left[1+\sqrt{\frac{m+n}{N}}\right], \tag{10}$$

*where $\Gamma_{\widetilde{\mathcal{F}}_r^p,\ell} := \frac{6(m+n)N(\ell+1)(r+1)}{\delta}$.*

*In particular, if the sampling distribution is uniform, the same result holds for $\mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$. Furthermore, the same upper bound holds for $\sup_{Z\in\widetilde{\mathcal{F}}_r^p} \mathbb{E}(\mathrm{l}(Z,\widetilde{G})) - \widehat{\mathbb{E}}(\mathrm{l}(Z,\widetilde{G}))$. (In fact, the generalization bound holds with a factor of $1/2$ on the right with probability $\geqslant 1-\delta$.)*

*Proof.* This follows immediately from Theorem D.2, and Theorem F.11.

□

Very similarly, we have the following result which applies with an additional constraint on the maximum entry:

**Theorem C.2.** *Let $\mathrm{l} \in \mathcal{L}_{\ell,\mathcal{B}}$ be a loss function. Consider the following function class:*

$$\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p := \left\{ Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}};\ \|Z\|_\infty \leqslant \mathcal{B}_0 \right\}. \tag{11}$$

*Let $\hat{Z} := \min_{Z\in\widetilde{\mathcal{F}}_r^p} \widehat{E}(\mathrm{l}(Z_\xi, \widetilde{G}))$. With probability greater than $1-\delta$, we have the following excess risk bound:*

$$\mathbb{E}(\mathrm{l}(\hat{Z}_\xi, \widetilde{G}_\xi)) - \mathbb{E}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G}_\xi)) \leqslant 12\,\mathcal{B}\sqrt{\frac{\log(4/\delta)}{2N}} + 4\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 44\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(2\Gamma_{\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p,\ell})\left[1+\sqrt{\frac{m+n}{N}}\right], \tag{12}$$

*where $\Gamma_{\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p,\ell} := \frac{3Nmn^3[\mathcal{B}_0+1][\ell+1]+1}{\delta}$. In particular, in the case of a uniform distribution, the same result holds for $\mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$. The same result also holds for the generalization error $\sup_{Z\widetilde{\mathcal{F}}_{r,\mathcal{B}_0}^p} \mathbb{E}(\mathrm{l}(Z_\xi,\widetilde{G}_\xi)) - \bar{\mathbb{E}}(\mathrm{l}(Z_\xi,\widetilde{G}_\xi))$.*

*Proof.* This follows immediately from Theorem D.4, and Theorem F.11. □

Next, we consider the case of a non-uniform distribution with non-weighted trace norm constraints:

**Theorem C.3.** *Consider the following function class:*

$$\mathcal{F}_{\mathrm{t}}^{\mathrm{p}} := \left\{ Z \in \mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M} = [r\sqrt{mn}]^{\frac{2-p}{2p}} \right\} \tag{13}$$

*With probability $\geqslant 1-\delta$, excess risk bound for $\hat{Z} \in \arg\min_{Z \in \mathcal{F}_{\mathrm{t}}^{\mathrm{p}}} \widehat{\mathbb{E}}(\mathrm{l}(Z, \widetilde{G}))$:*

$$\mathbb{E}(\mathrm{l}(\hat{Z}_\xi, \widetilde{G}_\xi)) - \mathbb{E}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G}_\xi)) \leqslant 12\,\mathcal{B}\sqrt{\frac{\log(4/\delta)}{2N}} + 4\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 4(\sqrt{18C}+\sqrt{2})\,\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}\log\left(2\Gamma_{\mathcal{F}_{\mathrm{t}}^{\mathrm{p}},\ell}\right)}{Np}}. \tag{14}$$

*where $\Gamma_{\mathcal{F}_{\mathrm{t}}^{\mathrm{p}},\ell} := 6N(m+n)(r+1)(\ell+1)$.*

*Furthermore, if we consider instead the optimization over the function class*

$$\mathcal{F}^p_{r,\mathcal{B}_0} := \left\{ Z \in \mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M} = [r\sqrt{mn}]^{\frac{2-p}{2p}},\ \|Z\|_\infty \leqslant \mathcal{B}_0 \right\} \tag{15}$$

*then we have instead*

$$\mathbb{E}(\mathrm{l}(\hat{Z}_\xi, \widetilde{G}_\xi, \xi)) - \mathbb{E}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G}_\xi, \xi)) \leqslant 12\,\mathcal{B}\sqrt{\frac{\log(4/\delta)}{2N}} + 4\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 4(\sqrt{18C}+\sqrt{2})\,\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}\log\left(2\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\ell}\right)}{N}}. \tag{16}$$

*where $\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\ell} := 6(m+n)N[\mathcal{B}_0+1][\ell+1]$. Furthermore,the same upper bounds hold for $\sup_{Z\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}} \mathbb{E}(\mathrm{l}(Z,\widetilde{G},\xi)) - \widehat{\mathbb{E}}(\mathrm{l}(Z,\widetilde{G},\xi))$.*

*Proof.* This follows immediately from Theorems D.3, D.5 and F.11. □

Next, we provide an excess risk result for the empirically weighted version.

**Theorem C.4.** *Assume that $p = \frac{2}{d}$ for some integer $d$. Let $\hat{Z} \in \arg\min\left(\widehat{\mathbb{E}}(\mathrm{l}(Z_\xi, \widetilde{G})) : Z \in \check{\mathcal{F}}^p_{2r}\right)$. If we assume that the ground truth $\mathrm{G}$ belongs to $\widetilde{\mathcal{F}}^p_r$, we have the following excess risk bound, which holds with probability $\geqslant 1-\delta$ under the condition that $N \geqslant 140(m+n)\log\left(\frac{3(m+n)}{\delta}\right)$:*

$$\mathbb{E}(\mathrm{l}(\widehat{Z}_\xi, \widetilde{G}, \xi)) - \mathbb{E}(\mathrm{l}(\mathrm{G}, \widetilde{G}, \xi)) \leqslant 12\,\mathcal{B}\sqrt{\frac{\log(12/\delta)}{2N}} + 4\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 100\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\log(6\Gamma_{\widetilde{\mathcal{F}}^p_r,\ell}). \tag{17}$$

*Furthermore, the upper bound also holds for the generalisation error (with the same error probability), and an analogous result holds for the class $\check{\mathcal{F}}^p_{r,\mathcal{B}_0}$, with $\Gamma_{\widetilde{\mathcal{F}}^p_r,\ell}$ replaced by $\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}$ and with the factor of $p$ removed.*

*Proof.* By lemma E.5 we have, with probability $\geqslant 1-\delta/3$,

$$\|\widetilde{\mathrm{Z}^0}\|^{2/d}_{\mathrm{sc},2/d} \leqslant \left(1+\sqrt{\frac{6(m+n)\log\left(\frac{3(m+n)}{\delta}\right)}{N}}\right)\|\widetilde{\mathrm{Z}^0}\|^{\frac{2}{d}}_{\mathrm{sc},2/d}. \tag{18}$$

In particular, as long as $N \geqslant 24(m+n)\log\left(\frac{3(m+n)}{\delta}\right)\left[2^{1-\frac{p}{2}}-1\right]^{-2}$ (note that since $p$ is at most 1, this is satisfied as long as $N \geqslant 140(m+n)\log\left(\frac{(m+n)}{\delta}\right)$) we certainly have

$$\|\widetilde{\mathrm{Z}^0}\|^p_{\mathrm{sc},p} \leqslant 2^{1-\frac{p}{2}}\|\widetilde{\mathrm{Z}^0}\|^p_{\mathrm{sc},p} \leqslant [2r]^{1-\frac{p}{2}}. \tag{19}$$

This implies that

$$\mathrm{G} \in \check{\mathcal{F}}^p_{2r}. \tag{20}$$

Similarly, by Lemma F.10, we also have with probability $\geqslant 1-\delta/3$ (as long as $N \geqslant 8(m+n)\log(\frac{3(m+n)}{\delta})$, which is already required by the stronger condition for (19)):

$$\check{p}_i \geqslant \frac{\tilde{p}_i}{2} \qquad \text{and} \qquad \check{q}_j \geqslant \frac{\tilde{q}_j}{2}. \tag{21}$$

Under this condition, for any matrix $Z \in \check{\mathcal{F}}^p_{2r}$,

$$\|\widetilde{Z}\|^p_{\mathrm{sc},p} = \|\sqrt{\mathrm{diag}(\tilde{p})\,\mathrm{diag}(\check{p})^{-1}}\check{Z}\sqrt{\mathrm{diag}(\tilde{q})\,\mathrm{diag}(\check{q})^{-1}})\|^p_{\mathrm{sc},p} \leqslant 2^p\|\check{Z}\|^p_{\mathrm{sc},p} \leqslant 2^p[2r]^{1-\frac{p}{2}} \leqslant [2^{\frac{2p}{2-p}}2r]^{1-\frac{p}{2}} \leqslant [4r]^{1-\frac{p}{2}}.$$

Hence, we certainly have:

$$\check{\mathcal{F}}^p_{2r} \subset \widetilde{\mathcal{F}}^p_{4r}. \tag{22}$$

Now, by Theorem C.1 we have w.p. $\geqslant 1-\delta/3$ simultaneously over all $Z \in \widetilde{\mathcal{F}}^p_{4r}$:

$$\begin{aligned}\mathbb{E}(\mathrm{l}(Z_\xi, \widetilde{G}_\xi)) - \widehat{E}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G}_\xi)) &\leqslant 6\,\mathcal{B}\sqrt{\frac{\log(12/\delta)}{2N}} + 2\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} \\ &\quad + 44\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\log(6\Gamma_{\widetilde{\mathcal{F}}^p_{4r},\ell})\left[1+\sqrt{\frac{m+n}{N}}\right].\end{aligned} \tag{23}$$

Thus after a union bound, equations (19), (22) an (23) hold simultaneously with probability $1-\delta$ and applying this to $\widehat{Z}$ and the ground truth G, we obtain:

$$\begin{aligned}&\mathbb{E}(\mathrm{l}(\widehat{Z}_\xi, \widetilde{G})) - \mathbb{E}(\mathrm{l}(\mathrm{G}, \widetilde{G})) \leqslant \mathbb{E}(\mathrm{l}(\widehat{Z}_\xi, \widetilde{G})) - \widehat{\mathbb{E}}(\mathrm{l}(\widehat{Z}_\xi, \widetilde{G})) + \widehat{\mathbb{E}}(\mathrm{l}(\widehat{Z}_\xi, \widetilde{G})) - \widehat{\mathbb{E}}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G})) + \widehat{\mathbb{E}}(\mathrm{l}(\mathrm{G}_\xi, \widetilde{G})) - \mathbb{E}(\mathrm{l}(\mathrm{G}, \widetilde{G})) \\ &\leqslant 12\,\mathcal{B}\sqrt{\frac{\log(12/\delta)}{2N}} + 4\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 88\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}\log(6\Gamma_{\widetilde{\mathcal{F}}^p_{r},\ell})\left[1+\sqrt{\frac{m+n}{N}}\right],\end{aligned} \tag{24}$$

as expected. □

### C.2. Generalization and Excess Risk Bounds for Functionally Rescaled Schatten quasi-norm Matrix Completion (FRMC-FSd and FRMC-FSWd)

We consider the following function class:

$$\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} : \left\{ g : [m]\times[n] \to \mathbb{R} : \exists f \in \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f},\ Z \in \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} :\ g(i,j) = f(Z_{i,j}) \right\}. \tag{25}$$

**Theorem C.5.** *With probability greater than $1-\delta$ over the draw of the training set we have the following bound on the empirical Rademacher complexity of the class $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_r$:*

$$\begin{aligned}\widehat{\mathfrak{R}}(\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}) &\leqslant 150\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\mathcal{B}_\mathrm{f} + {\mathcal{B}_\mathrm{f}}^2+1}}{\sqrt{N}}\log_2(N) \\ &\quad + 11\,{\mathcal{B}_\mathrm{f}}^{\frac{2-2p}{2-p}}\,\mathrm{L_f}^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log^2(2N\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\left[1+\sqrt{\frac{(m+n)}{N}}\right]\end{aligned} \tag{26}$$

*where* $\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\mathrm{L_f}} := \frac{3Nmn^3[\mathcal{B}_0+1][\mathrm{L_f}+1]+1}{\delta}$.

*In particular, for any fixed loss function* $\mathrm{l} \in \mathcal{L}_{\ell,\mathcal{B}}$*, with probability greater than* $1-\delta$ *over the draw of the training set, we have the following generalization bound for any* $f \circ Z \in \mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_r$*:*

$$\mathbb{E}(\ell(f(Z_\xi),\widetilde{G})) - \frac{1}{N}\sum_{o=1}^{N}\ell_o(f(Z_{\xi^o})) \leqslant 6\,\mathcal{B}\sqrt{\frac{\log(4/\delta)}{2N}} + 300\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}\log_2(N) \tag{27}$$
$$+\,22\,\mathcal{B}^{\frac{2-2p}{2-p}}(\mathrm{L_f}\,\ell)^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log^2(4N\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\left[1+\sqrt{\frac{(m+n)}{N}}\right].$$

*Furthermore, we also have the following excess risk bound, which holds with probability greater than* $1-\delta$*:*

$$\mathbb{E}(\ell(\hat{g}(i,j),\widetilde{G}) - \mathbb{E}(\ell(g^*(i,j),\widetilde{G})$$
$$\leqslant 12\,\mathcal{B}\sqrt{\frac{\log(4/\delta)}{2N}} + 600\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}\log_2(N) \tag{28}$$
$$+\,44\,\mathcal{B}^{\frac{2-2p}{2-p}}(\mathrm{L_f}\,\ell)^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log^2(4N\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\left[1+\sqrt{\frac{(m+n)}{N}}\right],$$

*where* $g^*$ *and* $\hat{g}$ *denote* $\min_{g\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}}\mathbb{E}(\mathrm{l}(g_\xi,\widetilde{G}))$ *and* $\min_{g\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}}\widehat{\mathbb{E}}(\mathrm{l}(g_\xi,\widetilde{G}))$ *respectively.*

*In particular, if the distribution is uniform, the same results hold for the function class* $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0}$*.*

*Proof.* By Proposition F.13 with $d=1$, for every $\epsilon>0$, there exists a uniform cover $\mathcal{C}(\epsilon)$ of $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ with cardinality satisfying

$$\log\left(|\,\mathcal{C}(\epsilon)\,|\right) \leqslant 3\left[\left[\frac{2\,\mathcal{B}_0\,\mathrm{L_f}}{\epsilon}\right]+1\right]. \tag{29}$$

Note that since this is a cover of $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ with respect to the uniform norm, it satisfies the properties of Lemma E.4, to wit, for any matrix $Z\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$, $|(f-\bar{f})(Z_{i,j})|\leqslant\epsilon$ (and therefore $|\,\mathrm{l}(f(Z_{i,j}),\widetilde{G},(i,j))-\mathrm{l}(\bar{f}(Z_{i,j}),\widetilde{G},(i,j))|\leqslant\epsilon\,\ell$ holds uniformly over any matrix $Z$ and any input $(i,j)$ (this is the condition from Lemma E.3 (cf. Eq. (174))), which is stronger than that in Lemma E.4 (cf. Eq. (183))). Thus, we can apply our Lemma E.4 with $\Theta_1=\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ and $\Theta_2=\widetilde{\mathcal{F}}^p_r$.

$$\widehat{\mathfrak{R}}(\mathrm{l}\circ\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}) \leqslant \mathbb{E}_\sigma\sup_{\theta_{1,2}\in\Theta_{1,2}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1,\theta_2) \tag{30}$$
$$\leqslant \log_2\left(\frac{1}{\alpha}\right)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right)+4\alpha+4\sqrt{10}\int_\alpha^{\mathcal{B}}\sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f},\epsilon/\ell))}{N}}d\epsilon+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}}. \tag{31}$$

For the first term, note that by Theorem D.2, with probability $\geqslant 1-\delta$ over the draw of the training set, we actually have

$$\sup_{\bar{\mathrm{f}}\in\mathcal{C}(\epsilon)}\widehat{\mathfrak{R}}_S(\mathrm{l}\circ\bar{\mathrm{f}}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}) \leqslant \sup_{\mathrm{f}\in\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}}\widehat{\mathfrak{R}}_S(\mathrm{l}\circ\mathrm{f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}) \tag{32}$$
$$\leqslant\sqrt{\frac{7\,\mathcal{B}^2+1}{N}}+11\,\mathcal{B}^{\frac{2-2p}{2-p}}\left[\mathrm{L_f}\,\ell\right]^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}})\left[1+\sqrt{\frac{(m+n)}{N}}\right]. \tag{33}$$

Regarding the second term in Equation (31), we have the following simple calculation:

$$4\sqrt{10}\int_\alpha^{\mathcal{B}}\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon))}{N}}d\epsilon = 4\sqrt{10}\int_\alpha^{\mathcal{B}}\sqrt{\frac{6\left[\frac{\mathcal{B}_0\,\mathrm{L_f}\,\ell}{\epsilon}+1\right]}{N}}d\epsilon \tag{34}$$
$$\leqslant 8\sqrt{15}\sqrt{\frac{\mathcal{B}_0\,\mathrm{L_f}\,\ell}{N}}\left[2\sqrt{\epsilon}\right]_\alpha^{\mathcal{B}}+4\sqrt{10}\frac{\mathcal{B}-\alpha}{\sqrt{N}} \leqslant 16\sqrt{\frac{15\,\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}}{N}}+4\sqrt{10}\frac{\mathcal{B}}{\sqrt{N}} \tag{35}$$
$$\leqslant 128\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2}}{\sqrt{N}}.$$

Plugging this back into Equations (31) and (32) we get $\widehat{\mathfrak{R}}(\mathrm{l}\circ\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}\circ\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0})\leqslant$

$$\begin{aligned}
&\leqslant\sqrt{\frac{7\,\mathcal{B}^2+1}{N}}\log_2(N)+11\,\mathcal{B}^{\frac{2-2p}{2-p}}\left[\mathrm{L_f}\,\ell\right]^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\log_2(N)\left[1+\sqrt{\frac{(m+n)}{N}}\right]\\
&\qquad\qquad+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}}+128\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2}}{\sqrt{N}}\qquad(36)\\
&\leqslant 11\,\mathcal{B}^{\frac{2-2p}{2-p}}\left[\mathrm{L_f}\,\ell\right]^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\log_2(N)\left[1+\sqrt{\frac{(m+n)}{N}}\right]+150\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}\log_2(N),
\end{aligned}$$

where we have assumed w.l.o.g. that $N\geqslant 2$ (the Theorem statement is obvious for $N=1$). Setting $\mathrm{l}=\mathrm{Id}$ establishes the first inequality (since in this case $\mathcal{B}=\mathcal{B}_\mathrm{f}$ and $\ell=1$). The generalization bound then follows from Theorem F.11 and a union bound over the two failure probabilities.

□

We now move on to prove a distribution-free result for the class $\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0}$.

**Theorem C.6.** *Consider the following function class:*

$$\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0}:=\left\{g:[m]\times[n]\to\mathbb{R}:\exists f\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f},\ Z\in\mathcal{F}^p_{r,\mathcal{B}_0}:\ g(i,j)=f(Z_{i,j})\right\}.$$

*We have the following excess risk bound, which holds with probability greater than* $1-\delta$*:*

$$\begin{aligned}
&\mathbb{E}(\ell(\hat{g}(\xi),\widetilde{G})-\mathbb{E}(\ell(g^*(\xi),\widetilde{G})\leqslant 12\,\mathcal{B}\sqrt{\frac{\log(2/\delta)}{2N}}+\qquad(37)\\
&4\log_2(N)\left[150\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}+\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\,\mathrm{L_f}^{\frac{p}{2}}\sqrt{\frac{19C_1\,\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{N}}\sqrt{\log\left(\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\mathrm{L_f}\,\ell}\right)}\right]
\end{aligned}$$

*where* $\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\ell}:=6(m+n)N[\mathcal{B}_0+1][\ell+1]$, $C_1=\max(C,1)$ *(*$C$ *being the constant from (Latała, 2005))* $g^*$ *and* $\hat{g}$ *denote* $\min_{g\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0}}\mathbb{E}(\mathrm{l}(g_\xi,\widetilde{G}))$ *and* $\min_{g\in\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0}}\widehat{\mathbb{E}}(\mathrm{l}(g_\xi,\widetilde{G}))$ *respectively.*

*Proof.* By the same arguments (cf. Equation (31)) as in the proof of Theorem C.5, we have the following bound on the Rademacher complexity of $\ell\circ\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0}$:

$$\widehat{\mathfrak{R}}(\ell\circ\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0})\leqslant\log_2(N)\sup_{f\in\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}}\widehat{\mathfrak{R}}(\mathrm{l}\circ f\circ\mathcal{F}^p_{r,\mathcal{B}_0})+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}}+128\frac{\sqrt{\mathcal{B}_0\,\ell\,\mathrm{L_f}\,\mathcal{B}+\mathcal{B}^2}}{\sqrt{N}}.$$

Next, by Theorem D.5, we can continue $\widehat{\mathfrak{R}}(\mathrm{l}\circ\mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}\circ\mathcal{F}^p_{r,\mathcal{B}_0})\leqslant$

$$\begin{aligned}
&\log_2(N)\left[\sqrt{\frac{7\,\mathcal{B}^2+1}{N}}+\mathcal{B}^{1-\frac{p}{2}}(\mathrm{L_f}\,\ell)^{\frac{p}{2}}\sqrt{\frac{2\,\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{N}}\left(3\sqrt{C}+\sqrt{\log\left(\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\mathrm{L_f}\,\ell}\right)}\right)\right]\qquad(38)\\
&\qquad\qquad+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}}+128\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2}}{\sqrt{N}}\\
&\leqslant\log_2(N)\left[150\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\mathcal{B}\,\ell+\mathcal{B}^2+1}}{\sqrt{N}}+\mathcal{B}^{1-\frac{p}{2}}\,\mathrm{L_f}^{\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{19C_1\,\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{N}}\sqrt{\log\left(\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\mathrm{L_f}\,\ell}\right)}\right],
\end{aligned}$$

which holds for any training sample. In particular, we can apply Theorem F.11 to yield the result immediately. □

We next turn our attention to the slightly more delicate case of results for the minimizer of the empirically weighted trace norm.

**Theorem C.7.** *Assume that $p = \frac{2}{d}$ for some integer $d$. Let $\widehat{g} \in \arg\min \left( \widehat{\mathbb{E}}(\mathrm{l}(g_\xi, \widetilde{G})) : g \in \mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \check{\mathcal{F}}^p_{2r,\mathcal{B}_0} \right)$, where $\check{\mathcal{F}}^p_{2r,\mathcal{B}_0}$ is the data dependent function class $\left\{ Z \in \mathbb{R}^{m\times n} : \|\check{Z}\|^p_{\mathrm{sc},p} \leqslant [2r]^{1-\frac{p}{2}},\ \|Z\|_\infty \leqslant \mathcal{B}_0 \right\}$.*

*If we assume that the ground truth $\mathrm{G}$ belongs to $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$, we have the following excess risk bound, which holds with probability $\geqslant 1-\delta$ under the condition that $N \geqslant 140(m+n)\log\left(\frac{m+n}{\delta}\right)$:*

$$\mathbb{E}(\ell(\hat{g}(i,j),\widetilde{G}) - \mathbb{E}(\ell(\mathrm{G}_{i,j},\widetilde{G}) \leqslant 6\,\mathcal{B}\sqrt{\frac{\log(12/\delta)}{2N}} + 600\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}\log_2(N) + \\ + 100\,\mathcal{B}^{\frac{2-2p}{2-p}}(\mathrm{L_f}\,\ell)^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log^2(12N\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}}). \tag{39}$$

*Proof.* The proof is similar to the proof of Theorem C.4. Let us write the ground truth as

$$\mathrm{G} = f \circ \mathrm{Z}^0 \tag{40}$$

with $f \in \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f}$ and $\mathrm{Z}^0 \in \widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$.

As in the proof of Theorem C.4, as long as $N \geqslant 140(m+n)\log\left(\frac{m+n}{\delta}\right)$ we certainly have w.p. $\geqslant 1-\delta/3$

$$\|\widetilde{\mathrm{Z}^0}\|^p_{\mathrm{sc},p} \leqslant 2^{1-\frac{p}{2}}\|\widetilde{\mathrm{Z}^0}\|^p_{\mathrm{sc},p} \leqslant [2r]^{1-\frac{p}{2}}. \tag{41}$$

This implies that

$$\mathrm{Z}^0 \in \check{\mathcal{F}}^p_{2r,\mathcal{B}_0}. \tag{42}$$

Similarly, by Lemma F.10, we also have with probability $\geqslant 1-\delta/3$ (as long as $N \geqslant 8(m+n)\log(\frac{3(m+n)}{\delta})$, which is already required by the stronger condition for (41)):

$$\check{p}_i \geqslant \frac{\tilde{p}_i}{2} \qquad \text{and} \qquad \check{q}_j \geqslant \frac{\tilde{q}_j}{2}. \tag{43}$$

Under this condition, we certainly have, by the same argument as in Equation (22) in the proof of Theorem C.4:

$$\check{\mathcal{F}}^p_{2r,\mathcal{B}_0} \subset \widetilde{\mathcal{F}}^p_{4r,\mathcal{B}_0}. \tag{44}$$

This, together with equation (42), implies that

$$\hat{g} \in \mathcal{F}_{\mathrm{lip},\mathrm{L_f},\mathcal{B}_f} \circ \widetilde{\mathcal{F}}^p_{4r,\mathcal{B}_0}. \tag{45}$$

Thus, we can apply Theorem C.5 (with $r \leftarrow 4r$, $\delta \leftarrow 3\delta$, $\ell \leftarrow \ell\,\mathrm{L_f}$) to obtain that an additional failure probability of $\delta$, we have the following for every $g \in \widetilde{\mathcal{F}}^p_{4r,\mathcal{B}_0}$:

$$\mathbb{E}(\ell(g_\xi,\widetilde{G})) - \frac{1}{N}\sum_{o=1}^N \ell_o(g_{\xi^o}) \leqslant 3\,\mathcal{B}\sqrt{\frac{\log(12/\delta)}{2N}} + 300\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}\log_2(N) \\ + 44\,\mathcal{B}^{\frac{2-2p}{2-p}}[\mathrm{L_f}\,\ell]^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(12N\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\left[1+\sqrt{\frac{(m+n)}{N}}\right]. \tag{46}$$

In particular, by a union bound, equations (45), (44) an (46) hold simultaneously with probability $1-\delta$ and applying this to $\hat{g}$ and the ground truth G, we obtain:

$$\begin{aligned}\mathbb{E}(\mathrm{l}(\hat{g}_\xi,\widetilde{G}))-\mathbb{E}(\mathrm{l}(\mathrm{G},\widetilde{G}))&\leqslant \mathbb{E}(\mathrm{l}(\hat{g}_\xi,\widetilde{G}))-\widehat{\mathbb{E}}(\mathrm{l}(\hat{g}_\xi,\widetilde{G}))+\widehat{\mathbb{E}}(\mathrm{l}(\hat{g}_\xi,\widetilde{G}))-\widehat{\mathbb{E}}(\mathrm{l}(\mathrm{G}_\xi,\widetilde{G}))+\widehat{\mathbb{E}}(\mathrm{l}(\mathrm{G}_\xi,\widetilde{G}))-\mathbb{E}(\mathrm{l}(\mathrm{G}_\xi,\widetilde{G}))\\ &\leqslant 6\,\mathcal{B}\sqrt{\frac{\log(12/\delta)}{2N}}+600\frac{\sqrt{\mathcal{B}_0\,\mathrm{L_f}\,\ell\,\mathcal{B}+\mathcal{B}^2+1}}{\sqrt{N}}\log_2(N)+\\ &\quad+88\,\mathcal{B}^{\frac{2-2p}{2-p}}\left[\mathrm{L_f}\,\ell\right]^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(12N\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell\,\mathrm{L_f}})\left[1+\sqrt{\frac{(m+n)}{N}}\right],\end{aligned}\tag{47}$$

as expected. □

### C.3. Generalization and Excess Risk Bounds for a Sum of a Latent Matrix and a Neural Encoding (Sd+NN)

**Theorem C.8.** *Fix a loss function* $\mathrm{l}\in\mathcal{L}_{\ell,\mathcal{B}}$ *and consider the following function class:*

$$\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}\tag{48}$$

*where we assume the output dimension in the class* $\mathcal{N}_{0,\mathrm{W,c}}$ *is* $K=0$*. Assume that* $N\geqslant 8(m+n)\log(\frac{3(m+n)}{\delta})$*.*

*Let* $\hat{g}\in\arg\min\left(\widehat{\mathbb{E}}(\mathrm{l}(g_{i,j},\widetilde{G})):g\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}\right)$ *and* $g^*\in\arg\min\left(\mathbb{E}(\mathrm{l}(g_{i,j},\widetilde{G})):g\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}\right)$*. Define*

$$\bar{B}:=\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}+\mathcal{B}\sqrt{\frac{d(m+n)}{N}}+\mathcal{B}\,\frac{s_0 c\,\mathrm{R_W}}{\sqrt{N}}.$$

*With probability greater than* $1-\delta$ *over the draw of the training set, we have the following:*

$$\widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\mathcal{N}_{0,\mathrm{W,c}}))\leqslant\widetilde{O}\left(\bar{B}\right)\tag{49}$$

$$\sup_{g\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}}\mathbb{E}(\mathrm{l}(g_\xi,\widetilde{G},\xi))-\widehat{\mathbb{E}}(\mathrm{l}(g_\xi,\widetilde{G},\xi))\leqslant\widetilde{O}\left(\bar{B}\right)+O\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right)\tag{50}$$

$$\mathbb{E}(\mathrm{l}(\hat{g},\widetilde{G},\xi))\leqslant\mathbb{E}(\mathrm{l}(g^*,\widetilde{G},\xi))+\widetilde{O}\left(\bar{B}\right)+O\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right),\tag{51}$$

*where the* $\widetilde{O}$ *notation hides polylogarithmic factors of all relevant quantities (*$\mathcal{B}$*,* $\mathcal{B}_0$*,* $\mathrm{l}$*,* $N,m,n,c,s_0,\mathrm{R_W},\prod_{\ell=1}^L s_\ell$ *etc.). In particular, if the distribution is uniform, the same result holds for the class* $\mathcal{F}^p_{r,\mathcal{B}_0}+\mathcal{N}_{0,\mathrm{W,c}}$ *instead. Furthermore, the same results hold for the class* $\widetilde{\mathcal{F}}^p_r+\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}$ *with* $\bar{B}$ *replaced by* $\underline{\mathrm{B}}$ *where*

$$\underline{\mathrm{B}}:=\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{Np}}+\mathcal{B}\sqrt{\frac{d(m+n)}{N}}+\mathcal{B}\,\frac{s_0 c\,\mathrm{R_W}}{\sqrt{N}}.$$

*Proof.* We aim to use Lemma E.4 with $\Theta_1=\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}$ and $\Theta_2=\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$. Assume that equations (270) are satisfied (this happens with probability $\geqslant 1-\delta/3$ as long as $N\geqslant 8(m+n)\log(\frac{3(m+n)}{\delta})$, by Lemma F.10) ). Then we can let $\mathcal{C}$ be a cover of granularity $\frac{\epsilon}{\ell}$ of the class $\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}$, as guaranteed by Proposition E.6. By Proposition E.6, we have

$$\log(|\mathcal{C}|)\leqslant\left[2d(m+n)+32s_0^2c^2\left[\frac{1}{\epsilon^2}+1\right]{\mathrm{R_W}}^2\right]\log\left(\Gamma_{W,\epsilon/\ell}\right).\tag{52}$$

Now, for any $\Psi\in\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}$, we write $\bar{\Psi}$ for the associated cover element. For any $g\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}$, we can write $g$ as $g=Z+\Psi$. We define an associated cover element in $\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\mathcal{C}$ as $Z+\bar{\Psi}$. For any value of $Z\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}$, it is certainly the

case that

$$\frac{1}{N}\sum_{o=1}^{N}(\mathrm{l}(g_{\xi^o},\widetilde{G}_o)-\mathrm{l}(\bar{g}_{\xi^o},\widetilde{G}_o))^2 \leqslant \ell^2\,\frac{1}{N}\sum_{o=1}^{N}(g_{\xi^o}-\bar{g}_{\xi^o})^2$$
$$=\ell^2\,\frac{1}{N}\sum_{o=1}^{N}(g_{\xi^o}-\bar{g}_{\xi^o})^2=\ell^2\,\frac{1}{N}\sum_{o=1}^{N}(Z_{\xi^o}+\Psi_{\xi^o}-[Z_{\xi^o}+\bar{\Psi}_{\xi^o}])^2=\ell^2\,\frac{1}{N}\sum_{o=1}^{N}(\Psi_{\xi^o}-\bar{\Psi}_{\xi^o})^2\leqslant\epsilon^2. \tag{53}$$

Thus, the condition (183) is satisfied and we can apply Lemma (E.4) to obtain $\widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}))\leqslant$

$$\log_2\left(\frac{1}{\alpha}\right)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right)+4\alpha+4\sqrt{10}\int_{\alpha}^{\mathcal{B}}\sqrt{\frac{\log(\mathcal{C})}{N}}d\epsilon+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}} \tag{54}$$
$$=\log_2\left(\frac{1}{\alpha}\right)\sup_{\Psi\in\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}}\widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\Psi))+4\alpha+4\sqrt{10}\int_{\alpha}^{\mathcal{B}}\sqrt{\frac{\log(|\mathcal{C}|)}{N}}d\epsilon+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}}.$$

Now, we tackle both main terms in equation (54) separately.

For the first term, it is clear that by applying Theorem D.4, (with an additional failure probability of $\delta/3$):

$$\sup_{\Psi\in\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}}\widehat{\mathfrak{R}}(\ell\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\Psi))$$
$$\leqslant\sqrt{\frac{7\,\mathcal{B}^2+1}{N}}+11\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(3\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell})\left[1+\sqrt{\frac{(m+n)}{N}}\right]$$
$$\leqslant\sqrt{\frac{7\,\mathcal{B}^2+1}{N}}+22\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(3\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}). \tag{55}$$

That is because Theorem D.4 explicitly holds uniformly over all loss functions in $\mathcal{L}_{\ell,\mathcal{B}}$, which includes all the "loss functions" $\mathrm{l}':(y,\widetilde{G},\xi)\mapsto\mathrm{l}'(y,\widetilde{G},\xi)=\mathrm{l}(y+\Psi_\xi,\widetilde{G},\xi)$ for any $\ell\in\mathcal{L}_{\ell,\mathcal{B}}$ and any $\Psi:[m]\times[n]\to\mathbb{R}$. Note that at the third line, we have used the condition $N\geqslant 8(m+n)\log(\frac{3(m+n)}{\delta})$.

For the second main term, we simply calculate the integral relying on Equation (52) (setting $\alpha=\frac{1}{N}$):

$$4\sqrt{10}\int_{\alpha}^{\mathcal{B}}\sqrt{\frac{\log(\mathcal{C})}{N}}d\epsilon\leqslant 4\sqrt{10}\int_{\alpha}^{\mathcal{B}}\sqrt{\frac{\left[2d(m+n)+32s_0^2c^2\left[\frac{1}{\epsilon^2}+1\right]\mathrm{R_W}^2\right]\log\left(\Gamma_{W,\epsilon/\ell}\right)}{N}}d\epsilon$$
$$\leqslant 8\sqrt{5}\sqrt{\log(\Gamma_{W,1/(N\ell)})}\left[\sqrt{\frac{d(m+n)}{N}}+\frac{4s_0c\,\mathrm{R_W}}{\sqrt{N}}\left[\log(\mathcal{B}\,N)+\mathcal{B}\right]\right]. \tag{56}$$

Plugging Equations (55) and (56) back into Equation (54), we obtain, with overall probability $\geqslant 1-2\delta/3$, that $\widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}))\leqslant$

$$8\sqrt{5}\sqrt{\log(\Gamma_{W,1/(N\ell)})}\left[\sqrt{\frac{d(m+n)}{N}}+\frac{4s_0c\,\mathrm{R_W}}{\sqrt{N}}\left[\log(\mathcal{B}\,N)+\mathcal{B}\right]\right]+\frac{4}{N}+4\,\mathcal{B}\sqrt{\frac{5\pi}{N}}$$
$$+\log_2(N)\left[\sqrt{\frac{7\,\mathcal{B}^2+1}{N}}+22\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(3\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell})\right]. \tag{57}$$

The results then follow immediately from Theorem F.11. The proof for $\widetilde{\mathcal{F}}^p_r+\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$ is the same except we are using Theorem D.2 instead of Theorem D.4. □

We now turn our attention to the case of the non-weighted regularization in the arbitrary sampling case:

**Theorem C.9.** *We now consider the following function class:* $\mathcal{F}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}$. *We also let* $\hat{g} \in \arg\min_g \left(\widehat{\mathbb{E}}(\mathrm{l}(g_\xi, \widetilde{G}, \xi)) : g \in \mathcal{F}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}\right)$ *and* $g^* \in \arg\min_g \left(\mathbb{E}(\mathrm{l}(g_\xi, \widetilde{G}, \xi)) : g \in \mathcal{F}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}\right)$. *Let*

$$\underline{C} := \mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{N}} + \mathcal{B}\sqrt{\frac{d(m+n)}{N}} + \mathcal{B}\,\frac{s_0 c\sqrt{m+n}\,\mathrm{R_W}}{\sqrt{N}}.$$

*With probability* $\geqslant 1-\delta$, *we have*

$$\widehat{\mathfrak{R}}(\mathrm{l}\circ(\mathcal{F}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}})) \leqslant \widetilde{O}\left(\bar{C}\right) \tag{58}$$

$$\sup_{g\in\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}+\mathcal{N}_{0,\mathrm{W},\mathrm{c}}} \mathbb{E}(\mathrm{l}(g_\xi, \widetilde{G}, \xi)) \leqslant \widetilde{O}\left(\bar{C}\right) + O\left(\mathcal{B}\sqrt{\frac{\log(4/\delta)}{N}}\right) \tag{59}$$

$$\mathbb{E}(\mathrm{l}(\hat{g}, \widetilde{G}, \xi)) \leqslant \mathbb{E}(\mathrm{l}(g^*, \widetilde{G}, \xi)) + \widetilde{O}\left(\bar{C}\right) + O\left(\mathcal{B}\sqrt{\frac{\log(4/\delta)}{N}}\right), \tag{60}$$

*where the* $\widetilde{O}$ *notation hides polylogarithmic factors of all relevant quantities (*$\mathcal{B}$, $\mathcal{B}_0$, l, $N, m, n, c, s_0, \mathrm{R_W}, \prod_{\ell=1}^L s_\ell$ *etc.). Furthermore, if we consider instead the class* $\mathcal{F}^{\mathrm{p}}_{\mathrm{t}} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}}$, *then the same result holds with* $\bar{C}$ *replaced by*

$$\underline{C} := \mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{Np}} + \mathcal{B}\sqrt{\frac{d(m+n)}{N}} + \mathcal{B}\,\frac{s_0 c\sqrt{m+n}\,\mathrm{R_W}}{\sqrt{N}}.$$

*Proof.* By the same arguments as in the proof of Theorem C.8, W.p. $\geqslant 1-\delta/2$, as long as $N \geqslant 8(m+n)\log(\frac{3(m+n)}{\delta})$ there exists a cover $\mathcal{C}$ of $\mathcal{N}_{0,\mathrm{W},\mathrm{c}}$ satisfying condition (183) and

$$\log(|\mathcal{C}|) \leqslant \left[2d(m+n) + 32 s_0^2 c^2 (m+n)\left[\frac{1}{\epsilon^2}+1\right]\mathrm{R_W}^2\right]\log\left(\Gamma_{W,\epsilon/\ell}\right). \tag{61}$$

Furthermore,

$$\widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}})) \leqslant \log_2(N) \sup_{\Psi\in\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}} \widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} + \Psi)) + \widetilde{O}\left(\mathcal{B}\sqrt{\frac{d(m+n)}{N}} + \frac{s_0 c\sqrt{m+n}\,\mathrm{R_W}}{\sqrt{N}}\log(\mathcal{B}N)\right). \tag{62}$$

For the first term, we now have by Theorem D.5:

$$\begin{aligned}&\log_2(N) \sup_{\Psi\in\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}} \widehat{\mathfrak{R}}(\mathrm{l}\circ(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} + \mathcal{N}_{0,\mathrm{W},\mathrm{c}})) \\ &\leqslant \log_2(N)\left[\sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{19 C_1\,\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{N}\log(\Gamma_{\mathcal{F}^p_{r,\mathcal{B}_0},\ell})}\right].\end{aligned} \tag{63}$$

Plugging this back into Equation (62) yields the result.

□

Finally, we consider excess risk bounds for the empirically weighted version of our algorithm. In this case, the "doubling argument" must be used for both components of the model.

**Theorem C.10.** *Assume that the ground truth* G *belongs to* $\breve{\mathcal{F}}^p_{r,\mathcal{B}_0} + \widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$. *Let* $\hat{g} \in \arg\min_g \left(\widehat{\mathbb{E}}(\mathrm{l}(g_\xi, \widetilde{G}, \xi)) : g \in \breve{\mathcal{F}}^p_{2r,\mathcal{B}_0} + \widetilde{\mathcal{N}_{0,W,2c}}\right)$. *Let*

$$\underline{\mathrm{D}} := \mathcal{B}\left[\sqrt{\frac{d(m+n)}{N}} + \frac{s_0 c\,\mathrm{R_W}}{\sqrt{N}}\right] + \ell^{\frac{p}{2-p}}\,\mathcal{B}^{\frac{2-2p}{2-p}}\sqrt{\frac{r(m+n)}{N}} \tag{64}$$

*With probability greater than* $1-\delta$ *over the draw of the training set, we have*

$$\sup_{g\in\check{\mathcal{F}}^p_{2r,\mathcal{B}_0}+\widetilde{\mathcal{N}_{0,W,2c}}} \mathbb{E}(\mathrm{l}(g_\xi,\widetilde{G},\xi)) - \widehat{\mathbb{E}}(\mathrm{l}(g_\xi,\widetilde{G},\xi)) \leqslant \widetilde{O}\left(\underline{\mathrm{D}}\right) + O\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right) \tag{65}$$

$$\mathbb{E}(\mathrm{l}(\hat{g},\widetilde{G},\xi)) - \mathbb{E}(\mathrm{l}(g^*,\widetilde{G},\xi)) \leqslant \widetilde{O}\left(\underline{\mathrm{D}}\right) + O\left(\mathcal{B}\sqrt{\frac{\log(6/\delta)}{N}}\right), \tag{66}$$

*where the* $\widetilde{O}$ *notation hides logarithmic factors of all relevant quantities (*$\mathcal{B}$, $\mathcal{B}_0$, $\mathrm{l}$, $N, m, n, c, s_0, \mathrm{R_W}, \prod_{\ell=1}^L s_\ell$ *etc.).*

*Proof.* Let us write the ground truth as

$$\mathrm{G} = \mathrm{Z}^0 + \Psi^0 \tag{67}$$

Just as in the proof of theorem C.4, by lemma E.5 we have, with probability $\geqslant 1-2\delta/3$, and as long as $N \geqslant 140(m+n)\log\left(\frac{m+n}{\delta}\right)$, that the following are all simultaneously satisfied:

$$\mathrm{Z}^0 \in \check{\mathcal{F}}^p_{2r}. \tag{68}$$

$$\check{p}_i \geqslant \frac{\tilde{p}_i}{2} \qquad \text{and} \qquad \check{q}_j \geqslant \frac{\tilde{q}_j}{2}, \tag{69}$$

and

$$\check{\mathcal{F}}^p_{2r} \subset \widetilde{\mathcal{F}}^p_{4r}. \tag{70}$$

Also, by the proof of Lemma F.10, on the same high probability event as above, we have (c.f. Equation 208) $\check{p}_i \leqslant 2\tilde{p}_i$ and $\check{q}_j \leqslant 2\tilde{q}_j$ (for all $i,j$). Thus since $\Psi^0 \in \widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$, we also have

$$\Psi^0 \in \widetilde{\mathcal{N}_{0,W,2c}} \subset \mathcal{N}_{0,\mathrm{W},4\mathrm{c}}\,. \tag{71}$$

Thus, we have

$$\mathrm{G} \in \check{\mathcal{F}}^p_{2r,\mathcal{B}_0} + \widetilde{\mathcal{N}_{0,W,2c}} \subset \widetilde{\mathcal{F}}^p_{4r,\mathcal{B}_0} + \widetilde{\mathcal{N}_{0,\mathrm{W},4\mathrm{c}}} \tag{72}$$

And thus, the theorem follows by applying Theorem C.8 with $\delta \leftarrow \delta/3$. □

**Remarks about the Norm Based Bounds on Neural Encodings:** The results in this Section comes with some caveats regarding the improvements offered by the weighting $\|\operatorname{diag}(\tilde{p})^{\frac{1}{2}}U\|^2_{\mathrm{Fr}} + \|\operatorname{diag}(\tilde{q})^{\frac{1}{2}}V\|^2_{\mathrm{Fr}}$. Indeed, the term $\sqrt{\frac{d(m+n)}{N}}$ still contains a parametric dependency on the dimension $d$ of $A^0(u_i,v_j)^\top$, so it cannot be said that the bounds in Theorems C.8 and C.9 capture any rank sparsity inducing properties of the regularizer on $U,V$. In fact, the weighting merely serves to increase the uniformity of the input *norms* of the embeddings, which improves the behavior of norm-based bounds. However, one can also control the complexity of the neural encoding with a parameter counting strategy, which would remove any difference between the weighted and unweighted scenarios. This is what we do in the Section G, which deals with the case where multiple hidden matrices are present.

## D. Our Results on the Complexity of Matrix Classes with the Schatten quasi-Norms

This section compiles our first end-product results: Rademacher complexity bounds for classes of matrices with low Schatten quasi-norms. This section is divided into two very similar sections where we treat the two cases where a separate upper bound on the entries is enforced or not. Indeed, a such a separate condition is required to prevent the bounds from blowing up as $p \to 0$. This condition is very mild, but we still cover both cases since it may be interesting to study the dependence on $p$. The proofs rely on the important tools developed in Section E.

### D.1. Without Constraints on Entries

First, we will need the following simple parameter-counting lemma (the proofs use standard techniques) for matrices of fixed rank. Similar results hark back to (Srebro & Shraibman, 2005; Srebro et al., 2004; Vandermeulen & Ledent, 2021), but the rest of our proofs will require the variation below, which is uniform over any draw of the sample set $\xi^1, \dots, \xi^N \in [m] \times [n]$.

**Lemma D.1.** *Consider the following function class over matrices in $\mathbb{R}^{m\times n}$:*

$$\mathcal{E}_{\mathrm{r,t}} := \left\{R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant t,\ \mathit{rank}(R) \leqslant r\right\}. \tag{73}$$

*For all $u \leqslant N$ let $\mathrm{l}_u : \mathbb{R} \to \mathbb{R}$ be $\ell$-Lipschitz functions which are bounded by $\mathcal{B}$. The covering number of $\mathcal{E}_{\mathrm{r,t}}$ is bounded as follows:*

$$\log \mathcal{N}_\infty\left(\mathcal{E}_{\mathrm{r,t}}, \epsilon\right) \leqslant (m+n)r \log\left(\frac{3\sqrt{2t}}{\epsilon} + 1\right). \tag{74}$$

*Furthermore, we have the following parameter counting bound on the empirical Rademacher complexity of $\mathrm{l} \circ \mathcal{E}_{\mathrm{r,t}}$.*

$$\widehat{\mathfrak{R}}(\mathrm{l} \circ \mathcal{E}_{\mathrm{r,t}}) := \sup_{Z \in \mathcal{E}_{\mathrm{r,t}}} \frac{1}{N}\sum_{o=1}^{N} \mathrm{l}_o(\mathrm{Z}_{\xi^o}) \leqslant \frac{1}{N} + \sqrt{\frac{2(m+n)r}{N} \log\left(3N\,\ell\sqrt{2t}+1\right)}\,\mathcal{B}. \tag{75}$$

*Proof.* First, note that by Lemma F.18 for any $R \in \mathcal{E}_{\mathrm{r,t}}$, we can find two matrices $A \in \mathbb{R}^{m\times o}$ and $B \in \mathbb{R}^{n\times o}$ such that

$$AB^\top = R \tag{76}$$

$$\|A\|_{\mathrm{Fr}}^2 + \|B\|_{\mathrm{Fr}}^2 \leqslant 2t. \tag{77}$$

In particular, Equation (77) certainly implies that $\|A\|_{\mathrm{Fr}}, \|B\|_{\mathrm{Fr}} \leqslant \sqrt{2t}$.

Next, setting $\varepsilon = \frac{\epsilon}{4\sqrt{2t}}$, by Lemma F.17, there exist covers $\mathcal{C}_A \subset \left\{A \in \mathbb{R}^{m\times r} : \|A\|_{\mathrm{Fr}} \leqslant \sqrt{2t}\right\}$ and $\mathcal{C}_B \subset \left\{B \in \mathbb{R}^{n\times r} : \|B\|_{\mathrm{Fr}} \leqslant \sqrt{2t}\right\}$ (with respect to the Frobenius norm) such that for all $A \in \mathbb{R}^{m\times r}$ (resp. $B \in \mathbb{R}^{m\times r}$), there exists a $\bar{A} \in \mathcal{C}_A$ (resp. $\bar{B} \in \mathcal{C}_B$) such that $\|A - \bar{A}\|_{\mathrm{Fr}}, \|B - \bar{B}\|_{\mathrm{Fr}} \leqslant \varepsilon$ and

$$|\mathcal{C}_A| \leqslant \left(\frac{3\sqrt{2t}}{\epsilon} + 1\right)^{mr} \qquad \text{and} \tag{78}$$

$$|\mathcal{C}_B| \leqslant \left(\frac{3\sqrt{2t}}{\epsilon} + 1\right)^{nr}. \tag{79}$$

The cover $\mathcal{C} := \left\{R \in \mathbb{R}^{m\times n} : R = AB^\top : A_1 \in \mathcal{C}_A,\ B_1 \in \mathcal{C}_B\right\} \subset \mathbb{R}^{m\times n}$ is an (external) $\epsilon/2$-cover of $\mathcal{E}_{\mathrm{r,t}}$ with respect to the $L_\infty$ norm. Indeed, for any $R \in \mathcal{E}_{\mathrm{r,t}}$, we can write $R = AB^\top$ for some $A \in \mathcal{C}_A, B \in \mathcal{C}_B$. Then, writing $\bar{A}$ (resp. $\bar{B}$) for the corresponding nearest cover elements in $\mathcal{C}_A$ (resp. $\mathcal{C}_B$), and writing $\bar{R} = \bar{A}\bar{B}^\top$ we have for any $i \leqslant m, j \leqslant n$:

$$\begin{aligned}\left(\bar{R} - R\right)_{i,j} \leqslant \left\|\bar{R} - R\right\|_{\mathrm{Fr}} &= \left\|\bar{A}\bar{B}^\top - AB^\top\right\|_{\mathrm{Fr}} = \left\|\bar{A}\left(\bar{B}^\top - B^\top\right) + \left(\bar{A} - A\right)B^\top\right\|_{\mathrm{Fr}} \\ &\leqslant \left\|\bar{A}\right\|_{\mathrm{Fr}}\left\|\bar{B}^\top - B^\top\right\|_{\mathrm{Fr}} + \left\|\bar{A} - A\right\|_{\mathrm{Fr}}\left\|\bar{B}\right\|_{\mathrm{Fr}} \qquad (80)\\ &\leqslant \sqrt{2t}\varepsilon + \sqrt{2t}\varepsilon = \frac{\epsilon}{2}. \qquad (81)\end{aligned}$$

In particular (by Equations (78)), there must exist an internal $\epsilon$-cover $\mathcal{C}' \subset \mathcal{E}_{\mathrm{r,t}}$ with

$$\begin{aligned}\left|\mathcal{C}'\right| \leqslant |\mathcal{C}| &\leqslant |\mathcal{C}_A|\,|\mathcal{C}_B| \qquad (82)\\ &\leqslant \left(\frac{3\sqrt{2t}}{\epsilon} + 1\right)^{mr}\left(\frac{3\sqrt{2t}}{\epsilon} + 1\right)^{nr} = \left(\frac{3\sqrt{2t}}{\epsilon} + 1\right)^{(m+n)r}. \qquad (83)\end{aligned}$$

Next, we need a simple argument analogous to proofs of Dudley's entropy integrals (however, since the covering number has mild dependence on $\epsilon$, it is not necessary to use a full chaining argument via Lemma E.4, E.3 or E.1). For any $R \in \mathcal{E}_{\mathrm{r,t}}$

let $\bar{R}$ be the closest cover element in $\mathcal{C}'$, we have, for any sample set $\xi^1, \dots, \xi^N \in [m] \times [n]$;

$$\begin{aligned}
&\mathbb{E}_\sigma \sup_{R \in \mathcal{E}_{\mathrm{r,t}}} \frac{1}{N} \sum_{u=1}^{N} \sigma_u \, \mathrm{l}_u(R_{\xi^u}) && (84)\\
&\leqslant \mathbb{E}_\sigma \sup_{R \in \mathcal{E}_{\mathrm{r,t}}} \frac{1}{N} \sum_{u=1}^{N} \sigma_u (\mathrm{l}_u(R_{\xi^u}) - \mathrm{l}_u(\bar{R}_{\xi^u})) + \mathbb{E}_\sigma \sup_{R \in \mathcal{E}_{\mathrm{r,t}}} \frac{1}{N} \sum_{u=1}^{N} \sigma_u \, \mathrm{l}_u(\bar{R}_{\xi^u}) \\
&\leqslant \mathbb{E}_\sigma \sup_{R \in \mathcal{E}_{\mathrm{r,t}}} \frac{1}{N} \sum_{u=1}^{N} \sigma_u (\mathrm{l}_u(R_{\xi^u}) - \mathrm{l}_u(\bar{R}_{\xi^u})) + \mathbb{E}_\sigma \sup_{R \in \mathcal{C}'} \frac{1}{N} \sum_{u=1}^{N} \sigma_u \, \mathrm{l}_u(\bar{R}_{\xi^u}) \\
&\leqslant \epsilon \, \ell + \sqrt{2 \log(|\mathcal{C}'|)} \frac{\mathcal{B}}{\sqrt{N}} && (85)\\
&\leqslant \epsilon \, \ell + \sqrt{\frac{2(m+n)r}{N} \log\left(\frac{3\sqrt{2t}}{\epsilon} + 1\right)} \mathcal{B},
\end{aligned}$$

where the fifth line (85) follows from Proposition F.14. Setting $\epsilon = \frac{1}{N\,\ell}$ yields the result. □

**Theorem D.2.** *Consider the following function class:*

$$\widetilde{\mathcal{F}}_r^p := \left\{ Z \in \mathbb{R}^{m \times n} : \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant r^{1-\frac{p}{2}} \right\}. \tag{86}$$

*With probability $\geqslant 1 - \delta$ over the draw of the training set, the following bound holds on the Rademacher complexity of $\mathrm{l} \circ \widetilde{\mathcal{F}}_r^p$ holds simultaneously over all choices of $\mathrm{l} \in \mathcal{L}_{\ell,\mathcal{B}}$ :*

$$\mathbb{E}_\sigma \sup_{Z \in \widetilde{\mathcal{F}}_r^p} \frac{1}{N} \sum_{o=1}^{N} \sigma_o \, \mathrm{l}_o(Z_{\xi^o}) \leqslant \sqrt{\frac{7\,\mathcal{B}^2 + 1}{N}} + 11\,\mathcal{B}^{\frac{2-2p}{2-p}} \, \ell^{\frac{p}{2-p}} \sqrt{\frac{r(m+n)}{Np}} \log(\Gamma_{\widetilde{\mathcal{F}}_r^p}) \left[1 + \sqrt{\frac{m+n}{N}}\right],$$

*where* $\Gamma_{\widetilde{\mathcal{F}}_r^p} := \frac{6(m+n)N(\ell+1)(r+1)}{\delta}$.

*Proof.* Let us write $t$ for $t = r^{1-\frac{p}{2}}$. Then we have

$$\widetilde{\mathcal{F}}_r^p = \left\{ Z \in \mathbb{R}^{m \times n},\ \|\widetilde{Z}\|_{\mathrm{sc},p}^p \leqslant t \right\}.$$

For any matrix $X$ of rank $k$, let us write $\rho_1(X), \dots, \rho_k(X)$ for the singular values of $X$, ordered from largest to smallest. We can also define, for any $\tau \geqslant 0$, the quantity $U(X) := |\{\kappa \leqslant k : \rho_\kappa \geqslant \tau\}|$. By Markov's inequality, we certainly have for any $X$:

$$U(X) \leqslant \frac{\|X\|_{\mathrm{sc},p}^p}{\tau^p}. \tag{87}$$

Note also that for if $\rho_\kappa \leqslant \tau$ for all $\kappa \geqslant U + 1$, then we certainly have

$$\sum_{\kappa=U+1}^{k} \rho_\kappa = \sum_{\kappa=U+1}^{k} \rho_\kappa^p \rho_\kappa^{1-p} \leqslant \tau^{1-p} \sum_{\kappa=U+1}^{k} \rho_\kappa^p. \tag{88}$$

Thus, applying the decomposition to $X = \widetilde{Z}$ we have the following super-decomposition of the function class $\widetilde{\mathcal{F}}_r^p$:

$$\widetilde{\mathcal{F}}_r^p \subset \mathcal{R}_\tau + \mathcal{T}_\tau, \tag{89}$$

where $\mathcal{R}_\tau$ consists in the contribution from singular values greater than $\tau$ and $\mathcal{T}_\tau$ consists in the contribution from singular values less than $\tau$. Thus, recalling Equation (73),

$$\mathcal{R}_\tau := \left\{ R \in \mathbb{R}^{m \times n} : \|\widetilde{R}\|_* \leqslant \bar{t},\ \mathrm{rank}(R) \leqslant \bar{U} \right\} \tag{90}$$

where $\widetilde{R} = \operatorname{diag}(\sqrt{\tilde{p}}) R \operatorname{diag}(\sqrt{\tilde{q}})$ and

$$\bar{U} := \left\lfloor \frac{r^{\frac{2-p}{2}}}{\tau^p} \right\rfloor \qquad \text{and} \qquad \bar{t} = t^{1/p} n = n r^{\frac{2-p}{2p}}, \tag{91}$$

where the expression for $\bar{U}$ follows from Equation (87) and the expression for $\bar{t}$ follows from the fact that for any matrix $X \in \mathbb{R}^{m\times n}$, $\|X\|_{\mathrm{sc},p} n \geqslant \|X\|_{\mathrm{sc},p} \mathrm{rank}(X) \geqslant \mathrm{rank}(X)\|X\| \geqslant \|X\|_*$. Note that we upper bound the rank by $n$ rather than $\min(m,n)$ for cosmetic reasons only: the corresponding terms will only give rise to logarithmic factors in any case.

For $\mathcal{T}_\tau$, we have

$$\mathcal{T}_\tau := \widetilde{\mathcal{F}}^1_{\tilde{t}} := \left\{ Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_* \leqslant \tilde{t} \right\}, \tag{92}$$

where from Equation (88) we obtain the suitable expression for $\tilde{t}$ as

$$\tilde{t} := t\,\tau^{1-p} = r^{\frac{2-p}{2}}\,\tau^{1-p}. \tag{93}$$

Since Equation (89) holds with the prescribed values of $\bar{U}, \bar{t}$ and $\tilde{t}$ we can upper bound the Rademacher complexity via Lemma E.3 for $\mathcal{F} = \left\{ \left(\mathrm{l}_o((Z^1+Z^2)_{\xi^o})\right)_{o\leqslant N} : Z^1 \in \mathcal{R}_\tau,\ Z^2 \in \mathcal{T}_\tau \right\}$, $\Theta_1 = \mathcal{R}_\tau$ and $\Theta_2 = \mathcal{T}_\tau$:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}^p_r} \frac{1}{N}\sum_{o=1}^N \sigma_o\, \mathrm{l}_u(Z_{\xi^u}) \leqslant \mathbb{E}_\sigma \sup_{Z^1\in\mathcal{R}_\tau, Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \sigma_o\, \mathrm{l}_u(Z^1_{\xi^u} + Z^2_{\xi^u}) \tag{94}$$

$$\leqslant \epsilon + \sup_{\bar{Z}^1\in\mathcal{C}(\epsilon)} \hat{\mathfrak{R}}\left(\mathcal{F}_{\bar{Z}^1}\right) + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1,\epsilon))}{N}}, \tag{95}$$

where $\mathcal{C}(\epsilon)$ is an $\epsilon$-uniform cover of $\Theta_1 = \mathcal{R}_\tau$ in the sense of Lemma E.3, and for the avoidance of doubt,

$$\mathcal{F}_{\bar{Z}^1} = \left\{ \left(\mathrm{l}_o((\bar{Z}^1+Z^2)_{\xi^o})\right)_{o\leqslant N} :\ Z^2 \in \mathcal{T}_\tau \right\}.$$

The above equation (95) holds for any $\epsilon$, and since the loss function is uniformly Lipschitz, it is certainly true that if $\bar{Z}^1$ is a cover element such that $\left|\left(\bar{Z}^1 - Z^1\right)_{\xi^o}\right| \leqslant \epsilon$ for any $o$, then we also have that

$$\left|\mathrm{l}_o((Z^1+Z^2)_{\xi^o}) - \mathrm{l}_o((\bar{Z}^1+Z^2)_{\xi^o})\right| \leqslant \epsilon\,\ell \tag{96}$$

holds uniformly over all $Z^2 \in \mathcal{T}_\tau$, all $o$s and all loss functions ls satisfying the required conditions.

Thus

$$\mathcal{N}(\mathcal{F}_1,\epsilon) \leqslant \mathcal{N}_\infty\left(\mathcal{R}_\tau, \frac{\epsilon}{\ell}\right). \tag{97}$$

Note also that since $\tilde{p}_i \geqslant \frac{1}{2m}$ and $\tilde{q}_j \geqslant \frac{1}{2n}$, for any matrix $X$ with $\|\widetilde{X}\|_* \leqslant \bar{t}$, we have $\|X\|_* \leqslant \sqrt{n}\|\widetilde{X}\|_{\mathrm{Fr}} 2\sqrt{mn} \leqslant 2\sqrt{mn^2\bar{t}}$. Thus we have $\mathcal{R}_\tau \subset \left\{ R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant 2\sqrt{mn^2\bar{t}},\ \mathrm{rank}(R) \leqslant \bar{U} \right\}$. Thus, using Lemma (D.1) and plugging the result

back into Equation (95) above, we obtain:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o \, \mathrm{l}_u(Z_{\xi^u}) \tag{98}$$

$$\leqslant \epsilon + \sup_{\bar{Z}^1\in\mathcal{C}(\epsilon)} \mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \sigma_o \, \mathrm{l}_o(Z^2_{\xi^o} + \bar{Z}^1_{\xi^o}) \tag{99}$$

$$+ \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(\frac{3\sqrt{4\sqrt{mn^2}\bar{t}}\,\ell}{\epsilon}+1\right)}{N}}$$

$$\leqslant \epsilon + \ell\,\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \sigma_o Z_{\xi^u} + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(\frac{6\sqrt{\sqrt{mn^2}nr^{\frac{2-p}{2p}}}\,\ell}{\epsilon}+1\right)}{N}} \tag{100}$$

$$\leqslant \frac{1}{N} + \ell\,\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \sigma_o Z_{\xi^u} \tag{101}$$

$$+ \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{2r^{1-\frac{p}{2}}(m+n)\log\left(6(m+n)N(\ell+1)(r+1)\right)}{Np\,\tau^p}}$$

where the second line (99) we used Lemma D.1, at the next line (100) we have used the Talagrand contraction Lemma (once for each value of $\bar{Z}^1$), and at the final line (101), we have set $\epsilon = \frac{1}{N}$, used the fact that $\bar{U} \leqslant \frac{t^p}{\tau^p} = \frac{r^{1-\frac{p}{2}}}{\tau^p}$ and the fact that $p \leqslant 1$.

Next, by applying Theorem F.5 (with $r \leftarrow \widetilde{t}^2$) we know that with probability $\geqslant 1-\delta$ over the draw of the training set we have

$$\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \sigma_o Z_{\xi^u} \leqslant 4\sqrt{\frac{2\widetilde{t}^2(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} + \frac{16\sqrt{mn\widetilde{t}}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{102}$$

$$\leqslant 4\,\tau^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} + \frac{16\,\tau^{1-p}\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{103}$$

where at line (103) we used the definition of $\widetilde{t}$ (cf. Equation (93)). Plugging Equation (103) back into Equation (101) we obtain the following bound for the Rademacher complexity of $\widetilde{\mathcal{F}}_r^p$, which holds with probability $\geqslant 1-\delta$ over the draw of the training set:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o \, \mathrm{l}_u(Z_{\xi^u}) \tag{104}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\ell\,\tau^{1-p}\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{105}$$

$$+ \mathcal{B}\sqrt{2\frac{r^{1-\frac{p}{2}}(m+n)\log\left(6(m+n)N(\ell+1)(r+1)\right)}{Np\,\tau^p}}$$

$$+ 4\,\ell\,\tau^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)}.$$

Ignoring logarithmic factors, the dominant terms scale as follows:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o \, \mathrm{l}_u(Z_{\xi^u}) \leqslant \widetilde{O}\left(\mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)}{Np\,\tau^p}} + \ell\,\tau^{1-p}r^{1-\frac{p}{2}}\sqrt{\frac{m+n}{N}}\right). \tag{106}$$

Thus, optimizing over the choice of $\tau$, we select

$$\tau = \mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}p^{-\frac{1}{2-p}}. \tag{107}$$

Substituting this back into Equation (105) we obtain $\mathfrak{R}(\widetilde{\mathcal{F}}_r^p) \leqslant$

$$\begin{aligned}
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\ell\,\tau^{1-p}\,\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) + 4\,\ell\,\tau^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} \\
&\qquad + \mathcal{B}\sqrt{\frac{2r^{1-\frac{p}{2}}(m+n)\log\left(6(m+n)N(\ell+1)(r+1)\right)}{Np\,\tau^p}}. \\
&= \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\ell[\mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}p^{-\frac{1}{2-p}}]^{1-p}\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) \\
&\qquad + \mathcal{B}\sqrt{\frac{2r^{1-\frac{p}{2}}(m+n)\log\left(6(m+n)N(\ell+1)(r+1)\right)}{Np[\mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}p^{-\frac{1}{2-p}}]^p}} \\
&\qquad + 4\,\ell[\mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}p^{-\frac{1}{2-p}}]^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} \\
&= \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\,r^{\frac{1}{2}}\sqrt{mn}}{3Np^{\frac{1-p}{2-p}}}\log\left(\frac{m+n}{\delta}\right) \\
&\quad + \mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\left(\sqrt{\frac{2r(m+n)\log\left(6(m+n)N(\ell+1)(r+1)\right)}{Np^{\frac{2-2p}{2-p}}}} + 4\sqrt{\frac{2r(m+n)}{3Np^{\frac{2-2p}{2-p}}}\log\left(\frac{m+n}{\delta}\right)}\right) \\
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\,r^{\frac{1}{2}}\sqrt{mn}\log\left(\frac{m+n}{\delta}\right)}{3Np^{\frac{1-p}{2-p}}} \qquad (108) \\
&\qquad + 5\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)\log\left(\frac{6(m+n)N(\ell+1)(r+1)}{\delta}\right)}{Np^{\frac{2-2p}{2-p}}}} \\
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 11\frac{\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\,p^{\frac{p}{2(2-p)}}}{p^{\frac{1}{2}}}\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}_r^p})\left[1+\sqrt{\frac{m+n}{N}}\right] \qquad (109) \\
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 11\frac{\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}}{p^{\frac{1}{2}}}\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}_r^p})\left[1+\sqrt{\frac{m+n}{N}}\right], \qquad (110)
\end{aligned}$$

as expected. □

Next, we consider the situation where the distribution is arbitrary, but the Schatten quasi norm is not weighted:

**Theorem D.3.** *Consider the following function class:*

$$\mathcal{F}_{\mathrm{t}}^{\mathrm{p}} := \left\{Z \in \mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M}\right\}. \tag{111}$$

*With probability $\geqslant 1-\delta$, we have the following bound on the Rademacher complexity of $\mathrm{l}\circ\mathcal{F}_{\mathrm{t}}^{\mathrm{p}}$, where $\mathrm{l}$ is any set of $N$ $\ell$-Lipschitz functions uniformly bounded by $\mathcal{B}$:*

$$\widehat{\mathfrak{R}}(\widetilde{\mathcal{F}}_r^p) = \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + (\sqrt{18C}+\sqrt{2})\,\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}\log\left(\Gamma_{\mathcal{F}_{\mathrm{t}}^{\mathrm{p}},\ell}\right)}{Np}}. \tag{112}$$

*where $\Gamma_{\mathcal{F}_{\mathrm{t}}^{\mathrm{p}},\ell} := 6N(m+n)(r+1)(\ell+1)$.*

*Proof.* Similarly to the above proofs we can write

$$\widetilde{\mathcal{F}}_r^p \subset \mathcal{R}_\tau + \mathcal{T}_\tau, \tag{113}$$

where

$$\mathcal{R}_\tau := \left\{R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant \bar{t},\ \text{rank}(R) \leqslant \bar{U}\right\} \tag{114}$$

with

$$\bar{U} = \left\lfloor \frac{r^{1-\frac{p}{2}}\sqrt{mn}^p}{\tau^p} \right\rfloor \quad \text{and} \qquad \bar{t} = \sqrt{mn}r^{\frac{2-p}{2p}}n. \tag{115}$$

And similarly

$$\mathcal{T}_\tau := \left\{Z \in \mathbb{R}^{m\times n} : \|Z\|_* \leqslant \widetilde{t}\right\} \tag{116}$$

with

$$\widetilde{t} = r^{1-\frac{p}{2}}\sqrt{mn}^p\,\tau^{1-p}\,. \tag{117}$$

Thus, by the same argument as in the proof of Theorem D.2 we have

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o\, \mathrm{l}_o(Z_{\xi^o}) \tag{118}$$

$$\leqslant \frac{1}{N} + \sup_{\bar{Z}^1\in\mathcal{C}(1/N)} \widehat{\mathfrak{R}}\left(\mathcal{F}_{\bar{Z}^1}\right) + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(3N\sqrt{4\sqrt{mn^2\bar{t}}\,\ell+1}\right)}{N}} \tag{119}$$

where

$$\mathcal{F}_{\bar{Z}^1} = \left\{\left(\mathrm{l}_o((\bar{Z}^1 + Z^2)_{\xi^o})\right)_{o\leqslant N} :\ Z^2 \in \mathcal{T}_\tau\right\}. \tag{120}$$

with the values for $\mathcal{T}_\tau, \bar{U}, \bar{t}, \widetilde{t}$ defined as per Equations (116), (115) and (117) respectively. Replacing the appropriate values for $\bar{U}$ and $\bar{t}$, we have the following

$$\mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(3N\sqrt{4\sqrt{mn^2\bar{t}}\,\ell+1}\right)}{N}} \tag{121}$$

$$\leqslant \mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}\sqrt{mn}^p(m+n)\log\left(3N\sqrt{4\sqrt{mn^2}\sqrt{mn}r^{\frac{2-p}{2p}}n}\,\ell+1\right)}{N\,\tau^p}} \tag{122}$$

$$\leqslant \mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{p+1}\log\left(3N\sqrt{4\sqrt{mn^2}\sqrt{mn}r^{\frac{2-p}{2p}}n}\,\ell+1\right)}{N\,\tau^p}} \tag{123}$$

$$\leqslant \mathcal{B}\sqrt{2\frac{r^{1-\frac{p}{2}}(m+n)^{p+1}\log\left(6N(m+n)(r+1)(\ell+1)\right)}{Np\,\tau^p}}. \tag{124}$$

For any $\bar{Z}^1 \in \mathcal{C}(\epsilon)$ we can apply Proposition F.3 from (Shamir & Shalev-Shwartz, 2011) to obtain the following inequality, which is valid for **any** training set:

$$\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \mathrm{l}_o(\bar{Z}^1 + Z^2) \leqslant \sqrt{\frac{9C\,\mathcal{B}\,\ell\,\widetilde{t}(\sqrt{m}+\sqrt{n})}{N}} = \sqrt{\frac{9C\,\mathcal{B}\,\ell\,r^{1-\frac{p}{2}}\sqrt{mn}^p\,\tau^{1-p}(\sqrt{m}+\sqrt{n})}{N}}.$$

Taking a supremememum over $\bar{Z}^1$ we certainly have

$$\sup_{\bar{Z}^1 \in \mathcal{C}(\epsilon)} \mathbb{E}_\sigma \sup_{Z^2 \in \mathcal{T}_\tau} \frac{1}{N} \sum_{o=1}^{N} \mathrm{l}_o(\bar{Z}^1 + Z^2) \leqslant \sqrt{\frac{9C\,\mathcal{B}\,\ell\, r^{1-\frac{p}{2}} \sqrt{mn}^p\, \tau^{1-p}(\sqrt{m}+\sqrt{n})}{N}}. \tag{125}$$

Plugging Equations (125) and (124) back into Equation (119) we obtain:

$$\mathbb{E}_\sigma \sup_{Z \in \widetilde{\mathcal{F}}_r^p} \frac{1}{N} \sum_{o=1}^{N} \sigma_o\, \mathrm{l}_o(Z_{\xi^o}) \tag{126}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \sqrt{\frac{18C\,\mathcal{B}\,\ell\, r^{1-\frac{p}{2}}\, \tau^{1-p}(\sqrt{m+n})^{1+2p}}{N}} \tag{127}$$

$$+ \mathcal{B}\sqrt{2\frac{r^{1-\frac{p}{2}}(m+n)^{p+1}\log\left(6N(m+n)(r+1)(\ell+1)\right)}{Np\,\tau^p}}.$$

this motivates the following choice of threshold

$$\tau := \mathcal{B}\,\ell^{-1}\,p^{-1}\sqrt{m+n}, \tag{128}$$

which yields:

$$\mathbb{E}_\sigma \sup_{Z \in \widetilde{\mathcal{F}}_r^p} \frac{1}{N} \sum_{o=1}^{N} \sigma_o\, \mathrm{l}_o(Z_{\xi^o}) \tag{129}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \sqrt{\frac{18C\,\mathcal{B}^{2-p}\,\ell^p\, r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}}{Np^{1-p}}} \tag{130}$$

$$+ \mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{2\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}\log\left(6N(m+n)(r+1)(\ell+1)\right)}{Np^{1-p}}}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + (\sqrt{18C}+\sqrt{2})\,\mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)^{1+\frac{p}{2}}\left(6N(m+n)(r+1)(\ell+1)\right)}{Np}},$$

as expected. □

### D.2. With Constraints on Entries

**Theorem D.4.** *Consider the following function class:*

$$\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} := \left\{ Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|^p_{\mathrm{sc},p} \leqslant r^{1-\frac{p}{2}};\ \|Z\|_\infty \leqslant \mathcal{B}_0 \right\}. \tag{131}$$

*With probability $\geqslant 1-\delta$, we have the following bound on the Rademacher complexity of $\mathrm{l}\circ\widetilde{\mathcal{F}}^p_r$, where $\mathrm{l}$ a Lipschitz function of $\xi^o$ and $\widetilde{G}_o$ but is $\ell$-Lipschitz with respect to the first argument and uniformly bounded by $\mathcal{B}$:*

$$\hat{\mathfrak{R}}(\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0}) = \mathbb{E}_\sigma \sup_{Z \in \widetilde{\mathcal{F}}_r^p} \frac{1}{N} \sum_{o=1}^{N} \sigma_o\, \mathrm{l}_o(Z_{\xi^o}) \tag{132}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 11\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell})\left[1+\sqrt{\frac{(m+n)}{N}}\right].$$

*Proof.* Let us write $t$ for $t = r^{1-\frac{p}{2}}$. Then we have

$$\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0} = \left\{ Z \in \mathbb{R}^{m\times n},\ \|\widetilde{Z}\|^p_{\mathrm{sc},p} \leqslant t, \|Z\|_\infty \leqslant \mathcal{B}_0 \right\}.$$

As before, we decompose our function class into two parts

$$\widetilde{\mathcal{F}}^{p}_{r,\mathcal{B}_0} \subset \mathcal{R}_\tau + \mathcal{T}_\tau, \tag{133}$$

where $\mathcal{R}_\tau$ is a class of rank $U$ matrices where the contribution from singular values greater than $\tau$ belong and $\mathcal{T}_\tau$ contains in the contribution from singular values less than $\tau$.

More precisely,

$$\mathcal{R}_\tau := \left\{R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant \bar{t},\ \text{rank}(R) \leqslant \bar{U}\right\} \tag{134}$$

where $\widetilde{R} = \text{diag}(\sqrt{\bar{p}})R\,\text{diag}(\sqrt{\bar{q}})$ and

$$\bar{U} := \left\lfloor \frac{r^{\frac{2-p}{2}}}{\tau^p} \right\rfloor \qquad \text{and} \qquad \bar{t} = \sqrt{mnn}\,\mathcal{B}_0, \tag{135}$$

where the expression for $\bar{U}$ comes from Markov's inequality (as before) and (differently from before) the expression for $\bar{t}$ comes from the fact that for any matrix $R$,

$$\sum_{\rho_i \geqslant \tau} \rho^i u^i (v^i)^\top \leqslant \|Z\|\text{rank}(Z) \leqslant \sqrt{mn}\,\mathcal{B}_0\,\text{rank}(Z) \leqslant \sqrt{mnn}\,\mathcal{B}_0\,. \tag{136}$$

Note also that exactly as before, for if $\rho_\kappa \leqslant \tau$ for all $\kappa \geqslant U+1$, then we certainly have

$$\sum_{\kappa=U+1}^{k} \rho_\kappa = \sum_{\kappa=U+1}^{k} \rho_\kappa^p \rho_\kappa^{1-p} \leqslant \tau^{1-p} \sum_{\kappa=U+1}^{k} \rho_\kappa^p. \tag{137}$$

Thus, for $\mathcal{T}_\tau$, we have

$$\mathcal{T}_\tau := \widetilde{\mathcal{F}}^{1}_{\widetilde{t}} := \left\{Z \in \mathbb{R}^{m\times n} : \|\widetilde{Z}\|_* \leqslant \widetilde{t}\right\}, \tag{138}$$

where from equation (137) we obtain the suitable expression for $\widetilde{t}$ as

$$\widetilde{t} := t\,\tau^{1-p} = r^{\frac{2-p}{2}}\,\tau^{1-p}\,. \tag{139}$$

Since Equation (133) holds with the prescribed values of $\bar{U}, \bar{t}$ and $\widetilde{t}$ we can upper bound the Rademacher complexity via Lemma E.3 for $\mathcal{F} = \left\{\left(\mathrm{l}_o((Z^1+Z^2)_{\xi^o})\right)_{o\leqslant N} : Z^1 \in \mathcal{R}_\tau,\ Z^2 \in \mathcal{T}_\tau\right\}$, $\Theta_1 = \mathcal{R}_\tau$ and $\Theta_2 = \mathcal{T}_\tau$:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}^p_r} \frac{1}{N}\sum_{o=1}^{N} \sigma_o\,\mathrm{l}_u(Z_{\xi^u}) \leqslant \mathbb{E}_\sigma \sup_{Z^1\in\mathcal{R}_\tau, Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^{N} \sigma_o\,\mathrm{l}_u(Z^1_{\xi^u} + Z^2_{\xi^u}) \tag{140}$$

$$\leqslant \epsilon + \sup_{\bar{Z}^1\in\mathcal{C}(\epsilon)} \widehat{\mathfrak{R}}\,(\mathcal{F}_{\bar{Z}^1}) + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1,\epsilon))}{N}}. \tag{141}$$

By the same arguments as in the proof of Theorem D.2,

$$\mathcal{N}(\mathcal{F}_1,\epsilon) \leqslant \mathcal{N}_\infty\left(\mathcal{R}_\tau, \frac{\epsilon}{\ell}\right). \tag{142}$$

Thus, using Lemma D.1 and plugging the result back into Equation (141) above, we obtain:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^{N}\sigma_o \mathrm{l}_u(Z_{\xi^u}) \tag{143}$$

$$\leqslant \epsilon + \sup_{\bar{Z}^1\in\mathcal{C}(\epsilon)} \mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^{N}\sigma_o \mathrm{l}_o(Z^2_{\xi^o}+\bar{Z}^1_{\xi^o})\,(\mathcal{F}_{\bar{Z}^1}) \tag{144}$$

$$+\mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(\frac{3\sqrt{\sqrt{mn^3}\,\mathcal{B}_0}\,\ell}{\epsilon}+1\right)}{N}}$$

$$\leqslant \frac{1}{N} + \ell\,\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^{N}\sigma_o Z_{\xi^u} + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)\log\left(3Nmn^3[\mathcal{B}_0+1][\ell+1]+1\right)}{N\,\tau^p}} \tag{145}$$

at the final line (145), we have set $\epsilon = \frac{1}{N}$, used the fact that $\bar{U} \leqslant \frac{t^p}{\tau^p} = \frac{r^{1-\frac{p}{2}}}{\tau^p}$ .

Next, the application of Theorem F.5 (with $r \leftarrow \widetilde{t}^2$) to $\mathcal{T}_\tau$ is unchanged from the proof of Theorem D.2, thus we know as before that with probability $\geqslant 1-\delta$ over the draw of the training set we have

$$\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^{N}\sigma_o Z_{\xi^u} \leqslant 4\sqrt{\frac{2\widetilde{t}^2(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} + \frac{16\sqrt{mn}\widetilde{t}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{146}$$

$$\leqslant 4\,\tau^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} + \frac{16\,\tau^{1-p}\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{147}$$

where at line (147) we used the definition of $\widetilde{t}$ (cf. Equation (139)). Plugging Equation (147) back into Equation (145) we obtain the following bound for the Rademacher complexity of $\widetilde{\mathcal{F}}_r^p$, which holds with probability $\geqslant 1-\delta$ over the draw of the training set:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^{N}\sigma_o \mathrm{l}_u(Z_{\xi^u}) \tag{148}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\ell\,\tau^{1-p}\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{149}$$

$$+\mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)\log\left(3Nmn^3[\mathcal{B}_0+1][\ell+1]+1\right)}{N\,\tau^p}} + 4\,\ell\,\tau^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)}.$$

Thus, optimizing over the choice of $\tau$ (ignoring logarithmic factors), we select

$$\tau = \mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}. \tag{150}$$

Substituting this back into Equation (149) and writing $\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}$ for $\frac{3Nmn^3[\mathcal{B}_0+1][\ell+1]+1}{\delta} \geqslant 3Nmn^3[\mathcal{B}_0+1][\ell+1]+1$ we

obtain $\mathfrak{R}(\widetilde{\mathcal{F}}_r^p) \leqslant$

$$
\begin{aligned}
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\ell\,\tau^{1-p}\,\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right) \qquad (151)\\
&\quad + \mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)\log\left(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}\right)}{N\,\tau^p}} + 4\,\ell\,\tau^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)}.\\
&= \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\ell[\mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}]^{1-p}\sqrt{mn}r^{\frac{2-p}{2}}}{3N}\log\left(\frac{m+n}{\delta}\right)\\
&\quad + \mathcal{B}\sqrt{\frac{r^{1-\frac{p}{2}}(m+n)\log\left(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}\right)}{N[\mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}]^p}} + 4\,\ell[\mathcal{B}^{\frac{2}{2-p}}\,\ell^{-\frac{2}{2-p}}\,r^{-\frac{1}{2}}]^{(1-p)}\sqrt{\frac{2r^{2-p}(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)}\\
&= \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\,r^{\frac{1}{2}}\sqrt{mn}}{3N}\log\left(\frac{m+n}{\delta}\right)\\
&\qquad + \mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\left(\sqrt{\frac{r(m+n)\log\left(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}\right)}{N}} + 4\sqrt{\frac{2r(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)}\right)\\
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \frac{16\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\,r^{\frac{1}{2}}\sqrt{mn}\log\left(\frac{m+n}{\delta}\right)}{3N} + 5\,\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\sqrt{\frac{r(m+n)\log\left(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell}\right)}{N}},\\
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + 11\left[\mathcal{B}^{\frac{2-2p}{2-p}}\,\ell^{\frac{p}{2-p}}\right]\sqrt{\frac{r(m+n)}{N}}\log(\Gamma_{\widetilde{\mathcal{F}}^p_{r,\mathcal{B}_0},\ell})\left[1+\sqrt{\frac{(m+n)}{N}}\right]
\end{aligned}
$$

as expected. □

**Theorem D.5.** *Consider the following function class:*

$$\mathcal{F}^p_{r,\mathcal{B}_0} := \left\{Z\in\mathbb{R}^{m\times n} : \|Z\|_{\mathrm{sc},p} \leqslant \mathcal{M} = \left[r\sqrt{mn}\right]^{\frac{2-p}{2p}},\ \|Z\|_\infty \leqslant \mathcal{B}_0\right\} \qquad (152)$$

*With probability $\geqslant 1-\delta$, we have the following bound on the Rademacher complexity of $\mathrm{l}\circ\mathcal{F}^{\mathrm{p}}_{\mathrm{t}}$, where $\mathrm{l}$ is any set of $N$ $\ell$-Lipschitz functions uniformly bounded by $\mathcal{B}$:*

$$
\begin{aligned}
&\hat{\mathfrak{R}}(\mathcal{F}^p_{r,\mathcal{B}_0}) = \mathbb{E}_\sigma \sup_{Z\in\mathcal{F}^{\mathrm{p}}_{\mathrm{t}}}\frac{1}{N}\sum_{o=1}^N \sigma_o\,\mathrm{l}_o(Z_{\xi^o}) \qquad (153)\\
&\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{2\,\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{N}}\left(3\sqrt{C}+\sqrt{\log\left(6(m+n)N[\mathcal{B}_0+1][\ell+1]\right)}\right),
\end{aligned}
$$

*where $C$ is the absolute constant from (Latała, 2005).*

*Note that similarly to Theorem D.3, the sample complexity is*

$$\widetilde{O}\left(\mathcal{M}^p(m+n)^p\right) = \widetilde{O}\left(r^{1-\frac{p}{2}}n^{1+\frac{p}{2}}\right), \qquad (154)$$

*where $r$ is the rank-like quantity $\left[\frac{\mathcal{M}}{\sqrt{mn}}\right]^{\frac{2p}{2-p}}$.*

*Proof.* Similarly to the above proofs we can write

$$\widetilde{\mathcal{F}}_r^p \subset \mathcal{R}_\tau + \mathcal{T}_\tau, \tag{155}$$

where

$$\mathcal{R}_\tau := \left\{R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant \bar{t},\ \text{rank}(R) \leqslant \bar{U}\right\} \tag{156}$$

with

$$\bar{U} = \left\lfloor \frac{\mathcal{M}^p}{\tau^p} \right\rfloor \quad \text{and} \qquad \bar{t} = \mathcal{B}_0\sqrt{mn}n. \tag{157}$$

And similarly

$$\mathcal{T}_\tau := \left\{Z \in \mathbb{R}^{m\times n} : \|Z\|_* \leqslant \widetilde{t}\right\} \tag{158}$$

with

$$\widetilde{t} = \mathcal{M}^p\,\tau^{1-p}\,. \tag{159}$$

Thus, by the same argument as in the proof of Theorem D.2 we have

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o\, \mathrm{l}_o(Z_{\xi^o}) \tag{160}$$

$$\leqslant \frac{1}{N} + \sup_{\bar{Z}^1\in\mathcal{C}(1/N)} \widehat{\mathfrak{R}}\left(\mathcal{F}_{\bar{Z}^1}\right) + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(3N\sqrt{2\bar{t}}\,\ell+1\right)}{N}} \tag{161}$$

where

$$\mathcal{F}_{\bar{Z}^1} = \left\{\left(\mathrm{l}_o((\bar{Z}^1 + Z^2)_{\xi^o})\right)_{o\leqslant N} :\ Z^2 \in \mathcal{T}_\tau\right\}. \tag{162}$$

with the values for $\mathcal{T}_\tau, \bar{U}, \bar{t}, \widetilde{t}$ defined as per Equations (158), (157) and (159) respectively. For any $\bar{Z}^1 \in \mathcal{C}(\epsilon)$ we can apply Proposition F.3 from (Shamir & Shalev-Shwartz, 2011) to obtain the following inequality, which is valid for **any** training set:

$$\mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \mathrm{l}_o(\bar{Z}^1 + Z^2) \leqslant \sqrt{\frac{9C\,\mathcal{B}\,\ell\,\widetilde{t}(\sqrt{m}+\sqrt{n})}{N}} = \sqrt{\frac{9C\,\mathcal{B}\,\ell\,\mathcal{M}^p\,\tau^{1-p}(\sqrt{m}+\sqrt{n})}{N}}. \tag{163}$$

Taking a supremememum over $\bar{Z}^1$ we certainly have

$$\sup_{\bar{Z}^1\in\mathcal{C}(\epsilon)} \mathbb{E}_\sigma \sup_{Z^2\in\mathcal{T}_\tau} \frac{1}{N}\sum_{o=1}^N \mathrm{l}_o(\bar{Z}^1 + Z^2) \leqslant \sqrt{\frac{9C\,\mathcal{B}\,\ell\,\mathcal{M}^p\,\tau^{1-p}(\sqrt{m}+\sqrt{n})}{N}} \tag{164}$$

where $C$ is the absolute constant in (Latała, 2005). Furthermore, replacing the appropriate values for $\bar{U}$ and $\bar{t}$, we have the following

$$\mathcal{B}\sqrt{\frac{\bar{U}(m+n)\log\left(3N\sqrt{2\bar{t}}\,\ell+1\right)}{N}} \leqslant \mathcal{B}\sqrt{\frac{2\,\mathcal{M}^p(m+n)\log\left(6(m+n)N[\mathcal{B}_0+1][\ell+1]\right)}{N\,\tau^p}}. \tag{165}$$

Plugging Equations (164) and (165) back into Equation (161) we get:

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o \, \mathrm{l}_o(Z_{\xi^o}) \tag{166}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \sqrt{\frac{18C\,\mathcal{B}\,\ell\,\mathcal{M}^p\,\tau^{1-p}(\sqrt{m+n})}{N}} \tag{167}$$

$$+ \mathcal{B}\sqrt{\frac{2\,\mathcal{M}^p(m+n)\log\left(6(m+n)N[\mathcal{B}_0+1][\ell+1]\right)}{N\,\tau^p}}.$$

This motivates the following choice of threshold:

$$\tau = \ell^{-1}\,\mathcal{B}^1\,\sqrt{m+n}, \tag{168}$$

which yields

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^p} \frac{1}{N}\sum_{o=1}^N \sigma_o \, \mathrm{l}_o(Z_{\xi^o}) \tag{169}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \sqrt{\frac{18C\,\mathcal{B}\,\ell\,\mathcal{M}^p\,\tau^{1-p}(\sqrt{m+n})}{N}} + \mathcal{B}\sqrt{2\frac{\mathcal{M}^p(m+n)\log\left(6(m+n)N[\mathcal{B}_0+1][\ell+1]\right)}{N\,\tau^p}}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \sqrt{\frac{18C\,\mathcal{B}^{2-p}\,\ell^p\,\mathcal{M}^p(\sqrt{m+n})^{1-\frac{p}{2}}}{N}}$$

$$+ \sqrt{2\frac{\mathcal{M}^p(m+n)^{1-\frac{p}{2}}\,\mathcal{B}^{2-p}\,\ell^p\log\left(6(m+n)N[\mathcal{B}_0+1][\ell+1]\right)}{N}}$$

$$\leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}} + \mathcal{B}^{1-\frac{p}{2}}\,\ell^{\frac{p}{2}}\sqrt{\frac{\mathcal{M}^p(m+n)^{1-\frac{p}{2}}}{N}}\left(\sqrt{18C} + \sqrt{2\log\left(6(m+n)N[\mathcal{B}_0+1][\ell+1]\right)}\right),$$

where as usual at the second line we have used the fact that $\frac{1}{N} + \mathcal{B}\sqrt{\frac{2\pi}{N}} \leqslant \sqrt{\frac{7\,\mathcal{B}^2+1}{N}}$.

□

## E. Important Tools

In this section, we collect some of the main building blocks of our proofs, which include estimates of perturbations of Schatten quasi-norms in the estimation of the marginals, various tailor-made generalizations of chaining arguments, and generalization bounds for classes of neural embeddings.

### E.1. Generalizations of Dudley's Entropy Theorem

This subsection details some of our results on how to calculate Rademacher complexities of composite function classes when a covering number is only available for one of the classes.

Recall the following standard form of Dudley's entropy integral:

**Lemma E.1.** *Let $\mathcal{F}$ be a real-valued function class taking values in $[0,1]$, and assume that $0\in\mathcal{F}$. Let $S$ be a finite sample of size $n$. We have the following relationship between the Rademacher complexity $\mathfrak{R}(\mathcal{F}|_S)$ and the $L^2$ covering number $\mathcal{N}(\mathcal{F}|S,\epsilon,\|\cdot\|_2)$.*

$$\mathfrak{R}(\mathcal{F}|_S) \leqslant \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{N}}\int_\alpha^1 \sqrt{\log\mathcal{N}(\mathcal{F}|S,\epsilon,\|\cdot\|_2)}\right)d\epsilon,$$

*where the norm $\|\cdot\|_2$ on $\mathbb{R}^N$ is defined by $\|x\|_2^2 = \frac{1}{N}(\sum_{i=1}^N |x_i|^2)$.*

We will also to extend the above Lemma to various settings where several a function class over two parameters. For this, we will first need to establish the following slight extension of Lemma F.19:

**Lemma E.2** (Scale sensitive concentration of Rademacher complexity)**.** *For any fixed $x_1, \dots, x_N$ and any function class $\mathcal{F} \subset \mathbb{R}^{[N]}$. Assume that there are $N$ numbers $s_1, s_2, \dots, s_N > 0$ such that for all $f \in \mathcal{F}$ we have*

$$|f_i| \leqslant s_i \tag{170}$$

*with*

$$\frac{1}{N}\sum_{i=1}^{N} s_i^2 = c^2$$

*for some $c > 0$. We have with probability $\geqslant 1 - \delta$ over the draw of the Rademacher variables $\sigma_1, \dots, \sigma_N$,*

$$\left|\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f_i - \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f_i\right| < c\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{171}$$

*Proof.* This is a direct application of the Mc Diarmid inequality with the variables being the $\sigma_1, \dots, \sigma_N$. Indeed, if $\sigma, \bar{\sigma} \in \{-1, 1\}^N$ differ only in the ith component $\sigma_i$, then we certainly have

$$\left|\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f_i - \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \bar{\sigma}_i f_i\right| \leqslant 2s_i, \tag{172}$$

which means that Mc Diarmid's inequality implies

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f_i - \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f_i\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{2\epsilon^2}{4c^2}\right) = 2\exp\left(-\frac{\epsilon^2}{2c^2}\right). \tag{173}$$

The lemma follows upon rearranging. □

The two Lemmas below perform a generalized "chaining argument" (Guermeur, 2018; Vershynin, 2018; Boucheron et al., 2004) with multiple component function classes. They are extensions of Proposition A.4. (page 3 of the supplementary) of (Ledent et al., 2021b) and allow one to bound the Rademacher complexity of a composition of two function classes in terms of the uniform *covering number* of the *second* class and the *Rademacher complexity* of the *first* one.

**Lemma E.3** (Multi-class chaining: simple compositional extension of Dudley's entropy theorem)**.** *Let $\mathcal{F} := \{f_i(\theta_1, \theta_2) : \theta_a \in \Theta_1, \theta_2 \in \Theta_2\}$ be a class of functions on $\{1, 2, \dots, n\}$ with values in $[-\mathcal{B}, \mathcal{B}]$ and dependent on two parameters $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$.*

*Let $\epsilon \geqslant 0$ and assume that $\Theta_1$ admits an $\epsilon$-uniform cover $\mathcal{C}(\epsilon) \subset \Theta_1$ (of size $\mathcal{N}(\mathcal{F}_1, \epsilon)$, dependent on $\epsilon$) in the following sense: For any $\theta_1 \in \Theta_1$ there exists a $\bar{\theta} \in \mathcal{C}(\epsilon)$ such that for all $\theta_2 \in \Theta_2$ and for all $i \leqslant n$ we have*

$$\left|f_i(\theta_1, \theta_2) - f_i(\bar{\theta}, \theta_2)\right| \leqslant \epsilon. \tag{174}$$

*Then we have the following result on the Rademacher complexity of the function class $\mathcal{F}$:*

$$\widehat{\mathfrak{R}}(\mathcal{F}) = \mathbb{E}_\sigma \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f_i(\theta_1, \theta_2) \leqslant \epsilon + \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \widehat{\mathfrak{R}}\left(\mathcal{F}_{\bar{\theta}}\right) + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1, \epsilon))}{N}}. \tag{175}$$

*where for all $\bar{\theta} \in \Theta_1$ we define $\mathcal{F}_{\bar{\theta}} := \left\{f_\bullet(\bar{\theta}, \theta_2) : \theta_2 \in \Theta_2\right\}$, and as usual the $\sigma_i$s are i.i.d. Rademacher variables.*

*Proof.* Without loss of generality, we can assume $\mathcal{B} = 1$.

For any $\theta_1 \in \Theta_1$ let us denote by $\bar{\theta}_1$ the cover element associated to $\theta_1$ as in equation (174), by the assumption on the cover we have

$$\begin{aligned}
\widehat{\mathfrak{R}}(\mathcal{F}) &= \mathbb{E}_\sigma \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\theta_1,\theta_2) \\
&\leqslant \mathbb{E}_\sigma \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta}_1,\theta_2) + [f_i(\theta_1,\theta_2) - f_i(\bar{\theta}_1,\theta_2)] \\
&\leqslant \epsilon + \mathbb{E}_\sigma \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta}_1,\theta_2) \\
&\leqslant \epsilon + \mathbb{E}_\sigma \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2). \qquad (176)
\end{aligned}$$

By Lemma F.19 we have, for any choice of $\bar{\theta} \in \Theta_1$, that with probability $\geqslant 1-\delta$ over the draw of the Rademacher variables $\sigma_1,\ldots,\sigma_N$,

$$\left| \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) - \mathbb{E}_\sigma \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) \right| \leqslant \sqrt{\frac{2\log(2/\delta)}{N}}. \qquad (177)$$

Thus, by a union bound we have that with probability $\geqslant 1-\delta$ over the draw of the Rademacher variables:

$$\begin{aligned}
&\sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) - \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \widehat{\mathfrak{R}}\left(\mathcal{F}_{\bar{\theta}}\right) \\
&\leqslant \left| \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) - \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \mathbb{E}_\sigma \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) \right| \\
&\leqslant \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \left| \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) - \mathbb{E}_\sigma \sup_{\theta_2\in\Theta_2} \frac{1}{n}\sum_{i=1}^n \sigma_i f_i(\bar{\theta},\theta_2) \right| \\
&\leqslant \sqrt{\frac{2\log(2/\delta)}{N}} + \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1,\epsilon))}{N}}. \qquad (178)
\end{aligned}$$

Let $X$ denote the random variable

$$\widehat{\mathfrak{R}}(\mathcal{F}) - \epsilon - \sup_{\bar{\theta}\in\mathcal{C}(\epsilon)} \widehat{\mathfrak{R}}\left(\mathcal{F}_{\bar{\theta}}\right) - \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1,\epsilon))}{N}} \qquad (179)$$

(with the randomness arising from the Rademacher variables $\sigma_i$s).

By equations (178) and (176) we have for all $\varepsilon > 0$

$$\mathbb{P}(X \geqslant \varepsilon) \leqslant 2\exp\left(-\frac{\varepsilon^2 N}{2}\right). \qquad (180)$$

Integrating over $\varepsilon$ we obtain

$$\begin{aligned}
\mathbb{E}(X) &\leqslant \int_0^\infty 2\exp\left(-\frac{\varepsilon^2 N}{2}\right) d\varepsilon \qquad (181) \\
&= \frac{2\sqrt{2}}{\sqrt{N}} \int_0^\infty \exp(-\theta^2) d\theta = \sqrt{\frac{2\pi}{N}}, \qquad (182)
\end{aligned}$$

which, plugged back into the definition of $X$ (i.e. Equation (179)) gives the result.

□

Whilst Lemma E.3 above works well when the function class $\Theta_1$ enjoys a log covering number with very mild dependence on the granularity $\epsilon$ (e.g. $\log(1/\epsilon)$), it is insufficient to handle the typical $1/\epsilon^2$ dependency of norm-based bounds. The generalization below is more suitable in this case.

**Lemma E.4** (Multi-class chaining: full compositional generalization of Dudley's entropy theorem)**.** *Let $\mathcal{F} := \{f_i(\theta_1, \theta_2) : \theta_a \in \Theta_1, \theta_2 \in \Theta_2\}$ be a class of functions on $\{1, 2, \dots, n\}$ with values in $[-\mathcal{B}, \mathcal{B}]$ and dependent on two parameters $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$. Assume that there exists a $\theta_o \in \Theta_1$ such that $f_i(\theta_0, \theta_2) = 0$ for all $i$ and for all $\theta_2 \in \Theta_2$. Assume that for all $\epsilon > 0$, $\Theta_1$ admits an $\epsilon$-uniform cover $\mathcal{C}(\epsilon) \subset \Theta_1$ (of minimum size $\mathcal{N}(\Theta_1, \epsilon)$, dependent on $\epsilon$) in the following sense: For any $\theta_1 \in \Theta_1$ there exists a $\bar{\theta} \in \mathcal{C}(\epsilon)$ such that **for all** $\theta_2 \in \Theta_2$ we have*

$$\frac{1}{N}\sum_{i=1}^{N}\big(f_i(\theta_1, \theta_2) - f_i(\bar{\theta}, \theta_2)\big)^2 \leqslant \epsilon^2. \tag{183}$$

*The Rademacher complexity of the function class $\mathcal{F}$ is bounded as follows:*

$$\begin{aligned}\widehat{\mathfrak{R}}(\mathcal{F}) &:= \mathbb{E}_\sigma \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1, \theta_2) \\ &\leqslant \log_2\left(\frac{1}{\alpha}\right)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right) + 4\alpha + 4\sqrt{10}\int_{\alpha}^{\mathcal{B}}\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon))}{N}}d\epsilon + 4\,\mathcal{B}\sqrt{\frac{5\pi}{N}},\end{aligned} \tag{184}$$

*where for any fixed $\theta_1 \in \Theta_1$, $\mathcal{F}_{\theta_1}$ is the function class $\{f(\theta_1, \theta_2) : \theta_2 \in \Theta_2\}$.*

*Proof.* W.l.o.g. we assume $\mathcal{B} = 1$. Let $H$ be arbitrary, and let $\epsilon_h = 2^{-(h-1)}$ for $h = 1, 2, \dots, H$. For all $h$, let $V_h \subset \Theta_1$ denote the cover achieving (183), where we can choose $v_1 = \{\theta_0\}$. For each $\theta_1 \in \Theta_1$ let us also write $v^h[\theta_1]$ for the cover element n $V_h$ which achieves (183). Using a similar decomposition to classic proofs of the standard Dudley entropy theorem, we have for any value of the Rademacher variables $\sigma_1, \dots, \sigma_N$:

$$\begin{aligned}&\sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1, \theta_2) \\ &\leqslant \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(v^1[\theta_1], \theta_2) + \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i[f_i(\theta_1, \theta_2) - f_i(v^H[\theta_1], \theta_2)] \\ &+ \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H-1}\sigma_i[f_i(v^h[\theta_1], \theta_2) - f_i(v^{h+1}[\theta_1], \theta_2)]. \\ &\leqslant \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(v^1[\theta_1], \theta_2) + \epsilon_H \\ &\qquad + \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H-1}\sigma_i[f_i(v^h[\theta_1], \theta_2) - f_i(v^{h+1}[\theta_1], \theta_2)] \\ &\leqslant \epsilon_H + \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H-1}\sigma_i[f_i(v^h[\theta_1], \theta_2) - f_i(v^{h+1}[\theta_1], \theta_2)],\end{aligned} \tag{185}$$

where the second inequality follows from the definition of the cover and the Cauchy-Schwartz inequality, and the last inequality follows from $v_1 = \{\theta_0\}$. Next let us define the function class

$$W_h = \{w \in \mathbb{R}^{[N]\otimes\Theta_2} : \exists\theta_1 \in \Theta_1 \text{ s.t. } w_i(\theta_2) = f_i(v^h[\theta_1], \theta_2) - f_i(v^{h+1}[\theta_1], \theta_2)\}.$$

Then note that we have

$$|W_h| \leqslant |V_h||V_{h+1}| \leqslant |V_{h+1}|^2 = \mathcal{N}(\Theta_1, \epsilon_{h+1})^2. \tag{186}$$

Note that for all $h \leqslant H$ we have

$$\mathbb{E}_\sigma \sup_{\theta_{1,2}\in\Theta_{1,2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i[f_i(v^h[\theta_1], \theta_2) - f_i(v^{h+1}[\theta_1], \theta_2)] \leqslant \mathbb{E}_\sigma \sup_{\theta_2\in\Theta_2}\sup_{w\in W_{h,\theta_2}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i w_i. \tag{187}$$

Let $W = \bigcup_{h \leqslant H} W_h$. By definition of the cover (cf. Equation (183)), we have, for any $\theta_2 \in \Theta_2$:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}[f_i(v^h[\theta_1],\theta_2) - f_i(v^{h+1}[\theta_1],\theta_2)]^2 \qquad (188)\\
&= \frac{1}{N}\sum_{i=1}^{N}[f_i(v^h[\theta_1],\theta_2) - f_i(\theta_1,\theta_2) + f_i(\theta_1,\theta_2) - f_i(v^{h+1}[\theta_1],\theta_2)]^2\\
&\leqslant \frac{2}{N}\sum_{i=1}^{N}\left[f_i(v^h[\theta_1],\theta_2) - f_i(\theta_1,\theta_2)\right]^2 + \frac{2}{N}\sum_{i=1}^{N}\left[f_i(\theta_1,\theta_2) - f_i(v^{h+1}[\theta_1],\theta_2)\right]^2\\
&\leqslant 2\epsilon_h^2 + 2\epsilon_{h+1}^2 \leqslant 10\epsilon_{h+1}^2. \qquad (189)
\end{aligned}$$

Hence, we can apply Lemma E.2 to conclude that, for any choice of $w \in W$, with probability $\geqslant 1-\delta$ over the draw of the Rademacher variables $\sigma_1,\ldots,\sigma_N$, we have (simultaneously over all $\theta_2$):

$$\left|\sup_{\theta_2\in\Theta_2}\frac{1}{N}\sum_{i=1}^{N}\sigma_i w_i(\theta_2) - \mathbb{E}_\sigma\sup_{\theta_2\in\Theta_2}\frac{1}{N}\sum_{i=1}^{N}\sigma_i w_i(\theta_2)\right| < \sqrt{10}\epsilon_{h+1}\sqrt{\frac{2\log(2/\delta)}{N}}. \qquad (190)$$

Thus, by a union bound running over $w \in W$ together with equation (186) we have that with probability $\geqslant 1-\delta$ over the draw of the Rademacher variables, the following holds for all values of $w \in W_h$ simultaneously:

$$\left|\sup_{\theta_2\in\Theta_2}\sum_{i=1}^{N}\frac{\sigma_i}{N}w_i(\theta_2) - \mathbb{E}_\sigma\sup_{\theta_2\in\Theta_2}\sum_{i=1}^{N}\frac{\sigma_i}{N}w_i(\theta_2)\right| \leqslant \sqrt{10}\epsilon_{h+1}\left[\sqrt{\frac{2\log(2/\delta)}{N}} + \sqrt{\frac{4\log(\mathcal{N}(\Theta_1,\epsilon_{h+1}))}{N}}\right]. \qquad (191)$$

Furthermore, for any $w \in W$, we certainly have:

$$\mathbb{E}_\sigma\sup_{\theta_2\in\Theta_2}\sum_{i=1}^{N}\frac{\sigma_i}{N}w_i(\theta_2) \leqslant 2\sup_{\theta_1\in\Theta_1}\mathbb{E}_\sigma\sup_{\theta_2\in\Theta_2}\sum_{i=1}^{N}\frac{\sigma_i}{N}f_i(\theta_1,\theta_2) = 2\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right). \qquad (192)$$

Plugging equations (191) and (192) back into equations (187) and (185), we have with probability $\geqslant 1-\delta$ over the draw of the Rademacher variables $\sigma_1,\ldots,\sigma_N$:

$$\begin{aligned}
&\sup_{\theta_{1,2}\in\Theta_{1,2}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1,\theta_2) \qquad (193)\\
&\leqslant \epsilon_H + 2(H-1)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right) + \sup_{\theta_{1,2}\in\Theta_{1,2}}\sum_{i=1}^{N}\sum_{h=1}^{H-1}\frac{\sigma_i}{N}[f_i(v^h[\theta_1],\theta_2) - f_i(v^{h+1}[\theta_1],\theta_2)]\\
&\leqslant \epsilon_H + 2(H-1)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right) + \sum_{h=1}^{H-1}\sqrt{10}\left[\epsilon_{h+1}\sqrt{\frac{2\log(2/\delta)}{N}} + \epsilon_{h+1}\sqrt{\frac{4\log(\mathcal{N}(\Theta_1,\epsilon_{h+1}))}{N}}\right]\\
&\leqslant \epsilon_H + 2(H-1)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right) + 4\sqrt{\frac{5\log(2/\delta)}{N}} + 4\sqrt{10}\sum_{h=1}^{H}[\epsilon_h - \epsilon_{h+1}]\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon_h))}{N}}.
\end{aligned}$$

Finally, take $H$ to be the largest integer such that $\epsilon_{H+1} > \alpha$, then $\epsilon_H = 4\epsilon_{H+2} \leqslant 4\alpha$ and we can continue to show that (w.p. $\geqslant 1-\delta$ over the draw of $\sigma$):

$$\begin{aligned}
&\sup_{\theta_{1,2}\in\Theta_{1,2}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1,\theta_2)\\
&\leqslant \epsilon_H + 2(H-1)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right) + 4\sqrt{\frac{5\log(2/\delta)}{N}} + 4\sqrt{10}\int_{\epsilon_{H+1}}^{1}\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon))}{N}}d\epsilon\\
&\leqslant 2\log_2\left(\frac{1}{\alpha}\right)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right) + 4\alpha + 4\sqrt{10}\int_{\alpha}^{1}\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon))}{N}}d\epsilon + 4\sqrt{\frac{5\log(2/\delta)}{N}}. \qquad (194)
\end{aligned}$$

Next, let $X$ denote the random variable

$$\frac{1}{4\sqrt{5}}\left[\sup_{\theta_{1,2}\in\Theta_{1,2}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1,\theta_2)-4\alpha-4\sqrt{10}\int_{\alpha}^{1}\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon))}{N}}d\epsilon-2\log_2\left(\frac{1}{\alpha}\right)\right], \tag{195}$$

with the randomness arising from the Rademacher variables $\sigma_i$s.

By Equation (194) we have

$$\mathbb{P}(X\geqslant\varepsilon)\leqslant 2\exp\left(-\varepsilon^2 N\right). \tag{196}$$

Integrating over $\varepsilon$ we obtain

$$\mathbb{E}(X)\leqslant\int_0^{\infty}2\exp\left(-\varepsilon^2 N\right)d\varepsilon \tag{197}$$

$$=\frac{2}{\sqrt{N}}\int_0^{\infty}\exp(-\theta^2)d\theta=\sqrt{\frac{\pi}{N}}. \tag{198}$$

Plugging this back into the definition of $X$ (eq (179)) after taking expectations with respect to $\sigma_1,\ldots,\sigma_N$, we get:

$$\widehat{\mathfrak{R}}(\mathcal{F})=\mathbb{E}_{\sigma}\sup_{\theta_{1,2}\in\Theta_{1,2}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(\theta_1,\theta_2) \tag{199}$$

$$\leqslant\log_2\left(\frac{1}{\alpha}\right)\sup_{\theta_1\in\Theta_1}\widehat{\mathfrak{R}}\left(\mathcal{F}_{\theta_1}\right)+4\alpha+4\sqrt{10}\int_{\alpha}^{1}\sqrt{\frac{\log(\mathcal{N}(\Theta_1,\epsilon))}{N}}d\epsilon+4\sqrt{\frac{5\pi}{N}}, \tag{200}$$

as expected.

□

### E.2. On the Impact of the Estimation of the Marginals on the Schatten quasi-norms of $\widetilde{Z}$

In this subsection, we present the following result, which is useful when proving our excess risk bounds.

**Lemma E.5** (Generalization of Lemma 4 in (Foygel et al., 2011))**.** *Let $d\geqslant 2$ be an integer. For any $N\geqslant 6(m+n)\log(\frac{m+n}{\delta})$) we have the following inequality with probability greater than $1-\delta$ over the draw of the training set:*

$$\|\check{Z}\|_{\mathrm{sc},2/d}^{2/d}\leqslant\left(1+\sqrt{\frac{6(m+n)\log\left(\frac{m+n}{\delta}\right)}{N}}\right)\|\widetilde{Z}\|_{\mathrm{sc},2/d}^{\frac{2}{d}} \tag{201}$$

*Proof.* Let $A,B,D_1,\ldots,D_{d-1}$ be such that

$$d\|\widetilde{Z}\|_{\mathrm{sc},2/d}^{\frac{2}{d}}=\sum_{i=1}^{d-2}\|D_i\|_{\mathrm{Fr}}^2+\|\operatorname{diag}(\widetilde{p})^{\frac{1}{2}}A\|_{\mathrm{Fr}}^2+\|\operatorname{diag}(\widetilde{q})^{\frac{1}{2}}B\|_{\mathrm{Fr}}^2 \tag{202}$$

$$A\prod_{i=1}^{d-2}D_iB^{\top}=Z. \tag{203}$$

By Corollary (F.23) we have:

$$d\|\check{Z}\|_{\mathrm{sc},2/d}^{2/d}\leqslant\sum_{i=1}^{d-2}\|D_i\|_{\mathrm{Fr}}^2+\|\operatorname{diag}(\check{p}/\tilde{p})\widetilde{A}\|_{\mathrm{Fr}}^2+\|\operatorname{diag}(\check{q}/\tilde{q})\widetilde{B}\|_{\mathrm{Fr}}^2 \tag{204}$$

where $\widetilde{A}:=\operatorname{diag}(\tilde{p})^{\frac{1}{2}}A$ and $\widetilde{B}:=\operatorname{diag}(\tilde{p})^{\frac{1}{2}}B$.

By a Bernstein bound, for any $\epsilon \leqslant 1$,

$$\mathbb{P}\left(\frac{\frac{\hat{p}_i - p_i}{2}}{\frac{1}{2m} + \frac{p_i}{2}} \geqslant \epsilon\right) \leqslant \exp\left(-\frac{\epsilon^2 N}{6m}\right) \tag{205}$$

and similarly

$$\mathbb{P}\left(\frac{\frac{\hat{q}_j - q_j}{2}}{\frac{1}{2n} + \frac{q_j}{2}} \geqslant \epsilon\right) \leqslant \exp\left(-\frac{\epsilon^2 N}{6n}\right). \tag{206}$$

Thus we have for any $\epsilon \leqslant 1$, using a union bound:

$$\mathbb{P}\left(\exists i : \frac{\frac{\hat{p}_i - p_i}{2}}{\frac{1}{2m} + \frac{p_i}{2}} \geqslant \epsilon \vee \exists j : \frac{\frac{\hat{q}_j - q_j}{2}}{\frac{1}{2n} + \frac{q_j}{2}} \geqslant \epsilon\right) \leqslant (m+n)\exp\left(-\frac{\epsilon^2 N}{6(n+m)}\right). \tag{207}$$

Thus, we know that with probability greater than $1-\delta$ we have simultaneously over all $i, j$:

$$\begin{aligned} \frac{\frac{1}{2m} + \frac{\hat{p}_i}{2}}{\frac{1}{2m} + \frac{p_i}{2}} &\leqslant 1 + \sqrt{\frac{6(m+n)\log\left(\frac{m+n}{\delta}\right)}{N}} \qquad \text{and} \\ \frac{\frac{1}{2n} + \frac{\hat{q}_i}{2}}{\frac{1}{2n} + \frac{q_i}{2}} &\leqslant 1 + \sqrt{\frac{6(m+n)\log\left(\frac{m+n}{\delta}\right)}{N}}. \end{aligned} \tag{208}$$

Plugging this back into Equation (204) we obtain (w.p. $1-\delta$):

$$d\|\breve{Z}\|_{\mathrm{sc},2/d}^{2/d} \leqslant \sum_{i=1}^{d-2} \|D_i\|_{\mathrm{Fr}}^2 + \|\operatorname{diag}(\breve{p}/\tilde{p})\widetilde{A}\|_{\mathrm{Fr}}^2 + \|\operatorname{diag}(\breve{q}/\tilde{q})\widetilde{B}\|_{\mathrm{Fr}}^2 \tag{209}$$

$$\leqslant \left(1 + \sqrt{\frac{6(m+n)\log\left(\frac{m+n}{\delta}\right)}{N}}\right)\left[\|\widetilde{A}\|_{\mathrm{Fr}}^2 + \|\widetilde{B}\|_{\mathrm{Fr}}^2\right] + \sum_{i=1}^{d-2}\|D_i\|_{\mathrm{Fr}}^2 \tag{210}$$

$$\leqslant \left(1 + \sqrt{\frac{6(m+n)\log\left(\frac{m+n}{\delta}\right)}{N}}\right)\left[\|\widetilde{A}\|_{\mathrm{Fr}}^2 + \|\widetilde{B}\|_{\mathrm{Fr}}^2 + \sum_{i=1}^{d-2}\|D_i\|_{\mathrm{Fr}}^2\right] \tag{211}$$

$$= \left(1 + \sqrt{\frac{6(m+n)\log\left(\frac{m+n}{\delta}\right)}{N}}\right) d\|\breve{Z}\|_{\mathrm{sc},2/d}^{\frac{2}{d}}, \tag{212}$$

as expected.

□

### E.3. Covering Number Bounds for Neural Embeddings

In this section, we provide covering number bounds for neural embeddings, including both with weighted and un-weighted versions of the constraints on the zeroth layer learnable embeddings. This section relies on results from subsection F.3 such as Proposition F.8.

We consider the following function classes.

$$\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}(a,s,c) := \Bigg\{ g:[m]\times[n]\to\mathbb{R}^1 \big| \exists f\in\mathcal{N}_{1,\mathrm{W}}(a,s), U\in\mathbb{R}^{m\times\bar{\mathrm{m}}}, V\in\mathbb{R}^{n\times\bar{\mathrm{m}}}: \tag{213}$$

$$\|\operatorname{diag}(\widetilde{p})^{\frac{1}{2}}U\|_{\mathrm{Fr}}^2 + \|\operatorname{diag}(\widetilde{q})^{\frac{1}{2}}V\|_{\mathrm{Fr}}^2 \leqslant c, \|A^0\|\leqslant s_0 :\ g(i,j)=f(A^0(u_i,v_j)^\top)\ \ \forall i,j \Bigg\},$$

$$\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}(a,s,c) := \Bigg\{ g:[m]\times[n]\to\mathbb{R}^1 \big| \exists f\in\mathcal{N}_{1,\mathrm{W}}(a,s), U\in\mathbb{R}^{m\times\bar{\mathrm{m}}}, V\in\mathbb{R}^{n\times\bar{\mathrm{m}}}: \tag{214}$$

$$\|\operatorname{diag}(\widetilde{p})^{\frac{1}{2}}U\|_{\mathrm{Fr}}^2 + \|\operatorname{diag}(\widetilde{q})^{\frac{1}{2}}V\|_{\mathrm{Fr}}^2 \leqslant c, \|A^0\|\leqslant s_0 :\ g(i,j)=f(A^0(u_i,v_j)^\top)\ \ \forall i,j \Bigg\},$$

$$\mathcal{N}_{0,\mathrm{W},\mathrm{c}}(a,s,c) := \Bigg\{ g:[m]\times[n]\to\mathbb{R}^1 \big| \exists f\in\mathcal{N}_{1,\mathrm{W}}(a,s), U\in\mathbb{R}^{m\times\bar{\mathrm{m}}}, V\in\mathbb{R}^{n\times\bar{\mathrm{m}}}: \tag{215}$$

$$\|U\|_{\mathrm{Fr}}^2 + \|V\|_{\mathrm{Fr}}^2 \leqslant c^2\max(m,n), \|A^0\|\leqslant s_0 :\ g(i,j)=f(A^0(u_i,v_j)^\top)\ \ \forall i,j \Bigg\},$$

where $u_i$ and $v_j$ denote the $i$th and $j$th rows of $U$ and $V$ respectively and $A^0\in\mathbb{R}^{d\times(2\,\bar{\mathrm{m}})}$.

**Proposition E.6** ($L^2$ covers of $\mathcal{N}_{0,\mathrm{W},\mathrm{c}}, \widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$)**.** *Assume as usual that $s_l\geqslant 1$ for all $l$ and $\chi^2\geqslant 1$. For any sample $\xi^1,\dots,\xi^N\in[m]\times[n]$ such that $\forall i,j$, $\hat{p}_i\leqslant 2p_i$ and $\hat{q}_j\leqslant 2q_j$, any $\epsilon>0$, there exists a cover $\mathcal{C}(\epsilon)\subset\mathcal{N}_{0,\mathrm{W},\mathrm{c}}(a,s,c)$ (resp. $\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}(a,s,c)$) with the following properties.*

1. *For $g\in\mathcal{N}_{0,\mathrm{W},\mathrm{c}}(a,s,c)$ (resp. $\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$) there exists a $\bar{g}$ in $\mathcal{C}(\epsilon)$ such that*

$$\frac{1}{N}\sum_{o=1}^{N}(g-\bar{g})_{\xi^o}\leqslant\epsilon^2. \tag{216}$$

2.

$$\log(|\mathcal{C}(\epsilon)|)\leqslant\left[2d(m+n)+32s_0^2c^2\left[\frac{1}{\epsilon^2}+1\right]\mathrm{R_W}^2\right]\log\left(\Gamma_{W,\epsilon}\right)\quad \textit{for}\quad \widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}} \tag{217}$$

$$\left[2d(m+n)+32s_0^2c^2(m+n)\left[\frac{1}{\epsilon^2}+1\right]\mathrm{R_W}^2\right]\log\left(\Gamma_{W,\epsilon}\right)\quad \textit{for}\quad \mathcal{N}_{0,\mathrm{W},\mathrm{c}} \tag{218}$$

*where* $\Gamma_{W,\epsilon}=\frac{96Ws_0(m+n)\sqrt{mn}\prod_{\ell=1}^{L}s_\ell}{\epsilon}+1$.

*Proof.* We write a single proof for both cases as they are very similar.

**Step 1: Uniform cover $\mathcal{C}_1$ of the embeddings** $A^0(u_i,v_j)^\top$

For any $A^0,U,V$, we write $\bar{U}(A^0,U,V)$ for the tensor in $\mathbb{R}^{d\times m\times n}$ such that $\bar{U}(A^0,U,V)_{u,i,j}=\left[A^0(u_i,v_j)^\top\right]_u$ for any $i\leqslant m, j\leqslant n, u\leqslant d$. Let $\mathcal{U}:=\left\{\bar{U}\in\mathbb{R}^{d\times mn}:\exists(A^0,U,V)\in\mathcal{A}_{a,s}\right\}$ where $\mathcal{A}_{a,s}$ is the set of admissible $U,V,A^0$ defined according to the corresponding equations (213), (214) depending on whether we are proving the Theorem for the class $\mathcal{N}_{0,\mathrm{W},\mathrm{c}}, \widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$.

Note that for any $A^0$, we can always write $A^0(u_i,v_j)^\top=A_1^0u_i^\top+A_2^0v_j^\top$ where $A_1^0$ and $A_2^0$ represent the first and last $\bar{\mathrm{m}}$ columns of $A^0$ respectively. Thus, we have $\mathcal{U}\subset\mathcal{U}_1+\mathcal{U}_2$ where $\mathcal{U}_1:=\left\{A_1U^\top\middle|\exists V,A_2:((A_1,A_2),U,V)\in\mathcal{A}_{a,s}\right\}$ and $\mathcal{U}_2:=\left\{A_2V^\top\middle|\exists U,A_1:((A_1,A_2),U,V)\in\mathcal{A}_{a,s}\right\}$.

For each admissible $U,V,A^0$, it is certainly the case that $\|A^0u_i^\top\|,\|A^0v_j^\top\|\leqslant\|A^0\|\left(\|U\|_{\mathrm{Fr}}^2+\|V\|_{\mathrm{Fr}}^2\right)^{\frac{1}{2}}\leqslant 2s_0c(m+n)$ for any $i,j$, where we have used the fact that $\tilde{p}_i\geqslant\frac{1}{2m}$ and $\tilde{q}_j\geqslant\frac{1}{2n}$.

Thus, we have $\mathcal{U}_1 \subset \widetilde{\mathcal{U}}_1 := \left\{\bar{U} \in \mathbb{R}^{d\times m} : \|\bar{U}\|_{\mathrm{Fr}} \leqslant 2s_0(m+n)\sqrt{m}\right\}$ and $\mathcal{U}_2 \subset \widetilde{\mathcal{U}}_2 := \left\{\bar{U} \in \mathbb{R}^{d\times n} : \|\bar{U}\|_{\mathrm{Fr}} \leqslant 2s_0(m+n)\sqrt{n}\right\}$.

Thus, by Lemma F.17 applied to $\widetilde{\mathcal{U}}_1$ and $\widetilde{\mathcal{U}}_2$, we can obtain internal covers $\widetilde{\mathcal{C}}_{11}$ and $\widetilde{\mathcal{C}}_{12}$ of $\widetilde{\mathcal{U}}_1$ and $\widetilde{\mathcal{U}}_2$ with respect to the $\|\cdot\|_{\mathrm{Fr}}$ norm with granularity $\epsilon/(8\prod_{\ell=1}^{L} s_\ell)$ such that

$$\begin{aligned}&\log\left(\min\left[\left|\widetilde{\mathcal{C}}_{11}\right|, \left|\widetilde{\mathcal{C}}_{12}\right|\right]\right) \\ &\leqslant (m+n)d\left(\frac{6s_0(m+n)\sqrt{mn}}{\frac{\epsilon}{(8\prod_{\ell=1}^{L} s_\ell)}}+1\right) = (m+n)d\left(\frac{48s_0(m+n)\sqrt{mn}\prod_{\ell=1}^{L} s_\ell}{\epsilon}+1\right),\end{aligned} \tag{219}$$

which immediately yields an external cover of $\mathcal{U}$ with granularity $\epsilon/(4\prod_{\ell=1}^{L} s_\ell)$, and finally an internal cover $\mathcal{C}_1$ of the same, with granularity $\epsilon/(2\prod_{\ell=1}^{L} s_\ell)$ and cardinality

$$\log\left(|\mathcal{C}_1|\right) \leqslant 2d(m+n)\log\left(\frac{48s_0(m+n)\sqrt{mn}\prod_{\ell=1}^{L} s_\ell}{\epsilon}+1\right). \tag{220}$$

**Step 2: $L^2$ covers of the network class $\mathcal{N}_{1,\mathrm{W}}$ relative to the cover elements in $\mathcal{C}_1$**

First, note that for any admissible $U, V, A^0$, we certainly have

$$\begin{aligned}&\frac{1}{N}\sum_{o=1}^{N}\|A^0(u_{\xi_1^o}, v_{\xi_2^o})^\top\|^2 \leqslant s_0^2\frac{1}{N}\sum_{o=1}^{N}\left[\|u_{\xi_1^o}\|^2+\|v_{\xi_2^o}\|^2\right] = s_0^2\sum_i \hat{p}_i\|u_i\|^2 + s_0^2\sum_j \hat{q}_j\|v_j\|^2 \\ &\leqslant 2s_0^2\min\left(\sum_i \tilde{p}_i\|u_i\|^2+\sum_j \tilde{q}_j\|v_j\|^2, \max_i\|u_i\|^2+\max_j\|v_j\|^2\right) \\ \leqslant& 4s_0^2c^2, \quad \text{for} \quad \widetilde{\mathcal{N}_{0,\mathrm{W,c}}}, \quad \text{and} \qquad (221) \\ &4s_0^2c^2(\max(m,n)) \quad \text{for} \quad \mathcal{N}_{0,\mathrm{W,c}}. \qquad (222)\end{aligned}$$

Thus, for each cover element $\bar{U} = (A^0, U, V) \in \mathcal{C}_1$, there we can apply Proposition F.8 to show that there is a cover $\mathcal{C}_2(\bar{U}) \subset \mathcal{N}_{1,\mathrm{W}}$ such that the following properties are satisfied:

1. For any $f \in \mathcal{N}_{1,\mathrm{W}}$ there exists $\bar{f} \in \mathcal{C}_2(\bar{U})$ such that

$$\frac{1}{N}\sum_{o=1}^{N}\left\|[f-\bar{f}](A^0(u_i, v_j)^\top)\right\|^2 \leqslant \epsilon^2/4, \tag{223}$$

and

2.

$$\log\left(\mathcal{C}_2\right) \leqslant 32s_0^2c^2\left[\frac{1}{\epsilon^2}+1\right]{\mathrm{R_W}}^2\log(2W), \tag{224}$$

where as usual $\mathrm{R_W}$ is defined by Equation (264).

**Step 3: $L^2$ cover of $\mathcal{N}_{0,\mathrm{W,c}}$ and $\widetilde{\mathcal{N}_{0,\mathrm{W,c}}}$**

We now finally define the cover $\mathcal{C} \subset \mathcal{N}_{0,\mathrm{W,c}}$ via

$$\mathcal{C} = \left\{f \circ \bar{U} \middle| \bar{U} \in \mathcal{C}_1, f \in \mathcal{C}_2(\bar{U})\right\}, \tag{225}$$

where by abuse of notation, $f \circ \bar{U}$ denotes the function $g : [m]\times[n] \to \mathbb{R}^1$ such that $g(i,j) = f(\bar{U}_{\bullet,i,j}) = f(A^0(u_i, v_j)^\top)$ (where $A^0, U, V$ realise the element $\bar{U}$ of $\mathcal{U}$ and as usual $u_i$ and $v_j$ denote the $i$th and $j$th rows of $U$ and $V$ respectively).

We now have that $\mathcal{C}$ is an $\epsilon$ cover with respect to the $L^2$ norm and the sample $\xi^1, \dots, \xi^N$. Indeed, for $g = f \circ \bar{U} \in \mathcal{N}_{0,\mathrm{W},\mathrm{c}}$, let $\bar{g} = \bar{f} \circ \bar{\bar{U}}$ be the associated cover element in $\mathcal{C}$, we have

$$\frac{1}{N}\sum_{o=1}^{N}(g-\bar{g})_{\xi^o}^2 \leqslant 2\frac{1}{N}\sum_{o=1}^{N}\left([f-\bar{f}]\circ\bar{U}\right)_{\xi^o}^2 + 2\frac{1}{N}\sum_{o=1}^{N}\left(\bar{f}\circ\left[\bar{U}-\bar{\bar{U}}\right]\right)_{\xi^o}^2 \tag{226}$$

$$\leqslant 2\frac{\epsilon^2}{4} + 2\left[\prod_{\ell=1}^{L}s_\ell\right]^2\frac{\epsilon^2}{4\left[\prod_{\ell=1}^{L}s_\ell\right]^2} = \epsilon^2 \tag{227}$$

where at the first line (226) we have used the AM-GM inequality and at the last line (227) we have used the properties of both covers. This established the validity of the cover (Equation (216)). We now only have to estimate the cardinality to establish Equation (217):

$$\log\left(|\mathcal{C}(\epsilon)|\right) \leqslant \sum_{\bar{U}\in\mathcal{C}_1}\left|\mathcal{C}_2(\bar{U})\right| \tag{228}$$

$$\leqslant 2d(m+n)\log\left(\frac{48s_0(m+n)\sqrt{mn}\prod_{\ell=1}^{L}s_\ell}{\epsilon}+1\right) + 32s_0^2\underline{c}^2\left[\frac{1}{\epsilon^2}+1\right]\mathrm{R_W}^2\log(2W) \tag{229}$$

$$\leqslant \left[2d(m+n) + 32s_0^2\underline{c}^2\left[\frac{1}{\epsilon^2}+1\right]\mathrm{R_W}^2\right]\log\left(\frac{96Ws_0(m+n)\sqrt{mn}\prod_{\ell=1}^{L}s_\ell}{\epsilon}+1\right), \tag{230}$$

where $\underline{c}$ stands for $c$ if we are covering $\widetilde{\mathcal{N}_{0,\mathrm{W},\mathrm{c}}}$ and $c\sqrt{\max(m,n)}$ if we are covering $\mathcal{N}_{0,\mathrm{W},\mathrm{c}}$, and (at Equation (229)) we have used Equations (224) and (220). The result follows. □

## F. Variations on Known Results

This section compiles some minor variations of known results. We include some proofs both for completeness and because the precise results we need sometimes deviate slightly from the known version. For instance, we often require high probability versions of results which were previously presented in expectation over the training set.

**Remark:** The modification is necessary for incorporation with the neural network bounds of Section E.3 and even for incorporation with the Lipschitz constant bounds of Section C.2 via Lemma E.4 and Lemma E.3. Indeed, although the fact that the construction of the L2 cover of the class $\mathcal{N}_{1,\mathrm{W}}$ is non constructive and dependent on the sample $\xi^1, \dots, \xi^N$ may not be a strong obstacle to applying an expectation version of Lemma E.4 or Lemma E.3, the supremum over $\theta_1$ is a bigger issue. Taking the example of Lemma F.1, we need to know that with high probability over the draw of the training set, Equation (231) will be satisfied, allowing us to show that for this particular training set, the inequality in Theorem D.2 holds *uniformly* over all bounded Lipschitz functions $\ell$.

### F.1. On the Complexity of Classes of Matrices with Nuclear Norm Constraints

We will need the following classic Lemma:

**Lemma F.1** (Cf. (Foygel et al., 2011), proof of Theorem 1, (Tropp, 2012), Remarks 6.4 and 6.5, cf. also (Ledent et al., 2021b))**.** *Consider the matrix $R_N = \frac{1}{N}\sum_{i=1}^N \sigma_i e_{\zeta_i}$ where the $\sigma_i$s are Rademacher variables and for all $i, j$, $e_{(i,j)}$ is the matrix with a $1$ in the entry $(i, j)$ and zeros in all other entries and the entries are selected i.i.d from a distribution with uniform marginals. With probability greater than $1-\delta$ over the draw of the training set and the draw of the Rademacher variables $\sigma_1, \dots, \sigma_N$, we have*

$$\|R_N\| \leqslant \sqrt{8/3}\sqrt{\frac{1}{N\min(m,n)}}\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{8}{3N}\log\left(\frac{m+n}{\delta}\right). \tag{231}$$

*Proof.* $R_N$ is an average of $N$ i.i.d. matrices $X_i = \sigma_i e_{\zeta_i}$. Thus, we can apply the non commutative Bernstein inequality (Proposition F.20). We have $\|X_i\| \leqslant 1/N$ with probability 1 for all $i$, thus "M" is $1/N$. Furthermore, we can compute the $\rho^2$ as follows:

$$\left\|\mathbb{E}\left(X_i X_i^\top\right)\right\| = \frac{1}{N^2}\|\operatorname{diag}(p_1, p_2, \dots, p_m)\| = \frac{1}{mN^2} \tag{232}$$

and

$$\left\|\mathbb{E}\left(X_i^\top X_i\right)\right\| = \frac{1}{N^2}\|\operatorname{diag}(q_1, q_2, \dots, q_n)\| = \frac{1}{nN^2}, \tag{233}$$

where the $p_i$s (resp. $q_j$s) denote the marginal probabilities for each row (column), which are uniform by our assumption. Hence $\rho_k^2$ is $\frac{1}{N^2\min(m,n)}$ and $\sigma^2 = \frac{1}{N\min(m.n)}$. From this, it follows by applying Proposition F.21 that with probability $\geqslant 1-\delta$ over the draw of the training set, we have

$$\|R_N\| \leqslant \sqrt{8/3}\sigma\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{8M}{3}\log\left(\frac{m+n}{\delta}\right) \tag{234}$$

$$\leqslant \sqrt{8/3}\sqrt{\frac{1}{N\min(m,n)}}\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{8}{3N}\log\left(\frac{m+n}{\delta}\right), \tag{235}$$

as expected.

□

**Lemma F.2** (Cf. (Foygel et al., 2011), proof of Theorem 1 cf. also (Ledent et al., 2021b))**.** *Consider the matrix $R_N = \frac{1}{N}\sum_{o=1}^N \sigma_o \frac{e_{\xi^o}}{\sqrt{\widehat{p}_{\xi_1^o}\widehat{q}_{\xi_2^o}}}$ where the $\sigma_o$s are Rademacher variables, for all $i, j$, $e_{(i,j)}$ is the matrix with a $1$ in the entry $(i, j)$ and zeros in all other entries, and as usual. With probability greater than $1-\delta$ over the draw of the training set and the draw of the Rademacher variables $\sigma_1, \dots, \sigma_N$, we have*

$$\|R_N\| \leqslant 4\sqrt{2/3}\sqrt{\frac{m+n}{N}}\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{16\sqrt{mn}}{3N}\log\left(\frac{m+n}{\delta}\right). \tag{236}$$

*Proof.* $R_N$ is a sum of $N$ i.i.d. matrices $X_o = \frac{e_{\xi^o}}{\sqrt{p_{\xi_1^o,\xi_2^o}}}$. Thus, we can apply the non commutative Bernstein inequality (Proposition F.20). Since $\widetilde{p}_i \geqslant \frac{1}{2m}$ and $\widetilde{q}_i \geqslant \frac{1}{2n}$ for all $i, j$, we have $\|X_o\| \leqslant \frac{2\sqrt{mn}}{N}$ with probability 1 for all $o$, thus "M" is $\frac{2\sqrt{mn}}{N}$. Furthermore, we can compute the $\rho^2$ as follows. for any $i \leqslant m$, we have

$$N^2\mathbb{E}\left(X_o X_o^\top\right)_{i,i} = \frac{1}{\widetilde{p}_i}\sum_{j=1}^{n} p_{i,j}\frac{1}{\widetilde{q}_j} \leqslant \frac{2n}{\widetilde{p}_i}\sum_{j=1}^{n} p_{i,j} = 2n\frac{p_i}{\widetilde{p}_i} \leqslant 4n, \tag{237}$$

where as usual the $p_i$s (resp. $q_j$s) denote the marginal probabilities for each row (column), which are uniform by our assumption. Similarly,

$$N^2\mathbb{E}\left(X_o^\top X_o\right)_{j,j} = \frac{1}{\widetilde{q}_j}\sum_{i=1}^{m} p_{i,j}\frac{1}{\widetilde{p}_i} \leqslant \frac{2m}{\widetilde{q}_j}\sum_{i=1}^{m} p_{i,j} = 2\frac{q_j}{\widetilde{q}_j} \leqslant 4m \tag{238}$$

Hence $\rho_k^2$ can be taken as $\frac{4(m+n)}{N^2}$ and $\sigma^2$ can be taken as $= \frac{4(m+n)}{N}$. From this, it follows by applying Proposition F.21 that with probability $\geqslant 1-\delta$ over the draw of the training set, we have

$$\|R_N\| \leqslant \sqrt{8/3}\sigma\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{8M}{3}\log\left(\frac{m+n}{\delta}\right) \tag{239}$$

$$\leqslant 4\sqrt{2/3}\sqrt{\frac{(m+n)}{N}}\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{16\sqrt{mn}}{3N}\log\left(\frac{m+n}{\delta}\right), \tag{240}$$

as expected.

□

For the distribution-free case with the nuclear norm, recall the following theorem:

**Proposition F.3** (Theorem 2 page 3405 in (Shamir & Shalev-Shwartz, 2011))**.** *Consider the following function class:*

$$\mathcal{F}_{\mathrm{t}}^1 := \left\{\mathbb{R}^{m\times n} \ni Z : \|Z\|_* \leqslant t\right\}. \tag{241}$$

*Let* $\mathrm{l} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ *be a function which is uniformly bounded by* $\mathcal{B}$ *and* $\ell$*-Lipschitz w.r.t. the second argument, for any training set* $\xi^1, \ldots, \xi^N \in [m]\times[n]$ *we have the following bound on the empirical Rademacher complexity of* $\mathrm{l}\circ\mathcal{F}_t^1$*:*

$$\mathbb{E}_\sigma \sup_{Z\in\mathcal{F}_{\mathrm{t}}^1} \frac{1}{N}\sum_{o=1}^{N} \sigma_o\, \mathrm{l}(Z_{\xi^o}, \widetilde{G}_o) \leqslant \sqrt{\frac{9C\,\mathcal{B}\,\ell\, t(\sqrt{m}+\sqrt{n})}{N}}, \tag{242}$$

*where* $C$ *is the (absolute) constant in (Latała, 2005).*

**Proposition F.4** (High probability version of Theorem 1 in (Foygel et al., 2011))**.** *Consider the following function class:*

$$\mathcal{F}_{\mathrm{t}}^1 := \left\{\mathbb{R}^{m\times n} \ni Z : \|Z\|_* \leqslant t\right\}. \tag{243}$$

*Assume that the sampling distribution has uniform marginals. For any* $\delta > 0$ *we have the following bound on the empirical Rademacher complexity of* $\mathcal{F}_t^1$ *with probability* $\geqslant 1-\delta$ *over the draw of the dataset:*

$$\mathbb{E}_\sigma \sup_{Z\in\mathcal{F}_{\mathrm{t}}^1} \frac{1}{N}\sum_{o=1}^{N} \sigma_o Z_{\xi^o} \leqslant 8\sqrt{\frac{t^2}{3N\min(m,n)}\log\left(\frac{m+n}{\delta}\right)} + \frac{16t}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{244}$$

*Proof.* This follows immediately (as a particular case) from Proposition F.5 below. □

**Proposition F.5** (High probability version of Theorem 3 in (Foygel et al., 2011))**.** *Consider the following function class:*

$$\widetilde{\mathcal{F}}_r^1 := \left\{ \mathbb{R}^{m\times n} \ni Z : \|\widetilde{Z}\|_* \leqslant \sqrt{r} \right\}, \tag{245}$$

*where as usual* $\widetilde{Z} = \operatorname{diag}(\tilde{p}) Z \operatorname{diag}(q)$*). For any* $\delta > 0$ *we have the following bound on the empirical Rademacher complexity of* $\mathcal{F}_t^1$ *with probability* $\geqslant 1-\delta$ *over the draw of the dataset:*

$$\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^1} \frac{1}{N}\sum_{o=1}^N \sigma_o Z_{\xi^o} \leqslant 4\sqrt{\frac{2r(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} + \frac{16\sqrt{mnr}}{3N}\log\left(\frac{m+n}{\delta}\right) \tag{246}$$

*Proof.* Similarly to the original proof, expanding the definition of the Rademacher complexity, we have

$$\begin{aligned}
\mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^1} \frac{1}{N}\sum_{o=1}^N \sigma_o \mathrm{Z}_{\xi^o} &= \mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^1} \frac{1}{N}\left\langle R_N, \operatorname{diag}(\sqrt{\tilde{p}}) Z \operatorname{diag}(\sqrt{\tilde{q}})\right\rangle \\
&\leqslant \mathbb{E}_\sigma \sup_{Z\in\widetilde{\mathcal{F}}_r^1} \|R_N\| \|\operatorname{diag}(\sqrt{\tilde{p}}) Z \operatorname{diag}(\sqrt{\tilde{q}})\|_* \qquad (247)\\
&\leqslant \mathbb{E}_\sigma \|R_N\| \sup_{Z\in\widetilde{\mathcal{F}}_r^1} \|\widetilde{Z}\|_* = \sqrt{r}\mathbb{E}_\sigma\|R_N\| \\
&\leqslant 4\sqrt{\frac{2r(m+n)}{3N}\log\left(\frac{m+n}{\delta}\right)} + \frac{16\sqrt{mnr}}{3N}\log\left(\frac{m+n}{\delta}\right), \qquad (248)
\end{aligned}$$

where at the second line (247) we have used the duality between the nuclear norm and the spectral norm, and at the last line (248), we have used Lemma F.2. □

### F.2. Computational Lemmas

This subsection compiles basic calculations which are useful when translating a high-probability bound into a bound in expectation.

**Lemma F.6.** *Let* $F$ *be a random variable that depends only on the draw of the training set. Assume that with probability* $\geqslant 1-\delta$*,*

$$\mathbb{E}(F) \leqslant f(\delta), \tag{249}$$

*for some given monotone increasing function* $f$*. Then we have, in expectation over the training set:*

$$\mathbb{E}(F) \leqslant \sum_{i=1}^\infty f(2^{-i}) 2^{1-i}, \tag{250}$$

*In particular, if* $f(\delta) = C_1\sqrt{\log(\frac{1}{\delta})} + C_2$*, then we have in expectation over the draw of the training set:*

$$\mathbb{E}(F) \leqslant \frac{C_1}{\sqrt{2}-1} + C_2 \leqslant \frac{5}{2}C_1 + C_2. \tag{251}$$

*Further, if* $f(\delta) = C_3\log(\frac{1}{\delta})$*, then we have*

$$\mathbb{E}(F) \leqslant 6C_3. \tag{252}$$

*Proof.* By assumption we have for any $\delta$:

$$\mathbb{P}\left(X \geqslant f(\delta)\right) \leqslant \delta \tag{253}$$

Let us write $A_i$ for the event $A_i = \{F \leqslant f(\delta_i)\}$ where we set $\delta_i = 2^{-i}$ for $i = 1, 2, \ldots$ . We also set $\tilde{A}_i = A_i \backslash A_{i-1}$ for $i = 1, 2, \ldots$ with the convention that $A_0 = \varnothing$ so that $\tilde{A}_1 = A_1$.

We have, for $i \geqslant 2$, $\mathbb{P}(\widetilde{A}_i) \leqslant \mathbb{P}(A_{i-1}^c) \leqslant \delta_{i-1}$, and for $i = 1$, $\mathbb{P}(\widetilde{A}_1) \leqslant 1 = \delta_{i-1}$. Thus we can write

$$\mathbb{E}(F) \leqslant \sum_{i=1}^{\infty} \mathbb{E}(X|\widetilde{A}_i)\mathbb{P}(\widetilde{A}_i) \leqslant \sum_{i=1}^{\infty} \mathbb{E}(X|\widetilde{A}_i)\delta_{i-1} \leqslant \sum_{i=1}^{\infty} f(\delta_i)\delta_{i-1}, \tag{254}$$

yielding identity (250) as expected.

Next, assuming $f(\delta) = C_1\sqrt{\log(\frac{1}{\delta})} + C_2$, we can continue as follows:

$$\mathbb{E}(F - C_2) \leqslant \sum_{i=1}^{\infty} f(\delta_i)\delta_{i-1} \leqslant \sum_{i=1}^{\infty} [C_1\sqrt{\log(2^i)}]2^{1-i} \tag{255}$$

$$\leqslant \sum_{i=1}^{\infty} [C_1\sqrt{i}]2^{1-i} \leqslant C_1 \sum_{i=1}^{\infty} \sqrt{2}^{1-i} = \frac{C_1}{\sqrt{2}-1} \tag{256}$$

where at the second line we have used the fact that for any natural number $i$, $\sqrt{i} \leqslant \sqrt{2}^{i-1}$.

If we assume that $f(\delta) = C_3 \log(\frac{1}{\delta})$ we have instead

$$\mathbb{E}(f(\delta)) \leqslant \sum_{i=1}^{\infty} f(\delta_i)\delta_{i-1} \leqslant \sum_{i=1}^{\infty} [C_3 \log(2^i)]2^{1-i} \tag{257}$$

$$\leqslant C_3 \sum_{i=1}^{\infty} i2^{1-i} \leqslant C_3 \sum_{i=1}^{\infty} \frac{3}{2^{5/6}}\sqrt{2}^{i-1}2^{1-i} \leqslant \frac{3C_3}{2-\sqrt{2}} \leqslant 6C_3 \tag{258}$$

□

**Lemma F.7.** *Let $a, b, s > 0$ be three positive real numbers: we have*

$$\log(1 + ab^s) \leqslant \log((1 + a)(1 + b^s)) \leqslant \log(1 + a) + \log(1 + (1 + b)^s) \leqslant \log(1 + a) + \log(2(1 + b)^s)$$
$$\leqslant \log(2(1 + a)) + s \log(1 + b).$$

### F.3. Covering Numbers for Neural Networks

In this subsection, we collect variations of known results on covering numbers of classes of neural networks. These results are useful in Subsection E.3, where apply them to construct covers of the space of neural embeddings over elements of $[m] \times [n]$.

In line with much of the literature, we consider fully-connected neural networks of the following form:

$$f(x) = \sigma_L \left(A^L \sigma^{L-1} \left(\sigma^{L-2} \dots \sigma^1 \left(A^1 x\right) \dots\right)\right), \tag{259}$$

where the input $x \in \mathbb{R}^d$ and the output $f(x) \in \mathbb{R}^1$ and the activations $\sigma^\ell$ (for $\ell \leqslant L$) are assumed to be 1-Lipschitz. We write W for the total number of parameters of the network.

For a fixed architecture defined by the intermediary widths $w_1, \dots, w_L = 1$ (so that $\mathrm{W} = \sum_{\ell=1}^{L} w_L W_{L-1}$), and for a fixed set of constants $a_1, \dots, a_L$ and $s_1, s_2, \dots, s_L$ and initialization matrices $M^1, M^2, \dots, M^L$ we consider the $\mathcal{N}_{1,\mathrm{W}}(s, a)$ class of networks $f$ that satisfy the following conditions:

$$\left\|\left(A^\ell - M^\ell\right)^\top\right\|_{2,1} \leqslant a_\ell \qquad \forall \ell \leqslant L \qquad \text{and} \tag{260}$$

$$\left\|A^\ell\right\| \leqslant s_\ell \qquad \forall \ell \leqslant L. \tag{261}$$

The following is a particular case of (Graf et al., 2022), Theorem C.15 (cf also (Bartlett et al., 2017; Ledent et al., 2021c)) applied to fully-connected neural networks.

**Proposition F.8** ($L^2$ covering number for $\mathcal{N}_{1,\mathrm{W}}$)**.** *Assume that $\chi^2 = \frac{1}{N}\sum_{o=1}^{N}\|x_o\|^2 \geqslant 1$ and $\forall l, s_l \geqslant 1$. For any $\epsilon > 0$, there is an $L^2$ cover $\mathcal{C}(\epsilon)$ of $\mathcal{N}_{1,\mathrm{W}}$ with the following properties:*

1. *For any $f \in \mathcal{N}_{1,\mathrm{W}}$ there exists a $\bar{f}$ in $\mathcal{C}(\epsilon)$ such that*

$$\frac{1}{N}\sum_{o=1}^{N}\left|f(x_o) - \bar{f}(x_o)\right|^2 \leqslant \epsilon^2. \tag{262}$$

2.

$$\log(|\mathcal{C}|) \leqslant \left\lceil \frac{1}{\epsilon^2} \right\rceil \chi^2 \,\mathrm{R_W}^2 \log(2W), \tag{263}$$

*where*

$$\mathrm{R_W} = 2\prod_{\ell=1}^{L} s_\ell \left[\sum_{\ell=1}^{L}\left\lceil \frac{a_\ell}{s_\ell}\right\rceil^{2/3}\right]^{3/2} \qquad \textit{and} \qquad \chi^2 = \frac{1}{N}\sum_{o=1}^{N}\|x_o\|^2. \tag{264}$$

**Notes:** We omit the improved dependency on the output dimension which can be derived from the techniques of (Ledent et al., 2021c; Wu et al., 2021; Mustafa et al., 2021) since the output of our network is one dimensional. Such techniques can also be easily used to make the above cover uniform over the samples at the cost of additional logarithmic factors.

**Proposition F.9** (Uniform $L^\infty$ covering number for $\mathcal{N}_{2,W}$)**.** *Consider the class $\mathcal{N}_{2,W}(s,a)$ of fully connected neural networks with the same architecture as those in $\mathcal{N}_{1,\mathrm{W}}$, but where the weight matrices only need satisfy the following constraints:*

$$\|A^\ell\| \leqslant s_\ell \qquad \|A^\ell - M^\ell\| \leqslant a_l. \tag{265}$$

*We further assume that $s_\ell \geqslant 1$ for all $\ell \leqslant L$. For any $1 > \epsilon > 0$, and any $\ell$ Lipschitz loss function $\mathrm{l}$ there is an $L^2$ cover $\mathcal{C}(\epsilon)$ of $\mathcal{N}_{2,W}$ with the following properties:*

1. *For any $f \in \mathcal{N}_{1,\mathrm{W}}$ there exists a $\bar{f}$ in $\mathcal{C}(\epsilon)$ such that for **any** $x \in \mathbb{R}^d$ with $\|x\| \leqslant \chi$, we have:*

$$\left|\mathrm{l}\circ f(x) - \mathrm{l}\circ\bar{f}(x)\right| \leqslant \epsilon \tag{266}$$

2.

$$\log(|\mathcal{C}|) \leqslant W\log\left(\frac{6\chi\,\ell\left[\prod_{\ell=1}^{L} s_\ell\right]\left[\sum_\ell a_\ell\right]}{\epsilon} + 1\right). \tag{267}$$

**Remark**: This result and its proof are very similar to analogous results in (Long & Sedghi, 2020; Graf et al., 2022) (no claim of originality is made here), but the requirement on the cover is stricter than in the Theorem statement in (Graf et al., 2022). The control on the bounds is also looser, but we do not invest too much into the technicalities of obtaining tighter logarithmic factors, since the aim of this section is merely to illustrate how to combine our bounds on Schatten quasi-norm regularized matrices with neural network bounds.

*Proof.* For the sake of completeness, we repeat the main parts of the proof here, which mostly follow (Long & Sedghi, 2020).

Note that by a standard argument (see, e.g. pages 4,5 in (Long & Sedghi, 2020)), we have for any sets of matrices $A^1,\ldots,A^L$ and $\bar{A}^1,\ldots,\bar{A}^L$ satisfying the conditions (261) and corresponding networks $f$ and $\bar{f}$:

$$\left|f(x) - \bar{f}(x)\right| \leqslant \|x\|\left[\prod_{\ell=1}^{L} s_\ell\right]\sum_{\ell=1}^{L}\frac{\left\|A^\ell - \bar{A}^\ell\right\|}{s_\ell} \leqslant \|x\|\left[\prod_{\ell=1}^{L} s_\ell\right]\sum_{\ell=1}^{L}\left\|A^\ell - \bar{A}^\ell\right\| \tag{268}$$

with the inequality holding uniformly over any $x \in \mathbb{R}^d$.

Thus, as long as $\bar{\mathcal{C}}(\epsilon)$ is a cover of the space $\left\{\mathcal{A} = (A^1, \dots, A^L) : \|A^\ell - M^\ell\| \leqslant a_\ell \quad \forall \ell\right\}$ with respect to the norm $\|\mathcal{A}\| := \sum_{\ell=1}^L \|A^\ell\|$ with granularity $\varepsilon := \frac{\epsilon}{\chi\, \ell \prod_{\ell=1}^L s_\ell}$, the associated cover $\mathcal{C}(\epsilon) \subset \mathcal{N}_{2,W}$ gives a uniform $\epsilon$-cover of $1 \circ \mathcal{N}_{2,W}$. Furthermore, by Lemma F.17 and a doubling argument (to ensure the condition $\|A_\ell\| \leqslant s_\ell$ is also satisfied by each element of the cover) such a cover exists with cardinality $\leqslant$

$$\left(\frac{6 \sum_\ell a_\ell}{\varepsilon} + 1\right)^W = \left(\frac{6\chi\, \ell \left[\prod_{\ell=1}^L s_\ell\right] \left[\sum_\ell a_\ell\right]}{\epsilon} + 1\right)^W \tag{269}$$

and the result follows.

□

### F.4. A Result on the Estimation of the Marginals

**Lemma F.10** (Variation on Lemma 2 in (Foygel et al., 2011) and Lemma E.1 in (Ledent et al., 2021b))**.** *For any $\delta > 0$, if $N \geqslant 8(m+n)\log(\frac{(m+n)}{\delta})$ then with probability $\geqslant 1 - \delta$, the following holds for all $i \leqslant m$ and $j \leqslant n$:*

$$\check{p}_i \geqslant \frac{\tilde{p}_i}{2} \qquad \text{and} \qquad \check{q}_j \geqslant \frac{\tilde{q}_j}{2}. \tag{270}$$

*Proof.* The proof is almost identical to that of Lemma 2 in (Foygel et al., 2011) but we repeat it for completeness as we need our variant with arbitrarily high probability. Note that this lemma is also a particular case of the inductive case from Lemma E.1 in (Ledent et al., 2021b) with identity side information matrices.

If $p_i \leqslant \frac{1}{m}$ (resp. $q_j \leqslant \frac{1}{n}$), then $\tilde{p}_i \leqslant \frac{1}{m}$ and $\check{p}_i \geqslant \frac{1}{2m}$ (resp. $\tilde{q}_j \leqslant \frac{1}{n}$ and $\check{q}_j \geqslant \frac{1}{2n}$). On the other hand, if $p_i > \frac{1}{n}$ then by a multiplicative Chernoff bound (Lemma F.15), we have for any $i \leqslant m$:

$$\mathbb{P}\left(\hat{p}_i < \frac{p_i}{2}\right) \leqslant \exp\left(-\frac{N p_i}{8}\right) \leqslant \exp\left(-\frac{N}{8m}\right) \leqslant \frac{\delta}{(m+n)}, \tag{271}$$

where at the last inequality we have used the fact that $N \geqslant 8(m+n)\log(\frac{(m+n)}{\delta})$.

Similarly, for all $j \leqslant n$:

$$\mathbb{P}\left(\hat{q}_i < \frac{q_j}{2}\right) \leqslant \exp\left(-\frac{N q_j}{8}\right) \leqslant \exp\left(-\frac{N}{8n}\right) \leqslant \frac{\delta}{(m+n)}. \tag{272}$$

The result then follows immediately from a union bound. □

### F.5. Basic Covering Numbers, Concentration Inequalities and Classic Results in Learning Theory

In this section, we summarize some existing results which are useful to our study.

**Theorem F.11** (Generalization bound from Rademacher complexity, cf. e.g., (Bartlett & Mendelson, 2001), (Scott, 2014), (Guermeur, 2020) etc.)**.** *Let $Z, Z_1, \dots, Z_N$ be iid random variables taking values in a set $\mathcal{Z}$. Consider a set of functions $\mathcal{F} \in [0,1]^{\mathcal{Z}}$. $\forall \delta > 0$, we have with probability $\geqslant 1 - \delta$ over the draw of the sample $S$ that*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(f(Z)) \leqslant \frac{1}{N}\sum_{i=1}^N f(z_i) + 2\,\mathfrak{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2N}},$$

*where $\mathfrak{R}_S(\mathcal{F})$ can be either the empirical or expected Rademacher complexity. In particular, if $f^* \in \arg\min_{f\in\mathcal{F}} \mathbb{E}(f(Z))$ and $\hat{f} \in \arg\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N f(z_i)$, then*

$$\mathbb{E}(\hat{f}(Z)) \leqslant \mathbb{E}(f^*(Z)) + 4\,\mathfrak{R}_S(\mathcal{F}) + 6\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Recall the following theorem on the covering number of classes of Lipschitz functions.

**Proposition F.12** (Covering number of Lipschitz function balls, see (von Luxburg & Bousquet, 2004), Theorem 17 page 684, see also (Tikhomirov, 1993))**.** *Let $\mathcal{X}$ be a connected and centred metric space (i.e. for all $A \subset \mathcal{X}$ with $\operatorname{diam}(A) \leqslant 2r$ there exists an $x \in \mathcal{X}$ such that $d(x,a) \leqslant r$ for all $a \in A$). Let $B$ denote the set of 1-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ which are uniformly bounded by $\operatorname{diam}(\mathcal{X})$. For any $\epsilon > 0$ we have the following bound on the covering number of the class $B$ as a function of the covering number of $\mathcal{X}$:*

$$\mathcal{N}(B, \epsilon, \|\cdot\|_\infty) \leqslant \left( \left\lceil \frac{2 \operatorname{diam}(\mathcal{X})}{\epsilon} \right\rceil + 1 \right) \times 2^{\mathcal{N}(\mathcal{X}, \frac{\epsilon}{2}, d)}. \tag{273}$$

Applying the above to the $d$ dimensional euclidean space, we immediately obtain:

**Proposition F.13.** *Let $\|\cdot\|_{\max}$ denote max norm on $\mathbb{R}^d$, i.e. $\|x\|_d = \max_i |x_i|$. Let $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}$ denote the set of $\mathrm{L_f}$-Lipschitz functions from $[-\mathcal{B}_0, \mathcal{B}_0]^d$ to $\mathbb{R}$. We have the following bound on the covering number of $\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f}$ with respect to the $L^\infty$ (uniform) norm on functions:*

$$\log\left(\mathcal{N}(\mathcal{F}_{lip,\mathrm{L_f},\mathcal{B}_f})\right) \leqslant \log\left( \left\lceil \frac{4\,\mathcal{B}_0\,\mathrm{L_f}}{\epsilon} \right\rceil + 1 \right) + \left[ \left\lceil \frac{2\,\mathcal{B}_0\,\mathrm{L_f}}{\epsilon} \right\rceil + 1 \right]^d \log(2) \tag{274}$$

$$\leqslant 3 \left[ \left\lceil \frac{2\,\mathcal{B}_0\,\mathrm{L_f}}{\epsilon} \right\rceil + 1 \right]^d \tag{275}$$

*Proof.* W.l.o.g, let $\mathrm{L_f} = 1$. Then, take the following $\epsilon/2$ cover of $[-\mathcal{B}_0, \mathcal{B}_0]^d$: $[[-\mathcal{B}_0, \mathcal{B}_0] \cap \epsilon\mathbb{Z}]^d$, which has cardinality less than $\left[\left\lceil \frac{2\mathcal{B}_0}{\epsilon} \right\rceil + 1\right]^d$. □

**Proposition F.14** (Massart's finite class lemma)**.** *Let $A \subset \mathbb{R}^N$ be a finite class of functions from $[N]$ to $\mathbb{R}$. Let $r = \max_{u \in A} \|u\|_2$. We have the following bound on the Rademacher complexity of $A$ over the sample $[N]$:*

$$\mathbb{E}_\sigma \left( \frac{1}{N} \sup_{u \in A} \sum_{i=1}^N \sigma_i u_i \right) \leqslant \frac{r\sqrt{2 \log \#(A)}}{N}. \tag{276}$$

**Lemma F.15** (Multiplicative Chernoff bound, well known, Cf e.g. Corollary 13.3 (pp. 13-3 and 13-4) in (Sinclair). i.i.d. case originally from (Angluin & Valiant, 1979), Proposition 2.4 p 158, cf. also (Boucheron et al., 2004) (exercise 2.10 on p. 48) and (Hagerup & Rüb, 1990))**.** *Suppose $X_1, \ldots, X_N$ are independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{i=1}^N X_i$ denote their sum. For any $\delta > 0$ we have*

$$\mathbb{P}(X \geqslant (1+\delta)\mathbb{E}(X)) \leqslant \exp\left( -\frac{\delta^2 \mathbb{E}(X)}{2+\delta} \right). \tag{277}$$

*In addition, for all $0 < \delta < 1$ we have*

$$\mathbb{P}(X \leqslant (1-\delta)\mathbb{E}(X)) \leqslant \exp\left( -\frac{\delta^2 \mathbb{E}(X)}{2} \right). \tag{278}$$

We will need the following further consequence in our analysis:

**Corollary F.16.** *Let $m \in \mathbb{N}$ and $\eta_i$ be independent categorical variables on the domain $\{1, 2, \ldots, m\}$. For all $j \leqslant m$ let us write $X^j := \sum_{i=1}^N 1(\eta_i = j)$ for the number of $\eta_i$s which assume the value $j$. For any $0 < \delta < 1$ have*

$$\mathbb{P}\left(\exists j,\ \textit{s.t. } X^j \leqslant (1-\delta)\mathbb{E}(X^j)\right) \leqslant m \exp\left( -\frac{N\delta^2\mu}{2} \right), \tag{279}$$

*where $\mu \min_j \mathbb{E}(X^j)$.*

**Lemma F.17** (Lemma 8 in (Long & Sedghi, 2020))**.** *The (internal) covering number $\mathcal{N}$ of the ball of radius $\kappa$ in dimension $d$ (with respect to any norm $\|\cdot\|$) can be bounded by:*

$$\mathcal{N} \leqslant \left\lceil \frac{3\kappa}{\epsilon} \right\rceil^d \leqslant \left( \frac{3\kappa}{\epsilon} + 1 \right)^d \tag{280}$$

Recall the following result from (Mazumder et al., 2010):

**Lemma F.18** (Lemma 6 in (Mazumder et al., 2010))**.** *For any matrix $X \in \mathbb{R}^{m\times n}$, the following holds:*

$$\|X\|_* = \min_{A,B,AB^\top = X} \frac{1}{2}\left[\|A\|_{\mathrm{Fr}}^2 + \|B\|_{\mathrm{Fr}}^2\right]. \tag{281}$$

Recall the following useful result, which is an immediate consequence of the McDiarmid inequality. A similar result was presented in (Bartlett & Mendelson, 2001) (cf. Theorem 11 page 469) for the expected Rademacher complexity. See also (Ledent et al., 2021b) for the exact result.

**Lemma F.19.** *For any fixed $x_1, \ldots, x_N$ and any function class $\mathcal{F}$ mapping to $[-1,1]$ we have with probability $\geqslant 1-\delta$ over the draw of the Rademacher variables $\sigma_1, \ldots, \sigma_N$,*

$$\left|\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \sigma_i f(x_i) - \widehat{\mathfrak{R}}_{(x_1,\ldots,x_n)}(\mathcal{F})\right| \leqslant \sqrt{\frac{2\log(2/\delta)}{N}}. \tag{282}$$

**Proposition F.20** (Non commutative Bernstein inequality, Cf. (Recht, 2011))**.** *Let $X_1, \ldots, X_S$ be independent, zero mean random matrices of dimension $m \times n$. For all $k$, assume $\|X_k\| \leqslant M$ almost surely, and denote $\rho_k^2 = \max(\|\mathbb{E}(X_k X_k^\top)\|, \|\mathbb{E}(X_k^\top X_k)\|)$ and $\nu^2 = \sum_k \rho_k^2$. For any $\tau > 0$,*

$$\mathbb{P}\left(\left\|\sum_{k=1}^S X_k\right\| \geqslant \tau\right) \leqslant (m+n)\exp\left(-\frac{\tau^2/2}{\sum_{k=1}^S \rho_k^2 + M\tau/3}\right). \tag{283}$$

**Proposition F.21** (High probability version of Bernstein inequality)**.** *Let $X_1, \ldots, X_S$ be independent, zero mean random matrices of dimension $m \times n$. For all $k$, assume $\|X_k\| \leqslant M$ almost surely, and denote $\rho_k^2 = \max(\|\mathbb{E}(X_k X_k^\top)\|, \|\mathbb{E}(X_k^\top X_k)\|)$. Writing $\sigma^2 = \sum_{k=1}^S \rho_k^2$, for any $\delta > 0$, we have, with probability greater than $1-\delta$:*

$$\left\|\sum_{k=1}^S X_k\right\| \leqslant \sqrt{8/3}\sigma\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{8M}{3}\log\left(\frac{m+n}{\delta}\right). \tag{284}$$

*Proof.* From Proposition F.20, we can make the following conclusions splitting into two cases depending on whether $M\tau \leqslant \sigma^2$ or $M\tau \geqslant \sigma^2$:

If $M\tau \leqslant \sigma^2$, we have, with probability $\geqslant 1-\delta$:

$$\left\|\sum_{k=1}^S X_k\right\| \leqslant \sqrt{8/3}\sigma\sqrt{\log\left(\frac{m+n}{\delta}\right)}. \tag{285}$$

Similarly, if $M\tau \geqslant \sigma^2$, we have, with probability $\geqslant 1-\delta$:

$$\left\|\sum_{k=1}^S X_k\right\| \leqslant \frac{8M}{3}\log\left(\frac{m+n}{N}\right). \tag{286}$$

Thus, in all cases, we certainly have, with probability greater than $1-\delta$:

$$\left\|\sum_{k=1}^S X_k\right\| \leqslant \sqrt{8/3}\sigma\sqrt{\log\left(\frac{m+n}{\delta}\right)} + \frac{8M}{3}\log\left(\frac{m+n}{\delta}\right), \tag{287}$$

as expected. □

### F.6. Deterministic Results

In this subsection, we show how some of the most popular recent neural network models indirectly contain Schatten norm regularized matrix components.

One of the first descriptions of non-linear matrix factorization methods appears in (Dziugaite & Roy, 2015), which describes the following very general class of models:

$$g(i,j) = f_\theta\left(u_i^1, v_j^1, u_i^2 \circ v_j^2, \dots, u_i^{\underline{m}} \circ v_j^{\underline{m}}\right), \tag{288}$$

where $\circ$ denotes an element wise product, $f_\theta$ is a trainable neural network and $u^1, \dots, u^{\underline{m}}$ (resp. $v^1, \dots, v^{\underline{m}}$) are low dimensional row (resp. column) embeddings. A particularly famous example of a more specific architectural variation is the model presented in (He et al., 2017), which has achieved state of the art performance in various recommender systems datasets. From an architectural perspective, the model can be described as follows.

$$g(i,j) = \left\langle \text{concat}(u_i' \circ v_j', f_1(u_i, v_j)), \text{concat}(d, d') \right\rangle, \tag{289}$$

where $f_1$ is a neural network with a mutli-dimensional output, and $\text{concat}(d, d')$ is a vector representing the last linear layer. Since then, further models have been proposed which rely on expressing the set of observed entries as a graph through which one can propagate the embeddings (He et al., 2020; Zhang & Chen, 2020; Mao et al., 2021), to cite but a few.

In Corollary F.23 below, we show that a global minimum of (1) can always be attained with the additional constraint that $D_1, \dots D_{d-2}$ be diagonal matrices. In particular, this further cements the validity of the Schatten norm as a regularizer: the models (288) and (1) also hide implicit Schatten norm regularized components. For instance, the model in (He et al., 2017) is in fact equivalent to

$$g(i,j) = Z_{i,j} + \Psi(i,j), \tag{290}$$

where $\Psi(i,j) = \langle D', f_2(u_i, v_j)\rangle$ is a neural network encoding derived from $f_1$ with an additional linear layer, and the matrix $Z$ satisfies $Z_{i,j} = \langle (u_i \circ v_j), d\rangle = u_i^\top \operatorname{diag}(d) v_j^\top$ so that $Z = UDV^\top$ for $D = \operatorname{diag}(d)$ where the rows of $U$ (resp. $V$) collect the row (resp. column) embeddings. Thus, imposing $L2$ regularization (which is arguably implicitly present in popular optimization schemes such as gradient descent) on the parameters $d, U, V$ is equivalent to imposing regularization on the Schatten quasi-norm of $Z$ with $p = \frac{2}{3}$.

We first recall the following reformulation of Theorem 1 in (Dai et al., 2021), which can be interpreted as a generalization of Lemma F.18 (i.e. Lemma 6 in (Mazumder et al., 2010)):

**Theorem F.22** (Theorem 1 in (Dai et al., 2021))**.** *Let $Z \in \mathbb{R}^{m\times n}$, for any integers $d \in \mathbb{N}$ and $o \in \mathbb{N}$ with $o \geqslant \min(m,n)$ we have*

$$\min_{\substack{W_d \in \mathbb{R}^{o\times n}, W_1 \in \mathbb{R}^{m\times o} \\ W_k \in \mathbb{R}^{o\times o} \forall k \neq 1,d}} \left( \sum_{k=1}^d \|W_k\|_{\text{Fr}}^2 \;\middle|\; \prod_{k=1}^d W_k = Z \right) = d \sum_{u=1}^r \sigma_u^{2/d} = d\|Z\|_{\text{sc},2/d}^{2/d}, \tag{291}$$

*where $r = rank(Z)$, the $\sigma_u$s are the singular values of $Z$ and $\|\cdot\|_{\text{sc},p}$ is the $p$-Schatten quasi-norm.*

From this, we obtain the following result:

**Corollary F.23.** *Let $Z \in \mathbb{R}^{m\times n}$, for any integers $d \in \mathbb{N}$ and $o \in \mathbb{N}$ with $o \geqslant \min(m,n)$ and $d \geqslant 2$ we have*

$$\begin{aligned} &\min_{\substack{A\in\mathbb{R}^{m\times o}, B\in\mathbb{R}^{n\times o} \\ \forall k \leqslant d-2,\ D_k = \operatorname{diag}(d^k),\ d_k \in \mathbb{R}^o}} \left( \|A\|_{\text{Fr}}^2 + \|B\|_{\text{Fr}}^2 + \sum_{k=1}^{d-2} \|D_k\|_{\text{Fr}}^2 \;\middle|\; A\left[\prod_{k=1}^{d-2} D_k\right] B^\top = Z \right) \\ &= d \sum_{k=1}^r \sigma_k^{2/d} = d\|Z\|_{2/d}^{SC}, \end{aligned} \tag{292}$$

*where the minimum runs over all matrices $A \in \mathbb{R}^{m\times o}, B \in \mathbb{R}^{n\times o}$ and $D_1, \dots, D_{d-2} \in \mathcal{Q}$ where $\mathcal{Q}$ is either the set of all matrices in $\mathbb{R}^{o\times o}$ or the set of all diagonal matrices in $\mathbb{R}^{o\times o}$. (As in Theorem F.22 $r = rank(Z)$, the $\sigma_u$s are the singular values of $Z$ and $\|\cdot\|_{\text{sc},p}$ is the $p$-Schatten quasi-norm.)*

*In particular, for $d = 3$, we have*

$$\min_{\substack{A\in\mathbb{R}^{m\times o}, B\in\mathbb{R}^{n\times o}, \\ \mathbb{R}^{o\times o} D = \operatorname{diag}(d), d\in\mathbb{R}^o}} \left( \|A\|_{\text{Fr}}^2 + \|B\|_{\text{Fr}}^2 + \|D\|_{\text{Fr}}^2 \middle| ADB^\top = Z \right) = 3 \sum_{k=1}^r \sigma_k^{2/3} = 3\|Z\|_{\text{sc},2/3}^{2/3}. \tag{293}$$

*Proof.* We prove the theorem in two directions.

**LHS $\geqslant$ RHS:**

This follows immediately from Theorem F.22, since our set of admissible factor matrices $(A, D_1, \ldots, d_{d-2}, B^\top)$ is a subset of the set of admissible $(W_1, W_2, \ldots, W_d)$ in Theorem F.22 (as a result of the additional constraint that the matrices $D_1, \ldots, D_{d-2}$ must be in $\mathcal{Q}$).

**LHS $\leqslant$ RHS:**

This follows by constructing a set of matrices $A, B, D_1, \ldots, D_{d-2}$ which achieves the minimum whilst satisfying the strictest constraint that $D_1, \ldots, D_{d-2}$ are diagonal matrices. For this, let $Z = U\Sigma V^\top$ be the singular value decomposition of $Z$ with additional dimensions chosen such that $U \in \mathbb{R}^{m\times o}, \Sigma \in \mathbb{R}^{o\times o}, V \in \mathbb{R}^{n\times o}$. We now choose:

$$\begin{aligned} A &= U\Sigma^{1/d} \\ B &= V\Sigma^{1/d} \\ D_k &= \Sigma^{1/d} \quad (\forall k \leqslant d-2). \end{aligned} \tag{294}$$

It is clear that

$$A\left[\prod_{k=1}^{d-2} D_k\right] B^\top = U\Sigma^{1/d}\Sigma^{(d-2)/d}\Sigma^{1/d}V^\top = U\Sigma V^\top = Z.$$

In addition, we by the invariance of the Frobenius norm to rotations, we also have

$$\begin{aligned} \|A\|_{\mathrm{Fr}}^2 + \|B\|_{\mathrm{Fr}}^2 + \sum_{k=1}^{d-2}\|D_k\|_{\mathrm{Fr}}^2 &= \|U\Sigma^{1/d}\|_{\mathrm{Fr}}^2 + \|V\Sigma^{1/d}\|_{\mathrm{Fr}}^2 + \sum_{k=1}^{d-2}\|\Sigma^{1/d}\|_{\mathrm{Fr}}^2 \\ &= \|\Sigma^{1/d}\|_{\mathrm{Fr}}^2 + \|\Sigma^{1/d}\|_{\mathrm{Fr}}^2 \sum_{k=1}^{d-2}\|\Sigma^{1/d}\|_{\mathrm{Fr}}^2 = d\|\Sigma^{1/d}\|_{\mathrm{Fr}}^2 \\ &= d\sum_{k=1}^{r}\sigma_k^{2/d} = d\|Z\|_{2/d}^{SC} = \text{RHS}, \end{aligned} \tag{295}$$

as expected. The result follows. □

We have the following immediate variant of Corollary F.23, which shows how we can add the weights to our regularizers in our practical experiments.

**Corollary F.24.** *Let $\check{p} \in \mathbb{R}^m$ and $\check{q} \in \mathbb{R}^n$ be arbitrary vectors and let $Z \in \mathbb{R}^{m\times n}$, for any integers $d \in \mathbb{N}$ and $o \in \mathbb{N}$ with $o \geqslant \min(m, n)$ and $d \geqslant 2$ we have*

$$\begin{aligned} &\min_{\substack{A\in\mathbb{R}^{m\times o}, B\in\mathbb{R}^{n\times o} \\ \forall k\leqslant d-2,\ D_k\in\mathcal{Q}}} \left( \|\check{A}\|_{\mathrm{Fr}}^2 + \|\check{B}\|_{\mathrm{Fr}}^2 + \sum_{k=1}^{d-2}\|D_k\|_{\mathrm{Fr}}^2 \;\middle|\; A\left[\prod_{k=1}^{d-2} D_k\right] B^\top = Z \right) \\ &= d\sum_{k=1}^{r}\sigma_k^{2/d} = d\|\check{Z}\|_{2/d}^{SC}, \end{aligned} \tag{296}$$

*where $\check{A} := \mathrm{diag}(\sqrt{\check{p}})A$, $\check{B} := \mathrm{diag}(\sqrt{\check{q}})B$, $\check{Z} = \mathrm{diag}(\sqrt{\check{p}})Z\,\mathrm{diag}(\sqrt{\check{q}})$ and the minimum runs over all matrices $A \in \mathbb{R}^{m\times o}, B \in \mathbb{R}^{n\times o}$ and $D_1, \ldots, D_{d-2} \in \mathcal{Q}$ where $\mathcal{Q}$ is either the set of all matrices in $\mathbb{R}^{o\times o}$ or the set of all diagonal matrices in $\mathbb{R}^{o\times o}$.*

## G. Extention: Multiple Latent Matrices

In this section, we develop tools to extend some of our results to situations where there are multiple latent matrices, possibly with different Schatten indices. This section is mostly illustrative: the models themselves are quite complicated, making the bounds less meaningful than in other sections. In addition, the dependence on the number of latent factors obtainable with the techniques below is quite strong. However, the tools developed show a general strategy which could be used for a broad class of similar models.

### G.1. Extension: Multi-latent Lipschitz Decomposition Lemmas

In this subsection, we prove some new results, analogous both Talagrand's concentration lemma and Dudley's entropy theorem, aimed at bounding the Rademacher complexity of $\ell(\mathcal{F}_1, \ldots, \mathcal{F}_m)$ where $\mathcal{F}_v$ (for $v \leqslant m$) are function classes and $\ell$ is Lipschitz. The aim is to be able to bound the Rademacher complexity of $\mathcal{F}$ even if a covering number is not available for any of the $\mathcal{F}_v$.

We begin by remind the reader of the following classic.

**Lemma G.1** (Talagrand contraction lemma (cf. (Ledoux & Talagrand, 1991) see also (Meir & Zhang, 2003) page 846))**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz. Consider the set of functions $\{f_i(\theta), i \leqslant N\}$ (on $\{1, 2, \ldots, N\}$) depending on a parameter $\theta \in \Theta$.*

*For any function $c(x, \theta)$ where $x \in X$ and any probability distribution on $X$, we have*

$$\mathbb{E}_\sigma \mathbb{E}_X \sup_{\theta \in \Theta} \left\{ c(X, \theta) + \sum_{i=1}^{N} \sigma_i g(f_i(\theta)) \right\} \leqslant \mathbb{E}_\sigma \mathbb{E}_\eta \mathbb{E}_X \sup_{\theta \in \Theta} \left\{ c(X, \theta) + \sum_{i=1}^{N} \sigma_i f_i(\theta) \right\}, \tag{297}$$

*where the $\sigma_i$s are i.i.d. Rademacher variables.*

We then present our first extension of the above:

**Lemma G.2.** *Let $g : \mathbb{R}^m \to \mathbb{R}$ be a function satisfying the following Lipschitz condition:*

$$\left| g(y^2) - g(y^1) \right| \leqslant \sum_{k \leqslant \mathrm{m}} \left| y_k^2 - y_k^1 \right| \lambda_k \tag{298}$$

*with $\sum_{k=1}^{\mathrm{m}} \lambda_k = \ell$. Consider $\left\{ f_i^1(\theta) \right\}, \left\{ f_i^2(\theta) \right\}, \ldots, \{ f_i^m(\theta) \}$, $i \leqslant N$ functions (on $\{1, 2, \ldots, N\}$) depending on a parameter $\theta \in \Theta$.*

*Define the function $g$ on $\{1, 2 \ldots, N\}$ by $g_i(\theta) = g(f_i^1(\theta), f_i^2(\theta), \ldots, f_i^m(\theta))$ for all $i \leqslant N$.*

*For any function $c(x, \theta)$ where $x \in X$ and any probability distribution on $X$, we have*

$$\mathbb{E}_\sigma \mathbb{E}_X \sup_{\theta \in \Theta} \left\{ c(X, \theta) + \sum_{i=1}^{N} \sigma_i g_i(\theta) \right\} \leqslant \mathbb{E}_\sigma \mathbb{E}_\eta \mathbb{E}_X \sup_{\theta \in \Theta} \left\{ c(X, \theta) + \ell \sum_{i=1}^{N} \sigma_i f_i^{\eta_i}(\theta) \right\}, \tag{299}$$

*where the $\sigma_i$s are i.i.d. Rademacher variables and the $\eta_i$ are i.i.d. categorical variables on the domain $\{1, 2, \ldots, m\}$ with corresponding probabilities $\lambda_1/\ell, \ldots, \lambda_m/\ell$.*

*N.B.:* The case $m = 1$ is the standard Talagrand concentration Lemma.

*Proof.* W.l.o.g., we assume $\ell = 1$. The proof is by induction over $N$. The initial case $N = 1$ certainly holds. Assume it holds for $N$, we will show it holds for $N + 1$.

$$
\begin{aligned}
&\mathbb{E}_{\sigma_1,\ldots,\sigma_{N+1}}\mathbb{E}_X \sup_{\theta\in\Theta}\left\{c(X,\theta)+\sum_{i=1}^{N+1}\sigma_i g_i(\theta)\right\}\\
&\leqslant \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X \sup_{\theta_1,\theta_2\in\Theta}\left\{\frac{c(X,\theta_1)+c(X,\theta_2)}{2}+\sum_{i=1}^{N}\sigma_i\frac{g_i(\theta_1)+g_i(\theta_2)}{2}+\frac{g_{N+1}(\theta_1)-g_{N+1}(\theta_2)}{2}\right\}\\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X \sup_{\theta_1,\theta_2\in\Theta}\left\{\frac{c(X,\theta_1)+c(X,\theta_2)}{2}+\sum_{i=1}^{N}\sigma_i\frac{g_i(\theta_1)+g_i(\theta_2)}{2}+\frac{|g_{N+1}(\theta_1)-g_{N+1}(\theta_2)|}{2}\right\}\\
&\leqslant \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X \sup_{\theta_1,\theta_2\in\Theta}\left\{\frac{c(X,\theta_1)+c(X,\theta_2)}{2}\right.\\
&\qquad\qquad\left.+\sum_{i=1}^{N}\sigma_i\frac{g_i(\theta_1)+g_i(\theta_2)}{2}+\sum_{j=1}^{m}\frac{\lambda_j|f^j_{N+1}(\theta_1)-f^j_{N+1}(\theta_2)|}{2}\right\}\\
&\leqslant \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X \sum_{j=1}^{m}\sup_{\theta_1,\theta_2\in\Theta}\left\{\frac{\lambda_j c(X,\theta_1)+\lambda_j c(X,\theta_2)}{2\,\ell}\right.\\
&\qquad\qquad\left.+\sum_{i=1}^{N}\sigma_i\lambda_j\frac{g_i(\theta_1)+g_i(\theta_2)}{2\,\ell}+\frac{\lambda_j|f^j_{N+1}(\theta_1)-f^j_{N+1}(\theta_2)|}{2}\right\}\\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X \sum_{j=1}^{m}\sup_{\theta_1,\theta_2\in\Theta}\left\{\frac{\lambda_j c(X,\theta_1)+\lambda_j c(X,\theta_2)}{2\,\ell}\right.\\
&\qquad\qquad\left.+\sum_{i=1}^{N}\sigma_i\lambda_j\frac{g_i(\theta_1)+g_i(\theta_2)}{2\,\ell}+\frac{\lambda_j\left[f^j_{N+1}(\theta_1)-f^j_{N+1}(\theta_2)\right]}{2}\right\}\\
&\leqslant \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X\mathbb{E}_{\sigma_{N+1}}\sum_{j=1}^{m}\sup_{\theta\in\Theta}\left\{\frac{\lambda_j c(X,\theta)}{\ell}+\sum_{i=1}^{N}\sigma_i\frac{\lambda_j g_i(\theta)}{\ell}+\sigma_{N+1}\lambda_j f^j_{N+1}(\theta)\right\}\\
&\leqslant \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\mathbb{E}_X\mathbb{E}_{\sigma_{N+1}}\mathbb{E}_{\substack{\eta\in[m]^N\\ \eta\sim\lambda}}\sup_{\theta\in\Theta}\left\{c(X,\theta)+\ell\sum_{i=1}^{N}\sigma_i f_i^{\eta_i}(\theta)+\sum_{j=1}^{m}\sigma_{N+1}\lambda_j f^j_{N+1}(\theta)\right\} \qquad (300)\\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_N,\sigma_{N+1}}\mathbb{E}_X\mathbb{E}_{\substack{\eta\in[m]^{N+1}\\ \eta\sim\lambda}}\sup_{\theta\in\Theta}\left\{c(X,\theta)+\ell\sum_{i=1}^{N+1}\sigma_i f_i^{\eta_i}(\theta)\right\},
\end{aligned}
$$

where the line (300) follows from the induction hypothesis. This completes the proof.

□

Using this, we then obtain the following result.

**Proposition G.3.** *Let $g:\mathbb{R}^m\to\mathbb{R}$ be a function satisfying the following condition:*

$$
\left|g(y^2)-g(y^1)\right|\leqslant\sum_{k\leqslant \mathrm{m}}\left|y^2_k-y^1_k\right|\lambda_k \qquad (301)
$$

*with $\sum_{k=1}^{\mathrm{m}}\lambda_k=\ell$. Consider $m$ function classes $\mathcal{F}_1,\ldots,\mathcal{F}_m$ from $[N]$ to $[-\mathcal{B},\mathcal{B}]$ and define $\mathcal{G}:=\left\{\tilde{g}:x\to\tilde{g}(x)=g(f^1(x),f^2(x),\ldots,f^m(x)):f_1\in\mathcal{F}_1,\ldots,f_m\in\mathcal{F}_m\right\}$. We have the following bound on the Rademacher complexity of $\mathcal{G}$:*

$$\widehat{\mathfrak{R}}(\mathcal{G}) = \mathbb{E}_\sigma \sup_{f_1,\dots,f_m} \frac{1}{N}\sum_{o=1}^{N} \sigma_o g(f^1(\mathrm{o}), f^2(\mathrm{o}), \dots, f^m(\mathrm{o})) \tag{302}$$

$$\leqslant \mathcal{B}\,\ell \frac{13m\log(2mN)}{N} + \ell \sum_{j\in J_1} \frac{\lambda_j}{\ell} \max_{k \geqslant \frac{N\lambda_j}{2\ell}} \widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j), \tag{303}$$

*where* $\widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j) := \mathbb{E}_{\substack{C\subset[N]\\|C|=k}} \mathbb{E}_{\sigma\in\{-1,1\}^C} \sup_{f\in\mathcal{F}^j} \frac{1}{k}\sum_{o\in C}\sigma_o f(o)$.

*Proof.* By Lemma G.2 we have

$$\mathfrak{R}(\mathcal{G}) \leqslant \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f_1,\dots,f_m} \frac{1}{N}\sum_{o=1}^{N} \sigma_o f^{\eta^i}(\mathrm{o}) \tag{304}$$

$$= \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f_1,\dots,f_m} \mathbb{E}_{\mathrm{o}}\sigma_o f^{\eta^i}(\mathrm{o}), \tag{305}$$

where the expectation over o runs over the uniform distribution over $[N]$. Now, let $C_j = \{i : \eta_i = j\}$. By Corollary F.16, we have that with probability greater than $1-\delta/2$ over the draw of the variables $\eta_1,\dots,\eta_N$,

$$|C_j| \geqslant \frac{\lambda_j}{2\,\ell} \qquad \text{for all} \qquad j \quad \text{s.t.} \qquad \lambda_j \geqslant \gamma := \frac{8\ell\log(\frac{2m}{\delta})}{N}. \tag{306}$$

Similarly, by a Chernoff bound (see Lemma F.15), w.p. $\geqslant 1-\delta/2$, the following holds for all $j \leqslant m$:

$$\frac{|C_j|}{N} \leqslant \frac{\lambda_j}{\ell} + \sqrt{\frac{\lambda_j}{\ell}\frac{3\log\left(\frac{2m}{\delta}\right)}{N}}. \tag{307}$$

Thus, with overall probability greater than $1-\delta$ over the draw of the $\eta^i$s, both Equations (306) and (307) hold under their respective constraints, which implies that we can continue the calculation from Equation (305) as follows, where we write $J_1$ (resp. $J_2$) for the sets of indices satisfying (resp. not satisfying) the second equation in Equation (306)

$$\begin{aligned}
&\mathfrak{R}(\mathcal{G}) \leqslant \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f_1,\dots,f_m} \mathbb{E}_{\mathrm{o}}\sigma_o f^{\eta^i}(\mathrm{o})\\
&\leqslant \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f_1,\dots,f_m} \sum_j \frac{|C_j|}{N}\mathbb{E}_{o\in C_j}\sigma_o f^{\eta^i}(\mathrm{o})\\
&\leqslant \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f_1,\dots,f_m} \sum_{j\in J_1} \frac{|C_j|}{N}\mathbb{E}_{o\in C_j}\sigma_o f^{j}(\mathrm{o}) + \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f_1,\dots,f_m} \sum_{j\in J_2} \frac{|C_j|}{N}\mathbb{E}_{o\in C_j}\sigma_o f^{j}(\mathrm{o})\\
&\leqslant \ell \sum_{j\in J_1} \mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f^j} \frac{|C_j|}{N}\mathbb{E}_{o\in C_j}\sigma_o f^{j}(\mathrm{o}) + \sum_{j\in J_2} \ell\,\mathbb{E}_\sigma\mathbb{E}_\eta \sup_{f^j} \frac{|C_j|}{N}\mathbb{E}_{o\in C_j}\sigma_o f^{j}(\mathrm{o})\\
&\leqslant \delta\,\mathcal{B}\,\ell + \mathcal{B}\,\ell \sum_{j\in J_2}\left[\frac{\lambda_j}{\ell} + \sqrt{\frac{\lambda_j}{\ell}\frac{3\log\left(\frac{2m}{\delta}\right)}{N}}\right] + \ell\sum_{j\in J_1}\frac{\lambda_j}{\ell}\max_{k\geqslant\frac{N\lambda_j}{2\ell}} \mathbb{E}_{\substack{C_j:\\|C_j|=k}} \mathbb{E}_\sigma \sup_{f^j} \mathbb{E}_{o\in C_j}\sigma_o f^{j}(\mathrm{o}) \qquad (308)\\
&\leqslant \frac{\mathcal{B}\,\ell}{N} + \mathcal{B}\,\ell\left[\frac{8\log(2mN)}{N} + \sqrt{\frac{8\log(2mN)}{N}\frac{3\log(2mN)}{N}}\right] + \ell\sum_{j\in J_1}\frac{\lambda_j}{\ell}\max_{k\geqslant\frac{N\lambda_j}{2\ell}}\widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j) \qquad (309)\\
&\leqslant \mathcal{B}\,\ell\frac{13m\log(2mN)}{N} + \ell\sum_{j\in J_1}\frac{\lambda_j}{\ell}\max_{k\geqslant\frac{N\lambda_j}{2\ell}}\widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j) \qquad (310)
\end{aligned}$$

where at line (309), we have simply set $\delta = \frac{1}{N}$ and replaced the definition of $\mathfrak{R}_{N,k}(\mathcal{F}_j)$. The result follows.

□

Finally, the next corollary is the result we need for our analysis of NNmSd(+NN).

**Corollary G.4.** *Let $g : \mathbb{R}^m \to \mathbb{R}$ satisfy the conditions of Proposition G.3 and consider $m$ functions classes $\mathcal{F}_1, \ldots, \mathcal{F}_m$ from $\mathcal{X}$ to $[-\mathcal{B}, \mathcal{B}]$. As in Proposition G.3, define $\mathcal{G} := \left\{\tilde{g} : x \to \tilde{g}(x) = g(f^1(x), f^2(x), \ldots, f^m(x)) : f_1 \in \mathcal{F}_1, \ldots, f_m \in \mathcal{F}_m\right\}$. Assume that the Rademacher complexities of the individual function classes $\mathcal{F}_j$ satisfy the following inequality for any $k \leqslant N$:*

$$\mathfrak{R}(\mathcal{F}_j) = \mathbb{E}_X \mathbb{E}_\sigma \sup_{f \in \mathcal{F}_j} \frac{1}{k} \sum_{o=1}^{k} \sigma_o f(x_o) \leqslant \sqrt{\frac{R_j(k)}{k}}, \tag{311}$$

*where $R_j(k)$ is an increasing function of $k$ for all $j \leqslant m$. We have the following bound on the Rademacher complexity of the function class $\mathcal{G}$:*

$$\mathfrak{R}(\mathcal{G}) = \mathbb{E}_X \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{k} \sum_{o=1}^{k} \sigma_o g(x_o) \leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sqrt{\frac{2\sum_{j \leqslant m} R_j(N)}{N}}.$$

*Proof.* By Proposition G.3 for any sample $x_1, \ldots, x_N$ we have

$$\widehat{\mathfrak{R}}(\mathcal{G}) \leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sum_{j \in J_1} \frac{\lambda_j}{\ell} \max_{k \geqslant \frac{N\lambda_j}{2\ell}} \widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j). \tag{312}$$

Taking expectations with respect to the sample $x_1, \ldots, x_N$ on both sides we obtain:

$$\mathfrak{R}(\mathcal{G}) \leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sum_{j \in J_1} \frac{\lambda_j}{\ell} \max_{k \geqslant \frac{N\lambda_j}{2\ell}} \mathbb{E}_X \widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j). \tag{313}$$

Since the distribution of a uniformly random subset of size $k$ of $\{x_1, \ldots, x_N\}$ where the $x_o$s are drawn i.i.d. from $X$ is distributed as $k$ i.i.d. samples from $X$, we have

$$\mathbb{E}_X \widehat{\mathfrak{R}}_{N,k}(\mathcal{F}_j) = \mathbb{E}_X \widehat{\mathfrak{R}}_k(\mathcal{F}_j) \leqslant \sqrt{\frac{R_j(k)}{k}} \leqslant \sqrt{\frac{R_j(N)}{k}}. \tag{314}$$

Plugging this back into Equation (313) we obtain:

$$\mathfrak{R}(\mathcal{G}) \leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sum_{j \in J_1} \frac{\lambda_j}{\ell} \max_{k \geqslant \frac{N\lambda_j}{2\ell}} \sqrt{\frac{R_j(N)}{k}} \tag{315}$$

$$\leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sum_{j \in J_1} \frac{\lambda_j}{\ell} \sqrt{\frac{2R_j(N)\ell}{N\lambda_j}} \tag{316}$$

$$\leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sum_{j \leqslant m} \sqrt{\frac{\lambda_j}{\ell}} \sqrt{\frac{2R_j(N)}{N}} \tag{317}$$

$$\leqslant \mathcal{B}\,\ell \frac{13m \log(2mN)}{N} + \ell \sqrt{\frac{2\sum_{j \leqslant m} R_j(N)}{N}}, \tag{318}$$

where at Line (317) we have used the fact that $\gamma := \frac{8\ell \log(2mN)}{N}$ (cf. Equation (306) with $\delta = 1/N$, or the end of the statement of Proposition G.3) and at the last line (318), we have used the Cauchy-Schwarz inequality. □

## G.2. Extension: an Example Generalization and Excess Risk Bound for a Composite Model with Multiple Latent Matrices

In this subsection, we prove an additional result for a model which takes several latent matrices as input. In this preliminary version, we allow a factor of the embedding size of the neural network embeddings in the final bounds and reserve removing this factor to future work. We consider the following function class:

$$\mathcal{H} := \mathcal{N}_{2,W}(a', s') \circ (\mathrm{concat}_{v=1}^{\mathrm{m}}(\widetilde{\mathcal{F}}_{r_v, \mathcal{B}_0}^{p_v}), \mathcal{N}_{1,W,\mathrm{id}}(a, s)) \tag{319}$$

where the input of the network $\Psi \in \mathcal{N}_{1,W,\mathrm{id}}(a,s)$ is a concatenation of a user ID and an item ID: $\mathcal{N}_{1,W,\mathrm{id}}$ being the class of matrices in $\mathbb{R}^{m\times n}$ which can be represented as $\tilde{\phi}(x_\xi)$ where $\tilde{\phi}$ is a network form (259) satisfying Conditions (265) and $x_{i,j} := \mathrm{concat}(e_i, e_j)$. In this subsection, we assume that $s_\ell, a_\ell, s'_\ell, a'_\ell, \mathcal{B}_0, \mathcal{B} \geqslant 1$.

**Theorem G.5.** *Define* $\hat{g} \in \arg\min\left(\widehat{\mathbb{E}}(\mathrm{l}(g_\xi, \widetilde{G}, \xi)) : g \in \mathcal{H}\right)$ *and* $g^* \in \arg\min\left(\mathbb{E}(\mathrm{l}(g_\xi, \widetilde{G}, \xi)) : g \in \mathcal{H}\right)$. *Define* $\bar{r} = \sum_{v=1}^{\underline{\mathrm{m}}} r_v$ *and assume that* $s_\ell, a_\ell, s'_\ell, a'_\ell, \mathcal{B}_0, \mathcal{B} \geqslant 1$. *With probability greater than* $1-\delta$ *over the draw of the training set we have*

$$\widehat{\mathfrak{R}}(\mathcal{H}), \sup_{g\in\mathcal{H}} \mathbb{E}(\mathrm{l}(g, \widetilde{G}, \xi)) - \widehat{\mathbb{E}}(\mathrm{l}(g, \widetilde{G}, \xi)), \mathbb{E}(\mathrm{l}(\hat{g}, \widetilde{G}, \xi)) - \mathbb{E}(\mathrm{l}(g^*, \widetilde{G}, \xi)) \leqslant \bar{\mathcal{R}}, \tag{320}$$

*where*

$$\bar{\mathcal{R}} = \widetilde{O}\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B}\sqrt{\frac{W+W'}{N}} + \mathcal{B}_0\,\mathcal{S}'\,\ell\sqrt{\frac{\underline{\mathrm{m}}^2\,\bar{r}(m+n)}{N}} + \mathcal{B}_0\,\mathcal{S}'\,\ell\sqrt{\frac{\underline{\mathrm{m}}^3}{N}}\right) \tag{321}$$

*where* $\mathcal{S}' = \prod_{\ell=1}^{L} s'_\ell$, *and the* $\widetilde{O}$ *notation hides polylogarithmic factors of* $m, n, N, \underline{\mathrm{m}}, \mathcal{B}_0, \mathcal{B}, \ell, \prod_\ell s_\ell, \prod_\ell s'_\ell, \prod_\ell a_\ell, \prod_\ell a'_\ell$ *(in particular, the depth* $L$ *is considered constant).*

**Remark:** Note that the dependence on $\underline{\mathrm{m}}$ is very strong. In particular, the last term alone contributes a sample complexity of $\widetilde{O}(\underline{\mathrm{m}}^3)$. This is due to the need to bound the Lipschitz constant of the network $\Psi$ in each dimension individually, resulting in an additional factor of $\underline{\mathrm{m}}$ outside the square root in the last two terms. We leave the delicate question of mitigating this dependence, perhaps via an improved version of Corollary G.4, to future work.

*Proof.* Similarly to the proof of Theorem C.8, we use Lemma E.3 to join the Rademacher complexities of $\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}$ for all values of $v$ with the covering numbers of the classes $\mathcal{N}_{1,W,\mathrm{id}}(a,s)$ and $\mathcal{N}_{2,W}(a',s')$.

To that aim, we begin by using Proposition F.9 with $\mathrm{l} = Id$, $\ell = 1$ and $\chi = \prod_{\ell=1}^{L} s_\ell + \underline{\mathrm{m}}\,\mathcal{B}_0$, and $\epsilon \leftarrow \frac{\epsilon}{2}$ to obtain a cover of $\mathcal{C}_1 \subset \mathcal{N}_{2,W}(a',s')$ such that for any $\phi \in \mathcal{N}_{2,W}$ there exists a $\bar{\phi} \in \mathcal{C}_1$ such that for any $y \in \mathbb{R}^1$ with $\|y\| \leqslant \chi$ we have

$$|(\phi - \bar{\phi})(y)| \leqslant \frac{\epsilon}{2} \tag{322}$$

and

$$\log(|\mathcal{C}_1|) \leqslant W' \log\left(\frac{12\left[\prod_{\ell=1}^{L} s_\ell + \mathcal{B}_0\,\underline{\mathrm{m}}\right]\left[\prod_{\ell=1}^{L} s'_\ell\right]\left[\sum_\ell a'_\ell\right]}{\epsilon} + 1\right) \tag{323}$$

$$\leqslant W \log(\Gamma_{W,\underline{\mathrm{m}}}), \tag{324}$$

after setting $\epsilon = 1/N$ and $\Gamma_{W,\underline{\mathrm{m}}} := 12N\left[\prod_{\ell=1}^{L} s_\ell + \underline{\mathrm{m}}\,\mathcal{B}_0\right]\left[\prod_{\ell=1}^{L} s'_\ell\right]\left[\prod_{\ell=1}^{L} s_\ell\right]\left[\sum_\ell a'_\ell\right]\left[\sum_\ell a_\ell\right] + 1$. Next we can invoke Proposition F.9 again to construct a cover $\mathcal{C}_2$ of $\mathcal{N}_{2,W}(a,s)$ such that for any $\Psi \in \mathcal{N}_{2,W}(a,s)$ there exists a $\bar{\Psi} \in \mathcal{C}_2$ such that

$$\left|\Psi(\xi) - \bar{\Psi}(\xi)\right| \leqslant \frac{\epsilon}{2\left[\prod_{\ell=1}^{L} s'_\ell\right]} \tag{325}$$

and

$$\log(|\mathcal{C}_2|) \leqslant W' \log\left(\frac{12\prod_{\ell=1}^{L} s_\ell\left[\prod_{\ell=1}^{L} s'_\ell\right]\left[\sum_\ell a_\ell\right]}{\epsilon} + 1\right) \leqslant W' \log(\Gamma_{W,\underline{\mathrm{m}}}), \tag{326}$$

where we set $\epsilon = \frac{1}{N}$. Note that for any fixed set of matrices $Z_v \in \widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}$ ($v \leqslant \underline{\mathrm{m}}$), we have

$$\left|\phi(Z_1,\ldots,Z_{\underline{\mathrm{m}}},\Psi) - \bar{\phi}(Z_1,\ldots,Z_{\underline{\mathrm{m}}},\bar{\Psi})\right| \tag{327}$$

$$\leqslant \left|\phi(Z_1,\ldots,Z_{\underline{\mathrm{m}}},\Psi) - \bar{\phi}(Z_1,\ldots,Z_{\underline{\mathrm{m}}},\Psi)\right| + \left|\bar{\phi}(Z_1,\ldots,Z_{\underline{\mathrm{m}}},\Psi) - \bar{\phi}(Z_1,\ldots,Z_{\underline{\mathrm{m}}},\bar{\Psi})\right|$$

$$\leqslant \epsilon \tag{328}$$

where at the last line we have used Equations (322) and (325).

Thus, we are in a position to apply Lemma E.3 with $\epsilon = \frac{1}{N}$ to obtain:

$$\widehat{\mathfrak{R}}(\mathcal{H}) \leqslant \frac{1}{N} + \sup_{\bar{\phi},\bar{\Psi}} \widehat{\mathfrak{R}}\left(\phi \circ (\text{concat}(\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}), \Psi)\right) + \mathcal{B}\sqrt{\frac{2\pi}{N}} + \mathcal{B}\sqrt{\frac{\log(|\mathcal{C}_1(1/N)||\mathcal{C}_2(1/N)|)}{N}}. \tag{329}$$

Note that by Theorem D.4, and Lemma F.6, we have for any $N' \geqslant 9(m+n)$,

$$\mathbb{E}_\xi \widehat{\mathfrak{R}}_{N'}(\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}) \leqslant \sqrt{\frac{7\,{\mathcal{B}_0}^2+1}{N'}} + 88\,\mathcal{B}_0\sqrt{\frac{r_v(m+n)}{N'}}\log(\underline{\Gamma}), \tag{330}$$

where $\underline{\Gamma} =: 6Nmn^3[\mathcal{B}_0+1]+1$, and thus for any $N'$,

$$R_v \leqslant 2\times 88^2\log^2(\underline{\Gamma})\left[{\mathcal{B}_0}^2\, r_v(m+n) + ({\mathcal{B}_0}^2+1)\right]. \tag{331}$$

Thus by Corollary G.4 we have, for any $\ell\left[\prod_{\ell=1}^L s'_\ell\right]$-Lipschitz function $G:\mathbb{R}^{\underline{\mathrm{m}}}\to\mathbb{R}$:

$$\begin{aligned}
&\mathfrak{R}(G(\text{concat}^{\underline{\mathrm{m}}}_{v=1}(\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}))) \leqslant \mathcal{B}_0\,\ell\left[\prod_{\ell=1}^L s'_\ell\right]\frac{13\,\underline{\mathrm{m}}^2\log(2\,\underline{\mathrm{m}}\,N)}{N} + \ell\left[\prod_{\ell=1}^L s'_\ell\right]\sqrt{\frac{2\,\underline{\mathrm{m}}^2\sum_{v\leqslant\underline{\mathrm{m}}}R_v(N)}{N}}\\
&\leqslant \mathcal{B}_0\,\ell\left[\prod_{\ell=1}^L s'_\ell\right]\frac{13\,\underline{\mathrm{m}}^2\log(2\,\underline{\mathrm{m}}\,N)}{N} + 176\log(\underline{\Gamma})\,\ell\left[\prod_{\ell=1}^L s'_\ell\right]\sqrt{\frac{\underline{\mathrm{m}}^2\,{\mathcal{B}_0}^2\,\bar{r}(m+n)+\underline{\mathrm{m}}^3({\mathcal{B}_0}^2+1)}{N}}\\
&=: \mathcal{R},
\end{aligned} \tag{332}$$

and therefore, for any $\delta$, with probability greater than $1-\delta$ we have

$$\widehat{\mathfrak{R}}_N(G(\text{concat}^{\underline{\mathrm{m}}}_{v=1}(\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}))) \leqslant \mathcal{R} + O\left(\mathcal{B}\sqrt{\frac{\log(2/\delta)}{N}}\right). \tag{333}$$

By a union bound, inequality (334) below holds with probability $\geqslant 1-\delta$ over all the choices of $G$ given by $G(x)=\bar{\phi}(x,\bar{\Psi})$ for $\bar{\phi}\in\mathcal{C}_2$ and $\bar{\Psi}\in\mathcal{C}_1$:

$$\widehat{\mathfrak{R}}_N(G(\text{concat}^{\underline{\mathrm{m}}}_{v=1}(\widetilde{\mathcal{F}}^{p_v}_{r_v,\mathcal{B}_0}))) \leqslant \mathcal{R} + O\left(\mathcal{B}\sqrt{\frac{\log(2/\delta)}{N}}\right) + 2\,\mathcal{B}\sqrt{\frac{\log(|\mathcal{C}_1||\mathcal{C}_2|)}{N}}. \tag{334}$$

Plugging this back into Equation (329) (after setting $\epsilon=\frac{1}{N}$) we obtain (w.p. $\geqslant 1-\delta$)

$$\widehat{\mathfrak{R}}(\mathcal{H}) \leqslant \frac{1}{N} + \mathcal{R} + 2\,\mathcal{B}\sqrt{\frac{\log(2/\delta)}{N}} + \mathcal{B}\sqrt{\frac{2\pi}{N}} + O\left(\mathcal{B}\sqrt{\frac{\log(|\mathcal{C}_1||\mathcal{C}_2|)}{N}}\right) \tag{335}$$

$$\leqslant O\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right) + O\left(\mathcal{B}\sqrt{\frac{(W+W')\log\left(\Gamma_{W,\underline{\mathrm{m}}}\right)}{N}}\right) + \mathcal{R} =: \bar{\mathcal{R}}, \tag{336}$$

as expected. □

## H. Additional Experimental Details

To assess the proposed methodology in this paper and compare it with related matrix completion approaches, we conducted experiments using both synthetic and real-world datasets. On the one hand, the synthetic experiments were designed to evaluate the performance of the methods and related mechanisms by varying proportions of observed entries in the incomplete matrix targeted for completion. On the other hand, the real-world datasets were employed in practical scenarios to gain insights into the methods' behavior in real-world applications of matrix completion.

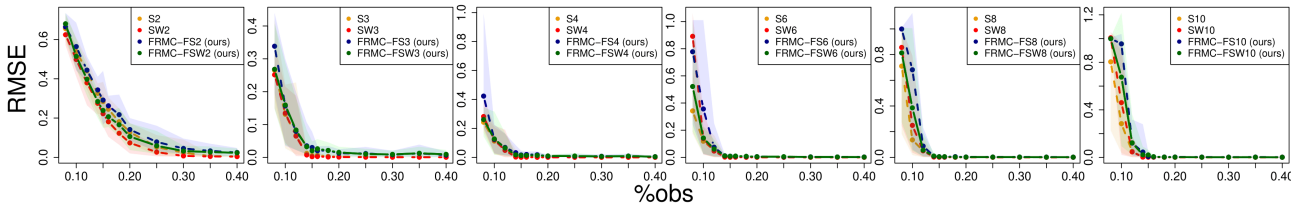


Figure 3: Summary of the results from the synthetic data experiments with ground truth was generated by considering $f(x) = x$.

### H.1. Experiments on Synthetic Data

As described in the main paper, we generate synthetic square data matrices in $\mathbb{R}^{n\times n}$ with a given rank $r$. For each matrix, we vary the proportion of observed entries ($\%$obs=$\mathbb{E}[N/n^2]$) as well as performed non-uniform sampling distribution. Regarding the proportion of observed entries $\%$obs, we explore values in the set $\{0.08, 0.10, \ldots, 0.20, 0.25, \ldots, 0.40\}$. Concerning the sampling distribution, we divide the matrix into four equal-sized regions of size $n/2 \times n/2$. In the first region, the probability of entry sampling equals $\alpha$. In regions 2 and 3, it is $2\alpha$, and in region 4, it is $4\alpha$. For a given function $f$ or generation procedure, it is described as follows:

1. Randomly generate matrices $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{n\times r}$ with each entry $(i, j)$ sampled from a normal distribution $\mathcal{N}(0, 1)$.
2. Make $\tilde{R} = UV^\top$ and rescale the product as $\bar{R} = n \times \tilde{R}/\|\tilde{R}\|_{\mathrm{Fr}}$.
3. Apply the function $f$ element-wise to $\tilde{R}$ and obtain $R$. Return the ground truth generation as $R$.

We generated 25 independent instances by considering the aforementioned generation procedure. For each matrix, we varied the observations accordingly.

**Remark:** We observe that the number of degrees of freedom for an $n \times n$ matrix of rank $r$ is $nr$, which (up to logarithmic terms) is loosely connected to the sample complexity. Consequently, the proportion of observed entries necessary for the prediction task to be (statistically) feasible is linked to the choice of $n$ and $r$. In our synthetic experiments, we opt for smaller matrices due to the high number of simulations. Therefore, we set $n = 100$ and $r = 3$ in line with the aforementioned observation strategy. As a choice for $f$, we considered the identity function $g(x) = x$ and the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$.

**Further Results:** Figures 3 provide detailed results of our synthetic data experiments when the ground truth function is the identity. In this case traditional matrix completion methods perform similarly to ours: the unneeded additional representative power doesn't hurt the performance, since it is small enough to come with negligible additional sample complexity as per Thm C.6 and C.6.

**Validation, Optimization and Hardware Specifications:** For all synthetic data experiments, we employed a $\%$obs of the data for training (specified in each case), allocating $10\%$ for validation, and utilizing the remaining portion as the test set. In the validation procedure, we selected the regularization parameter from the set $\lambda \in \{10^{-7}, 10^{-6}, \ldots, 10^{2}\}$ and fixed the size of the embeddings to 20. We optimize all models with ADAM using Nesterov optimization through TensorFlow 2. We consider a maximum of 100 epochs with early stopping and a patience of 5 in the validation loss, returning the best weights. Regarding hardware specifications, all synthetic experiments were executed on a CPU cluster with 128 threads and 512GB of RAM.

### H.2. Experiments on Real-world Data

**Validation, Optimization and Hardware Specifications:** Similar to the synthetic data experiments, we optimized the models using ADAM with Nesterov optimization in TensorFlow 2. We set a maximum of 100 epochs with early stopping, employing a patience of 15 based on the validation loss for all models. The best weights were selected during training. Real-world experiments were conducted on Nvidia DGX-A100 graphics cards with 40GB of GPU RAM.

Regarding our validation procedure, we randomly split the observed entries uniformly, resulting in $90\%$ for training and $5\%$ for each validation and test set. The parameter $\lambda$ was selected from a sequence exponentially distributed between $10^{-7}$ and $10^2$. For DOUBAN and LASTFM, this sequence has a size of 50, and for MVL25, it has a size of 15. For all datasets and methods, we fixed the embeddings to have a size of 128.

### H.3. Datasets

**DOUBAN** [$m = 2718, n = 34893$ and $\%$obs$= 1.2\%$]: Douban serves as a platform for users to curate movies. Within this matrix, users are interconnected within the social network, and the items represent movies. User ratings, ranging from $0.5$ to $5$ (in intervals of $0.5$), are denoted by the entry $(i, j)$, corresponding to the rating of user $i$ for movie $j$.

**LASTFM** [$m = 1892, n = 17632$ and $\%$obs$= 0.27\%$]:: Last.fm, profiles users' musical preferences and habits. In contrast to other datasets, entries $(i, j)$ in this matrix signify the log-scaled number of views user $i$ has for band/artist $j$.

**MVL25** [$m = 162541, n = 57971$ and $\%$obs$= 0.27\%$]: The MovieLens 25M dataset, a widely adopted and stable benchmark dataset, originates from a non-commercial movie recommendation website. Similar to Douban, entries $(i, j)$ here indicate the rating of user $i$ for movie $j$, but on a scale from 1 to 5.

## I. More Detailed Related Works

**Note:** in Matrix Completion, by "*approximate recovery*", we mean results which bound the *excess risk* in the form of a function of architectural parameters and the number of samples, with decay rate typically of the form $1/\sqrt{N}$ (but sometimes $1/N$, or $1/\sqrt[4]{N}$ if expressing the bound in terms of the Frobenius norm error rather than the square loss). For instance, in the realisable case, if the noise is independent of the entries and has standard deviation $\varepsilon$ and the loss function is the square loss, this means that the normalized Frobenius norm of the error scales like $\varepsilon^2 + \sqrt{\frac{R}{N}}$ where $R$ is some architectural quantity. By "*exact recovery*", we mean results which guarantee that the ground truth matrix is recovered exactly with high probability when the number of samples $N$ is large enough as long as there is no noise in the observations. By "*perturbed recovery*", we mean results which guarantee that for large enough $N$, an error of the type $\varepsilon\sqrt{\frac{\bar{R}}{N}}$ is achievable for with high probability for some other architectural quantity $\bar{R}$. The quantity $\bar{R}$ typically has much worse dependence on architectural parameters than the quantity $R$, and as long as that is the case, approximate recovery and perturbed/exact recovery are not subordinate to each other, even if we ignore the minor difference in the optimization problem and sampling regime. To the best of our knowledge: the only existing result which achieves the extremely impressive task of providing a *perturbed recovery* result where the architectural dependence of $\bar{R}$ is as tight as that of $R$ in competing approximate recovery results is (Chen et al., 2020), which only deals with the nuclear norm ($p = 1$) and does not include non-linearities. In addition, like all exact and perturbed recovery results we are aware of, the results in (Chen et al., 2020) are limited to the uniform sampling case. Our results concern *approximate recovery* with the Schatten (quasi) norm, but we still compare to some exact and noisy recovery results for illustrative purposes.

**Approximate Recovery in Matrix Completion:** There is a lot of literature on the sample complexity of matrix completion with bounded Lipschitz losses and norm constraints. In particular, our work takes much inspiration from the pioneering works of (Foygel et al., 2011) and (Shamir & Shalev-Shwartz, 2011; 2014), which proved analogues of our results (without a learnable function) in the case $p = 1$. The explicitly rank-restricted case was studied in classification settings in (Srebro et al., 2004; Srebro & Shraibman, 2005; Srebro & Jaakkola, 2005). In general, the sample complexity is $\widetilde{O}(rn)$.

**Alternative Learning Settings and Models:** There is also a substantial amount of work on other soft relaxations of the rank, such as the max norm. In particular, the early work of (Srebro & Shraibman, 2005) shows a sample complexity of $\widetilde{O}(nM^2)$, where $M$ is a constraint on the max norm. A perturbed recovery result was achieved for the max norm in the classic work of (Cai & Zhou, 2016), which was further extended in (Wang et al., 2021) to provide bounds on the *uniformly weighted* Frobenius error of the recovered matrix in the *non-uniform sampling* regime (under some approximate uniformity assumption on the sampling probabilities). With nuclear norm regularizers, other works which provide uniform Frobenius error bounds without uniform sampling include the "missing not at random" setting (Ma & Chen, 2019; Sportisse et al., 2020), which adopts a Bayesian approach. Further, the pioneering work of (Gui et al., 2023) computes entry-wise confidence intervals in low-rank matrix completion with an arbirary backbone model, substantially extending the entry-wise guarantees provided in the known rank case in (Chen et al., 2021).

**Exact Recovery in Matrix Completion:** The problem of exactly recovering the entries of a uniformly sampled matrix

can be traced back to the work of Tao (Candès & Tao, 2010) and has been widely studied since (Recht, 2011; Gross, 2011; Candès & Recht, 2009; Chen, 2013): by minimizing the nuclear norm, the matrix can be recovered with high probability after sampling $\widetilde{O}(nr)$ entries where $r$ is the rank of the ground truth matrix.

**Perturbed Recovery in Matrix Completion:** The first bounds with a refined dependence on the variance of the noise can be traced back to the early work of (Candes & Plan, 2010), which roughly speaking shows an excess risk bound of order $\widetilde{O}([1+\sqrt{\frac{n^3}{N}}]\sigma)$ where $\sigma$ is the standard deviation of the perturbation, sampling is without replacement and $n \gg \widetilde{O}(nr)$. Thus, the architectural dependence on the matrix size $n$ is very strong inside the term which involves the variance parameter $\sigma$. Much later, a nearly optimal bound of $\widetilde{O}(\sigma\sqrt{\frac{nr}{N}})$) (also for sampling without replacement) was achieved in (Chen et al., 2020).

**Inductive Matrix Completion:** Inductive Matrix Completion studies predictors of the form $AD_1D_2B^\top$ where $A$ and $B$ are **fixed** matrices which collect "side information" about the rows and columns. Thus, this can be viewed as an analogue of deep matrix factorization with $d=4$ with $A$ and $B$ fixed. However, since $A, B$ are fixed, the problem behaves more similarly to matrix completion with nuclear norm constraints. To the best of our knowledge, the first bounds for this model in the approximate recovery setting are from (Chiang et al., 2018; 2015), giving bounds of order $\mathcal{M}\sqrt{\frac{1}{N}}$ where $\mathcal{M}$ is a bound on the nuclear norm of $D_1D_2$. Expressed in terms of rank-like quantities, this yields $\widetilde{O}(\sqrt{\frac{ra^2}{N}})$ where $a$ is the number of columns of $A$ and $B$. Later, stronger results were provided in (Ledent et al., 2021b) which match the non-inductive literature with $a$ playing the role of $n$ in standard MC. For instance, the distribution-free sample complexity rate is $\widetilde{O}(a^{\frac{3}{2}}\sqrt{r})$. For exact recovery, a sample complexity rate of $\widetilde{O}(ar)$ was provided in (Xu et al., 2013). Later, (Ledent et al., 2023) provided a perturbed recovery bound of $\widetilde{O}\left(\sigma\sqrt{\frac{a^4}{N}}\right)$. Furthermore, several works study more specific settings where the rows ans columns have implicit cluster structure (Qiaosheng et al., 2019; Zhang et al., 2022; Ledent et al., 2021a; Alves et al., 2021). Such assumptions are also becoming common in the field of low rank bandits (Pal & Jain, 2022; Pal et al., 2023). However, none of these works consider the situation where the matrices $A$ and $B$ are trainable (which corresponds to the case $d=4$ in our setting).

**Orthogonal Tensor Recovery with the Schatten Quasi-Norm** Beyond the examples above, we are not aware of any work on the approximate recovery for Schatten norm constrained matrix completion. However, similar problems have been studied with different losses or sampling regimes. In particular, (Fan et al., 2020; Fan, 2021) studies approximate *tensor* recovery with **Schatten regularization**. The results are far reaching and go well beyond the more restricted setting of *matrix* completion which we study here. However, in the case of a 2-way tensor (i.e. a matrix), the results can be interpreted as a Lagrangian formulation of the empirical risk minimization problems we study. The loss function is the square loss and sampling is uniformly at random without replacement, which means the results are not directly comparable. The achieved excess Frobenius norm bounds scale like $\sqrt[4]{\frac{n^{\frac{2-2p}{2-p}}\mathcal{M}^{\frac{2p}{2-p}}}{Np}}$(cf. (Fan, 2021), Theorem 4), where $\mathcal{M}$ is an upper bound on the $\|\cdot\|_{\mathrm{sc},p}$ [3] Expressed in terms of our rank-like quantity $r$, this turns into $\sqrt[4]{\frac{rn^{\frac{2}{2-p}}}{Np}}$. In contrast, our result is $\widetilde{O}\left(\mathcal{B}^{\frac{2-2p}{2-p}}\ell^{\frac{p}{2-p}}\sqrt{\frac{\mathcal{M}^{\frac{2p}{2-p}}n^{\frac{2-3p}{2-p}}}{Np}}\right)$, which translates to $\widetilde{O}\left(\sqrt{\frac{rn}{Np}}\right)$. Firstly, note both results scale like $\widetilde{O}(rn)$ when $p\to 0$ (though the constant blows up like $1/p$ in both cases). Secondly, our rate is uniformly tighter since $\frac{2}{2-p}>1$. And lastly, the bound in (Fan, 2021) is vacuous for $p=1$, scaling like $\widetilde{O}(rn^2)$ in that case, compared to $\widetilde{O}(rn)$ in our result.

**Matrix sensing with Schatten Quasi-Norm:** While exact and perturbed recovery for matrix completion with the nuclear norm (and inductive matrix completion) is a very well-studied problem, for $p<1$, there appears to be little to no existing work in the case of randomly sampled *entries*. However, there is a lot of work on the sample complexity of *compressed sensing* for matrix completion, including with Schatten norm minimization (Zhang et al., 2013; Arora et al., 2019; Liu et al., 2014; Recht et al., 2010). In compressed sensing, instead of observing *entries*, we observe *measurements* in the form of Frobenius inner products of the ground truth with certain matrices Although MC can be expressed in the language of compressed sensing by saying that the measurement matrices are indicator function of entries (and inductive matrix completion can be expressed by saying that the measurement matrices are all the possible outer products of row and column

[3] We express our bounds in terms of excess risk with a bounded loss (which could be the truncated square loss), so the decay rate in $N$ can be understood as comparable: the main difference lies in the architectural sample complexity.

side information vectors), it is not easy to deduce even existing results for matrix completion or IMC from their compressed sensing analogues: indeed, the conditions on the measurement matrices are typically expressed deterministically via the Restricted Isometry Property, which cannot be easily checked for indicator measurement matrices, though it holds with high probability for certain classes of measurement matrices. For instance, (Zhang et al., 2013; Liu et al., 2014) show a sample complexity of $nr$ for perturbed recovery with the Schatten norm for a broad class of measurement matrices called "nearly isometric families" (cf. (Recht et al., 2010)), which includes measurement matrices with i.i.d. Gaussian entries but not indicator measurements: in that case, Property 4.3 from (Recht et al., 2010) only holds for uniform sampling, and property 4.1 only holds for bounded $X$, which violates the definition (which requires the property to be satisfied for all $X$), though the fact it does hold for bounded $X$ may offer insights on the relationship between the proof techniques. It is clear from the uniform sampling complexity of $\widetilde{O}(nr)$ that this setting, although much more general in many ways, cannot capture the detailed effects of the sampling distribution on the function class capacity of matrices with constrained norms offered by (Shamir & Shalev-Shwartz, 2011; 2014; Ledent et al., 2021b) and the present work.

Earlier works on **deep matrix factorization** often focus on the optimization and algorithmic aspects (Trigeorgis et al., 2016; Zhao et al., 2017) without providing sample complexity bounds, though some include non-linear components (Xue et al., 2017; Fan & Cheng, 2018; Wang et al., 2017; De Handschutter et al., 2021; Wei et al., 2020; Lara-Cabrera et al., 2020). Note also that the non-linear components in those works are interpsersed between each matrix in the product (by analogy with the activation functions in feedforward neural netowrks), rather than entry-wise and after the matrix multiplication (as in FRMC), which implies the models are also different.

The observation that deep matrix factorization is equivalent to Schatten norm regularization was made in other works, including (Arora et al., 2019), which studies the optimization landscape of the problem in a compressed sensing setting where the measurement matrices commute (which does not apply to indicator measurements). The implications this has on the implicit rank-restriction in which occurs when training deep neural networks is currently the subject of a large amount of interest in the community (Dai et al., 2021; Jacot, 2022; Wang & Jacot, 2023). However, those works typically do not study sample complexity, perhaps because it is only non trivial when the matrix is not flat, which implies a multi-*output* scenario in the neural network context. Nevertheless, the potential to generalize our results to that situation is a tantalizing direction for future work which may shed a different light on implicit rank-restriction in DNN training.

## J. Future Directions

There are plenty of unanswered questions which can be studied in future work. For instance:

1. Can the strong dependence on $\underline{\mathrm{m}}$ in the results in Section G be improved through a more refined handling of the L1 Lipschtiz constant in Proposition G.3?

2. Our results concern matrix completion. However, the equivalence between Schatten quasi-norm regularization and L2 regularization of factor matrices is valid in the case of neural networks as well: in fact, there is a large amount of renewed enthusiasm for this problem in the community in recent years from the optimization perspective (Dai et al., 2021; Wang & Jacot, 2023; Giampouras et al., 2020; Arora et al., 2019). Do our results extend to this case? A simple question is how the sample complexity of linear networks of the form

$$\mathbb{R}^m \ni f(x) = A^L \dots A^1 x \qquad (x \in \mathbb{R}^n) \tag{337}$$

   behaves similarly to our bounds where the quantity $\mathcal{M}$ would be replaced by an upper bound on $\sum \|A^\ell\|_{\mathrm{Fr}}^2$. The two problems are still technically distinct, and adaptations of the techniques would be necessary. The question can also be extended to non zero reference matrices, which appears to be a highly non trivial problem. more generally, the relationship between our results and those of (Dai et al., 2021) could be investigated further in this context.

3. Perhaps a unifying question regarding both points above is whether the results of Section C.2 can be obtained through a covering number approach.

4. Can our chaining and Talagrand type arguments in Lemmas E.4 and E.3, as well as proposition G.3 be used to improve existing generalization bounds for neural networks (with activations), at least by removing certain logarithmic terms?

5. Do our results extend unbounded losses?

6. What can be said in the transductive case? Since has been studied in the case of nuclear norm regularization before (Shamir & Shalev-Shwartz, 2011), it is not unreasonable to assume that similar results could hold for our setting.

7. Our results concern excess risk bounds which correspond to traditional performance measures (e.g. RMSE). However, Recommendation Systems typically rely on measures more sensitive to higher predictions than lower ones (e.g. recall, NDCG). Can generalization bounds be proved in those settings?

8. In recommendation systems settings, do our results extend to Graph neural networks such as LightGCN (He et al., 2020)?